AUTHOR INDEX

Author	Session	Title
A, Jayakrishnan	C1.1	*MSDF-Net: A Multi-Scale Deep Fusion Network with Dilated Convolutions for Cloud Removal from Sentinel-2 Imagery*
Abdulla, Waleed	A4.1	*ResNet-Based Camera Model Identification with Adaptive Preprocessing Module and Weight Fusion of Global Information*
	I5.1	*Urban Noise Monitoring Using Edge Computing with CNN-LSTM on Jetson Nano*
	P5.16	*Enhancing Retinal Disease Classification with Dual Scale Twin Vision Transformers Using OCT Imaging*
Abe, Masanobu	F3.3	*Speech Synthesis Using Ambiguous Inputs from Wearable Keyboards*
	P5.3	*Speech-Emotion Control for Text-To-Speech in Spoken Dialogue Systems Using Voice Conversion and x-Vector Embedding*
Abe, Narishige	I4.3	*Privacy-Preserving Image Transformation Method for Person Detection and Re-ID*
Abe, Takuya	A1.5	*Detection Model of Sister Chromatid Cohesion Defects Based on Vision Transformer*
Abed-Meraim, Karim	J1.2	*On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems*
Abukawa, Shoya	E5.4	*Approximate Modeling of Malware Diffusion on Wireless Mobile Devices*
Adapa, Eswar	P1.1	*Fatigue Classification and Onset Estimation Using Surface EMG Signals During Strength Training*
Adiono, Trio	I1.3	*DQN Algorithm Design for Fast Efficient Shortest Path System*
Adiono, Trio	I1.2	*Human Activity Recognition Based on FMCW Radar Using CNN and Transfer Learning*
	I1.1	*A Comparative Evaluation of Video Codecs for rPPG-Based Heart Rate Estimation*
	I1.5	*Signal Quality Assessment for Wearable Multichannel Photoplethysmography Signals*
Adiono, Trio	I1.4	*Comparison of MPPT Based on Deep Reinforcement Learning by DQN, DDPG and TD3*
Adiono, Trio	I3.2	*Algorithm Development for Step-Wise Valve Deflation Method in Blood Pressure Measurement*
	I3.4	*Sparsity Exploration for Structured and Unstructured Weight Formations in CNN Architecture*

Adiono, Trio	I3.1	*A QoS Throughput Performance Measurement Comparison Between UGS and BE Services of a Real-Time FPGA Based OFDM Multi-User System Design Implementation*
	I3.3	*SUMO Based Hardware/Software Co-Simulation for Two-Intersection Adaptive and Collaborative Traffic Signal Controller*
Aditya, Ghanes	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Agrawal, Dharmeshkumar	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Ahmadi, Nur	I1.2	*Human Activity Recognition Based on FMCW Radar Using CNN and Transfer Learning*
	I1.1	*A Comparative Evaluation of Video Codecs for rPPG-Based Heart Rate Estimation*
	I1.5	*Signal Quality Assessment for Wearable Multichannel Photoplethysmography Signals*
Ahuja, Mitali	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Aizawa, Jumpei	A1.3	*Manipulation of Neuronal Network Firing Patterns Using Temporal Deep Unfolding-Based MPC*
Akai, Reon	C2.4	*A Study on Eliminating Biased Node in Federated Learning*
Akhter, MD Tousin	P1.8	*FID-RPRGAN-VC: Fréchet Inception Distance Loss Based Region-Wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion*
Akhyar, Fityanul	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Albu, Felix	P5.13	*Proportionate NLMS with Variable Step-Size for Adaptive Feedback Cancellation in Hearing Aids*
Amadeus, Clarence	I3.2	*Algorithm Development for Step-Wise Valve Deflation Method in Blood Pressure Measurement*
An, Wenjuan	C4.5	*Feature Augmentation Reconstruction Network for Few-Shot Image Classification*
An, Yuting	P1.26	*A Visually Meaningful Image Encryption Algorithm with Attention Mechanism and Artificial Bee Colony Optimization*
Ang, Yi Yang	P3.6	*Heart Rate Acquisition and Processing Techniques for a Miniature Wearable Microphone Sensor*
Anggraini, Irin	E4.2	*OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises*
Anh, Anh	A3.4	*An Automatic Pipeline for Building Emotional Speech Dataset*

Anistyasari, Yeni	B5.1	*ASF-LLRDA: Locality-Regularized Linear Regression Discriminant Analysis with Approximately Symmetrical Face Preprocessing for Face Recognition*
Aoki, Remi	P3.1	*Remarks on an Optimal Predictive Control Using a Quaternion Neural Network and a Metaheuristics Optimisation Method*
Aoki, Takafumi	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
Arablouei, Reza	G3.2	*Continual Local Updates for Federated Learning with Enhanced Robustness to Link Noise*
Arakawa, Takahiko	I4.5	*Development of a Robust Ear Recognition Algorithm Using Planar Approximation*
Araki, Shoko	B3.1	*Modified Parametric Multichannel Wiener Filter for Low-Latency Enhancement of Speech Mixtures with Unknown Number of Speakers*
	A5.5	*Spatially-Regularized Switching Independent Vector Analysis*
Armunanta, Dwi	P3.24	*Efficient Dictionary and Grid-Based Framework for Answering Durable k-Nearest Neighbor Queries on Time Series Data*
Arvaneh, Mahnaz	P1.6	*Predicting Outcomes of Cognitive Behavioral Therapy for Depression Using Data Driven Approaches*
Asami, Taichi	H5.1	*Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation*
Baguio, Tyler	G3.4	*Distributed On-Line Anomaly Detection Using Kernel Methods*
Bah, Mohamed	D2.1	*Interpretable Image Recognition in Hyperbolic Space*
Bai, Mingsian	H3.2	*Multichannel Learning-Based Spatially Extended Active Noise Control via Model Matching and Sensor Transfer Function Interpolation*
Bai, Zhongxin	B3.5	*A Weighted Binary Cross-Entropy for Sound Event Representation Learning and Few-Shot Classification*
Bando, Yoshiaki	C5.1	*DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection*
Banerjee, Padmanabha	P1.8	*FID-RPRGAN-VC: Fréchet Inception Distance Loss Based Region-Wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion*
Bao, Changchun	P5.26	*GAN-Based Time-Domain Packet Loss Concealment Method with Consistent Mapping Approach*
Barhumi, Imad	P3.3	*Compressive Sensing Based Algorithms for Limited-View PAT Image Reconstruction*
Barua, Hrishav	J2.2	*ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation*
Baskaran, Vishnu Monn	F4.5	*HDR Image Watermarking Based on Saliency Detection and Quantization Index Modulation*

Benesty, Jacob	B3.5	*A Weighted Binary Cross-Entropy for Sound Event Representation Learning and Few-Shot Classification*
	P4.29	*A Controlled Noise Reduction Wiener Filter Based on the Quadratic Eigenvalue Problem*
Benner, Seico	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Bhardwaj, Mohit	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Bhat, Savita	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Boodhoo, Kirish	P3.6	*Heart Rate Acquisition and Processing Techniques for a Miniature Wearable Microphone Sensor*
Bu, Bing	P2.28	*Single Source Zone Detection in the Spherical Harmonic Domain for Multisource Localization*
Buaruk, Suphachok	D4.6	*Enhancing Snoring Detection with Statistical Analysis of Audio Features*
Buttelli, Olivier	J1.1	*After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals*
Cai, Tao	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Cai, Yunrui	P2.29	*Robust Representation Learning for Speech Emotion Recognition with Moment Exchange*
Cao, Bowen	P5.24	*NADiffuSE: Noise-Aware Diffusion-Based Model for Speech Enhancement*
Chai, Abel Yu Hao	D2.4	*Unveiling Robust Feature Spaces: Image vs. Embedding-Oriented Approaches for Plant Disease Identification*
Chai, AiFang	J2.6	*ELEGANT: End-To-End Language Grounded Speech Denoiser for Efficient Generation of Talking Face*
Chai, Kok Chin	D2.6	*How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?*
	D2.3	*MoMo Strategy: Learn More from More Mistakes*
Chaiwongyen, Anuwat	G5.1	*Deepfake-Speech Detection with Pathological Features and Multilayer Perceptron Neural Network*
Chalothorn, Tawunrat	BS1.4	*Residual, Mixer, and Attention: The Three-Way Combination for Streaming Wake Word Detection Framework*
Chan, Bo-Cheng	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*

Chan, Kai-Hsuan	G5.4	*Reduction of Annotation Effort in Medical Image Analysis Based on Self-Supervised Learning*
Chan, Mei-Ling	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	B4.4	*Point Cloud Inpainting Based on Delaunay Triangulation*
	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Chang, Arthur	C3.2	*Automated Carina Detection in Chest X-Ray Images Using Non-Overlapping and Cross-Squeeze Convolutional Neural Networks*
Chang, Cheng-Yuan	H3.3	*A Study of the Microphone Protection of Active Noise Control for Axial Fan*
Chang, Ching Hsiang	P1.15	*Old Damaged Photo Recovery with Style Transfer-Based Data Augmentation*
Chang, Ching-Chun	C2.2	*Image-Text Out-Of-Context Detection Using Synthetic Multimodal Misinformation*
Chang, Chuan-Yu	C3.5	*Integration of Contextual Information with YOLOv5 for Detecting Tracheobronchial Foreign Objects*
Chang, Dongliang	P1.29	*Multi-Frequency Feature Enhancement for Multi-Granularity Visual Classification*
Chang, Jie-Fan	H2.4	*Light-Weight Zero-Reference-Based Image Enhancement for Low-Light Images*
Chang, Jih-Sheng	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*
Chang, Kai Chieh	P5.17	*Classification of Infant Sleep/Wake States: Cross-Attention Among Large Scale Pretrained Transformer Networks Using Audio, ECG, and IMU Data*
Chang, Ku-Yaw	H2.4	*Light-Weight Zero-Reference-Based Image Enhancement for Low-Light Images*
Chang, Li-Jen	BS1.6	*Mask2Hand: Learning to Predict the 3D Hand Pose and Shape from Shadow*
Chang, Min-Kuan	D5.5	*Autoencoder-Enhanced Federated Learning with Reduced Overhead and Lower Latency*
Chang, Shih-Chieh	G1.2	*Gait Analysis in Powered Exoskeleton-Assisted Walking in Patients with Stroke: A Case Series Cohort*
Chang, Ting-Yu	B5.4	*Multitask Record for Badminton Match*
Chang, Wen-Yi	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*

Chang, Yang Loong	D2.6	*How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?*
	D2.3	*MoMo Strategy: Learn More from More Mistakes*
Chang, Yao-Chung	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
	P2.10	*Improving Regularization of Deep Learning Models in Fundus Analysis*
Chang, Yen-Ching	P2.14	*Asynchronous Federated Learning Method in the 6G Massive Multimedia Transmission System*
Chang, Yi-Cheng	F5.2	*Study on the Reduction of Background Fringes for Detect Detection on Specular Surface*
Chang, Yu-Jen	H1.1	*Machine Learning Based Action Recognition with Modular CNN*
Chanpornpakdi, Ingon	BS1.3	*Event-Related Potential in Rapid Serial Visual Presentation-Based Partial Face Cognition Depends on Visible Face Components*
Chao-Tsung, Huang	P3.15	*A Biased Mixed-Precision Convolution Engine for Hardware-Efficient Computational Imaging CNN*
Charlton, Peter	I1.1	*A Comparative Evaluation of Video Codecs for rPPG-Based Heart Rate Estimation*
	I1.5	*Signal Quality Assessment for Wearable Multichannel Photoplethysmography Signals*
Chen, Berlin	P5.11	*AVATAR: Robust Voice Search Engine Leveraging Autoregressive Document Retrieval and Contrastive Learning*
Chen, Boru	A4.1	*ResNet-Based Camera Model Identification with Adaptive Preprocessing Module and Weight Fusion of Global Information*
Chen, Chen	E3.3	*Study of Generative Adversarial Networks for Noisy Speech Simulation from Clean Speech*
Chen, Chi-Yuan	C3.2	*Automated Carina Detection in Chest X-Ray Images Using Non-Overlapping and Cross-Squeeze Convolutional Neural Networks*
Chen, Chia-Ping	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*
Chen, Chien-Chih	I5.2	*Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind*
Chen, Chien-Ching	B1.4	*Collaborative Pseudo Labeling for Prompt-Based Learning*
Chen, Ching-Chieh	P4.9	*Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods*
Chen, Chiung-An	P2.8	*VLSI Design of Lossless Image Compression Using Improved LZW*

	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	B4.3	*Identification of Victims Wearing Vibrant Clothing Using MATLAB*
	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
	B4.2	*The VLSI Design DSP and DDR4 for Multi-Sensor in Biomedical System*
Chen, Cynthia	G3.3	*Multi-Task Gaussian Process Learning to Fill in Gaps in LBS Mobile Data*
Chen, Duan-Yu	P1.12	*Lightweight Models Distillation with Learnable Teaching Material: An Application for Smart Table Tennis System*
Chen, Fan	P5.22	*Time Sensitive Hash and Adaptive Image Recovery Based Self-Embedding Fragile Watermarking Scheme in Encrypted Images*
Chen, Fang-Ching	H1.5	*Utilizing Unlabeled Data and Synthetic Data for Bird Sound Detection: Consistency Training, Mean Teacher, and Domain Adaptation Techniques*
Chen, Fei	B2.2	*Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-To-End CNN with Fine-Tuning*
	P3.23	*Assessing the Effects of Filtering Processing on Pulse Wave Transit Time Measured by Photoplethysmography from Earlobe*
Chen, Gong	B3.5	*A Weighted Binary Cross-Entropy for Sound Event Representation Learning and Few-Shot Classification*
Chen, Hang	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
	P2.26	*Improved Data2vec with Soft Supervised Hidden Unit for Mandarin Speech Recognition*
	P2.25	*Progressive Multi-Scale Self-Supervised Learning for Speech Recognition*
	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Chen, Heliu	P1.23	*CoA-DLinkNet: Connectivity-Enhanced Dual-Branch Road Extraction Network Based on D-LinkNet*
Chen, Hui-Lin	P3.19	*DOC: A Novel DOuble-Contour-Based Macro Placement Framework for Mixed-Size Designs*
Chen, Hwann-Tzong	BS1.6	*Mask2Hand: Learning to Predict the 3D Hand Pose and Shape from Shadow*
Chen, Jeffrey	B5.4	*Multitask Record for Badminton Match*
Chen, Jia-Hong	P1.14	*Analysis of the Interaction Effect on Pruning and Transfer Learning in Model Training*

Chen, Jingdong	P1.20	*A Subband Approach to Personal Sound Zone with Joint Optimization of Sound Pressure and Particle Velocity*
Chen, Jingdong	B3.5	*A Weighted Binary Cross-Entropy for Sound Event Representation Learning and Few-Shot Classification*
	E3.2	*On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering*
Chen, Jingdong	P4.29	*A Controlled Noise Reduction Wiener Filter Based on the Quadratic Eigenvalue Problem*
Chen, Jun-Cheng	BS1.1	*Enhancing Real-Time Semantic Segmentation with Textual Knowledge of Pre-Trained Vision-Language Model: A Lightweight Approach*
Chen, Junjie	P4.25	*Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition*
Chen, Kuan-Dar	P4.9	*Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods*
Chen, Lei	P1.30	*Improving Aspect Sentiment Classification via Retrieving from Training Data*
Chen, Ming-Yen	B1.5	*Learning Meta Soft Prompt for Few-Shot Language Models*
Chen, Pei-Yuan	I5.2	*Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind*
Chen, Po-Kai	F5.5	*Question Answering System Based on Pre-Training Model and Retrieval Reranking for Industry 4.0*
Chen, Po-See	G5.2	*Temporal and Type Correlation in Digital Phenotyping for Bipolar Disorder State Prediction Using Multitask Self-Supervised Learning*
Chen, Qian	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Chen, Shih-Lun	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
	B4.4	*Point Cloud Inpainting Based on Delaunay Triangulation*
	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
	B4.2	*The VLSI Design DSP and DDR4 for Multi-Sensor in Biomedical System*
Chen, Sze-Ann	G4.5	*Virtual Garment Fitting Through Parsing and Context-Aware Generative Adversarial Networks with Discriminator Group*

Chen, Ta-Fu	F5.5	*Question Answering System Based on Pre-Training Model and Retrieval Reranking for Industry 4.0*
Chen, Ting-Wei	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
Chen, Tsung-Ping	F1.3	*Learning Multifaceted Self-Similarity for Musical Structure Analysis*
Chen, Tsung-Yi	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Chen, Tzu-Chieh	C3.1	*A Reconfigurable Hardware Architecture for Graph Convolution Network in Action Recognition*
Chen, Wei-Yu	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*
Chen, Wei-Yu	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*
Chen, Wen-Ru	E5.5	*Impacts of 5G-TDD Time Slot Configurations on the Downlink and Uplink Data Rates*
Chen, Xianhong	P2.23	*Multi-Task Piano Transcription with Local Relative Time Attention*
Chen, Xin-Yu	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Chen, Xingyan	C4.3	*Privacy-Oriented Coded Caching in Mobile Information-Centric Network*
Chen, Xingyu	P4.28	*Sound Field Estimation Around a Rigid Sphere with Physics-Informed Neural Network*
Chen, Yan	P2.17	*STrack: Velocity Estimation Using Single Antenna WiFi Devices*
Chen, Yan-Hao	P1.16	*A Deep Learning Based Sustainable Energy Scheduling System*
Chen, Yan-Ru	B4.6	*Safety Enhancement for Mobility Scooter with Rule-Based Danger Prevention*
Chen, Yanli	P3.26	*Fine-Grained Face Anti-Spoofing Based on Recursive Self-Attention and Multi-Scale Fusion*
Chen, Yen-Hsin	A4.6	*Exploring a CLIP-Enhanced Automated Approach for Video Description Generation*
Chen, Yi-Hsin	G4.3	*Transformer-Based Image Compression with Variable Image Quality Objectives*
Chen, Yong-Tai	P3.15	*A Biased Mixed-Precision Convolution Engine for Hardware-Efficient Computational Imaging CNN*
Chen, You-Siang	H3.2	*Multichannel Learning-Based Spatially Extended Active Noise Control via Model Matching and Sensor Transfer Function Interpolation*

Chen, Young-Long	E4.4	*Learned String Quartet Music with Variational Auto Encoder*
Chen, Yu-Guang	P3.19	*DOC: A Novel DOuble-Contour-Based Macro Placement Framework for Mixed-Size Designs*
Chen, Yu-Hsuan	P1.12	*Lightweight Models Distillation with Learnable Teaching Material: An Application for Smart Table Tennis System*
Chen, Yu-Jen	P1.13	*Selecting Suitable Data Input for Deep-Learning Sign-Language Recognition with a Small Dataset*
Chen, Yueru	J2.3	*LSR++: An Efficient and Tiny Model for Image Super-Resolution*
Chen, Yunlin	P5.23	*Multi-Granularity Semantic and Acoustic Stress Prediction for Expressive TTS*
Cheng, Cheng-Hsu	G1.2	*Gait Analysis in Powered Exoskeleton-Assisted Walking in Patients with Stroke: A Case Series Cohort*
Cheng, Cuiping	P1.21	*An Multi-Evidence Fusion Based on C-Distance with Uncertain Reasoning for Classification*
	P2.27	*Investigation of Ensemble of Self-Supervised Models for Speech Emotion Recognition*
Cheng, Fang-Yi	I5.4	*Impact of the Weighted Loss Function on the Innovative CMAQ-CNN PM2.5 Forecasting Model*
Cheng, Haowei	P4.2	*Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection*
Cheng, Hsu-Yung	D5.2	*In-Air Handwriting for Chinese Character Recognition from Monocular Camera: A Deep Learning Based Approach with Fingertip Detection and Virtual Strokes Elimination*
Cheng, Huai-Hsun	A4.6	*Exploring a CLIP-Enhanced Automated Approach for Video Description Generation*
Cheng, Jen-Chieh	P5.15	*Multitaper Adaptive Time-Frequency Windowed Fourier Transform Based on the Reliable Region of Window Widths*
Cheng, Lulu	P4.16	*Efficient Reversible Data Hiding for 3D Mesh Models Based on Multi-LSB Substitution and Ring-Prediction*
Cheng, Shi	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Cheng, Wen-Huang	G4.2	*Task-Specific Pruning: Efficient Parameter Reduction in Multi-Task Object Detection Models*
Cheng, Yu-Han	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*
Cheung, Tsun-Hin	P2.2	*FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking*

Chi, Jung-Chun	H1.4	*Deep-Learning-Based Lattice Reduction Preprocessing for Time-Correlated MIMO Systems*
Chi, Tai-Shih	B3.4	*A Dual-Channel Three-Stage Model for DoA and Speech Enhancement*
	P5.14	*Residual Echo Suppression Using Spatial Feature for Stereo Acoustic Echo Cancellation*
Chi, Wenjiang	P5.23	*Multi-Granularity Semantic and Acoustic Stress Prediction for Expressive TTS*
Chiang, Chen Yu	P2.6	*Tone Labeling by Deep Learning-Based Tone Recognizer for Mandarin Speech*
Chiang, Chen-Kuo	H2.5	*Class-Wise Self-Paced Self-Training for Semi-Supervised Image Classification*
Chiang, Jui -Chiu	G4.6	*Sparse Tensor-Based Point Cloud Attribute Compression Using Augmented Normalizing Flows*
Chien, Cheng	G4.3	*Transformer-Based Image Compression with Variable Image Quality Objectives*
Chien, Feng-Tsun	D5.5	*Autoencoder-Enhanced Federated Learning with Reduced Overhead and Lower Latency*
Chien, Jen-Tzung	P1.7	*Learning Adapters for Code-Switching Speech Recognition*
	B1.5	*Learning Meta Soft Prompt for Few-Shot Language Models*
	B1.4	*Collaborative Pseudo Labeling for Prompt-Based Learning*
	P2.7	*Learning to Enhance the Position Embedding and Coherence*
Chien, Ming-Chun	P4.15	*Breast Cancer Detection Auxiliary System Leveraging Deep Learning and Mixed Reality*
Chin, Chen-I	G4.5	*Virtual Garment Fitting Through Parsing and Context-Aware Generative Adversarial Networks with Discriminator Group*
Chin, Yen-Ju	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
Ching-Hui, Huang	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Chiu, Ching-Te	H1.1	*Machine Learning Based Action Recognition with Modular CNN*
Chiu, Wei-Chen	G4.3	*Transformer-Based Image Compression with Variable Image Quality Objectives*
Chng, Eng-Siong	B1.2	*Adopting Neural Translation Model in Data Generation for Inverse Text Normalization*
	E3.3	*Study of Generative Adversarial Networks for Noisy Speech Simulation from Clean Speech*
	H4.5	*ASR Model Adaptation for Rare Words Using Synthetic Data Generated by Multiple Text-To-Speech Systems*

	A5.1	*Analysis of Speech Separation Performance Degradation on Emotional Speech Mixtures*
Cho, Nam Ik	G2.1	*Resolution-Adaptive Lossless Image Compression Using Frequency Decomposition Network*
Chodingala, Piyushkumar K.	E1.2	*Relevance of Quadrature Phase for Replay Detection in Voice Assistants (VAs)*
Choi, Hyuk	P3.9	*Investigation of the Neuromodulation Effect of tVNS on Electrical Brain Activity*
Choi, Seonghwa	A2.5	*Holo-QoI: A Human Factor-Based Dataset and Prediction Framework for Assessing Quality of Interaction in Augmented Reality*
Chong, Chun Yong	D2.5	*Facial Expression Recognition as Markers of Depression*
Chou, Cheng-Hou	P4.10	*Detecting the Underwater Distance and Swimming Direction of Tilapia Using YOLO*
Chou, He-Sheng	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
	B4.4	*Point Cloud Inpainting Based on Delaunay Triangulation*
	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
Chou, Hsuan-Cheng	B3.4	*A Dual-Channel Three-Stage Model for DoA and Speech Enhancement*
	P5.14	*Residual Echo Suppression Using Spatial Feature for Stereo Acoustic Echo Cancellation*
Chu, Minseok	D4.5	*An Improved Check Digit-Based Participant Identification System for Human Biorepositories*
Chuang, Hsiang-Feng	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*
Chudiwal, Utkarsh	E3.3	*Study of Generative Adversarial Networks for Noisy Speech Simulation from Clean Speech*
Chulif, Sophia	D2.6	*How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?*
	D2.3	*MoMo Strategy: Learn More from More Mistakes*
Chun-Wei, Li	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
Chung, Meng-Ju	I5.2	*Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind*
Chung, Ming-Yu	E4.1	*Un-Rectifying in ReLU Networks and Applications*
Chung, Yu-Ching	E3.1	*An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids*

Chung-Nan, Lee	E5.3	*A UAV Indoor Obstacle Avoidance System Based on Deep Reinforcement Learning*
Ciou, Tai-Cyuan	B5.4	*Multitask Record for Badminton Match*
Constandinou, Timothy	I1.1	*A Comparative Evaluation of Video Codecs for rPPG-Based Heart Rate Estimation*
Cooper, Erica	P5.8	*Exploring Isolated Musical Notes as Pre-Training Data for Predominant Instrument Recognition in Polyphonic Music*
Cui, Jianwei	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
Da Zhi, Gao	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Dai, Dongyang	P2.29	*Robust Representation Learning for Speech Emotion Recognition with Moment Exchange*
Dai, Li-Rong	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Dai, Yuchao	P1.23	*CoA-DLinkNet: Connectivity-Enhanced Dual-Branch Road Extraction Network Based on D-LinkNet*
Dang, Cong-Ty	G4.4	*From Synthetic to Real: Enhancing Deep Learning Models with Generative Adversarial Networks for Efficient Data Utilization in Automatic Retail Stores*
Dang, Tung	B2.3	*Decoding Time-Course of Saliency Network of fMRI Signals by EEG Signals Using Optimized Forward Variable Selection: A Concurrent EEG-fMRI Study*
Das, Swagatam	P1.8	*FID-RPRGAN-VC: Fréchet Inception Distance Loss Based Region-Wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion*
Deepaisarn, Somrudee	D4.6	*Enhancing Snoring Detection with Statistical Analysis of Audio Features*
Delgadillo, Jaime	P1.6	*Predicting Outcomes of Cognitive Behavioral Therapy for Depression Using Data Driven Approaches*
Deng, Pan	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
Deng, Tengyu	BS1.5	*Audio-To-Score Singing Transcription Based on Joint Estimation of Pitches, Onsets, and Metrical Positions with Tatum-Level CTC Loss*
Dhall, Abhinav	J2.2	*ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation*
Dhar, Sandipan	P1.8	*FID-RPRGAN-VC: Fréchet Inception Distance Loss Based Region-Wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion*
Dharmaratne, Anuja	P3.4	*Towards AST-LLDs for the Analysis of Depression in Speech Signals*

Ding, Jian-Jiun	P5.15	*Multitaper Adaptive Time-Frequency Windowed Fourier Transform Based on the Reliable Region of Window Widths*
	D5.3	*EffSegmentNet: Efficient Design for Real-Time Semantic Segmentation*
Ding, Shihang	A4.2	*Structural Quality Assured Global Optimization for CTU-Level Rate Control of Screen Content Coding*
Dinh, Binh-Duong	J3.1	*1M Parameters are Enough? A Lightweight CNN-Based Model for Medical Image Segmentation*
Do, Hai	A3.4	*An Automatic Pipeline for Building Emotional Speech Dataset*
Doan, Shintaro	P2.3	*Auditory Representation Effective for Estimating Vocal Tract Information*
Doi, Kenya	H3.4	*SFANC with Compensation Filter Based on MEFxDCTLMS Algorithm*
Dong, Jianmin	P2.16	*Poisoning Attacks Against Gait-Based Identity Recognition*
Dong, Wenxiang	P1.18	*Modeling and Analysis of the Epidemic-Behavior Co-Evolution Dynamics with User Irrationality*
Du, Jun	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Du, Muyang	P3.17	*Efficient Incremental Text-To-Speech on GPUs*
Du, Zhihao	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Duangpummet, Suradej	P2.1	*Segmentation Enhancement for Iris Recognition Using Unit Gradient Vectors*
	G5.1	*Deepfake-Speech Detection with Pathological Features and Multilayer Perceptron Neural Network*
Duc-Tan, Tran	J3.2	*Imaging Ultrasound Scattering Targets Using Density-Enhanced Chaotic Compressive Sampling*
Duong, Bach	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*
Duong, Thanh	J3.4	*Investigating the Role of Human Action Detector in Visual-Guide Audio Source Separation System*
Echizen, Isao	C2.2	*Image-Text Out-Of-Context Detection Using Synthetic Multimodal Misinformation*
	P3.20	*Hindering Adversarial Attacks with Multiple Encrypted Patch Embeddings*
Eguchi, Ryunosuke	I2.4	*JPEG Artifact Removal for Hyperspectral Images Based on Spatial-Spectral Regularization*

	D3.4	*Convolutional Multidimensional Amplitude Spectrum Nuclear Norm for Frequency-Domain Robust Principal Component Analysis*
Emoto, Atsuya	P3.7	*A Study on Counter-Movement Detection on Dumbbell Curl Using Convolutional Autoencoder*
Endrawati, Devi	I3.4	*Sparsity Exploration for Structured and Unstructured Weight Formations in CNN Architecture*
Fan, Chih-Peng	E4.2	*OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises*
Fan, Cunhang	A1.1	*Mixed Emotion Recognition Based on EEG Signals*
Fan, Wen-Lin	C3.2	*Automated Carina Detection in Chest X-Ray Images Using Non-Overlapping and Cross-Squeeze Convolutional Neural Networks*
Fei, Xuezhi	C4.5	*Feature Augmentation Reconstruction Network for Few-Shot Image Classification*
Felouat, Hichem	C2.2	*Image-Text Out-Of-Context Detection Using Synthetic Multimodal Misinformation*
Feng, Chih-Yung	I5.4	*Impact of the Weighted Loss Function on the Innovative CMAQ-CNN PM2.5 Forecasting Model*
Feng, Junlan	P1.19	*Noise-Robust Pitch Detection Based on Super-Resolution Harmonics*
Feng, Linqing	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Feng, Xiaoqin	P5.23	*Multi-Granularity Semantic and Acoustic Stress Prediction for Expressive TTS*
Feng, Yan	C1.2	*Instance Implant-Aided Non-Uniformly Cropping for Person Detection in Aerial Images*
	P3.25	*Retinex-Based Low-Light Image Enhancement*
Fu, Meijiang	P1.29	*Multi-Frequency Feature Enhancement for Multi-Granularity Visual Classification*
Fu, Qiang	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
Fujio, Masakazu	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
Fujisawa, Shuntaro	D4.3	*Deep Learning-Based MRI Super-Resolution Using Non-Uniform Segmented Phase-Scrambling Fourier Transform Signals*
Fujita, Yoto	C5.1	*DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection*
Fujiwara, Koichi	G1.3	*Prediction Model of Postoperative Pain Exacerbation Using a Wearable Electrocardiogram Sensor*

	P1.4	*Soft-Sensor Construction Method Based on Adaptive Modeling and Transfer Learning for Manufacturing Process Including Maintenance Periods*
	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Fukuma, Seishi	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
Fukumori, Takahiro	P4.4	*Investigating the Effectiveness of Speaker Embeddings for Shout Intensity Prediction*
	A5.3	*Universal Sound Separation Using Replay-Based Data Sampling in Incremental Learning*
Fukushima, Norishige	I2.1	*Local Contrast Enhancement with Multi-Scale Filtering*
Funabiki, Nobuo	C2.3	*Gait Recognition Scheme Focusing on Operating Characteristics at Feature Points Detected by OpenPose*
	C2.4	*A Study on Eliminating Biased Node in Federated Learning*
	F4.4	*Study on Face Landmark-Based Analysis for Synthetic Media Identification Generated by Adversarial Generative Networks*
	E4.2	*OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises*
Furunaga, Shinya	B3.3	*Moving Interference Speaker Removal Using Geometrically Constrained Independent Vector Analysis*
Furusawa, Soma	P3.22	*Numerical Performance Evaluation of ¥(¥ell_1-¥ell_2¥) Sparse Reconstruction Using Optical Analog Circuit*
Gan, Woon Seng	H3.5	*Practical Active Noise Control: Restriction of Maximum Output Power*
Gandhi, Vineet	E1.5	*Adversarial Robustness of Mel Based Speaker Recognition Systems*
	A4.4	*RobustL2S: Speaker-Specific Lip-To-Speech Synthesis Exploiting Self-Supervised Representations*
Gao, Peng	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Gao, Yingying	P1.19	*Noise-Robust Pitch Detection Based on Super-Resolution Harmonics*
Gao, Yu	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Garcia Perera, Leibny Paola	H4.3	*Synthetic Data Augmentation for ASR with Domain Filtering*
Geng, Yuting	P5.9	*Speech Quality Improvement Utilizing Out-Of-Focus Areas in Rolling-Shutter Video on Speech Extraction*
	P5.4	*Narrow-Edged Acoustical Beamforming Utilizing Phase Inversion for Frequency Modulation-Based Parametric Array Loudspeaker*

Gian, Toan	J1.4	*Transformer-Based Deep Learning Detector for Dual-Mode Index Modulation 3D-OFDM*
Ginting, Kendrik Emkel	I3.3	*SUMO Based Hardware/Software Co-Simulation for Two-Intersection Adaptive and Collaborative Traffic Signal Controller*
Gogineni, Vinay Chakravarthi	G3.2	*Continual Local Updates for Federated Learning with Enhanced Robustness to Link Noise*
Gu, Jason	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
	P1.28	*Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion*
Gue, Jia Xuan	D2.5	*Facial Expression Recognition as Markers of Depression*
Guo, Aoqi	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Guo, Jing-Ming	B5.5	*Deep Residual and Classified Neural Networks for Inverse Halftoning*
	B5.4	*Multitask Record for Badminton Match*
Guo, Ning	B3.1	*Modified Parametric Multichannel Wiener Filter for Low-Latency Enhancement of Speech Mixtures with Unknown Number of Speakers*
Guo, Qinwen	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Guo, Shen	P4.23	*Enhancing Whisper Model for Pronunciation Assessment with Multi-Adapters*
Guo, Sheng	B4.5	*Dense Three-Dimensional Color Reconstruction for Large-Scale Outdoor Scenes*
Guo, Taiyang	H5.2	*Contribution of Modulation Spectral Features for Cross-Lingual Speech Emotion Recognition Under Noisy Reverberant Conditions*
Guo, Xiaowei	C4.6	*Dual Feature Reconstruction Network Based on Few-Shot Image Classification*
Guo, Yu	C4.3	*Privacy-Oriented Coded Caching in Mobile Information-Centric Network*
Guo, Yusheng	P4.18	*Preemptive Image Protection Against Steganography*
Gupta, Priyanka	E1.2	*Relevance of Quadrature Phase for Replay Detection in Voice Assistants (VAs)*
	P3.2	*Classification of Normal vs. Pathological Infant Cries Using Morse Wavelets*
Ha-Phan, Ngoc-Quan	H2.3	*A Reliable Feature-Based Framework for Vehicle Tracking in Advanced Driver Assistance Systems*
Hai, Vu	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*

Hamdulla, Askar	P4.30	*Target Speaker Extraction with Attention Enhancement and Gated Fusion Mechanism*
Han, Ji-Yan	E3.1	*An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids*
Han, Meng	P3.26	*Fine-Grained Face Anti-Spoofing Based on Recursive Self-Attention and Multi-Scale Fusion*
Han, Xinhui	P3.28	*Robust Watermark Imaging via Graph-Signal Optimization*
Han, Yoonchang	C5.4	*Cross-Domain Sound Recognition for Efficient Underwater Data Analysis*
Haneda, Yoichi	P5.21	*Subjective Evaluation of a Focused Sound Source Reproducing at the Positions of a Listener's Moving Hand*
Haoyang, Li	H4.5	*ASR Model Adaptation for Rare Words Using Synthetic Data Generated by Multiple Text-To-Speech Systems*
Hara, Sunao	F3.3	*Speech Synthesis Using Ambiguous Inputs from Wearable Keyboards*
	P5.3	*Speech-Emotion Control for Text-To-Speech in Spoken Dialogue Systems Using Voice Conversion and x-Vector Embedding*
Harada, Noboru	E3.5	*Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer*
	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Harashima, Ryoya	D3.4	*Convolutional Multidimensional Amplitude Spectrum Nuclear Norm for Frequency-Domain Robust Principal Component Analysis*
Harfiya, Latifa Nabila	F5.3	*On the Optimal Self-Supervised Multi-Fault Detector for Temperature Sensor Data*
Harjoko, Agus	G5.5	*STUA-Net: A Fingerprint Reconstruction with Swin Transformer and Soft Collective Attention*
Hartono, Rudi	P3.18	*Dual-Path Residual Attention Convolution Networks for Color-Embedded-Grayscale Image*
Hasegawa-Johnson, Mark	P5.17	*Classification of Infant Sleep/Wake States: Cross-Attention Among Large Scale Pretrained Transformer Networks Using Audio, ECG, and IMU Data*
Hashimoto, Masafumi	P3.1	*Remarks on an Optimal Predictive Control Using a Quaternion Neural Network and a Metaheuristics Optimisation Method*
Hatzinakos, Dimitrios	P2.18	*SEformer: Dual-Path Conformer Neural Network is a Good Speech Denoiser*
Hau, Joan	F4.6	*Quick Response (QR) Codes Embedding in VVC Using Quantisation Parameter Manipulation*

Hayakawa, Ryo	P3.22	*Numerical Performance Evaluation of ¥(¥ell_1-¥ell_2¥) Sparse Reconstruction Using Optical Analog Circuit*
Hayashi, Kazunori	P3.22	*Numerical Performance Evaluation of ¥(¥ell_1-¥ell_2¥) Sparse Reconstruction Using Optical Analog Circuit*
Hayashi, Kohei	I2.1	*Local Contrast Enhancement with Multi-Scale Filtering*
He, Changjun	P2.24	*Real and Imaginary Part Interaction Network for Monaural Speech Enhancement and De-Reverberation*
He, Chun-Yi	P1.7	*Learning Adapters for Code-Switching Speech Recognition*
He, Hongjie	P5.22	*Time Sensitive Hash and Adaptive Image Recovery Based Self-Embedding Fragile Watermarking Scheme in Encrypted Images*
He, Mingyi	P1.23	*CoA-DLinkNet: Connectivity-Enhanced Dual-Branch Road Extraction Network Based on D-LinkNet*
He, Renjie	P1.23	*CoA-DLinkNet: Connectivity-Enhanced Dual-Branch Road Extraction Network Based on D-LinkNet*
He, Shulin	P5.29	*ScaleFormer: Transformer-Based Speech Enhancement in the Multi-Scale Time Domain*
Hidayat, Taufiqurrakhman	P3.18	*Dual-Path Residual Attention Convolution Networks for Color-Embedded-Grayscale Image*
Hidayati, Shintami Chusnul	B5.1	*ASF-LLRDA: Locality-Regularized Linear Regression Discriminant Analysis with Approximately Symmetrical Face Preprocessing for Face Recognition*
Higashi, Hiroshi	B2.1	*Supervised Single-Channel EEG Decomposition Using Detector-Kernel Networks for Noise Reduction*
	I2.2	*Marine Snow Removal Benchmarking Dataset*
	BS1.2	*EEG Study on Anticipation of Difficulty for Upcoming Auditory Task*
Hirata, Kouji	E5.4	*Approximate Modeling of Malware Diffusion on Wireless Mobile Devices*
Hiroe, Atsuo	P4.5	*Is the Ideal Ratio Mask Really the Best?---Exploring the Best Extraction Performance and Optimal Mask of Mask-Based Beamformers*
Hirohashi, Miho	P5.21	*Subjective Evaluation of a Focused Sound Source Reproducing at the Positions of a Listener's Moving Hand*
Hladek, Daniel	F5.4	*Application of Wafer Defect Pattern Classification Model in the Semiconductor Industry*
Ho, Chian C.	C3.4	*Pose-Based Visual Servoing with Lightweight Deep-Learning Binarization for Autonomous Mobile Robot Application*
Ho, Jinn	E4.1	*Un-Rectifying in ReLU Networks and Applications*
Ho, Thi-Nga	B1.2	*Adopting Neural Translation Model in Data Generation for Inverse Text Normalization*

Ho, Tuan Vu	H4.3	*Synthetic Data Augmentation for ASR with Domain Filtering*
Hoa, Nguyen	J1.4	*Transformer-Based Deep Learning Detector for Dual-Mode Index Modulation 3D-OFDM*
Hoang Bach, Nguyen	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*
Hojo, Keigo	P5.2	*Combining Multiple End-To-End Speech Recognition Models Based on Density Ratio Approach*
Hon, Hock Woon	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Honda, Jun	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
Hong, Ki-Hwan	P3.9	*Investigation of the Neuromodulation Effect of tVNS on Electrical Brain Activity*
Hong, Le	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*
Hong, Qingyang	D1.3	*CASA-Net: Cross-Attention and Self-Attention for End-To-End Audio-Visual Speaker Diarization*
Hora, Baveet	E1.3	*Exploring Residual Cepstral Features for Spoken Language Identification*
Horiguchi, Shota	H4.3	*Synthetic Data Augmentation for ASR with Domain Filtering*
Horii, Koharu	P4.6	*Language Modeling for Spontaneous Speech Recognition Based on Disfluency Labeling and Generation of Disfluent Text*
Hou, Tanshuai	P5.27	*MHM-Based Reversible Data Hiding with Adaptive Feature Selection*
Hsia, Chih-Hsien	P4.27	*Improve Lightweight Convolutional Neural Network for Finger Vein Recognition System*
	P4.12	*Wavelet and Cutout in YOLO Architecture for Road Pothole Detection*
	P4.14	*Lightweight CNN and Image Enhancement Using in Palm Vein Recognition*
Hsiao, Hsu-Feng	G4.5	*Virtual Garment Fitting Through Parsing and Context-Aware Generative Adversarial Networks with Discriminator Group*
Hsieh, Chi-Kai	D5.5	*Autoencoder-Enhanced Federated Learning with Reduced Overhead and Lower Latency*
Hsieh, I-Ting	G5.3	*Data Selection Based on Phoneme Affinity Matrix for Electrolarynx Speech Recognition*
Hsieh, Yi-Zeng	P4.10	*Detecting the Underwater Distance and Swimming Direction of Tilapia Using YOLO*
Hsu, Chung-Chian	C3.2	*Automated Carina Detection in Chest X-Ray Images Using Non-Overlapping and Cross-Squeeze Convolutional Neural Networks*

Hsu, Heng-Cheng	H2.5	*Class-Wise Self-Paced Self-Training for Semi-Supervised Image Classification*
Hsu, Jia-Hao	G5.2	*Temporal and Type Correlation in Digital Phenotyping for Bipolar Disorder State Prediction Using Multitask Self-Supervised Learning*
Hsu, Wan-Lin	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
Hsu, Wei-Cheng	P4.9	*Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods*
Hsu, Wei-Wen	P2.10	*Improving Regularization of Deep Learning Models in Fundus Analysis*
	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
Hsu, Yan-Cheng	F5.3	*On the Optimal Self-Supervised Multi-Fault Detector for Temperature Sensor Data*
Hsu, Yu-Teng	P5.10	*Personalized Audio Quality Preference Prediction*
Hu, Chia-Yu	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
Hu, Shiyong	B1.1	*GRALA: Modeling Social Information for Microblog Sentiment Analysis from the View of Balancing Sparsity and Smoothness of Social Contexts*
Hu, Wei	P3.28	*Robust Watermark Imaging via Graph-Signal Optimization*
Huang, Chen-Hsiu	G4.1	*CPIPS: Learning to Preserve Perceptual Distances in End-To-End Image Compression*
Huang, Ching-Chun	H2.3	*A Reliable Feature-Based Framework for Vehicle Tracking in Advanced Driver Assistance Systems*
	H2.2	*Efficient Constraint-Aware Neural Architecture Search for Object Detection*
	G4.4	*From Synthetic to Real: Enhancing Deep Learning Models with Generative Adversarial Networks for Efficient Data Utilization in Automatic Retail Stores*
Huang, Choa-Ting	P1.16	*A Deep Learning Based Sustainable Energy Scheduling System*
Huang, Chung-I	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*
Huang, Guo-Wei	P1.16	*A Deep Learning Based Sustainable Energy Scheduling System*
Huang, Hsin-I	E4.4	*Learned String Quartet Music with Variational Auto Encoder*
Huang, Jian-Jia	G1.4	*Directional Neural Connectivity During Robot Mirror Therapy in Patients with Stroke*
	G1.2	*Gait Analysis in Powered Exoskeleton-Assisted Walking in Patients with Stroke: A Case Series Cohort*

Huang, Kuan-Hua	P1.10	*Learning Single Image Rain Streak Removal Based on Deep Attention Mechanism*
Huang, Shi-Zong	H1.1	*Machine Learning Based Action Recognition with Modular CNN*
Huang, Shilei	P1.30	*Improving Aspect Sentiment Classification via Retrieving from Training Data*
Huang, Shuai	G3.3	*Multi-Task Gaussian Process Learning to Fill in Gaps in LBS Mobile Data*
Huang, Wan-Chia	E4.2	*OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises*
Huang, Wen-Chin	F3.1	*Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion*
Huang, Wenmin	P3.30	*AI-Generated Image Detection Using a Cross-Attention Enhanced Dual-Stream Network*
Huang, Yu-Chuan	P2.10	*Improving Regularization of Deep Learning Models in Fundus Analysis*
Huang, Yu-Fang	E4.5	*SOAda-YOLOR: Small Object Adaptive YOLOR Algorithm for Road Object Detection*
Huang, Yu-Shun	B5.4	*Multitask Record for Badminton Match*
Huang, Yuan-Hao	H1.4	*Deep-Learning-Based Lattice Reduction Preprocessing for Time-Correlated MIMO Systems*
Huang, Zi Hang	D5.2	*In-Air Handwriting for Chinese Character Recognition from Monocular Camera: A Deep Learning Based Approach with Fingertip Detection and Virtual Strokes Elimination*
Hung, Jen-Wen	G1.4	*Directional Neural Connectivity During Robot Mirror Therapy in Patients with Stroke*
Hung, L v	B5.5	*Deep Residual and Classified Neural Networks for Inverse Halftoning*
Hung, Wei-Jie	H2.2	*Efficient Constraint-Aware Neural Architecture Search for Object Detection*
Hussin, Zarina	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Hwang, Han-Jeong	P3.11	*Optimal Input Data Search of Deep Learning Model for Cognitive Task Classification Based on NIRS*
	P3.10	*Gender-Specific EEG Feature-Based Diagnostic Tool for Male Depression Based on Machine-Learning Techniques*
	P3.9	*Investigation of the Neuromodulation Effect of tVNS on Electrical Brain Activity*
Hwang, Wen-Liang	E4.1	*Un-Rectifying in ReLU Networks and Applications*
Hyanda, Muhammad	I1.1	*A Comparative Evaluation of Video Codecs for rPPG-Based Heart Rate Estimation*
Ibrahum, Ahmed	D4.4	*An Extreme Gradient Boosting-Based Prediction for Depression*

Ichinohe, Noritaka	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Igarashi, Ami	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Iimura, Yasushi	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Iinuma, Yuta	P5.18	*Dynamic Characteristics of Electroencephalogram Reflecting Driving-Experience-Dependent Performance Using Microstates*
Ikehara, Masaaki	J2.5	*Enhanced Residual Fourier Transformation Network for Lightweight Image Super-Resolution*
Ikeshita, Rintaro	A5.5	*Spatially-Regularized Switching Independent Vector Analysis*
Imaizumi, Shoko	C2.1	*A Reversible Image Processing Method for Color Tone Control Using Data Hiding*
Imoto, Keisuke	C5.1	*DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection*
	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Inagaki, Keiichiro	P1.2	*P300 Event-Related Potential in Perception of Multiple Traffic Objects During Vehicle Driving*
	P5.18	*Dynamic Characteristics of Electroencephalogram Reflecting Driving-Experience-Dependent Performance Using Microstates*
Inoue, Koh	P3.7	*A Study on Counter-Movement Detection on Dumbbell Curl Using Convolutional Autoencoder*
Irino, Toshio	P2.3	*Auditory Representation Effective for Estimating Vocal Tract Information*
Ishida, Taito	P4.4	*Investigating the Effectiveness of Speaker Embeddings for Shout Intensity Prediction*
Ishrat, Hamza Ahmed	D2.4	*Unveiling Robust Feature Spaces: Image vs. Embedding-Oriented Approaches for Plant Disease Identification*
Islam, Bashima	P5.17	*Classification of Infant Sleep/Wake States: Cross-Attention Among Large Scale Pretrained Transformer Networks Using Audio, ECG, and IMU Data*
Ito, Koichi	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
Ito, Satoshi	D4.3	*Deep Learning-Based MRI Super-Resolution Using Non-Uniform Segmented Phase-Scrambling Fourier Transform Signals*
Itoh, Yoshiaki	P2.4	*Accurate and Practical Query-By-Example by Using Multiple Deep Learning Models and Frame Compression Methods*
	P5.6	*Ensemble of Transformer and Convolutional Recurrent Neural Network for Improving Discrimination Accuracy in Automatic Chord Recognition*

Itoh, Yoshito	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Itoyama, Katsutoshi	P4.5	*Is the Ideal Ratio Mask Really the Best?---Exploring the Best Extraction Performance and Optimal Mask of Mask-Based Beamformers*
Iwai, Kenta	H3.1	*Radiated Sound Field Reproduction for Surrounding Loudspeaker Array Based on Higher-Order Ambisonics*
	E3.4	*Step Size Control of Shared-Error Normalized Least Mean Square Algorithm for Acoustic Echo and Noise Canceller*
	P5.9	*Speech Quality Improvement Utilizing Out-Of-Focus Areas in Rolling-Shutter Video on Speech Extraction*
Iwasaki, Matsuri	F3.3	*Speech Synthesis Using Ambiguous Inputs from Wearable Keyboards*
Jana, Nanda	P1.8	*FID-RPRGAN-VC: Fréchet Inception Distance Loss Based Region-Wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion*
Jang, Jyh-Shing	P5.10	*Personalized Audio Quality Preference Prediction*
Jang, Won-Dong	G2.2	*Implementation and Analysis on Backpropagating Refinement Scheme for Interactive Image Segmentation*
Jeong, Young-Min	G2.5	*Low-Light Image Enhancement via Distillation of NIR-To-RGB Conversion Knowledge*
Jia, Maoshen	P2.28	*Single Source Zone Detection in the Spherical Harmonic Domain for Multisource Localization*
Jia, Pan	P2.25	*Progressive Multi-Scale Self-Supervised Learning for Speech Recognition*
	P2.26	*Improved Data2vec with Soft Supervised Hidden Unit for Mandarin Speech Recognition*
Jiahao, Yang	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Jiang, Jeng-Ting	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
Jiang, Siyi	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Jiang, Ya	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Jiang, Yufei	B1.2	*Adopting Neural Translation Model in Data Generation for Inverse Text Normalization*
Jonathan, Michael	I3.1	*A QoS Throughput Performance Measurement Comparison Between UGS and BE Services of a Real-Time FPGA Based OFDM Multi-User System Design Implementation*

Josy John, Mary	P3.3	*Compressive Sensing Based Algorithms for Limited-View PAT Image Reconstruction*
Jung, Alexander	D3.5	*Moreau Envelope ADMM for Decentralized Weakly Convex Optimization*
Junhai, Xu	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Kachhi, Aastha	P3.2	*Classification of Normal vs. Pathological Infant Cries Using Morse Wavelets*
Kadlay, Satej	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Kaga, Yosuke	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
	I4.2	*IPFed: Identity Protected Federated Learning for User Authentication*
Kagamiishi, Eri	P3.1	*Remarks on an Optimal Predictive Control Using a Quaternion Neural Network and a Metaheuristics Optimisation Method*
Kai, Atsuhiko	A1.2	*Attention-Based CNN and Relative Phase Feature Modeling for Improved Imagined Speech Recognition*
Kajikawa, Yoshinobu	H3.4	*SFANC with Compensation Filter Based on MEFxDCTLMS Algorithm*
Kamakura, Daichi	F1.2	*CTC2: End-To-End Drum Transcription Based on Connectionist Temporal Classification with Constant Tempo Constraint*
	F1.1	*Joint Drum Transcription and Metrical Analysis Based on Periodicity-Aware Multi-Task Learning*
Kamarudin, Nur Afiqah	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Kamble, Madhu	P5.19	*Quefrency Domain Features with Residual Networks for Spoof Speech Detection*
Kameda, Kaito	P3.22	*Numerical Performance Evaluation of ¥(¥ell_1-¥ell_2¥) Sparse Reconstruction Using Optical Analog Circuit*
Kameoka, Hirokazu	F3.2	*DisC-VC: Disentangled and F0-Controllable Neural Voice Conversion*
	A5.4	*Multiple Sound Source Tracking Based on Generative Modeling and Recursive Bayesian Filtering of Spatial Gradient Spectra*
Kamiyama, Hosana	H5.1	*Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation*
Kanaizuka, Yuma	G1.4	*Directional Neural Connectivity During Robot Mirror Therapy in Patients with Stroke*
Kanayama, Noriaki	G1.5	*Evaluation of Neural Response Recorded Using Scalp EEG in Virtual Reality Environment*
Kaneko, Reina	I2.2	*Marine Snow Removal Benchmarking Dataset*

Kang, Hong-Goo	E1.4	*Consideration of Varying Training Lengths for Short-Duration Speaker Verification*
	P5.20	*PDF-NET: Pitch-Adaptive Dynamic Filter Network for Intra-Gender Speaker Verification*
Kang, Jewon	G2.3	*Implicit Neural Representation for Video Coding Through Progressive Feature Extraction*
Kang, Li-Wei	P1.10	*Learning Single Image Rain Streak Removal Based on Deep Attention Mechanism*
Kano, Manabu	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Kao, Chia-Hao	G4.3	*Transformer-Based Image Compression with Variable Image Quality Objectives*
Karn, Prakash Kumar	P5.16	*Enhancing Retinal Disease Classification with Dual Scale Twin Vision Transformers Using OCT Imaging*
Karnjana, Jessada	P3.12	*Detection and Correction of Defective Relative Humidity Data Collected from the Greenhouse Environment Using Nested Kalman Filters with Standard Deviation Analysis*
	G5.1	*Deepfake-Speech Detection with Pathological Features and Multilayer Perceptron Neural Network*
Katayama, Kaito	P1.4	*Soft-Sensor Construction Method Based on Adaptive Modeling and Transfer Learning for Manufacturing Process Including Maintenance Periods*
Kato, Shota	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Kato, Tsuneo	P4.7	*Transformer-Based Automatic Speech Recognition of Simultaneous Interpretation with Auxiliary Input of Source Language Text*
Kawahara, Hideki	F1.4	*Simultaneous Measurement of Multiple Acoustic Attributes Using Structured Periodic Test Signals Including Music and Other Sound Materials*
Kawakami, Shokei	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
Kawamura, Masaki	F4.2	*Application for Generating Re-Accessible Screenshots of Web Pages Using Histogram Shrinkage*
	F4.1	*Rotation Angle Detection Using a Pilot Signal from Rotated Stego-Image*
Kawano, Rinka	F4.2	*Application for Generating Re-Accessible Screenshots of Web Pages Using Histogram Shrinkage*

	F4.1	*Rotation Angle Detection Using a Pilot Signal from Rotated Stego-Image*
Kawauchi, Hideto	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Ke, Wei-Hsun	G4.2	*Task-Specific Pruning: Efficient Parameter Reduction in Multi-Task Object Detection Models*
Ke, Wenjun	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Ke, Xiaoquan	P5.1	*Jointly Modelling Transcriptions and Phonemes with Optimal Features to Detect Dementia from Spontaneous Cantonese*
Ke, Zhang	C4.4	*MKTformer: Fine-Grained Meter Classification Based on Multi-Modal Knowledge Transfer*
Keigo, Hirakawa	I2.5	*Data Driven Multiband Image Fusion That Preserves Wavelength-Specific Image Features*
Keiji, Yasuda	P4.7	*Transformer-Based Automatic Speech Recognition of Simultaneous Interpretation with Auxiliary Input of Source Language Text*
Khan, Ahmed	F4.5	*HDR Image Watermarking Based on Saliency Detection and Quantization Index Modulation*
Khanh, Nguyen	J1.5	*GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer*
Khysru, Kuntharrgyal Khysru	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Kidani, Shunsuke	H5.2	*Contribution of Modulation Spectral Features for Cross-Lingual Speech Emotion Recognition Under Noisy Reverberant Conditions*
Kilfoyle, Riley	G3.5	*Communication-Efficient Design of Leaning System for Energy Demand Forecasting of Electrical Vehicles*
Kim, Ahyun	P3.11	*Optimal Input Data Search of Deep Learning Model for Cognitive Task Classification Based on NIRS*
Kim, Chang-Su	G2.2	*Implementation and Analysis on Backpropagating Refinement Scheme for Interactive Image Segmentation*
Kim, Jong-Ok	G2.5	*Low-Light Image Enhancement via Distillation of NIR-To-RGB Conversion Knowledge*
Kim, Miseul	P5.20	*PDF-NET: Pitch-Adaptive Dynamic Filter Network for Intra-Gender Speaker Verification*
Kim, Seongjean	A2.5	*Holo-QoI: A Human Factor-Based Dataset and Prediction Framework for Assessing Quality of Interaction in Augmented Reality*
Kim, WooJin	P3.9	*Investigation of the Neuromodulation Effect of tVNS on Electrical Brain Activity*

Kimura, Tomotaka	E5.4	*Approximate Modeling of Malware Diffusion on Wireless Mobile Devices*
Kinoshita, Yuma	B5.3	*Maximization of 2D Cross-Correlation Based on Auxiliary Function Method for Image Alignment*
	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Kishi, Keita	A3.1	*Speech Emotion Recognition by Late Fusion of Linguistic and Acoustic Features Using Deep Learning Models*
Kitagishi, Yuki	H5.1	*Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation*
Kitamori, Aoi	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
Kitamura, Daichi	B3.2	*Blind Source Separation Using Independent Low-Rank Matrix Analysis with Spectrogram-Consistency Regularization*
Kitamura, Tatsuya	F1.4	*Simultaneous Measurement of Multiple Acoustic Attributes Using Structured Periodic Test Signals Including Music and Other Sound Materials*
Kitaoka, Norihide	P4.6	*Language Modeling for Spontaneous Speech Recognition Based on Disfluency Labeling and Generation of Disfluent Text*
	H4.6	*Streaming End-To-End ASR Using CTC Decoder and DRA for Linguistic Information Substitution*
	P5.7	*Construction of Automatic Speech Recognition Model That Recognizes Linguistic Information and Verbal/Non-Verbal Phenomena*
	P5.2	*Combining Multiple End-To-End Speech Recognition Models Based on Density Ratio Approach*
Kiya, Hitoshi	P3.20	*Hindering Adversarial Attacks with Multiple Encrypted Patch Embeddings*
	F4.3	*Domain Adaptation for Efficiently Fine-Tuning Vision Transformer with Encrypted Images*
	B5.3	*Maximization of 2D Cross-Correlation Based on Auxiliary Function Method for Image Alignment*
	H5.3	*Vocal Tract Length Perturbation-Based Pseudo-Speaker Augmentation for Speaker Embedding Learning*
Ko, WooSeok	E1.4	*Consideration of Varying Training Lengths for Short-Duration Speaker Verification*

Kobayashi, Akio	P5.5	*Corpus Construction for Deaf Speakers and Analysis by Automatic Speech Recognition*
Kobayashi, Iori	I2.4	*JPEG Artifact Removal for Hyperspectral Images Based on Spatial-Spectral Regularization*
Kohara, Shunichi	P5.3	*Speech-Emotion Control for Text-To-Speech in Spoken Dialogue Systems Using Voice Conversion and x-Vector Embedding*
Koi Yee, Ng	C2.5	*Can StArtGAN Withstand Image Processing Attacks?*
Kojima, Kazunori	P2.4	*Accurate and Practical Query-By-Example by Using Multiple Deep Learning Models and Frame Compression Methods*
	P5.6	*Ensemble of Transformer and Convolutional Recurrent Neural Network for Improving Discrimination Accuracy in Automatic Chord Recognition*
Komatani, Kazunori	H4.2	*Out-Of-Vocabulary Word Detection in Spoken Dialogues Based on Joint Decoding with User Response Patterns*
	A4.3	*Multi-Modal Emotion Recognition Based on 2D Kernel Density Estimation for Multiple Labels Fusion*
	P4.3	*Flexible Evidence Model to Reduce Uncertainty Mismatch Between Speech Enhancement and ASR Based on Encoder-Decoder Architecture*
Komeiji, Shuji	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Kongprawechnon, Waree	P2.1	*Segmentation Enhancement for Iris Recognition Using Unit Gradient Vectors*
	P3.12	*Detection and Correction of Defective Relative Humidity Data Collected from the Greenhouse Environment Using Nested Kalman Filters with Standard Deviation Analysis*
	G5.1	*Deepfake-Speech Detection with Pathological Features and Multilayer Perceptron Neural Network*
Kosaka, Tetsuo	A3.1	*Speech Emotion Recognition by Late Fusion of Linguistic and Acoustic Features Using Deep Learning Models*
Kosgi, Saiteja	E1.5	*Adversarial Robustness of Mel Based Speaker Recognition Systems*
Krishnasamy, Ganesh	J2.2	*ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation*
Kuh, Anthony	G3.4	*Distributed On-Line Anomaly Detection Using Kernel Methods*
Kumagai, Masaya	E5.1	*Deep Unfolding-Based Distributed MIMO Detection*
Kuo, C.-C. Jay	J2.3	*LSR++: An Efficient and Tiny Model for Image Super-Resolution*
Kuo, Chih-Hung	P1.9	*Deformable Aligned Fusion for Video Super Resolution*
Kuo, Tien-Ying	P1.15	*Old Damaged Photo Recovery with Style Transfer-Based Data Augmentation*
	P1.14	*Analysis of the Interaction Effect on Pruning and Transfer Learning in Model Training*

Kuo, Yu-Yi	P4.15	*Breast Cancer Detection Auxiliary System Leveraging Deep Learning and Mixed Reality*
Kuo-Chen, Li	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Kuribayashi, Minoru	C2.3	*Gait Recognition Scheme Focusing on Operating Characteristics at Feature Points Detected by OpenPose*
	C2.4	*A Study on Eliminating Biased Node in Federated Learning*
	F4.4	*Study on Face Landmark-Based Analysis for Synthetic Media Identification Generated by Adversarial Generative Networks*
	F4.5	*HDR Image Watermarking Based on Saliency Detection and Quantization Index Modulation*
Kyochi, Seisuke	D3.4	*Convolutional Multidimensional Amplitude Spectrum Nuclear Norm for Frequency-Domain Robust Principal Component Analysis*
La, Hyoung Sul	C5.4	*Cross-Domain Sound Recognition for Efficient Underwater Data Analysis*
Lai, Chin-Feng	P2.14	*Asynchronous Federated Learning Method in the 6G Massive Multimedia Transmission System*
Lai, Chun-Liang	C3.5	*Integration of Contextual Information with YOLOv5 for Detecting Tracheobronchial Foreign Objects*
Lai, Hong-Lun	E5.5	*Impacts of 5G-TDD Time Slot Configurations on the Downlink and Uplink Data Rates*
Lai, Hsin-Hung	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
Lai, Junjie	P3.17	*Efficient Incremental Text-To-Speech on GPUs*
Lai, KuanTing	H2.4	*Light-Weight Zero-Reference-Based Image Enhancement for Low-Light Images*
Lai, Shang-Hong	P3.13	*Pedestrian Crossing Intention Prediction with Multi-Modal Transformer-Based Model*
Lai, Wen-Ping	E5.5	*Impacts of 5G-TDD Time Slot Configurations on the Downlink and Uplink Data Rates*
Lai, Ying-Hui	E3.1	*An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids*
Lam, Kin-Man	P2.2	*FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking*

Lari, Ehsan	G3.2	*Continual Local Updates for Federated Learning with Enhanced Robustness to Link Noise*
Lau, Phooi Yee	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Law, Ngai Fong	P4.1	*Word Encoding for Word-Looking DGA-Based Botnet Classification*
Le, Ngoc-Hoang-Lam	G4.4	*From Synthetic to Real: Enhancing Deep Learning Models with Generative Adversarial Networks for Efficient Data Utilization in Automatic Retail Stores*
Le, Nhat Minh	A3.4	*An Automatic Pipeline for Building Emotional Speech Dataset*
Le, Thanh-Dat	J3.4	*Investigating the Role of Human Action Detector in Visual-Guide Audio Source Separation System*
Le, Thi-Lan	J1.3	*Accurate Continuous Action and Gesture Recognition Method Based on Skeleton and Sliding Windows Techniques*
	J3.4	*Investigating the Role of Human Action Detector in Visual-Guide Audio Source Separation System*
Le, Trung-Hieu	J1.5	*GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer*
Le, Viet-Duc	J1.3	*Accurate Continuous Action and Gesture Recognition Method Based on Skeleton and Sliding Windows Techniques*
Lebedeva, Irina	D2.1	*Interpretable Image Recognition in Hyperbolic Space*
Lee, Akinobu	F3.4	*Accent-Preserving Voice Conversion Between Native-Nonnative Speakers for Second Language Learning*
Lee, Chaewon	G2.2	*Implementation and Analysis on Backpropagating Refinement Scheme for Interactive Image Segmentation*
Lee, Chin-Hui	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Lee, Chin-Wei	F5.4	*Application of Wafer Defect Pattern Classification Model in the Semiconductor Industry*
Lee, Chul	G2.4	*Deep Unfolded Underwater Image Enhancement Based on Extreme Channels Prior*
Lee, Hyung-Tak	P3.11	*Optimal Input Data Search of Deep Learning Model for Cognitive Task Classification Based on NIRS*
Lee, Jihoo	G2.3	*Implicit Neural Representation for Video Coding Through Progressive Feature Extraction*

Lee, Ming Zhan	B4.4	*Point Cloud Inpainting Based on Delaunay Triangulation*
	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
Lee, Ming-Sui	J2.3	*LSR++: An Efficient and Tiny Model for Image Super-Resolution*
Lee, Sanghoon	A2.5	*Holo-QoI: A Human Factor-Based Dataset and Prediction Framework for Assessing Quality of Interaction in Augmented Reality*
Lee, Seung-Hwan	P3.10	*Gender-Specific EEG Feature-Based Diagnostic Tool for Male Depression Based on Machine-Learning Techniques*
Lee, Shi-wook	P2.4	*Accurate and Practical Query-By-Example by Using Multiple Deep Learning Models and Frame Compression Methods*
Lee, Shuai	P1.25	*Hiding Patient Information in Medical Images: A High-Capacity and Reversible Hiding Algorithm for E-Healthcare*
Lee, Sin-Hong	P1.9	*Deformable Aligned Fusion for Video Super Resolution*
Lee, Su-Ling	D5.1	*Distributed Computation of Heat Kernel Smoothing Using Series Expansion Method*
Lee, Sue Han	D2.6	*How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?*
	D2.3	*MoMo Strategy: Learn More from More Mistakes*
	D2.4	*Unveiling Robust Feature Spaces: Image vs. Embedding-Oriented Approaches for Plant Disease Identification*
Lee, Wei-Min	P2.10	*Improving Regularization of Deep Learning Models in Fundus Analysis*
Lee, Yi-Ju	I5.4	*Impact of the Weighted Loss Function on the Innovative CMAQ-CNN PM2.5 Forecasting Model*
Lee, Yun-Sung	P3.9	*Investigation of the Neuromodulation Effect of tVNS on Electrical Brain Activity*
Lei, Xuejing	J2.3	*LSR++: An Efficient and Tiny Model for Image Super-Resolution*
Levita, Liat	P1.6	*Predicting Outcomes of Cognitive Behavioral Therapy for Depression Using Data Driven Approaches*
Li, Bingjie	A1.1	*Mixed Emotion Recognition Based on EEG Signals*
Li, Chengfei	P4.22	*Hybrid Syllable and Character Representations for Mandarin ASR*
Li, Haizhou	P4.19	*Zero-Shot Multi-Speaker Accent TTS with Limited Accent Data*
Li, Hao	B2.2	*Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-To-End CNN with Fine-Tuning*
Li, Haowen	P4.24	*DoA Estimation of Room Reflections Using NN-Based MUSIC Algorithm*
Li, Jiamu	P2.17	*STrack: Velocity Estimation Using Single Antenna WiFi Devices*
Li, Jin	E1.1	*Modeling Suprasegmental Information Using Finite Difference Network for End-To-End Speaker Verification*
Li, Jing	P4.23	*Enhancing Whisper Model for Pronunciation Assessment with Multi-Adapters*

Li, Kai	P4.2	*Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection*
Li, Linghan	P1.23	*CoA-DLinkNet: Connectivity-Enhanced Dual-Branch Road Extraction Network Based on D-LinkNet*
Li, Lixiang	P4.13	*Improved One-Class Learning for Voice Spoofing Detection*
Li, Ming	D1.5	*Low-Complexity Multi-Channel Speaker Extraction with Pure Speech Cues*
Li, Ming-Hsu	I5.2	*Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind*
Li, Mingtao	B2.2	*Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-To-End CNN with Fine-Tuning*
Li, Na	P1.22	*On Uncertainty Principles for Lowband Graph Signals*
Li, Pengcheng	P2.26	*Improved Data2vec with Soft Supervised Hidden Unit for Mandarin Speech Recognition*
Li, Qinglong	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Li, Rongjun	C4.1	*Dictionary-Driven Chinese ASR Entity Correction with Controllable Decoding*
Li, Rui	P4.23	*Enhancing Whisper Model for Pronunciation Assessment with Multi-Adapters*
Li, Sheng	C3.3	*Identifying the Style of Chatting*
Li, Sixia	H5.2	*Contribution of Modulation Spectral Features for Cross-Lingual Speech Emotion Recognition Under Noisy Reverberant Conditions*
Li, Taihao	B1.1	*GRALA: Modeling Social Information for Microblog Sentiment Analysis from the View of Balancing Sparsity and Smoothness of Social Contexts*
	A1.1	*Mixed Emotion Recognition Based on EEG Signals*
	P1.21	*An Multi-Evidence Fusion Based on C-Distance with Uncertain Reasoning for Classification*
	P2.30	*Few Shot Learning Guided by Emotion Distance for Cross-Corpus Speech Emotion Recognition*
	A2.1	*CH-MEAD: A Chinese Multimodal Conversational Emotion Analysis Dataset with Fine-Grained Emotion Taxonomy*
	P2.16	*Poisoning Attacks Against Gait-Based Identity Recognition*
	P2.19	*Complex Feature Information Enhanced Speech Emotion Recognition*

	P2.27	*Investigation of Ensemble of Self-Supervised Models for Speech Emotion Recognition*
	D2.1	*Interpretable Image Recognition in Hyperbolic Space*
Li, Tao	D1.3	*CASA-Net: Cross-Attention and Self-Attention for End-To-End Audio-Visual Speaker Diarization*
Li, Te	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
	P1.28	*Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion*
Li, Wenpan	P1.29	*Multi-Frequency Feature Enhancement for Multi-Granularity Visual Classification*
Li, WuHao	P2.6	*Tone Labeling by Deep Learning-Based Tone Recognizer for Mandarin Speech*
Li, Xiaolong	P3.29	*A Print-Scan-Resilient Watermarking Scheme for Trademark Images*
Li, Xiaoxu	C4.5	*Feature Augmentation Reconstruction Network for Few-Shot Image Classification*
	C4.6	*Dual Feature Reconstruction Network Based on Few-Shot Image Classification*
Li, Yangfu	P5.30	*UniVR: A Unified Framework for Pitch-Shifted Voice Restoration in Speaker Identification*
Li, YaoFeng	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
Li, Yaqing	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Li, Yi-Mei	H1.4	*Deep-Learning-Based Lattice Reduction Preprocessing for Time-Correlated MIMO Systems*
Li, Yongwei	P4.25	*Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition*
Li, Yung-Hui	F5.3	*On the Optimal Self-Supervised Multi-Fault Detector for Temperature Sensor Data*
Li, Zhen	C4.5	*Feature Augmentation Reconstruction Network for Few-Shot Image Classification*
Li, Zhidu	A4.2	*Structural Quality Assured Global Optimization for CTU-Level Rate Control of Screen Content Coding*
Li, Zhifei	P5.23	*Multi-Granularity Semantic and Acoustic Stress Prediction for Expressive TTS*
Liang, Xiujian	P3.27	*StyleStegan: Leak-Free Style Transfer Based on Feature Steganography*

Liang, Zechen	B1.3	*Mismatched Semi-Supervised Learning with Feature Similarity Consistency*
Liao, Chen	P1.30	*Improving Aspect Sentiment Classification via Retrieving from Training Data*
Liao, Shangdi	P3.23	*Assessing the Effects of Filtering Processing on Pulse Wave Transit Time Measured by Photoplethysmography from Earlobe*
Liao, Sheng-Siang	P2.14	*Asynchronous Federated Learning Method in the 6G Massive Multimedia Transmission System*
Liao, Yi-Kai	P1.11	*A Transformer-Based Framework for Tiny Object Detection*
Liao, Yu-Cheng	BS1.6	*Mask2Hand: Learning to Predict the 3D Hand Pose and Shape from Shadow*
Liao, Yuan-Fu	F5.4	*Application of Wafer Defect Pattern Classification Model in the Semiconductor Industry*
Lie, Wen-Nung	G2.6	*3D Human Skeleton Estimation from Single RGB Image Based on Fusion of Predicted Depths from Multiple Virtual-Viewpoints*
	G4.6	*Sparse Tensor-Based Point Cloud Attribute Compression Using Augmented Normalizing Flows*
Liew, Hui Sze	D2.2	*Low-Light is More than Darkness: An Empirical Study on Illumination Types and Enhancement Methods*
Liew, Sea Ran Cleon	P4.1	*Word Encoding for Word-Looking DGA-Based Botnet Classification*
Lim, Chern Hong	H2.6	*CapFormer: A Space-Time Video Description Model Using Joint-Attention Transformer*
	P3.4	*Towards AST-LLDs for the Analysis of Depression in Speech Signals*
Lim, Hyungseob	P5.20	*PDF-NET: Pitch-Adaptive Dynamic Filter Network for Intra-Gender Speaker Verification*
Lim, Khai Yin	P3.14	*Revolutionizing Formative Assessment in STEM Fields: Leveraging AI and NLP Techniques*
Lim, Mei Kuan	D2.5	*Facial Expression Recognition as Markers of Depression*
Lin, Binghuai	P5.25	*Multi-Accent Pronunciation Assessment Based on Domain Adversarial Training*
Lin, Che-Hsuan	P4.15	*Breast Cancer Detection Auxiliary System Leveraging Deep Learning and Mixed Reality*
Lin, Chen-Yen	H1.3	*QoS-Aware Downlink Beamforming for Joint Transmission in Multi-Cell Networks*
Lin, Chia Hui	BS1.6	*Mask2Hand: Learning to Predict the 3D Hand Pose and Shape from Shadow*
Lin, Chia-Hsiang	H2.1	*GNN-Based Small-Data Learning with Area-Control Mechanism for Hyperspectral Satellite Change Detection*

Lin, Chia-Yi	BS1.1	*Enhancing Real-Time Semantic Segmentation with Textual Knowledge of Pre-Trained Vision-Language Model: A Lightweight Approach*
Lin, Chih-Yang	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Lin, Chun-An	E4.5	*SOAda-YOLOR: Small Object Adaptive YOLOR Algorithm for Road Object Detection*
Lin, Esther Ching-Lan	G5.2	*Temporal and Type Correlation in Digital Phenotyping for Bipolar Disorder State Prediction Using Multitask Self-Supervised Learning*
Lin, Fang-Pang	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*
Lin, Feng	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Lin, Gong-Si	P1.11	*A Transformer-Based Framework for Tiny Object Detection*
Lin, Guo-Shiang	H2.4	*Light-Weight Zero-Reference-Based Image Enhancement for Low-Light Images*
Lin, Hsuan	I2.5	*Data Driven Multiband Image Fusion That Preserves Wavelength-Specific Image Features*
Lin, Jia-Yi	P2.15	*The Development of an AI-Assisted Diagnosis System for Adult Glioma Subtyping Prediction*
Lin, Jinwei	P3.5	*ecVoice: Audio Text Extraction Optimization of Video Based on Idioms Similarity Replacement*
Lin, Li	D1.3	*CASA-Net: Cross-Attention and Self-Attention for End-To-End Audio-Visual Speaker Diarization*
Lin, Mu-Feng	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
Lin, Szu-Yin	P4.15	*Breast Cancer Detection Auxiliary System Leveraging Deep Learning and Mixed Reality*
Lin, Ting-Lan	B4.4	*Point Cloud Inpainting Based on Delaunay Triangulation*
	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
Lin, Tzu-Han	P4.9	*Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods*
Lin, Tzu-Hsuan	H2.1	*GNN-Based Small-Data Learning with Area-Control Mechanism for Hyperspectral Satellite Change Detection*
Lin, Tzu-Po	G4.6	*Sparse Tensor-Based Point Cloud Attribute Compression Using Augmented Normalizing Flows*
Lin, Wei-Chi	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
Lin, Xiaodan	P5.30	*UniVR: A Unified Framework for Pitch-Shifted Voice Restoration in Speaker Identification*

Lin, Yi-Xing	F5.5	*Question Answering System Based on Pre-Training Model and Retrieval Reranking for Industry 4.0*
Lin, Yu-Chun	P5.10	*Personalized Audio Quality Preference Prediction*
Lin, Yuan-Jin	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
Lin, Yuan-Jin	B4.1	*3D Point Cloud Denoising Based on Color Attribute*
Lin, Yuan-Jin	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Lin, Zheyuan	P1.28	*Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion*
Ling, Tongtao	P1.30	*Improving Aspect Sentiment Classification via Retrieving from Training Data*
Linh, Nguyen	J3.5	*A Combination of Time and Frequency Synchronization with Doppler Compensation in Coded OFDM-Based UWA Systems*
Linh Trung, Nguyen	J1.2	*On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems*
Liu, Bingshan	P3.27	*StyleStegan: Leak-Free Style Transfer Based on Feature Steganography*
Liu, Chia-Chuan	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*
Liu, Chuan	P3.17	*Efficient Incremental Text-To-Speech on GPUs*
Liu, Kuan-Hsien	E4.5	*SOAda-YOLOR: Small Object Adaptive YOLOR Algorithm for Road Object Detection*
	E4.6	*Badminton Self-Training System Based on Virtual Reality*
Liu, Kuang-Hao (Stanley)	H1.3	*QoS-Aware Downlink Beamforming for Joint Transmission in Multi-Cell Networks*
Liu, Mingkuan	P2.23	*Multi-Task Piano Transcription with Local Relative Time Attention*
Liu, Peiqi	P4.20	*Speech Enhancement with Multi-Granularity Vector Quantization*
Liu, Ruibo	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
Liu, Tan	P2.25	*Progressive Multi-Scale Self-Supervised Learning for Speech Recognition*
Liu, TehSin	P2.6	*Tone Labeling by Deep Learning-Based Tone Recognizer for Mandarin Speech*
Liu, Tsung-Jung	E4.5	*SOAda-YOLOR: Small Object Adaptive YOLOR Algorithm for Road Object Detection*
Liu, Xuefei	P4.25	*Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition*

Liu, Yi-Wen	H1.5	*Utilizing Unlabeled Data and Synthetic Data for Bird Sound Detection: Consistency Training, Mean Teacher, and Domain Adaptation Techniques*
	P4.9	*Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods*
Liu, Yu	P5.28	*Feature Selection Based on Clonal Selection Algorithm for Image Steganalysis*
Liu, Yu-Lin	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
	B4.4	*Point Cloud Inpainting Based on Delaunay Triangulation*
	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Liu, Zixiao	B4.5	*Dense Three-Dimensional Color Reconstruction for Large-Scale Outdoor Scenes*
Lo, Chun-Huang	E5.3	*A UAV Indoor Obstacle Avoidance System Based on Deep Reinforcement Learning*
Loh, Yuen Peng	C2.5	*Can StArtGAN Withstand Image Processing Attacks?*
	D2.2	*Low-Light is More than Darkness: An Empirical Study on Illumination Types and Enhancement Methods*
Lourenco Correa, Iago	F3.4	*Accent-Preserving Voice Conversion Between Native-Nonnative Speakers for Second Language Learning*
Lu, Cheng-Yu	H2.5	*Class-Wise Self-Paced Self-Training for Semi-Supervised Image Classification*
Lu, Chung-Li	P3.16	*A Lightweight Speaker Verification Model for Edge Device*
	P5.12	*Regression-Based Sound Event Detection with Semi-Supervised Learning*
Lu, Da-Wen	P2.10	*Improving Regularization of Deep Learning Models in Fundus Analysis*
Lu, Hao-Cheng	B4.3	*Identification of Victims Wearing Vibrant Clothing Using MATLAB*
Lu, Hsin-Tse	P2.12	*Anomaly Detection of Rolling Element Bearings Using Domain Adaptation*
Lu, Ligang	G3.5	*Communication-Efficient Design of Leaning System for Energy Demand Forecasting of Electrical Vehicles*
Luo, Bin	P4.17	*MAEDefense: An Effective Masked AutoEncoder Defense Against Adversarial Attacks*
	P4.16	*Efficient Reversible Data Hiding for 3D Mesh Models Based on Multi-LSB Substitution and Ring-Prediction*

Luo, Ge	C3.3	*Identifying the Style of Chatting*
Luo, Rui	P3.25	*Retinex-Based Low-Light Image Enhancement*
Luo, Weiqi	P3.30	*AI-Generated Image Detection Using a Cross-Attention Enhanced Dual-Stream Network*
Luo, Zhaojie	A4.3	*Multi-Modal Emotion Recognition Based on 2D Kernel Density Estimation for Multiple Labels Fusion*
Luong, Thien	J1.4	*Transformer-Based Deep Learning Detector for Dual-Mode Index Modulation 3D-OFDM*
	J1.1	*After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals*
Luong Thi, Theu	J3.2	*Imaging Ultrasound Scattering Targets Using Density-Enhanced Chaotic Compressive Sampling*
Luu, Thien	J1.1	*After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals*
Lv, Zhao	A1.1	*Mixed Emotion Recognition Based on EEG Signals*
Lyu, Wanli	P1.27	*High-Quality Triggers Based Fragile Watermarking for Optical Character Recognition Model*
	P4.17	*MAEDefense: An Effective Masked AutoEncoder Defense Against Adversarial Attacks*
	P4.16	*Efficient Reversible Data Hiding for 3D Mesh Models Based on Multi-LSB Substitution and Ring-Prediction*
M, Venkatesan	C1.1	*MSDF-Net: A Multi-Scale Deep Fusion Network with Dilated Convolutions for Cloud Removal from Sentinel-2 Imagery*
Ma, Bin	A5.1	*Analysis of Speech Separation Performance Degradation on Emotional Speech Mixtures*
Ma, Fei	P4.28	*Sound Field Estimation Around a Rigid Sphere with Physics-Informed Neural Network*
Ma, Yuchen	C3.3	*Identifying the Style of Chatting*
Ma, Zhanyu	P1.29	*Multi-Frequency Feature Enhancement for Multi-Granularity Visual Classification*
Maben, Leander	E3.3	*Study of Generative Adversarial Networks for Noisy Speech Simulation from Clean Speech*
Machhar, Shrey	E1.3	*Exploring Residual Cepstral Features for Spoken Language Identification*
Machizawa, Maro	G1.5	*Evaluation of Neural Response Recorded Using Scalp EEG in Virtual Reality Environment*
	B2.3	*Decoding Time-Course of Saliency Network of fMRI Signals by EEG Signals Using Optimized Forward Variable Selection: A Concurrent EEG-fMRI Study*

Maeda, Yoshihiro	I2.1	*Local Contrast Enhancement with Multi-Scale Filtering*
Mahdy, Abid	P3.18	*Dual-Path Residual Attention Convolution Networks for Color-Embedded-Grayscale Image*
Mai, Trang	J1.1	*After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals*
Mai, Truong	G2.4	*Deep Unfolded Underwater Image Enhancement Based on Extreme Channels Prior*
Mak, Man Wai	E1.1	*Modeling Suprasegmental Information Using Finite Difference Network for End-To-End Speaker Verification*
	P5.1	*Jointly Modelling Transcriptions and Phonemes with Optimal Features to Detect Dementia from Spontaneous Cantonese*
Mak, Man-wai	H5.5	*Cross-Domain Adaptation in Distance Space for Speaker Verification*
Makino, Shoji	B3.3	*Moving Interference Speaker Removal Using Geometrically Constrained Independent Vector Analysis*
Makino, Shoji	E3.2	*On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering*
Makino, Shoji	E3.5	*Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer*
	A5.5	*Spatially-Regularized Switching Independent Vector Analysis*
	A5.2	*Geometrically Constrained Blind Moving Source Extraction Based on Constant Separation Vector and Auxiliary Function Technique*
Manabe, Takahiro	G1.4	*Directional Neural Connectivity During Robot Mirror Therapy in Patients with Stroke*
Mao, Jiarong	P1.26	*A Visually Meaningful Image Encryption Algorithm with Attention Mechanism and Artificial Bee Colony Optimization*
Masato, Akagi	F3.5	*Increasing Speech Intelligibility by Mimicking Professional Announcers' Voices and Its Physical Correlates*
Masuda, Kohta	H4.4	*Multi-Self-Supervised Learning Model-Based Throat Microphone Speech Recognition*
Masuyama, Yoshiki	P2.5	*Fundamental Frequency Estimation Based on Finite-Order Harmonic Constraint Differential Equation*
Matoba, Satoaki	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Matsumoto, Shinya	A1.5	*Detection Model of Sister Chromatid Cohesion Defects Based on Vision Transformer*
Matsuoka, Ryo	I2.4	*JPEG Artifact Removal for Hyperspectral Images Based on Spatial-Spectral Regularization*

	P3.7	*A Study on Counter-Movement Detection on Dumbbell Curl Using Convolutional Autoencoder*
MaungMaung, AprilPyone	P3.20	*Hindering Adversarial Attacks with Multiple Encrypted Patch Embeddings*
Mawalim, Candy	D4.2	*Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids*
	P4.2	*Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection*
McElwain, Nancy	P5.17	*Classification of Infant Sleep/Wake States: Cross-Attention Among Large Scale Pretrained Transformer Networks Using Audio, ECG, and IMU Data*
Meam, Limhourlaurent	P2.1	*Segmentation Enhancement for Iris Recognition Using Unit Gradient Vectors*
Meng, Edwin Tiong Kwong	P4.15	*Breast Cancer Detection Auxiliary System Leveraging Deep Learning and Mixed Reality*
Meng, Helen	P2.29	*Robust Representation Learning for Speech Emotion Recognition with Moment Exchange*
	P5.1	*Jointly Modelling Transcriptions and Phonemes with Optimal Features to Detect Dementia from Spontaneous Cantonese*
Mijit, Ablimit	P4.30	*Target Speaker Extraction with Attention Enhancement and Gated Fusion Mechanism*
Minematsu, Nobuaki	P5.8	*Exploring Isolated Musical Notes as Pre-Training Data for Predominant Instrument Recognition in Polyphonic Music*
Ming-Jiang, Wang	P2.24	*Real and Imaginary Part Interaction Network for Monaural Speech Enhancement and De-Reverberation*
Mirzaeifard, Reza	D3.5	*Moreau Envelope ADMM for Decentralized Weakly Convex Optimization*
	G3.1	*Robust Networked Federated Learning for Localization*
Misawa, Sota	B3.2	*Blind Source Separation Using Independent Low-Rank Matrix Analysis with Spectrogram-Consistency Regularization*
Mitsuhashi, Takumi	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Miura, Hideyoshi	E5.4	*Approximate Modeling of Malware Diffusion on Wireless Mobile Devices*
Miura, Shumpei	H3.1	*Radiated Sound Field Reproduction for Surrounding Loudspeaker Array Based on Higher-Order Ambisonics*
Miyakoshi, Makoto	G1.5	*Evaluation of Neural Response Recorded Using Scalp EEG in Virtual Reality Environment*

Miyamoto, Ryuji	P3.21	*Implementation of PLIM on 429MHz LoRa/FSK with Improved Conversion Table*
Miyazaki, Noboru	H5.1	*Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation*
Mizumachi, Mitsunori	F1.4	*Simultaneous Measurement of Multiple Acoustic Attributes Using Structured Periodic Test Signals Including Music and Other Sound Materials*
Mo, Kaien	E3.2	*On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering*
Mohan, Alkha	C1.1	*MSDF-Net: A Multi-Scale Deep Fusion Network with Dilated Convolutions for Cloud Removal from Sentinel-2 Imagery*
Mohd Talib, Muhammad Syukri	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Momma, Toshifumi	P5.6	*Ensemble of Transformer and Convolutional Recurrent Neural Network for Improving Discrimination Accuracy in Automatic Chord Recognition*
Mori, Daiki	P5.2	*Combining Multiple End-To-End Speech Recognition Models Based on Density Ratio Approach*
Moriya, Takehiro	B3.1	*Modified Parametric Multichannel Wiener Filter for Low-Latency Enhancement of Speech Mixtures with Unknown Number of Speakers*
Morsalin, S M Salahuddin	C3.2	*Automated Carina Detection in Chest X-Ray Images Using Non-Overlapping and Cross-Squeeze Convolutional Neural Networks*
Moussa, Mahamat	H2.6	*CapFormer: A Space-Time Video Description Model Using Joint-Attention Transformer*
Mukhtar, Roshana	C3.5	*Integration of Contextual Information with YOLOv5 for Detecting Tracheobronchial Foreign Objects*
Mulyawan, Rahmat	I3.1	*A QoS Throughput Performance Measurement Comparison Between UGS and BE Services of a Real-Time FPGA Based OFDM Multi-User System Design Implementation*
Munakata, Hokuto	H4.2	*Out-Of-Vocabulary Word Detection in Spoken Dialogues Based on Joint Decoding with User Response Patterns*
Murayama, Kazuaki	D3.3	*Asymptotic Estimation Performance of Linear Regression Model with Sparse Bayesian Learning as Both Samples and Signals Approach Infinity*
Nadhif, Abrar	P3.18	*Dual-Path Residual Attention Convolution Networks for Color-Embedded-Grayscale Image*
Nagamori, Teru	F4.3	*Domain Adaptation for Efficiently Fine-Tuning Vision Transformer with Encrypted Images*

Naganuma, Kazuki	I2.3	*Cross-Frame Foreground Structural Similarity Modeling by Convolutional Sparse Representation*
Nagappan, Sidharrth	P3.4	*Towards AST-LLDs for the Analysis of Depression in Speech Signals*
Nagatoshi, Takuya	A2.4	*Human Flow Measurement System Using Floor Estimation of Depth Images for Low-End IoT Devices*
Naidu, Surya	P1.1	*Fatigue Classification and Onset Estimation Using Surface EMG Signals During Strength Training*
Naiki, Shota	H3.1	*Radiated Sound Field Reproduction for Surrounding Loudspeaker Array Based on Higher-Order Ambisonics*
Nakadai, Kazuhiro	P4.5	*Is the Ideal Ratio Mask Really the Best?---Exploring the Best Extraction Performance and Optimal Mask of Mask-Based Beamformers*
Nakagawa, Hiroshi	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
Nakagawa, Seiichi	A1.2	*Attention-Based CNN and Relative Phase Feature Modeling for Improved Imagined Speech Recognition*
Nakai-Kasai, Ayano	E5.1	*Deep Unfolding-Based Distributed MIMO Detection*
Nakamura, Eita	F1.1	*Joint Drum Transcription and Metrical Analysis Based on Periodicity-Aware Multi-Task Learning*
	F1.2	*CTC2: End-To-End Drum Transcription Based on Connectionist Temporal Classification with Constant Tempo Constraint*
	A2.2	*Evolutionary Analysis and Cultural Transmission Models of Color Style Distributions in Painting Arts*
	BS1.5	*Audio-To-Score Singing Transcription Based on Joint Estimation of Pitches, Onsets, and Metrical Positions with Tatum-Level CTC Loss*
	E4.3	*Multimodal Multifaceted Music Emotion Recognition Based on Self-Attentive Fusion of Psychology-Inspired Symbolic and Acoustic Features*
Nakamura, Wakako	A1.4	*Goodness of Fit to the Convolution Model of fMRI Data and Determination of the Regularization Parameter*
Nakamura, Wataru	I4.1	*A Biometric Signature Scheme with Template Protection and Authenticated Sample Recoverability*
Nakanishi, Toshiyuki	G1.3	*Prediction Model of Postoperative Pain Exacerbation Using a Wearable Electrocardiogram Sensor*
Nakano, Hayata	P5.9	*Speech Quality Improvement Utilizing Out-Of-Focus Areas in Rolling-Shutter Video on Speech Extraction*
Nakatani, Tomohiro	B3.1	*Modified Parametric Multichannel Wiener Filter for Low-Latency Enhancement of Speech Mixtures with Unknown Number of Speakers*
	A5.5	*Spatially-Regularized Switching Independent Vector Analysis*

Nakaya, Daichi	C2.1	*A Reversible Image Processing Method for Color Tone Control Using Data Hiding*
Nakayama, Masato	P5.4	*Narrow-Edged Acoustical Beamforming Utilizing Phase Inversion for Frequency Modulation-Based Parametric Array Loudspeaker*
Ng, Dianwen	A5.1	*Analysis of Speech Separation Performance Degradation on Emotional Speech Mixtures*
Ng, Yi Pei	C1.3	*Addressing Distortion in Social Distance Monitoring: Calibration for Accurate Distance Calculation*
Nghiem, Thi-Lich	J1.3	*Accurate Continuous Action and Gesture Recognition Method Based on Skeleton and Sliding Windows Techniques*
	J3.4	*Investigating the Role of Human Action Detector in Visual-Guide Audio Source Separation System*
Nguyen, Hieu	J1.1	*After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals*
Nguyen, Hieu	P5.13	*Proportionate NLMS with Variable Step-Size for Adaptive Feedback Cancellation in Hearing Aids*
Nguyen, Huy	C2.2	*Image-Text Out-Of-Context Detection Using Synthetic Multimodal Misinformation*
Nguyen, Quoc Khuong	J3.5	*A Combination of Time and Frequency Synchronization with Doppler Compensation in Coded OFDM-Based UWA Systems*
Nguyen, Thanh-Thu	J3.1	*1M Parameters are Enough? A Lightweight CNN-Based Model for Medical Image Segmentation*
Nguyen, Trung-Hieu	J3.4	*Investigating the Role of Human Action Detector in Visual-Guide Audio Source Separation System*
Nguyen, Trungtan	J1.4	*Transformer-Based Deep Learning Detector for Dual-Mode Index Modulation 3D-OFDM*
Nguyen Chi, Thanh	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*
Nian, Zhaoxu	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Niimi, Michiharu	A2.3	*Ultimatelink Between Characters Having a Certain Meaning in Physical Space to URL in Cyberspace with Robust Print and Scan*
	A2.4	*Human Flow Measurement System Using Floor Estimation of Depth Images for Low-End IoT Devices*
	J2.4	*High-Quality Font Generation Based on StyleGAN2 and FSFont Font Generation Model*
Niimura, Yoshiki	A1.2	*Attention-Based CNN and Relative Phase Feature Modeling for Improved Imagined Speech Recognition*

Niizumi, Daisuke	E3.5	*Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer*
	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Nishida, Masafumi	H4.4	*Multi-Self-Supervised Learning Model-Based Throat Microphone Speech Recognition*
Nishikawa, Kiyoshi	A1.5	*Detection Model of Sister Chromatid Cohesion Defects Based on Vision Transformer*
Nishimura, Masafumi	H4.4	*Multi-Self-Supervised Learning Model-Based Throat Microphone Speech Recognition*
Nishimura, Ryota	P4.6	*Language Modeling for Spontaneous Speech Recognition Based on Disfluency Labeling and Generation of Disfluent Text*
Nishioji, Kenichi	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Nishiura, Takanobu	H3.1	*Radiated Sound Field Reproduction for Surrounding Loudspeaker Array Based on Higher-Order Ambisonics*
	E3.4	*Step Size Control of Shared-Error Normalized Least Mean Square Algorithm for Acoustic Echo and Noise Canceller*
	P5.9	*Speech Quality Improvement Utilizing Out-Of-Focus Areas in Rolling-Shutter Video on Speech Extraction*
	P5.4	*Narrow-Edged Acoustical Beamforming Utilizing Phase Inversion for Frequency Modulation-Based Parametric Array Loudspeaker*
Niu, Shutong	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Nobukawa, Sou	P1.2	*P300 Event-Related Potential in Perception of Multiple Traffic Objects During Vehicle Driving*
	P5.18	*Dynamic Characteristics of Electroencephalogram Reflecting Driving-Experience-Dependent Performance Using Microstates*
Noguchi, Jun	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Nojima, Takashi	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Novamizanti, Ledya	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Nur Hakim, Farhan	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Ogata, Jun	H4.4	*Multi-Self-Supervised Learning Model-Based Throat Microphone Speech Recognition*

Ogawa, Atsunori	P4.6	*Language Modeling for Spontaneous Speech Recognition Based on Disfluency Labeling and Generation of Disfluent Text*
	H5.1	*Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation*
	P5.2	*Combining Multiple End-To-End Speech Recognition Models Based on Density Ratio Approach*
Ogawa, Atunori	H4.6	*Streaming End-To-End ASR Using CTC Decoder and DRA for Linguistic Information Substitution*
Ogura, Masaki	A1.3	*Manipulation of Neuronal Network Firing Patterns Using Temporal Deep Unfolding-Based MPC*
Ohishi, Yasunori	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Ohki, Tetsushi	I4.5	*Development of a Robust Ear Recognition Algorithm Using Planar Approximation*
Ohno, Shuichi	D3.2	*¥(¥Delta¥Sigma¥) Modulators for Discrete-Time Closed Loop Control Systems with Quantization and Saturation*
Ohta, Kengo	P4.6	*Language Modeling for Spontaneous Speech Recognition Based on Disfluency Labeling and Generation of Disfluent Text*
	P5.2	*Combining Multiple End-To-End Speech Recognition Models Based on Density Ratio Approach*
Okada, Shogo	H5.2	*Contribution of Modulation Spectral Features for Cross-Lingual Speech Emotion Recognition Under Noisy Reverberant Conditions*
Okubo, Kan	A1.5	*Detection Model of Sister Chromatid Cohesion Defects Based on Vision Transformer*
Ong, Huey Fang	C1.4	*Unbiased Decision-Making Framework in Long-Video Macro & Micro-Expression Spotting*
Ong, Simying	C2.5	*Can StArtGAN Withstand Image Processing Attacks?*
	D2.2	*Low-Light is More than Darkness: An Empirical Study on Illumination Types and Enhancement Methods*
Onishi, Masaki	C5.1	*DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection*
Ono, Kentaro	B2.3	*Decoding Time-Course of Saliency Network of fMRI Signals by EEG Signals Using Optimized Forward Variable Selection: A Concurrent EEG-fMRI Study*
Ono, Nobutaka	P2.5	*Fundamental Frequency Estimation Based on Finite-Order Harmonic Constraint Differential Equation*
	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*

	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Ono, Shunsuke	I2.3	*Cross-Frame Foreground Structural Similarity Modeling by Convolutional Sparse Representation*
	I2.4	*JPEG Artifact Removal for Hyperspectral Images Based on Spatial-Spectral Regularization*
Ono, Yumie	G1.4	*Directional Neural Connectivity During Robot Mirror Therapy in Patients with Stroke*
Oshio, Miki	H4.2	*Out-Of-Vocabulary Word Detection in Spoken Dialogues Based on Joint Decoding with User Response Patterns*
Ou, Ruizhe	C4.2	*A Method of Efficient Synthesizing Post-Disaster Remote Sensing Image with Diffusion Model and LLM*
Ou, Yu-Feng	P3.15	*A Biased Mixed-Precision Convolution Engine for Hardware-Efficient Computational Imaging CNN*
Ouchi, Yumo	I4.3	*Privacy-Preserving Image Transformation Method for Person Detection and Re-ID*
Oyama, Takehisa	F1.1	*Joint Drum Transcription and Metrical Analysis Based on Periodicity-Aware Multi-Task Learning*
Ozawa, Takuto	P5.18	*Dynamic Characteristics of Electroencephalogram Reflecting Driving-Experience-Dependent Performance Using Microstates*
Pal, Arghya	C1.4	*Unbiased Decision-Making Framework in Long-Video Macro & Micro-Expression Spotting*
	J2.6	*ELEGANT: End-To-End Language Grounded Speech Denoiser for Efficient Generation of Talking Face*
Pan, Chao	B3.5	*A Weighted Binary Cross-Entropy for Sound Event Representation Learning and Few-Shot Classification*
Pan, Ningning	P4.29	*A Controlled Noise Reduction Wiener Filter Based on the Quadratic Eigenvalue Problem*
Panggabean, Jayandi	I1.4	*Comparison of MPPT Based on Deep Reinforcement Learning by DQN, DDPG and TD3*
Panneer, Prabhavathy	C1.1	*MSDF-Net: A Multi-Scale Deep Fusion Network with Dilated Convolutions for Cloud Removal from Sentinel-2 Imagery*
Park, Jeong-Hyeok	G2.5	*Low-Light Image Enhancement via Distillation of NIR-To-RGB Conversion Knowledge*
Park, Jeongsoo	C5.4	*Cross-Domain Sound Recognition for Efficient Underwater Data Analysis*
Park, Tae-Sung	G2.5	*Low-Light Image Enhancement via Distillation of NIR-To-RGB Conversion Knowledge*

Parmar, Krishna	E1.3	*Exploring Residual Cepstral Features for Spoken Language Identification*
Patil, Hemant	E1.2	*Relevance of Quadrature Phase for Replay Detection in Voice Assistants (VAs)*
	E1.3	*Exploring Residual Cepstral Features for Spoken Language Identification*
	P3.2	*Classification of Normal vs. Pathological Infant Cries Using Morse Wavelets*
	A3.5	*Analysis of Emotions in Speech Using AESDD*
Pedanekar, Niranjan	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Pei, Guanxiong	A1.1	*Mixed Emotion Recognition Based on EEG Signals*
	P2.16	*Poisoning Attacks Against Gait-Based Identity Recognition*
Pei, Yu-Cheng	G1.2	*Gait Analysis in Powered Exoskeleton-Assisted Walking in Patients with Stroke: A Case Series Cohort*
Peng, Bo	I5.1	*Urban Noise Monitoring Using Edge Computing with CNN-LSTM on Jetson Nano*
Peng, Da-tian	P2.16	*Poisoning Attacks Against Gait-Based Identity Recognition*
Peng, Haipeng	P4.13	*Improved One-Class Learning for Voice Spoofing Detection*
Peng, Wei	C4.1	*Dictionary-Driven Chinese ASR Entity Correction with Controllable Decoding*
Peng, Wen-Hsiao	G4.3	*Transformer-Based Image Compression with Variable Image Quality Objectives*
	G4.6	*Sparse Tensor-Based Point Cloud Attribute Compression Using Augmented Normalizing Flows*
Peng, Yu-Wen	P2.9	*The Color Demosaicing and Image Scaling Based on Improve Hamilton-Adams*
Peng, YuShan	C3.5	*Integration of Contextual Information with YOLOv5 for Detecting Tracheobronchial Foreign Objects*
Pham, Cuong	J1.5	*GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer*
Pham, Duc-Huy	J3.4	*Investigating the Role of Human Action Detector in Visual-Guide Audio Source Separation System*
Pham, Thuy	G2.4	*Deep Unfolded Underwater Image Enhancement Based on Extreme Channels Prior*
Pham, Van-Cuong	J1.4	*Transformer-Based Deep Learning Detector for Dual-Mode Index Modulation 3D-OFDM*
Pham, Van-Truong	J3.1	*1M Parameters are Enough? A Lightweight CNN-Based Model for Medical Image Segmentation*

Phan, Raphael	C1.4	*Unbiased Decision-Making Framework in Long-Video Macro & Micro-Expression Spotting*
	J2.6	*ELEGANT: End-To-End Language Grounded Speech Denoiser for Efficient Generation of Talking Face*
Piao, Zhenyu	E1.4	*Consideration of Varying Training Lengths for Short-Duration Speaker Verification*
	P5.20	*PDF-NET: Pitch-Adaptive Dynamic Filter Network for Intra-Gender Speaker Verification*
Pleva, Matus	F5.4	*Application of Wafer Defect Pattern Classification Model in the Semiconductor Industry*
Poliakov, Egor	H2.2	*Efficient Constraint-Aware Neural Architecture Search for Object Detection*
Prasetyo, Heri	P3.18	*Dual-Path Residual Attention Convolution Networks for Color-Embedded-Grayscale Image*
Pratomo, Baskoro	P3.24	*Efficient Dictionary and Grid-Based Framework for Answering Durable k-Nearest Neighbor Queries on Time Series Data*
Praveen, Kiran	E1.3	*Exploring Residual Cepstral Features for Spoken Language Identification*
Prihatmoko, Muhammad Dzaky	I1.5	*Signal Quality Assessment for Wearable Multichannel Photoplethysmography Signals*
Pun, Man-On	B4.5	*Dense Three-Dimensional Color Reconstruction for Large-Scale Outdoor Scenes*
Pun, Sio Hang	B2.2	*Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-To-End CNN with Fine-Tuning*
Purwarianti, Ayu	I1.1	*A Comparative Evaluation of Video Codecs for rPPG-Based Heart Rate Estimation*
Putri, Wenny	G5.5	*STUA-Net: A Fingerprint Reconstruction with Swin Transformer and Soft Collective Attention*
Qi, Jiaxing	P3.17	*Efficient Incremental Text-To-Speech on GPUs*
Qi, Wenfa	P3.29	*A Print-Scan-Resilient Watermarking Scheme for Trademark Images*
	P3.28	*Robust Watermark Imaging via Graph-Signal Optimization*
Qian, Zhenxing	P3.27	*StyleStegan: Leak-Free Style Transfer Based on Feature Steganography*
	C3.3	*Identifying the Style of Chatting*
	P4.18	*Preemptive Image Protection Against Steganography*
Qiao, Tong	P3.26	*Fine-Grained Face Anti-Spoofing Based on Recursive Self-Attention and Multi-Scale Fusion*
Qiao, Xiaosong	P2.20	*Incorporating Pinyin into Pipeline Named Entity Recognition from Chinese Speech*

Qin, Chuan	P1.24	*Black-Box Lossless Fragile Watermarking Based on Hidden Space Search for DNN Integrity Authentication*
Qin, Mengjie	P1.28	*Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion*
Qu, Leyuan	P2.30	*Few Shot Learning Guided by Emotion Distance for Cross-Corpus Speech Emotion Recognition*
Quang-Huy, Tran	J3.2	*Imaging Ultrasound Scattering Targets Using Density-Enhanced Chaotic Compressive Sampling*
Radhakrishnan, Balaji	E1.3	*Exploring Residual Cepstral Features for Spoken Language Identification*
Rajanala, Sailaja	C1.4	*Unbiased Decision-Making Framework in Long-Video Macro & Micro-Expression Spotting*
	J2.6	*ELEGANT: End-To-End Language Grounded Speech Denoiser for Efficient Generation of Talking Face*
Ramadhani, Reina	I3.2	*Algorithm Development for Step-Wise Valve Deflation Method in Blood Pressure Measurement*
Ramdhon, Ferdi	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Rao, Haiyang	P5.27	*MHM-Based Reversible Data Hiding with Adaptive Feature Selection*
Raswa, Farchan	G5.5	*STUA-Net: A Fingerprint Reconstruction with Swin Transformer and Soft Collective Attention*
Ravier, Philippe	J1.1	*After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals*
Ren, Kai	C4.6	*Dual Feature Reconstruction Network Based on Few-Shot Image Classification*
Ren, Yeqing	P4.13	*Improved One-Class Learning for Voice Spoofing Detection*
Ren, Yuling	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Rhee, Hochang	G2.1	*Resolution-Adaptive Lossless Image Compression Using Frequency Decomposition Network*
Ruan, Yu-Ping	A2.1	*CH-MEAD: A Chinese Multimodal Conversational Emotion Analysis Dataset with Fine-Grained Emotion Taxonomy*
Ryu, Keunho	D4.5	*An Improved Check Digit-Based Participant Identification System for Human Biorepositories*
S, Uthiraa	A3.5	*Analysis of Emotions in Speech Using AESDD*
Sahipjohn, Neha	E1.5	*Adversarial Robustness of Mel Based Speaker Recognition Systems*

	A4.4	*RobustL2S: Speaker-Specific Lip-To-Speech Synthesis Exploiting Self-Supervised Representations*
Saito, Yasuyuki	A2.2	*Evolutionary Analysis and Cultural Transmission Models of Color Style Distributions in Painting Arts*
Sakakibara, Ken-Ichi	F1.4	*Simultaneous Measurement of Multiple Acoustic Attributes Using Structured Periodic Test Signals Including Music and Other Sound Materials*
Sakamoto, Ayaka	F4.2	*Application for Generating Re-Accessible Screenshots of Web Pages Using Histogram Shrinkage*
Sakano, Hitoshi	I4.5	*Development of a Robust Ear Recognition Algorithm Using Planar Approximation*
Sakdejayont, Theerat	BS1.4	*Residual, Mixer, and Attention: The Three-Way Combination for Streaming Wake Word Detection Framework*
Samsuri, Muhammad Hafiz	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Sankarasrinivasan, S	B5.5	*Deep Residual and Classified Neural Networks for Inverse Halftoning*
Santoso, Bagus Jati	P3.24	*Efficient Dictionary and Grid-Based Framework for Answering Durable k-Nearest Neighbor Queries on Time Series Data*
Saruwatari, Hiroshi	B3.2	*Blind Source Separation Using Independent Low-Rank Matrix Analysis with Spectrogram-Consistency Regularization*
Sasaoka, Takafumi	B2.3	*Decoding Time-Course of Saliency Network of fMRI Signals by EEG Signals Using Optimized Forward Variable Selection: A Concurrent EEG-fMRI Study*
Sasou, Akira	A3.3	*Ensembling Multilingual Pre-Trained Models for Predicting Multi-Label Regression Emotion Share from Speech*
	A3.2	*Multilingual, Cross-Lingual, and Monolingual Speech Emotion Recognition on EmoFilm Dataset*
Sato, Kiyohide	A3.1	*Speech Emotion Recognition by Late Fusion of Linguistic and Acoustic Features Using Deep Learning Models*
Sato, Yuya	I2.2	*Marine Snow Removal Benchmarking Dataset*
Sato, Yuya	I4.5	*Development of a Robust Ear Recognition Algorithm Using Planar Approximation*
Seko, Yuya	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Ser, Wee	P3.6	*Heart Rate Acquisition and Processing Techniques for a Miniature Wearable Microphone Sensor*

Setiawan, Erwin	I3.1	*A QoS Throughput Performance Measurement Comparison Between UGS and BE Services of a Real-Time FPGA Based OFDM Multi-User System Design Implementation*
Shah, Neil	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Shah, Neil	A4.4	*RobustL2S: Speaker-Specific Lip-To-Speech Synthesis Exploiting Self-Supervised Representations*
Shalabi, Fatma	C2.2	*Image-Text Out-Of-Context Detection Using Synthetic Multimodal Misinformation*
Shang, Linbo	P1.22	*On Uncertainty Principles for Lowband Graph Signals*
Shao, Jinliang	P1.17	*A Computational Efficient Direct Position Determination Approach of Narrow-Band Emitter*
Shen, Kung-Yao	E3.1	*An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids*
Shen, XiaoYi	H3.5	*Practical Active Noise Control: Restriction of Maximum Output Power*
Shen, Yi-Tsung	H3.3	*A Study of the Microphone Protection of Active Noise Control for Axial Fan*
Shen, Yih Liang	B3.4	*A Dual-Channel Three-Stage Model for DoA and Speech Enhancement*
	P5.14	*Residual Echo Suppression Using Spatial Feature for Stereo Acoustic Echo Cancellation*
Sheng, Hanmin	P1.17	*A Computational Efficient Direct Position Determination Approach of Narrow-Band Emitter*
Shi, Chaojun	C4.4	*MKTformer: Fine-Grained Meter Classification Based on Multi-Modal Knowledge Transfer*
Shi, Dong-Yuan	H3.5	*Practical Active Noise Control: Restriction of Maximum Output Power*
Shi, Mohan	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Shih, Bo-Wun	B3.4	*A Dual-Channel Three-Stage Model for DoA and Speech Enhancement*
	P5.14	*Residual Echo Suppression Using Spatial Feature for Stereo Acoustic Echo Cancellation*
Shih, Cheng-Liang	E4.2	*OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises*
Shim, Miseon	P3.10	*Gender-Specific EEG Feature-Based Diagnostic Tool for Male Depression Based on Machine-Learning Techniques*
Shimamura, Yuki	J2.4	*High-Quality Font Generation Based on StyleGAN2 and FSFont Font Generation Model*
Shimonishi, Kanta	A5.3	*Universal Sound Separation Using Replay-Based Data Sampling in Incremental Learning*

Shimono, Masanori	A1.3	*Manipulation of Neuronal Network Firing Patterns Using Temporal Deep Unfolding-Based MPC*
Shin, Ishii	BS1.2	*EEG Study on Anticipation of Difficulty for Upcoming Auditory Task*
Shinoda, Koichi	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Shione, Nagito	P5.7	*Construction of Automatic Speech Recognition Model That Recognizes Linguistic Information and Verbal/Non-Verbal Phenomena*
Shiota, Sayaka	F4.3	*Domain Adaptation for Efficiently Fine-Tuning Vision Transformer with Encrypted Images*
	H5.3	*Vocal Tract Length Perturbation-Based Pseudo-Speaker Augmentation for Speaker Embedding Learning*
Shu, Ting-Jia	P2.7	*Learning to Enhance the Position Embedding and Coherence*
Sinaga, Sindy	I3.2	*Algorithm Development for Step-Wise Valve Deflation Method in Blood Pressure Measurement*
Singkul, Sattaya	BS1.4	*Residual, Mixer, and Attention: The Three-Way Combination for Streaming Wake Word Detection Framework*
Sinh Huy, Nguyen	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*
Siripool, Nutchanon	P3.12	*Detection and Correction of Defective Relative Humidity Data Collected from the Greenhouse Environment Using Nested Kalman Filters with Standard Deviation Analysis*
Sirisanwannakul, Kraithep	P3.12	*Detection and Correction of Defective Relative Humidity Data Collected from the Greenhouse Environment Using Nested Kalman Filters with Standard Deviation Analysis*
Sivaprasad, Sarath	E1.5	*Adversarial Robustness of Mel Based Speaker Recognition Systems*
Sobue, Kazuya	G1.3	*Prediction Model of Postoperative Pain Exacerbation Using a Wearable Electrocardiogram Sensor*
Soeta, Yoshiharu	H3.1	*Radiated Sound Field Reproduction for Surrounding Loudspeaker Array Based on Higher-Order Ambisonics*
Son, Do	J1.2	*On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems*
Song, Changhe	P2.29	*Robust Representation Learning for Speech Emotion Recognition with Moment Exchange*
Song, Jae-Jun	P3.9	*Investigation of the Neuromodulation Effect of tVNS on Electrical Brain Activity*
Song, Qi	C4.5	*Feature Augmentation Reconstruction Network for Few-Shot Image Classification*
	C4.6	*Dual Feature Reconstruction Network Based on Few-Shot Image Classification*

Song, Wei	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
Songlin, Sun	B5.2	*Joint Optimization Algorithm for Adaptive Bit Allocation Based on Temporal-Spatial Information*
Srivastava, Ritu	E1.5	*Adversarial Robustness of Mel Based Speaker Recognition Systems*
Srivastava, Vivek	P4.26	*It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data*
Stefanov, Kalin	J2.2	*ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation*
Studiawan, Hudan	P3.24	*Efficient Dictionary and Grid-Based Framework for Answering Durable k-Nearest Neighbor Queries on Time Series Data*
Su, Li	F1.3	*Learning Multifaceted Self-Similarity for Musical Structure Analysis*
Su, Ming-Hsiang	F5.5	*Question Answering System Based on Pre-Training Model and Retrieval Reranking for Industry 4.0*
	F5.4	*Application of Wafer Defect Pattern Classification Model in the Semiconductor Industry*
Su, Po-Chyi	P1.13	*Selecting Suitable Data Input for Deep-Learning Sign-Language Recognition with a Small Dataset*
Su, Wei-Hong	G4.5	*Virtual Garment Fitting Through Parsing and Context-Aware Generative Adversarial Networks with Discriminator Group*
Sudo, Yui	P4.3	*Flexible Evidence Model to Reduce Uncertainty Mismatch Between Speech Enhancement and ASR Based on Encoder-Decoder Architecture*
Sugano, Hidenori	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Sumarudin, A	I1.3	*DQN Algorithm Design for Fast Efficient Shortest Path System*
Sumiyoshi, Takashi	H4.3	*Synthetic Data Augmentation for ASR with Domain Filtering*
Sun, Chen-Kai	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*
Sun, Qibin	P2.17	*STrack: Velocity Estimation Using Single Antenna WiFi Devices*
Suo, Hongbin	D1.5	*Low-Complexity Multi-Channel Speaker Extraction with Pure Speech Cues*
Sutisna, Nana	I1.4	*Comparison of MPPT Based on Deep Reinforcement Learning by DQN, DDPG and TD3*
	I1.3	*DQN Algorithm Design for Fast Efficient Shortest Path System*
	I3.4	*Sparsity Exploration for Structured and Unstructured Weight Formations in CNN Architecture*
	I3.1	*A QoS Throughput Performance Measurement Comparison Between UGS and BE Services of a Real-Time FPGA Based OFDM Multi-User System Design Implementation*

	I3.3	*SUMO Based Hardware/Software Co-Simulation for Two-Intersection Adaptive and Collaborative Traffic Signal Controller*
Suzuki, Hiroharu	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Suzuki, Kenji	P3.12	*Detection and Correction of Defective Relative Humidity Data Collected from the Greenhouse Environment Using Nested Kalman Filters with Standard Deviation Analysis*
Suzuki, Yusei	I4.2	*IPFed: Identity Protected Federated Learning for User Authentication*
Sven, Nordholm	P5.13	*Proportionate NLMS with Variable Step-Size for Adaptive Feedback Cancellation in Hearing Aids*
Syafalni, Infall	I1.4	*Comparison of MPPT Based on Deep Reinforcement Learning by DQN, DDPG and TD3*
	I1.3	*DQN Algorithm Design for Fast Efficient Shortest Path System*
	I3.4	*Sparsity Exploration for Structured and Unstructured Weight Formations in CNN Architecture*
	I3.1	*A QoS Throughput Performance Measurement Comparison Between UGS and BE Services of a Real-Time FPGA Based OFDM Multi-User System Design Implementation*
	I3.3	*SUMO Based Hardware/Software Co-Simulation for Two-Intersection Adaptive and Collaborative Traffic Signal Controller*
Sze, Jyh-Rou	F5.2	*Study on the Reduction of Background Fringes for Detect Detection on Specular Surface*
Ta, Thang	A3.4	*An Automatic Pipeline for Building Emotional Speech Dataset*
Tai, Tzu-Chiang	F5.5	*Question Answering System Based on Pre-Training Model and Retrieval Reranking for Industry 4.0*
Tai, Wei-Shen	E4.6	*Badminton Self-Training System Based on Virtual Reality*
Takaba, Kiyotsugu	D3.2	*¥(¥Delta¥Sigma¥) Modulators for Discrete-Time Closed Loop Control Systems with Quantization and Saturation*
Takagi, Tatsunari	H4.6	*Streaming End-To-End ASR Using CTC Decoder and DRA for Linguistic Information Substitution*
Takahashi, Kazuhiko	P3.1	*Remarks on an Optimal Predictive Control Using a Quaternion Neural Network and a Metaheuristics Optimisation Method*
Takahashi, Kenta	I4.1	*A Biometric Signature Scheme with Template Protection and Authenticated Sample Recoverability*
	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
	I4.2	*IPFed: Identity Protected Federated Learning for User Authentication*
Takahashi, Tomohiro	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*

Takamune, Norihiro	B3.2	*Blind Source Separation Using Independent Low-Rank Matrix Analysis with Spectrogram-Consistency Regularization*
Takamura, Seishi	I2.6	*Shot-Noise-Aware Image Signal Restoration for Photoelectronic Charge-Based Sensors*
Takazawa, Keisuke	A5.4	*Multiple Sound Source Tracking Based on Generative Modeling and Recursive Bayesian Filtering of Spatial Gradient Spectra*
Takeda, Keita	P3.21	*Implementation of PLIM on 429MHz LoRa/FSK with Improved Conversion Table*
Takeda, Ryu	P4.3	*Flexible Evidence Model to Reduce Uncertainty Mismatch Between Speech Enhancement and ASR Based on Encoder-Decoder Architecture*
	H4.2	*Out-Of-Vocabulary Word Detection in Spoken Dialogues Based on Joint Decoding with User Response Patterns*
Takemoto, Jun	A1.2	*Attention-Based CNN and Relative Phase Feature Modeling for Improved Imagined Speech Recognition*
Takeuchi, Daiki	E3.5	*Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer*
	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Takyu, Osamu	P3.21	*Implementation of PLIM on 429MHz LoRa/FSK with Improved Conversion Table*
Tamamori, Akira	P1.3	*Kernel Random Projection Depth for Outlier Detection*
Tambrahalli, Vishal	A4.4	*RobustL2S: Speaker-Specific Lip-To-Speech Synthesis Exploiting Self-Supervised Representations*
Tamura, Akihiro	P4.7	*Transformer-Based Automatic Speech Recognition of Simultaneous Interpretation with Auxiliary Input of Source Language Text*
Tan, Chi Wee	C1.3	*Addressing Distortion in Social Distance Monitoring: Calibration for Accurate Distance Calculation*
	P3.14	*Revolutionizing Formative Assessment in STEM Fields: Leveraging AI and NLP Techniques*
Tan, Li Peng	F4.6	*Quick Response (QR) Codes Embedding in VVC Using Quantisation Parameter Manipulation*
Tan, Pei Sze	C1.4	*Unbiased Decision-Making Framework in Long-Video Macro & Micro-Expression Spotting*
Tan, Rex Xiao	P3.6	*Heart Rate Acquisition and Processing Techniques for a Miniature Wearable Microphone Sensor*
Tan, Yuan	A4.2	*Structural Quality Assured Global Optimization for CTU-Level Rate Control of Screen Content Coding*

Tanabe, Nari	D3.1	*Real-Time Noise Suppression Using Harmonic/Percussive Separation with Morphological Operations for Hammering Test*
Tanaka, Chinatsu	C2.3	*Gait Recognition Scheme Focusing on Operating Characteristics at Feature Points Detected by OpenPose*
Tanaka, Toshihisa	BS1.3	*Event-Related Potential in Rapid Serial Visual Presentation-Based Partial Face Cognition Depends on Visible Face Components*
	B2.4	*Multimodal Recognition of Speech and Electrocorticogram*
Tanaka, Yuichi	I2.2	*Marine Snow Removal Benchmarking Dataset*
Tang, Boshi	P2.29	*Robust Representation Learning for Speech Emotion Recognition with Moment Exchange*
Tang, Tao	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Tang, Tong	A4.2	*Structural Quality Assured Global Optimization for CTU-Level Rate Control of Screen Content Coding*
Taniguchi, Shuta	P4.7	*Transformer-Based Automatic Speech Recognition of Simultaneous Interpretation with Auxiliary Input of Source Language Text*
Tao, Jianhua	P4.25	*Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition*
Tao, Liang	P2.28	*Single Source Zone Detection in the Spherical Harmonic Domain for Multisource Localization*
Tao, Shimin	P2.20	*Incorporating Pinyin into Pipeline Named Entity Recognition from Chinese Speech*
Tawara, Naohiro	H5.1	*Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation*
Tee Kit Tsun, Mark	D2.6	*How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?*
	D2.3	*MoMo Strategy: Learn More from More Mistakes*
Tew, Yiqi	C1.3	*Addressing Distortion in Social Distance Monitoring: Calibration for Accurate Distance Calculation*
	F4.6	*Quick Response (QR) Codes Embedding in VVC Using Quantisation Parameter Manipulation*
Then, Patrick	D2.4	*Unveiling Robust Feature Spaces: Image vs. Embedding-Oriented Approaches for Plant Disease Identification*
Then, Yi Lung	D2.6	*How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?*
	D2.3	*MoMo Strategy: Learn More from More Mistakes*
Thuy Quynh, Tran Thi	J1.2	*On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems*

	J3.2	*Imaging Ultrasound Scattering Targets Using Density-Enhanced Chaotic Compressive Sampling*
Tian, Qi	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Tian, Zhenzhen	F5.1	*Explainable Deep Mechanical Troubleshooting*
Ting, Chee-Ming	J2.6	*ELEGANT: End-To-End Language Grounded Speech Denoiser for Efficient Generation of Talking Face*
Titalim, Benita	D4.2	*Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids*
Toda, Tomoki	F3.1	*Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion*
	P4.8	*An Analysis of Personalized Speech Recognition System Development for the Deaf and Hard-Of-Hearing*
Tonosaki, Taito	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
Tran, Dung	F3.5	*Increasing Speech Intelligibility by Mimicking Professional Announcers' Voices and Its Physical Correlates*
Tran, Duy	J1.2	*On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems*
Tran, Linh	P5.13	*Proportionate NLMS with Variable Step-Size for Adaptive Feedback Cancellation in Hearing Aids*
Tran, Quang	J2.1	*Generative Adversarial Network-Based Frame Interpolation with Multi-Perspective Discrimination*
Tran, Thanh-Hai	J1.5	*GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer*
Tran, Thi-Thao	J3.1	*1M Parameters are Enough? A Lightweight CNN-Based Model for Medical Image Segmentation*
Tran, Trung-Kien	J1.5	*GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer*
Tran, Vu-Hoang	H2.3	*A Reliable Feature-Based Framework for Vehicle Tracking in Advanced Driver Assistance Systems*
	G4.4	*From Synthetic to Real: Enhancing Deep Learning Models with Generative Adversarial Networks for Efficient Data Utilization in Automatic Retail Stores*
Triani, Listi	I1.2	*Human Activity Recognition Based on FMCW Radar Using CNN and Transfer Learning*
Tris Atmaja, Bagus	A3.3	*Ensembling Multilingual Pre-Trained Models for Predicting Multi-Label Regression Emotion Share from Speech*
	A3.2	*Multilingual, Cross-Lingual, and Monolingual Speech Emotion Recognition on EmoFilm Dataset*

Truong, Thanh-Nguyen	H2.3	*A Reliable Feature-Based Framework for Vehicle Tracking in Advanced Driver Assistance Systems*
Tsai, Kun-Lin	P1.16	*A Deep Learning Based Sustainable Energy Scheduling System*
Tsai, Pei-Yun	H1.2	*Real-Time Processing for Weighted Pulse Decomposition of Photoplethysmography Signals Based on Interior Point Method in Wearable Devices for Hemodynamic State*
Tsai, Shu-Wei	G5.3	*Data Selection Based on Phoneme Affinity Matrix for Electrolarynx Speech Recognition*
Tsai, Tsung-Han	C3.1	*A Reconfigurable Hardware Architecture for Graph Convolution Network in Action Recognition*
Tseng, Chien-Cheng	D5.1	*Distributed Computation of Heat Kernel Smoothing Using Series Expansion Method*
Tseng, Hua-Wei	G5.2	*Temporal and Type Correlation in Digital Phenotyping for Bipolar Disorder State Prediction Using Multitask Self-Supervised Learning*
Tseng, Shih-Ting	P1.16	*A Deep Learning Based Sustainable Energy Scheduling System*
Tseng, Yi-Li	G1.1	*Cognitive Assessment of Autism Spectrum Disorder Using an EEG-Based Social Interaction Platform*
Tseng, Yu-Wen	G4.2	*Task-Specific Pruning: Efficient Parameter Reduction in Multi-Task Object Detection Models*
Tsou, Yung-An	P4.9	*Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods*
Tsubaki, Shunsuke	C5.3	*Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach*
Tu, Hao-Jiun	P3.15	*A Biased Mixed-Precision Convolution Engine for Hardware-Efficient Computational Imaging CNN*
Tung, Shih-Shuo	E4.1	*Un-Rectifying in ReLU Networks and Applications*
Turlapaty, Anish	P1.1	*Fatigue Classification and Onset Estimation Using Surface EMG Signals During Strength Training*
Tyszczuk, Lily	P1.6	*Predicting Outcomes of Cognitive Behavioral Therapy for Depression Using Data Driven Approaches*
Uchida, Hidetsugu	I4.3	*Privacy-Preserving Image Transformation Method for Person Detection and Re-ID*
Uchiyama, Ryugo	D3.1	*Real-Time Noise Suppression Using Harmonic/Percussive Separation with Morphological Operations for Hammering Test*
Ueda, Takumi	I2.2	*Marine Snow Removal Benchmarking Dataset*
Ueda, Tetsuya	B3.3	*Moving Interference Speaker Removal Using Geometrically Constrained Independent Vector Analysis*

	E3.2	*On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering*
	A5.5	*Spatially-Regularized Switching Independent Vector Analysis*
	A5.2	*Geometrically Constrained Blind Moving Source Extraction Based on Constant Separation Vector and Auxiliary Function Technique*
Ueno, Natsuki	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
Ueno, Sei	F3.4	*Accent-Preserving Voice Conversion Between Native-Nonnative Speakers for Second Language Learning*
Uesaka, Minato	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Ugurel, Ekin	G3.3	*Multi-Task Gaussian Process Learning to Fill in Gaps in LBS Mobile Data*
Um, Seyun	E1.4	*Consideration of Varying Training Lengths for Short-Duration Speaker Verification*
Unoki, Masashi	F3.5	*Increasing Speech Intelligibility by Mimicking Professional Announcers' Voices and Its Physical Correlates*
	D4.2	*Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids*
	P4.2	*Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection*
	G5.1	*Deepfake-Speech Detection with Pathological Features and Multilayer Perceptron Neural Network*
	H5.2	*Contribution of Modulation Spectral Features for Cross-Lingual Speech Emotion Recognition Under Noisy Reverberant Conditions*
Ura, Akinobu	F4.4	*Study on Face Landmark-Based Analysis for Synthetic Media Identification Generated by Adversarial Generative Networks*
Usman, Koredianto	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Van Duc, Nguyen	J3.5	*A Combination of Time and Frequency Synchronization with Doppler Compensation in Coded OFDM-Based UWA Systems*
Van Loi, Nguyen	J3.3	*Segmentation and Observation of Hand Rehabilitation Exercises by Supporting of Acceleration Signals*
Vann, Veasna	G2.6	*3D Human Skeleton Estimation from Single RGB Image Based on Fusion of Predicted Depths from Multiple Virtual-Viewpoints*
Venkategowda, Naveen	D3.5	*Moreau Envelope ADMM for Decentralized Weakly Convex Optimization*
	G3.1	*Robust Networked Federated Learning for Localization*

Violeta, Lester Phillip	P4.8	*An Analysis of Personalized Speech Recognition System Development for the Deaf and Hard-Of-Hearing*
Wadayama, Tadashi	E5.1	*Deep Unfolding-Based Distributed MIMO Detection*
Wagatsuma, Nobuhiko	P1.2	*P300 Event-Related Potential in Perception of Multiple Traffic Objects During Vehicle Driving*
	P5.18	*Dynamic Characteristics of Electroencephalogram Reflecting Driving-Experience-Dependent Performance Using Microstates*
Wakamatsu, Tomoka	H5.3	*Vocal Tract Length Perturbation-Based Pseudo-Speaker Augmentation for Speaker Embedding Learning*
Wakamiya, Naoki	A1.3	*Manipulation of Neuronal Network Firing Patterns Using Temporal Deep Unfolding-Based MPC*
Wan, Genshun	P2.25	*Progressive Multi-Scale Self-Supervised Learning for Speech Recognition*
	P2.26	*Improved Data2vec with Soft Supervised Hidden Unit for Mandarin Speech Recognition*
Wan, Minhong	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
	P1.28	*Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion*
Wan, Timothy	G1.2	*Gait Analysis in Powered Exoskeleton-Assisted Walking in Patients with Stroke: A Case Series Cohort*
Wan, Yulong	D1.5	*Low-Complexity Multi-Channel Speaker Extraction with Pure Speech Cues*
Wang, Bo-Sin	E3.1	*An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids*
Wang, Chang	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Wang, Chenxi	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
	P2.25	*Progressive Multi-Scale Self-Supervised Learning for Speech Recognition*
Wang, Chih-Hao	P1.15	*Old Damaged Photo Recovery with Style Transfer-Based Data Augmentation*
Wang, Chung-Che	P5.10	*Personalized Audio Quality Preference Prediction*
Wang, Cyun-Bo	D5.3	*EffSegmentNet: Efficient Design for Real-Time Semantic Segmentation*
Wang, Hongxia	P5.28	*Feature Selection Based on Clonal Selection Algorithm for Image Steganalysis*

Wang, Hsin-Wei	P5.11	*AVATAR: Robust Voice Search Engine Leveraging Autoregressive Document Retrieval and Contrastive Learning*
Wang, Jia Wen	P2.8	*VLSI Design of Lossless Image Compression Using Improved LZW*
Wang, Jia-Ching	F5.5	*Question Answering System Based on Pre-Training Model and Retrieval Reranking for Industry 4.0*
	G5.5	*STUA-Net: A Fingerprint Reconstruction with Swin Transformer and Soft Collective Attention*
	F5.3	*On the Optimal Self-Supervised Multi-Fault Detector for Temperature Sensor Data*
Wang, Jiameng	P3.29	*A Print-Scan-Resilient Watermarking Scheme for Trademark Images*
Wang, Jie	D1.3	*CASA-Net: Cross-Attention and Self-Attention for End-To-End Audio-Visual Speaker Diarization*
Wang, Jihua	F5.1	*Explainable Deep Mechanical Troubleshooting*
Wang, Kai	P2.18	*SEformer: Dual-Path Conformer Neural Network is a Good Speech Denoiser*
Wang, Kevin I-Kai	I5.1	*Urban Noise Monitoring Using Edge Computing with CNN-LSTM on Jetson Nano*
Wang, Lan	E1.1	*Modeling Suprasegmental Information Using Finite Difference Network for End-To-End Speaker Verification*
Wang, Lang	C4.5	*Feature Augmentation Reconstruction Network for Few-Shot Image Classification*
Wang, Lijun	P4.2	*Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection*
Wang, Liyuan	P5.25	*Multi-Accent Pronunciation Assessment Based on Domain Adversarial Training*
Wang, Qi	P2.23	*Multi-Task Piano Transcription with Local Relative Time Attention*
Wang, Qing	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Wang, Ruoyu	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Wang, Shaokang	B5.2	*Joint Optimization Algorithm for Adaptive Bit Allocation Based on Temporal-Spatial Information*
Wang, Shenjian	D3.2	*¥(¥Delta¥Sigma¥) Modulators for Discrete-Time Closed Loop Control Systems with Quantization and Saturation*
Wang, Sijie	P4.30	*Target Speaker Extraction with Attention Enhancement and Gated Fusion Mechanism*
Wang, Tai-chi	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*

Wang, Tianrui	P1.19	*Noise-Robust Pitch Detection Based on Super-Resolution Harmonics*
Wang, Ting Wei	P3.13	*Pedestrian Crossing Intention Prediction with Multi-Modal Transformer-Based Model*
Wang, Wei	J2.3	*LSR++: An Efficient and Tiny Model for Image Super-Resolution*
Wang, Wen	P5.24	*NADiffuSE: Noise-Aware Diffusion-Based Model for Speech Enhancement*
Wang, Xianrui	E3.2	*On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering*
Wang, Xin	P5.22	*Time Sensitive Hash and Adaptive Image Recovery Based Self-Embedding Fragile Watermarking Scheme in Encrypted Images*
Wang, Xinyu	B2.2	*Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-To-End CNN with Fine-Tuning*
Wang, Yi-Cheng	P5.11	*AVATAR: Robust Voice Search Engine Leveraging Autoregressive Document Retrieval and Contrastive Learning*
Wang, Yu-Chieh	P4.15	*Breast Cancer Detection Auxiliary System Leveraging Deep Learning and Mixed Reality*
Wang, Yuan-Gen	B1.3	*Mismatched Semi-Supervised Learning with Feature Similarity Consistency*
Wang, Yujun	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Wang, Zhe	D1.2	*Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization*
Watanabe, Chihiro	F3.2	*DisC-VC: Disentangled and F0-Controllable Neural Voice Conversion*
Watanabe, Satoshi	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Watanabe, Shinji	H4.3	*Synthetic Data Augmentation for ASR with Domain Filtering*
Wei, An-Chi	F5.2	*Study on the Reduction of Background Fringes for Detect Detection on Specular Surface*
Wei, Chung-Yen	P2.11	*Design of Interactive System for Acupoint Analysis Based on Augmented Reality*
Wei, Jianguo	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Wei, Kangkang	P3.30	*AI-Generated Image Detection Using a Cross-Attention Enhanced Dual-Stream Network*
Wei, Sha	P1.27	*High-Quality Triggers Based Fragile Watermarking for Optical Character Recognition Model*
Wei, Yina	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*

Wei, Yu-Jen	P1.15	*Old Damaged Photo Recovery with Style Transfer-Based Data Augmentation*
	P1.14	*Analysis of the Interaction Effect on Pruning and Transfer Learning in Model Training*
Wen, Zhengqi	P4.25	*Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition*
Weng, Shaowei	P5.27	*MHM-Based Reversible Data Hiding with Adaptive Feature Selection*
Wenhuan, Lu	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Werner, Stefan	D3.5	*Moreau Envelope ADMM for Decentralized Weakly Convex Optimization*
	G3.1	*Robust Networked Federated Learning for Localization*
	G3.2	*Continual Local Updates for Federated Learning with Enhanced Robustness to Link Noise*
Wicaksana, Prabowo Yoga	G5.5	*STUA-Net: A Fingerprint Reconstruction with Swin Transformer and Soft Collective Attention*
Widyadhana, Arya	B5.1	*ASF-LLRDA: Locality-Regularized Linear Regression Discriminant Analysis with Approximately Symmetrical Face Preprocessing for Face Recognition*
Wong, Chin Wee	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Wong, KokSheik	J2.2	*ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation*
	H2.6	*CapFormer: A Space-Time Video Description Model Using Joint-Attention Transformer*
	F4.5	*HDR Image Watermarking Based on Saliency Detection and Quantization Index Modulation*
Wong, Ting-Jui	H1.2	*Real-Time Processing for Weighted Pulse Decomposition of Photoplethysmography Signals Based on Interior Point Method in Wearable Devices for Hemodynamic State*
Wu, Chia-Hsuan	P4.10	*Detecting the Underwater Distance and Swimming Direction of Tilapia Using YOLO*
Wu, Chung-Hsien	G5.3	*Data Selection Based on Phoneme Affinity Matrix for Electrolarynx Speech Recognition*
	G5.2	*Temporal and Type Correlation in Digital Phenotyping for Bipolar Disorder State Prediction Using Multitask Self-Supervised Learning*
Wu, Ja-Ling	BS1.1	*Enhancing Real-Time Semantic Segmentation with Textual Knowledge of Pre-Trained Vision-Language Model: A Lightweight Approach*

	G4.1	*CPIPS: Learning to Preserve Perceptual Distances in End-To-End Image Compression*
Wu, Jiasheng	P3.26	*Fine-Grained Face Anti-Spoofing Based on Recursive Self-Attention and Multi-Scale Fusion*
Wu, Jijie	C4.6	*Dual Feature Reconstruction Network Based on Few-Shot Image Classification*
Wu, Junnan	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Wu, Meng-Hsuan	B3.4	*A Dual-Channel Three-Stage Model for DoA and Speech Enhancement*
	P5.14	*Residual Echo Suppression Using Spatial Feature for Stereo Acoustic Echo Cancellation*
Wu, Mengjiang	P4.17	*MAEDefense: An Effective Masked AutoEncoder Defense Against Adversarial Attacks*
Wu, Ming	C4.2	*A Method of Efficient Synthesizing Post-Disaster Remote Sensing Image with Diffusion Model and LLM*
Wu, Pei-Chen	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Wu, Tianci	P5.29	*ScaleFormer: Transformer-Based Speech Enhancement in the Multi-Scale Time Domain*
Wu, Ting	P3.26	*Fine-Grained Face Anti-Spoofing Based on Recursive Self-Attention and Multi-Scale Fusion*
Wu, Tsung-Hsi	I5.2	*Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind*
Wu, Wen	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Wu, Yanfeng	P2.27	*Investigation of Ensemble of Self-Supervised Models for Speech Emotion Recognition*
	P2.30	*Few Shot Learning Guided by Emotion Distance for Cross-Corpus Speech Emotion Recognition*
Wu, Zhiyong	P2.29	*Robust Representation Learning for Speech Emotion Recognition with Moment Exchange*
Wu, Zhizheng	P4.19	*Zero-Shot Multi-Speaker Accent TTS with Limited Accent Data*
Wumaier, Aishan	P4.23	*Enhancing Whisper Model for Pronunciation Assessment with Multi-Adapters*
Xi, Ziyi	P3.30	*AI-Generated Image Detection Using a Cross-Attention Enhanced Dual-Stream Network*

Xiao, Yanqi	E4.2	*OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises*
Xiao, Yongxiong	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
Xie, Lei	P5.23	*Multi-Granularity Semantic and Acoustic Stress Prediction for Expressive TTS*
Xie, Shichuang	P3.26	*Fine-Grained Face Anti-Spoofing Based on Recursive Self-Attention and Multi-Scale Fusion*
Xing, Haoran	E3.5	*Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer*
Xiong, Wenmeng	P2.23	*Multi-Task Piano Transcription with Local Relative Time Attention*
Xiong, Zixiang	G3.5	*Communication-Efficient Design of Leaning System for Energy Demand Forecasting of Electrical Vehicles*
Xu, Bo-yuan	P2.11	*Design of Interactive System for Acupoint Analysis Based on Augmented Reality*
Xu, Jiacong	G3.5	*Communication-Efficient Design of Leaning System for Energy Demand Forecasting of Electrical Vehicles*
Xu, Jian	P2.17	*STrack: Velocity Estimation Using Single Antenna WiFi Devices*
Xu, Kedi	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Xu, Shiyun	P2.24	*Real and Imaginary Part Interaction Network for Monaural Speech Enhancement and De-Reverberation*
Xue, Jing-Hao	B1.5	*Learning Meta Soft Prompt for Few-Shot Language Models*
Xue, Liumeng Xue	P5.23	*Multi-Granularity Semantic and Acoustic Stress Prediction for Expressive TTS*
Xue, Xiaopeng	P4.13	*Improved One-Class Learning for Voice Spoofing Detection*
Yamada, Kenta	P2.5	*Fundamental Frequency Estimation Based on Finite-Order Harmonic Constraint Differential Equation*
Yamadera, Keiji	A2.3	*Ultimatelink Between Characters Having a Certain Meaning in Physical Space to URL in Cyberspace with Robust Print and Scan*
Yamaga, Hikaru	P2.4	*Accurate and Practical Query-By-Example by Using Multiple Deep Learning Models and Frame Compression Methods*
	P5.6	*Ensemble of Transformer and Convolutional Recurrent Neural Network for Improving Discrimination Accuracy in Automatic Chord Recognition*
Yamagishi, Junichi	P5.8	*Exploring Isolated Musical Notes as Pre-Training Data for Predominant Instrument Recognition in Polyphonic Music*
Yamaguchi, Kanji	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*

Yamamoto, Kazuki	P1.4	*Soft-Sensor Construction Method Based on Adaptive Modeling and Transfer Learning for Manufacturing Process Including Maintenance Periods*
Yamamoto, Yuki	P1.2	*P300 Event-Related Potential in Perception of Multiple Traffic Objects During Vehicle Driving*
Yamamoto, Yuya	H4.1	*Toward Leveraging Pre-Trained Self-Supervised Frontends for Automatic Singing Voice Understanding Tasks: Three Case Studies*
Yamaoka, Kouei	P2.5	*Fundamental Frequency Estimation Based on Finite-Order Harmonic Constraint Differential Equation*
	B5.3	*Maximization of 2D Cross-Correlation Based on Auxiliary Function Method for Image Alignment*
	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Yamashita, Yoichi	P4.4	*Investigating the Effectiveness of Speaker Embeddings for Shout Intensity Prediction*
	A5.3	*Universal Sound Separation Using Replay-Based Data Sampling in Incremental Learning*
Yamasue, Hidenori	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
Yamato, Kazuki	D4.3	*Deep Learning-Based MRI Super-Resolution Using Non-Uniform Segmented Phase-Scrambling Fourier Transform Signals*
Yamawaki, Shigeto	B2.3	*Decoding Time-Course of Saliency Network of fMRI Signals by EEG Signals Using Optimized Forward Variable Selection: A Concurrent EEG-fMRI Study*
Yan, Bi-Cheng	P5.11	*AVATAR: Robust Voice Search Engine Leveraging Autoregressive Document Retrieval and Contrastive Learning*
Yan, Haotian	C4.2	*A Method of Efficient Synthesizing Post-Disaster Remote Sensing Image with Diffusion Model and LLM*
Yan, Nan	E1.1	*Modeling Suprasegmental Information Using Finite Difference Network for End-To-End Speaker Verification*
Yan, Wei	F5.1	*Explainable Deep Mechanical Troubleshooting*
Yang, Binchen	C4.3	*Privacy-Oriented Coded Caching in Mobile Information-Centric Network*
Yang, Chao-Han	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Yang, Dongchao	P5.24	*NADiffuSE: Noise-Aware Diffusion-Based Model for Speech Enhancement*

Yang, Hao	P2.20	*Incorporating Pinyin into Pipeline Named Entity Recognition from Chinese Speech*
Yang, Ruiguo	P3.28	*Robust Watermark Imaging via Graph-Signal Optimization*
Yang, Shih-Hsuan	J2.1	*Generative Adversarial Network-Based Frame Interpolation with Multi-Perspective Discrimination*
Yang, Tzu-Ting	P5.11	*AVATAR: Robust Voice Search Engine Leveraging Autoregressive Document Retrieval and Contrastive Learning*
Yang, Xiaokang	P2.22	*Effective Fine-Tuning Method for Tibetan Low-Resource Dialect Speech Recognition*
Yang, Xue	P5.26	*GAN-Based Time-Domain Packet Loss Concealment Method with Consistent Mapping Approach*
Yang, Yichen	E3.2	*On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering*
Yang, Yunming	J2.5	*Enhanced Residual Fourier Transformation Network for Lightweight Image Super-Resolution*
Yang, Zhih-Min	I5.4	*Impact of the Weighted Loss Function on the Innovative CMAQ-CNN PM2.5 Forecasting Model*
Yang-Mao, Shih-Fang	BS1.6	*Mask2Hand: Learning to Predict the 3D Hand Pose and Shape from Shadow*
Yano, Kota	P3.8	*Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model*
Yao, Dingding	P2.28	*Single Source Zone Detection in the Spherical Harmonic Domain for Multisource Localization*
Yasu, Keiichi	P5.5	*Corpus Construction for Deaf Speakers and Analysis by Automatic Speech Recognition*
Yasumura, Yoshiko	I4.4	*Eye Biometrics Combined with Periocular and Iris Recognition Using CNN*
Yatabe, Kohei	F1.4	*Simultaneous Measurement of Multiple Acoustic Attributes Using Structured Periodic Test Signals Including Music and Other Sound Materials*
	B3.2	*Blind Source Separation Using Independent Low-Rank Matrix Analysis with Spectrogram-Consistency Regularization*
Ye, Qichen	P5.24	*NADiffuSE: Noise-Aware Diffusion-Based Model for Speech Enhancement*
Ye, Zheng-Kai	I5.2	*Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind*

Ye, Zhongfu	P2.26	*Improved Data2vec with Soft Supervised Hidden Unit for Mandarin Speech Recognition*
	P2.25	*Progressive Multi-Scale Self-Supervised Learning for Speech Recognition*
Ye, Zhongyi	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
Yeh, Mei-Chen	P1.11	*A Transformer-Based Framework for Tiny Object Detection*
	A4.6	*Exploring a CLIP-Enhanced Automated Approach for Video Description Generation*
Yeh, Yun-Ching	B5.4	*Multitask Record for Badminton Match*
Yen, Tzu-Te	E4.4	*Learned String Quartet Music with Variational Auto Encoder*
Yi, EunGyoung	P3.10	*Gender-Specific EEG Feature-Based Diagnostic Tool for Male Depression Based on Machine-Learning Techniques*
Yi, Liu	P1.30	*Improving Aspect Sentiment Classification via Retrieving from Training Data*
Yi, Lu	H5.5	*Cross-Domain Adaptation in Distance Space for Speaker Verification*
Yi-Qian, Li	P2.13	*Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement*
Yim, Monyneath	G4.6	*Sparse Tensor-Based Point Cloud Attribute Compression Using Augmented Normalizing Flows*
Yin, Heng	P1.27	*High-Quality Triggers Based Fragile Watermarking for Optical Character Recognition Model*
Yin, Yujie	P1.27	*High-Quality Triggers Based Fragile Watermarking for Optical Character Recognition Model*
Yin, Zhaoxia	P1.27	*High-Quality Triggers Based Fragile Watermarking for Optical Character Recognition Model*
	P4.17	*MAEDefense: An Effective Masked AutoEncoder Defense Against Adversarial Attacks*
	P4.16	*Efficient Reversible Data Hiding for 3D Mesh Models Based on Multi-LSB Substitution and Ring-Prediction*
Ying, Qichao	P3.27	*StyleStegan: Leak-Free Style Transfer Based on Feature Steganography*
Yip, Jia Qi	A5.1	*Analysis of Speech Separation Performance Degradation on Emotional Speech Mixtures*
Yoshii, Kazuyoshi	F1.1	*Joint Drum Transcription and Metrical Analysis Based on Periodicity-Aware Multi-Task Learning*
	F1.3	*Learning Multifaceted Self-Similarity for Musical Structure Analysis*
	F1.2	*CTC2: End-To-End Drum Transcription Based on Connectionist Temporal Classification with Constant Tempo Constraint*

	BS1.5	*Audio-To-Score Singing Transcription Based on Joint Estimation of Pitches, Onsets, and Metrical Positions with Tatum-Level CTC Loss*
	E4.3	*Multimodal Multifaceted Music Emotion Recognition Based on Self-Attentive Fusion of Psychology-Inspired Symbolic and Acoustic Features*
	C5.1	*DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection*
Yoshizawa, Tsubasa	P5.9	*Speech Quality Improvement Utilizing Out-Of-Focus Areas in Rolling-Shutter Video on Speech Extraction*
Young, Si-Sheng	H2.1	*GNN-Based Small-Data Learning with Area-Control Mechanism for Hyperspectral Satellite Change Detection*
Yu, Chih Chang	D5.2	*In-Air Handwriting for Chinese Character Recognition from Monocular Camera: A Deep Learning Based Approach with Fingertip Detection and Virtual Strokes Elimination*
Yu, Fan	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Yu, Hui Hung	I5.3	*A Framework for Reusing Earth Science Data on Data and Model Marketplaces*
Yu, Tsai Chun	P1.9	*Deformable Aligned Fusion for Video Super Resolution*
Yu, Zhipeng	P1.30	*Improving Aspect Sentiment Classification via Retrieving from Training Data*
Yu-Chieh, Chu	P4.11	*Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments*
Yuan, Zichen	P3.29	*A Print-Scan-Resilient Watermarking Scheme for Trademark Images*
Yue, Pengcheng	P1.21	*An Multi-Evidence Fusion Based on C-Distance with Uncertain Reasoning for Classification*
	P2.30	*Few Shot Learning Guided by Emotion Distance for Cross-Corpus Speech Emotion Recognition*
	P2.19	*Complex Feature Information Enhanced Speech Emotion Recognition*
	P2.27	*Investigation of Ensemble of Self-Supervised Models for Speech Emotion Recognition*
Yuen, Kwok	H4.5	*ASR Model Adaptation for Rare Words Using Synthetic Data Generated by Multiple Text-To-Speech Systems*
Yuen, Shang Li	P1.5	*Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset*
Yukawa, Masahiro	A5.4	*Multiple Sound Source Tracking Based on Generative Modeling and Recursive Bayesian Filtering of Spatial Gradient Spectra*
Yukoh, Wakabayashi	H4.6	*Streaming End-To-End ASR Using CTC Decoder and DRA for Linguistic Information Substitution*

	P5.7	*Construction of Automatic Speech Recognition Model That Recognizes Linguistic Information and Verbal/Non-Verbal Phenomena*
	C5.5	*Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring*
	H5.4	*Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development*
	P5.2	*Combining Multiple End-To-End Speech Recognition Models Based on Density Ratio Approach*
Zeng, Bang	D1.5	*Low-Complexity Multi-Channel Speaker Extraction with Pure Speech Cues*
Zeng, Yi-Chong	A4.5	*Realizing Nipple in Profile Recognition and Nipple Detection Using a Single Classification*
	G5.4	*Reduction of Annotation Effort in Medical Image Analysis Based on Self-Supervised Learning*
Zhang, Chao	A1.1	*Mixed Emotion Recognition Based on EEG Signals*
Zhang, Chuang	C4.2	*A Method of Efficient Synthesizing Post-Disaster Remote Sensing Image with Diffusion Model and LLM*
Zhang, Chunlong	P1.28	*Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion*
Zhang, Dongheng	P2.17	*STrack: Velocity Estimation Using Single Antenna WiFi Devices*
Zhang, Fengrun	P4.22	*Hybrid Syllable and Character Representations for Mandarin ASR*
Zhang, Haihong	P1.6	*Predicting Outcomes of Cognitive Behavioral Therapy for Depression Using Data Driven Approaches*
Zhang, Jia-Sheng	B4.2	*The VLSI Design DSP and DDR4 for Multi-Sensor in Biomedical System*
Zhang, Jie	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
	P4.20	*Speech Enhancement with Multi-Granularity Vector Quantization*
	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Zhang, Lijun	P4.24	*DoA Estimation of Room Reflections Using NN-Based MUSIC Algorithm*
Zhang, Manman	C3.3	*Identifying the Style of Chatting*
Zhang, Min	P2.20	*Incorporating Pinyin into Pipeline Named Entity Recognition from Chinese Speech*
Zhang, Mingyang	P4.19	*Zero-Shot Multi-Speaker Accent TTS with Limited Accent Data*
Zhang, Ruifeng	A5.2	*Geometrically Constrained Blind Moving Source Extraction Based on Constant Separation Vector and Auxiliary Function Technique*
Zhang, ShiLei	P1.19	*Noise-Robust Pitch Detection Based on Super-Resolution Harmonics*

Zhang, Shiliang	P4.21	*A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-Party Meetings*
Zhang, Shiqi	E3.5	*Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer*
Zhang, Siang-Ling	A4.6	*Exploring a CLIP-Enhanced Automated Approach for Video Description Generation*
Zhang, Weitai	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
Zhang, Wen	P1.20	*A Subband Approach to Personal Sound Zone with Joint Optimization of Sound Pressure and Particle Velocity*
	P4.24	*DoA Estimation of Room Reflections Using NN-Based MUSIC Algorithm*
Zhang, Wenbin	C5.2	*Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios*
Zhang, Xiangqing	C1.2	*Instance Implant-Aided Non-Uniformly Cropping for Person Detection in Aerial Images*
Zhang, Xinpeng	P3.27	*StyleStegan: Leak-Free Style Transfer Based on Feature Steganography*
	C3.3	*Identifying the Style of Chatting*
	P4.18	*Preemptive Image Protection Against Steganography*
Zhang, Xinyu	F5.1	*Explainable Deep Mechanical Troubleshooting*
Zhang, Yao-Zhong	P2.8	*VLSI Design of Lossless Image Compression Using Improved LZW*
Zhang, Zehua	P2.24	*Real and Imaginary Part Interaction Network for Monaural Speech Enhancement and De-Reverberation*
Zhang, Zhichao	P1.22	*On Uncertainty Principles for Lowband Graph Signals*
	D5.4	*Universal Optimal Parameters of the Closed-Form Linear Canonical Wigner Distribution*
Zhao, Gejian	P1.24	*Black-Box Lossless Fragile Watermarking Based on Hidden Space Search for DNN Integrity Authentication*
Zhao, H. Vicky	P1.18	*Modeling and Analysis of the Epidemic-Behavior Co-Evolution Dynamics with User Irrationality*
Zhao, Jiahao	E4.3	*Multimodal Multifaceted Music Emotion Recognition Based on Self-Attentive Fusion of Psychology-Inspired Symbolic and Acoustic Features*
Zhao, Jiangjiang	C2.6	*Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition*
Zhao, Mengmeng	P4.13	*Improved One-Class Learning for Voice Spoofing Detection*
Zhao, Shuyuan	P2.30	*Few Shot Learning Guided by Emotion Distance for Cross-Corpus Speech Emotion Recognition*
Zhao, Xiaoying	P4.20	*Speech Enhancement with Multi-Granularity Vector Quantization*

Zhao, Yanqing	P2.20	*Incorporating Pinyin into Pipeline Named Entity Recognition from Chinese Speech*
Zhao, Yingke	P1.20	*A Subband Approach to Personal Sound Zone with Joint Optimization of Sound Pressure and Particle Velocity*
Zhao, Yu-Xaing	P2.11	*Design of Interactive System for Acupoint Analysis Based on Augmented Reality*
Zhao, Yuan	P1.17	*A Computational Efficient Direct Position Determination Approach of Narrow-Band Emitter*
Zhao, Yunhao	P5.26	*GAN-Based Time-Domain Packet Loss Concealment Method with Consistent Mapping Approach*
Zhao, Ziping	P4.25	*Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition*
Zheng, Peijia	P3.30	*AI-Generated Image Detection Using a Cross-Attention Enhanced Dual-Stream Network*
Zheng, Shukai	P2.19	*Complex Feature Information Enhanced Speech Emotion Recognition*
	P2.30	*Few Shot Learning Guided by Emotion Distance for Cross-Corpus Speech Emotion Recognition*
Zheng, Wei-Zhong	E3.1	*An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids*
Zheng, Yixiao	P1.29	*Multi-Frequency Feature Enhancement for Multi-Granularity Visual Classification*
Zheng, Zhaoye	C4.4	*MKTformer: Fine-Grained Meter Classification Based on Multi-Modal Knowledge Transfer*
Zhong, Lifan	P5.8	*Exploring Isolated Musical Notes as Pre-Training Data for Predominant Instrument Recognition in Polyphonic Music*
Zhong, Nan	P4.18	*Preemptive Image Protection Against Steganography*
Zhong, Pei-Lin	H3.2	*Multichannel Learning-Based Spatially Extended Active Noise Control via Model Matching and Sensor Transfer Function Interpolation*
Zhou, Haodong	D1.3	*CASA-Net: Cross-Attention and Self-Attention for End-To-End Audio-Visual Speaker Diarization*
Zhou, Jing	P5.26	*GAN-Based Time-Domain Packet Loss Concealment Method with Consistent Mapping Approach*
Zhou, Jinrun	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Zhou, Xiajie	D4.2	*Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids*
Zhou, Xiaoyi	P1.25	*Hiding Patient Information in Medical Images: A High-Capacity and Reversible Hiding Algorithm for E-Healthcare*

	P1.26	*A Visually Meaningful Image Encryption Algorithm with Attention Mechanism and Artificial Bee Colony Optimization*
Zhou, Xinyuan	P2.21	*Learning Semantic Information from Machine Translation to Improve Speech-To-Text Translation*
Zhou, Yeping	P4.20	*Speech Enhancement with Multi-Granularity Vector Quantization*
Zhou, Yi	P4.19	*Zero-Shot Multi-Speaker Accent TTS with Limited Accent Data*
Zhu, Dongjie	P1.19	*Noise-Robust Pitch Detection Based on Super-Resolution Harmonics*
Zhu, Nan	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Zhu, Qin	D4.1	*A Cloud-Based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis*
Zhu, Qiushi	P4.20	*Speech Enhancement with Multi-Granularity Vector Quantization*
Zhu, Shiqiang	D1.1	*Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation*
Zhu, Weibin	P1.19	*Noise-Robust Pitch Detection Based on Super-Resolution Harmonics*
Zhu, Wenbo	D1.4	*Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction*
Zhuang, Cheng-Xuan	H2.4	*Light-Weight Zero-Reference-Based Image Enhancement for Low-Light Images*
Zhuo, Yin-Rong	P3.19	*DOC: A Novel DOuble-Contour-Based Macro Placement Framework for Mixed-Size Designs*
Zichen, Song	BS1.2	*EEG Study on Anticipation of Difficulty for Upcoming Auditory Task*
Zidni Ilman, Mukhamad	E5.2	*A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera*
Zixun, Guo	E3.3	*Study of Generative Adversarial Networks for Noisy Speech Simulation from Clean Speech*
Zou, Xiaomei	B1.1	*GRALA: Modeling Social Information for Microblog Sentiment Analysis from the View of Balancing Sparsity and Smoothness of Social Contexts*
Zou, Yuexian	P5.24	*NADiffuSE: Noise-Aware Diffusion-Based Model for Speech Enhancement*

PROCEEDING

Mixed Emotion Recognition Based on EEG Signals

Guanxiong Pei[*§], Bingjie Li[†§], Taihao Li[*ʃ], Cunhang Fan[‡], Chao Zhang[‡], Zhao Lv[‡ʃ]
[*]Zhejiang Lab, Hangzhou, China

E-mail: lith@zhejianglab.com Tel: +86-0571-58005295
[†]National University of Singapore, Singapore, Singapore
E-mail: bjlistat@nus.edu.sg Tel: +65-93978342
[‡]Anhui University, Anhui, China
E-mail: kjlz@ahu.edu.cn Tel: +86-0551-63861956
[§]Guanxiong Pei and Bingjie Li contribute equally to the article.
[ʃ]Correspondence: lith@zhejianglab.com, kjlz@ahu.edu.cn

***Abstract*— Advanced emotion monitoring and intervention systems are critical for human-machine cooperation to finish specific tasks in which recognizing mixed emotions is essential. This study is the first to propose intelligent recognition methods of mixed emotions based on EEG signals, thus endowing machines with higher emotional intelligence. We extracted differential entropy (DE), differential asymmetry (DASM), and rational asymmetry (RASM) as features. To better serve the system application and promotion, this study selected three lightweight classifiers, including Random Forest (RF), Probabilistic Neural Networks (PNN), and Generalized Regression Neural Networks (GRNN), all of which achieved good classification performance. The highest classification accuracy is 96.7%. The results of this study can be applied in the fields of mental state assessment, psychological assistance, and clinical psychotherapy.**

I. INTRODUCTION

Emotions are not discretely experienced but often co-occur as mixed emotions simultaneously [1]. Just like mixing painting colors on a palette, mixed emotions are the concrete manifestation of the richness and diversity of human emotions and can more truly reflect the actual emotional state and colorful life experience of individuals [2,3]. In Chinese proverbs, this state is described as “Five flavors in the heart”. As a universal affective experience, mixed emotions are characterized by the co-activation of different dimensions of emotions [1,4,5]. Typically, homogeneously mixed emotions consist of basic emotions with the same valence, such as happiness-pleasure or sadness-fear. In contrast, heterogeneously mixed emotions are blended with basic emotions with different valences, such as happiness-sadness or excitement-fear [6]. From the perspective of discrete emotion models, too many basic emotion combinations lead to low operability and feasibility for further research, such as Plutchik's Wheel of Emotions [7]. Therefore, our research divides mixed emotions into four categories (positive mixed emotions vs. negative mixed emotions vs. neutral mixed emotions vs. heterogeneously mixed emotions) from the perspective of the dimensional emotion model [8].

The recognition of mixed emotions is of great significance. In goal-oriented tasks such as driving a car, working at heights, and emergency rescue, individuals are always in a mixed emotional state. At the same time, they must select and retain task-relevant information and screen out interference from task-irrelevant distractors [9]. It requires balancing adaptation to complex affective states and cognitive control of task-relevant information. In particular, different mixed emotions may have dissociable effects on cognitive control [2,5]. Therefore, the primary goal of advanced emotion monitoring and intervention systems is to develop intelligent methods for mixed emotion recognition, enabling machines to judge and understand complex individual emotional states [10].

Previous studies have confirmed that mixed emotions can be recognized from different perspectives [6,11]. For example, from a behavioral evidence perspective, Larsen and McGraw [6] developed a button press measure for mixed emotions. The results showed that participants who watched the bittersweet movie clips reported mixed emotional experiences. From the perspective of machine vision evidence, one work defines fifteen mixed emotion categories based on the combination of separable action units of facial expressions [11,12]. As the regulating organ of the human body functions and the source of EEG signals, the brain’s neural activity determines the individual’s keystroke behavior and facial expression changes. Therefore, it is possible to describe the patterns and boundaries of mixed emotions based on the electrophysiological evidence. In addition, EEG signals are not easily disguised (so-called social masking) and thus more objective and reliable than behavioral and facial expression data in emotion recognition.

One of the challenging issues for EEG-based emotion recognition using machine learning is a lack of large and available dataset. In particular, there is no open-source EEG dataset of mixed emotion. Therefore, we have to autonomously collect EEG data for algorithm training and testing under the condition of inducing mixed emotional states. There are two main approaches to activate mixed emotions: stimulus-induced and task-induced. The former is a way to present participants

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

with stimuli containing mixed emotional cues, such as movie clips [6,13], paired pictures [14,15] and music [16]. In stimulus induction, participants undergo a mixed-emotional self-assessment procedure either during or after the presentation. There is a need for a scientific manipulation examination of the material's effectiveness in eliciting mixed emotions. Task induction is accomplished by making participants engage in and perform specific tasks, such as performing computer game tasks, freehand drawing tasks, recall tasks, etc., during which participants self-assess their emotions [4,17,18]. The difference between the two methods is whether the participant passively receives the mixed emotional stimulus or actively participates in the eliciting activity. While the latter is considered ecologically more efficient, the former is more mature and standardized. The paired-picture paradigm was used in this study [14,15], with pictures from the International Affective Picture System (IAPS).

The research on mixed emotion classification and recognition based on EEG signals is limited to our knowledge. Therefore, we review the relevant research on basic emotion recognition and signal monitoring system design, hoping to draw valuable lessons. Directly using raw or preprocessed EEG data for classification is computationally intensive, with poor classification effect and low interpretability. Therefore, most studies are based on the inevitable feature extraction of preprocessed EEG signals, followed by classification using classifiers [19,20]. Drawing on previous basic emotion classification and recognition methods, we extract differential entropy (DE), differential asymmetry (DASM), and rational asymmetry (RASM) as features, which are usually used to classify emotions of different valences with outstanding performance [21-23].

To better serve the design of emotion monitoring and intervention systems, this study tried three lightweight classifiers, namely Random Forest (RF), Probabilistic Neural Networks (PNN), and Generalized Regression Neural Networks (GRNN), hoping to take into account both classification performance and computational efficiency, to provide adequate support for future practical applications. RF is a decision tree-based ensemble learning algorithm that can directly process data containing high-dimensional features. RF has high prediction accuracy, low computational overhead, good tolerance to outliers and noise, and is not prone to overfitting [24,25]. PNN combines Bayesian decision theory and density function estimation in radial basis neural networks [26]. During the modeling process, the decision boundary realized by PNN asymptotically approaches the Bayes optimal decision surface. Its structure is simple and easy to implement. It also has the characteristics of fast operation [27]. As a generalization of PNN networks, GRNN networks do not require an iterative process for training, and a high degree of parallelism is presented in their structure [28], which has strong nonlinear mapping ability, local approximation ability, and fast learning ability, as well as the high degree of fault tolerance and robustness [29]. The classification performance of the three methods was compared.

II. Materials And Methods

A. Participants

A total of 25 participants were recruited. Data from two participants were deleted due to excessive recording artifacts. There were 23 valid participants (including 12 females) aged 18-27 (mean = 22.23, SD = 2.20) for final data analysis. They were right-handed, native Chinese speakers with normal or corrected-to-normal vision but no color blindness or weakness. They had no history of neurological or psychiatric disorders. The ethics committee approved the study. Before the experiment, all subjects signed a written informed consent form informing the experiment time, purpose, content, equipment, and procedures. Before the formal experiment began, participants were told they would receive ¥ 45 (~ $ 7.0) as compensation for their participation.

B. Materials

This study employed the paired-picture paradigm [15] and selected stimulus material from the IAPS. The IAPS is widely used in emotional psychology research and contains parameters of various dimensions (e.g., valence, arousal, dominance) obtained from the subjective evaluation results of Western participants. Mixed emotions are elicited and differentiated based on individual evaluations of different pictures, heavily influenced by cultural and social norms [30]. It is inappropriate to directly apply the IAPS rating results of Western participants to Chinese participants. We chose to use the evaluation summary table (704 pictures version) in Mainland China produced by the team of Professor Zhou Renlai from the Department of Psychology of Nanjing University to select the images needed for the experimental stimulus. 120 positive, negative, and neutral images are used, respectively. We performed ANOVA and compared potency for positive, negative, and neutral images. The difference in image valence between the pictures was very significant ($F_{2,357} = 4482.56$, $p < 0.001$), as shown in Table 1. The average valance score was 7.08 (SD = 0.400) for positive images, 5.00 (SD = 0.212) for neutral images, and 2.68 (SD = 0.428) for negative images (Table I).

TABLE I
THE EVALUATION RESULTS OF THE STIMULUS MATERIALS

Index	Measurements	Positive Pictures	Negative Pictures	Neutral pictures
Valence	Mean	7.080	2.680	5.000
	SD	0.400	0.428	0.212
Arousal	Mean	5.750	5.370	4.890
	SD	0.572	0.635	0.736

Index	Measurements	Positive Pictures	Negative Pictures	Neutral pictures
Dominance	Mean	3.640	6.180	4.710
	SD	0.458	0.532	0.501

C. Experimental Procedure

After reading the printed instructions, participants were directed to a soundproofed and electrically shielded room. The stimuli were presented in the center of the computer screen at a viewing angle of 7.50°×5.40°, with a distance of 100 cm. A keyboard was provided for the subjects to complete the Stroop task. The Stroop task has two functions in this study: on the one hand, it is to confirm that participants have completed the entire experimental task seriously; on the other hand, it is to perform emotional erasure to reduce the interference between trials. The experiment consisted of eight blocks, each containing 40 trials. The formal experiment started after ten practice trials to familiarize the participants with the experimental task.

At the beginning of each trial, a fixation appeared in the center of the screen for 500 ms, prompting participants to pay attention. As a priming phase, two emotional stimuli images were presented on the screen for 1500 ms. According to the different emotional valences evoked by the stimulus pictures, it is divided into four different categories: positive mixed emotions (a combination of positive-positive images), negative mixed emotions (a combination of negative-negative images), heterogeneously mixed emotions (a combination of positive and negative images) and neutral mixed emotions (a combination of neutral-neutral images). The word "red" or "green" appeared on the screen for 1000 ms as the target phase after a 500 ms inter-stimulus interval. In some trials, goal-unrelated information was consistent with goal-related information (e.g., the word "green" was printed in green ink); in other trials, the two were inconsistent (e.g., the word "green" was printed in red ink). Then, the prompt sentence "Please click on the text color" appeared. Participants were asked to judge the color of the text by pressing "1" or "3" as quickly as possible. The colors corresponding to the buttons were balanced among the participants. The inter-trial interval is 500 ms. Participants had 2-3 min between blocks to rest and relax (Fig. 1).

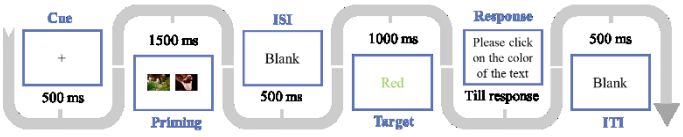


Fig. 1 Trial flow chart. It showed the procedure of one trial, and there were 320 trials in total.

Note. ISI = inter-stimulus interval; ITI = inter-trial interval.

D. Behavioral Data Acquisition and Analysis

The main purpose of our detection of behavioral data is to confirm that the subjects participated in this experiment seriously. Behavioral data were collected using E-prime 2.0 software (Psychology Software Tools Inc., Pittsburgh, PA, USA) and exported. After obtaining the original behavioral data, we used SPSS 25.0 statistical software (SPSS, Inc., Chicago, IL, USA) to perform the repeated-measures analysis of variance (ANOVAs) [4 (priming conditions) × 2 (Stroop conditions)].

E. EEG Data Acquisition and Analysis

Raw electroencephalographic (EEG) data were acquired using a Neuroscan Synamp2 Amplifier (Neurosoft Labs, Inc., Sterling, Virginia, USA) with Ag/AgCl electrodes placed at 64 scalp sites according to the extended international 10-20 system, with an online bandpass of 0.5-70 Hz and at a sampling rate of 500 Hz. The left mastoid served as the online reference, and the EEG was re-referenced offline to the mathematically averaged mastoids. Two pairs of electrodes recorded vertical and horizontal electrooculograms (EOGs). One pair was placed above and below the left eye parallel with the pupil, and the other was placed 10 mm from the lateral canthi. Impedances were kept below 5 kΩ throughout the experiment. EEG data were further processed offline using the Matlab EEGLab toolbox (Mathworks, Inc., Natick, MA, USA). Bandpass filters the EEG data at 0.5-50 Hz using a zero-phase-shift FIR filter. A notch filter removes electrical interference from 50 Hz line noise [31]. EEG data were down-sampled at a sampling frequency of 256 Hz to speed up calculations. Brain-independent artifacts such as significant muscle activity and eye movements were removed using independent component analysis [32]. The signal was then segmented for the emotion recognition phase to analyze the period from 500 ms before the presentation of mixed pictures to 1500 ms after the presentation. We divide the [0 ms, 1500 ms] data into 15 sub-epochs of 100 ms. Hanning filters are used to extract the delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-49 Hz) bands. Baseline correction was applied to each frequency band in a time window from -500 to 0 ms relative to stimulus onset. Three features (DE, DASM, and RASM) were extracted. DE can be used to describe the chaotic nature of EEG signals as entropy the degree of variability and complexity of the system. DE is equivalent to the logarithmic energy spectrum, where the time series X obeys the Gaussian distribution $N(\mu,\delta 2)$.

$$h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx = \frac{1}{2} log(2\pi e\sigma^2) \quad (1)$$

DASM and RASM are the differences and ratios between DE of 27 pairs of hemispheric asymmetry electrodes (left-right: *FP1-FP2, F7-F8, F3-F4, FT7-FT8, FC3-FC4, T7-T8, P7-P8, C3-C4, TP7-TP8, CP3-CP4, P3-P4, O1-O2, AF3-AF4, F5-F6, F7-F8, FC5-FC6, FC1-FC2, C5-C6, C1-C2, CP5-CP6, CP1-CP2, P5-P6, P1-P2, PO7-PO8, PO5-PO6, PO3-PO4, and CB1-CB2*), which can be calculated as

$$DASM = h(X_i^{left}) - h(X_i^{right}) \quad (2)$$

$$RASM = h(X_i^{left})/h(X_i^{right}) \quad (3)$$

where $h(X)$ is defined in Equation 1, and i is the pair number.

The processed features were further applied to mixed emotion classification (positive mixed emotions vs. negative mixed emotions vs. neutral mixed emotions vs. heterogeneously mixed emotions) using RF, PNN, and GRNN. RF is an ensemble learning method. In this study, the outputs of multiple decision trees were combined to obtain a single result, which was trained using bagging. In our model, the number of trees in the forest is 100, and the maximum depth of the trees is 500. PNN is a neural network technique based on Bayesian decision rules. In this study, PNN includes four network layers: input, pattern, summation, and output. When an input is presented, the pattern layer computes distances from the input vector to the training input vectors. It produces a vector whose elements indicate how close the input is to a training input. The summation layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities. Finally, a compete transfer function on the output of the summation layer picks the maximum of these probabilities, and produces a 1 for that class and a 0 for the other classes.

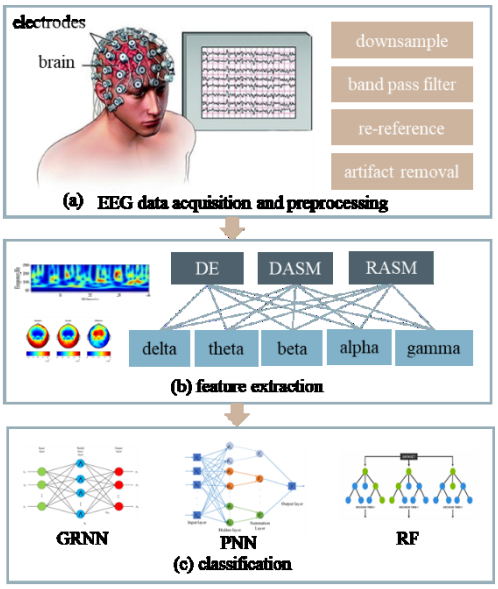


Fig. 2 Flow chart of EEG data processing for mixed emotion classification

GRNN is a generalization of PNN networks. Different from PNN, GRNN adopts linear output instead of competitive output. The input layer receives the input signal and transmits it to the hidden layer; the hidden layer is composed of basis functions, which produce different responses to the input signal; the output layer is composed of linear functions. We applied the 10-fold cross-validation strategy to estimate the generalization ability of our classifiers. We selected several performance indicators for the classification models, including accuracy, precision, recall, and F1 (macro-averaged F1-score).

III. Results

A. Behavioral Results

We conducted a two-way 4 (mixed emotion priming conditions) × 2 (Stroop conditions) repeated measures ANOVA for reaction accuracy (see Table II). There was a significant main effect of the Stroop condition ($F_{1,22} = 8.751$, $p = 0.007$), and the accuracy rate in the congruent condition (M = 0.995, SD = 0.012) was significantly greater than that in the incongruent condition (M = 0.985, SD = 0.024), while the main effect of the priming condition ($F_{3,20} = 1.366$, $p = 0.282$) and the interaction effect of these two factors ($F_{3,20} = 0.349$, $p = 0.790$) were not significant.

TABLE II
DESCRIPTIVE STATISTICS OF REACTION ACCURACY UNDER DIFFERENT CONDITIONS

Conditions	Min	Max	Mean	SD
negative - incongruent	0.93	1.00	0.985	0.022
negative - congruent	0.95	1.00	0.993	0.014
neutral - incongruent	0.90	1.00	0.985	0.033
neutral - congruent	0.98	1.00	0.999	0.005
heterogeneous - incongruent	0.95	1.00	0.989	0.017
heterogeneous - congruent	0.98	1.00	0.996	0.010
positive - incongruent	0.93	1.00	0.982	0.026
positive - congruent	0.95	1.00	0.991	0.018

B. EEG Results

The performance of classifiers based on features from different frequency bands is shown in Table III. Mixed emotion classification methods proposed have achieved good classification results, and the classification accuracy can generally reach 96%. Significantly, based on the gamma frequency band features, the classification accuracy of PNN reached the highest, 96.7%. In addition, the features from the various frequency bands do not show significant differences in mixed emotion classification. Generally speaking, the accuracy has reached the expected effect and has good applicability.

TABLE III
PERFORMANCE OF CLASSIFIERS IN DIFFERENT FREQUENCY BANDS

Classifier	band	Accuracy	Precision	Recall	F1
RF	delta	0.961	0.923	0.921	0.922
	theta	0.961	0.925	0.875	0.895
	alpha	0.963	0.925	0.924	0.924
	beta	0.965	0.934	0.876	0.899
	gamma	0.963	0.929	0.876	0.896
GRNN	delta	0.962	0.923	0.912	0.917
	theta	0.963	0.927	0.876	0.897
	alpha	0.964	0.928	0.923	0.925

Classifier	band	Accuracy	Precision	Recall	F1
PNN	beta	0.960	0.916	0.882	0.897
	gamma	0.963	0.926	0.925	0.925
	delta	0.951	0.921	0.860	0.861
	theta	0.958	0.921	0.918	0.909
	alpha	0.961	0.923	0.939	0.925
	beta	0.965	0.924	0.954	0.935
	gamma	0.967	0.922	0.962	0.939

IV. DISCUSSION

Mixed emotion recognition embodies advanced emotional intelligence in emotion monitoring and intervention systems. This study is the first to propose an intelligent recognition method of mixed emotions based on EEG signals so that the machine can perceive and identify one's mixed emotion state.

For the behavioral data results, the reaction accuracy under each condition was greater than 93%, indicating that all participants accurately understood the experimental task and completed it with high quality, with relatively intense concentration throughout the experiment. The reaction accuracy results show a significant main effect of the Stroop condition. The accuracy in the congruent condition was significantly greater than that in the incongruent condition. This result is consistent with previous studies [33,34], where semantic interference with color judgment has significantly reduced the subject keypress task's accuracy. This interference effect reflects greater competition for attentional and cognitive resources in response to incongruent conditions. The interaction of mixed emotion priming and Stroop condition was not significant. Mixed emotions may interfere with individual cognitive control [2,4]. However, because the Stroop task is relatively easy to implement, forming an explicit representation in behavioral data may be difficult.

For EEG feature extraction, existing studies mainly focus on the frequency domain of EEG signals [21,23,35] while failing to utilize dynamic temporal changes and electrode channel information. We fill this gap using DE features incorporating time and frequency domain information. Meanwhile, DASM and RASM further contain spatial information, exploiting the relationship between different brain hemispheres to learn more informative and discriminative representations. The composite feature extraction based on time domain, frequency domain, and spatial domain information is one of the characteristics of this research. The results show that different classifiers based on the above features generally achieve good classification results, indicating that this study's feature extraction is successful.

It is worth mentioning that, unlike previous studies on pure emotion classification and recognition [21,23], the classification accuracy of mixed emotions does not show significant differences in different frequency band features. This may be related to the mixed state of emotions. When the brain is in a mixed emotional state, it may generate EEG signals with multispectral activity, potentially valuable in emotion classification. This is an important discovery that distinguishes mixed emotion recognition from pure emotion recognition, reflecting the complex and chaotic state of the brain induced by mixed emotions.

The classifiers used in this study are RF, PNN, and GRNN. The common advantage of the three classifiers is the combination of high computational efficiency and classification performance. The RF training speed is fast because the trees are independent of each other during training, so that they can be implemented in parallel. PNN is based on a Bayesian strategy feed-forward neural network, which has the advantages of easy training and fast convergence. GRNN has strong nonlinear mapping ability and learning speed and has the characteristics of self-adaptive determination of the structure. The output has nothing to do with the initial weight. There are few artificially adjusted parameters, and the network's learning mainly depends on data samples, which determines that the network can avoid the influence of subjective human assumptions on the prediction results to the greatest extent. The three classifiers are lightweight, which is helpful for the wide-scale application and promotion of the research results. Combined with different application scenarios and software systems, the classification model proposed in this study can be selectively used.

Mixed emotion recognition using EEG signals is of great interest and wide application, especially its crucial role in human-machine collaboration. For example, mixed emotions may profoundly affect cognitive flexibility and control while driving a car [36]. We can design an adaptive driver-assistance system to dynamically detect their emotional and cognitive states and react to drivers to adjust their emotions to improve driving safety. In addition, mixed emotion recognition may be applied in multiple fields, including the potential economic and social value in human-computer interaction, brain-computer interface, psychological assistance, and clinical psychotherapy.

An interesting direction for future research would be to develop multimodal frameworks for mixed emotion recognition and neural mechanism delineation. It is illustrated that one modality's weaknesses are offset by another's strengths, which can effectively increase emotion recognition accuracy and robustness compared to unimodal frameworks [37,38]. Taken together, the results of this study provide important insights into the mixed emotion recognition and priming effects on cognitive control, which are helpful for us to build a more advanced emotion monitoring and intervention system, thereby enhancing the ability to resist distraction and temptations,

execute voluntary and goal-directed behavior, and produce satisfactory performance.

ACKNOWLEDGMENT

This work was supported by the Zhejiang Provincial Natural Science Foundation of China [grant number LQ22C090007], the National Natural Science Foundation of China [grant number T2241018], the National Science and Technology Major Project of the Ministry of Science and Technology of China [grant number 2021ZD0114303], the Scientific Project of Zhejiang Lab [grant number 2020KB0AC01], the Open Research Project of Key Laboratory of Brain-Machine Intelligence for Information Behavior (Ministry of Education and Shanghai) [grant number 2022KFKT002].

REFERENCES

[1] F. Ali, and S. C. Tan, "Emotions and lifelong learning: synergies between neuroscience research and transformative learning theory," *International journal of lifelong education* vol. 41, pp.76-90, 2022.

[2] S. D. Kreibig, and J. J. Gross, "Understanding mixed emotions: paradigms and measures," *Current opinion in behavioral sciences* vol. 15, pp.62-71, 2017.

[3] E. Burkitt, D. Watling, and F. Cocks, "Mixed emotion experiences for self or another person in adolescence," *J Adolescence* vol. 75, pp.63-72, 2019.

[4] R. Berrios, P. Totterdell, and S. Kellett, "Eliciting mixed emotions: a meta-analysis comparing models, types, and measures," *Front Psychol* vol. 6, 428, 2015.

[5] S. E. Henderson, and C. J. Norris, "Counterfactual thinking and reward processing: An fMRI study of responses to gamble outcomes," *Neuroimage* vol. 64, pp.582-589, 2013.

[6] J. T. Larsen, and B. J. Stastny, "It's a bittersweet symphony: Simultaneously mixed emotional responses to music with conflicting cues.," *Emotion* vol. 11, 1469, 2011.

[7] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to Emotion* vol. 1984, pp.2-4, 1984.

[8] L. F. Barrett, "Discrete emotions or dimensions? The role of valence focus and arousal focus," *Cognition Emotion* vol. 12, pp.579-599, 1998.

[9] S. Xue, Y. Li, X. Kong, Q. He, J. Liu, and J. Qiu, "The dissociable neural dynamics of cognitive conflict and emotional conflict control: An ERP study," *Neurosci Lett* vol. 619, pp.149-154, 2016.

[10] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Comput Hum Behav* vol. 65, pp.267-275, 2016.

[11] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences* vol. 111, pp.E1454-E1462, 2014.

[12] S. Du, and A. M. Martinez, "Compound facial expressions of emotion: from basic research to clinical applications," *Dialogues Clin Neuro* vol. 17, 443, 2015.

[13] I. Mavrou, and J. M. Dewaele, "Emotionality and pleasantness of mixed-emotion stimuli: The role of language, modality, and emotional intelligence," *International Journal of Applied Linguistics* vol. 30, pp.313-328, 2020.

[14] U. Schimmack, and S. Colcombe, "Eliciting mixed feelings with the paired-picture paradigm: A tribute to Kellogg," *Cognition and Emotion* vol. 21, pp.1546-1553, 2007.

[15] F. Pan, Q. Lu, Y. Chen, X. Wu, and Q. Li, "Emotional conflict occurs at a late stage: evidence from the paired-picture paradigm," *Transl Neurosci* vol. 7, pp.56-61, 2016.

[16] P. G. Hunter, E. G. Schellenberg, and U. Schimmack, "Mixed affective responses to music with conflicting cues," *Cognition Emotion* vol. 22, pp.327-352, 2008.

[17] E. Burkitt, and D. Watling, "How do children who understand mixed emotion represent them in freehand drawings of themselves and others?" *Educ Psychol-Uk* 36, pp.935-955, 2016.

[18] D. L. Cothran, and R. Larsen, "Comparison of inhibition in two timed reaction tasks: The color and emotion Stroop tasks," *The Journal of psychology* vol. 142, pp.373-385, 2008.

[19] M. R. Islam, M. M. Islam, M. M. Rahman, C. Mondal, S. K. Singha, M. Ahmad, A. Awal, M. S. Islam, and M. A. Moni, "EEG channel correlation based model for emotion recognition," *Comput Biol Med* vol. 136, 104757, 2021.

[20] P. Patel, and R. N. Annavarapu, "EEG-based human emotion recognition using entropy as a feature extraction measure," *Brain informatics* vol. 8, 20, 2021.

[21] R. Duan, J. Zhu, and B. Lu, "Differential entropy feature for EEG-based emotion classification," *In 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81-84, 2013.

[22] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, "Asymmetric spatial pattern for EEG-based emotion detection," *In The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7, 2012.

[23] G. Pei, B. Li, T. Li, R. Xu, J. Dong, and J. Jin, "Decoding Emotional Valence from EEG in Immersive Virtual Reality," *In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, (IEEE, 2022), pp. 1469-1476, 2022.

[24] G. Biau, and E. Scornet, "A random forest guided tour," Test-Spain vol. 25, pp.197-227, 2016.

[25] R. Prytz, S. Nowaczyk, T. Rögnvaldsson, and S. Byttner, "Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data," *Eng Appl Artif Intel* vol. 41, pp.139-150, 2015.

[26] M. R. Berthold, and J. Diamond, "Constructive training of probabilistic neural networks," *Neurocomputing* vol. 19, pp.167-183, 1998.

[27] M. Kusy, and P. A. Kowalski, "Weighted probabilistic neural network," *Inform Sciences* vol. 430, pp.65-76, 2018.

[28] J. Qi, G. Jiang, G. Li, Y. Sun, and B. Tao, "Surface EMG hand gesture recognition system based on PCA and GRNN," *Neural Computing and Applications* vol. 32, pp.6343-6351, 2020.

[29] Y. Q. Ni, and M. Li, "Wind pressure data reconstruction using neural network techniques: A comparison between BPNN and GRNN," *Measurement* vol. 88, pp.468-476, 2016.

[30] T. Brosch, G. Pourtois, and D. Sander, "The perception and categorisation of emotional stimuli: A review," *Cognition and emotion* vol. 24, pp.377-400, 2010.

[31] S. J. Luck, "Ten simple rules for designing and interpreting ERP experiments," *Event-related potentials: A methods handbook* vol. 4, 2005.

[32] M. Chaumon, D. V. Bishop, and N. A. Busch, "A practical guide to the selection of independent components of the electroencephalogram for artifact correction," *J Neurosci Meth* vol. 250, pp.47-63, 2015.

[33] D. Algom, and E. Chajut, "Reclaiming the stroop effect back from control to input-driven attention and perception," *Front Psychol* vol. 10, 1683, 2019.

[34] F. Scarpina, and S. Tagini, "The stroop color and word test," *Front Psychol* vol. 8, 557, 2017.

[35] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, D. Hazry, and I. Zunaidi, "Time-frequency analysis of EEG signals for human emotion detection,", *In 4th Kuala Lumpur International Conference on Biomedical Engineering 2008* (Springer, 2008), pp. 262-265, 2008.

[36] N. Habibifar, and H. Salmanzadeh, "Relationship between driving styles and biological behavior of drivers in negative emotional state," *Transportation research part F: traffic psychology and behaviour* vol. 85, pp.245-258, 2022.

[37] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE T Cybernetics* vol. 49, pp.1110-1122, 2018.

[38] G. Pei, and T. Li, "A literature review of EEG-based affective computing in marketing," *Front Psychol* vol. 12, 602843, 2021.

Attention-based CNN and Relative Phase Feature Modeling for Improved Imagined Speech Recognition

Yoshiki Niimura*, Jun Takemoto*, Atsuhiko Kai* and Seiichi Nakagawa†
* Graduate School of Integrated Science and Technology, Shizuoka University, Japan
E-mail: {niimura.yoshiki.19, kai.atsuhiko}@shizuoka.ac.jp
† Chubu University, Japan
E-mail: nakagawa@isc.chubu.ac.jp

***Abstract*—In EEG (Electroencephalography) based imagined speech recognition research, deep learning is gradually being applied, but has not yet achieved sufficient performance. The main reasons for this are the lack of ability to acquire generalized feature representations and the insufficient amount of training data. For the former, we propose two methods, one is a CNN-based feature extractor that introduces an attention mechanism, and the other is an improved feature extractor that introduces the inter-channel relative phase obtained from the multi-channel EEG signal as an additional spectral feature instead of a CNN-based feature extractor. For the latter, we propose a pre-training method using multi-subject EEG data to construct subject-adapted models for an imagined speech recognition task. The effectiveness of the proposed methods is verified by using an imagined speech recognition task for English phonemes, classification for English vowels and consonants. As a result, the CNN model with attention mechanism successfully improved the feature extraction capability of the CNN and achieved a classification accuracy rate of 71.9%, which are approximately 20% higher than that of the CNN only. On the other hand, a simple neural network based classifier approach using spectral features as input achieved 71.1% by using the proposed inter-channel relative phase features over the conventional relative phase features, and 75.0% by applying ensemble processing to the test data. Pre-training and subject adaptation also achieved a 6% improvement in accuracy, indicating that it is possible to mitigate the impact of insufficient training data.**

Index Terms: EEG, imagined speech, CNN, attention, transfer learning, relative phase

I. Introduction

Some patients with severe motor paralysis have paralysis of the speech organs, making communication by speech impossible. Imagined speech recognition, which recognizes words from EEG when the patient performs a speech imagining action without uttering the words aloud, has been studied as a technology that can provide such patients with an alternative means of communication to speech, and research is currently being conducted on methods to improve recognition accuracy for practical applications.

Studies on EEG (Electroencephalography) recognition and decoding are gradually making use of deep learning frameworks [1]. In addition, CNNs have been used for artifact noise removal, which is an important part of such EEG-based methods [2]. However, few studies have demonstrated the effectiveness of deep learning in the task of imagined speech recognition. In the past, methods for extracting features of phonemes, syllables, and words from brain waves during speech imagery have used analytical features such as statistics from raw signals or spectral features such as DWT and provide them as input to a discriminative model [3][4][5]. Recently, however, deep learning models have shown higher accuracy than conventional machine learning methods in the fields of speech and image recognition, and research on deep learning models is increasing in this field. Among deep learning models, there has been a particularly large amount of researches on CNNs, where feature extraction can be omitted by directly inputting multi-channel EEG signals as 2D data, as in computer vision and speech recognition applications, and CNNs specializing in EEG have also been proposed [6][7]. However, there are limited reports on the effectiveness of deep learning models for imagined speech recognition tasks, while there have been many reports on the effectiveness of deep learning models for other EEG-based classification tasks, such as event related potential (ERP) and motor imagery.

As an example of research using CNNs, Clayton et al. [8] reported that SVM had a recognition accuracy of 69.0% for the classification between English vowels and consonants for imagined speech, whereas CNN was about 49.0%, that was chance rate. In addition, Kamble et al. using machine learning methods reported results of 36.7% for the identification of EEG during imagining Spanish five vowels [9], and Cooney et al. using CNN reported results slightly lower at 35.6% [10].

This can be attributed to two problems: incomplete feature extraction capability of the trained CNN model and insufficient amount of training data. Due to the low signal-to-noise ratio of EEG signals in the speech imagery task, the features of the imagined speech are buried in noise, and a typical CNN consisting of a convolutional layer and a pooling layer is not able to acquire feature extraction well [7]. In addition, the amount of training data is insufficient for CNNs, as conventional methods are trained on one specific person's data due to the subject dependence of EEG [8].

To improve the feature extraction capability, this paper pro-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

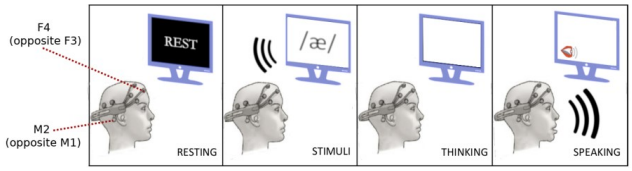


Fig. 1. Illustration of the recording procedure [8]. Participants listen to five repetitions of a phone (recorded in their own voice), then imagine speaking the phone five times (with the same rhythm), then overtly speak the phone five times.

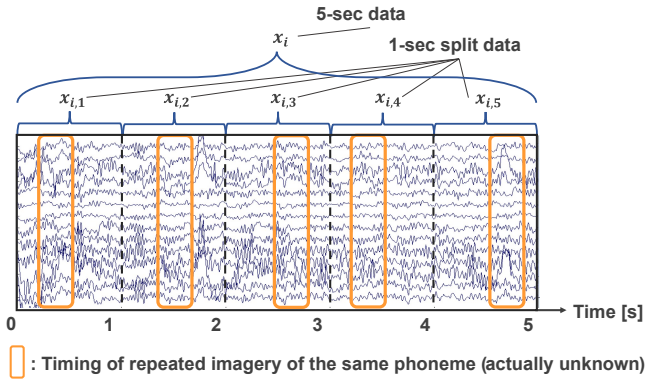


Fig. 2. Two different definitions of sample units for the "thinking" step assumed in our classification task (5-sec v.s. 1-sec split). Although it is unclear at what timing the features of EEG appear during imaging, we assume that each features imagining 5 times repeated within the 5-sec data would manifest within 1 second.

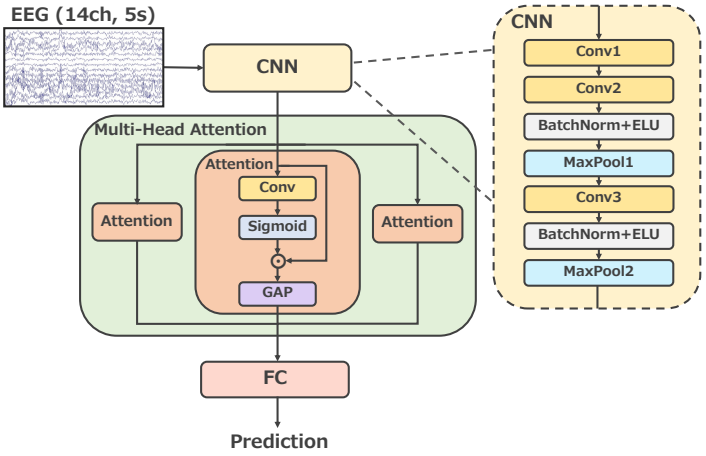


Fig. 3. Proposed classifier with attention-based CNN. CNN:Convolutional Neural Network, GAP:Global Average Pooling, FC: Fully Connected Layer.

TABLE I
PARAMETERS OF THE CNN PART

Layers	Parameters
Conv1	n_filters = 5, filter_size = (1, 10), stride = 1
Conv2	n_filters = 50, filter_size = (ch_n, 1), stride = 1
MaxPool1	kernel_size = (1, 3), stride = 3
Conv3	n_filters = 15, filter_size = (1, 5), stride = 1
MaxPool2	kernel_size = (1, 3), stride = 3

poses two methods: one is to introduce an attention mechanism into the CNN-based feature extractor, and the other is to use an improved feature extractor that introduces as an additional spectral feature inter-channel relative phase obtained from the multi-channel EEG signal instead of a CNN-based feature extractor. Attention is a mechanism that automatically learns which of the intermediate feature outputs of a deep learning model are important, and our intention in this study is to input the features output by the CNN into the attention mechanism to emphasize the features of recalled speech that are expected to be observed only for a short period of time. To compensate for the lack of ability in feature representation learning, we also propose a method to improve the spatial feature extraction of EEG with relative phase feature, which is an improved method of our previous work [11]. To cope with the lack of training data, pre-training and subject adaptation are proposed. The amount of insufficient training data is compensated by pre-training with multiple subjects' data, and the model is trained again with the target subject's data to cope with subject dependence. The effectiveness of these methods is verified in this study.

II. DATASET

The dataset used in this study [8] contains EEG data from 21 subjects using a commercial-grade light-weight mobile headset EEG device (Emotiv EPOC+). The EEG signals were recorded using the 14 channels with sampling frequency of 256Hz. There are 160 samples for each subject, and each sample is the EEG of a trial consisting of four steps ("resting", "stimulus", "thinking", and "speaking") of 5 seconds each (Fig. 1). During the "stimulus" step, a randomly selected English phoneme is presented repeatedly for 5 times. Next, in the "thinking" step, subjects are instructed to imagine the same phoneme being repeated five times as it was played in the "stimulus" step. There are 16 English phonemes to be imagined (4 vowels and 12 consonants), with 10 sample trials per phoneme recorded for each subject. Only the 5-second portion of the "thinking" step of the dataset is used in our experimental task. Also, since this dataset contains subjects with significant noise contamination, only data from 12 subjects, excluding those, are used.

In this study, each trial in this dataset is treated in two different ways (Fig. 2). In the experiment using a convolutional neural network (section III-A), the EEG data imagined five times in 5 second is used as input as it is. Also, in the experiment using spectral features (section III-C), each trial data is divided into 5 parts for increasing the training samples. These divided parts are treated as separate samples for training and classification. Although it is unclear at what timing the features of EEG appear during imaging, we assume that each features imagining 5 times repeated within the 5-sec data would manifest within 1 second.

III. METHOD

A. Attention-Based CNN

Attention is an architecture that is used for deep learning models and allows them to learn which features are important and highlight them among the intermediate outputs of the deep learning model [12]. In this study, an attention is placed after CNN to capture the moment of occurrence of the imagined speech features in the entire EEG features. The structure of the model used in this study is shown in Fig. 3. The CNN part that plays the role of Encoder is based on the CNN structure used in the previous study [7]. In Ref. [8], they reported a loss of accuracy due to insufficient number of training samples, so we made adjustments to the number of layers and the parameters of each layer. The details of the configuration of the CNN part are shown in Table I. Convoluting the input 2D EEG data consisting of electrode dimension (14 channels) and time dimension (5 seconds) is performed by three convolution layers and two pooling layers. Then, an attention is placed after the last pooling layer.

The model for the attention block in Fig. 3 is based on [13] and composed of a convolutional layer with a filter size of 1×1, a sigmoid layer, and an average-pooling layer. First, the attention block generates attention mask $\mathbf{M}$. Let $\mathbf{M}$ be the attention mask for a certain filter of the final convolutional layer, the attention mask $\mathbf{M}$ is obtained as follows:

$$\mathbf{M} = \sigma(f^{1\times1}(\mathbf{F}^{H\times W})) \tag{1}$$

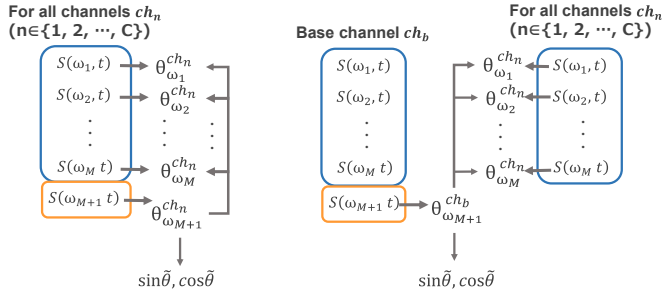
where $\mathbf{F}^{H\times W}$ is the input feature (H:14 ch, W:1280 (5-sec×256 Hz)), $f^{1\times1}$ is the convolution with filter size 1×1, and σ is the sigmoid function. $\mathbf{M}$ has the same $H \times W$ size as the input features, and update with learning to highlight the important feature. Then, important features are highlighted as:

$$\mathbf{F'}^{H\times W} = \mathbf{F}^{H\times W} \odot \mathbf{M} \tag{2}$$

Finally, global average pooling is applied to compress each feature map to a size of 1×1. In addition, this study also employs a structure called multi-head attention [14]. Assuming the number of attention heads is denoted as H_{Att} and the number of filters as C, the $H_{Att} \times C$ features are passed to the final linear layer, that is, fully connected (FC) layer.

B. CNN Model Pre-training and Subject Adaptation

We pre-train the CNN model with data of multiple subjects to compensate sufficient training data and fine-tune (FT) it with the data of a target subject for subject adaptation. When subject $n(1 \leq n \leq N)$ is selected from the N subjects dataset $\{X_1, \cdots, X_N\}$, the target domain (subject adaptation) dataset is defined as $D_{\mathbf{T}} = \{X_n\}$. The source domain (pre-training) dataset is the combination of all subject data except for the target subject and defined as $D_{\mathbf{S}} = \cup_{i\neq n}\{X_i\}$. The pre-training task and subject adaptation task are set to the same classification (vowel or consonant). In this study, all parameters of the pre-trained model is not fixed in FT. Subject adaptation allows the pre-trained model to take care of the dependence on the target subject.



(a) Intra-channel(fixed base-freq) (b) Inter-channel with fixed base-channel & freq (ICRP_FF)

Fig. 4. Relative phase features of EEG signals (conventional methods). $(M + 1)$ is the base frequency bin.

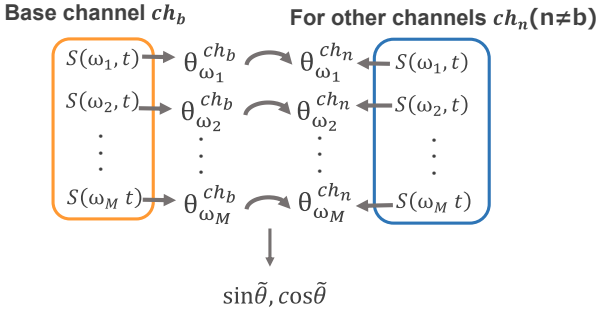


Fig. 5. Proposed inter-channel relative phase within the same frequency bins (ICRP_SF) of multi-channel EEG signals

C. Neural Network Using Spectral Features

We propose a classification method using improved front-end feature extraction, as a comparison to the attention-based CNN model. In conventional speech signal processing, Mel-Frequency Cepstral Coefficients (MFCC) are primarily used as the feature representation. However, MFCC only utilizes the amplitude spectrum and ignores the phase spectrum of the obtained spectrogram. In Ref. [15][16], the effectiveness of phase has been shown in speaker classifier modeling. In the feature representation of multi-channel EEG signals, the relative phase feature is considered effective in extracting spatial feature representation.

For per-channel EEG feature extraction, the amplitude and phase spectra are computed for each frequency bin in the analysis interval using the Discrete Fourier Transform (DFT). This can be expressed as follows:

$$S(\omega, t) = A(\omega, t) \times e^{j\theta(\omega,t)} \tag{3}$$

Here, $A(\omega, t)$ and $\theta(\omega, t)$ represent the amplitude spectrum and the phase spectrum. The phase spectrum depends strongly on the position of window, so we use the relative phase. In the case of relative phase, the phase of a certain frequency component is used as a reference, and the phases at other frequencies are calculated relative to this reference phase using the following equation:

$$\tilde{\theta}(\omega', t) = \theta(\omega', t) - \frac{\omega'}{\omega}\theta(\omega, t) \tag{4}$$

TABLE II
FEATURE REPRESENTATIONS COMPARED FOR SIMPLE NEURAL NETWORK CLASSIFIER SETTING (FEATURE "CONV-RP" REPRESENTS ONE OF THE TWO CONVENTIONAL RELATIVE PHASE FEATURE EXTRACTION METHODS SHOWN IN FIGURE 4)

Features	Input Dimensions
Amplitude	140 (14ch×5 dim×2(mean,std))
CONV-RP	280 (14ch×5 dim×2(sin, cos)×2 (mean,std))
ICRP_SF	260 (13ch×5 dim×2(sin, cos)×2 (mean,std))
Amplitude + CONV-RP	420 (140 + 280)
Amplitude + ICRP_SF	400 (140 + 260)

To address the issue of phase discontinuity, we utilize the corresponding coordinate values $[\cos\tilde{\theta}, \sin\tilde{\theta}]$ instead of $\tilde{\theta}$. In previous studies on imagined speech recognition, Sakai et al. conducted research using relative phase information. They set the reference frequency as one higher order than the frequency range of interest and utilized the relative phase of the lower-order 5 dimensions (8 to 40 Hz). As a result, they reported an accuracy of 51.5% in a binary classification task using only relative phase (Fig. 4a) and 58.5% using amplitude spectrum and relative phase features together [11]. Sakai et al. also proposed inter-channel relative phase with fixed base channel and base frequency (Fig. 4b) to capture the phase difference between each channel and reported minor improvement.

In this study, our intention is to extend the phase feature representation to capture the spatial features of multi-channel EEG signals, that is, correlation between channels. Therefore, we focus on the inter-channel relative phase within the same frequency bins (ICRP_SF, see Fig. 5). As a baseline for the EEG classifier, we compare with amplitude and phase features and investigate the effectiveness of combining these features. Let designate ch_b as the base channel and ch_n as the channel for which the relative phase is calculated. The base frequency is identical to the frequency used for calculating the relative phase. Therefore, we can express it as follows, using the formula for relative phase (Eq. 4):

$$\begin{aligned}\tilde{\theta}^{ch_n}(\omega', t) &= \theta^{ch_b}(\omega', t) - \frac{\omega'}{\omega}\theta^{ch_n}(\omega, t) \\ &= \theta^{ch_b}(\omega, t) - \theta^{ch_n}(\omega, t)\end{aligned} \tag{5}$$

It is a simple phase subtraction in each frequency dimension. Therefore, the case $ch_b = ch_n$ is excluded (because the relative phase is zero). We calculate DFT with a window length of 125ms and utilize the average and standard deviation along the time dimension of the lower-order 5-dimensional (8 to 40 Hz) amplitude and phase spectra as feature parameters. Table II shows a comparison of the types of feature parameters and their dimensions. When calculating the relative phase, we exclude the target channel if it is the same as the base channel (calculated relative phase becomes 0). This features is given as input to a linear layer with 100 nodes in the hidden layer for classification.

IV. RESULTS

We examined the recognition accuracy of a binary classification task, in which the task is to recognize from EEG whether an imagined phoneme is a vowel or a consonant. Although there were originally 160 samples per subject, the number of samples for consonants was reduced to 40, the same as vowels, for a total of 80 samples, in order to eliminate class bias (consonants:vowels = 120:40 → 40:40). The number of multi-head attention H_{Att} was set to 5 in our experiments since preliminary experiments showed the best performance. However, almost the same performance was obtained with $H_{Att} = 1$, indicating that multi-head does not significantly affect the performance of this experimental task.

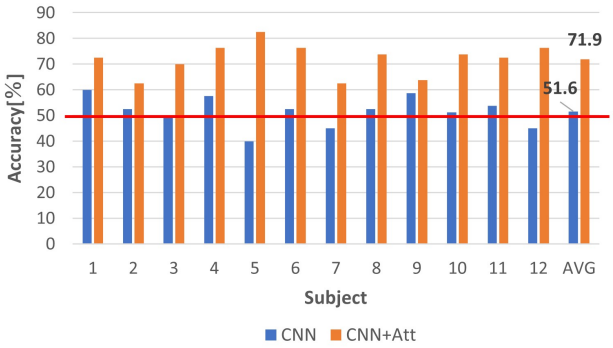


Fig. 6. Classification accuracy of the English phoneme by the typical CNN and the attention-based CNN (red line: chance rate)

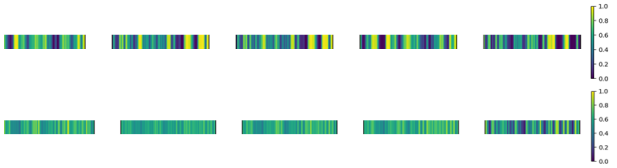


Fig. 7. Example of attention mask generated for correct (top) and incorrect (bottom) predicts

A. Attention-Based CNN

In the baseline CNN and CNN+Attention models, raw EEG data is utilized. Therefore, the input consists of 2D data with dimensions of 14 channels × 1280 (256Hz × 5s). Training and evaluation were performed independently for each subject, and 5-fold cross validation was applied.

The results of vowel and consonant classification using the English phoneme dataset from subjects 1 to 12 with the baseline CNN and the CNN+Attention models are shown in Fig. 6. The CNN+Attention model outperforms the baseline CNN model in terms of classification accuracy for all subjects. In most subjects, the CNN+Attention model achieves a classification accuracy that is approximately 10-20% higher than the baseline CNN model (in other words, the error reduction rate is approximately 20-40%). Notably, even though the baseline CNN falls below 50% accuracy for subject 5, the CNN+Attention model achieves an accuracy of over 80%. Finally, Fig. 7 shows the results of visualizing the attention masks generated in the attention mechanism for correct and incorrect responses. It can be seen that there are repeated

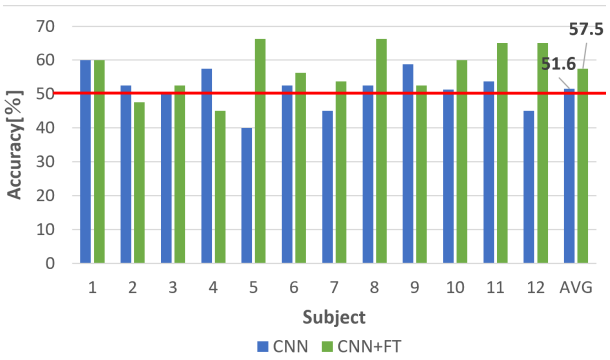


Fig. 8. Classification accuracy of the English phoneme by the CNN with pre-training+subject adaptation (FT) and the CNN without them(red line: chance rate)

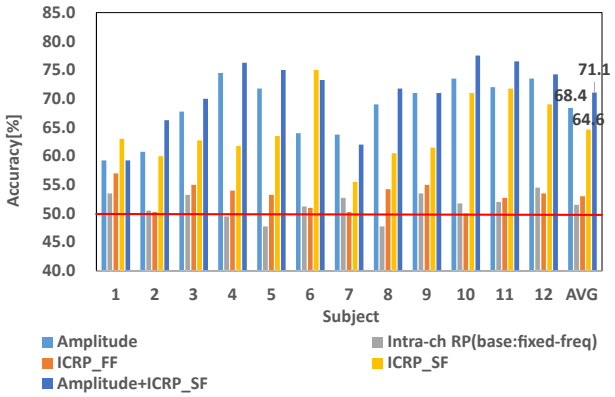


Fig. 9. Accuracy of classification by the NN model for each subject when using the amplitude spectrum, relative phase and their combination. In conventional inter, intra-channel relative phase, the fixed base frequency is the 6th bin of the base electrode. The base electrode for inter-channel relative is selected as the one that was best on average for the subjects.

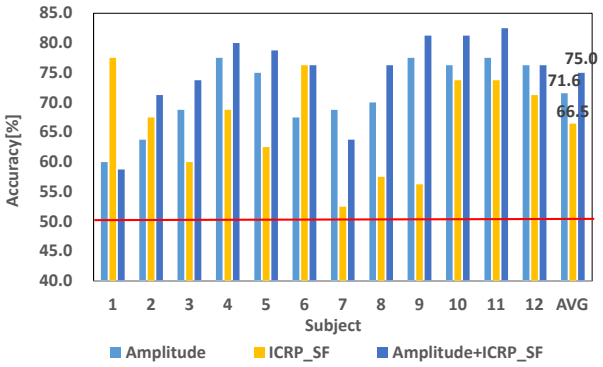


Fig. 10. Accuracy of **ensemble classification** by the NN model for each subject when using the amplitude spectrum, relative phase and their combination. The base electrode for inter-channel relative phase is selected as the one that was best on average for the subjects.

highlighted areas in the attention mask if the prediction is correct. On the other hand, the values are not repeatedly highlighted if the prediction is wrong. From this, it can be said that the attention mechanism is functioning to make correct predictions.

B. CNN model Pre-training + Subject Adaptation

In the case of conducting pre-training with data from multiple subjects, 11 out of the 12 subjects (excluding the target subject) are used for pre-training (880 samples in total, 11 subjects $\times$ 80 samples). After pre-training with the combined data from 11 subjects, the 5-fold cross validation procedure is applied for the binary classification task and the model with the lowest loss is used.

The results of the classification of the English phoneme dataset by the baseline CNN and the CNN with pre-training+subject adaptation (CNN+FT) are shown in Fig. 8. Assuming that the result is chance rate when the accuracy is in the 45-55% range, three subjects (2, 3, 7) have chance rate, but the CNN+FT model outperformed the baseline CNN in the other six of the remaining nine subjects (5, 6, 8, 10, 11, 12). It can be inferred that the increase in the number of training samples by 11 times has contributed to this improvement. In terms of average classification accuracy, applying FT leads to an improvement of approximately 6% (in other words, error reduction rate is 12%). Based on these results, it can be concluded that the proposed approach of pre-training and subject adaptation is effective, and it can mitigate the impact of insufficient training data.

C. Neural Network Using Spectral Features

In the first experiment for neural network classification using spectral features, each trial data which the subject imagined 5 times repeatedly in 5 seconds is divided into 5 parts for increasing the training samples. These divided parts are treated as separate samples for training and classification. Therefore, it is important to note that the sample sizes are different compared to the performance evaluation of the CNN-based models described earlier. The training and evaluation of the model employed 5-fold cross validation, using the data of each subject independently.

Fig. 9 shows the classification accuracy using the amplitude spectrum, the conventional relative phase features [11] ("Intra-channel (fixed base-freq)" and "ICRP_FF" in Fig. 4), the relative phase proposed in this study (Fig. 5), and the combined use of the relative phase and amplitude spectrum. Here, the base electrode is selected as the one that was best on average for the subjects. From the comparison of the relative phase between the conventional and proposed methods, it can be observed that the relative phase used in this study achieved 64.6% accuracy, significantly surpassing the conventional method. Furthermore, it is evident that the proposed method outperforms the amplitude spectrum in participants 1 and 6. On average across subjects, the relative phase exhibits a difference of approximately 4% in classification accuracy compared to the amplitude spectrum. This suggests that relative phase captures EEG features to a slightly lesser extent than amplitude features. In addition, the combined use of the amplitude spectrum and proposed relative phase shows improved accuracy compared to

TABLE III
COMPARISON OF THE EFFECTS WITH AND WITHOUT DATA DIVISION

	5-sec data	1-sec split data
CNN (raw)	51.6	50.3
CNN (raw)+Att	71.9	56.5
NN (Amp+ICRP_SF)	52.9	71.1

using the amplitude spectrum alone for all subjects except 1, 7, and 9. Notably, the subject 4 achieved a classification accuracy of approximately 76.5% using the combined features. Despite using only statistical simple features as inputs to the neural network, the model achieved performance comparable to the aforementioned attention-based CNN model which classified each sample by using 5 seconds.

Next, in the inference stage, the classification results were evaluated using a ensemble scheme (majority vote) based on dividing the 5-second trial samples that were imagined 5 times into 5 segments and using each segment classification. The results of this classification are presented in Fig. 10. It is the same as the CNN and CNN+Attention models in that the entire 5s is used for classification prediction. The experiment is unchanged in that 1s of data divided into 5s is given as input in the training stage, 5-fold cross validation is applied and the data is used for each subject independently. From the results, it can be observed that by utilizing both amplitude and phase information, classification accuracies of 80% or higher were achieved for subjects 4, 9, 10, and 11. Furthermore, a classification accuracy of 75% was achieved on average for all subjects and the accuracy is improved by 2-4% compared to the case where no ensemble process (in other words, the error reduction rate is 5-13%). However, it is important to note that these comparisons have different experimental designs (number of test samples) in the evaluation.

D. Comparison

1) Comparison of the effects with and without data division: In the CNN and CNN+Attention models, the 5-second raw signal, which was imagined five times, was directly used as input. On the other hand, in the experiments using spectral features, the trial data was divided into five equal parts and treated as separate samples. In this point of view, we compare the difference in classification accuracy with and without data division for each of the methods. The results are shown in Table III. In the case of CNN+Attention, it is observed that the classification accuracy is higher when data division is not performed. This suggests that the Attention mechanism is effective for the data that is repeated imagery. Furthermore, in the case of spectral features, it is observed that data division improves the classification accuracy. A possible reason for the poor classification accuracy when data division is not applied is that the speech imagery features that appear in a short period of time are buried because we extract feature as mean statistic for entire segment, and the feature might be influenced remarkably by non-imagined parts.

TABLE IV
AVERAGE CLASSIFICATION ACCURACY FOR EACH METHOD IN THE TWO-CLASS CLASSIFICATION OF THE ENGLISH PHONEME DATASET[%]. RP:RELATIVE PHASE, FT:FINE TUNING, ICRP_SF:PROPOSED RELATIVE PHASE, *:1-SEC SPLIT DATA WERE USED

Model (input)	Accuracy[%]
SVM (statistical feature)[8]	69.0
CNN (raw)[8]	49.0
CNN (raw)	51.6
CNN (raw)+FT	57.5
CNN (raw)+Att	71.9
CNN(raw)+Att+FT	**75.3**
NN (Amp)*	68.4
NN (Intra-ch RP base 6th dim)[11]*	51.5
NN (Amp+Intra-ch RP base 6th dim)[11]*	58.5
NN (ICRP_SF)*	64.6
NN (Amp+ICRP_SF)*	71.1
NN (Amp)+test data ensemble	71.6
NN (ICRP_SF)+test data ensemble	66.5
NN (Amp+ICRP_SF)+test data ensemble	**75.0**

2) Comparison with previous research: The average classification accuracy of the method proposed in this study and the previous studies [8] and [11] for the same database are summarized in Table IV. Although the experimental conditions were different, the classification accuracy of CNN models with attention (CNN+Att, CNN+Att+FT) exceeded 70% in both cases, surpassing the classification accuracy of 69.0% achieved by SVM in [8]. The classification with spectral features in this study achieved a classification accuracy of 75.0% by the test data ensemble process, which is the highest accuracy for classification with handcrafted features using this data set. These show the effectiveness of the two proposed classification methods.

V. CONCLUSIONS

In this study, we investigated methods to solve two problems related to deep learning in EEG-based imagined speech recognition. To address the lack of feature extraction capability of CNNs for imagined speech, we showed that the attention mechanism is capable of capturing repeated imagined speech features. For the lack of training data, pre-training with multiple subject data (excluding the target subject's) and subject adaptation with specific subject data were shown to be effective for improvement. In addition, we proposed the classification model using both the inter-channel relative phase and the amplitude spectrum features. The classification accuracy in this case reached 75%, which was the highest for classification using handcrafted features in the dataset used. As a future work, it is possible to try to further improve classification accuracy by using proposed spectral features as input for the attention-based CNN model.

REFERENCES

[1] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16,

no. 3, p. 031 001, Apr. 2019. DOI: 10.1088/1741-2552/ab0ab5. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/ab0ab5.

[2] X. Chen, C. Li, A. Liu, M. J. McKeown, R. Qian, and Z. J. Wang, "Toward Open-World Electroencephalogram Decoding Via Deep Learning: A comprehensive survey," *IEEE Signal Processing Magazine*, vol. 39, no. 2, pp. 117–134, 2022. DOI: 10.1109/MSP.2021.3134629.

[3] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[4] K. Brigham and B. V. K. V. Kumar, "Imagined Speech Classification with EEG Signals for Silent Communication: A Preliminary Investigation into Synthetic Telepathy," in *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, 2010, pp. 1–4. DOI: 10.1109/ICBBE.2010.5515807.

[5] D. Rojas, O. Ramos, and J. Saby, "Recognition of Spanish vowels through imagined speech by using spectral analysis and SVM," vol. 7, pp. 889–897, Jan. 2016.

[6] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056 013, Jul. 2018. DOI: 10.1088/1741-2552/aace8c. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/aace8c.

[7] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017. DOI: https://doi.org/10.1002/hbm.23730. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23730. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730.

[8] J. Clayton, S. Wellington, C. Valentini-Botinhao, and O. Watts, "Decoding Imagined, Heard, and Spoken Speech: Classification and Regression of EEG Using a 14-Channel Dry-Contact Mobile Headset," in *Proc. Interspeech 2020*, 2020, pp. 4886–4890. DOI: 10.21437/Interspeech.2020-2745. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2745.

[9] A. Kamble, P. Ghare, and V. Kumar, "Machine-learning-enabled adaptive signal decomposition for a brain-computer interface using EEG," *Biomedical Signal Processing and Control*, vol. 74, p. 103 526, 2022, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2022.103526. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422000489.

[10] C. Cooney, R. Folli, and D. Coyle, "Optimizing Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 1311–1316. DOI: 10.1109/SMC.2019.8914246.

[11] R. Sakai, A. Kai, and S. Nakagawa, "Classification of Imagined and Heard Speech Using Amplitude Spectrum and Relative Phase of EEG," in *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, 2021, pp. 373–375. DOI: 10.1109/LifeTech52111.2021.9391883.

[12] D. Bahdanau, K. Cho, and Y. Bengio, *"Neural Machine Translation by Jointly Learning to Align and Translate"*, 2016. arXiv: 1409.0473 [cs.CL].

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *the European Conf. Computer Vision*, 2018, pp. 3–19.

[14] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *"Attention Is All You Need"*, 2017. arXiv: 1706.03762 [cs.CL].

[15] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing Speech Detection Using Modified Relative Phase Information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 660–670, 2017. DOI: 10.1109/JSTSP.2017.2694139.

[16] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2012. DOI: 10.1109/TASL.2011.2172422.

Manipulation of Neuronal Network Firing Patterns using Temporal Deep Unfolding-based MPC

Jumpei Aizawa*, Masaki Ogura*, Masanori Shimono† and Naoki Wakamiya*
* Osaka University, Japan
E-mail: {j-aizawa, m-ogura, wakamiya}@ist.osaka-u.ac.jp Tel: +81-06-6879-4358
† Kyoto University, Japan
E-mail: shimono.masanori.7w@kyoto-u.ac.jp Tel: +81-075-751-4173

***Abstract*—Because neuronal networks are intricate systems composed of interconnected neurons, their control poses challenges owing to their nonlinearity and complexity. In this paper, we propose a method to design control input to a neuronal network to manipulate the firing patterns of modules within the network. We propose a methodology for designing a control input based on temporal deep unfolding-based model predictive control (TDU-MPC), a control methodology based on the deep unfolding technique actively investigated in the context of wireless signal processing. During the method development, we address the unique characteristics of neuron dynamics, such as zero gradients in firing times, by approximating input currents using a sigmoid function. The effectiveness of the proposed method is confirmed via numerical simulations. In networks with 15 and 30 neurons, the control was achieved to switch the firing frequencies of two modules without directly applying control inputs. This study includes a tailored methodology for networked neurons, the extension of TDU-MPC to nonlinear systems with reset dynamics, and the achievement of desired firing patterns in neuronal networks.**

I. Introduction

The brain, a highly complex organ, consists of a vast network of interconnected neurons. When an individual neuron within the network generates an electrical impulse called firing, neighboring neurons receive an electrical influx that serves as a stimulus to trigger their firing. This phenomenon propagates throughout the entire network and contributes to the complex pattern of the firing dynamics in the brain. Hence, the brain exhibits a wide range of electrically active patterns and dynamic behaviors that account for its remarkable capabilities and functions [1].

Studies focused on brain dynamics control [2] have the potential to yield significant advancements in the enhancement of human cognitive function and the performance of brain–machine interfaces and artificial nerves. For example, advancements in brain dynamics control are expected to pave the way for the development of novel therapies for psychiatric disorders [3]. In addition, a greater degree of seamless communication between the human brain and external devices can be achieved by precisely controlling the firing patterns of neurons to optimize the brain–machine interfaces [4]. The prospect of using brain dynamics control to improve human welfare, both in terms of cognitive enhancements and psychiatric disorder treatments, demonstrates the potential of such research [2].

Owing to the complex nature of a neuronal network, controlling its firing pattern is challenging [5]. A neuron is often modeled as a nonlinear dynamical system characterized by abrupt jumps in membrane potential as neurons are fired [6]. Such nonlinearity adds a layer of complexity to the control problem [7]. In addition, a neuronal network as a whole represents a large system consisting of several interconnected neurons with complex interactions. Network size further complicates control efforts by introducing a large number of variables and inter-dependencies, which are to be considered. Addressing these challenges requires developing sophisticated control strategies capable of accommodating the unique characteristics and complexities inherent to neuronal network dynamics.

In this study, we aim to develop a methodology for designing control input currents to effectively manipulate the firing frequency of multiple modules within a network, even when neurons not belonging to the modules are under control. Therefore, we employed temporal deep unfolding-based model predictive control (TDU-MPC) introduced in [8] as a control methodology. TDU-MPC offers solutions for tackling optimization problems that arise during each step of MPC by employing emerging technologies, such as deep unfolding [9]–[13]. Deep unfolding involves the temporal unfolding of the mathematical model of a control object by treating it as a feed-forward network. Using backpropagation, the control inputs at each instant of time step are acquired via learning, and the iterative process is repeated for each step.

Although TDU-MPC has been recognized for its effectiveness in the control of nonlinear dynamical systems [8], the dynamic characteristics of neurons pose challenges. The main challenge lies in the fact that once a neuron model is discretized, the firing times of the neurons can have a zero gradient with respect to the control, as has been observed by our preliminary experiments. This property of the gradient complicates the process of learning control inputs because TDU-MPC relies on deep learning techniques. To overcome such hurdles, we developed an approach to approximate the input current of neighboring neurons by means of a sigmoid function, which enabled us to avoid the aforementioned problem of zero gradient.

This work was supported by JSPS KAKENHI Grant Numbers JP22H00514 and JP21H01352.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

The contributions of this study are summarized as follows. First, the paper proposes the methodology for designing control input to neuronal networks to achieve a desired firing pattern of modules within the networks. Second, unlike other methodologies primarily focused on single neurons [5], our methodology is tailored to *networks* of neurons. Third, the applicability of TDU-MPC can be extended to include nonlinear systems exhibiting reset dynamics.

This paper is organized as follows: Section II states the problem studied in this paper, while Section III proposes a method for solving the control problem. Section IV presents the results of numerical simulations. Finally, the paper is concluded in Section V.

II. Problem Statement

In this section, we formally state our problem of controlling the neuronal dynamics. In Subsection II-A, we state our model of neuronal networks. Subsequently, in Subsection II-B, we state the control problem addressed in this paper.

A. Neuronal Network

In this subsection, we state the model of a neuronal network. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the directed graph representing the connectivity of the neuronal network. The node set $\mathcal{V} = \{1, \dots, N\}$ represents the set of neurons, while the edge set $\mathcal{E}$ represents the connectivity among the neurons. For each $i \in V$, $\mathcal{N}_i$ denotes the set of in-neighbors of node i. Under this notation, we assume that when a neuron in $\mathcal{V}_i$ fires, the neuron i receives a current.

The type of firing neuron determines the effect on its neighboring neurons. For this study, we assume, as is common in the literature, that neurons are divided into two types of neurons: excitatory neurons and inhibitory neurons. The set of the former neurons is denoted by $\mathcal{V}_{\text{ex}}$, while the latter is denoted by $\mathcal{V}_{\text{in}}$. An excitatory neuron induces a positive current I_{ex} to its neighbors, while an inhibitory neuron induces a negative current I_{in}.

We assume that each neuron $i \in \mathcal{V}$ follows the Izhikevich model [6] described by the differential equation

$$\dot{v}_i(t) = 0.04v_i^2(t) + 5v_i(t) + 140 - u_i(t) + I_i(t), \quad (1)$$

$$\dot{u}_i(t) = a(bv_i(t) - u_i(t)) \quad (2)$$

with the reset dynamics described by

$$\begin{bmatrix} v_i(t^+) \\ u_i(t^+) \end{bmatrix} = \begin{bmatrix} c \\ u_i(t) + d \end{bmatrix}, \quad \text{if } v_i(t) \geq 30. \quad (3)$$

For this study, we assume that a neuron i fires at time t if $v_i(t) \geq 30$. The real numbers a, b, c, and d in the dynamical model are considered dimensionless parameters. By varying the values of these dimensionless parameters, a rich variety of neuronal behaviors, including regular spiking, bursting, and fast spiking, can be achieved. Therefore, the Izhikevich model is extensively used to study and replicate various types of neuronal activity seen in different regions of the brain.

In the differential equations (1) and (2), the scalar variable $v_i(t)$ represents the membrane potential of neuron i at time t, while $u_i(t)$ represents the value of the membrane recovery variable of neuron i at time t. Further, $I_i(t)$ represents the current input to neuron i and is assumed as

$$I_i(t) = I_{i,\text{control}}(t) + I_{i,\text{internal}}(t),$$

where $I_{i,\text{control}}(t)$ is the control input signal that we need to design, and $I_{i,\text{internal}}(t)$ represents the inter-network current induced by firing events of neurons. In this paper, for simplicity, the inter-network current is supposed to be of the additive form defined as

$$I_{i,\text{internal}}(t) = \sum_{j \in \mathcal{N}_i} I_{ij}(t), \quad (4)$$

where $I_{ij}(t)$ represents the current input from neuron j to neuron i and is assumed to be of the form

$$I_{ij}(t) = \begin{cases} I_{\text{ex}}, & \text{if neuron } j \text{ fired in } [t-\tau, t] \text{ and } j \in \mathcal{V}_{\text{ex}}, \\ I_{\text{in}}, & \text{if neuron } j \text{ fired in } [t-\tau, t] \text{ and } j \in \mathcal{V}_{\text{in}}, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

where $\tau > 0$ is a parameter.

B. Control Objective

For this study, we consider the selective and dynamic enhancement problems caused by the firing activity of a specified module of neurons. Specifically, the network is assumed to consist of three modules, with stronger within-module connectivities than inter-module connectivities. Further, we assume that the control current is applied only to the neurons within a specific module. We consider a closed-loop control of the neurons; therefore, we assume that we can use the membrane potentials $v_i(t)$ for all $i \in \mathcal{V}$ to determine the input current $I_{i,\text{control}}(t)$ for $i \in \mathcal{V}_{\text{control}}$. Under this supposition, our objective is to control the firing frequency of other modules.Specifically, we aim to design the following firing pattern: From the initial to a pre-specified time, a module $\mathcal{V}_1 \subset \mathcal{V}$ fires frequently while another module $\mathcal{V}_2 \subset \mathcal{V}$ exhibits low firing activities. Thereafter, the module $\mathcal{V}_2$ should fire frequently, while module $\mathcal{V}_1$ fires less frequently.

Thus, the aforementioned control problem can be formally stated as follows:

Problem 1. *Assume that the set $\mathcal{V}$ of neurons is divided into disjoint subsets $\mathcal{V}_{control}$, $\mathcal{V}_1$, and $\mathcal{V}_2$. Assume*

$$I_{i,control}(t) = 0 \text{ for all } i \in \mathcal{V}_1 \cup \mathcal{V}_2.$$

Let T_s and T_e be real numbers satisfying $0 < T_s < T_e$ and define intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ by

$$\mathcal{I}_1 = [0, T_s], \ \mathcal{I}_2 = [T_s, T_e].$$

For a subset $\mathcal{W}$ of $\mathcal{V}$ and a time interval $\mathcal{I}$, define

$$f_{\mathcal{W},\mathcal{I}} = \sum_{i \in W} \text{(number of fires of neuron } i \text{ on } \mathcal{I}).$$

Construct a feedback controller of the form

$$I_{i,control}(t) = K_i(v_1(t), \dots, v_N(t)) \quad (6)$$

for all $i \in \mathcal{V}_{control}$ such that the objective function

$$(f_{\mathcal{V}_1,\mathcal{I}_1} - f_{\mathcal{V}_2,\mathcal{I}_1}) + (f_{\mathcal{V}_2,\mathcal{I}_2} - f_{\mathcal{V}_1,\mathcal{I}_2}) \tag{7}$$

is maximized.

Remark 1. *In the practical implementation of the control rules* (6)*, real-time calculation of the control input $I_{i,control}$ is preferred. However, in this research, real-time capability is not a requirement, and its realization is left for future investigation. Similarly, the assumption of membrane potential observability, though challenging in practice, aligns with existing practice in the literature (see, e.g., [5]).*

Because Problem 1 requires the control of a nonlinear reset system [14], a standard and effective solution involves the use of model predictive control (MPC) [15]. MPC is a type of feedback control method that aims to achieve optimal control by predicting the behavior of a controlled object at a fixed interval in the future from the current time. In MPC, the controlling input at each instance is determined by solving an optimization problem. The optimization problem consists of a mathematical model of the control target, a cost function obtained from the outputs in the prediction horizon, and constraints. Owing to its effectiveness in optimizing control inputs, handling constraints, and dealing with nonlinear systems, MPC has various applications, including process control, energy management, and automotive systems. Among various approaches available for performing MPC, we chose to use TDU-MPC [8] for this study. TDU-MPC offers several advantages such as easy implementation and the ability to find high-quality solutions.

III. Proposed Method

In this section, we propose a method to solve Problem 1 leveraging TDU-MPC. In Subsection III-A, we present an overview of TDU-MPC, while Subsection III-B presents the proposed methodology for solving Problem 1.

A. TDU-MPC

To understand TDU-MPC, we begin by reviewing deep unfolding [9]–[13] that TDU-MPC is built upon. Deep unfolding is a method for tuning the parameters of iterative algorithms using techniques from deep learning. In deep unfolding, a given iterative algorithm is first unfolded in the time direction and considered a feed-forward network with the number of layers equal to the number of iterations. Next, the parameters of each step of the iterative algorithm are embedded as learnable parameters in the network and tuned via backpropagation and stochastic gradient descent methods. Incremental learning, a widely used technique, is adopted to avoid gradient loss that can occur during parameter learning.

TDU-MPC has been proposed as a method to overcome the disadvantages of existing methodologies for performing MPC. In this subsection, we provide a brief overview of TDU-MPC, considering a discrete-time dynamical system such as

$$\Sigma : x[k+1] = f(x[k], u[k], w[k]),\ k = 0, 1, 2, \dots$$

where x, u, and w represent the state variable, control input, and disturbance signal, respectively. In TDU-MPC, the system is considered as an iterative algorithm, and the control input is regarded as a learnable parameter.

Specifically, we first make the feed-forward network with T layers. Each layer is the state x at each instance of time of unfolded T-step state transitions in the temporal direction. Next, the control inputs at each instance can be learned using deep learning techniques as in deep unfolding. The disturbances at each instance of time are used as data, and learning is performed using the objective function of the optimization problem in MPC as the cost function. Thereafter, the control input minimizing the following objective function

$$\sum_{\ell=0}^{T-1} p_\ell(x[k+\ell], u[k+\ell]) + p_T(x[k+T])$$

is designed using the aforementioned scheme. In Fig. 1, we illustrate a schematic picture illustrating TDU-MPC. For the details of TDU-MPC, readers are referred to [8].

B. Discrete-time Model

As described in the previous subsection, TDU-MPC is a methodology for controlling discrete-time dynamical systems, while Problem 1 involves a continuous-time model. Therefore, for applying TDU-MPC to Problem 1, the dynamics presented as (1)–(3) and the objective function (7) must be appropriately discretized.

Let $\Delta t > 0$ denote the time step for discretization. For this study, we discretized the differential equations (1) and (2) as

$$v_i[k+1] = \begin{cases} v_i[k] + (0.04v_i^2[k] + 5v_i[k] \\ \qquad +140 - u_i[k] + I_i[k])\Delta t, & \text{if } v_i[k] < 30, \\ v_i[k] - (30 - c), & \text{otherwise} \end{cases} \tag{8}$$

and

$$u_i[k+1] = \begin{cases} u_i[k] + (a(bv_i[k] - u_i[k]))\Delta t, & \text{if } v_i[k] < 30, \\ u_i[k] + d, & \text{otherwise,} \end{cases} \tag{9}$$

where the input current

$$I_i[k] = I_{i,\text{control}}[k] + I_{i,\text{internal}}[k]$$

consists of the control signal $I_{i,\text{control}}[k]$ to be designed and the input current $I_{i,\text{internal}}[k]$ resulting from firing events at neighbor neurons. As in the continuous-time case, we say that a neuron i fires at time k if $v_i[k] \geq 30$.

Similar to (4), we assume the input current $I_{i,\text{internal}}[k]$ of the form

$$I_{i,\text{internal}}[k] = \sum_{j \in \mathcal{N}_i} I_{ij}[k],$$

where $I_{ij}[k]$ represents the current input from neuron j to neuron i and is given by

$$I_{ij}[k] = \begin{cases} S(v_j[k])I_{\text{ex}}, & \text{if } j \in \mathcal{V}_{\text{ex}}, \\ S(v_j[k])I_{\text{in}}, & \text{otherwise} \end{cases} \tag{10}$$

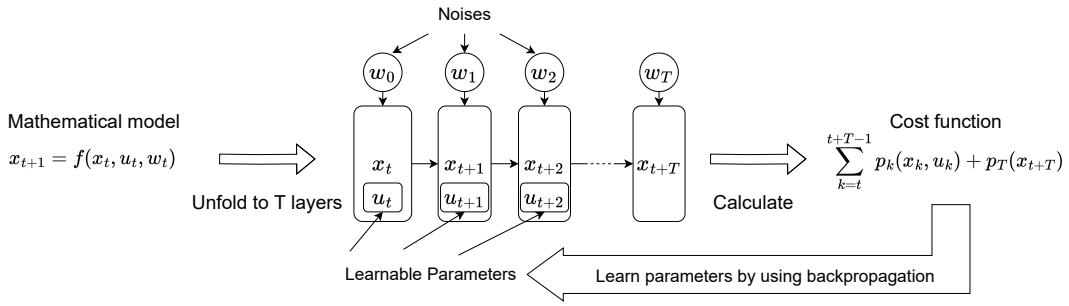


Fig. 1: Schematic picture of temporal deep unfolding-based model predictive control (TDU-MPC)

in which the introduced coefficient $S(v_j[t])$ serves the purpose of introducing a soft-threshold and is defined as

$$S(v_j[k]) = \frac{1}{1 + \exp(-\sigma(v_j[k] - 20))}. \tag{11}$$

One of the main reasons for introducing the soft-threshold mechanism is to ensure the differentiability of the objective function. To effectively perform deep learning and deep unfolding, the differentiability of the objective function is an important requirement. Conversely, the differentiability of the cost function with respect to the input current $I_{i,\text{control}}$ to be designed cannot be guaranteed without soft-thresholding. Therefore, introducing the soft-thresholding mechanism is necessary. In fact, we confirmed through our preliminary experiments that for a system model with an inter-node input current of the form (5), TDU-MPC often fails to yield an effective control input signal.

C. Cost Function

For reasons similar to soft-thresholding in (10) and (11), the cost function has to be differentiable with the control input to be designed. Thus, to design a control input within the first interval $\mathcal{I}_1$ in which firing must be promoted within the first module, we adopt an cost function such that

$$\sum_{k=l}^{l+T}\Bigg(\sum_{i\in\mathcal{V}_1} (1-\delta_i[k])(30 - c - (v_i[k+1] - v_i[k])) + \sum_{j\in\mathcal{V}_2} |v_j[k+1] - v_j[k]|\Bigg), \tag{12}$$

where

$$\delta_i[k] = \begin{cases} 1, & \text{if neuron } i \text{ fires at time } k, \\ 0, & \text{otherwise} \end{cases}$$

is a binary variable encoding the firing event of neurons. Minimization of the aforementioned cost function leads to a smaller first term, which then makes the amount of change in the membrane potential of the neurons in module $\mathcal{V}_1$ bigger, thereby stimulating neurons in the module $\mathcal{V}_1$ to fire. Conversely, we introduce the second term in (12) to suppress firing within the other module $\mathcal{V}_2$. For the aforementioned reasons, the cost function after the switching time T_s is set as

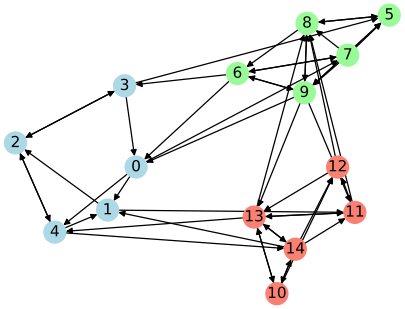
$$\sum_{k=l}^{l+T}\Bigg(\sum_{i\in\mathcal{V}_1} |v_i[k+1] - v_i[k]| + \sum_{j\in\mathcal{V}_2} (1-\delta_j[k])(30 - c - (v_j[k+1] - v_j[k]))\Bigg).$$

IV. Results and Discussions

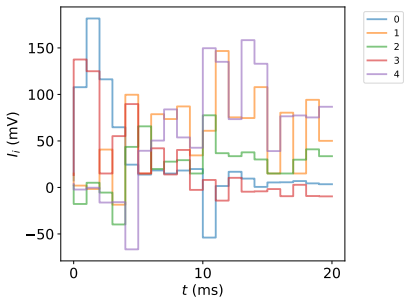
In this section, we discuss the effectiveness of the proposed approach via numerical simulations. We considered neuronal networks consisting of "fast spiking" neurons described in [6]. Therefore, the model parameters in (8) and (9) were set to $a = 0.1$, $b = 0.2$, $c = -65$, and $d = 2$. Additionally, we set the amount of current flux incoming from neighboring neurons as $I_{\text{ex}} = 15\,\text{mV}$ and $I_{\text{in}} = -3\,\text{mV}$. For the neuronal network, we assumed that 80% of the neurons were of the excitatory type, while 20% were inhibitory neurons. Further, we assumed that the network consisted of three different modules (or clusters) such as $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_{\text{control}}$. As stated in Problem 1, we aimed to design the control input into the neurons in $\mathcal{V}_{\text{control}}$ to control the firing patterns in modules $\mathcal{V}_1$ and $\mathcal{V}_2$.

The (hyper-)parameters of the proposed method were chosen as follows. First, Adam was chosen as the optimizer, an algorithm used for updating the control input, based on existing literature [16]. Second, similar to [8], incremental learning was adopted during the process of control input learning. Third, for the Python implementation of TDU-MPC, we used libraries such as PyTorch [17] and NumPy [18]. Last, the time step for discretization was set to $\Delta t = 1\,\text{ms}$, while the prediction horizon within TDU-MPC was set as $T = 10$ step. Furthermore, for the soft-thresholding function (11), we chose parameter $\sigma = 0.38$.

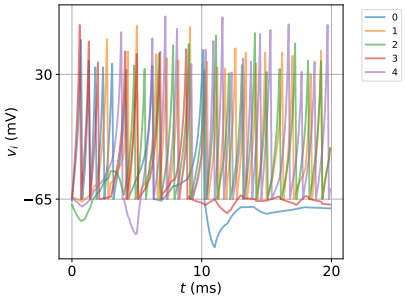
As aforementioned, we first generated a network with $N = 15$ neurons and performed a control experiment via numerical simulations. To construct a network with modules, we used the stochastic block model [19] with Modules $\mathcal{V}_{\text{control}} = \{0, 1, 2, 3, 4\}$, $\mathcal{V}_1 = \{5, 6, 7, 8, 9\}$, and $\mathcal{V}_2 = \{10, 11, 12, 13, 14\}$. Out of the 15 neurons, neurons 6, 9, and 13 were inhibitory, while the others were excitatory. The edge probability within the module was set to $1/2$, while the one between modules was set to $1/8$. Utilizing these parameters,



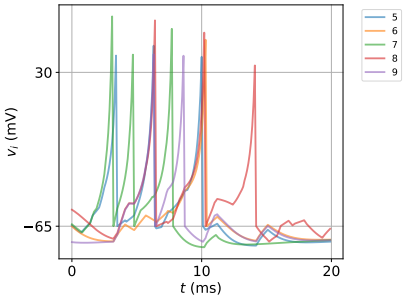
(a) Neuronal network $\mathcal{V}$. Blue nodes: $\mathcal{V}_{\text{control}}$, green nodes: $\mathcal{V}_1$, red nodes: $\mathcal{V}_2$.



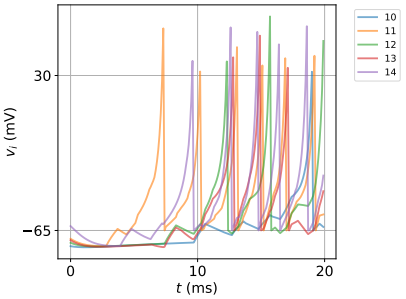
(b) Control inputs to neurons in $\mathcal{V}_{\text{control}}$



(c) Membrane potentials of neurons in $\mathcal{V}_{\text{control}}$.

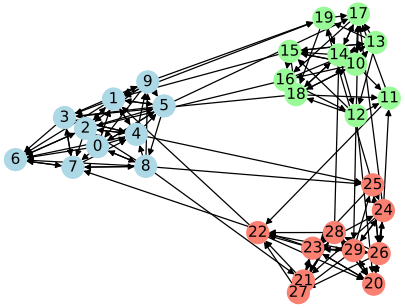


(d) Membrane potentials of neurons in $\mathcal{V}_1$

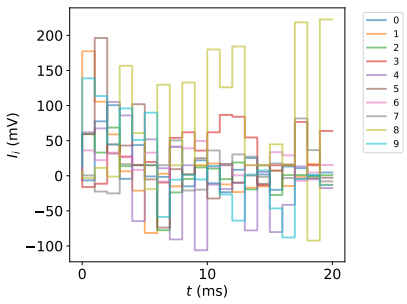


(e) Membrane potentials of neurons in $\mathcal{V}_2$

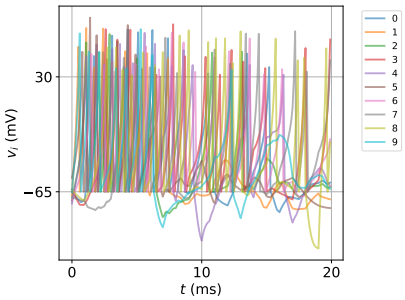
Fig. 2: Control of firing pattern in a network with $N = 15$ neurons



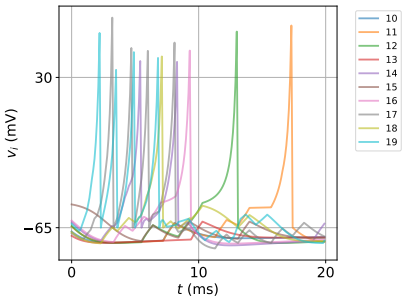
(a) Neuronal network $\mathcal{V}$. Blue nodes: $\mathcal{V}_{\text{control}}$, green nodes: $\mathcal{V}_1$, red nodes: $\mathcal{V}_2$.



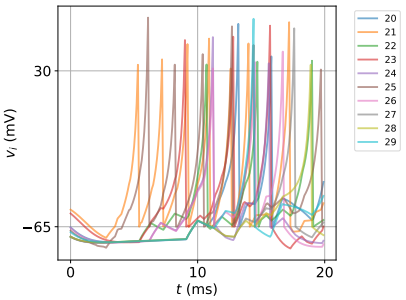
(b) Control inputs to neurons in $\mathcal{V}_{\text{control}}$



(c) Membrane potentials of neurons in $\mathcal{V}_{\text{control}}$.



(d) Membrane potentials of neurons in $\mathcal{V}_1$.



(e) Membrane potentials of neurons in $\mathcal{V}_2$.

Fig. 3: Control of firing pattern in a network with $N = 30$ neurons

TABLE I: Numbers of fires in each of the modules

(a) $N = 15$

Neuron No. \ time (ms)	$0 \leq$ time < 10	$10 \leq$ time ≤ 20
5-9	**8**	4
10-14	2	**16**

(b) $N = 30$

Neuron No. \ time (ms)	$0 \leq$ time < 10	$10 \leq$ time ≤ 20
10-19	**12**	2
20-29	6	**23**

we obtained the network shown in Fig. 2a. Within the resulting network, although the number of connections between the modules are fewer than the ones within the modules, the inter-module connections are still bidirectional, which makes the control problem nontrivial.

For initializing the membrane potentials v_i and membrane recovery variables u_i, we first set the potentials to c and the recovery variables to 0. Thereafter, a random current was applied to all neurons for $10\,\mathrm{ms}$. After initialization, we performed neuronal network control using the method described in Section III. The switching and simulation times were set to $T_s = 10\,\mathrm{ms}$ and $T_e = 20\,\mathrm{ms}$, respectively. Our objective was to design control inputs $I_{0,\mathrm{control}}, \ldots, I_{4,\mathrm{control}}$ such that the firing pattern of $\mathcal{V}_1$ was more active than that of $\mathcal{V}_2$ in the time interval $\mathcal{I}_1 = [0\,\mathrm{ms}, 10\,\mathrm{ms}]$, and that of $\mathcal{V}_2$ was more active than $\mathcal{V}_1$ in the time interval $\mathcal{I}_2 = [10\,\mathrm{ms}, 20\,\mathrm{ms}]$.

The results of the control experiment are presented in Fig. 2. In Fig. 2b, the control input signals to neurons in $\mathcal{V}_{\mathrm{control}}$ obtained from our method have been illustrated. From the membrane potentials of neurons in $\mathcal{V}_1$ and $\mathcal{V}_2$ shown in Figs. 2d and 2e, respectively, we confirmed that module $\mathcal{V}_1$ fired frequently from the initial time to $t = 10\,\mathrm{ms}$ as compared with $\mathcal{V}_2$. Additionally, module $\mathcal{V}_2$ fired frequently from $t = 10\,\mathrm{ms}$ to $t = 20\,\mathrm{ms}$ when compared with $\mathcal{V}_1$, as desired. We observed that the average computation time for a control input for the duration of $1\,\mathrm{ms}$ was $106\,\mathrm{s}$. Thus, improvements to the current implementation for a shorter computational time are of practical importance.

Further, the simulation results with $N = 30$ neurons were observed. The modules were set as $\mathcal{V}_{\mathrm{control}} = \{0, 1, 2, \ldots, 9\}$, $\mathcal{V}_1 = \{10, 11, 12, \ldots, 19\}$, and $\mathcal{V}_2 = \{20, 21, 22, \ldots, 29\}$. Out of the 30 neurons, neurons 3, 4, 6, 20, 26, and 27 were inhibitory, while the others were excitatory. The edge probability within the network was set to $1/2$, while the one between modules was set to $1/25$. With these parameters, we constructed a neuronal network with 30 neurons, as shown in Fig. 3a. Similar to our numerical experiment for the network with $N = 15$ nodes, we designed the control input $I_{i,\mathrm{control}}$ ($i \in \mathcal{V}_{\mathrm{control}}$) for the new network. The obtained time series of the control input has been shown in Fig. 3b. Further, the membrane potentials of $\mathcal{V}_1$ and $\mathcal{V}_2$ under the control input are shown in Figs. 3e and 3e, respectively. Modules $\mathcal{V}_1$ and $\mathcal{V}_2$ were confirmed to fire in the desired manner, and the average computational time for one step of the control input was $353\,\mathrm{s}$.

The results presented in Table I provide a summary of our numerical experiment for the two networks. Within the table, the number of fires in each of the modules for time intervals $\mathcal{I}_1$ and $\mathcal{I}_2$. The number of fires within the targeted module (the one required to be more active than the other) was confirmed to be no less than double the number of the one within the other module, thereby illustrating the effectiveness of the proposed method.

In both simulation results, initial time interval $\mathcal{I}_1$ corresponds to the period where module $\mathcal{V}_1$ should display higher firing activity. During this interval, the applied control input primarily activated the excitatory neurons in V_{control} having edges outgoing to $\mathcal{V}_1$. For instance, when considering $N = 15$ as an example, we can observe from Fig. 2b that neurons 0 and 3 were predominantly stimulated at the beginning of the time interval. As the switching time T_s approaches, the control input gradually diminished, effectively preventing firing within $\mathcal{V}_1$ in the subsequent time interval $\mathcal{I}_2$ when $\mathcal{V}_2$ was expected to exhibit high firing activity. Similarly, in the second time interval $\mathcal{I}_2$, we observed a comparable pattern to the aforementioned case. The excitatory neurons in V_{control} having edges outgoing to $\mathcal{V}_2$ were primarily stimulated. For instance, considering the case of $N = 15$ again, neurons 1 and 4 demonstrated increased activation during this time interval. These observed stimulation patterns align well with our intuitive understanding and expectations.

V. Conclusions

This study addressed the challenges associated with neuronal network control. We proposed a novel method for designing control inputs to manipulate the firing patterns of modules within a network. Based on TDU-MPC, our approach exploits deep unfolding techniques commonly used in wireless signal processing. The effectiveness of our proposed method was demonstrated via extensive numerical simulations.

Future research directions include evaluating the effectiveness of the proposed method via numerical experiments on other networks consisting of neurons other than the fast-spiking type. Additionally, robustness evaluation of the proposed method with respect to the modeling errors of the neurons can be of practical relevance, as the current research is conducted on the assumption that the dynamics of neurons are available and, therefore, errors in the model are assumed to be not significant. Another important research direction is the investigation of a method to design control inputs without the need to observe all membrane potentials.

References

[1] R.M. Hutchison, T. Womelsdorf, E.A. Allen, P.A. Bandettini, V.D. Calhoun, M. Corbetta, S.D. Penna, et.al, "Dynamic functional connectivity: promise, issues, and interpretations", *Neuroimage*, vol. 80, pp. 360–378, 2013.

[2] E. Tang and D.S. Bassett, "Colloquium: Control of dynamics in brain networks", *Reviews of Modern Physics*, vol. 90, 2018.
[3] D.J. Lee, C.S. Lozano, R.F. Dallapiazza and A.M. Lozano, "Current and future directions of deep brain stimulation for neurological and psychiatric disorders: JNSPG 75th Anniversary Invited Review Article", *Journal of Neurosurgery*, vol. 131, pp. 333–342, 2019.
[4] J.D. Wander and R.P.N. Rao, "Brain–computer interfaces: a powerful tool for scientific inquiry", *Current Opinion in Neurobiology*, vol. 25, pp. 70–75, 2014.
[5] A. Iolov, S. Ditlevsen and A. Longtin, "Stochastic optimal control of single neuron spike trains," *Journal of Neuronal Engineering*, vol. 11, 2014.
[6] E.M. Izhikevich, "Simple model of spiking neurons" *IEEE Transactions on Neuronal Networks*, vol. 14, pp. 1569–1572, 2003.
[7] A.E. Motter, "Networkcontrology", *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, 2015.
[8] M. Kishida and M. Ogura, "Temporal deep unfolding for constrained nonlinear stochastic optimal controls," *IET Control Theory & Applications*, vol. 16, pp. 139–150, 2022.
[9] J.R. Hershey, J.L. Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," arXiv:1409.2574, 2014.
[10] D. Ito, S. Takabe, and T. Wadayama, "Trainable ISTA for sparse signal recovery," *IEEE Transactions on Signal Processing*, vol. 67, pp. 3113–3125, 2019.
[11] V. Monga, Y. Li and Y.C. Eldar, "Algorithm unrolling: Interpretable efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, pp. 18–44, 2021.
[12] A. Jagannath, J. Jagannath and T. Melodia, "Redefining wireless communication for 6G: Signal processing meets deep learning with deep unfolding," *IEEE Transactions on Artificial Intelligence*, vol. 2, pp. 528–536, 2021.
[13] M. Ogura, K. Kobayashi, and K. Sugimoto, "Static output feedback synthesis of time-delay linear systems via deep unfolding," In *17th IFAC Workshop on Time Delay Systems*, pp. 214–215, 2022.
[14] J. Carrasco, A. Baños and A van der Schaft, "A passivity-based approach to reset control systems stability", *Systems & Control Letters*, vol. 59, pp. 18–24, 2010.
[15] C.E. Garcia, D.M. Prett and M. Morari, "Model predictive control: Theory and practice—A survey", *Automatica*, vol. 25, pp. 335–348, 1989.
[16] K. Bae, H. Ryu, and H. Shin, "Does adam optimizer keep close to the optimal point?" *arXiv preprint arXiv:1911.00289*, 2019.
[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury et al., "Pytorch: An imperative style, high-performance deep learning library" *Advances in Neuronal Information Processing Systems*, vol. 32, 2019.
[18] S. Van Der Walt, S.C. Colbert and G. Varoquaux, "The NumPy array: a structure for efficient numerical computation" *Computing in Science & Engineering*, vol. 13, pp. 22–30, 2011.
[19] C. Lee and D.J. Wilkinson, "A review of stochastic block models and extensions for graph clustering", *Applied Network Science*, vol. 4, pp. 1–50, 2019.

Goodness of Fit to the Convolution Model of fMRI Data and Determination of the Regularization Parameter

Wakako Nakamura*
* Shimane University, Japan
E-mail: wakako@riko.shimane-u.ac.jp Tel/Fax: +81-852-32-6065

***Abstract*—In the analysis of fMRI data, the time when the event of interest occurred in the experiment is usually used for analysis. However, in recent years, measurement of fMRI data which is not necessarily related to a certain simple event, such as measurement in resting state, is often performed. In this paper, we propose a method to analyze fMRI data without using time when events of interest occurred. In the proposed method, we fit a generation model of convolution form to fMRI data and obtain values of goodness of fit to the generation model. If the goodness of fit is sufficiently large for data from a voxel, the data can be considered to include a time course from neural activities. We suggest that this goodness of fit can be used to know whether the MR time series from a certain voxel contains a signal induced by neural activities or not. We propose a method to decide the value of the regularization parameter of the generation model using difference between data from voxels in gray matter and white matter. We compare the analysis result using the measure we proposed, the result by the fractional amplitude of low-frequency fluctuation (fALFF) and the result by generalized linear model using time course of events in the experiment. To evaluate analysis methods, we analyzed event related fMRI data. Both the index we proposed and the fALFF measure were similar to the result of analysis with generalized linear model, that is, usual analysis of variance. However, the measure we proposed showed higher correlation coefficient because of constant linear relation including range with large neural activities.**

I. Introduction

In conventional experiments of fMRI(functional Magnetic Resonance Imaging), data are obtained alternating a time period for the experimental condition and a time period for the control condition. In the analysis, data recorded with two conditions are compared. In recent years, fMRI data without a control condition are often recorded, such as data recorded when subjects are just resting or watching movies.

In this study, we propose a method to analyze fMRI data when the times of occurrences of events supposed to cause neural activities are not available. In the method, we assume a generation model of convolution form of fMRI data. We also assume the sparsity of occurrences of events which cause neural activities. With these assumptions, we fit the generation model to fMRI data and separate the data into a signal caused by neural activities and the sum of other components such as noises or artifacts. Further, the obtained signal is separated into a hemodynamic response function and a time series representing the time when each event occurred. Therefore, the time series of events is estimated from the fMRI data not knowing the design of the experiment. In this algorithm, a measure of goodness of fit to the convolutive generation model is defined and estimated from fMRI data measured in one voxel. For the voxel with a sufficiently large value of the measure, we assume that the fMRI data contain a fluctuation caused by neural activities. We suggest that this goodness of fit can be used to know whether some neural activity exists in a voxel or not.

The proposed measure of goodness of fit includes two terms, that is, a term of the residual between data and the model and a term of the sparsity of event occurrences. We need to decide the value of the regularization parameter which balances these two terms. We propose a method to decide the value using difference between data from voxels in gray matter and white matter.

The index called the fractional amplitude of low-frequency fluctuation (fALFF) proposed by Zou et al. (2008) [1] is an analysis method that does not use the time series of event occurences in the experiment. It uses the ratio of low frequency fluctuations as a measure of neural activities. Caballero-Gaudes et al. (2013) [2] proposed a method using the Dantzig selector to estimate a time series of events from fMRI data. After estimating the time series of events for each voxel, they proposed a method to define a time course of events common to all voxels and used it to obtain a map of neural activation.

For evaluation, we compare the analysis result by the proposed method and the result by the fractional amplitude of low-frequency fluctuation (fALFF). We analyzed a set of event related fMRI data. For the data, we knew when each event in the experiment occurred but we did not use the information for these two methods. We also analyzed the same data by the conventional analysis of variance using time courses of events in the experiment. The method is standard for fMRI and it assumes the generalized linear model. It uses t statistic showing the magnitude of fluctuations from neural activities caused by events of the experiment. We compare analysis results by these three methods.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Algorithm

A. The estimation of a_τ and $s(t)$

We use an algorithm similar to Olshausen and Field (1996) [3]. The algorithm is also based on an algorithm by Nakamura and Inouye (2009) [4]. We define $x(t)$ as an fMRI time series measured at a certain voxel. t represents the time or the index of fMRI scan. $s(t)$ is a time series and it takes 1 when an event such as presentation of a visual stimulus occurs, and takes 0 otherwise. We define a_τ as the fluctuation in fMRI data that an event caused. a_τ is a hemodynamic response function. With these variables, we suppose that the following equation represents the generation process of fMRI data

$$x(t) = \sum_{\tau=1}^{\tau_{\max}} a_\tau s(t-\tau+1) + \epsilon(t) \tag{1}$$

$\epsilon(t)$ represents additive noise and it follows a normal distribution and is uncorrelated in time. We assume that estimation of $s(t)$ can take any real number. We assume that the probability distribution of $s(t)$ is sparse as Olshausen and Field (1996) [3]. In summary, we assume the generation model of eq. (1) and the sparse distribution for the time series of stimuli. Then we estimate a_τ and $s(t)$ to minimize the next cost function.

$$L = \sum_t \left\{ x(t) - \sum_\tau a_\tau s(t-\tau+1) \right\}^2 + \lambda_s \sum_t \log(1+\beta s(t)^2) \tag{2}$$

This cost function can be interpreted as a log-likelihood function when $x(t)$ is given. Here we assume that the probability distribution of $s(t)$ is a truncated Cauchy distribution. This distribution is sparse.

$$p(s) = \begin{cases} \left(\frac{2}{\sqrt{\beta}} \tan^{-1}(\sqrt{\beta}\gamma)\right)^{-1} \frac{1}{1+\beta s^2}, & |s| \le \gamma \\ 0, & |s| > \gamma \end{cases} \tag{3}$$

We assume that $x(t) - \sum_\tau a_\tau s(t-\tau+1)$ follows a normal distribution. The probability distribution of $x(t)$ is

$$\begin{aligned} &p(x(1), x(2), \cdots, x(t), \cdots, x(T)) \\ &= \prod_{t=1}^{T} p(s(t)) \exp\left\{ \frac{(x(t) - \sum_\tau a_\tau s(t-\tau+1))^2}{\lambda_s} \right\} \end{aligned} \tag{4}$$

and eq. (2) is its log-likelihood function. The algorithm we use for estimation of parameters is the maximization of the posterior probability.

First, we determine the initial values of a_τ and $s(t)$. Then we iterate procedures for updating a_τ and $s(t)$ alternately. When we estimate both a_τ and $s(t)$ from data, we cannot estimate magnitudes of these variables. Therefore, we assume that $\sum_{\tau=1}^{\tau_{\max}} a_\tau^2 = 1$.

We update values of a_τ while values of $s(t)$ are fixed. Values of a_τ that minimizes the cost function L under the constraint $\sum_{\tau=1}^{\tau_{\max}} a_\tau^2 = 1$ can be obtained from the following equation. It is derived by Lagrange's undetermined multiplier method.

$$\begin{aligned} &\frac{\partial L}{\partial a_u} + 2\mu a_u = \\ &2\sum_t \left(\sum_\tau a_\tau s(t-\tau+1) - x(t) \right) s(t-u+1) + 2\mu a_u = 0 \\ &u = 1, 2, \cdots \tau_{\max} \end{aligned} \tag{5}$$

We can write the equation with vectors and matrices as follows. We define $\boldsymbol{a} = (a_1, a_2, \cdots, a_{\tau_{\max}})^T$. $\boldsymbol{c}_{sx}$ is a vector of length $\tau_{\max}$ and its k-th element is

$$\{\boldsymbol{c}_{sx}\}_k = \sum_t x(t) s(t-k+1) \tag{6}$$

The kl element of $\tau_{\max}$-by-$\tau_{\max}$ matrix C_{ss} is

$$\{C_{ss}\}_{kl} = \sum_t s(t-k+1) s(t-l+1). \tag{7}$$

Then, we can write eq. (5) as follows.

$$\frac{\partial L}{\partial \boldsymbol{a}} + 2\mu \boldsymbol{a} = 2C_{ss}\boldsymbol{a} - 2\boldsymbol{c}_{sx} + 2\mu\boldsymbol{a} = 0$$

By multiplying both sides by $\boldsymbol{a}^T$, we obtain the following equation.

$$\mu = -\boldsymbol{a}^T C_{ss} \boldsymbol{a} + \boldsymbol{a}^T \boldsymbol{c}_{sx} \tag{8}$$

To derive update rules by the Newton method, we calculate the second derivatives.

$$\frac{\partial}{\partial a_v}\left(\frac{\partial L}{\partial a_u} + 2\mu a_u \right) = 2\sum_t s(t-v+1) s(t-u+1) + 2\mu\delta_{uv} \tag{9}$$

Here, δ_{uv} is the Kronecker delta. From these equations, the next equation can be obtained.

$$\frac{\partial}{\partial \boldsymbol{a}}\left(\frac{\partial L}{\partial \boldsymbol{a}} + 2\mu\boldsymbol{a} \right) = 2C_{ss} + 2\mu I \tag{10}$$

I is the identity matrix.

Finally, the next update rules are obtained.

$$\begin{aligned} \boldsymbol{a} \leftarrow& \boldsymbol{a} - \left(\frac{\partial}{\partial \boldsymbol{a}} \left(\frac{\partial L}{\partial \boldsymbol{a}} + 2\mu\boldsymbol{a} \right) \right)^{-1} \left(\frac{\partial L}{\partial \boldsymbol{a}} + 2\mu\boldsymbol{a} \right) \\ &= (C_{ss} + (-\boldsymbol{a}^T C_{ss}\boldsymbol{a} + \boldsymbol{a}^T \boldsymbol{c}_{sx}) I)^{-1} \boldsymbol{c}_{sx} \end{aligned} \tag{11}$$

$$\boldsymbol{a} \leftarrow \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \tag{12}$$

Next, we fix values of a_τ and update $s(t)$ simply by the gradient descent method with the cost function L.

$$s(t) \leftarrow s(t) - c\left(\frac{\partial L}{\partial s(t)} \right) \tag{13}$$

$$\begin{aligned} \frac{\partial L}{\partial s(t)} =& 2\sum_u \left(\sum_\tau a_\tau s(t+u+1-\tau) - x(t+u) \right) a_u \\ &+ \lambda_s \frac{2\beta s(t)}{1+\beta s(t)^2} \end{aligned} \tag{14}$$

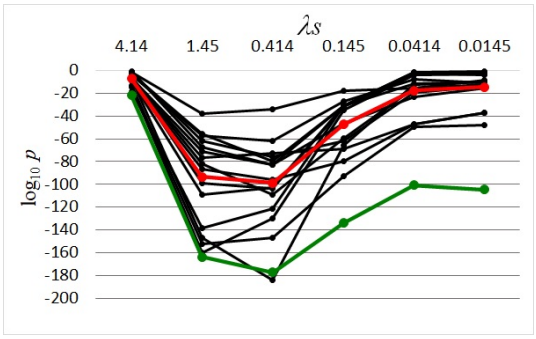


Fig. 1. p-values for the alternative hypothesis that values of L are smaller for voxels in the gray matter than the white matter. The common logarithms of p-values are shown. A black line is a result for each subject. The red line shows the mean of the common logarithm of p-values of all subjects. The green line shows the common logarithm of the mean of p-values of all subjects.

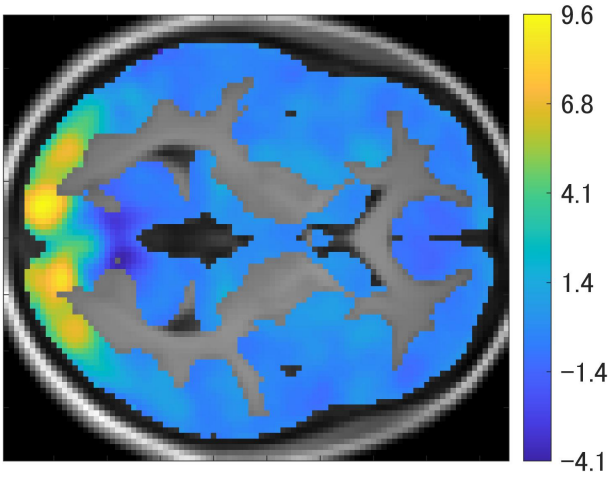


Fig. 2. Values of t-statistic using the usual analysis of variance approach. Mean values of 15 subjects. The slice of z=2.0mm.

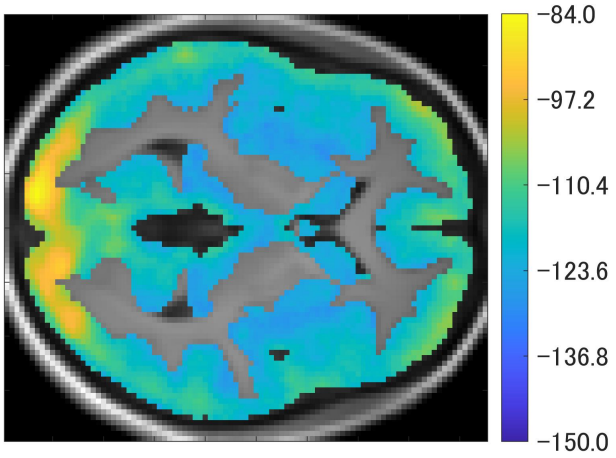


Fig. 3. Values of the measure of goodness of fit times -1, $-L$. Mean values of 15 subjects. The slice of z=2.0mm.

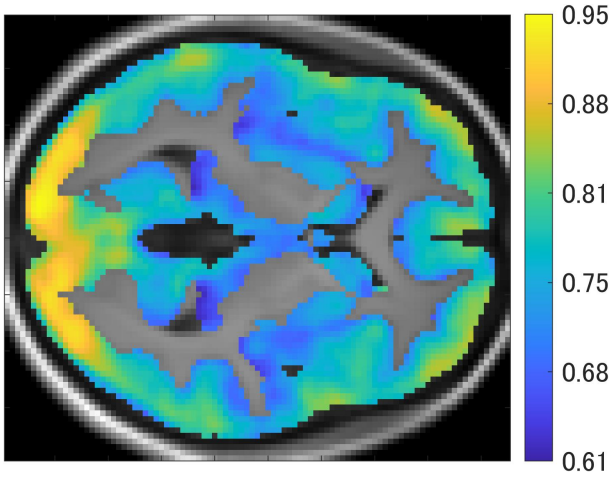


Fig. 4. Values of fALFF. Mean values of 15 subjects. The slice of z=2.0mm.

When an updated value of $s(t)$ is outside the range specified by its assumed probability distribution $p(s)$, we substitute it by the closest value in the range. That is, when $s(t)$ is updated by the gradient descent method, if $s(t) < -\gamma$, we substitute the update rule by $s(t) = -\gamma$. If $s(t) > \gamma$, then we apply $s(t) = \gamma$. We iterate the step updating a_τ with fixed $s(t)$ and the step updating $s(t)$ with fixed a_τ alternately, until the change of the cost function due to the update becomes sufficiently small. As described above, we estimate a_τ and $s(t)$ that minimize L.

The first term of L represents the residual between the data and the generation model and the second term represents sparseness of $s(t)$. λ_s is the parameter which balances these two terms. λ_s is a type of hyperparameter called regularization parameter. In the next section, we explain the method to determine the value of this parameter.

B. Estimation of the regularization parameter λ_s

We determine the value of λ_s as follows. First, we choose candidate values of λ_s. For each candidate value of λ_s, we estimate a_τ and $s(t)$ using the algorithm in the previous section that minimizes the cost function L. With estimated a_τ and $s(t)$, we examine the difference between values of L of voxels in white matter and voxels in gray matter. We calculate p-values of t-test corresponding to the alternative hypothesis that L is smaller in voxels in gray matter and select the value of λ_s with the smallest p-value.

With this λ_s, we determine β as follows. First, we assume that the mean and the variance of the data $x(t)$ are normalized to μ_x and one. If we assume the following approximation,

$$\sum_{t=1}^{T}\left\{\sum_{\tau=1}^{\tau_{\max}} a_\tau s(t-\tau+1)\right\}^2 \approx \left\{\sum_{\tau=1}^{\tau_{\max}} a_\tau^2\right\}\left\{\sum_{t=1}^{T} s(t)^2\right\} \tag{15}$$

then, from $\sum_{\tau=1}^{\tau_{\max}} a_\tau^2 = 1$, we obtain

$$\sum_{t=1}^{T}\left\{\sum_{\tau=1}^{\tau_{\max}} a_\tau s(t-\tau+1)\right\}^2 \approx \left\{\sum_{t=1}^{T} s(t)^2\right\} \tag{16}$$

Since $\epsilon(t)$ follows normal distribution with mean 0 and vari-

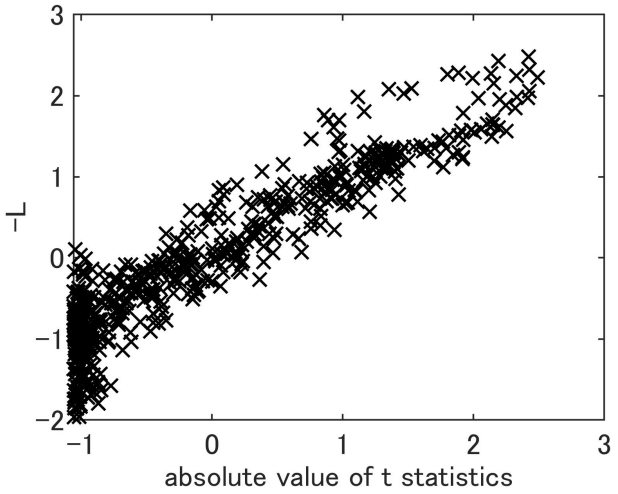


Fig. 5. The plot of $-L$ against the absolute value of t statistic. Both values are means of 15 subjects and normalized to mean 0 and variance 1. They are calculated from data recorded in gray matter on the z=2.0mm slice

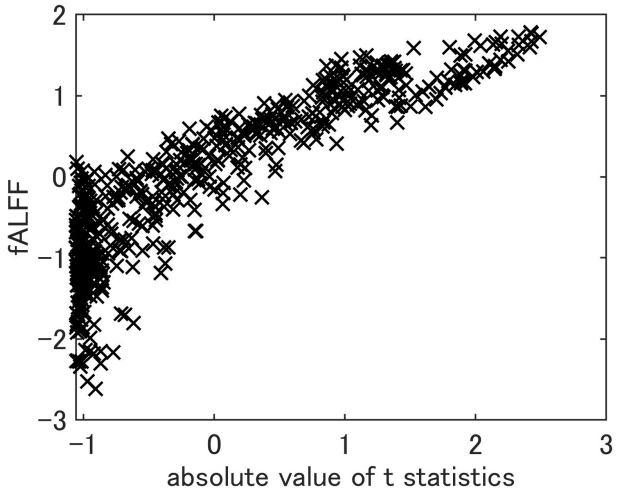


Fig. 6. The plot of fALFF against the absolute value of t statistic. Both values are means of 15 subjects and normalized to mean 0 and variance 1. They are calculated from data recorded in gray matter on the z=2.0mm slice

ance $\lambda_s/2$ and it is independent of $s(t)$,

$$\begin{aligned} \mathrm{E}\left[x(t)^2\right] &= \mathrm{E}\left[\left(\sum_{\tau=1}^{\tau_{\max}} a_\tau s(t-\tau+1)+\epsilon(t)\right)^2\right] \\ &= \mathrm{E}\left[\left(\sum_{\tau=1}^{\tau_{\max}} a_\tau s(t-\tau+1)\right)^2\right] + \mathrm{E}\left[\epsilon(t)^2\right] \\ &\approx \mathrm{E}\left[s(t)^2\right] + \frac{\lambda_s}{2} \end{aligned} \tag{17}$$

From the probability density function of s of eq. (3), the average of s^2 is

$$\mathrm{E}[s^2] = \frac{\gamma}{\sqrt{\beta}} \frac{1}{\tan^{-1}\left(\sqrt{\beta}\gamma\right)} - \frac{1}{\beta} \tag{18}$$

Since

$$\mathrm{E}\left[x(t)^2\right] = 1 + \mu_x^2 \tag{19}$$

we obtain the following equation.

$$\frac{\gamma}{\sqrt{\beta}} \frac{1}{\tan^{-1}\left(\sqrt{\beta}\gamma\right)} - \frac{1}{\beta} + \frac{\lambda_s}{2} = 1 + \mu_x^2 \tag{20}$$

By using eq. (20), if the values of λ_s and μ_x are given, we can calculate β.

III. Results

A. Data

We analyzed the data of Aminoff and Tarr (2015)[5]. They are available on the OpenfMRI[6] which is a website for data sharing. The data set includes data of 15 subjects. They are 12 females and 3 males and one subject is left-handed. They are 18 to 33 years old, with an average of 24. Aminoff and Tarr (2015)[5] studied information processing of the relationship between objects in the visual field. For this reason, they used an image in which cells are arranged in 4 by 4 and various objects are contained in some of these cells. In addition, images of landscapes and objects that are often seen in daily life and phase randomized images are used. In an image presentation block, four images are selected from these images, one image is presented twice and one presentation lasts 2[s]. One image presentation block lasts 16[s] and this block and a block that shows only a fixation point for 10 [s] are alternately arranged. One fMRI scan session includes 18 image presentation blocks. In our analysis, we focused on the difference between the time period in which images were presented and the time period in which only the fixation point was presented. The image presentations are treated as events of the experiment. We assume that a fluctuation is caused in fMRI data by the neural activity related to the events. We estimate the fluctuation and use it in the analysis of variance approach. We analyzed fMRI data measured in the slice of z= 2.0mm in MNI coordinates. The slice is in the primary visual cortex.

B. Determination of a value of λ_s

In the analysis of fMRI data, we used the following candidate values for λ_s. $\{4.14, 1.45, 0.414, 0.145, 0.0414, 0.0145\}$. With each value of λ_s, we calculate values of L for all voxels on the slice. Next, we calculate the p-value for a one-sided test of the alternative hypothesis that values of L are smaller for voxels in the gray matter than the white matter. Obtained p-values for each subject and the mean are shown in Fig.1.

For 9 subjects, the p-value was the smallest at $\lambda_s = 0.414$ and for 6 subjects, it was the smallest at $\lambda_s = 1.45$. Two kinds of means were both the smallest at $\lambda_s = 0.414$, but the difference from values with $\lambda_s = 1.45$ was small. We mainly present the results with $\lambda_s = 0.414$.

C. Measure of goodness of fit

We present results of the measure of goodness of fit L, fALFF, and the t statistic of the usual analysis of variance approach. Before we calculate a value of L, a time series $x(t)$

is normalized to have a mean of 1.25 and a variance of 1.00. Values of parameters are $\gamma = 10$ and $\tau_{\max} = 8$.

We calculated the means of 15 subjects of these variables and present them. In Fig. 2, we present the values of t statistic from the usual analysis of variance approach. In Fig. 3, we present the measure of goodness of fit to the convolutive generation model with sparse time series of event occurrences.The measure is multiplied by -1. In Fig. 4, we show values of fALFF. Both values of fALFF and values of $-L$ look similar to the values of t statistic. In some areas where t statistic is small, fALFF and $-L$ show large values.

We show the plot between these variables. The plot of $-L$ against the absolute value of the t statistic is shown in Fig. 5 and it shows high correlation. Both values are normalized to mean 0 and variance 1. The absolute value of the t statistic is used because a small negative value of the t statistic also means a relation with the experimental condition. The plot of fALFF against the absolute value of the t statistic is shown in Fig. 6 and it also shows high correlation. When the t statistic increases, fALFF stops increasing with it at some point and the slope between two values changes. The measure of goodness of fit keeps the same slope with large values of t statistic. The correlation coefficient between fALFF and the absolute value of the t statistic was 0.883 and the one between $-L$ and the absolute value of the t statistic was 0.929. With $\lambda_s = 1.45$, the latter was 0.954.

IV. Discussion

We proposed a method to obtain a measure of goodness of fit to the convolution model of fMRI. The measure showed high correlation with the conventional t statistic. The value of fALFF also showed high correlation with the t statistic, but the measure of goodness of fit showed higher correlation. The time course of the experiment of analyzed data is periodic. The method of fALFF may have worked well because of this periodicity. The measure L which we proposed does not utilize periodicity and may work also well for data with which we cannot assume periodicity of occurrences of events. The measure could be used to detect candidate areas of neural activation without making an assumption about what kind of event caused them.

For the proposed measure, we did not show how to decide the threshold of significance. The value of the regularization parameter λ_s determined by our method did not give the highest correlation among the candidate values. These problems should be considered in the future.

Acknowledgment

We thank Dr. Aminoff, Dr. Tarr and their colleagues for making their data public. The data were obtained from OpenfMRI website[6] and the accession number is ds000149.

References

[1] Q.-H. Zou, C.-Z. Zhu, Y. Yang, *et al.*, "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fmri: Fractional ALFF," *Journal of neuroscience methods*, vol. 172, no. 1, pp. 137–141, 2008.

[2] C. Caballero Gaudes, N. Petridou, S. T. Francis, I. L. Dryden, and P. A. Gowland, "Paradigm free mapping with sparse regression automatically detects single-trial functional magnetic resonance imaging blood oxygenation level dependent responses," *Human brain mapping*, vol. 34, no. 3, pp. 501–518, 2013.

[3] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[4] W. Nakamura and Y. Inouye, "Measures of goodness of fit to convolution model for analysis of fMRI data," in *17th European Signal Processing Conference*, Glasgow, Scotland, UK, 2009, pp. 2387–2391.

[5] E. M. Aminoff and M. J. Tarr, "Associative processing is inherent in scene perception," *PLoS One*, vol. 10, no. 6, e0128840, 2015.

[6] OpenfMRI, https://openfmri.org/.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Detection model of sister chromatid cohesion defects based on Vision Transformer

Shinya Matsumoto*, Kan Okubo*, Takuya Abe, and Kiyoshi Nishikawa*
Department of Science, Tokyo Metropolitan University, Japan
E-mail: 0330abe@tmu.ac.jp Tel: +81-426772542
* Department of Systems Design, Tokyo Metropolitan University, Japan
E-mail: kanne@tmu.ac.jp, kiyoshi@tmu.ac.jp Tel: +81- 425858423

***Abstract*— To evaluate the function of cohesin and its regulators, classifications of chromosomes based on the positional relationship between sister chromatids are used. Although the analysis has been done manually by researchers, we recently developed CNN-based classification models to automate the analysis. Our purpose in this study is to improve the accuracy of classification. For this purpose, we applied Vision Transformer (ViT) since ViT is expected to achieve better performance in the vision processing system than CNN. We show that the performance varies depending on the set of layers used in fine tuning. After proper fine tuning, the ViT-based model achieves higher classification accuracy than CNN.**

I. INTRODUCTION

In this paper, we propose a model for detecting sister chromatids cohesion defects based on Vision Transformer (ViT).

Sister chromatids are two identical copies of the same chromosome that are formed after DNA replication and are physically attached to each other until the onset of cell division. Subsequently, sister chromatids are equally distributed to daughter cells during somatic cell division. The physical attachment of sister chromatids until somatic cell division occurs is mediated by a ring-shaped protein complex called cohesin and proteins that regulate cohesin function [1]. Defects in proteins that regulate cohesin function result in the loss of sister chromatid cohesion and abnormalities in equal partitioning during mitosis [2]. It has been reported that such abnormalities cause various diseases, including cancer, developmental syndromes and infertility [3], [4]. For this reason, the mechanism of cohesin regulation has been studied since its discovery in 1997 [5]. To evaluate the involvement of proteins in the regulation of cohesin, a common method used in such studies is to classify the sister chromatid shapes after depletion of the protein of interest and to investigate the proportion of these shapes.

Traditional process for evaluating sister chromatid shape consists of three major steps: preparation of chromosome spreads, shape classification, and calculation of the ratio of each class. Currently, shape classification is done manually by researchers. The classification of sister chromatid shapes by manual inspection by researchers has two problems: it is time-consuming, and the classification results are dependent of subjective judgment of researchers. In our previous studies [6][14], we developed convolutional neural network (CNN) based models to enable automatic classification of sister chromatid shapes. Especially, two models based on DenseNet161 [7] and ShuffleNetV2 [8] pre-trained models achieved high accuracy classification. However, whether these trained models can be applied to different datasets (captured in different conditions) has not been examined.

In this study, we applied novel image classification models based on ViT. ViT, an application of the Transformer [11] used in the field of natural language processing, has attracted much attention because of its ability to achieve higher classification accuracy than CNN models [10]. We used four pre-trained models based on ViT and two pre-trained models based on CNN, and compared the performance of these models with varying number of layers used in fine tuning.

II. CONVENTIONAL MODELS

A. *Conventional Models [6][14]*

Our goal is to perform shape classification according to the degree of sister chromatid spreads. Although there are no standard classification criteria, we consider a model using CNN for classifying the sister chromatids into three classes in [6]. Namely, we classified them into classA, classB, classC : classA for images in which the two chromosomes are tightly connected, classB for images in which the arms are separated, and classC for images in which the two chromosomes are almost separated at the centromere. Illustrated images of each class are shown in Fig. 1.

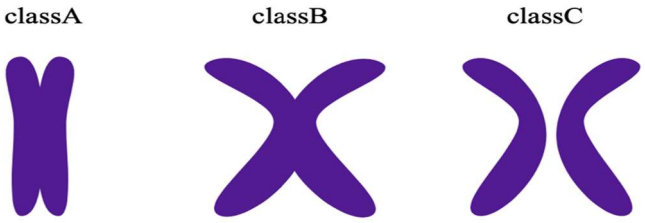


Fig. 1. Illustrated images of sister chromatids of each class

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In Ref. [6][14], image classification models are constructed by transfer learning and fine tuning using various CNN models which are pre-trained using images from ImageNet [9]. Because it is difficult to collect a large number of labeled images of sister chromatids for learning without pre-training. Based on the comparison of the accuracies, we proposed two image classification models based on DenseNet161 [7] or ShuffleNetV2 [8] and it is shown that they achieve high classification accuracy.

B. *Subject of This Study*

Although our previous models achieve the high accuracy rate, i.e., more than 90%, for the dataset described in the reference, in this study, we consider the models which can achieve higher classification validity. In previous studies, we only considered the models based on CNN. Here, the models based on the networks other than CNN, especially, the Vision Transformer, are considered.

III. PROPOSED MODEL

In this section, the proposed model is described. The proposed model is based on the Vision Transformer (ViT) instead of CNN.

A. *Vision Transformer (ViT) [10]*

For improving the accuracy of the previous method, we consider applying ViT as the base network instead of CNNs. ViT is derived from the Transformer [11] used in natural language processing (NLP) which employs the self-Attention mechanism instead of the convolutions used in CNNs.

Fig. 2 shows an architecture of ViT. The standard ViT consists of linear projection, Transformer encoder and Multi-Layer Perceptron (MLP) head. Besides, Transformer encoder consists of Multi Head Attention (MHA) blocks, which implement Attention mechanism, and MLP blocks. MHA extracts image features using the relevance of each image patch. The MLP consists of two layers with GELU (Gaussian Error Linear Unit). Transformer encoder in ViT receives each patch as an input, and the final layer of the Transformer encoder outputs the classification tokens used to make predictions.

B. *Multi Head Attention (MHA)*

MHA used in ViT is a mechanism to capture important parts of an image based on the similarity between each patch. MHA generates three vectors (query, key, and value) from each patch to the Transformer encoder of ViT. Then, we find a patch *Pr* with a large relevance based on the inner product between the query of a patch and the keys of all the patches. Next, we add the feature values of the patches with large relevance to the value of patch *Pr*. These operations are applied in parallel. As a result, ViT can recognize features of images globally. The number of operations performed in parallel is called Attention Heads.

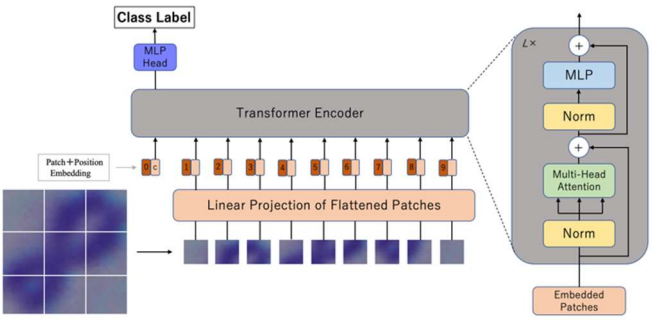


Fig. 2. Architecture of ViT [10]

C. *Model Variants*

In the following, we compare four pre-trained models of ViT available on the GitHub [12] or Torchvision [13]. Namely, we obtained ViT_Tiny and ViT_Small models from GitHub and the ViT_Base and ViT_Large models from Torchvision. These models of ViT are pre-trained using images from ImageNet [9]. Table I shows a comparison between those models. We use these models as the basic network of the proposed method.

Then, we perform fine tuning to each model with varying number of layers of Transformer encoder for which the fine tuning would apply. From the results, we investigate which layers are more effective for fine tuning.

Fig. 3 and Fig. 4 show the set of layers for which fine tuning would apply in each model. We note that ViT_Tiny, ViT_Small and ViT_Base have the same number of layers. These models differ in the number of attention heads and MLP size as Table I shows.

In the figures, LS stands for “layer set for fine tuning” and represents the set of layers used for fine tuning. For examples, ‘LS 1’ means that we use convolution layer through to fully connected layer of ViT_Tiny, ViT_Small and ViT_Base models for fine tuning. Similarly, ‘LS 2’ means that we use from transformer encoder 0 to fully connected layer.

Table I. The differences between the ViT models

Model	Input size	Patch size	Layers of Transformer encoder	MLP size	Attention Heads	Params
ViT_Tiny	224 × 224	16 × 16	12	768	3	5.5M
ViT_Small	224 × 224	16 × 16	12	1536	6	22M
ViT_Base	224 × 224	16 × 16	12	3072	12	86M
ViT_Large	224 × 224	16 × 16	24	4096	16	307M

D. *Fine Tuning*

As shown in Fig. 3 and Fig. 4, we divided the layers for fine tuning of ViT_Tiny, ViT_Small, and ViT_Base models into five patterns, and the layers for fine tuning of ViT_Large model into seven patterns.

During fine tuning, following parameters were used for all models: batch size is 32; number of epochs is 50; learning rate is 0.00003. Also, we used the Adam optimization algorithm, and cross-entropy as the loss function.

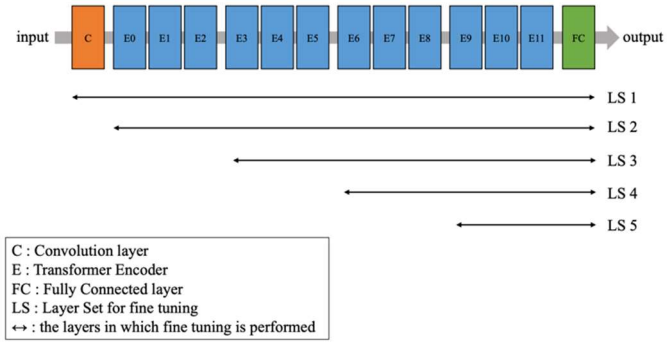


Fig. 3. The layer structure of ViT_Tiny, ViT_Small and ViT_Base models and the layers in which fine tuning is performed.

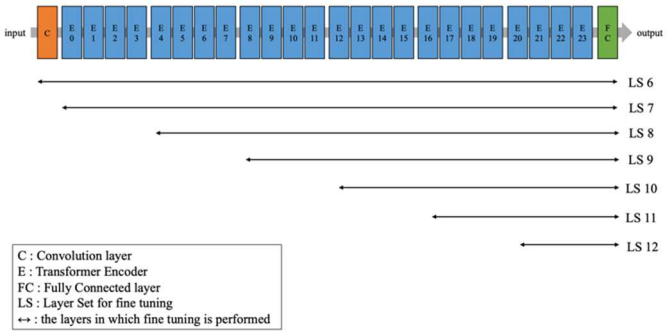


Fig. 4. The layer structure of ViT_Large model and the layers in which fine tuning is performed.

IV. EXPERIMENTS

Here, in order to verify the effectiveness of the proposed method, we show the results of experiments and the comparison with the previous method [6][14].

A. Datasets

Table II shows the composition of the datasets used in the experiments. These datasets are the same as used in [6]. We used the images of the sister chromatids taken by an author using a microscope. The images of these datasets were taken under same conditions at the same day. Note that, we applied the same pre-processing described in [6] to all images before the processing.

For the proposed method, fine tuning was performed using sister chromatids data consists of 667 images for training and 100 for validation in each class. To verify the effectiveness of the proposed method, we compared the accuracy of the proposed and the previous methods.

Table II. the composition of the datasets

	classA	classB	classC
Training Dataset	667 images (for training)	667 images (for training)	667 images (for training)
	100 images (for validation)	100 images (for validation)	100 images (for validation)
Dataset 1	100 images	100 images	100 images

B. Results

Here, we show the results of applying the four ViT models after fine tuning to Dataset 1 shown in Table II and compared their accuracy. Tables III and IV show the classification accuracy of Dataset 1 for each ViT model with different number of layers used for fine tuning.

Table III shows the results for the ViT_Tiny, ViT_Small and ViT_Base models. We can see from the table that, for ViT_Tiny and ViT_Base, the best results are obtained by the fine-tuning using LS 1. On the other hand, for ViT_Small, they are obtained when LS 2 is fine tuned. Table IV shows the results for the ViT_Large model and the best results are obtained when LS 6 is fine tuned.

Table III. The classification accuracy of the Dataset1 for ViT_Tiny, ViT_Small and ViT_Base models with different number of layers used for fine tuning.

	ViT_Tiny	ViT_Small	ViT_Base
LS 1	**0.9080**	0.9287	**0.9302**
LS 2	0.9053	**0.9300**	0.9244
LS 3	0.8998	0.9249	0.9227
LS 4	0.8553	0.9120	0.9118
LS 5	0.8340	0.8751	0.8802

Table IV. The classification accuracy of the Dataset 1 for ViT_Large model with different number of layers used for fine tuning.

	ViT_Large
LS 6	**0.9198**
LS 7	0.9164
LS 8	0.9176
LS 9	0.9116
LS 10	0.8993
LS 11	0.8938
LS 12	0.9007

Next, we compared the accuracy of the proposed and the conventional method [6][14]. Table V shows the best accuracy for each ViT model and the conventional method [6][14]. In Table V, we can see that the ViT_Base model has the highest accuracy.

Table V. Comparison of classification accuracy of each ViT model and the conventional method [6][14]

Proposed method	ViT_Tiny	0.9080
	ViT_Small	0.9300
	ViT_Base	**0.9302**
	ViT_Large	0.9198
Conventional method	DenseNet161	0.9002
	ShuffleNetV2	0.9198

C. Effectiveness of the proposed method

As described above, in our proposed method, we use the pre-trained ViT models, and then, apply the fine tuning to those models. Here, we compare the proposed method with a ViT without pre-training for verifying the effectiveness of the proposed method.

For that, we trained ViT_Base model from scratch using the training dataset shown in Table II and compared the classification accuracy for the same data. The results are shown in Table VI. We can see from the table that the accuracy of ViT model without pre-training is 0.8458 although the same network with pre-training achieves 0.9302 in the proposed method as shown in Tabel V. Hence, it is confirmed that the pre-training is effective in our situation where the number of images of the sister chromatids are not enough for training the models from scratch even if the ViT models are trained using the natural images from the dataset like ImageNet.

Note that we do not deny the possibility that the performance of a model trained from scratch may achieve better accuracy when a large number of images of sister chromatids are available. The comparison here shows the result in a situation where the number of images of sister chromatids available for training is limited.

Table VI. The classification accuracy of the Dataset 1 for ViT_Base model after fine tuning and non-pre-trained

After fine tuning	**0.9302**
Non-pre-trained	0.8458

D. Results with Different Dataset

To compare the performances of the proposed and the previous methods [6][14] for different evaluation data, we show the results of applying the methods to another dataset, which is referred as Dataset 2 in the following. As shown in Table VII, Dataset 2 consists of 104 images of class A, 345 of classB and 40 of class C. Note that Dataset 2 were taken using the same microscope but on a different day from the dataset shown in Table II were taken. In the following, we call the dataset in Table II as Dataset 1 to distinguish from Dataset 2. Fig. 5 shows examples from Dataset 1 (left) and from Dataset 2 (right).

Table VII. The composition of the dataset 2

	classA	classB	classC
Dataset 2	104 images	345 images	40 images

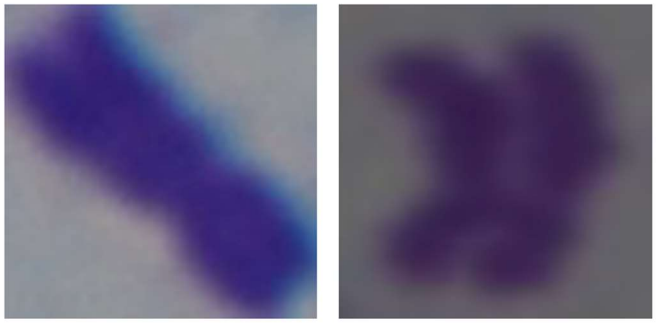


Fig. 5. Examples of the images from Dataset 1 (left) and Dataset 2 (right).

In Table VIII and IX, we show the results when Dataset 2 are used for the evaluation. Table VIII shows the results of the ViT_Tiny, ViT_Small and ViT_Base models. We can see from the table that, for ViT_Tiny and ViT_Small, the best results are obtained by the fine-tuning using LS 2. On the other hand, for ViT_Base, they are obtained when LS 4 is fine tuned. In Table IX, we show the results for the ViT_Large model and the best results are obtained when LS 6 is fine tuned.

Table VIII. The classification accuracy of the Dataset 2 for ViT_Tiny, ViT_Small and ViT_Base models with different number of layers used for fine tuning.

	ViT_Tiny	ViT_Small	ViT_Base
LS 1	0.6312	0.6964	0.6833
LS 2	**0.6511**	**0.7057**	0.6010
LS 3	0.6164	0.6830	0.6613
LS 4	0.4571	0.6669	**0.7016**
LS 5	0.3895	0.6083	0.6840

Table IX. The classification accuracy of the Dataset 2 for ViT_Large model with different number of layers used for fine tuning.

	ViT_Large
LS 6	**0.6660**
LS 7	0.6642
LS 8	0.6371
LS 9	0.5783
LS 10	0.5760
LS 11	0.6005
LS 12	0.5336

Finally, we compare the accuracy of the proposed and the conventional method for Dataset 1 and Dataset 2. Table X shows the best accuracy for each ViT model and the conventional method when we use Dataset 1 and Dataset 2. Note that, although the ViT_Small model has the highest accuracy for Dataset 2, we observed that the accuracies of both the proposed and the conventional methods decreased compared to them for Dataset 1. These results may suggest that the generalization of the network is not enough to achieve better classification regardless of the dataset from which evaluation data are taken. Hence, we need to increase the number of data used for learning to improve the robustness of the classification in the future.

Table X. Comparison of classification accuracy of each ViT model and the conventional method [6] for Dataset 1 and Dataset 2

		Dataset 1	Dataset 2
Proposed method	ViT_Tiny	0.9080	0.6511
	ViT_Small	0.9300	**0.7057**
	ViT_Base	**0.9302**	0.7016
	ViT_Large	0.9198	0.6660
Conventional method	DenseNet161	0.9002	0.6312
	ShuffleNetV2	0.9198	0.6191

V. Conclusions

In this paper, we proposed an image classification model for detecting the sister chromatids cohesion defects using the pre-trained ViT models. We applied fine-tuning to ViT with varying the set of layers to be tuned and compared in terms of the classification accuracy. By applying the model to the actual images of sister chromatids, we confirmed that the proposed model achieves the higher classification accuracy compared to our previous CNN based models. We also confirmed that the accuracy depends on the set of layers used for fine tuning.

VI. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K12170 to KN and JSPS KAKENHI Grant Number JP 22H05072 to TA.

References

[1] Yasuto Murayama, Catarina P. Samora, Yumiko Kurokawa, Hiroshi Iwasaki and Frank Uhlmann, "Establishment of DNA-DNA Interactions by the Cohesin Ring," *Cell*, vol. 172, no. 3, 2018, pp. 465-477.e15.

[2] Silvia Remeseiro, Ana Cuadrado, and Ana Losada, "Cohesin in development and disease," *Development*, vol. 140, no. 18, 2013, pp. 3715-3718.

[3] Yotaro Ochi, et al., "Combined Cohesin–RUNX1 Deficiency Synergistically Perturbs Chromatin Looping and Causes Myelodysplastic SyndromesMDS Induced by Combined Cohesin–Runx1 Deficiency," *Cancer discovery*, vol. 10, no. 6, 2020, pp. 836-853.

[4] David A. Solomon, et al., "Mutational inactivation of STAG2 causes aneuploidy in human cancer," *Science*, vol. 333, no. 6045, 2011, pp. 1039-1043.

[5] R. Kawasumi, T. Abe, et al, "ESCO1/2's roles in chromosome structure and interphase chromatin organization," *Genes and Dev*, vol. 2017, no. 31, pp. 2136-2150.

[6] D. Ikemoto, T. Taniguchi, K. Hirota, K. Nishikawa, K. Okubo, T. Abe, "Application of neural network-based image analysis to detect sister chromatid cohesion defects," *Scientific Reports*, 2023; 13, 2133.

[7] Gao Huang, et al, "Densely Connected Convolutional Netwoks," *CVPR*, 2017, pp. 4700-4708.

[8] Ningning Ma, et al, "ShuffleNet V2 : Practical Guidelines for Efficient CNN Architecture Design," *ECCV*, 2018, pp. 116-131.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *CVPR*, 2009, pp. 248–255.

[10] A. Dosovitskiy, et al., "An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale," *In ICLR*, 2021.

[11] A. Vaswani, et al., "Attention Is All You Need," *In NIPS*, 2017.

[12] GitHub:https://github.com/huggingface/pytorch-image-models/tree/main/timm

[13] torchvision https://pytorch.org/vision/stable/index.html

[14] D. Ikemoto, T. Abe, K. Okubo, K. Nishikawa, "Automatic classification of human chromosome shape by the convolutional neural network models," to be submitted.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

GRALA: modeling social information for microblog sentiment analysis from the view of balancing sparsity and smoothness of social contexts

Xiaomei Zou* and Shiyong Hu† Taihao Li‡
* Zhejiang Lab, China
E-mail: zouxiaomei@zhejianglab.edu.cn
† Hangzhou Applied Acoustics Research Institute, China
E-mail: hsy0578@163.com
‡ Zhejiang Lab, China
E-mail: lith@zhejianglab.edu.cn, corresponding author

***Abstract*—As online social networks provide people with a popular and convenient platform to express opinions, microblog sentiment analysis has become a hot research area. Currently, social relations between microblogs are extracted and integrated into content-based microblog sentiment analysis models. However, these methods take all links between microblogs into consideration, ignoring that some links are noisy and some neighbor microblogs are ambiguous semantically, resulting in the insufficiency of sentiment analysis. To deal with this problem, we propose a method that introduces sparse modeling of social contexts and tries to balance the sparsity and smoothness of social contexts to integrate social information prior knowledge into the content-based models in this paper. Specifically, a microblog graph is built based on sentiment consistency and emotional contagion, then we not only formalize the smoothness assumption of this graph but also introduce a graph-structured sparse regularization term to automatically select microblogs with high quality among all neighbors to facilitate sentiment analysis. Extensive experiments on two real-world public datasets show that our method achieves better performance than other baseline methods, indicating the effectiveness of sparse modeling of graph structure for microblog sentiment analysis.**

I. Introduction

With the rapid development of Web 2.0 and mobile communication devices, online social networks like microblogging services have become very popular for users to share information and opinions about various topics swiftly and conveniently. Twitter and Sina Weibo are among the most representative platforms. Almost every aspect of people's daily lives is published on these platforms, as a result, more and more user-generated content is flooded on the internet. These activities make mobile social networks become a huge resource repository, which draws the attention from governments, enterprises, and scholars. Microblog sentiment analysis has become a hot research topic in natural language processing and data mining. It arises in many applications such as recommendation systems, stock price prediction, and timely crisis management [1], [2], [3]. For example, due to the popularity, wide coverage, and rapid spread of microblogging platforms, microblog sentiment analysis can provide technical support for public opinion monitoring and early warning.

The problem of microblog sentiment analysis is to predict the sentiment (such as positive and negative) for microblog text. The existing works mainly take this problem as a classification task. Usually, they follow approaches of text classification: features are extracted from texts and represented as a vector by the bag of features, then machine learning methods are applied by implicitly assuming that microblog texts are independent [4], [5], [6], [7]. However, different from other texts such as news reports, microblog texts are very short and ambiguous. Slang, abbreviations, misspelled words, and other informal expressions are very common. It presents numerous difficulties in practice by exploiting textual data for microblog sentiment analysis. As a result, the accuracy of content-based methods drops sharply in real applications.

Microblogs lie in the environment of online social networks. These platforms provide rich social context information other than texts. For example, according to the theory of sentiment consistency [8], a user's sentiments have a higher probability to be the same in a short period about a specific topic. In addition, due to the theory of emotional contagion, a user's sentiment state can affect people around him or her [9], [10], [11], [12], which has been observed in online social networks [13], [14]. Inspired by these theories, some studies explore the utilization of social context information to facilitate microblog sentiment analysis [15], [16], [17], [18], [19], [20], [21]. They develop different methods to extract social context information between microblogs and use the extracted information to build a microblog graph and assume the graph is smooth, then they formalize the smoothness by graph Laplacian regularization or graph neural networks to integrate social knowledge information with microblog content. These methods utilize all the connections between microblogs in the graph by assuming that texts of all neighbors can facilitate sentiment analysis of the central microblog. However, this assumption does not always hold. Some neighbor microblogs are unfavorable to analyzing the sentiment of the central microblog if their text information is ambiguous. Besides, when extracting social

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

context information by constructing the microblog graph, some noises may be introduced. It would be helpful to remove these noisy connections when analyzing sentiment. To solve these problems, we investigate how to remove noises and choose high-quality neighbors to facilitate microblog sentiment analysis automatically. specifically, we first build a microblog graph to extract social knowledge information, then we discuss how to integrate the graph with microblog texts and extract beneficial relationships from the graph in a sparse modeling method. In addition, we also take the smoothness of the built microblog graph into consideration by balancing it with the sparse term. The main contributions of our work are: 1) building a microblog graph based on the sentiment consistency and emotional contagion theory. 2) proposing a graph structured sparse regularization to automatically choose helpful microblogs from all neighbors and use related microblogs to perform feature selection. 3) balancing the graph structured sparse regularization term and the graph smoothness regularization term. 4) elaborating the effectiveness of the proposed method by conducting extensive experiments on two datasets.

II. Related Works

Microblog sentiment analysis has been a hot research topic, attracting enormous attention from researchers, companies, and governments. Existing methods mostly regard microblog texts as the only source and focus on extracting better representations for sentiment analysis by lexicons or machine learning methods. Some researchers utilized word lexicons to incorporate sentiment analysis. For example, Wu et al. built a new lexicon for Chinese microblog sentiment analysis by exploiting word emoticons relationships, word sentiment relationships, and existing sentiment lexicons [22]. A sentiment lexicon was built in [23] by selecting representative seeds and propagating sentiment labels among a graph of sentiment units built by social relations, topic, and so on. There are also some researchers who exploit machine learning methods for microblog sentiment analysis. [24] utilized probabilistic word sampling and word identification for zero masking or contextual replacement to augment sentiment datasets. Zhang et al. argued that the existing pre-trained methods considered features and fine-tuning methods separately, so they proposed a multi-task representation method [25]. However, these methods are content-centric, some researchers argue that their limitations come from focusing on content information only while ignoring language's social factors [26], [27].

Social contexts are explored in many social network analysis tasks such as recommendation systems [28], information retrieval [29], and fake news detection [30]. There are also few scholars studying the influence of social contexts on microblog sentiment analysis. [15] used the direct relationship between users, and integrated social relations with the least square method. [19] introduced community information for microblog sentiment analysis by multi-task learning. Although these methods integrated social contexts with microblog sentiment analysis, they ignore that social contexts may be noisy. There are also works that exploited graph neural networks to incorporate social contexts with text information [20], [21]. Although these works can assign different weights to different neighbors in the graph, they did not investigate how the sparse modeling of a graph can influence sentiment analysis and are not well explainable. [17] utilized user information and friend information to refine the sentiment labels of microblogs predicted by the logistic regression model or support vector machine. This work is the most similar to ours, as they also dealt with the noises in social contexts. Our work differs from this work in several fundamental aspects. [17] explored social contexts in the testing stage, while our work explores social contexts in a sparsity view in the training stage and investigates how sparse modeling influence the feature selection. In addition, our work not only exploits sparse modeling, we also take the smoothness of social contexts into consideration.

III. Problem Fromulation

A. Background & Notations

Matrixes are used uppercase bold letters such as $\mathbf{X}$ to represent, and vectors are represented by lowercase bold letters like $\mathbf{x}$. Lowercase letters like α denote scalar values. $\mathbf{X}_{i*}$ and $\mathbf{X}_{*j}$ denote the i-th row and the j-th column of the matrix $\mathbf{X}$ respectively. $\mathbf{A}_{ij}$ represents the element located at the i-th row and j-th column. $\mathrm{Tr}(.)$ computes the trace of a matrix. The norm $||.||_1$ is the sum of vector magnitudes in a space, the norm $||.||_2$ means to make the value of every dimension to be small, which can also be explained by minimizing the distance between the origin and the coordinate in the vector space.

Given that $\mathbf{x}^i \in R^m$ represents the i-th microblog feature vector, m is the number of features, then the sentiment prediction problem can be defined as a loss function $\mathcal{L}(\mathbf{x}^i, \mathbf{w}, y^i)$ which quantifies the loss between the prediction $h_{\mathbf{w}}(\mathbf{x}^i)$ of a classifier and the true class label $y^i \in \mathcal{Y}$ associated with the sample $\mathbf{x}^i \in \mathcal{X}$. $\mathcal{X}$ represents the microblog texts set, $\mathcal{Y}$ represents the sentiment label set. $\mathbf{w}$ is the parameters of the classifier, which represents the feature weight vector. Given the microblog training set $\{(\mathbf{x}^i, y^i)\}_{i=1,\cdots,n}$, we can get a classifier $\mathbf{w}$ by optimize the set of feature weights using:

$$\mathbf{w}^* = \mathrm{argmin}_{\mathbf{w}} \quad \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}^i, \mathbf{w}, y^i), \tag{1}$$

where n represents the number of microblogs in the traning set. $\mathcal{L}(\mathbf{x}^i, \mathbf{w}, y^i)$ can be logistic regression, least squares or other losses. In this paper, we use the squared loss as an example: $\mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^i\mathbf{w} - y^i)^2$. It also can be rewritten in the form of matrix calculation: $\mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) = \frac{1}{2}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$, where $X \in R^{n\times m}$ is the feature matrix of the training set, m represents the number of features, $\mathbf{x}^i = \mathbf{X}_{i*}$ represents the feature vector of the i-th microblog. $\mathbf{y} \in R^{n\times 1}$ represents the label vector. We focus on binary sentiment analysis, so for any microblog, its sentiment label satisfies: $y^i \in \{-1, +1\}$. For a positive microblog, $y^i = +1$.

B. Regularization

Regularization is proven to be of paramount importance in language modeling and text classification [31]. It can not only avoid overfitting in the training stage, but also take prior knowledge into consideration [32]. Unlike texts in traditional media, microblogs are extremely short and ambiguous, containing lots of abbreviations, slang, and personalized expressions. This characteristic often leads to a high-dimension feature vector space when representing microblog texts by the unigram model. In addition, it makes the feature vector very sparse. So the sparse regularization L_1 norm is utilized to reconstruct a sparse space from the high-dimension one:

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \quad \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + \beta||\mathbf{w}||_1, \tag{2}$$

where β is a hyper-parameter. The L_1 norm assumes a multivariate Laplacian prior on the feature weight vector $\mathbf{w}$. Minimizing the L_1 norm will set some feature weights to 0, which means automatically selecting features. The L_1 norm regularization is used as an additional constraint to the classification problem, by relaxing into a penalty term controlled via a hyper-parameter β.

Emotional contagion

IV. Microblog Sentiment Analysis Via Graph Structured Sparse Regularization

A. Modeling microblog content informaiton

In a sentiment prediction problem, pre-processing is the first step and its quality can influence the analysis results. To better represent microblog texts, references [33] and [34] conduct experiments to compare the effects of different preprocessing methods and features. Following these studies, we use the unigram model with term presence and do not remove punctuation and stop words as they carry sentiment information. In addition, online social networks provide some mechanisms for users to better express their emotions and opinions, such as hashtags, @-mentions, and emoticons. In microblogs, hashtags are marked with the symbol # and used to indicate the topic of microblogs. For example, the microblog "I love #ipad" infers this is about "ipad". As topic information is very important for sentiment analysis, we preserve hashtags in microblog texts. Different from formal texts in which body language and verbal tone do not translate, emoticons are a popular way for people to convey their nuanced meaning in online social networks. So emoticons are extracted as features by regular expressions. @-mentions and URLs are also preserved in this paper as they may contain some sentiment clues.

B. Extracting and Modeling Social Knowledge Information

In addition to text information, online social networks also contain the information of the user (such as following and followee information) who posted the text. The environmental information may provide sentiment clues. In this section, we construct a microblog graph to reveal their sentiment clues by using the social knowledge information from sociological theories. The microblog graph is represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V}$ is composed of microblogs, and the edge set $\mathcal{E}$ contains connections between microblogs. We use matrix $\mathbf{A}$ to represent its adjacency matrix. Firstly, the sentiment of a specific user tends to be consistent in a short period, which is called the sentiment consistency theory [8]. This phenomenon means that microblogs posted by the same user within a short time interval are more likely to have the same sentiment. To extract this prior social knowledge information, we construct a microblog graph $\mathcal{G}_1$ whose adjacency matrix $\mathbf{A}_{sc}$ can be computed by $\mathbf{A}_{sc} = \mathbf{U}^T \times \mathbf{U}$. $\mathbf{U} \in R^{d \times n}$ denote the user-microblog matrix, n is the number of microblogs, d is the number of users. $\mathbf{U}_{ij} = 1$ if the i-th user posted the j-th microblog. Otherwise, $\mathbf{U}_{ij} = 0$. In this case, if microblog i and microblog j are posted by the same user, then $\mathbf{A}_{sc_{ij}} = 1$, otherwise, $\mathbf{A}_{sc_{ij}} = 0$. Secondly, sentiment diffusion has been reported in online social networks based on emotional contagion and has close relationships with microblog sentiment polarities in social sciences [9], [35]. The analysis indicates that emotional contagion may be helpful for microblog sentiment analysis. Emotional contagion reveals that if two microblogs are published by friends, their sentiment polarities are more likely to be similar. Based on this theory, we use the friend information $\mathbf{F} \in R^{d \times d}$ provided by microblogging services to build the microblog graph $\mathcal{G}_2$. Its adjacency matrix can be computed by $\mathbf{A}_{ec} = \mathbf{U}^T \times \mathbf{F} \times \mathbf{U}$. $\mathbf{F}_{ij} = 1$ if and only if the i-th user is a friend of the j-th user, otherwise, $\mathbf{F}_{ij} = 0$. Therefore, $\mathbf{A}_{ec_{ij}} = 1$ when the i-th microblog and the j-th microblog are posted by connected users, otherwise, $\mathbf{A}_{ec_{ij}} = 0$. After all these steps, we get a new graph $\mathcal{G}$ whose adjacency matrix is $\mathbf{A} = \mathbf{A}_{sc} + \mathbf{A}_{ec}$.

Based on the discussion above, the built micorblog graoh is smooth. Microblog texts connected by edges in a graph $\mathcal{G}$ are published by the same user or friends, which means that they usually convey similar sentiment. Therefore, we try to construct a connection to make the sentiment of connected microblogs as close as possible. Similar sentiment scores are assigned to connected microblogs in our method. **Graph Laplacian regularization** is used to model the smoothness:

$$\begin{aligned} &\text{argmin}_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}||\hat{y^i} - \hat{y^j}||_2^2 \\ =&\text{argmin}_{\mathbf{w}} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}||\mathbf{x}^i\mathbf{w} - \mathbf{x}^j\mathbf{w}||_2^2 \\ =&\text{argmin}_{\mathbf{w}} \quad \text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}), \end{aligned} \tag{3}$$

where $\hat{y^i}$ represents the predicted sentiment label, $\mathbf{L}$ denotes the Laplacian matrix of graph $\mathcal{G}$ and can be computed by $\mathbf{L} = \mathbf{D} - \mathbf{A}$. $\mathbf{D}$ represents the diagonal matrix whose entries represent the number of neighbors of the i-th microblog.

In addition to graph Laplacian regularization, we introduce a novel method (**graph structured sparse regularization**, also called **graphical lasso regularization**) to model the social knowledge information prior in this paper. In graph Laplacian regularization, the influence of all neighbors of a microblog is taken into consideration by assuming that the texts of all neighbors can help sentiment analysis. Similar to the L_2 norm regularization, graph Laplacian regularization results in

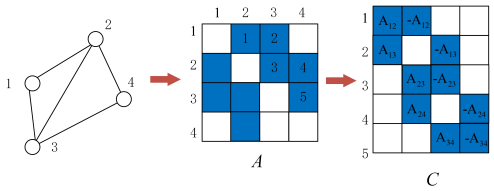


Fig. 1. An example of construction of C

reducing values. However, when building the microblog graph, some noises are introduced and not all neighbors are helpful for microblog sentiment. Some microblogs are predicted with high confidence coefficient because they have words with clear sentiment polarities, while many other microblogs without clear sentiment words are predicted as incorrect labels with a high probability. Under this scenario, it would be helpful to choose the microblog texts containing sentiment words among all the neighbors. To solve this problem, A graph structured sparse regularization term is proposed. It can be formulated as:

$$\begin{aligned} & \text{argmin}_{\mathbf{w}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}||\hat{y}^i - \hat{y}^j||_1 \\ = \ & \text{argmin}_{\mathbf{w}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}||\mathbf{x}^i\mathbf{w} - \mathbf{x}^j\mathbf{w}||_1 \\ = \ & \text{argmin}_{\mathbf{w}} \quad ||\mathbf{C}\mathbf{X}\mathbf{w}||_1, \end{aligned} \tag{4}$$

where $\mathbf{C} \in R^{|\mathcal{E}|\times n}$, $|\mathcal{E}|$ represents the number of edges in graph $\mathcal{G}$. Fig. 1 shows an example of the construction this matrix, blue boxes denote the nonzero elements. First, due to the fact that the adjacency matrix $\mathbf{A}$ is symmetric, we number the nonzero entries of the matrix $\mathbf{A}$ above the diagonal in ascending order from top to bottom and left to right. Second, if $\mathbf{A}_{ij}$ is nonzero and located in the n-th position, then $\mathbf{C}_{ni} = \mathbf{A}_{ij}$ and $\mathbf{C}_{nj} = -\mathbf{A}_{ij}$. The rest of elements are set to zeros in this row.

To integrate content-based methods and social knowledge information for microblog sentiment analysis, a unified framework **GRALA** is proposed:

$$\begin{aligned} \text{argmin}_{\mathbf{w}} \quad \mathcal{F} =& \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + \lambda||\mathbf{C}\mathbf{X}\mathbf{w}||_1 \\ &+ \frac{\alpha}{2}\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}) + \beta||\mathbf{w}||_1, \end{aligned} \tag{5}$$

where $\mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y})$ is the loss function of content-based microblog sentiment methods, $||\mathbf{C}\mathbf{X}\mathbf{w}||_1$ and $\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w})$ are the regularization terms to explore social information prior knowledge from the sparsity and smoothness view respectively. $||\mathbf{C}\mathbf{X}\mathbf{w}||_1$ is the graphical lasso regularization term which automatically chooses helpful neighbour microblogs, while $\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w})$ tries to make the sentiment of connected microblogs as close as possible. α, β, and λ are the hyper-parameters to balance the weights of different regularization terms.

C. Optimization Algorithm

The optimization problem in (5) is convex. However, as $||\mathbf{C}\mathbf{X}\mathbf{w}||_1$ and $||\mathbf{w}||_1$is not smooth, it is not easy to solve it directly. The ADMM algorithm is utilized to solve it. Equation (5) can be reformulated as:

$$\text{argmin}_{\mathbf{w}} \quad \mathcal{F} = \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + ||\mathbf{H}\mathbf{w}||_1 + \tfrac{\alpha}{2}\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}), \tag{6}$$

where $\mathbf{H} \in R^{(1+|\mathcal{E}|)\times m}$, $\mathbf{H} = \begin{bmatrix} \beta \\ \lambda\mathbf{C}\mathbf{X} \end{bmatrix}$. Let $\mathbf{H}\mathbf{w} = \mathbf{v}$, (6) can be rewritten as:

$$\begin{aligned} \text{argmin}_{w} \quad & \mathcal{L} = \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + ||\mathbf{v}||_1 + \frac{\alpha}{2}\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}) \\ \text{s.t.} \quad & \mathbf{v} = \mathbf{H}\mathbf{w}. \end{aligned} \tag{7}$$

An augmented Lagrangian problem is used to solve (7):

$$\begin{aligned} \text{argmin}_{w} \quad \mathcal{F} \ = \ & \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + ||\mathbf{v}||_1 + \frac{\alpha}{2}\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}) \\ & + \mu^T(\mathbf{H}\mathbf{w} - \mathbf{v}) + \frac{1}{\rho}||\mathbf{H}\mathbf{w} - \mathbf{v}||_2^2. \end{aligned} \tag{8}$$

Let $\mathbf{u} = \frac{\mu}{\rho}$ as the scaled dual variable, ρ is a positive penalty coefficient, (8) can be reformulated as:

$$\begin{aligned} \text{argmin}_{w} \quad \mathcal{F} \ = \ & \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + ||\mathbf{v}||_1 + \frac{\alpha}{2}\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}) \\ & + \frac{\rho}{2}||\mathbf{H}\mathbf{w} - \mathbf{v} + \mathbf{u}||_2^2 - \frac{\rho}{2}||\mathbf{u}||_2^2. \end{aligned} \tag{9}$$

Essentially, the problem becomes the iterative update of $\mathbf{w}$, $\mathbf{v}$ and $\mathbf{u}$:

$$\begin{aligned} \mathbf{w}_{t+1} \ = \ & \text{argmin}_{\mathbf{w}} \quad \mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y}) + \frac{\rho}{2}||\mathbf{H}\mathbf{w} - \mathbf{v}_t + \mathbf{u}_t||_2^2 \\ & + \frac{\alpha}{2}\text{Tr}((\mathbf{X}\mathbf{w})^T\mathbf{L}\mathbf{X}\mathbf{w}), \end{aligned} \tag{10}$$

$$\begin{aligned} \mathbf{v}_{t+1} \ &= \ \text{argmin}_{\mathbf{v}} \quad ||\mathbf{v}||_1 + \frac{\rho}{2}||\mathbf{H}\mathbf{w}_{t+1} - \mathbf{v} + \mathbf{u}_t||_2^2 \\ &= \ S_{1/\rho}(\mathbf{H}\mathbf{w}_{t+1} + \mathbf{u}_t), \end{aligned} \tag{11}$$

where S is soft-thresholding operator.

$$\mathbf{u}_{t+1} \ = \ \mathbf{u}_t + \rho * (\mathbf{H}\mathbf{w}_{t+1} - \mathbf{v}_{t+1}). \tag{12}$$

To minimize (10), we set the gradient with $\mathbf{w}$ to zero, as $\mathcal{L}(\mathcal{X}, \mathbf{w}, \mathcal{Y})$ in this paper is the least squares, then we can update $\mathbf{w}$ by:

$$\mathbf{w}_{t+1} = (\mathbf{X}^T\mathbf{X} + \rho\mathbf{H}^T\mathbf{H} + \alpha\mathbf{X}^T\mathbf{L}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y} + \rho\mathbf{H}^T(\mathbf{v}_t - \mathbf{u}_t)). \tag{13}$$

The detail of our proposed optimization method is shown in Algorithm 1.

V. EXPERIMENTS

Two real public datasets (OMD [36] and HCR [37]) are used to conduct experiments in this paper. Friend information is from the complete follower graph built by [38]. We randomly select 80% microblogs as the training set, the left 20% microblogs are used as the testing set. Results of different methods are evaluated by classification accuracy and the weighted average F_1-score. The hyper-parameters are chosen by cross validation.

Algorithm 1: The proposed optimization algorithm for microblog sentiment analysis

Input: $\mathbf{X}$, $\mathbf{y}$, $\mathbf{A}$, α, β, λ, ρ.
Output: $\mathbf{w}$

Initialize Initialize $\mathbf{w}_0 = \mathbf{w}_1 = \mathbf{0}$, $\mathbf{u}_0 = 0$, $\mathbf{v}_0 = 0$, $t = 0$;
while *not convergent* **do**
  $t = t + 1$;
  compute $\mathbf{H}$;
  compute $\mathbf{L}$;
  update $\mathbf{w}_t$ using (13);
  $\mathbf{v}_t = S_{1/\rho}(\mathbf{H}\mathbf{w}_t + \mathbf{u}_{t-1})$;
  $\mathbf{u}_t = \mathbf{u}_{t-1} + \rho * (\mathbf{H}\mathbf{w}_t - \mathbf{v}_t)$;
end
$\mathbf{w} = \mathbf{w}_t$.

TABLE I
THE EFFECTIVENESS OF DIFFERENT PARTS

Method	F1(%)	Accu(%)
LS baseline	73.468	75.000
LS+L1	76.401	76.695
LS+L1+graph laplacian	77.413	78.390
LS+L1+graphical lasso	78.833	79.237
GRALA(LS+L1+graphical lasso+ graph laplacian)	**79.067**	**79.661**

A. The effectiveness of different regularization

The effectiveness of different parts in the loss function is evaluated. The OMD dataset is used as an example. The results are shown in Table I. "LS+L1" represents the content-based method with the L_1 norm regularization. "LS+L1+graph laplacian" represents the content-based method with graph Laplacian and the L_1 norm regularization. "LS+L1+graphical lasso" represents the content-based method with the graphical lasso and the L_1 norm regularization. "GRALA" represents our proposed method. From this figure, we can see that compared with the baseline content-based method, adding L_1 norm regularization can improve the sentiment analysis results. This indicates that not all features are useful for sentiment classification. Integrating social knowledge information hasa positive influence on microblog sentiment analysis as both gralap and gralas outperform Lasso. The method GRALA achievesthe best performance because it combines theL_1 normregularization, graph Laplacian regularization, and graphicallasso regularization. GRALA can automatically choosehelpfulmicroblogs among all neighbors and make the sentiment ofconnected microblogs as close as possible at the same time.

B. Performance Comparisions

Our proposed method GRALA is compared with other state-of-the-art sentiment analysis models. These methods include the least squares (LS), logistic regression (LR), support vector machine (SVM), Naive Bayes classifier (NB), Convolutional Neural Network (CNN) [39], Long short-term memory neural network (LSTM) [40], SANT [15], SASS [18], GCN [20]. All microblogs in the training set are used to train all the classifiers. Results are shown in Table II. The LS, LR, SVM, NB, CNN, and LSTM are content-based methods. SANT, SASS and GCN are methods integrated with social knowledge information. From the experimental results, we can see that SANT, SASS, GCN and GRALA achieve better performance than other methods, which suggests that leveraging social network knowledge can improve microblog sentiment analysis results. Among all methods, our method performs best on both datasets, which indicates the effectiveness of graph structured sparse regularization.

TABLE II
THE SENTIMENT ANALYSIS RESULTS OF DIFFERENT METHODS

Method	OMD		HCR	
	F1(%)	Accu(%)	F1(%)	Accu(%)
LS	73.468	75.000	74.097	75.524
LR	77.412	77.966	74.505	76.224
NB	73.005	75.000	72.722	74.476
SVM	75.756	75.847	76.725	76.923
CNN	73.264	73.729	70.827	73.077
LSTM	75.852	76.271	75.612	75.874
SANT	77.413	78.390	77.498	78.671
SASS	78.193	79.237	76.630	79.720
GCN	76.695	76.543	78.066	79.021
GRALA	**79.067**	**79.661**	**78.494**	**79.720**

VI. CONCLUSION

In this paper, we formulate microblog sentiment analysis by integrating social knowledge information and content information. More importantly, we investigate how sparse modeling of social contexts can influence microblog sentiment analysis. Particularly, we extract latent sentiment clues between microblogs based on the sentiment consistency and emotional contagion theory and assume that the sentiment of connected microblogs is similar. Then we propose a framework to balance the sparsity and smoothness of social contexts. Graph Laplacian regularization is utilized to represent the smoothness of the built graph, and graph structured sparse regularization is presented to model the sparsity of social information. As a result, microblogs with clear sentiment words are selected among neighbors automatically. The results of experiments conducted on two real-world public datasets show that our method performs better than other baseline methods and indicate the effectiveness of graph structured sparse regularization on microblog sentiment analysis. This work can suggest some interesting directions for future work. For example, it would be promising to integrate sparse modeling into graph neural networks to overcome the over-smoothing problem.

ACKNOWLEDGMENT

This paper is supported by 1) Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ23F020039, 2) National Science and Technology Major Project of China under Grant No. 2021ZD0114303, 3) National Natural Science Foundation of China under Grant No. 62176087 and No. T2241018.

REFERENCES

[1] David Valle-Cruz, Vanessa Fernandez-Cortez, Asdrúbal López-Chau, and Rodrigo Sandoval-Almazán, "Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods," *Cognitive computation*, vol. 14, no. 1, pp. 372–387, 2022.

[2] Renata Lopes Rosa, Gisele Maria Schwartz, Wilson Vicente Ruggiero, and Demstenes Zegarra Rodrłguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2124–2135, 2019.

[3] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu, "Opinionflow: Visual analysis of opinion diffusion on social media," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1763–1772, 2014.

[4] Sergio Consoli, Luca Barbaglia, and Sebastiano Manzan, "Fine-grained, aspect-based sentiment analysis on economic and financial lexicon," *Knowledge-Based Systems*, vol. 247, pp. 108781, 2022.

[5] Abdalsamad Keramatfar, Hossein Amirkhani, and Amir Jalaly Bidgoly, "Modeling tweet dependencies with graph convolutional networks for sentiment analysis," *Cognitive Computation*, Feb 2022.

[6] Xiaofei Zhu, Jie Wu, Ling Zhu, Jiafeng Guo, Ran Yu, Katarina Boland, and Stefan Dietze, "Exploring user historical semantic and sentiment preference for microblog sentiment classification," *Neurocomputing*, vol. 464, pp. 141–150, 2021.

[7] Jonnathan Carvalho and Alexandre Plastino, "On the evaluation and combination of state-of-the-art features in twitter sentiment analysis," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1887–1936, 2021.

[8] R. P. Abelson, "Whatever became of consistency theory?," *Personality & Social Psychology Bulletin*, vol. 9, no. 1, pp. 37–64, 1983.

[9] Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson, "Emotional contagion," *Current Directions in Psychological Science*, vol. 2, no. 3, pp. 96–100, 1993.

[10] Paul F Lazarsfeld, Robert K Merton, et al., "Friendship as a social process: A substantive and methodological analysis," *Freedom and control in modern society*, vol. 18, no. 1, pp. 18–66, 1954.

[11] Miller McPherson, Lynn Smith-Lovin, and James M Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[12] Lei Wang, Jianwei Niu, and Shui Yu, "Sentidiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 2026–2039, 2020.

[13] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[14] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li, "User-level sentiment analysis incorporating social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 1397–1405, ACM.

[15] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu, "Exploiting social relations for sentiment analysis in microblogging," in *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, Rome, Italy, 2013, pp. 537 – 546.

[16] Tao-Jian Lu, "Semi-supervised microblog sentiment analysis using social relation and text similarity," in *2015 International conference on big data and smart computing (BigComp)*. IEEE, 2015, pp. 194–201.

[17] Fangzhao Wu, Yongfeng Huang, and Yangqiu Song, "Structured microblog sentiment classification via social context regularization," *Neurocomputing*, vol. 175, no. PartA, pp. 599 – 609, 2016.

[18] Xiaomei Zou, Jing Yang, and Jianpei Zhang, "Microblog sentiment analysis using social and topic context," *PloS one*, vol. 13, no. 2, pp. e0191163, 2018.

[19] "Collaborative community-specific microblog sentiment analysis via multi-task learning," *Expert Systems with Applications*, vol. 169, pp. 114322, 2021.

[20] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017, pp. 1–14.

[21] Jie Chen, Nan Song, Yansen Su, Shu Zhao, and Yanping Zhang, "Learning user sentiment orientation in social networks for sentiment analysis," *Information Sciences*, vol. 616, pp. 526–538, 2022.

[22] Fangzhao Wu, Yongfeng Huang, Yangqiu Song, and Shixia Liu, "Towards building a high-quality microblog-specific chinese sentiment lexicon," *Decision Support Systems*, vol. 87, pp. 39–49, 2016.

[23] Chuanjun Zhao, Suge Wang, and Deyu Li, "Exploiting social and local contexts propagation for inducing chinese microblog-specific sentiment lexicons," *Computer Speech and Language*, vol. 55, pp. 57–81, 2019.

[24] "Tailored text augmentation for sentiment analysis," *Expert Systems with Applications*, vol. 205, pp. 117605, 2022.

[25] Tong Zhang, Xinrong Gong, and C. L. Philip Chen, "Bmt-net: Broad multitask transformer network for sentiment analysis," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 6232–6243, 2021.

[26] Dirk Hovy and Diyi Yang, "The importance of modeling social factors of language: Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 588–602.

[27] Dong Nguyen, Laura Rosseel, and Jack Grieve, "On learning and representing social meaning in nlp: a sociolinguistic perspective," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 603–612.

[28] Changhao Song, Bo Wang, Qinxue Jiang, Yehua Zhang, Ruifang He, and Yuexian Hou, "Social recommendation with implicit social influence," in *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*, 2021, pp. 1788–1792.

[29] Hamid Khalifi, Sarah Dahir, Abderrahim El Qadi, and Youssef Ghanou, "Enhancing information retrieval performance by using social analysis," *Social Network Analysis and Mining*, vol. 10, pp. 1–7, 2020.

[30] Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser, "Tackling fake news detection by continually improving social context representations using graph neural networks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1363–1380.

[31] Stanley F Chen and Ronald Rosenfeld, "A survey of smoothing techniques for me models," *IEEE transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 37–50, 2000.

[32] Konstantinos Skianis, François Rousseau, and Michalis Vazirgiannis, "Regularizing text categorization with clusters of words," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Nov. 2016, pp. 1827–1837, Association for Computational Linguistics.

[33] Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," *Expert Systems with Applications*, vol. 110, pp. 298–310, 2018.

[34] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[35] Stefan Stieglitz and Linh Dang-Xuan, "Emotions and information diffusion in social mediałsentiment of microblogs and sharing behavior," *Journal of management information systems*, vol. 29, no. 4, pp. 217–248, 2013.

[36] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *Sigmm Workshop on Social Media*, 2009, pp. 3–10.

[37] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proceedings of the First Workshop on Unsupervised Learning in NLP*, USA, 2011, EMNLP '11, p. 53C63, Association for Computational Linguistics.

[38] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010, WWW '10, pp. 591–600, ACM.

[39] Yoon Kim, "Convolutional neural networks for sentence classification," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1746 – 1751.

[40] Sepp Hochreiter and Jrgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Adopting Neural Translation Model in Data Generation for Inverse Text Normalization

Yufei Jiang, Thi-Nga Ho, and Eng-Siong Chng
Nanyang Technological University, Singapore
E-mail: jian0133@e.ntu.edu.sg

***Abstract*—This research proposes a novel data generation pipeline for the Inverse Text Normalization task with a focus on improving the data alignment technique. The pipeline adopts Automatic Speech Recognition (ASR) and Neural Word Alignment techniques to facilitate the needs. The pipeline is a three-stage process from data crawling and spoken transcript generation to data alignment. Firstly, audio recordings and their corresponding written-form transcripts were collected in a semi-automatic manner. Subsequently, the audio recordings are sent to an ASR system to obtain spoken transcripts. Lastly, a cutting-edge word alignment technique is utilized to pair words between written and spoken transcripts to produce a word-level alignment ITN dataset. The proposed data alignment technique achieves at least 3.3% improvement in decimal word transformation compared to the representative baselines.**

I. Introduction

The objective of the research is to develop high-quality datasets for ITN research. A typical ITN dataset comprises two distinct sequences of sentences or words, i.e. the spoken and the corresponding written sequences. The spoken form is a verbatim representation of the audio in text, while the written form preserves the audio information but is organized in written format for better readability. For instance, a spoken transcript can be "nineteen eighty five", while its written form will be "1985". A few examples of spoken-written transcripts are illustrated in Table I.

TABLE I
Examples of spoken from written form pairs

Spoken Form	Written Form
Doctor is in room two zero three	Doctor is in room 203
It's twelve thrity now	It's 12:30 now
Please give me five apple	Please give me 5 apple
The car costs me three thousand dollors	The car costs me 3000$
He paid one hundred percent effort	He paid 100% effort
In twenty twenty one	In 2021

Despite advancements in recent ITN research, the scarcity of ITN training corpora remains an issue. The Google Text Normalization Dataset is the only sizeable public dataset for English, Russian, and Polish[1]. The dataset is structured at word-level with classification labels assigned to each word pair. Figure 1 shows a screenshot of labeled text data from Google Text Normalization Corpus with three columns. The first column represents the ITN class of a word, e.g. *PLAIN* means there is no corresponding ITN class to the word or sequence in the middle column. The second and the last column is the written and spoken form respectively. If there is no difference between written form and spoken form, the last column will be marked as *<self>*. The Google Text Normalization Dataset is generated by a Weighted Finite-State Transducer (WFST), however, it has been treated as a gold standard corpus Inverse Text Normalization task due to the lack of human-labeled corpus for the task.

```
PLAIN   Brillantaisia   <self>
PLAIN   is  <self>
PLAIN   a   <self>
PLAIN   genus   <self>
PLAIN   of  <self>
PLAIN   plant   <self>
PLAIN   in  <self>
PLAIN   family  <self>
PLAIN   Acanthaceae <self>
PUNCT   .   sil
<eos>   <eos>
DATE    2006    two thousand six
LETTERS IUCN    i u c n
PLAIN   Red <self>
PLAIN   List    <self>
PLAIN   of  <self>
```

Fig. 1. Data in Google text normalization dataset

There are multiple attempts in augmenting data for the ITN task in recent years, for example, [1] and [2]. [1] utilizes existing datasets and applies data augmentation strategies for more diversity. However, this method may introduce errors and is constrained by a set of limited augmentation rules. Another method is to use Text-to-Speech (TTS) and Text Normalization (TN) systems[2] to generate spoken form sentences from written forms. This approach is frequently employed in conjunction with existing datasets, such as the supplementary dataset created from publicly available sources as outlined in previous works [3], [4], [5], [6]. The drawback is the performance of TTS and TN systems greatly influences the quality of the generated dataset. Furthermore, the generated dataset is structured at the sentence level, which restricts its utility for advancing further research.

Our study designs and implements an ITN dataset generation pipeline including semi-automatic data source selection and crawling, data pre-processing, and a novel approach for data

[1]available at https://www.kaggle.com/datasets/richardwilliamsproat/text-normalization-for-english-russian-and-polish

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

alignment as a post-process to achieve the desired ITN dataset. Specifically, we propose to employ a neural network-based word alignment method for the data alignment module. The neural models offer advantages such as the ability to improve performance with data training and eliminate the need for manual rule designing. Our experimental results show that the neural network-based pipeline outperforms the reference rule-based method.

The rest of the paper is organized as follows. Section II explains the proposed pipeline. Section III describes the proposed data alignment approach, and Section IV narrates the experiment setup and training details for the proposed data alignment approach. Section V presents an analysis of the experimental results. Finally, Section VI concludes the work and points out future improvements.

II. Data Preparation Pipeline

Figure 2 depicts the architecture of the pipeline, which consists of three primary stages, i.e. data crawling, data processing, and data alignment. The data crawling stage involves collecting audio and their corresponding written form transcripts. Subsequently, in the data processing stage, an ASR system is utilized to transcribe the audio into spoken form transcripts. Ultimately, in the data alignment stage, the two transcripts are inputted into the word alignment model and aligned to create the written-spoken parallel corpus for ITN. An example of written-spoken forms in the output corpus is shown in Figure 3. The first column presents words in their written form, while the second column is the corresponding spoken form of the words. Words that have no difference between their written and spoken forms are represented as *<self>* in the second column.

In the next subsections, we describe the first two stages in more detail, while the last stage, i.e. Data Alignment with the novel word alignment approach will be presented in Section III.

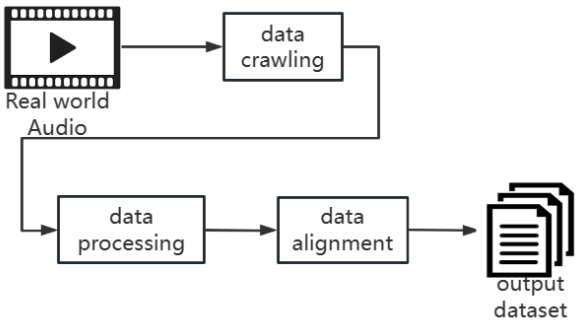


Fig. 2. Overall architecture of the pipeline

A. Data Crawling

A crawler is built to fetch audio and transcripts from web-based sources. The transcripts are equipped with time alignments to the audio and are already converted into their written forms. The transcripts may come in various formats, but predominantly in Web Video Text Tracks (VTT) format

```
today's <self>
date    <self>
is      <self>
the     <self>
19th    nineteen
of      <self>
march   <self>
2022    twenty twenty two
and     <self>
this    <self>
is      <self>
my      <self>
site    <self>
select  <self>
```

Fig. 3. Example of output dataset

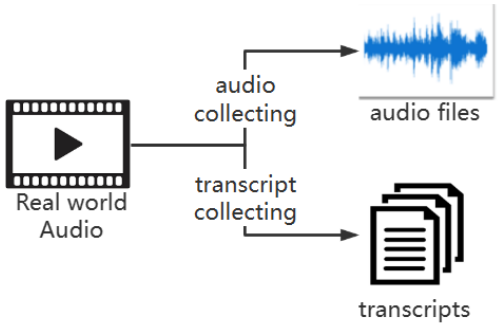


Fig. 4. Data crawling section

shown in Figure 5. These transcripts are then reformatted into sequences of sentences to simplify future operations.

B. Data Processing

The second stage of the pipeline consists of two primary tasks, re-segmentation and transcript generation, that is to properly segment and transcribe the corresponding audio form of each written transcript obtained in the first stage. Firstly, the written-form transcripts are re-segmented to achieve more appropriate divisions. Subsequently, an ASR system is employed to generate transcripts in the spoken form.

```
00:00:10.500 --> 00:00:12.530 align:start position:0%
article right the Chinese newspaper of
Singapore<00:00:10.860><c> mainstream</c><00:00:11.519><c> media</c><00:00:11.820><c> talk</c>
<00:00:12.420><c> about</c>

00:00:12.530 --> 00:00:12.540 align:start position:0%
Singapore mainstream media talk about


00:00:12.540 --> 00:00:14.930 align:start position:0%
Singapore mainstream media talk about
this<00:00:12.900><c> article</c><00:00:13.320><c> about</c><00:00:13.679><c> whether</c>
<00:00:14.040><c> cpf</c><00:00:14.639><c> is</c><00:00:14.759><c> going</c>

00:00:14.930 --> 00:00:14.940 align:start position:0%
this article about whether cpf is going


00:00:14.940 --> 00:00:17.269 align:start position:0%
this article about whether cpf is going
```

Fig. 5. Example of collected transcripts

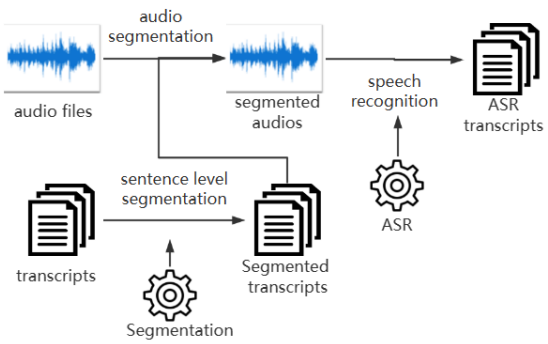


Fig. 6. Data processing section

```
2nKg1q5aVJM_10000-14480 0.26 4.26 money cost s one e two he plans to retire at thirty five
2nKg1q5aVJM_14480-17200 0.32 2.52 with one million dollars in his portfolio
2nKg1q5aVJM_17200-20480 0.26 3.0 e p two when this podcast was released
2nKg1q5aVJM_77280-81200 2.06 3.18 at thirty five years old
2nKg1q5aVJM_93520-96800 0.32 2.7600000000000002 you going to thai soon one point four million
2nKg1q5aVJM_118880-121600 0.38 2.22 not enough to retire at thirty five
2nKg1q5aVJM_121600-123840 0.08 2.04 are you going to survive on just one million dollars
2nKg1q5aVJM_123840-126320 0.14 2.22 for forty to fifty years of your life even
2nKg1q5aVJM_126320-129120 0.08 2.34 if you manage to win the eight million dollars lottery
2nKg1q5aVJM_135760-138160 0.14 2.1 like this guy it's not one million dollars
2nKg1q5aVJM_138160-141360 0.92 2.94 is one million dollars in portfolio
```

Fig. 7. Example of ASR transcripts

1) Transcript Re-segmentation: The written-form transcripts obtained from the data crawling stage usually lack suitable segmentation as they are primarily intended for on-screen display. Additionally, punctuations are absent from the transcripts. To address this, we attempt to re-segmentation the transcripts using a deep learning toolkit namely Deepsegment[2]. Deepsegment is a segmentation tool that uses neural networks to produce efficient, high-quality segmentation. Importantly, the output from this segmentation process includes starting and ending points of each segment in the audio form, which are beneficial for the second task, i.e. speech recognition.

2) Speech Recognition: To obtain spoken-form transcripts, we employed a fine-tuned WeNet ASR system[7], [8] to transcribe audios. Conventionally, if we send an audio to the system, if it is too long, the ASR system will automatically divide it into segments without considering if a spoken word has been completed. This can be a problem as it may inadvertently split words into two segments. Thus, instead of sending the entire crawled audio to the system, we use the output from the first task to segment the audio correctly. The segmented audio is then inputted into the system to produce the spoken form transcripts. An example of an ASR transcript produced by the Wenet ASR system is displayed in Figure 7.

III. Data Alignment

With the pairs of written and spoken form transcripts that were crawled and generated in the first and second stages respectively, it is now essential to accurately match words at the word level between the pairs to form an ITN corpus. Especially the part where there are differences between the written and spoken forms. This can be done by employing a word alignment approach. Word alignment, also known as bi-text word alignment, which is commonly used in machine translation, seeks to identify potential translations between two texts [9]. It computes the possibility of connections between words and assigns matches if the word pairs translate to each other. Statistical models like Fast Align [10] and GIZA++ [11] were commonly employed in early works and were able to offer satisfaction at certain levels. In recent years, neural-based word alignment models have outperformed their statistical counterparts in numerous bilingual situations. For example, Luong et al. [12] proposed to use force-decoding in a global attention model and achieved competitive results at 34.0 of Alignment Error Rate (AER) for English-German alignment compared with Berkeley Aligner[13]. Kobayashi et al. [14] introduced simple adjustments based on the purposes of attention utilization, Sabet et al. [15] developed Sim Align, Awesome Align that was expanded from Sim Align by Dou et al. [16], and recently, Cross-Align developed by Lai et al. [17].

A. Cross-Align

This work assumes that aligning a written form sequence to its corresponding spoken form is a translation task from one language to another. We adopt Cross-Align [17] for the data alignment stage. Cross-Align is an attention-based method that distinguishes itself from earlier alignment approaches by simultaneously considering source and target language sentences, with a focus on the relationship between two languages. Essentially, Cross-Align accounts for cross-lingual information and leads to performance improvement.

Figure 8 demonstrated the adopted Cross-Align model which consists of three stages. The initial stage independently trains the paired inputs and results in self-attentions that encompass monolingual information for both languages. In the second stage, the monolingual information is exchanged between the languages and leads to the acquisition of cross-attention that carries cross-lingual information. During the final stage, the output embeddings of the two languages are mapped and compared by employing an extraction algorithm. The extraction algorithm works as follows. The similarity between the source and target languages is calculated by taking the dot product of the corresponding cross-attention embedding sequence, yielding similarity matrices. A softmax function is then applied to convert these similarity matrices into probability matrices. Next, a threshold filtration method is used to assign a value of 1 to entries above the threshold and a value of 0 to those below it, generating the final decision metrics. Whenever there is a 1 in the matrix, the two corresponding words should be connected in alignment. The alignment is carried out at the byte pair encoding (BPE) level, as explored in previous studies [18], [19], [20]. Therefore, alignment between source and target words is generated based on these prior works. The output follows the *"Pharaoh format"*, where "i-j" signifies that the i-th word in the source is aligned with the j-th word in the target sentence [21].

[2]Available at https://github.com/notAI-tech/deepsegment/tree/v2

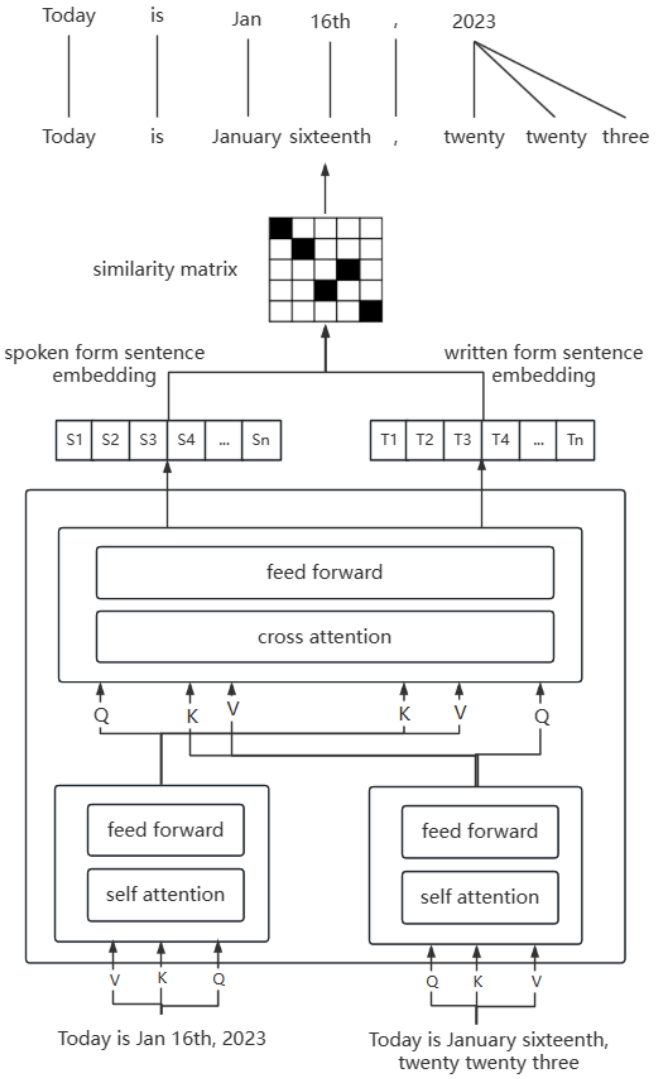


Fig. 8. The proposed Cross-Align model

B. Training Strategy

The training process is divided into two stages to enhance performance. The first stage employs Translation Language Modeling (TLM), a technique designed to learn both monolingual and cross-lingual information. The second stage involves self-supervised alignment, a strategy that fine-tunes the model to better achieve the intended alignment goal.

1) Translation Language Modeling: TLM is a prevalent method utilized in the extraction of monolingual information. Recent research, though, has identified that this approach can also positively impact the learning of cross-lingual information [22]. As a result, the Cross-Align model employs TLM to bridge the gap and enhance understanding between spoken and written forms[17].

The approach involves masking 15% of the input words for the training process, a technique that aligns with previous research[23]. To elaborate, during training, 15% of the words in both spoken and written sentences are selected randomly. Among these chosen words, 80% is substituted with a mask token, 15% is replaced by a random token, and the remaining 5% are left unchanged. The training goal is to predict the words obscured by the mask based on the context provided by the adjacent text. The objective is formally structured in the following way:

$$argmin\lambda_{TLM} = argmin(-logP(x|\hat{x}, \hat{y}, \theta_s, \theta_c) \\ -logP(x|\hat{x}, \hat{y}, \theta_s, \theta_c))$$

In the formula, $\hat{x}$ and $\hat{y}$ represent the masked version of x and y. θ_s and θ_c represent all the parameters in the self-attention module and cross-attention module [17].

2) Self-supervised Alignment: The objective of this stage is to train the model for sentence alignment, as the previous stage focused on predicting masked words rather than the main goal. Consequently, we shift the model's focus to alignment inference. In this training phase, the output from the prior stage serves as the golden dataset, and the model is trained for the word alignment task.

The preceding section made use of the masking technique to acquire both monolingual and cross-lingual information. In contrast, this section reorients the training goal back to alignment production. The model learnt from the preceding section is employed as a pre-trained model. Given that alignment for training isn't always accessible, the training process is designed to be self-supervised. Following the work by Garg et al.[24], the objective is structured as follows:

$$\lambda_{SS} = -\frac{1}{S}\Sigma_{s=1}^{S}\Sigma_{t=1}^{T}G_{st}log(P_{st}^{S}) - \frac{1}{T}\Sigma_{t=1}^{T}\Sigma_{s=1}^{S}G_{st}log(P_{st}^{T})$$

In the equation, S and T are spoken and written form inputs, G indicates alignment label metrics and P stands for connection possibility metrics.

Studies have discovered that the output embeddings from intermediate layers outperform those from the final layer in alignment tasks [16]. As a result, the output from the middle layer of the cross-attention module is leveraged as the word embeddings to enhance the overall performance of the model. To maintain the parameters learnt during TLM, the self-attention and cross-attention layers are frozen in alignment training. Knowledge for alignment is stored and updated within the alignment layers.

IV. Experiment

A. Datasets

We evaluate the proposed alignment approach on two datasets. The first dataset is derived from the Google Text Normalization Dataset. The parallel words within this dataset are organized into parallel sentences to align with the model's input requirements. This dataset is employed for both training and testing the Cross-Align model. Alignments for the testing corpus are generated to assess the model's performance. The format of the data is depicted in Table II.

The second dataset consisting of real-world audio is built by our team to test the performance of the function of the

TABLE II
DATASET FORMAT

Data Type	Format
google text normalization dataset	PLAIN Archived <self> PLAIN from <self> PLAIN the <self> PLAIN original <self> PLAIN on <self> DATE 3 November 2007 the third of November PUNCT . sil <eos><eos>
model input	Archived from the original on 3 November 2007 ||| Archived from the original on the third of two thousand seven
corresponding alignment	1-1 2-2 3-3 4-4 5-5 8-12

whole pipeline. The audios are fetched from different public audio and video sharing websites. Considering the difficulty of generating word-level alignment for the parallel sentence, the dataset is only used for testing pipeline performance.

B. Implementation Details

Our implementation primarily follows the Cross-Align paper's guidelines [17]. The model's code is based on Awesome Align developed by Zi-Yi Dou and Graham Neubig [16]. Parameters for the cross-attention module are randomly initialized, while the self-attention module and alignment layer parameters are initialized using the pre-trained mBERT-base model [23]. We employ AdamW as the optimizer, with learning rates $lr_{self} = 5e^{-4}$ and $lr_{cross} = 1e^{-5}$. To expedite training and adapt the model for local environments, we set the batch size to 32. Training occurs on both RTX 3070ti (locally) and NVIDIA Tesla V100 (remotely) GPUs. The self-attention layers are set to 10, and the cross-attention module is set to 2. We train for 3 epochs in the first stage and 2 epochs in the second stage, with thresholds τ set to 0.001 and 0.15, respectively.

The performance of our model is benchmarked against the following approaches:

1) Fast Align [10]. A renowned statistical alignment tool developed by IBM, based on the IBM Model 2 but re-parameterized.
2) GIZA++ [25]. Another notable statistical aligner from IBM. GIZA++ is an extension of GIZA, incorporating various models such as IBM Models 1, 3, 4, and the HMM model.
3) MASK-ALIGN [26]. A self-supervised NMT-based aligner. Unlike other NMT models, MASK-ALIGN considers the target side's context bidirectionally, resulting in improved alignment quality.
4) Awesome Align [16]. A fine-tuned, multilingual, pre-trained language model that enhances its performance by training on parallel data.
5) Rule-based baseline. A WFST-based rule system is proposed as a baseline, with performance improvements achieved through the use of word-level timestamps.

C. Evaluation Metrics

We employed Alignment Error Rate (AER) and Spoken Form Accuracy (SFA) as evaluation metrics to evaluate the proposed alignment approach.

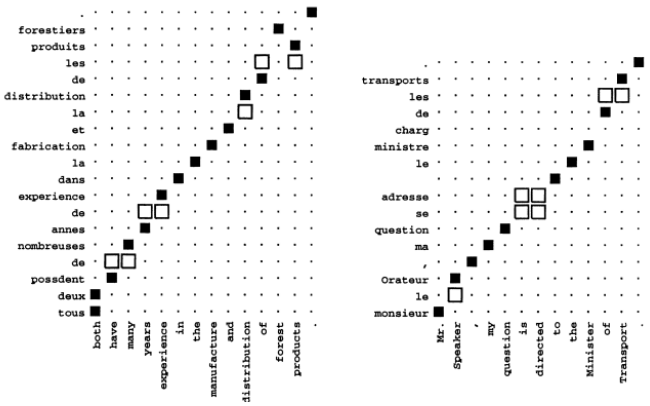


Fig. 9. Example sentence alignment explaining the S(filled squares) and P (unfilled squares)

1) Alignment Error Rate: AER [11] is a widely used evaluating method for word alignment. AER requires standard alignments to compare with and divide the standard alignment into S (sure, robust) alignment and P (possible, ambiguous) alignment. Examples of S and P alignment are illustrated in Figure 9.

Taking in the model output alignment A, the precision and recall are defined as follows:

$$recall = \frac{|A \cap S|}{|S|}$$

$$precision = \frac{|A \cap P|}{|A|}$$

AER is then defined in reference to the F-measure:

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

AER has a maximum value of 1.0, indicating that when the model's output alignment matches the standard alignment perfectly, it reaches this maximum value. As the disparity between the model's output and the standard alignment increases, the AER value decreases.

2) Spoken Form Accuracy: SFA is designed as an alternative assessment method. Upon the model's generation of alignments for the test set, word pairs involving transformations between spoken and written forms are extracted and sampled. These sampled words are then compared with standard pairs to determine the overall accuracy, denoted as SFA.

$$T_{align} = gen(A, T_{para})$$

$$SFA(T_{align}, P_{gold}) = \frac{extract(T_{align})}{P_{gold}}$$

SFA is does not address sentence structure. instead, it exclusively concentrates on form transformation pairs. A transformation pair is considered correct solely if the transformation is coherent. Incorrectly transformed pairs, including plain word conversions, will contribute to a reduction in SFA value.

V. Result and Analysis

A. Result on Google Text Normalization Dataset

TABLE III
AER result for the different alignment methods on the test set. Better performance is indicated for lower AER.

Method	AER
Fast Align	18.0
GIZA++	12.0
Mask Align	12.8
Awesome Align	12.7
ours	11.8

The results showed on Table III indicates that on spoken-written form text alignment task, our Cross-Align based model performs the best. The Google Text Normalization Dataset is a well-structured and high-quality dataset that possesses significant characteristics. The majority of word pairs in the dataset are identical, as it is primarily composed of plain words. The only distinct pairs are those that involve transformations between spoken and written forms. This straightforward structure carries potent features that statistical models can easily latch onto. For example, the widely used GIZA++ can achieve an AER of 12.0, demonstrating competitive performance when compared to neural methods. While neural methods can also identify these features, they may be more susceptible to noise interference and potentially overlook the primary characteristics. This problem becomes more pronounced when the neural network model is not fully trained. More information can be found in Section V-B.

The AER evaluation method might not precisely represent the performance of models in the experiment. The standard alignment for the test set is derived from the original word-level pairs in the Google dataset. This alignment merely indicates the starting point of the spoken and written forms, and ignores potential associations within the spoken and written form phrases. As a consequence, when the model creates an alignment that partitions these phrases, the AER escalates. Table IV illustrates examples of such intricate connections identified by the model. The model assimilates more detailed features and generates a more accurate alignment. These could be regarded as correctable errors and lead to the possibility that the model's performance might surpass what the AER score suggests.

TABLE IV
Examples of extra inner connections grasped by the model.

gold alignment suggest	ours output
model model ((0-4-0Tnr o sil four sil o T nr . .	model model ((0-4-0 o sil four sil o T T nr nr . .
from from the the original original on on 3 of the third of November 2007 . .	from from the the original original on on 3 the third November two thousand seven November of November 2007 two thousand seven . .

B. Result on The Real Audio Dataset

The Google dataset is well-organized, allowing inner features to be more prominent. However, this may render it unsuitable for assessment purposes. To test the model with noisier data and realistic situations, the real-world audio dataset is utilized for additional evaluation.

TABLE V
SFA results for the different alignment methods on the real audio dataset. Better performance is indicated for higher numeric word accuracy.

Method	SFA
Fast Align	77.0
GIZA++	83.6
Mask Align	85.2
Awesome Align	85.2
ours (rule based model)	80.3
ours (neural model)	88.5

SFA is selected as the evaluation metric to eliminate the influence of words that are supposed to remain the same. Thus, the transformation between spoken and written forms is the only consideration. The real-world audio test set holds significantly more noise compared to the Google dataset, as illustrated by instances of such inconsistency in Table VI. This issue arises due to the frequently observed discrepancies between parallel transcripts. Although the WeNet ASR is fine-tuned, its performance is not assured under complex scenarios such as the source audios are noisy, and the speakers may have unique accents. The volume of the speaker also significantly impacts the ASR performance. In practice, only the transformation between spoken and written forms is documented in the output dataset, making SFA a suitable evaluation.

Table V presents the SFA results and demonstrates that the

TABLE VI
EXAMPLES OF INCONSISTENCY BETWEEN ORIGINAL TRANSCRIPTS AND WENET TRANSCRIPTS

original transcript	WeNet ASR system transcript
hey folks welcome back to honey money sg now	hey folks welcome back to honey money as she
i guess that's it	i guess
that is your cspx which tracks the s p 500 and then for qqq	that is your c s p x which tracks the s and p five hundred and then for q
so that's for my software portfolio	right so that's for my s portfo

proposed Cross-Align based method exceeds all benchmark baselines in performance. The proposed approach surpasses Awesome Align and proves that the enhancements and alterations made to Cross-Align, built upon Awesome Align, do contribute to the boost in performance. As the test set becomes noisier, GIZA++'s performance fails to remain competitive with neural models. This validates the hypothesis posited in the previous section that a simpler dataset, with its strong features, is amenable to statistical models. This further bolsters our confidence in deploying neural network-based models in the data generation pipeline.

TABLE VII
EXAMPLE OF ERROR IN WORD ALIGNMENT RESULT

Our approach output	original courpus
portfolio portfolio	portfolio portfolio
return return	return return
is is	is is
	128 hundred and twenty eight
	(5 other words)
we we	we we

The proposed method is not without limitations. Text missing can be observed in some instances, as illustrated in Table VII. This issue arises from the incomplete separation of the text. Ideally, for each word in the source text, there should be at least one corresponding link from the target text. However, there are scenarios where the output alignment is not fully segmented into words, indicating that the alignment label for the spoken and written form is imperfect. This could be due to the model being under-trained and affect roughly 4% of the text. The model's inability to generate fine-grained alignment results in the output dataset simply disregarding the unpaired words. While data loss happens during the process, it does not significantly affect the generated dataset as it only leads to minor corpus loss.

VI. CONCLUSIONS

The research concentrates on ITN data generation and aims to create a clean and well-formatted dataset to support ITN research. This area is underdeveloped, as previous studies do not pay much on data quality. The scarcity of publicly available well-constructed datasets for ITN model training and testing is a critical problem. This paper proposed a new approach using ASR systems and word alignment techniques to generate high-quality ITN datasets.

The proposed approach establishes a pipeline to convert audio and transcripts into a word-level written-spoken parallel corpus. The corpus retains a similar format to the Google Text Normalization Dataset. The pipeline involves stages of data crawling, data processing, and data alignment. It employs the WeNet ASR system for audio transcription and the Cross-Align[17] technique for word-level text pairing to enhance the overall performance of the pipeline.

Assessments are carried out using both AER and SFA. On the Google dataset, the proposed method yields an AER of 11.8 and surpasses other benchmarks. On a more complex real-world audio dataset, the method attains an SFA of 88.5, and also outperforms the referenced baselines by at least 3.3%. Despite these results, there is still room for improvement in the pipeline. ITN can be considered data-sensitive and requires high data quality. Consequently, the pipeline needs enhancements to ensure superior data quality. Therefore, further refinements remain crucial.

ACKNOWLEDGMENT

This research is supported by ST Engineering Mission Software & Services Pte. Ltd under a collaboration programme (Research Collaboration No: REQ0149132). We also would like to acknowledge the High Performance Computing Centre of Nanyang Technological University Singapore, for providing the computing resources, facilities, and services that have contributed significantly to this work.

REFERENCES

[1] Laxmi Pandey, Debjyoti Paul, Pooja Chitkara, Yutong Pang, Xuedong Zhang, Kjell Schubert, Mark Chou, Shu Liu, and Yatharth Saraf. Improving data driven inverse text normalization using data augmentation. *arXiv preprint arXiv:2207.09674*, 2022.

[2] Peter Ebden and Richard Sproat. The kestrel tts text normalization system. *Natural Language Engineering*, 21(3):333–353, 2015.

[3] Sunil Kumar, Debanjan Sadhya, Santosh Singh Rathore, et al. Efficient text normalization via hybrid bi-directional lstm. In *2021 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6. IEEE, 2021.

[4] Tuan Manh Lai, Yang Zhang, Evelina Bakhturina, Boris Ginsburg, and Heng Ji. A unified transformer-based framework for duplex text normalization. *arXiv preprint arXiv:2108.09889*, 2021.

[5] Yashesh Gaur, Nick Kibre, Jian Xue, Kangyuan Shu, Yuhui Wang, Issac Alphanso, Jinyu Li, and Yifan Gong. Streaming, fast and accurate on-device inverse text normalization for automatic speech recognition. *arXiv preprint arXiv:2211.03721*, 2022.

[6] Sharman Tan, Piyush Behre, Nick Kibre, Issac Alphonso, and Shuangyu Chang. Four-in-one: a joint approach to inverse text normalization, punctuation, capitalization, and disfluency for automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 677–684. IEEE, 2023.

[7] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547*, 2021.

[8] Binbin Zhang, Di Wu, Zhendong Peng, Xingchen Song, Zhuoyuan Yao, Hang Lv, Lei Xie, Chao Yang, Fuping Pan, and Jianwei Niu. Wenet 2.0: More productive end-to-end speech recognition toolkit. *arXiv preprint arXiv:2203.15455*, 2022.

[9] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. The mathematics of statistical machine translation: Parameter estimation. 1993.

[10] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, 2013.

[11] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

[12] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[13] Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. 2006.

[14] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*, 2020.

[15] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*, 2020.

[16] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*, 2021.

[17] Siyu Lai, Zhen Yang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Cross-align: Modeling deep cross-lingual interactions for word alignment. *arXiv preprint arXiv:2210.04141*, 2022.

[18] Anoop Kunchukuttan and Pushpak Bhattacharyya. Learning variable length units for smt between related languages via byte pair encoding. *arXiv preprint arXiv:1610.06510*, 2016.

[19] Jenn Leana Fernandez and Kristine Mae M Adlaon. Exploring word alignment towards an efficient sentence aligner for filipino and cebuano languages. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 99–106, 2022.

[20] Jingyi Zhang and Josef van Genabith. A bidirectional transformer based alignment model for unsupervised word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 283–292, 2021.

[21] Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*, pages 115–124, 2004.

[22] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[24] Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China, November 2019. Association for Computational Linguistics.

[25] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing*, pages 49–57, 2008.

[26] Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. Accurate word alignment induction from neural machine translation. *arXiv preprint arXiv:2004.14837*, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Mismatched Semi-supervised Learning with Feature Similarity Consistency

Zechen Liang*, Qiaosong Fan†, and Yuan-Gen Wang‡
* School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China
E-mail: liangzechen@e.gzhu.edu.cn, Tel: +86-20-39366670
† Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 511442, China
E-mail: 202164010202@mail.scut.edu.cn, Tel: 018127960317
‡ School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China
E-mail: wangyg@gzhu.edu.cn, Tel: +86-20-39366670
Z. Liang and Q. Fan contribute equally to this work, and Y.-G. Wang serves as the corresponding author of this work.

***Abstract*—Effective learning from a small amount of labeled data and a large amount of unlabeled data remains a prominent issue in machine learning. Current semi-supervised learning methods often neglect mismatched problems within class distribution, and do not fully use feature-level encoding for unlabeled data. In this article, we introduce a novel semi-supervised learning method called FSimC, which effectively uses information from unlabeled data at both the semantic- and feature-levels and is successfully applied to mismatched distribution scenarios. At the semantic-level, FSimC consists of two classifiers predicting pseudo and complementary labels for unlabeled data. The mutual alignment of the predictive distributions of the two classifiers results in more accurate pseudo and complementary labels. Then, the two classifiers undergo separate consistency regularization. At the feature-level, we create a cache to store the feature encodings of labeled data. Next, we compute the feature similarity vectors between feature encodings of the weakly and strongly augmented unlabeled data with feature encodings in the cache. Finally, we perform feature similarity consistency on the feature similarity vectors. Extensive experiments are conducted on various benchmark datasets, and the results demonstrate that FSimC achieves outstanding performance in matched and mismatched distribution scenarios. The code is publicly available at *https://github.com/GZHU-DVL/FSimC*.**

I. Introduction

Deep neural networks have achieved impressive performance in various computer vision tasks by leveraging large-scale labeled datasets. However, data annotation is an arduous task that requires substantial manpower, material, and financial investments. Semi-supervised learning (SSL) is a powerful approach that can leverage a small amount of labeled data and a large amount of unlabeled data to train deep neural networks, showing promising research direction.

Recently, some SSL algorithms [1], [2], [3], [4], [5], [6], [7] have greatly improved the performance of the trained models by introducing pseudo labels and distribution alignment. [1] uses consistency regularization, forcing the model to produce the same result for different augmentations of the same unlabeled data. FixMatch [3] introduces a high fixed threshold to filter out unlabeled data with low confidence. To minimize the entropy and push the model predictions away from the decision boundary [8], FixMatch converts the pseudo labels into one-hot vectors. However, during the training, the model showed significant differences in prediction performance for different classes. To address this issue, the dynamic threshold is introduced in [4]. However, these methods focus only on semantic-level information, while [16] successfully uses the information at both semantic-level and feature-level.

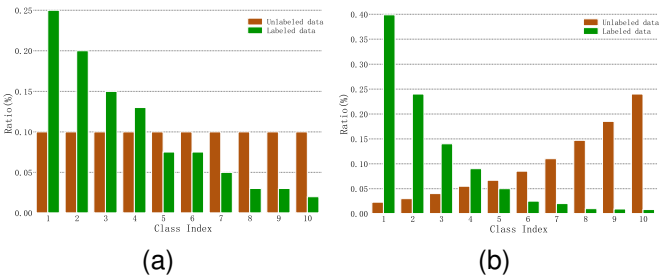


Fig. 1. Illustration of the CIFAR-10 dataset. (a) The distribution of labeled and unlabeled data when the number of labeled data is 40 and $N_{x0} = 40$. (b) The distribution of labeled and unlabeled data when the number of labeled data is 40, $N_{x0} = 40$ and $\gamma_u = 10$.

All the abovementioned methods use the distribution alignment technique. If the distributions of labeled and unlabeled data do not match (i.e., mismatched distribution), as shown in Figure 1, the distribution alignment technique will not work well [2]. RDA [2] proposes a new distribution alignment algorithm that can be applied to mismatched distribution scenarios. However, when using unlabeled data, the information at the feature-level plays a crucial role. Consequently, this vital aspect has not been considered in RDA.

Motivated by the above limitations, this paper develops a novel SSL algorithm called FSimC. At the semantic-level, we introduce the primary classifier ($\mathcal{M}$) and the complementary classifier ($\mathcal{C}$) to predict the pseudo labels and the complementary labels (i.e., not belonging to this class) of unlabeled data respectively, then interact the information of the pseudo labels and complementary labels. At the feature-level, we first establish a cache to store the feature vectors of the labeled data. Then, we calculate the similarity between the feature vectors

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of weakly and strongly augmented unlabeled data and those stored in the cache, to obtain the weakly and strongly augmented feature similarity vectors. Finally, we perform feature similarity consistency on the weakly and strongly augmented feature similarity vectors to update the convolutional layer. This feature similarity consistency forces the convolutional layer to produce similar feature vectors for the same unlabeled data under different augmentations. Our contributions can be summarized in the following three aspects:

- We propose FSimC, a novel SSL algorithm that successfully leverages semantic and feature information from unlabeled data.
- FSimC achieves similar consistency on the same unlabeled data under different augmentation methods, improving model robustness and forecast prediction accuracy.
- Extensive experiments demonstrate that FSimC can be applied to matched and mismatched distribution scenarios with excellent performance and robustness against disruption.

II. PROPOSED METHOD

In this section, we will introduce FSimC and provide detailed description of our algorithm.

A. Semantic-Level

Denote a batch of B labeled data as $\mathcal{X} = \{(x_b, y_b)\}_{b=1}^{B}$, where $y \in \mathcal{Y} = \{1, \cdots, n\}$ is the ground truth label of x, and n represents the number of categories in the dataset which x belongs to. For the complementary label $\bar{y}$ of x, it is randomly chosen from $\mathcal{Y}/\{y\}$. We perform weak augmentation on labeled data denoted as $w(x)$, then extract feature vector $z = \mathcal{N}(w(x))$ using the convolutional layer. Finally, we feed the feature vector into two different fully connected layers. One called the main classifier ($\mathcal{M}$) to predict the pseudo label ($P_{\mathcal{M}}(z)$), and the other called the complementary classifier ($\mathcal{C}$) to predict the complementary label ($P_{\mathcal{C}}(z)$). Next, we compute the cross-entropy loss $H(\cdot,\cdot)$ between $P_{\mathcal{M}}(z)$ and y, and between $P_{\mathcal{C}}(z)$ and $\bar{y}$, which yields the main supervised loss and the complementary supervised loss, respectively. The main supervised loss is used for the classifier $\mathcal{M}$, while the complementary supervised loss is used for the classifier $\mathcal{C}$:

$$\mathcal{L}_{S\mathcal{M}} = \frac{1}{B}\sum_{n=1}^{B} H\left(y, P_{\mathcal{M}}(z)\right), \mathcal{L}_{S\mathcal{C}} = \frac{1}{B}\sum_{n=1}^{B} H\left(\bar{y}, P_{\mathcal{C}}(z)\right). \tag{1}$$

We employ the unlabeled data to perform consistency regularization on $\mathcal{M}$ and $\mathcal{C}$. Besides, we also take advantage of the interaction between the two classifiers to make the pseudo labels predicted by $\mathcal{M}$ less dependent on the distribution of labeled data, thus alleviating the impact of distribution mismatch. For a batch of unlabeled data $\mathcal{U} = \{(u_b)\}_{b=1}^{B}$ containing B samples, we first apply weak augmentation $w(\cdot)$ and strong augmentation $s(\cdot)$ to them. Then, we use the convolutional layer to extract the feature vectors $z_w = \mathcal{N}(w(u_b))$ and $z_s = \mathcal{N}(s(u_b))$. Finally, we input these feature vectors to $\mathcal{M}$ and $\mathcal{C}$ respectively, to obtain pseudo labels $p_w = P_{\mathcal{M}}(z_w)$, $p_s = P_{\mathcal{M}}(z_s)$, and complementary labels $q_w = P_{\mathcal{C}}(z_w)$ and $q_s = P_{\mathcal{C}}(z_s)$. Next, we perform transpose operation on the pseudo label p_w and the complementary label q_w to obtain their corresponding complementary label $\bar{p}_w$ and pseudo label $\bar{q}_w$:

$$\bar{p}_w = Norm\left(\mathbb{1} - p_w\right), \bar{q}_w = Norm\left(\mathbb{1} - q_w\right), \tag{2}$$

where $\mathbb{1}$ denotes a vector of all-one, and $Norm(\cdot)$ represents the normalization operation which can be expressed as $Norm(x_i) = x_i / \sum_{j=1}^{n} x_j$, $i \in (1, \cdots, n)$. Although we perform transposition operations on p_w and q_w, the relative relationships between classes remain unchanged. Therefore, we can still extract valuable information from $\bar{p}_w$ and $\bar{q}_w$.

Next, we use $\bar{p}_w$ and $\bar{q}_w$ as a guide to scale the predicted vectors of $\mathcal{M}$ and $\mathcal{C}$ (i.e., p_w and q_w):

$$\tilde{p}_w = Norm\left(p_w \times \frac{\mathcal{E}(\bar{q}_w)}{\mathcal{E}(p_w)}\right), \tag{3}$$

$$\tilde{q}_w = Norm\left(q_w \times \frac{\mathcal{E}(\bar{p}_w)}{\mathcal{E}(q_w)}\right), \tag{4}$$

where $\mathcal{E}(\cdot)$ calculates the moving average of the last 128 batches to make the obtained vectors smoother and avoid significant fluctuations caused by outliers, thus enhancing the robustness of the model. Subsequently, to obtain low entropy prediction results, we first calculate $arg\max(\tilde{p}_w)$ to get the one-hot vector $\hat{\tilde{p}}_w$. Then, we use $\hat{\tilde{p}}_w$ as a pseudo label to perform consistency regularization on the main classifier $\mathcal{M}$:

$$\mathcal{L}_{\mathcal{UM}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} H\left(\hat{\tilde{p}}_{wi}, p_{si}\right). \tag{5}$$

Unlike $\mathcal{M}$, when applying regularization to $\mathcal{C}$, we use the soft label $\tilde{q}_w$ as the complementary label to calculate the cross-entropy loss with q_s:

$$\mathcal{L}_{\mathcal{UC}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} H\left(\tilde{q}_{wi}, q_{si}\right). \tag{6}$$

In the process of extracting semantic-level information, FSimC does not set any threshold or hyperparameter to filter unlabeled data. Thus, FSimC can exploit all unlabeled data.

B. Feature-Level

In order to extract more valuable information from the data, in addition to using the semantic-level information, we also integrate feature-level information by performing feature similarity consistency in the process of model training. To begin with, we establish a K size cache to store the features of weakly augmented labeled data during the training process as $z_k = \{(z_1, \cdots, z_K)\}$. It should be noted that the feature vectors stored in the cache are continuously updated and follow the first-in-first-out principle.

For a batch of unlabeled data $\mathcal{U}$, the corresponding feature vectors of weakly augmented and strongly augmented versions are denoted as $\left\{\left(z_w^b\right)\right\}_{b=1}^{B}$ and $\left\{\left(z_s^b\right)\right\}_{b=1}^{B}$, respectively. For

the feature vector z_w^b of weakly augmented unlabeled data, we calculate its similarity with each feature vector in the cache to generate its corresponding feature similarity vector, followed by a $soft\max$ operation:

$$Q_w = soft\max\left(sim\left(z_w^b, z_k\right)/t\right), \tag{7}$$

where $sim\left(\cdot,\cdot\right)$ is used to calculate the similarity between two vectors, which can be represented by $sim\left(u,v\right) = uv^T/\left\|u\right\|\left\|v\right\|$. The parameter t is used to control the sharpness of the distribution. Next, we calculate the feature similarity vector between the feature vector of strongly augmented unlabeled data z_s^b with z_k in the same way as above:

$$Q_s = soft\max\left(sim\left(z_s^b, z_k\right)/t\right). \tag{8}$$

Finally, we perform the feature similarity regularization by calculating the cross-entropy loss between Q_w and Q_s:

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{uB}\sum_{b=1}^{uB} H\left(Q_w, Q_s\right). \tag{9}$$

Here, $\mathcal{L}_{\mathcal{S}}$ acts as the convolutional layer to reduce the differences in the feature distributions obtained by this model for the same unlabeled data under different augmentation methods.

C. Overall Loss

Based on the above designs, we can derive the overall loss function:

$$\mathcal{L} = \mathcal{L}_{\mathcal{SM}} + \mathcal{L}_{\mathcal{SC}} + \mathcal{L}_{\mathcal{UM}} + \mathcal{L}_{\mathcal{UC}} + \mathcal{L}_{\mathcal{S}}. \tag{10}$$

At the semantic-level, $\mathcal{L}_{\mathcal{SM}}$ and $\mathcal{L}_{\mathcal{SC}}$ use labeled data to train $\mathcal{M}$ and $\mathcal{C}$ respectively, while $\mathcal{L}_{\mathcal{UM}}$ and $\mathcal{L}_{\mathcal{UC}}$ use unlabeled data to perform regularization operation on $\mathcal{M}$ and $\mathcal{C}$ separately, to address the problem of mismatched distribution. At the feature-level, $\mathcal{L}_{\mathcal{S}}$ uses unlabeled data to perform feature similarity consistency on the convolutional layer.

III. EXPERIMENTS

A. Implementation Details

For all experiments, we use Wide ResNet 28-2 [9] and Wide ResNet 28-8 [9] as the backbone for CIFAR-10 [10] and CIFAR-100 [10], and Resnet-18 [11] as the backbone for STL-10 [12]. We apply RandAugment [13] to strongly augment the unlabeled data. We use SGD [14] as the optimizer, with the weight decay set to 0.0005, the learning rate set to 0.03, and use the decay strategy with cosine learning rate. The model was trained for 1,024 epochs. Unless otherwise specified, we set all loss coefficients to 1 and cache size to K = 1,024 and B = 64.

B. Datasets and Baselines

We evaluate our method on three widely used SSL datasets: 1) **CIFAR-10**, which includes 60,000 32×32 images with 10 classes, each containing 5,000 training samples and 1,000 testing samples. 2) **CIFAR-100**, which is a more complex dataset than CIFAR-10, with 100 classes, 500 training samples, and 100 test samples for each class. 3) **STL-10**, which consists of 113,000 96×96 samples with 10 classes, each having 500 labeled training samples and 800 labeled testing samples. Additionally, there are 100,000 unlabeled samples, including samples outside of these 10 classes, which better reflect the real-world data distribution and can help evaluate the performance of tested SSL methods.

TABLE I
PERFORMANCE COMPARISON ON CIFAR-10, CIFAR-100, AND STL-10 WITH BALANCED, MATCHED Dx AND Du.

	CIFAR-10		CIFAR-100	STL-10
Methods	80 labels	100 labels	1000 labels	1000 labels
MixMatch [7]	80.76%	84.17%	55.32%	63.21%
AlphaMatch [15]	92.45%	92.91%	60.21%	69.57%
FixMatch [3]	91.99%	93.14%	59.73%	65.38%
CoMatch [16]	94.08%	94.55%	60.81%	79.80%
RDA [2]	94.24%	94.35%	61.61%	82.63%
FSimC	**94.93%**	**95.03%**	**63.53%**	**87.92%**

TABLE II
PERFORMANCE COMPARISON ON CIFAR-10 AND CIFAR-100 WITH IMBALANCED Dx AND BALANCED Du.

	CIFAR-10	CIFAR-100		
	100 labels	400 labels	1000 labels	
Methods	N_{x0}=80	40	60	80
FixMatch [3]	71.57%	25.66%	42.81%	40.22%
CoMatch [16]	26.91%	25.66%	30.17%	28.35%
RDA [2]	79.00%	30.86%	44.02%	41.29%
FSimC	**88.40%**	**32.12%**	**45.29%**	**44.81%**

We adopt six SSL methods as baselines for comparison, including Mixmatch [7], AlphaMatch [15], FixMatch [3], CoMatch [16], RDA [2], and DARP [5], which are representatives of state-of-the-arts.

C. Dx and Du

To demonstrate the effectiveness of FSimC in different data distribution scenarios, we conduct the experiments in the following four cases: 1) The labeled data (Dx) and unlabeled data (Du) are balanced in distributions, and Dx and Du are matched. 2) Dx is imbalanced while Du is balanced, as shown in Figure 1a. 3) Dx is balanced while Du is imbalanced. 4) Dx and Du are both imbalanced and mismatched, as shown in Figure 1b. It is worth noting that when both Dx and Du are imbalanced, Dx is a transposition of Du. This scenario is a more challenging experimental setting that can better evaluate the robustness and generalization ability of the compared methods.

To construct the imbalanced or mismatched Dx and Du scenarios, when the dataset has n classes, we determine the number of labeled data for each class in the following way: $N_{xi} = N_{x0} \times \gamma_x^{-\frac{i-1}{n-1}}, i \in (1, \cdots, n)$. N_{x0} can control the specific number of labeled data for each class. The value of γ_x ensures that $\sum_{i=1}^{n} N_{xi}$ is equal to the total number of labeled data. The number of unlabeled data in each category is: $N_{ui} = N_{u0} \times \gamma_u^{-\frac{i-1}{n-1}}$. In CIFAR-10 and CIFAR-100, N_{u0} is set to 5000 and 500, respectively. The value of γ_u ensures

TABLE III
PERFORMANCE COMPARISON ON CIFAR-10 AND CIFAR-100 WITH IMBALANCED, MISMATCHED Dx AND Du.

	CIFAR-10			CIFAR-100
	40 labels	100 labels		1000 labels
Methods	N_{x0}, γ_u=40,10	40, 5	40, 10	40, 10
FixMatch [3]	57.78%	58.72%	57.49%	29.73%
RDA [2]	68.54%	79.33%	70.93%	35.11%
FSimC	**70.17%**	**80.62%**	**72.02%**	**36.31%**

TABLE IV
PERFORMANCE COMPARISON ON CIFAR-10 WITH BALANCED Dx AND IMBALANCED Du.

	CIFAR-10			
	50 labels	100 labels		200 labels
Methods	γ_u=70	100	200	100
FixMatch [3]	46.21%	50.08%	41.27%	52.63%
CoMatch [16]	45.73%	47.11%	40.31%	50.91%
RDA [2]	51.31%	52.99%	45.17%	56.34%
FSimC	**53.12%**	**55.86%**	**46.36%**	**58.17%**

TABLE V
PERFORMANCE COMPARISON ON CIFAR-10 AND STL-10 WITH DROPOUT SETTING.

	CIFAR-10 (γ_x=100)		STL-10 ($\gamma_x \neq \gamma_u$)	
Methods	$\gamma_u = 50$	150	$\gamma_x = 10$	20
FixMatch [3]	73.90%	69.60%	72.90%	63.40%
DARP [5]	77.30%	72.90%	77.80%	69.90%
RDA [2]	79.77%	74.48%	83.61%	82.86%
FSimC	**80.97%**	**76.13%**	**84.85%**	**84.13%**

TABLE VI
ABLATION STUDY ON CIFAR-100 WITH THE WIDE RESNET 28-2 AS THE BACKBONE AND 1000 LABELS.

	CIFAR-100 (1000 labels)	
Methods	$N_{x0} = 80$	$N_{x0} = 40, \gamma_u = 10$
w/o $\mathcal{L}_S$	37.03%	32.15%
w/o Eq. (3)	34.08%	30.18%
w/o Eq. (4)	37.63%	31.54%
FSimC	**39.26%**	**33.47%**

that $\sum_{i=1}^{n} N_{ui}$ is equal to the total number of unlabeled data in the dataset.

D. Experimental Results

1) Dx and Du are balanced and matched: We test our approach on CIFAR-10, CIFAR-100 and STL-10 with different numbers of labeled data, and compare them with existing baseline methods. As shown in Table I, we can see that FSimC outperforms RDA by 2% on the complex CIFAR-100 dataset. In addition, on the more realistic STL-10 dataset, FSimC is 1.3% better than RDA. In CIFAR-10, we test our method with 10 and 20 labeled data. The result shows that our FSimC achieves higher classification accuracy than the baseline methods. The performance gain of FSimC is due to not only the improved pseudo labels through the interaction between complementary labels and pseudo labels, but also the effective use of the semantic and feature information from all unlabeled data.

2) Imbalanced Dx and balanced Du: Given a fixed total number of labeled data, we adjust the imbalance degree of Dx by varying N_{x0}, while keeping Du balance. We test CIFAR-10 and CIFAR-100 with different values of N_{x0}, and the results are reported in Table II. We can see from Table II that FSimC outperforms other baseline methods in case of mismatched distribution. Especially, when the number of labeled data is 100 and N_{x0} is 80, FSimC exceeds RDA by 9.4% on CIFAR-10. When the number of labeled data is 1,000 and N_{x0} = 80, FSimC respectively exceeds RDA and FixMatch by 3.5% and 4.6% on CIFAR-100.

It can be observed that all SSL methods show a significant decrease in classification accuracy in the mismatched distribution scenario. This is because the traditional consistency regularization requires the match between Dx and Du. While FSimC introduces a new distribution alignment approach at the semantic-level, and performs feature similarity consistency on unlabeled data under different augmentation methods at the feature-level. These strategies improve the accuracy of pseudo labels and the robustness of the model under mismatched Dx and Du, further boosting the performance of FSimC.

3) Dx and Du are imbalanced and mismatched: This is a more challenging test environment, where we conduct experiments on CIFAR-10 and CIFAR-100 under different imbalanced and mismatched distribution settings. The results are shown in Table III. In this more severe mismatched distribution condition, although the accuracy of all the compared methods decreases, FSimC still obtains satisfactory performance and outperforms all the baseline methods.

4) Balanced Dx and imbalanced Du: In this context, we conduct experiments on CIFAR-10, as shown in Table IV. These results show that despite the change in experimental settings, FSimC still achieves excellent performance, and its accuracy surpasses all the baseline methods. This demonstrates that FSimC has great robustness when handling different mismatched distribution scenarios.

5) Mismatched distribution scenario with the DRAP protocol: In different experimental scenarios that meet the DRAP protocol, we conduct experiments on CIFAR-10 and STL-10, and the results are shown in Table V. The experimental results show that FSimC performs excellently in all experimental scenarios, outperforming the DRAP and RDA methods.

E. Ablation Study

In this part, we conduct ablation experiments on the CIFAR-100 by using the Wide ResNet 28-2 as the backbone with a limited set of 1,000 labeled samples to assess the contribution of each component to the proposed FSimC. The experimental results are presented in Table VI. Specifically, we consider different experimental setups, include N_{x0} = 80 or N_{x0} = 40, $\gamma_u = 10$. Notably, we observe varying performance degradation when we remove either the feature similarity

consistency or the interaction between pseudo labels and complementary labels. This observation clearly demonstrates the importance and effectiveness of the individual components within FSimC, highlighting their essential contributions to the overall performance.

IV. Conclusion

In this article, we have presented a new SSL algorithm called FSimC, which efficiently exploits all unlabeled data's semantic- and feature-level information. First, we interact with the distribution information of pseudo and complementary labels at the semantic-level to handle the mismatched distribution scenarios. And at the feature-level, we perform feature similarity consistency on the same unlabeled data under different augmentation strategies, to improve the robustness of the model and further increase its performance in various scenarios. Finally, we conduct extensive experiments with different experimental parameters, and the results show that our FSimC surpasses existing methods in both the matched and mismatched distribution scenarios.

Acknowledgements: This work was supported by the National Natural Science Foundation of China under Grant 62272116. The authors acknowledge the Network Center of Guangzhou University for providing HPC computing resources.

References

[1] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *International Conference on Learning Representations*, pp. 1–13, 2020.

[2] Y. Duan, L. Qi, L. Wang, L. Zhou, and Y. Shi, "Rda: Reciprocal distribution alignment for robust semi-supervised learning," *European Conference on Computer Vision*, pp. 533–549, 2022.

[3] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[4] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[5] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 567–14 579, 2020.

[6] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu, "Simmatch: Semi-supervised learning with similarity matching," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14 471–14 481, 2022.

[7] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5049–5059, 2019.

[8] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1–8, 2004.

[9] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[10] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technique Report*, 2009.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[12] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," *Proceedings of Machine Learning Research*, vol. 15, pp. 215–223, 2011.

[13] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

[14] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[15] C. Gong, D. Wang, and Q. Liu, "Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13 683–13 692, 2021.

[16] J. Li, C. Xiong, and S. C. Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," *IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Collaborative Pseudo Labeling for Prompt-Based Learning

Jen-Tzung Chien and Chien-Ching Chen
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Prompt-based learning has been recently popular as an emerging data efficient approach to parameter fine-tuning for various few-shot classification tasks. The knowledge transfer from a large-scale pre-trained language model could be systematically performed. It is challenging to conduct transfer learning from unlabeled data. Pseudo labeling is feasible to enhance the model generalization by unsupervised learning where additional pseudo labeled data are augmented. However, this method likely suffers from the model bias problem due to the incorrect pseudo labels. This paper presents a new collaborative pseudo labeling for prompt learning which is implemented through a teacher-student model where teacher and student are learned collaboratively. Accordingly, the resulting pseudo labels do not mislead the model while the few-shot classification task can be improved. The experiments on few-shot sentiment classification and natural language inference show the merit of collaborative pseudo labeling and prompt tuning compared with the other methods.**

I. Introduction

In the past few years, the pre-trained language models (PLMs) [1], [2] have been popularly utilized and fine-tuned via supervised methods where the end performance has been significantly advanced for various downstream tasks. An efficient and effective knowledge transfer has been made for the tasks such as natural language inference and machine translation. Basically, the success of these works depend on sufficiency of training data [3]. However, collecting the training samples and their labels is usually time consuming and cost expensive. The real-world tasks likely suffers from the sparseness of labeled data which considerably degrades the performance of supervised learning. To handle these issues, the recent works on prompt tuning [4] based on PLMs have been emerging and receiving numerous attention. Such an approach is feasible to activate the self guided learning via natural language prompt so as to optimize the model capacity. For example, the input sentence "The kids are laughing." with manual prompt template "It is [MASK]" can be converted to the prompt sentence "The kids are laughing. It is [MASK]". Then, the masked language model (MLM) is learned to calculate the blank or mask token [MASK] with a meaningful word which is close to the predefined label word "great" or "good". Recently, a method called better few-shot fine-tuning of language models (LM-BEF) [5] was proposed to implement the prompt tuning for small language models based on BERT [1], [6] and RoBERTa [2] in a way to simulate the concept of in-context learning in the large-scaled model based on the generative pre-trained transformer (GPT-3) [4]. After the prompt tuning with a number of annotated samples, even small language models could achieve the performance comparable with large-scaled model using GPT-3. However, it was difficult to obtain clear improvement by using few-shot supervised data. Accordingly, the rich information from unlabeled data was incorporated to carry out semi-supervised learning in [7], [8]. The resulting implementation was run by using those unlabeled data combined with high-confidence pseudo labels. The generated pseudo labels therefore served as additional supervised data to mitigate the issue of limited data to train generative model [9], [10]. Previously, the pseudo labeling method had an severe problem called the confirmation bias where the incorrect pseudo labels would mislead the model training. In order to address the aforementioned issue, this paper combines the hard prompt tuning driven by the idea of meta pseudo labels (MPL) [11]. A teacher-student framework is introduced to learn the information from pseudo labels so as to improve the supervised performance by using labeled data. Prompt tuning and pseudo labeling are collaboratively performed. The experiments on natural language tasks illustrate the goodness of using this work to cope with the issue of sparse labeled data in text representation.

II. Background Survey

This study presents a new approach to semi-supervised learning where hard prompt tuning is merged with the scheme of pseudo labeling in presence of labeled and unlabeled data.

A. Semi-Supervised Leaning

An inevitable weakness of training a deep neural network is the scarcity of labeled data. The semi-supervised learning [12], [13], [14] has been developed as an emerging and challenging research direction. There have been various methods deployed in real-world applications where the labeled and unlabeled data were both used. A recent method based on the consistency training [15], [16], [17], [7] was proposed to utilize the unlabeled data to enforce the consistency of the model outputs given the similar inputs. The concept of consistency training aims to make prediction of a training sample which is consistent even when data augmentation or noise addition over training data has been done. In order to increase the amount of labeled data for model training, this study leverages the supervision signals offered by labeled data through generation of new pseudo labeled data from unlabeled data. Basically, pseudo labeling [18], [19] focuses on finding the pseudo labels of unlabeled data with high confidence scores which are then augmented in the training dataset for semi-supervised training. Recently, a hybrid model [20], [8] combining the consistency

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

training and pseudo labeling was accordingly developed to improve the performance of semi-supervised learning [21].

B. Prompt-Based Learning

Prompt-based learning is known as an effective approach to enhance model capacity based on a pre-trained language model (PLM) through prompting and learning in a supervised setting. Different from the standard PLM fine-tuning, the prompt tuning is implemented by directly calculating the probability of text data by utilizing PLM. In order to exploit the prior knowledge from PLM, the input sentence $\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_L\}$ is transformed to form a prompt sentence $\mathbf{x}_{\text{prompt}}$ with a hard prompt template $\mathcal{T}(\cdot)$ with the masked token [MASK] as

$$\mathbf{x}_{\text{prompt}} = \mathcal{T}(\mathbf{x}) = \{\mathbf{x}, \text{It is } [\text{MASK}]\}. \tag{1}$$

After obtaining the hidden representation $\mathbf{h}_{[\text{MASK}]}$ of prompt sentence $\mathbf{x}_{\text{prompt}}$ by using PLM, the masked language model has to predict the word distribution which should fill the masked token with the correct word v in vector $\mathbf{w}_v$ by using

$$p([\text{MASK}] = v|\mathbf{x}_{\text{prompt}}) = \frac{\exp(\mathbf{w}_v^T \cdot \mathbf{h}_{[\text{MASK}]})}{\sum_{v' \in \mathcal{V}} \exp(\mathbf{w}_{v'}^T \cdot \mathbf{h}_{[\text{MASK}]})}. \tag{2}$$

The verbalizer $\mathcal{M} : \mathcal{Y} \rightarrow \mathcal{V}$ is the mapping from the task label space $\mathcal{Y}$ to individual words in vocabulary space $\mathcal{V}$ which are masked [22], [23]. The distribution of predicting the word label $v \in \mathcal{M}(y)$ can be expressed by using the prompt-based language model with parameter θ in a form of

$$\begin{aligned} p_\theta(y|\mathbf{x}) &= p_\theta([\text{MASK}] = \mathcal{M}(y)|\mathbf{x}_{\text{prompt}}) \\ &= \sum_{v \in \mathcal{M}(y)} p_\theta([\text{MASK}] = v|\mathbf{x}_{\text{prompt}}). \end{aligned} \tag{3}$$

The objective for prompt-based learning is seen as a classification problem where language model parameter θ is fine-tuned by minimizing the cross-entropy loss

$$\mathcal{L}_{\text{prompt}}(\mathbf{x}, y, \theta) = -\sum_{y \in \mathcal{Y}} y \log p_\theta(y|\mathbf{x}). \tag{4}$$

III. Pseudo Labeling & Prompt Tuning

This study carries out a semi-supervised scheme based on pseudo labeling which is implemented for prompt tuning and learning in a number of natural language tasks. In general, prompt tuning is exploited to estimate the masked language model head so as to predict the words behind the masked tokens. Such a process is similar to the fine-tuning process in training a standard classifier. This paper would like to incorporate pseudo labeling in prompt tuning and language modeling where a teacher-student framework is built with two masked language model heads. The overall system architecture is illustrated in Figure 1 which is detailed in what follows.

A. Teacher-Student Framework for Pseudo Labeling

A teacher-student framework is built to carry out pseudo labeling for prompt-based learning where semi-supervised learning is performed by using the labeled and unlabeled data. A pre-trained language model (PLM) is utilized and

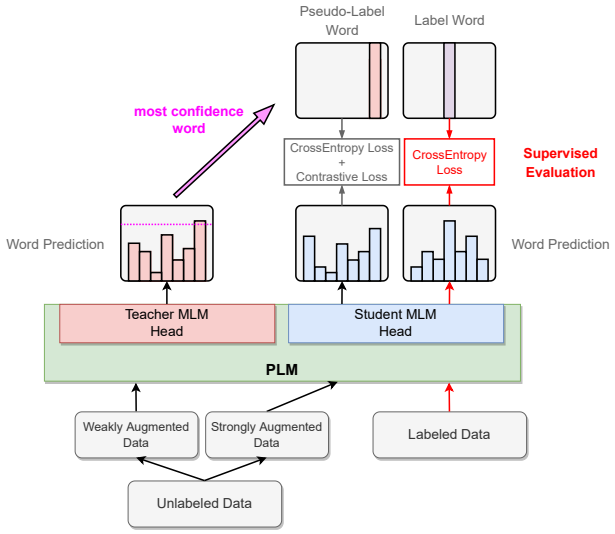


Fig. 1. Overview of semi-supervised learning via pseudo labeling where the teacher-student MLM heads in a pre-trained language model are estimated by prompt-based learning.

the masked language model (MLM) heads for a teacher and a student are configured on top of PLM. These two MLM heads are estimated through prompt-based learning. In the process, the weakly and strongly augmented data from unlabeled samples are first collected and then fed in PLM to find the predicted words of masked tokens in prompt-based representation from the teacher and student MLM head, respectively. On the other hand, the student MLM head is also utilized to predict the masked tokens by using the labeled sentences. The prompt-based cross-entropy loss from the head outputs is minimized to conduct a supervised evaluation from purely labeled data. In particular, another cross-entropy loss, which is calculated from the outputs using weakly and strongly augmented data, is minimized to perform an unsupervised evaluation from purely unlabeled data. The predicted words of the masked tokens from the weakly augmented sentences with the highest confidence scores are treated as the pseudo labels for a kind of collaborative learning [24], [25] under a teacher-student framework. The learning is performed to pursue the *label consistency* from the head outputs of teacher and student MLMs. Teacher MLM head is learned to provide the pseudo labels corresponding to the unlabeled sentences which are utilized to learn the student MLM head. In what follows, the optimization problem is formulated.

B. Objectives from Unlabeled and Labeled Data

In this study, the PLM $\mathcal{F}_\phi$ with the pre-trained parameters ϕ is acted as a feature extractor or encoder. The controllable parameters include the masked language model heads for teacher f_{θ_t} and student f_{θ_s} which are seen as the classifier layers for word prediction of masked tokens. Let the hidden representation ofthe weakly augmented data of the unlabeled prompt input $\mathbf{x}_{\text{prompt}} \sim \mathcal{D}_u$ be calculated by using the PLM-

based feature extractor as

$$\mathbf{h}_{[\text{MASK}]} = \mathcal{F}_\phi(\mathbf{x}_{\text{prompt}}). \tag{5}$$

This representation is then used to predict the pseudo label $\widehat{y}_{\theta_t}$ of a masked token based on MLM scheme via

$$\widehat{y}_{\theta_t} = \underset{i}{\text{argmax}} f_{\theta_t,i}(\mathbf{h}_{[\text{MASK}]}) \tag{6}$$

where $f_{\theta_t,i}$ denotes the output node of teacher MLM head corresponding to i^{th} vocabulary word. The strongly augmented data from unlabeled data $\mathbf{x}_u \sim \mathcal{D}_u$ are fed to PLM by using the student MLM head f_{θ_s} where the overall prompt-based cross-entropy loss is calculated by

$$\mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s) = \sum_{n=1}^{|\mathcal{D}_u|} \mathcal{L}_{\text{prompt}}(\mathbf{x}_u^{(n)}, \widehat{y}_{\theta_t}^{(n)}, \theta_s) \tag{7}$$

where n is the sample index. Notably, the pseudo labels from teacher MLM head $\{\widehat{y}_{\theta_t}^{(n)}\}$ are used. Then, the student MLM head θ_s is estimated according to the stochastic gradient descent (SGD) algorithm with a learning rate α

$$\theta_s' = \theta_s - \alpha \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s). \tag{8}$$

The pseudo labels $\{\widehat{y}_{\theta_t}^{(n)}\}$ does affect the supervised evaluation of student MLM head by using the labeled data $\mathbf{x}_l \sim \mathcal{D}_l$. The updated student parameter θ_s' always depends on the teach parameter θ_t through its relation to pseudo labels $\{\widehat{y}_{\theta_t}^{(n)}\}$. The prompt-based cross-entropy loss from labeled data $\mathcal{D}_l = \{\mathbf{x}_l^{(n)}, y_l^{(n)}\}$ is therefore calculated by

$$\begin{aligned} &\mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s') = \sum_{n=1}^{|\mathcal{D}_l|} \mathcal{L}_{\text{prompt}}(\mathbf{x}_l^{(n)}, y_l^{(n)}, \theta_s') \\ &\text{where} \quad \theta_s'(\theta_t) = \theta_s' \end{aligned} \tag{9}$$

C. Optimization and Implementation

This study aims to improve the supervised evaluation of using student MLM head through a scheme called collaborative pseudo labeling (CPL) where the teacher MLM head provides the pseudo supervision. A constrained optimization problem is then formulated and expressed by

$$\begin{aligned} &\min_{\theta_t} \ \mathcal{L}_{\text{prompt}}\left(\mathcal{D}_l, \theta_s(\theta_t)\right) \\ &\text{subject to} \ \ \theta_s'(\theta_t) = \underset{\theta_s}{\text{argmin}} \ \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s). \end{aligned} \tag{10}$$

To facilitate the teacher-student learning framework, the idea of meta learning for pseudo labels [11] is merged to rewrite the optimization problem as an unconstrained optimization.

$$\begin{aligned} &\min_{\theta_t} \ \mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s'(\theta_t)) \\ &= \min_{\theta_t} \ \mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s - \alpha \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s)). \end{aligned} \tag{11}$$

In the implementation, the gradient of the prompt-based loss with respect to the teacher MLM head θ_t can be derived as

$$\begin{aligned} &\nabla_{\theta_t} \mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s - \alpha \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s)) \\ &= -\alpha \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s') \nabla_{\theta_t} [\nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s)] \\ &= -\alpha \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s') \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s) \\ &\quad \times \nabla_{\theta_t} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_t) \triangleq g_t(\mathcal{D}_l, \mathcal{D}_u). \end{aligned} \tag{12}$$

The double gradient term in Eq. (12) with the defined function $g_s(\widehat{y}_{\theta_t}) = \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathbf{x}_u, \widehat{y}_{\theta_t}, \theta_s)$ can be derived as

$$\begin{aligned} &\nabla_{\theta_t} \mathbb{E}_{\mathbf{x}_u \sim \mathcal{D}_u, \widehat{y}_{\theta_t} \sim f_{\theta_t}(\mathbf{x}_u)} [\nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathbf{x}_u, \widehat{y}_{\theta_t}, \theta_s)] \\ &= \mathbb{E}_{\mathbf{x}_u \sim \mathcal{D}_u, \widehat{y}_{\theta_t} \sim f_{\theta_t}(\mathbf{x}_u)} [\nabla_{\theta_t} g_s(\widehat{y}_{\theta_t})] \\ &= \mathbb{E}_{\mathbf{x}_u \sim \mathcal{D}_u, \widehat{y}_{\theta_t} \sim f_{\theta_t}(\mathbf{x}_u)} [g_s(\widehat{y}_{\theta_t}) \nabla_{\theta_t} \widehat{y}_{\theta_t}] \\ &= \mathbb{E}_{\mathbf{x}_u \sim \mathcal{D}_u, \widehat{y}_{\theta_t} \sim f_{\theta_t}(\mathbf{x}_u)} [g_s(\widehat{y}_{\theta_t}) \nabla_{\theta_t} \log p(\widehat{y}_{\theta_t} | \mathbf{x}_u, \theta_t)] \\ &= \mathbb{E}_{\mathbf{x}_u \sim \mathcal{D}_u, \widehat{y}_{\theta_t} \sim f_{\theta_t}(\mathbf{x}_u)} [g_s(\widehat{y}_{\theta_t}) \nabla_{\theta_t} \mathcal{L}_{\text{prompt}}(\mathbf{x}_u, \widehat{y}_{\theta_t}, \theta_t)] \end{aligned} \tag{13}$$

which is obtained by considering the prompt-based loss using cross-entropy loss. The gradient over teacher MLM head θ_t is finally derived as a product of three gradients. Therefore, CPL is implemented by a learning procedure as shown in Algorithm 1. The feedback from teacher MLM head via the pseudo labels $\widehat{y}_{\theta_t}$, is received to jointly update the student and teacher heads $\{\theta_s, \theta_t\}$.

Algorithm 1: Collaborative pseudo labeling for prompt-based language models.

Require: $\{\mathcal{D}_l, \mathcal{D}_u\}$: labeled and unlabeled training dataset. B: batch size. I: no. of iterations. ϕ: PLM parameters. $\{\theta_t, \theta_s\}$: teacher and student MLM head parameters. $\{\alpha, \beta\}$: learning rates of student and teacher heads.

for $i = 0, \ldots, I-1$ **do**

- sample labeled batch $\{\mathbf{x}_l^{(n)}, y_l^{(n)}\}_{n=1}^B \sim \mathcal{D}_l$
- sample unlabeled batch $\{\mathbf{x}_u^{(n)}\}_{n=1}^B \sim \mathcal{D}_u$
- sample pseudo labels $\widehat{y}_{\theta_T}^{(n)} \sim p(\widehat{y}_{\theta_T}^{(n)} | \mathbf{x}_u^{(n)}, \theta_t)$
- compute student loss with pseudo labels $\mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s) = \sum_{n=1}^B \mathcal{L}_{\text{prompt}}(\mathbf{x}_u^{(n)}, \widehat{y}_{\theta_t}^{(n)}, \theta_s)$
- update student $\theta_s' \leftarrow \theta_s - \alpha \nabla_{\theta_s} [\mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s)]$
- compute teacher loss $\mathcal{L}_{\text{prompt}}(\mathcal{D}_l, \theta_s - \alpha \nabla_{\theta_s} \mathcal{L}_{\text{prompt}}(\mathcal{D}_u, \theta_s))$
- compute teacher gradient $g_t(\mathcal{D}_l, \mathcal{D}_u)$ in Eq. (12)
- update teacher $\theta_t' \leftarrow \theta_t - \beta g_t(\mathcal{D}_l, \mathcal{D}_u)$

end

IV. Experiments

A. Experimental Settings

In the experiments, the proposed collaborative pseudo labeling (CPL) was evaluated for a number of public-domain down-stream tasks including sentiment classification using the Stanford sentiment treebank (SST-2 for two classes and SST-5 for five classes) [26], the customer review (CR) [27], the movie review (MR), the subjectivity (Subj) and natural language inference using the multi-genre NLI (MNLI), the Stanford NLI (SNLI), and the question answering NLI (QNLI). The standard settings in various public-domain datasets were adopted. For comparative study, the related methods including the standard fine-tuning (FT), the LM-BEF [5] and the state-of-the-art prompt-based few-shot learner (SFLM) [28] were carried out. The accuracy was reported in the comparison. The semi-supervised learning was evaluated by examining the

performance by tuning the ratio between unlabeled and labeled data, denoted by μ. Here, k denotes the number of labeled instances per class. Batch size B is 10. AdamW optimizer [29] was used. The learning rates α and β were both set as 10^{-5}. In the experiments, the PLM using BERT with 100M number of parameters is also fine-tuned in the learning process. The computation overhead of calculating teacher-student MLM heads relative to fine-tuning the whole BERT model [1] was only 3.5%. The ablation studies on the effects of supervision, data efficiency, latent visualization and attention score were assessed in various sets of comparison.

TABLE I
ACCURACY (%) ON SENTIMENT CLASSIFICATION.

Model	**SST-2**	**SST-5**	**CR**	**MR**	**Subj**
FT [1]	78.3	36.4	78.6	69.5	89.6
LM-BFF [5]	89.9	45.8	89.5	84.1	88.3
SFLM [28]	91.0	47.7	90.8	86.6	92.4
CPL	**91.4**	**48.1**	**92.1**	**86.9**	**92.5**

TABLE II
ACCURACY (%) ON NATURAL LANGUAGE INFERENCE.

Model	**MNLI**	**SNLI**	**QNLI**
FT [1]	41.2	43.9	60.3
LM-BFF [5]	60.2	65.8	60.6
SFLM [28]	62.6	67.4	61.0
CPL	**62.9**	**69.2**	**61.5**

B. Experimental Results

First of all, the classification accuracies of using FT, LM-BEF, SFLM and the proposed CPL in the tasks of SST-2, SST-5, CR, MR and Subj are shown in Table I. The results of MNLI, SNLI and QNLI are reported in Table II. The absolute and consistent improvements over state-of-the-art method SFLM [28] 0.4%, 0.4%, 1.9%, 0.3% 0.1%, 0.3%, 1.8% and 0.5% in the tasks of SST-2, SST-5, CR, MR, Subj, MNLI, SNLI and QNLI are obtained, respectively. In what follows, a set of ablation studies are conducted to evaluate various effects in the learning procedure based on the proposed CPL.

C. Ablations and Analysis

1) Effect of supervision: In this ablation study, the effect on the number of labeled data N in the performance of semi-supervised learning is examined. Table III shows the results of few-shot learning with N to be 20, 100, 1000 on SST-2 and MNLI. For SST-2, the total number of training examples is 6920. The case of using N=20 for fine-tuning PLM and estimating teacher-student MLM heads is relatively few. In this set of experiments, the number of unlabeled examples was fixed as 800. It is obvious that increasing number of labeled examples significantly raises the classification accuracy in different tasks. However, the ratio between unlabeled and labeled data also affects the end performance. The results shown in Tables I and II were the case of using 100 labeled and 800 unlabeled samples.

TABLE III
ACCURACY (%) ON SST-2 AND MNLI TASKS. N DENOTES THE NUMBER OF LABELED DATA FOR MODEL TRAINING.

Model	N	SST-2	MNLI
CPL	20	90.9	61.2
CPL	100	91.4	62.9
CPL	1000	93.7	83.0

2) Effect of data efficiency: This work uses labeled data and unlabeled data to learn for few-shot natural language classification tasks. However, the amounts of labeled and unlabeled data are required in the joint training is an issue. There are four sentiment classification tasks (SST-2, SST-5, CR and MR) analyzed in this evaluation. Figure 2 illustrates the performance of CPL with different combinations of k and μ. It is found that the error rate reduces as μ increases. This implies that the proposed CPL can work better with the increasing number of unlabeled samples. But, the performance improvement in SST-5 task is not significant. The benefit of increasing the amount of supervision is obvious. With the increase of labeled data, the error rate of all four tasks reduces significantly. The model is truly sensitive to the number of labeled data. The hyperparameters k=16 and μ=4 are fixed hereafter.

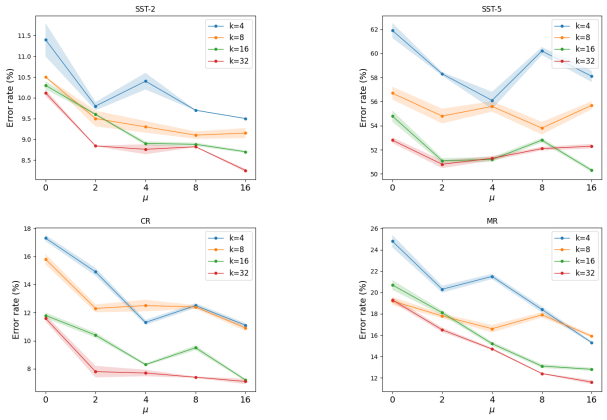


Fig. 2. Classification error rate (%) of using CPL under different numbers of instances per class k and ratios between unlabeled and labeled data μ where four tasks are evaluated.

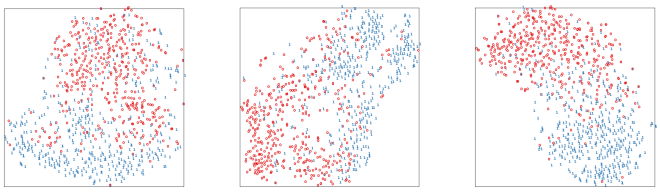


Fig. 3. Comparison of visualization of latent representations via t-SNE by using different models in presence of two classes.

3) Latent representation: In addition, Figure 3 shows the latent visualizations by using t-distributed stochastic neighbor embedding (t-SNE) among three methods where SST-2 task is evaluated. This comparison shows the visualization of latent embeddings or a kind of latent semantics for two classes as shown by red and blue. Here, the samples of special token [CLS] as a kind of class information for original input sentences are used. This token took from the final layer of PLM. For the prompt sentences, the embeddings of masked token [MASK] from the masked language model are used to predict the masked words in prompt-based fine-tuning. It can be seen that CPL and SFLM obtain better separation of two classes compared with FT. Latent representation using CPL look more separate than that using SFLM.

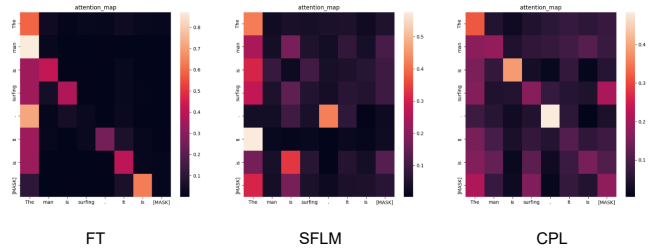


Fig. 4. Comparison of attention scores on SST-2 task.

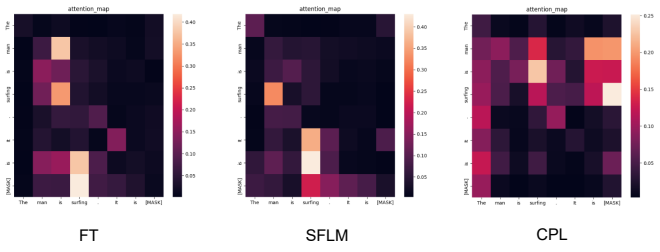


Fig. 5. Comparison of attention scores on SST-5 task.

4) Attention Analysis: The comparison of FT, SFLM and the proposed CPL is further extended to that for attention analysis. Figures 4 and 5 draw the attention maps of three methods for an example of sentence "The main is surfing. It is [MASK]" where SST-2 and SST-5 tasks are evaluated, respectively. These results show that CPL are more attentive to the masked token [MASK] (the right most token or column in the map) than FT and CPL.

V. Conclusions

This paper has presented the collaborative pseudo labeling for prompt-based learning, which was feasible to jointly learn the pseudo label words from the unlabeled data and improve the supervised evaluation of labeled data simultaneously. The experiments on a number of sentiment classification and natural language inference tasks showed that the potential improvement compared with fine-tuning method and state-of-the-art method could be achieved. Such an approach based on the teacher-student framework is able to be employed together with other methods to learn the pseudo label words from the unlabeled data so as to improve the supervised evaluation of labeled data simultaneously.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] Shinji Watanabe and Jen-Tzung Chien, *Bayesian Speech and Language Processing*, Cambridge University Press, 2015.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[5] Tianyu Gao, Adam Fisch, and Danqi Chen, "Making pre-trained language models better few-shot learners," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 3816–3830.

[6] M.-Y. Chen, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, "Meta learning for domain agnostic soft prompt," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[7] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[8] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[9] Jen-Tzung Chien and Chih-Jung Tsai, "Amortized mixture prior for variational sequence generation," in *Proc. of International Joint Conference on Neural Networks*, 2020, pp. 1–6.

[10] Tien-Ching Luo and Jen-Tzung Chien, "Variational dialogue generation with normalizing flows," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7778–7782.

[11] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le, "Meta pseudo labels," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11557–11568.

[12] Patrick Nguyen, Philippe Gelin, Jean-Claude Junqua, and Jen-Tzung Chien, "N-best based supervised and unsupervised adaptation for native and non-native speakers in cars," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 173–176.

[13] Hung-Yu Chen and Jen-Tzung Chien, "Deep semi-supervised learning for domain adaptation," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2015, pp. 1–6.

[14] Jen-Tzung Chien and Jean-Claude Junqua, "Unsupervised hierarchical adaptation using reliable selection of cluster-dependent parameters," *Speech Communication*, vol. 30, no. 4, pp. 235–253, 2000.

[15] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," in *Proc. of International Conference on Learning Representations*, 2017.

[16] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[18] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. of ICML Workshop on challenges in representation learning*, 2013.

[19] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, "Self-training with noisy student improves imagenet classification," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687–10698.

[20] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] Xiaodong Cui, Jing Huang, and Jen-Tzung Chien, "Multi-view and multi-objective semi-supervised learning for hmm-based automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1923–1935, 2012.

[22] Hou Lio, Shang-En Li, and Jen-Tzung Chien, "Adversarial mask transformer for sequential learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4178–4182.

[23] Jen-Tzung Chien and Yu-Han Huang, "Bayesian transformer using disentangled mask attention," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 1761–1765.

[24] Jen-Tzung Chien and Wei-Hsiang Chang, "Collaborative regularization for bidirectional domain mapping," in *Proc. of International Joint Conference on Neural Networks*, 2021, pp. 1–8.

[25] Jen-Tzung Chien and Wei-Hsiang Chang, "Dualformer: A unified bidirectional sequence-to-sequence learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7718–7722.

[26] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[27] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews," in *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.

[28] Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li, "Revisiting self-training for few-shot learning of language model," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9125–9135.

[29] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. of International Conference on Learning Representations*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Learning Meta Soft Prompt for Few-Shot Language Models

Jen-Tzung Chien* Ming-Yen Chen* Jing-Hao Xue†
* Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
† Department of Statistical Science, University College London, London, United Kingdom

***Abstract*—Prompt-based learning is powerful to utilize the large-scaled pre-trained language model (PLM) for language understanding where the input sentences are augmented by either adding the hard prompt using word tokens or the soft prompt in a form of trainable tokens. However, the learned soft prompt in training domain may not really help a frozen PLM to handle domain shift in test domain. This paper presents an approach to incorporate meta learning into domain adaptation to train new soft prompt which sufficiently generalizes the frozen PLM to a number of domains. The meta soft prompt is then developed for few-shot unsupervised domain adaptation where a frozen PLM can be quickly adapted to a target domain. This soft prompt is optimized according to meta learning where the domain adaptation loss and the prompt-based classification loss are jointly minimized. The experiments on multi-domain natural language understanding show the benefits of the proposed meta soft prompt in pre-trained language model by using BERT under the few-shot setting.**

***Index Terms*—Meta learning, few-shot learning, prompt tuning, domain adaptation, language model**

I. Introduction

Traditionally, deep neural networks were trained by supervised learning and have achieved desirable performance in a variety of speech and language tasks [1]. Recently, self-supervised or unsupervised schemes have been successfully developed to improve the end performance even without additional training data. A famous example is the bidirectional encoder representations from transformers (BERT) [2] which is a pre-trained language model [3], [4] based on transformer [5] learned from a large corpus based on the unsupervised tricks consisting of the masked language modeling and the next sentence prediction. PLMs have been further extended to boost the performance of challenging tasks by enhancing the model capacity with very limited controlled parameters through utilizing the adapter learning [6] or model reprogramming [7]. With the emerging of the powerful generative pre-trained transformer like GPT-3 [8] or ChatGPT, it becomes crucial to conduct the cutting-edge research on how to utilize the pre-trained model for various downstream tasks via prompt-based learning or tuning. In general, the previous PLMs did not handle the issue that specific domain of input sentences is unseen and far from those in training sentences. The performance of PLMs will drop drastically. This phenomenon known as the domain shift is common in practice [9], [10]. The larger the domain shift, the higher the influence on the end performance. In addition, collecting a large amount of labeled data over a variety of domains is difficult and time-consuming. The unknown domain is hard to predict.

To activate fast learning in multi-domain language modeling, this study focuses on zero-shot or few-shot domain adaptation for prompt-based learning where there is few data or even no training data provided in target domain. This paper further copes with the issue of sub-optimal performance in domain adaptation which is caused by inconsistent objectives in both pre-training and fine-tuning stages [11], [12]. In order to address the aforementioned issues, this work proposes the so-called meta soft prompt to enhance the adaptability of PLM to an incoming target domain. In addition to domain adaptation [13], this method is designed by considering two learning paradigms. The first one is the soft prompt learning which converts the actual word tokens in hard prompt into some real-valued vectors as the continuous tokens in augmented input sentence. Prompt optimization is based on gradient updates by using the labeled data [14], [15]. Compared to the hard prompt, soft prompt looks more attractive since both specific domain expertise and trial-and-error process can be avoided [16], [17], [18]. The second one is the model-agnostic meta learning (MAML) [19], [20] which is performed to address the adaptability of soft prompt across multiple domains. A gradient-based meta learning algorithm [21], [22] is implemented to estimate the parameters of meta learner through a set of task-specific learners for individual learning over various meta-training tasks. In this paper, each individual meta-training task is treated as an unsupervised domain adaptation problem which is tackled by learning the soft prompt from source-domain labeled data as well as target-domain unlabeled data. After completing the meta training across diverse meta-training tasks, the trained meta soft prompt is feasible to smoothly adapt to any incoming target domain with a few unlabeled data. In the experiments, the competitive result is obtained by using the proposed prompt-based language model for multi-domain sentiment classification.

II. Soft Prompt & Domain Adaptation

A. Prompt-based language model

Pre-trained language models (PLMs) have shown convincing results in recent years. To leverage the benefit of PLM, the task-specific layer or head was added on top of PLM to adjust the PLM parameters to a specific downstream task. More recently, prompt-based learning is recognized as a powerful method to improve the fine-tuning of PLM after GPT-3 [8]

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

has been publicly released. This method allows training a model with a cloze-style input sentence which adds a textual string prompt to original sentence that has some unfilled slots. The resulting language model needs to predict the word for unfilled slots instead of predicting the class label. Due to the success of GPT-3, such a learning scheme was further introduced to employ smaller PLMs [16], [17], [18] by fine-tuning the pre-trained model. Using this approach, hard prompt is formed by a string of "discrete" word tokens from the vocabulary. But, determining a suitable prompt for a specific domain or task requires the domain expertise and needs many trial-and-errors. Although hard prompt template is fixed, the PLM and the masked language model (MLM) head should be fine-tuned for a downstream task. Alternatively, the soft prompt was constructed in the continuous space with good adaptability. Since the soft prompt $(v_1 v_2 \ldots)$ can be easily updated by gradient descent, the optimal prompt is obtained in a handcraft-free way. Using soft prompt, the parameters of PLM and MLM head are frozen. Only the soft prompt tokens are estimated. Relative to fine-tuning and hard prompt, the number of controlled parameters in soft prompt is greatly reduced. Figure 1 compares the hard (left) and soft (right) prompt-based language models.

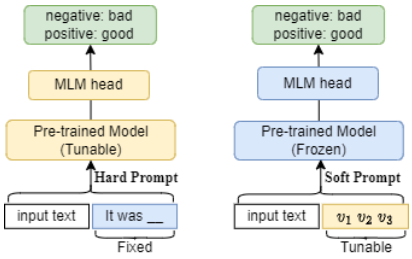


Fig. 1. Hard and soft prompts using pre-trained models.

B. Domain adaptation

In general, fine-tuning, hard prompt or soft prompt using PLM only works well for a specific test domain, and could not easily handle the diverse input sentences which may come from a variety of domains in practical scenarios. As a result, the researches on unsupervised domain adaptation (UDA) [23], [24] have become crucial for handling multi-domain tasks. UDA aims to mitigate the issue of domain shift caused by different distributions from source and target data where the class labels of adaptation data in target domain are missing. Previously, UDA was employed in BioBERT [25] where the pre-training phase was continued by adaptation using domain-specific data. In [26], UDA was performed by simultaneously running the domain-specific pre-training as well as task-specific pre-training by using task-specific unlabeled data [27]. The learned representation was closer to task distribution. In [28], UDA through language model was recently proposed and implemented by combining MLM task and downstream task where the model robustness and the sample efficiency were improved during the adaptation to target domain. Nevertheless, there is existing a difference between the pre-training objective for PLM task and the fine-tuning objective for downstream task, which may constrain the utilization of knowledge based on PLM [11]. This study introduces UDA to adapt the soft prompt across different domains via meta learning in accordance with a consistent hybrid objective for domain adaptation and prompt learning.

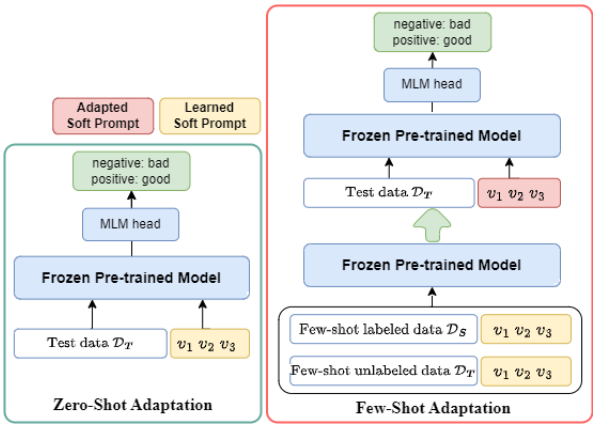


Fig. 2. Soft prompt language models with zero-shot and few-shot domain adaptation where the frozen PLM is used.

C. Joint soft prompt and few-shot domain adaptation

In this paper, soft prompt learning is collaborated with unsupervised domain adaptation where zero-shot and few-shot settings are considered as illustrated in Figure 2 where the frozen PLM is utilized and the labeled data in target domain $\mathcal{D}_T$ are missing. The left-hand-side subfigure shows the zero-shot adaptation where there is no additional adaptation data. Only the test sentence in target domain is provided. Soft prompt is updated by using test data. On the other hand, the right-hand-side subfigure displays the scenario of joint soft prompt estimation and few-shot domain adaptation. In this scenario, there are two stages where the frozen PLM is both applied. The first stage is to learn soft prompt where the labeled data in source domain $\mathcal{D}_S$ and the few-shot unlabeled data in target domain $\mathcal{D}_T$ are provided for prompt-based learning. Then, the second stage is to further adapt the soft prompt by using the test sentence in target domain. To cope with multi-domain tasks, such an adaptive soft prompt is hereafter estimated through meta learning.

III. Few-Shot Learning & Meta Soft Prompt

In practical situation, the input query is usually originated from various domains. It becomes crucial to develop a general solution to soft prompt across different domains and present a fast adaptation scheme to a new domain where multi-domain tasks are handled. The underlying idea of this paper is to enhance the adaptability of soft prompt language model to a new target domain where meta learning or learning to learn is performed to acquire multi-domain knowledge from a number of tasks which are designed to solve the problem on hands.

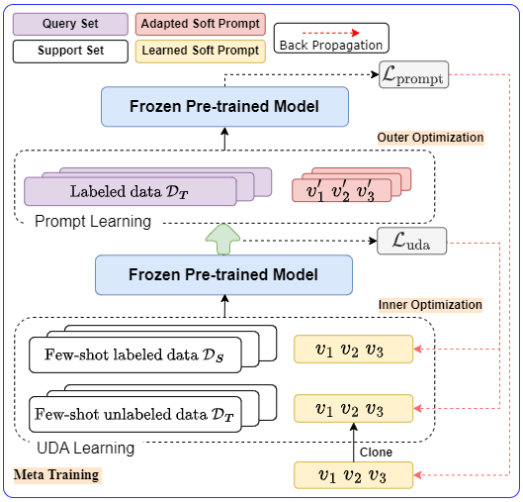


Fig. 3. Multi-domain learning for meta soft prompt language model via few-shot unsupervised domain adaptation (UDA).

A. System overview and meta learning

This study presents the model agnostic soft prompt for few-shot domain adaptation. An overview of the proposed method is shown in Figure 3. There are three parts in the learning process including input data, soft prompt and PLM using BERT. Input data contain *support set* and *query set*. In meta training, *support set* is used to update the prompt parameters of learner (v_1, v_2, v_3). *Query set* is then used to calculate the gradients for updating the parameters of learner (v_1, v_2, v_3) followed by updating the parameters of meta learner (v_1', v_2', v_3'). Soft prompt is treated as the trainable parameters, and is appended to an input sequence and a mask token denoted by [MASK]. PLM is responsible to encode the input sentence and extract the features of mask token. These features are used to calculate MLM head outputs which are employed to measure the probability distribution over vocabulary words to implement the meta soft prompt language model. Unlike traditional meta learning approaches [19], which train the whole model from a large number of few-shot classification problems to allow the model to quickly adapt to the unseen few-shot classification tasks, the proposed meta learning framework aims to tackle domain shift issue by introducing soft prompt learning conditioned on a frozen PLM. All of the meta-training tasks are designed to simulate the scenarios which will be encountered during test stage. Therefore, the frozen PLM is feasible to generalize multiple source domains to a target domain by only tuning the parameters of soft prompt.

B. Soft prompt language model

Soft prompt language model adopts a soft prompt template consisting of a set of trainable vectors that are added to the input sentence as the description of target task. The label words are defined as the highest probability words that PLM will predict. In a binary sentiment classification task, the input sentence is formed by L words as $\mathbf{x} = \{x_1, \cdots, x_L\}$ and the corresponding label is $y \in \mathcal{Y} = \{\text{positive}, \text{negative}\}$. Given the template function $T(\cdot)$, the input sentence $\mathbf{x}$ can be converted to a MLM input to BERT-based PLM, $\mathbf{x}_{\text{prompt}} = T(\mathbf{x} = \{e(\mathbf{x}), v_1, \cdots, v_k, e([\text{MASK}])\})$, where the $e(\cdot)$ is the embedding function of PLM $\mathcal{F}$. For the label words, a verbalizer $\mathcal{M} : \mathcal{Y} \mapsto \mathcal{V}^*$ is given to map the label space to a set of label words $\mathcal{V}^* \subset \mathcal{V}$, where $\mathcal{V}$ is the vocabulary of $\mathcal{F}$. Then, we can treat $\mathcal{F}$ as a function of mapping $\mathbf{x}_{\text{prompt}}$ to a vocabulary distribution of mask token as $p(\mathcal{V}_y(\mathbf{x}_{\text{prompt}})) = \mathcal{F}(\mathbf{x}_{\text{prompt}}) \triangleq \mathbf{v}_{\text{mask}}$. The conditional likelihood of the predicted label word $y^* \in \mathcal{V}_y^*$ with respect to mask token given $\mathbf{x}_{\text{prompt}}$, $p(\mathcal{V}_y^* \leftarrow [\text{MASK}]|\mathbf{x}_{\text{prompt}}, \theta)$, is yielded by

$$p(y^*|\mathbf{x}_{\text{prompt}}, \theta) = \frac{\exp(\mathbf{v}_{\text{mask}}(\mathcal{V}_y^*))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{v}_{\text{mask}}(\mathcal{V}_{y'}^*))} \tag{1}$$

where $\mathbf{v}_{\text{mask}}(\mathcal{V}_y^*)$ denotes the probability of label word $\mathcal{V}_y^*$ in the vocabulary distribution $\mathbf{v}_{\text{mask}}$. The overall training objective for soft prompt learning is expressed by a cross-entropy loss as

$$\mathcal{L}_{\text{prompt}}(\mathbf{x}, y, \theta) = -\sum_{y' \in \mathcal{Y}} y' \log p(y^*|\mathbf{x}_{\text{prompt}}, \theta) \tag{2}$$

where θ denotes the trainable parameters for soft prompt.

C. Unsupervised domain adaptation

Unsupervised domain adaptation as shown in Figure 3 implements the inner optimization that adapts the soft prompt language model to a target domain. This study presents an approach to UDA by combining soft prompt learning and masked language modeling using labeled data from source domain $\mathcal{D}_S$ and unlabeled data from target domain $\mathcal{D}_T$. The soft prompt (v_1, v_2, v_3) in UDA is updated by using the objectives for masked language model $\mathcal{L}_{\text{mlm}}$ (via cross-entropy loss for predicting mask tokens) and soft prompt $\mathcal{L}_{\text{prompt}}$ where the unlabeled data $\mathcal{D}_T$ and labeled data $\mathcal{D}_S$ are adopted, respectively. The adapted soft prompt is able to simultaneously capture the target domain information from MLM objective and the task language knowledge from prompt objective. UDA objective is built as $\mathcal{L}_{\text{uda}} = \mathcal{L}_{\text{prompt}} + \lambda \mathcal{L}_{\text{mlm}}$ with a hyperparameter λ. α is implemented via the popular optimizer using AdamW.

D. Nested optimization procedure

The overall learning and adaptation for meta soft prompt are performed via a nested loop. In the inner loop, the learners start with the parameters of meta learner and update the parameters based on $\mathcal{L}_{\text{uda}}$ by using *support set* in each meta-training task. In the outer loop, the meta learner is optimized according to the performance of those learners evaluated on *query set* in each meta-training task. The whole process is repeated until reaching convergence. Parameter updating for learners is expressed as

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\text{uda}}(\mathcal{D}_i^{\text{sup}}, \theta), \quad 1 \le i \le m. \tag{3}$$

$\mathcal{D}_i^{\text{sup}}$ represents the *support set* of meta-training task $\mathcal{T}_i$. Meta learner $\widehat{\theta}$ is estimated by minimizing the prompt loss over

Algorithm 1: Meta soft prompt learning & adaptation

Require: $\mathcal{D}$: multiple source and target domains
$\mathcal{D}_{\text{train}} = \{\mathcal{T}_i\}_{i=1}^m$: a set of training tasks from $p(\mathcal{T})$
α: step size of updating soft prompt
Initialize: soft prompt parameters θ
repeat
  for *each meta-training task* $\mathcal{T}_i \in \mathcal{D}_{train}$ **do**
    sample $\mathcal{D}_S \sim \mathcal{D}$ as source domain
    sample $\mathcal{D}_T \sim \{\mathcal{D} - \mathcal{D}_S\}$ as target domain
    sample *support set*
      $\mathcal{D}_i^{\text{sup}} = \{\mathbf{x}_S^j, y_S^j\}_{j=1}^{k_s} \cup \{\mathbf{x}_T^j\}_{j=1}^{k_t}$
    sample *query set* $\mathcal{D}_i^{\text{qry}} = \{\mathbf{x}_T^j, y_T^j\}_{j=1}^{k_q}$
    adapt θ to each domain θ_i' using *support set*
      $\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{uda}}(\mathcal{D}_i^{\text{sup}}, \theta)$
    calculate gradient of $\mathcal{L}_{\text{prompt}}$ using *query set*
      $g_i \leftarrow \nabla_{\theta_i'} \mathcal{L}_{\text{prompt}}(\mathcal{D}_i^{\text{qry}}, \theta_i')$
  end
  update soft prompt $\widehat{\theta} \leftarrow \theta - \alpha \sum_{i=1}^m g_i$
until *training converged*

individual learners θ_i'. Meta learning is performed across various meta-training tasks $\mathcal{T}_i$ sampled from $p(\mathcal{T})$

$$\widehat{\theta} = \arg\min_{\theta} \sum_{\mathcal{T}_i} \mathcal{L}_{\text{prompt}}(\mathcal{D}_i^{\text{qry}}, \theta - \alpha \nabla_\theta \mathcal{L}_{\text{uda}}(\mathcal{D}_i^{\text{sup}}, \theta)). \quad (4)$$

$\mathcal{D}_i^{\text{qry}}$ represents the *query set* of $\mathcal{T}_i$. The overall learning and adaptation of meta soft prompt is shown in Algorithm 1. This procedure is a kind multi-task learning over various domains.

IV. Experiments

The proposed meta soft prompt language model was evaluated for multi-domain text classification by using the FDU-MTL dataset [29] which is known as a challenging dataset consisting of 16 domains, as shown in Table III, which were broadly categorized into Amazon product reviews and movie reviews. In the evaluation, one domain was chosen as target domain, and the remaining domains were seen as source domains. A binary sentiment classification with the classes of positive and negative for reviews was performed. Classification accuracy was reported in different sets of evaluation.

A. Evaluation for latent visualization

First, the goodness of meta soft prompt language model for unseen domain data is evaluated. Figure 4 compares 2-D visualization using t-SNE [30] for latent representation of the mask tokens by using the prompted data in 'books' reviews. Basically, the mask tokens using the proposed language model are separate in the classes of positive and negative reviews, and are domain-matching in source and target domains. In the experimental setting, the soft prompt length 5 was used and λ=0.8 was fixed.

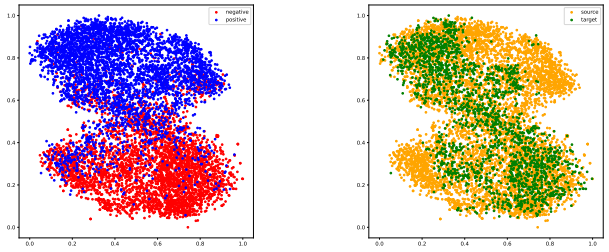


Fig. 4. Visualization on reviews of 'books' target domain using the proposed meta soft prompt. (left) Positive and negative reviews are in blue and red, and (right) source and target domain reviews are in orange and green, respectively.

TABLE I
Accuracy (%) on 4 domains of FDU-MTL dataset. Number in brackets shows the soft prompt length (S).

Domain	**FT**	**SP** (2)	**MSP** (2)	**SP** (5)	**MSP** (5)
books	80.8	**85.2**	83.6	86.8	**88.0**
DVD	79.3	83.2	**83.4**	84.4	**85.6**
electronics	79.4	82.1	**82.6**	84.8	**85.1**
kitchen	79.5	84.5	**86.2**	86.1	**87.6**

B. Evaluation for document classification

The proposed meta soft prompt (MSP) is further compared with traditional fine-tuning (FT) method and the previous soft prompt (SP) [14]. The experiments on review or document classification were conducted on four domains: books, DVD, electronics and kitchen. For a fair comparison, the number of learned parameters were set identically in all methods. For the conventional fine-tuning, the number of trainable parameters in the task-specific layer 768×2 was chosen. As a consequence, the length of soft prompt (S) was set to 2 as each prompt was encoded as a 768-dimensional word embedding. The case S=5 is also evaluated. The accuracy is shown in Table I which indicates that the proposed MSP outperforms FT and SP in various domains especially when the soft prompt length S is increased to 5.

TABLE II
Accuracy (%) and number of trainable parameters (N) for the methods with/without UDA using S=10 with 4 or 8 shots. MoE-Tr [10] requires a large-scaled trainable model.

Domain	MoE-Tr	SP	MSP	SP-4	MSP-4	MSP-8
books	90.0	87.5	**88.6**	87.9	88.7	**89.0**
DVD	89.3	86.2	**86.9**	87.0	**88.5**	88.1
electronics	90.6	87.4	**88.4**	87.9	89.2	**90.3**
kitchen	90.8	88.5	**89.8**	89.2	90.5	**90.7**
UDA	yes	no	no	yes	yes	yes
N	264M	7.68K	7.68K	7.68K	7.68K	7.68K

In Table II, the comparison is extended to a recent work called the transformer-based multi-source domain adaptation (MoE-Tr) [10], which introduced a mixture of experts by using multiple trainable PLMs in adaptation where the number of trainable parameters is N=264M. MoE-Tr performed the domain adversarial learning [31] where many shots were used. Instead, MSP has N=7.68K and only adopts one single frozen

PLM. Unsupervised domain adaptation works for SP and MSP. The improvements by increasing shot number and soft prompt length are obtained.

TABLE III
ACCURACY (%) ON 16 DOMAINS OF FDU-MTL DATASET.

Domain	MT-DNN	ASP-MTL	MAN-L2	MAN-NLL	BERT	MSP
books	82.2	84.0	87.6	86.8	87.0	**89.0**
DVD	84.2	85.5	88.1	**88.6**	85.6	88.1
electronics	81.7	86.8	87.4	88.8	88.3	**90.3**
kitchen	80.7	86.2	89.8	89.9	91.0	**90.7**
apparel	85.0	87.0	87.6	87.6	90.0	**92.0**
camera	86.2	89.2	**91.4**	90.7	90.0	90.8
health	85.7	88.2	89.8	89.4	88.3	**91.3**
music	84.7	82.5	85.9	85.5	86.8	**87.8**
toys	87.7	88.0	90.0	90.4	**91.3**	90.8
video	85.0	84.5	89.5	**89.6**	88.0	88.4
baby	88.0	88.2	90.0	90.2	**91.5**	91.3
magazine	89.5	92.2	92.5	**92.9**	92.8	90.2
software	85.7	87.2	90.4	90.9	89.3	**90.9**
sports	83.2	85.7	89.0	89.0	90.8	**91.8**
IMDb	83.2	85.5	86.6	87.0	85.8	**88.3**
MR	75.5	76.7	76.1	76.7	74.0	**80.4**
AVG	84.3	86.1	88.2	88.4	88.1	**89.5**

Next, the proposed MSP language model is compared with the previous multi-task learning methods over various domains of FDU-MTL dataset. The experiments were setup by choosing one domain as the target for testing and the remaining domains as the sources for training. After the meta training on different source domains is done, the learned meta soft prompt is adapted to a target domain by using 8 shots of unlabeled reviews. The adapted meta soft prompt is collaborated with the frozen PLM to conduct evaluation on the target domain. The results are shown in Table III. These results are consistently compared with strong baselines including the multi-task deep neural network (MT-DNN) [32], the adversarial multi-task learning (ASP-MTL) [29], the multinomial adversarial network with the least square loss (MAN-L2), the negative log-likelihood loss (MAN-NLL) [33], and the BERT model [2] which is fine-tuned on each domain. ASP-MTL [29] used abundant unlabeled samples from the target domain. Instead, MSP only adopted 8 unlabeled samples for domain adaptation. The results reveal that MSP obtains the highest averaged accuracy in most of domains.

V. CONCLUSIONS

This paper has presented the meta soft prompt language model with few-shot domain adaptation. The learned meta soft prompt was appended to input data and adapted to different domains by using the frozen pre-trained model with few-shot unlabeled samples in target domain. The results have shown that the meta soft prompt could successfully boost a frozen pre-trained model to capture domain-specific information and achieved desirable results by only training a few parameters. For future work, the proposed method could be integrated with other unsupervised domain adaptation in the inner optimization for meta learning. Furthermore, the proposed method is feasible to collaborate with not only masked language model but also sequence-to-sequence model or autoregressive model.

REFERENCES

[1] J.-T. Chien, "Deep Bayesian natural language processing," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2019, pp. 4171–4186.

[3] J.-T. Chien and C.-H. Chueh, "Latent dirichlet language model for speech recognition," in *Proc. of IEEE Spoken Language Technology Workshop*, 2008, pp. 201–204.

[4] J.-T. Chien, "Hierarchical Pitman–Yor–Dirichlet language model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1259–1272, 2015.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. of International Conference on Machine Learning*, 2019, pp. 2790–2799.

[7] C.-H. Yang, Y.-Y. Tsai, and P.-Y. Chen, "Voice2series: Reprogramming acoustic models for time series classification," in *Proc. of International Conference on Machine Learning*, 2021, pp. 11808–11819.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[9] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware BERT for cross-domain sentiment analysis," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2020, pp. 4019–4028.

[10] D. Wright and I. Augenstein, "Transformer based multi-source domain adaptation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 7963–7974.

[11] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," *arXiv preprint arXiv:2103.10385*, 2021.

[12] L.-J. Yang, I-P. Yeh, and J.-T. Chien, "Low-resource speech synthesis with speaker-aware embedding," in *Proc. of International Symposium on Chinese Spoken Language Processing*, 2022, pp. 235–239.

[13] W. Lin, M.-W. Mak, and J.-T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

[14] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[15] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5203–5212.

[16] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proc. of Conference of European Chapter of the Association for Computational Linguistics*, 2021, pp. 255–269.

[17] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2339–2352.

[18] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proc. of International Joint Conference on Natural Language Processing*, 2021, pp. 3816–3830.

[19] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. of International Conference on Machine Learning*, 2017, pp. 1126–1135.

[20] M.-Y. Chen, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, "Meta learning for domain agnostic soft prompt," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[21] J.-T. Chien and W. X. Lieow, "Meta learning for hyperparameter optimization in dialogue system," *Proc. of Annual Conference of International Speech Communication Association*, pp. 839–843, 2019.

[22] J.-T. Chien and W. Lai, "Variational skill embeddings for meta reinforcement learning," in *Proc. of International Joint Conference on Neural Networks*, 2023, pp. 1–8.
[23] J.-T. Chien and Y.-Y. Lyu, "Partially adversarial learning and adaptation," in *Proc. of European Signal Processing Conference*, 2019, pp. 1–5.
[24] J.-T. Chien and C.-W. Huang, "Stochastic adversarial learning for domain adaptation," in *Proc. of International Joint Conference on Neural Networks*, 2020, pp. 1–7.
[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
[26] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2020, pp. 8342–8360.
[27] X. Han and J. Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 4238–4248.
[28] C. Karouzos, G. Paraskevopoulos, and A. Potamianos, "UDALM: unsupervised domain adaptation through language modeling," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2579–2590.
[29] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2017, pp. 1–10.
[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[31] J.-C. Tsai and J.-T. Chien, "Adversarial domain separation and adaptation," in *Proc. of International Workshop on Machine Learning for Signal Processing*, 2017, pp. 1–6.
[32] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2015, pp. 912–921.
[33] X. Chen and C. Cardie, "Multinomial adversarial networks for multi-domain text classification," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 1226–1240.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

MSDF-Net: A Multi-Scale Deep Fusion Network with Dilated Convolutions for Cloud Removal from Sentinel-2 Imagery

Jayakrishnan A*, Venkatesan M†, Prabhavathy P‡ and Alkha Mohan §
* Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal-609609, Puducherry, India E-mail: jaykrizz@gmail.com
† Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal-609609, Puducherry, India E-mail: venkisakthi77@gmail.com
‡ School of Information Technology, Vellore Institute of Technology, Vellore-632014, Tamil Nadu, India E-mail: pprabhavathy@vit.ac.in
§ Department of Computer Science and Engineering, Indian Institute of Information Technology Tiruchirappalli Tiruchirappalli-620012, Tamil Nadu, India E-mail: alkham@iiitt.ac.in

***Abstract*—Advancements in satellite imaging catalysed exploratory leaps in Earth observation activities. The high spatial and temporal resolution of remotely sensed imagery aid land use land cover identification , cropland monitoring, change detection and disaster management. However, keeping the quality of these observations is still a challenge. The presence of obstacles, such as clouds and aerosols, usually impairs observations from space, and this is not desirable for computer-based applications as they rely hugely on the available data. The cloud removal problem has been studied for decades, but the randomness in the presence and type of cloud made reconstruction tasks challenging. Fusing Synthetic Aperture Radar (SAR) auxiliary information with optical imagery guided the cloud reconstruction problem. The proposed work introduces a Multi-Scale Deep Fusion Network (MSDF-Net) that learns the spatial and temporal dependency from seen samples and removes clouds from multispectral Sentinel-2 observations. The network extracts physical properties from the SAR auxiliary data and aids the cloud-free reconstruction. The proposed method shows its capability to remove both thin and thick clouds. The network is trained, tested, and validated on the SEN12MS-CR dataset, a global real cloud-removal dataset consisting of triplets of Sentinel-2 cloudy, Sentinel-2 cloud-free and Sentinel-1 SAR data. The network is compared against state-of-the-art techniques and found to exhibit superior performance.**

I. Introduction

The spatial-spectral data gathered from remote sensing satellites play a significant role in earth observation applications such as natural resource management, land use land cover classification, weather forecasting etc. The main challenge faced by space-borne optical frequency imaging is the presence of abundant and random clouds in the earth's atmosphere. Thick clouds block the ground observation entirely, whereas thin clouds allow partial visibility. In any case, these anomalies in reflectance products affect the quality and quantity of temporal observations. There is no guarantee that the subsequent observations can provide cloud-free data because of the uncertainty of cloud presence in the earth's atmosphere. Statistical studies on 12 years of Moderate Resolution Imaging Spectroradiometer (MODIS) observations reveal that the global cloud distortion is 67%, whereas landmass cloud distortion is 55% [1]. The spatial data analysis of the Brazilian Amazonian Landsat archive from 1984 to 1997 also uncovered that cloud cover degrades the accuracy of land observations and complicates biogeophysical understandings. The increasing need for quality temporal land observations accelerated the efforts to design an optimal cloud removal technique.

Spatial filtering was introduced as the preliminary technique for removing clouds from optical data. [2] studied the atmospheric cloud's spatial-frequency properties, resulting in the design of various Homomorphic Filtering (HF) schemes for removing clouds from Landsat data. A modified moving averaging method that replaces cloud pixels with the average value of cloud-free ensemble points was proposed [3]. The Harmonic ANalysis of Time Series (HANTS) algorithm perform Fourier analysis to generate the cloud mask and performs temporal interpolations with the aid of remaining observations to generate a cloud-free composite [4].Another advancement in the research's approach was the idea of image enhancement for quality improvement [5]. The filtering techniques are based on the assumption that the land cover change is negligible for subsequent satellite revisits.

Cloud removal methods based on classical machine learning approaches, such as Regression Trees, Expectation Maximization (EM), Support Vector Machines (SVM), Support Vector Regression (SVR), and Multioutput Support Vector Regression (MSVR), Independent Component Analysis (ICA), came into the picture [6][7][8]. However, these techniques failed to accommodate spectral characteristics. The spatially weighted linear regression on spatiotemporally selected similar pixels succeeded in integrating complementary information from spectral bands [9]. The shortcomings of Haze-Optimized Transformation (HOT) to discriminate haze over bright land surfaces were resolved by the Iterative HOT (IHOT) based on multivariate regression [10]. A Noise-Adjusted Principal Component Transform (CR-NAPCT) outperformed all the existing

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

PCA and ICA techniques [11].

Inpainting approaches consider global and local neighbourhood relationships to restore the deteriorated portions of a given image. Early inpainting techniques based on Bandelet transform [12] and Dictionary-based learning [13] only study spatial relationships. Later the inpainting technique incorporating the spatial, spectral and temporal dependencies evolved [14]. The selection of the most suitable pixels for gap filling is the main challenge of inpainting methods and can be tackled with the help of auxiliary data [15].

The emergence of deep learning enabled the analysis of massive volumes of remotely sensed data, and several automated cloud detection and removal methods have been successively introduced. The Deep residual symmetrical concatenation network (RSC-Net) with an encoder-decoder architecture successfully removed haze from Landsat-8 OLI observations [16]. Multi-spectral Generative Adversarial Networks (GANs) and Attention-based GANs surpass conventional GANs in encoding spatial dependencies and channel relationships [17] [18].However, robust deep learning on spectral data is challenged by the issues of limited samples and higher dimensionality [19][20][21].

The longer wavelength characteristics of microwave sensors make SAR imagery resistant to clouds, dust and adverse climatic scenarios such as heavy rainfall and cyclones . Advancements in multi-sensor data fusion techniques enabled scientists to extract complementary spatial-spectral information from SAR and optical data [22]. [23] came up with a unified Spatial-Temporal-Spectral Convolutional Neural Network (STSCNN). They fused multisource information but failed to extract SAR-guided physical properties. An extension of existing conditional GANs (cGANs) performed cloud removal by learning the corresponding SAR data and temporal cloud-free optical data [24]. The first encoder-decoder architecture fusing Sentinel-1 SAR and Sentinel-2 multispectral optical data has an architecture of U-Net with three encoders and one decoder [25]. Pixel-Level merging of temporal optical sequence and SAR observations performed using a hybrid ConvLSTM+cGAN network [26]. They secured a better pixel-wise restoration of cloudy data.

The state-of-the-art deep learning based data fusion assisted cloud removal techniques include c-GAN [27], Simulation-Fusion GAN [28], DSen2-CR [29], and GLF-CR [30]. The deep learning methods in literature use different variations of convolutional neural networks (CNN) but fail to extract information from all available spectral dimensions at multiple scales. These techniques are trained and validated on simulated or limited datasets, which questions the model's generalizability in addressing real-world scenarios. GANs based models in the literature are complex to train and have the identical drawback of generalizability. This research paper addresses existing techniques' shortcomings and proposes a fusion based multi-scale deep convolutional neural network for multispectral cloud removal. The significant contributions of this work are:

1) Introduces a multi-scale SAR-optical fusion-based deep CNN that extracts non-linear features for automated cloud removal.
2) The dilated convolutions in the proposed model increase the receptive field without any added overhead on computational complexity and extract intrinsic features
3) The proposed cloud-constrained loss function derived from novel cloud map evaluates loss by giving weightage to cloudy pixels and improving model fitness

The remainder of the paper is organised as follows. The proposed multi-scale SAR-optical fusion-based deep convolutional neural network is explained in Section II, and the experimental setup and results are covered in Section III. Finally, the article is summarised in Section IV with insightful remarks.

II. Proposed Methodology

A. Dilated Convolution

One major drawback of CNNs is the failure to perform global processing and extract contextual information from the scene. The solution is to use larger filters or stack up more hidden layers. Both these techniques increase the total number of trainable parameters and impose overheads in computational complexity and generalization capability. The dilated convolution inflates the kernel, expands the receptive field according to the dilation rate, and helps capture more contextual information. The output y_{mn} of a 2D dilated convolution with a dilation rate of d on input x and kernel k is given in (1).

$$y_{mn} = f\left(\sum_{r}\sum_{i=0}^{h-1}\sum_{j=0}^{w-1} k_{ij}x_{(m-d*i)(n-d*j)} + b_{mn}\right) \quad (1)$$

The dilation operation causes exponential growth in the receptive field in each network level by a power of 2 ($d = 2^l$, where l is the level of network). For example, if $d = 4$, the dilations will be performed on two levels. The receptive field of dilated convolution at different levels is derived by (2), where $l_0 = 1$ by default.

$$l_i = l_{i-1} + (k-1)d \quad (2)$$

B. Cloud-Shadow Map Generation

Segregating the cloud and shadow regions of an image is critical for the proposed cloud reconstruction methodology. The generated Cloud-Shadow Map(CM) guides the cloud removal method to reconstruct only the cloud-shadow-affected regions actively. The cloud and shadow map are found separately and later merged to produce a combined cloud-shadow map . The physical and reflectance properties of cloud and shadow are exploited to generate a cloud-shadow map.

Cloud map generation: Cloud map is generally constructed considering High-Reflectance, Normalized Difference Moisture Index (NDMI), and Normalized Difference Snow Index (NDSI) characteristics [31]. Initially, all pixels of the image are considered cloudy and cloud map is initialized to 1 ($\Psi = 1$). Subsequently, the update procedure verifies whether each pixel is covered by clouds or not. The cloud map construction is based on the min-max normalized Sentinel-2 TOA data. Cloud

has high reflectance, and hence cloudy pixels usually appear bright. These high reflectance characteristics are extracted from Coastal aerosol, Blue, Green, Red and Cirrus bands of Sentinel-2 data. The high reflectance-based update Ψ_{HR} is computed for each pixel by sequentially replacing the previous cloud map with the values derived from (3).

$$\begin{aligned} \Psi &= min(R_{\downarrow}(B2)_{[0.1:0.5]}, Cloud\ Map) \\ \Psi &= min(R_{\downarrow}(B1)_{[0.1:0.3]}, \Psi) \\ \Psi &= min(R_{\downarrow}(B10)_{[0.5:0.7]}, \Psi) \\ \Psi_{HR} &= min(R_{\downarrow}(B2+B3+B4)_{[0.2:0.8]}, \Psi) \end{aligned} \tag{3}$$

The existing cloud map generation methods consider only NDMI as an index to approximate water contents and produce reliable cloud identification over a complete vegetation area. However, this approximation fails for diverse regions of vegetation and water bodies. This research work incorporates NDWI as an alternative index, which uses the GREEN-NIR combination to effectively identify cloudy pixels in such mixed regions [32][33]. The proposed NDWI-based cloud map update is given in (4).

$$\Psi_{NDWI} = min\left(R_{\downarrow}\left(\frac{B3-B8}{B3+B8}\right)_{[-1:1]}, \Psi_{HR}\right) \tag{4}$$

The NDMI index is considered for cloud detection for moist pixel approximation. NDMI-based cloud pixel evaluation is carried out as follows.

$$\Psi_{NDMI} = min\left(R_{\downarrow}\left(\frac{B8-B11}{B8+B11}\right)_{[-0.1:0.1]}, \Psi_{NDWI}\right) \tag{5}$$

Snow-based assumptions are also crucial for discriminating cloudy pixels. The effect of condensed water particles in the cloud reflected in NDSI, and NDSI-based cloud map update is done as follows.

$$\Psi_{NDSI} = min\left(R_{\downarrow}\left(\frac{B3-B11}{B3+B11}\right)_{[0.8:0.6]}, \Psi_{NDMI}\right) \tag{6}$$

The rescaling operation $R_{\downarrow}$ performs a pixel remapping and is defined by the rescaling in the range [l, u]. Equation (7) remaps the pixel value Ψ in the [l, u] range.

$$\Psi_{R_{\downarrow}} = \frac{\Psi - l}{u - l} \tag{7}$$

The updated cloud map undergoes opening and closing morphological operations with a 3×3 structuring element for accurate cloud filling. The opening operation removes any isolated high reflectance responses from highly bright surfaces such as building rooftops, whereas the closing operation fills the random gaps and cracks within high bright cloud map regions. The resulting cloud map undergoes rescaling to the range [0,1] and a binary thresholding with a threshold of $T_a = 0.2$ to produce the final cloud map. The final cloud map update is given as

$$\Psi = \begin{cases} 1, & if \quad \Psi > T_a \\ 0, & otherwise \end{cases} \tag{8}$$

Shadow map generation: The procedure takes the same min-max normalized Sentinel-2 TOA image as the input. The low reflectance characteristics and low NDVI characteristics are considered to generate a shadow map. The proposed shadow map generation algorithm distinguishes the cloud shadow from other materials by utilizing Coastal aerosol, NIR, SWIR-1 and SWIR-2 bands [31][34]. The algorithm identifies low reflectance pixels by thresholding ($T_b = 0.75$) the average value of bands B1, B8, B11 and B12 . These low-reflectance pixels are further discarded from the shadow map if they exhibit lower NDVI values ($T_c = 0.8$), assuming those are dark pixels generated due to water presence. The thresholds T_a, T_b and T_c are selected from the empirical cloud detection studies conducted for Sentinel-2 multispectral data [34]. A closing morphological operation is performed to close gaps in shadows and generate a final shadow map. The NDVI computation and shadow map update Φ are derived from (9) and (10).

$$NDVI = \frac{B8-B4}{B8+B4} \tag{9}$$

$$\Phi = \begin{cases} 1, & \left(\frac{B1+B8+B11+B12}{4}\right) < T_b\ AND \\ & NDVI < T_c \\ 0, & otherwise \end{cases} \tag{10}$$

The cloud map and shadow map undergo pixel-wise OR operation to produce a combined cloud-shadow map . Fig. 1 portray the flowchart for the complete cloud-shadow map generating process.

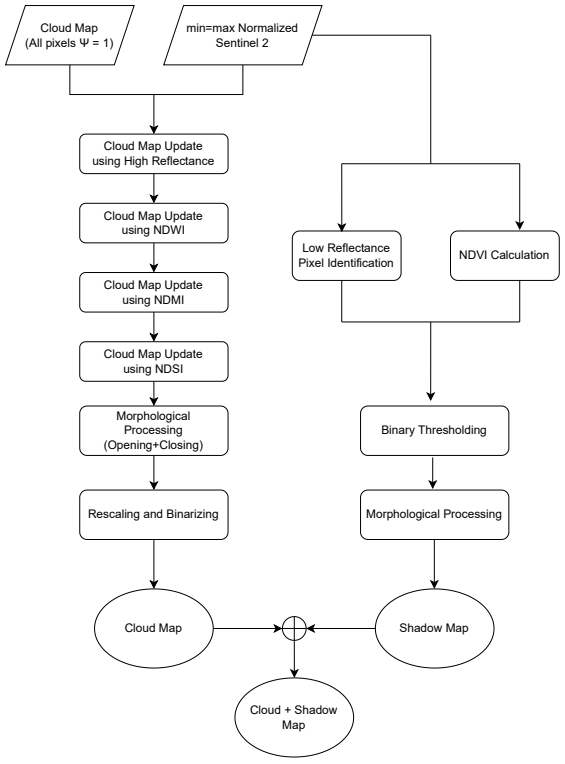


Fig. 1: Combined Cloud-Shadow Map Generation Process

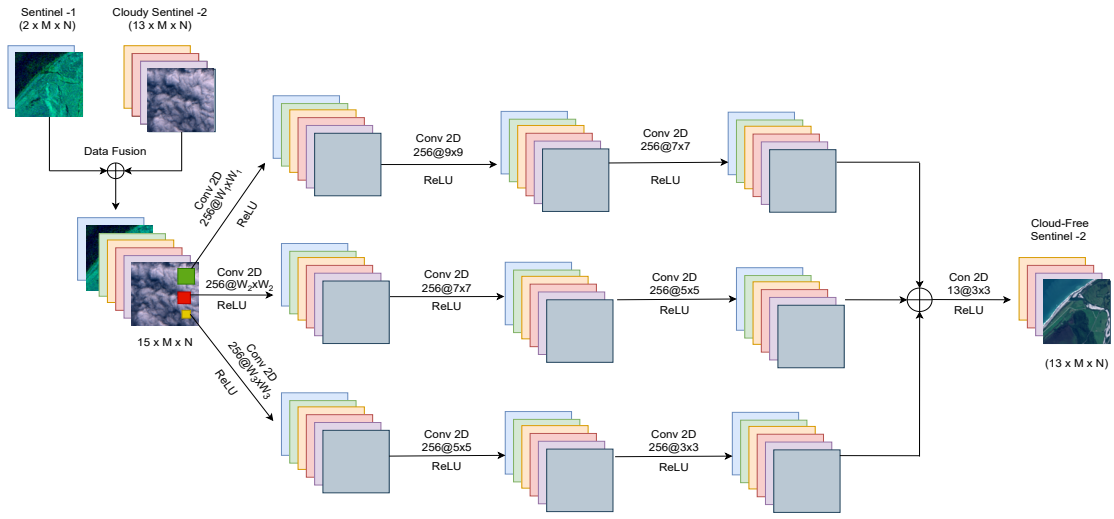


Fig. 2: Proposed MSDF-Net Cloud Removal Architecture

C. Cloud Constrained Loss Function

L_1 loss is a robust loss function commonly used for regression and reconstruction problems [35]. The learning algorithms try to minimize the L_1 loss, which is the total of all absolute differences between the actual and reconstructed pixel values. The L_1 loss function is given in (11), where X, Y stand for the reference and reconstructed images, whereas N denotes the total number of observations.

$$L_1 = \frac{\sum_{i=1}^{n} |X_i - Y_i|}{N} \tag{11}$$

However, the L_1 loss alone cannot guarantee a reliable reconstruction because the L_1 loss is computed for the total image without giving any weightage for the cloud and shadow-affected regions. The uncloudy reference observations are usually captured in temporal revisits of the satellite, and there is no guarantee that the landscape will remain the same during the revisit. Therefore, retaining all available cloud-unaffected information of the optical input data is critical. Since the reconstruction happens only for the cloud-shadow-affected region, an explicit regularized novel loss function comprising the previously generated cloud-shadow map is proposed. The proposed cloud-constrained loss function L_{ccl} is defined as

$$L_{ccl} = \frac{\sum_{i=1}^{n} CM \cdot |X_i - Y_i|}{N} + \frac{\sum_{i=1}^{n} |X_i - Y_i|}{N} \tag{12}$$

The first part of L_{ccl} selectively evaluates loss by giving weightage to cloudy pixels and improve model fitness . The second part of the equation is the actual L_1 loss, that act as a regularization term that deal with overfitting.

D. Proposed Model

The auxiliary information in the form of SAR guides the cloud reconstruction process by providing useful texture and geophysical information. The proposed MSDF-Net cloud removal utilizes the Sentinel-1 SAR information as the complimentary data for missing region reconstruction. Both Sentinel-1 and Sentinel-2 data are 3D cubes of sizes $B \times M \times N$, where B represents the number of bands and M, N denotes the spatial dimensions. A channel-wise concatenation of the Sentinel-1 VV, VH bands with all the available 13 bands of Sentinel-2 is performed to produce a fused input of size $15 \times M \times N$ and given as input to the model. Multi-scale feature extraction is the core concept of the proposed model, where each convolution branches extract feature maps using variable window sizes. The model is designed to have three convolution branches with varying window sizes, W_1, W_2, and W_3, respectively. Each convolution branch performs two additional sequential 2D convolutions. Branch-1 perform the 2D sequential convolutions with kernels of sizes 9×9 and 7×7. Branch-2 engage convolution kernels of sizes 7×7 and 5×5, whereas 5×5 and 3×3 convolutions are carried out in Branch-3. The model is designed to extract 256 feature maps at each convolution and employs Rectified Linear Unit (ReLU) as an activation function. The same padding technique at each convolution layer ensures consistent spatial resolution through out the reconstruction. The feature maps obtained from the individual multi-scale branch of the model are concatenated to generate a final 756 (3×256) feature map. A cloud-free 13 band Sentinel-2 image is reconstructed from this final feature map using a 3×3 2D convolution with 13 filters. The model is trained to minimize the cloud-constrained loss function L_{ccl}. Fig. 2 depict the detailed architecture of the proposed model.

III. Results and Discussions

A. Experimental Setup and Evaluation parameters

All experiments and model training are carried out in NVIDIA DGX A100 GPU SERVER consisting of eight A100 GPUs with a total of 320GB memory. The proposed model is tested on a large-scale all-season all-region multi-model uni-temporal dataset SEN12MS-CR [36]. The saturated values may happen due to clouds as well as highly elective objects. In the experiment they are treated as equals and the Sentinel-2 data is restricted to a range [0,10000] that avoids saturated values,

whereas the Sentinel-1 undergo a value clipping in the range [-25.0, -32.5] to avoid invariant pixels and noise anomalies. . The dataset is split in the ratio of 70:15:15 to generate the train test validation sets. The training triplets undergo data normalisation to the range of [0,1] before passing to the proposed deep fusion model and thereby reducing computational complexity. Data augmentation and shuffling were performed to mitigate any bias prospect and improve generalizability. The spatial resolution of the input patches is set to 256×256, that is, $M = N = 256$. The multi-scale model operates at different branch window sizes and is set to $W_1 = 11 \times 11$, $W_2 = 9 \times 9$, and $W_3 = 7 \times 7$. The model performs dilated convolutions with dilation rate d chosen from the set $1, 2, 4, 8, 16$, where $d = 1$ indicates regular convolution. The hyperparameters, such as batch size and learning rate, are fixed to 16 and $7e-6$.The quantitative evaluation of predicted cloud-free images uses various Pixel-wise and spectral-structural closeness metrics [37].

B. Result Analysis

The proposed MSDF framework is compared against four different state-of-the-art data fusion-based cloud removal techniques, including c-GAN, Simulation-Fusion GAN, DSen2-CR, and GLF-CR. Fig.3 visualizes the cloud-constrained loss obtained during the training and validation phase of the proposed model. The validation loss shows how well the model fits new data, whereas the training loss shows how well it fits training data. It is evident that there is little variation between the validation and training curves, and as a result, the model is reliable and generalizable.The loss curve The validation accuracy spiked over the training accuracy during the initial stages of training. However, as the training progressed, it was observed that the validation loss reduced and settled in a steady state with training accuracy. At epoch 198, the model exhibited a validation loss lower than the training loss. The training is stopped at this point resulting in an optimal model. The model comparison of various evaluation parameters such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio), SAM (Spectral Angle Mapper) and SSIM (Structural Similarity Index) for state-of-the-art techniques against the proposed method are tabulated in Table I.

TABLE I: Comparison of evaluation parameters of proposed MSDF-Net against state-of-the-art

Methods	MAE	MSE	PSNR	RMSE	SAM(°)	SSIM
c-GAN	**0.0441**	0.1074	25.29	0.3278	14.43	0.7594
Simulation Fusion GAN	0.0455	0.0886	24.55	0.2978	15.59	0.6947
DSen2-CR	0.0841	0.0182	28.70	0.1352	8.04	0.8750
GLF-CR	0.0821	0.0219	29.07	0.1481	7.64	0.8855
MSDF-Net (d=4)	0.0778	**0.0163**	**32.15**	**0.1280**	**4.42**	**0.9205**

The proposed model with a dilation rate of 4 surpassed all the state-of-the-art techniques under comparison. As per the results, c-GAN scored a minimal value for MAE but

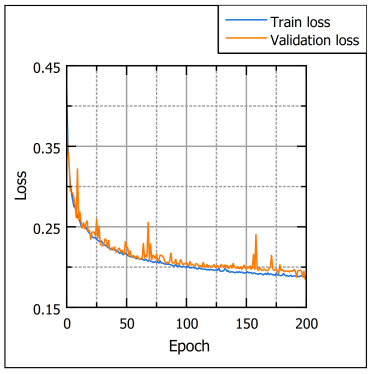


Fig. 3: Characteristic training and validation loss curve for the proposed MSDF-Net

failed to achieve credible scores for other evaluation metrics. Compared with c-GAN, all other state-of-the-art techniques exhibit superior performance except for MAE. GAN models demonstrated their capacity to reconstruct with simulated data, while real-world observations with unpredictable clouds and noise presented difficulties for GAN-based reconstruction. The proposed cloud removal framework demonstrate a lower MAE score than the recently released DSEN2-CR and GLF-CR models. The DSen2-CR model employed residual connection to achieve a better reconstruction in terms of PSNR, SAM, and SSIM. However, the residual connections increased the number of learnable parameters and failed to extract multi-scale features. The GLF-CR model employed separate global and local feature extraction networks, outperforming DSen2-CR in terms of PSNR, SAM and SSIM but scoring higher MSE value. Implementing dilated convolution in the proposed model extracts more global features without a notable increase in the number of learnable parameters. The multi-scale feature extraction procedure incorporates global and local features in a single network structure. Hence the proposed model outperformed all the state-of-the-art techniques by scoring higher PSNR, SSIM and lower SAM. Fig. 4 and 5 show the subjective evaluations of the proposed model against state-of-the-art techniques for different regions under different cloud conditions. The subjective evaluation of the results in Fig. 4 and Fig. 5 indicates that the model works for thin and thick cloud scenarios and maintains the scene's physical, texture, and colour details.

1) Effect of SAR Auxiliary Information in Reconstruction: The proposed model is trained in two settings, one fusing actual cloudy with SAR auxiliary information and the other without auxiliary information. This ablation study shed light on understanding the need and effect of SAR in cloud-free reconstruction. Fig. 6 illustrates the effect of SAR data in reconstructing cloudy images taken over diverse land regions . The penetration power of SAR through clouds aids in extracting physical information, including boundaries, textures and patterns and helps in better reconstruction. A quantitative comparison of various evaluation parameters with and without

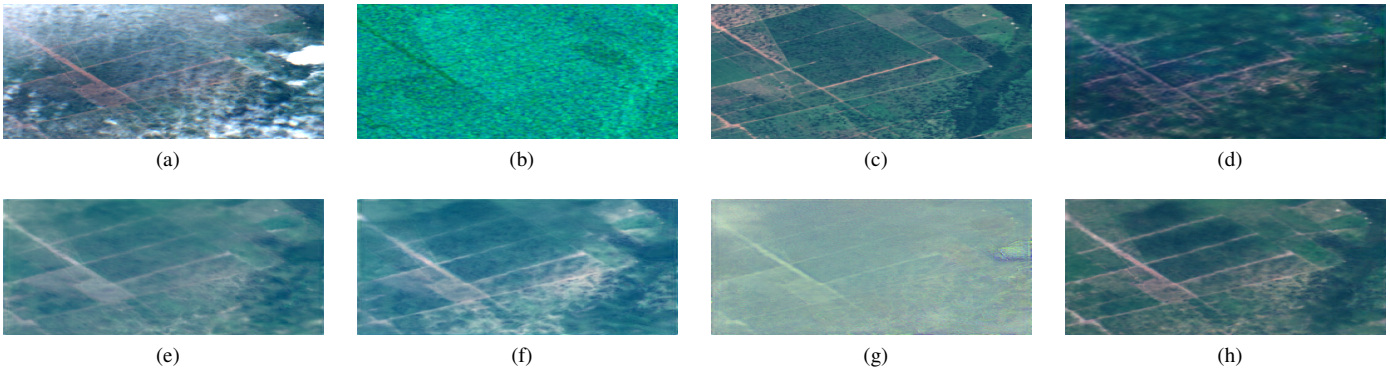


Fig. 4: Subjective comparison of proposed model with state-of-the-art techniques for an agricultural area under thin cloud, (a) cloudy input, (b) SAR auxiliary data, (c) cloud free observation, (d-g) reconstructed output for c-GAN, Simulation Fusion-GAN, DSen2-CR, GLF-CR, (h) cloud free reconstruction from proposed MSDF-Net

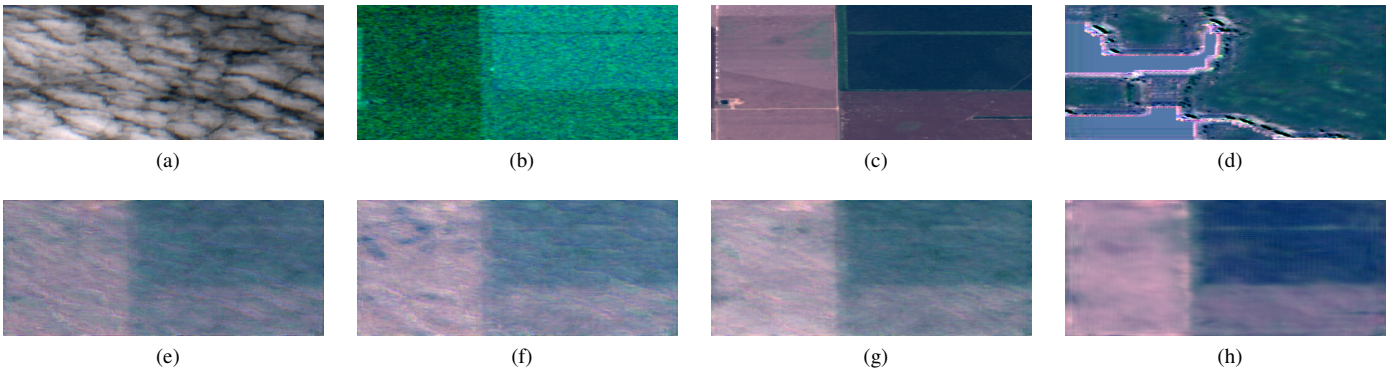


Fig. 5: Subjective comparison of proposed model with state-of-the-art techniques for for an area under thick cloud (a) cloudy input, (b) SAR auxiliary data, (c) cloud free observation, (d-g) reconstructed output for c-GAN, Simulation Fusion-GAN, DSen2-CR, GLF-CR, (h) cloud free reconstruction from proposed MSDF-Net

TABLE II: Comparison of evaluation parameters for proposed MSDF-Net with and without auxiliary data

Model Setting	MAE	MSE	PSNR	RMSE	SAM (°)	SSIM
Without SAR	0.1011	0.0269	30.03	0.1643	7.65	0.8864
With SAR	**0.0778**	**0.0163**	**32.15**	**0.1280**	**4.42**	0.9205

SAR supplementary information is given in Table II. While comparing the values in table, it is clear that reconstruction using proposed deep fusion network without SAR is poor and exhibits high MAE. The higher value of SAM and lower SSIM indicate the minimal spectral similarity compared to the model with auxiliary data. The subjective evaluation of Fig. 6 for various cloud conditions also supports the same.

2) Selection of Dilation Factor: Selecting a proper dilation factor is essential, and there is no guarantee that higher dilation rates can achieve better reconstruction. A higher dilation rate may fail to collect local characteristics by giving more importance to global features. A lower dilation rate, however, succeeds in extracting local neighbourhood characteristics but fails to do so for global features. Hence an empirical evaluation and selection of the correct dilation factor is mandatory. The proposed model is tested under various dilation rates ($d = 1, 2, 4, 8, 16$) and chosen $d = 4$ by analysing evaluation parameters. As the dilation rate further increased from 4, the performance started degrading. Table III tabulates the effect of different dilation rates on evaluation parameters.

TABLE III: Comparison of evaluation parameters for various dilation rate in proposed MSDF-Net

Dilation Setting	MAE	MSE	PSNR	RMSE	SAM(°)	SSIM
d=1	0.0779	0.0185	31.6	0.1362	4.57	0.9175
d=2	0.0823	0.0178	31.79	0.1337	4.54	0.9182
d=4	**0.0778**	**0.0163**	**32.15**	**0.1280**	**4.42**	**0.9205**
d=8	0.0804	0.0171	31.96	0.1309	4.52	0.9172
d=16	0.091	0.0213	31.02	0.1460	5.04	0.9083

IV. Conclusions

The presence of clouds in the atmosphere shrinks satellite observation quality and impacts numerous ground applications. This research article introduced a reliable multi-scale fusion framework with dilated convolution for cloud removal from

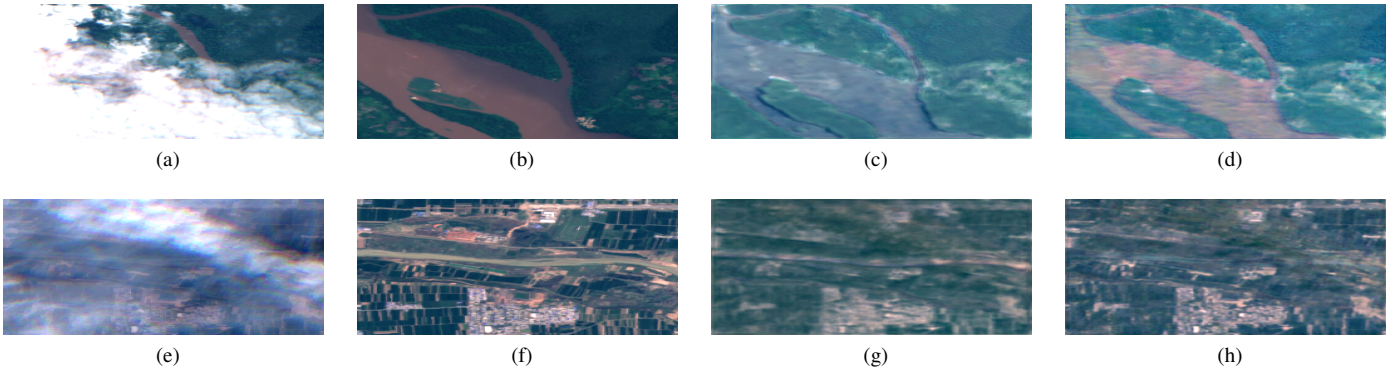


Fig. 6: Subjective comparison of proposed model with and without SAR auxiliary data for different cloud conditions (a,e) cloudy input, (b,f) cloud free observation, (c,g) reconstruction without SAR, (d,h) reconstruction with SAR

Sentinel-2 data. The model fused SAR auxiliary data with optical observations to extract underlying geophysical features. The multi-scale dilated convolutions and the cloud-constrained loss function derived from the novel cloud mask enabled global feature extraction. The quantitative study against various state-of-the-art techniques reveals superior model performance. The subjective analysis revealed the model's potential in reconstructing cloud-free data retaining colour, texture and geometrical fidelity. Also, the model excelled in handling all kinds of thick and thin clouds. This research concretes the need and effect of multi-sensor fusion for cloud removal tasks.

References

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3826–3852, 2013.

[2] O. R. Mitchell, E. J. Delp, and P. L. Chen, "Filtering to remove cloud cover in satellite imagery," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 137–141, 1977.

[3] D. Long, D. Daum, and P. Hardin, "Spatial resolution enhancement of ssm/i data: vegetation studies of the amazon basin," in *IGARSS '96. 1996 International Geoscience and Remote Sensing Symposium*, vol. 3, 1996, pp. 1606–1608 vol.3.

[4] G. J. Roerink, M. Menenti, and W. Verhoef, "Reconstructing cloudfree ndvi composites using fourier analysis of time series," *International Journal of Remote Sensing*, vol. 21, no. 9, pp. 1911–1917, 2000.

[5] F. Albu, C. Vertan, C. Florea, and A. Drimbarean, "One scan shadow compensation and visual enhancement of color images," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 3133–3136.

[6] F. Melgani, "Contextual reconstruction of cloud-contaminated multitemporal multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 2, pp. 442–455, 2006.

[7] G. Hu, X. Sun, D. Liang, and Y. Sun, "Cloud removal of remote sensing image based on multi-output support vector regression," *Journal of Systems Engineering and Electronics*, vol. 25, no. 6, pp. 1082–1088, 2014.

[8] Y. Shen, Y. Wang, H. Lv, and J. Qian, "Removal of thin clouds in landsat-8 oli data with independent component analysis," *Remote Sensing*, vol. 7, no. 9, pp. 11 481–11 500, 2015.

[9] B. Chen, B. Huang, L. Chen, and B. Xu, "Spatially and temporally weighted regression: A novel method to produce continuous cloud-free landsat imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 27–37, 2017.

[10] S. Chen, X. Chen, J. Chen, P. Jia, X. Cao, and C. Liu, "An iterative haze optimized transformation for automatic cloud/haze detection of landsat imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2682–2694, 2016.

[11] M. Xu, X. Jia, M. Pickering, and S. Jia, "Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 215–225, 2019.

[12] A. Maalouf, P. Carre, B. Augereau, and C. Fernandez-Maloigne, "A bandelet-based inpainting technique for clouds removal from remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2363–2371, 2009.

[13] F. Meng, X. Yang, C. Zhou, and Z. Li, "A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery," *Sensors*, vol. 17, no. 9, 2017.

[14] T.-Y. Ji, N. Yokoya, X. X. Zhu, and T.-Z. Huang, "Nonlocal tensor completion for multitemporal remotely sensed images' inpainting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3047–3061, 2018.

[15] Q. Cheng, H. Shen, L. Zhang, Q. Yuan, and C. Zeng, "Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal mrf

model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 92, pp. 54–68, 2014.

[16] W. Li, Y. Li, D. Chen, and J. C.-W. Chan, "Thin cloud removal with residual symmetrical concatenation network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 153, pp. 137–150, 2019.

[17] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1533–1541.

[18] L. Sun, Y. Zhang, X. Chang, Y. Wang, and J. Xu, "Cloud-aware generative network: Removing cloud from optical remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 4, pp. 691–695, 2020.

[19] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, Apr 2023. [Online]. Available: https://doi.org/10.1186/s40537-023-00727-2

[20] A. Mohan and M. Venkatesan, "Hybridcnn based hyperspectral image classification using multiscale spatiospectral features," *Infrared Physics & Technology*, vol. 108, p. 103326, 2020.

[21] ——, "Hybrid dimensionality reduction technique for hyperspectral images using random projection and manifold learning," in *Trends and Applications in Knowledge Discovery and Data Mining*, W. Lu and K. Q. Zhu, Eds. Cham: Springer International Publishing, 2020, pp. 116–127.

[22] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for sar and optical images: A review," *Information Fusion*, vol. 59, pp. 13–29, 2020.

[23] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4274–4288, 2018.

[24] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220–1224, 2019.

[25] R. Cresson, D. Ienco, R. Gaetano, K. Ose, and D. H. Tong Minh, "Optical image gap filling using deep convolutional autoencoder from optical and radar images," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 218–221.

[26] A. Sebastianelli, E. Puglisi, M. P. Del Rosso, J. Mifdal, A. Nowakowski, P. P. Mathieu, F. Pirri, and S. L. Ullo, "Plfm: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[27] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1220–1224, 2019.

[28] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and sar images using generative adversarial networks," *Remote Sensing*, vol. 12, no. 1, 2020.

[29] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[30] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu, "Glf-cr: Sar-enhanced cloud removal with global–local fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 268–278, 2022.

[31] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Aggregating cloud-free sentinel-2 images with google earth engine," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, pp. 145–152, 2019.

[32] X. Li, F. Ling, X. Cai, Y. Ge, X. Li, Z. Yin, C. Shang, X. Jia, and Y. Du, "Mapping water bodies under cloud cover using remotely sensed optical images and a spatiotemporal dependence model," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102470, 2021.

[33] Y. Lin, L. He, Y. Zhang, and Z. Wu, "Cloud detection of gaofen-2 multi-spectral imagery based on the modified radiation transmittance map," *Remote Sensing*, vol. 14, no. 17, 2022.

[34] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 144, pp. 235–253, 2018.

[35] W. Yu, X. Zhang, and M.-O. Pun, "Cloud removal in optical remote sensing imagery using multiscale distortion-aware networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[36] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5866–5878, 2021.

[37] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for sar and optical images: A review," *Information Fusion*, vol. 59, pp. 13–29, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Instance Implant-Aided Non-uniformly Cropping for Person Detection in Aerial Images

Xiangqing Zhang*, Yan Feng†, Shun Zhang† and Yuning Wang†
* Northwestern Polytechnical University, Xi'an, China and Yan'an University, Yan'an, China
E-mail: zxq1991@mail.nwpu.edu.cn
† Northwestern Polytechnical University, Xi'an, China
E-mail: sycfy@nwpu.edu.cn, szhang@nwpu.edu.cn, yuning2023@mail.nwpu.edu.cn

***Abstract*—Human inspection of aerial-based scenarios is a time-consuming, labor-intensive, and error-prone task. However, current aerial-vision-based object detection techniques have not adequately improved the detection performance of persons with sparse and weak features. In this article, a novel approach is proposed to automatically detect persons in aerial images. Specifically, numerous and multiscale instances are first implanted on fresh background images with a copy-paste mechanism to collect rich training data. Then, cropping high-resolution images into irregular sizes with density maps in the training stage. Lastly, hyper-inferencing on the raw and sliced blocks of the image by merging these overlap results to find persons. Experiments were run on the original Heridal datasets against eight approaches of popular object detectors. The metrics of precision, AP^{test}@0.5, and AP^{test}@0.5:0.95 achieved 0.9277 (+2.77%), 0.8012, and 0.547, respectively. Meanwhile, our proposed method is being evaluated on the modified Heridal and a number of publicly available datasets to make a more robust person detector in aerial images.**

I. Introduction

Object detection is a critical research problem in the field of remote sensing and has become a research hotspot. With the remarkable advances in unmanned aerial vehicles (UAVs) and computer performance, UAVs acquire vast amounts of aerial images carrying detailed geographic and graphical information. Such information has significant practical applications in military operations, civilian search and rescue initiatives, monitoring, and inspection[1]. However, high-resolution images present numerous challenges, including difficulties in identifying persons from the background, large-scale backgrounds with various object categories, and limited memory available for training networks. While person detection in fixed indoor or confined monitoring locations has improved significantly and achieved relatively good results, aerial person detection[2], [3], [4] for tiny clues about the missing person's location remains a challenge, currently relying mainly on manual and error-prone human efforts.

Data augmentation is a promising approach to address a few unbalanced samples in single scenes. It generates more representations from the original data, which improves the quantity and quality of the original data without significantly increasing data volume. Copy-paste is a robust data augmentation method that has recently received increased attention[5]. It copies an object from its original location to a new location and uses random transformations to ensure that the pasted objects do not overlap with any existing objects. However, [6] argued that it only copies the pixels that correspond to an object, not all pixels within the object's bounding box. This method places the copied instance randomly without modeling the surrounding visual context. These methods are based on the MS-COCO dataset, and their effectiveness with aerial images remains uncertain.

Many researchers have proposed solutions to address the challenges of dense small objects with uneven distribution in drone-based scenarios. For instance, [7], [8] chunks the original image into overlapping slices for detection and fuses detection results using non-maximum suppression (NMS). However, this method inevitably reduces computational speed. [9] and [10] integrate clustering and density estimation into the image cropping module, which reduces the number of image partitions but hinders portability and deployment. [11] efficiently detects smaller objects without requiring retraining the model or a larger GPU memory allocation. It chunks the images for detection and combines all inferred images into a large graph by NMS, which is time-consuming and suffers from boundary errors in detection.

This article proposes an aerial person detection scheme based on YOLOv5[12] to address the limited training samples, low accuracy, and time-consuming inference. Extensive experiments demonstrate that our algorithm achieves high accuracy and robustness.

In summary, this work has the following contributions:

- The Copy-Paste Mechanism for Object Implantation involves implanting objects from a patches library and the Heridal dataset's background. This method calculates the IoU match in detected pseudo-labels of instances, manually generates object masks, and performs multiple copy-paste to increase the number of tiny persons accurately.
- Non-uniform image cropping involves finding highly concentrated blocks of the original image with instances. Guided by density maps, this method displays the location and density of targets in a given area to create better image blocks for the next stage of object detection.

This research is funded by the National Natural Science Foundation of China (No. 62271409, No.62171381, and No. 62262067), and Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No.23JK0738); Crresponding author: Yan Feng.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

- The YOLOv5 C3 module has been improved to C3TR to account for tiny persons. The loss function of YOLOv5 has been altered, with different weights assigned to the real labels and pseudo-labels.
- Multiple experiments have been conducted on synthetic datasets created based on the optimized Heridal dataset. The proposed method's detection results have also been compared with state-of-the-art methods using publicly available datasets.

This article is organized as follows: Section 2 provides a review of related works and our improvement tips. Section 3 presents our methodology and the experiments conducted with the Heridal dataset in detail. Section 4 reports on the experimental results and analysis. The conclusion is presented in Section 5.

II. RELATE WORKS

A. Anchor-Based Detector

In this work, the proposed method was conducted based on the YOLOv5 framework and gradually was optimized by modular networks of YOLOv5 to improve detection accuracy. The YOLOv5 framework consists of four main components: data preprocessing at the input, feature extraction network for the Backbone, feature enhancement network at the Neck, and the detection head. The feature extraction network comprises CSPDarknet53, which includes Conv modules and C3 modules (our experimental tests show that the C3TR module performs better on the Heridal dataset). A feature enhancement network is added between the feature extraction network and detection head to mine the semantic information within the image better and make it more applicable for classification. The YOLOv5 framework utilizes SPP or SPPF network structure to convert feature maps of arbitrary sizes into fixed-size feature vectors and employs the PANet network structure, which has a top-down bidirectional feature fusion backbone. For predicting the bounding box, the detection head network adopts the YOLOv3 Head structure and assigns three anchors to points on each layer of the image.

B. Improved Loss Function

The YOLOv5 loss function includes three parts: Bounding Box regression loss, object confidence loss, and category loss. Every network separates the feature map into cells that output a vector of [t_x, t_y, t_w, t_h, p_o, c_1, c_2, ...]. t_x and t_y represent the centroid's offset between the prediction box and the corresponding anchor box, while t_w and t_h are employed in computing the box width and height. p_o expresses the chance that the cell (prediction box) contains the object, and c_1, c_2, ..., are the predictions' values for the corresponding categories. All three loss parts are calculated by linking to positive sample pairs, and each output feature map is summed up straight to obtain the final loss of each part. Each network's feature map cells are autonomous of each other.

$$\mathcal{L}_{v,5}\left(\mathbf{t}_p, \mathbf{t}_{gt}\right) = \sum_{k=0}^{K} \left[\begin{array}{l} \alpha_k^{balance} \alpha_{box} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{kij}^{obj} \mathcal{L}_{\text{CloU}} \\ +\alpha_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{kij}^{obj} \mathcal{L}_{\text{obj}} \\ +\alpha_{cls} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{I}_{kij}^{\text{obj}} \mathcal{L}_{cls} \end{array} \right] \quad (1)$$

Where K, S^2, and B in Eq. (1) are the output feature maps, cells, and a number of anchors on each cell, respectively. α_* (α_{box}, α_{obj}, α_{cls}) is the corresponding item's weight, and the default values in hyp.scratch-high.yaml is (α_{box} = 0.05, α_{obj} = 0.3, α_{cls} = 0.7), $\mathbb{I}_{kij}^{obj}$ indicating whether the k output feature map, i cell, and j anchor box are positive samples, with 1 indicating positive and 0 indicating negative. t_p, t_{gt} is the prediction and ground-truth vector; it is used to balance the weights of the output feature maps for each scale, and the default values are [4.0, 1.0, 0.4], which size of feature maps correspond to 80×80, 40×40, and 20×20, respectively.

$$\begin{aligned} \mathcal{L}_{v5}^{\text{total}}\left(\mathbf{t}_p, \mathbf{t}_{gt}\right) = & \beta_{\text{truth}} \mathcal{L}_{v5}^{\text{truth}}\left(\mathbf{t}_p, \mathbf{t}_{gt}\right) + \\ & \beta_{\text{pseudo}} \mathcal{L}_{v5}^{\text{pseudo}}\left(\mathbf{t}_p, \mathbf{t}_{gt}\right) \end{aligned} \quad (2)$$

The total loss is defined as Eq. (2), where weights β_{truth} and β_{pesudo} are used to adjust the number of unreliable pseudo-labels, $\mathcal{L}_{v5}^{\text{truth}}$ and $\mathcal{L}_{v5}^{\text{pseudo}}$ are the same as in Eq. (1).

III. OUR PROPOSED METHOD

Since the Heridal dataset includes only a few instances, direct training from scratch results in slow network convergence and inferior detection performance. Thus, the primary experimental strategy for the YOLOv5 pre-trained model involves fine-tuning the training model on the Heridal dataset.

A. General Framework

As shown in Fig. 1, our method employs different strategies to optimize the algorithms in the training and inference phases. It consists of five main parts: object implantation (green dashed box), image cropping (yellow dashed box), hyperparameter evolution (brown dashed box), slice-assisted detection (purple dashed box), and one-stage detector (blue dashed box). The object implantation module generates pseudo-labels between the Heridal background images and object libraries. The image cropping module is then used to crop blocks of images containing dense regions of persons and add them to the training subset using non-uniform slicing. The detector combines the detection results from cropped and original images using NMS. Moreover, optimizing the pre-trained hyperparameters using a genetic algorithm on the Heridal dataset in the preprocessing stage can significantly improve model training convergence speed and accuracy.

B. Object Implantation with Copy-Paste Augmentation

In Search and Rescue (SaR) scenarios, persons captured by drones often appear in non-urban and remote areas, leading to a scarcity of available real-world samples. The object implantation algorithm aims to increase the number of labeled training samples by generating pseudo-labels in unlabeled images to enhance the predictive performance of the detection

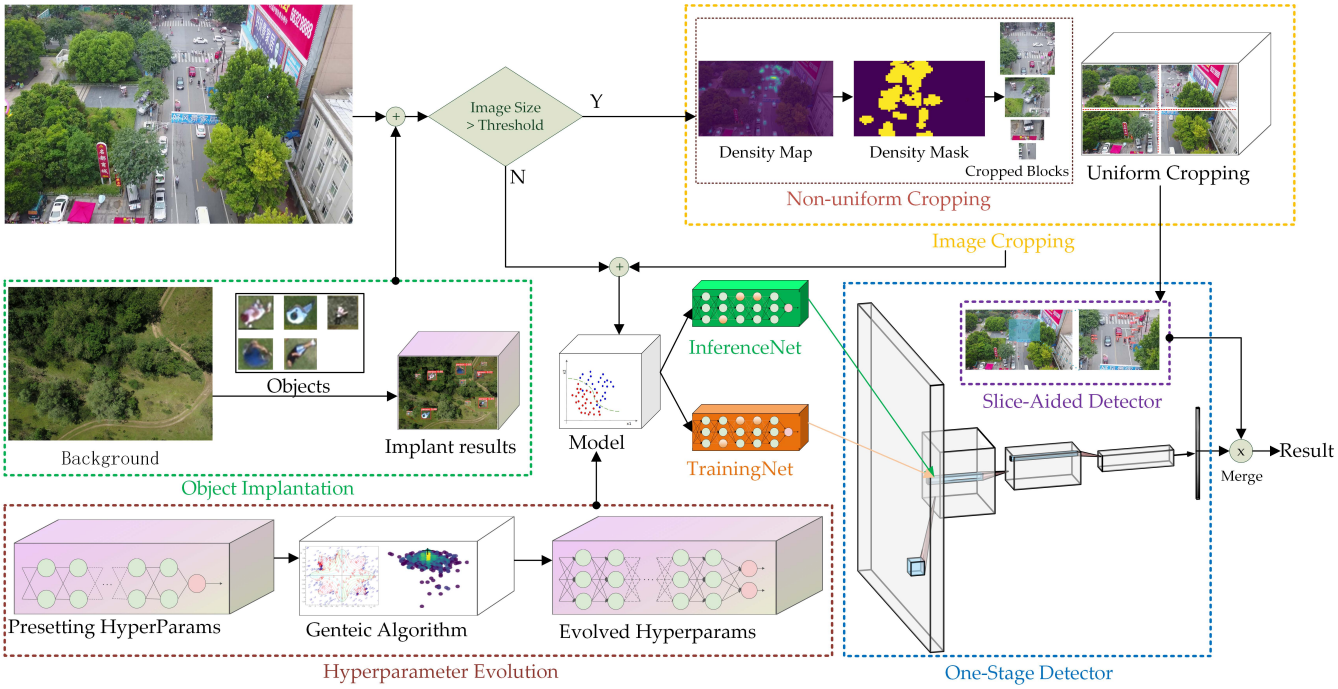


Fig. 1. General framework of our proposed scheme. The main modules have been drawn with different dashed boxes.

model[13]. In this study, a high-confidence detection model is employed to infer unlabeled images and derive pseudo-labels. The detector first infers under the pre-trained model to obtain predicted samples with high confidence, followed by an IoU matching calculation between predicted and real labels. The mask generator then extracts the mask of the pseudo-labels to create an object instance library. Lastly, pseudo-instances are formed by randomly combining objects and backgrounds. The trainer utilizes both real and fake samples to optimize the model's parameters and obtain the best performance.

C. Image Cropping Guided by Density Map

The cropping strategy can potentially increase the training samples; however, random or mosaic cropping cannot guarantee that the cropped region contains valid data, resulting in an imbalance between negative-to-positive sample ratios. As illustrated in Fig. 2, our experiments with high-resolution aerial images (obtained from the VisDrone2019[14] Val subset) used a density estimation network to crop images into non-overlapping blocks before mapping the training labels regression into the cropped blocks. A dense multi-cropping object detector network based on YOLOv5[12] was proposed, which utilizes Density Map Guided Object Detection network ((DMNet)[10] and the Twice Observed Network (YOLT) [7]) to enhance detection accuracy.

The non-uniform cropping strategy employs density maps that generate object-specific densities using MCNN[15] and class-wise-kernel masks. Initially, pixels with a density mask value of 1 are selected, and they are subsequently connected into a large candidate region using an eight-linked region. Finally, the candidate region's outer rectangle is used to crop the original image, but regions with lower resolution than the density value are removed. As illustrated in Fig.2, the DMNet processes the original image to create cropped boxes with numerous positive samples.

D. Slice-assisted Detection

Various factors such as altitude can impact the detection algorithm in UAV-captured images, leading to reduced efficiency in search and rescue operations. To address this, the following approach was proposed:

- Identify the shortest flight pathway to the object centroid by inferring it once on the original map.
- Lower flight altitude or adjust the lens to capture close-range images of the pathway for improved detection accuracy.
- Merge detection results for the region.

To create a dataset of sliced samples, images are sliced to a fixed size using sliding windows with overlap. The Heridal dataset contains few person samples and many background samples following the slicing process. To address this, samples with boundary values during the slicing process are removed. The trainer settings are kept consistent with the original image, but the input size is based on the size of the slices. As demonstrated in Fig. 3, the detection results are combined from both the original and zoomed-in cropped image using the NMS method to obtain the final results.

IV. EXPERIMENTAL RESULTS ANS ANALYSIS

All experimental results in this study were conducted on the NVIDIA Tesla V100 * 2 hardware platforms with 16G×2 memory. YOLOv5 version 6.0 was used, along with CUDA

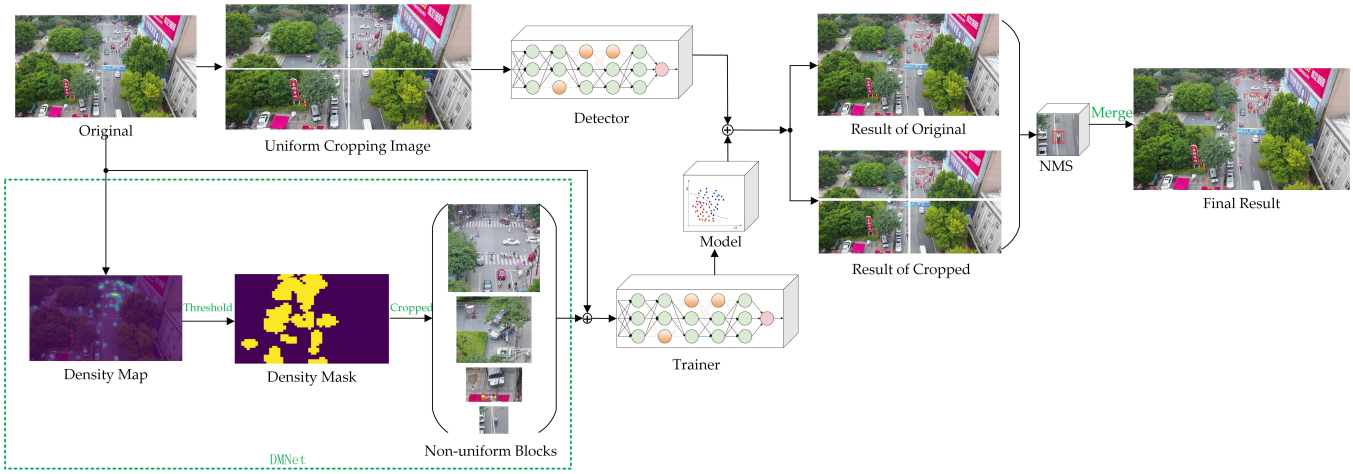


Fig. 2. Cropping of images is done in both the training and inference stages. Density map-guided non-uniform cropping is performed, while uniform cropping of slices without overlapping regions is used during inference. The detection results obtained from the original and cropped images are then combined using NMS to obtain the final detection results. The threshold is also utilized to create a density mask for the connectivity map.

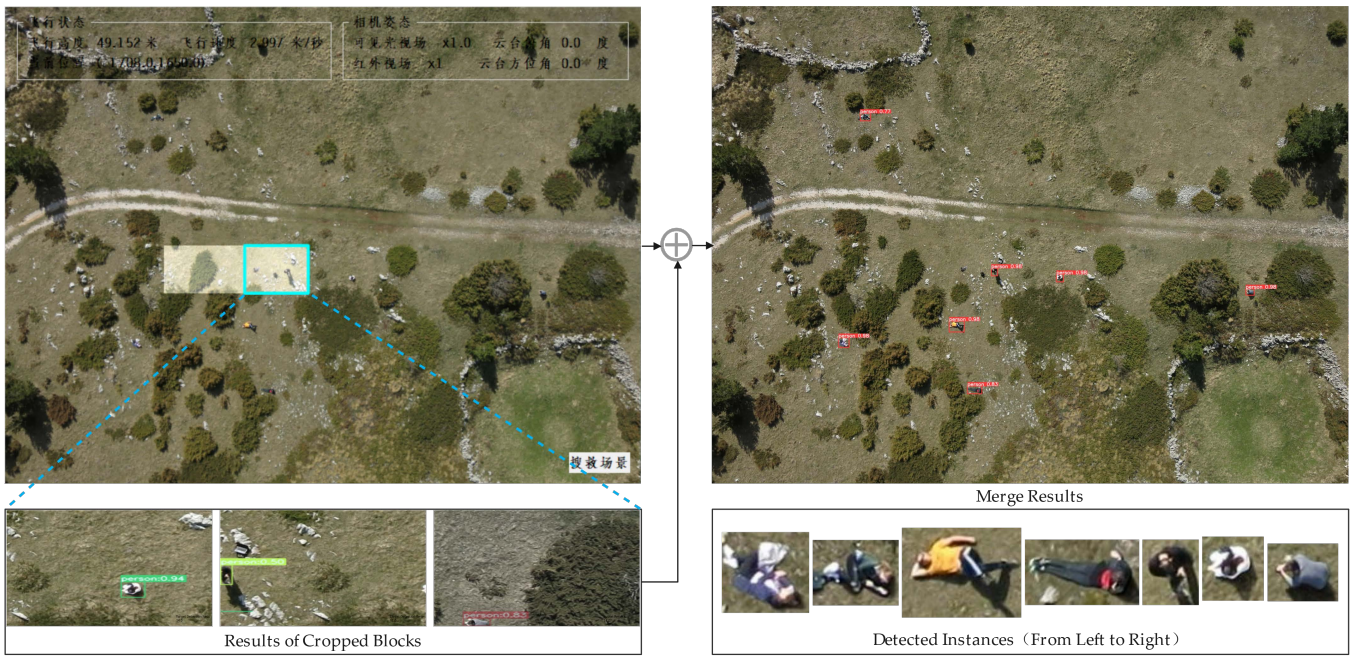


Fig. 3. Slice-assisted for fine-tuning detection results.

version 10.2, SGD as the gradient descent algorithm, and ResNet101 as the Backbone.

A. Dataset and Setting

1) Optimized Hyperparameters: The correct initialization of setting parameters is crucial for producing better final results. As the default parameters of the pre-trained model were evolved on the COCO dataset, they do not well fit SaR scenarios. YOLOv5 has 28 hyperparameters (hyps) for various training settings. To optimize these hyps, a base_hyp is selected to generate mutations based on the trained hyps, and then the previous hyps are weighted based on the results of each previous evolution. Therefore, optimizing hyperparameters is crucial for correctly preprocessing hyps before training.

2) Datasets: The Heridal dataset serves as the primary experimental data for this study. It comprises 1650 high-resolution images (4000×3000) containing instances of persons in various non-urban areas such as mountains, forests, oceans, deserts, and so on. The Department of Electronics and the Polytechnic Institute of the University of Split collected these images using unmanned helicopters. TABLE I provides an overview of the number of training subsets and test subsets. The original dataset was optimized for the experiments and primarily consisted of the following.

- Removing unlabeled background samples and error la-

TABLE I
NUMBER OF SAMPLES ON THE ORIGINAL AND OPTIMIZED HERIDAL DATASETS.

Dataset	Description	Input_Size	Overlap_Size	Numebers of Sample			
				Train_Set	Val_Set	Test_Set	Total
Heridal	Original images	4000x3000	-	1547	0	103	1650
Heridal_1536	Remove background and mislabeled images	1536x1536	-	900	78	101	1079
Heridal_Implant_1536	Embed instances in background	1536x1536	-	1399	150	101	1650
Heridal_Crop_320	Evenly cropped with overlapping	320x320	200x200	8501	793	1175	10469
Heridal_Crop_400	Evenly cropped with overlapping	400x400	200x200	5458	494	773	6725
Heridal_Crop_412	Evenly cropped with overlapping	412x412	200x200	3158	302	436	3896
Heridal_Crop_416	Evenly cropped with overlapping	416x416	200x200	2659	247	358	3002

TABLE II
TRAINING METRICS FOR THE OPTIMIZED HERIDAL DATASET.

Dataset	Precision	Recall	AP^{val} 0.5	AP^{test} 0.5	AP^{val} 0.5:0.95	AP^{test} 0.5:0.95	Params /M	GFlOPs /M	Speedtest(V100 b1/ms)			
									Preprocessing	Inference	NMS	Total_Time
Heridal_1536	0.8262	0.6914	0.7452	0.7306	4587	0.4498	6.7	15.2	1.9	12.6	2.3	16.8
Heridal_Impant_1536	0.8876	0.7023	0.7884	0.7763	0.5205	0.5012	6.7	15.8	1.8	12.6	2.1	16.5
Heridal_Crop_320	0.9512	0.8691	0.9282	0.9088	0.6164	0.5927	6.7	14.3	0.3	7.1	3	10.4
Heridal_Crop_400	0.9479	0.9203	0.9317	0.9201	0.6299	0.6103	6.7	14.3	0.3	6.4	2.1	**8.8**
Heridal_Crop_412	0.9307	0.8816	0.9304	0.9189	0.627	0.6147	6.7	14.3	0.3	6.6	3.8	10.7
Heridal_Crop_416	**0.9724**	**0.9466**	**0.9799**	**0.9548**	**0.7729**	**0.7576**	6.7	14.3	0.3	7.1	4.3	11.7

Bolded fonts represent the best training metric values for various Heridal datasets.

bels.

- Adding samples from the Heridal Patches library randomly to the background, and the labels default to the aspect ratio of the images in the Patches.
- Cropping blocks that contain objects with uniform overlapping sliding windows of sizes 320×320, 400×400, 412×412, and 416×416.

B. Model Training and Analysis

1) Results of Cropped Heridal Dataset: The Heridal dataset initially contained numerous unlabeled background images. To address this, the object implantation method is employed to add samples from the patches' library, resulting in the Heridal_Impant_1536 dataset. All results were obtained using the YOLOv5s model with a batch_size of 32. TABLE II displays the training metrics for the Heridal, Heridal_Impant_1536, and cropped blocks size of samples: [320×320], [400×400], [412×412], and [416×416] datasets. The precision of Heridal_Impant_1536 increased from 0.8262 to 0.8876 (↑ 6.14%), while mAP@0.5 increased from 0.7452 to 0.7884 (↑ 4.32%). Among all the datasets, Heridal_Crop_416 achieved the best training metrics, with precision of 0.9724, recall of 0.9466, mAP@0.5 of 0.9799, and mAP@0.5:0.95 of 0.7729. Despite the decrease in the average inference speed on the test_set by 1.3ms, 2.9ms, and 1ms, respectively, it still meets the real-time detection requirements.

2) Results of Different Pretrained Model: In this experiment, different model depths with various batch sizes were utilized to enhance training speed, and the input size was set to 1536×1536. Fig. 4 displays the training precision, recall, mAP@0.5, and mAP@0.5:0.95 curves for the Heridal dataset on various YOLOv5 models, including s, s6, m, m6, l, l6, x, and x6. Methods 1-8 were performed on the Heridal_1536 dataset using a training strategy that adds non-uniformly cropped blocks and original images to the trainer. Methods 9-16 were performed on the Heridal_Implant_1536 dataset using object implantation. The training metrics were significantly improved on the Heridal_Implant_1536 dataset, with the best metrics being precision (0.9277), recall (0.73), mAP0.5 (0.802), and mAP0.5:0.95 (0.547). When compared to the deepest model YOLOv5x6 (method 16), the shallowest YOLOv5s model (method 8) showed a 4.01% decrease in precision, 1.68% increase in recall, 1.28% decrease in mAP0.5, and 2.65% decrease in mAP0.5:0.95, with the YOLOv5s model only having 15.8M parameters, making it more suitable for deployment on embedded devices.

C. Model Inference and Visualization

The trained model is compiled by TensorRT to create a serialization engine on embedded devices, which includes dynamic batch inference and model quantization. During training, data with different batch sizes can be iterated upon, and input tensor sizes (shapes) are uncertain during inference. Therefore, making full use of multiple batches can significantly improve the model's detection speed. Model quantization involves converting a signal's continuous values into a limited number of discrete values. In training CNNs, the data precision of FP32 is commonly used to represent weights, biases, activation values, and so on. However, previous experience has shown that low precision (FP16, INT8) sacrifices some training accuracy but reduces memory usage and shortens model training time. In this work, the quantization accuracy is set to FP16.

1) Comparative Tests: The results of the detection speed on Jetson Tx2 and Xavier NX are presented in Table III. These tests were conducted at the maximum frequency by default. With the same YOLOv5s model and batch size of 4, Xavier NX's inference speed is 19.1ms faster than the baseline, with nearly identical parameters. Although the quantized model loses about 2-3% accuracy, inference times are significantly

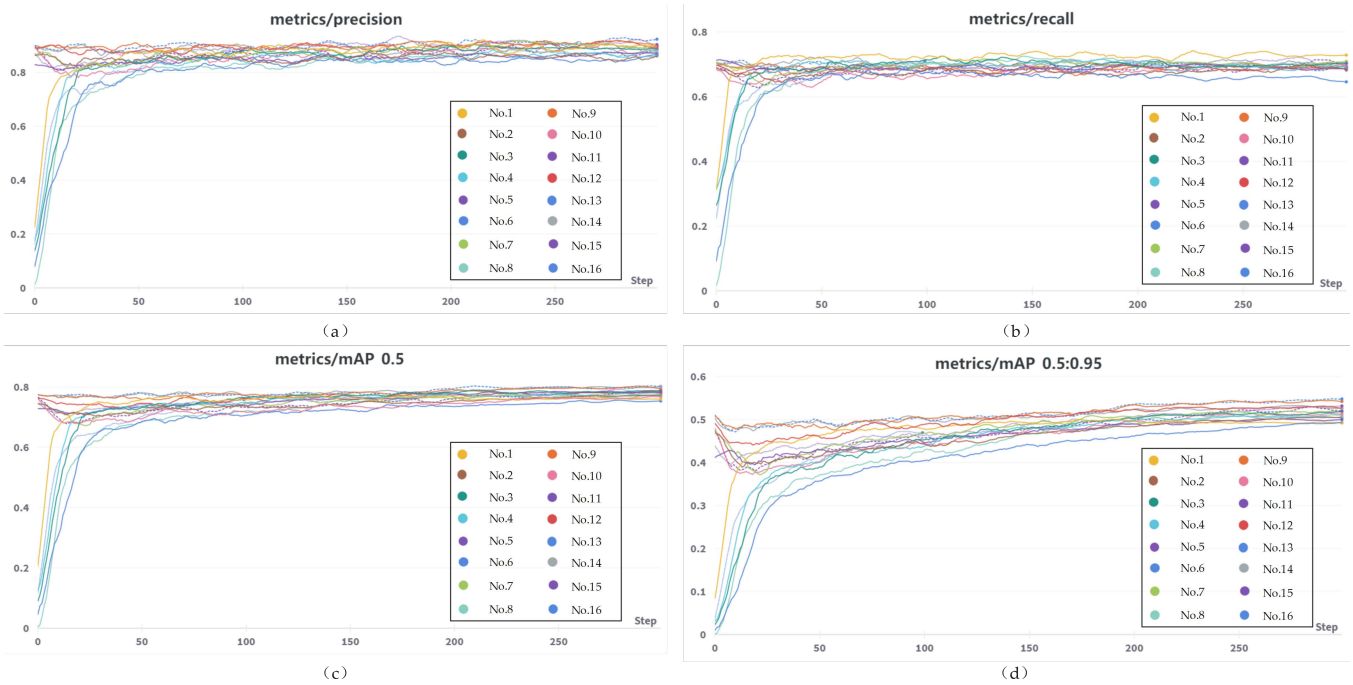


Fig. 4. Different depth models' training curves on the Heridal dataset. (a) Training curves of precision for various depth models. (b) Training curves of recall for various depth models. (c) Training curves of mAP@0.5 for various depth models. (d) Training curves of mAP@0.5:0.95 for various depth models. where No. 1–8 methods based on object implantation with s, s6, m, m6, l, l6, x, and x6 models. No. 9–16 methods based on image cropping with s, s6, m, m6, l, l6, x, and x6 models.

TABLE III
YOLOV5S MODEL TESTING AND INFERENCE METRICS ON VARIOUS DEVICES.

Method	Model	Batch Size	Input Size	APtest@0.5	Params/M	GFLOPs/M	Speedtest (Jetson Tx2 b1/ms)	Speedtest (Xavier NX b1/ms)	Speedtest(V100 b1/ms)			
									Preprocessing	Inference	NMS	Total Time
Baseline	YOLOv5s	1	640x640	0.8366	6.7	15.8	283.4	75.4	0.6	7.2	2.1	9.9
	YOLOv5s6	1	640x640	0.8366	11.5	16.2	363.1	98.6	0.6	11.1	2.2	13.9
Batch Inference	YOLOv5s	2	640x640	0.8366	6.7	15.8	272.9	68.2	0.4	7.9	1.7	10
	YOLOv5s	4	640x640	0.8366	6.7	15.8	265.8	56.3	0.3	7.5	1.6	9.4
Parameter Quantification	YOLOv5s	4	640x640	0.8366 (-0.8183)	4.3	19.8	141.7	26.5	-	-	-	-
	YOLOv5s	4	1024x1024	0.8419 (0.8203)	8.2	24.7	157.5	28.7	-	-	-	-
	YOLOv5s	4	1280x1280	0.8761 (-0.8505)	11.6	28.6	159.6	30.3	-	-	-	-
	YOLOv5s	**4**	**1536x1536**	**0.8876 (-0.8653)**	**17.5**	**32.4**	**162**	**34.2**	-	-	-	-

reduced. Xavier NX achieves real-time detection at 34.2ms when the model input size is 1536×1536.

TABLE IV
THE COMPARISON EXPERIMENTAL RESULTS BETWEEN OUR PROPOSED MODEL WITH OTHER POPULAR ALGORITHMS

Method	Precision	Recall	AP^{test}@0.5/%	AP^{test}@0.5:0.95/%
[16]	34.8	88.9	-	-
[2]	67.3	88.3	-	-
Edge Boxes	26.91	63.5	-	-
Mean Shift	52.76	51.03	-	-
[17]	4.33	94.36	-	-
RPN	41.54	**95.54**	-	-
FPN	76.84	86.64	-	-
[18]	68.89	94.65	-	-
[3]	90.5	87.8	-	-
[19]	90.68	68.14	80.27	53.86
[12]	90.17	67.45	79.01	51.22
[20]	88.72	66.27	78.28	51.45
Ours (best)	**92.77**	68.55	**80.12**	**54.7**

Table IV displays the results of our proposed and existing algorithms on the Heridal dataset. The baseline method was first proposed on the Heridal dataset, but its detection accuracy was only 34.8%, which could not meet the application requirements. In recent years, more complex networks such as Faster RCNN, RetinaNet, and SSD have significantly improved detection accuracy, but at the expense of slower detection speed. In this study, reliable training results are achieved based on YOLOv4, YOLOv5, and YOLOv6, taking into account detection accuracies and inference speeds on embedded devices. As shown in Table IV, the optimal precision was 92.77%, but the recall of the YOLO series was lower, primarily due to the IoU evaluation's unfavorable impact on small object detection, and a slight deviation of the bounding box could result in a marked deterioration of detection performance.

2) Detection Results: Fig.5(a) displays the YOLOv5s detection results on the Xavier NX, indicating that it can effectively detect persons in various jungle scenes. The detection results of the object implantation method on the Heridal_Implant_1536 val_set are shown in Fig.5(b), where synthetic samples of per-



Fig. 5. Detection results of different methods on the Heridal test_set. (a) Detection results of the No. 8 method from Table 4 or Table 5 on the Heridal test set. (b) Detection results of the object implantation method on the Heridal_Implant_1536 val_set. (c) Merge detection results of original images and cropped blocks on the Heridal test_set. To facilitate the target clarity, a small target of each different image sequence was marked and enlarged simultaneously by a red box.

sons generated using copy-paste techniques can be accurately detected. The detection results of merging the original image with sliding windows are shown in Fig.5(c). A detector with a model size of 416×416 on cropped images can compensate for the flaws found in the original images, particularly detecting persons, thereby improving the object detection accuracy.

D. Ablation Study

1) Numbers of Objects based Copy-Paste Mechanism: The positive and negative samples used in this study were obtained from the patch libraries in the Heridal dataset and patches implanted by the Copy-Paste Mechanism to the background. There were 29,050 positive and 39,700 negative instances, all of which were 81×81 pixels. Additionally, the size of synthesized positive and negative instances was set to be less than 32×32 pixels (i.e., small objects in MS-COCO). According to Table V, the number of implanted positive instances varied from 3 to 7, while the number of implanted negative instances was 0. Based on the results, the optimal combination was determined to be [positive = 5, negative = 0], with non-repeated samples and the following performance metrics: [precision = 0.8876, recall = 0.7023, APval@0.5=0.7884, APval@0.5:0.95=0.5205]. These experiments showed that increasing the number of positive instances significantly improved the accuracy of the model in high-resolution images, while adding negative samples repeatedly had little impact on model performance, possibly due to numerous negative instances or backgrounds contained in the Heridal dataset.

2) Sizes of Blocks with Non-uniformly Cropping: DMNet dynamically adjusts the number of objects to cover all regions

TABLE V
THE RELATIONSHIP BETWEEN THE NUMBER OF IMPLANTED POSITIVE OBJECTS AND THE PERFORMANCE OF DETECTION MODELS.

Num of Pos	Num of Neg	Repeat	Precision	Recall	AP^{test}@0.5	AP^{test}@0.5:0.95
3	0		0.8769	0.6917	0.7756	0.5174
3	0	✓	0.8617	0.6822	0.7596	0.5163
5	0		**0.8876**	0.7023	**0.7884**	0.5205
5	0	✓	0.8648	0.6859	0.7573	0.5194
7	0		0.8642	0.6948	0.7007	**0.5327**
7	0	✓	0.8701	0.6856	0.7568	0.5182
5	2		0.8865	**0.7043**	0.7826	0.52
5	2	✓	0.8675	0.6844	0.7521	0.5189

TABLE VI
THE IMPACT OF NON-UNIFORM CROP BLOCK SIZES ON THE PERFORMANCE OF THE DETECTION MODEL

Threshold	Crop_Size	Train_Set	Val_Set	Precision	Recall	AP^{test}0.5
Baseline				0.8262	**0.6914**	0.7452
0.08	320x320	506	87	0.8037	0.6806	0.7356
1	320x320	472	69	0.8112	0.6842	0.7379
5	500x500	203	41	0.8496	0.6867	0.7428
10	500x500	146	29	0.8625	0.6888	0.7516
15	640x640	117	21	0.8709	0.6903	0.7598
20	640x640	97	16	**0.8721**	0.6884	**0.7603**

in the image. As the threshold is increased, the boundary becomes more irregular, and pixels on the boundary are more likely to be filtered out. Additionally, a sliding window is applied to the density map, and its size is set equal to the average size of objects in the training set, while the step size is set to the size of the sliding window (without overlap). The sample size in this study was primarily impacted by adjusting the density threshold and crop window size, as shown in Table VI, and the test results were obtained from the Heridal_Implant_1536 dataset. According to Table VI, the optimal training metrics were achieved when the density threshold was set to 20 and the cropping window size was [320×320]. A lower density threshold resulted in no connected domain and fewer usable samples to crop. Additionally, a crop block that is too small can cause severe deformation of instances when combined with the original images. Thus, selecting appropriate density thresholds and cropping window sizes can significantly impact the model's performance.

V. CONCLUSION

In this study, the Copy-Paste Mechanism for object implantation and a slice-assisted inference approach with multiscale feature fusion was proposed, which achieves better detection performance on the optimized Heridal dataset. Comparative experiments using YOLOv5 demonstrate that our algorithm achieves better average detection precision, with the best metrics being 92.77% for precision, 68.55% for recall, 80.12% for mAPtest@0.5, and 54.7% for mAPtest@0.5:0.95. Additionally, dynamic batch inference and parameter quantization were tested on the embedded devices Jetson Tx2 and Xavier NX, exhibiting promising lightweight results. In the future, our work will involve the design of better pseudo-label generators and more efficient person detectors.

REFERENCES

[1] A. Jain, R. Ramaprasad, P. Narang, M. Mandal, V. Chamola, F. R. Yu, and M. Guizan, "AI-enabled object detection in UAVs: Challenges, design choices, and research directions," *IEEE Network*, vol. 35, no. 4, pp. 129–135, Jul 2021.

[2] D. Božić-Štulić, Ž. Marušić, and S. Gotovac, "Deep learning approach in aerial imagery for supporting land search and rescue missions," *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1256–1278, Mar 2019.

[3] P. Pyrrö, H. Naseri, and A. Jung, "Rethinking drone-based search and rescue with aerial person detection," *arXiv preprint arXiv:2111.09406*, 2021.

[4] T. Niedzielski, M. Jurecka, B. Miziński, W. Pawul, and T. Motyl, "First successful rescue of a lost person using the human detection system: A case study from beskid niski (SE poland)," *Remote Sensing*, vol. 13, no. 23, p. 4903, Dec 2021.

[5] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," in *9th International Conference on Advances in Computing and Information Technology (ACITY 2019)*. Aircc Publishing Corporation, Dec 2019.

[6] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2021.

[7] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," *arXiv preprint arXiv:1805.09512*, 2018.

[8] X. Zhang, Y. Feng, S. Zhang, N. Wang, and S. Mei, "Finding nonrigid tiny person with densely cropped and local attention object detector networks in low-altitude aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4371–4385, May 2022.

[9] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct 2019.

[10] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Jun 2020.

[11] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," *arXiv preprint arXiv:2202.06934*, 2022.

[12] G. J. et. al., "Ultralytics/yolov5: V6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support," Zenodo, 2021.

[13] X. Zhang, Y. Feng, S. Zhang, N. Wang, S. Mei, and M. He, "Semi-supervised person detection in aerial images with instance segmentation and maximum mean discrepancy distance," *Remote Sensing*, vol. 15, no. 11, p. 2928, jun 2023.

[14] Y. Cao, Z. He, L. Wang, W. Wang, Y. Yuan, D. Zhang, J. Zhang, P. Zhu, L. V. Gool, J. Han, S. Hoi, Q. Hu, M. Liu, C. Cheng, F. Liu, G. Cao, G. Li, H. Wang, J. He, J. Wan, Q. Wan, Q. Zhao, S. Lyu, W. Zhao, X. Lu, X. Zhu, Y. Liu, Y. Lv, Y. Ma, Y. Yang, Z. Wang, Z. Xu, Z. Luo, Z. Zhang, Z. Zhang, Z. Li, and Z. Zhang, "VisDrone-DET2021: The vision meets drone object detection challenge results," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, Oct 2021.

[15] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2016.

[16] Ž. Marušić, D. Božić-Štulić, S. Gotovac, and T. Marušić, "Region proposal approach for human detection on aerial imagery," in *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*. IEEE, 2018, pp. 1–6.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[18] M. Kundid Vasić and V. Papić, "Multimodel deep learning for person detection in aerial images," *Electronics*, vol. 9, no. 9, p. 1459, 2020.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[20] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

A Futuristic Proposed Solution to a Global Challenge: The Promise of AI for Social Distancing Compliance

Yi Pei Ng
Faculty of Computing and Information Technology
Tunku Abdul Rahman University of Management and Technology
Kuala Lumpur, Malaysia
ngyp-wa16@student.tarc.edu.my

Chi Wee Tan
Faculty of Computing and Information Technology
Tunku Abdul Rahman University of Management and Technology
Kuala Lumpur, Malaysia
chiwee@tarc.edu.my

Yiqi Tew
Faculty of Computing and Information Technology
Tunku Abdul Rahman University of Management and Technology
Kuala Lumpur, Malaysia
yiqi@tarc.edu.my

***Abstract*—Infectious disease is a terrible pandemic around the world, the economy and even people's lifestyles are affected while implementing a lot of standard operating procedures (SOPs) to control the spread of the virus. One of the most effective ways is to control the physical distance between people. Therefore, a Social Distancing Monitoring (SDM) system can be used to track the safe distance between people. In this study, a comparison between algorithms will be made to identify the most efficient algorithm based on human detection results. As a result, the OpenPose algorithm has outstanding human detection results compared with YOLOv4 and YOLOv4-pedestrian. Furthermore, a distinctive bounding box method for human detection has been proposed, which is defined from the human skeleton using the OpenPose algorithm. Hence, SDM will be efficient after a perspective transformation to calculate the distance between people, since most of the people in the frames have been detected. Finally, this study will discuss how the power of AI enhances the SDM system in detecting humans and calculating the distance between people detected.**

***Index Terms*—infectious disease, human detection, social distancing monitoring, perspective transformation, calibration**

I. Introduction

Every epidemic can be controlled by two important measures: quarantine and physical distancing. Since 2003 the SARS virus began spreading worldwide, and these measures have always been part of the SOPs for ensuring public health. In 2019, the World Health Organization will promote a physical distance of at least one metre between people as infectious diseases such as COVID-19 will be transmitted through close contact and the air [1]. Thus, an SDM system can determine the distance between people to determine whether they are at a safe distance.

Technology is advancing rapidly, and Artificial Intelligence (AI) has become a part of everyone's daily lives. Hence, researchers have found that machine learning is effective in automating tasks such as automatically detecting the physical distance among people in real-time using a surveillance camera [2]. There are, however, many concerns associated with a real-time SDM system. To solve the perspective distortion, we performed a perspective transformation with cv2.getPerspectiveTransform and cv2.warpPerspective [3] from the OpenCV package, which implements homogenization [4]. For example, the perspective transformation for calculating the distance between people should proceed in real time by turning a frame into a bird's-eye view. Furthermore, a calibration process must be done to ensure the pixel measured in the frame can be converted to a centimetre measurement. From there, the distance measurement will always be converted to real distance to ensure the violation is detected correctly. Besides, the violated distance measurement, 180cm, will always be consistent, even if the videos are captured from different angles. Therefore, this study adopts YOLOv4-pedestrian, YOLOv4 and OpenPose models to find suitable and effective human detection algorithms.

This paper provided new insights into the studies about some algorithms used for human detection, which focus on rendering bounding boxes for each human detected. In addition, distance calculation using Euclidean Distance [5] after perspective transformation to avoid the distortion issue of the video captured, as well as the calibration process to ensure the distance measurement in the real world.

The paper first briefly overviews the recent history of SDM and introduces the algorithm used for human detection and distance measurement. After that, Section 2 discusses literature reviews on the algorithms proposed and the main contributions of the related SDM papers. Furthermore, Section 3 will discuss the process to execute the SDM system, and the results of the system will be proposed in Section 4. Last but not least, the summary of the experiment and future works will be presented in Section 5.

II. Literature Reviews

Many pieces of research discuss various deep learning algorithms to detect people and several methods of calculating the physical distance between people in real time. Some experiments achieved high accuracy while detecting people using the YOLO algorithm. For example, Ahmed et al. proposed a framework of the YOLOv3 algorithm to perform

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

deep learning-based SDM without any occlusion, as they used overhead perspective video as input data [6]. With the transfer learning of an overhead human dataset with a pre-trained YOLOv3 algorithm, the accuracy of the detection model increased by 3%, which is 95%. Using the Euclidean distance method, the distance among people can be calculated by the centroid of the bounding boxes detected. On the other side, Ahmed et al. also proposed the pre-trained Faster Recurrent Convolutional Neural Network (RCNN) model with the same method and achieved an accuracy of 96% [7].

With a similar approach, Rezaei and Azarmi conducted people detection in indoor and outdoor environments by using large datasets, including MSCOCO and Google Open Image (GOI) datasets, to train and test the YOLOv4-based Deep Neural Network (DNN) model [8]. They used a bird's eye view to calculate the distance for calibration purposes so that the high-risk zone of infectious disease would be identified. Furthermore, Shao et al. trained YOLOv3 with the Merge-Head dataset, which was a larger dataset, and fine-tuned it with a smaller dataset called the UAV-Head dataset sourced from UAV-Head to save resources as it is a similar view of surveillance video [9]. Training results achieved an average precision of 92.22% and 88.5% for testing with the UAV-Head dataset. This experiment only detected the human head and applied a perspective transformation to ignore a complex background.

In another study conducted by Punn et al., they detected humans using the YOLOv3 algorithm while fine-tuning Deepsort techniques to assign ID for detecting people and achieved a balanced mean Average Precision (mAP) and training time with the highest score on FPS [10]. However, they had mentioned their distance measured using the L2 norm but did not illustrate clearly the distance measurement process. In addition, Qin and Xu also used Euclidean distance to calculate the distance between people. They trained the SSD 300 model using the VOC2007 dataset to identify the human object [11]. To measure the distance between two humans, the top-left point of the bounding boxes would be used.

From the experiment of Ansari and Singh [12], a CNN-based human detection has been proposed. Two CNN models have been compared: one of the models, Model 1, consists of two convolutional layers with two pooling layers, while another model, Model 2, consists of three convolutional layers with three pooling layers. Ultimately, Model 2 performs better than Model 1. As one of the methods for detecting humans using the skeleton, Al-Sa'd et al. used OpenPose on the EPFL-MPV, EPFL-Wildtrack, and OxTown public datasets and achieved an accuracy of 94.9%, 99.7% and 95.6% respectively [13]. A perspective transformation was also been carried out to evaluate the crowd density map and social distancing monitoring, which is useful for avoiding occlusion and dealing with missing values.

III. Proposed Method

According to most experiments, the YOLO model had a good result in detecting humans, as it achieved an average accuracy. Therefore, we proposed to utilise the YOLOv4-pedestrian model [14] to compare with other algorithms such as YOLOv4 and OpenPose [15] for human detection. After comparison, we used the OpenPose algorithm to test distance calculation using different points. Fig. 1 showed the entire process flow of our proposed method. At first, the OpenPose algorithm will perform the human feature extraction to an input video, and each human detected was covered by bounding boxes. To calculate the distance precisely, the video frames' perspective was transformed into a bird's eye view to avoid distortion. After calculating the distance between people, the people were classified into three different risk levels in different colours, as shown in Eq. (1).

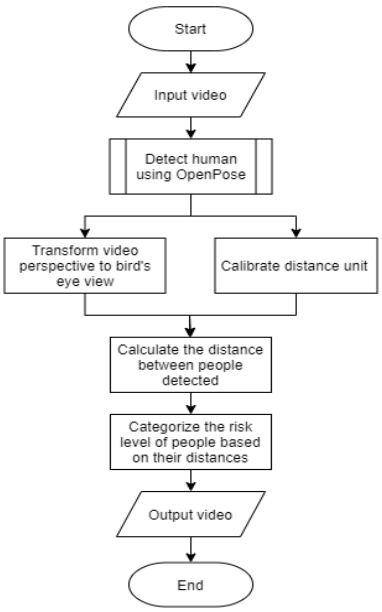


Fig. 1: Our proposed process flow

$$\alpha = \begin{cases} \text{safe}, & \text{if } \delta > 230. \\ \text{careful}, & \text{if } 180 < \delta \leq 230. \\ \text{danger}, & \text{otherwise.} \end{cases} \quad (1)$$

where α is represented as the risk level of certain people and δ refers to distance between people.

A. Human Detection

OpenPose algorithm detected the skeleton of 25 human parts from head to foot. Thus, we integrated the detection mechanism, bounding box, as utilised by the YOLOv4-pedestrian model and YOLOv4 model detections as shown in Fig. 2.

To render a bounding box, the x and y minimum values of the human skeleton detected by OpenPose have been assigned as the bounding box's top-left coordinates. The width and height of the bounding box are calculated by taking the maximum value minus the minimum value of the same detected human skeleton.

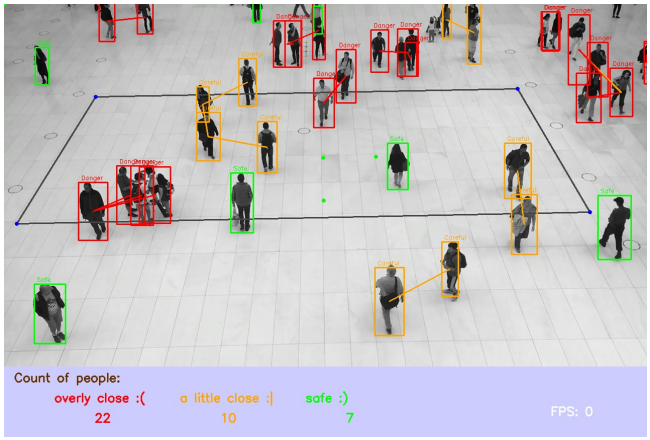


Fig. 2: Output of OpenPose detection.

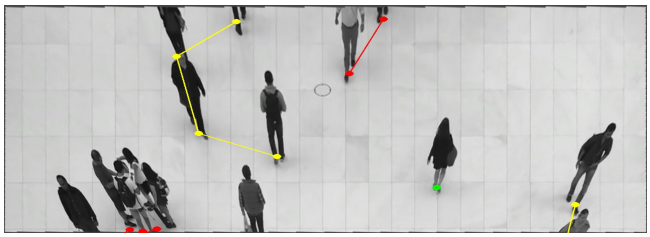


Fig. 3: Detection output in top view perspective.

B. Distance Monitoring

To improve the detection results of human detection in social distance measurement, we performed a perspective transformation process before distance measurement. There are seven points that require manual definition: the first four points (blue dots) will be used in perspective transformation, and the next three points (green dots) for calibration are shown in Fig. 2. These three dots formed two orthogonal lines, and each line represented a length of 180cm in the real world. By doing this, the distance between people can be accurately calculated when the frames have been turned to a top-down perspective, as shown in Fig. 3. Besides that, the distance calculated by pixel in frames is mapped and calibrated to the distance in centimetre measurements by using the two orthogonal lines.

There were three categories of risk level, which represented whether the person had a higher chance of getting infected by a disease with a relatively short social distance between each other. To ease the identification of one's risk level, each detected person was rendered with a bounding box in different colours. Based on Eq. (1), the bounding box in red represented the risk level, α equal to danger, and it turned to orange when the risk level was a careful status, while the green box indicated people in a safe distance from others. If the count of the red bounding box is high in a certain area, the chance of people around that area getting infected may possibly be higher.

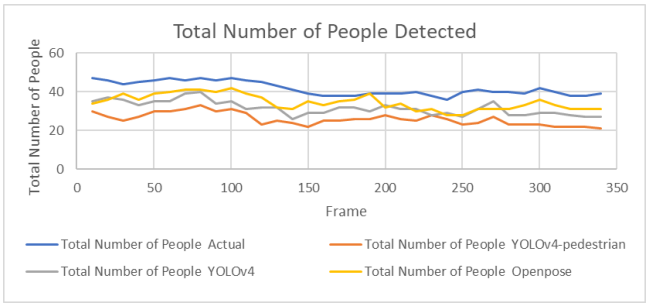


Fig. 4: Comparison among algorithms used in detecting total number of people.

IV. Results

We performed a quantitative metric based on the number of humans detected to evaluate the system's performance. A comparative approach was adopted to evaluate the accuracy of detecting humans. The advantage of using this approach is identifying the efficient algorithm used for the current video or situation.

A. Total Number of People Detected

A comparison chart for the detection results among three detection algorithm categories is shown in Fig. 4. In the figure's legend, the Actual (blue line) referred to the number of people counted by using naked eyes (ground truth), and others were utilising different approaches to the human detection results. In this study, the input video did not contain any occlusions or obstacles, so most of the human activity has been clearly detected. By looking at the yellow line in the chart in Fig. 4, the OpenPose detection result was the most accurate, as most of the detection results were close to the actual count of people detected.

B. OpenPose Detection Results

To find a suitable centre point for distance calculation between people, this experiment used two types of centre points after human detection using OpenPose: the bounding box's bottom centre point and the neck part of the skeleton as centre point. Thus, Fig. 5 compared the number of people detected using OpenPose at the different centre points and categorised them by risk level. In the end, OpenPose detection using the bounding box's bottom centre point detected most people in the violated distance, while OpenPose using the neck part of the skeleton as the centre point detected many people in the safe distance.

Based on Fig. 6, the shape of the human detected was stretched after a perspective transformation. Thus, the neck centre as in Fig. 6(a) might not be an accurate centre point for calculating the distance between people. Furthermore, at the beginning of the system, the last three points selected by users indicated that a horizontal calibration of 180 cm was the equivalent of three boxes of floor tiles, while a vertical calibration of 180 cm was one and a half boxes of floor tiles. Hence, Fig. 6(b) which was the OpenPose detection using

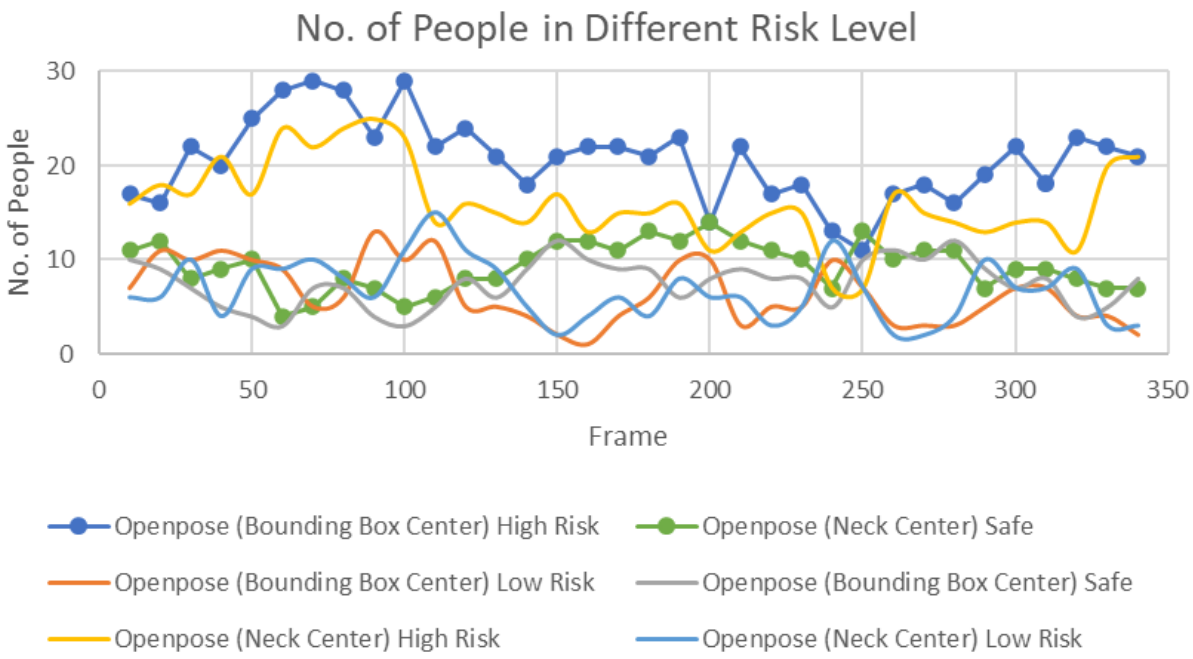
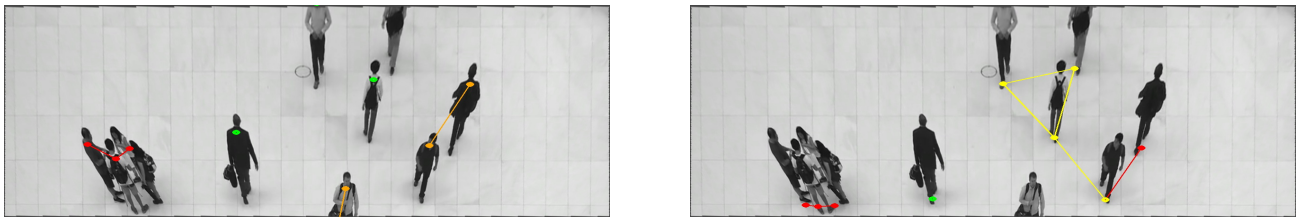


Fig. 5: OpenPose detection results in various risk levels.



(a) Neck centre point

(b) Bottom centre point

Fig. 6: Bird's eye view: OpenPose detection in different centre points.

the bounding box's bottom centre point, had a more accurate distance calculation between people based on the calibrations.

V. Conclusion

In this study, the SDM system was able to calculate the distance between people after a perspective transformation and then categorise people into three risk levels. For human detection, the state-of-the-art algorithm was YOLOv4-pedestrian, which is a pre-trained YOLOv4 model that only detects pedestrians. This study compared the detection results of YOLOv4-pedestrian with two pre-trained algorithms, YOLOv4 and OpenPose. Nonetheless, OpenPose can produce a good human detection result as the video did not contain any occlusion. Hence, this suggests that future research should be undertaken to improve the performance of the system to test the videos in different conditions. For example, videos which contain stairs, objects occluding or in dark conditions, may affect the results of human detection and also cause the distance measurement to be defective.

This system achieved the objectives mentioned in the introduction. At first, a perspective transformation was applied to the video to transform the video's perspective to a bird's eye view, as the distance between people can be calculated accurately. Moreover, the centimetre unit is used for distance measurement, as pixels are calibrated into centimetres by using the last three points selected at the beginning of the system. An improvement in future, the four points of perspective transformation can be selected automatically instead of pointed manually. Furthermore, this paper served as fundamental research on this topic, which does not own a dataset. In future, the system can be tested with a dataset which contains videos with different angles of depression to test the calibration process and the reliability of the system.

Acknowledgment

We would like to express our great appreciation to Tunku Abdul Rahman University of Management and Technology (TAR UMT), under internal grant project no. UC/I/G2021-00073, for the financial and resource support needed to carry out this project.

REFERENCES

[1] World Health Organization, "Covid-19: physical distancing," 2019.

[2] Moein Razavi, Hamed Alikhani, Vahid Janfaza, Benyamin Sadeghi, and Ehsan Alikhani, "An automatic system to monitor the physical distance and face mask wearing of construction workers in covid-19 pandemic," *SN Computer Science*, vol. 3, pp. 27, 1 2022.

[3] Willow Garage and ItSeez, "Open source computer vision: Geometric image transformations," 2023.

[4] J.F. Hughes and J.D. Foley, *Computer Graphics: Principles and Practice*, chapter 10.6, The systems programming series. 2014.

[5] Vikram Singh, "How to compute euclidean distance in python," 2023.

[6] Imran Ahmed, Misbah Ahmad, Joel J.P.C. Rodrigues, Gwanggil Jeon, and Sadia Din, "A deep learning-based social distance monitoring framework for covid-19," *Sustainable Cities and Society*, vol. 65, pp. 102571, 2 2021.

[7] Imran Ahmed, Misbah Ahmad, and Gwanggil Jeon, "Social distance monitoring framework using deep learning architecture to control infection transmission of covid-19 pandemic," *Sustainable Cities and Society*, vol. 69, pp. 102777, 6 2021.

[8] Mahdi Rezaei and Mohsen Azarmi, "Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic," *Applied Sciences*, vol. 10, pp. 7514, 10 2020.

[9] Zhenfeng Shao, Gui Cheng, Jiayi Ma, Zhongyuan Wang, Jiaming Wang, and Deren Li, "Real-time and accurate uav pedestrian detection for social distancing monitoring in covid-19 pandemic," *IEEE Transactions on Multimedia*, 2021.

[10] Narinder Singh Punn, Sanjay Kumar Sonbhadra, Sonali Agarwal, and Gaurav Rai, "Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deepsort techniques," 5 2020.

[11] Jingchen Qin and Ning Xu, "Reaserch and implementation of social distancing monitoring technology based on ssd," *Procedia Computer Science*, vol. 183, pp. 768–775, 1 2021.

[12] Mohd. Aquib Ansari and Dushyant Kumar Singh, "Monitoring social distancing through human detection for preventing/reducing covid spread," *International Journal of Information Technology*, vol. 13, pp. 1255–1264, 6 2021.

[13] Mohammad Al-Sa'd, Serkan Kiranyaz, Iftikhar Ahmad, Christian Sundell, Matti Vakkuri, and Moncef Gabbouj, "A social distance estimation and crowd monitoring system for surveillance cameras," *Sensors*, vol. 22, pp. 418, 1 2022.

[14] Hanlai Chen, Yexiong Xu, and Qinxi Yu, "Real-time social distance detection system," 2020.

[15] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

2023 Asia Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)

Unbiased Decision-Making Framework in Long-Video Macro & Micro-Expression Spotting

Pei-Sze Tan†, Sailaja Rajanala†, Arghya Pal†, Raphaël C.-W. Phan†‡, Huey-Fang Ong†
† CyPhi (ΨΦ) AI Lab, School of Information Technology, Monash University, Malaysia campus
‡ Dept of Software Systems & Cybersecurity, Faculty of IT, Monash University, Clayton campus
Corresponding Authors E-mail: {sailaja.rajanala, arghya.pal}@monash.edu

***Abstract*—Spotting macro and micro expressions proves more demanding than subsequent tasks such as recognizing or classifying expressions. This intricate process entails pinpointing the precise commencement, conclusion, and pinnacle frames of brief facial expressions within video footage. Our study delves into the susceptibility of expression spotting to biases ingrained in deep learning models, often stemming from a lack of diversity in the training data. These biases manifest diversely, such as an overemphasis on specific expression categories, genders, or racial backgrounds, leading to subpar performance on other samples. These spurious biases in deep learning models adversely affect classification accuracy. Our work aims to unearth common biases in micro and macro expression spotting and suggests methods to detect them. In addressing this challenge, we introduce a causal graph that elucidates the connections among a trained deep learning model, the feature space it acquires, and the resulting outcomes. This causal graph plays a pivotal role in pinpointing potential origins of bias during the training phase. We utilise counterfactual inputs to evaluate model biases, providing advantages in terms of computational efficiency and interpretability when contrasted with in-training debiasing approaches. This makes counterfactual debiasing a promising avenue for addressing biases in machine learning models. Experimental results underscore the effectiveness of this approach in successfully mitigating biases in several state-of-the-art macro and micro-expression spotting methods.**

I. Introduction

Facial emotion and micro-expression detection are fields of study that are very closely related. While they both genuinely uncover the emotions exhibited by the individuals, the actual difference lies in how the emotions are conveyed. Since micro-expressions are spontaneously generated in an unconscious state, it is difficult to conceal or disguise and are usually directly related to genuine emotions, so micro-expressions are more reliable in sentiment analysis and have broad application prospects. The success rate is not high, so a computer is required to recognize micro-expressions automatically. Some early works such as [1] note that microexpressions are essentially leaked facial expressions that are brief and sometimes considered more genuine. Specifically, the study notes that microexpressions are defined by their total duration being less than 500 ms or their relatively short onset duration of 260 ms.

Macro and Micro-expression spotting [2] refers to detecting whether a macro or micro-expression is included in a video stream and marking the start point (onset), peak value (apex) and endpoint (offset) of that specific expression. The onset frame refers to the time when that expression appears; the apex frame refers to the time when the magnitude of that expression is the largest; the offset frame refers to the time when the expression disappears. Identifying and interpreting expressions, especially micro-expressions, can provide insights into emotional states, deception detection, and assessing an individual's genuine reactions.

Bias in machine learning model [3] leads to unfairness and discriminative outcomes, amplifying the inequality towards specific groups of people. Label annotation for images or videos by experts or professionals who aim to assign an accurate label, but biases arise towards some groups will also present underlying labels that are biased [4]. Besides the label, racial bias needs to be mitigated to ensure the network results are objective [5]. Besides that, the medicine field [6], human-computer interaction [7] and software research topics [8] also take fairness under their consideration. All the potential biases above in deep learning models should be taken into account during the experimental and evaluation stages. It promotes transparency, as understanding the sources and types of bias helps explain the decision-making process in the black-box predictions of the machine learning models.

In this paper, our contributions can be summarized as follows:

1) We proposed a method to identify the potential or unseen biases in all variables involved in the experiments.
2) We proposed a novel technique for mitigating the bias in the macro and micro-expressions spotting machine learning model.
3) We verify the effectiveness of the presented method by comparing it with state-of-the-art methods.

II. Literature Review

A. Spotting task in Micro expression

Most present works propose their novelty in the feature extraction and classification model. Optical flow as input features is mainly used in existing works [9], [10], [11], [12], [13] to extract the motion features of each long video. Using optical flow, which estimates the apparent motion of pixels between consecutive frames, allows us to focus on the dynamic changes that occur. Using a convolutional model, Yap et al. [14] proposed a 3D-CNN-based approach that compares each frame with two temporally local reference frames. Another 3D-CNN-based model called SOFTNet is presented by Liong et al.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

[15], which estimates a score that represents the probability of a frame falling within an expression interval. The graph-based model is yet to be investigated. Only [9] presented an action units (AUs) aware model with the graph embedding function in their model. Yu et al. [16] employed a Temporal Multi-Receptive Field Feature Fusion Module (TMRF3M) and a Local Suppression and Global Enhancement Module (LSGEM) to enhance the precision of spotting short intervals and suppress irrelevant background information. Yang et al. [17] proposed an advanced Concat-CNN model which is capable of capturing the inherent correlations among distinct frames.

B. Casuality and Fairness in Facial Expression

Casual analysis [18] aims to comprehend the causal relationship between variables, focusing on how one variable directly affects another. Gender and racial bias [5], [19] are often examined in previous facial affect classification work. This is partly because of the imbalanced distribution of ethnicities and gender identity in the published datasets. Biased label identification and reduction of this bias are also studied by a few works [20], [4]. Action unit recognition in human faces is crucial in annotating and classifying individuals' genuine emotions. In [21], they utilized domain-incremental learning settings to enhance system fairness while effectively addressing biases arising from uneven data distributions. Cheong et al. [22] presented a framework that used the fast causal inference algorithm to construct a casual relation graph in facial expression recognition. A proposal utilizing Siamese Networks to minimize the disparities between predictions made on the original images and their corresponding counterfactual counterparts is presented in [23]. Chen et al. [24] mitigate the adverse impacts caused by dataset bias with their novel, the Interventional Emotion Recognition Network, which aims to perform backdoor adjustment. This adjustment technique is a fundamental de-confounding method in causal inference. Debiasing in expression spotting tasks is still a new research direction yet to be widely discovered.

III. Methodology

The primary aim of this unbiased decision-making framework is to identify and reduce the bias in spotting models effectively.

A. Macro and Macro expression Spotting

Formally, the problem of micro and macro-expression can be defined as: Given a video I_n we aim to learn a model $M_\theta(.)$ that can identify the onset L_{on}, apex L_{apex}, and the offset L_{off} frames of expression within I_n, θ being the parameters of the model and, n, the number of frames. The output results of the $M_\theta(.)$ given the I_n can be thus represented as as $M_\theta(I_n) = P(L_{on}, L_{apex}, L_{off}|I_n)$. From here on, we refer to the triplet $(L_{on}, L_{apex}, L_{off})$ containing the onset, apex and offset frame as L and it is obtained from the ground truth. The predictions of $M_\theta(.)$ can be falsely impacted by many unwanted factors known as confounders or otherwise, the bias variables.

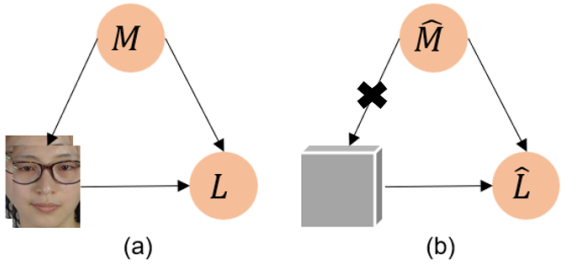


Fig. 1. A causal relation graph akin to Pearl's [25] is established among the key elements: the *trained model* M, the extracted features, and *the outcome* L. (a) It is important to observe that this causal graph illustrates how the trained model M can exert influence over both the learned features and the outcome L." Thus, it acts as a confounder in the relationship between the "feature space" and L. (b) To gain insights into the genuine causal connection of "feature space" $\rightarrow L$," we employ backdoor adjustment, an approach that intervenes using the counterfactual blank frames denoted by the shaded feature in this context.

B. Causal Formulation of Bias

The biased predictions of a model can be formulated as the following conditional probability shown in the equation below:

$$P(L|I_n) = \sum_{n=1}^{C} P(L|I_n, c)P(c|I_n) \tag{1}$$

The factorization described above is aware of the causal graph depicted in Figure 1, which illustrates the relationship between various components. It demonstrates that the expected outcome $P(L|I_n)$ is influenced by confounding factors c. Confounders are the factors that influence both the independent and dependent variables, i.e., the expected outcome. They are often factors that are not directly measured or controlled for in the study but can influence the outcome. The graph in Figure 1 represents the interconnections among the trained model $M_\theta(.)$, the learned features B, and the expected outcome L. It indicates that the trained model $M_\theta(.)$ is a confounder that impacts both the learned features B and the outcome L. Specifically, the relationship depicted as $B \leftarrow M \rightarrow L$ suggests the existence of a backdoor path or confounding path, where $M_\theta(.)$ indirectly affects the outcome L through B. This indirect path complicates understanding the complete or isolated effect of B on L. Moreover, when a biased model $M_\theta(.)$ is employed, the learned features B also tend to be biased.

To address this issue and isolate the effect of $M_\theta(.)$ from that of $L \rightarrow B$, we utilize Pearl's backdoor adjustment strategy [26]. This strategy allows us to disentangle the influence of $M_\theta(.) \rightarrow B$ by explicitly assigning a value to B. In causal theory, this intervention is known as the "do(.)" operation, where we do or, in other terms, set the variable of interest to a particular value. Below, we showcase the do operation-based debiasing technique. Please note that in practice, we need more complete knowledge of all the confounding factors contributing to bias. Thus, it is impossible

to determine the entire set of confounders, and we can only rely on a limited subset that is known or observed. In this work, we work on debiasing the **dataset-based biases**.

C. Mitigation of bias

This study examined how the model performs in classifying offset and onset frames. We discovered the presence of dataset biases within the models, where certain characteristics of the input training data are incorrectly associated with specific outcomes. To delve deeper into this matter, we investigated by manipulating the features B. Specifically, we set them to blank feature frames I_{0n} that do not contain any meaningful feature information. This manipulation is denoted as $M(L|do(I_n = I_{0n}))$. In this context, I_{0n} is referred to as a counterfactual input, which enables us to assess how the model performs when presented with no information at all [27].

Our expectation was that, given the absence of relevant information, the model $M_\theta(.)$ would not be able to predict L_{on}, L_{apex}, and L_{off}. However, our experimental results indicate that the model still predicts these labels for blank frames, suggesting that the model has learned the biases present in the dataset.

Following are the biases that we observe in the current models. During the prediction of the offset frame L_{off}, the model tends to select frames that are closer to the end of the total frame length. As a result, even if the expression in the video ends earlier, the model would still primarily predict the last frames as potential candidates for the offset frame. In addition, the models tend to classify the onset frame as taking place approximately 30-40 frames prior to the offset frame. Consequently, rather than identifying them as micro-expressions, the models are more inclined to categorize these I_{0n} frames as macro-expressions. The fact that there is a consistent pattern of assigning specific frames and types of expressions in the prediction results implies that the models rely on external factors that are unrelated to the inherent characteristics of the expressions for their classification. Both $M(L|I_n)$ and $M(L|do(I_n = I_{0n}))$ generate a probability distribution P for each frame, indicating the likelihood of it being classified into any of the frame types within L. The shape of P depends on the model outcomes. In this paper, we refer to these probabilities as:

$$P_I = \{pr_{s,n}, pr_{e,n}, pr_{a,n}, pr_{t,n}\} \tag{2}$$

I is the type of testing inputs which is I_{0n} or I_n. $pr_{s,n}, pr_{e,n}, pr_{a,n}, pr_{t,n}$ denotes the set of probabilities for onset, offset, apex and expression type. With the method above, after addressing a bias, we could use Pearl's backdoor criterion [25] to increase the fairness in the model. In order to remove the bias, we first record the original biased probabilities P_{I_n} for a given input video I_n and then debias the output probability distribution by subtracting $P_{I_{0n}}$ from it,

$$f(l|\mathbf{v}) = M(L|I_n) - \lambda M(\hat{L}|do(I_n = \hat{I_{0n}})) \tag{3}$$

here, λ_{param} is the hyperparameter that mitigates the plausible sources of biases. λ values differ for every set of P to maximize the pertinency of the debias method. This method is known as a counterfactual debiasing method and is performed during the inference time, which is less time-consuming and flexible. This is also a more interpretable method of debiasing.

IV. Experiments

A. Datasets Available

1) Real-World Data: The human facial expression is usually hard to capture in the wild because the unstable environment will influence one's emotions and increase inaccuracy when we label the captured samples. Therefore, the human emotion datasets will usually be recorded in an uninterrupted studio environment and specific videos will be used to trigger particular emotions. Collecting such datasets for micro-expressions (ME) is even more difficult as they are challenging to induce naturally. Therefore, the current samples from the real world are limited. There are a few macro and micro-expression spotting databases that are constantly being utilized which are CAS(ME)^2 [28], CAS(ME)^3 [29], SMIC-E-Long Video (SMIC-E)[30], SAMM-Long Video (SAMM-LV)[31] and 4D micro-expression dataset(4DME) [32].

2) Trainable Synthetic ME Dataset: Since the data from the real persons have an imbalanced number of samples in categories and the dataset size is too small for training purposes, Y Liu et al. [33] introduced the MiE-X dataset synthesized from a generative adversarial network model. This dataset is not suitable for training or testing in a spotting task. Each individual and each of the three emotion classes only consists of three micro-expression samples and three macro-expression samples for each combination.

This experiment used the CAS(ME)^2 and SAMM-LV dataset as training and testing inputs. CAS(ME)^2 includes long videos and cropped expression samples. It has 87 long videos capturing spontaneous macro-expressions and micro-expressions. SAMM-LV consists of 147 videos that contain a total of 343 macro-expressions and 159 micro-expressions. The emotion labels in the databases are based on a combination of action units (AUs) and the emotion types of the videos that evoke the emotions. Both datasets are collected in a controlled environment.

B. Baselines

We employ two state-of-the-art methods in our experimental setup: the ABPN framework and the AUW-GCN architecture. These methods are widely recognized for their effectiveness in the field.

- **Apex and Boundary Perception Network (ABPN) [10]**. Part of this work's important novelty is that calculating optical flow differences mitigates the influence of head movement and other facial areas on microexpression spotting. Additionally, they employ 1D convolutional layers to extract temporal features and introduce the probability evaluation module to estimate the auxiliary probability of each frame belonging to an apex or boundary frame.

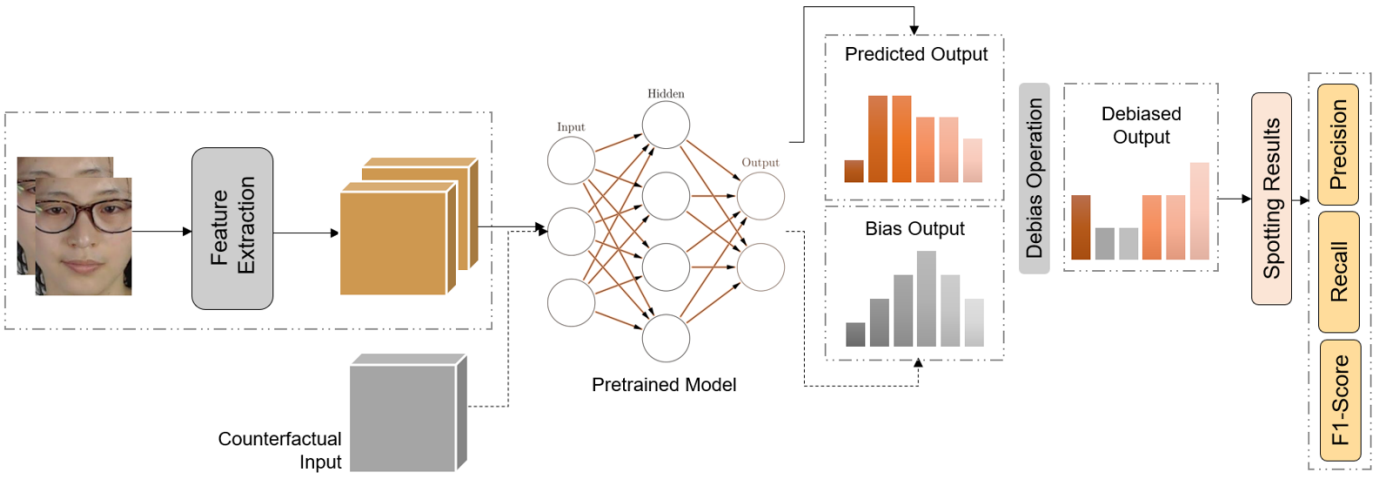


Fig. 2. The diagram illustrates the structure of our proposed debiasing technique. During inference, we gather the predicted output by passing a test sample through a trained model. Subsequently, we introduce the counterfactual input, as detailed in Section III, and capture its predicted output, which represents the biased output. To achieve a debiased outcome, we subject the predicted output distributions of the test sample and the counterfactual input to a debiasing operation.

- **Action-Unit-aWare Graph Convolutional Network (AUW-GCN) [9]**. This work proposed an action units aware model in which AU-related statistics are encoded into the GCN. AUs are used to build the adjacency matrix and multiply it with the features of the video frames to complete the spatial encoding phase. The spatial-temporal learning architecture in this paper minimized the loss in the learning process and increased the results significantly.

C. Data-Preprocessing

Video frames pre-processing methods and model parameters remain unchanged from the original works to ensure optimal performance without introducing unaccounted biases. The implementations of both models are based on the PyTorch library.

Before implementing Equation 3, we prepared blank frames I_{n0} as the counterfactual inputs and used them as test inputs for the pre-trained model. After passing I_{n0} to the model, the $P_{I_{n0}}$ of the counterfactual testing samples are recorded. Figure 2 concludes the steps to mitigate the identified bias and re-run the experiments with the pre-trained weights.

D. Metrics

In line with the baselines, Leave-One-Subject-Out (LOSO) validation is used in our experiments and the evaluation metric is the F1 score. It is commonly used for imbalanced sample datasets because it provides a balanced measure of a model's performance by considering both the ability to correctly identify positive instances (precision) and capture all positive instances (recall). This choice allows for a fair and reliable assessment of our model's performance.

Overall results in table I and II refer to the F1 score for both macro and micro-expressions. The overall F1 score is as shown in Equation 6:

$$F1_{ME} = \frac{TP_{ME}}{TP_{ME} + \frac{1}{2}(FP_{ME} + FN_{ME})} \tag{4}$$

$$F1_{MaE} = \frac{TP_{MaE}}{TP_{MaE} + \frac{1}{2}(FP_{MaE} + FN_{MaE})} \tag{5}$$

TP, FP and FN are the number of true positives (TP), false positives (FP), and false negatives (FN) produced by the model on the test data. MaE and ME denote the macro expression and micro-expression, respectively.

$$F1_{overall} = F1_{ME} + F1_{MaE} \tag{6}$$

E. Results and Dicussion

Table I and II shows the results in the F1 score before and after debiasing the ABPN and AUW-GCN model. After applying the proposed method and fine-tuning the hyperparameters, the F1 scores for both methods show improvement in CAS(ME)^2 and SAMM-LV datasets. CAS(ME)^2 reflects a better performance with the framework partly because it is an RGB dataset, unlike the SAMM-LV dataset. The contextual information in a video's environment can influence a model's learning process, potentially leading to biases in the dataset and preventing the model from fully capturing essential features in the frames. The surrounding environment, including lighting conditions, background elements, and spatial relationships, can impact the appearance and interpretation of facial expressions.

Our findings highlight the importance of addressing dataset biases in expression spotting tasks, particularly in micro expressions' video sequences where facial part motion intensity is minimal. The susceptibility of microexpressions to dataset biases calls for greater scrutiny when collecting training datasets, ensuring they are diverse and representative of real-world scenarios. Furthermore, fine-tuning hyperparameters can enhance performance by enabling models to overcome biases and improve their ability to spot accurate frames.

Method	CAS(ME)^2						SAMM-LV					
	Macro		Micro		Overall		Macro		Micro		Overall	
ABPN (Proposed)	0.3357		0.1590		0.3117		0.3349		0.1689		0.2908	
ABPN (Reproduce)	**0.3940**	**-0.51%**	0.1282	**+0.51%**	0.3224	**+0.91**	**0.3438**	**-0.24%**	0.1854	**+0.00%**	0.2760	**+0.35%**
Ours (After Debias)	0.3889		**0.1333**		**0.3315**		0.3414		0.1854		**0.2795**	

TABLE I
RESULTS OF F1 SCORE BEFORE AND AFTER DEBIASING IN ABPN MODEL FOR ME SPOTTING TASK

Method	CAS(ME)^2						SAMM-LV					
	Macro		Micro		Overall		Macro		Micro		Overall	
AUW-GCN (Proposed)	0.4235		0.1538		0.3834		0.4293		0.1984		0.3728	
AUW-GCN (Reproduce)	0.4084	**+1.52%**	0.0506	**+5.25%**	0.3678	**+0.96**	0.4212	**+0.16%**	0.2372	**+0.09%**	0.3767	**+0.14%**
Ours (After Debias)	**0.4236**		**0.1031**		**0.3774**		**0.4228**		**0.2381**		**0.3781**	

TABLE II
RESULTS OF F1 SCORE BEFORE AND AFTER DEBIASING IN AUW-GCN MODEL FOR MAE AND ME SPOTTING TASK

Method	CAS(ME)^2			SAMM-LV		
	MaE	ME	Total	MaE	ME	Total
MDMD [34]	0.1196	0.0082	0.0376	0.0629	0.0364	0.0445
STF [35]	0.2131	0.0547	0.1403	0.1331	0.0725	0.0999
3DCNN[14]	0.0686	0.0119	0.0497	0.1863	0.0409	0.1193
SOFT [36]	0.2410	0.1173	0.2022	0.2169	0.1520	0.1881
LSS [37]	0.3770	0.0420	0.3250	0.2810	0.1310	0.2380
ABPN [10]	0.3357	**0.1590**	0.3117	0.3349	0.1689	0.2908
MTSN [38]	0.4104	0.0808	0.3620	0.3459	0.0878	0.2867
AUW [9]	0.4168	0.0506	0.3739	**0.4310**	0.1955	0.3736
Ours	**0.4236**	0.1031	**0.3774**	0.4228	**0.2381**	**0.3781**

TABLE III
RESULTS OF F1 SCORE IN BASELINE METHODS AND THE DEBIAS RESULTS IN AUW-GCN MODEL.

F. Limitations

Various spotting experiments have been conducted in previous works with real-world data. Nevertheless, the number of samples in each dataset is much smaller than the facial expression recognition benchmark size, as natural micro-expressions are challenging to trigger. Consequently, we need to synthesize more samples to achieve fairer experiment results. Moreover, the causal relation between all variables is denoted with prior knowledge or a hand-crafted method. A more robust model to prove the source of bias is needed to gain reliable outcomes.

V. CONCLUSION

This study introduces an innovative debiasing approach tailored for the task of facial expression spotting. We commence by addressing prevalent biases commonly observed in micro and macro-expression spotting, with a primary focus on mitigating dataset-induced bias. Subsequently, we outline methods designed to detect the presence of such bias. Employing a causal graphical model, we delve into potential sources of bias and utilize counterfactual inference to scrutinize the model for any inherent bias, as postulated in. Notably, we observe that a model trained on a biased dataset acquires bias within its feature space. To rectify this, we employ non-informative counterfactual blank frames to execute a back-door adjustment. Crucially, the counterfactual intervention is executed during inference, rendering our debiasing method efficient in terms of both time and space. The insights gained through counterfactual inference enable us to identify and subsequently correct biases present in the model's predicted outputs. Our experimental results vividly illustrate the efficacy of our approach in successfully alleviating biases across a range of state-of-the-art macro and micro-expression spotting methods.

ACKNOWLEDGMENT

This research was supported in part by the Malaysia Ministry of Higher Education (MoHE)'s Fundamental Research Grant Scheme (FRGS) under the AdverSurreal project FRGS/1/2021/ICT02/MUSM/01/1. Pei-Sze Tan is supported by the project Uxpress funded by the Prader-Willi Syndrome Association of Victoria (PWSA) and Faculty of IT, Monash University, Australia. All the experiments were conducted on Montage, the high-performance computing (HPC) platform at Monash University, Malaysia campus.

REFERENCES

[1] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37:217–230, 2013.

[2] Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, Xiaobai Li, and Guoying Zhao. Micro-expression spotting: A new benchmark. *Neurocomputing*, 443:356–368, 2021.

[3] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[4] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[5] Abdallah Hussein Sham, Kadir Aktas, Davit Rizhinashvili, Danila Kuklianov, Fatih Alisinanoglu, Ikechukwu Ofodile, Cagri Ozcinar, and Gholamreza Anbarjafari. Ethical ai in facial expression analysis: Racial bias. *Signal, Image and Video Processing*, 17(2):399–406, 2023.

[6] Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. Mitigating bias in machine learning for medicine. *Communications medicine*, 1(1):25, 2021.

[7] Kerui Gu, Linlin Yang, Michael Bi Mi, and Angela Yao. Bias-compensated integral regression for human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[8] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

[9] Shukang Yin, Shiwei Wu, Tong Xu, Shifeng Liu, Sirui Zhao, and Enhong Chen. Au-aware graph convolutional network for macro-and micro-expression spotting. *arXiv preprint arXiv:2303.09114*, 2023.

[10] Wenhao Leng, Sirui Zhao, Yiming Zhang, Shiifeng Liu, Xinglong Mao, Hao Wang, Tong Xu, and Enhong Chen. Abpn: Apex and boundary perception network for micro- and macro-expression spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 7160–7164, New York, NY, USA, 2022. Association for Computing Machinery.

[11] He Yuhong. Research on micro-expression spotting method based on optical flow features. MM '21, New York, NY, USA, 2021. Association for Computing Machinery.

[12] Yuan Zhao, Xin Tong, Zichong Zhu, Jianda Sheng, Lei Dai, Lingling Xu, Xuehai Xia, Yu Jiang, and Jiao Li. Rethinking optical flow methods for micro-expression spotting. MM '22, New York, NY, USA, 2022. Association for Computing Machinery.

[13] Jun Yu, Zhongpeng Cai, Zepeng Liu, Guochen Xie, and Peng He. Facial expression spotting based on optical flow features. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 7205–7209, New York, NY, USA, 2022. Association for Computing Machinery.

[14] Chuin Hong Yap, Moi Hoon Yap, Adrian Davison, Connah Kendrick, Jingting Li, Su-Jing Wang, and Ryan Cunningham. 3d-cnn for facial micro- and macro-expression spotting on long video sequences using temporal oriented reference frame. MM '22, New York, NY, USA, 2022. Association for Computing Machinery.

[15] Gen-Bing Liong, John See, and Lai-Kuan Wong. Shallow optical flow three-stream cnn for macro- and micro-expression spotting from long videos. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2643–2647, 2021.

[16] Wang-Wang Yu, Jingwen Jiang, Kai-Fu Yang, Hong-Mei Yan, and Yong-Jie Li. Lgsnet: A two-stream network for micro- and macro-expression spotting with background modeling. *IEEE Transactions on Affective Computing*, pages 1–18, 2023.

[17] Bo Yang, Jianming Wu, Kazushi Ikeda, Gen Hattori, Masaru Sugano, Yusuke Iwasawa, and Yutaka Matsuo. Deep learning pipeline for spotting macro- and micro-expressions in long video sequences based on action units and optical flow. *Pattern Recognition Letters*, 165:63–74, 2023.

[18] Leland Gerson Neuberg. Causality: Models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

[19] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. Gender stereotyping impact in facial expression recognition. *arXiv preprint arXiv:2210.05332*, 2022.

[20] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition, 2021.

[21] Nikhil Churamani, Ozgur Kara, and Hatice Gunes. Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. *IEEE Transactions on Affective Computing*, pages 1–15, 2022.

[22] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Causal structure learning of bias for fair affect recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 340–349, 2023.

[23] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 245–261. Springer, 2023.

[24] Yuedong Chen, Xu Yang, Tat-Jen Cham, and Jianfei Cai. Towards unbiased visual emotion recognition via causal intervention, 2022.

[25] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009.

[26] Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010.

[27] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, 2021.

[28] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. Cas (me) ˆ2: A database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 9(4):424–436, 2017.

[29] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[30] Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, Xiaobai Li, and Guoying Zhao. Micro-expression spotting: A new benchmark. *Neurocomputing*, 443:356–368, 2021.

[31] Chuin Hong Yap, Connah Kendrick, and Moi Hoon Yap. Samm long videos: A spontaneous facial micro- and macro-expressions dataset. pages 771–776, 11 2020.

[32] Xiaobai Li, Shiyang Cheng, Yante Li, Muzammil Behzad, Jie Shen, Stefanos Zafeiriou, Maja Pantic, and Guoying Zhao. 4dme: A spontaneous 4d micro-expression dataset with multimodalities. *IEEE Transactions on Affective Computing*, 2022.

[33] Yuchi Liu, Zhongdao Wang, Tom Gedeon, and Liang Zheng. How to synthesize a large-scale and trainable micro-expression dataset? In *ECCV*, 2022.

[34] Ying He, Su-Jing Wang, Jingting Li, and Moi Hoon Yap. Spotting macro- and micro-expression intervals in long video sequences. 09 2020.

[35] Li-Wei Zhang, Jingting Li, Su-Jing Wang, Xian-Hua Duan, Wen-Jing Yan, Hai-Yong Xie, and Shu-Cheng Huang. Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences. 09 2020.

[36] Gen-Bing Liong, John See, and Lai-Kuan Wong. Shallow optical flow three-stream cnn for macro- and micro-expression spotting from long videos. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2643–2647, 2021.

[37] Wang-Wang Yu, Jingwen Jiang, and Yong-Jie Li. Lssnet: A two-stream convolutional neural network for spotting macro- and micro-expression in long videos. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4745–4749, New York, NY, USA, 2021. Association for Computing Machinery.

[38] Gen Bing Liong, Sze-Teng Liong, John See, and Chee-Seng Chan. Mtsn: A multi-temporal stream network for spotting facial macro- and micro-expression with hard and soft pseudo-labels. In *Proceedings of the 2nd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis*, FME '22, page 3–10, New York, NY, USA, 2022. Association for Computing Machinery.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Adaptive Beamforming Based on Interference-Plus-Noise Covariance Matrix Reconstruction for Speech Separation

Yongxiong Xiao*†, Shiqiang Zhu*†, Te Li*†, Minhong Wan*†[1], Wei Song*†[2], Jason Gu*†‡, Qiang Fu*†

* Research Center for Intelligent Robotics, Research Institute of Interdisciplinary Innovation, Zhejiang Lab, Hangzhou, 311100, China.

† Zhejiang Engineering Research Center for Intelligent Robotics, Hangzhou, 311100, China.

‡ Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS Canada B3H 4R2.

***Abstract*— Estimating the interference-plus-noise covariance matrix (INCM) is critical for the robustness of the minimum variance distortionless response (MVDR) beamformer. Existing INCM reconstruction methods are computationally intensive and not suitable for real-time speech separation. We propose a singular value decomposition (SVD) based INCM reconstruction method for speech separation. The spatial covariance matrix (SCM) for each source is obtained by rank-1 approximation using the nominal steering vector (SV) or the pre-measured relative transfer function (RTF). The INCM used to separate each source is reconstructed as the sum of the covariance matrices of interference and spherical isotropic noise. The proposed method is evaluated using the mixed signal received by a circular array with six microphones placed in a simulated reverberation chamber. The results show that the proposed method has comparable sound quality performance to the reference method, but requires much less computation.**

I. INTRODUCTION

Speech separation refers to the problem of extracting one or more target utterances while eliminating interfering utterances and noise [1]. It is a key pre-processing step for noise-robust automatic speech and speaker recognition, and its applications include enabling intelligent robots to understand speech content and interaction intention when multiple people are speaking simultaneously.

In recent years, deep learning methods has been widely studied for speech separation by exploring single-channel features [2]-[4], multi-channel features [5][6], and multi-modal features [7][8]. Most devices today feature a microphone array consisting of two or more sensors, but not all of them have cameras. Adaptive beamforming as a linear spatial filter is still one of the most efficient and effective front-end processing techniques for robust speech separation in practical scenarios [9]-[11].

Various robust adaptive beamformers with different optimization criteria have been proposed, such as multi-channel Wiener filter (MWF), minimum variance distortionless response (MVDR), and linear constraint minimum variance (LCMV). Estimating acoustic parameters, such as the acoustic transfer function (ATF) and the interference-plus-noise covariance matrix (INCM), from the mixed speech observations are critical for the robustness of these beamformers [12] [13]. Although many methods for improving the robustness of beamformers can be found in the literature, such as diagonal loading [14], eigenspace projection [15], worst-case optimization [16], and covariance matrix tapering [17], estimating the ATF and INCM in the presence of multiple sources is still an open research question.

Covariance subtraction-based and covariance whitening-based methods for estimating relative transfer functions (RTFs) typically require knowledge of time frames in which only one

[1] Corresponding authors, E-mail: wanmh@zhejianglab.com Tel: +86-13309810721

[2] Co-corresponding authors, E-mail: weisong@zhejianglab.com Tel: +86-13867218612

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

desired source is active [18] or estimation of speech presence probabilities (SPPs) [19]. However, estimating the SPP per speaker in dynamic scenes raises more problems, such as clustering and permutation alignment [20].

The INCM used to separate the desired source is defined as the sum of the covariance matrices of interference and all possible noises. Gu and Leshem [21] proposed an INCM reconstruction method by integrating a Capon spectral estimator over an angular sector excluding the direction of arrival (DOA) of the signal of interest. Yuan and Gan [22] proposed an INCM reconstruction method using principal eigenvalue vectors. Although many efforts have been made to reduce the computational complexity [23] [24], existing INCM reconstruction methods are still not suitable for real-time speech separation.

Generalized singular value decomposition (SVD) for real-time sound source localization has been shown to significantly reduce computation while maintaining noise robustness in localization [25]. Observing that SVD provides the same orthogonality as eigenvalue decomposition as well as a more efficient way of computing matrix inverses, we propose in this work an SVD-based INCM reconstruction method for real-time speech separation. First, we employ a recursive smoothing method to estimate the covariance matrix of the mixed speech observations. The SVD of the covariance matrix is then used to compute its inverse and estimate the variance of each source and noise. Furthermore, the spatial covariance matrix (SCM) for each source is obtained by its rank-1 approximation. The final INCM used to separate each source is full rank by summing the covariance matrices of the interference and spherical isotropic noise. Finally, the coefficients of the separation matrices are calculated by the MVDR method.

The remainder of this paper is organized as follows. In Section II, we describe the proposed adaptive beamforming method for speech separation based on INCM reconstruction. We then provide the experimental setup and evaluation results in Section III. Conclusions are given in Section IV.

II. Proposed Adaptive Beamformer

A. Signal Model

Suppose a recorded sound scene consists of D desired sources and a microphone array with M isotropic transducers. The microphone index and source index are denoted by i and j, respectively. The short-time Fourier transform (STFT) of the multichannel signal received by the microphone array can be expressed as

$$\boldsymbol{y}(t,f) = \sum_{j=1}^{D} \boldsymbol{a}_j(f) S_j(t,f) + \boldsymbol{N}(t,f), \quad (1)$$

where $S_j(t,f)$, $j = 1,2,\dots D$, is the j-th desired source, $\boldsymbol{a}_j(f)$ is the ATF of j-th source, $\boldsymbol{N}(t,f)$ is the noise vector including all undesired sources, noise and late reverberation, and

$$\boldsymbol{y}(t,f) = [Y_1(t,f) \quad Y_2(t,f) \quad \dots \quad Y_M(t,f)]^T, \quad (2)$$

where $Y_i(t,f)$, $i = 1,2,..M$ is the signal received by the i-th microphone at time-frame index t and frequency index f, T denotes the regular matrix transpose. Note that we assume that each speaker's ATF is time-invariant.

Speech separation is applied by using spatial filters as

$$\boldsymbol{Z}(t,f) = \boldsymbol{W}^H(t,f)\boldsymbol{y}(t,f), \quad (3)$$

where $\mathbf{W}(t,f)$ is the separation matrix with weighting coefficients, H denotes the complex conjugate matrix transpose, and $\boldsymbol{Z}(t,f)$ is the vector of separated signals.

B. INCM Reconstruction

At the initial time step, the $M \times M$ covariance matrix $\boldsymbol{\Phi}_{\text{mix}}$ of the mixed signal is initialized as

$$\boldsymbol{\Phi}_{\text{mix}}(0,f) = \boldsymbol{\Gamma}_{0,\pi}(f) + \varepsilon(f)\boldsymbol{I}_M, \quad (4)$$

where $\boldsymbol{I}_M$ is an $M \times M$ identity matrix, $\varepsilon(f)$ is a frequency-dependent regularization parameter, $\boldsymbol{\Gamma}_{0,\pi}(f)$ is the covariance matrix of spherical isotropic noise [11],

$$\boldsymbol{\Gamma}_{0,\pi}(f) = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} \boldsymbol{d}(\theta,\phi,f)\boldsymbol{d}^H(\theta,\phi,f)\sin\theta\, d\theta\, d\phi, \quad (6)$$

where $\boldsymbol{d}(\theta,\phi,f)$ is the array manifold vector, θ is the evaluation angle measured from the positive z-axis, and ϕ is the azimuth angle measured from the x-axis to the y-axis. The covariance matrix of spherical isotropic noise models late reverberation effects [1].

At time step t, $\boldsymbol{\Phi}_{\text{mix}}$ is obtained by recursive smoothing as

$$\boldsymbol{\Phi}_{\text{mix}}(t,f) = \alpha_1\boldsymbol{\Phi}_y(t-1,f) + (1-\alpha_1)\boldsymbol{\Phi}_y(t,f), \quad (5)$$

where $\boldsymbol{\Phi}_y(t,f) = \boldsymbol{y}(t,f)\boldsymbol{y}^H(t,f)$, α_1 is a smoothing parameter, $0 < \alpha_1 < 1$.

The parameter $\varepsilon(f)$ tries to find a good compromise between supergain and white noise amplification. Since white noise amplification is much worse at low frequencies than at high frequencies [11], we choose larger values of $\varepsilon(f)$ at low frequencies and smaller values at high frequencies.

SVD of the covariance matrix of the mixed signal gives

$$\boldsymbol{\Phi}_{\text{mix}}(t,f) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H, \quad (7)$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are left- and right-singular vectors, respectively, which are unitary and mutually orthogonal, $\boldsymbol{\Sigma} = \mathrm{diag}\{\varsigma_1, \varsigma_2, \ldots \varsigma_M\}$, is a diagonal matrix containing the singular values in descending order.

For the noise covariance matrix, the noise power is simply estimated as the mean of the $(M-D)$ smallest singular values as

$$\sigma_n^2(t,f) = \frac{\mathbf{1}}{M-D}\sum_{i=M-D}^{M}\varsigma_i. \tag{8}$$

In order to reduce the amount of calculation, the inverse matrix of the $\boldsymbol{\Phi}_{\mathrm{mix}}$ is obtained from singular values and vectors as

$$\boldsymbol{\Phi}_{\mathrm{mix}}^{-1}(t,f) = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^H, \tag{9}$$

where $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}\{1/\varsigma_1, 1/\varsigma_2, \ldots 1/\varsigma_M\}$.

The signal variance of the j-th source is estimated as

$$\boldsymbol{\sigma}_j(t,f) = 1/\left(\bar{\boldsymbol{a}}_j{}^H(f)\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^H\bar{\boldsymbol{a}}_j(f)\right), \tag{10}$$

where $\bar{\boldsymbol{a}}_j$ can be pre-measured RTF vectors or just the nominal SVs. In this work, the pre-measured RTFs are obtained when only the one of desired sources is active, and normalized as [26]

$$\bar{\boldsymbol{a}}_j(f) = e^{-\mathbb{j}\angle \boldsymbol{u}_{1j}(f)}\,\boldsymbol{u}_j(f)/\left\|\boldsymbol{u}_j(f)\right\|, \tag{11}$$

where $\boldsymbol{u}_j(f)$ is the first column of the left-singular vectors corresponding to the largest singular value, $\angle\boldsymbol{u}_{1j}(f)$ is the phase of the first element of $\boldsymbol{u}_j(f)$, $\|\cdot\|$ is the Euclidean norm and $\mathbb{j} = \sqrt{-1}$. The nominal SVs is written as

$$\bar{\boldsymbol{a}}_j(f) = \boldsymbol{d}(\theta_j, \phi_j, f)/d_1(\theta_j, \phi_j, f), \tag{12}$$

where d_1 is the SV of the first microphone, which has been chosen as the reference channel. The performance of two choices of $\bar{\boldsymbol{a}}_j$ is evaluated in Section III.

The SCM of the j-th source is obtained by the rank-1 approximation [27] as

$$\boldsymbol{\psi}_j(t,f) = \boldsymbol{\sigma}_j(t,f)\bar{\boldsymbol{a}}_j(f)\bar{\boldsymbol{a}}_j{}^H(f). \tag{13}$$

The INCM for the k-th source is reconstructed as

$$\boldsymbol{\Phi}_k(t,f) = \sum_{j=1,j\neq k}^{D}\boldsymbol{\psi}_j(t,f) + \sigma_n^2(t,f)\boldsymbol{I}_M. \tag{14}$$

The weighting coefficients used to separate the k-th source are given by the MVDR formula,

$$\boldsymbol{w}_k(t,f) = \eta^{-1}\bar{\boldsymbol{\Phi}}_k^{-1}(t,f)\bar{\boldsymbol{a}}_k(f), \tag{15}$$

where η is a scalar,

$$\eta = \bar{\boldsymbol{a}}_k{}^H(f)\bar{\boldsymbol{\Phi}}_k^{-1}(t,f)\bar{\boldsymbol{a}}_k(f), \tag{16}$$

and

$$\bar{\boldsymbol{\Phi}}_k(t,f) = \alpha_2\boldsymbol{\Phi}_k(t,f) + (1-\alpha_2)\boldsymbol{\Phi}_{\mathrm{mix}}(0,f), \tag{17}$$

where $0 \le \alpha_2 \le 1$. This allows for better adaptation to reverberant environments via a full-rank spatial covariance matrix. When $\alpha_2 = 0$, the method degenerates to a fixed beamformer. The separation matrix contains $\boldsymbol{w}_k$ as columns,

$$\mathbf{W}(t,f) = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \ldots \quad \boldsymbol{w}_D]. \tag{18}$$

The interfering speech attenuation factor (ISAF), ξ_{n}, is defined as

$$\xi_{\mathrm{n}} = 10\log_{10}(\sigma_I^2/\sigma_{FI}^2), \tag{19}$$

where σ_I^2 and σ_{FI}^2 are the variance of the interfering speech before and after beamforming. A larger ISAF value indicates better interference suppression.

III. Experimental Setup and Evaluation

A. Experimental Setup

A circular array with six microphones was placed in a 40 m × 50 m × 4 m reverberation room. The radius of the array was 0.035 m. The *x*0*y* plane was set on the floor of the room. The *x* and *y* coordinates of the center of the microphone array were chosen randomly in the range $[10,30]$ meters. The height of the microphone array was set to 1.65 m. The distance from the center of the microphone array to all sound sources was 1 m. The height of all sources was set to 1.70 m. Two sound sources were mixed in the experiment. The azimuth angle of the first source was chosen randomly. The azimuth of the second source was chosen to be 45° from the first source. The signal-to-interference ratio (SIR) was chosen randomly in the range $[0,3]$ dB. The reverberant room with a T_{60} time of 600 ms was simulated by the image source method [28] [29]. All audio was trimmed to 3 seconds at a sampling frequency of 16 kHz.

For far-field point sources in an anechoic chamber, the array manifold vector of a circular array can be written as

$$\boldsymbol{d}(\theta,\phi,f) = \begin{bmatrix} e^{(\mathbb{j}H_e\sin\theta\cos(\phi-\varphi_1))} \\ \vdots \\ e^{(\mathbb{j}H_e\sin\theta\cos(\phi-\varphi_M))} \end{bmatrix}, \tag{20}$$

where $H_e = 2\pi f r_{\mathrm{mic}}/c_0$, r_{mic} is the radius of the array, φ_i are the azimuth angles of the microphones,

$$\varphi_i = 60(i-1)\pi/180, i = 1,2,..M. \tag{21}$$

Parameters for the beamformer were $\alpha_1 = (L-1)/(L+1)$ and $\alpha_2 = 0.9$, where $L = 320$ was the number of STFT points. Half-overlapping Hanning windows were used. $\varepsilon(f)$ was chosen empirically to decrease linearly from 0.01 to 0.001 as the frequency increases from 0 to 8 kHz. The sound speed is 343 m/s. It is assumed in this work that the ground truth DOA of each sound source is known.

B. Results and Discussion

The proposed method was evaluated and compared with reference methods using five performance metrics: perceptual evaluation of speech quality (PESQ) score [30], short-time objective intelligibility (STOI) [31], scale-invariant signal-to-distortion ratio improvement (SI-SDRi) [32], ISAF, and online computational time (OCT). OCT is the average time for online processing a 6-channel, 200 ms signal on an Intel Xeon processor Cascade Lake@3.0GHz PC. In practice, the covariance matrix of the observed signal is updated every 10 ms, while the beamformer weights are updated every 200 ms. All of the results were averaged over 100 cases with different speeches and DOA combinations.

The methods for comparison were:

a) Adaptive MVDR beamformer without INCM reconstruction, i.e., $\alpha_2 = 0$.
b) Adaptive MVDR beamformer with INCM reconstruction as described in [22].
c) The proposed adaptive MVDR beamformer with INCM reconstruction using nominal SVs.
d) The proposed adaptive MVDR beamformer with INCM reconstruction using pre-measured RTFs.

For a fair comparison, the coefficients of the separation matrices for all these methods were initialized using the initial SCM, $\mathbf{\Phi}_{\text{mix}}(0, f)$ in (4). The reference signal for performance evaluation was chosen as the first channel of the desired speech received by the microphone array.

Table 1 lists the mean PESQ, STOI, SI-SDRi, and OCT for each method. From Table 1, we find that the quality of the speech obtained by MVDR beamformers with INCM reconstruction is significantly improved compared to that without INCM reconstruction.

TABLE I. PERFORMANCE COMPARISON

Method	Metrics				
	PESQ	STOI	ISAF (dB)	SI-SDRi (dB)	OCT (ms)
MVDR without reconstruction	1.41	0.86	8.50	1.88	**4.56**
MVDR with reconstruction as in [22]	1.75	0.93	13.41	**7.86**	951.20
Proposed method with reconstruction using SVs	1.78	0.93	14.11	7.41	60.15
MVDR with reconstruction using pre-measured RTFs	1.78	0.93	**14.61**	7.40	60.15

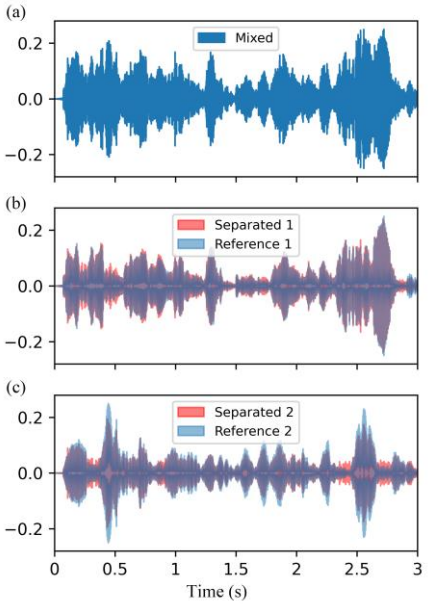


Fig. 1 Waveforms of the (a) mixed signal, (b) separated and clean reference signal of utterance 1, and (b) separated and clean reference signal of utterance 2.

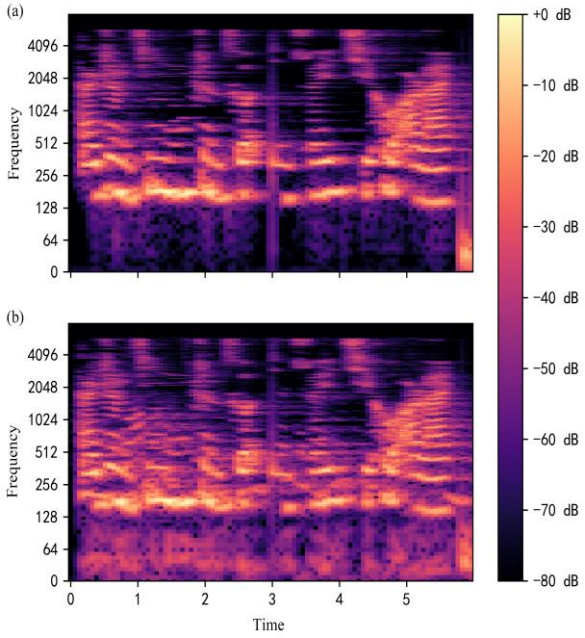


Fig. 2 Log-frequency power spectrogram of the desired speech: (a) the clean reference signal; (b) the signal estimated from the mixture.

The sound quality of the separated signals obtained by the proposed method is comparable to that of the adaptive MVDR

beamformer with INCM reconstruction described in [22]. However, the OCT of the proposed method is greatly reduced and less than the length of the audio, which is essential for real-time speech separation. Furthermore, we find that the proposed adaptive MVDR beamformer achieves the best ISAF performance using pre-measured RTFs.

Fig. 1 shows the time-domain waveforms of the mixed signal, separated utterance 1, separated utterance 2, and the clean reference signal for utterance 1 and 2, respectively. The waveforms for utterance 1 and utterance 2 are mostly recovered from the mixture, although there are some mismatches in some time frames.

Fig. 2 shows the log-frequency power spectrogram plots of the clean reference signal for utterance 1 and the separated utterance 1. The power spectrogram plot reveals a small amount of low-frequency noise and interference remaining after beamforming.

IV. Conclusions

We proposed an SVD-based INCM reconstruction method for beamforming-based speech separation. The SCM for each source is obtained by the rank-1 approximation using the nominal SVs or the pre-measured RTFs. The inverse of the covariance matrix of the mixed speech observations is obtained directly from the results of SVD. The INCM used to separate each source is reconstructed as the sum of the covariance matrices of interference and spherical isotropic noise. Experimental results show that the proposed adaptive beamformer based on INCM reconstruction for speech separation exhibits comparable sound quality in terms of PESQ, STOI, and SI-SDRi. ISAF is improved by using the pre-measured RTFs for INCM reconstruction. Most importantly, the computational time is greatly reduced and less than the frame length of the audio. Therefore, the proposed method is suitable for real-time speech separation.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (NSFC No. 52105128), in part by the Key Research Project of Zhejiang Lab (No. G2021NB0AL03), and in part by the Youth Foundation Project of Zhejiang Lab (K2023BA0AA03).

References

[1] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[2] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis" in *Speech Enhancement*, Berlin, Germany: Springer, pp. 371–402, 2005.

[3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] C. Xu, W. Rao, E. S. Chng and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[5] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018.

[6] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," arXiv:1905.06286v2, 2019.

[7] J. Lee, S. Chung, S. Kim, H. Kang, K. Sohn, "Looking into Your Speech: Learning Cross-modal Affinity for Audio-visual Speech Separation," in *Proc. IEEE/CVF CVPR*, pp. 1336–1345, 2021.

[8] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency", in *Proc. IEEE/CVF CVPR*, pp. 15 490–15 500, 2021.

[9] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing* (Vol. 1), Berlin, Germany: Springer-Verlag, 2008.

[10] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh and I. Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Proc. APSIPA-ASC*, Hollywood, CA, USA, 2012, pp. 1–10.

[11] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*, John Wiley & Sons, 2017.

[12] S. A. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *Signal Process.*, vol. 93, pp. 3264–3277, 2013.

[13] B. Widrow, K. Duvall, R. Gooch, and W. Newman, "Signal cancellation phenomena in adaptive antennas: causes and cures," *IEEE Trans. Antennas Propagat.*, vol. 30, no. 3, pp. 469–78, May. 1982.

[14] Y. T. Tsai, B. Su, Y. Tsao and S. -S. Wang, "Adaptive subspace-constrained diagonal loading," in *Proc. APSIPA-ASC*, Jeju, Korea (South), 2016, pp. 1-4.

[15] L. Chang and C. C. Yeh, "Performance of DMI and eigenspace-based beamformers," *IEEE Trans. Antennas Propagat.*, vol. 40, no. 11, pp. 1336–1347, Nov. 1992.

[16] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003.

[17] J. R. Guerci, "Theory and application of covariance matrix tapers for robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 47, pp. 977–985, Apr. 1999.

[18] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[19] M. Taseska and E. A. P. Habets, "Nonstationary Noise PSD Matrix Estimation for Multichannel Blind Speech Extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2223–2236, Nov. 2017.

[20] M. Ito, S. Araki, and T. Nakatani, "Permutation-free clustering of relative transfer function features for blind source separation," in *Proc. European Sig. Proc. Conf.*, 2015, pp. 409–413.

[21] Y. Gu and A. Leshem, "Robust adaptive beamforming based on Interference covariance matrix reconstruction and steering vector estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3881–3885, July 2012.

[22] X. Yuan and L. Gan, "Robust adaptive beamforming via a novel subspace method for interference covariance matrix reconstruction," *Signal Process.*, vol. 130, pp.233–242, Jul. 2016.

[23] J. Liu, W. Xie, C. Wang, and Q. Wan, "Robust widely beamforming via the technique of shrinkage for steering vector estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*., 2018, pp. 3349–3353.

[24] S. Mohammadzadeh, V. H. Nascimento, R. C. de Lamare and O. Kukrer, "Maximum Entropy-Based Interference-Plus-Noise Covariance Matrix Reconstruction for Robust Adaptive Beamforming," *IEEE Signal Process. Lett.*, vol. 27, pp. 845–849, Jun. 2020.

[25] K. Nakamura, K. Nakadai and G. Ince, "Real-time super-resolution Sound Source Localization for robots," *IEEE/RSJ Int. Conf. Intell. Robot Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 694-699.

[26] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Sig. Proc.*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[27] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Au., Sp., Lang. Proc.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[29] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A Python library for room impulse response simulation with GPU acceleration," *Multimed. Tools Appl., vol* 80, pp. 5653–5671, 2021.

[30] "P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec," ITU-T Std. P.862.2, 2007.

[31] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE Trans. Audio, Speech, Lang. Process*., 2016.

[32] R. J. Le, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done? ," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Correlated Multi-Level Speech Enhancement for Robust Real-World ASR Applications Using Mask-Waveform-Feature Optimization

Hang Chen[†], Jun Du[†*], Zhe Wang[†], Chenxi Wang[†], Yuling Ren[‡], Qinglong Li[‡], Ruibo Liu[‡], Chin-Hui Lee[§]
[†] University of Science and Technology of China, China
[‡] China Mobile Online Services Company Limited, China
[§] Georgia Institute of Technology, USA
✉ jundu@ustc.edu.cn

***Abstract*—Our proposed correlated multi-level optimization approach enhances speech recognition performance for high-performance acoustic models in real-world applications. By combining mean squared error of mask, scale-invariant source-to-noise ratio, and cross-entropy loss functions, as well as adopting Pearson correlation coefficient as a part of the optimization goal to measure the correlation between them, our approach aims to not only reduce the value of each loss during training but also increase the correlation between them. Experimental results on continuous Mandarin recognition in mobile phone scenarios show that our approach achieves a relative reduction of about 25.29% in the average character error rate across five signal-to-noise ratio levels. Notably, our approach improves objective perception qualities and intelligibility measures, as well as recognition accuracies, surpassing some advanced speech enhancement techniques in the context of automatic speech recognition.**

I. Introduction

Speech enhancement (SE) aims to extract clean speech from noisy speech by reducing or suppressing background noises [1]. However, conventional SE techniques [2]–[6] often struggle to estimate noise characteristics and track non-stationary noises in real-world acoustic conditions. In recent years, data-driven SE approaches based on deep neural networks (DNNs) have gained popularity [7]–[9]. Thanks to their powerful modeling capability, DNNs can effectively capture the complex relationship between noisy and clean speech [10], and estimate a mask to suppress noise [11]–[13]. These approaches are typically optimized using minimum-mean-squared-error (MMSE) based loss functions and have improved subjective and objective metrics related to perceptual quality.

SE can also be used as a front-end to automatic speech recognition (ASR) systems, which often suffer from degraded performance in real-world environments with adverse acoustic conditions, such as channel distortion and ambient noise. However, most DNN-based SE models have limited impact on reducing ASR word error rate (WER) [7], [14]–[16]. In fact, some studies have reported that SE can even increase WERs when the back-end is a DNN-based model with multi-condition training [17]–[20], due to unseen distortions in enhanced speech that cause mismatches between the training and testing.

One potential solution is to retrain the back-end ASR system by including enhanced speech in the training set [21]. Another approach is to jointly train the front-end SE model and the back-end acoustic model [22]–[24], using methods such as teacher-student learning [25], adversarial training [26], or a combination of both [27]. However, in many real-world applications, the ASR models may be provided by a third party and their parameters cannot be re-estimated. Therefore, it is important to explore advanced loss functions that can optimize the SE models while using fixed ASR models.

Given the aforementioned constraints, reinforcement learning (RL) [28] has emerged to optimize SE models directly based on recognition results. For example, in [29], two DNNs using MetricGAN [30] were developed, one for speech enhancement and the other for mimicking the output CERs obtained from an ASR system. In [31], a cross-entropy-guided measure (CEGM) was proposed to evaluate ASR performance and optimize SE model. CEGM calculates the cross entropy of state posteriors between enhanced and clean speech, which are obtained from the outputs of the DNN-HMM (Hidden Markov Model) acoustic model. However, it was shown that CEGM is challenging for enhancement models to improve recognition performance when the back-end acoustic models are more powerful, as is typically the case in real-world ASR applications. Additionally, CEGM-enhanced speech is no longer intelligible. Therefore, there is a pressing need for SE models that can improve both recognition performance and perceptual quality.

This paper addresses the issue that using the conventional CEGM alone is inadequate for optimizing SE models with high-performance back-end ASR models. We propose a correlated multi-level (CML) optimization approach that utilizes a sequential loss function comprising four terms: mean squared error between the estimated and ideal ratio masks (MSE-M), scale-invariant source-to-noise ratio (SISNR) between the reconstructed and clean speech waveforms, cross-entropy between the state posteriors of the acoustic model evaluated using the reconstructed and clean speech features (CEGM), and Pearson correlation coefficients (PCCs) [32] between MSE-M, SISNR, and CEGM. By utilizing these various loss functions,

*Corresponding author

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

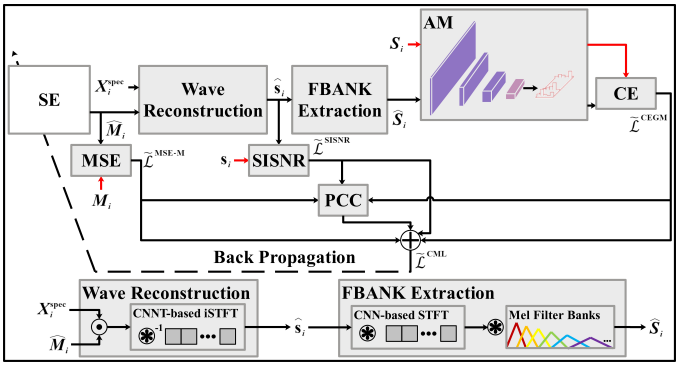


Fig. 1. Illustration of correlated multi-level speech enhancement.

our approach aims to enhance the perceptual quality and recognition accuracy of the enhanced speech.

CEGM facilitates the transfer of acoustic knowledge from pre-trained models to the SE model, reducing distortions that impact recognition accuracy. MSE-M and SISNR address optimization challenges brought by the back-end acoustic model, utilizing magnitude masks and reconstructed waveforms. PCC considers the interplay between different loss functions, working to increase their correlation. Moreover, our proposed CML approach overcomes the limitations of CEGM by not solely relying on back-end models and explicitly modeling the correlation between different loss functions. Our experimental results show that the proposed approach significantly improves speech recognition performance and our reconstructed speech exhibits improved perceptual quality compared to the conventional MMSE-enhanced speech.

II. Correlated Multi-Level Speech Enhancement

Different from CEGM in which the front-end SE model is limited to output the desired acoustic features to the back-end ASR model, CML processes the noisy log power spectra (LPS) feature $X_i \in \mathbb{R}^{T\times F}$ and outputs a magnitude mask $\widehat{M}_i \in \mathbb{R}^{T\times F}$:

$$\widehat{M}_i = \mathcal{F}(X_i; W) \tag{1}$$

where i is the sample index in the current batch. $\mathcal{F}$ and W denote the SE model and the corresponding parameter set, respectively. The values of $\widehat{M}_i$ range from 0 to 1. T and F denote the time length and the number of frequency bins of spectrum feature, respectively. In this study, F is set as 201 by default. Then the wave reconstruction module uses $\widehat{M}_i$ to filter the noisy spectrum $X_i^{\mathrm{S}} \in \mathbb{C}^{T\times F}$ and the filtered spectrum is fed to a 1D transposed convolution layer to reconstruct waveform:

$$\widehat{\mathbf{s}}_i = \mathcal{U}(X_i^{\mathrm{spec}}, \widehat{M}_i; W_{\mathrm{stft}}) \tag{2}$$

where $\widehat{\mathbf{s}}_i \in \mathbb{R}^L$ denotes the reconstructed wave of length L. $\mathcal{U}$ denotes the wave reconstruction module and W_{stft} denotes the forward weight of STFT which is also the parameter set of the 1D transposed convolution. The feature extraction module $\mathcal{E}$ takes $\widehat{\mathbf{s}}_i$ as input and outputs the FBANK feature $\widehat{S}_i \in \mathbb{R}^{T\times F_{\mathrm{mel}}}$:

$$\widehat{S}_i = \mathcal{E}(\widehat{\mathbf{s}}_i; W_{\mathrm{fe}}) \tag{3}$$

where W_{fe} denotes the corresponding parameter set of $\mathcal{E}$, respectively. F_{mel} denotes the numbers of triangular filters and is set as 40 by default in this study. As shown in Figure 1, the FBANK extraction module consists of two 1D convolution layers, where the first 1D convolution layer is used to perform online STFT, and the second 1D convolution layer is used as Mel filter banks. Finally, we adopt the acoustic model of the hybrid DNN-HMM ASR model to perform deep feature extraction. The low-level acoustic feature is mapped to the high-level presentation, i.e., the clustered state posterior probability:

$$\mathrm{p}(\mathbf{h}|\widehat{S}_i) = \mathcal{G}(\widehat{S}_i; W_{\mathrm{am}}) \tag{4}$$

where $\mathbf{h} = \{h_0, ..., h_{I-1}\}$ denotes the clustered HMM states, $\mathrm{p}(\mathbf{h}|\widehat{S}_i) \in \mathbb{R}^{T\times I}$ denotes the posteriors of $\mathbf{h}$ given $\widehat{S}_i$. $\mathcal{G}$ and W_{am} denote the acoustic model and the corresponding parameter set, respectively.

The proposed correlated multi-level loss function $\mathcal{L}_{\mathrm{CML}}$ is calculated as follows:

$$\mathcal{L}^{\mathrm{CML}} = \sum_{i=0}^{N-1} \frac{\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}} - \widetilde{\mathcal{L}}_i^{\mathrm{SISNR}} + \widetilde{\mathcal{L}}_i^{\mathrm{CEGM}}}{N} - \mathcal{L}^{\mathrm{PCC}} \tag{5}$$

where $\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}}$ is the normed MSE between $\widehat{M}_i$ and ideal ratio mask (IRM) $M_i \in \mathbb{R}^{T\times F}$:

$$\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}} = \mathcal{L}_i^{\mathrm{MSE-M}} / \max_{k\in \text{train set}} (|\mathcal{L}_k^{\mathrm{MSE-M}}|) \tag{6}$$

$$\mathcal{L}_i^{\mathrm{MSE-M}} = \sum_{t=0}^{T-1}\sum_{j=0}^{F-1} \frac{(\widehat{m}_{i,t,j} - m_{i,t,j})^2}{TF} \tag{7}$$

where $\widehat{m}_{i,t,j}$ and $m_{i,t,j}$ are the values at t-th time step and j-th frequency bin of $\widehat{M}_i$ and M_i, respectively. N denotes the batch size and is set as 32 by default in this study. k is the sample index in the whole train set. To prevent the loss weight from being affected by the magnitude scale, we divide each element of $\mathcal{L}_i^{\mathrm{MSE-M}}$ by the maximum amplitude calculated from the whole training set, ensuring that the values of $\widetilde{\mathcal{L}}_i^{\mathrm{MSE-M}}$ are also in the range of 0 to 1.

$\widetilde{\mathcal{L}}_i^{\mathrm{SISNR}}$ denotes the normed SISNR between $\widehat{\mathbf{s}}_i$ and the corresponding clean waveform $\mathbf{s}_i \in \mathbb{R}^L$:

$$\widetilde{\mathcal{L}}_i^{\mathrm{SISNR}} = \mathcal{L}_i^{\mathrm{SISNR}} / \max_{k\in \text{train set}} (|\mathcal{L}_k^{\mathrm{SISNR}}|) \tag{8}$$

$$\mathcal{L}_i^{\mathrm{SISNR}} = 10\mathbf{log}\frac{\sum_{\tau=0}^{L-1} s_{i,\tau}^2}{\sum_{\tau=0}^{L-1}[\widetilde{s}_{i,\tau} - s_{i,\tau}]^2} \tag{9}$$

where $\widetilde{s}_{i,\tau} = \widehat{s}_{i,\tau}(\sum_{\tau=0}^{L-1} s_{i,\tau}^2)/(\sum_{\tau=0}^{L-1} \widehat{s}_{i,\tau} s_{i,\tau})$. $\widehat{s}_{i,\tau}$ and $\mathrm{s}_{\mathrm{i},\tau}$ are values at the τ-th time step of $\widehat{\mathbf{s}}_i$ and $\mathbf{s}_i$, respectively. We also apply the same normalization procedure to $\mathcal{L}_i^{\mathrm{SISNR}}$ as we do to $\mathcal{L}_i^{\mathrm{MSE-M}}$.

$\widetilde{\mathcal{L}}_i^{\mathrm{CEGM}}$ is the normed cross entropy between enhanced and clean representations, i.e. $\mathrm{p}_t(h_n|\widehat{S}_i)$ and $\mathrm{p}_t(h_n|S_i)$, which are posteriories of h_n given $\widehat{S}_i$ and S_i at the t-th time step:

$$\widetilde{\mathcal{L}}_i^{\mathrm{CEGM}} = \mathcal{L}_i^{\mathrm{CEGM}} / \max_{k\in \text{train set}} (|\mathcal{L}_k^{\mathrm{CEGM}}|) \tag{10}$$

$$\mathcal{L}_i^{\mathrm{CEGM}} = -\sum_{t=0}^{T-1}\sum_{n=1}^{I-1}\frac{\mathrm{p}_t(h_n|S_i)\mathbf{log}\mathrm{p}_t(h_n|\widehat{S}_i)}{T}. \quad (11)$$

The same normalization procedure is applied to $\mathcal{L}_i^{\mathrm{CEGM}}$ as we do to $\mathcal{L}_i^{\mathrm{MSE}}$ and $\mathcal{L}_i^{\mathrm{SISNR}}$.

Optimizing multiple learning goals simultaneously can result in suboptimal performance. Focusing on reducing one or more loss values to improve performance metrics in evaluation can come at the expense of other metrics. Especially in our work, optimizing for MSE and SISNR can significantly enhance the perceptual quality of speech but may not directly benefit recognition accuracy. Conversely, optimizing for CEGM may improve accuracy but compromise the enhanced speech's perceptual quality. To take into account the interplay between different loss functions, we propose using the PCC as a part of optimization goal to measure the correlation between two loss functions. $\mathcal{L}^{\mathrm{CML}}$ aims to not only reduce the value of each loss during training but also increase the correlation between them. $\mathcal{L}_{\mathrm{PCC}}$ is calculated using all sample in the current batch:

$$\begin{aligned}\mathcal{L}^{\mathrm{PCC}} =&\rho(\widetilde{\mathcal{L}}^{\mathrm{MSE-M}}, \widetilde{\mathcal{L}}^{\mathrm{SISNR}}) + \rho(\widetilde{\mathcal{L}}^{\mathrm{MSE-M}}, \widetilde{\mathcal{L}}^{\mathrm{CEGM}}) \\ &+ \rho(\widetilde{\mathcal{L}}^{\mathrm{CEGM}}, \widetilde{\mathcal{L}}^{\mathrm{SISNR}})\end{aligned} \quad (12)$$

$$\rho(\mathcal{L}^a, \mathcal{L}^b) = \frac{\sum_{i=1}^{N}(\mathcal{L}_i^a - \bar{\mathcal{L}^a})(\mathcal{L}_i^b - \bar{\mathcal{L}^b})}{\sqrt{\sum_{i=1}^{N}(\mathcal{L}_i^a - \bar{\mathcal{L}^a})^2}\sqrt{\sum_{i=1}^{N}(\mathcal{L}_i^b - \bar{\mathcal{L}^b})^2}} \quad (13)$$

Where $\bar{\mathcal{L}^a} = \sum_{i=1}^{N}\mathcal{L}_i^a/N$ and $\bar{\mathcal{L}^b} = \sum_{i=1}^{N}\mathcal{L}_i^b/N$. $\mathcal{L}^a, \mathcal{L}^b \in \{\widetilde{\mathcal{L}}^{\mathrm{MSE-M}}, \widetilde{\mathcal{L}}^{\mathrm{SISNR}}, \widetilde{\mathcal{L}}^{\mathrm{CEGM}}\}$ and $\mathcal{L}^a \neq \mathcal{L}^b$. PCC measures the strength and direction of the correlation.

In this study, we selected the No Embedding Aware Speech Enhancement (NoEASE) [33] model as our SE model. We trained the model using the Adam optimizer [34] for 100 epochs, with early stopping when the validation loss did not improve for 10 epochs. We set the initial learning rate to 0.0003 and reduced it by half during training if the validation loss did not improve for 3 epochs. The best model was selected based on the lowest validation loss.

III. Experiments and Result Analysis

A. Data Set Description

To evaluate the effectiveness of the proposed CML approach, we conducted experiments using two sets of simulated noisy speech data. The first dataset, TcdTimit-Noisy, was used in a previous study [33] and comprises noisy speech with all video components removed. The second dataset is an in-house Mandarin corpus collected using mobile phones in real-world conditions and is referred to as Mandarin-Noisy. This corpus consists of 7,081 utterances from different speakers, with 6,900 utterances randomly selected for training, 85 for validation, and 96 for testing. The noise data used to generate noisy-clean pairs were recorded from real bathrooms, kitchens, balconies, and living rooms, resulting in approximately 81 hours of training data, 3 hours for validation, and 3 hours for testing. Each subset contains different speakers and noise recording rooms, and there are five SNR levels (15 dB, 10 dB, 5 dB, 0 dB, -5 dB) in the training, validation, and test sets.

Table 1
Average PER comparison of SE models trained with different criteria on the TcdTimit-Noisy test set at different SNRs.

Criterion	**PER of InASR** (in %) ↓					
SNR(in dB)	−5	0	5	10	15	Avg.
Noisy	74.61	62.75	47.35	36.71	29.72	50.23
MSE-M	74.45	61.89	47.23	37.01	29.82	50.08
SISNR	72.04	60.42	47.50	36.38	29.33	49.13
CEGM	72.71	58.85	43.93	33.41	28.35	47.25

B. Back-end Acoustic Model

In [31], we use a hybrid DNN-HMM ASR model as the back-end, which is trained on the same TcdTimit-Noisy training set as the SE model. The back-end model consists of an acoustic model with the same structure as [35] and a 2-gram phone-based language model, denoted as InASR.

We then evaluate the enhanced speech using a real-world speech recognition application, denoted as RealASR. The RealASR model is also a hybrid DNN-HMM ASR model, consisting of an acoustic model, a language model, and a pronunciation dictionary. The acoustic model comprises 20 ResBlocks followed by a 12-layer variant of Transformer [36], where two ResBlocks follow each Multi-Head Attention. The ResBlock has a structure similar to the basic block in [37], but with a minor change where the Leaky ReLU activation replaces the ReLU activation. We use over 100,000 hours of Mandarin data collected in real scenarios to train the acoustic model. We adopt a 4-gram word-based language model trained with more than 500 million sentences and a large pronunciation dictionary containing over 600,000 Chinese words for decoding.

C. Evaluation Metrics

In this study, we evaluate the SE models in two ways: perceptual quality and recognition performance. Perceptual quality is assessed using the Perceptual Evaluation of Speech Quality (PESQ) [38] and Short-Time Objective Intelligibility (STOI) [39] metrics. PESQ is a speech quality estimator that predicts the mean opinion score of a speech quality listening test under certain degradations. The PESQ score ranges from -0.5 to 4.5, with a higher score indicating better speech quality. STOI measures speech intelligibility, with scores typically ranging from 0 to 1, where higher scores indicate better intelligibility.

Recognition performance is evaluated using the phone error rate (PER) and character error rate (CER) metrics. Lower PER and CER values indicate better recognition performance.

D. Recognition Performance Degradation for Real-World ASR Back-end

We first evaluated the recognition performance of SE models optimized by MSEM, SISNR and CEGM with InASR. All SE models were trained using the TcdTimit-Noisy training set, and the results are shown in Table 1. We found that

Table 2
AVERAGE CER COMPARISON OF SE MODELS TRAINED WITH DIFFERENT CRITERIA ON THE MANDARIN-NOISY TEST SET AT DIFFERENT SNRS.

Criterion	**CER of RealASR** (in %) ↓					
SNR(in dB)	−5	0	5	10	15	Avg.
Noisy	30.21	15.58	8.57	6.00	4.87	13.05
MSE-M	31.07	14.58	8.03	5.66	4.57	12.78
SISNR	30.93	14.64	7.89	5.77	4.88	12.82
CEGM	32.38	19.16	9.87	7.50	6.11	15.00
CML	22.02	11.28	6.25	4.82	4.39	9.75
w/o PCC	24.37	12.72	6.58	5.02	4.55	10.64
w/o CEGM	28.73	14.06	6.87	5.17	4.76	11.92
w/o SISNR	25.13	14.24	6.67	5.00	4.58	11.12
w/o MSE-M	24.92	14.22	6.57	5.01	4.56	11.06

CEGM outperformed MSEM and SISNR in PER across all SNR levels. Previous work [31] suggested that reducing distortions in the high-level representations extracted from low-level acoustic features by the back-end ASR model improves ASR performance.

We then used RealASR to evaluate the recognition performance of the same set of SE models. Results are shown in Table 2. MSEM and SISNR showed minimal performance improvements over unprocessed speech ("Noisy") at most SNR levels, with a performance degradation at −5 dB. This observation is consistent with previous studies [29], [31], indicating that the MMSE-based loss function directly improves perceptual quality rather than the back-end ASR performance. CEGM also yielded higher CER values than Noisy at all SNRs, suggesting that using CEGM to optimize SE models under high-performance acoustic models is challenging.

In contrast to the CEGM criterion, CML estimates a magnitude mask rather than the desired acoustic features of the back-end ASR model. The unbounded property of the acoustic feature makes it often difficult to optimize due to the large search space. On the other hand, the value of each bin in the mask ranges from 0 to 1. Two supervisions from low-level acoustic features (magnitude masks and reconstructed waveforms) can effectively avoid over-fitting acoustic knowledge from the back-end being passed to the SE model. Coupled with the active modeling of the correlation between distortions on low-level and high-level features, the whole optimization difficulties have been significantly reduced. In contrast to the CEGM criterion, the CML approach estimates a magnitude mask rather than the desired acoustic features of the back-end ASR model. This is beneficial as the unbounded property of acoustic features often makes optimization challenging due to the large search space. In comparison, the value of each bin in the mask ranges from 0 to 1, making optimization more feasible. Furthermore, the use of two supervisions from low-level acoustic features, namely magnitude masks and reconstructed waveforms, effectively prevents over-fitting of acoustic knowledge from the back-end being passed to the SE model. Additionally, the active modeling of the correlation between distortions on low-level and high-level features significantly reduces the overall optimization difficulties. Our

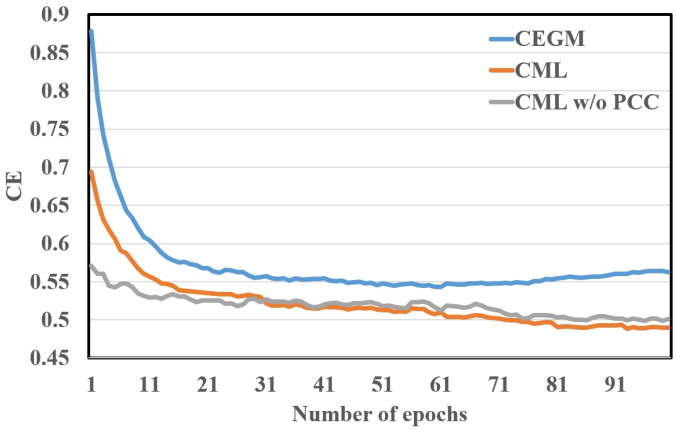


Fig. 2. A comparison of learning curves between CEGM and CML.

experimental results in Table 2 show that CML yielded more improvements when compared to MSEM, SISNR, and CEGM at all SNRs, especially at lower SNRs.

E. Ablation Study

We conducted an ablation study, as presented in Table 2, to analyze the contribution of each loss item to the overall performance. Our results show that each loss item significantly contributed to the performance improvement. Specifically, removing PCC, CEGM, SISNR, and MSE-M resulted in a relative increase in CER by 9.19%, 22.21%, 14.06%, and 13.36%, respectively. These findings demonstrate the importance of leveraging multiple loss functions through our proposed CML approach to achieve the best possible outcome.

To evaluate the optimization performance of our proposed method, we compared the learning curves of the cross-entropy (CE) between CEGM and CML on the validation set, as shown in Fig. 2. We observed that the learning curve of CEGM started to rise after the 66-th epoch, whereas the learning curve of CML continued to decline and consistently generated smaller CEs on the validation set than CEGM. Notably, when we removed PCC from the CML loss function, the learning curve converged to a higher CE value earlier, indicating the importance of explicitly modeling the interplay between different loss functions in optimizing the SE model.

F. Perceptual Quality Evaluation

To evaluate the perceptual quality of the reconstructed speech, we compared the average PESQ and STOI scores among the four different enhancement methods (Noisy, MSE-M, SISNR, and CML) at different SNR levels. We plotted the results in Figure 3.

From the plot, we can observe that SISNR achieves a higher average STOI score compared to MSE-M (79.02% vs. 78.02%), while the reciprocal results are observed for PESQ (2.53 vs. 2.56). This suggests that MSE-M is better at preserving the original speech signal's perceptual quality, while SISNR is better at improving intelligibility.

Furthermore, the complementarity of the methods is also reflected in their performance across different SNR levels. Specifically, SISNR performs better than MSE-M at low SNRs

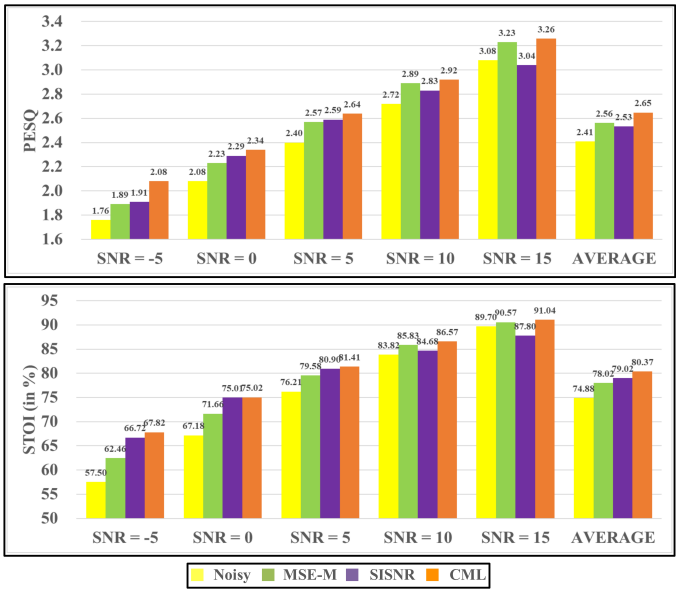


Fig. 3. PESQ and STOI comparisons of SE models trained with different criteria on the Mandarin-Noisy test set at different SNRs.

(-5 and 0 dBs) across both PESQ and STOI metrics, while reciprocal results are observed at high SNRs (10 and 15 dBs). This indicates that SISNR is better suited for improving the perceptual quality of speech signals with a high level of noise, while MSE-M is more effective at improving the quality of speech signals with low to moderate noise levels.

Finally, the CML method takes advantage of the complementarity of MSE-M and SISNR and achieves the best performance across both PESQ and STOI metrics at all SNR levels. This suggests that combining the two methods can improve the overall quality of the reconstructed speech signals, regardless of the noise level.

IV. Conclusion

This paper proposed a novel correlated multi-level (CML) optimization approach for speech enhancement, which addresses the challenge of optimizing SE models with high-performance back-end ASR models. By leveraging multiple loss functions that model distortions on different representations and explicitly modeling their interplay, the proposed CML approach significantly improves the recognition performance and perceptual quality of the enhanced speech. Our empirical results demonstrate the effectiveness of CML in the context of a high-performance acoustic model from a real-world application. Future work could focus on exploring network structures to reduce complexity and enable real-time operations on low-power portable devices. In addition, extending the proposed CML approach to handle other types of distortions, such as reverberation, would be an interesting direction for future research.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427

References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. 2007.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 466–475, 2003.

[7] T. Higuchi, N. Ito, T. Yoshioka, *et al.*, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *Proc. ICASSP 2016*, 2016, pp. 5210–5214.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP 2016*, 2016, pp. 196–200.

[9] H. Erdogan, J. R. Hershey, S. Watanabe, *et al.*, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *Proc. Interspeech 2016*, 2016, pp. 1981–1985.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[11] Y. Wang and D.-L. Wang, "Towards scaling up classification-based speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[12] Y. Wang, A. Narayanan, and D.-L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] Y. Zhao, Z.-Q. Wang, and D.-L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 53–62, 2019.

[14] H.-J. Hsieh, B. Chen, and J.-W. Hung, "Employing median filtering to enhance the complex-valued acoustic spectrograms in modulation domain for noise-robust

speech recognition," in *Proc. ISCSLP 2016*, 2016, pp. 1–5.

[15] D. Yu, L. Deng, J. Droppo, *et al.*, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *Proc. ICASSP 2008*, 2008, pp. 4041–4044.

[16] M. L. Seltzer, D. Yu, and Y.-Q. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7398–7402.

[17] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[18] S.-J. Chen, A. S. Subramanian, H.-N. Xu, *et al.*, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Proc. Interspeech 2018*, 2018, pp. 1571–1575.

[19] T. Menne, R. Schlüter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust asr," in *Proc. ICASSP 2019*, 2019, pp. 6660–6664.

[20] Y.-H. Tu, J. Du, L. Sun, *et al.*, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," in *Proc. Interspeech 2017*, 2017, pp. 394–398.

[21] Z.-Q. Wang and D. L. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[22] Z.-Q. Wang and D. L. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 2839–2843.

[23] W.-Y. Zhang, J. Shi, C.-D. Li, *et al.*, "Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions," in *Proc. WASPAA 2021*, 2021, pp. 146–150.

[24] Y. Koizumi, S. Karita, A. Narayanan, *et al.*, "SNRi Target Training for Joint Speech Enhancement and Recognition," in *Proc. Interspeech 2022*, 2022, pp. 1173–1177.

[25] S. Watanabe, T. Hori, J. Le Roux, *et al.*, "Student-teacher network learning with enhanced features," in *Proc. ICASSP 2017*, 2017, pp. 5275–5279.

[26] B. Liu, S. Nie, S. Liang, *et al.*, "Jointly Adversarial Enhancement Training for Robust End-to-End Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 491–495.

[27] L. Wu, H. T. Chen, L. Wang, *et al.*, "Speaker-Invariant Feature-Mapping for Distant Speech Recognition via Adversarial Teacher-Student Learning," in *Proc. Interspeech 2019*, 2019, pp. 431–435.

[28] Y.-L. Shen, C.-Y. Huang, S.-S. Wang, *et al.*, "Reinforcement learning based speech enhancement for robust speech recognition," in *Proc. ICASSP 2019*, 2019, pp. 6750–6754.

[29] R. Sawata, Y. Kashiwagi, and S. Takahashi, "Improving character error rate is not equal to having clean speech: Speech enhancement for asr systems with black-box acoustic models," in *Proc. ICASSP 2022*, 2022, pp. 991–995.

[30] S.-W. Fu, C.-F. Liao, Y. Tsao, *et al.*, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML 2019*, 2019, pp. 2031–2041.

[31] L. Chai, J. Du, Q.-F. Liu, *et al.*, "A cross-entropy-guided measure (cegm) for assessing speech recognition performance and optimizing dnn-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 106–117, 2021.

[32] P. Y. Chen, M. Smithson, P. M. Popovich, *et al.*, *Correlation: Parametric and nonparametric measures*. Sage, 2002.

[33] H. Chen, J. Du, Y. Hu, *et al.*, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Network*, vol. 143, pp. 171–182, 2021.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR 2015*, 2015.

[35] H. Chen, J. Du, Y.-S. Dai, *et al.*, "Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis," in *Proc. INTERSPEECH 2022*, 2022.

[36] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[37] K. M. He, X. Y. Zhang, S. Q. Ren, *et al.*, "Deep residual learning for image recognition," in *Proc. CVPR 2016*, 2016, pp. 770–778.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, *et al.*, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP 2001*, vol. 2, 2001, pp. 749–752.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

CASA-Net: Cross-attention and Self-attention for End-to-End Audio-visual Speaker Diarization

Haodong Zhou[1], Tao Li[2], Jie Wang[1], Lin Li[1], Qingyang Hong[2]
[1] School of Electronic Science and Engineering, Xiamen University, China
[2] School of Informatics, Xiamen University, China
E-mail: {lilin, qyhong}@xmu.edu.cn

***Abstract*—Audio-visual speaker diarization (AVSD) is a critical technique that segments audio-visual signals and assigns them to multiple speakers in practical scenarios. Thus, how to efficiently extract the cross-modal information from audio-visual signals is essential for AVSD. In this paper, we present CASA-Net, an embedding fusion method for end-to-end AVSD system. CASA-Net incorporates cross-attention (CA) module to capture cross-modal information in audio-visual signals, and utilizes self-attention (SA) module to learn the context information among audio-visual frames. On the development set of the Multimodal Information Based Speech Processing (MISP) Challenge 2022, the proposed CASA-Net achieved a diarization error rate (DER) of 13.18%, which is 1.56% lower compared to the concatenation (Concat) method. To further enhance performance, we utilized beamforming to integrate the available multi-channel audio information, along with data augmentation. By fusing multiple systems, we ultimately achieved a DER of 12.60% on the development set.**

I. Introduction

Audio-only speaker diarization segments and classifies a speech signal from multiple speakers, which assigns different parts of the signal to different speakers. This process can be viewed as a technique for speech signal separation and clustering, which solves the problem of "who spoke when" and is essential for practical scenarios such as audio conferences, conversations and interviews [1][2]. Several systems [3][4] provides feasible solutions for the audio-only speaker diarization.

The discovery that visual cues can aid in human speech perception [5] has sparked interest in exploring video information [6]–[9] to improve the challenge of audio-only speaker diarization. The Multimodal Information Based Speech Processing (MISP) Committee has released the first large-scale multi-microphone conversational Chinese audio-visual corpus [10]. Additionally, the MISP 2022 Challenge introduced the track1 challenge, specifically focusing on addressing the Audio-Visual Speaker Diarization (AVSD) task by labeling speech timestamps according to the identify of speakers. This challenge aims to encourage participants to leverage audio and visual information and explore the fusion technology of multiple modalities to enhance the AVSD system's robustness in complex scenarios.

The study conducted by [8] reveals that Visual Voice Activity Detection (V-VAD) is more adept at handling overlapped speech segments than audio-only speaker diarization within the MISP corpus. V-VAD is an advanced technology that utilizes visual information to detect voice activity. Its objective is to determine the presence, absence, or temporal boundaries of speech activities by analyzing facial expressions, lip movements, and other visual features in video images [11][12]. Simultaneously utilizing temporal context information [13][14] in videos is beneficial to learn visual representations. However, the effectiveness of V-VAD can be compromised when visual features are absent due to occlusions and off-screen speakers. Additionally, the observation of lip movements can be impeded by the speaker's head movements and environmental factors, whereas the audio remains unaffected. Therefore, it is crucial to explore methods that capture the correlation between audio and lip movements, as well as the context information from audio-visual frames, to enhance the performance of audio-visual speaker diarization.

Motivated by the attention mechanism for active speaker detection(ASD) [15] and context information extraction [16], we proposed our Audio-visual speaker diarization system for the MISP 2022 track1 challenge. We utilized the outputs of the audio and visual modality encoders as inputs to the Cross-Attention (CA) module to capture cross-modal information, and then used Self-Attention (SA) module to learn context relationships between temporal frames. As the MISP corpus employs multi-channel recording equipment, we employed beamforming to integrate multi-channel far-field audio to enhance robustness. Furthermore, to address audio noise and lip image interference, we applied various data augmentation techniques such as external audio noise and image transformations during the training process. We presented the results based on the publicly accessible development set. With CASA-Net, we achieved a diarization error rate (DER) of 13.18% on the development set, with a 1.56% reduction compared to the Concat method. Subsequently, our performance was further improved by beamforming and data augmentation. Finally, using DOVER-lap [17] to fuse our three systems, we obtained a DER result of 12.60%.

The following sections of our paper will be organized as follows: In Section 2, we introduce some related works of the baseline, including encoders, extractor and decoder. Section 3 describes the proposed network framework and optimization. In Section 4, we demonstrate the experimental setup and analyze the results, respectively. The Section 5 concludes the research.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

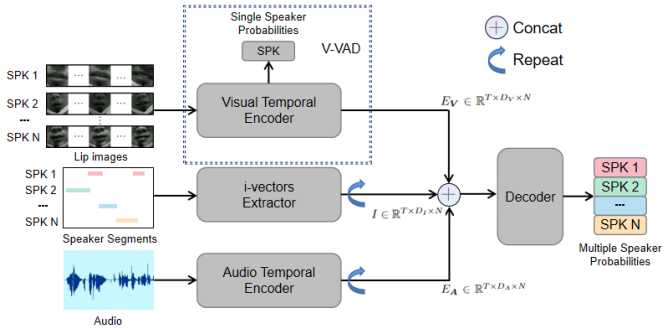


Fig. 1. The baseline [8] of MISP 2022 Challenge.

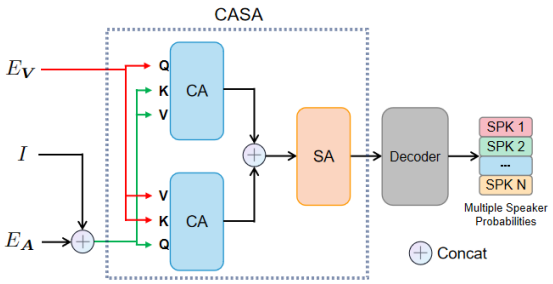


Fig. 2. The framework of CASA network.

II. Related Works

The MISP 2022 Challenge adopts [8] as the baseline, which includes multiple modules: visual and audio temporal encoders, *i-vector* extractor and decoder. Fig. 1 shows its structure.

A. Visual temporal encoder

The visual temporal encoder functions as visual voice activity detection (V-VAD). We cut the video into multiple blocks, and the lip images within each block are served as inputs $X_{\boldsymbol{V}} \in \mathbb{R}^{T\times W\times H\times N}$ to visual temporal encoder. Then we obtained the visual embedding outputs as $E_{\boldsymbol{V}} \in \mathbb{R}^{T\times D_V\times N}$, where T is the total number of frames, W and H are the width and height of image, D is the dimension, and N is the number of speakers.

B. Audio temporal encoder

We used the FBANKs as inputs for the audio temporal encoder, utilizing four-layer CNNs that consist of 2D Convolution, BatchNorm and ReLU for each layer. Then we got the outputs $\hat{E}_{\boldsymbol{A}} \in \mathbb{R}^{T\times D_A}$ through a fully connected layer. Finally, we repeated N times to obtain $E_{\boldsymbol{A}} \in \mathbb{R}^{T\times D_A\times N}$ to align with $E_{\boldsymbol{V}}$.

C. I-vector extractor

In [8], *i-vector* is claimed to solve alignment problems when lips are missing due to occluded cameras or unreliable lip detection. We use a 100-dimensional *i-vector* extractor trained on CN-Celeb [18] to obtain speaker embeddings. These *i-vectors* are computed using oracle labels (provided by the MISP organizing committee) for each speaker. The outputs $\hat{I} \in \mathbb{R}^{D_I\times N}$ is repeated T times to get $I \in \mathbb{R}^{T\times D_I\times N}$ to align with $E_{\boldsymbol{V}}$.

D. Decoder

The structure of the decoder is similar to [4], where it concatenates all embeddings as inputs $E_{\boldsymbol{all}} \in \mathbb{R}^{T\times(D_A+D_V+D_I)\times N}$ and the outputs of prediction is $S \in \mathbb{R}^{T\times N}$. In addition, a softmax operation is adopted to calculate the loss between the outputs and the labels.

III. Proposed Methods

Fig. 2 shows our CASA network structure. The lip images, *i-vectors*, and audio data are input into visual temporal encoder, *i-vector* extractor and audio temporal encoder to obtain the corresponding embeddings. As both *i-vector* and audio are closely associated with the audio modality, we concatenated them as complementary sources. Instead of using the Concat method in [8], we propose the CASA network, which is composed of cross-attention (CA) module and self-attention (SA) module to learn the intrinsic cross-modal information from the embeddings.

A. Cross-Attention (CA)

Although it is impossible to fully align the time of audio-visual data, we hope to learn audio-visual instantaneous interactive information at the frame level through cross-attention. Audio and visual representations are inputs for query (Q), key (K) and value (V) for audio-visual interaction.

Since both $E_{\boldsymbol{A}}$ and I are generated from audio data, we concatenate them as $F_a \in \mathbb{R}^{T\times(D_A+D_I)\times N}$, while $E_{\boldsymbol{V}}$ is equivalent to $F_v \in \mathbb{R}^{T\times D_V\times N}$. The multi-head attention mechanism allows the model to focus on different aspects of useful information and extract richer features. As formulated in Eq. (1) and Eq. (2), the query is generated by F_v, and the key and value are generated by F_a for $F_{a\to v} \in \mathbb{R}^{T\times D_V\times N}$, and vice versa. The left side of Fig. 3 shows the detailed structure of the $F_{a\to v}/F_{v\to a}$.

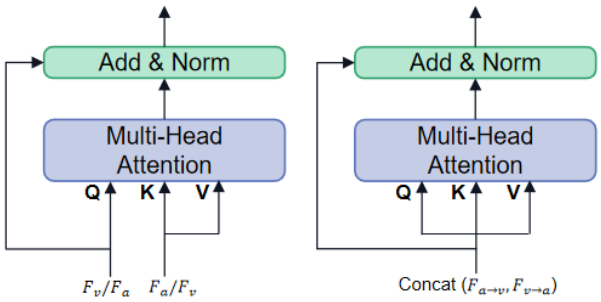
$$F_{a\to v} = \text{softmax}\left(\frac{Q_v K_a^T}{\sqrt{d}}\right) V_a \tag{1}$$

$$F_{v\to a} = \text{softmax}\left(\frac{Q_a K_v^T}{\sqrt{d}}\right) V_v \tag{2}$$

B. Self-Attention (SA)

Self-attention module can learn the context relationship between temporal audio-visual frames and increase their correlation..

The Fig. 3(b) shows the detailed structure of the SA module. The query, key and value are all generated by the concatenation of $F_{a\to v}$ and $F_{v\to a}$.



(a) Cross-Attention (CA) (b) Self-Attention (SA)

Fig. 3. Structure of CA and SA

C. Optimization

We proceeded a two-step training strategy.

- **Step 1.** Train a V-VAD model (Visual Temporal Encoder) using lip images. The loss function is presented in Eq. (3), where $BCE()$ is the binary cross entropy (BCE) loss function [19], $S^V_{t,n}$ and $Y_{t,n}$ are the predicted and the ground truth labels of person n in the t^{th} video frame, T refers to the number of video frames, and N is the number of people in the scenario.

$$Loss_V = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} BCE\left(Y_{t,n}, S^V_{t,n}\right) \quad (3)$$

- **Step 2.** Freeze the V-VAD network for extracting visual features, align visual features and audio data frames, then train the audio temporal encoder, CASA network and the decoder. The loss function is presented in Eq. (4), where $S^{AV}_{t,n}$ is the predicted labels of person n in the t^{th} audio-visual frame.

$$Loss_{AV} = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} BCE\left(Y_{t,n}, S^{AV}_{t,n}\right) \quad (4)$$

IV. Experiments and Result

A. Dataset and metrics

Dataset. The MISP 2022 Challenge dataset [10] includes more than 100 hours of audio-visual data, from 320+ sessions. Each session consists of about 20 minutes of discussion. The dataset has been split into training (Train), development (Dev) and evaluation sets. The evaluation dataset is not publicly released, so we only discussed the results on the development set. The information of Train and Dev sets as show in Table I.

TABLE I
The MISP 2022 challenge dataset

Dataset	Train	Dev
duration(h)	106.09	2.60
Session	301	10

The dataset faces several technical challenges. Firstly, the dataset exhibits a high speech overlap rate, particularly when two groups of speakers are present, with an overlap rate of up to 70%. Secondly, reverberation negatively impacts the audio quality, and TV noise interference can be misclassified as the audio of the target speaker. Additionally, a fixed microphone array is used as the recording device in this scenario, to provide multi-channel signals that increase the difficulty. For the training set, the released videos include far and middle scenes, while audio includes far, middle and near scenes. But for Dev set, there are only video and audio from far scenes. There is no intersection between datasets.

We used the training set to train our all experiments and it included audio signal and lip images. We dereverberated the raw audio signal with NARA-WPE [20] and utilized beamforming technology in Kaldi[21] toolkit to effectively use multi-channel audio information, then applied additive noises (i.e., Babble, Noise, and Music from Musan [22]) and RIRs corpora as the audio augmentation. For audio signals, we got a total 11-fold training dataset, including multi-channel data (6 folds), beamforming (1 fold) and additive noises for beamforming (4 folds). To augment the lip images, we applied random cropping, rotation, flipping, and brightness adjustments. In addition, we extracted non-spoken lip videos from the training set's RTTM (Reference Time and Text Markup) to fill empty lip videos within consecutive temporal audio-visual frames and empty speakers, because each scene has a fixed input of speaker number, so may exist empty speaker inputs, it makes the system to have the ability to distinguish between valid and invalid inputs.

Metrics. We measure the performance of our model by Diarization Error Rate (DER) [23], which is represented with Eq. (5).

$$DER = \frac{FA + MISS + SpkErr}{TOTAL} \quad (5)$$

where FA is the speech durations of false alarm, $MISS$ is the speech durations of missed detection, $SpkErr$ is the speech durations of speaker error, and $TOTAL$ is the sum of durations of all reference speakers' utterances.

B. Implementation details

1) Training: We fixed 6 speakers as input from one scene. The training audio and video signals are segmented into 8-second data blocks with a 4-second step. For the audio input, we used 40-dimensional FBanks as the acoustic feature, with a frame length of 25 ms and a frame shift of 10 ms, and each block has 800 audio frames. In the video modality, lip region images with a resolution of 96×96 and an FPS (Frames Per Second) of 25 are utilized, and each block has 200 visual frames.

Initially, the V-VAD model is trained using BCE loss and the Adam optimizer, with a learning rate of 1e-4, then the video features are extracted through the freezed V-VAD model for each frame and serve as input for the video modality. In the second step, the audio modality input is combined and passed through CASA and decoder modules. BCE loss and the Adam optimizer are employed, with a learning rate of 1e-4 until convergence.

TABLE II
THE PERFORMANCE OF DIFFERENT SYSTEMS ON THE MISP 2022 DEVELOPMENT SET.

NO	Modality	System	Fusion Method	Lip_aug	Beamforming	Audio_aug	MISP_2022_DEV			
							FA(%)	MISS(%)	SpkErr(%)	DER(%)
1	Visual	Baseline	-	×	-	-	6.64	8.17	3.89	18.69
2		Baseline_aug	-	✓	-	-	6.67	7.65	3.65	17.97
3	Audio-visual	Baseline	Concat(Joint training)	×	×	×	4.01	5.86	3.22	13.09
4		Ours	Concat	✓	×	×	3.44	6.53	4.77	14.74
5			CASA	✓	×	×	2.61	6.85	3.72	13.18
6				✓	✓	×	2.43	6.72	3.68	12.83
7				✓	✓	✓	2.39	**6.62**	**3.62**	12.63
8	Fusion	DOVER-Lap of NO.5, 6 and 7	-	-	-	-	**2.32**	6.63	3.65	**12.60**

2) Evaluation: All development sets were divided into 8-second blocks with a step of 6 seconds, and the scores of the overlapped parts between the blocks are averaged as the result. And the competition provided the Oracle VAD and allowed participants to use it.

C. Results

Table II illustrates the DERs (%) comparison of our systems and the baseline. System 1 and 3 are the official baselines in Visual-only and Audio-visual modality respectively. We made lip images augmentation based on System 1 to obtain System 2, Baseline_aug. System 4 is our reconstitution on System 3 without joint training. System 5-7 are our proposed structures with CASA using different data augmentation methods.

Compared to the Visual-only systems, all Audio-visual systems can effectively integrate audio-visual information, and achieve improved performances with lower DERs. The results of Systems 1 and 2 show that effective lip images augmentation can simulate more complex real scenarios and enhance the ability of V-VAD to detect speech and silence. We observe that the CASA fusion method in System 5 achieved a DER of 1.56% improvement compared to System 4, which indicates that CASA can capture more cross-modal interaction information and learn the context relationship between temporal audio-visual frames than the Concat method. Furthermore, the performance of System 6 shows that beamforming effectively utilizes multi-channel speech information, and surpasses that of System 3. Additionally, the audio augmentation in System 7 obtained a better DER of 12.63% based on the results of System 6. Finally, by employing DOVER-lap for fusing our multiple systems, we achieved a DER of 12.60%.

V. CONCLUSION

This paper introduces an embedding fusion method incorporating cross-attention and self-attention mechanisms for audio-visual speaker diarization. Our two-step approach enables the training and integration of both audio and visual modalities, leveraging V-VAD and CASA to enhance system performance and optimization. We have made a series of improvement measures, including beamforming, data augmentation and system-fusion technology, to address the current difficulties faced by AVSD. Finally, our system achieved DER of 12.60%, surpassing the baseline. This result highlights the effectiveness of our approach. We did not conduct joint training in this paper, which will be our future research work.

VI. ACKNOWLEDGEMENTS

Thanks to the National Natural Science Foundation of China (Grant No. 62001405, No. 62276220 and No. S61876160) and Fundamental Research Funds for the Central Universities (No. 20720210087) for funding.

REFERENCES

[1] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[2] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: Speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[3] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 296–303.

[4] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," *arXiv preprint arXiv:2005.07272*, 2020.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[6] R. Ahmad, S. Zubair, and H. Alquhayz, "Speech enhancement for multimodal speaker diarization system," *IEEE Access*, vol. 8, pp. 126 671–126 680, 2020.

[7] E. Z. Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "Ava-avd: Audio-visual speaker diarization in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3838–3847.

[8] M.-K. He, J. Du, and C.-H. Lee, "End-to-end audio-visual neural speaker diarization," *Proc. Interspeech 2022*, pp. 1461–1465, 2022.

[9] T. Liu, S. Fan, X. Xiang, *et al.*, "Msdwild: Multimodal speaker diarization dataset in the wild," *Proc. Interspeech 2022*, pp. 1476–1480, 2022.

[10] H. Chen, J. Du, Y. Dai, *et al.*, "Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022, 2022, pp. 1766–1770.

[11] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2857–2866. DOI: 10.1109/WACV48630.2021.00290.

[12] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, "Training strategies for improved lip-reading," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8472–8476. DOI: 10.1109/ICASSP43922.2022.9746706.

[13] J. Wang, X. Yang, H. Li, L. Liu, Z. Wu, and Y.-G. Jiang, "Efficient video transformers with spatial-temporal token selection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, Springer, 2022, pp. 69–86.

[14] Y. Wei, H. Liu, T. Xie, Q. Ke, and Y. Guo, "Spatial-temporal transformer for 3d point cloud sequences," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1171–1180.

[15] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935.

[16] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[17] D. Raj, L. P. Garcia-Perera, Z. Huang, *et al.*, "Dover-lap: A method for combining overlap-aware diarization outputs," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 881–888.

[18] Y. Fan, J. Kang, L. Li, *et al.*, "Cn-celeb: A challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7604–7608.

[19] S. Mannor, D. Peleg, and R. Rubinstein, "The cross entropy method for classification," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 561–568.

[20] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, VDE, 2018, pp. 1–5.

[21] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[22] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[23] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers 3*, Springer, 2006, pp. 309–322.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhanced Neural Beamformer with Spatial Information for Target Speech Extraction

Aoqi Guo[1], Junnan Wu[2], Peng Gao[2], Wenbo Zhu[2], Qinwen Guo[1], Dazhi Gao[1,*] and Yujun Wang[2]
[1] Ocean University of China, China
[2] Xiaomi Inc., Beijing, China
E-mail: guoaoqi@stu.ouc.edu.cn dzgao@ouc.edu.cn

***Abstract*—Recently, deep learning-based beamforming algorithms have shown promising performance in target speech extraction tasks. However, most systems do not fully utilize spatial information. In this paper, we propose a target speech extraction network that utilizes spatial information to enhance the performance of neural beamformer. To achieve this, we first use the UNet-TCN structure to model input features and improve the estimation accuracy of the speech pre-separation module by avoiding information loss caused by direct dimensionality reduction in other models. Furthermore, we introduce a multi-head cross-attention mechanism that enhances the neural beamformer's perception of spatial information by making full use of the spatial information received by the array. Experimental results demonstrate that our approach, which incorporates a more reasonable target mask estimation network and a spatial information-based cross-attention mechanism into the neural beamformer, effectively improves speech separation performance.**

I. Introduction

The target speech extraction task is derived from the speech separation task, which aims to extract the speech of a specific target speaker from a mixed signal. Unlike conventional speech separation, which deals with the permutation order of signals, target speech extraction only focuses on extracting the target signal, thus avoiding the permutation problem [1]. In recent years, multi-channel beamforming algorithms have shown great potential in the task of target speech extraction. Beamforming algorithms achieve spatial domain filtering of the observed signal by enhancing the signal in the desired direction and suppressing the signal in other directions, which naturally meets the task goal of target speech extraction. Classical beamforming algorithms require accurate estimation of the wave direction and determination of the speech or noise part in the mixed signal [2], which is difficult to obtain in noisy and reverberant environments using traditional algorithms. Therefore, obtaining accurate target orientation information in real scenes, improving the estimation accuracy of the covariance matrix, and solving the numerical instability problem in matrix operations are the key to improving beamforming performance.

In recent years, the development of deep learning technology has provided new research ideas for solving the above problems. First, using neural networks for wave direction estimation has achieved good experimental results [3], providing more accurate target orientation information for beamforming algorithms. At the same time, combining neural networks with beamforming algorithms improves the estimation accuracy of the covariance matrix and effectively improves the performance upper limit of beamforming algorithms. Among them, Mask Based Beamforming, as a classic representative of the combination of deep learning and beamforming algorithms, uses a neural network to predict the mask of speech and noise to calculate the spatial covariance matrix of speech and noise, respectively, and then inputs it into traditional beamforming algorithms such as GEV or MVDR for numerical calculation [4], [5]. For the numerical instability problem that traditional beamforming algorithms still have difficulty in solving even after diagonal loading [6] in matrix numerical operations, ADL-MVDR [7] used an RNN to replace the matrix inversion and eigendecomposition processes in the MVDR beamforming algorithm, avoided the numerical instability caused by singular matrices in matrix operations, and achieved significant performance improvement for the neural beamformer. In recent experiments, the combination of deep learning and beamforming algorithms has been further deepened. GRNNBF [8] used an RNN to model the covariance matrices of speech and noise, replaced all matrix numerical operations in traditional beamforming algorithms, and directly predicted beamforming weights using neural networks, achieved good separation results. SARNN [9] added self-attention mechanism to the beamforming module, enhanced the neural network's modeling ability for the spatial and temporal information contained in the covariance matrix. EABNet [10] adopted a more aggressive approach, directly modelled the array observation signals to obtain higher-dimensional information representations than the covariance matrix. The neural network implicitly includes all the calculation steps of beamforming and directly outputs beamforming weights. The above methods combined with deep learning technology have all achieved good performance improvements for beamforming algorithms.

In scenarios such as in-car and remote conferences, the position of the target speaker relative to the microphone array is often fixed, and its angle information is easily obtained through the device [1]. To leverage the angle information of the target sound source, Chen et al. proposed the wave direction feature calculated using the angle information of the target sound source, which enhanced the directionality of the speech separation system [11]. ADL-MVDR [7], GRNNBF [8], SARNN [9], UFE [12], etc. inputted the angle feature

Work conducted when the first author was intern at Xiaomi Inc.
* Corresponding author.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

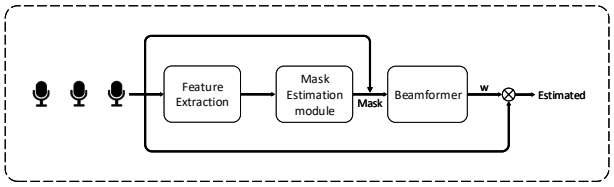


Fig. 1. General structure of a neural beamformer.

into the neural network, improved the separation module's perception of spatial information, thereby improved the separation performance of the system. Gu et al. extended the angle feature from a two-dimensional plane to a three-dimensional space, improved the separation performance of the model when the azimuth angle between the target sound source and the interference source was small [13].

In this paper, we designed a model that uses target spatial orientation information to enhance the performance of neural beamforming. The system consists of a front-end pre-separation module and a back-end beamforming module. Compared with other models, our improvements mainly focus on the following two points:

• To improve the estimation accuracy of the spatial covariance matrices of speech and noise, we stack the input features in the channel dimension and use the UNet-TCN structure to model the input features. This avoids the loss of feature information caused by directly reducing the input features after connecting them in the frequency dimension.

• To better utilize the spatial information obtained by the array, we transform the spatial features of the input signal and the covariance matrix calculated by the pre-separation module into a feature space with the same dimensionality. We then use a cross-attention mechanism to enhance the beamforming network's perception of spatial orientation, thereby improving the prediction accuracy of the beamforming weights.

Experimental results show that after the improvements to the pre-separation module and beamforming module mentioned above, the target speech separation performance of the neural beamformer could be effectively improved.

The remainder of this paper is organized as follows. Section II presents the signal model for the target speech extraction task and the beamforming algorithm. Section III details our proposed neural beamforming algorithm. Section IV provides an overview of the experimental setup and presents an analysis of the experimental results. Finally, Section V concludes the paper.

II. Signal Modeling and Beamforming

Consider the far-field time-domain signal model of the real scene, described as:

$$y(t) = x(t) * a_1(t) + n(t) * a_2(t) + s(t) \quad (1)$$

$$= x^{'}(t) + n^{'}(t) + s(t) \quad (2)$$

where $y(t) = [y^{(0)}(t), y^{(1)}(t), ..., y^{(M-1)}(t)]^T$ indicating the time-domain signal received by the M-channel microphone array. $x(t), n(t)$ represent the reverberation-free speech signals from the target speaker and the interfering speaker, respectively. $a_1(t), a_2(t)$ are the room impulse response(RIR) between the speaker and the array elements of the microphone. $*$ denotes the convolution operation. $x^{'}(t), n^{'}(t)$ represent the speech signals of the target speaker and the interfering speaker with reverberation, respectively. $s(t)$ is the background noise. When not focusing on the dereverberation task, the task goal of target speech extraction is to extract the target speaker's speech $x^{'}(t)$ from the noisy signal $y(t)$.

After the signal model is transferred to the time-frequency domain by short-time fourier transform(STFT), it is expressed as:

$$Y(t,f) = X(t,f) + N(t,f) + S(t,f) \quad (3)$$

The purpose of the beamforming algorithm is to obtain the filter weight w for the array observation signal, and extract the desired signal from the array observation signal by performing spatial filtering on the observation signal, that is:

$$X(t,f) = w^H Y(t,f) \quad (4)$$

where $(\cdot)^H$ denotes the conjugate transpose operation.

Taking the traditional MVDR beamforming algorithm as an example, the purpose is to minimize the output noise power of the beamforming algorithm without distorting the signal, that is:

$$min(w^H \Phi_{NN} w^H), \quad s.t \quad w^H \alpha(\theta) = 1 \quad (5)$$

For the above formula, the desired filter weight w is solved by the lagrange multiplier method:

$$w = \frac{\Phi_{NN}^{-1}\alpha(\theta)}{\alpha^H(\theta)\Phi_{NN}^{-1}\alpha(\theta)} \quad (6)$$

where $\Phi_{NN} = N(t,f)N^H(t,f), \Phi_{NN} \in M \times M$ is the covariance matrix of the noise signal, $\alpha(\theta) = [e^{-j\omega\tau_m}], m = [0, 1, ..., M-1]$ is the steering vector that calculated using target sound source's DOA. When the steering vector is difficult to obtain due to the interference of environmental factors, PCA can be performed on the covariance matrix of the speech signal to obtain the estimated value of the steering vector [14].

Therefore, it can be seen that the key to the beamforming algorithm lies in the estimation of the covariance matrix. Traditional beamforming methods often rely on certain algorithms to determine the time periods of speech and noise in the signal, but these approach is not optimal. With the development of deep learning technology, using deep neural networks to determine the speech and noise components in the signal has achieved good experimental results.

Fig. 1 shows the general framework of the combination of neural network and beamforming. It first predicts a set of time-frequency masks representing the corresponding relationship between the desired signal and the original mixed signal through the neural network, such as IBM [4], IRM [5], CRM [15], CRF [16], etc. Taking the IRM as an example, it defines

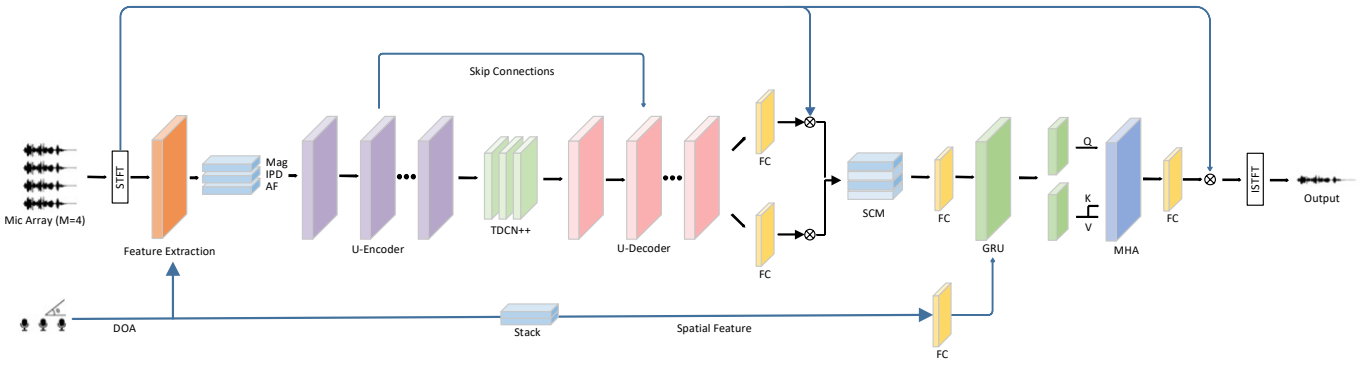


Fig. 2. The overall structure of our proposed model.
Both the input features and the covariance matrix are stacked along the channel dimension when inputting to the network.

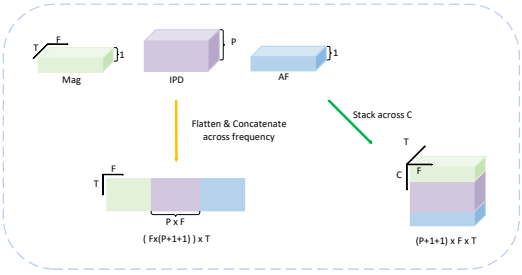


Fig. 3. The combination of input features.

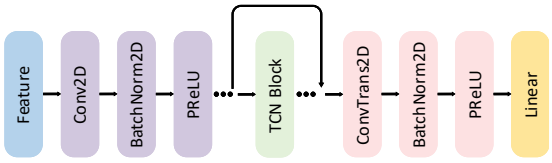


Fig. 4. Pre-separator based UNet-TCN structure

the energy ratio relationship between the desired signal and the mixed signal:

$$IRM_x = \frac{|X(t,f)|}{|Y(t,f)|} \tag{7}$$

Then the speech and noise components in the mixed signal can be recovered through mask, so as to calculate the speech and noise space covariance matrix respectively [4]:

$$\Phi = \frac{\sum_{t=1}^{T} M^2(t,f) Y(t,f) Y^H(t,f)}{\sum_{t=1}^{T} M^2(t,f)} \tag{8}$$

Once the covariance matrix of the target signal and noise is calculated, it can be substituted into traditional beamforming algorithms such as GEV and MVDR to obtain the beamforming weight w. Recent experiments have demonstrated that utilizing neural networks to model covariance matrix information and predict beam weights can lead to improved target speech separation compared to traditional algorithms.

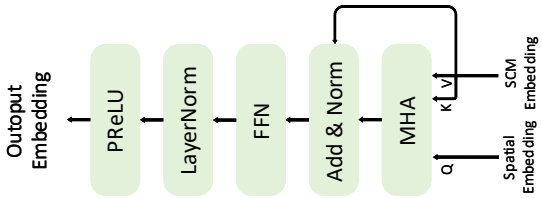


Fig. 5. A cross attention module with spatial information as Query and covariance matrix information as Key and Value. All features are on the time frame level.

III. Enhanced Neural Beamformer with Spatial Information

Fig. 2 shows the overall architecture of our proposed model. First, the observation signal of the array is transformed into the time-frequency domain by STFT to extract the spatial features of the signal. The pre-separation network then estimates the mask of the target speech and noise based on the input features and calculates the covariance matrix. Following this, the beamforming network leverages the spatial features to predict the beamforming weight, which is subsequently used to beamform the observed signal. Finally, the original time domain signal is restored through ISTFT. Specifically as follows:

A. Feature Extraction

The first microphone from the multi-channel mixed signal is selected as the reference microphone, and its received signal's amplitude spectrum is calculated as the input feature for the pre-separation network.

$$Y_1^{mag}(t,f) = |Y_1(t,f)| \tag{9}$$

To obtain the spatial information of the microphone array, we select P pairs of microphones and calculate the phase difference between each pair as the spatial feature for the input of our model. This approach has been widely used in speech enhancement and source separation tasks, as it provides a concise and effective way to capture the spatial information of the microphone array.

$$IPD_{i,j}(t,f) = \angle Y_i(t,f) - \angle Y_j(t,f) \tag{10}$$

$$i, j \in (0, 1, ..., M-1), i \neq j$$

In order to enhance the spatial information acquired by the model, the cosine distance between the phase delay φ_1 of the signal in the desired direction on the array and the phase difference φ_2 observed by a certain pair of microphones is calculated. Then sum over all P pairs of microphones. That is, the angular characteristics of the target signal serve as supplementary input.

$$AF(\theta, f) = \sum_{p=1}^{P} cos(IPD_{i,j}(t, f) - \frac{2\pi f d cos\theta}{c}) \quad (11)$$

When the phase angle of the mixed signal and the desired signal is closer, the value of the angle characteristic tends to 1 approximately [12]. All these features are stacked along the channel dimension to serve as the network's input features.

$$Feature = [Y_1^{mag}(t, f), cosIPD_{i,j}(t, f), AF(\theta, f)]$$

B. Pre-separation Module

The pre-separation network adopts an Encoder-TCN-Decoder structure, which includes 2D convolutional layers, linear layers, and a variant of TCN [17]. To process the input features, it's a common practice to concatenate them on the frequency dimension, followed by 1D convolution to reduce dimensionality before feeding them into the TCN or RNN module [7], [8], [9]. However, to avoid the loss of feature information due to direct dimensionality reduction of the input features, we employ the UNet-TCN structure to model them. By stacking the input features along the channel dimension, we obtain $Y_{input} \in R^{N \times F \times T}(N = P + 1 + 1)$, which is then fed into the UNet-TCN layer to model the nonlinear mapping of the time dimension from the source signal domain to the separated signal domain and noise domain. Finally, the predicted cRM is restored through the linear layer to recover the target speech.

$$X_{cRM}(t, f) = (cRM_r + jcRM_i)(Y_r + jY_i) \quad (12)$$

The noise signal is also calculated in the above way. All activation functions in the pre-separation network use PReLU to accelerate the convergence speed of the network. Compared with ReLU, it allows negative values to appear, which is synergistic with the negative values between the input phase difference and the output mask.

C. Neural Beamforming Module

The beamforming network is composed of a recurrent neural network and a cross-attention module. First, the speech and noise signals generated by the pre-separation of the front-end network are used to calculate the spatial covariance matrix $\Phi_{SS}(t, f) \in C^{M \times M}, \Phi_{NN}(t, f) \in C^{M \times M}$ at the time frame level respectively.

$$\Phi_{SS}(t, f) = X_{cRM}(t, f) X_{cRM}^H(t, f) \quad (13)$$

$$\Phi_{NN}(t, f) = N_{cRM}(t, f) N_{cRM}^H(t, f) \quad (14)$$

Since the spatial covariance matrix is a complex-valued Hermitian matrix, and the defined neural network is a real network, the real and imaginary parts of the speech and noise signal covariance matrices are concatenated on the channel dimension, and LayerNorm is used to the covariance matrix of is standardized. In order to enhance the spatial constraints on the beamformer, we use the spatial information received by the array, that is, the phase difference between channels and the angle characteristics of the target sound source, to perform a cross attention mechanism on the covariance matrix of speech and noise. First, the linear layer and GRU are used to transform the spliced covariance matrix and cosIPD, Angle-Feature(AF) to model the inter-channel information at the time frame level and transform it into embedding of the same dimension. Subsequently, the modeled spatial information is used as the query of multi-head attention, and the covariance matrix information is used as the key and value, and the spatial orientation information is used to enhance the ability of attention to model inter-channel information at the time frame level.

$$Query = RNN - DNN(cosIPD, AF) \quad (15)$$

$$Key, Value = RNN - DNN(\Phi_{SS}, \Phi_{NN}) \quad (16)$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (17)$$

Then, the observed signals from the microphones are beamformed using the beamforming weights output by the linear layer to obtain the spectrogram of the model prediction output. The time-domain signal is then recovered through ISTFT, resulting in the desired time-domain speech signal.

IV. Experiment setup and result analysis

A. Dataset settings

The source speech signals for the target and interfering speakers were obtained from the Aishell2 [18] dataset, while the background noise data was sourced from DNS2023 [19]. For an indoor reverberation scene, the room size was set to a random size between [3, 3, 1.5] and [8, 8, 2.5] meters for length, width, and height, and the reverberation time(rt60) was set to 0.1-0.6 seconds. The simulation was carried out using the image source method, generating room impulse responses for the target and interference signals [20]. The microphone array consisted of a four-element linear array with an element spacing of 3 cm. To ensure a certain degree of spatial independence in the signals, we set the minimum angle between the two sources relative to the microphone array to 5°. The signal-to-noise ratio between the target and interference signals was set to [-6, 6] db. To enhance the model's robustness, a background noise of [-5, 20] db was added to the simulated data. In total, the simulation generated 120,000 pieces of training set data, 14,000 pieces of validation set data, and 7,000 pieces of test set data. All audio was downsampled to 16kHz and each piece of data is 4 seconds long.

TABLE I
SI-SDR(DB), PESQ AND STOI RESULTS OF IRM-MVDR, GRNNBF, SARNN AND OUR PROPOSED MODELS.

Model	Para.(M)	PESQ↑	STOI↑	SISDR↑
No processing	-	1.148	0.563	-1.76
IRM MVDR**[5]**	5.33	1.586	0.757	5.25
GRNNBF**[8]**	15.73	2.176	0.845	8.42
SARNN**[9]**	16.99	2.324	0.859	9.48
Proposed	8.64	**2.499**	**0.88**	**10.51**

TABLE II
SI-SDR(DB), PESQ AND STOI RESULTS OF ABLATION EXPERIMENTS WITH OUR PROPOSED MODEL

Model	Para.(M)	PESQ↑	STOI↑	SISDR↑
TCN IRM (cat)	5.33	1.586	0.755	5.95
+ UNet IRM (stack)	6.76	1.846	0.799	6.82
TCN cRM (cat)	5.4	1.591	0.759	5.97
+ UNet cRM (stack)	7.09	1.851	0.801	6.86
+ BF(MHSA 128)	8.05	2.466	0.876	10.11
+ BF(MHSA 256)	8.89	2.481	0.878	10.41
+ BF(MHCA)	8.64	**2.499**	**0.88**	**10.51**

B. Experiment settings

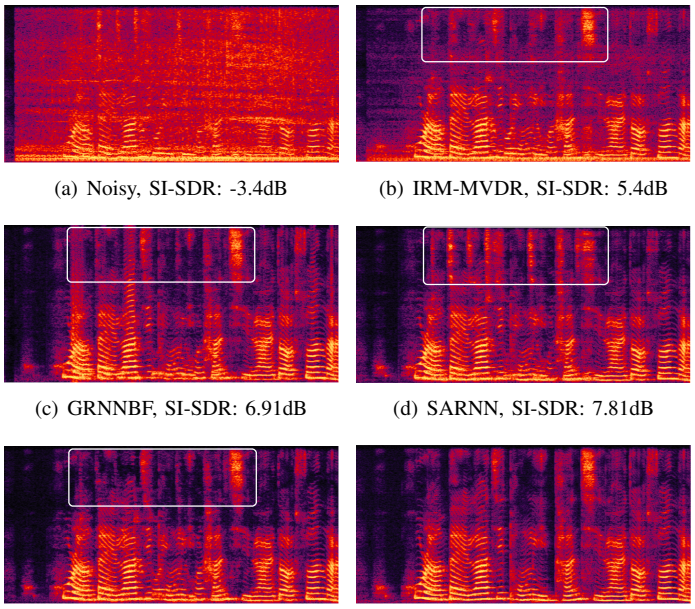
During the training process, the window length of STFT and the number of FFT points are set to 32ms, and the frame shift is set to 16ms. For the array, we select (0,1), (0,2), (0,3) three microphone pairs to calculate cosIPD. Therefore, the channel dimension of the input feature is 5, that is $Feature \in R^{5\times F\times T}$. For the pre-separation module, the number of channels of Conv2D in the UNet structure is (5, 32, 64, 128). The TCN structure is a 3x8 stacked TCN Block with 128 channels. For the beamforming module, the number of hidden layer units of the double-layer GRU network is 256, the dimension of the cross attention module is 128, and the last linear layer outputs complex-valued beamforming weights, so the output dimension is set to 8 (channels*2), the proposed The model is trained end-to-end.

The experiment uses two A100 graphics cards to train 60 epochs on the simulation data set, and the Batch Size is set to 24. The Adam optimizer is used for optimization, the initial learning rate is set to 2e-3 and decays continuously with the number of epochs, and the decay coefficient is 0.98. The maximum gradient clipping is set to 10 to speed up the network convergence process.

We use the MVDR method based on the IRM and the GRNNBF [8] and SARNN[9] models trained according to the MISO model as the baseline for experimental comparison. For IRM-MVDR, we use the TCN network to predict the IRM, which is then substituted into the MVDR algorithm for beamforming. For GRNNBF and SARNN, the 4x8 TCN Block is used to predict the CRF mask, which is input to the GRU or self-attention module with 500 hidden units after LayerNorm to output beamforming weights. For all models, we use the joint loss function of SI-SDR and MSE for training, and the weights of the two are equal.

C. Results

Table 1 shows the experimental results of our proposed model and baseline model on the test set. In terms of model parameters, our model has increased compared with IRM-MVDR, which is mainly due to the difference in the prediction mask and the replacement of the MVDR algorithm with a neural beamforming module. Compared with GRNNBF and SARNN, our parameter reduction is more obvious. This is because after adding the cross-attention mechanism based on spatial information, the model does not require a large number of hidden layer units to achieve a satisfactory separation effect.



(a) Noisy, SI-SDR: -3.4dB (b) IRM-MVDR, SI-SDR: 5.4dB

(c) GRNNBF, SI-SDR: 6.91dB (d) SARNN, SI-SDR: 7.81dB

(e) Proposed, SI-SDR: 9.97dB (f) Target

Fig. 6. Spectrum of a piece of data in the test set after model processing. SIR=-2.2dB, SNR=-4.8dB. The angle between the speaker is 9°.

Therefore, the experimental results show that our proposed model can significantly improve the separation performance of the target speech compared with the baseline while reducing the number of parameters.

In Table 2, we present the results of our ablation experiments on the proposed model. The experimental results show that for the pre-separation module, the pre-separation effect obtained by first encoding the input features through the Unet structure and then inputting them to the TCN block to output the mask is better than directly inputting the input features to the TCN through Conv1D. For the neural beamformer, the UNet-TCN structure makes the model more fully utilize the input feature information, reduces the information loss in the dimension reduction process, and makes the estimation of the covariance matrix more accurate. In terms of mask prediction, whether it is TCN or UNet-TCN structure, the effect of network prediction cRM on recovering speech signals is slightly better than that of IRM, but the gap is not large. This is because for frequency-domain speech separation, when the frame length is 32ms, the effect of phase estimation on the separation effect is limited

[21].

For the neural beamforming module, we take multi-head self-attention(MHSA) module with 128 and 256 hidden layer units as an example. It is evident that incorporating the embedding modeled by the covariance matrix and spatial information into the multi-head cross-attention(MHCA) module leads to better speech separation performance than the MHSA module with an equivalent number of hidden layers. Experimental results demonstrate the effectiveness of introducing spatial information into the beamforming module via the cross-attention mechanism.

D. Spectrogram analysis

We analyze and compare the spectrograms of the target speech generated using different separation methods, and discuss the results further. In order to verify the robustness of the algorithm, we selected a piece of data with a low SNR and a small angle between the target speaker and the interference speaker from the test set for visualization, as shown in Fig. 6. Due to numerical instability in the training process and the influence of the MVDR algorithm itself, the IRM-MVDR model has a poor suppression effect on interference noise, and the separated spectrum is severely damaged by noise. In GRNNBF and SARNN using neural beamers, the separation performance of the system has been improved to a certain extent, and both have a good suppression effect on background noise, but there may still be some residual interference noise present in the separated signal. After the proposed model uses the UNet-TCN structure to improve the pre-separation network and adds spatial information to the beamforming module, the background and interference noise are further suppressed, and the degree of spectral distortion is lower than that of the baseline model, achieved better experimental results.

E. Beam pattern

To further analyze the spatial filtering performance of the model, Fig. 7 visualizes the beam patterns of different neural beamformers, including GRNNBF, SARNN and our proposed model. The speech is divided into four segments, segment (a) has only background noise, segment (b) has two speakers speaking at the same time, segment (c) has only the interfering speaker active, segment (d) has only the target speaker active. All beam patterns are averaged over the frequency dimension.

All three models have a strong suppression near 90° when there is no sound source activity, which is a little strange, we don't know if this is due to the influence of background noise. When two sound sources speak at the same time, all three models produce a strong suppression effect on the interference direction. Our proposed model suppresses the disturber to a greater extent when only the disturber is active. When only the target sound source is active, the directivity of our proposed model is better than that of GRNNBF and SARNN, which reduces the degree of distortion of the model-predicted spectrum.

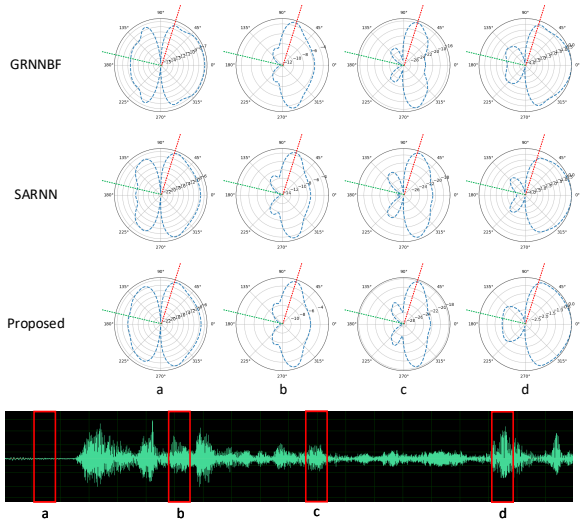


Fig. 7. An example of beam patterns derived by GRNNBF, SARNN and our proposed module. The DOA of the target speaker and the interference speaker are 79 and 164, and marked with red and green respectively.(a) is a silent segment, (b) is two speakers are active at the same time, (c) is only the interference speaker, (d) is only the target speaker.

V. Conclusion

In this study, we propose a novel UNet-TCN structure to model input features, resulting in improved estimation accuracy of the covariance matrix. Furthermore, we introduce spatial information through the cross-attention module in the neural beamforming module to enhance beamformer performance, enabling better noise rejection and reduced spectral distortion. Objective analysis and subjective evaluation show significant improvements in separation effect. In the future, we plan to achieve better target speech extraction while reducing model complexity and output spectrum distortion.

REFERENCES

[1] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[3] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *Interspeech*, 2016, pp. 1981–1985.

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5210–5214.

[6] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE transactions on signal processing*, vol. 51, no. 7, pp. 1702–1715, 2003.

[7] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6089–6093.

[8] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized spatio-temporal rnn beamformer for target speech separation," *arXiv preprint arXiv:2101.01280*, 2021.

[9] X. Li, Y. Xu, M. Yu, *et al.*, "Mimo self-attentive rnn beamformer for multi-speaker speech separation," *arXiv preprint arXiv:2104.08450*, 2021.

[10] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6487–6491.

[11] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 558–565.

[12] J. Wu, Z. Chen, J. Li, *et al.*, "An end-to-end architecture of online multi-channel speech separation," *arXiv preprint arXiv:2009.03141*, 2020.

[13] R. Gu, S.-X. Zhang, M. Yu, and D. Yu, "3d spatial features for multi-channel target speech separation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 996–1002.

[14] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[15] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[16] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7407–7411.

[17] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[18] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[19] H. Dubey, A. Aazami, V. Gopal, *et al.*, "Icassp 2023 deep speech enhancement challenge," *arXiv preprint arXiv:2303.11510*, 2023.

[20] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 351–355.

[21] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *JASA Express Letters*, vol. 2, no. 10, p. 104 802, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Low-complexity Multi-Channel Speaker Extraction with Pure Speech Cues

Bang Zeng*†, Hongbin Suo‡, Yulong Wan‡, Ming Li*†
* Data Science Research Center, Duke Kunshan University, Kunshan, China
† School of Computer Science, Wuhan University, Wuhan, China
E-mail: bangzeng@whu.edu.cn, ming.li369@dukekunshan.edu.cn
‡ Data&AI Engineering System, OPPO, Beijing, China
E-mail: suohongbin@oppo.com, wanyulong@oppo.com

***Abstract*—Most multi-channel speaker extraction schemes use the target speaker's location information as a reference, which must be known in advance or derived from visual cues. In addition, memory and computation costs are enormous when the model deals with the fusion input. In this paper, we propose Speaker-extraction-and-filter Network (SeafNet), which is a low-complexity multi-channel speaker extraction network with only speech cues. Specifically, the SeafNet separates the mixture by utilizing the correlation between an estimation of target speaker on reference channel and the mixed input on rest channels. Experimental results show that compared with the baseline, the SeafNet model achieves 6.4% relative SISNRi improvement on the fixed geometry array and 8.9% average relative SISNRi improvement on the ad-hoc array. Meanwhile, the SeafNet achieves 60% relative reduction in the number of parameters and 42% relative reduction in the computational cost.**

I. Introduction

The cocktail party problem [1] shows the fact that the performance of the speech applications, such as speech recognition, is significantly affected by noise and reverberation in the real world. Speech separation is widely used to deal with this problem by improving the noise robustness of the back-end speech systems. Although most methods work well in the single-channel scenario [2]–[5], the performance declines significantly when the separation system faces a reverberation background. The multi-channel speech separation model with a beamformer is designed to address this issue. Conventional beamforming algorithms can be broadly classified as fixed and adaptive. The fixed beamformer [6] first calculates the time difference of arrival (TDOA) between the reference microphone and the remaining microphones to derive a time-shifted signal for each microphone and then sums up the time-shifted signals to obtain the final beamformer output. The adaptive beamforming algorithms determine the optimal weights of the beamformer through different criteria [7]–[9], the main one of which is the minimum variance distortionless response [8]. In recent years, multi-channel speech separation adopting a learning-based beamformer has become dominant. There are two main routes for these multi-channel speech separation schemes. One class [10]–[12] uses deep neural networks (DNNs) to pre-separate the signals of each channel first, and then estimates the beamformer coefficients based on the pre-separated outputs, or filters the pre-separated outputs directly with a conventional beamformer. The other class [13]–[16], which implicitly employs beamforming in DNN, directly models the mapping relationship between multi-channel inputs and target outputs. Among the aforementioned methods, Filter-and-Sum Network with Transform-Average-Concatenate (FaSNet-TAC) [16] belongs to the latter, and achieves good performance. FaSNet-TAC is a microphone permutation and number invariant model due to the use of pair-wise features. Moreover, in order to overcome the shortcomings of using pair-wise features, FaSNet-TAC adopts TAC operation to fully utilizes the information from all microphones. Experimental results in [16] shows that FaSNet-TAC works well in both fixed geometry array and ad-hoc array. However, the unknown number of speakers and the global permutation problem are still the main obstacles for multi-channel blind speech separation algorithms.

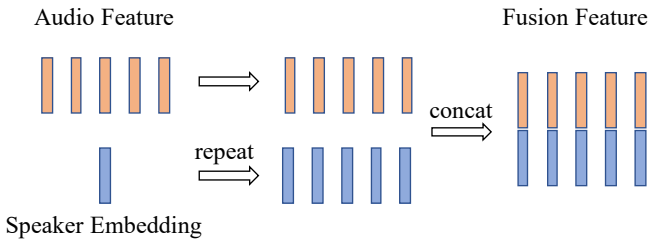


Fig. 1. The tranditional way to use the speaker embedding in target speaker extraction.

Speaker extraction [17], [18], also known as target speech separation, can extract a specific speaker's speech from the mixture with reference information. The speaker's location is commonly used as auxiliary reference information in multi-channel scenarios [19], [20]. However, the speaker's location must be known in advance or extracted from the visual features [21]. Using the speaker embedding from the target speaker's enrollment speech can deal with this issue, yet this will lead to another problem of bigger model size and computational costs due to the relatively high dimensionality of fusion features as shown in Figure 1. This problem is even more serious on multi-channel tasks. In this paper, we extend the FaSNet-TAC [16] model to the multi-channel speaker extraction task. Moreover, to resolve the aforementioned limitations, we propose a relatively low-complexity speaker extraction

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

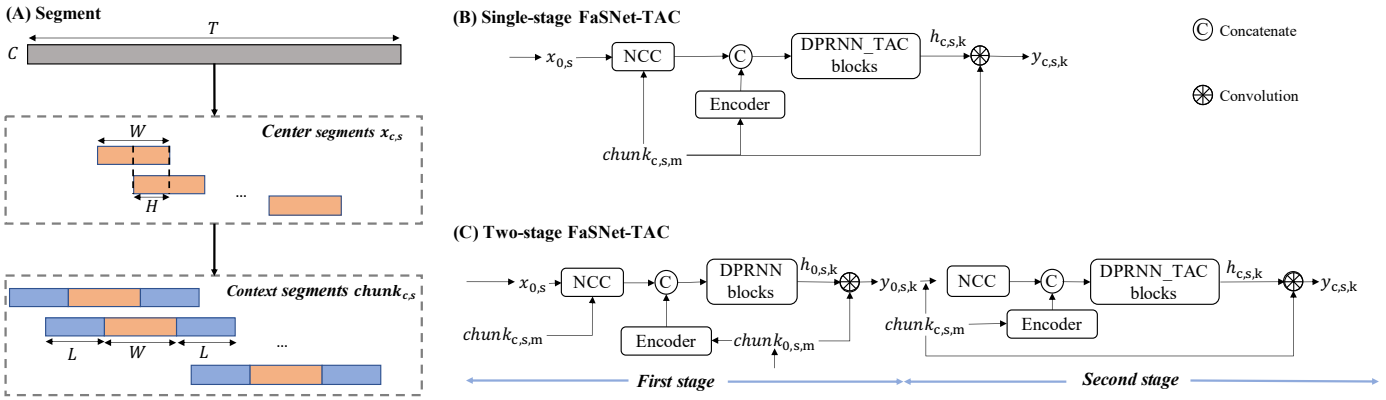


Fig. 2. System flowchart of the origin FaSNet-TAC models. (A) The FaSNet-TAC first splits the input into center segments and context segments. (B) The single-stage FaSNet-TAC model. (C) The two-stage FaSNet-TAC model

network only using speech cues, called Speaker-extraction-and-filter Network (SeafNet). Specifically, the SeafNet first extracts the target speaker's wave of the reference microphone signal by utilizing a reference speech. Then, the SeafNet model computes the Normalized Cross-Correlation (NCC) of this estimation and mixed waveform. The NCC feature contains the content difference between the reference microphone's target speaker estimation and the multi-channel mixed input. Finally, the SeafNet derives the separation result in the same way as the FaSNet-TAC model. In this way, the speaker features only need to be concatenated with the time-domain features of the reference microphone, rather than with the ones of all microphones' signal, significantly reducing the number of parameters and computational costs of the model.

II. Methods

A. Original FaSNet with TAC

FaSNet-TAC is an effective multi-channel blind speech separation solution. Figure 2 (B) and (C) shows the two-stage [15] and single-stage [16] FaSNet-TAC models. As shown in Figure 2 (A), FaSNet-TAC first split the input into center segments with the length of W:

$$\mathbf{x}_{c,s} = \mathbf{x}_c[sH : sH + W - 1] \tag{1}$$

where $\mathbf{x}_{c,s}$ is the center segments, $c \in 0, 1, ..., C-1$ is the c-th microphone of the input, $s \in \mathbf{Z}$ is the index of the segments, and $H \in [0, W-1]$ is the hop size. Then, $\mathbf{x}_{c,s}$ is concatenated with a context window with the length of L on both sides and results $\mathbf{chunk}_{c,s}$:

$$\mathbf{chunk}_{c,s} = \mathbf{x}_c[sH - L : sH + W + L - 1] \tag{2}$$

The two-stage FaSNet-TAC first calculates the Normalized Cross-Correlation (NCC) feature, defined as the cosine similarity between the reference microphone signal and the signal of each remaining microphones. The NCC feature can be formulated as:

$$\begin{cases} \mathbf{chunk}_{c,s,m} = \mathbf{x}_{c,s}[m : m + W - 1] \\ \mathbf{q}_{c,s,m} = \frac{\mathbf{x}_{0,s}(\mathbf{chunk}_{c,s,m})^T}{||\mathbf{x}_{0,s}||_2||\mathbf{chunk}_{c,s,m}||_2} \end{cases}, m = 1, ..., 2L + 1 \tag{3}$$

where $\mathbf{q}_{c,s,m}$ is the NCC feature between $\mathbf{x}_{0,s}$ and $\mathbf{chunk}_{c,s,m}$. $\mathbf{x}_{0,s}$ and $\mathbf{chunk}_{c,s,m}$ is the center segments of the reference microphone and the context segments of the c-th microphone, respectively. The mean of the obtained NCC features $M(\mathbf{q}_{(0,...,C-1),s,m})$ is then concatenated with the context segments' encoding of reference microphone $E(\mathbf{chunk}_{0,s,m})$. Then, the results are fed into N_1 Dual-Path RNN (DPRNN) [5] blocks $\mathscr{F}(*)$ to calculate the beamformer coefficients of the reference microphone $\mathbf{h}_{0,s,k}$ where $k = 0, 1, ..., K-1$. K is the number of the speakers. The estimation of the reference microphone $\mathbf{y}_{0,s,k}$ is then obtained by filtering $\mathbf{chunk}_{0,s,m}$ with the beamformer coefficients $\mathbf{h}_{0,s,k}$:

$$\mathbf{h}_{c,s,k} = \mathscr{F}_{TAC}([E(\mathbf{chunk}_{c,s,m}), \mathbf{g}_{c,s,m}]) \tag{4}$$

$$\mathbf{y}_{s,k} = \mathbf{y}_{0,s,k} + \sum_{c=1}^{C-1} \mathbf{chunk}_{c,s,m} \circledast \mathbf{h}_{c,s,k} \tag{5}$$

Unlike the above two-stage architecture, the single-stage FaSNet-TAC model jointly estimates the beamformer coefficients for all input channels and applies the TAC module between all DPRNN blocks.

B. FaSNet-TAC for Speaker Extraction

We propose two main options for using the FaSNet-TAC to extract the target source. The first solution, called FaSNet-TAC+Sim, is to use the original FaSNet-TAC to obtain the separation results of multiple speakers and then filter the target source through the cosine similarity in the inference stage. The second solution, called Tar-FaSNet-TAC, is to utilize the target speaker embedding during the training stage to obtain a speaker extraction model based on FaSNet-TAC, which can directly extract the target source.

Figure 3 (A) shows the structure schematic Tar-FaSNet-TAC model, which consists of a speaker embedding extraction module and a speaker extraction module. The ResNet34 used as the speaker embedding extraction module in this work is based on [22]. The speaker extraction model consists of N DPRNN blocks with TAC. The speaker embedding extraction module first extract the target speaker embedding $\mathbf{e} \in \mathbb{R}^{1 \times D}$, where D is the dimension of the frequency bins. Then the

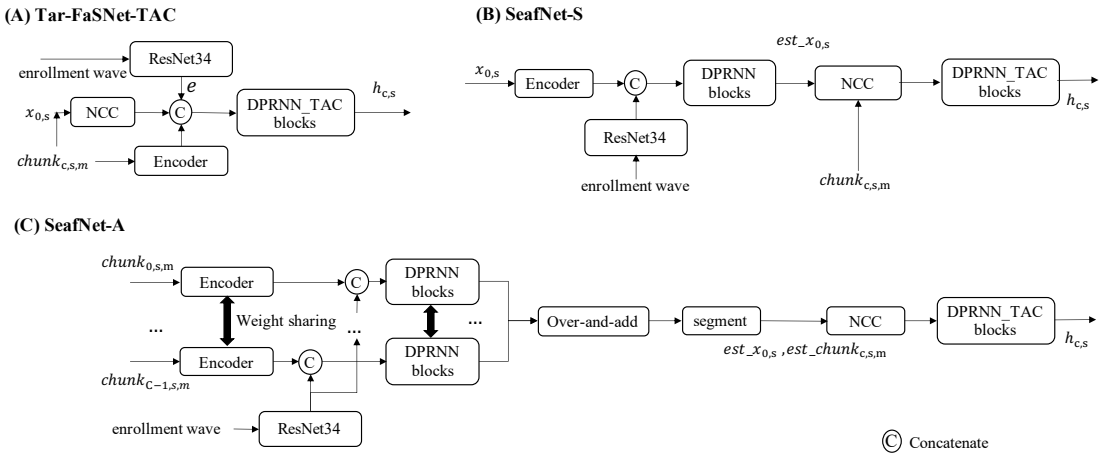


Fig. 3. System flowchart of the SeafNet models. (A) Multi-channel speaker extraction model based on FaSNet-TAC. (B) The SeafNet-S model applies speaker extraction on each microphone. (C) The SeafNet-A model only applies speaker extraction on reference microphone.

speaker extraction module calculates the NCC feature $\mathbf{g}_{c,s,m}$ between the reference signal and multi-channel inputs. The obtained $\mathbf{e}$ and $\mathbf{g}_{c,s,m}$ are then concatenated with the encoder of the context segments $\mathbf{chunk}_{c,s,m}$ to derive the target speaker's filter coefficients $\mathbf{h}_{c,s}$, where $c \in 0, ..., C-1$:

$$\mathbf{h}_{c,s} = \mathscr{F}_{TAC}([E(\mathbf{chunk}_{c,s,m}), \mathbf{g}_{c,s,m}, \mathbf{e}]) \tag{6}$$

Finally, the model generate the estimated target source by filtering multi-channel inputs with $\mathbf{h}_{c,s}$.

C. SeafNet

Although the solutions in Section II-B work well, the model's size and computation costs are significantly increased when processing fusion features. In order to address this problem, we further propose the SeafNet framework. Figure 3 shows the two different structures of SeafNet, named SeafNet-A and SeafNet-S, respectively.

1) SeafNet-A: SeafNet-A in Figure 3 (C) first applies single-channel speaker extraction on the context segments of all input microphones:

$$\mathbf{est}_{c,s,m} = \mathscr{F}([E(\mathbf{chunk}_{c,s,m}), \mathbf{e}]), \quad c = 0, ..., C-1 \tag{7}$$

where $\mathscr{F}(*)$ indicates the DPRNN blocks, C indicates the number of microphones, and $\mathbf{e}$ is the target speaker embedding. Next, we obtain the center segments $\mathbf{est_x}_{c,s}$ and context segments $\mathbf{est_chunk}_{c,s,m}$ of the pre-separated output through overlap-and-add and split operation. Then SeafNet-A calculates the NCC feature between $\mathbf{est_x}_{0,s}$ and $\mathbf{est_chunk}_{c,s,m}$ and estimates the target speaker's filter coefficients:

$$\mathbf{g}_{c,s,m} = \frac{\mathbf{est_x}_{0,s}(\mathbf{est_chunk}_{c,s,m})^T}{||\mathbf{est_x}_{0,s}||_2||\mathbf{est_chunk}_{c,s,m}||_2} \tag{8}$$

$$\mathbf{h}_{c,s} = \mathscr{F}_{TAC}([E(\mathbf{est_chunk}_{c,s,m}), \mathbf{g}_{c,s,m}]) \tag{9}$$

where $\mathscr{F}_{TAC}(*)$ indicates the DPRNN with TAC blocks. Finally, $\mathbf{h}_{c,s}$ are convolved with $\mathbf{est_chunk}_{c,s,m}$ and generate the separation results like the original FaSNet-TAC.

2) SeafNet-S: SeafNet-S in Figure 3 (B) first applies single-channel speaker extraction only on the reference microphone's center segments, not on the context segments of all microphones as in the SeafNet-A model:

$$\mathbf{est}_{0,s} = \mathscr{F}([E(\mathbf{x}_{0,s}), \mathbf{e}]) \tag{10}$$

where $E(\mathbf{x}_{0,s})$ indicates the encoder of the reference microphone's center segments. $\mathbf{e}$ is the target speaker embedding. Then SeafNet-S calculates the NCC feature $\mathbf{g}_{c,s,m}$ between $\mathbf{est}_{0,s}$ and all microphones' context segments $\mathbf{chunk}_{c,s,m}$. The target speaker's filter coefficients can be calculated as:

$$\mathbf{g}_{c,s,m} = \frac{\mathbf{est}_{0,s}(\mathbf{chunk}_{c,s,m})^T}{||\mathbf{est}_{0,s}||_2||\mathbf{chunk}_{c,s,m}||_2} \tag{11}$$

$$\mathbf{h}_{c,s} = \mathscr{F}_{TAC}([E(\mathbf{chunk}_{c,s,m}), \mathbf{g}_{c,s,m}]) \tag{12}$$

At last, $\mathbf{h}_{c,s}$ are convolved with $\mathbf{chunk}_{c,s,m}$ and generate the final estimation.

III. Experiment Setup

A. Dataset

Datasets for speaker embedding extraction module: We use the development set of the VoxCeleb2 [23], which contains over 1 million utterances from 5991 speakers, to train a ResNet34-based speaker verification model. We use the test set of the VoxCeleb1 [23] for testing. The test set contains 4715 utterances from 40 celebrities and has no overlap with the identities in the VoxCeleb2.

Datasets for speaker extraction module: We use the same dataset in [16], a multi-channel two-speaker noisy reverberant dataset with both ad-hoc and fixed geometry microphone arrays. Each class of microphone array includes 20000, 5000, and 3000 4-second long utterances for training, validation and test, respectively. The source utterances in these multi-channel mixed waveform are selected from the 100-hour Librispeech [24], and we randomly select an additional utterance as the enrollment waveforms for each source.

TABLE I

RESULTS ON THE 6-MIC FIXED GEOMETRY ARRAY. **SV**: WITH OR WITHOUT A SPEAKER VERIFICATION MODEL IN THE INFERENCE STAGE. **EMB**: WITH OR WITHOUT A LOCAL TARGET SPEAKER EMBEDDING EXTRACTED IN ADVANCE. **FASNET-TAC+SIM**: ORIGINAL SINGLE-STAGE FASNET-TAC COMBINED WITH TARGET SOURCE SELECTION THROUGH COSINE SIMILARITY IN THE INFERENCE STAGE. **TAR-FASNET-TAC**: MULTI-CHANNEL TARGET SPEAKER EXTRACTION MODEL BASED ON THE SINGLE-STAGE FASNET-TAC AND RESNET34 MODELS DESCRIBED IN SECTION II-B.

Methods	SV	Emb	# of Params(M)	FLOPS(G)	SISNRi(dB)
Baseline: FaSNet-TAC + Sim	w/	w/	9.0(2.9+6.1)	64.31(55.09 + 2*4.61)	10.92
Proposed: Tar-FaSNet-TAC	w/o	w/	6.3	148.10	**12.15**
Proposed: SeafNet-A	w/o	w/	3.6	82.30	11.64
Proposed: SeafNet-S	w/o	w/	**3.6**	**37.03**	11.62

TABLE II

RESULTS ON THE AD-HOC ARRAY. THE PARAMS, FLOPS, AND SISNRI OF THE METHODS ARRAYS WITH 2 / 4 / 6 MICROPHONES ARE REPORTED.

Methods	SV	Emb	# of Params(M)	FLOPs(G)	SISNRi(dB)
Baseline: FaSNet-TAC + Sim	w/	w/	9.0	29.2 / 46.75 / 64.31	9.72 / 9.67 / 11.74
Proposed: Tar-FaSNet-TAC	w/o	w/	6.5	50.98 / 99.54 / 148.10	**11.01 / 12.55 / 12.94**
Proposed: SeafNet-A	w/o	w/	3.6	28.24 / 55.27 / 82.30	10.5 / 11.59 / 12.26
Proposed: SeafNet-S	w/o	w/	**3.6**	**19.18 / 28.11 / 37.03**	10.10 / 11.55 / 12.15

B. Implementation details

Speaker embedding extraction module: We use the default ResNet34V2[1] to extract a 256-dimensional embedding for the target speaker. Specifically, we use the SAP [25] encoder and the ge2e [26] loss function in our training.

Speaker extraction module: We have compared the performance of four models: 1) FaSNet-TAC[2]+Sim, 2) Tar-FaSNet-TAC, 3) proposed SeafNet-A, and 4) proposed SeafNet-S. The results in 1) are the baseline. We use 4 DPRNN blocks in all models and the hyperparameters are the same as [16]. Specifically, we apply TAC between all the DPRNN blocks in 1) and 2), while in 3) and 4) we use it only in the last two blocks. The length of the center segment W is 4ms and the length of context segment L is 16ms. The scale-invariant SNR (SISNR) [27] is used as the training target. The SISNR improvement (SISNRi) is used as the separation performance metric. In addition, we evaluate the FLOPs and Params information of all models.

IV. RESULTS AND DISCUSSIONS

Table I shows the experimental results on the 6-mic fixed geometry array. Among all schemes, the Tar-FaSNet-TAC model obtained the maximum SI-SNRi improvement, which achieved 1.23 dB SISNRi improvement over the baseline. However, its FLOPs' cost increases significantly because the dimensionality of the fusion features is much larger than the one of the original model inputs. The SeafNet-A further reduces the number of model parameters, but the FLOPs cost is still higher than the baseline due to the iterative pre-separation of each microphone. The SeafNet-S model has a comparable separation performance to the SeafNet-A model, which achieves a 6.4% relative improvement over the FaSNet-TAC+Sim baseline model. It can be noticed that the proposed SeafNet-S method has the lowest model size and computational cost.

Table II shows the results on ad-hoc array and we only report the results on 2, 4, and 6 microphones to compare with [16]. The results in Table II shows similar trends as in Table I. Specifically, the Tar-FaSNet-TAC model achieves the largest SISNR improvement, but accompanied with an increase in FLOPs' costs. The SeafNet-S model achieves 3.5% relative SISNRi improvement, 42% relative FLOPs reduction, and 60% relative Params reduction against the baseline (6-mic), respectively. We note that in the 2-mic scenario, the SeafNet-S still outperforms the baseline, but there is a slight degradation in the performance compared to the SeafNet-A model. It may be because the output of pre-separation account for a notable contribution to the calculation of the NCC features, which could greatly affect the separation performance of the model.

Although the SISNRi performance of our proposed SeafNet-A and SeafNet-S is slightly lower than that of Tar-FaSNet-TAC, the SeafNet model outperforms the Tar-FaSNet-TAC model by a large margin in terms of Params' and FLOPs' cost. In addition, the results in Table I and Table II show that SeafNet-S outperforms the baseline on both fixed geometry array and ad-hoc array with lower Params' and FLOPs' cost. It means that for the speaker extraction task, the improvement brought by our proposed SeafNet models is quite effective.

V. CONCLUSION

In this work, we extend the FaSNet-TAC model to multi-channel target speaker extraction task and propose a Speaker extraction-and-filter Network (SeafNet). SeafNet first extracts the target source from the signal of the reference microphone by utilizing auxiliary speech. Then, SeafNet calculates the NCC feature between the estimated target source and multi-channel input. Finally, SeafNet derives a set of filter coefficients using the fusion feature of the input and NCC feature. The filter result of the input is our final output. Experimental results show that our proposed SeafNet achieves a favorable separation performance improvement while maintaining relatively small model size and low computational costs.

VI. ACKNOWLEDGEMENT

This research is funded in part by the Science and Technology Program of Suzhou City (SYC2022051) and OPPO. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

[1]https://github.com/clovaai/voxceleb_trainer

[2]https://github.com/yluo42/TAC

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. of ICASSP*, IEEE, 2017, pp. 246–250.

[3] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. of ICASSP*, IEEE, 2017, pp. 241–245.

[4] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of ICASSP*, IEEE, 2018, pp. 696–700.

[5] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. of ICASSP*, IEEE, 2020, pp. 46–50.

[6] T. Chou, "Frequency-independent beamformer with low response error," in *Proc. of ICASSP*, IEEE, vol. 5, 1995, pp. 2995–2998.

[7] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[8] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2009.

[9] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[10] X. Zhang, Z.-Q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust asr," in *Proc. of ICASSP*, IEEE, 2017, pp. 276–280.

[11] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *Proc. of Interspeech*, 2016, pp. 1981–1985.

[12] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Dnn-based speech mask estimation for eigenvector beamforming," in *Proc. of ICASSP*, IEEE, 2017, pp. 66–70.

[13] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. of ICASSP*, IEEE, 2017, pp. 271–275.

[14] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Processing Letters*, vol. 28, pp. 1370–1374, 2021.

[15] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. of ASRU*, IEEE, 2019, pp. 260–267.

[16] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. of ICASSP*, IEEE, 2020, pp. 6394–6398.

[17] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-net: Target speaker separation via attention-based neural network," in *Proc. of Interspeech*, 2020.

[18] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.

[19] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *Proc. of ICASSP*, IEEE, 2021, pp. 6089–6093.

[20] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized spatio-temporal rnn beamformer for target speech separation," in *Proc. of Interspeech*, 2021.

[21] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-spex: Localized target speaker extraction," in *Proc. of ICASSP*, IEEE, 2022, pp. 7287–7291.

[22] Y. Kwon, H. S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: Lessons from VoxSRC 2020," in *Proc. of ICASSP*, IEEE, 2021.

[23] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2020.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. of ICASSP*, IEEE, 2015, pp. 5206–5210.

[25] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Proc. of Interspeech*, 2017, pp. 1517–1521.

[26] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. of ICASSP*, IEEE, 2018, pp. 4879–4883.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *Proc. of ICASSP*, IEEE, 2019, pp. 626–630.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Modeling Suprasegmental Information Using Finite Difference Network for End-to-End Speaker Verification

Jin Li*, Man-Wai Mak*, Nan Yan†, Lan Wang†
* Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China
† Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

***Abstract*—In recent years, using raw waveforms as input to deep networks has been widely explored for speaker verification systems that process speech signals at the segmental level. A critical issue of such an approach is that the front-end network with a small kernel fails to capture the suprasegmental information, such as the intonation patterns and prosody that span longer than one second. This paper proposes a novel framework that can capture the segmental and suprasegmental information after the first convolutional layer. Concretely, suprasegmental information is obtained from the first-order finite difference of two consecutive suprasegmental envelopes estimated by Hilbert transforms. Experimental evaluations on the Voxceleb dataset show that combining segmental and suprasegmental features can reduce the EER of an end-to-end system by 27%. To our best knowledge, this is the first attempt to incorporate suprasegmental information for end-to-end speaker verification.**

I. Introduction

Speaker verification (SV) aims to verify whether the same speaker speaks the enrollment and test utterances. In previous studies of deep speaker embedding, acoustic features – such as filterbanks – were commonly used as the input to deep neural networks (DNNs) [1]. By contrast, many recent studies [2]–[6] explored using raw waveforms as the input and demonstrated satisfactory performance. For example, to extract useful information from waveforms, Ravanelli and Bengio [2] designed filters parametrized by sinc functions for the first layer of an end-to-end (e2e) network. In RawNet [6], convolutional residual blocks are used for frame-level processing, and GRUs are then used for aggregating the frame-level features into an utterance-level embedding. Feature map scaling and shifting have been applied to increase speaker discrimination in RawNet [3]. More recently, Kim *et al.* [5] used the ResNeXt structure to make the front end deeper and wider. Jung *et al.* [7] improved RawNet2 by exploring the combination of Res2Net and multi-layer feature aggregation. Wav2Spk [4] sets the stride and kernel size so that the output corresponds to 30ms and 10ms frameshift, similar to wav2vec [8]. Because these previous studies utilized small convolutional kernels, the resulting networks can only capture speakers' characteristics at the segmental level. This restriction means that depending on the network depth, the time scale of these networks is only one to several phones.

We propose a novel framework to increase the time scale to extract both subsegmental and suprasegmental features (see Figure 1). Like conventional CNN-based speaker embedding networks, segmental information – such as vowels and consonants – can be preserved. Importantly, the framework is also designed to preserve suprasegmental information such as intonation patterns, rhythm, and pitch [9]. Both segmental and suprasegmental features enable humans to distinguish speakers. The former is affected by speakers' physical traits, whereas

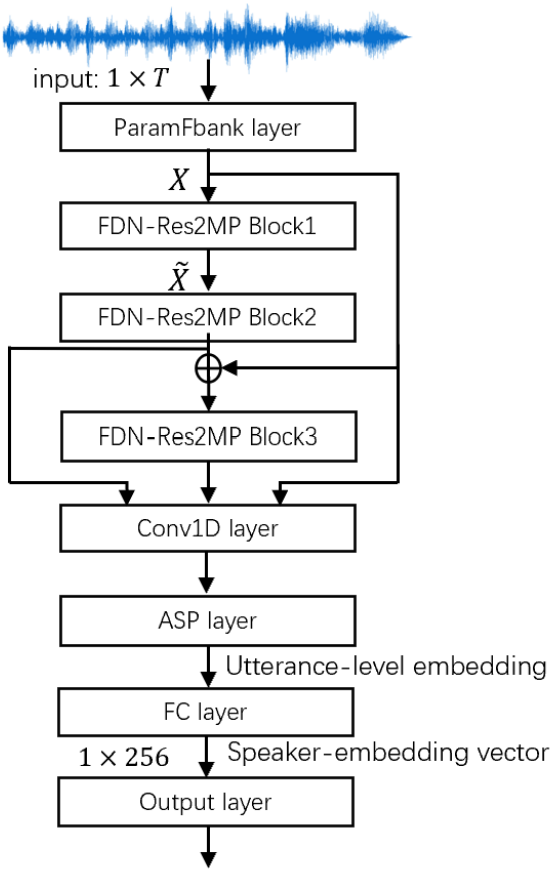


Fig. 1. The architecture of the proposed framework: FDN-Res2MP-Net. It uses pre-emphasized speech waveforms as input and produces speaker posteriors as output.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the latter is determined by how speakers use spoken languages to convey the messages [10]. Suprasegmental features have been used for speaker recognition [11], [12]. However, they have not been explored in an e2e manner with raw waveforms as input.

We extract suprasegmental information after the first convolutional layer because the first convolutional layer with small kernels emphasizes speaker information at low frequencies (below 1000 Hz) and models the fundamental frequency of speech signals [13]. The suprasegmental features are obtained by computing the first-order finite difference between two adjacent suprasegmental envelopes, each with around 3 seconds. The suprasegmental features are combined with segmental features at the output of each FDN-Res2MP block in Figure 1. Because the envelopes change slowly and have duration covering words, they contain suprasegmental information across phones, syllables, and phrases [14], [15]. As a result, it is possible to extract speaker-discriminative prosody information from the first-order finite difference between two adjacent envelopes. In addition to using envelopes, we have also attempted to capture the prosody from the temporal feature maps at the output of the first convolutional layer, as shown in Figure 2(b). Our experimental results show that the suprasegmental features extracted by this approach benefit speaker verification.

Our main contributions are as follows:

1) A novel plug-and-play module with first-order Finite Difference Networks (FDN) is proposed to capture suprasegmental features that preserve prosody information.
2) We demonstrate that useful suprasegmental features can be extracted from raw temporal feature maps without computing the feature maps' envelopes.
3) We demonstrate that combining segmental and suprasegmental features enables e2e SV systems to achieve very competitive performance.

II. Methodology

A. System Architecture

Figure 1 shows the structure of the proposed speaker embedding networks. The building blocks are similar to that of RawNet3 [7]. It comprises a parametrized complex-valued analytic filterbanks (ParamFbank) layer [16], residual blocks, an attentive statistics pooling (ASP) layer, a 1D convolution layer, a fully-connected (FC) layer, and an output layer. The ParamFbank is an extension of the SincNet filter that captures speakers' pitch and formants. The motivation for using complex-valued representations is that they allow perfect synthesis via the overlap-and-add method. The output of the first layer is input to three consecutive residual blocks, named FDN-Res2MP in Figure 1.* The three consecutive blocks capture information at different temporal scales. Unlike RawNet3, α-feature map scaling (AFMS) is not applied. Instead, an FDN is added to the front of each residual block to extract suprasegmental information. Next, an attentive statistics pooling layer aggregates the frame-level features to an utterance-level embedding, which is then fed into a fully-connected layer. The speaker embedding is extracted from the output of the fully-connected layer. The number of nodes in the output layer is the same as the number of training speakers.

During training, AAM-softmax loss [17] was employed to compute the error between the speaker labels and speaker posteriors.

B. Operations of FDN-Res2MP Blocks

In one mode of operation of the FDN-Res2MP Blocks in Figure 1, the FDN extracts suprasegmental information from the envelopes of the feature maps ($[x_1, \ldots, x_{C/\alpha}]$ in Figure 2(a)). We used Hilbert transforms [26] to compute the envelopes. We denote the FDN as "FDN-E" when the FDN-Res2M Blocks are operated in this mode. In another mode of operation, the FDN is directly operated on the raw feature maps without computing the envelopes, as shown in Figure 2(b). We denote the FDN as "FDN-W" when the FDN-Res2M Blocks are operated in this mode.

Let $X = \{x_i\}_{i=1}^{C} \in \mathbb{R}^{C \times T}$ be the feature maps of a frame-level representation after the first ParamFbank layer, where C is the number of channels and T is the sequence duration. The number of channels in X is first squeezed from C to $\frac{C}{\alpha}$ by 1D convolution with kernel size of 1. This squeezing operation is to reduce computation. Then, the 1D-convolutional filters' outputs are partitioned into overlapping suprasegments by a large sliding window along the time axis for each channel. Figure 2(a) shows an example of two suprasegments with length T'. Note that the number of suprasegments depends on the signal duration (T) and the window overlapping. For clarity, Figure 2(a) illustrates two suprasegments and indicates them by the segment labels a and b, respectively.† The envelopes of two segments $e_{\cdot,a}$ and $e_{\cdot,b}$ are obtained by Hilbert transforms [26]. The prosody is then calculated by computing the first-order finite difference between the two suprasegments for each channel:

$$\hat{e}_i = \mathsf{ReLU}(\mathsf{Conv}(e_{i,b}, W_2)) - \mathsf{ReLU}(\mathsf{Conv}(e_{i,a}, W_1)), \quad (1)$$

where W_1 and W_2 are learnable parameters for 1D convolution and i represents the i^{th} channel. The prosody across different channels is summarized by global average pooling along the time axis:

$$\bar{e}_i = \frac{1}{T'} \sum_{t=1}^{T'} \hat{e}_{i,t} \qquad i = 1, \ldots, \frac{C}{\alpha}. \quad (2)$$

The pooled vector $\bar{e}$, which contains the global prosody information, is projected to a C-dimensional space through a fully connected (FC) layer. A sigmoid function is then applied independently to individual dimensions to scale the values to a range compatible with the ParamFbank's output, i.e.,

*The name Res2MP is adopted from [7].

†The concept can be generalized to multiple suprasegments. In such a case, the finite difference is applied to adjacent suprasegments, followed by averaging their differences.

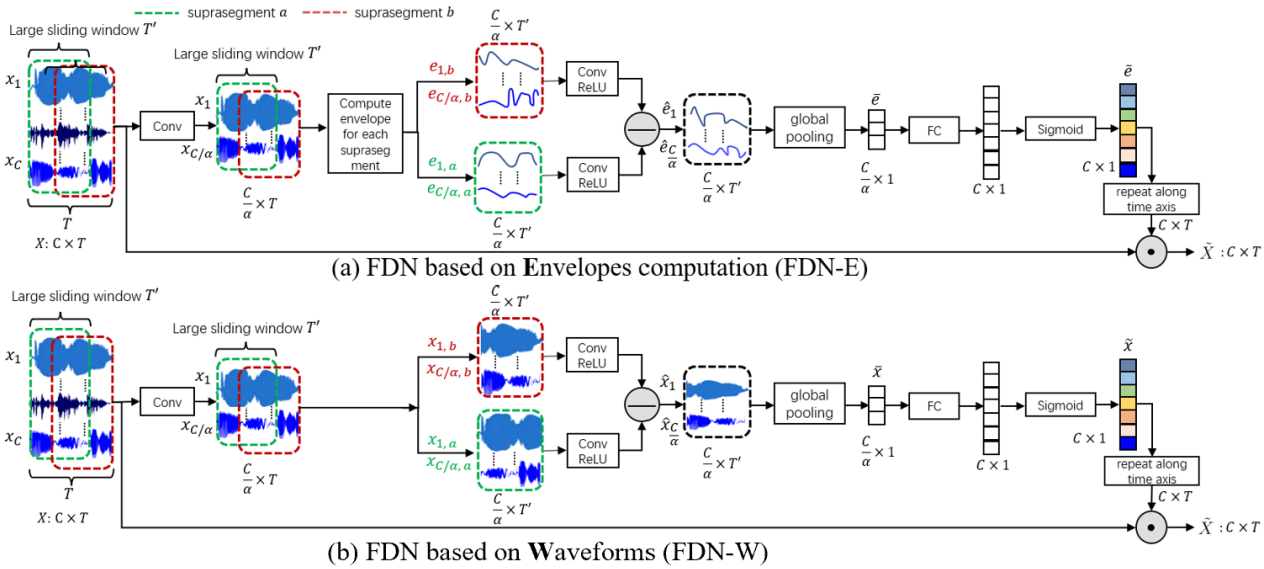


Fig. 2. Signal flow diagrams showing the operation of the FDN module in Figure 1. The module can be inserted into the front of each residual block in Figure 1. In (a), Conv represents 1D convolution aiming to reduce computation. It squeezes the number of channels from C to C/α. Then, the suprasegments are produced using a large sliding window with 3.07s and a frameshift of 0.47s. The prosody is computed by the first-order difference between the two envelopes obtained from the Hilbert transform. Global pooling is used to summarize the prosody corresponding to each channel. Then, the number of channels is recovered to the same as that of the input using a fully connected (FC) layer. The sigmoid activation function is then applied to the channels independently. Finally, the prosody is fused with the input waveforms X along each channel using a channel-wise dot product. In (b), the suprasegmental information is extracted directly from the feature maps.

$\tilde{e} = \sigma(f(\bar{e}, W_3))$, where $f()$ is the FC layers' operation and W_3 denotes its learnable weights. $\tilde{e}$ is repeated T times along the time axis before performing elementwise multiplication (Hadamard product) with the input feature map as shown in Figure 2(a).

The FDN-W in Figure 2(b) is similar to FDN-E, except that the finite differences are based on the feature maps instead of the envelopes.

III. Experiments

A. Datasets

We used the VoxCeleb2 development set [18] for training and the VoxCeleb1 test set [27] to evaluate the model performance. VoxCeleb2 consists of more than one million utterances from 6112 speakers, and VoxCeleb1 contains approximately 330 hours of recordings from 1251 speakers for text-independent tasks.

B. Experimental Configurations

1) Backbones: Two types of neural networks were used as the backbone networks: RawNet and Res2MP-Net.

RawNet. It is the backbone network for the first baseline system. This backbone is similar to the architecture in [6], but the speaker embedding dimension is 1024. A GRU layer with 1024 nodes aggregates frame-level features into an utterance-level embedding. Following [6], utterances were either cropped or duplicated (concatenated until the desired length has been reached) to 3.59s in the training phase. Softmax loss was used during training. There was no data augmentation during the training phase.

FDN-W-RawNet. It is a RawNet with FDNs, where an FDN was inserted in the front of the "Res block" in Table 1 of [6]. In the FDN, the sliding window size (T' in Figure 2) and shift are 3.2s and 0.5s, respectively. The channel reduction ratio α was set to 8. Other settings are the same as RawNet.

Res2MP-Net. It is the backbone network for the second baseline system. The network does not have any FDN module. It was trained using randomly cropped 4-second utterances. AAM-softmax loss [17] was applied with a margin of 0.1 and a scale of 30. We added MUSAN noise [28] and applied room impulse response (RIR) simulations [29] as data augmentation.

FDN-Res2MP-Net. It is an Res2MP-Net with an FDN for each Res2MP block in Figure 1. For the FDNs, the channel reduction ratio α was set to 4. The sliding window size and shift were set to 3.07s and 0.47s, respectively. Other settings are the same as Res2MP-Net.

For all networks, an Adam optimizer [30] with a learning rate of 0.001 was used, and weight decay of $5e^{-5}$ was set.

C. Results Analysis

Performance is reported in terms of Equal Error Rate (EER), where the false accept rate and false reject rate are equal, as well as the minimum Detection Cost Function (minDCF).

As shown in Table I, the system with FDN-W outperforms the baseline system (Res2MP-Net), which does not contain any FDN. This result demonstrates that suprasegmental information improves the system's accuracy and helps improve speaker

TABLE I
COMPARISON WITH SYSTEMS ON THE VOXCELEB1 TEST SET. (†: REIMPLEMENT USING OFFICIAL SOURCE CODE)

Model or Author	Front-end Network	Input	Training Set	Data Augmentation	EER(%)	minDCF
Chung *et al.* [18]	ResNet-34 variants	Spectrogram	Vox2-dev	No	4.19	N/A
Xie *et al.* [19]		Spectrogram	Vox2-dev	No	3.22	N/A
Kim *et al.* [20]		Raw waveform	Vox2-dev	No	3.15	N/A
Jung *et al.* [6]		Raw waveform	Vox2-dev	No	4.80	N/A
Jung *et al.* [3]		Raw waveform	Vox2-dev	No	2.48	N/A
Zhu *et al.* [21]		Raw waveform	Vox2-dev	No	2.60	0.2390
RawNet		Raw waveform	Vox2-dev	No	2.50	N/A
FDN-W-RawNet (Ours)		Raw waveform	Vox2-dev	No	2.44	N/A
Zhu *et al.* [22]	RawNet2 variants	Raw waveform	Vox2-dev	MUSAN + RIR	1.98	0.230
Peng *et al.* [23]		Raw waveform	Vox2-dev	MUSAN + RIR	1.92	0.1370
Jung *et al.* [7] †		Raw waveform	Vox2-dev	MUSAN + RIR	1.73	0.1099
Res2MP-Net		Raw waveform	Vox2-dev	MUSAN + RIR	1.74	0.1079
FDN-E-Res2MP-Net (Ours)		Raw waveform	Vox2-dev	MUSAN + RIR	1.61	0.0966
FDN-W-Res2MP-Net (Ours)		Raw waveform	Vox2-dev	MUSAN + RIR	1.54	0.0985
Desplanques *et al.* [24]	TDNN variants	MFCC	Vox2-dev	MUSAN + RIR	0.87	0.1066
Ravanelli *et al.* [25]		Fbank	Vox2-dev	MUSAN + RIR +SpecAugment	0.69	N/A
Lin *et al.* [4]		Raw waveform	Vox2-dev+Vox1-dev	MUSAN+RIR	1.95	0.2030
Jung *et al.* [7]		Raw waveform	Vox2-dev+Vox1-dev	MUSAN+RIR+Masking +Speed pertubation	0.89	0.0659
FDN-W-Res2MP-Net (Ours)		Raw waveform	Vox2-dev+Vox1-dev	MUSAN+RIR	1.42	0.0925

TABLE II
ABLATION STUDY OF APPLYING (✓) OR NOT APPLYING (✗) FDN-W TO DIFFERENT RES2MP BLOCKS IN FIGURE 1.

Block1	Block2	Block3	EER(%)	minDCF
✗	✗	✗	1.86	0.1199
✓	✗	✗	1.81	0.1118
✗	✓	✗	1.70	0.1112
✗	✗	✓	1.66	0.1077
✓	✓	✓	**1.54**	**0.0985**

TABLE III
NUMBER OF PARAMETERS (MILLION) AND PERFORMANCE OF FDN-W-RES2MP-NET AT DIFFERENT REDUCTION RATIOS α ON THE VOXCELEB1 TEST SET. (†: REIMPLEMENT USING OFFICIAL SOURCE CODE)

α	#Params (million)	EER (%)	minDCF
2	20.164	1.55	0.1011
4	16.647	**1.54**	**0.0985**
8	15.497	1.73	0.1099
16	15.074	1.64	0.1024
32	14.901	1.58	0.0987
Res2MP-Net	**14.753**	1.74	0.1079
Jung *et al.* [7] †	17.815	1.73	0.1099

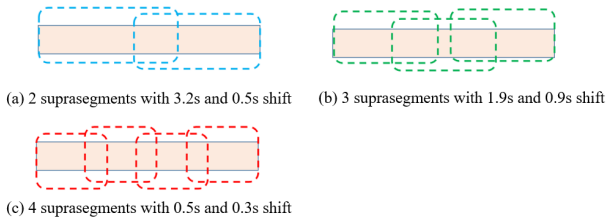
discrimination. Results also suggest that it is possible to train convolutional filters to capture prosody information from raw speech in a data-driven way. However, the performance does not improve when the network complexity increases. To balance the performance and the computation complexity, we found that $\alpha = 4$ is appropriate for the FDN-Res2MP-Net. Table I also compares the performance of FDN-based systems with other SV systems based on raw waveforms and hand-craft features. The results show that FDN-E improves performance on the VoxCeleb1 test set. However, the performance of FDN-E is poorer than FDN-W. The inferior performance of FDN-E is caused by the strong assumption that speech is a modulated signal.

Table II shows the ablation study that explores the positions

TABLE IV
PERFORMANCE OF DIFFERENT SEGMENT DURATIONS IN EACH FDN-W IN THE FDN-W-RAWNET ON THE VOXCELEB1 TEST SET.

Segment Duration	Number of Segments	EER (%)
3.2	2	**2.44**
1.9	3	2.48
0.5	4	2.49

and the number of FDN-W modules inserted into the network. The results show that FDN-W performs better when it is added to the deeper Res2MP block. In addition, inserting the FDN-W into all blocks achieves the best performance because more suprasegmental information can be obtained.



(a) 2 suprasegments with 3.2s and 0.5s shift (b) 3 suprasegments with 1.9s and 0.9s shift

(c) 4 suprasegments with 0.5s and 0.3s shift

Fig. 3. Different numbers of suprasegments and different durations for the FDN-W-RawNet.

To verify the effectiveness of the squeezing operation in FDNs, we varied the reduction ratio α in the FDN-Res2MP-Net. Table III shows the performance and the number of parameters when α increases. The results show that setting α to 4 achieves a good balance between network size and performance. This setting also makes the FDN-Res2MP-Net outperforms the baseline Res2MP-Net.

To investigate the importance of suprasegment duration, we varied the window size and window shift shown in Figure 3, where each suprasegment is denoted by $\hat{x}$ in Figure 2(b). We applied the windowed suprasegments in Figure 3 to FDN-W-

RawNet. When the number of suprasegments is larger than two, the differences between adjacent suprasegments were linearly combined by trainable attention weights. The results are shown in Table IV, which show that when the suprasegment duration decreases, the EER increases, suggesting that long suprasegments contain more prosody information.

IV. Discussion

Stadelmann and Freisleben [31] suggested that incorporating suprasegmental information is likely to enhance speaker clustering performance. In their experiments, they used a suprasegment length of up to 496 ms to capture speaker-specific sounds. Stadelmann *et al.* [32] further showed that increasing the suprasegment length from 400 ms to 1000 ms can provide useful prosodic information to reduce the misclassification rate of a speaker clustering task.

Our framework uses a finite difference network spanning across words to capture suprasegmental information. The difference between two overlapping suprasegmentals further emphasizes the suprasegmental information. The inclusion of suprasegmental information offers two advantages. Firstly, it enables the front-end network to focus on the pitch and F1 profiles, which cannot be extracted without the FDN in the front end. Secondly, the suprasegmental features compensate for the limitation of small convolutional kernels in end-to-end systems where raw waveforms are used as input.

V. Conclusions and Future Work

This paper proposes a novel framework that uses segmental and suprasegmental information for end-to-end speaker verification. The segmental features convey speaker information at the scale of phones, whereas suprasegmental features characterize the prosody of speakers. We achieve this goal through a finite difference network. Experimental results show that combining phone- and prosody-level features can improve the performance of the VoxCeleb1 test set. However, the average utterance duration of VoxCeleb is around 4 seconds, which limits the number of suprasegments within an utterance. In the future, we will explore other corpora that have longer utterances. In addition, we will visualize and analyze the influence of FDNs through the time-frequence response.

VI. Acknowledgements

This work was supported by the RGC of Hong Kong SAR, Grant No. 15210122, National Key R&D Program of China (2020YFC2004100), National Natural Science Foundation of China (NSFC 62271477), Shenzhen Science and Technology Program (KQTD20200820113106007) and Shenzhen Science and Technology Program, Grant Award (JCYJ20210324115810030, JCYJ20220818101217037, JCYJ20220818101411025).

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[2] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. SLT*, 2018, pp. 1021–1028.

[3] J. Jung, S. Kim, H. Shim, J. Kim, and H. Yu, "Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms," in *Proc. Interspeech*, 2020, pp. 1496–1500.

[4] W.-W. Lin and M.-W. Mak, "Wav2spk: A simple DNN architecture for learning speaker embeddings from waveforms.," in *Proc. Interspeech*, 2020, pp. 3211–3215.

[5] J. Kim, H. Shim, J. Heo, and H. Yu, "RawNeXt: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies," in *Proc. ICASSP*, 2022, pp. 7647–7651.

[6] J. Jung, H. Heo, J. Kim, H. Shim, and H. Yu, "RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," in *Proc. Interspeech*, 2019, pp. 1268–1272.

[7] J. Jung, Y. Kim, H. Heo, B. Lee, Y. Kwon, and J. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022, pp. 2228–2232.

[8] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[9] N. J. Veenendaal, M. A. Groen, and L. Verhoeven, "The contribution of segmental and suprasegmental phonology to reading comprehension," *Reading Research Quarterly*, 2016.

[10] T. B. Amin, P. Marziliano, and J. S. German, "Glottal and vocal tract characteristics of voice impersonators," *IEEE Transactions on Multimedia*, vol. 16, pp. 668–678, 2014.

[11] D. Pati and S. Prasanna, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information," *International Journal of Speech Technology*, vol. 14, pp. 49–64, 2011.

[12] F. Farahani, P. G. Georgiou, and S. S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," in *Proc. ICASSP*, 2004, pp. 89–92.

[13] H. Muckenhirn, M. M. Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proc. ICASSP*, 2018, pp. 4884–4888.

[14] B. R. Myers, M. D. Lense, and R. L. Gordon, "Pushing the envelope: Developments in neural entrainment to speech and the biological underpinnings of prosody perception," *Brain Sciences*, vol. 9, p. 70, 2019.

[15] M. A. Pitt and A. G. Samuel, "The use of rhythm in attending to speech.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 16, p. 564, 1990.

[16] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *Proc. ICASSP*, 2020, pp. 6364–6368.
[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.
[18] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
[19] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. ICASSP*, 2019, pp. 5791–5795.
[20] S. Kim, J. Jung, H. Shim, J. Kim, and H. Yu, "Segment aggregation for short utterances speaker verification using raw waveforms," in *Proc. Interspeech*, 2020, pp. 1521–1525.
[21] G. Zhu, F. Jiang, and Z. Duan, "Y-vector: Multiscale waveform encoder for speaker embedding," in *Proc. Interspeech*, 2021, pp. 96–100.
[22] G. Zhu, F. Cwitkowitz, and Z. Duan, "A study of the robustness of raw waveform based speaker embeddings under mismatched conditions," in *Proc. ICASSP*, 2022, pp. 7657–7661.
[23] J. Peng, X. Qu, J. Wang, *et al.*, "ICSpk: Interpretable complex speaker embedding extractor from raw waveform.," in *Proc. Interspeech*, 2021, pp. 511–515.
[24] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
[25] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.
[26] L. Marple, "Computing the discrete-time analytic signal via FFT," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2600–2603, 1999.
[27] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
[28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, 2015.
[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[31] T. Stadelmann and B. Freisleben, "Unfolding speaker clustering potential: A biomimetic approach," in *Proc. of the 17th ACM international conference on Multimedia*, 2009, pp. 185–194.
[32] T. Stadelmann, S. Glinski-Haefeli, P. Gerber, and O. Dürr, "Capturing suprasegmental features of a voice with RNNs for improved speaker clustering," in *Proc. Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop*, 2018, pp. 333–345.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Relevance of Quadrature Phase For Replay Detection in Voice Assistants (VAs)

Priyanka Gupta, Piyushkumar K. Chodingala, and Hemant A. Patil
DA-IICT, Gandhinagar, India.
E-mail: {priyanka_gupta, piyush_chodingala, hemant_patil}@daiict.ac.in

***Abstract*—There have been various studies involving Instantaneous Frequency (IF) estimation for Spoofed Speech Detection (SSD) task, such as the derivative of the phase obtained by Hilbert Transform (HT) approach and Energy Separation-based method. However, IF estimation by HT leads to lack of good temporal resolution. On the other hand, ESA-based method leads to excellent time resolution, however, lacks the relative phase information. Therefore, in this paper, we have proposed Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency using Quadrature Energy Separation Algorithm (CFCCIF-QESA) feature set, which merits of having an excellent time resolution as well as inclusion of the relative phase information. Hence, we illustrate the significance of incorporating the quadrature-phase component along with the in-phase component for SSD of replay detection in VAs. To that effect, we perform experiments on the Realistic Replay AttackF Microphone-Array Speech Corpus (ReMASC) dataset. Furthermore, the proposed CFCCIF-QESA feature set gives 28.71 and 29.89 %EER using GMM and CNN respectively, on Eval set. The proposed feature set is evaluated using various performance metrics, including EER and other confusion matrix-based metrics. Finally, the latency for CFCCIF-QESA and CFCCIF-ESA is presented, showing the better suitability of the proposed CFCCIF-QESA feature set w.r.t. practical deployment.**

Index Terms: Spoofed Speech Detection, Mutual Information, Quadrature Phase, Energy Separation Algorithm

I. Introduction

Voice enabled systems, such as Automatic Speaker Verification (ASV) systems and Voice Assistants (VAs) are widely used. However, these systems also undergo various security concerns, such as spoofing by mimicry, Voice Conversion (VC), Speech Synthesis (SS), and replay. Amongst these, replay attack is the easiest to mount but hard to detect, since it requires only simple recording and playback devices. In particular, an attacker without any technical expertise, can gain illegal access by mounting a replay attack. VAs, such as Google Home, Alexa, Siri are very commonly used these days in households. These voice-driven interfaces provide hands-free operation, where users can focus their attention to some other task, while giving voice commands to a VA. These systems have enhanced the ease of managing multiple tasks at a time, using voice commands, however, at the same time, they also need to have robust security [1]. Some of the attacks pertaining to VAs are hidden voice commands [2], adversarial attacks [3], and self-triggered attacks [4]. To that effect, countermeasures have been proposed in the literature to defend against these attacks [4], [5].

However, for the detection of spoofing attacks, particularly replay attacks, most of the countermeasures or Spoofed Speech Detection (SSD) systems have been designed in the context of ASV systems, leveraged by bi-annual organization of ASVSpoof challenges during INTERSPEEECH in 2015, 2017, 2019, and 2021. Notably, the countermeasures designed for ASV systems might not work as efficiently for VAs as well. This is due to substantial differences between an ASV system and a VA. In particular, an ASV system usually comprises a single microphone (i.e., one channel), and the user is in a controlled environment. However, this is not the case for VAs, which are designed with microphone array (i.e., multi-channel) to capture speech from various directions and varying distances. Given the usage scenarios of VAs, the user is not in a controlled environment and can have different acoustic environments.

To that effect, in order to detect replay attacks on VAs, Energy Separation Algorithm (ESA)-based Instantaneous Frequency (IF) has been proposed in the past in the context of ASV systems [6]. Another technique to estimate IF is by differentiating the phase of the analytic signal obtained using the Hilbert transform. However, this technique has a poor resolution in the time domain (as it requires a block (or frame) of speech data) [7]. To get a better time resolution, recently, an extension of Cochlear Filter Cepstral Coefficients-based Instantaneous Frequency (CFCCIF) was proposed as CFCCIF-ESA in [6]. The CFCCIF-ESA is based on cochlear filtering followed by estimation of IF using the conventional Teager Energy Operator (TEO)-based ESA [8], [7], [9]. TEO is known to estimate energy using only the *amplitude of real signals.* Consequently, in [6] the information pertaining to only the amplitude of the signal is used for the synthetic SSD task. Even though this approach is known to have better time resolution, it falls short of the contribution made by the quadrature-phase component, which is derived from HT. Therefore, in order to have both merits, we use the *extended-definition of TEO*, which leads to our proposition of CFCCIF-QESA (Quadrature-based ESA) feature set for replay detection in VAs.

Apart from this, the rationale behind choosing the quadrature-phase (90°) component along with the in-phase component is shown by a Mutual Information (MI)-based analysis, described in further detail in Section II. As a result, CFCCIF-QESA feature set is proposed in this work, and evaluated its performance on the Realistic Replay Attack Microphone-Array Speech Corpus (ReMASC) dataset. To fur-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

ther examine the efficiency of our proposed feature set, model-level comparison is performed on the Gaussian mixtures of the genuine *vs.* spoof class for the CFCCIF-ESA and CFCCIF-QESA feature sets. In this regard, the Kullback–Leibler (KL) divergence is estimated for both features. Lastly, we have estimated the latency period of the proposed feature set to investigate its capacity for practical SSD system deployment.

II. Proposed Work

This work proposes the incorporation of a quadrature phase in the existing CFCCIF-ESA feature set. This is done by using the extended definition of TEO.

A. Motivation for quadrature phase

So far, commonly used features have been based on the magnitude spectrum of the speech signal, however, phase information is also shown to be important for many speech applications [10], [11]. Here, we have first shown the significance of a quadrature phase ($90°$) over the other phase angles using MI-based analysis. Mathematically, using the joint and marginal probability density functions (*pdfs*) of S1 and S2, the MI is [12]:

$$I(S1;S2) = \int_{s1}\int_{s2} f_{S1S2}(s1,s2)\log_2\left(\frac{f_{S1S2}(s1,s2)}{f_{S1}(s1)f_{S2}(s2)}\right)ds1ds2. \tag{1}$$

Here, amount of information between a signal (as shown in eq. (2)) and its corresponding phase-shifted versions is measured by MI [12]. The signal in eq. (2) is an AM-FM signal (shown in Fig. 1 (a)), which is used to model speech signals [8], [14]. The MI obtained corresponding to varying relative phase of this signal is shown in Fig. 1 (b).

$$\begin{aligned} a(t) &= \left(1 + 0.7cos\left(\frac{\pi}{40}t\right)\right), \\ x(t) &= a(t)\cos\left(\frac{\pi}{3}t + 5sin\left(\frac{\pi}{30}t + \left(\frac{\pi}{4}\right)\right)\right). \end{aligned} \tag{2}$$

The *optimal* phase is the angle at which MI is observed to be the minimum. From Fig. 1 (b), it can be seen that the minimum MI obtained is achieved at 90 degrees, (i.e., quadrature).

Algorithm 1: Estimation of Mutual Information (MI)

Input: Input: Signals S1 and S2 having equal lengths n
Output: Output: MI
/* According to Freedman-Diaconis Rule [13] */
1 $width1_bin = 2\frac{IQR(u)}{\sqrt[3]{n}}$, where IQR is Inter-Quartile Range
2 $width2_bin = 2\frac{IQR(v)}{\sqrt[3]{n}}$
3 $avg_bins = avg(width1_bin, width2_bin)$
4 Generate histograms of S1 and S2 using avg_bins
5 Get H(U) and H(V) from the probability values computed from the histograms
6 Using the probability values, compute the joint probability
7 Estimate the MI

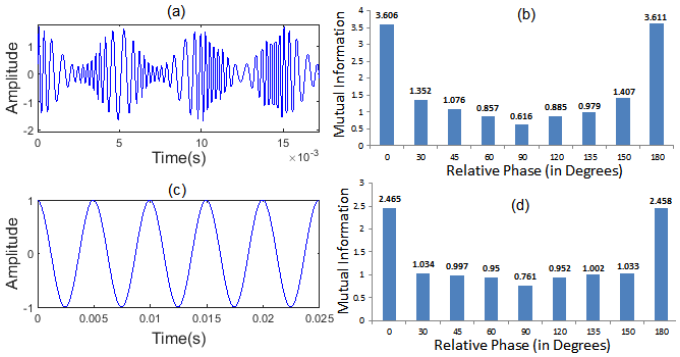


Fig. 1. The effect of varying the relative phase of (a) AM-FM signal, (b) its corresponding MI, (c) cosine signal, and (d) its corresponding MI.

Furthermore, a similar analysis by considering a cosine signal is also shown in Fig. 1 (c). The motivation for this is the significance of the quadrature phase even in the definition of Fourier transform. In particular, the Fourier transform of the signal $x(t)$ is given by [15]:

$$\begin{aligned} \mathcal{F}\{x(t)\} = X(\omega) &= \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt, \\ \therefore X(\omega) &= \int_{-\infty}^{\infty} x(t)\cos(\omega t)dt - j\int_{-\infty}^{\infty} x(t)\sin(\omega t)dt, \\ \therefore X(\omega) &= X_R(\omega) + jX_I(\omega) = |X(\omega)|e^{j\measuredangle X(\omega)}, \\ \therefore \measuredangle X(\omega) &= tan^{-1}\left(\frac{X_I(\omega)}{X_R(\omega)}\right). \end{aligned} \tag{3}$$

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X(\omega)e^{j\omega t}d\omega. \tag{4}$$

Using the fundamental theorem of integral calculus,

$$\begin{aligned} \therefore x(t) &= \frac{1}{2\pi}\lim_{\omega_0\to 0}\sum_{n=-\infty}^{+\infty}\left[X(\omega_n)\omega_0\right]e^{j\omega_n t} \\ x(t) &= \frac{1}{2\pi}\lim_{\omega_0\to 0}\sum_{n=-\infty}^{+\infty}\left(|X(\omega_n)|\omega_0\right)e^{j(\measuredangle(X\omega_n)+\omega_n t)}, \end{aligned} \tag{5}$$

where ω_0 is the step size (width), and $\omega_n = n.\omega_0, n \in Z$. Eq. (3) shows that $\measuredangle X(\omega) = 0$ for $X_I(\omega) = 0$. Furthermore, from eq. (5), it can be concluded that $\measuredangle X(\omega)$ provides the relative phases of the complex exponentials that *make up* the signal $x(t)$ [15]. As observed from Fig. 1 (c) and Fig. 1 (d)), we obtain a minimum MI at $\pi/2$ (i.e., $cos(\omega t)$ and $sin(\omega t)$. Given this MI-based analysis, we have proposed the inclusion of the quadrature-phase component of the detection of spoofed speech, and hence the *Quadrature*-based ESA (i.e., QESA). The proposed QESA uses the *extended* definition of the TEO in order to include the quadrature phase in the signal.

B. Using Extended Definition of TEO

The quadrature phase-shift is incorporated by estimating the analytic signal by computing HT. Given the analytic signal is complex-valued, it shows *causal* behavior in the frequency

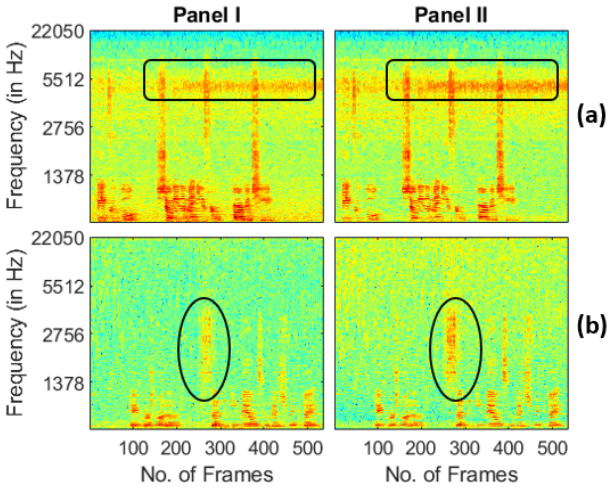


Fig. 2. Panel I and Panel II correspond to the spectrograms of CFCCIF-ESA and CFCCIF-QESA feature sets, respectively, where (a) is the genuine speech signal, and (b) is the spoofed (replay) speech signal.

domain. Further, to estimate the energy of this complex-valued signal, we exploit the extended definition of the TEO on a complex-valued signal $x_z(t)$, i.e., $\psi_c[x_z(t)]$ is given by [16]:

$$\psi_c[x_z(t)] = x_z(t)\dot{x_z}^*(t) - \frac{1}{2}[\ddot{x_z}(t)x_z^*(t) + x_z(t)\ddot{x_z}^*(t)], \quad (6)$$

where $\dot{}$ and $\ddot{}$ denote the differentiation and double differentiation operators, respectively. Given that $x_z(t)$ is complex, the TEO defined in eq. (6) is applied on real and imaginary parts of $x_z(t)$. In particular,

$$\psi_c[x_z(t)] = \dot{x_{z_R}}^2(t) + \dot{x_{z_I}}^2(t) - x_{z_R}(t)\ddot{x}_{z_R}(t) - x_{z_I}(t)\ddot{x}_{z_I}(t). \quad (7)$$

When $x_z(t)$ is complex, eq. (7) can be re-written as [17]:

$$\psi_c[x_z(t)] = \psi[x_{z_R}(t)] + \psi[x_{z_I}(t)]. \quad (8)$$

In this work, using eq. (8), we have extracted TEO-based energy of the analytic signal for improved estimation of energy as a part of ESA, which is discussed in detail in the next sub-Section.

C. CFCCIF-QESA Feature Extraction Procedure

The feature extraction procedure of the proposed CFCCIF-QESA feature set is illustrated in Fig. 3. The computational details of the hair cell representation and Nerve Spike Density (NSD) can be studied from [6]. Here, $x[n]$ is a speech signal and $d[n]$ denotes the subband filtered signal obtained from the cochlear filterbank [18] (as shown in Fig. 3). Then, quadrature phase and subsequently, IF is estimated using QESA as shown in Algorithm 2. It can be observed in Fig. 2 that the proposed CFCCIF-QESA is able to capture more energy as compared to the CFCCIF-ESA feature set, for both the cases of genuine and spoofed signals.

Algorithm 2: The Estimation of IF Using Quadrature-Based Energy Separation Algorithm (QESA)

Input: Input: Subband filter output $d[n]$
Output: Output: IF
$d_z[n] = d[n] + j.HT\{d[n]\}$
/* Using Equation(8) */
$T_r[n] \leftarrow \text{TEO}\{\text{real}(d_z[n])\}$
$T_i[n] \leftarrow \text{TEO}\{\text{imag}(d_z[n])\}$
$\psi\{d_z[n]\} = T_r[n] + T_i[n]$
$IF \leftarrow Cos^{-1}\left[\frac{1-\psi\{d_z[n]-d_z[n-1]\}}{2\psi\{d_z[n]\}}\right]$

III. Experimental Setup

A. Dataset, Classifiers, and Features Used

Dataset: The dataset used is ReMASC, which is particularly developed for anti-spoofing on VAs [19]. The dataset was designed using four acoustic environments, namely, Outdoor Environment (Env A), Indoor Environment 1 (Env B), Indoor Environment 2 (Env C), and Vehicle Environment (Env D). The details of each of the environments can be studied from [19]. The environment-related experiments performed in this work are w.r.t. environment-dependent scenario, where training and testing have common acoustic environments.

GMM Classifier: The GMM classifier is trained using 512 mixture components in this work [20].

CNN Classifier: The architecture of the CNN classifier used comprises 6 convolution blocks and 2 Fully-Connected (FC) layers. Each convolution block is made of a 2-D convolution layer, which is followed by a batch normalization and max-pooling layer. Kernels of sizes 5×5 and 2×2 are taken in the convolution and max-pooling layers, respectively. The padding and stride is fixed as 1. The size of the input to the first convolution layer is 30×400. The batch size and learning rate are taken to be 32 and 0.0001, respectively. The loss function and the activation function used are cross-entropy loss, and Rectified Linear Unit (ReLU), respectively.

Features Used: This work uses the CQCC, CFCCIF-ESA, and the proposed CFCCIF-QESA feature sets. The parameters for CQCC are taken from [21]. Furthermore, we have taken 200 subband filters and 14 dimension for both of the feature vectors, where only static features are taken into consideration.

B. Performance Metrics:

Equal Error Rate (EER) and Score Level Fusion: EER estimated from the Log-Likelihood Ratio (LLR), is the threshold at which the Probability of False Alarm (P_{FA}) is *equal* to the Miss Probability (P_{Miss}). Here, the complementary information of each system is extracted using score-level

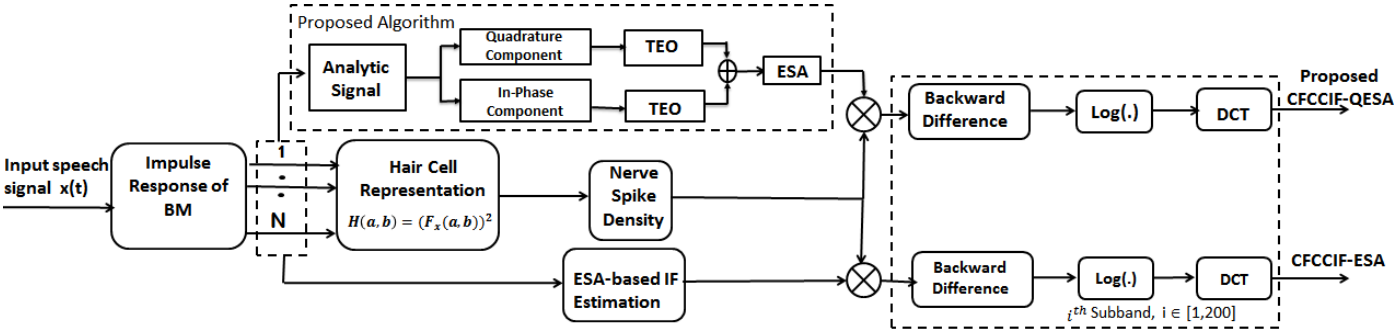


Fig. 3. The working of the proposed CFCCIF-QESA feature set, and the existing CFCCIF-ESA feature set.

fusion of LLR scores, which is computed as

$$LLR_{fused} = \alpha \cdot LLR_{feat1} + (1-\alpha) \cdot LLR_{feat2}, \alpha \in [0,1], \quad (9)$$

where LLR_{feat1} and LLR_{feat2} are LLR scores derived from feature set-1 and feature set-2, respectively.

Confusion Matrix-Based Metrics: Matthews Correlation Coefficient (MCC), F-measure, J-measure, Jaccard Index (JI), and Hamming Loss (HL) are some of the confusion matrix-based measures used to evaluate the performance of CFCCIF-QESA in this work. Out of these, a high value of MCC, F-measure, J-measure, and JI indicates better performance of the SSD system, whereas low value of HL indicates better performance of the SSD system. Details of each of these metrics can be studied from [22], [23], [24], [25].

IV. Experimental Results

Table I shows the environment-wise performance in terms of % EER. The environment-wise analysis shows the enhanced SSD capability of CFCCIF-QESA feature set over CFCCIF-ESA, particularly for the case of Env B, Env C, and Env D. Furthermore, for Env A, the performance of CFCCIF-ESA and CFCCIF-QESA is comparable. Next, Table II shows overall results in terms of %EER on GMM and CNN classifiers. It can be observed that in the case of classification by GMM, CFCCIF-QESA achieves overall EER of 28.71% on the Eval set, thereby achieving relative % decrease of 4.17% w.r.t. CFCCIF-ESA. Table III shows the confusion matrix-based performance measures. It can be observed that the proposed CFCCIF-QESA feature set shows improved performance as compared to the CFCCIF-ESA feature set. However, the CQCC feature set performs the best.

TABLE I
Environment-wise results in % EER using GMM as classifier

Feature Set	Environment-wise % EER			
	Env A	Env B	Env C	Env D
CQCC	22.58	48.48	43.45	14.47
CFCCIF-ESA	29.15	46.33	47.02	15.13
CFCCIF-QESA	29.22	45.53	46.81	12.75
CQCC ⊕ CFCCIF-ESA	**22.56**	46.33	43.45	12.42
CQCC ⊕ CFCCIF-QESA	22.58	**45.53**	**43.45**	**11.55**
⊕ indicates score-level fusion as per eq. (9)				

TABLE II
Results in %EER using GMM and CNN as classifiers

Feature Set	GMM		CNN	
	Dev	Eval	Dev	Eval
CQCC	19.94	22.56	15.36	25.33
CFCCIF-ESA	21.64	29.95	16.23	28.03
CFCCIF-QESA	26.93	28.71	15.47	29.89
CQCC⊕CFCCIF-ESA	**18.14**	22.56	11.20	**23.55**
CQCC⊕CFCCIF-QESA	18.69	**22.02**	**10.63**	23.84

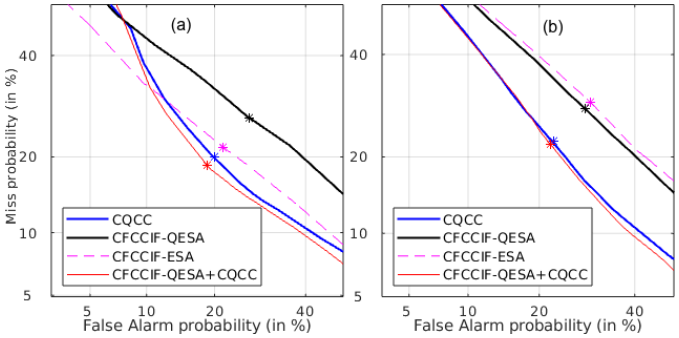


Fig. 4. DET plots on (a) Dev set, and (b) Eval set.

TABLE III
Confusion Matrix-Based Performance Measures

	MCC	F1	J	JI	HL
CQCC	0.5735	0.8373	0.6006	0.64	0.34
CFCCIF-ESA	0.3543	0.8097	0.3638	0.71	0.28
CFCCIF-QESA	**0.3813**	**0.8212**	**0.4096**	**0.65**	**0.34**

A. Model-Level Comparison of CFCCIF-ESA vs. CFCCIF-QESA

KLD and Jensen–Shannon Divergence (JSD) have been used between the genuine and spoofed classes [26], to estimate the discriminative ability of our trained model. KLD and JSD are model-level information-theoretic measures to explore the discriminative abilities of the CFCCIF-ESA *vs.* CFCCIF-QESA feature sets. KLD is a measure to quantify how much two probability distribution functions (PDFs) differ from each other. It is mathematically defined as [27] If u and v are two

discrete PDFs, then KLD is :

$$KLD(u||v) = -\int_x u(x) \ln \left\{ \frac{v(x)}{u(x)} \right\} dx. \quad (10)$$

As it can be deduced from eq. (10), $KLD(u||v) \neq KLD(v||u)$ and thus, it is an asymmetric measure, and therefore, to deal with the asymmetry, a symmetric measure JSD is used, which is computed as [28]:

$$JSD(u||v) = \frac{1}{2}KLD(u||m) + \frac{1}{2}KLD(v||m), \quad (11)$$

where m is $\frac{1}{2}(a+b)$. A relatively high value of KLD and JSD indicates better discriminative ability of the model. Fig. 5 (a) shows the KLD between genuine and spoof GMMs, whereas Fig. 5 (b) shows the KLD between spoof and genuine GMMs, for both the feature sets. Here, it can be seen that the KLD between the GMMs of CFCCIF-QESA is *more* as compared to the KLD w.r.t. CFCCIF-ESA and hence, re-confirming the better discriminative ability of CFCCIF-QESA w.r.t. CFCCIF-ESA. Similar trend w.r.t. JSD can be seen in Fig. 5 (c). Furthermore, the differences in JSD, are shown in Fig. 5 (d), where the highest difference in JSD is observed at 64 number of GMM mixtures.

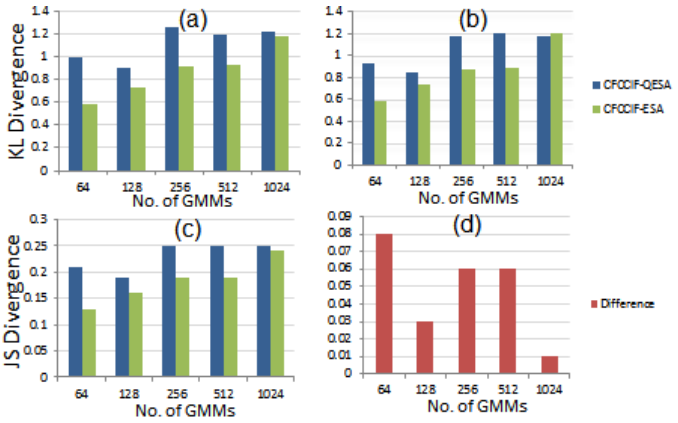


Fig. 5. Model-level measures: (a) $KLD(gen||sp)$, (b) $KLD(sp||gen)$, (c) $JSD(gen||sp)$, and (d) difference of JSD for varying number of GMM mixtures.

B. Analysis of Latency Period

In this study, we investigate the latency period for CFCCIF-QESA *w.r.t* CFCCIF-ESA feature set, which indicates how fast the SSD system is w.r.t. deployment in real-world applications. Latency is the performance evaluation in terms of %EER *w.r.t* different durations of speech segment in an utterance. Fig. 6 shows the latency analysis for the CFCCIF-ESA and the CFCCIF-QESA feature sets. The utterance duration ranges from 10 ms to 60 ms, with an interval of 10 ms. From Fig. 6, it is observed that CFCCIF-QESA shows relatively smaller latency as compared to the CFCCIF-ESA, i.e., CFCCIF-QESA reaches better performance than the CFCCIF-ESA even for a short duration (i.e., 10 ms) of speech. The low latency achieved by CFCCIF-QESA indicates that as compared to CFCCIF-ESA, CFCCIF-QESA has the ability to faster classification and thus, better suitability for practical system deployment.

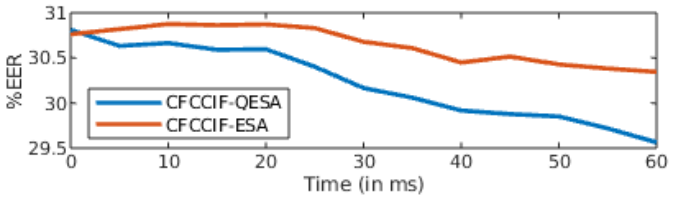


Fig. 6. Latency curves for CFCCIF-ESA *vs.* CFCCIF-QESA.

V. Summary and Conclusions

This paper investigated the importance of incorporating the quadrature phase for replay SSD w.r.t. VAs. An MI-based analysis is shown to highlight the importance of taking the quadrature phase. Furthermore, the extended definition of TEO is used to include the quadrature component in the signal. Given that the IF computation by the existing ESA method merits good time resolution of the TEO, the proposed QESA includes excellent time resolution, as well as the quadrature phase. With the similar approach as CFCCIF, except for the method of estimation of IF, a novel feature CFCCIF-QESA is designed. In this respect, the experiments are performed on the ReMASC dataset. Through our experiments, an improved performance is observed due to the inclusion of the quadrature phase using CFCCIF-QESA as compared to CFCCIF-ESA alone.

References

[1] X. Lei, G.-H. Tu, A. X. Liu, K. Ali, C.-Y. Li, and T. Xie, "The insecurity of home digital voice assistants–Amazon Alexa as a case study," *arXiv preprint arXiv:1712.03327*, 2017, {Last Accessed: 28 March, 2022}.

[2] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 30^{th} Oct-3^{rd} Nov, 2017, Dallas, USA, pp. 103–117.

[3] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Security and Privacy Workshops (SPW)*, San Francisco, CA, USA, 21-23 May, 2018, pp. 1–7.

[4] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proceedings of the* 4^{th} *ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, Arizona, USA, November, 2014, pp. 63–74.

[5] Y. Gong and C. Poellabauer, "Protecting voice controlled systems using sound source identification based on acoustic cues," in 27^{th} *International Conference on Computer Communication and Networks (ICCCN)*, July 30 -August 2, 2018, Hangzhou, China, pp. 1–9.

[6] Ankur T. Patil, et al., "Effectiveness of energy separation-based instantaneous frequency estimation for cochlear cepstral features for synthetic and voice-converted spoofed speech detection," *Computer Speech & Language*, vol. 72, p. 101301, 2022.

[7] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[8] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, USA, April 1990, pp. 381–384.

[9] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 1^{st} edition, Pearson Education India, 2015.

[10] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.
[11] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *INTERSPEECH, Portland, USA*, 9-13 September 2012.
[12] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal (BSTJ)*, vol. 27, no. 3, pp. 379–423, 1948.
[13] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.
[14] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.
[15] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*. Prentice Hall, 1997.
[16] R. Hamila, J. Astola, F. A. Cheikh, M. Gabbouj, and M. Renfors, "Teager energy and the ambiguity function," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 260–262, 1999.
[17] P. Maragos and A. C. Bovik, "Image demodulation using multidimensional energy separation," *Journal of the Optical Society of America A (JOSA)*, vol. 12, no. 9, pp. 1867–1876, 1995.
[18] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.
[19] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic replay attack corpus for voice controlled systems," *INTERSPEECH, Graz, Austria*, pp. 2355–2359, 15-19 September 2019.
[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[21] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant-Q cepstral coefficients." in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 21 June to 24 June, 2016, pp. 283–290.
[22] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.
[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[24] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
[25] L. Hamers *et al.*, "Similarity measures in scientometric research: The jaccard index *vs.* salton's cosine formula." *Information Processing and Management*, vol. 25, no. 3, pp. 315–18, 1989.
[26] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech," in *INTERSPEECH, San Francisco, CA, USA*, 8-12,September 2016, pp. 1710–1714.
[27] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2017.
[28] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Exploring Residual Cepstral Features for Spoken Language Identification

Baveet Singh Hora*, Krishna Parmar*, Shrey Machhar*, Hemant A. Patil*
Kiran Praveen† and Balaji Radhakrishnan †
* Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India
E-mail: {201901256, 201901155, 201901151, hemant_patil}@daiict.ac.in
† Samsung Research Institute, Bangalore, India
E-mail: {k.praveen.t, balaji.r}@samsung.com

***Abstract*—In this paper, we introduce a Spoken Language Identification (SLI) system, which uses Linear Prediction (LP) residual, decomposed with linear filterbank and followed by log and Discrete Cosine Transform (DCT). This feature set is called Linear Frequency Residual Cepstral Coefficients (LFRCC). Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) were also extracted and kept as baseline features. The Attentive Time Delay Neural Network (Attentive - TDNN) was employed to evaluate the efficiency, relevancy, and sufficiency of said LFRCC features. For experiments, a computationally challenging VoxLingua107 dataset was used. From experiments, we achieved 9.22%, 7.8%, and 6.86% EER for MFCC, LFCC, and LFRCC, respectively. Significant improvements in accuracies and equal error rate (EER) were observed when comparing the performance of MFCC and LFCC with data fusion strategies. The feature-level fusion of MFCC and LFRCC resulted in a notable enhancement of 3.16%, 2.12%, and 1.16%, accuracies compared to other combinations. Moreover, each combination of feature sets in score-level fusion has outperformed LFRCC and baselines. This paper provides a significant study on how prosodic information provided by LP residual contributes to language-specific information. We present the pitch strength analysis to understand better the contribution of LP residual in capturing language-specific information.**

I. Introduction

The process of extracting information about one's spoken language is called Spoken Language Identification (SLI). At present, Human-Computer Interaction (HCI) applications are getting wide adoption and transcription. The services, such as Voice Assistance and Text Transcript from speech, are getting more popular, and many efforts are being made to be accessible to people of different languages. In such systems, SLI is sometimes used as a fundamental and primitive block, on top of which the architecture of such services is built [1]. *Language Line Interpreter* is one example of such service first provided by AT&T [2]. As technology reaches more people, it needs to be inclusive and robust; hence, developing such an SLI system is essential.

India is known to be multilingual, with different languages spoken in different parts of the country. In the past decade, researchers have shown an increased interest in developing SLI systems in the context of Indian languages [3]. Many recent studies use features that provide information about the vocal tract system as acoustics of speech that can be extracted using Mel Frequency Cepstral Coefficients (MFCC) [4], [5]. These acoustic features contain different phonemes for each language and, thus, help distinguish various languages. However, MFCC features are notoriously affected by noisy environments and changes in speaking style. The prosodic features, however, reveal the excitation source information such as Glottal Closing Instants (GCIs), stress, and intonations unaffected by noise or change in speaking style [6]. The direct use of Linear Prediction (LP) residual for SLI has been explored in [6]. In our paper, we parameterize the LP residual on a linear filterbank to extract the cepstral features, namely, LFRCC features. In addition, we also explore how different order of LP (in LFRCC) affects the performance of the SLI system.

Many researchers in the past have used methods that utilize acoustic features [7], [8]. They utilize statistical methods, such as Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), i-Vector Universal Background Models (UBM), and Support Vector Machines (SVM). Further advancements in the field were made using deep learning architectures, such as Convolutional Neural Networks (CNN), due to their ability to capture local contexts [7]. Our study utilizes the Time Delay Neural Networks, which process sequential data better than CNN [9]; the paper introduced the TDNN, which is adopted in speech processing for their ability to capture the long temporal contexts [10], [11].

A. Related Work

The earlier studies reported in [12], [13] have applied LFRCC features to detect spoofed audio in the Automatic Speaker Verification system. Historically, LP residual represents the excitation source information of the speech, which contains the speaker's information and is helpful to differentiate speakers. In our paper, we exploit the LFRCC for the SLI. The LFRCC helps better preserve speech prosody, which leads to language-specific information. LFRCC contains pitch, rhythm, stress, and intonation, which are fundamental components of speech. Several methods have been explored in this area, such as [6], [8]. The method is computationally expensive in [6] as it utilizes an ASR system. In [8], they extract the pitch (fundamental frequency F_o) and energy of speech using the toolkit followed by mean normalization. In

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

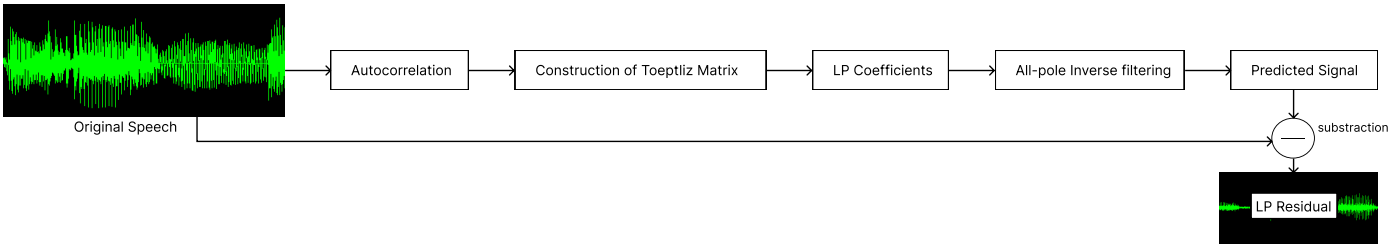


Fig. 1. Functional block diagram of extraction of LP residual.

our work, we gain such prosodic information from the LP residual and then parameterize it with a linear filterbank. The paper contributes the following.

- We study how the residual of the speech contributes to language-specific information, which helps in SLI.
- We show the visual inspection between the actual speech and residual speech
- For better understanding, we calculate the pitch strength and analyze the contribution of the LP residual in the context of syllable-timed vs. stress-timed languages
- As the SLI task should be robust against smaller speech segments, we show our model's latency period for the proposed LFRCC features.
- We also improved the results of the experiments with the help of two data-fusion strategies, namely, score-level and feature-level fusion.

The rest of the paper is organized as follows: Section II presents details of LP residual. Section III analyzes the pitch strength. Section IV shows the experimental setup used for the study. Section V presents the experimental results. Finally, section VI concludes the paper along with the potential future research directions.

II. Linear Prediction (LP) Residual

Linear prediction is a commonly employed and effective method for analyzing and predicting the behavior of dynamic systems. It involves using a specific number of previous samples to estimate the future sample, known as the prediction order, also referred to as the model order or LP order. Essentially, by setting the prediction order to p, the linear prediction filter utilizes p past samples to predict the next sample. The n^{th} sample of $X(n)$ can be predicted by a weighted summation of past samples, where a_p is the Linear Prediction Coefficient for k^{th} past sample [14]

$$\hat{X}(n) = -\sum_{k=1}^{p} a_k x(n-k) \ , \tag{1}$$

Calculating coefficients is an essential step in the linear prediction process required for the predicted signal. This involves several key procedures. Firstly, the autocorrelation between the preceding 'p' samples is computed. Subsequently, this auto-correlation information is utilized to generate a Toeplitz matrix. The linear predicted signal is ultimately obtained by applying the all-pole inverse filtering technique.

The error between the actual value $x(n)$ and the predicted signal $\tilde{x}(n)$ is given by,

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^{p} a_k x(n-k) \ , \tag{2}$$

where $e(n)$ represents the residual component or linear prediction residual (LP residual). The LFRCC features, as discussed in the subsequent Section, are specifically designed to capture the spectral envelope of the excitation source. This characteristic has been widely recognized as a reliable indicator for various tasks, such as speaker identification and language recognition [6].

To ensure precise predictions, selecting an optimal prediction order is of utmost importance. Small orders may lead to inaccurate predictions as the model fails to capture essential signal features w.r.t dependencies in the sequence of samples of speech signals. Conversely, an excessively high LP order can result in overfitting of the model, resulting in a significant loss of generalized information and increased prediction errors.

The LP residual serves as a critical indicator of the performance of the LP model. A high energy value in the LP residual suggests that the model cannot accurately capture the fundamental characteristics of the signal, and a low energy value indicates that the linear prediction model has successfully captured the key features of the signal. This difference is clearly observed in figure 2, between (c) and (d). In many areas in the graph, high-energy horizontal striations are not present in the residual waveform. However, they are present in the original speech signal.

The LP residual comprises the signal portion not accounted for by the linear prediction model. Figure 1 shows the functional block diagram for extraction of LP residual.

A. Spectographic Analysis of LP Residual

Figure 2 shows the speech, LP residual, and their corresponding spectrogram.
It can be observed that LP residual discards the vocal tract shape as it flattens the spectral envelope. On the other hand, it also enhances the vocal fold vibration patterns atleast implicitly. These vocal fold vibrations contain several prosodic features, such as fundamental frequency (F_0), stress, and

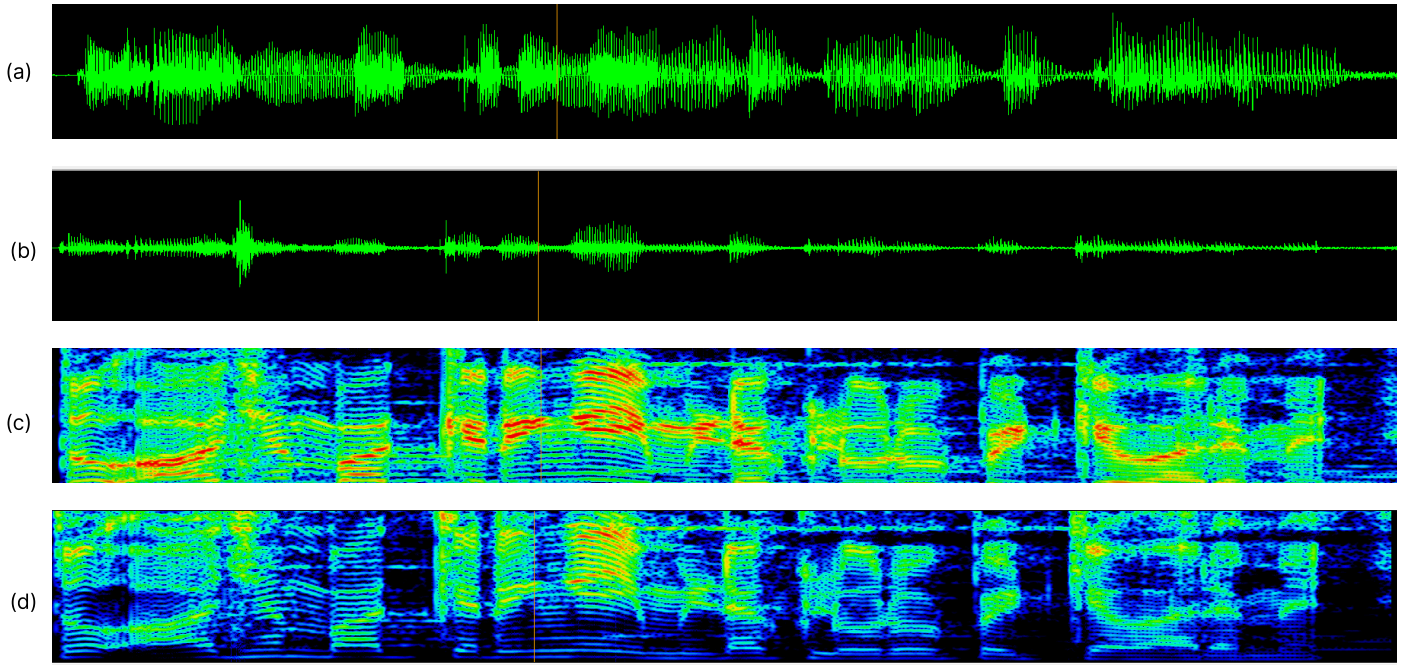


Fig. 2. (a) Speech waveform, (b) Corresponding LP residual speech for LP order 16, Spectrogram of (c) speech waveform shown in Fig 2(a) and (d) spectrogram of LP residual shown in Fig 2(b).

rhythm. These features depend on language structure and contain cues that can provide language-specific discriminatory information.

III. Analysis of Pitch Strength

We can define Pitch Strength as the distribution and preservation of pitch over the duration of the speech. The spoken languages can be divided into three groups: syllable-timed, stress-timed, and mora-timed [3], [5]. Languages such as English and Panjabi can be put into the stress-timed language category, where syllables have different lengths and maintain a rhythm. Indic languages such as Hindi, Bengali, and Tamil are said to be syllable-timed languages where the length of syllables is near-equal. This language structure depends on the prosodic information derived from the excitation source. To understand the prominence of the LP residual, we calculate the pitch strength of a syllable-timed and a stress-timed language using the auto-correlation method for LP Order 12 and LP Order 16. We can see changes in pitch strength for such in Figure 3. We can see that LP order 16 better preserves the speech structure, which significantly contributes to the language-specific information as prosodic information is mainly dependent on language [6]. We have calculated pitch strength explicitly for LP order 12 and LP order 16 because of the results (refer section V) for fine-tuning the prediction order p. In experiments, LP order 16 gives the highest accuracy, whereas LP order 12 gives the lowest.

IV. Experimental Setup

A. Dataset Used

Our study used the state-of-the-art, computationally challenging, and statistically prominent VoxLingua107 database [15]. The corpus is explicitly introduced for Spoken Language Recognition. The authors Valk J *et al.* use semi-random search phrases and scraped audio files from YouTube Videos, which were afterward validated using crowd-sourcing. The resulting database has an average of 62 hours of speech per language, totaling 107 languages. A total of 6628 hours of speech are provided in the database.

From the database, we use five language classes: English, Hindi, Bengali, Tamil, and Panjabi. Initially, the speech samples for each language class were divided by the groups of duration (i.e., sample duration less than 5 seconds, sample duration between 5 seconds and 10 seconds, sample duration between 10 seconds and 15 seconds, sample duration more than 15 seconds). Afterward, the same number of samples were taken from each duration class for each language. In one of those duration-wise groups for English, there were fewer speech samples ($<$ 1300), resulting in a total of 5132 speech samples. The detailed specifications of the database are provided in Table I. The resulting database was divided into 80% for training, 10% for validation (val set), and 10% for testing (test set).

B. Other Feature Sets Used

In our study, we have compared the performance of the LFRCC features with MFCC and LFCC features. The parametric details of these features can be found in Table II. Each

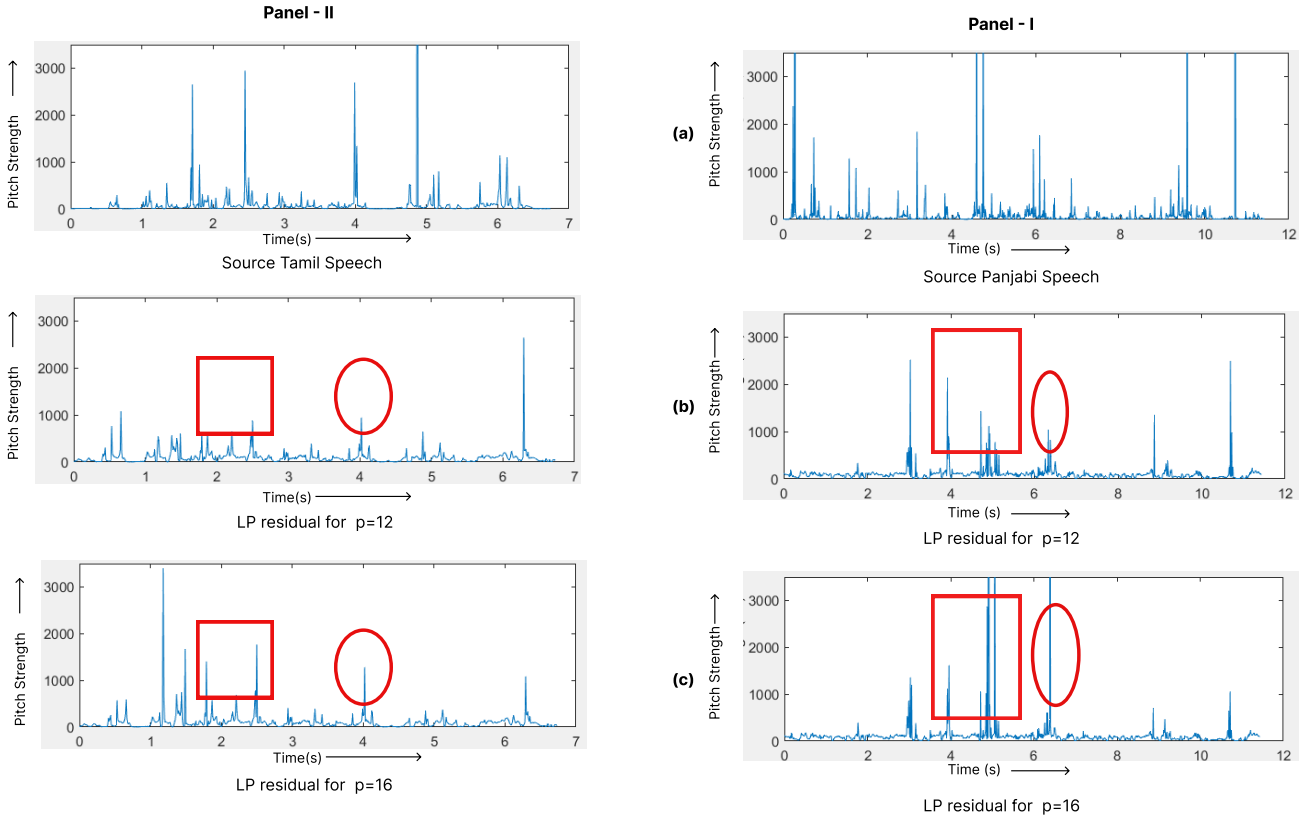


Fig. 3. Panel-I (a) shows the pitch strength of the stress-timed language, such as Panjabi; Panel-II (a) shows the pitch strength of syllable-timed languages, such as Tamil; Panel-I (b) and Panel-II (b) shows the pitch strength of the LP residual for p=12; Panel-I (c) and Panel-II (c) shows pitch strength of the LP residual for p=16

TABLE I
SPECIFICATIONS OF DATABASE FOR EACH LANGUAGE CLASS TAKEN

Language Class	Number of Samples	Total Length (in hrs)	Average sample length (in seconds)
English	5132	15.08	10.30
Hindi	5200	15.23	10.40
Tamil	5200	15.26	10.26
Bengali	5200	15.19	10.24
Panjabi	5200	15.28	10.36
Total	25932	76.04	10.312

TABLE II
DETAILS OF PARAMETERS OF FEATURE SETS

Parameters	MFCC	LFCC	LFRCC
Frequency Scale	Mel	Linear	Linear
Subband Filters	40	40	40
Window Length	30ms	30ms	30 ms
Hop Length	15ms	15ms	15ms
Dimension	13	13	13

feature vector was appended with 13 Δ and 13 $\Delta\Delta$ features to capture the temporal context. The extraction pipeline for each of these is provided in Figure 4. To gain LFRCC, first, the pre-emphasis is performed on the raw speech signal. Afterward, we obtain its LP residual using LP residual analysis. Then, the obtained speech is framed and windowed. Fast Fourier Transform is applied to gain their frequency representation on the framed and windowed speech. The resulting speech is afterward passed through a Linear filterbank of 40 subband filters. In the end, a log is applied to convert it to human perception. Discrete Cosine Transform decorrelates the filter bank coefficients and produces a set of cepstral coefficients. In MFCC, we do not pass the LP residual but instead pass the pre-emphasized speech and use the Mel Filterbank. In LFCC, similar to MFCC, we use the linear filterbank.

C. Model Architecture

The state-of-the-art TDNN [9] architecture is chosen for experimentation. The authors are inspired by the work [16]

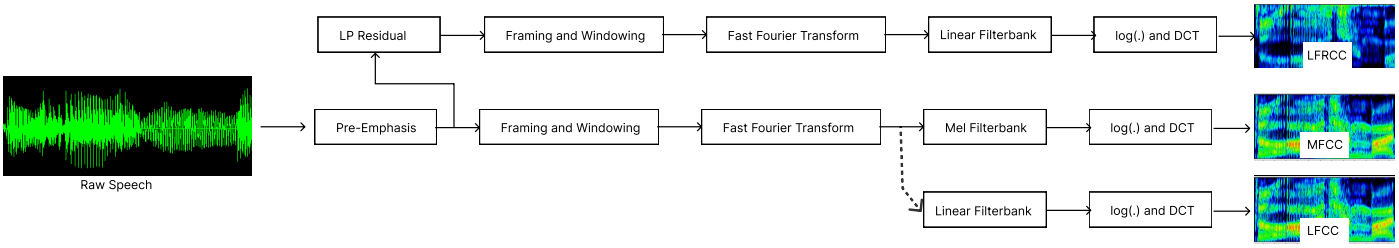


Fig. 4. Functional block diagram of extraction pipelines of various feature sets

and have performed the Attentive Statistics pooling in the TDNN network. The architecture closely follows the model described in [16]. The motivation behind using TDNN for our SLID is its ability to process sequential data and capture long temporal contexts [9]. The details of the TDNN layer used can be found in Table IV. We used TDNN layers and the classic attention and statistics pooling layers. The details of the classic attention module can be found in Table III. Table V provides an overview of the architecture, showcasing all the layers involved.

TABLE III
CLASSIC ATTENTION MODULE

Module Type	Classic Attention
Input Shape	(batch, seq_len, input_dim)
Output Shape	(batch, seq_len, input_dim)
Parameters	input_dim, embed_dim, attn_dropout
Linear Projection	Linear projection with kernel size (input_dim x embed_dim)
Activation	Tanh
Attention	Softmax along the sequence length dimension
	Output Shape: (batch, seq_len, input_dim)
Total Parameters: $(input_dim \times embed_dim) + embed_dim$	

TABLE IV
TIME DELAY NEURAL NETWORK (TDNN) MODULE

Module Type	TDNN
Input Shape	(batch, seq_len, input_dim)
Output Shape	(batch, new_seq_len, output_dim)
Parameters	context_size, stride, dilation, batch_norm, dropout_p
Linear Projection	Linear projection with kernel size (context_size x input_dim)
Activation	ReLU
Normalization	BatchNorm1d (if batch_norm=True)
Regularization	Dropout (if dropout_p > 0)
Forward pass	
1. Unfold input	Unfold input into smaller temporal contexts with kernel size (context_size x input_dim) **Output Shape:** (batch, output_dim*context_size, new_seq_len)
2. Linear projection	Compute output of shape (batch, new_seq_len, output_dim)
3. Activation	Apply ReLU activation
4. Regularization	Apply Dropout with probability dropout_p (if dropout_p > 0)
5. Normalization	Apply BatchNorm1d along the output dimension **Output Shape:** (batch, new_seq_len, output_dim)
Total Parameters	$(context_size * input_dim * output_dim) + output_dim$

1) Attentive Statistics Pooling: Statistics pooling is used in [17]. It calculates mean μ and standard deviation (second order statistics) σ over frames f_t *(t = 1,2,..T)*.

$$\boldsymbol{\mu} = \frac{1}{T}\sum_{t}^{T} \boldsymbol{f}_t \quad , \tag{3}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{T}\sum_{t}^{T} \boldsymbol{f}_t \odot \boldsymbol{f}_t - \boldsymbol{\mu} \odot \boldsymbol{\mu}} \quad , \tag{4}$$

TABLE V
TDNN NETWORK ARCHITECTURE

Layer	Kernel size	Filter size	Stride	Output shapes
TDNN	5	512	1	(batch_size, 512, 1)
TDNN	5	512	1	(batch_size, 512, 1)
TDNN	7	512	1	(batch_size, 512, 1)
TDNN	1	512	1	(batch_size, 512, 1)
TDNN	1	512	1	(batch_size, 512, 1)
Classic Attention	-	512	-	(batch_size, 512, 1)
Linear	1024	512	1	(batch_size, 512)
Attention statistics pooling				
Linear	512	512	1	(batch_size, 512)
Output	-	5	-	(batch_size, number of classes)

where $\odot$ is Hadamard product. Equations [3] and [4] are used to calculate the mean and standard deviation. This mean summarizes frame-level features and the standard deviation, which we believe encompasses various temporal variations in longer contexts associated with the language's characteristics other than the mean.

In language recognition tasks, certain frame-level features within an utterance may have significant prosodic and distinctive than others. To address this, studies such as [11] have incorporated attention mechanisms in their system to get weighted frame-level utterances. [18] applies this attention mechanism on TDNN for SLI.

An attention model Table III works with a Deep neural network to calculate a scalar score P_t for each frame-level feature

$$p_t = \boldsymbol{v}^T f\left(\boldsymbol{W}\boldsymbol{f}_t + \boldsymbol{b}\right) + k \quad , \tag{5}$$

where *f(.)* is tanh activation function. The score is further normalized using the *Softmax* function to obtain the normalized score N_t.

$$N_t = \frac{\exp\left(e_t\right)}{\sum_{\tau}^{T} \exp\left(e_\tau\right)} \, . \tag{6}$$

This score is then used as a weight in the pooling layer to calculate weighted mean vector $\tilde{\boldsymbol{\mu}}$.

$$\tilde{\boldsymbol{\mu}} = \sum_{t}^{T} N_t \boldsymbol{f}_t \, . \tag{7}$$

Instead of just mean, we have their weighted score, and we can use this to get the weighted standard deviation, $\tilde{\boldsymbol{\sigma}}$.

$$\tilde{\boldsymbol{\sigma}} = \sqrt{\sum_{t}^{T} N_t \boldsymbol{f}_t \odot \boldsymbol{f}_t - \tilde{\boldsymbol{\mu}} \odot \tilde{\boldsymbol{\mu}}} \quad , \tag{8}$$

where N_t is calculated by (6). By utilizing both statistics pooling and attention mechanisms, the weighted standard deviation method is thought to capture the long-term variations in feature representation while selecting frames based on their importance for language recognition.

V. Experimental Results

A. Effect of LP Order

This section presents the experimental results on classes described in Table I. According to the findings reported in [12], the study suggests an optimal range for LP orders is between 8 and 20. Experiments were conducted by generating the features for different LP orders and training them on the TDNN network. The outcomes of this experiment are depicted in Figure 5, which specifically showcases the results within the optimal LP order range; the results presented are on the test set. Upon analysis, it can be observed that the highest accuracy is achieved at LP order 16, and the lowest is achieved at LP order 12. This suggests that the LP order of 16 likely provides a more accurate representation than the other orders, leading to its superior performance. Based on the observed results, LP order 16 was taken for all subsequent experiments.

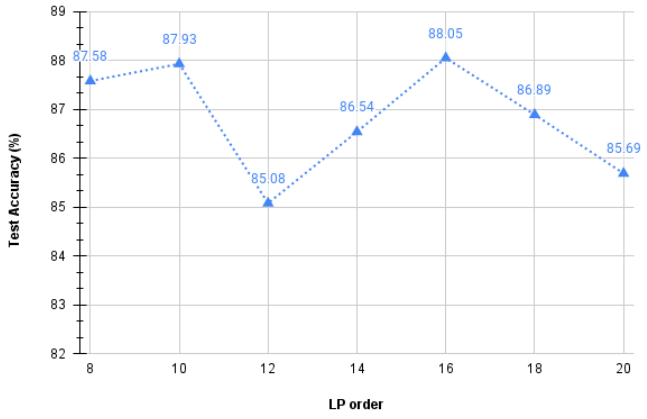


Fig. 5. Effect of LP order on SLI

B. Comparison Against Baselines

Further in the experiments, optimal LP order LFRCC features were compared against baselines MFCC and LFCC, trained on the Attentive-TDNN architecture described in Section IV. Table VI provides the results of LFRCC with the baselines. As LP residual in LFRCC preserves the speech harmonics better than any other features, it contributes to the language-specific information, which helps the model to learn discriminatory information between languages and, hence, outperforms the baselines.

C. Results of Score-Level Fusion

Score level fusion combines the scores of multiple features or classifiers to improve overall performance. We estimate log-likelihood scores of the features and fuse them by:

TABLE VI
Results in % of LFRCC against baselines for validation

Feature Set	Accuracy in %	EER in %
MFCC	85.23	9.22
LFCC	87.50	7.8
LFRCC	**89.01**	**6.86**

$$LLK_{\text{fused}} = \alpha LLK_{\text{feature 1}} + (1-\alpha) LLK_{\text{feature 2}} \quad (9)$$

where $LLK_{\text{feature 1}}$ is a log-likelihood score of MFCC or LFCC, whereas $LLK_{\text{feature 2}}$ represent the score of the LFRCC feature set. The fusion parameter (α) lies between $0 < \alpha < 1$ to decide the relative weight of scores. Equal Error Rate (EER) gauges the system's performance.

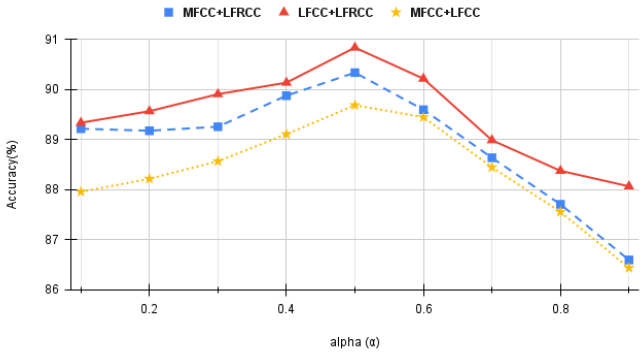


Fig. 6. Accuracy of score-level fusion in % for different α values

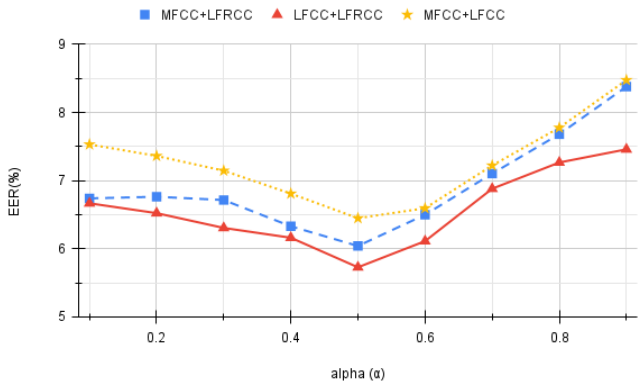


Fig. 7. EER of Score-Level Fusion in % for different α values

From Figure 6 and 7, we can see that the highest accuracy and lowest EER for each fusion have been achieved at 0.5 value.

D. Results of Feature-Level Fusion

In feature-level fusion, we fuse two different feature sets by stacking their coefficients. For this fusion, their time frames must match. Figure 8 shows the results of the feature-level

fusion over the Attentive-TDNN. We can see that the lowest EER has been achieved for the feature-level fusion of LFRCC LP order 16 and MFCC. As we know, MFCC provides the vocal-track information, and LFRCC provides the source information. Their fusion gives us information about the entire system from which speech is generated. Additionally, both of those contribute to discriminatory and essential language-specific information. Hence, it gives the best results.

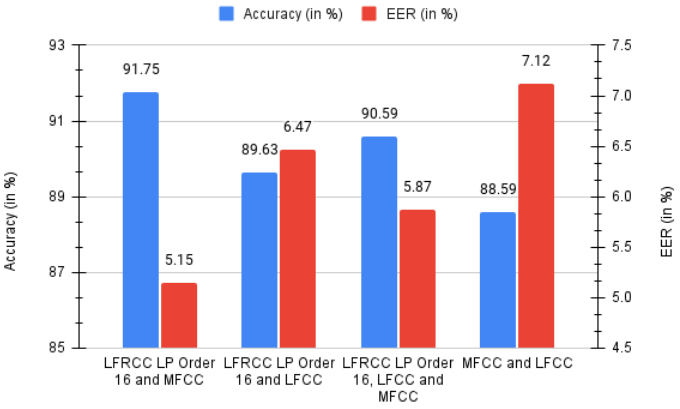


Fig. 8. Results of feature level Fusion

E. Analysis of Latency Period

In this study, computational and performance aspects of the LFRCC features against the baselines were also inspected using the latency period as shown in Figure 9. The latency period is calculated by % EER of classification for varying speech duration. Test segments of 1.5 seconds to 19.98 seconds were considered.

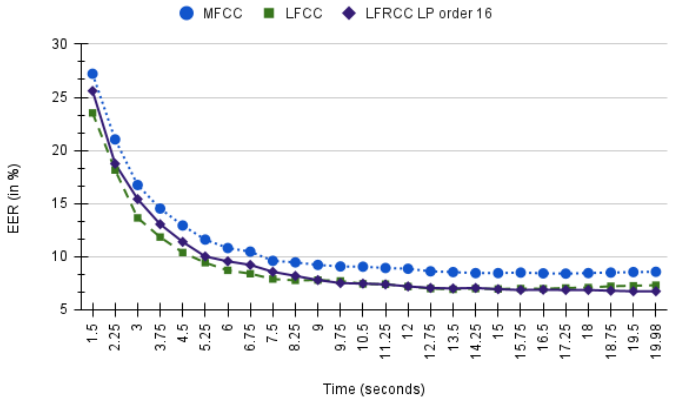


Fig. 9. Analysis of latency period for various feature sets

It can be observed from Fig 9 that the LFRCC features give the least % EER in classification. However, LFCC gives near-equal EER % classification as LFRCC for the speech duration < 9 seconds. In this context, the efficiency and credibility of the LFRCC features for practical use are clearly visible.

VI. Summary and Conclusion

The study presented an in-depth analysis of LFRCC features on SLI. To analyze the relevancy of LFRCC features, pitch strength analysis is introduced. The experiments are performed on the state-of-the-art VoxLingua107 database. To evaluate the performance, the attention pooling-based state-of-the-art TDNN architecture is employed. Additionally, MFCC and LFCC features were kept as baselines for a fair comparison. From the results, LFRCC features are observed to give better results than MFCC and LFCC. Additionally, score- and feature-level fusion improves the system's performance. The fusion of LFRCC with MFCC and LFCC has proven effective. Further, we also present the latency period analysis details on MFCC, LFCC, and LFRCC, which shows the potential of the LFRCC features for practical use. Future work will be directed towards including more language classes for the SLI. Additionally, in the future, the performance of the proposed LFRCC features can be evaluated in cross-database scenarios.

VII. Acknowledgement

The authors would like to thank the authorities of DA-IICT for providing the resources and support to conduct these research experiments. The authors would also like to thank Mr. Abhishek Pandey from Samsung Research Institute, Bangalore.

References

[1] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 33–41, 1994.

[2] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[3] A. Bakshi and S. Kumar Kopparapu, "Spoken indian language identification: A review of features and databases," *Sādhanā*, vol. 43, pp. 1–14, 2018.

[4] D. Deshwal, P. Sangwan, and D. Kumar, "Feature extraction methods in language identification: A survey," *Wireless Personal Communications*, vol. 107, no. 4, pp. 2071–2103, 2019.

[5] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[6] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782–796, 2008.

[7] P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, and A. Yoneyama, "Comparative study on spoken language identification based on deep learning," in *26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 2265–2269.

[8] D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, "Ivector-based prosodic system for language identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4861–4864.

[9] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 3214–3218.

[10] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1517–1521.

[11] F. A. R. R. Chowdhury, Q. Wang, I. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5359–5363.

[12] H. Tak and H. A. Patil, "Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 726–730.

[13] P. Gupta and H. A. Patil, "Linear frequency residual cepstral features for replay spoof detection on asvspoof 2019," in *30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022, pp. 349–353.

[14] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, 1975.

[15] J. Valk and T. Alumäe, "VoxLingua107: A dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[16] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 2252–2256.

[17] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 999–1003.

[18] T. Mandava and A. K. Vuppala, "Attention based residual-time delay neural network for indian language identification," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Noida, India, 2019, pp. 1–5.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Consideration of Varying Training Lengths for Short-Duration Speaker Verification

WooSeok Ko*, Seyun Um*, Zhenyu Piao* and Hong-goo Kang*
* Yonsei University, South Korea
E-mail: {wsko,syum,jkyung}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—We present an efficient training scheme for speaker verification (SV) networks in short-duration speech input scenarios. We analyze the effects of varying training lengths on SV performance, with a particular focus on short utterances. Despite the high demand for short-duration SV in real-world applications, state-of-the-art SV systems have primarily been evaluated on long utterances, and little research has been conducted on short-duration SV. By considering the innate characteristics of SV architectures and the performance discrepancies associated with varying training data lengths, we propose a training scheme that accounts for varying length conditions. We categorize speaker characteristics as coarse-grained and fine-grained features and demonstrate that training models to learn both features can result in length-robust speaker embeddings. Our proposed training scheme improves model performance by 28.7% and 37.9% in terms of equal error rate on short-duration speech scenarios compared to baseline models.**

***Index Terms*—Speaker verification, short utterance, training scheme**

I. INTRODUCTION

Speaker verification (SV) is the task of determining whether a speaker is enrolled in a given scenario based on a speech signal. It has applications in various fields, including security and personal authentication. With the development of neural networks, SV models have achieved excellent performance, making them applicable to many real-world scenarios [1]–[3].

Recently, there has been a rapid increase in demand for user-centered voice interface-based systems, such as wake-up words or short voice commands in AI speakers, mobile systems, or automatic driving systems [4], [5]. As a result, research on short-duration SV has gained attention, including the short-duration speaker verification (SDSV) challenge [6], [7] and various studies focusing on short-duration SV scenarios.

However, existing research on short-duration speaker verification primarily evaluates models on test utterances up to 8-second long, which is not short enough compared to the dominant speech datasets utilized for keyword spotting, which are approximately 1-second or shorter [8], [9]. Moreover, previous studies [10]–[13] that analyze short-duration speaker verification often rely on the revision of model architecture, complex fine-tuning processes or model fusion, leading to an increase in model parameters and computational complexity.

To address this issue, we investigate the inherent characteristics of short utterances compared to long utterances and define coarse- and fine-grained features and their relationship with the length of the utterance when extracting speaker embeddings. Accordingly, we propose an effective hard-condition training (HCT) scheme that does not require additional modules or model modification, allowing SV models to capture various speaker features on varying lengths. Thus, our proposed training scheme enables SV models to extract robust embeddings from utterances of varying lengths, particularly short utterances. Additionally, based on a deep analysis of two popular SV models, ECAPA-TDNN [3] and Thin ResNet34 [14], we propose adaptive training schemes tailored to their architectures and demonstrate the efficiency of the proposed training scheme. Our proposed training scheme improves the EER of ECAPA-TDNN and Thin ResNet34 by 28.7%, 37.9%, and 27.6%, 33.1% on two short utterance evaluation protocols, respectively.

Our contributions can be summarized as follows: 1) we propose an efficient training scheme for short utterance-driven neural SV systems by considering the characteristics of speaker-related features; 2) we redefine the type of speaker-related features depending on the length of input utterances, which is more suitable to categorize the behavior of short utterance-driven SV systems; 3) we perform a detailed analysis on the effect of the proposed training scheme and its contribution to performance improvement.

The remainder of this paper is organized as follows. Section II describes two main SV models and speaker-related features, and Section III explains our proposed training scheme along with a description of coarse-grained and fine-grained features. Section IV provides a detailed interpretation of the experiments and Section V offers an analysis of the results. Finally, we conclude the paper in Section VI.

II. BACKGROUND

A. Baseline speaker verification systems

As baseline systems for our experiments, we use the two famous deep learning models that utilize different neural network architectures, ECAPA-TDNN [3] and Thin ResNet34 [14]. Both models extract fixed-dimensional speaker embeddings from mel-spectrograms using a similar computational flow. However, they are different in the sense that ECAPA-TDNN maintains the resolution of time frames but Thin ResNet34 downsamples the input frames in the feature extraction process.

1) ECAPA-TDNN: ECAPA-TDNN and its variants [3], [15] are based on the time-delay neural network (TDNN) [16] and use 1-dimensional convolutional layers to estimate

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

speaker features. The frequency dimension of the input mel-spectrogram is first expanded to a large channel dimension using a 1-dimensional convolutional layer, and the hidden features maintain their dimension in both the channel and time axes up to the pooling layer. This time-preserving characteristic of ECAPA-TDNN enables the model to maintain the resolution of time-frames.

2) Thin ResNet34: ResNet systems utilize two-dimensional convolutional layers to process the features in both the time and frequency domains. Unlike ECAPA-TDNN, Thin ResNet34 down-samples the input mel-spectrogram by 1/4 and 1/8 in both the time and frequency axes, respectively.

B. Speaker-related features

Various speaker-related features have been studied for SV tasks. In [17], [18], such features are categorized into high-level and low-level features. High-level features include prosodic features such as pitch and energy, phonetic features such as pronunciation idiosyncrasies, and lexical features such as the preference of word usage, which reflect speaker habits and style not captured in the cepstral level. Low-level features, on the other hand, are related to the sound of a person's voice, such as cepstral features, acoustic correlates of timbre (color of sound), and features such as Mel-frequency cepstral coefficients (MFCC) [17].

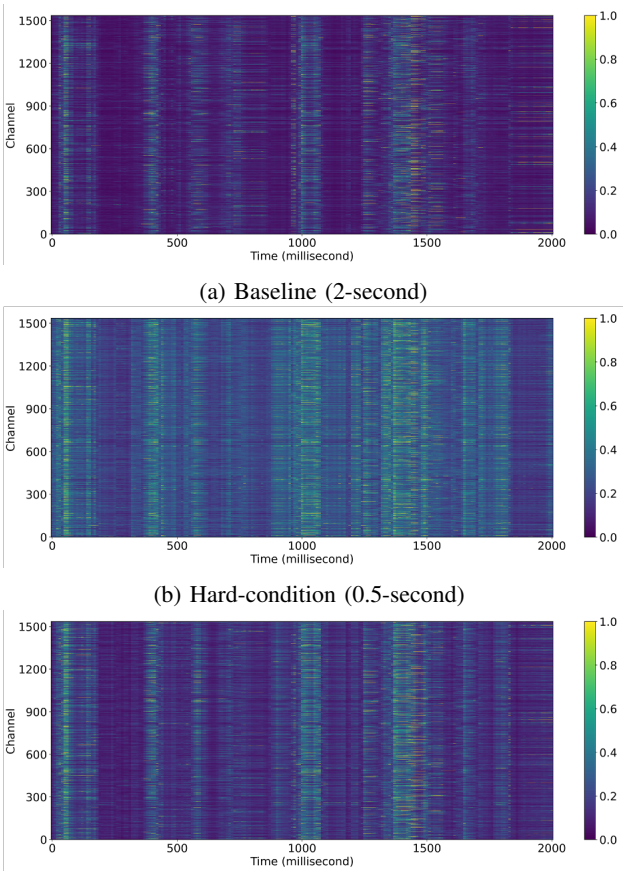
III. Proposed Training Scheme

A. Fine and coarse-grained features

The categorization of high-level or low-level feature styles mentioned above is unsuitable for recent short-duration SV tasks because it is not easy to extract high-level features from the short length of utterances. Herein, taking the length of popular short-command datasets into account, we set 2-second as the reference length to define short or long utterances. In this paper, we categorize speaker-related features into fine-grained and coarse-grained features based on the length of the utterances used for training.

Coarse-grained feature: When an SV model is trained using long utterances ($\geq$2s), it tends to exploit the characteristics more coarsely. We denote features with large attention weights as coarse-grained features, as shown in Fig. 1a. SV models mainly focus on the time frames where more distinguishable speaker characteristics exist. In Fig. 2, the attention weight of the frames with high energy is larger compared to other frames. In other words, the SV models tend to concentrate on higher energy regions, rather than finding additional speaker characteristics existing between adjacent frames.

Fine-grained feature: In contrast, when the SV models are trained on short utterances, they tend to exploit utterances in a more fine-grained manner as shown in 1b. Disregarded frames or speaker characteristics in coarse-grained features are also thoroughly investigated. Detailed analysis of the two features will be discussed in Section V-A.



(a) Baseline (2-second)

(b) Hard-condition (0.5-second)

(c) Hard-condition mix III

Fig. 1: Channel-time attention plot of the attentive statistics pooling (ASP) layers from ECAPA-TDNN (identical 2-second inputs): (a) baseline (2-second) (b) hard-condition (0.5-second) (c) hard-condition mix III (2 and 0.5-second)

B. Hard-condition training

Conventional training schemes for SV systems [2], [3], [14] utilize a fixed length of relatively long utterances (2-3 second) for training, which allows the model to learn coarse-grained features. However, this approach causes the model to overlook fine-grained features during training, leading to poor performance in short-duration SV scenarios [19], [20]. Inherently, the amount of speaker-related features is less abundant in short utterances compared to long utterances, requiring a different approach for extracting speaker-related features. On the other hand, when the model is trained using only short utterances, it learns to extract fine-grained features; however, it shows poor performance [20] with respect to long utterance SV scenarios. To address this, we propose a hard-condition training scheme that trains the SV model using both short (0.5-1 second) and long (2-3 second) utterances to exploit both coarse and fine-grained features.

To fully utilize the essence of our proposed training scheme, it is essential to consider the neural architecture and processing flow of the SV model. For time-maintaining SV models such

as ECAPA-TDNN, using the training data with the combination of 0.5 and 2-second long utterances shows the best performance. However, for SV models having a temporal-downsampling architecture such as Thin ResNet34, training on short utterances lasting under 1-second is not recommended because the downsampling to time-frames affects the training of the SV model.

IV. Experiments

A. Database

We utilize the popular benchmark dataset, namely VoxCeleb 2 [21] dev set for training, and VoxCeleb 1 [22] test subset for evaluation. We do not perform any voice activity detection (VAD) or data augmentation schemes.

B. System Description

The baseline systems for our experiment are ECAPA-TDNN, Thin ResNet34, and Thin ResNet34-X. Thin ResNet34-X is a modified version of the original Thin ResNet34 that is designed to maintain the time frames. We modify the dimension of the output embedding to 512 for all systems. 80 and 40 dimension filterbank features extracted from a 25 ms frame window with a 10 ms shift are utilized as an input for ECAPA-TDNN and Thin ResNets, respectively. The network is trained with the Adam optimizer [23]. Initial learning rate and weight decay are set to 0.001 and 2e-5, respectively. For speaker loss, AAM-Softmax [24] with a margin of 0.3 and a scale of 30 is used for ECAPA-TDNN; alternatively, AM-Softmax [25] with a margin of 0.2 and a scale of 30 is utilized for Thin ResNet34 and Thin ResNet34-X.

C. Training Strategy

In this section, we explain the proposed training scheme in greater detail. For each training scheme, an equal number of audio samples are utilized for training.

- **Baseline-training**: Randomly crop 2-second-long utterances for training.
- **Hard-condition training-mix**: Randomly crop both long and short utterances for training.

D. Evaluation protocol

To effectively look into the efficiency of our proposed method, we simulated two different evaluation scenarios. And for both scenarios, we computed the cosine similarity between the speaker embeddings to obtain equal error rates (EERs), as done in [21]. No additional score normalization technique was used for evaluation. Details of each evaluation scenario are as follows:

Case 1 (Full-Short): We first simulated a scenario that the users enroll a relatively long speech, and test using a short utterance for convenience. For enrollment, we divide the full utterance into ten 3-second-long segments with regular intervals to extract the speaker embedding from each segment. For the short test utterance, we randomly crop x-second, which is identical to the length of the input short utterance, and extract a single embedding from it for the test. Subsequently, the cosine similarity scores between ten embeddings that are extracted from a full enrollment utterance and one embedding that is extracted from a short test utterance (10×1) are utilized to compute EERs.

Case 2 (Short-Short): To further improve user convenience, we also experimented with utilizing a short utterance for both enrollment and test. For both enrollment and test, we randomly crop x-second from each utterance and extract a single embedding.

To eliminate the randomness caused by randomly cropping short segment speech, we repeated ten trials for each experiment with identical random seed sets and averaged the resultant EERs to obtain the final score.

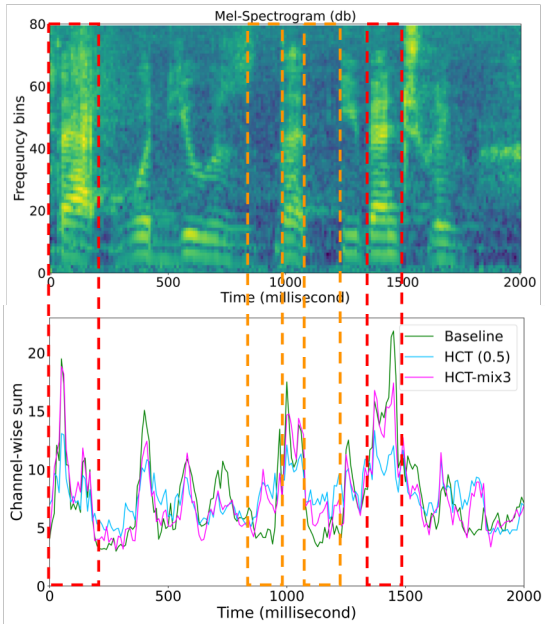


Fig. 2: Mel-spectrogram (upper) and corresponding channel-wise summation of attention weights obtained from ASP layers (lower).

V. Result and Analysis

A. Coarse-grained and fine-grained features

To analyze the coarse-grained and fine-grained features learned using different training schemes, we plot the channel-wise attention weights taken from the attentive statistic pooling layer in Fig. 1a. From the figure, we observed that the attention weights obtained from the baseline model selectively concentrate on certain time frames, whereas those of the model trained only with 0.5-second (Fig. 1b) focus on each frame in a more fine-grained manner. The HCT-mix III trained model (Fig. 1c) shows the intermediate behavior between the two.

To compare the characteristics of these features in detail, we plot the channel-wise sum of the three attentions in Fig. 1 along with the corresponding mel-spectrogram in Fig. 2. It is shown that the baseline model gives more weight to the voiced regions with high energy. On the other hand, the model

TABLE I: Performance was evaluated using the VoxCeleb1-O testset. We averaged ten evaluations for each experiment. HCT (0.5 sec) and HCT (1 sec) denote models trained with only 0.5 and 1-second, respectively. HCT-mix I, II, III denote training with a combination of [0.5, 1], [1, 2], and [0.5, 2] second with a 1:1 ratio, respectively.

Model	Training Scheme	Evaluation Protocol					
		Case 1 : Full-Short (EER)↓			Case 2 : Short-Short (EER)↓		
		2-second	1-second	0.5-second	2-second	1-second	0.5-second
ECAPA-TDNN	Baseline	2.660	5.701	13.462	4.374	12.241	27.049
	HCT (1 sec)	2.501	4.524	10.024	3.604	8.692	20.996
	HCT (0.5 sec)	3.240	4.655	**8.350**	4.120	**7.450**	**15.678**
	HCT-mix I	2.619	4.190	8.357	3.516	7.462	17.013
	HCT-mix II	2.468	4.698	10.660	3.790	9.641	22.620
	HCT-mix III	**2.402**	**4.064**	8.454	**3.340**	7.600	17.803
Thin ResNet34	Baseline	4.34	9.137	19.376	7.187	18.233	33.838
	HCT (1 sec)	4.679	7.408	14.136	6.417	13.079	26.149
	HCT (0.5 sec)	7.886	9.792	13.658	9.212	13.336	21.688
	HCT-mix I	5.881	8.157	**13.036**	7.239	11.894	**20.608**
	HCT-mix II	**3.914**	**6.614**	13.427	**5.502**	12.192	25.722
	HCT-mix III	4.764	7.032	14.004	6.194	**11.412**	22.224
Thin ResNet34-X	Baseline	**3.501**	7.108	15.496	5.600	14.548	29.474
	HCT (1 sec)	4.134	6.491	12.19	5.544	11.085	22.434
	HCT (0.5 sec)	6.051	8.021	12.404	7.230	11.274	**19.265**
	HCT-mix I	4.880	6.990	11.937	6.148	10.691	19.940
	HCT-mix II	3.804	6.421	12.614	5.392	11.692	23.436
	HCT-mix III	4.039	**6.236**	**11.481**	**5.357**	**10.586**	21.036

trained with hard conditioning utilizes every time frame with a similar weight or exploits an inter-frame relationship. To be specific, the red dashed box highlights the time frames with high energy, indicating that the baseline model tends to give higher attention to those frames. By contrast, the attention weights in the orange dashed box indicate that our proposed training scheme also considers those time frames that are less regarded by the baseline. This characteristic of the hard-condition trained model indicates the success of capturing both coarse-grained and fine-grained features.

B. Effect of hard-condition training

Table I shows that using short utterances for training is crucial for achieving good performance on short utterances in the evaluation protocol of Case 2. The best-performing models in this scenario are those trained using either 0.5-second only or a mix of 0.5-second and longer utterances. For instance, the ECAPA-TDNN model trained using HCT (0.5-second) achieves an EER of 7.450% on short utterances, which is the best result in this evaluation protocol. Given the limited amount of information available in short utterances, fine-grained features are essential for extracting speaker embedding. The experiment result of ECAPA-TDNN on HCT (1 sec) and HCT (0.5 sec) further emphasizes the importance of fine-grained features for short utterances. HCT (0.5 sec) shows a superior result over HCT (1 sec) even in 1-second where the train-test condition matches.

However, simply focusing on the fine-grained features during training is not enough to achieve good performance on long utterances in the evaluation protocol of Case 1. Considering the length discrepancy between the enrollment utterance and the test utterance, speaker embedding should be robust to the length of the utterance. Thus, the SV models should consider both coarse-grained and fine-grained features when extracting speaker embeddings. In Table I, HCT-mix trained models show better performance compared baseline model in general. Specifically, The ECAPA-TDNN model trained using HCT-mix III, which utilizes both 0.5 and 2-second for training, achieves the best performance on this evaluation protocol. This model achieves an EER of 4.064% on 1-second utterances, which is a 28.7% improvement over the baseline result, as shown in Fig. 3. In summary, these results demonstrate the importance of using short utterances for training to learn fine-grained features, as well as the need to use longer utterances to learn coarse-grained features.

C. Analysis on model architecture

*1) **ECAPA-TDNN vs Thin ResNet34**:* In general, the experimental results obtained with Thin ResNet34 and ECAPA-TDNN show similar trends. As shown in Table I, Thin ResNet34 trained using HCT-mix I, II and III resulted in promising performance on both evaluation scenarios compared to simply using a single fixed length for training. In Fig. 3, both ECAPA-TDNN and Thin ResNet34 have shown significant improvement, and the trends of relative improvement are also similar for both models. However, the lower relative improvement on 0.5-second utterances is due to using the downsampling processing of Thin ResNet34, as mentioned in Section II. The temporal downsampling characteristic of Thin ResNet34 may distract the model from learning speaker-

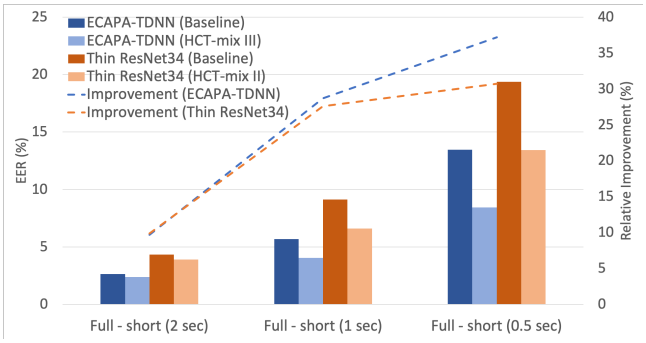


Fig. 3: SV performance evaluation protocol of Case 1 on two baseline models with relative improvement

related features on short training utterances. For example, in the case of HCT (0.5-second) for Thin ResNet34, the time frames are downsampled from 50 to 16 frames, resulting in poor performance on the evaluation protocol of Case 1.

*2) **Thin ResNet34 and Thin ResNet34-X**:* In Table I, the Thin ResNet34-X shows similar trends but superior results compared to the original Thin ResNet34 in all evaluation scenarios. This improvement is due to maintaining the length of time frames so that the model is capable for learning fine-grained features. Compared to the original Thin ResNet34, Thin ResNet34-X can process the hidden features in a more fine-grained manner and extract more detailed speaker-related features. Therefore, the result of Thin ResNet34-X resembles that of ECAPA-TDNN than the original Thin ResNet34, showing the best performance when trained using HCT-mix III.

VI. Conclusion

We thoroughly analyzed the varying feature characteristics of the speaker verification (SV) system depending on the length of training utterances. Pointing out the different speaker features on varying audio lengths and model characteristics, we emphasize the importance of using both the coarse-grained and fine-grained features for SV. By training the SV model with long and short utterances together, SV models are forced to focus not only on the coarse-grained features but also on the fine-grained features, resulting in much higher performance. As a result, our training scheme has achieved a relative improvement of 28.7%, 27.6% on the evaluation protocol of case 1 and 37.9%, 33.1% on case 2 for ECAPA-TDNN and Thin ResNet34, respectively. For future research, we plan to develop a new training framework to explore methods to effectively combine SV and keyword spotting systems.

VII. Acknowledgement

This research was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002).

References

[1] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[2] Y.-Q. Yu and W.-J. Li, "Densely connected time delay neural network for speaker verification." in *Proc. INTERSPEECH*. IEEE, 2020, pp. 921–925.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. INTERSPEECH*. IEEE, 2020, pp. 3830–3834.

[4] M. Kim, Z. Piao, S. Um, R. Lee, J. Joh, S. Lee, and H.-G. Kang, "Light-Weight Speaker Verification with Global Context Information," in *Proc. INTERSPEECH*. IEEE, 2022, pp. 5105–5109.

[5] A. Zehetner, M. Hagmüller, and F. Pernkopf, "Wake-up-word spotting for mobile systems," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 1472–1476.

[6] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (sdsv) challenge 2021: the challenge evaluation plan," *arXiv preprint arXiv:1912.06311*, 2019.

[7] ——, "Sdsv challenge 2020: Large-scale evaluation of short-duration speaker verification." in *INTERSPEECH*, 2020, pp. 731–735.

[8] B. Kim, M. Lee, J. Lee, Y. Kim, and K. Hwang, "Query-by-example on-device keyword spotting," 2019.

[9] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[10] B. Han, Z. Chen, Z. Zhou, and Y. Qian, "The sjtu system for short-duration speaker verification challenge 2021," *arXiv preprint arXiv:2208.01933*, 2022.

[11] T. Alumäe and J. Valk, "The taltech systems for the short-duration speaker verification challenge 2020." in *INTERSPEECH*, 2020, pp. 746–750.

[12] S. M. Kye, Y. Jung, H. B. Lee, S. J. Hwang, and H. Kim, "Meta-learning for short utterance speaker recognition with imbalance length pairs," 2020.

[13] J.-W. Jung, H.-S. Heo, H.-J. Shim, and H.-J. Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 335–341.

[14] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," 2020.

[15] T. Liu, R. K. Das, K. A. Lee, and H. Li, "Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7517–7521.

[16] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[17] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu *et al.*, "The supersid project: Exploiting high-level information for high-accuracy speaker recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 4. IEEE, 2003, pp. IV–784.

[18] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high-and low-level features for speaker recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[19] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[20] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva *et al.*, "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv preprint arXiv:2002.06033*, 2020.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. INTERSPEECH*. IEEE, 2018, pp. 1086–1090.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. INTERSPEECH*. IEEE, 2017, pp. 2616–2620.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular

margin loss for deep face recognition," in *Proc.CVF*, 2019, organization=IEEE, pp. 4690–4699.

[25] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. CVPR*. IEEE, 2018, pp. 5265–5274.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Adversarial Robustness of Mel Based Speaker Recognition Systems

Ritu Srivastava Saiteja Kosgi Sarath Sivaprasad Neha Sahipjohn Vineet Gandhi
CVIT, KCIS, IIIT Hyderabad
ritu.srivastava@research.iiit.ac.in

***Abstract*—Convolutional neural networks (CNN) applied to Mel spectrograms now have a dominant presence in the landscape of speaker recognition systems. Correspondingly, it is also important to evaluate their robustness to adversarial attacks that remains not thoroughly explored for end-to-end trained CNNs for speaker recognition. Our work addresses this gap and investigates variations of the iterative Fast Gradient Sign Method (FGSM) to perform adversarial attacks. We observe that a vanilla iterative FGSM can flip the identity of each speaker sample to that of every other speaker in the LibriSpeech dataset. Furthermore, we propose adversarial attacks specific to Mel spectrogram features by (a) limiting the number of pixels attacked, (b) restricting changes to specific frequency bands, (c) restricting changes to particular time duration, and (d) using a substitute model to craft the adversarial sample. Using thorough qualitative and quantitative results, we demonstrate the fragility and non-intuitive nature of the current CNN-based speaker recognition systems, where the predicted speaker identities can be flipped without any perceptible changes in the audio. The samples are available at "https://advdemo.github.io/speech/"**

I. Introduction

Convolutional neural networks have matured over the last decade and continue to show performance gains on numerous tasks [1]. However, one major concern with CNN models is that they are susceptible to small input perturbations. Previous works have shown that CNN models can be tricked to misclassify highly confident predictions by adding a small additive targeted noise to the input image [2]. Consequently, adversarial robustness has been an area of interest for deep learning research in recent years, and the robustness to adversarial attacks has become an important factor in evaluating systems. The study of adversarial robustness is specifically important in applications where the cost of misclassification is high, for example, biometric authentication. With access to the trained model (white-box attacks) or otherwise by multiple trials, an attacker can compute the gradients or its numerical approximation to break the authentication systems. To this extent, distortion-less adversarial attacks are malicious to authentication systems as they make it difficult to debug and trace the attack.

Speech signals are a popular modality for biometric authentication. Various features and modeling techniques have been proposed over the years to enhance the performance of Speaker Recognition (SR) systems. Earlier approaches relied on classifiers learned on top of common feature representations like i-vector [3] and x-vector [4]. The seminal work by Nagrani et al. [5] shows that an end-to-end CNN, when applied to raw spectrograms, gives promising results. Using Mel spectrograms as the input also circumvents the need to pre-process the audio data. Further, this approach of learning CNNs on Mel spectrograms has found success in numerous downstream tasks like automatic speech recognition, speaker recognition and style capture in speech generation.

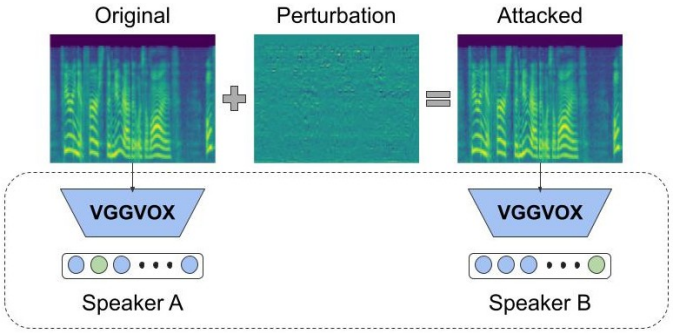


Fig. 1. Adversarial attack on Speaker Recognition Systems

Speaker recognition, among other applications, requires systems to be secure from malicious attacks and have been tested for their robustness to such attacks. Traditional testing methods focused on spoofing attacks like replaying recorded speech [6], mimicking human voice [7] and use of synthesized speech [8]. More recent forms of attacks attempt to fool models by supplying deceptive or adversarial input *i.e.*, to generate acoustic utterances that sound as one speaker (say speaker A) to the human ear but are classified by the system as another speaker (see Figure 1 for an illustration). Adversarial attacks on speaker identification systems have been studied for i-vector and x-vector systems, relying on Mel frequency cepstrum or MFCC features [9]–[11]. The adversarial robustness of current state-of-the-art, end-to-end CNN models on raw Mel spectrograms remains largely unexplored.

Our work bridges this gap and studies the adversarial robustness of end-to-end trained CNN models for speaker recognition. We perform our experiments on LibriSpeech dataset [12] using variants of Fast Gradient Sign Method (FGSM) [13] on the VGGVox architecture [5]. Our findings show that with limited change in the input pixels of Mel spectrogram, a vanilla iterative FGSM attack can change the prediction of each speaker sample to every other label in the dataset. We observe that the corresponding distortion in audio is proportional to the number of gradient updates in FGSM. We also evaluate the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

efficacy of adversarial attacks constrained to specific regions of the spectrogram. Predicted speaker labels can be flipped by limiting changes to a small time band, a small frequency band, or a few pixels (as illustrated in Figure 2). We also use a substitute model to compute the adversarial sample to show that the attack can be performed with minimal assumptions on the target model.

FGSM-like attacks become infeasible in the absence of access to the entire model. This mitigates the risk of adversarial attacks in many in-house developed real-world systems. However, since most models are trained using transfer learning, it is often the case that these backbone models are available in public. We show that adversarial samples can be crafted using a different model with the same backbone. Overall, our work makes the following contributions:

- We explore the robustness of VGGVox architecture to FGSM attack and show that it can change the prediction of a sample to any other class in the given dataset.
- We propose three constrained FGSM attacks and show that predictions can be altered with minimal changes to spectrogram that are mostly imperceptible to the human ear.
- We evaluate the performance of models when the adversarial samples are crafted without access to the actual model but using a different model with the same backbone.
- We present qualitative and quantitative results to demonstrate the efficacy of proposed methods with ablation studies analyzing the role of the number of iterations of FGSM, number of pixels, and different time/frequency bands.

II. Related Work

Early research in SR systems devoted significant resources to the study of crafting the suitable feature set for speaker recognition [14] [15]. In the pre-deep learning era, MFCC-like features were mostly used in speaker recognition systems. Classifiers using i-vectors [3], a low-dimensional speaker and channel-dependent space learned using factor analysis, achieved state-of-the-art results in SR. Subsequently, methods based on x-vector features [4] learned from deep networks outperformed earlier models [16]. x-vectors are extracted from a deep neural network trained for speaker classification [4].

Nagrani et al.[5] proposed modeling the SR system as an end-to-end convolutional network over 2D representation of Mel frequency intensities. End-to-end CNN models gave promising results and offer a methodically simple solution to SR. Since [5], using convolutions on Mel features has become popular in many speech and audio based tasks like audio tagging [17], acoustic scene classification [18], speech emotion classification [19], sound event detection [20], and style capture in speech generation [21].

Since CNNs are known to be susceptible to adversarial attack in image classification literature [2], it is natural to ask; how robust are speech and audio systems built with convolutions on spectrograms to adversarial attacks? Especially those like SR that are used in user authentication systems.

Different speaker recognition methods have been studied for their adversarial robustness. A detailed survey of existing research on the robustness of speaker recognition systems is presented in [22]. Li et al. [9] study the vulnerability of x-vector based SR system by developing adversarial samples that remain effective while being played over-the-air. This is achieved by modeling the air channel using the estimated room impulse response, then convolving it with the input, followed by FGSM. Wang et al. [11] also shows that x-vector based SR system can change speakers across genders with no perceptible distortion in the input. Chen et al. [10] studies the robustness of SR systems to adversarial attacks to understand their security weakness in the practical black-box setting. They propose an adversarial attack and explore its efficacy on i-vector based systems, GMM-UBM systems, and x-vector based systems.

The prior art to be seen as closest to our work is [11], which proposed an imperceptible white-box psychoacoustics-based method to attack an x-vector model. We study the adversarial robustness of speaker recognition systems, which are end-to-end CNN based models that take raw spectrogram as input with softmax layer to predict the speaker label. Our work focuses on making a comprehensive analysis by restricting the attacks in time and frequency bands, limiting the number of editable pixels, and crafting adversarial samples by making minimal assumptions about the model.

III. Method

Adversarial samples are inputs explicitly designed to confuse or mislead a neural network, resulting in the misclassification of the sample. FGSM uses the gradients of the neural network to create the adversarial example. This method is known to achieve the attack through imperceptible additive perturbations.

Given an input spectrogram $\mathbf{x}$ belonging to y, we wish to generate $\tilde{\mathbf{x}} \sim \mathbf{x} + \eta$ with perturbation η that model will misclassify into y^*. FGSM generates perturbations using gradients of the loss J at $\mathbf{x}$ using y^* to generate $\tilde{\mathbf{x}}$ as,

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \operatorname{sign}\left(\nabla_{\mathbf{x}} J(\mathbf{x}, y^*; \theta)\right) \tag{1}$$

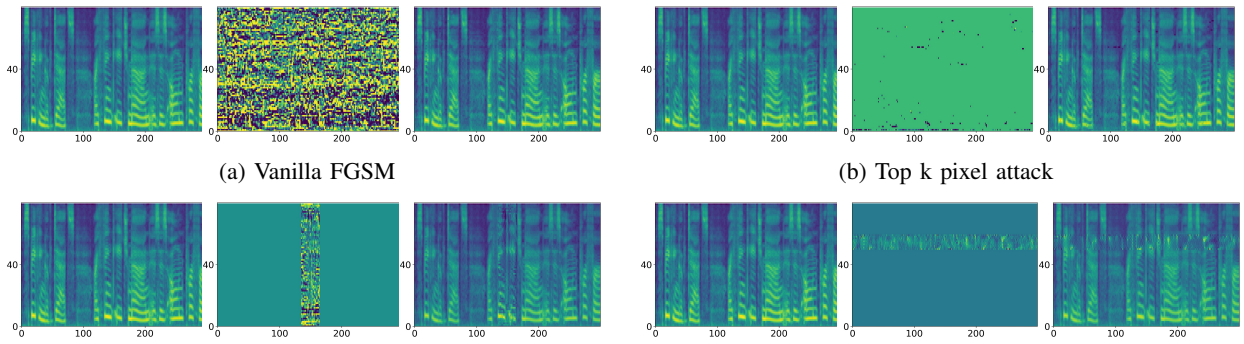
where θ is the network parameter and ϵ is a hyper-parameter controlling magnitude of the perturbation. We apply FGSM iteratively in our experiments with initialization $\tilde{\mathbf{x}}_0 = \mathbf{x}$ as,

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \epsilon \operatorname{sign}\left(\nabla_{\tilde{\mathbf{x}}_{t-1}} J(\tilde{\mathbf{x}}_{t-1}, y^*; \theta)\right) \tag{2}$$

and terminating the loop when the predicted class label is flipped to the desired $y^* = \operatorname{argmax}_y f(\tilde{\mathbf{x}}^*; \theta)$. We propose variants of iterative FGSM , all of which can be considered as different masks applied on the gradient before applying on $\tilde{\mathbf{x}}$. The masked variations is,

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \epsilon \left(\mathbf{M}_{t-1} \odot \operatorname{sign}\left(\nabla_{\tilde{\mathbf{x}}_{t-1}} J(\tilde{\mathbf{x}}_{t-1}, y^*; \theta)\right)\right) \tag{3}$$

where $\mathbf{M}_{t-1}$ is the binary mask and $\odot$ represents element-wise product. We evaluate three different masks in this work.



(a) Vanilla FGSM

(b) Top k pixel attack

(c) Selected time band

(d) Selected frequency band

Fig. 2. Figure shows the effect of the four attacks used on the spectrogram. For each experiment the three spectrograms in the figure shows: the original spectrogram, the computed perturbation and the attacked spectrogram respectively

Top-k pixels: In this variant, we restrict perturbations to a fraction k of all pixels in the spectrogram $\tilde{\mathbf{x}}$ with the largest gradient magnitude in the first iteration. Mask $[\mathbf{M}]_{ij} = 1$ is computed at the first iteration consisting of pixels with magnitude of the corresponding entry in gradient $[\text{abs}(\nabla_{\tilde{\mathbf{x}}} J(\tilde{\mathbf{x}}, y^*))]_{ij} > \tau$ and $[\ \cdot\ ]_{ij}$ is indexing operation into (i, j) element. Threshold τ is chosen such that $k * N$ pixels are allowed to be altered. N is the total number of pixels. In subsequent iterations, only pixels corresponding to $[\mathbf{M}]$ are updated. The resulting mask satisfies $\mathbf{1}^\top \mathbf{M} \mathbf{1} = k * N$ where $\mathbf{1}$ is a vector of ones.

Frequency band: In this case, we allow the attack only on a chosen frequency band of the spectrogram. The frequency band is a horizontal strip in the spectrogram, as shown in Figure 2(c). The mask here is a fixed matrix across all iterations set to $[\mathbf{M}_t]_{ij} = 1$ only for $i \in [\tilde{\nu} - \nu/2, \tilde{\nu} + \nu/2]$ for a chosen frequency $\tilde{\nu}$ and window length of ν around it. We report experiments for various frequencies $\tilde{\nu}$.

Time period: Similar to the frequency band restriction, we consider limiting attack to a specific time span of the speech signal. This corresponds to constraining the attack to a vertical strip of the spectrogram and the mask is fixed across all iterations to $[\mathbf{M}_t]_{ij} = 1$ only for $j \in [\tilde{\rho} - \rho/2, \tilde{\rho} + \rho/2]$ for a chosen time $\tilde{\rho}$ with window ρ.

Using a substitute model: We evaluate the adversarial robustness of SR models in two different scenarios. We train a model(M_1) on N speakers and craft an adversarial sample using this. The target model (M_2) has n speakers, such that $n << N$. M_2 is fine-tuned with the backbone of M_1. In the first setting, the two sets of speakers used to train source model M_1 and target model M_2 are mutually exclusive and exhaustive. In the second setting, one of the speakers is common between the training sets of M_1 and M_2.

IV. Experiments

We experiment with a CNN speaker recognition model as in [5] trained on LibriSpeech [12] corpus. While the complete LibriSpeech corpus consists of about 1000 hours of speech from 2400 speakers, we use only a clean subset of 360 hours comprising 927 speakers. We sample the waveforms at 22KHZ and convert them into Mel spectrograms with 80 Mel channels by short-time Fourier transform using a Hamming window of length 1024 and hop length of 256. We train the speaker discriminator with a simple training strategy using no augmentation.

We perform our FGSM attacks using the aforementioned trained model. For all our experiments, we compare the audio generated from the attacked Mel spectrogram with the audio generated from the original spectrogram using MelGan [23]. Also, we limit the maximum number of perturbation iterations allowed to 5000. In other words, we label the adversarial attack unsuccessful if the network prediction is not converted to the target label within 5000 updates.

We perform a human survey to qualitatively measure the distortion in the audio generated by the attacked Mel spectrogram. For each of our four experiments, we ask fifty users to evaluate twenty pairs of audio each. Each pair constitutes one audio generated from the original Mel spectrogram and the other generated from the attacked Mel spectrogram. We ask the participants to rate the distortion in the attacked audio compared to the former on a Likert scale of three; with zero corresponding to no distortion at all, one for minor distortions, and two for significant distortions. We report the Mean Opinion Score (MOS) of the participants for all four experiments.

Vanilla iterative attack vs. Top-k pixel attack: In the first experiment, we apply the vanilla iterative FGSM attack described in Equation 2 on one random sample of each 927 speakers. Each sample is perturbed to change the prediction to that of every other speaker in the dataset. We report the average number of iterations taken to change the label for each speaker. In the second experiment, we evaluate the efficacy of the top-k pixels attack strategy for two distinct values, $k = 1.25\%$ and $k = 2.5\%$. We measure the Perceptual Evaluation of Speech Quality (PESQ), Signal to Noise Ratio (SNR), and Mean Square Error (MSE) between the audio from the attacked Mel spectrogram and the audio from the original spectrogram for quantitative evaluation.

Frequency band and time period attack: We choose 50 speakers from the dataset for experiments with restricted

frequency and time band adversarial attack. We use stratified sampling to include same number of male and female speakers for this subset. We attack a sample from each speaker from this subset with all other speakers as target label. We choose a mask of width 1 KHz and evaluate the attack on ten such non-overlapping positions. Frequency values $\tilde{\nu}$ are chosen such that the 10 non-overlapping windows span the entire width of the spectrogram ranging from 0-8 KHz. In the fourth experiment, we evaluate the efficacy of time segment based attacks. We evaluate time segment attack by choosing a mask of width ρ as small as 2 frames to a maximum of 30 frames. We position the mask at the spectrogram's beginning, middle, and end.

Using a substitute model: We perform two experiments to evaluate the adversarial robustness of models when the adversarial sample is crafted using a substitute model. In the first experiment, we show the transferability of attacks using samples crafted from a model trained on a different set of labels. This experiment evaluates the adversarial robustness when a model with the same backbone is publicly available for crafting adversarial samples. We use 800 speakers to train the source model (M_1) and 121 different speakers to train the target model (M_2). M_2 is initialized with the backbone of M_1, and the backbone weights are frozen during the training. There are no common speaker identities between the trained set of M_1 and M_2, and we attack using the M_1 model to change the label to a speaker from the opposite gender and evaluate it on M_2. We flip between the two gender labels provided in the LibriSpeech dataset.

In this attack, we assume we cannot access the model M_2. For a sample $s \in testdata(M_2)$, we make a prediction using M_1. We then apply an iterative FGSM attack in M_1 to flip the identity so that the gender label flips. We call the attacked sample $\hat{s}$, following $(gender(M_1(s)) \neq gender(M_1(\hat{s})))$. We evaluate $\hat{s}$ on M_2 and call the attack successful if the prediction changes, in other words, $M_2(s) \neq M_2(\hat{s})$.

In the second experiment, we use the source model trained on 800 speakers and fine-tune the model on a new set of 127 speakers with one speaker C_s common to the substitute and source models. We perform targeted FGSM attack using the source model, such that the identity of every sample in the target model is changed to C_s.

V. Results and Discussion

Vanilla iterative attack vs. Top-k pixel attack: On training the CNN [5] model on the selected 927 speakers of LibriSpeech [12] corpus, we get an accuracy of 98.7% on the test set with an average confidence of 93% for prediction. Despite the network having such high accuracy and high confidence in the prediction, Table I shows that vanilla iterative FGSM can change samples from any 927 speakers to every other speaker in the dataset with minimal changes. The distortions in the audio caused by the attack are imperceptible in most cases, with a mean MOS of 0.08 in distinguishing the speech generated from original and attached spectrograms. We also show the results for the fifty samples which incurred most iterations in Table I. We observe higher distortion in these samples, with mean MOS jumping to 0.79. This correlation between the number of iterations and distortion can be observed from both qualitative and quantitative results Table I.

Table I further shows that allowing attack on only 2.5% pixels can still give an attack success rate of 89.2% in flipping every speaker to any other speaker in the dataset. It gives no significant added distortion to the audio as well. With a lower k, as low as 1.25% (about 300 pixels), we observe that 69% of the attacks are successful. As the MOS values indicate, both these attacks are almost imperceptible to humans.

To strengthen the claim of adversarial vulnerability of these models, we train the model with data augmentation (SpecAugment [24]) for randomly selected 200 speakers. We observe no significant improvement in robustness in simple iterative FGSM. We observe that the attack success rate remains the same post-augmentation, with the average number of iterations being 123.19 (without augmentation, the average number of iterations for the selected two hundred speakers is 113.11). The corresponding distortions are imperceptible.

Time period and frequency band attack: When the number of pixels modified is restricted to a time band as small as 2 frames (that is equivalent to 23ms), the average MOS score across different placements of the time band is 1.05. Table IV shows that the distortion is more perceptible when the time band appears in the middle of the audio clip compared to the beginning or end. In our study, the MOS when the mask is placed in the center is at least 50% more than the next best placement. It is interesting to note that the model prediction changes with a distortion limited to just 2ms output of an audio that spans three seconds. The distortion becomes less (and almost imperceptible) when the time window is increased to 30 frames (or 0.34s) with a mean MOS of 0.6.

Table III shows that the distortion caused by restricting

TABLE I
QUALITATIVE AND QUANTITATIVE EVALUATION. ASR REFERS TO ATTACK SUCCESS RATE AND ITER. IS THE NUMBER OF ITERATIONS.

	ASR	ITER.	MOS	Quantitative Results		
				PESQ	SNR	MSE
V-FGSM	100%	61	0.08	3.71	57.72	10.21
V-FGSM for 50 largest t	100%	182	0.79	2.56	59.17	17.34
Top-k, $k = 1.25$	63.9%	1401	0.17	3.20	56.47	10.91
Top-k, $k = 2.5$	89.2%	1048	0.15	3.22	58.41	12.65

TABLE II
QUALITATIVE AND QUANTITATIVE EVALUATION FOR SUBSTITUTE MODEL. ASR REFERS TO ATTACK SUCCESS RATE AND ITER. IS THE NUMBER OF ITERATIONS. THE ROWS SHOW ITERATIVE FGSM FOR SUBSTITUTE MODEL FOR MALE TO FEMALE AND FEMALE TO MALE RESPECTIVELY

	ASR	ITER.	MOS	Quantitative Results	
				PESQ	MSE
(F to M)	85.2%	112	0.14	3.12	0.0023
(M to F)	91.6%	110	0.26	2.74	0.0037

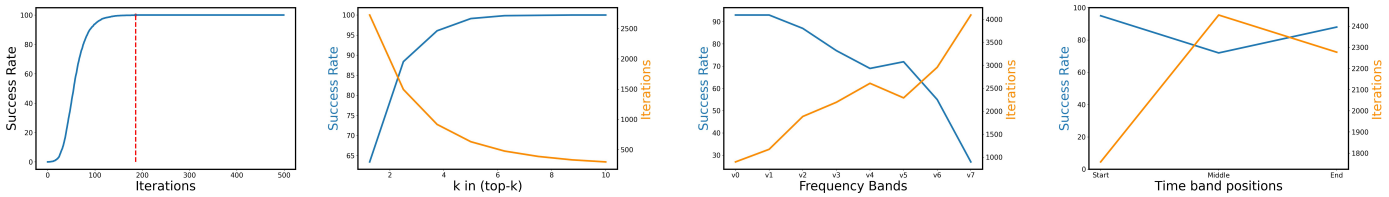


Fig. 3. Iterations vs. attack success rate trade-offs. See Section V on Results and Discussion for description, to be viewed in color.

TABLE III
MOS EVALUATION OF DIFFERENT FREQUENCY BANDS.

	ν_7	ν_6	ν_5	ν_4	ν_3	ν_2	ν_1	ν_0
MOS	0.82	0.07	0.10	0.28	0.17	0.14	0.21	0.60
PESQ	3.61	3.84	3.48	3.36	3.17	3.2	2.8	2.6

TABLE IV
MOS EVALUATION OF DIFFERENT TIME BANDS.

	Start	Middle	End
$\rho = 2$	0.92	1.38	0.84
$\rho = 30$	0.44	0.9	0.46

attacks to a specific frequency band is almost imperceptible irrespective of where we place the band except on the two bands on the extreme end of the spectrum. We hypothesize that since the lower frequencies(0-1 KHz) have higher activity with higher amplitudes, they are more perceptible with even minor distortions leading to significant audible differences. Our experiments show that with a sufficiently large band of 1 KHz, we can successfully attack the system in the range of 2-7 KHz without incurring much distortion in the audio.

Using a substitute model: On training the CNN [5] model such that target model (M2) has 121 speakers fine tuned with the backbone of pretrained model M1 on the selected 800 speakers of LibriSpeech [12] corpus, we get an accuracy of 100%. Despite the network having such high accuracy, Table II shows that iterative FGSM can change samples from any of the male speakers to every other female speaker in the dataset and vice versa with minimal change.

Transferability of attacks: We show the transferability of attacks using samples crafted from publicly available pre-trained models. We create a substitute model by training speaker detection on a portion of speakers. We fine-tune the model on a new set of 121 speakers with no speaker common to the substitute model label set. Here we observe an attack success rate of 84.2% when attack sample is crafted using substitute pretrained model and attack is performed on the fine-tuned model. The attack success rate is 91.6% when the source speaker is male and target speaker is female. Similarly, the attack success rate is 85.2% when source speaker is a female from the set of 800 speakers, and target speaker is a male from the set of 121 speakers.

We also fine-tune the model on a new set of 200 speakers with one speaker C_s common to the substitute model label set. We observe an attack success rate of 60% (with attack target C_s) when attack sample is crafted using substitute pretrained model and attack is performed on the fine-tuned model. This shows that with no access to the fine-tuned model, we can successfully attack a model with assumptions about the backbone.

Attack Cost vs. Success Rate In this section, we discuss the trade-off between Attack Success Rate (ASR) and the cost incurred for an attack measured in terms of the iterations required for the attack to succeed. In each run, we fix the maximum number of iterations required. We report the success rate over a range of iterations. Figure 3(a) shows the attack success and the iterations required for vanilla iterative FGSM attack. With just 186 iterations, we can achieve a 100% success rate with an average of 61 iterations across samples. For Top-k pixel attack, we can observe from Figure 3 (b) that with increase in k, the attack success rate increases. Which suggests it is easy to attack if we can modify more pixels. It can also be observed that attack success rates and iterations required are inversely correlated, suggesting if it is easy to attack a sample, it would require less number of iterations and attack success rate would be more. Conversely, if it is difficult to attack, the attack success rate would be less and requires large number of iterations.

We study the success-rate vs. iteration trade-off in frequency band and time period restricted attacks. Figure 3(c) shows the attack success rates vary across frequency bands. When limiting the number of iterations to 5000, the attack success rate largely decreases with frequency while the number of iterations required increases. Figure 3(d) shows the impact of the adversarial attack on time bands positioned at different positions (Start, Middle, and End) for a time band of 30 frames. The attack success rate remains above 75% across all three positions, while the number of iterations required is lower for the first position.

VI. CONCLUSIONS

We explored the adversarial robustness of the state-of-the-art speaker recognition system built from convolution operations. We observe that a vanilla FGSM attack can successfully flip speaker labels of every speaker to any other sample without perceivable changes to the audio. Further, the distortion in audio is proportional to the number of gradient updates in FGSM but is not correlated to the distance between the

embeddings learned by the speaker discriminator. We also evaluate the efficacy of adversarial attacks when the attack is restricted to a small set of pixels or a small band of time or frequency within the spectrogram, and empirical results show that it is possible to trick the classifier without major perturbations to the input. We would like to study further if these observations hold to the numerous downstream tasks which is very likely, and explore avenues for improving these high accuracy systems to be more robust to such attacks.

REFERENCES

[1] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, 2020.

[2] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, 2019.

[3] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, *et al.*, "i-vector based speaker recognition on short utterances," in *Interspeech*, 2011.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[5] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Interspeech*, 2017.

[6] H.-J. Shim, J.-W. Jung, H.-S. Heo, S.-H. Yoon, and H.-J. Yu, "Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes," in *TAAI*, 2018.

[7] M. Farrús Cabeceran, M. Wagner, D. Erro Eslava, *et al.*, "Automatic speaker recognition as a measurement of voice imitation and conversion," *International Journal of Speech, Language and the Law*, 2010.

[8] Z. Wu, T. Kinnunen, N. Evans, *et al.*, "Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech*, 2015.

[9] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *HotMobile*, 2020.

[10] G. Chen, S. Chenb, L. Fan, X. Du, *et al.*, "Who is real bob? adversarial attacks on speaker recognition systems," in *SP*, 2021.

[11] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," *arXiv:2005.10637*, 2020.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *ICLR*, 2015.

[14] S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *Signal processing letters*, 2005.

[15] M. McLaren, R. Vogt, B. Baker, *et al.*, "Experiments in svm-based speaker verification using short utterances.," in *Odyssey*, 2010.

[16] D. Snyder, D. Garcia-Romero, *et al.*, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP*, 2019.

[17] J. Pons, O. Nieto, M. Prockup, E. Schmidt, *et al.*, "End-to-end learning for music audio tagging at scale," *ISMIR*, 2018.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, *et al.*, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *TASLP*, 2020.

[19] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, 2019.

[20] T. Chen and U. Gupta, "Attention-based convolutional neural network for audio event classification with feature transfer learning," in *CVSSP*, 2018.

[21] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*, 2018.

[22] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," in *2021 IEEE symposium on security and privacy (SP)*, IEEE, 2021, pp. 730–747.

[23] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Neurips*, 2019.

[24] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Joint Drum Transcription and Metrical Analysis Based on Periodicity-Aware Multi-Task Learning

Daichi Kamakura, Eita Nanamura, Takehisa Oyama, and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Japan
E-mail: {kamakura, nakamura, ooyama}@sap.ist.i.kyoto-u.ac.jp, yoshii@i.kyoto-u.ac.jp

***Abstract*—This paper describes a multi-task learning method that detects the onset times of drums and the beat and downbeat times from a music signal. Since the drum part typically consists of repetitive patterns synchronized with the metrical structure in popular music, drum transcription and metrical analysis would benefit each other. The basic approach is to train a deep neural network (DNN) with a branching architecture that extracts latent features common to both tasks from a music signal and then feeds them to task-specific networks separately. The estimated frame-level posterior probabilities of drum onsets, beats, and downbeats, however, often have weak and inconsistent periodic structures. To solve this problem, we propose a regularized training method that encourages the three probability sequences to be highly auto- and cross-correlated. Specifically, our method aims to minimize the entropy of the auto-spectrum computed from each probability sequence and that of the cross-spectrum computed from each of the three sequence pairs, because a smaller entropy indicates a stronger element-wise periodicity or pairwise consistency. The experiment showed the mutual benefit of drum transcription and metrical analysis and the effectiveness of the periodicity- and consistency-based regularizations.**

I. Introduction

Automatic drum transcription (ADT) is one of the most fundamental subtasks of automatic music transcription (AMT) that aims to estimate the symbolic score of a music signal. Since the drum part affects the rhythmic characteristics of the song in popular music, drum transcription forms the basis of music information retrieval (MIR). Although the ultimate goal of ADT is audio-to-score transcription, we focus on audio-to-MIDI transcription that aims to estimate the onset times of drums in seconds as an intermediate goal.

In recent studies on audio-to-MIDI drum transcription, deep neural networks (DNNs) have successfully been used for estimating from a music spectrogram the posterior probabilities of the presence of drum onsets at the frame level [1]–[3]. Given a ground-truth binary sequence representing the presence or absence of drum onsets, one can train a DNN such that the posterior probability of the binary sequence is maximized, i.e., the cross-entropy loss is minimized. If only a limited amount of training data is available as is often the case with AMT [4], [5], however, the repetitive structures of the drum part observed at multiple metrical levels might not be captured well.

In this paper, we take the multi-task learning approach to joint drum transcription and metrical analysis (beat and downbeat detection). Both tasks are typically defined as binary classification problems that aim to estimate the presence or absence of drum onsets, beats, and downbeats at the frame level, and are expected to benefit each other by extracting common latent features with respect to periodicity from a music signal [2]. In this approach, one may train a DNN with a common feature extractor followed by task-specific binary classifiers. In such a basic branching architecture, however, the posterior probabilities of drum onsets, beats, and downbeats are estimated in a conditionally independent manner and are thus not guaranteed to be consistent in terms of periodicity.

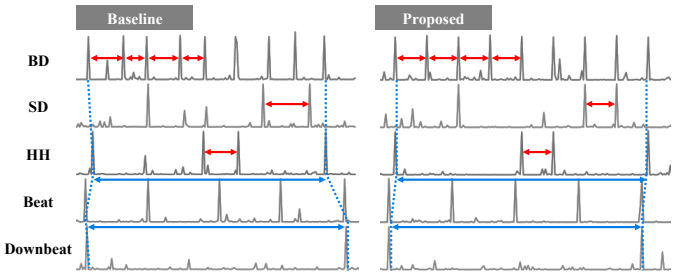


Fig. 1. The frame-level sequences of drum onset, beat, and downbeat probabilities estimated by the multi-task learning method without/with the periodicity-aware regularization. The red and blue arrows indicate the element-wise periodicity and pairwise consistency, respectively.

To mitigate this problem, we propose a regularized training method that encourages the sequences of drum onset, beat, and downbeat probabilities to have periodic patterns synchronously (Fig. 1). We focus on the fact that the tempo tends to be kept constant throughout the entire song in popular music. This requests that each sequence should have clear periodicity solely and that a sequence should share the same periodicity with another sequence. We thus propose element-wise and pairwise regularization terms described in detail below.

First, we focus on the auto-correlation function of each sequence. If the sequence of drum onset probabilities is repetitive at the measure level, its auto-correlation function exhibits sharp peaks spaced at the downbeat interval. The auto-spectrum obtained by applying the discrete Fourier transform (DFT) to the auto-correlation function exhibits a few peaks at the downbeat frequency. The *entropy of the auto-spectrum* can thus be used as a regularization term to be minimized. The same applies to the sequences of beat and downbeat probabilities exhibiting the periodicity at the beat and downbeat levels.

Second, we focus on the cross-correlation function for each of the three sequence pairs. In general, every time the sequence

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of drum onset probabilities is time-shifted by a downbeat interval, it matches well the sequence of downbeat probabilities in terms of peak positions. The cross-correlation function exhibits sharp peaks spaced at the downbeat interval. The *entropy of the cross-spectrum* can thus be another regularization term. The same applies to the other two pairs.

We experimentally show that the multi-task learning approach is especially effective for improving the performance of metrical analysis and the regularized training method has a potential of improving the performances of drum transcription and metrical analysis.

II. Related Work

Deep learning has been the central choice in modern ADT [2], [3], [5]–[10]. In general, a deep neural network (DNN) is trained in a supervised manner using paired data consisting of music signals and drum onset annotations. Due to the impulsive nature of drums sounds, CNNs have widely been used as basic building blocks for extracting local features. In particular, a CNN variant called the temporal convolutional network (TCN) [11], [12] has gained much attention because it can take into account both short- and long-term dependencies [13]. To consider long-term contextual information, a CNN is often followed by a recurrent neural network (RNN), resulting in a convolutional RNN (CRNN) [2].

To improve the performance with a limited amount of paired data, one may use data synthesis [14], data augmentation [5], or unsupervised learning [15]. However, these methods are still insufficient for dealing with considerable variations in real-world drum sounds. Alternatively, regularized training techniques based on prior knowledge of drum patterns have been explored [10] as well as phase consideration [16] and architectural innovations [7]. In this study, we focus on the periodicity of drum onsets for regularization.

For metrical analysis, DNN-HMM combinations have often been used [13], [17]–[20]. In this approach, a DNN is used for estimating sequences of beat and downbeat probabilities, and an HMM is then used for detecting beat and downbeat times from these sequences. Instead of the standard HMM, a hidden semi-Markov model (HSMM) that explicitly represents the duration of each state was proven to be effective for taking into account the tempo and meter consistency of beat and downbeat times. In this study, we use a similar HSMM-based decoder to obtain consistent results.

III. Proposed Method

This section describes the multi-task learning of drum transcription and metrical analysis with a periodicity-aware regularized training method.

A. Problem Specification

We tackle drum transcription and beat and downbeat detection from a music signal. We use the power spectrograms of the left and right channels of the audio signal $\mathbf{X} \in \mathbb{R}^{2\times F\times T}$ as input. Let $\mathbf{Y}^{\mathrm{D}} \in \{0,1\}^{K\times T}$, $\mathbf{Y}^{\mathrm{B}} \in \{0,1\}^{T}$, and $\mathbf{Y}^{\mathrm{W}} \in \{0,1\}^{T}$ denote the presence or absence of drum onsets, that of beats, and that of downbeats at the frame level, respectively, where F is the number of frequency bins, T is the number of frames, and K is the number of drum instruments. We focus on the three major drum classes ($K = 3$): bass drum (BD), snare drum (SD), and hi-hats (HH). Let $\mathbf{Y} \triangleq \{\mathbf{Y}^{\mathrm{D}}, \mathbf{Y}^{\mathrm{B}}, \mathbf{Y}^{\mathrm{W}}\}$. In the training phase, both $\mathbf{X}$ and $\mathbf{Y}$ are given as paired data. In the test phase, $\mathbf{Y}$ is estimated from $\mathbf{X}$.

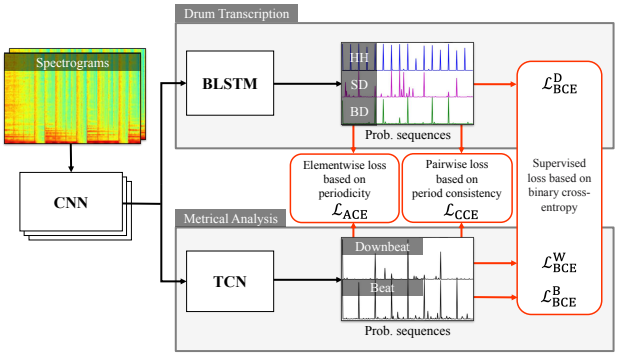


Fig. 2. The proposed joint drum transcription and metrical analysis based on periodicity-aware regularized multi-task learning.

B. Multi-Task Learning

We explain the loss function to be minimized for joint drum transcription and metrical analysis. Inspired by [17], [20], we use a CRNN with a branching architecture for jointly estimating a sequence of drum onset probabilities $\boldsymbol{\phi}^{\mathrm{D}} \in [0,1]^{K\times T}$, that of beat probabilities $\boldsymbol{\phi}^{\mathrm{B}} \in [0,1]^{T}$, and that of downbeat probabilities $\boldsymbol{\phi}^{\mathrm{D}} \in [0,1]^{T}$ from the input $\mathbf{X}$ (Fig. 2).

1) Drum Transcription Loss: The drum transcription loss is given by the binary cross-entropy (BCE) of $\boldsymbol{\phi}^{\mathrm{D}}$ for the ground-truth data $\mathbf{Y}^{\mathrm{D}}$:

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{BCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T}\left(Y^{\mathrm{D}}_{kt}\log\phi^{\mathrm{D}}_{kt} + (1-Y^{\mathrm{D}}_{kt})\log(1-\phi^{\mathrm{D}}_{kt})\right). \quad (1)$$

2) Metrical Analysis Losses: The metrical analysis losses are given by the BCEs of $\boldsymbol{\phi}^{\mathrm{B}}$ and $\boldsymbol{\phi}^{\mathrm{W}}$ for the ground-truth data $\mathbf{Y}^{\mathrm{B}}$ and $\mathbf{Y}^{\mathrm{W}}$:

$$\mathcal{L}^{\mathrm{B}}_{\mathrm{BCE}} = -\sum_{t=1}^{T}\left(Y^{\mathrm{B}}_{t}\log\phi^{\mathrm{B}}_{t} + (1-Y^{\mathrm{B}}_{t})\log(1-\phi^{\mathrm{B}}_{t})\right), \quad (2)$$

$$\mathcal{L}^{\mathrm{W}}_{\mathrm{BCE}} = -\sum_{t=1}^{T}\left(Y^{\mathrm{W}}_{t}\log\phi^{\mathrm{W}}_{t} + (1-Y^{\mathrm{W}}_{t})\log(1-\phi^{\mathrm{W}}_{t})\right). \quad (3)$$

3) Total Loss: For joint training, we aim to minimize the weighted sum of the individual losses:

$$\mathcal{L}_{\mathrm{BCE}} = \lambda^{\mathrm{D}}_{\mathrm{BCE}}\mathcal{L}^{\mathrm{D}}_{\mathrm{BCE}} + \lambda^{\mathrm{B}}_{\mathrm{BCE}}\mathcal{L}^{\mathrm{B}}_{\mathrm{BCE}} + \lambda^{\mathrm{W}}_{\mathrm{BCE}}\mathcal{L}^{\mathrm{W}}_{\mathrm{BCE}}, \quad (4)$$

where $\lambda^{\mathrm{D}}_{\mathrm{BCE}}$, $\lambda^{\mathrm{B}}_{\mathrm{BCE}}$, and $\lambda^{\mathrm{W}}_{\mathrm{BCE}}$ are adjustable weights.

C. Periodicity-Aware Regularization

We explain regularization terms that encourage element-wise periodicity and pairwise consistency. We first explain the core ideas underlying the regularization terms (Figs. 3(a) and 3(b)) and describe the regularization terms.

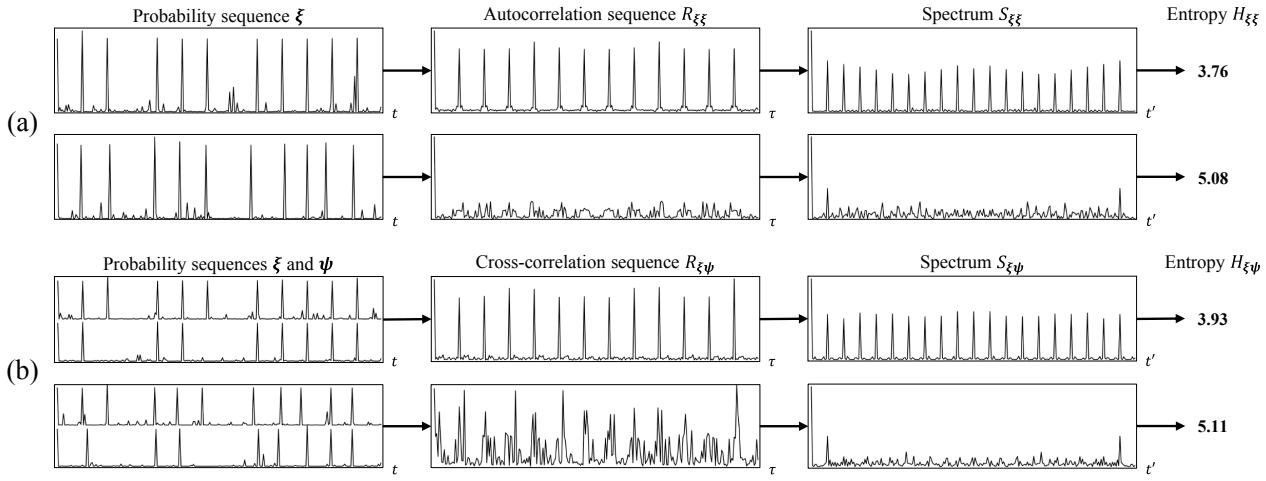


Fig. 3. (a) Examples of probability sequences and the corresponding auto-correlation functions and auto-spectra with entropies. A more periodic probability sequence results in the auto-spectrum with a smaller entropy. (b) Examples of probability sequence pairs and the corresponding cross-correlation functions and cross-spectra with entropies. A more consistent probability sequence pair results in the cross-spectrum with a smaller entropy.

1) Auto-Correlation Analysis: The auto-correlation function for a probability sequence $\boldsymbol{\xi} \in [0,1]^T$ is given by

$$R_{\boldsymbol{\xi}\boldsymbol{\xi}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \xi_t \xi_{(t+\tau-1)\%T+1}, \tag{5}$$

where $\%$ represents the modulo operation. Using the Wiener-Khinchin theorem, the spectrum of $R_{\boldsymbol{\xi}\boldsymbol{\xi}}$ is given by

$$S_{\boldsymbol{\xi}\boldsymbol{\xi}} = F R_{\boldsymbol{\xi}\boldsymbol{\xi}} = |F\boldsymbol{\xi}|^{\cdot 2}, \tag{6}$$

where $F \in \mathbb{C}^{T\times T}$ represents the discrete Fourier transform (DFT) matrix and $|\mathbf{a}|^{\cdot 2}$ represents the element-wise operation that takes the absolute square values of the elements of a vector $\mathbf{a}$. The entropy of the spectrum $S_{\boldsymbol{\xi}\boldsymbol{\xi}}$ is given by

$$H_{\boldsymbol{\xi}\boldsymbol{\xi}} = -\sum_{t=1}^{T} \bar{S}_{\boldsymbol{\xi}\boldsymbol{\xi}}(t) \log \bar{S}_{\boldsymbol{\xi}\boldsymbol{\xi}}(t), \tag{7}$$

$$\bar{S}_{\boldsymbol{\xi}\boldsymbol{\xi}} = \mathrm{Normalize}(S_{\boldsymbol{\xi}\boldsymbol{\xi}}), \tag{8}$$

where $\mathrm{Normalize}(\mathbf{a})$ represents the normalization operation that makes the sum of the elements of a vector $\mathbf{a}$ equal to 1. If the probability sequence $\boldsymbol{\xi}$ has clear periodicity, i.e., it has regular periodic patterns and the peak intervals are integer multiples of some basic time unit (e.g., tatum, beat, or downbeat), the auto-correlation function and the auto-spectrum have equally-spaced sharp peaks. The negative entropy $-H_{\boldsymbol{\xi}\boldsymbol{\xi}}$ can be used as an indicator of periodicity.

2) Cross-Correlation Analysis: The cross-correlation function for probability sequences $\boldsymbol{\xi}, \boldsymbol{\psi} \in [0,1]^T$ is given by

$$R_{\boldsymbol{\xi}\boldsymbol{\psi}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \xi_t \psi_{(t+\tau-1)\%T+1}. \tag{9}$$

Using the Wiener-Khinchin theorem, the magnitude spectrum of $R_{\boldsymbol{\xi}\boldsymbol{\psi}}$ is given by

$$S_{\boldsymbol{\xi}\boldsymbol{\psi}} = |F R_{\boldsymbol{\xi}\boldsymbol{\psi}}|^{\cdot} = |(F\boldsymbol{\xi})^* \odot (F\boldsymbol{\psi})|^{\cdot}, \tag{10}$$

where $\mathbf{a}^*$ represents the element-wise conjugate operation for a vector $\mathbf{a}$, $\odot$ represents the element-wise multiplication, and $|\mathbf{a}|^{\cdot}$ represents the element-wise operation that takes the absolute values of the elements of a vector $\mathbf{a}$. The entropy of the spectrum $S_{\boldsymbol{\xi}\boldsymbol{\psi}}$ is given by

$$H_{\boldsymbol{\xi}\boldsymbol{\xi}} = -\sum_{t=1}^{T} \bar{S}_{\boldsymbol{\xi}\boldsymbol{\psi}}(t) \log \bar{S}_{\boldsymbol{\xi}\boldsymbol{\psi}}(t), \tag{11}$$

$$\bar{S}_{\boldsymbol{\xi}\boldsymbol{\psi}} = \mathrm{Normalize}(S_{\boldsymbol{\xi}\boldsymbol{\psi}}). \tag{12}$$

If the probability sequences $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$ are consistent in terms of periodicity, i.e., they have synchronous periodic patterns, the cross-correlation function and the cross-spectrum have equally-spaced sharp peaks. The negative entropy $-H_{\boldsymbol{\xi}\boldsymbol{\psi}}$ can be used as an indicator of consistency.

3) Element-Wise Regularization: We propose element-wise regularization terms, denoted by $\mathcal{L}^{\mathrm{D}}_{\mathrm{ACE}}$, $\mathcal{L}^{\mathrm{B}}_{\mathrm{ACE}}$, and $\mathcal{L}^{\mathrm{W}}_{\mathrm{ACE}}$, based on the auto-correlation entropy (ACE) of the drum onset probability sequence $\boldsymbol{\phi}^{\mathrm{D}}$, that of the beat probability sequence $\boldsymbol{\phi}^{\mathrm{B}}$, and that of the downbeat probability sequence $\boldsymbol{\phi}^{\mathrm{W}}$:

$$\mathcal{L}^{\mathrm{D}}_{\mathrm{ACE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T} \bar{\phi}^{\mathrm{D}}_{kt} \log \bar{\phi}^{\mathrm{D}}_{kt}, \tag{13}$$

$$\bar{\boldsymbol{\phi}}^{\mathrm{D}}_{k} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{D}}_{k}|^{\cdot 2}), \tag{14}$$

$$\mathcal{L}^{\mathrm{B}}_{\mathrm{ACE}} = -\sum_{t=1}^{T} \bar{\phi}^{\mathrm{B}}_{t} \log \bar{\phi}^{\mathrm{B}}_{t}, \tag{15}$$

$$\bar{\boldsymbol{\phi}}^{\mathrm{B}} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{B}}|^{\cdot 2}), \tag{16}$$

$$\mathcal{L}^{\mathrm{W}}_{\mathrm{ACE}} = -\sum_{t=1}^{T} \bar{\phi}^{\mathrm{W}}_{t} \log \bar{\phi}^{\mathrm{W}}_{t}, \tag{17}$$

$$\bar{\boldsymbol{\phi}}^{\mathrm{W}} = \mathrm{Normalize}(|F\boldsymbol{\phi}^{\mathrm{W}}|^{\cdot 2}), \tag{18}$$

where $\boldsymbol{\phi}^{\mathrm{D}}_{k} \in [0,1]^T$ denotes the drum onset probability sequence for drum class k.

We aim to minimize the weighted sum of the individual regularization terms:

$$\mathcal{L}_{\text{ACE}} = \lambda^{\text{D}}_{\text{ACE}}\mathcal{L}^{\text{D}}_{\text{ACE}} + \lambda^{\text{B}}_{\text{ACE}}\mathcal{L}^{\text{B}}_{\text{ACE}} + \lambda^{\text{W}}_{\text{ACE}}\mathcal{L}^{\text{W}}_{\text{ACE}}, \tag{19}$$

where $\lambda^{\text{D}}_{\text{ACE}}$, $\lambda^{\text{B}}_{\text{ACE}}$, and $\lambda^{\text{W}}_{\text{ACE}}$ are adjustable weights.

4) Pairwise Regularization: We propose pairwise regularization terms, denoted by $\mathcal{L}^{\text{DB}}_{\text{CCE}}$, $\mathcal{L}^{\text{DW}}_{\text{CCE}}$, and $\mathcal{L}^{\text{BW}}_{\text{CCE}}$, based on the cross-correlation entropy (CCE) for each of the possible pairs between the drum onset probability sequence $\boldsymbol{\phi}^{\text{D}}$, the beat probability sequence $\boldsymbol{\phi}^{\text{B}}$, and the downbeat probability sequence $\boldsymbol{\phi}^{\text{W}}$:

$$\mathcal{L}^{\text{DB}}_{\text{CCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T} \bar{\phi}^{\text{DB}}_{kt} \log \bar{\phi}^{\text{DB}}_{kt}, \tag{20}$$

$$\bar{\boldsymbol{\phi}}^{\text{DB}}_{k} = \text{Normalize}(|F\boldsymbol{\phi}^{\text{D}}_{k}|^{\cdot} \odot |F\boldsymbol{\phi}^{\text{B}}|^{\cdot}), \tag{21}$$

$$\mathcal{L}^{\text{DW}}_{\text{CCE}} = -\frac{1}{K}\sum_{k,t=1}^{K,T} \bar{\phi}^{\text{DW}}_{kt} \log \bar{\phi}^{\text{DW}}_{kt}, \tag{22}$$

$$\bar{\boldsymbol{\phi}}^{\text{DW}}_{k} = \text{Normalize}(|F\boldsymbol{\phi}^{\text{D}}_{k}|^{\cdot} \odot |F\boldsymbol{\phi}^{\text{W}}|^{\cdot}), \tag{23}$$

$$\mathcal{L}^{\text{BW}}_{\text{CCE}} = -\sum_{t=1}^{T} \bar{\phi}^{\text{BW}}_{t} \log \bar{\phi}^{\text{BW}}_{t}, \tag{24}$$

$$\bar{\boldsymbol{\phi}}^{\text{BW}} = \text{Normalize}(|F\boldsymbol{\phi}^{\text{B}}|^{\cdot} \odot |F\boldsymbol{\phi}^{\text{W}}|^{\cdot}). \tag{25}$$

We aim to minimize the weighted sum of the individual regularization terms:

$$\mathcal{L}_{\text{CCE}} = \lambda^{\text{DB}}_{\text{CCE}}\mathcal{L}^{\text{DB}}_{\text{CCE}} + \lambda^{\text{DW}}_{\text{CCE}}\mathcal{L}^{\text{DW}}_{\text{CCE}} + \lambda^{\text{BW}}_{\text{CCE}}\mathcal{L}^{\text{BW}}_{\text{CCE}}, \tag{26}$$

where $\lambda^{\text{DB}}_{\text{CCE}}$, $\lambda^{\text{DW}}_{\text{CCE}}$, and $\lambda^{\text{BW}}_{\text{CCE}}$ are adjustable weights.

5) Total Loss: For regularized joint training, we aim to minimize the sum of the basic supervised loss $\mathcal{L}_{\text{BCE}}$ in (4), the element-wise regularization term $\mathcal{L}_{\text{ACE}}$ in (19), and the pairwise regularization term $\mathcal{L}_{\text{CCE}}$ in (26):

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{ACE}} + \mathcal{L}_{\text{CCE}}. \tag{27}$$

The weights can be adjusted for controlling the relative importance of different tasks.

D. Decoding

In the test phase, given $\mathbf{X}$, we jointly estimate the drum onset probability sequence $\boldsymbol{\phi}^{\text{D}}$, the beat probability sequence $\boldsymbol{\phi}^{\text{B}}$, and the downbeat probability sequence $\boldsymbol{\phi}^{\text{W}}$ using the trained CRNN. For improved performance, we use different decoding techniques for detecting the drum onsets $\mathbf{Y}^{\text{D}}$ and the beat and downbeat times $\mathbf{Y}^{\text{B}}$ and $\mathbf{Y}^{\text{W}}$.

1) Drum Transcription: The final output $\mathbf{Y}^{\text{D}}$ is obtained by performing peak-picking and thresholding for $\boldsymbol{\phi}^{\text{D}}$ as follows:

$$Y^{\text{D}}_{kt} = \begin{cases} 1 & (\phi^{\text{D}}_{kt} \geq \delta_k) \\ 0 & (\text{otherwise}) \end{cases}, \tag{28}$$

where δ_k is a threshold configured for each drum instrument.

2) Metrical Analysis: Since the naive thresholding has a limitation in performance, the final outputs $\mathbf{Y}^{\text{B}}$ and $\mathbf{Y}^{\text{W}}$ are detected from from $\boldsymbol{\phi}^{\text{B}}$ and $\boldsymbol{\phi}^{\text{W}}$ with a metrical analysis method based on a dynamic Bayesian network (DBN) [21].

We explain the decoding method for the beat probability sequence $\boldsymbol{\phi}^{\text{B}}$. The DBN is a hidden Markov model (HMM) with hidden states $\mathbf{z}_t = [c_t, \dot{c}_t]$ at frame t. Here, $c_t \in \{1, 2, \ldots, \dot{c}\}$ is a discrete random variable representing the position within a beat at frame t, and $\dot{c}_t \in \{\dot{c}_{\text{min}}, \dot{c}_{\text{min}}+1, \ldots, \dot{c}_{\text{max}}\}$ is the total number of discretized positions within one beat, which is a random variable representing the tempo at frame t, where $\dot{c}_{\text{min}}$ and $\dot{c}_{\text{max}}$ denote the minimum and maximum values of the tempo, respectively. Let Δ be the frame length of the audio signal and BPM_{t} represent the tempo in beats per minute at frame t. Then, $\dot{c}_t$ can be expressed as follows:

$$\dot{c}_t = \text{round}\left(\frac{4 \times 60}{\text{BPM}_{\text{t}} * \Delta}\right). \tag{29}$$

If we denote the observation sequence as $\{\mathbf{o}_t\}_{t=1}^{T}$, the probability model can be defined as follows:

$$p(\mathbf{z}_{1:T}, \mathbf{o}_{1:T}) = p(\mathbf{z}_{1:T})p(\mathbf{o}_{1:T}|\mathbf{z}_{1:T}). \tag{30}$$

Our goal is to estimate the hidden state sequence $\mathbf{z}^{*}_{1:T}$ with the maximum posterior probability as follows:

$$\mathbf{z}^{*}_{1:T} = \underset{z_{1:T}}{\text{argmax}}\, p(\mathbf{z}_{1:T}|\mathbf{o}_{1:T}), \tag{31}$$

$$p(\mathbf{o}_{1:T}|\mathbf{z}_{1:T}) \propto p(\mathbf{z}_1)\prod_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{o}_t|\mathbf{z}_t), \tag{32}$$

where $p(\mathbf{o}_t|\mathbf{z}_t)$ represents the observation model, and $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ represents the transition model. Equation (31) can be solved efficiently using the Viterbi algorithm [22]. The term $p(\mathbf{z}_{1:T})$ in (30) represents the generative process of the hidden states $\mathbf{z}_{1:T}$, which is given by

$$p(\mathbf{z}_{1:T}) = p(\mathbf{z}_1)\prod_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1}), \tag{33}$$

where $p(\mathbf{z}_1)$ represents the initial probability and $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ represents the transition probability given by

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = p(c_t|c_{t-1}, \dot{c}_{t-1})p(\dot{c}_t|\dot{c}_{t-1}), \tag{34}$$

$$p(c_t|c_{t-1}, \dot{c}_{t-1}) = \mathbb{1}_z. \tag{35}$$

where $\mathbb{1}_z$ is a function that takes 1 if $c_t = (c_{t-1}+1) \bmod \dot{c}_{t-1}$ and 0 otherwise. If $c_{t-1} = \dot{c}_{t-1}$, we have

$$p(\dot{c}_t|\dot{c}_{t-1}) = \exp\left(-\lambda\left|\frac{\dot{c}_t}{\dot{c}_{t-1}} - 1\right|\right), \tag{36}$$

where λ is a hyperparameter. Otherwise, we have

$$p(\dot{c}_t|\dot{c}_{t-1}) = \begin{cases} 1 & (c_t = c_{t-1} + 1) \\ 0 & \text{otherwise} \end{cases}. \tag{37}$$

Finally, for the hidden state sequence $\mathbf{x}^{*}_{1:T}$ with the maximum posterior probability, we detect all t's where $c_t = 1$. This sequence is the estimated beat time sequence.

TABLE I
THE F-MEASURES OF DRUM TRANSCRIPTION AND METRICAL ANALYSIS. VALUES IN BOLD ARE WITHIN 0.5 PTS OF THE BEST VALUE.

Multi-task Learning			Regularized Training		Drum Transcription				Metrical Analysis	
D	B	W	ACE	CCE	BD	SD	HH	Ave.	B	W
	✓	✓			-	-	-	-	90.5	87.3
✓					**78.0**	71.2	74.2	74.5	-	-
✓			✓		77.0	70.1	74.7	73.9	-	-
✓	✓				76.6	69.6	76.5	74.2	90.6	-
✓	✓		✓		76.6	71.2	75.2	74.3	90.3	-
✓	✓			✓	77.0	71.3	75.1	74.5	88.9	-
✓	✓		✓	✓	**77.6**	**73.5**	75.7	**75.6**	89.5	-
✓	✓	✓			75.1	68.7	72.4	72.1	93.9	**89.2**
✓	✓	✓	✓		77.1	74.1	73.7	75.0	**94.6**	**89.1**
✓	✓	✓		✓	74.2	70.7	76.4	73.8	**94.1**	87.2
✓	✓	✓	✓	✓	76.8	72.2	**77.1**	**75.4**	92.1	86.3

IV. EVALUATION

This section reports a comparative experiment conducted for evaluating the proposed regularized multi-task learning method through ablation study.

A. Experimental Conditions

The RWC Music Database: Popular Music [23] was used for evaluation. It consists of 100 Japanese popular songs (J-POP) with human-performed and synthesized drum tracks. We used 64 songs with accurate drum onset annotations and 10 songs without drum annotations. The stereo signals at a sampling rate of 44.1 kHz were processed using short-time Fourier transform (STFT) with a window size of 1024 points and a hop size of 441 points. The left and right channels were concatenated to form the input matrix $\mathbf{X}$.

Our CRNN consists of common convolutional layers for feature extraction followed by a bidirectional long short-term memory (BLSTM) network for drum transcription and a TCN for metrical analysis (Fig. 2). The convolutional layers, with a kernel size of 3×3, a padding size of 1×1, and a stride of 1, yielded a (512×4)-dimensional feature map, on which a dropout of 30% was applied before feeding it into a linear layer. The BLSTM network had three layers, each of which had hidden states of 200 dimensions. The TCN had eleven layers with a receptive field of about 80 [s]. The weights were set as follows: $\lambda^{\mathrm{D}}_{\mathrm{BCE}}=1$, $\lambda^{\mathrm{B}}_{\mathrm{BCE}}=\lambda^{\mathrm{W}}_{\mathrm{BCE}}=0.1$, $\lambda^{\mathrm{D}}_{\mathrm{ACE}}=1$, $\lambda^{\mathrm{B}}_{\mathrm{ACE}}=\lambda^{\mathrm{W}}_{\mathrm{ACE}}=0$, and $\lambda^{\mathrm{DB}}_{\mathrm{CCE}}=\lambda^{\mathrm{DW}}_{\mathrm{CCE}}=\lambda^{\mathrm{BW}}_{\mathrm{CCE}}=1$. We used AdamW [24] with a learning rate of $\gamma = 0.001$, weight decays parameters of $\lambda = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\varepsilon = 10^{-9}$ for optimization.

To evaluate the effectiveness of the individual components of the proposed method, we conducted an ablation study. We tackled drum transcription and metrical analysis jointly or separately with or without the element-wise and pairwise regularization terms. Five-fold cross-validation was conducted, where the performance was evaluated on the 64 songs with drum annotations. We also compared the proposed method with a state-of-the-art method [20].

The performances of drum transcription and metrical analysis were evaluated in terms of the F-measure with error tolerances of 30 [ms] and 70 [ms], respectively. The beat interval stability was also evaluated in terms of the CMLt and AMLt metrics [25] that consider double/half tempo ambiguity in beat interpretation. Note that the drum onset annotations might be inconsistent due to the characteristics of annotators. In the training phase, the ground-truth onset times were perturbed by being convolved with a Gaussian distribution with a mean of 0 [ms] and a standard deviation of 12 [ms]. For data augmentation, the time stretch operation was applied to each song with a scaling factor drawn from a Gaussian distribution with a mean of 1 and a standard deviation of 0.1.

TABLE II
THE PERFORMANCES OF METRICAL ANALYSIS FOR 100 SONGS OF RWC POPULAR MUSIC DATABASE. BOLD FONTS INDICATE THE BEST VALUES.

	Beat			Downbeat		
	F-meas.	CMLt	AMLt	F-meas.	CMLt	AMLt
Böck et al. [20]	89.5	81.8	91.5	83.1	80.3	88.4
Ours	**92.8**	**85.9**	**94.3**	**86.3**	**81.3**	**89.4**

B. Experimental Results

As shown in Table I, the multi-task learning approach improved the beat and downbeat detection performances by approximately 4 and 2 pts, respectively, compared with the single-task learning approach. Both the elementwise and pairwise regularizations improved the performance of drum transcription by approximately 1–3 pts. For metrical analysis, in contrast, no significant improvement was observed, mainly because the DBN-based decoding had a dominant impact on the results. If the naive thresholding was used instead, some performance gain was obtained (not reported in the table). As shown in Table II, the configuration that achieved the best performance of beat detection for all the 100 songs of the RWC Popular Music Database (the ninth row in Table I) outperformed the state-of-the-art method [20] by approximately 3 pts higher.

Fig. 4 shows the positive and negative effects of the multi-task learning. In the left example, since the drum onsets were located on the beat and tatum grids, the snare drum onsets on unnatural positions were suppressed successfully. In contrast, in the right example, although the drum onsets with swing rhythm were deviated from the beat and tatum grids, they were encouraged to be synchronized with the grids.

Fig. 5 shows the positive effect of the regularized training. The histogram of the inter-onset intervals (IOIs) of the hi-hat onsets had a single sharp peak, meaning that the estimated onsets were placed at a constant interval. Such periodicity-aware regularization, however, might fail to estimate non-regular drum patterns (e.g., fill-ins) and degrade the performance.

These results suggest the importance of the weight configuration according to the target data. One promising solution worth investigation would be to use the self-attention mechanism for automatically adjusting the weights according to the regularity of drum patterns.

V. CONCLUSION

In this paper, we presented a multi-task learning approach to joint drum transcription and metrical analysis based on the periodicity-aware element-wise and pairwise regularization terms. Specifically, our method aims to minimize the entropy

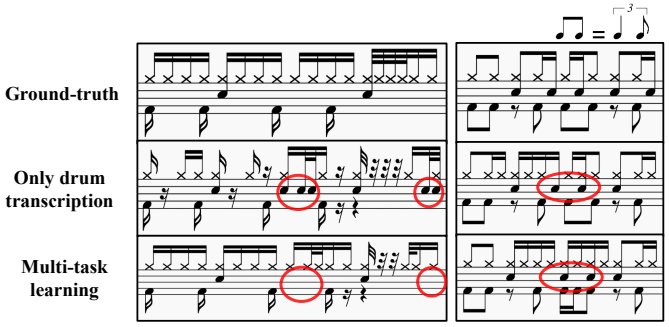


Fig. 4. Examples of drum transcription results obtained with the multi-task and single-task learning approaches. The performance was improved in the left example and was degraded in the right example.

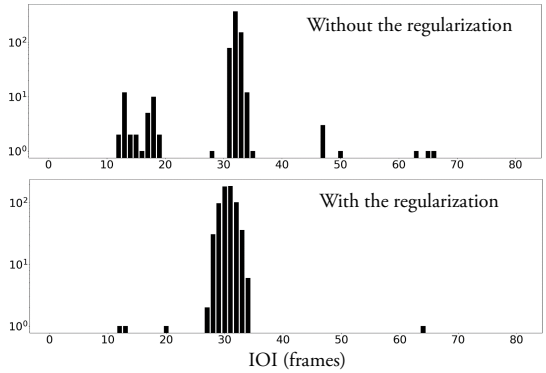


Fig. 5. Histograms of inter-onset intervals (IOIs) in frame units computed from hi-hat onsets detected without/with the regularization (RWC-MDB-P-2001 No. 11).

of the auto-spectrum computed from each probability sequence and that of the cross-spectrum computed from each of the three sequence pairs, because a smaller entropy indicates a stronger element-wise periodicity or pairwise consistency. We experimentally showed the effectiveness of using the multi-task learning and the regularized training for drum transcription. The performance of metrical analysis, however, was not maximized by the concurrent use of these methods. Further investigation is thus needed to determine the weights of the loss functions appropriately for each task, compare with other deep learning methods, and explore a peak detection method that maximizes the performance of the proposed method. In addition, we plan to investigate a new regularization term based on the fractal structure inherent in music.

Acknowledgment

This study was partially supported by JSPS KAKENHI Nos. 21K12187, 21K02846, 22H03661, 20H00602, and 21H03572, JST PRESTO No. JPMJPR20CB, and JST FOREST No. JPMJPR226X.

References

[1] C. W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Müller, and A. Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 9, pp. 1457–1483, 2018.

[2] R. Vogl, M. Dorfer, G. Widmer, and P. Knees. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 150–157. IEEE, 2017.

[3] C. Jacques and A. Roebel. Automatic drum transcription with convolutional neural networks. In *International Conference on Digital Audio Effects (DAFx)*, pp. 80–86, 2018.

[4] E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, Vol. 133, No. 3, pp. 1727–1741, 2013.

[5] C. Jacques and A. Roebel. Data augmentation for drum transcription with convolutional neural networks. In *European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

[6] R. Stables, J. Hockman, and C. Southall. Automatic drum transcription using bi-directional recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 591–597, 2016.

[7] S. Ueda, K. Shibata, Y. Wada, R. Nishikimi, E. Nakamura, and K. Yoshii. Drum transcription using convolutional non-negative matrix factorization based on deep drum score prior distribution (in japanese). *IPSJ SIG technical reports (EC)*, Vol. 2019, No. 26, pp. 1–6, 2019.

[8] R. Ishizuka, R. Nishikimi, E. Nakamura, and K. Yoshii. Tatum-level drum transcription based on a convolutional recurrent neural network with language model-based regularized training. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020.

[9] Y. Wang, J. Salamon, M. Cartwright, Nicholas J. Bryan, and J. P. Bello. Few-shot drum transcription in polyphonic music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 117–124, 2020.

[10] R. Ishizuka, R. Nishikimi, and K. Yoshii. Global structure-aware drum transcription based on self-attention mechanisms. *Signals*, Vol. 2, No. 3, pp. 508–526, 2021.

[11] C. Lea, R. Vidal, A. Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision (ECCV) Workshop*, pp. 47–54, 2016.

[12] S. Bai, J. Zico Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In *arXiv preprint arXiv:1803.01271*, 2018.

[13] M. E. P. Daveis and S. Böck. Temporal convolutional networks for musical audio beat tracking. In *European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

[14] M. Cartwright and J. P. Bello. Increasing drum transcription vocabulary using data synthesis. In *International Conference on Digital Audio Effects (DAFx)*, pp. 72–79, 2018.

[15] K. Choi and K. Cho. Deep unsupervised drum transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 183–191, 2019.

[16] H. Kawata, T. Hori, and K. Nakamura. Automatic drum transcription based on RNN considering phase information (in japanese). *IPSJ SIG technical reports (MUS)*, No. 27, pp. 1–4, 2019.

[17] T. Oyama, R. Ishizuka, and K. Yoshii. Phase-aware joint beat and downbeat estimation based on periodicity of metrical structure. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–499, 2021.

[18] S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 255–261, 2016.

[19] S. Böck, M. E. P. Davies, and P. Knees. Multi-task learning of tempo and beat: Learning one to improve the other. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 486–493, 2019.

[20] S. Böck and M. E. P. Daveis. Deconstruct, analysis, reconstruct: How to improve tempo, beat, and downbeat estimation. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 574–582, 2020.

[21] F. Krebs, S. Böck, and G. Widmer. An efficient state-space model for joint tempo and meter tracking. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 72–78, 2015.

[22] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE.*, pp. 77(2):257–286, 1989.

[23] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, classical and jazz music databases. In *International*

Society for Music Information Retrieval Conference (ISMIR), pp. 287–288, 2002.

[24] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, pp. 1–8, 2017.

[25] M. E. P. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical report, Centre for Digital Music, Queen Mary University of London, 2009.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

CTC2: End-to-End Drum Transcription Based on Connectionist Temporal Classification With Constant Tempo Constraint

Daichi Kamakura, Eita Nakamura, and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Japan
E-mail: {kamakura, enakamura}@sap.ist.i.kyoto-u.ac.jp, yoshii@i.kyoto-u.ac.jp

***Abstract*—This paper describes end-to-end automatic drum transcription for directly estimating a drum score from an audio signal of popular music using non-aligned paired data. We aim to convert a sequence of frame-level acoustic features into a sequence of tatum-level score fragments (three-dimensional multi-hot vectors) representing the presence or absence of the onsets of the bass and snare drums and the hi-hats. The main challenge of this task lies in estimating the correct number of inactive tatums having no onset between active tatums. One may use the connectionist temporal classification (CTC) for end-to-end training of a deep neural network (DNN) that infers a frame-level state sequence (alignment path) including the special "blank" states representing the tatum boundaries. At run-time, a drum score is obtained by annexing repeated states and removing all blank states from the most likely frame-level state sequence. This approach, however, tends to yield a shortened drum score in which repeated inactive tatums are annexed mistakenly because the blank state (tatum boundary) cannot be distinguished acoustically from the inactive state (onset absence) at the frame level. In this paper, we propose a sophisticated version of the CTC with constant tempo constraint, CTC2 in short, that encourages each tatum to be aligned with almost the same number of frames. Although the loss function can be computed efficiently as in the basic CTC, the backpropagation over the huge computation graph made through the forward algorithm is computationally prohibitive. To solve this problem, we propose to perform the backpropagation with only an alignment path stochastically drawn with Gibbs sampling. The experiment showed that the proposed method worked well as expected.**

I. Introduction

Automatic drum transcription (ADT) is a fundamental task that aims to estimate a drum score (MusicXML format) from a music signal. Although ADT plays a key role in computational music understanding, it has been tackled only partially; most studies aim to estimate a piano roll (MIDI format) by detecting the onset times of drums in seconds [1], [2]. To estimate a drum score, for example, one may estimate the beat and downbeat times in advance [3], [4] and then perform quantized onset detection [5]. Instead, one may take a more sophisticated approach in the same way as the state-of-the-art automatic piano transcription [6] based on the combination of audio-to-MIDI transcription [7] and MIDI-to-score transcription (rhythm transcription) [8]. Such a cascading approach, however, suffers from the error propagation problem.

Inspired by the great success of end-to-end automatic speech recognition (ASR), end-to-end automatic music transcription (AMT) has recently been investigated [9]–[13]. This approach can circumvent the error propagation problem and make effective use of non-aligned pairs of music signals and the corresponding scores as training data. Considering the monotonic nature of the audio-to-score mapping, we focus on the connectionist temporal classification (CTC) [14] (cf. [9]) because it is expected to work stably with a limited amount of training data, compared with the encoder-decoder model (cf. [10]) and that with the attention mechanism [15] (cf. [11]–[13]).

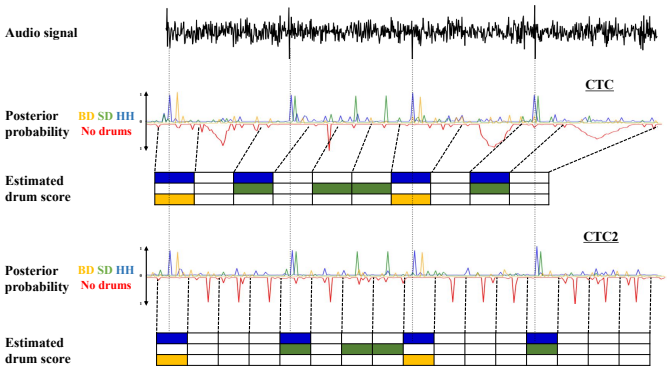


Fig. 1. The comparison of the basic CTC and the proposed CTC with constant tempo constraint (CTC2). The repeated inactive tatums having no drum onsets can be estimated correctly by encouraging the frames to be segmented into the tatums with almost the same length.

End-to-end AMT methods can also be characterized by their score representations: how to define symbols constituting the output sequence. The representation affects the difficulty unique to AMT in estimating the temporal information about note values (quantized durations) and metrical structures. The basic way is to convert a sequence of frame-level acoustic features into a sequence of *notation-level score components* (e.g., notes and bars) [12]. The note values (temporal attribute), however, are considerably harder to estimate than the note pitches (instantaneous attribute) with the basic input-output alignment mechanism. Another way is to estimate a sequence of *tatum-level score fragments* [11], where the tatum is a basic time unit on the score at the sub-beat level (typically at the sixteenth-note level). Instead of directly estimating the note values, this approach aims to estimate the presence or absence of the note onsets (instantaneous attribute) at the tatum level.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In this paper we tackle CTC-based ADT that aims to estimate a sequence of tatum-level score fragments (Fig. 1). Each fragment (tatum) is represented as a three-dimensional multi-hot vector indicating the presence or absence of the onsets of the bass and snare drums and the hi-hats (eight states in total). The main challenge unique to this task is to estimate the correct number of inactive tatums (a state corresponding to the all-zero vector) between active tatums (the other seven states). If the basic CTC is used for end-to-end training, the state of each tatum of the ground-truth sequence is associated with frames of the input sequence whose acoustic features are particularly relevant to the state through forced alignment. Such frames are typically only at the beginning of the tatum; the special blank state (denoted by "_") is associated with the other acoustically irrelevant frames. At run-time, however, successive inactive tatums are annexed mistakenly because the blank state (tatum boundary) and the inactive state (onset absence) are hard to distinguish acoustically at the frame level.

To solve this problem, we propose a sophisticated version of the CTC with constant tempo constraint, CTC2 in short, based on the reasonable assumption that the tempo is usually kept constant throughout the song in popular music (Fig. 1). In the basic CTC, the frame-to-tatum alignment is obtained through unconstrained time stretch because only the order of output symbols (states) matters and their durations are ignored, as in the inference of the latent state sequence of the hidden Markov model (HMM). Inspired by the hidden semi-Markov model (HSMM) with explicit duration modeling [16], we instead encourage the tatum boundaries (transitions between the main eight states and the blank state) to occur regularly with a constant interval, where adjacent tatums favor to have the same duration [3], [17]. This makes successive inactive frames split into an appropriate number of successive inactive tatums at the frames associated with the blank state.

We also propose an efficient training method that works with the CTC and its variants. The objective function of the CTC2 to be maximized is given by the sum of the posterior probabilities of all possible alignment paths as in the CTC. Although it can thus be computed efficiently with dynamic programming (forward algorithm) in the analogy of the HMM to the HSMM, the backpropagation over the huge computation graph is computationally prohibitive. We thus perform the backpropagation along only an alignment path stochastically drawn according to the posterior probability with Gibbs sampling. Technically, we run the forward filtering-backward sampling in the evaluation mode to draw a path and then make a compact computation graph along the path in the training mode.

II. Related Work

This section reviews related work on ADT and that on end-to-end AMT.

A. Automatic Drum Transcription

Deep neural networks (DNNs) have intensively been used for ADT [1], [2], [5], [18]–[22]. In general, the spectrogram of an audio signal is used as input, and the annotated drum onset times are used as training data. For feature extraction, convolutional neural networks (CNNs) have commonly been used. Regularization based on prior information about drum scores [5] and refined network architectures [23] have also been attempted, resulting in improved performance of transcription. For converting DNN outputs into musical scores, we use beat information for quantization, and it has been reported that using temporal convolutional networks (TCNs) [24], [25] for beat tracking is effective. However, one problem with these methods is that they require as training data a sufficient amount of music spectrograms with precise annotations of drum onset times. To address this issue, the use of synthetic datasets [26], data augmentation [19], and unsupervised learning [27] have been proposed. Nevertheless, these approaches still face challenges in terms of robustness against real performances and the ability to handle variations in drum sounds.

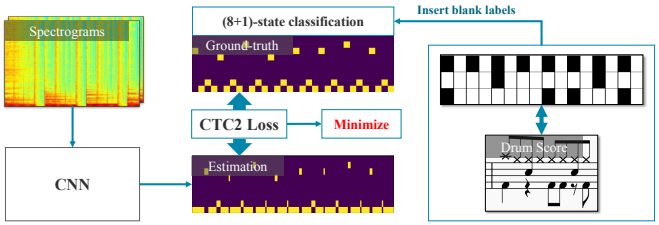


Fig. 2. The proposed ADT method based on the CTC2.

B. End-to-End Automatic Music Transcription

For end-to-end AMT that aims to directly estimate a musical score (a sequence of musical notations on the score), the attention mechanism [12] and the CTC [28], [29] have been investigated. In singing transcription based on the attention mechanism, the tempo consistency was taken into account [12], where the frame-to-tatum alignment is encouraged to be monotonic and regular by imposing a regularization term on the attention matrix. Such regularization, however, was found to prevent the initial progress of the training.

III. Proposed Method

This section describes the audio-to-score ADT method based on the CTC2-based end-to-end training (Fig. 2).

A. Problem Specification

Our goal is to estimate a drum score from a sequence of the stereo power spectra of a target musical piece, $\mathbf{X} \triangleq \{\mathbf{x}_t\}_{t=1}^{T}$ ($\mathbf{x}_t \in \mathbb{R}^{2F}$), where F is the number of frequency bins and T is the number of frames. We aim to estimate the presence or absence of the onsets of I drum instruments at the tatum level. In this study, we focus on the three common drum instruments: bass drum (BD), snare drum (SD), and hi-hats (HH) ($I = 3$). Let $y_n \in \{1, \ldots, K\}$ be the state of tatum n, which is defined as $y_n = 1 + 2^0\langle \mathtt{BD} \rangle_n + 2^1\langle \mathtt{SD} \rangle_n + 2^2\langle \mathtt{HH} \rangle_n$, where $K = 2^I = 8$, and $\langle \mathtt{DR} \rangle_n$ is a binary value indicating the presence or absence of an onset of DR at the tatum. Let $\mathbf{Y} \triangleq \{y_n\}_{n=1}^{N}$, where N is the number of tatums. The tatum is defined as the one-fourth of the beat, i.e., the tatum and beat correspond to the durations of the sixteenth and quarter notes, respectively.

B. Training and Inference

We describe the basic flow of end-to-end ADT. In addition to the main states indexed by $\{1, 2, \dots, K\}$ (Section III-A), we introduce the special blank state indexed by 0. Let $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t=1}^{T}$ be a redundant state sequence, where $\pi_t \in \{0, \dots, K\}$ represents the state of frame t. We use a DNN with parameters θ that outputs the posterior probabilities of the $(K+1)$ states at the frame level, denoted by $\boldsymbol{\phi} \triangleq \{\phi_{k,t}\}_{k=0,t=1}^{K,T}$, where $\phi_{k,t}$ represents the probability of state k at frame t.

1) Training: Given a non-aligned pair of $\mathbf{X}$ and $\mathbf{Y}$ as training data, we train the DNN. Let $\mathcal{B}(\boldsymbol{\pi})$ be a one-to-one reducer that returns $\mathbf{Y}$ by annexing repeated states and removing all blank states from $\boldsymbol{\pi}$, e.g., $\mathcal{B}(01110112) = \mathcal{B}(01100122) = 112$. Let $\mathcal{B}^{-1}(\mathbf{Y})$ be a one-to-many expander that returns a set of all possible redundant state sequences that reduce to $\mathbf{Y}$, e.g., $\mathcal{B}^{-1}(112) = \{01110112, 1100122, \dots\}$. The DNN parameters θ are optimized such that the following posterior probability is maximized:

$$\begin{aligned} \mathcal{L} &= \log p(\mathbf{Y}|\mathbf{X}, \theta) \\ &= \log \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{Y})} p(\boldsymbol{\pi}|\mathbf{X}, \theta). \end{aligned} \tag{1}$$

2) Inference: Given $\mathbf{X}$, we estimate $\mathbf{Y}$ using a trained DNN with the parameters θ. The redundant state sequence $\boldsymbol{\pi}$ is obtained by selecting the most likely state at each frame:

$$\pi_t = \operatorname*{argmax}_k \phi_{k,t}. \tag{2}$$

The final output $\mathbf{Y}$ is given by $\mathbf{Y} = \mathcal{B}(\boldsymbol{\pi})$.

C. Conventional CTC

As a baseline, we briefly explain the training method based on the basic CTC. The posterior probability of an alignment path $\boldsymbol{\pi}$ in (1) is given by the product of the frame-wise posterior probabilities of the states as follows:

$$p(\boldsymbol{\pi}|\mathbf{X}, \theta) = \prod_{t=1}^{T} \phi_{\pi_t, t}. \tag{3}$$

The loss function $\mathcal{L}$ or $p(\mathbf{Y}|\mathbf{X}, \theta)$ in (1) can be computed efficiently with a dynamic programming technique called the forward algorithm as in the HMM. Let $\mathbf{Y}' \triangleq \{y'_s\}_{s=0}^{S}$ be an expanded state sequence obtained by inserting the $N-1$ blank states between the N states of $\mathbf{Y}$ and pushing the blank states to the front and back of $\mathbf{Y}$, i.e., $S = 2N+1$, $y'_{2n-1} = y_n$, and $y'_s = y_{\lfloor \frac{s}{2}+1 \rfloor}$. Let $\alpha_t(s)$ be the forward probability obtained by accumulating the posterior probabilities of all possible paths that align x_t with y'_s:

$$\alpha_t(s) = \sum_{\pi_{1:t} \in \mathcal{B}^{-1}(y_{1:\lfloor \frac{s}{2} \rfloor})} \prod_{\tau=1}^{t} \phi_{\pi_\tau, \tau}, \tag{4}$$

where $a_{1:t} \triangleq \{a_1, \dots, a_t\}$. Since an alignment path $\boldsymbol{\pi}$ is valid if the last frame x_T is aligned with the last non-blank symbol $y'_{S-1} = y_n$ or the last blank symbol y'_S, we have

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \alpha_T(S-1) + \alpha_T(S). \tag{5}$$

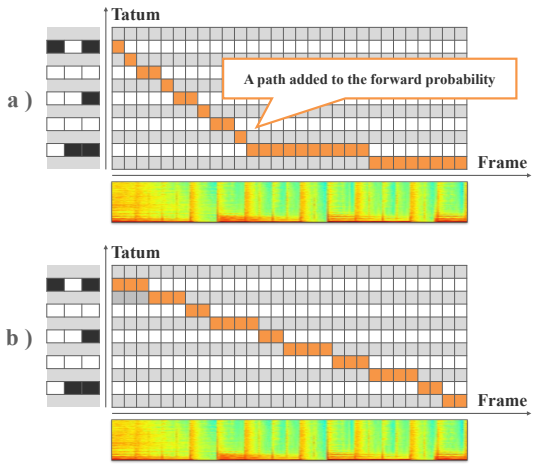


Fig. 3. Examples of the alignment path $\boldsymbol{\pi}$ estimated by a) the CTC- and b) CTC2-based training methods.

The forward probabilities can be computed recursively with the time complexity of $O(TS)$. First, we set the initial values:

$$\alpha_1(1) = \phi_{0,1}, \tag{6}$$

$$\alpha_1(2) = \phi_{y_1,1}, \tag{7}$$

$$\alpha_1(s) = 0, \forall s > 2. \tag{8}$$

We then use the following recursive formulas:

$$\alpha_t(s) = \begin{cases} \text{if } y'_s = \text{blank or } y'_{s-2} \\ \quad [\alpha_{t-1}(s) + \alpha_{t-1}(s-1)]\, \phi_{\pi_s, t}, \\ \text{else} \\ \quad [\alpha_{t-1}(s) + \alpha_{t-1}(s-1) + \alpha_{t-1}(s-2)]\, \phi_{\pi_s, t}. \end{cases} \tag{9}$$

In the inference stage, however, the optimal alignment path $\boldsymbol{\pi}$ estimated by (2) tends to deviate considerably from the diagonal line on the alignment map between the input sequence $\mathbf{X}$ and the estimated sequence $\mathbf{Y}'$ (Fig. 3-a). The durations of the N symbols in $\mathbf{Y} = \mathcal{B}(\boldsymbol{\pi})$ thus have a large variation, meaning that the tempo is allowed to frequently change at the tatum level in a musically-unnatural manner.

D. Training with Proposed CTC2

We then explain the training method based on the proposed CTC2 for explicit duration modeling with the constant tempo constraint (Fig. 4). Let $\mathbf{D} = \{d_n\}_{n=1}^{N}$ $(D_{\min} \le d_n \le D_{\max})$ denote the durations of the N symbols of $\mathbf{Y}$, where $D_{\min}$ and $D_{\max}$ are the minimum and maximum durations to be considered. When the frame shift is 10 [ms], the local tempo at tatum n is $1500/d_n$ [bpm]. Let $\mathcal{B}(\boldsymbol{\pi})$ be redefined as a one-to-one reducer that returns a pair $(\mathbf{Y}, \mathbf{D})$, e.g., $\mathcal{B}(01110112) = (112, 421)$, $\mathcal{B}(01100122) = (112, 412)$. Let $\mathcal{B}^{-1}(\mathbf{Y}, \mathbf{D})$ be redefined as a one-to-many expander that returns all possible sequences that reduce to $(\mathbf{Y}, \mathbf{D})$, e.g., $\mathcal{B}^{-1}(112, 412) = \{01110112, 1100122, \dots\}$. Note that even if both $\mathbf{Y}$ and $\mathbf{D}$ are given, $\mathcal{B}^{-1}(\mathbf{Y})$ may return multiple paths.

The posterior probability of $\mathbf{Y}$ in (1) is given by

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \sum_{\mathbf{D}} p(\mathbf{Y}|\mathbf{X}, \theta, \mathbf{D}) p(\mathbf{D}), \tag{10}$$

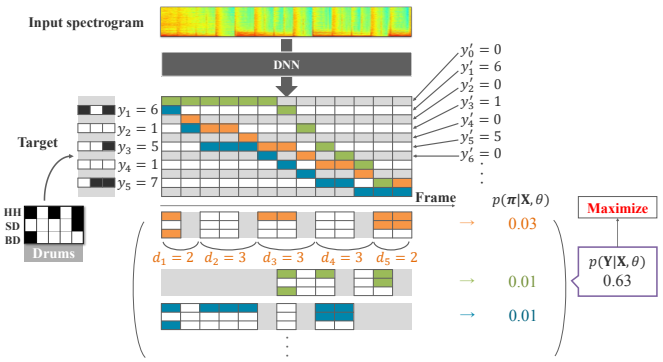


Fig. 4. The forward algorithm in the proposed CTC2-based training with explicit duration modeling.

where $p(\mathbf{Y}|\mathbf{X},\theta,\mathbf{D})$ is the posterior probability of $\mathbf{Y}$ conditioned by $\mathbf{D}$ and $p(\mathbf{D})$ is the generative probability of $\mathbf{D}$ given by a Markov model as follows:

$$p(\mathbf{D}) = p(d_1)\prod_{n=2}^{N} p(d_n|d_{n-1}), \tag{11}$$

$$p(d_1) \propto 1, \tag{12}$$

$$p(d_n|d_{n-1}) \propto \exp\left(-\lambda\left|\frac{d_n}{d_{n-1}} - 1\right|\right). \tag{13}$$

The loss function $\mathcal{L}$ or $p(\mathbf{Y}|\mathbf{X},\theta)$ in (1) can be computed efficiently as in the HSMM. Let $\alpha_t(s,d,c)$ be the forward probability obtained by accumulating the posterior probabilities of all possible paths that align x_t with y'_s:

$$\alpha_t(s,d,c) = \sum_{\pi_{1:t}\in\mathcal{B}^{-1}(y_{1:\lfloor\frac{s}{2}\rfloor},d_{1:\lfloor\frac{s}{2}\rfloor})} \prod_{\tau=1}^{t} \phi_{\pi_\tau,\tau}\cdot p(d_{1:\lfloor\frac{s}{2}\rfloor}), \tag{14}$$

$$p(d_{1:\lfloor\frac{s}{2}\rfloor}) = p(d_1)\prod_{n=2}^{\lfloor\frac{s}{2}\rfloor} p(d_n|d_{n-1}), \tag{15}$$

where d represents the duration of the current main state $y_{\lfloor\frac{s}{2}\rfloor}$ and c is a counter variable that is set to d when a new state starts and is then decremented until the state ends. We have

$$p(\mathbf{Y}|\mathbf{X},\theta) = \sum_{s=S-1}^{S}\sum_{d=D_{\min}}^{D_{\max}}\sum_{c=1}^{d}\alpha_T(s,d,c). \tag{16}$$

Note that if $\lambda = 0$ in (13), i.e., the durations of the output symbols are allowed to take any values with the same probability, i.e., $p(\mathbf{D}) = 1$, the CTC2 reduces to the CTC, i.e., (14) and (16) reduce to (4) and (5), respectively.

The forward probabilities can be computed recursively with the time complexity of $O(TSD^2)$, where $D \triangleq D_{\max} - D_{\min} + 1$. First, we set the initial values:

$$\alpha_1(1,d,c) = \phi_{0,1}, \forall d,c, \tag{17}$$

$$\alpha_1(2,d,c) = \phi_{y_1,1}, \forall d,c, \tag{18}$$

$$\alpha_1(s,d,c) = 0, \forall s > 2. \tag{19}$$

We then use the following recursive formulas:

$$\alpha_t(s,d,c) = \begin{cases} \text{if } d = c \\ \quad \text{if } y'_s = \text{blank} \\ \quad\quad 0, \\ \quad \text{else if } y'_s = y'_{s-2} \\ \quad\quad \sum_{d'}\alpha_{t-1}(s-1,d',1)\cdot\phi_{y'_s,t}\cdot p(d|d'), \\ \quad \text{else} \\ \quad\quad \sum_{d'}\bar{\alpha}_t(s,d,c)\cdot\phi_{y'_s,t}\cdot p(d|d'), \\ \text{else if } d > c \\ \quad \text{if } y'_s = \text{blank} \\ \quad\quad \bar{\bar{\alpha}}_t(s,d,c)\cdot\phi_{y'_s,t}, \\ \quad \text{else} \\ \quad\quad \alpha_{t-1}(s,d',c+1)\cdot\phi_{y'_s,t}, \\ \text{else} \\ \quad 0, \end{cases} \tag{20}$$

$$\bar{\alpha}_t(s,d,c) = \alpha_{t-1}(s-1,d,1) + \alpha_{t-1}(s-2,d,1), \tag{21}$$

$$\bar{\bar{\alpha}}_t(s,d,c) = \alpha_{t-1}(s,d,c+1) + \alpha_{t-1}(s-1,d,c+1). \tag{22}$$

In the inference stage, the optimal alignment path $\boldsymbol{\pi}$ estimated by (2) tends to roughly follow the diagonal line on the alignment map between the input sequence $\mathbf{X}$ and the estimated sequence $\mathbf{Y}'$ (Fig. 3-b). The durations $\mathbf{D}$ of the N symbols of $\mathbf{Y}$, which are given by $(\mathbf{Y},\mathbf{D}) = \mathcal{B}(\boldsymbol{\pi})$, are thus kept almost constant. This enables the active and inactive states (tatums) to have similar durations.

The CTC2, however, considers several orders of magnitude larger number of possible paths than the CTC and thus makes the exact backpropagation prohibitive in practice. Note that the forward computation in the inference mode without constructing the computation graph can still be performed quickly.

E. Computationally-Efficient Training

We propose an efficient training method that can be applied to the CTC2 (and the CTC) at a minimum sacrifice of performance. Instead of backpropagating the error through all possible paths on the huge computation graph constructed by the forward algorithm, we focus on only the most likely path found by the Viterbi algorithm or a path randomly selected according to its posterior probability with the Gibbs sampling. To determine such a path, the forward and backward recursions can be executed faster in the inference mode. Since the Gibbs sampling stochastically generates likely paths not limited to the most likely one, it is expected to be more robust against local optima than the Viterbi algorithm. Let $\mathbf{Z} \triangleq \{z_t\}_{t=1}^{T}$ be a latent state sequence, where $z_t \triangleq (s_t,d_t,c_t)$. An alignment path $\boldsymbol{\pi}$ is uniquely determined by $\mathbf{Z}$.

1) Viterbi Algorithm: In the forward recursion, the forward probabilities $\{\alpha_t(z_t)\}_{t=1}^{T}$ are computed as described in Section III-D except that the sum operations over d' of (20) and the additions of (21) and (22) are replaced with the max operations, where the indices and terms that take the maximum values are memorized. In the backward recursion, the states of most likely $\mathbf{Z}$ are determined in the reverse order by backtracing the memorized path from z_T such that $\alpha_T(z_T)$ is maximized, where $s_T \in \{S-1,S\}$, $d_T \in \{D_{\min},\ldots,D_{\max}\}$, and $c_T \in \{1,\ldots,d_T\}$.

2) Gibbs Sampling: We use the forward filtering-backward sampling algorithm, which was originally proposed for sampling the latent sequence of an HMM or HSMM [30]. In the forward filtering, the forward probabilities are computed as described in Section III-D. In the backward sampling, the states of $\mathbf{Z}$ are sampled in the reverse order according to

$$p(z_T) \propto \alpha_T(z_T), \tag{23}$$

$$p(z_t|z_{t+1:T}) \propto \alpha_t(z_t) A_{z_t z_{t+1}} \ (t = \{T-1, \ldots, 1\}), \tag{24}$$

where $A_{z_t z_{t+1}}$ represents the transition probability from z_t to z_{t+1} obtained by normalizing the following matrix:

$$A'_{z_t z_{t+1}} = \begin{cases} p(d_t|d_{t+1}) & \text{if condition 1 holds,} \\ 1 & \text{if condition 2 holds,} \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

condition 1: $c_t = 1$ and ($y'_s =$ blank and $s_{t+1} = s_t + 1$ or $y'_s \neq$ blank and $y'_s \neq y'_{s+2}$ and $s_{t+1} = s_t + 2$),

condition 2: $d_t > c_t$ and $d_t = d_{t+1}$ and $c = c_{t+1} + 1$ and ($y'_s =$ blank and $s_{t+1} = s_t + 1$ or $s_{t+1} = s_t$).

IV. Evaluation

This section reports a comparative experiment conducted for evaluating the performance of the proposed ADT method and the effectiveness of the efficient training.

A. Experimental Conditions

The 100 songs of the RWC Music Database: Popular Music [31] were randomly split into 60 and 40 songs for training and test data, respectively. The stereo signals of each song sampled at 44.1 kHz were analyzed by short-time Fourier transform (STFT) with a window size of 1024 pts ($F = 513$) and a hop size of 441 pts (10 [ms]). The left and right channels were concatenated to form the input data $\mathbf{X}$. The tempo was assumed to be between 50 and 250 [bpm], i.e., the duration of a tatum was between $D_{\min} = 6$ to $D_{\max} = 30$.

We trained a convolutional recurrent neural network (CRNN) with the basic CTC or the proposed CTC2. It has 11 convolutional layers with a kernel size of 3×3, a padding size of 1×1, and a stride of 1 yielding a (512×4)-dimensional feature map, on which a dropout of 30% was applied before feeding it into a linear layer, followed by three bi-directional long short-term memory (BLSTM) layers with 200-dimensional hidden states. We used AdamW [32] with a learning rate $\gamma = 0.001$, a weight decay parameters $\lambda = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\varepsilon = 10^{-9}$ for optimization. When the Viterbi algorithm or the Gibbs sampling (Section III-E) was used with the CTC, the basic training (Section III-C) was performed for initialization. For the CTC2, the basic training could not be tested due to the huge computational cost.

For comparison, we implemented to test a recent audio-to-MIDI transcription method based on multi-task learning that jointly detects drum onset times and beat times at the frame level [2]. This method was trained using music signals with frame-level onset annotations, whereas our method was trained end-to-end using music signals with *non-aligned* scores. For MIDI-to-score transcription, the detected drum onset times were quantized on the tatum grid made of the beat times estimated by an HMM-based beat tracker [3] as post-processing.

Our end-to-end transcription method with the CTC or CTC2 and the cascading method mentioned above [2], [3] were evaluated in terms of the tatum- and frame-level accuracies. For tatum-level evaluation of estimated *scores*, we used the tatum error rate (TER), an edit distance similar in spirit to the word error rate (WER) in speech recognition, as follows:

$$\mathrm{TER} = \frac{\#\mathrm{insersion} + \#\mathrm{deletion} + \#\mathrm{substitution}}{N}, \tag{26}$$

where $\#$insersion, $\#$deletion, $\#$substitution represent the numbers of insertion, deletion, and substitution errors, respectively, and N represents the total number of tatums in the ground-truth data. For frame-level evaluation of estimated *onset times*, we set the error tolerance to 70 [ms]. Note that the initial frames of the active states (tatums) determined by the estimated alignment path $\boldsymbol{\pi}$ were regarded as drum onset times for convenience in our end-to-end method. Since those onset times tend to have a constant offset from the ground-truth onset times, the best offset was found for each song such that the performance was maximized for fair comparison.

TABLE I
THE TATUM- AND FRAME-LEVEL EVALUATION RESULTS.

Method	Tatum error rate ↓			F-measure ↑		
	Basic	Viterbi	Gibbs	Basic	Viterbi	Gibbs
CTC [14]	34.4	35.1	35.6	79.4	78.6	76.1
CTC2 (ours)	-	**33.8**	**33.8**	-	79.2	79.0
Vogl et al. [2]		34.5			**83.1**	

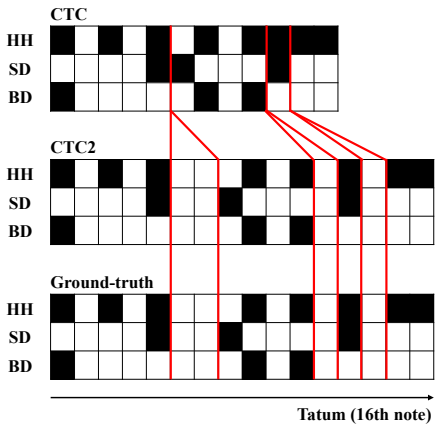


Fig. 5. Examples of tatum-level drum score fragments estimated by the CRNN trained with the basic CTC or the proposed CTC2. Whereas the CTC often failed to detect inactive states, the CTC2 correctly detected them.

B. Experimental Results

As shown in Table I, the CTC2 steadily outperformed the CTC in both metrics. Although the quantitative difference was small, the CTC2 yielded considerably better drum scores in the naturalness of rhythm despite the fact that the metrical structure was not considered. As shown in Fig. 5, the CTC2 successfully estimated a correct number of inactive tatums,

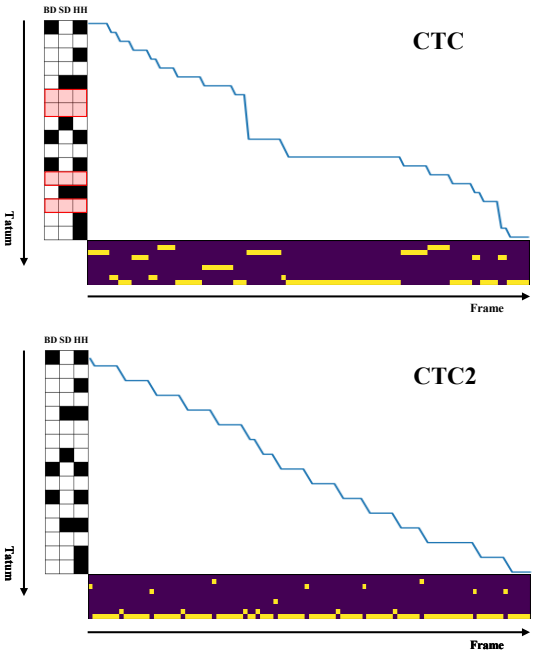


Fig. 6. Examples of alignment paths estimated by the CRNN trained with the basic CTC or the proposed CTC2 in the training phase. Whereas the CTC yielded a path with sudden tempo changes, the CTC2 yielded a diagonal path with an almost constant tempo.

which were missed by the CTC, between active tatums thanks to the constant tempo constraint. As shown in Fig. 6, in the training phase based on the forced alignment between the estimated and ground-truth state sequences, the CTC2 yielded an alignment path with an almost constant grade (tempo) around the diagonal line on the lattice. The CTC, in contrast, allowed an alignment path to discontinuously and frequently change the grade in a musically-unnatural manner. The Viterbi algorithm and the Gibbs sampling worked comparably in performance (Table I) and were several orders of magnitude faster than the basic backpropagation on the entire computation graph at a minimum sacrifice of performance (Fig. 7).

Compared with the conventional cascading method [2], our end-to-end method with the CTC2 performed slightly better by 0.7 pts in the TER, but worse by 6.0 pts in the F-measure. Note that the frame-level metric is advantageous for audio-to-MIDI transcription methods that use precise frame-level onset annotations for training. Considering that our method can be trained end-to-end with non-aligned audio-score pairs and does not aim at frame-level onset detection (the alignment path does not necessarily indicate precise onset times), this result is still considered to be promising.

V. Conclusion

In this paper, we proposed an end-to-end ADT method based on the CTC with constant tempo constraint, CTC2 in short, that estimates a sequence of tatum-level drum score fragments from a music signal. Although the loss function can be computed efficiently as in the vanilla CTC, the backpropagation

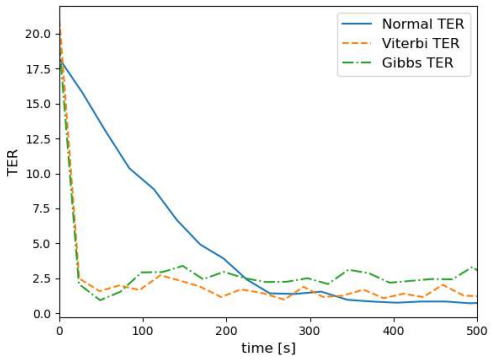


Fig. 7. Comparion of convergence speeds achieved by the basic training, the Viterbi algorithm, and the Gibbs learning.

over the huge computation graph made by the forward algorithm is computationally prohibitive. To solve this problem, we proposed a novel efficient training method that performs the backpropagation through only an alignment path found by the Viterbi algorithm or stochastically drawn with Gibbs sampling. Since non-aligned drum scores are easier to collect than labor-intensive frame-level onset annotations, the proposed end-to-end ADT method has a large potential for performance improvement based on large-scale training.

The CTC2 is a general technique for explicit duration modeling in a monotonic sequence-to-sequence mapping task not limited to ADT. It can be customized by formulating a duration model according to the task. In handwritten character recognition [33], for example, the CTC with constant "size" constraint would be useful for encouraging each character to have almost the same size. The Viterbi algorithm and the Gibbs sampling can also be applied to a wide range of tasks in various fields for accelerated CTC-based training. We are currently investigating how to effectively use the Viterbi algorithm in the inference phase as well as the training phase.

Acknowledgment

This study was partially supported by JSPS KAKENHI Nos. 21K12187, 21K02846, 22H03661, 20H00602, and 21H03572, JST PRESTO No. JPMJPR20CB, and JST FOREST No. JPMJPR226X.

References

[1] R. Stables, J. Hockman, and C. Southall. Automatic drum transcription using bi-directional recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[2] R. Vogl, M. Dorfer, G. Widmer, and P. Knees. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 150–157, 2017.

[3] F. Krebs, S. Böck, and G. Widmer. An efficient state-space model for joint tempo and meter tracking. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 72–78, 2015.

[4] S. Böck and M. E. P. Daveis. Deconstruct, analysis, reconstruct: How to improve tempo, beat, and downbeat estimation. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 574–582, 2020.

[5] R. Ishizuka, R. Nishikimi, E. Nakamura, and K. Yoshii. Tatum-level drum transcription based on a convolutional recurrent neural network with language model-based regularized training. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020.

[6] K. Shibata, E. Nakamura, and K. Yoshii. Non-local musical statistics as guides for audio-to-score piano transcription. *Information Sciences*, Vol. 566, pp. 262–280, 2021.

[7] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck. Onsets and frames: Dual-objective piano transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 50–57, 2018.

[8] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon. Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 101–105. IEEE, 2018.

[9] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza. An end-to-end framework for audio-to-score music transcription on monophonic excerpts. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 34–41, 2018.

[10] R. C. G. Carvalho and P. Smaragdis. Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 151–155. IEEE, 2017.

[11] R. Nishikimi, E. Nakamura, M. Goto, and K. Yoshii. End-to-end melody note transcription based on a beat-synchronous attention mechanism. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 26–30. IEEE, 2019.

[12] R. Nishikimi, E. Nakamura, S. Fukayama, M. Goto, and K. Yoshii. Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 161–165. IEEE, 2019.

[13] L. Liu, V. Morfi, and E. Benetos. Joint multi-pitch detection and score transcription for polyphonic piano music. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 281–285. IEEE, 2021.

[14] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International conference on Machine Learning (ICML)*, pp. 369–376, 2006.

[15] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, pp. 1–15, 2015.

[16] M. Russell and R. Moore. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 10, pp. 5–8. IEEE, 1985.

[17] E. Nakamura, K. Yoshii, and S. Sagayama. Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 4, pp. 794–806, 2017.

[18] C. Jacques and A. Roebel. Automatic drum transcription with convolutional neural networks. In *International Conference on Digital Audio Effects (DAFx)*, 2018.

[19] C. Jacques and A. Roebel. Data augmentation for drum transcription with convolutional neural networks. In *European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

[20] S. Ueda, K. Shibada, Y. Wada, R. Nishikimi, E. Nakamura, and K. Yoshii. Drum transcription using convolutional non-negative matrix factorization based on deep drum score prior distribution (in japanese). *IPSJ SIG technical reports (EC)*, Vol. 2019, No. 26, pp. 1–6, 2019.

[21] Y. Wang, J. Salamon, M. Cartwright, Nicholas J. Bryan, and J. P. Bello. Few-shot drum transcription in polyphonic music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 117–124, 2020.

[22] R. Ishizuka, R. Nishikimi, and K. Yoshii. Global structure-aware drum transcription based on self-attention mechanisms. *Signals*, Vol. 2, No. 3, pp. 508–526, 2021.

[23] D. Kamakura, T. Oyama, and K. Yoshii. Drum transcription and beat structure estimation based on multi-task learning (in japanese). *The 84th National Convention of IPSJ*, Vol. 2022, No. 1, pp. 517–518, 2022.

[24] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision (ECCV) Workshop*, pp. 47–54, 2016.

[25] M. E. P. Daveis and S. Böck. Temporal convolutional networks for musical audio beat tracking. In *European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

[26] M. Cartwright and J. P. Bello. Increasing drum transcription vocabulary using data synthesis. In *International Conference on Digital Audio Effects (DAFx)*, pp. 72–79, 2018.

[27] K. Choi and K. Cho. Deep unsupervised drum transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 183–191, 2019.

[28] R. G. C. Carvalho and P. Smaragdis. Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 151–155. IEEE, 2017.

[29] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza. An end-to-end framework for audio-to-score music transcription on monophonic excerpts. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 34–41, 2018.

[30] J. Van Gael, Y. Saatci, Yee Whye Y. W. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *International conference on Machine Learning (ICML)*, pp. 1088–1095, 2008.

[31] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, classical and jazz music databases. In *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 287–288, 2002.

[32] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, pp. 1–8, 2017.

[33] M. Ibrayim, W. Simayi, and A. Hamdulla. Unconstrained online handwritten uyghur word recognition based on recurrent neural networks and connectionist temporal classification. *International Journal of Biometrics*, Vol. 13, No. 1, pp. 51–63, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Learning Multifaceted Self-Similarity for Musical Structure Analysis

Tsung-Ping Chen*, Li Su†, and Kazuyoshi Yoshii*
*Graduate School of Informatics, Kyoto University, Japan
{tchen, yoshii}@sap.ist.i.kyoto-u.ac.jp
†Institute of Information Science, Academia Sinica, Taiwan
lisu@iis.sinica.edu.tw

***Abstract*—This paper describes a data-driven music structure analysis (MSA) method that performs segmentation and clustering of musical sections for a music signal. Since the intra-section homogeneity and inter-section difference are important clues for MSA, most studies on MSA have focused on self-similarity matrices (SSMs) computed from various acoustic features of a music signal. The performance of this approach, however, might be limited because the acoustic features used for computing SSMs are designed manually, and multiple SSMs are often integrated in a heuristic manner. To overcome these limitations, we propose a method that learns latent features useful for MSA with a stack of convolution-augmented multi-head self-attention (CAMHSA) layers that compute and fuse multiple self-attention maps representing multifaceted self-similarity. The estimated features are then clustered into an appropriate number of sections with a Gaussian mixture model (GMM). In the segmentation and clustering tasks, the proposed method outperformed baseline methods based on hand-crafted SSMs. In particular, it achieved state-of-the-art performance on the segmentation task. We found that the internal attention maps represent the section boundaries at the fine and course levels.**

I. Introduction

Audio-based music structure analysis (MSA) aims to split a musical recording into musically-meaningful segments based on the homogeneity and heterogeneity of the musical contents, and then assign a letter (e.g., A, B, or C) or functional label (e.g., verse, chorus, or bridge) to each of the segments. The process of MSA can be divided into two sub-tasks: *segmentation* and *clustering*. The former aims to detect the boundaries of musical sections. The latter aims to categorize the segments into multiple groups. In the field of music information retrieval (MIR), the two tasks are often tackled separately.

Given the repetitive nature of music, the self-similarity matrix (SSM) has been a common feature for tackling MSA [1]–[3], as repetitions and musically homogeneous parts typically result in diagonals and block patterns on the SSM. Nevertheless, the choice of audio representation for computing the SSM has a crucial impact on the performance of MSA. For instance, using timbre- or harmony-related features might yield two different SSMs highlighting possibly two distinct temporal structures for the same audio signal, as depicted in Fig. 1. Considering the multiple aspects of similarity, one may compute multiple SSMs from various kinds of acoustic features [4]. Several techniques have also been proposed for fusing different aspects of similarity [5], [6]. Such an approach, however, has a performance limitation due to a finite number of hand-crafted features used for computing SSMs. In addition, segmentation and clustering algorithms should be tailor-made according to the characteristics of each SSM.

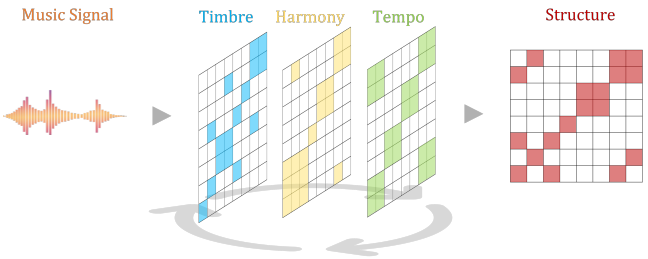


Fig. 1: The self-similarity of a music signal varies with the aspect across musical elements (e.g., timbre, harmony, and tempo). Learning a multifaceted view of self-similarity is thus crucial for music structure analysis.

To overcome these limitations, we propose an MSA method that performs segmentation and clustering on latent features learned by a deep neural network (DNN) with a convolution-augmented multi-head self-attention (CAMHSA) mechanism, an extension of the multi-head self-attention (MHSA) mechanism [7]. The main motivation for our study is that the self-attention maps computed internally through a stack of MHSA layers (i.e., the transformer encoder [7]) are expected to represent multifaceted self-similarity at different abstract levels. Our method is capable of learning to fuse such self-attention maps with additional convolutional operations in each layer. Specifically, the DNN is trained in a supervised manner to output latent features with which section boundaries are detected accurately and from which the derived SSM is made close to the ground-truth SSM. The latent features are then used for an unsupervised clustering method based on the Gaussian mixture model (GMM).

The main contributions of this study are twofold. We theoretically and experimentally show the effectiveness of the CAMHSA in learning a multifaceted self-similarity for MSA. We achieved a state-of-the-art performance of segmentation.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Related Work

This section briefly introduces audio-based music structure analysis and the self-attention mechanism.

A. Music Structure Analysis

Previous research in MIR often tackles MSA from either the segmentation or the clustering aspects. For the segmentation task, a bunch of deep learning approaches is proposed to identify musical boundaries within a given piece [8]–[12]. This task is often formulated as a binary classification problem, where spectrogram-based features, such as chromagram and mel-spectrogram, and SSMs derived from these features are commonly used as the inputs. Considering the size of SSMs grows quadratically with the duration of the corresponding audio signal, self-similarity lag matrices (SSLMs) [13] can be used instead to represent the similarities of each audio frame to a limited number of the preceding frames.

In contrast, the clustering task is typically approached as a clustering problem. This involves utilizing frame-wise acoustic features or affinity matrices (either hand-crafted or learned) as inputs for a clustering algorithm or a matrix decomposition method in order to group the audio frames [5], [6], [14]. Alternatively, this task can be formulated as a classification problem, and a model is employed to label each audio frame with a predefined vocabulary [15]. With manually annotated datasets such as the SALAMI dataset [16] and the Beatles dataset (released originally in the Isophonics dataset [17]), previous studies mostly applied supervised learning frameworks to the segmentation and clustering tasks. Recently, contrastive learning approaches are also explored in learning audio representations for MSA in an unsupervised or self-supervised fashion [18], [19]. For a detailed review of the audio-based MSA in MIR, we refer readers to [20].

B. Self-Attention Mechanism

The intra-attention mechanism [21], [22], also known as the *self-attention* (SA) [7], is proposed to encode compositional relationships between a set of elements (e.g., words of a sentence), and has demonstrated its effectiveness in various research fields such as natural language processing (NLP) [23]–[25] and computer vision (CV) [26]–[28]. In the field of MIR, several studies have utilized the SA [29]–[31]. However, the potential of the SA for MSA has yet to be investigated.

Formally, given a sequence of t elements with d-dimensional features, $\mathbf{X} \in \mathbb{R}^{t \times d}$, the SA yields a new representation of the sequence, i.e., $\mathrm{SA}(\mathbf{X}) \in \mathbb{R}^{t \times d}$, by internally computing an affinity matrix or an *attention map*, $\mathbf{A} \in \mathbb{R}^{t \times t}$, indicating the similarity between each pair of the elements in the sequence as follows:

$$\begin{aligned} \mathrm{SA}(\mathbf{X}) &= \mathbf{A} f_v(\mathbf{X}) \\ &= \mathrm{softmax}(f_q(\mathbf{X}) f_k(\mathbf{X})^{\mathrm{T}}) f_v(\mathbf{X}), \end{aligned} \tag{1}$$

where $f_{\{q,k,v\}} \colon \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ are learnable layers projecting the *query*, *key*, and *value* [7] (all denote the same input for self-attention) to a common latent space in which similarity is measured and the input is represented. With this definition, the attention map can be regarded as a generalization of a softmax-normalized SSM computed in the latent space, and it becomes an exact SSM if $f_q(\cdot) = f_k(\cdot)$.

Based on the idea of the SA, the multi-head self-attention (MHSA) mechanism [7] consists of multiple SAs for extracting multifaceted similarity information from various representation subspaces as follows:

$$\mathrm{MHSA}(\mathbf{X}) = [\mathrm{SA}(\mathbf{X}^{(1)}), \cdots, \mathrm{SA}(\mathbf{X}^{(h)})]\mathbf{W}, \tag{2}$$

where $h \in \mathbb{N}^+$ is the number of heads, $\mathbf{X}^{(\mathrm{h}=1:h)} \in \mathbb{R}^{t \times \frac{d}{h}}$ is a division of $\mathbf{X}$ within the feature dimension, $[\ ]$ denotes a matrix concatenation along the feature dimension, and $\mathbf{W} \in \mathbb{R}^{d \times d}$ is a learnable parameter used to fuse information from the attention heads. In practice, the attention heads can be computed in a parallel manner via manipulating the dimensionality of $f_{\{q,k,v\}}(\mathbf{X})$. Let $\hat{f}_{\{q,k,v\}}(\mathbf{X}) \in \mathbb{R}^{h \times t \times \frac{d}{h}}$ be the dimensionality-manipulated tensor of $f_{\{q,k,v\}}(\mathbf{X})$. We can rewrite (2) as follows:

$$\mathrm{MHSA}(\mathbf{X}) = \mathrm{restore}(\hat{\mathbf{A}} \hat{f}_v(\mathbf{X}))\mathbf{W}, \tag{3}$$

where $\hat{\mathbf{A}} = \mathrm{softmax}(\hat{f}_q(\mathbf{X}) \hat{f}_k(\mathbf{X})^{\mathrm{T}})$ and $\mathrm{restore}(\cdot)$ get back the dimensionality from $\mathbb{R}^{h \times t \times \frac{d}{h}}$ to $\mathbb{R}^{t \times d}$.

The MHSA can attend to various relations in different latent spaces by referring to multiple SSMs. Previous studies have pointed out that the attention heads of the MHSA are capable of capturing different types of implicit relationships between the elements of a sequence [32], [33]. The MHSA mechanism is thus favorable for MSA as it can alleviate the feature selection issue with learned features and affinity matrices.

III. Proposed Method

To cope with the feature selection issue when employing SSMs for MSA, we propose the *convolution-augmented multi-head self-attention*, which learns to *construct* and *fuse* multiple affinity matrices. Besides, we tackle the MSA with a two-stage framework. In the first stage, we employ a DNN equipped with the proposed attention mechanism to jointly predict musical boundaries and frame-wise representations for a given input. In the second stage, we obtain structural groups by clustering the frame-wise representations.

A. Data Representation

The input is a standardized audio signal sampled at 32 kHz. Three types of acoustic features are extracted from the signal: *mel-spectrogram*, *chromagram*, and *tempogram* [34]. For mel-spectrogram ($\mathbf{X}_m \in \mathbb{R}^{t \times 80}$), we use a mel-scaled filterbank of 80 triangular filters from 80 Hz to 16 kHz and scale magnitudes logarithmically. For chromagram ($\mathbf{X}_c \in \mathbb{R}^{t \times 12}$), the constant-Q spectrogram of 12 bins per octave is used. For tempogram ($\mathbf{X}_t \in \mathbb{R}^{t \times 384}$), the rhythmic content is encoded via the local auto-correlation of the onset strength envelope. All the features are calculated with *librosa* [35]. To analyze the entire piece once, which is crucial for learning long-term music structure, we compute all three acoustic features with a large hop size of

1600 points and downsample them with a factor of 10 using a median filter, resulting in a frame size of 0.5 sec.

The boundary annotation of each piece is given as a binary sequence ($\mathbf{Y}_b$) where 1 indicates a boundary and 0 otherwise; for the clustering task, the annotations of section names are converted into a finite alphabet ($\mathbf{Y}_s$) and the variation signs (e.g., '′' and '″') are ripped off. For instance, the annotation `[verse,verse′,chorus]` would become `[A,A,B]`.

B. Convolution-Augmented Multi-Head Self-Attention

While the MHSA is capable of capturing various relantionships between a set of elements, the calculations of the attention heads are independent of each other. More precisely, each attention map is computed individually without being related to other attention maps. In fact, it has been demonstrated that the tasks of MSA can benefit from integrating multiple affinity matrices into a united representation [6], [9]. For this reason, we propose to equip the MHSA with convolutional layers to fuse information from multiple attention maps. Concretely, we perform 2-D convolutions on the stack of the attention maps before the softmax function is applied. The convolution-augmented MHSA, denoted by $\mathrm{CAMHSA}(\mathbf{X}) \in \mathbb{R}^{t\times d}$, can thus be formulated as follows (Fig. 2):

$$\mathrm{CAMHSA}(\mathbf{X}) = \mathrm{restore}(\hat{\mathbf{A}}^{*} f_v(\hat{\mathbf{X}}))\mathbf{W}, \tag{4}$$

where $\hat{\mathbf{A}}^{*} = \mathrm{softmax}(\mathrm{conv}(f_q(\hat{\mathbf{X}}) f_k(\hat{\mathbf{X}})^{\mathrm{T}}))$ is an attention map and $\mathrm{conv}(\cdot)$ is a stack of two successive convolutional layers with a layer normalization [36] in between.

In fact, a recent work, i.e., the Conformer [37], has explored the combination of CNNs and the MHSA mechanism for modeling both local and global dependencies of a given sequence. Our approach differs from this work in how convolutions are employed. As shown in Fig. 3, the Conformer employs a 1-D CNN *outside* the MHSA block to capture local dependencies of the output sequence, whereas the CAMHSA performs 2-D convolutions *inside* the MHSA block to merge structural information of different attention maps. By utilizing convolutions, the CAMHSA can learn the correlations among various attention maps, leading to the enhancement of the attention heads.

C. Model Architecture

The model for MSA is composed of three parts, as shown in Fig. 4. The three types of input features are separately encoded by a 2-D CNN. The encoded features, $\mathbf{E}_{\{m,\ c,\ t\}} \in \mathbb{R}^{t\times d}$ for $\mathbf{X}_{\{m,\ c,\ t\}}$ respectively, are fused with another 2-D CNN. The fused feature, $\mathbf{E} \in \mathbb{R}^{t\times d}$, is then fed into a 2-layer transformer encoder with 8 attention heads [7], in which the MHSA mechanism is replaced by the CAMHSA. Finally, the latent representation by the transformer encoder, $\mathbf{H} \in \mathbb{R}^{t\times d}$, is used to predict segmentation, $\mathbf{P}_b \in \mathbb{R}^{t}$, for the segmentation task and frame-wise representations, $\mathbf{P}_s \in \mathbb{R}^{t\times d}$, for the clustering task with two 1-D CNNs. We set $d = 80$ for the model. The details of the computation blocks are depicted in Fig. 5 and elaborated as follows.

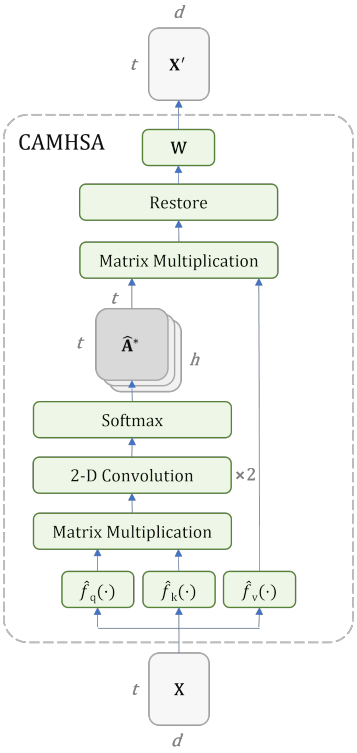


Fig. 2: Schematic diagram of the CAMHSA mechanism. The CAMHSA transforms $\mathbf{X}$ into $\mathbf{X}'$ similarly to the MHSA but with the addition of 2-D convolutions fusing information of the attention maps after the first matrix multiplication.

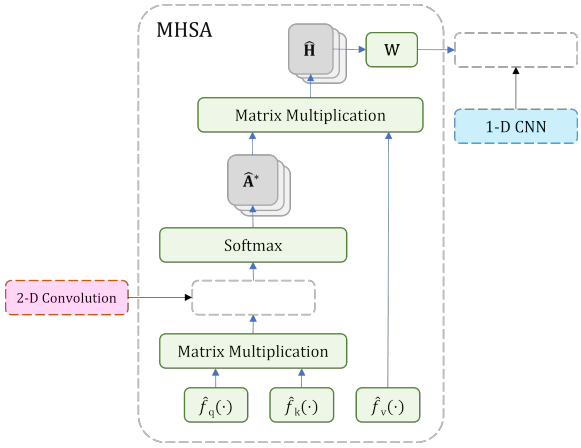


Fig. 3: Comparison between the CAMHSA and the Conformer. The CAMHSA improves the MHSA by performing 2-D convolutions after the matrix multiplication, whereas the Conformer utilizes a 1-D CNN for the MHSA output.

1) 2-D CNN (Fig. 5c): The 2-D CNNs are composed of three successive 2-D convolutions, each of which is followed by AdaNorm [38], the ReLU activation, and the *Squeeze-and-Excitation* (SE) operation [39], two branches of the dimension reduction performing *Max Pool* and *Dense* on the inner dimension, and a fusion of the two branches (*Dense & AdaNorm*

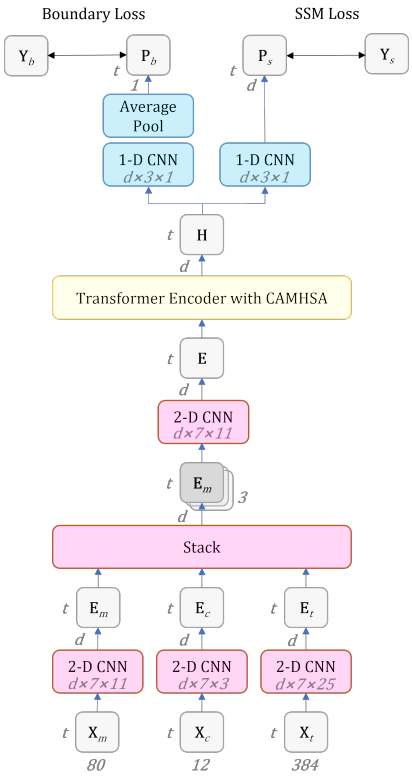
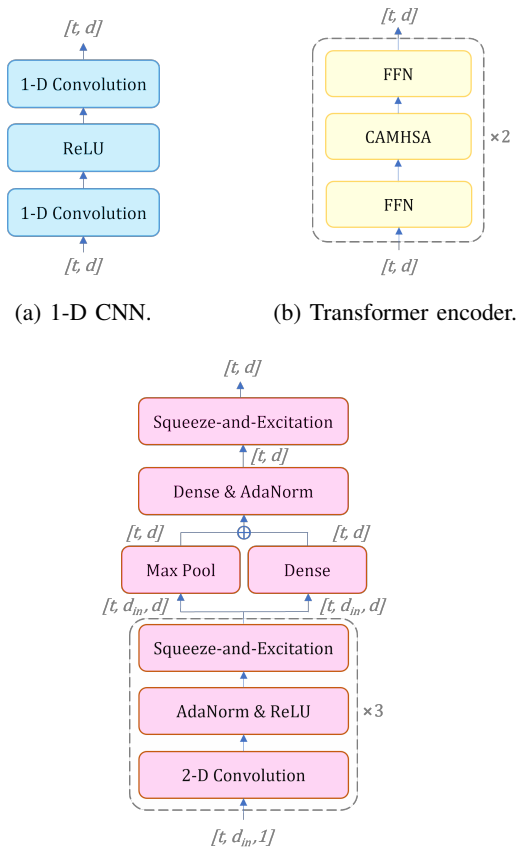


Fig. 4: Proposed model for MSA. The 2-D CNNs are used to encode and fuse the inputs; the Transformer encoder captures structural information via the CAMHSA mechanism; finally, the 1-D CNN functions as an output layer for boundary prediction. For each CNN, the number of convolution filters (c) and the kernel size, represented by height (h) and width (w), are denoted as $c \times h \times w$.



(a) 1-D CNN. (b) Transformer encoder.

(c) 2-D CNN.

Fig. 5: Computation blocks of the segmentation model. The MHSA block of the Transformer encoder is replaced by the CAMHSA. The dimensionality is indicated in the brackets.

followed by *SE*). The two branches of the dimension reduction function as global pooling layers summarizing the information of the *feature dimension* of the input (i.e., d_{in} in the figure).

2) Transformer Encoder with CAMHSA (Fig. 5b): We use a variant of the transformer architecture, i.e., the Macaron Net [37], [40], which has two feed-forward networks (FFNs) sandwiching the CAMHSA. Moreover, we employ relative positional embeddings to represent the distances between elements [41]. Besides, the residual connection and layer normalization techniques are also utilized in both the CAMHSA and the FFN (as used in the transformer blocks) despite not being displayed in the figure.

3) 1-D CNN (Fig. 5a): The 1-D CNNs consist of two 1-D convolutions with a ReLU activation in between.

D. Loss Functions

We employ two losses to jointly optimize the model (Fig. 4). The *boundary loss* is used for the segmentation task, while the *SSM loss* is used for the clustering task. The total loss is thus a summation of the two losses:

$$\text{Total Loss} = \text{Boundary Loss} + \text{SSM Loss}. \tag{5}$$

The boundary loss is a combination of the binary cross entropy loss (BCE) and the dice loss (DL) [42]:

$$\text{Boundary Loss} = \text{BCE}(\mathbf{P}_b, \mathbf{Y}_b) + \text{DL}(\mathbf{P}_b, \mathbf{Y}_b), \tag{6}$$

where $\text{DL}(\mathbf{A}, \mathbf{B}) = 1 - \frac{2 \times \sum \mathbf{A} \odot \mathbf{B}}{\sum \mathbf{A}^2 + \sum \mathbf{B}^2}$ ($\odot$ denotes the Hadamard product). We integrate the DL with the commonly used BCE, for it allows the model to pay attention to the overlapped regions between the predicted segmentation and the ground truth annotation [43], [44], and thus can alleviate the data imbalance issue in the segmentation task (i.e., the imbalance between boundary and non-boundary audio frames).

In contrast, we employ the SSM loss to regulate the predicted features, i.e., $\mathbf{P}_s$. Specifically, we compute two SSMs, $\mathbf{M}^{\mathbf{P}_s} \in \mathbb{R}^{t \times t}$ and $\mathbf{M}^{\mathbf{Y}_s} \in \mathbb{R}^{t \times t}$, with $\mathbf{P}_s$ and $\mathbf{Y}_s$ respectively, and then calculate the mean squared error:

$$\text{SSM Loss} = \frac{1}{t^2} \sum_{i,j=1:t} (\mathbf{M}^{\mathbf{P}_s} - \mathbf{M}^{\mathbf{Y}_s})^2, \tag{7}$$

where $\mathbf{M}_{ij}^{\mathbf{P}_s}$ is a similarity score ($\in [0, 2]$) measuring the normalized Euclidean distance between $\mathbf{P}_{s,i}$ and $\mathbf{P}_{s,j}$, and $\mathbf{M}_{ij}^{\mathbf{Y}_s} = 2$ if $\mathbf{Y}_{s,i} = \mathbf{Y}_{s,j}$ otherwise 0.

E. Clustering

We use the learned features by the proposed model, i.e., $\mathbf{P}_s$, to obtain section groups. We first perform the principal component analysis (PCA) to reduce the feature dimension of $\mathbf{P}_s$ from 80-D to 60-D. The reduced features are then clustered with Gaussian mixture models (GMMs). Specifically, we employ a set of GMMs with the number of mixture components ranging from 1 to k_{max} for each input piece, and then choose the best GMM based on the Bayesian information criterion (BIC) [45]. Finally, given the boundary prediction, we smooth the clustering results by reassigning the most common label in each segment to all frames within the segment.

IV. Evaluation

In this section, we present the comparative experiment conducted to evaluate the performance of the proposed method in terms of segmentation and clustering.

A. Data

To evaluate the proposed method, two datasets have been included. 1) The Beatles-TUT dataset[1] is a refined version of the Beatles dataset [17] by members of the Tampere University of Technology. It contains 174 songs by the Beatles and the corresponding *flat* (one-level, *upper*) annotations of structure. 2) The SALAMI dataset [16] is the largest publicly available set which contains *hierarchical* (two-level, *upper* and *lower*) annotations for 1,359 tracks. We use version 2.0 of the dataset[2] for the evaluation. We acquired 441 audio tracks from the Internet Archive[3] and found 555 matching tracks on YouTube[4].

We train and evaluate the proposed model with the two datasets individually. In the case of the Beatles-TUT dataset, we exclude 14 songs from the first album for testing and use the remaining 160 songs for training purposes. As for the SALAMI dataset, we utilize the 555 tracks from YouTube for training and the 441 tracks from the Internet Archive for testing. For both datasets, we augment the training data by 3 times via shifting the pitch of each track by ± 1 semitone.

B. Evaluation Metrics

For the segmentation task, we report the Hit Rate measure [46] with a time tolerance of ± 0.5 sec and ± 3 sec (which are the most frequently used values in the literature). This metric computes precision (P), recall (R), and F1 score (F1) by checking if a predicted boundary is close enough to an annotated boundary according to the given time tolerance.

For the clustering task, two metrics are employed. The *pairwise agreement* [47] is used to evaluate the frame-clustering segmentation regarding a flat annotation, while the *L-measure* [48] assesses all segmentation levels as a whole concerning a hierarchical annotation. Similarly to the Hit Rate measure, both the pairwise agreement and the L-measure compute the P, R, and F1 scores. Note that the L-measure is applied to the SALAMI dataset only as there are no hierarchical annotations in the Beatles-TUT dataset. All the evaluations are done using the *mir_eval* package [49].

C. Experiment Settings

The proposed model is evaluated separately on the segmentation task and the clustering task. Since the SALAMI dataset contains annotations at two structure levels, we equip the proposed model with an additional branch when training with the SALAMI dataset. Concretely, we create a branch by duplicating all computation blocks following the stack operation (see Fig. 4), while the first three 2-D CNNs are shared by the branches.

To validate our approach regarding the CAMHSA mechanism, we build a baseline model by removing the transformer encoder from the proposed model and including as additional inputs the SSMs derived from $\{\mathbf{X}_m, \mathbf{X}_c, \mathbf{X}_t\}$. Since it is memory-consuming to apply deep learning models to multiple track-level SSMs, we alternatively build three baseline models (denoted as *baseline-*$\{m, c, t\}$), each of which uses one of $\{\mathbf{X}_m, \mathbf{X}_c, \mathbf{X}_t\}$ and its corresponding SSM as inputs.

In comparison to previous work using supervised learning approaches, we report the best results in [9] (segmentation on SALAMI) and [11] (segmentation on Beatles and the upper level of SALAMI). To our knowledge, the former achieves the best segmentation performance on the SALAMI dataset while the latter on the Beatles dataset. We also report the results by [19] (segmentation and clustering on both Beatles and SALAMI) in comparison with the unsupervised learning approach. Note that the comparisons to these works should be taken with a grain of salt for there are many differences between these works and ours, such as dataset version (e.g., Beatles-Isophonics and Beatles-TUT), data usage (e.g., training and testing sets), and data representation.

Finally, we conduct ablation studies on the CAMHSA mechanism and the loss functions. Specifically, we build a model (denoted as *proposed w/o conv*) by replacing the CAMHSA blocks of the proposed model with the MHSA to validate the incorporation of convolutions; we also train the proposed model without including the DL in the boundary loss (denoted as *proposed w/o dl*) to investigate whether the DL provides any benefits to the segmentation task.

D. Results and Discussions

1) The Segmentation Task (Table I): On the Beatles-TUT dataset (Table Ia), our approach consistently outperformed all the baselines. Likewise, on the SALAMI dataset (Table Ib), our method was superior in all cases except for *Hit Rate(3)* at the lower level. When compared with the previous works, our method achieved new state-of-the-art segmentation performances on both datasets. More strikingly, our approach can rival the unsupervised approach [19] where the amount of audio tracks for training is two orders of magnitude larger than ours. Based on these findings, it appears that acquiring knowledge of multifaceted self-similarity is a more effective approach than relying on pre-established affinity matrices.

[1] https://pythonhosted.org/msaf/datasets.html.
[2] https://github.com/DDMAL/salami-data-public.
[3] https://archive.org/.
[4] https://github.com/jblsmith/matching-salami.

TABLE I: Performance on the segmentation task. *Hit Rate (0.5)* and *Hit Rate (3)* denote using the metric with a time threshold of {0.5, 3} sec. *baseline-{m, c, t}* indicates the baseline model with {mel-spectrogram, chromagram, tempogram} and the corresponding SSM as inputs. *proposed w/o dl* is the proposed model without using the dice loss for training. '–' is marked if a measure is not provided in the previous work.

Level	Model	Hit Rate (0.5)			Hit Rate (3)		
		P	R	F1	P	R	F1
upper	proposed	**0.555**	**0.617**	**0.578**	0.675	0.757	0.705
	proposed w/o conv	0.536	0.582	0.553	**0.700**	**0.769**	**0.729**
	proposed w/o dl	0.546	0.606	0.573	0.689	0.762	0.722
	baseline-m	0.436	0.560	0.488	0.559	0.718	0.625
	baseline-c	0.424	0.466	0.441	0.640	0.720	0.672
	baseline-t	0.425	0.511	0.460	0.608	0.733	0.659
	Maezawa [11]	0.507	0.520	0.492	–	–	–
	Buisson et al. [19]	–	–	–	–	–	0.718

(a) Evaluation on the Beatles-TUT dataset.

Level	Model	Hit Rate (0.5)			Hit Rate (3)		
		P	R	F1	P	R	F1
upper	proposed	**0.577**	0.586	0.559	**0.680**	0.688	**0.658**
	proposed w/o conv	0.534	**0.631**	**0.560**	0.629	**0.737**	0.658
	proposed w/o dl	0.533	0.537	0.518	0.665	0.666	0.645
	baseline-m	0.509	0.549	0.512	0.635	0.684	0.638
	baseline-c	0.504	0.375	0.417	0.636	0.472	0.525
	baseline-t	0.423	0.339	0.363	0.613	0.490	0.526
	Grill & Schlüter [9]	–	–	0.508	–	–	–
	Maezawa [11]	0.301	0.347	0.306	–	–	–
	Buisson et al. [19]	–	–	–	–	–	0.627
lower	proposed	**0.650**	**0.657**	**0.621**	0.806	0.820	**0.774**
	proposed w/o conv	0.590	0.611	0.574	0.785	0.821	0.768
	proposed w/o dl	0.582	0.598	0.561	0.790	0.812	0.763
	baseline-m	0.553	0.553	0.528	0.801	0.810	0.771
	baseline-c	0.483	0.581	0.502	0.728	**0.886**	0.763
	baseline-t	0.405	0.349	0.357	**0.827**	0.742	0.746
	Grill & Schlüter [9]	–	–	0.485	–	–	–
	Buisson et al. [19]	–	–	–	–	–	0.643

(b) Evaluation on the SALAMI dataset.

2) The Clustering Task (Table II): Similarly to the segmentation task, our approach surpassed the baselines in almost all cases. In general, the clustering performance on the SALAMI dataset (Table IIb) is worse than that on the Beatles-TUT (Table IIa) since the SALAMI dataset is more diverse in terms of music genres and styles. Upon evaluation, it is apparent that clustering at the lower level is more challenging than at the upper level for the SALAMI dataset. This can be attributed to the greater number of boundaries in the lower level compared to the upper level. Furthermore, the pairwise agreement measure is sensitive to the precise placement of the boundaries between the prediction and the ground truth [50]. Besides, it is worth pointing out that [19] obtained remarkable performances on the clustering task due to its contrastive learning objective aligning with the evaluation metrics.

3) Ablation studies (proposed w/o conv and w/o dl): When the CAMHSA is replaced by the MHSA (*proposed w/o conv*), we noticed a decline in both segmentation and clustering performances. This suggests that the correlations among the attention maps of different heads could potentially play a role in MSA. As for the ablation on the loss functions (*proposed w/o dl*), we found that employing the DL positively impacted the overall performance.

TABLE II: Performance on the clustering task. Note that the L-measure can be applied to the SALAMI dataset only, and it computes a unified score over the two levels although we report the results at the *upper* level.

Level	Model	Pairwise Agreement			L-measure		
		P	R	F1	P	R	F1
upper	proposed	**0.553**	0.803	0.648	–	–	–
	proposed w/o conv	0.536	**0.828**	0.639	–	–	–
	proposed w/o dl	0.512	0.753	0.604	–	–	–
	baseline-m	0.509	0.659	0.565	–	–	–
	baseline-c	0.542	0.752	0.624	–	–	–
	baseline-t	0.553	0.575	0.554	–	–	–
	Buisson et al. [19]	–	–	**0.723**	–	–	–

(a) Evaluation on the Beatles-TUT dataset.

Level	Model	Pairwise Agreement			L-measure		
		P	R	F1	P	R	F1
upper	proposed	0.561	0.695	0.592	0.420	0.559	0.476
	proposed w/o conv	0.550	**0.700**	0.589	0.412	0.552	0.469
	proposed w/o dl	0.581	0.666	0.594	**0.442**	0.570	0.494
	baseline-m	0.590	0.589	0.559	0.417	0.506	0.454
	baseline-c	0.538	0.690	0.577	0.406	0.532	0.457
	baseline-t	**0.596**	0.477	0.501	0.405	0.450	0.423
	Buisson et al. [19]	–	–	**0.714**	0.432	**0.694**	**0.527**
lower	proposed	0.428	0.550	0.457	–	–	–
	proposed w/o conv	0.416	**0.554**	0.450	–	–	–
	proposed w/o dl	**0.490**	0.454	0.445	–	–	–
	baseline-m	0.438	0.408	0.397	–	–	–
	baseline-c	0.415	0.503	0.430	–	–	–
	baseline-t	0.444	0.289	0.328	–	–	–
	Buisson et al. [19]	–	–	**0.580**	–	–	–

(b) Evaluation on the SALAMI dataset.

4) Feature Selection Issue (baseline-{m, c, t}): When comparing the baselines, it becomes clear that there is no prevailing feature for MSA. The mel-spectrogram exhibited superiority in the segmentation task, while the chromagram appeared to be preferable to the clustering task. Besides, the tempogram can also provide structural information to some extent. These results highlight the importance of learning higher-level representations and affinity matrices instead of relying on manually crafted features.

5) Visualization of Attention Maps: To examine the knowledge gained by CAMHSA on MSA, we analyzed the attention maps of the testing data from the SALAMI dataset. There were three primary observations exemplified in Fig. 6: 1) the attention heads display a variety of block patterns; 2) the size of the block patterns tends to be larger at the upper level compared to the lower level; 3) the SSM computed from the learned representation ($\mathbf{M}^{\mathbf{P}_s}$) is similar to the annotated one ($\mathbf{M}^{\mathbf{Y}_s}$). The first observation is in line with previous works suggesting that the attention heads can capture diverse relationships. The second observation reflects the fact that the musical segments tend to have longer duration at the upper

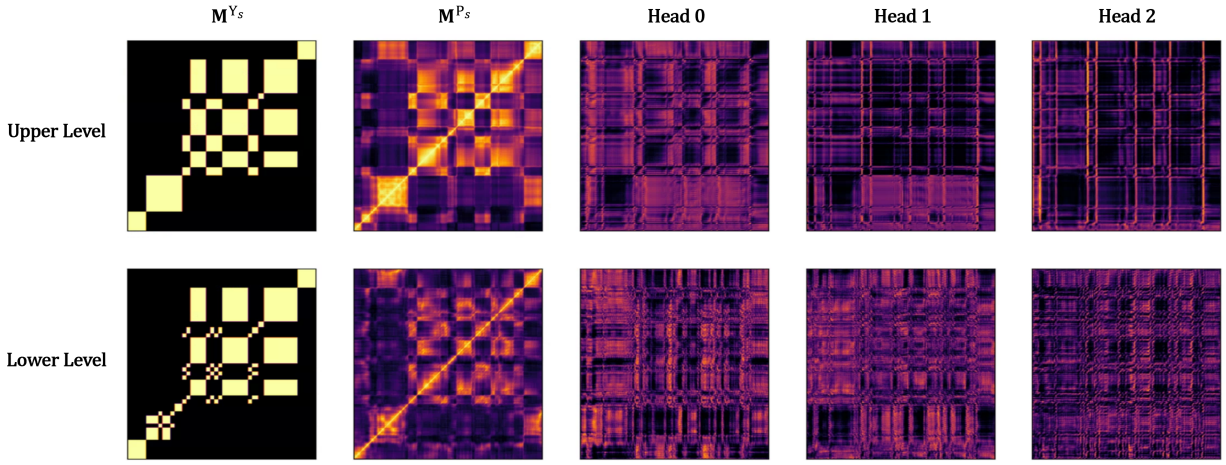


Fig. 6: Visualization of the attention maps for the song "Shim Sham" by Caravan of Thieves (song id 1038 of the SALAMI dataset). $\mathbf{M}^{\{\mathbf{Y}_s,\ \mathbf{P}_s\}}$ is the SSM computed with $\{\mathbf{Y}_s,\ \mathbf{P}_s\}$, respectively. Only the first three heads (index $\{0,\ 1,\ 2\}$) at the second layer of the transformer encoder are shown.

level. The third observation shows that the CAMHSA can effectively combine attention maps from different sources to create a cohesive representation ($\mathbf{P}_s$) that closely resembles the ground truth ($\mathbf{Y}_s$) in terms of SMMs.

V. Conclusion

This research focused on utilizing the multi-head self-attention (MHSA) mechanism to learn a multifaceted view of self-similarity across multiple abstraction levels. To achieve this, we have developed the convolution-augmented MHSA (CAMHSA) mechanism, which fuses various perspectives of self-similarity to extract latent features for audio-based MSA. Through our experiments, we have clearly demonstrated the effectiveness of the CAMHSA by comparing it with other methods that rely on hand-crafted affinity matrices. Especially for the segmentation task, we obtained new state-of-the-art results on the Beatles-TUT and the SALAMI datasets. For the clustering task, we adopted a two-stage framework, where the learning of audio representations is detached from the successive clustering stage. The clustering performance would thus heavily depend on the capability of the employed clustering algorithm. To mitigate this potential incompatibility, we plan to involve a deep clustering method in the current framework. Finally, by fusing different acoustic features, we are able to find a unified structure for each input signal. Examining the differences between the attention maps and SSMs derived from these features can provide musical insights into the relationship between the unified structure and each specific feature.

Acknowledgment

This work was partially supported by JST PRESTO No. JPMJPR20CB and KAKENHI Nos. 20H00602 and 21H03572.

References

[1] H. Grohganz, M. Clausen, N. Jiang, and M. Müller, "Converting path structures into block structures using eigenvalue decompositions of self-similarity matrices," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 209–214, 2013.

[2] B. McFee and D. P. W. Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5197–5201, 2014.

[3] J. B. L. Smith and M. Goto, "Multi-part pattern analysis: Combining structure analysis and source separation to discover intra-part repeated sequences," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 716–723, 2017.

[4] T. Grill and J. Schlüter, "Music boundary detection using neural networks on spectrograms and self-similarity lag matrices," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1296–1300, 2015.

[5] T. Cheng, J. B. L. Smith, and M. Goto, "Music structure boundary detection and labelling by a deconvolution of path-enhanced self-similarity matrix," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 106–110, 2018.

[6] C. J. Tralie and B. McFee, "Enhanced hierarchical music structure annotations via feature level similarity fusion," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 201–205, 2019.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[8] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 417–422, 2014.

[9] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 531–537, 2015.

[10] A. Cohen-Hadria and G. Peeters, "Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks," in *Proceedings of the AES International Conference Semantic Audio*, 2017.

[11] A. Maezawa, "Music boundary detection based on a hybrid deep model of novelty, homogeneity, repetition and duration," in *Proceedings of*

the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 206–210, 2019.

[12] C. Hernandez-Olivan, J. R. Beltrán, and D. Diaz-Guerra, "Music boundary detection using convolutional neural networks: A comparative analysis of combined input features," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, p. 78, 2021.

[13] M. Goto, "A chorus-section detecting method for musical audio signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 437–440, 2003.

[14] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 405–410, 2014.

[15] G. Shibata, R. Nishikimi, and K. Yoshii, "Music structure analysis based on an LSTM-HSMM hybrid model," in *Proceedings of the 21th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 23–29, 2020.

[16] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 555–560, 2011.

[17] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, "OMRAS2 metadata project 2009," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR) - Late-Breaking Session*, 2009.

[18] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 346–350, IEEE, 2019.

[19] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour, "Learning multi-level representations for hierarchical music structure analysis," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, pp. 591–597, 2022.

[20] O. Nieto, G. J. Mysore, C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee, "Audio-based music structure analysis: Current trends, open challenges, and applications," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, pp. 246–263, 2020.

[21] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 551–561, 2016.

[22] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2249–2255, 2016.

[23] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Adversarial transfer learning for chinese named entity recognition with self-attention mechanism," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 182–192, 2018.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pp. 4171–4186, 2019.

[25] R. Al-Rfou, D. Choe, N. Constant, M. Guo, and L. Jones, "Character-level language modeling with deeper self-attention," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3159–3166, 2019.

[26] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, 2018.

[27] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, pp. 4052–4061, 2018.

[28] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97, pp. 7354–7363, 2019.

[29] J. Park, K. Choi, S. Jeon, D. Kim, and J. Park, "A bi-directional Transformer for musical chord recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, (ISMIR)*, pp. 620–627, 2019.

[30] T. Chen and L. Su, "Harmony Transformer: Incorporating chord segmentation into harmony recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 259–267, 2019.

[31] Y. Huang and Y. Yang, "Pop music Transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia (ACM Multimedia)*, pp. 1180–1188, 2020.

[32] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pp. 5797–5808, 2019.

[33] J. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[34] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram - A mid-level tempo representation for musicsignals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5522–5525, 2010.

[35] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th python in science conference (SCIPY)*, vol. 8, 2015.

[36] L. J. Ba, R. Kiros, and G. E. Hinton, "Layer normalization," in *arXiv preprint arXiv: 1607.06450*, 2016.

[37] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 5036–5040, 2020.

[38] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 4383–4393, 2019.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[40] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T. Liu, "Understanding and improving Transformer from a multi-particle dynamic system point of view," *CoRR abs/1906.02762*, 2019.

[41] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pp. 464–468, 2018.

[42] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 4th International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[43] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–7, 2020.

[44] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 465–476, 2020.

[45] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461 – 464, 1978.

[46] D. Turnbull, G. R. G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," in *Proceedings of the 8th International Conference on Music Information (ISMIR)*, 2007.

[47] M. Levy and M. B. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[48] B. McFee, O. Nieto, M. M. Farbood, and J. P. Bello, "Evaluating hierarchical structure in music annotations," *Frontiers in Psychology*, vol. 8, 2017.

[49] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common MIR metrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 367–372, 2014.

[50] O. Nieto and J. P. Bello, "Music segment similarity using 2d-fourier magnitude coefficients," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 664–668, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Simultaneous Measurement of Multiple Acoustic Attributes Using Structured Periodic Test Signals Including Music and Other Sound Materials

Hideki Kawahara*, Kohei Yatabe†, Ken-Ichi Sakakibara‡, Mitsunori Mizumachi§, Tatsuya Kitamura¶,
* Wakayama University, Japan E-mail: kawahara@wakayama-u.ac.jp
† Tokyo University of Agriculture and Technology, Japan E-mail: yatabe@go.tuat.ac.jp
‡ Health Sciences University of Hokkaido, Japan E-mail: kis@hoku-iryo-u.ac.jp
§ Kyushu Institute of Technology, Japan E-mail: mizumach@ecs.kyutech.ac.jp
¶ Konan University, Japan E-mail: t-kitamu@konan-u.ac.jp

***Abstract*—We introduce a general framework for measuring acoustic properties such as liner time-invariant (LTI) response, signal-dependent time-invariant (SDTI) component, and random and time-varying (RTV) component simultaneously using structured periodic test signals. The framework also enables music pieces and other sound materials as test signals by "safeguarding" them by adding slight deterministic "noise." Measurement using swept-sin, MLS (Maxim Length Sequence), and their variants are special cases of the proposed framework. We implemented interactive and real-time measuring tools based on this framework and made them open-source. Furthermore, we applied this framework to assess pitch extractors objectively.**

I. Introduction

Computing devices today are 10^9 times more powerful than those available a half-century ago [1]. This power makes it possible to process huge amount of speech materials (for example: [2], [3]). This advancement motivated us to establish a framework for making speech materials more usable based on our recent findings [4]–[6].

The speech materials mentioned above do not necessarily fulfill recommended conditions for scientific research [7]. However, the abovementioned conditions are too strict depending on the research purpose. There is room to make speech materials usable by providing tools for assisting data acquisition and retrospective assesment [8].

The framework we introduce in this article provides a solid basis for tools making speech materials reusable. We start with a big picture of the framework, followed by descriptions of constituent algorithms. Then, we introduce acoustic measurement tools followed by applications other than acoustic measurement. Finally, we discuss issues and related works.

The main contribution of this paper is the fundamental reformulation of the underlying algorithms and the complete revision of the infrastructure, computationally efficient implementation based on FFT and inverse FFT. Our previous works consisted of many conceptual confusions and ad hoc and inefficient procedures [4]–[6], [17], [18], [20]–[22]. This paper enables us to renew all applications and tools while maintaining their functionality.

II. Measurement of acoustic attributes

Figure 1 shows a schematic representation of the whole types of acoustic attributes measurement that we will introduce in this article. The following three subsections outline this framework followed by sections describing technical details and applications.

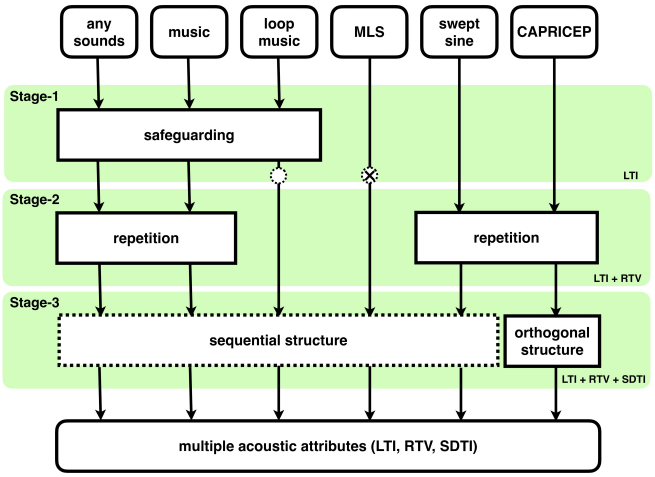


Fig. 1. Acoustic attributes and test signals. (LTI: Linear Time-Invariant, RTV: Random and Time-Varying, and SDTI: Signal Dependent Time-Invariant). The terminal symbol placed at the bottom of Stage-1 on the line from "MLS" represents that results are erroneous. The terminal symbol on the line from "loop music" represents that it destroys the signal design, looping.

A. Linear time-invariant (LTI) response: Stage-1

For linear time invariant (LTI) systems, response to input impulse (impulse response) uniquely determines the target system. In the frequency domain, dividing the Fourier transform of the target system output by the Fourier transform of the input signal provides the transfer function of the target system. The symbol "swept sine" [9] and "MLS" (Maximum Length Sequence [10], [11]) are commonly used test signals. They are members of TSP (Time Stretched Pulse) and have flat power spectrum. The symbol "CAPRICEP" (Casceded All-Pass filters with RandomIzed CEnter frequencies and Phase polority) is also a new member of TSP we proposed [5].

Spectral division using any other signals also provides the transfer function, in principle. However, their spectrum generally consists of very weak component(s) that makes

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

spectral division impractical. We introduced "safeguarding" method for making such signal usable for transfer function measurement [6].

B. Random and time-varying component: Stage-2

Two neighboring periods excerpted from the repeatedly concatenated periodic segments are identical. However, periods excerpted from different parts of the acquired signal are not identical because the background noise and sensitivity fluctuations do not have the same periodicity. The difference between observed periods provides background noise and sensitivity fluctuations information. This way, the repetitive presentation of periodic segments makes simultaneous measurement of the LTI response and the disturbance in the measurement.

C. Signal dependent component: Stage-3

Estimated impulse responses using different input signals are identical when the system is strictly LTI. However, estimated impulse responses of an acoustic system in the real world are not identical for different input signals because of non-linearity, even after suppressing the random and time-varying component by averaging many repetitive measurement results. The differences in estimated impulse responses are the second type of disturbing component. The component depends on the used test signals. It is necessary to describe the used test signals.

Repetition (used in Stage-2) is one type of structuring for the test signal. Sequentially aligning different repetitive test signals is the second type of test signal structuring. Using CAPRICEP, we introduce the third type of structuring, simultaneous structuring, by using the orthogonal nature of the Walsh-Hadamard matrix.

III. Algorithm

This section describes the underlying algorithms in each stage. The test and the acquired signals are discrete-time signals and share the same sampling clock. (In the appendix, we introduce a solution for handling signals sampled by independent clocks.)

A. LTI response: Stage-1, linear convolution

Let $x[n], y[n]$, and $h[n]$ represent an input signal, an output signal, and the impulse response of a system. The time index n is an integer. Assume that $x[n]$ is non-zero for $0 \leq n \leq N-1$ and $h[n]$ is non-zero for $0 \leq n \leq M-1$. Then, the output signal $y[n]$ is non-zero for $0 \leq n \leq M+N-1$. This setting, usually called as "zero-padding", is a common practice to calculate linear convolution using DFT-based circular convolution [12]. (See Fig. 2)

By using the discrete Fourier transform $\mathcal{F}[\cdot]$ of length $L > M+N-1$, the following equation provides the transfer function $H[k]$ of the system. The symbol k represents the index of the discrete frequency.

$$H[k] = \frac{\mathcal{F}[y[n]]}{\mathcal{F}[x[n]]}. \tag{1}$$

Without additive noise, the inverse transform $\mathcal{F}^{-1}[\cdot]$ of the transfer function provides the estimate of the impulse response $h_{\text{est}}[n]$.

$$h_{\text{est}}[n] = \mathcal{F}^{-1}[H[k]]. \tag{2}$$

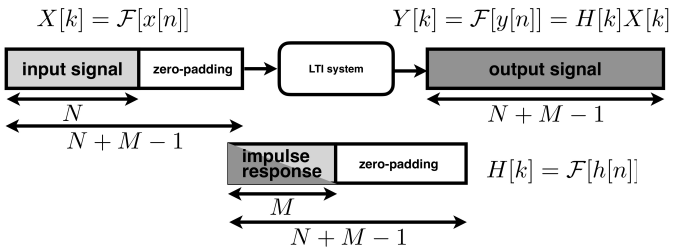


Fig. 2. Convolution of discrete-time signals. Implementation using cyclic convolution with zero-padding.

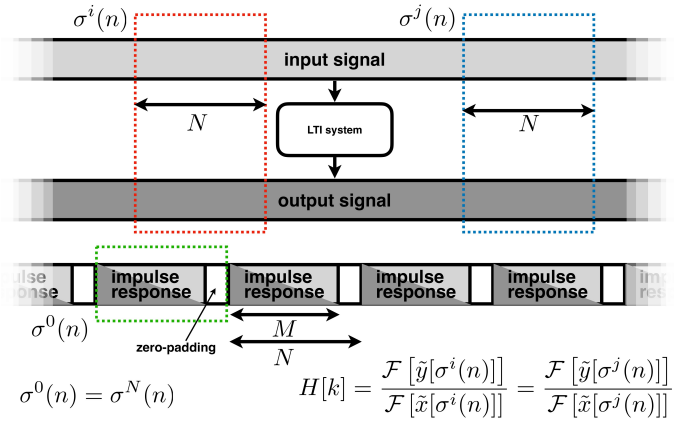


Fig. 3. Convolution of periodic signals. Implementation using cyclic convolution without zero-padding for input and output signals. The same cyclic permutation to input and output does not change the estimated transfer function.

The typical TSP (Time Stretched-Pulse) signal, swept-sine, and the recent member of TSP, unit-CAPRICEP, (usually) have frequency independent gain, such as $|H[k]| = 1$ for all k. Because of the background noise, frequency-dependent gain improves estimation accuracy [13].

1) Terminal symbols: Figure 1 has two terminal symbols for "loop music" and "MLS" at the bottom of Stage-1. We placed the symbol because MLS is a periodic sequence and is orthogonal only under cyclic convolution. The separated single cycle of the MLS sequence is not a time-stretched pulse under linear convolution. For loop music, isolating one repeated phrase is usable for calculating the LTI response, although it is not looping and destroys the original design purpose.

2) Signal safeguarding: Equation (2) holds for input signals with frequency-dependent spectral values. However, the estimated transfer functions are not usable because small absolute values in the denominator magnify the effects of observation noise. Signal-safeguarding, our proposal, significantly suppresses observation noise effects and makes any sounds appropriate for acoustic measurement [6]. With safeguarding, if necessary, all signals are relevant for acoustic measurement.

B. LTI response, and random and time-varying component: Stage-2, cyclic convolution

Figure 3 summarizes this stage. Zero padding is inefficient. The periodic nature of the discrete Fourier transform removes the zero-padding for input and output in Fig. 2.

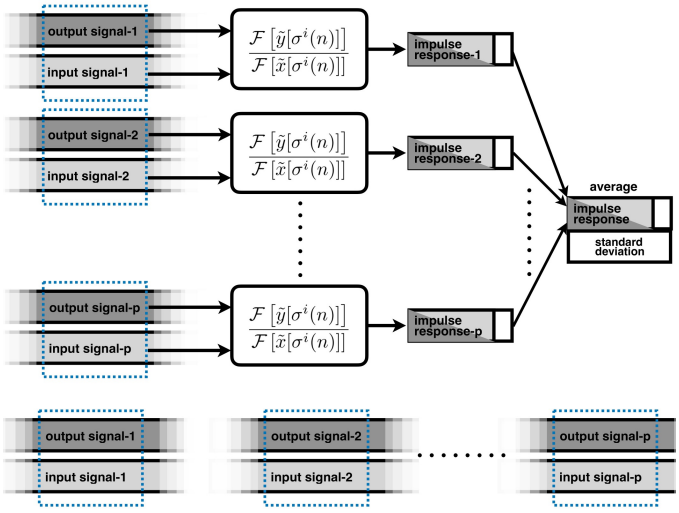


Fig. 4. Serial structuring illustration. Signal-dependent component isolation by averaging and calculating sample standard deviation of estimated impulse responses using different test signals.

Assume that $x[n], y[n]$, and $h[n]$ are periodic and share the same period, L. Then, there is no need for excessive zero-padding for test signals. Using a swept-sine signal in this repetitive presentation, removing zero-valued parts does not cause any problems. The following equation provides the transfer function.

$$H[k] = \frac{\mathcal{F}\left[\tilde{y}[n]\right]}{\mathcal{F}\left[\tilde{x}[n]\right]}, \tag{3}$$

where $\tilde{x}[n]$ and $\tilde{y}[n]$ represent that the signals are periodic.

Periodic test signal enables repeated measurement. Disturbance in measurement, such as background noise and interfering sounds, does not share the same periodicity with the test signal. Then, it is safe to assume that disturbing sounds are independent at each repetition. Averaging of the estimated impulse responses reduces the estimation variance. This averaging and calculation of observed standard deviation provide an estimate of the random and time-varying component (RTV in Fig. 1).

Periodic test signal has additional merits. Because it is periodic, there is no need to repeat the calculation of the Fourier transform of the test signal. Because it is periodic, there is no need to find the initial position of the analysis segment. Wherever the initial position is, the signal is periodic as far as the length is identical to the period. In short, the following holds.

$$\frac{\mathcal{F}\left[\tilde{y}[\sigma^i(n)]\right]}{\mathcal{F}\left[\tilde{x}[\sigma^i(n)]\right]} = \frac{\mathcal{F}\left[\tilde{y}[\sigma^j(n)]\right]}{\mathcal{F}\left[\tilde{x}[\sigma^j(n)]\right]}, \tag{4}$$

where $\sigma^i(n)$ and $\sigma^j(n)$ represent i-th and j-th cyclic permutations of the discrete time sequence n. Figure 3 illustrates these merits.

C. LTI response, and random and time-varying, and signal dependent components: Stage-3, using different test signals

The estimated impulse response is identical when the system is LTI strictly, and no noise exists. However, in actual acoustic measurement, impulse responses estimated using different input test signals are different. Averaging impulse responses using many different test signals provides less distorted LTI response. The difference between each impulse response and the averaged one provides an estimate of the magnitude of signal-dependent deviation. Note that we assume each estimated impulse response for a test signal is an average of repetition, and this averaging suppresses random variations.

Figure 4 illustrates one implementation of this strategy, serial structuring. This serial structuring is the only option for test signals other than CAPRICEP.

D. LTI response, and random and time-varying, and signal dependent components: Stage-3, structured test signal

CAPRICEP has another option, orthogonal structuring. Each unit-CAPRICEP has a raised-cosine-shaped power envelope. Periodic allocation of each unit-CAPRICEP with 50% overlap yields a constant power envelope. We use the Walsh-Hadamard matrix B of order 4 to determine the polarity of three different unit-CAPRICEPs.

$$B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} b_{1,1} & \cdots & \cdots & b_{1,4} \\ \vdots & b_{2,2} & & \vdots \\ \vdots & & \ddots & \vdots \\ b_{4,1} & \cdots & \cdots & b_{4,4} \end{bmatrix}. \tag{5}$$

1) Periodic signal generation: The following part describes a procedure to make a periodic test signal $\tilde{s}[n]$ from three different unit-CAPRICEPs. The discrete Fourier transform of $\tilde{s}[n]$ yields discrete spectral representation $S[k]$. The following procedure makes the absolute value $|S[k]|$ a constant, that is independent of the discrete frequency k.

Multiplying coefficient $b_{i,j}$ to each unit-CAPRICEP $g_{\mathrm{uC}}^{(q)}[n]$ and applying the overlap-and-add operation with $N/4$ shift for three different unit-CAPRICEPs provide a base unit sequence $u[n]$.

$$u[n] = \sum_{q=1}^{3} C_q \sum_{r=1}^{4} b_{q,r}\; g_{\mathrm{uC}}^{(q)}\left[n - \frac{(q-1)N}{4}\right] \tag{6}$$

where the super-script q of $g_{\mathrm{uC}}^{(q)}[n]$ is the identifier of each unit-CAPRICEP. The coefficient C_q value is 1 for $q = 1, 2$, and $\sqrt{2}$ for $q = 3$.

The next step is to make a periodic test signal $\tilde{s}[n]$ from the base unit sequence $u[n]$. Since the width of a unit-CAPRICEP is four to six times wider than the allocation interval $N/4$ (see Appendix), it is necessary to allocate (overlap-and-add) $u[n]$ more than $6 + 1$ times to make a periodic segment. Let P represent the number of repetitions. Considering the condition and for simplicity, we select an even number $P \geq 8$.

The periodic test signal $\tilde{s}[n]$ is an excerpt from the following intermediate signal $s_{\mathrm{tmp}}[n]$.

$$s_{\mathrm{tmp}}[n] = \sum_{p=1}^{P} u[n - (p-1)N], \tag{7}$$

Excerpting a segment with length N around the center location of $s_{\mathrm{tmp}}[n]$ provides the periodic test signal $\tilde{s}[n]$.

$$\tilde{s}[n] = s_{\mathrm{tmp}}\left[n - \left(\frac{P}{2} - 1\right) N\right], \quad (n = 0, \ldots, N-1) \tag{8}$$

2) Simultaneous multiple impulse response measurement: The following procedure provides three different impulse responses from the output signal $\tilde{y}[n]$. First, similar to the second stage, using the ratio of the discrete Fourier transform of input test signal and output provides an estimate of the impulse response $\hat{h}_{\mathrm{L}}[n]$.

$$\hat{h}_{\mathrm{L}}[n] = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\tilde{y}[n]]}{\mathcal{F}[\tilde{s}[n]]}\right], \tag{9}$$

where the subscript L of $\hat{h}_{\mathrm{L}}[n]$ represents that it is a longer impulse response obtained from the measurement. The length of $\hat{h}_{\mathrm{L}}[n]$ is L. Although, in actual measurement, other than the initial part, the noise floor (due to background noise and non-linearity, mainly) masks the impulse response corresponds to the LTI part.

The orthogonal structure of $\tilde{s}[n]$ provides three shorter impulse responses by using the virtual target signal $\tilde{v}_{\mathrm{S}}^{(q)}[n]$.

$$\tilde{v}_{\mathrm{S}}^{(q)}[n] = \sum_{r=1}^{4} b_{q,r}\ \delta\left[n, \frac{(q-1)N}{4}\right], \tag{10}$$

where $\delta[i, j]$ is the Kronecker delta and q identifies the corresponding unit-CAPRICEP.

The first $N/4$ elements of the signal defined by the following equation provide a short impulse response $\hat{h}_{\mathrm{S}}^{(q)}[n]$.

$$\hat{h}_{\mathrm{S}}^{(q)}[n] = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\tilde{v}_{\mathrm{S}}^{(q)}[n]]\mathcal{F}[\tilde{y}[n]]}{\mathcal{F}[\tilde{s}[n]]}\right]. \tag{11}$$

As noted, the initial $N/4$ elements are unique and relevant for impulse response estimates. Deviations from the same initial part of $\hat{h}_{\mathrm{L}}[n]$ represent the signal-dependent component.

Note that $\mathcal{F}[\tilde{v}_{\mathrm{S}}^{(q)}[n]]$ functions as a selector of discrete frequency components. Figure 5 illustrates how it works in the discrete frequency domain. For example, $\hat{h}_{\mathrm{S}}^{(1)}[n]$ selects the bleu-colored component, and $\hat{h}_{\mathrm{S}}^{(2)}[n]$ and $\hat{h}_{\mathrm{S}}^{(3)}[n]$ select red-colored and yellow-colored components.

There is a caveat that the procedure above suppresses the signal-dependent component resulting from even-symmetric nonlinearity. It is necessary to estimate impulse responses using the negated test signal $-\tilde{s}[n]$ to detect components caused by even-symmetric nonlinearity.

E. Implementation details

The descriptions mentioned above are simplified summaries. We placed descriptions of technical details in Appendix sections for readability. The discussed details are as follows: a) Optimized weighting shape for reducing artifacts due to truncation of acquired interfering signals. (Appendix A) b) Phase manipulation function that makes better localization of the generated unit-CAPRICEP. (Appndix B) c) Evaluation of signal safeguarding merits. (Appendix C)

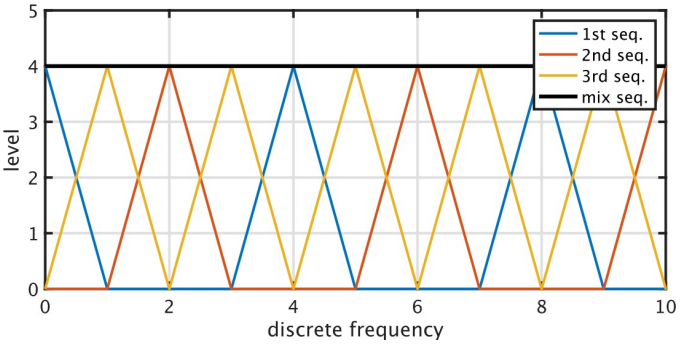


Fig. 5. Discrete spectral levels of the whole test signal $\tilde{s}[n]$ and constituent sequences corresponding first, second, and the third unit-CAPRICEP. The plot shows the initial 11 discrete frequency components.

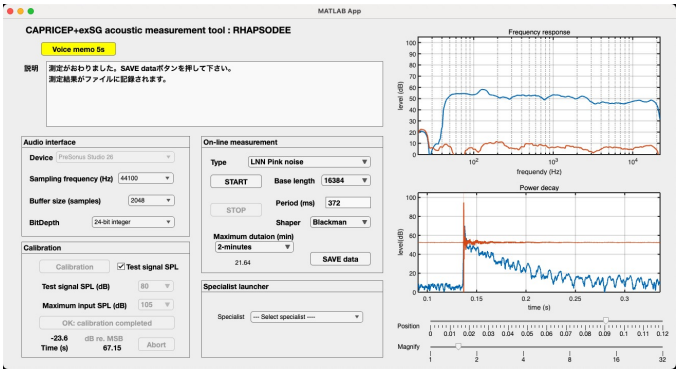


Fig. 6. Snapshot of the GUI of the "Control panel" for acoustic calibration and application launcher [15].

IV. Acoustic Measurement Tools

We developed acoustic measurement tools based on the proposed framework. This section introduces several examples. The tools introduced here are accessible in the first author's GitHub repositories [14].

A. Control panel for acoustic applications

Figure 6 shows a snapshot of the GUI of the "Control panel" for acoustic calibration and application launcher [15]. The top-left wide field is for showing prompt instructions for users. The left-middle and bottom sub-panels are for setting the audio interface and calibrating input sensitivity. The center sub-panel is for controlling the acoustic measurement. The center-bottom sub-panel is the application launcher.

Figure 7 shows a collection of GUI snapshots of speech-related applications we developed [15]–[18]. Selecting an item from the dropdown menu in the application launcher sub-panel launches the application.

The screenshot of Fig. 6 shows an interactive and real-time measurement of the sound recording environment using a periodic test signal with a modified pink-noise spectral shape. A powered loudspeaker (IKmultimedia iLound Micro Monitor) simulates a talker, and a measuring microphone (Earthworks M50) placed 20 cm in front of the loudspeaker acquires the output sound. The top-right graph shows the smoothed gain (blue line) of the transfer function $H[k]$ and the smoothed

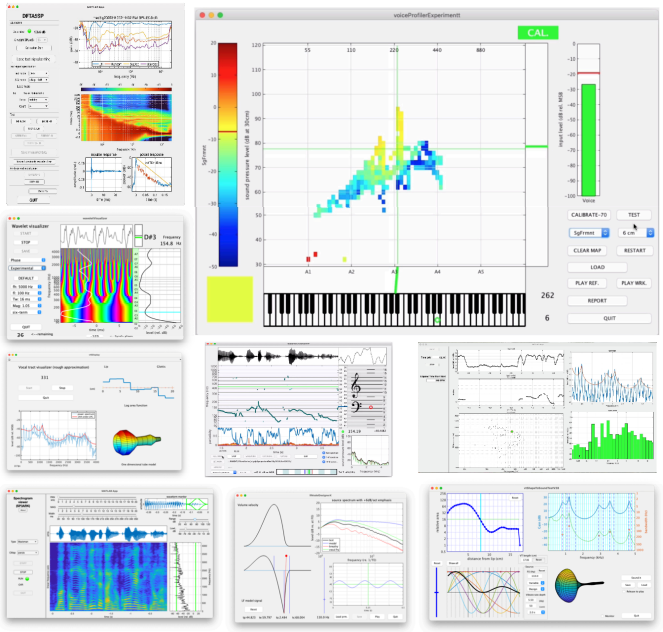


Fig. 7. Snapshots of application GUIs [15]–[18]. This controller launches applications by selecting the dropdown menu in the application launcher sub-panel.

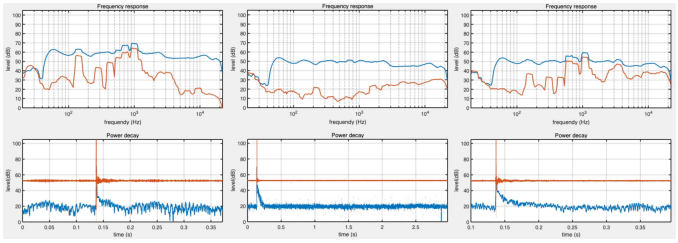


Fig. 8. Snapshots of monitored acoustic conditions. The left shows when the experimenter disturbed the measurement by one's voice /a/. The center shows the monitored LTI response and background noise using a test sound that is a safeguarded utterance (a Japanese sentence /bakuoN ga ginsekai no kougen ni hirogaru/ "A roaring sound spreads across the silvery snow-covered plateau.") spoken by a Japanese male. These are snapshots excerpted from a demonstration movie.

disturbing component (red line). The following graph shows the impulse response waveform (red line) and the smoothed power of the impulse response represented in dB (blue line).

Figure 8 shows monitored results in different conditions. These snapshots are excerpts from a demonstration movie[1].

The left graph shows when the experimenter disturbs the measurement by producing a loud /a/ voice. The disturbing component in the red line pushes the LTI estimate at several parts. The smoothed impulse response power also shows distortions in low-level parts.

The center graph shows the result using a safeguarded read sentence. The length of the period of the test signal used in this measurement is 2^{17}. Figure 16 in Appendix C shows the spectra of the original and safeguarded signals. The measured LTI gain is virtually identical to the result in Fig. 6. However, the level of disturbing components is higher than Fig. 6. It is because the power spectral level in the higher frequency region of the safeguarded test signal decreases steeper than that of the pink noise. Note that the noise floor, due to background noise, masks the smoothed power level of the estimated impulse response other than the initial part of the impulse response.

The right graph shows when the experimenter disturbs the measurement by producing a loud /a/ voice, similar to the left graph. Similarly, the disturbing component in the red line pushes the LTI estimate at several parts. There are differences in the shape of the disturbing components in the left and right graphs. The disturbing level is lower in the low-frequency region and higher in the high-frequency region. These differences are due to the spectral difference of the test signals. The pink noise-shaped test signal has higher energy in the high-frequency region than the safeguarded speech sounds.

V. Other Applications

We developed several tools other than acoustic measurement based on the proposed framework. This section introduces two applications closely related to each other.

The first one is measurement of voice f_o (fundamental frequency[2] This research measures the frequency modulation transfer function using the f_o frequency modulation by the spectrally shaped test signal made from CAPRICEP as the test signal. Then the f_o modulation of the produced voice sound as the output.

The second one is an objective measurement of pitch extractors using the f_o modulation transfer function [22]. In this investigation, we replaced the human with the pitch extractor. The movie that compares 16 pitch extractors using this method is informative[3].

VI. Discussion and Related Works

The proposed framework does not replace existing acoustic measurement methods. Instead, it provides them additional values and is an efficient computational infrastructure. For example, exponential swept-sine analyzes nonlinearity in a diagnostic way [24]. Our proposed framework provides information on the target system's behavior handling signals they are designed to use. In this manner, the proposed framework plays a complementing role in acoustic measurement.

Reverberation radius [23] represents the distance where the energy of the direct and the reverberant sounds equals. It is an essential attribute in, for example, classroom acoustics. Our method enables us to measure the reverberation radius and other acoustic attributes using actual teaching materials in a classroom while teaching students. Moreover, using the smartphones of each student, simultaneous measurements of each student's listening acoustic conditions are possible.

VII. Conclusions

We introduced a general framework for measuring acoustic properties such as liner time-invariant (LTI) response, signal-dependent component, and random and time-varying component simultaneously using structured periodic test signals. The framework also enables music pieces and other sound materials

[1] https://youtu.be/-nxD-8hbCv4

[2] See [19] for the reason why we are using f_o instead of using F0) response to auditory stimulation [20], [21].

[3] https://youtu.be/iXnP1tIuVic

as test signals by "safeguarding" them by adding slight deterministic "noise." Measurement using swept-sin, MLS (Maximum Length Sequence), and their variants are special cases of the proposed framework. We implemented interactive and real-time measuring tools based on this framework and made them open-source. Furthermore, we applied this framework to assess pitch extractors objectively. The proposed framework is general enough and applicable to many other fields.

ACKNOWLEDGMENT

KAKENHI by JST 21K19794, 21H03468, 21H00497, 21K11957 supported this line of research. The authors appreciate for careful and constructive reviewers' comments.

REFERENCES

[1] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuszmaul, B. W. Lampson, and et al., "There's plenty of room at the top: What will drive computer performance after moore's law?" *Science*, vol. 368, no. 6495, p. eaam9744, 2020.

[2] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, and ˙ Xu, "Libri-Light: A benchmark for ASR with limited or no supervision," in *Proc. ICASSP 2020*, May 2020, pp. 7669–7673.

[3] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *An international journal on information fusion*, p. 101869, Jun. 2023.

[4] H. Kawahara, K.-I. Sakakibara, M. Mizumachi, M. Morise, and H. Banno, "Simultaneous measurement of time-invariant linear and nonlinear, and random and extra responses using frequency domain variant of velvet noise," in *Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2020, pp. 174–183.

[5] H. Kawahara and K. Yatabe, "Cascaded all-pass filters with randomized center frequencies and phase polarity for acoustic and speech measurement and data augmentation," in *Proc. ICASSP 2021*. ieeexplore.ieee.org, 2021, pp. 306–310. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9415057/

[6] ——, "Safeguarding test signals for acoustic measurement using arbitrary sounds: Measuring impulse response by playing music," *Acoustical science and technology / edited by the Acoustical Society of Japan*, vol. 43, no. 3, pp. 209–212, 2022. [Online]. Available: http://dx.doi.org/10.1250/ast.43.209

[7] R. R. Patel, S. N. Awan, J. Barkmeier-Kraemer, M. Courey, D. Deliyski, T. Eadie, and et al., "Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing association expert panel to develop a protocol for instrumental assessment of vocal function," *American journal of speech-language pathology*, vol. 27, no. 3, pp. 887–905, Aug. 2018.

[8] K. Sakakibara and H. Kawahara and M. Mizumachi, "Recommended protocols for acoustic recording of speech data for high reusability," *Journal of the Acoustical Society of Japan*, vol. 76, no. 6, pp. 343–350, 2020.

[9] N. Aoshima, "Computer ‐ generated pulse signal applied for sound measurement," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1484–1488, May 1981. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.385782

[10] J. Borish and J. B. Angell, "An efficient algorithmfor measuringthe impulse response using pseudorandom noise," *J. Audio Engineering Society*, vol. 31, no. 7-8, pp. 478–488, 1983. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=4564

[11] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, 1 Jun. 1989. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=6086

[12] A. V. Oppenheim, R. W. Schafer, and R. W. Shaffer, *Discrete-time signal processing*, ser. Prentice Hall signal processing series. London, England: Prentice-Hall, Dec. 1989.

[13] Y. Nakahara, Y. Iiyama, Y. Ikeda, and Y. Kaneda, "Shortest impulse response measurement signal that realizes constant normalized noise power in all frequency bands," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 70, no. 1/2, pp. 24–35, 26 Jan. 2021. [Online]. Available: http://dx.doi.org/10.17743/jaes.2021.0048

[14] H. Kawahara, "GitHub repository for speech and hearing research/education tools," 2023, (retrieved 20 June 2023). [Online]. Available: https://github.com/HidekiKawahara

[15] H. Kawahara, K.-I. Sakakibara, and K. Yatabe, "Singing voice range profiling toolbox with real-time interaction and its application to make recording data reusable," in *Proceedings of the Stockholm Music Acoustics Conference 2023*, S. D'Amario, A. Friberg, and S. Ternström, Eds., 14 Jun. 2023, pp. 160–166. [Online]. Available: https://smcnetwork.org/smc2023/

[16] H. Kawahara, K.-I. Sakakibara, M. Morise, H. Banno, T. Toda, and T. Irino, "A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis," in *Proc. Interspeech 2017*, Stocholm, Aug. 2017, pp. 1358–1362. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-15

[17] H. Kawahara, K.-I. Sakakibara, E. Haneishi, and K. Hagiwara, "Real-time and interactive tools for vocal training based on an analytic signal with a cosine series envelope," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 907–910.

[18] H. Kawahara, K. Yatabe, K.-I. Sakakibara, M. Mizumachi, M. Morise, H. Banno, and T. Irino, "Interactive and Real-Time acoustic measurement tools for speech data acquisition and presentation: Application of an extended member of time stretched pulses," in *Show and Tell Proc. Interspeech 2021*, 2021, pp. 4853–4854. [Online]. Available: https://www.isca-speech.org/archive/pdfs/interspeech_2021/kawahara21b_interspeech.pdf

[19] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D. G. Miller, H. Noé, R. C. Scherer, J. R. Smith, B. H. Story, J. G. Švec, S. Ternström, and J. Wolfe, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015. [Online]. Available: http://dx.doi.org/10.1121/1.4919349

[20] H. Kawahara, T. Matsui, K. Yatabe, K.-I. Sakakibara, M. Tsuzaki, M. Morise, and T. Irino, "Mixture of Orthogonal Sequences Made from Extended Time-Stretched Pulses Enables Measurement of Involuntary Voice Fundamental Frequency Response to Pitch Perturbation," in *Proc. Interspeech 2021*, 2021, pp. 3206–3210.

[21] H. Kawahara, T. Matsui, K. Yatabe, K. I. Sakakibara, M. Tsuzaki, M. Morise, and T. Irino, "Implementation of interactive tools for investigating fundamental frequency response of voiced sounds to auditory stimulation," in *Proc. APSIPA ASC*, Tokyo, Dec. 2021, pp. 897–903.

[22] H. Kawahara, K. Yatabe, K.-I. Sakakibara, T. Kitamura, H. Banno, and M. Morise, "An objective test tool for pitch extractors' response attributes," in *Proc. Interspeech 2022*, 2022, pp. 659–663. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2022-800

[23] M. Mijić and D. Mašović, "Reverberation radius in real rooms," *Telfor journal*, vol. 2, no. 2, pp. 86–91, 2010. [Online]. Available: https://journal.telfor.rs/Published/Vol2No2/Vol2No2_A6.pdf

[24] A. Novak, L. Simon, and P. Lotton, "Analysis, synthesis, and classification of nonlinear systems using synchronized Swept-Sine method for audio effects," *EURASIP journal on advances in signal processing*, vol. 2010, no. 1, pp. 1–8, 20 Jul. 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/793816

APPENDIX

A. Reducing truncation artifacts

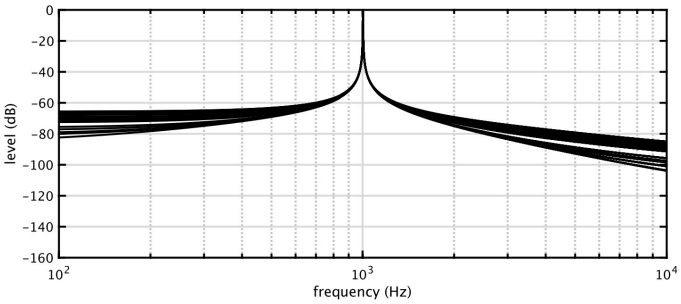


Fig. 9. Power spectra of truncated signal without any shaping the both ends.

Figure 9 shows an example of signal truncation asynchronous to the signal. This example uses a sinusoid having

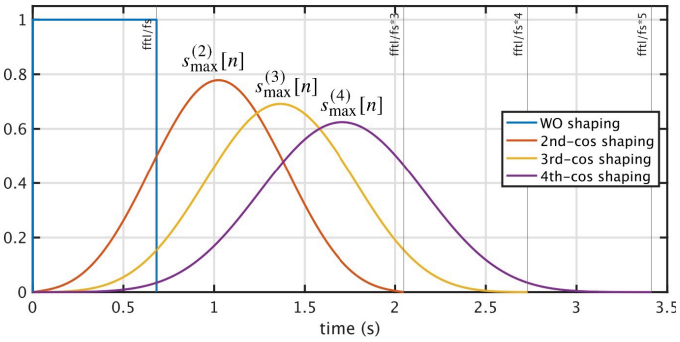


Fig. 10. Periodicity preservation tests of shaping functions. Plot shows the weighting function shapes.

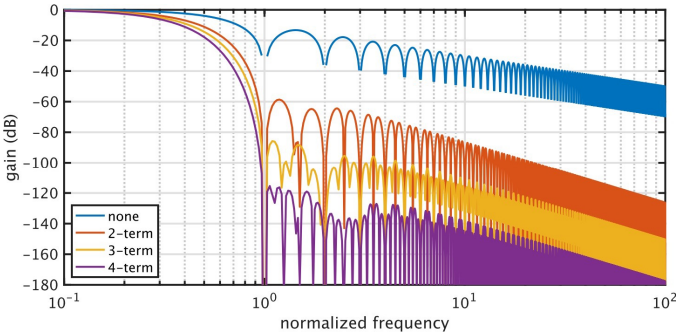


Fig. 11. Optimized weighting functions made from several cosine series convolved with a rectangular function with the repetition period length. They are numerically optimezed to minimize the maximum side-lobe level.

a frequency $1000 + \pi$ Hz. As shown in the figure, truncation produces spectral spread.

Using more than one cycle of the periodic signal and a weighting function reduces this issue. The weighting function has to be a constant value one when wrapped. Figure 10 shows such weighting functions. They are cosine-based window functions convolved with a rectangular function having the period width. The annotation $s^{(k)}_{\max}[n]$ indicates that the convolved cosine series has k-terms. The coefficients are numerically optimized to minimize the maximum side-lobe level.

Figure 11 shows their frequency responses.

Figure 12 shows the same asynchronous signal truncated using the optimized weighting function $s^{(4)}_{\max}[n]$. Note that

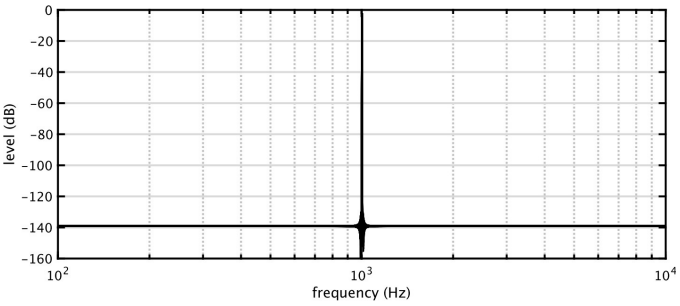


Fig. 12. Power spectra of truncated signal with optimized shaping weight.

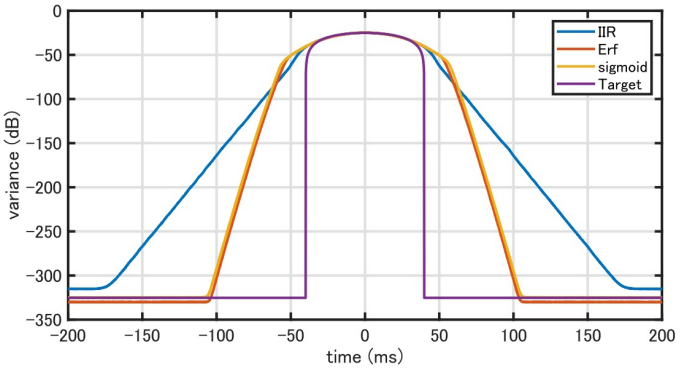


Fig. 13. Designed outline shapes using a raised cosine shape as the target. The vertical axis represents the variance in logarithmic (dB) scale.

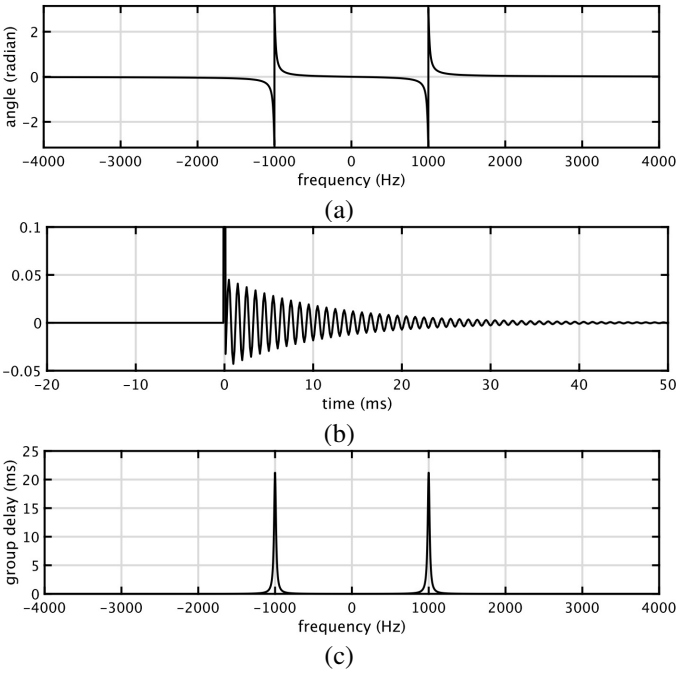


Fig. 14. Attributes of IIR-type all-pass filter. (a) phase-frequency response, (b) impulse response, and (c) group delay response.

weighting effectively suppressed the spectral spreads.

B. *Phase manipulation function*

Let us start with the result. Figure 13 shows the RMS average of the unit-CAPRICEP designed using a phase manipulation function of the original proposal [5] (IIR) and the error function, integrated Gaussian function (Erf) and its approximating function sigmoid (sigmoid).

Figure 14 shows an example of relations between phase manipulation, impulse response, and the group delay of the initially proposed phase function [5]. It is the phase of an IIR all-pass filter. This manipulation example and the following examples use phase manipulation at one point, 1000 Hz. The actual unit-CAPRICEP of 0.25 s effective width consists of about 7000 phase manipulation points.

Figure 15 shows impulse responses using the error function

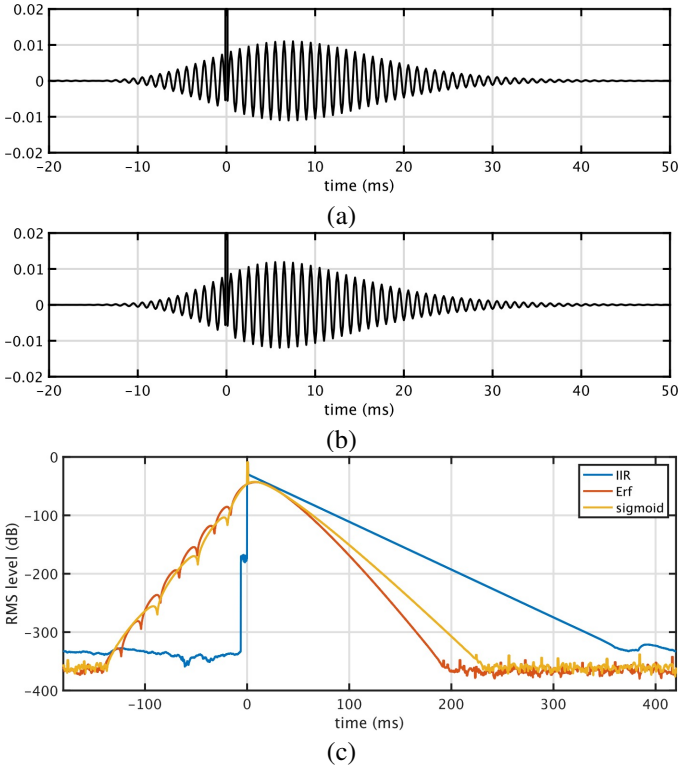


Fig. 15. Impulse response of the all-pass filters made from error function (a) and the sigmoid (b). The plot (c) shows smoothed RMS levels (in dB) of the impulse response of the IIR (original CAPRICEP [5]), the error function (Erf), and the sigmoid (sigmoid).

and the sigmoid. The bottom plot compares IIR, Erf, and Sigmoid's power decay. The decay of Erf is the steepest, and we selected Erf to design unit-CAPRICEP in our implementation.

C. Safeguarded signal and its merits

This is a revised excerpt form [6]. Let $x[n]$ be a periodic discrete-time signal with a period L. Convolution of $x[n]$ and the impulse response $h[n]$ of the target system yields the output $y[n]$. Because the signal is periodic, the DFT (Discrete Fourier transform) of $x[n]$ and $y[n]$ segments (their length is L) are invariant other than the phase rotation proportional to frequency. Let $X[k]$ and $Y[k]$ represent their DFT, where k, $(k = 0, \ldots, L-1)$, is the discrete frequency. Then, the ratio $Y[k]/X[k]$ is independent of the location of the segment. This ratio agrees with the DFT $H[k]$ of the impulse response $h[n]$, where $X[k] \neq 0$ for all k values is the condition of this relation to provide physically meaningful results.

However, this simple solution is sensitive to noise when the absolute value $|X[k]|$ is very small relative to absolute values $|H[k]|$ of other k values. We propose to limit the absolute value $|X[k]|$ to have larger value than a threshold[4]. We use the following equation to derive the DFT $X_s[m]$ of the safeguarded signal $\tilde{x}_s[n]$.

$$X_s[k] = \begin{cases} \dfrac{\theta_L[k]X[k]}{|X[k]|} & \text{for} \quad 0 < |X[k]| < \theta_L[k] \\ X[k] & \theta_L \leq |X[k]| \end{cases}, \qquad (12)$$

where we set $X_s[k] = \theta_L[k]$ when $X[k] = 0$. Then, we derive the safeguarded transfer function $H_s[k]$ as follows.

$$H_s[k] = \frac{Y_s[k]}{X_s[k]}, \qquad (13)$$

where $Y_s[k]$ represents the DFT of the output of the target system for periodic test signal $\tilde{x}_s[n]$. Because the safeguarded signal $\tilde{x}_s[n]$ is periodic, we can make *the safeguarded test signal* for acoustic measurement by concatenating it as many times as required.

Figure 16 shows example of safeguarding. The level represents the absolute values of discrete Fourier transform of a whole sentence length samples (the original and safeguarded ones). Figure 17 illustrates the merit of safeguarding.

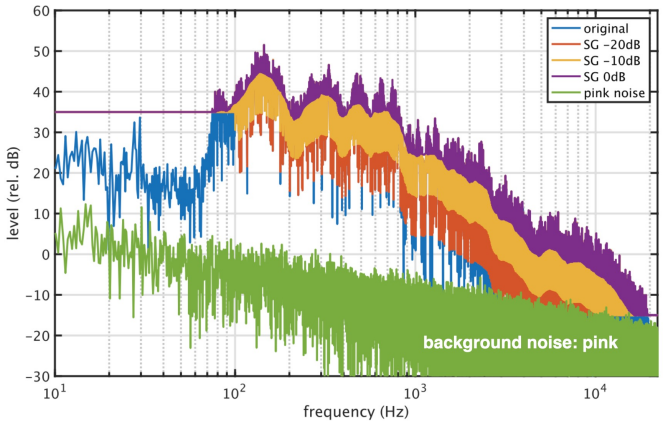


Fig. 16. Absolute values of the original (a Japanese sentence /bakuon ga ginsekai no kougeN ni hirogaru/ spoken by a male speaker) and safeguarded speech samples. The signal period is 2^{17} samples at 44100 Hz. Frequency-dependent thresholding uses a smoothed power spectrum with 1/3 octave width as a reference. The light green line represents the background pink noise for testing noise tolerance.

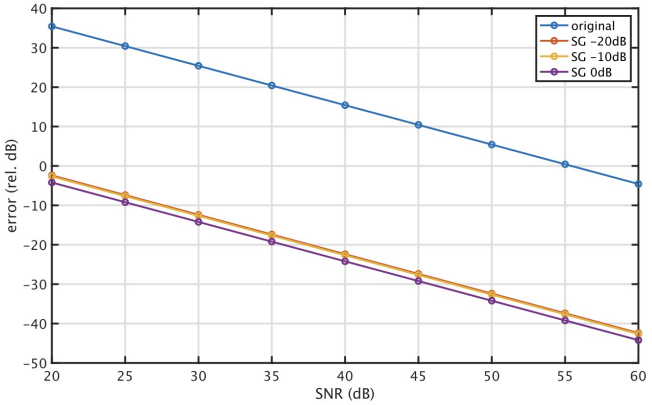


Fig. 17. Estimation error of the impulse responses using the original and the safeguarded signals for the test signal. For noise tolerance test, we added a background pink noise as shown in Fig. 16.

[4]In actual implementation we use frequency dependent threshold $\theta_L[k]$ using power spectra of the original signal and the background noise. We also set the minimum level and low frequency limit for $\theta_L[k]$.

Cognitive Assessment of Autism Spectrum Disorder Using an EEG-based Social Interaction Platform

Yi-Li Tseng[1], Yi-Ling Chien[2], Tse-Min Chuang[1], Yen-Nan Chiu[2], and Wen-Che Tsai[2]
[1]Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[2]Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan
E-mail: yilitseng@mail.nsysu.edu.tw Tel: +886-7-5252000#4186

Abstract — **Previous studies have suggested a strong correlation between social interaction skills and the social environment, as well as personal experiences. Neuroimaging studies have further demonstrated the link between social interaction and brain oscillations. In this study, we developed a game-based social training platform that recorded electroencephalography (EEG) signals to provide neurophysiological indicators for cognitive assessment in autistic teenagers. Both typically developed (TD) participants and individuals with autism spectrum disorder (ASD) were recruited to participate in the social training game, with 12-channel EEG signals recorded during the sessions. The EEG signals underwent preprocessing, and the event-related potentials (ERPs) and event-related spectral perturbations (ERSPs) were analyzed and compared between the two groups. Six machine learning frameworks were employed for computer-aided assessment of ASD. The results revealed that the support-vector machine with a polynomial kernel function achieved the highest balanced accuracy of 91.67% when both ERPs and ERSPs were utilized as features for evaluating autism. These findings suggest the potential use of EEG features incorporated into a game-based social training interface for the cognitive assessment of ASD.**

I. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder classified in the Diagnostic and Statistical Manual of Mental Disorders Fifth Ed. (DSM-5) [1]. Previous research suggests that individuals with ASD often struggle with facial processing and recognizing emotions, which impacts their ability to express emotions and affects their social cognition [2, 3]. Recent studies propose that challenges in social interaction among individuals with ASD can be attributed to impaired social motivation and deficits in basic perception [3-5]. Early interventions and cognitive assessments based on developmental and neuroimaging findings are essential for effectively evaluating the social and cognitive abilities of individuals with autism.

The use of game-based training platforms for behavioral interventions has gained popularity due to the reported enjoyment of individuals with autism in using computer technology [6]. In the preceding decade, a multitude of serious games have been created with the purpose of delivering individualized training and conducting cognitive assessment within domains encompassing emotions, social interactions, and communication [7-10]. These games cover various aspects of training, including behavioral, affective, cognitive functions, and social skills [10]. However, the effectiveness of these games has primarily been evaluated through behavioral questionnaires or verbal feedback. Progress in human-machine interfaces has opened avenues for integrating physiological indicators like heart rate, electroencephalography (EEG), and eye-tracking into forthcoming interfaces based on gaming. Utilizing these physiological signals will enable a deeper understanding of the underlying mechanisms and identification of reliable biomarkers for computer-aided diagnosis, cognitive assessment, and the effectiveness of game interventions.

To address the subjectivity of current behavior-dependent diagnosis and cognitive assessment processes, various neurophysiological and neuroimaging methods have been employed to investigate the relationship between brain functionality and autism behavior [11]. Among these methods, EEG stands out as a valuable tool for investigating and diagnosing ASD [12, 13]. Several studies have successfully detected ASD using EEG [13]. The physiological features extracted from these previous investigations can be broadly categorized into groups, including spectral power [12, 14-16], functional connectivity [17-19], and information dynamics [20].

In this study, we have developed a social cognitive paradigm specifically tailored for autistic teenagers. To evaluate brain oscillations and functional connectivity during the game, we recorded EEG signals using a mobile EEG device. Our

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

previous research [21] has demonstrated that deficits in joint attention and basic gaze perception can hinder the development of social cognition in individuals with ASD. Building upon these findings, our hypothesis for this study was that altered brain oscillations could be observed in autistic preadolescents during social interaction, and these patterns could serve as promising biomarkers for assessing social abilities in autistic preadolescents.

II. Materials and Methods

A. Participants

This study recruited 18 participants, consisting of 12 typically developing (TD) preadolescents (aged 11.83 ± 1.27; seven boys) and 6 autistic preadolescents and adolescents (aged 13.83 ± 3.71; four boys). Participants with a history of head trauma, neurological, and physiological diseases were excluded. All autistic participants were recruited from National Taiwan University Hospital, and diagnosed according to the diagnostic criteria of the DSM-5 [1] and International Classification of Diseases (ICD-10; WHO 1994). Approval for this study was granted by the Research Ethics Committee at National Taiwan University Hospital, Taiwan (REC no. 201709023RIPD). Written informed consent was obtained from all legal guardians of the participants before the commencement of the experiment.

B. Game-based Interface for Social Cognitive Training

A game-based training interface was designed including several interactive scenarios with different topics for autistic preadolescents and teenagers to learn, such as recognition of facial emotional expressions and social interaction skills. This game-based social cognitive interface is divided into events to fit the criteria of designing an EEG experimental paradigm. The Unity game engine (Unity Technologies 2018.1.14f) was used to develop the games and synchronize the timing with a mobile EEG system. As depicted in Fig. 1A, the game-based interface comprises two modules, each operating with distinct modes. The initial module introduces an intricate version of the facial emotional recognition game, encompassing fundamental facial expression techniques and attention-sharing elements. The second module integrates the first one and uses role-playing and first-person scenarios to simulate social interactions close to real life (Fig. 1B). Proper rewards are offered in the game based on the ages and preferences of the participants. The resultant game for social cognitive training is a 30-minute, personalized interactive computer-based activity. It encompasses 84 trials focused on facial emotion recognition and 45 trials centered around selecting social behaviors. These



Fig. 1 The game-based social cognitive paradigm with scenarios for the selection of (A) emotional expressions and (B) appropriate behaviors.

trials are synchronized with the wireless EEG system through triggers conveyed via the lab streaming layer. The game aims to evaluate the abilities of facial emotional recognition and social behavioral reactions. Within each trial, a 12-second scenario description is presented. Subsequently, participants are directed to establish eye contact with a character displayed on the screen and correctly identify the corresponding facial emotional expression aligned with the given social context (taking 1-3 seconds). Lastly, participants are prompted to choose a suitable behavior or reaction in accordance with the character's social interactions (allowing 7-10 seconds). Subsequent analysis focused on the participants' behavioral and electrophysiological responses within these two conditions.

C. EEG Recording and Signal Processing

A mobile and dry-electrode EEG headset (Quick-20 Dry EEG Headset, Cognionics, Inc., San Diego, CA) was used for EEG recording while the participants were playing the social training game. According to the 10–20 system, EEG signals from 12 scalp locations were recorded, including F3, F4, F7, F8, Fz, C3, C4, P3, P4, P7, P8, and Pz, with a reference on A1.

The sampling rate of the EEG signals was 500 Hz, and the signals were then amplified with a band-pass filter with a bandwidth of 0 to 50 Hz. The stimuli triggers were synchronized with EEG data streams and consolidated into a unified file employing the LSL data acquisition and synchronization framework. EEG data were analyzed using MATLAB R2021b (The MathWorks, Inc., Natick, Mass, USA). For the preprocessing of EEG data, independent component analysis and artifact subspace reconstruction from the EEGLAB toolbox [22-24] were used to identify and remove artifact components and trials for each subject. The cleaned and epoched EEG data then underwent Morlet wavelet transformation to retrieve the event-related potentials or oscillations (ERPs or ERSPs). Group-level averaged ERSPs were computed relative to the baseline period for each channel location. The parametric paired t-test was conducted for each time-frequency interval to examine the difference of brain oscillations between the two groups at distinct frequency bands.

D. Computer-aided assessment

The machine learning-based classification process involves feature extraction and classification. To extract representative features, the F score is employed to rank all potential features extracted from ERPs and ERSPs. For the classification task, six different classifiers are employed, namely k-Nearest Neighbors (kNN) with a neighbor number of ten (k = 10), Binary Decision Tree (BDT), Gentle Adaptive Boosting (GAB), Support Vector Machine (SVM) with linear, radial, and polynomial kernel functions. he optimization for classification was resolved using the SVM toolbox function LibSVM [25] incorporated in MATLAB R2021b. The performance of the classification methods was evaluated by leave-one-subject-out validation due to the limited number of participants. Groups of training sets and testing sets were fed into the classifiers. The categorization performance of the proposed framework is measured using sensitivity, specificity, F1 score, and balanced accuracy averaged after 100 iterations.

III. Results and Discussions

A. Rule-based feature extraction

The ERPs were compared between autistic and TD participants in Fig. 2 during facial emotional recognition. The average ERPs in the parietal regions (Pz and P7) were compared. Significant differences were observed around 200-240 ms and 550-750 ms between the two groups. Increased amplitudes of P2 components between 200 and 240 ms after the stimulus onset were observed at P7 in autistic preadolescents and adolescents (Fig. 2A). Fig. 2A presents the

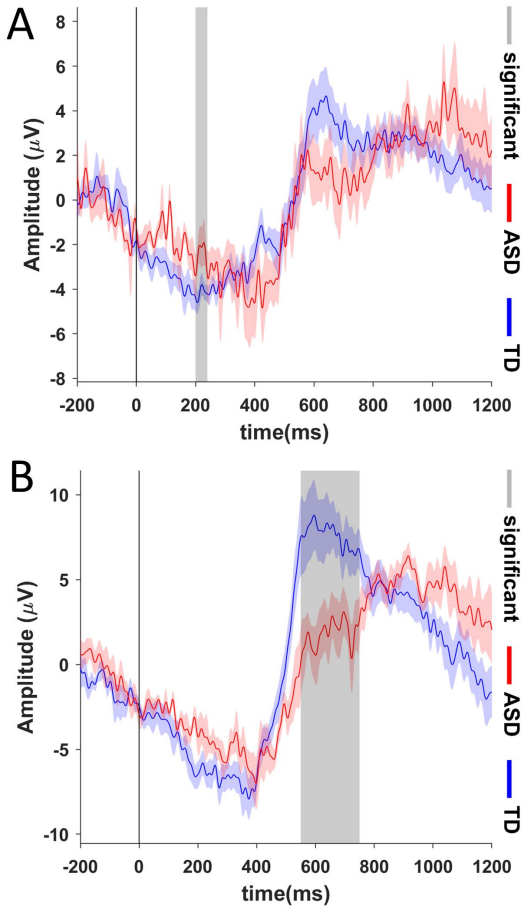


Fig. 2 The average ERPs during facial emotional recognition of the social cognitive game in the following time intervals in the parietal regions: (A) 200-240 ms (P2 component) at P7, and (B) 550-750 ms (LPP component) at Pz.

collective mean amplitudes and grand averages of the P2 waveform at channel P7. When contrasted with the TD group's mean of -4.18 ± 1.53 μV, the ASD preadolescents' mean of -2.44 ± 1.72 μV indicated a relatively augmented P2 ($p = 0.041$). This finding suggests an impairment in the structural encoding of facial features among the ASD group. The grand averages of the LPP waveforms at the Pz channel and the corresponding mean amplitudes are shown in Fig. 2B, which can be considered as a component to capture the allocation of attention toward motivational salient objects [26-31]. The autistic participants exhibited a smaller LPP (1.79 ± 3.58 μV) than the TD group (7.60 ± 4.72 μV, $p = 0.024$). This implies that preadolescents and adolescents with autism encountered challenges in directing attention towards the target stimulus.

The results of ERSPs revealed that brain oscillations in the delta, theta, and beta frequency bands showed significant differences between the two groups during facial emotional

recognition of the social-cognitive game. In TD preadolescents, there was a notable increment of theta (5-8 Hz) synchronization around 400 ms in the left frontal (F7) regions. However, this phenomenon was not observed in autistic participants ($p < 0.001$). In contrast, larger delta (1-4 Hz) synchronization is observed in autistic preadolescents in the left frontal ($p < 0.001$), right central ($p = 0.022$), and right parietal ($p < 0.001$) regions, as well as larger beta desynchronization between 15 and 20 Hz in the left frontal ($p < 0.001$) and right central ($p = 0.006$) regions after 500 ms.

Table 1 Performance for different machine learning classifiers (kNN: k-nearest neighbors; BDT: Binary decision tree; GAB: Gentle adaptive boosting)

Classifier	*Sensitivity(%)*	*Specificity(%)*	*F1 score*	*Balanced accuracy(%)*
kNN	50	100	0.67	75.00
BDT	83	75	0.71	79.17
GAB	67	83	0.67	75.00
SVM-linear	83	92	0.83	87.50
SVM-poly	**83**	**100**	**0.91**	**91.67**
SVM-radial	67	92	0.73	79.17

B. Results of machine learning

Table 1 presents the overall performance of the six classifiers mentioned. Among them, the SVM-based approach with a polynomial kernel function demonstrates the highest overall performance with a balanced accuracy of 91.67%, while the kNN and GBA-based methods show the lowest overall performance with balanced accuracy of 75.00%. In terms of sensitivity, the BDT classifier and SVM framework with linear and polynomial kernel functions achieve the highest value of 83%, and kNN has the lowest sensitivity of 50%. On the other hand, kNN and SVM with a polynomial kernel exhibit the highest average specificity value of 100%, and BDT has the lowest specificity value of 75%. In summary, SVM achieves the highest performance in terms of balanced accuracy and maintains a good balance between sensitivity and specificity, resulting in F1 scores and accuracy values of 0.91.

IV. Conclusions

Rather than relying on behavioral questionnaires, we employed an EEG-compatible social interaction game to assess cognitive abilities. We analyzed the altered ERPs and ERSPs in preadolescents with autism during the game. Our results indicate that these atypical EEG responses can be dependable indicators for assessing emotional recognition, social information processing, and cognitive functions in autistic teenagers. Specifically, we noted modified amplitudes of the P2 and LPP components, along with decreased theta synchronization and beta frequency band desynchronization in individuals with autism. Using SVM, computer-aided assessment of autism achieved a high performance with a balanced accuracy of up to 91.67%. This study highlights the potential of incorporating EEG as an indicator for computer-aided diagnosis and evaluation of cognitive performance during real-life game playing, given its portability and flexibility compared to other neuroimaging techniques.

Acknowledgment

This work was supported in part by the Taiwan National Science and Technology Council under Grant 110-2222-E-110-007-MY2 and 112-2221-E-110-021.

References

[1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[2] T. F. Gross, "The perception of four basic emotions in human and nonhuman faces by children with autism and other developmental disabilities," *J Abnorm Child Psych,* vol. 32, no. 5, pp. 469-480, Oct 2004.

[3] M. Behrmann, C. Thomas, and K. Humphreys, "Seeing it differently: visual processing in autism," *Trends Cogn Sci,* vol. 10, no. 6, pp. 258-64, Jun 2006.

[4] Y.-L. Tseng, H. H. Yang, A. N. Savostyanov, V. S. Chien, and M. Liou, "Voluntary attention in Asperger's syndrome: Brain electrical oscillation and phase-synchronization during facial emotion recognition," *Research in Autism Spectrum Disorders,* vol. 13, pp. 32-51, 2015.

[5] V. S. Chien, A. C. Tsai, H. H. Yang, Y.-L. Tseng, A. N. Savostyanov, and M. Liou, "Conscious and Non-conscious Representations of Emotional Faces in Asperger's Syndrome," *JoVE (Journal of Visualized Experiments),* no. 113, pp. e53962-e53962, 2016.

[6] S. V. Wass and K. Porayska-Pomsta, "The uses of cognitive training technologies in the treatment of autism spectrum disorders," *Autism,* p. 1362361313499827, 2013.

[7] E. Vallefuoco, C. Bravaccio, G. Gison, L. Pecchia, and A. Pepino, "Personalized Training via Serious Game to Improve Daily Living Skills in Pediatric Patients with Autism Spectrum Disorder," *IEEE J Biomed Health Inform,* vol. PP, Mar 1 2022, doi: 10.1109/JBHI.2022.3155367.

[8] C. Grossard, O. Grynspan, S. Serret, A.-L. Jouen, K. Bailly, and D. Cohen, "Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (ASD)," *Computers & Education,* vol. 113, pp. 195-211, 2017, doi: 10.1016/j.compedu.2017.05.002.

[9] J. S. Y. Tang, N. T. M. Chen, M. Falkmer, S. Bölte, and S. Girdler, "A systematic review and meta-analysis of social emotional computer based interventions for autistic individuals using the serious game framework," *Research in Autism Spectrum Disorders,* vol. 66, 2019, doi: 10.1016/j.rasd.2019.101412.

[10] G. M. Silva, J. J. S. Souto, T. P. Fernandes, I. Bolis, and N. A. Santos, "Interventions with Serious Games and Entertainment Games in Autism Spectrum Disorder: A Systematic Review," *Dev Neuropsychol,* vol. 46, no. 7, pp. 463-485, Oct 2021, doi: 10.1080/87565641.2021.1981905.

[11] L. Billeci *et al.*, "On the application of quantitative EEG for characterizing autistic brain: a systematic review," *Front Hum Neurosci,* vol. 7, p. 442, 2013, doi: 10.3389/fnhum.2013.00442.

[12] L. J. Gabard-Durnam, C. Wilkinson, K. Kapur, H. Tager-Flusberg, A. R. Levin, and C. A. Nelson, "Longitudinal EEG power in the first postnatal year differentiates autism outcomes," *Nat Commun,* vol. 10, no. 1, p. 4188, Sep 13 2019, doi: 10.1038/s41467-019-12202-9.

[13] O. Gurau, W. J. Bosl, and C. R. Newton, "How Useful Is Electroencephalography in the Diagnosis of Autism Spectrum Disorders and the Delineation of Subtypes: A Systematic Review," *Front Psychiatry,* vol. 8, p. 121, 2017, doi: 10.3389/fpsyt.2017.00121.

[14] S. Alhassan, A. Soudani, and M. Almusallam, "Energy-Efficient EEG-Based Scheme for Autism Spectrum Disorder Detection Using Wearable Sensors," *Sensors (Basel),* vol. 23, no. 4, Feb 16 2023, doi: 10.3390/s23042228.

[15] J. Zhao, J. Song, X. Li, and J. Kang, "A study on EEG feature extraction and classification in autistic children based on singular spectrum analysis method," *Brain Behav,* vol. 10, no. 12, p. e01721, Dec 2020, doi: 10.1002/brb3.1721.

[16] C. L. Wilkinson, A. R. Levin, L. J. Gabard-Durnam, H. Tager-Flusberg, and C. A. Nelson, "Reduced frontal gamma power at 24 months is associated with better expressive language in toddlers at risk for autism," *Autism Res,* vol. 12, no. 8, pp. 1211-1224, Aug 2019, doi: 10.1002/aur.2131.

[17] E. V. Orekhova *et al.*, "EEG hyper-connectivity in high-risk infants is associated with later autism," (in English), *Journal of Neurodevelopmental Disorders,* vol. 6, Nov 7 2014, doi: Artn 40 10.1186/1866-1955-6-40.

[18] R. Haartsen, E. J. H. Jones, E. V. Orekhova, T. Charman, M. H. Johnson, and B. team, "Functional EEG connectivity in infants associates with later restricted and repetitive behaviours in autism; a replication study," *Transl Psychiatry,* vol. 9, no. 1, p. 66, Feb 4 2019, doi: 10.1038/s41398-019-0380-2.

[19] A. Dickinson *et al.*, "Multivariate Neural Connectivity Patterns in Early Infancy Predict Later Autism Symptoms," *Biol Psychiatry Cogn Neurosci Neuroimaging,* vol. 6, no. 1, pp. 59-69, Jan 2021, doi: 10.1016/j.bpsc.2020.06.003.

[20] W. J. Bosl, H. Tager-Flusberg, and C. A. Nelson, "EEG Analytics for Early Detection of Autism Spectrum Disorder: A data-driven approach," *Sci Rep,* vol. 8, no. 1, p. 6828, May 1 2018, doi: 10.1038/s41598-018-24318-x.

[21] Y. L. Chien *et al.*, "Game-Based Social Interaction Platform for Cognitive Assessment of Autism Using Eye Tracking," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 31, pp. 749-758, 2023, doi: 10.1109/TNSRE.2022.3232369.

[22] S. Makeig, "Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones," *Electroencephalogr Clin Neurophysiol,* vol. 86, no. 4, pp. 283-93, Apr 1993.

[23] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods,* vol. 134, no. 1, pp. 9-21, 2004.

[24] T. R. Mullen *et al.*, "Real-Time Neuroimaging and Cognitive Monitoring Using Wearable Dry EEG," *IEEE Trans Biomed Eng,* vol. 62, no. 11, pp. 2553-67, Nov 2015, doi: 10.1109/TBME.2015.2481482.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," ed, 2001.

[26] R. T. Knight, "Decreased response to novel stimuli after prefrontal lesions in man," *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section,* vol. 59, no. 1, pp. 9-20, 1984.

[27] J. A. Rushby, R. J. Barry, and R. J. Doherty, "Separation of the components of the late positive complex in an ERP dishabituation paradigm," *Clinical Neurophysiology,* vol. 116, no. 10, pp. 2363-2380, 2005.

[28] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical neurophysiology,* vol. 118, no. 10, pp. 2128-2148, 2007.

[29] G. Hajcak, J. P. Dunning, and D. Foti, "Motivated and controlled attention to emotion: time-course of the late positive potential," *Clinical neurophysiology,* vol. 120, no. 3, pp. 505-510, 2009.

[30] E. S. Kappenman, A. MacNamara, and G. H. Proudfit, "Electrocortical evidence for rapid allocation of attention to

threat in the dot-probe task," *Social cognitive and affective neuroscience,* vol. 10, no. 4, pp. 577-583, 2015.

[31] A. Weinberg and G. Hajcak, "The late positive potential predicts subsequent interference with target processing," *Journal of cognitive neuroscience,* vol. 23, no. 10, pp. 2994-3007, 2011.

Gait Analysis in Powered Exoskeleton-Assisted Walking in Patients with Stroke: A Case Series Cohort

Jian-Jia Huang [1,2*], Shih-Chieh Chang [1], Cheng-Hsu Cheng [3], Timothy Wan [4], and Yu-Cheng Pei [1,2 †]

[1] Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taiwan

[2] Master of Science Degree Program in Innovation for Smart Medicine, Chang Gung University, Taoyuan City, Taiwan

[3] Department of Electrical Engineering, Chang Gung University, Taoyuan City, Taiwan

[4]Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA

[†]E-mail: yspeii@gmail.com.tw Tel: +886-3-328-1200 (ext. 3846)

[*]E-mail: jjhuang0919@gmail.com.tw Tel:+886-3-328-1200 (ext. 3846)

Abstract* — *This case series cohort study investigated gait characteristics based on sensor data obtained from personalized active powered lower limb exoskeletal robot-assisted gait training and assessed its putative therapeutic effects in patients with stroke. Preliminary findings suggest that the application of the exoskeletal robot did not improve gait symmetry under robot assistance and could facilitate inter-joint coordination within the hip-knee joint movement on the affected side. These findings suggest that robot-assisted gait training could improve gait by guiding motor learning.

I. INTRODUCTION

Cerebrovascular disease, a leading cause of disability worldwide, particularly causes hemiplegia even following a full course of rehabilitation, a condition inducing a partial or complete loss of ambulatory ability and significant healthcare burdens. However, advancements in the field of exoskeletal rehabilitation robots present a promising opportunity for aiding stroke patients in their ambulation training.

Among these developments, the personalized lower limb exoskeletal robot (PLER) offers advantages in the realm of stroke rehabilitation. It provides personalized and highly repetitive training, which fits the unique needs of each patient and compensates for their specific gait. This adaptability sets the PLER apart from conventional rehabilitation methods. However, it is important to acknowledge that the application and setup of the PLER involve a complex and laborious process. The gait control of the robot relies on a variety of parameters, and each patient presents varying degrees of strength, posture, spasticity, and symmetry. Consequently, configuring these parameters during the early fitting phase is challenging and thus demands significant expertise from engineers and clinical professionals. In this sense, a rapid determination of the optimal control parameters within a short timeframe will be pivotal for enhancing usability, patient satisfaction, and overall clinical effectiveness.

It is important to apply appropriate methods for analyzing the gait function when a subject is using a robotic device. Spatiotemporal gait parameters, such as gait velocity[1] and gait symmetry[2, 3], are widely used and are helpful in measuring walking performance. Gait symmetry is a vital complementary measure to gait velocity[3]. Clinically, the reduction of asymmetry is commonly addressed by therapists particularly in the post-stroke population[4]. Gait symmetry can be considered an indicator of the degree of gait control as it is a summary measure of gait parameters between the two lower limbs. Although previous studies have introduced several symmetry indexes for gait analysis, these indexes have certain limitations, such as low sensitivity [2], reliance on a single point or a limited set of points from waveforms to determine symmetry, and failure to analyze the entire curve or gait cycle comprehensively.

The inter-joint coordination can be quantified by a range of techniques that provide differing complexity and accuracy of results [6]. The cyclogram, also called the angle-angle diagram, proposed by Grieve[5], might be used to represent inter-joint coordination, the functional relationship between joint pairs during a whole gait cycle. The cyclogram is formed by plotting two angles directly, such as the alteration of hip and knee joint angles, in an X-Y plane without time information. It is possible that the cyclogram can be used clinically by analyzing geometric properties, such as its range of motion, area, and perimeter, as studies have supported the utility of the cyclogram for orthopedic [7, 8] and neurologic conditions [9, 10].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In the present case series cohort study, the strategy of personalized parameter setting of PLER was conducted by asking the patients to perform their best gait symmetry during the robot-assisted walking. We also recorded the gait signal before and during each robot-assisted walking, and then analyzed these signals using cyclogram-based spatiotemporal analysis. Finally, we analyzed the treatment effect by clinical assessment. The specific aims of this work are (1) to investigate the effects of the use of personalized parameter setting with performing a better gait symmetry in robot-assisted training, and (2) to explore indexes for monitoring the change of gait induced by the robot-assisted training.

II. Materials and Methods

A. Participants

The participants with stroke, whose age is ≧ 20 years, and without exclusion criteria of: (1) Individuals who are unable to stand and walk even with the support of lower limb exoskeleton robots; (2) vulnerable populations prone to injuries; (3) non-native speakers; (4) participants with conditions such as aphasia, moderate to severe cognitive impairment (e.g., moderate to severe dementia), or mental disorders (e.g., delirium) that hinder their ability to cooperate with the experiment; (5) individuals with severe musculoskeletal disorders of the lower limbs (e.g., fractures, tendon ruptures) that make them unable to withstand external forces; (6) individuals with skin disorders (e.g., pressure sores, wounds, cellulitis) that prevent them from using lower limb exoskeleton rehabilitation robots; (7) individuals with complete paralysis of the lower limbs or unhealed ulcers; (8) individuals with unilateral or bilateral deep vein thrombosis of the lower limbs or amputations; (9) individuals with severe hypertension who cannot maintain resting blood pressure within 180/100 millimeters of mercury (mmHg) under medical control; (10) Individuals with severe heart diseases, including a history of hospitalization for myocardial infarction or cardiac surgery within the past 3 months, congestive heart failure, severe and unstable arrhythmias, hypertrophic cardiomyopathy, severe aortic valve stenosis, angina, or respiratory difficulties during rest or daily activities; (11) individuals with joint diseases of the hip or knee joints that limit sufficient active range of motion (AROM) for walking, sitting, or squatting. Hip joint AROM < 60° or knee joint AROM < 90°, or knee joint flexion contracture > 30°; (12) individuals with spasticity of the lower limbs, with a modified Ashworth Scale (MAS) score greater than 3; (13) individuals with severe pain when standing, with a pain scale score (Numerical Rating Scale, NRS) greater than 6; (14) individuals known to have severe osteoporosis, with bone density more than 2.5 standard deviations below the average value for young adults, and with one or more osteoporotic fractures; (15) individuals experiencing epileptic seizures.

B. Set-ups and Protocols

The active personalized lower limb exoskeletal robot (Keeogo, Wistron Medical Technology, Taiwan) was utilized in this study. The device was fitted to the participants' limbs with the aim of providing power assistance to their walking ability. It consists of bilateral motors that assist the left and right knees (Fig. 1). The knee motor offers assistance to the participant during specific phases of walking without initiating or terminating movements, while the hip motion of the device is passive [11]. The right and left hips and knees are equipped with angle, angular velocity, angular acceleration, and gyro sensors. We recorded these signals during walking with a sampling period of 120ms.

The participant would receive a training a robot-assisted gait training 40 min/ session, 1 session a day, 2-3 sessions a week, 6 sessions in total. Before each training, the parameters of knee motors would be set according to the participant's motor ability for well supporting (Fig. 2A-C).

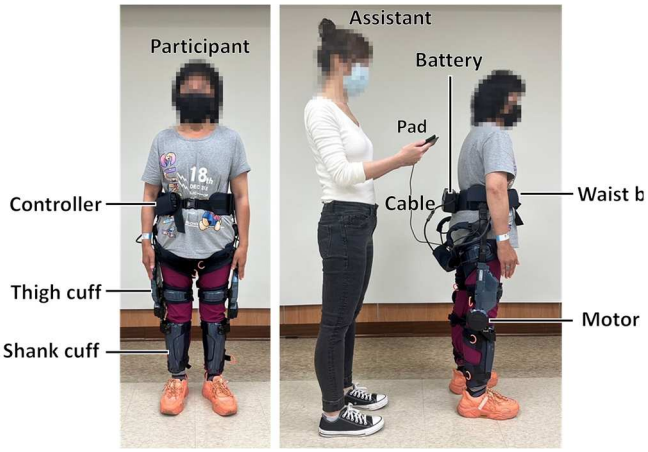


Fig. 1 The active personal exoskeleton robotic device.

C. Kinematics Analysis in Robot-assisted Training Using Cyclograms

The curves depicting angle variations at the hip and knee joints in the sagittal plane (Fig. 2D) were subjected to processing to generate separate hip-knee cyclograms for the right and left limbs. Additionally, the subsequent analysis involved determining various properties from these cyclograms, including the following values:

1. Symmetry of joints[12, 13]: the dimensionless parameter quantifies the similarity between two waveforms,

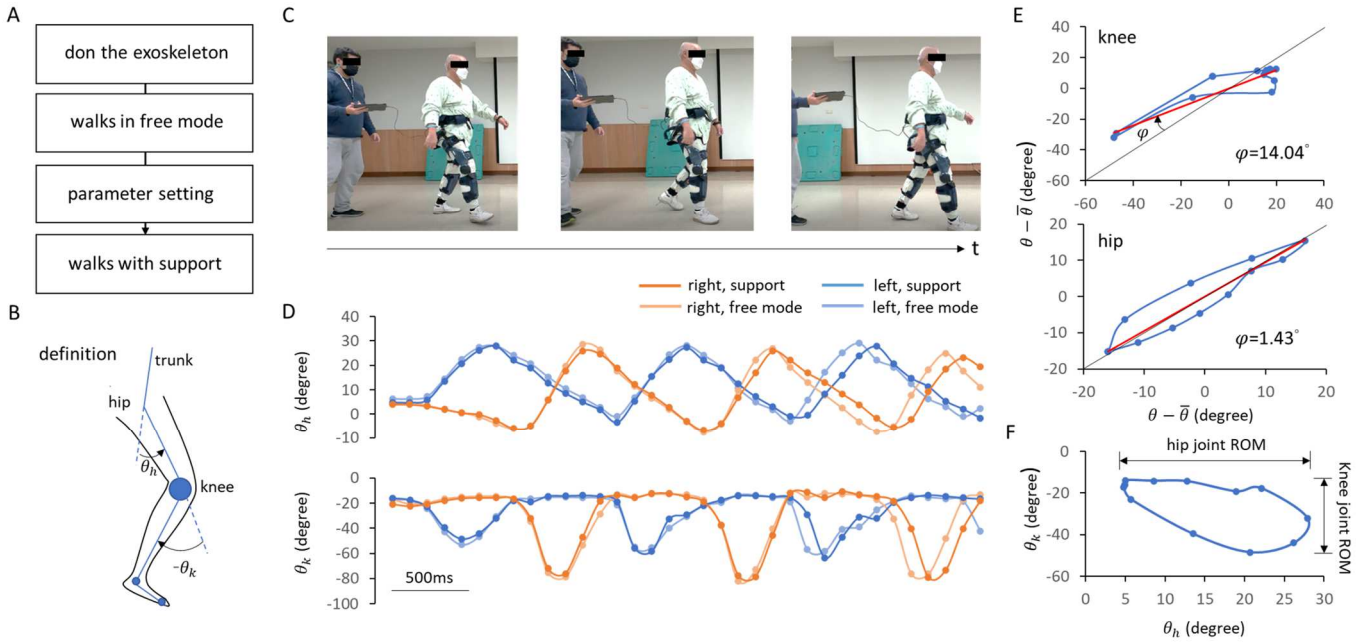


Fig. 2 The experimental protocol and gait analysis. (A) Experimental protocol; (B) Hip and knee angle definition; (C) Signal collection during robot-assisted walking; (D) Signals collected in free mode (no power support) and power supporting mode; (E) Symmetry analysis of hip and knee joints; (F) Inter-joint coordination in a single gait cycle

specifically the angular trends of the right and left legs across gait cycles for each joint of the hip and knee. It is calculated using eigenvector analysis, where the variability around the eigenvector is divided by the variability along the eigenvector. Notably, this parameter remains unaffected by waveform shifts or magnitude differences between the two waveforms (Fig. 2E). A value of 0 indicates perfect symmetry and increasing values of the trend symmetry (TS) parameter indicate higher levels of asymmetry.

$$\begin{Bmatrix} X_R \\ X_L \end{Bmatrix} = \begin{Bmatrix} X_{Ri} \\ X_{Li} \end{Bmatrix} - \begin{Bmatrix} X_{mR} \\ X_{mL} \end{Bmatrix} \quad (1)$$

$$\mathrm{TS} = \frac{emin}{emax} * 100 \quad (2)$$

where the symbols R and L represent the right and left sides, respectively. The variable 'm' denotes the mean value calculated from *'i'* data points within a waveform, where *'i'* represents each individual data point. Additionally, 'emin' and 'eman' represent the minimum and maximum eigenvectors, respectively.

2. Cyclogram orientation (CO)[12, 13]: this feature is characterized by two aspects, the absolute value of angle ϕ formed by the 45° line, which represents perfect interlimb symmetry, and the orientation of the principal axis of inertia, which corresponds to the minimum moment of inertia of the cyclogram. The latter was determined by calculating the eigenvector direction of the inertia matrix derived from the distribution of cyclogram points in the x-y reference system (representing the left joint angle–right joint angle, Fig. 2E). Smaller values of this angle indicate a higher degree of inter joint symmetry, as below:

$$\mathrm{CO} = 45 - (\frac{180}{\pi})\arctan(\frac{e1}{e2}) \quad (3)$$

where e1 and e2 denote the elements of eigenvectors.

3. Cyclogram perimeter (L_i)[5, 9]: The length of the trajectory reflects the average joint velocity and the distance covered by the hip and knee joints. As such, the perimeter can be considered as a measure of the coordination between the hip and knee joints during the gait cycle (Fig. 2F), as below:

$$L_i = \sqrt{\left(\theta_{h_i} - \theta_{h_{i+1}}\right)^2 + \left(\theta_{k_i} - \theta_{k_{i+1}}\right)^2}$$

$$= \Delta t \sqrt{{\omega_{h_i}}^2 + {\omega_{k_i}}^2} \quad (4)$$

where $\theta_{h_i}$and θ_{k_i} represent the hip and knee joint angles, respectively, at point i. The variables ω_{h_i} and ω_{k_i} represent the average angular velocities of the hip and knee joints, respectively, within a specific time interval Δt.

4. *Cyclogram area[5, 9]:* the area of the closed trajectory corresponds to the simultaneous angular variation occurring at the two joints of interest during the gait cycle. A larger area typically signifies a greater conjoint range of angular movements experienced at a specific joint pair throughout a complete gait cycle (Fig. 2F), as below:

$$\text{Area} = \frac{1}{2}\sum_i(\theta_{h_i}\theta_{k_{i+1}} - \theta_{h_i+1}\theta_{k_i}) \quad (5)$$

D. Assessment and Data Analysis

The training efficacy was assessed with outcomes to measure performance change before and after robot-assisted training. The assessment was performed by the other medical licensed physical therapists on the participant without wearing an exoskeleton robot.

The 6-minute walking test (6MWT, distance in meters) was used to evaluate gait function and endurance [14], higher scores (distance walked) means better gait endurance. The lower extremity portion of the Fugl-Meyer assessment (FMA-LE) is the most commonly used clinical assessment tool for identifying motor recovery of the paretic lower limb in patients with stroke [15]. The Berg Balance Test (BBS) is an excellent measurement tool with high reliability and can be used to measure static and dynamic balance in patients with stroke [16]. Timed Up and Go Test (TUG) is also a reliable tool for assessing lower limb mobility in people after stroke [17]. Barthel Index (BI) is one of the most common scales used in clinics to understand the independence of daily activities, ranging from 0 (completely dependent) to 100 (independent) [18]. The EuroQoL-5 Dimension (EQ-5D), a self-reported health status assessment, comprises five dimensions corresponding to mobility, self-care, routine activities, pain or discomfort and anxiety or depression as part of a total health state, with higher scores indicating more severe or frequent problems [19]. The Lawton–Brody instrumental activities of daily living (IADL), which comprises eight items corresponding to making phone calls, shopping, preparing food, housekeeping, laundering, taking pills, using transportation and managing finances, is used to evaluate how patients engage in activities of daily living [20]. The five-meter walking speed (5MWS) assesses longitudinal changes in walking speed after stroke [21].

III. RESULTS

A. Participant Characteristics

Table 1 presents the characteristics of the participants. We recruited three stroke patients whose stroke onset periods were 1, 9, and 23 months, respectively. The affected side for participants #1 and #3 was the left, while participant #2 had a right-side stroke. All participants successfully completed the six training sessions, and no adverse events were observed in the study.

Table 1. Characteristics of stroke patients

Participants	#1	#2	#3
Gender	M	M	M
Age (years)	51	26	54
Affected limb	L	R	L
Time since stroke(months)	1	9	23
mRS	4/3	3/2	3/3
BI	85/95	65/85	70/70
EQ-5D	7/5	8/7	10/10
IADL	3/4	2/2	2/2
BBS	36/47	37/43	39/43
5WMS (s)	7.53/4.78	10.1/7.39	9.12/7.5
6MWT (m)	162/432	31/189	225/270
FMA-LE	21/30	10/18	18/33
TUG (s)	20.35/12.36	25.5/16.34	22.85/15.08

The results of pre-/post- intervention shown in the assessment data.

B. Kinematics Analysis in Robot-assisted Gait Training

Figure 3 illustrates the inter-joint coordination during walking with robot assistance. We compared the robot-assisted walking in free mode versus the power-supported mode, as it was observed that the power-supported mode enhanced the range of motion for hip-knee coordinated gait on the affected side. Additionally, the ranges of inter-joint movement between the

hip and knee increased over time. For instance, the movement ranges during the 3rd and 6th training sessions were greater than those observed during the 1st training session, inparticular the range of knee joint movement.

We conducted a further analysis of the symmetry of the hip and knee joints on both sides, as well as the area, perimeter, and range of joint movement, based on the data presented in Figure 3 and the averaged data collected from the 1st, 3rd, and 6th training sessions as shown in Table 2. The results indicate that the joint symmetry of the hip and knee pairs, as well as the perimeter of inter-joint movement, did not exhibit an increase when power support was provided to the affected side knee joints compared to walking in free mode. However, the results showed an increased area of the hip-knee cyclogram on the affected side in the power support mode. Furthermore, a slight increase in the ranges of joint movements of the hip and knee were observed in participants #1 and #3 on the affected side. These findings suggest that the power provided by the exoskeleton robot to the knees may facilitate specific characteristics of gait function in stroke patients.

C. Effects After Robot-assisted Training

Finally, we analyzed the functional recovery in patients with acute and subacute stroke may exhibit superior outcomes compared to those in the chronic stroke stage, as indicated in

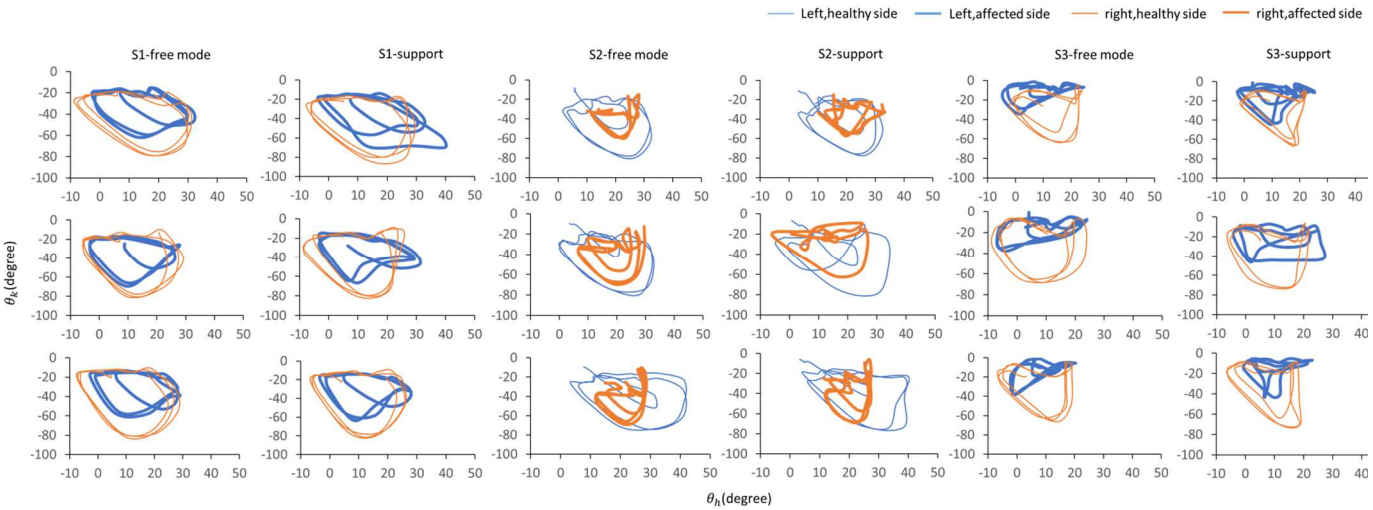


Fig. 3 The cyclograms illustrates the inter-joint coordination across three training sessions, namely the 1st, 3rd, and 6th training sessions, from top to bottom. The thick and thin lines indicate data obtained from the affected and healthy sides, respectively.

Table 2. Gait parameters in the three patient subjects. The results of the kinetics analysis are depicted, with the data presented as the mean (SD).

		#1		#2		#3	
Property		free mode	support	free mode	support	free mode	support
knee	TS	3.48(1.52)	5.98(3.12)	18.88(13.18)	19.60(14.93)	3.74(1.93)	7.09(4.25)
	CO (degree)	12.77(2.39)	12.09(4.29)	-8.7(7.46)	7.96(44.53)	22.69(5.95)	17.56(5.41)
hip	TS	4.46(4.39)	6.94(3.69)	10.49(3.59)	10.46(14.75)	6.80(6.09)	7.83(6.00)
	Co (degree)	5.10(4.90)	0.55(4.00)	-18.84(9.88)	-17.11(10.48)	7.97(10.48)	4.59(7.49)
perimeter	L	133.25(17.70)	138.88(19.12)	153.77(39.41)	153.49(35.91)	104.76(39.35)	108.21(32.76)
	R	180.74(8.32)	184.92(5.20)	157.25(21.36)	152.02(25.51)	170.43(19.02)	177.22(16.41)
area	L	725.55(206.20)	758.63(234.87)	897.62(550.86)	787.68(512.83)	271.74(136.63)	339.76(213.10)
	R	1430.63(141.70)	1412.80(110.80)	378.84(128.02)	435.29(260.35)	882.18(260.75)	797.59(250.85)
knee joint ROM	L (degree)	40.79(8.07)	44.31(7.47)	51.8(10.95)	50.87(10.23)	26.72(7.99)	31.21(9.86)
	R (degree)	63.61(4.99)	65.12(2.29)	46.2(7.68)	40.91(12.28)	52.89(7.20)	56.71(4.54)
hip joint ROM	L (degree)	26.8(4.56)	30.69(5.45)	27.51(6.32)	28.41(5.35)	20.63(6.82)	21.29(5.88)
	R (degree)	33.98(2.31)	31.32(2.81)	16.97(3.26)	20.00(7.06)	24.39(3.30)	24.48(2.37)

Table 1. Figure 3 provides a comprehensive overview of the effects on clinical outcomes before and after the training intervention. Following the training program, significant improvements in lower limb function were observed based on clinical assessments, including the Barthel Index (pre-: 73.33±10.4; post-: 83.33±12.58), Berg Balance Scale (pre-: 37.33±15.3; post-: 44.33±23.1), 5-Meter Walk Speed (pre-: 8.92±1.3s; post-: 6.56±1.54s), 6-Minute Walk Test (pre-: 139.33±98.97m; post-: 297±123.73m), Fugl-Meyer Assessment for Lower Extremity (pre-: 16.33±5.69; post-: 27±7.94), and Timed Up and Go Test (pre-: 22.9±25.8s; post-: 14.59±2.03s). However, notable improvements in the modified Rankin Scale, Instrumental Activities of Daily Living, and EuroQol-5D were exclusively observed in patients with acute stroke (Table 1).

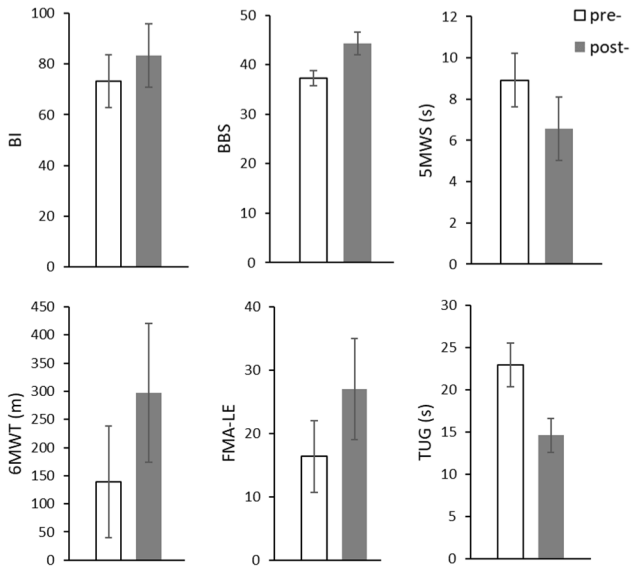


Fig.4 Clinical outcome measurements at baseline and after the 6th robot-assisted walking training. The data are presented as the mean±SD.

IV. Discussion

The present study observed and characterized improvements in spatiotemporal and sagittal joint kinematics as well as the results of clinical assessments following the robot-assisted gait training in participants with stroke. Specifically, lower limb function improved in the results of BBS and TGU but not in mRS, EQ-5D, or IADL."

We found that BBS and TUG demonstrated more effective improvements as compared with other parameters, such as BI, 5MWS, 6MWT, and FMA-LE. The motors of the lower limb exoskeleton robot were designed to compensate for the insufficient torque of knee extensor concentric or eccentric movements in the affected limbs of patients. One previous study has demonstrated that knee extension strength plays a crucial role in determining performance in both static and dynamic balance tests [22]. Furthermore, TUG differs to other walking-speed tests, such as 5MWS and 6MWT, in its functional involvement of the sit-to-stand process [23]. This process requires support through concentric and eccentric movements of the knee extensors, as a previous study have highlighted the crucial role of knee extension force in sit-to-stand [24]. Regarding other assessments of lower limb functions, such as 5MWS, 6MWT, and FMA-LE, which involve coordinated movements of all of the hip, knee, and ankle joints, the assistance of robot that focuses solely on knee support can only partially improve gait functional. BI, IADL, EQ-5D, and mRS are comprehensive evaluations of overall independence. Therefore, robotic support for the concentric and eccentric movements of the knee extensors alone is insufficient to generate a comprehensive impact on overall independence.

In the present study, motor relearning or employment of compensatory mechanisms are observed during testing and training. A sufficient adjustment period is necessary for participants who have not previously used a robotic device for walking assistance. Interestingly, we observed that an adjustment period is also needed following updates to the parameters for knee motor control. Consequently, following robot-assisted gait training, bilateral hip joint excursions and targeted modifications in knee flexion-extension movement may facilitate the adaptation to the use of robotic devices as well as a change of robotic control parameters. However, it should be noted that powered robot-assisted training can also influence gait performance on the unaffected side. The unaffected side may exhibit spontaneous resistance to the external power, resulting in unstable gait and a decrease of symmetry. Additionally, the unaffected side may attempt to compensate or support the affected side, leading to variations in both hip-knee cyclograms on each side.

In recent years, there has been a growing anticipation that a powered exoskeleton could enhance physical performance, even with a limited familiarity with this technology. Our preliminary data in the present study demonstrated a noteworthy improvement in motor function as they gradually gained proficiency in the use of the exoskeleton. This progress has played a crucial role in narrowing the disparity between their initial abilities in operating the robotic device and their performance after the adjustment of robotic parameters. However, further evaluation is needed to determine the optimal

parameters for stroke patients at different stages of functional recovery of gait.

The present study has several limitations that should be addressed carefully. First, the variance in stroke onset time among the enrolled participants may have influenced the study outcomes. To minimize this potential confounding factor, future studies could consider narrowing down the inclusion criteria regarding motor ability, an approach that could reduce bias resulting from spontaneous functional recovery. Second, it is challenging to identify the most suitable parameters for motor control in stroke patients with varying motor functional abilities. Currently, the assessment of gait symmetry during robot-assisted training and parameter testing heavily relies on expert professionals. Development of standardized protocols or algorithms to determine optimal parameters would enhance the objectivity and reproducibility of the findings. Third, the exoskeleton robot may affect patient's gait function event in the free mode and the present study did not explore the relationship between functional recovery and kinematic results, such as the correlation between gait functions and the hip-knee cyclograms obtained without robot before and after treatment. The understanding of the correspondence between functional improvements and specific kinematic patterns will provide valuable insights into the underlying mechanisms of recovery and guide the development of targeted rehabilitation interventions. Finally, the sample size and the types of training tasks employed in this study may limit the ability to draw definitive conclusions regarding treatment effects.

Future research should include a control group and incorporate a wider range of training tasks, an approach that will allow for a more comprehensive evaluation of the treatment effects and enhance our understanding of the efficacy of the intervention. By addressing these limitations in future studies, we can advance our knowledge in the field, improve the validity of the findings, and contribute to the development of more effective rehabilitation strategies for stroke patients.

V. Conclusions

By implementing specific parameter settings in the exoskeleton robot for stroke patients and providing instructions aimed at improving gait symmetry, we conducted a comprehensive investigation of gait characteristics and parameterization derived from active lower limb robot-assisted training. Our study demonstrated a positive impact on facilitating inter-joint coordination within the hip-knee joint movement on the affected side. The observed facilitated inter-joint movement induced by the robot suggests its potential in guiding motor learning processes that are essential for enhancing gait function.

Acknowledgment

This study was supported by the Ministry of Science and Technology, Taiwan (NMRPG3M0442, NSTC112-2218-E-182A-001), and Chang Gung Medical Foundation (BMRPB67). These funding agencies did not involve in the design of the study, data collection and analysis, and drafting of the manuscript. We also appreciated assistants of Hui-Xuan Qiu and Cheng-En Chien for assisting the current experiment.

References

[1] E. J. Roth, C. Merbitz, K. Mroczek, S. A. Dugan, and W. W. Suh, "HEMIPLEGIC GAIT: Relationships Between Walking Speed and Other Temporal Parameters: 1," *American journal of physical medicine & rehabilitation,* vol. 76, no. 2, pp. 128-133, 1997.

[2] K. K. Patterson, W. H. Gage, D. Brooks, S. E. Black, and W. E. McIlroy, "Evaluation of gait symmetry after stroke: a comparison of current methods and recommendations for standardization," *Gait & posture,* vol. 31, no. 2, pp. 241-246, 2010.

[3] K. K. Patterson, W. H. Gage, D. Brooks, S. E. Black, and W. E. McIlroy, "Changes in gait symmetry and velocity after stroke: a cross-sectional study from weeks to years after stroke," *Neurorehabilitation and Neural Repair,* vol. 24, no. 9, pp. 783-790, 2010.

[4] J. C. Wall, and G. I. Turnbull, "Gait asymmetries in residual hemiplegia," *Archives of physical medicine and rehabilitation,* vol. 67, no. 8, pp. 550-553, 1986.

[5] A. Goswami, "A new gait parameterization technique by means of cyclogram moments: Application to human slope walking," *Gait & posture,* vol. 8, no. 1, pp. 15-36, 1998.

[6] T. Krasovsky, and M. F. Levin, "Toward a better understanding of coordination in healthy and poststroke gait," *Neurorehabilitation and neural repair,* vol. 24, no. 3, pp. 213-224, 2010.

[7] J. A. Longworth, S. Chlosta, and K. C. Foucher, "Inter-joint coordination of kinematics and kinetics before and after total hip arthroplasty compared to asymptomatic subjects," *Journal of biomechanics,* vol. 72, pp. 180-186, 2018.

[8] J. H. Park, H. Lee, J.-s. Cho, I. Kim, J. Lee, and S. H. Jang, "Effects of knee osteoarthritis severity on inter-joint coordination and gait variability as measured by hip-knee cyclograms," *Scientific Reports,* vol. 11, no. 1, pp. 1789, 2021.

[9] H. S. Lee, H. Ryu, S.-U. Lee, J.-s. Cho, S. You, J. H. Park, and S.-H. Jang, "Analysis of gait characteristics using hip-knee cyclograms in patients with hemiplegic stroke," *Sensors,* vol. 21, no. 22, pp. 7685, 2021.

[10] E. C. Field-Fote, and D. Tepavac, "Improved intralimb coordination in people with incomplete spinal cord injury following training with body weight support and electrical stimulation," *Physical therapy,* vol. 82, no. 7, pp. 707-715, 2002.

[11] C. A. McGibbon, A. Sexton, A. Jayaraman, S. Deems-Dluhy, P. Gryfe, A. Novak, T. Dutta, E. Fabara, C. Adans-Dester, and P. Bonato, "Evaluation of the Keeogo exoskeleton for assisting ambulatory activities in people with multiple sclerosis: an open-label, randomized, cross-over trial," *Journal of neuroengineering and rehabilitation,* vol. 15, pp. 1-14, 2018.

[12] R. Pilkar, A. Ramanujam, K. Chervin, G. F. Forrest, and K. J. Nolan, "Cyclogram-based joint symmetry assessment after utilization of a foot drop stimulator during post-stroke hemiplegic gait," *Journal of biomechanical engineering,* vol. 140, no. 12, 2018.

[13] S. J. Crenshaw, and J. G. Richards, "A method for analyzing joint symmetry and normalcy, with an application to analyzing gait," *Gait & posture,* vol. 24, no. 4, pp. 515-521, 2006.

[14] N. D. Harada, V. Chiu, and A. L. Stewart, "Mobility-related function in older adults: assessment with a 6-minute walk test," *Archives of physical medicine and rehabilitation,* vol. 80, no. 7, pp. 837-841, 1999.

[15] J. Sanford, J. Moreland, L. R. Swanson, P. W. Stratford, and C. Gowland, "Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke," *Physical therapy,* vol. 73, no. 7, pp. 447-454, 1993.

[16] K. Berg, S. Wood-Dauphinee, and J. Williams, "The Balance Scale: reliability assessment with elderly residents and patients with an acute stroke," *Scandinavian journal of rehabilitation medicine,* vol. 27, no. 1, pp. 27-36, 1995.

[17] P. P. Chan, J. I. S. Tou, M. T. Mimi, and S. S. Ng, "Reliability and validity of the timed up and go test with a motor task in people with chronic stroke," *Archives of physical medicine and rehabilitation,* vol. 98, no. 11, pp. 2213-2220, 2017.

[18] C. S. Jacelon, "The Barthel Index and other indices of functional ability," *Rehabilitation Nursing,* vol. 11, no. 4, pp. 9-11, 1986.

[19] R. Rabin, and F. d. Charro, "EQ-SD: a measure of health status from the EuroQol Group," *Annals of medicine,* vol. 33, no. 5, pp. 337-343, 2001.

[20] M. P. Lawton, and E. M. Brody, "Assessment of older people: self-maintaining and instrumental activities of daily living," *The gerontologist,* vol. 9, no. 3_Part_1, pp. 179-186, 1969.

[21] N. M. Salbach, N. E. Mayo, J. Higgins, S. Ahmed, L. E. Finch, and C. L. Richards, "Responsiveness and predictability of gait speed and other disability measures in acute stroke," *Archives of physical medicine and rehabilitation,* vol. 82, no. 9, pp. 1204-1212, 2001.

[22] N. D. Carter, K. M. Khan, A. Mallinson, P. A. Janssen, A. Heinonen, M. A. Petit, and H. A. McKay, "Knee extension strength is a significant determinant of static and dynamic balance as well as quality of life in older community-dwelling women with osteoporosis," *Gerontology,* vol. 48, no. 6, pp. 360-368, 2002.

[23] S. S. Ng, and C. W. Hui-Chan, "The timed up & go test: its reliability and association with lower-limb impairments and locomotor capacities in people with chronic stroke," *Archives of physical medicine and rehabilitation,* vol. 86, no. 8, pp. 1641-1647, 2005.

[24] R. W. Bohannon, "Knee extension strength and body weight determine sit-to-stand independence after stroke," *Physiotherapy theory and practice,* vol. 23, no. 5, pp. 291-297, 2007.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Prediction Model of Postoperative Pain Exacerbation Using a Wearable Electrocardiogram Sensor

Toshiyuki Nakanishi*, Koichi Fujiwara†, and Kazuya Sobue‡
* Nagoya City University and Nagoya University, Japan
E-mail: nakanishi.anest@gmail.com
† Nagoya University, Japan
‡ Nagoya City University, Japan

***Abstract*—Few studies have evaluated postoperative pain in real patients using biosignal sensors. We aimed to develop a prediction model for postoperative pain exacerbation using a wearable electrocardiogram (ECG) sensor and an intravenous patient-controlled analgesia (IV-PCA) device. We collected ECG and IV-PCA data from eight patients who underwent laparoscopic gastrointestinal surgery. Finally, an AutoEncoder model predicted postoperative pain exacerbation with a true positive rate of 60% and a false positive rate of 1.94 times/h.**

I. Introduction

Annually, an estimated 300 million patients undergo surgical procedures [1]. As a result of advancements in surgical procedures and medical technology, minimally invasive surgical techniques, such as laparoscopic surgery, have been established as the gold standard, contributing to a reduction in postoperative mortality and morbidity [2]. However, even laparoscopic surgery can occasionally result in severe postoperative pain [3]. Despite advancements in perioperative analgesia techniques, a recent study revealed that 13% of patients experienced pain ranging from severe to extreme and 46% experienced moderate to extreme pain during the first two weeks following discharge [4]. Unmitigated postoperative pain can contribute to a decrease in quality of life, prolonged hospitalization, and the manifestation of chronic pain, resulting in substantial socioeconomic repercussions [5], [6]. Consequently, the management of postoperative pain continues to be a critical challenge warranting assiduous attention.

Traditional methods for assessing pain, like the Numeric Rating Scale (NRS) and the Visual Analog Scale (VAS), rely on the patient's subjective input [7]. These approaches are difficult to implement in certain groups, like pediatric patients or individuals with altered levels of consciousness. Moreover, they require active participation from healthcare staff, making constant monitoring challenging. Therefore, there is an urgent need for novel, objective ways to evaluate pain in real-time during the postoperative phase.

Previous studies have attempted to assess pain based on physiological signal changes or facial expressions of healthy volunteers exposed to thermal or electrical stimuli [8]–[14]. In contrast, few studies employed real-patient data to evaluate postoperative pain objectively. Two recent studies evaluated postoperative pain in actual patients using physiological data and machine-learning models [15], [16]. These investigations evaluated pain intensity induced by electrical stimuli or bodily movements in post-surgical patients, predicated on 30 minutes of physiological monitoring. The long-term course of postoperative pain intensity in real patients has yet to be studied.

In postoperative patients, small wearable sensors may be desirable for continuous monitoring as they are not in bed all day. To evaluate the long time course of pain intensity, we selected an electrocardiogram (ECG) as a feature to detect pain exacerbation because, in our experience, it is non-invasiveness and less noisy than other physiological signals. In addition, we utilized an intravenous patient-controlled analgesia (IV-PCA) device, a common analgesic modality used in post-surgical patients. When experiencing severe pain, patients can self-administer analgesic dosage by pressing a button of an IV-PCA device. We hypothesized that IV-PCA dose recordings could be used as a pain exacerbation lable during postoperative phase. We conducted a prospective observational study to construct a prediction model for pain exacerbation in post-surgical patients using ECG and IV-PCA records.

II. Methods

After the approval by the Nagoya City University Graduate School of Medical Sciences and Nagoya City University Hospital Institutional Review Board (60-22-0072, October 24, 2022), written informed consent was obtained from all the participants.

A. Participants

We included patients aged 20–90 with an American Society of Anesthesiologists-Physical Status (ASA-PS) of 1–3 who underwent laparoscopic or robotic-assisted gastrointestinal surgery at Nagoya City University Hospital. Patients incapable of conveying pain or utilizing an IV-PCA device as a result of dementia or cerebrovascular disease complications were excluded. Patients with comorbidities that affected their ECG, such as arrhythmia, Parkinson's disease, obstructive sleep apnea syndrome, epilepsy, and diabetic peripheral neuropathy, were also excluded.

B. Data Collection

The collection of data is depicted in Fig. 1. The ECG measurement and IV-PCA infusion were initiated in the post-anesthesia care unit (PACU) and persisted until postoperative

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

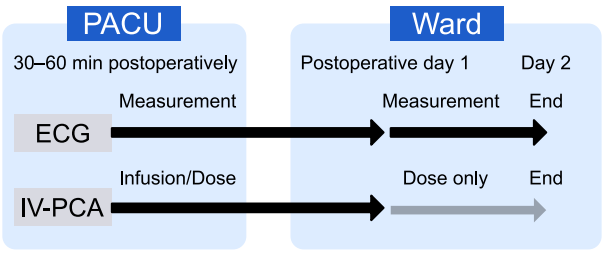


Fig. 1. Flow of Data Collection

day 2. We utilized a GSH761 ECG sensor (Golden Smart Home Technology Corp., Hsinchu, Taiwan, ROC) with a sampling rate of 250 Hz and a COOPDECH Amy PCA (Daiken Medical Co. Ltd., Osaka, Japan) that recorded the time of analgesic doses on a second-by-second basis. Bluetooth was utilized on a smartphone to synchronize the time between the wearable ECG sensor and the IV-PCA device. Patients were managed under general anesthesia with endotracheal intubation intraoperatively. For IV-PCA, fentanyl was administered at a rate of 0.3 μg/kg/h for continuous infusion and 0.3 μg/kg for bolus doses, with a lockout interval of 10 minutes.

C. Preprocessing of ECG

Subsequent to the extraction of ECG data from the sensor, artifacts were removed where the ECG waveforms had not been accurately captured. R waves within the ECG were detected with an algorithm predicated on first-order derivatives for peak identification. R-R intervals (RRIs) were ascertained through the measurement of temporal duration between consecutive R waves. A temporal window of three minutes was deployed for the computation of heart rate variability (HRV) features corresponding to each R wave.

D. Input Features of HRV

The three domains of HRV were used for input features.

- **LFnu :** Power of the low-frequency band (0.04–0.15 Hz) divided by total power.
- **HFnu :** Power of the high-frequency band (0.15–0.4 Hz) divided by total power.
- **LF/HF :** The ratio of the low-frequency band power to the high-frequency band power.

When the patient experiences pain exacerbation, the dominant sympathetic nervous tone would induce increased LFnu and LF/HF and decreased HFnu. If a bolus dose of fentanyl were administered, the sympathetic nervous tone would be attenuated.

E. Definition of Pain Exacerbation

Timing of pain exacerbation was defined as the times when patients self-administered IV-PCA doses. Given that intravenous fentanyl is anticipated to provide analgesia within a few minutes, we developed a prediction model that could detect HRV abnormalities 15 minutes prior to the IV-PCA dose.

III. MODEL

A. Architecture

Since we believed that patients would not experience severe pain for long periods of time with proper use of IV-PCA, the AutoEncoder (AE) model, an anomaly detection algorithm, was employed to classify pain exacerbation or not. As shown in Fig. 2, AE has inputs and outputs of the same dimension and hidden layers between them [17]. The AE model is trained to minimize a mean squared error (MSE), which is shown as follows:

$$L = \frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{n}\sum_{j=1}^{n}(x_{i,j}-\hat{x}_{i,j})^2\right\}$$

In the above equation, k is the number of input features, $x_{i,j}$ is the input features $\{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$, n is the length of the input sequence, and $\hat{x}_{i,j}$ is the restored features $\{\hat{x}_{i,1}, \hat{x}_{i,2}, \ldots, \hat{x}_{i,n}\}$ by the AE model.

We used the three HRV features with 60 sequential length. The network architecture had a 60×3 input, a 60×3 output that duplicates the input.

B. Training

The AE model was trained to minimize the MSE loss function. During the validation of each patient, the model was trained using HRV data derived from all patients except the one under validation (a leave-one-out-validation method). The training dataset included only the HRV data where IV-PCA doses were not administered within 60 minutes. We constructed the AE model with three hidden layers, with dimensions of 40, 20, and 40 units, respectively. The model was trained using the Adam optimizer, with a learning rate set at 0.002.

C. Validation and Test

For evaluation, reconstruction error (RE) was defined as the mean absolute error (MAE) between the input and output features.

$$RE = \frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{n}\sum_{j=1}^{n}|y_{i,j}-\hat{y}_{i,j}|\right\}$$

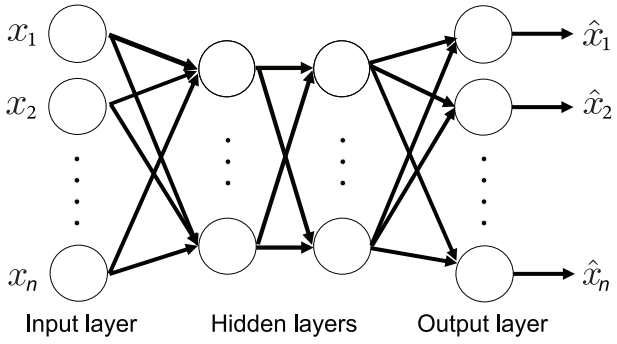


Fig. 2. AutoEncoder Model

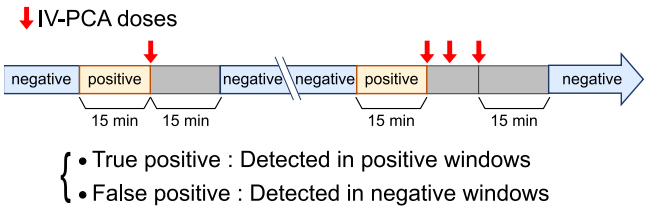


Fig. 3. Positive and Negative Windows of the Dataset

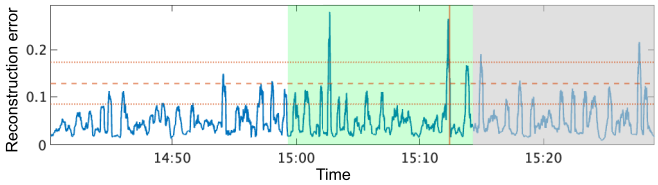


Fig. 4. Detection of Pain Exacerbation in Patient 5

The RE threshold ($\overline{RE}$) in each patient is required to detect pain exacerbation. Given the preliminary nature of this analysis, the limited cohort size might entail a substantial dispersion in individual RE values. Thus, we partitioned the time-series data for each patient into two segments in a 1:3 ratio to determine the ($\overline{RE}$). The initial 25% of the ECG data were employed to determine the $\overline{RE}$ for pain detection. Then, the latter 75% of the ECG data were utilized to evaluate the model performance.

If pain was detected, subsequent detection was delayed 15 minutes. In addition, we ensured that there would be no detection 15 minutes after each IV-PCA dose. The positive and negative windows of the test dataset are depicted in Fig. 3. The positive window was defined as the 15-minute period before each IV-PCA dose. If a second dose of IV-PCA was administered within 15 minutes, only the initial dose was used for the positive window. The negative window was defined as the period during which no IV-PCA doses were administered within 15 minutes before or after. For evaluating the model, we calculated the true positive rate (TPR) when pain was detected in a positive window and the false positive rate (FPR) when pain was detected in a negative window calculated as times/h.

We utilized an algorithm that detects pain exacerbation when the RE exceeds 90%–95% of its confidence interval ($\overline{RE}$) for 10–30 seconds. Through trial and error with different threshold patterns, the upper and lower limits of the confidence interval and duration were determined. In each patient, the GridSearch was employed to identify the $\overline{RE}$ to maximize the difference between the TPR and the FPR within the validation dataset.

The preprocessing of ECG was conducted using MATLAB 2022b (Mathworks Inc.), while other computations were carried out using Python 3.10.8 and PyTorch 1.13.1.

IV. RESULTS

We collected ECG and IV-PCA data from eight patients between December 2022 and February 2023 to construct a prediction model. Table I displays patient characteristics, the number of IV-PCA doses self-administered, and the duration of recorded ECG data. Due to postoperative nausea and vomiting, ECG measurements and IV-PCA was stopped for Patient 4. In patient 7, available ECG data were limited due to inaccurate measurements. The overall performance of the model was 60% for TPR and 1.94 times/h for FPR (Table II).

Fig. 4 shows the correct detection of pain exacerbation in patient 5. The green box indicates the positive window. The gray box indicates the 15 minutes after IV-PCA dose use. The vertical red line indicates the time the model detected pain exacerbation. In this patient, 94% of $\overline{RE}$ (a red horizontal dashed line) and 13 seconds duration were used. The horizontal red dotted lines indicate the standard deviation of the RE.

V. DISCUSSION

The AE model predicted postoperative pain exacerbation with a TPR of 60% and an FPR of 1.94 times/h. We previously reported a Self-Attentive AE model with a TPR of 54% and an FPR of 1.76 times/h using the same dataset [18]. The current results using the AE model achieved a higher TPR, but also a higher FPR. Sixty percent of TPR seems to be a reasonable performance in this preliminary result despite of the small dataset. Considering that the standard lockout time for IV-PCA with fentanyl is 10 minutes, suggesting six times per hour as a safe maximum dose, our proposed model is clinically safe in terms of relatively high FPRs. Nevertheless, FPRs may need to be decreased for clinical application. Since there is a trade-off between the TPRs and FPRs, the beneficial performance for post-surgical patients needs to be clarified in the future.

Few studies have evaluated postoperative pain in real patients using biosignal sensors. Naeini *et al.* [15] classified postoperative pain into five levels, downsampled from 11 levels of the NRS, using an ECG-derived machine learning model in post-surgical patients. Aqajari *et al.* [16] also classified postoperative pain using a machine learning model using skin resistance features. We used an IV-PCA record as a pain label assess continuous pain intensity in contrast to subjective scales like the NRS and the VAS [18]. The strength of our modality is its ability to evaluate pain over extended time periods.

Some limitations of our study should be noted. First, the number of participants were small. Second, the method of determining the $\overline{RE}$ used in this study may not be robust. We used the initial 25% of the ECG data and a small number of IV-PCA doses data to determine the $\overline{RE}$. The blood concentration

TABLE I
DATASET

No.	Age	Sex	# doses	Recorded duration
1	57	F	11	42 h 0 min
2	68	M	22	43 h 7 min
3	78	F	2	43 h 6 min
4	53	M	29	20 h 7 min
5	75	M	25	42 h 54 min
6	70	M	22	43 h 16 min
7	61	F	12	27 h 2 min
8	48	F	8	42 h 0 min

TABLE II
TEST RESULTS

No.	TPR [n/N (%)]	FPR [times/h]
1	0 / 1 (0%)	1.88
2	4 / 8 (50%)	2.62
3	0 / 1 (0%)	1.39
4	1 / 3 (33%)	1.14
5	9 / 13 (69%)	1.38
6	10 / 13 (77%)	2.99
7	5 / 8 (63%)	2.42
8	0 / 1 (0%)	1.53
Total	29 / 48 (60%)	1.94

of fentanyl could not attain a steady state immediately after the start of IV-PCA. There may also be residual effects of general anesthesia. We are investigating an algorithm to determine thresholds for each patient on a preoperative ECG. Finally, we excluded some patients who could not express pain, could not use an IV-PCA device, or had comorbidities that affected their ECG.

VI. CONCLUSIONS

We constructed a prediction model for postoperative pain exacerbation using a timely dataset of ECG and IV-PCA records. In this preliminary analysis, the AE model achieved a higher TPR, but also a higher FPR than our previous model [18]. We are going to increase data and compare multiple models, such as principal component analysis and other deep learning models.

ACKNOWLEDGMENT

This work was supported by JST, ACT-X Grant Number JPMJAX22AB, Japan.

REFERENCES

[1] Thomas G Weiser, Alex B Haynes, George Molina, Stuart R Lipsitz, Micaela M Esquivel, Tarsicio Uribe-Leitz, Rui Fu, Tej Azad, Tiffany E Chao, William R Berry, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *The Lancet*, 385:S11, 2015.

[2] Federico Coccolini, Fausto Catena, Michele Pisano, Federico Gheza, Stefano Fagiuoli, Salomone Di Saverio, Gioacchino Leandro, Giulia Montori, Marco Ceresoli, Davide Corbella, et al. Open versus laparoscopic cholecystectomy in acute cholecystitis. systematic review and meta-analysis. *International journal of surgery*, 18:196–204, 2015.

[3] Hans J Gerbershagen, Sanjay Aduckathil, Albert JM van Wijck, Linda M Peelen, Cor J Kalkman, and Winfried Meissner. Pain intensity on the first day after surgery: a prospective cohort study comparing 179 surgical procedures. *Anesthesiology*, 118(4):934–944, 2013.

[4] Asokumar Buvanendran, Jacqueline Fiala, Karishma A Patel, Alexandra D Golden, Mario Moric, and Jeffrey S Kroin. The incidence and severity of postoperative pain following inpatient surgery. *Pain medicine*, 16(12):2277–2283, 2015.

[5] Christine Miaskowski, James Crews, L Brian Ready, Steven M Paul, and Brian Ginsberg. Anesthesia-based pain services improve the quality of postoperative pain management. *Pain*, 80(1-2):23–29, 1999.

[6] Henrik Kehlet, Troels S Jensen, and Clifford J Woolf. Persistent postsurgical pain: risk factors and prevention. *The lancet*, 367(9522):1618–1625, 2006.

[7] Harald Breivik, Petter-Christian Borchgrevink, Sara-Maria Allen, Leiv-Arne Rosseland, Luis Romundstad, EK Breivik Hals, G Kvarstein, and A Stubhaug. Assessment of pain. *BJA: British Journal of Anaesthesia*, 101(1):17–24, 2008.

[8] Roi Treister, Mark Kliger, Galit Zuckerman, Itay Goor Aryeh, and Elon Eisenberg. Differentiating between heat pain intensities: the combined effect of multiple autonomic parameters. *PAIN®*, 153(9):1807–1814, 2012.

[9] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C Traue. Automatic pain recognition from video and biomedical signals. In *2014 22nd International Conference on Pattern Recognition*, pages 4582–4587. IEEE, 2014.

[10] Daniel Lopez-Martinez and Rosalind Picard. Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5624–5627. IEEE, 2018.

[11] Thomas Ledowski. Objective monitoring of nociception: a review of current commercial solutions. *British journal of anaesthesia*, 123(2):e312–e321, 2019.

[12] Ghazal Bargshady, Xujuan Zhou, Ravinesh C Deo, Jeffrey Soar, Frank Whittaker, and Hua Wang. Ensemble neural network approach detecting pain intensity from facial expressions. *Artificial Intelligence in Medicine*, 109:101954, 2020.

[13] Ehsan Othman, Philipp Werner, Frerk Saxen, Ayoub Al-Hamadi, Sascha Gruss, and Steffen Walter. Automatic vs. human recognition of pain intensity from facial expression on the x-ite pain database. *Sensors*, 21(9):3273, 2021.

[14] Yibo Huang, Linbo Qing, Shengyu Xu, Lu Wang, and Yonghong Peng. Hybnet: a hybrid network structure for pain intensity estimation. *The Visual Computer*, pages 1–12, 2022.

[15] Emad Kasaeyan Naeini, Ajan Subramanian, Michael-David Calderon, Kai Zheng, Nikil Dutt, Pasi Liljeberg, Sanna Salantera, Ariana M Nelson, and Amir M Rahmani. Pain recognition with electrocardiographic features in postoperative patients: method validation study. *Journal of Medical Internet Research*, 23(5):e25079, 2021.

[16] Seyed Amir Hossein Aqajari, Rui Cao, Emad Kasaeyan Naeini, Michael-David Calderon, Kai Zheng, Nikil Dutt, Pasi Liljeberg, Sanna Salanterä, Ariana M Nelson, and Amir M Rahmani. Pain assessment tool with electrodermal activity for postoperative patients: method validation study. *JMIR mHealth and uHealth*, 9(5):e25258, 2021.

[17] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

[18] Toshiyuki Nakanishi, Koichi Fujiwara, and Kazuya Sobue. Prediction model of postoperative pain exacerbation using an intravenous patient-controlled analgesia device and a wearable electrocardiogram sensor. *Annu Int Conf IEEE Eng Med Biol Soc.*, 2023.

Directional Neural Connectivity during Robot Mirror Therapy in Patients with Stroke

Yuma Kanaizuka*, Takahiro Manabe**, Jian-Jia Huang†‡, Jen-Wen Hung† and Yumie Ono**
* Electrical Engineering Program, Graduate School of Science and Technology, Meiji University, Japan
E-mail: ce231084@meiji.ac.jp Tel: +81-80-9270-9101
** School of Science and Technology, Meiji University, Japan
E-mail: yumie@meiji.ac.jp Tel: +81-44-934-7302
† Department of Rehabilitation, Chang Gung Memorial Hospital-Kaohsiung Medical Center, Kaohsiung, Taiwan
‡ School of Medicine, College of Medicine, Chang Gung University, Taoyuan, Taiwan

Abstract — **This study explored directional connectivity networks during hand movement training in stroke patients using 8 conditions combining mirror therapy, robot-assisted bimanual therapy, and object manipulation. The findings revealed that mirror therapy and robot-assisted bimanual therapy decreased interhemispheric inward connectivity to the affected motor cortex in the left-hand paralyzed patient, reducing inhibitory control. Conversely, these interventions increased interhemispheric inward connectivity in the right-hand paralyzed patient, suggesting enhanced excitatory connectivity. The results emphasized the influence of neurorehabilitation methods, hemiparesis severity, and affected side on interhemispheric connectivity. Further research is needed to develop personalized rehabilitation strategies based on directional connectivity measures.**

I. Introduction

Stroke, the second leading cause of death worldwide, results in approximately 5.5 million annual fatalities [1]. Even for survivors, up to 50% experience long-term disabilities such as hemiplegia and higher-order cognitive dysfunction. The recovery of hand functions induced by stroke is slower and more difficult than the lower limb due to the high dexterity [2, 3]. Therefore, it is vital of importance to investigate efficient rehabilitation strategies to facilitate stroke recovery.

In the field of hand motor rehabilitation for stroke patients, innovative approaches such as mirror therapy (MT), bimanual therapy, robot-assisted therapy, and combination therapy have demonstrated effectiveness in promoting the recovery of hand function [4-6]. However, there remains a lack of comprehensive understanding regarding the specific impact of each neurorehabilitation method on brain recovery, the identification of the patients who benefit the most from these interventions, and the optimal application methods [7].

Based on the previous studies highlighting imbalanced interhemispheric interaction in stroke patients [8-11], the objective of this study is to elucidate the effects of MT, robot-assisted bimanual therapy (RT), and robot mirror therapy (RMT) on the directional interhemispheric connectivity to the affected motor cortex using EEG during hand motor training. Interhemispheric interactions refer to the mechanism of neural circuitry in which the two hemispheres of the brain inhibit each other to perform unilateral movements, such as moving only one hand. After a stroke, this balance of interhemispheric inhibition is disrupted, resulting in increased inhibition from the unaffected hemisphere and decreased activity in the affected hemisphere. This disruption of interhemispheric interactions in bilateral motor cortices has been considered as the cause of stroke-induced hemiplegia, leading to impaired voluntary movements of the affected hand. Based on this hypothesis, an effective neurorehabilitation strategy could reduce excessive motor cortical connectivity from the unaffected side to the affected side.

To investigate these concepts, we examined EEG changes in the directional cortical connectivity of both healthy participants and stroke patients during hand movements under various therapeutic interventions. Additionally, we conducted subgroup analyses based on the severity of hemiplegia and the laterality of the affected hand to explore appropriate target patient populations for these therapeutic strategies.

II. Methods

A. Participants

Electroencephalography (EEG) data were collected from a total of 37 healthy young adult individuals and 46 stroke patients at Chang Gung Memorial Hospital, Kaohsiung

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

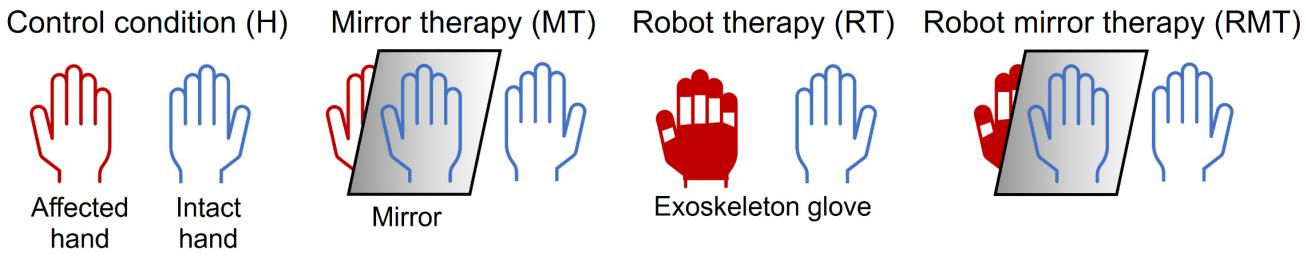


Fig. 1 Experimental conditions for therapy and control. Three types of therapy were employed: Mirror therapy (MT), Robot therapy (RT), and Combined robot-mirror therapy (RMT). The control condition involved the hand without mirror and robot (H). Each condition was tested with and without object manipulation, resulting in a total of 8 conditions.

Medical Center. Detailed inclusion and exclusion criteria for patients can be found in our previous report [7]. The stroke patient group was further divided into subgroups based on the laterality and severity of motor impairment. Specifically, there were 23 patients with right-hand hemiplegia (left-lesional strokes) and 23 patients with left-hand hemiplegia (right-lesional strokes). Patients within each hemiplegia group were classified into subgroups of severe and moderate motor impairment, based on their Brunnstorm recovery stage (BRS) scores in the distal part of the upper extremities [12]. The severe subgroup included patients with BRS II and III, with 12 patients for right-hand hemiplegia and 15 patients for left-hand hemiplegia. The moderate subgroup comprised patients with BRS IV and V, with 11 patients for right-hand hemiplegia and 8 patients for left-hand hemiplegia. The dominant hand of all subjects, including both healthy controls and patients, was the right hand.

The experimental protocol received approval (IRB number: 201900537A3), and written informed consent was obtained from all participants. All methods adhered to the regulations of the Human Subjects Research Act and guidelines of the Declaration of Helsinki 1975.

B. Experimental conditions

The EEG data were recorded using a mobile dry electrode EEG recording device (DSI-24 system; San Diego, CA, USA) with a sampling rate of 300 Hz. The scalp electrodes were positioned according to the standard 10-20 system channel arrangement, with reference electrodes placed on both earlobes.

Participants were seated in front of a table and placed both hands on the table. They performed slow, bilateral finger extension and flexion movements for a minimum duration of 60 s under eight different intervention strategy conditions. We employed a 2 (with vs. without mirror) × 2 (with vs. without robot assistance) × 2 (with vs. without object manipulation) factorial design to determine the intervention conditions (Fig. 1). The order of the training conditions was randomized, and each condition was applied once. For healthy participants, the assignment of moving either their right or left hand was randomized to mimic the other side of the hand as the affected side. Throughout the exercise, all participants were instructed to maintain their gaze on the affected or assumed affected side of the hand.

Fig. 1 depicts the combinations of mirror therapy (MT), robot therapy (RT), mirror and robot therapy (RMT) utilized in this study. MT involved participants focusing on their unaffected hand's reflection in the mirror while performing finger movements, creating an illusion of movement in the affected hand. The visual feedback provided by MT, simulating synchronized movement of the affected hand with the unaffected hand, is considered to promote the reorganization of the motor neural network [4]. RT involved participants wearing a sensor glove on their intact hand and a wearable exoskeleton hand robot (HS 001, Rehabotics Medical Technology Corporation, Taiwan) on their affected hand, facilitating passive movement. The RMT involved participants performing hand movements as in the RT condition. However, a mirror was positioned between the two hands, allowing participants to observe the reflected image of the intact hand during the exercise. This combination provided both visual and proprioceptive feedback, further strengthening the benefits of the intervention in stroke motor rehabilitation [5]. Additionally, the object condition (O) involved participants simultaneously manipulating plastic cups with each hand. These interventions aimed to explore differences in cortical connectivity during motor rehabilitation.

C. Pre-processing and directional connectivity analysis

To investigate the directional connectivity between brain regions, we employed a generalized orthogonalized partial directed coherence (gOPDC) approach [13]. This analysis was applied to the selected channels of EEG data recorded from frontal, central, parietal, and occipital cortices (F3, F4, C3, C4, P3, P4, O1, and O2). The C3 and C4 electrodes are positioned

above the primary sensory-motor cortices of the right and left hands, respectively, which were chosen as the target motor regions that receive interhemispheric neural signals to perform finger movement.

The recorded EEG data was firstly pre-processed using EEGLAB toolbox [14] under a MATLAB (MathWorks Inc., Natick, MA, USA) environment to remove EEG artifacts such as electrooculogram. Specifically, the raw EEG signals underwent a high-pass filter at 1 Hz, which would results in the lowest rejection threshold for non-EEG independent components [15] and improves the quality of the independent component analysis (ICA) [16] decomposition performed thereafter. Then, the line noise was removed with the 'cleanline' and 'clean_rawdata' under the condition, and Artifact Subspace Reconstruction was conducted with a default parameters to automatically modify the bad data. Furthermore, ICA, a feature extraction method often used for EEG artifact removal, was performed to decompose the EEG data into multiple components. Here, 'ICLabel' function was used to automatically label the components into several signal-type categories based on dominant type of signals components have. Of classified components, only component labeled 50 % or more as Brain were selected and other components were rejected as artifacts. Finally, cleaned data was segmented into 1-second epochs, overlapping 0.5 second for each (i.e., an n-second signal will be cut into 2n-1 epochs).

After pre-processing the data, a multivariate autoregressive (MVAR) model [17] was calculated for each condition. The weight matrix A_r of the p-th order MVAR model for the n-point time series data $y_i(n)$ $(i = 1, \dots, M)$ recorded from M measurement points (channels) is determined by the following equation:

$$\begin{bmatrix} y_1(n) \\ \vdots \\ y_M(n) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} y_1(n-r) \\ \vdots \\ y_M(n-r) \end{bmatrix} + \begin{bmatrix} w_1(n) \\ \vdots \\ w_M(n) \end{bmatrix} \quad (1)$$

$$A_r = \begin{bmatrix} a_{11}^r & \dots & a_{1M}^r \\ \vdots & \ddots & \vdots \\ a_{M1}^r & \dots & a_{MM}^r \end{bmatrix} \quad (2)$$

Here, w_i represents the error between the estimated and observed measurement values. The model order p was determined for each dataset using the ARfit toolbox [18]. The estimated optimal model orders ranged from 3 to 6, with a median value of $p = 4$, indicating that the neural network model could be established within the time lag of 13 ms.

The gOPDC is an extension of the partial directed coherence, designed to capture effective connectivity in neuro-physiological signals. It addresses the limitations of the partial directed coherence by accounting for variations in signal amplitude and mitigating the influence of volume conduction [13]. The gOPDC is defined as follows:

$$gOPDC = \frac{1}{\lambda_{kk}^2} \frac{|Real\{A_{kl}(n,f)\}|}{\sqrt{a_l^H(n,f)\,\Sigma_w^{-1}\,a_l(n,f)}} \cdot \frac{|Imag\{A_{kl}(n,f)\}|}{\sqrt{a_l^H(n,f)\,\Sigma_w^{-1}\,a_l(n,f)}} \quad (3)$$

$$A(n,f) = I - \sum_{r=1}^{p} A_r(n) z^{-r}\big|_{z=e^{j2\pi f}} \quad (4)$$

I is the identity matrix, a_l represents the l-th row vector of A_r, H denotes the Hermitian transpose, and A_{kl} represents the element in the k-th row and l-th column of $A(n,f)$, indicating the weight from channel l to channel k.

We assessed interhemispheric connectivity by calculating cumulative gOPDC values targeting the affected motor area using channels in the contralateral hemisphere. Specifically, in a comparison of right hand hemiplegia patients and right hand motor healthy controls, for the motor cortex corresponding to the right hand (C3), we summed the inward connectivity from channels F4, C4, P4, and O2 to C3 and subtracted the outward connectivity from C3 to F4, C4, P4, and O2. In a comparison of left hand hemiplegia patients and left hand motor healthy controls, for the motor cortex corresponding to the left hand (C4), we summed the inward connectivity from channels F3, C3, P3, and O1 to C4 and subtracted the outward connectivity from C4 to F3, C3, P3, and O1. This approach allowed us to evaluate interhemispheric connectivity patterns focused on the affected motor area, without being influenced by the specific lesion locations found in individual patients.

The strength of the inward interhemispheric information flow was assessed for each condition in both the alpha band (8-12 Hz) and the beta band (13-30 Hz) within the subgroups of severe or moderate motor impairment. The obtained strengths of the connectivity among two patient groups and healthy participants were shown in a bar graph and qualitatively assessed. We further conducted a three-way repeated-measures analysis of variance (ANOVA) and multiple comparisons to analyze the data of the patients, with factors including the presence of mirror, robot, and object in the intervention. All statistical analyses were performed using MATLAB and Microsoft Excel. All data were represented as mean and standard error and a p-value less than 0.05 was considered statistically significant.

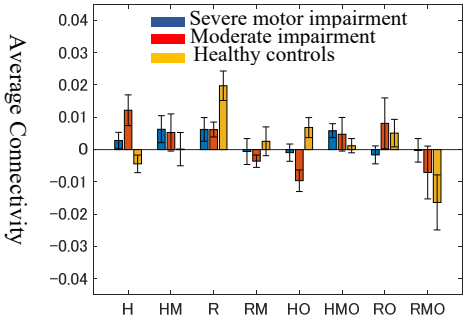
III. Results

A. gOPDC measures of interhemispheric connectivity

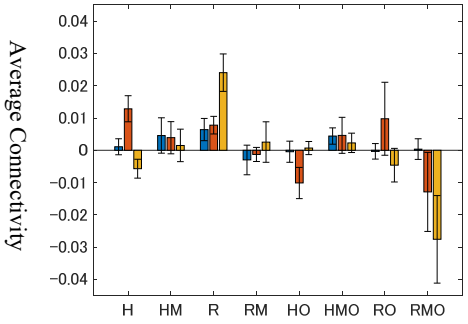
Fig. 2 illustrates the strength of directed interhemispheric connectivity across all combinations of conditions. Considering the impact of excessive inhibitory control from the unaffected side to the affected side motor cortex on stroke-induced motor impairment [8-10], we defined the meaningful effect after rehabilitation as an intervention that reduces inward interhemispheric connectivity to the affected motor cortex. Consistent with this notion, the inward connectivity tended to be higher in stroke patient group (blue and red bars) compared to healthy control group (yellow bars) in the most combinations of intervention strategies, indicating increased control of the affected motor cortex by the unaffected hemisphere. Nevertheless, the results also showed the connectivity changes varied among different combinations of intervention strategies; it is necessary to investigate the individual effects of each intervention.

B. Effects of different intervention strategies on interhemispheric connectivity to affected motor cortex

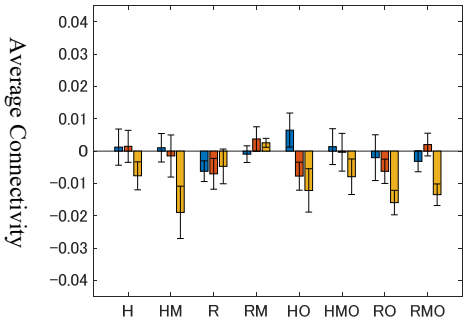
The impact of different intervention strategies on interhemispheric connectivity to the affected motor cortex varied depending on the severity of paralysis and the laterality of the affected hand. We observed a significant main effect only in the case of MT and an interaction between MT and RT. However, no significant interactions were observed across mirror therapy, robot therapy, and the use of objects. Since there were no significant main effects and interactions related to the use of objects, the subsequent results presented in Figs. 3 and 4 depict the interactions between MT and RT in the absence of object use. We observed a significant interaction between MT and RT in the condition of left-hand movement, specifically in the alpha wave connectivity, among patients with severe motor impairment ($F = 6.937$, $p = 0.021$; Fig. 3a). Although not statistically significant, a similar trend was also observed in the beta-band connectivity. Post-hoc multiple comparisons revealed a significant reduction in inward connectivity in RMT, compared to MT alone without incorporating RT. However, in the case of right-hand movement, there was no statistically significant interaction with any combination of the factors (Fig. 3b).



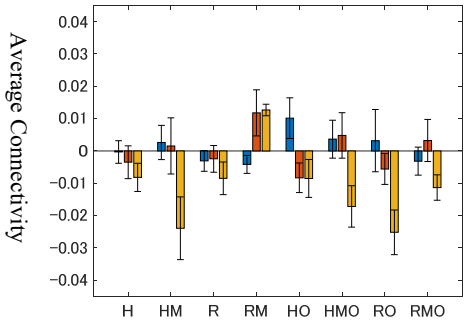
(a) Alpha band connectivity during left hand movement



(b) Beta band connectivity during left hand movement

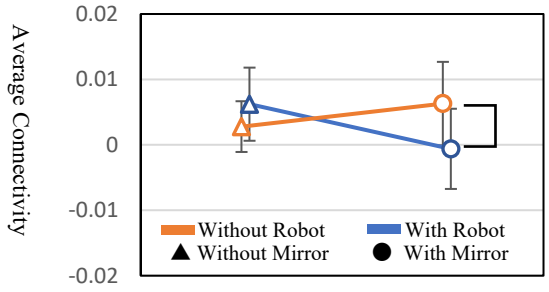


(c) Alpha band connectivity during right hand movement

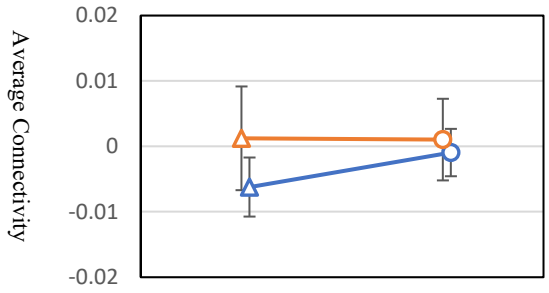


(d) Beta band connectivity during right hand movement

Fig. 2 Strength of inward interhemispheric connectivity to the affected motor area. The bar colors represent different participant groups: blue for patients with severe motor impairment, red for patients with moderate impairment, and yellow for healthy controls. The intervention conditions are labeled as follows: Control condition (H), mirror therapy condition (HM), robot therapy (R), combined mirror and robot therapy (RM). Object manipulation condition is indicated with(O).

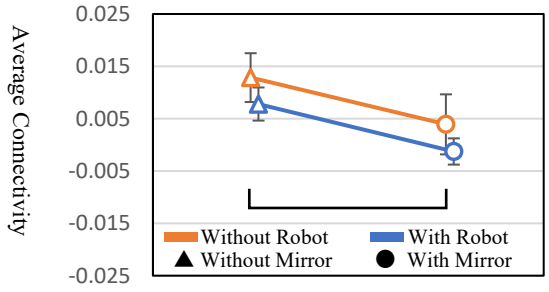


(a) Alpha band connectivity during left hand movement in patients with severe motor impairment without object manipulation

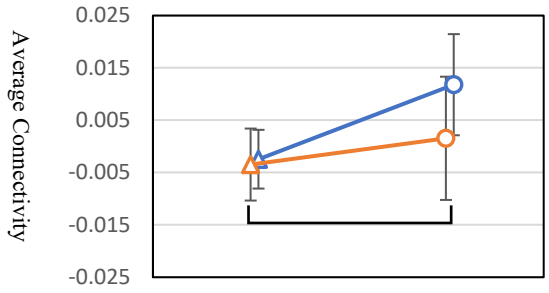


(b) Alpha band connectivity during right hand movement in patients with severe motor impairment without object manipulation

Fig. 3 Alpha-band inward interhemispheric connectivity to the affected motor area during hand movement in patients with severe motor impairment. Left paralyzed patients (a) show a significant interaction, while right paralyzed patients (b) do not. Circle and triangle markers represent data with and without MT, respectively. Blue and orange lines represent data with and without RT, respectively. Blankets indicate significant differences in post-hoc multiple comparisons.



(a) Beta band connectivity during left hand movement in patients with moderate motor impairment without object manipulation



(b) Beta band connectivity during right hand movement in patients with moderate motor impairment without object manipulation

Fig. 4 Beta-band inward interhemispheric connectivity to the affected motor area during hand movement in patients with moderate motor impairment. Left paralyzed patients (a) show a significant decrease of connectivity with a use of RT, while right paralyzed patients (b) showed opposite main effect. Circle and triangle markers represent data with and without MT, respectively. Blue and orange lines represent data with and without RT, respectively. Blankets indicate significant main effect.

In patients with moderate motor impairment, we observed a significant main effect of MT on beta band connectivity for both sides of hand movement. However, the impact of MT on the directional connectivity differed between the affected sides. During left-hand movement, MT significantly reduced the inward connectivity to the affected motor cortex, regardless of the involvement of RT ($F = 7.694$, $p = 0.028$; Fig. 4a). A similar trend was observed in alpha wave connectivity. In contrast, MT increased the inward connectivity to the affected motor cortex during right-hand movement, regardless of RT ($F = 5.179$, $p = 0.046$; Fig. 4b).

IV. Discussion

The directional connectivity analysis of EEG data during hand motor task revealed the influence of neurorehabilitation methods, hemiparesis severity, and affected side on interhemispheric connectivity in healthy participants and stroke patients. These findings suggest that the combination of MT and RT significantly decreased interhemispheric inward connectivity to the affected motor cortex in the left-hand paralyzed patient, suggesting the reduced inhibitory control from the intact hemisphere. These findings may provide additional benefits of RMT in improving neural networks in patients with severe paralysis. In case of patients with moderate paralysis on the left hand, MT was found to reduce the interhemispheric connectivity strength regardless of RT, while the RMT showed a greater reduction in interhemispheric connectivity in severely paralyzed patients compared to MT alone. This difference can be attributed to the range of hand motion in patients with different motor impairments. Severe paralysis severely limits finger mobility, and the combined feedback from proprioception and visual input during robot assisted finger movement may have significantly reduced inward interhemispheric connectivity beyond the effects of visual feedback alone. In patients with moderate paralysis, either visual or proprioceptive feedback could enhance incomplete finger movement, leading to a reduction in inward interhemispheric connectivity. There have been debates regarding the effectiveness of MT [19] and robot-assisted

therapy [6, 20] for motor recovery. Inconsistent results may arise from the residual motor ability of patients at baseline.

The contrasting trend of increased inward connectivity during right-hand movement suggests differences in the cortical connectivity network based on hand dominance. Previous studies using transcranial brain stimulation have reported variations in interhemispheric connectivity strength depending on the dominant or non-dominant motor sides [21, 22]. Our findings with healthy participants support this trend, demonstrating enhanced inward connectivity to the contralateral motor area during right-hand movement (Figs. 2cd) compared to the left-hand movement (Figs. 2ab). In the case of dominant hand movement, both inhibitory and excitatory interhemispheric control mechanisms [8, 10] may contribute to the summed inward connectivity. However, it is important to acknowledge the limitation of our study, as the gOPDC measure cannot differentiate between the polarity of signal transfer, such as GABAergic inhibitory or glutamatergic excitatory neuron activities. Further investigation using longitudinal datasets is necessary to gain a deeper understanding of the characteristics of interhemispheric directional connectivity throughout the different stages of stroke motor impairment recovery.

V. Conclusions

We examined the differences in directional connectivity networks during finger motor functional training using various intervention strategies, including MT, RT, and object manipulation. The contribution using the gOPDC method to analyze motor cortical networks in stroke patients is state-of-the-art in the present study. Our findings revealed that the choice of neurorehabilitation methods, the severity of motor function, and the affected side in the brain can influence interhemispheric connectivity during motor training. Analyzing information transmission between affected and unaffected motor cortices through network connections offers valuable insights for designing personalized rehabilitation programs and promoting recovery.

However, it is important to acknowledge that the current connectivity measure used in this study does not differentiate between inhibitory and excitatory interhemispheric interactions. Additionally, the tendencies in causal connectivity indices differed between patients with right hemiparesis and left hemiparesis. Further research is needed to develop patient-oriented rehabilitation strategies that leverage directional connectivity measures to optimize outcomes.

Conflicts of Interest

Jian-Jia Huang is the designer of the Mirror Hand and the funder of Rehabotics Medical Technology, Taiwan. Other authors had no conflict of interest.

References

[1] E. S. Donkor, "Stroke in the 21(st) Century: A Snaps hot of the Burden, Epidemiology, and Quality of Life, " *Stroke Res Treat,* vol. 2018, p. 3238165, 2018, doi: 10.1155/2018/3238165.

[2] P. W. Duncan, L. B. Goldstein, R. D. Horner, P. B. L andsman, G. P. Samsa, and D. B. Matchar, "Similar Motor Recovery of Upper and Lower-Extremities aft er Stroke," *Stroke,* vol. 25, no. 6, pp. 1181-1188, Jun 1994, doi: Doi 10.1161/01.Str.25.6.1181.

[3] H. M. Feys *et al.*, "Effect of a therapeutic interventio n for the hemiplegic upper limb in the acute phase aft er stroke - A single-blind, randomized, controlled mu lticenter trial," *Stroke,* vol. 29, no. 4, pp. 785-792, Ap r 1998, doi: Doi 10.1161/01.Str.29.4.785.

[4] P. Langhorne, J. Bernhardt, and G. Kwakkel, "Stroke rehabilitation," *The Lancet,* vol. 377, no. 9778, pp. 1 693-1702, 2011.

[5] K. C. Lin *et al.*, "Effect of mirror therapy combined with somatosensory stimulation on motor recovery a nd daily function in stroke patients: A pilot study," *J Formos Med Assoc,* vol. 113, no. 7, pp. 422-8, Jul 20 14, doi: 10.1016/j.jfma.2012.08.008.

[6] D. H. Kim, K. D. Lee, T. C. Bulea, and H. S. Park, "I ncreasing motor cortex activation during grasping via novel robotic mirror hand therapy: a pilot fNIRS stu dy," *J Neuroeng Rehabil,* vol. 19, no. 1, p. 8, Jan 24 2 022, doi: 10.1186/s12984-022-00988-7.

[7] J. J. Huang, Y. C. Pei, Y. Y. Chen, S. S. Tseng, and J. W. Hung, "Bilateral Sensorimotor Cortical Commun ication Modulated by Multiple Hand Training in Stro ke Participants: A Single Training Session Pilot Stud y," *Bioengineering (Basel),* vol. 9, no. 12, Nov 24 20 22, doi: 10.3390/bioengineering9120727.

[8] N. Takeuchi, Y. Oouchida, and S. I. Izumi, "Motor C ontrol and Neural Plasticity through Interhemispheric Interactions," *Neural Plast,* vol. 2012, 2012, doi: Art n 82328510.1155/2012/823285.

[9] N. Takeuchi and S. I. Izumi, "Noninvasive Brain Sti mulation for Motor Recovery after Stroke: Mechanis ms and Future Views," *Stroke Research and Treatme*

nt, vol. 2012, 2012, doi: Artn 58472710.1155/2012/584727.

[10] F. Meng *et al.*, "Cerebral Plasticity After Subcortical Stroke as Revealed by Cortico-Muscular Coherence," (in English), *Ieee T Neur Sys Reh,* vol. 17, no. 3, pp. 234-243, Jun 2009, doi: 10.1109/Tnsre.2008.2006209.

[11] S. Eldeeb *et al.*, "EEG-based functional connectivity to analyze motor recovery after stroke: A pilot study," *Biomed Signal Proces,* vol. 49, pp. 419-426, Mar 2019, doi: 10.1016/j.bspc.2018.12.022.

[12] S. Naghdi, N. N. Ansari, K. Mansouri, and S. Hasson, "A neurophysiological and clinical study of Brunnstrom recovery stages in the upper limb following stroke," *Brain Injury,* vol. 24, no. 11, pp. 1372-1378, 2010, doi: 10.3109/02699052.2010.506860.

[13] A. Omidvarnia, G. Azemi, B. Boashash, J. M. O'Toole, P. B. Colditz, and S. Vanhatalo, "Measuring Time-Varying Information Flow in Scalp EEG Signals: Orthogonalized Partial Directed Coherence," *Ieee T Bio-Med Eng,* vol. 61, no. 3, pp. 680-693, Mar 2014, doi: 10.1109/Tbme.2013.2286394.

[14] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J Neurosci Meth,* vol. 134, no. 1, pp. 9-21, Mar 15 2004, doi: 10.1016/j.jneumeth.2003.10.009.

[15] M. Klug and K. Gramann, "Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments," *Eur J Neurosci,* vol. 54, no. 12, pp. 8406-8420, Dec 2021, doi: 10.1111/ejn.14992.

[16] C. Jutten and J. Herault, "Blind Separation of Sources .1. An Adaptive Algorithm Based on Neuromimetic Architecture," *Signal Process,* vol. 24, no. 1, pp. 1-10, Jul 1991, doi: Doi 10.1016/0165-1684(91)90079-X.

[17] L. A. Baccala and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biol Cybern,* vol. 84, no. 6, pp. 463-74, Jun 2001, doi: 10.1007/PL00007990.

[18] T. Schneider and A. Neumaier, "Algorithm 808: ARfit - A matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models," *Acm T Math Software,* vol. 27, no. 1, pp. 58-65, Mar 2001, doi: Doi 10.1145/382043.382316.

[19] Lo AC, Guarino PD, Richards LG, Haselkorn JK, Wittenberg GF, Federman DG, Ringer RJ, Wagner TH, Krebs HI, Volpe BT, Bever CT Jr, Bravata DM, Duncan PW, Corn BH, Maffucci AD, Nadeau SE, Conroy SS, Powell JM, Huang GD, Peduzzi P： Robot-assisted therapy for long-term upper-limb impairment after stroke. N Engl J Med 2010；362： 1772-1783

[20] Y. M. Chen, S. S. Lai, Y. C. Pei, C. J. Hsieh, and W. H. Chang, "Development of a Novel Task-oriented Rehabilitation Program using a Bimanual Exoskeleton Robotic Hand," *Jove-J Vis Exp,* no. 159, May 2020, doi: ARTN e6105710.3791/61057.

[21] J. Netz, U. Ziemann, and V. Homberg, "Hemispheric-Asymmetry of Transcallosal Inhibition in Man," *Exp Brain Res,* vol. 104, no. 3, pp. 527-533, Jun 1995. [Online]. Available: <Go to ISI>://WOS:A1995RG46000016.

[22] J. Duque *et al.*, "Intermanual differences in movement-related interhemispheric inhibition," *J Cognitive Neurosci,* vol. 19, no. 2, pp. 204-213, Feb 2007, doi: DOI 10.1162/jocn.2007.19.2.204.

Evaluation of neural response recorded using scalp EEG in virtual reality environment

Noriaki Kanayama[1,2], Makoto Miyakoshi[3,4] and Maro Machizawa[2,5]
[1] National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
E-mail: kanayama.n@aist.go.jp, Makoto.Miyakoshi@cchmc.org, machizawa@hiroshima-u.ac.jp Tel: +81 29-861-8167
[2] Hiroshima University, Hiroshima, Japan
[3] Division of Child and Adolescent Psychiatry, Cincinnati Children's Hospital Medical Center
[4] Department of Psychiatry, University of Cincinnati College of Medicine
[5] Xiberlinc Inc. Tokyo, Japan

***Abstract* — With the recent development of virtual reality technology, sensory information in virtual space could be a candidate for experimental stimuli in the field of cognitive neuroscience. Specifically, auditory and tactile stimuli, which is spatiotemporally synchronized to visual information of the space and body delivered using a head-mounted display (HMD), can be presented in controlled way using headphones, tactile stimulators (e.g. Kanayama et al., 2019), etc.. During such a sensory stimulus presentation, we could measure any neuronal responses, and realize cognitive neuroscience study in the experimental setting impossible in real world. Such a method involves fixing the HMD to participant's head and using the actuator near participant's body, which suggests possible contamination of greater noise to recorded neuronal responses compared to that in experiments without virtual reality. In the case of high-density EEG measurements on the scalp, there should be interference to the electrodes from the fixation band of HMD to the participant's head, electromyogram (EMG) noise induced by participant's head movements, and noise emitted by the HMD device itself.**

I. INTRODUCTION

In recent decade, virtual reality technology has been dramatically developed. On December 2012, Oculus VR, LCC has been released a head-mounted display (HMD), Oculus Developer Kit for video game developers, and then on January 2016, the consumer version has started to sell. Before the consumer versions of HMDs were sold, some neuroscience researchers, who especially has been interested in human body recognition, used HMD for their psychological experiments [1,2] controlling body image in virtual reality environment. However, for the neuroscience study using scalp EEG, HMD was not easily used because of interference with electrode attached onto the participant's head, which led the neuroscience researcher avoid using HMD for their VR experiment with EEG measurement (e.g. using a large rear-projection screen [3]). Otherwise, scalp-EEG measurement was omitted from the experimental settings for embodiment study using HMD [4]. Although HMD and Scalp-EEG measurement are not compatible with each other, fMRI as alternative way to conduct cognitive neuroscience study has more critical limitation that the participant can't rotate own head to see around the VR environment. Because cognitive neuroscience study using fMRI also has another limitations such as the requirement to lie on a bed during scanning, it is considered that the most feasible method at present is to increase the EEG measurement accuracy on the combination use of scalp-EEG and HMD.

Altogether, when we conduct any cognitive neuroscience study for VR using HMD, the following two points are important.

1. Avoiding interference on scalp EEG electrodes by HMD
2. Quantitatively clarify the extent to which noise is affected by HMDs

For the first point, our previous study has made the settings avoiding the HMD interference on EEG electrode by fixation of HMD on the back of the neck and hanging it on the ceiling (Fig.1) [5]. For the second point, some previous research has demonstrated that HMD does not necessarily add any external noise to make the recorded EEG data unusable. A study has found a significant condition difference on visual- and somatosensory evoked potentials and reported the signal-to-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

noise ratios [6]. An our previous study also demonstrated the target event-related spectrum power (ERSP) showed statistical difference of the experimental condition (spatially congruency of visuo-tactile stimulation; congruent vs. incongruent) as the similar previous studies showed. However, there is no study to show how much the noise contaminated by using HMD and VR contents has any impact on the detection of condition difference. Actually, in almost all cases, we tried to exclude the noise contamination, and have clean data during the experiment, which requires the intentional data record with typical noise as a template. In this study, using the noise waveforms recorded as a template of VR related noise (electrode movement noise based on the head/neck movement), the support-vector machine classification performances were evaluated using original EEG data and noise contaminated EEG data.

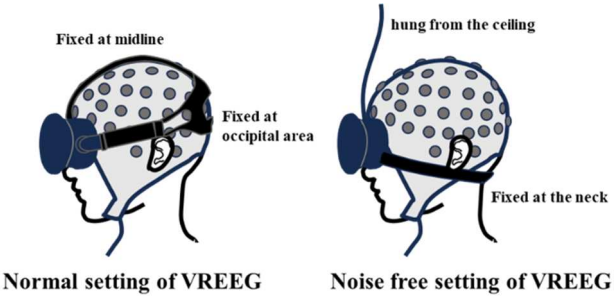


Fig. 1 Schematic illustration depicting the setup of wearing VR and EEG electrode cap.

II. Material and Methods

A. Experimental Procedures

The EEG datasets were obtained from 32 healthy individuals by the previous study [5] that aimed to detect the theta and gamma oscillatory activities during rubber hand illusion paradigm. To induce the illusion, we have used LEDs as visual stimuli and vibration using bone-conducting earphone as tactile stimuli. These stimuli were presented simultaneously, while the spatial position was changed between condition. In the congruent condition, if LED as a visual stimulation emitted at the index finger, tactile stimulation was delivered on the identical finger (the index finger), whereas in the incongruent condition, if LED as a visual stimulation emitted at the index finger, tactile stimulation was delivered on the different finger (the ring finger). The almost same experiment was conducted with HMD. The details of the experimental design and procedure can be available in the method section in the published previous paper [5].

B. EEG Data Analysis

The recorded EEG waveforms using actiCHamp (Brainproducts GmbH, Munich, Germany) from 63 electrodes with 1000 Hz sampling rate. At the recording the hardware filter was applied with time constant of 10 seconds. The electrode locations were selected based on the 10–10 international standard EEG electrode placement with a reference electrode at Cz and a ground electrode at Fpz.

EEG signal processing was conducted using EEGLAB version 14_1_2b (Delorme and Makeig 2004) under Matlab R2022b (MathWorks, Natick, MA, USA). A high pass filter (1Hz) and commercial power supply noise reduction using cleanline plug-in were applied to the continuous waveforms. The segmented data was made with the following settings: zero-time point was the visuo-tactile stimulation onset (0 ms), segment range was −600 to 1200 ms. Infomax independent component analysis (ICA) (Comon 1994; Bell and Sejnowski 1995) was conducted on the segmented data. Based on the IC waveforms, artifact-contaminated trials were detected and discarded. The artifact detection criteria were identical to the previous study [5]. Using the artifact rejected datasets, another ICA was conducted. The dipole estimation was conducted on these ICs using dipfit2.3 (EEGLAB plug-in using FieldTrip toolbox functions (Oostenveld et al. 2011)).

Cortical activation-related ICs were selected from the all the ICs obtained by the second ICA, using ICLabel plugin (Pion-Tonachini et al. 2019) which automatically classifies the components into brain, muscle, eye and other. In this study, only the ICs classified as “brain” were included the further analysis.

The data were down-sampled to 128 Hz. Using the down-sampled data, scalegrams were calculated using Morse wavelet using Matlab function cwt() at 0–800 ms post-stimulus and 3–60 Hz to exclude high gamma frequency band.

C. Noise Data Measurement

To obtain the template noise data, EEG waveforms were recorded from one participant. The participant was not included in the datasets in the main experiment. The participant wore HMD and EEG-cap with 63 electrodes, in similar manner with the main experiment. The difference is the way to use fixation belt. In this template noise recording session, the fixation belt provided officially from the developer was attached to the participant’s head, in order to put the template noise in the recorded data (in the main experiment, belt touched only on the neck). Under this condition, the EEG electrodes

Table 1. Classification accuracy results change for each participant by noise addition

	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14	S15	**S16**
Original	57.3	55.4	55.1	57.2	58.1	59.0	57.3	57.0	53.1	58.3	55.5	58.8	55.7	58.5	57.3	**61.5**
Mov Noise	52.2	50.5	53.8	53.4	52.9	54.1	54.1	53.0	52.2	55.4	52.0	50.7	51.7	54.2	54.5	**55.5**
Mov+Line Noise	52.0	52.1	52.2	52.7	50.7	53.6	52.6	52.9	52.4	53.3	53.0	50.3	52.5	53.4	54.4	**53.4**

	S17	S18	S19	S20	S21	S22	S23	S24	S25	S26	S27	S28	S29	S30	S31	S32
Original	57.8	59.0	56.7	59.6	58.4	54.9	57.7	53.4	54.2	56.7	53.0	54.3	55.4	55.5	55.5	56.3
Mov Noise	52.5	54.6	52.9	55.4	54.3	51.1	53.6	51.4	52.1	52.1	51.9	50.0	50.6	52.0	52.8	50.5
Mov+Line Noise	50.0	51.9	51.9	54.2	55.2	50.7	53.1	51.0	50.5	53.2	50.6	51.1	49.5	50.6	51.3	52.6

Mov Noise: neck movement noise, Line Noise: commercial power supply noise

must be contacted by the fixing belt, which made intense noise when the participant move own neck or head. The participants sat on the chair and continuously rotate the head for 1 minute. One stroke of head rotate was almost 2 s, which was previously trained using metronome. Furthermore, to obtain the commercial power line noise contaminated EEG data, all devices for using HMD were connected into power outlet. In the main experiment, the HMD and tracking camera were connected to the mobile battery for power supply without outlet in order to avoid contamination of the power line noise.

Signal processing of the obtained EEG data with template noise was almost identical to the main experiment.

D. Adding the Noises to the Test Signals

The various types of noise template data were segmented into 1-s epochs, then a Tukey window was applied (0.25 rise, 0.5 plateau, 0.25 fall). The segmented noise epochs were randomly selected, allowing overlapping selections, to match user-specified rate of the contaminated data (range 0-1). The length-matched noise data were re-scaled using user-specified re-scaling factor. The noise data were linearly added to the randomly selected segments of the test signal. The same process was repeated for the number of the types of noises.

E. Evaluation using support vector machine (SVM)

Evaluation of EEG quality was conducted by classification accuracy calculated by SVM. SVM was conducted using to binary classification of the stimulation pattern (spatially congruent/incongruent visuo-tactile stimulation). Spatially congruent stimulation pattern generally induces the illusory feeling of body ownership, whereas spatially incongruent stimulation pattern does not induce the illusory feeling. For the SVM classification, data from all the time-frequency window in the all brain ICs were used. Iteration number was 250. As a kernel function radial basis function was adopted. Scale parameter was automatically selected using heuristic method. Cross validation of the model was conducted by k-fold cross validation. The number of separation was 5. Based on the calculated accuracy values, the mean parameter of the normal distribution was evaluated.

III. Results

Champion data in binary classification of original data was observed in a participant No 16 (classification accuracy was 61.5 illustrated in Fig.2 and Table1). The averaged classification accuracy across all participants was 56.67 (SD = 1.98). After neck movement noise addition, the averaged classification accuracy across all participants was 52.75 (SD = 1.52). Additionally, commercial power supply noise added on the data, the averaged classification accuracy across all participants was 52.15 (SD = 1.36). Subtraction the accuracy in neck movement noise added datasets from original accuracy percentages revealed 3.9 (SD = 1.48) percent reduction by the noise addition in average. Subtraction the accuracy in neck movement and commercial power supply noise added datasets from the accuracy in neck movement noise added datasets revealed 0.6 (SD = 1.21) percent reduction by the noise addition in average.

IV. Discussion

In this study, we aimed to evaluate the impact of noise generally contaminated during VR experiment using HMD on scalp EEG. We calculated classification accuracy using SVM

for each individual person, in terms of stimulation pattern (spatially congruent/incongruent stimulation). As the results, we have revealed about 4 % reduction of classification accuracy rate in average. Especially, we observed an example of 8 % reduction (S12), which suggest that the neck movement noise has very intense impact on the EEG waveforms when we want to use the EEG spectrum power to predict any psychological event.

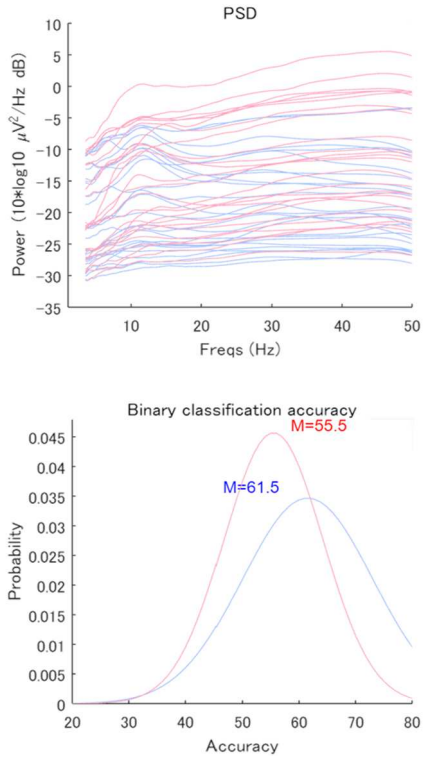


Fig. 2 Modulation of power spectrum density and binary classification accuracy by adding the neck movement noise on original EEG waveforms.
Top: Power spectrum density for each independent component, calculated by averaging ERSPs across all trials. Blue line: Original data. Red line: Noise added data. Bottom: Binary classification accuracy results. The unit of accuracy and Probability was percentage. PSD: Power spectrum density. Blue line: Original data. Red line: Noise added data

Although the neck movement noise has great impact on the EEG quality in terms of classification accuracy, an impact of the power supply noise, which was traditionally cared in psychophysiological research fields, was limited. Subtraction between with/without power supply noise conditions revealed less than 1% reduction of classification accuracy. For some participants, classification accuracy rate increased after power supply noise addition. These results suggested spectrum power of stationary wave at the high frequency range does not take important role when machine learning automatically separate the data. It seems that even if a stationary wave with little temporal fluctuation was added to the waveform for each condition, this did not significantly affect the difference in features between the conditions. Moreover, most of EEG components activated by any experimental stimulation related body in virtual reality environment were at the low frequency band, like alpha [2, 7, 8]. In addition, recent review on EEG study in VR also suggested that some studies demonstrated lower frequency band activities like alpha band can have significant role, while only one study has observed high frequency range activity like gamma band [9].

However, this does not suggest that we can ignore the power supply noise in any case. Some studies demonstrated gamma band oscillation could be an important role for motion recognition [10], motor control [11], and motion imagery [12]. It should be noted that we should evaluate the impact of noise for each experimental setting.

As limitations of this study, the accuracy of the 2-class classification using SVM is low. It should be caused by the abstractness of the 2-classes. In this study, we have tried to classify the EEG responses corresponding to spatial congruency of the visuo-tactile stimuli during the rubber hand illusion regarding body ownership, which can induce very highly cognitive process. Although, in the neuroscience with VR, the body recognition/ownership is very important theme, it is more useful to use simpler classification (e.g. right and left arm movement) for evaluation of noise contamination itself. In future study, we should choose more simplified classification.

Moreover, as an another limitation, we mention about the difference of indices between this study and traditional neuroscience and psychophysiological studies. In this study, classification accuracy was adopted to evaluate the noise impact, whereas traditional studies used a significant difference calculated by statistical test, for example, t-test. The results in this study does not have any direct impact on the traditional EEG studies which aimed at finding statistically significant differences. From the other perspective, usage of simple machine learning, SVM, is a limitation. Recently we could easily access any deep learning technique to classify the EEG data, many researchers tried to use deep learing for EEG data classification [13, 14]. In the case of EEG classification using deep learning, which automatically extracts feature values, it

seems that the results will dramatically change. In future, we should try to simulate the noise impact on various EEG classification methods.

ACKNOWLEDGMENT

This project was funded by JST COI (JPMJCA2208; JPMJCE1311) and Tateisi Science and Technology Foundation (A2221009). The authors would like to thank Chiaki Hisamatsu and Izumi Kitami (AIST, Japan) for their assistance with the experiments and data acquisition.

REFERENCES

[1] Ehrsson, H. H. (2007). The experimental induction of out-of-body experiences. Science, 317, pp.1048.

[2] Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007). Video Ergo Sum: Manipulating Bodily Self-Consciousness. Science, 317, pp.1096–1099.

[3] Lenggenhager, B., Halje, P., & Blanke, O. (2011). Alpha band oscillations correlate with illusory self-location induced by virtual reality. European Journal of Neuroscience, 33(10), pp.1935–1943.

[4] Spanlang, B., Normand, J.-M., Borland, D., Kilteni, K., Giannopoulos, E., Pomés, A. s, González-Franco, M., Perez-Marcos, D., Arroyo-Palacios, J., Muncunill, X. N., & Slater, M. (2014). How to Build an Embodiment Lab: Achieving Body Representation Illusions in Virtual Reality. Frontiers in Robotics and AI, 1, 9.

[5] Kanayama, N., Hara, M., & Kimura, K. (2021). Virtual reality alters cortical oscillations related to visuo-tactile integration during rubber hand illusion. *Scientific Reports*, *11*, 1436.

[6] Harjunen, V. J., Ahmed, I., Jacucci, G., Ravaja, N., & Spapé, M. M. (2017). Manipulating Bodily Presence Affects Cross-Modal Spatial Attention: A Virtual-Reality-Based ERP Study. *Frontiers in Human Neuroscience*, *11*.

[7] Evans, N. & Blanke, O. (2013). Shared electrophysiology mechanisms of body ownership and motor imagery. NeuroImage 64(1), 216-228.

[8] Kober, S. E., Kurzmann, J., & Neuper, C. (2012). Cortical correlate of spatial presence in 2D and 3D interactive virtual reality: An EEG study. *International Journal of Psychophysiology*, *83*(3), 365–374.

[9] Choi, J. W., Kwon, H., Choi, J., Kaongoen, N., Hwang, C., Kim, M., Kim, B. H., & Jo, S. (2023). Neural Applications Using Immersive Virtual Reality: A Review on EEG Studies. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 1645–1658.

[10] Zarka, D., Cevallos, C., Petieau, M., Hoellinger, T., Dan, B., & Cheron, G. (2014). Neural rhythmic symphony of human walking observation: Upside-down and uncoordinated condition on cortical theta, Alpha, Beta and gamma oscillations. Frontiers in Systems Neuroscience, 8, 103150.

[11] Cheron, G., Petit, G., Cheron, J., Leroy, A., Cebolla, A., Cevallos, C., Petieau, M., Hoellinger, T., Zarka, D., Clarinval, A. M., & Dan, B. (2016). Brain Oscillations in Sport: Toward EEG Biomarkers of Performance. Frontiers in Psychology, 7, 246.

[12] Uemoto, K., Yoshioka, M., Liang, H., & Zhu, C. (2017). Effect of Motor Intensity on Motion Imagery with Electroencephalogram Signal Analysis in Mirror Neuron System. Journal of Neuroscience and Neuroengineering, 4(1), 38–43.

[13] Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of Neural Engineering, 16(3), 031001.

[14] Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering, 16(5), 051001.

Machine Learning Based Action Recognition with Modular CNN

Shi-Zong Huang, Ching-Te Chiu and Yu-Jen Chang

sh2827228@gmail.com, ctchiu@cs.nthu.edu.tw, arthurchang09@gmail.com

Department of Communications, National Tsing Hua University, Hsinchu, Taiwan

***Abstract*—When building models for action recognition, 3D convolutional neural networks (CNNs) are commonly used. However, 3D CNNs also increase the model parameters significantly. We propose two methods, image segmentation and dynamic sampling learning to reduce network parameters and required memory access. Using image segmentation to keep the location of the action and remove the background of each image reduces the size of the feature map. Dynamic sampling learning allows the model to learn from low sampling rates without adding additional parameters, and to maintain performance while reducing the number of images. In order to implement the overall model in hardware for edge devices, we limit the kernel sizes of the 2D convolution layers and 3D convolution layers in the model to only 3x3 and 3x3x3 respectively. We perform experiments on HMDB51 [1] and UCF101 [2] datasets respectively with our proposed model. The accuracy of our proposed method achieve 7.2% and 5.9% reduction compared with DS-GRU2021 [3]. However, the number of parameters of our model is 30% fewer and execution speed x180 faster than DS-GRU2021 [3].**

***Index Terms*—Action Recognition, Deep Convolutional Networks, Real-time Computing, Dynamic Sampling Learning.**

I. Introduction

In recent years, convolutional neural networks(CNN) have been used in various application in computer vision, such as object detection, face recognition, and action recognition. However, the computing resources and memory units on the edge devices are limited. Traditional CNNs have high model computation and memory requirements, which makes it difficult to compute models on edge devices. Compared with common image recognition models, the action recognition model uses more complex 3D CNNs than 2D CNNs and the input of the model is increased to multiple images.

The current action recognition model uses three different strategies. The first one is to use all 2D CNNs in the model for computation [4]; the second one is to use all 3D CNNs in the model for computation [5]–[9]; and the last one is to combine these approaches to construct the model [10].

In our work, we focus on solving the problem of the large number of parameters and memory requirements of action recognition models. We propose two methods to reduce the memory used during model calculation, namely using image segmentation models on the network front-end and dynamic sampling learning. The first method involves using an image segmentation module on the front-end of the model to reduce the size of the feature map. The second method involves using dynamic sampling learning to reduce the number of images input to the model.

The rest of this paper is organized as follows. In Section II, we present the proposed approach to reduce the computation of 3D CNN by using eCNN modules and by image segmentation, and to reduce the number of input images to the model by using dynamic sampling rate. We also improve the hardware accelerator to allow the model to be computed efficiently. In Chapter III, we present the experimental results, and compare the proposed model with the results of other papers.

Our contribution in this thesis is as follows:

• We propose a hardware-friendly and less image based action recognition model.Our estimation indicates that the model only requires 2 images for each action, and the number of parameters is 5.11M, which is 30% less than [3].

• The accuracy of our proposed model is 65.1% for HMDB51 [1] and 89.6% for UCF101 [2] dataset. While these accuracies are lower than those achieved by Amin *et al.* [3], with a difference of 7.2% and 5.9% respectively, our model boasts a significantly faster execution speed, being x180 times faster.

II. Proposed Methods

A. Overall Scenario of Proposed Method

Our proposed action recognition model consists of four main components, as illustrated in Figure 1. The first component is an image segmentation module built on the eCNN [11] architecture. By employing image segmentation at the front end of the model, the subsequent 3D CNNs can focus on the action's location. This image segmentation process reduces the size of the feature image initially. The second component is a 2D CNN, which constructs a pathway to learn the appearance of an object and preserve the object's features after image segmentation. This step enhances the model's performance towards the end of the network. The third component comprises a 3D CNN that enables the model to learn temporal information between actions using the feature maps obtained from the previous image segmentation as input. The fourth component utilizes dynamic learning to facilitate self-learning within the model and improve its performance under varying sampling rates.

B. Video Down Sampling

Before proceeding with the identification process, we needed to convert the RGB video into RGB images. The

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

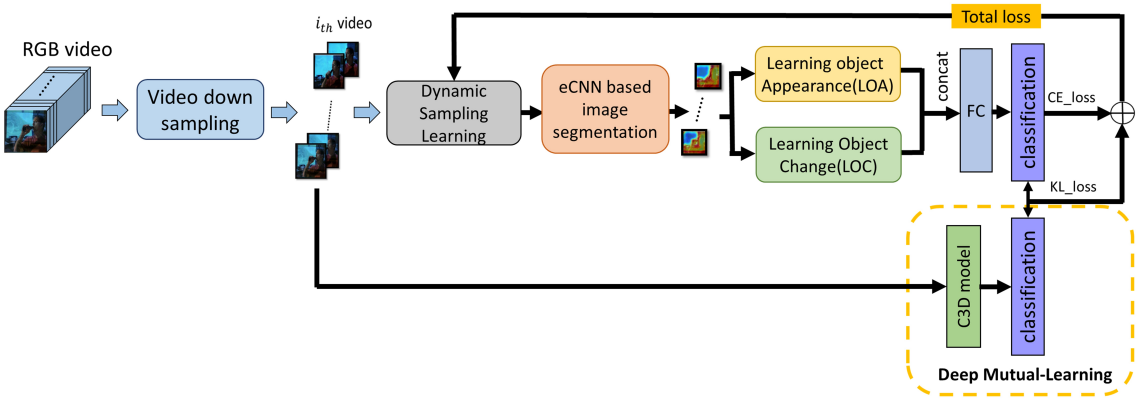


Fig. 1: Overview of the proposed architecture for video recognition system.

sampling rate for each video varied based on its length, with longer videos being sampled at larger intervals per frame. However, to maintain a sufficient amount of information for each video after sampling, we employed different sampling rates for each video, while ensuring that a minimum number of frames were sampled from each video.

C. Image segmentation based on eCNN

The image segmentation module can be broadly divided into two parts. The first part involves increasing the number of channels in the feature map. This is done to capture more spatial information. Specifically, when calculating the maximum pooling layer, the number of channels in the feature map is increased.

The second part consists of a stack of basic 2D eCNN modules. Each 2D eCNN basic block consists of two layers of 2D convolution. Furthermore, the kernel size of the convolutional layers in the 2D eCNN basic block is fixed at 3x3. This choice simplifies hardware calculations as the same method can be applied uniformly.

D. Dynamic Sampling Learning

Our work drew inspiration from the model proposed by Feichtenhofer *et al.* [12] and also took into account the experimental results presented by Wu *et al.* [13]. When evaluating the same action using the same model, we observed that certain actions performed better at lower sampling rates, while others performed better at higher sampling rates. This suggests that the model learns different performance characteristics for the same action at different sampling rates during the learning process.

Figure 2 illustrates the details of the dynamic sampling learning module. The process begins with an initial sampling interval, denoted as r_{curr}. Using this interval, images are sampled at regular intervals of r_{curr} from the processed video. These sampled images form a set called f_i. The model then processes these images and generates an estimation, from which a loss value is obtained.To determine the progress of learning, the obtained loss value is compared with the previous loss value. If the difference between the two values is less than the predetermined threshold, it indicates that the current model has already learned well at the current sampling interval. Consequently, the sampling interval for the next model to be learned is updated as $r_{next} = r_{curr} + 1$. This approach allows for adaptive adjustment of the sampling interval based on the model's learning progress.

E. Learning object Appearance(LOA)

To preserve the object information at the end of each frame after the image has been segmented, a separate pathway consisting of a 2D convolutional network is employed. This pathway is designed to prevent the previous spatial information from being influenced by the simultaneous learning of both temporal and spatial information within the 3D convolutional network.

By sending the segmented images as inputs to the pathway responsible for learning object appearance, the spatial information is retained and carried towards the backend of the network. This approach ensures that the spatial features of the objects are preserved and effectively learned within the model.

F. 3D based Short Sequence Action Module

The 3D-based Short Sequence Action Module is designed to enable the model to learn the temporal information between images using a network composed of 3D convolutional layers.

In the input section of this module, the output of the previous image segmentation module is stacked and then fed into the 3D convolution operation. By utilizing the image segmentation module at the front end of the network, we have successfully reduced the size of the feature image. This reduction in size not only saves memory but also reduces the computational requirements for 3D convolution within this module.

However, since the front end of the model utilizes 2D convolutional layers, the network has not yet learned temporal information. Therefore, this section of the module consists of a total of 6 stacked basic 3D blocks. Each 3D basic block comprises 2 layers of 3D convolution, 2 layers of batch normalization, and a final layer of ReLU activation.

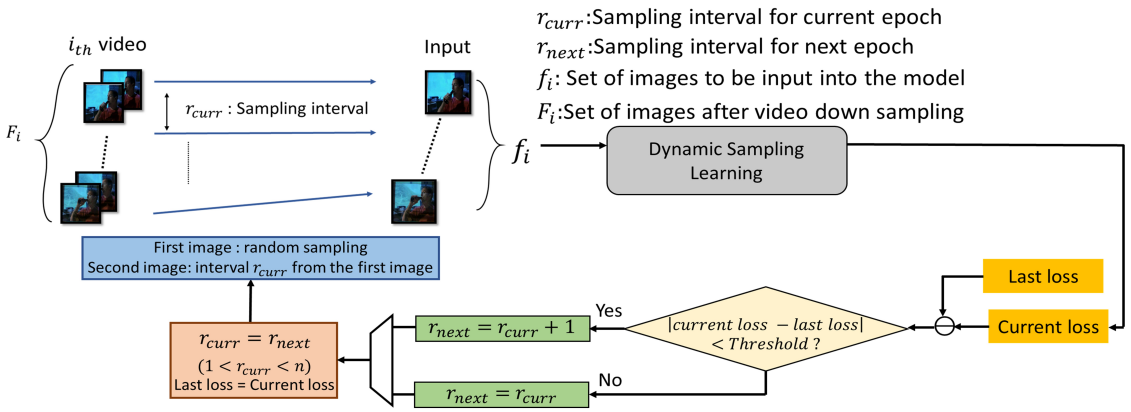


Fig. 2: Overall Framework for Dynamic Learning Sampling.

Additionally, the kernel sizes of the 3D convolution layers within the basic blocks are all set to 3x3x3.

G. Hardware Architecture

We have made improvements to the hardware architecture originally designed by Lin *et al.* [14] to support our action recognition model. Figure 3 illustrates the overall hardware architecture, which consists of three main components: the storage unit, the computing unit, and the SRAM controller.

The storage units include the Weight SRAM (WSRAM), which stores the model parameters, the Feature SRAM (FSRAM), which stores the input images for model calculations, the Fully Connected Layer SRAM (FCSRAM), which stores inputs for the fully connected layer, the Temporal SRAM (TSRAM), which temporarily stores the output after the image segmentation model has been computed, and the Output Reuse SRAM (ORSRAM), which temporarily stores the output to facilitate the calculation of the max-pooling layer.

The convolutional core module is used for performing 2D and 3D convolutional operations. It can process 64 input channels and generate 64 output channels simultaneously, with 9 processing elements (PEs) and stacks within each PE array. By matching the number of PEs with the input channels of the model and implementing a ring stream flow to reuse the feature map, the architecture reduces the need for additional storage units for each layer, resulting in more efficient computation and minimizing the storage requirements for partial sum calculations.

III. Experimental Results

A. Implement Details

The whole experiment was implemented and validated using the PyTorch framework [15]. During the model training phase, we relied on the experiments conducted by Kataoka *et al.* [16] and their corresponding results. These experiments demonstrated that models utilizing pre-training tend to outperform those trained from scratch. Consequently, our model incorporates the dataset proposed by Goyal *et al.* [17] for pre-training purposes.

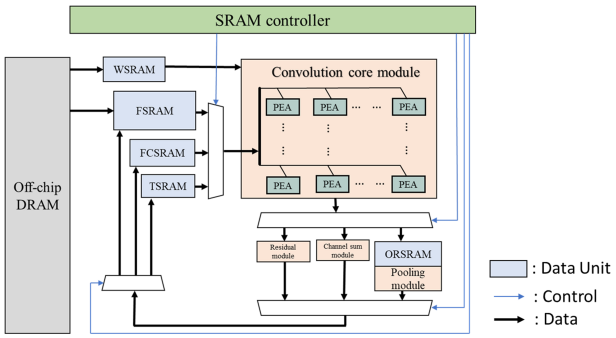


Fig. 3: Overview of the proposed architecture for action recognition accelerator.

B. Training & Testing Datasets

To pre-train our model, we utilize the Something-Something V1 dataset [17] dataset to pre-train our model. The estimated performance of the model is then evaluated using two other datasets: HMDB51 [1] and UCF101 [2].

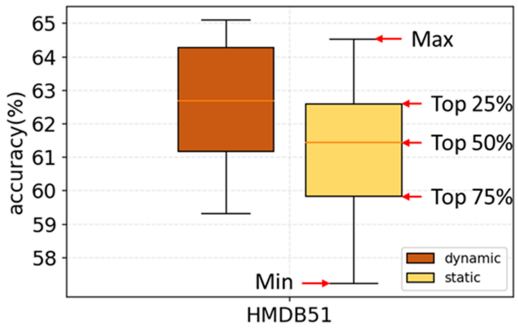


Fig. 4: Box plots on HMDB51 dataset using different training methods.

C. Ablation study

We employ two different approaches to train the model. The first approach involves training the model with a fixed

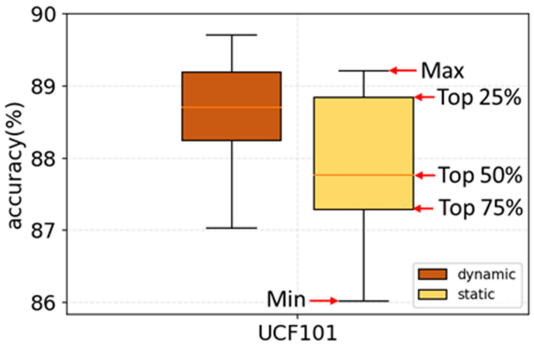


Fig. 5: Box plots on UCF101 dataset using different training methods.

sampling interval for each video, where we set the sampling interval to 5 in this experiment. The second approach is our proposed dynamic learning of the sampling rate, where the model adjusts the sampling interval based on the current situation. During the estimation, the sampling interval of the video is randomly changed. We conduct 30 experiments using these two training methods and present the results as box-plots in Figure 4, Figure 5 and Table I. Comparing the two methods, the dynamic learning of the sampling rate yields more centralized accuracy compared to the fixed method, indicating a more stable model. This suggests that the model trained with dynamic sampling exhibits improved consistency and reliability, as evidenced by the more concentrated results.

TABLE I: Testing accuracy on HMDB51 and UCF101 dataset using different training methods.

Data-set	Training method	mean	std	min	Q1 (top75%)	Q2 (top50%)	Q3 (top25%)	max
HMDB51	Dynamic (%)	62.58	0.018	59.32	61.17	62.68	64.28	65.11
	Fixed (%)	61.15	0.020	57.22	59.83	51.45	62.58	64.53
UCF 101	Dynamic (%)	88.64	0.007	87.03	88.24	88.7	89.14	89.6
	Fixed (%)	87.94	0.009	86.02	87.28	87.77	88.84	89.21

TABLE II: Pre-training + Mutual-learning(with C3D) on HMDB51 [1] dataset and UCF101 [2] dataset .

	HMDB51		UCF101	
Ratio (CE_loss/KL_loss)	C3D [5]	This work	C3D [5]	This work
Original	66.3%	62.5%	95.0%	89.1%
0.2/0.8	62.7%	63.4%	95.1%	89.2%
0.4/0.6	66.4%	63.7%	95.5%	89.1%
0.6/0.4	65.4%	62.6%	96.8%	89.4%
0.8/0.2	67.0%	65.1%	97.1%	89.6%

During the training process, we incorporate a method called Deep Mutual Learning, proposed by *et al.* [25], to enhance the learning of our model. This approach involves utilizing more than two models, including our own model and the C3D model proposed by Tran *et al.* [5]. We adjust the loss ratio between the models to optimize their learning performance. In our experiments, we tested four different loss scales on the two datasets, as presented in Table II. The results indicate that the best performance is achieved when the ratio between the Cross-Entropy (CE) loss and the Kullback-Leibler (KL) loss is set to 1:4.

D. Compare with other works

In this section, we evaluate our results on two datasets, HMDB51 and UCF101, using a GPU, in order to compare them with the results of other works. Our work stands out with the lowest number of parameters and computational cost among the works listed in Table III. Other works, such as Liu *et al.* [18], Tran *et al.* [6], Parmar *et al.* [19],Wu *et al.* [22], and Carrecira *et al.* [7] require larger model sizes and more computation, yet their accuracies on HMDB51 and UCF101 are lower compared to our work.

When compared with Amin *et al.* [3], our model surpasses their performance in terms of the number of parameters and speed, even though there is a slight decrease in accuracy, as demonstrated in Table III. Wu *et al.* [21] achieves the highest accuracy on UCF101 and the second-highest accuracy on HMDB51, but their parameter size is five times larger than our model.

IV. Conclusion

We have developed an action recognition model that addresses the issue of high memory requirements by integrating dynamic sampling learning and an image segmentation module at the front-end of the network. This approach enables us to reduce the size of feature images and the number of input images to the model. Additionally, our model exclusively utilizes RGB images as input, eliminating the need for optical flow images and significantly reducing pre-processing time, thereby achieving real-time performance. Notably, our model is capable of inference on our dedicated hardware accelerator.

Through our experimental results, we observed that employing dynamic sampling learning had a significant impact on the model's performance. The overall average accuracy and maximum accuracy were higher when using this method.

Ultimately, our proposed model achieves an accuracy of 65.1% for HMDB51 [1] and 89.6% for UCF101 [2] dataset. While these accuracy figures are 7.2% and 5.9% lower, respectively,compared to Amin *et al.* [3], our model offers significant advantages. We achieve a 30% reduction in the number of parameters, and the execution speed is 180 times faster.

TABLE III: Comparison with other works in terms of number of parameters and accuracy and other items.

Work	Input frames	Input size	#para (M)	GFLOP	Accuracy (HMDB51)	Accuracy (UCF101)	pretrained	fps
DS-GRU [3] (2021)	-	320x320	7.29	-	72.3%	95.5%	Caltec , SPID	One video per second
R-STAN [18] (2019)	12	224x224	-	3.8	53.40%	85.05%	ImageNet	-
ConvNet [6] (2017)	8	112x112	33.2	19.3	54.8%	85.8%	Sports-1M	-
HalluciNet-ing [19] (2020)	6	224x224	25	3.5	39.42%	79.83%	ImageNet	333.33
C3D [5]e (2015)	8	112x112	78.2	38.5	-	85.2%	Sports-1M	-
TVN [20] (2021)	2	224x224	11.1	13	72.1%	-	MiT	200
MSTSM [21] (2021)	8	224x224	22.4	4.01	72.19%	95.57%	ImageNet	-
CoViAR [22] (2019)	25 (Residual, Motion, I-frame)	224x224	33.6	3.5	56.16%	88.5%	ILSVRC 2012-CLS	145*25
R(2+1)D-18 [23] (2019)	32	112x112	33.2	2.4	54.9%	85.7%	Places 365 /ImageNet 1K	-
3DInception-v1 [7] (2017)	64	224x224	25	-	49.8%	84.5%	Kinetics 400	25
STAM [24] (2021)	64	256x256	96	1080	-	97%	Kinetics 400	307 (Nvidia V100)
This work	2	112x112	5.12	1.6	65.1%	89.6%	Something V1	360 (Nvidia 2080ti)

REFERENCES

[1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556–2563.

[2] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," in *Computer Vision and Pattern Recognition (cs.CV)*, 2012.

[3] A. U. M. D. P. U. H. W. Baik, "Efficient activity recognition using lightweight cnn and ds-gru network for surveillance applications," in *Applied Soft Computing, 2021*, vol. 103, 2021, pp. 1568–4946.

[4] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2740-2755, 2019.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[6] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *Computer Vision and Pattern Recognition (cs.CV), FOS: Computer and information sciences, FOS: Computer and information sciences*, 2017.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Z. Shou, X. Lin, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, S.-F. Chang, and Z. Yan, "Dmc-net: Generating discriminative motion cues for fast compressed video action recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1268–1277.

[9] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "Videomoco: Contrastive video representation learning with temporally adversarial examples," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 200–11 209.

[10] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," pp. 6450–6459, 2018.

[11] C.-T. Huang, Y.-C. Ding, H.-C. Wang, C.-W. Weng, K.-P. Lin, L.-W. Wang, and L.-D. Chen, "ecnn: A block-based and highly-parallel cnn accelerator for edge inference," *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," pp. 6201–6210, 2019.

[13] Z. Wu, C. Xiong, C.-Y. Ma, R. Socher, and L. S. Davis, "Adaframe: Adaptive frame selection for fast video recognition," pp. 1278–1287, 2019.

[14] K.-T. Lin, C.-T. Chiu, J.-Y. Chang, and S.-C. Hsiao, "High utilization energy-aware real-time inference deep convolutional neural network accelerator," pp. 1–5, 2021.

[15] S. A.Paszke, G. S.Chintala, Z. E.Yang, A. D. Z.Lin, and A. L.Antiga, "Automatic differentiation in pytorch," 2017.

[16] H. Kataoka, T. Wakamiya, K. Hara, and Y. Satoh, "Would mega-scale datasets further enhance spatiotemporal 3d cnns?" in *Computer Vision and Pattern Recognition (cs.CV), FOS: Computer and information sciences, FOS: Computer and information sciences*, 2020.

[17] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5843–5851.

[18] Q. Liu, X. Che, and M. Bie, "R-stan: Residual spatial-temporal attention network for action recognition," *IEEE Access*, vol. 7, pp. 82 246–82 255, 2019.

[19] P. Parmar and B. Morris, "Hallucinet-ing spatiotemporal representations using a 2d-cnn," *Computer Vision and Pattern Recognition(CVPR)*, 2019.

[20] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Tiny video networks," *Computer Vision and Pattern Recognition(CVPR)*, 2019.
[21] C.-T. C. Kun-Hsuan Wu, "Action recognition using multi-scale temporal shift module and temporal feature difference extraction based on 2d cnn," 2021.
[22] M.-C. Wu, C.-T. Chiu, and K.-H. Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," pp. 2202–2206, 2019.
[23] R. Girdhar, D. Tran, L. Torresani, and D. Ramanan, "Distinit: Learning video representations without a single labeled video," pp. 852–861, 2019.
[24] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?" *Computer Vision and Pattern Recognition(CVPR)*, 2021.
[25] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

Real-Time Processing for Weighted Pulse Decomposition of Photoplethysmography Signals Based on Interior Point Method in Wearable Devices for Hemodynamic State

Ting-Jui Wong and Pei-Yun Tsai

Department of Electrical Engineering,

National Central University, Taiwan

***Abstract* — Waveform decomposition technique can be applied to analyze photoplethysmography (PPG) signals from wearable devices for revealing the latent property of hemodynamic state. Solving for the component waves is considered as a constrained nonlinear optimization problem. The interior point method is used with a refined step size during line search. Exact derivatives in Hessian matrix and gradient are adopted in this work to acquire precise results. Inter-dependency and intra-dependency of the derivatives are exploited to reduce 93.7% computation complexity with a storage of only 62 ingredients. The block LDL decomposition with Bunch Kaufman pivoting strategy that takes advantage of symmetry property of Hessian matrix is employed to handle the linear equations described for the step size. Both single-precision and double-precision arithmetic are supported. From the experimental results, failure rate of 1.93% and one-cycle processing time of 0.1s are achieved by our program, which outperforms the commercial solver and the real-time processing requirement can be satisfied to demonstrate its applicability in the wearable devices.**

***Index Terms*—**waveform decomposition, interior point, line search, photoplethsymogram.

I. Introduction

Waveform decomposition technique is widely used in various applications, such as light detection and ranging (LiDAR) [1]-[3] and PPG signal analysis [4][5], to reveal the latent properties of the superimposed component waves. PPG signals derived from low-cost optical module equipped in all kinds of wearable devices to sense the blood volume change are commonly adopted for estimation of pulse wave velocity [6] and blood pressure [7]. With the waveform decomposition technique, the PPG pulse can be decomposed into the forward waves and the reflected waves which are mainly caused by the juncture between thoracic and abdominal aorta as well as the juncture between abdominal aorta and common iliac arteries. As a result, further intrinsic information of hemodynamic state can be observed.

Waveform decomposition can be considered as a constrained nonlinear optimization problem. In [1], the modified Levenberg-Marquardt (LM) method is used while Gauss-Newton algorithm is utilized in [2]. The interior point algorithms with boundary constraints derived from PPG features are adopted in [4] and [5]. All of these take into consideration Gaussian function as the basis of component waves because its morphology resembles the pulsatile wave. Although software packages have been implemented [8][9], we aim to develop the waveform decomposition technique to fulfill real-time processing requirements with both single-precision and double-precision arithmetic so as to support health monitoring in the wearable devices.

Usually, one PPG cardiac cycle can be decomposed into five Gaussian component waves, three in systole and two in diastole. Solving the optimization problem involves the iterative procedure. Thus, the precision effect influences either the processing time or the accuracy. We employ the exact derivatives of the Hessian matrix and the gradient vector during line search. Inter-dependency and intra-dependency are exploited to accelerate the computation. Only 62 ingredients are reserved. The required partial derivatives are generated by their paired products and 93.7% floating-point operations (flops) in raster scan order can be reduced. The strategy of refined step size for better adjustment is also proposed. We also take advantage of the symmetry of the Hessian matrix by using LDL decomposition to calculate the step size and to deal with the symmetric indefinite issue. From simulation results, more PPG pulses are successfully decomposed with fast processing time by our algorithm than by a Matlab generic solver. Also, high correlation exists between decomposed component waves from the synchronized finger and wrist PPG pulses.

II. System Model of Waveform Decomposition

Waveform decomposition can be described as an optimization problem that aims to minimize an objective function $f_{\mathrm{WMSE}}: \mathbb{R}^R \to \mathbb{R}$, where $\mathbf{x} \in \mathbb{R}^R$ denotes the vector of optimization variables with the inequality constraints $\boldsymbol{g}: \mathbb{R}^R \to \mathbb{R}^M$ as follows,

$$\min_{\mathbf{x}} f_{\mathrm{WMSE}}(\mathbf{x}) \tag{1}$$

$$\text{s.t. } \boldsymbol{g}(\mathbf{x}) \leq \mathbf{0} \tag{2}$$

Without loss of generality, the objective function is defined as the weighted mean square error between the normalized PPG pulse $p(n)$ and the fitting function $f(n, \mathbf{x})$.

$$f_{\mathrm{WMSE}}(\mathbf{x}) = \frac{1}{N} \times \sum_{n=1}^{N} (w(n) \times [p(n) - f(n, \mathbf{x})]^2) \tag{3}$$

where N is the length of one cardiac cycle, $w(n)$ denotes the weight. The fitting function $f(n, \mathbf{x})$ can be further described by the summation of Gaussian component waves,

$$f(n, \mathbf{x}) = \sum_{j=1}^{5} f_j\left(n, \mathbf{x}_j\right) \tag{4}$$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

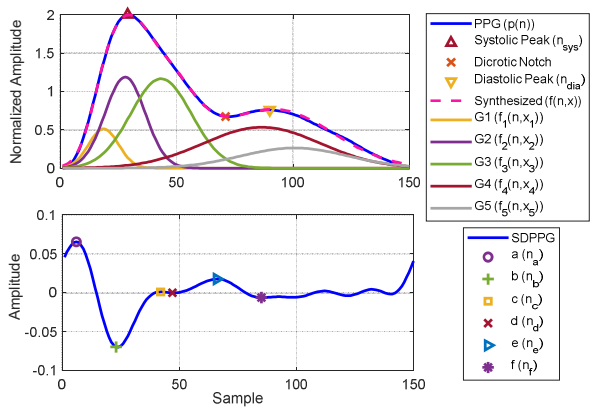


Fig. 1 Features from PPG as well as SDPPG and five component waves from waveform decomposition.

Table I: Boundary constraints of variables $[\alpha_j\ \beta_j\ \gamma_j]^T$, where $U_{b_{\beta 4}} = n_f/3 + 2N/3$.

	α_j		β_j		γ_j	
	L_{α_j}	U_{α_j}	L_{β_j}	U_{β_j}	L_{γ_j}	U_{γ_j}
$j=1$	0	A_b	n_a	n_b	0	$n_b/3$
$j=2$	0	A_{sys}	n_a	n_c	$n_a/3$	$n_d/3$
$j=3$	0	$0.8A_{sys}$	n_b	n_{d_n}	$n_b/3$	$n_{d_n}/3$
$j=4$	0	A_{dia}	n_{f_n}	U_{β_4}	0	$N/3$
$j=5$	0	A_{dia}	n_f	N	0	$N/3$

with $\mathbf{x} = [\mathbf{x}_1^T \quad \dots \quad \mathbf{x}_5^T]^T$ and

$$f_j(n, \mathbf{x}_j) = \alpha_j exp\left(-\frac{(n-\beta_j)^2}{2\gamma_j^2}\right), \tag{5}$$

where $\mathbf{x}_j = [\alpha_j\ \beta_j\ \gamma_j]^T$; α_j is the amplitude gain; β_j is the peak position; γ_j is the wave width of the jth Gaussian component wave (Gj).

From [5], the boundary constraints $\boldsymbol{g}(\mathbf{x})$ are given in Table I, where n_ω represents the position of feature ω ; A_ω indicates the PPG amplitude of feature ω; L_y and U_y are the lower bound and upper bound of variable y with $L_y \le y \le U_y$ and $y \in \{\alpha_j, \beta_j, \gamma_j\}$. The features of systolic peak (sys) and diastolic peak (dia) are local maxima of PPG while feature points a, b, c, d, e, and f are the extrema extracted from the second-derivative PPG (SDPPG) as shown in Fig. 1. Point d_n (f_n) is either zero-crossing point if it exists or middle point between points d (f) and e. Besides 30 constraints in Table I, four additional constraints are

$$\beta_j < \beta_{j+1} \qquad \text{for } j = 1, \dots, 4. \tag{6}$$

for delineating the ordering of peak positions. Hence, $R = 15$, $M = 34$ in the optimization problem. In addition, the weight is chosen to distinguish the informative portion of the PPG pulse between n_a and n_f, which is given by

$$w(n) = \begin{cases} 5 & n = n_a, \dots, n_f \\ 1 & else \end{cases}.$$

III. Waveform Decomposition Based on Interior Point Method

A. Interior Point Method (IPM)

Interior point methods can solve constrained optimization problems. The inequality constraints are transformed into equality constraints by introducing a vector of slack variables $\mathbf{s} = [s_1 \quad \cdots \quad s_M]^T$ as follows.

$$\min_{\mathbf{x},\mathbf{s}} \boldsymbol{\varphi}_\mu(\mathbf{x}, \mathbf{s})$$

$$\boldsymbol{\varphi}_\mu(\mathbf{x}, \mathbf{s}) \equiv f_{\text{WMSE}}(\mathbf{x}) - \mu \textstyle\sum_m \ln(s_m) \tag{7}$$

$$\text{s.t. } \boldsymbol{g}(\mathbf{x}) + \mathbf{s} = \mathbf{0}. \tag{8}$$

A logarithmic barrier term is included in the objective function to create barrier near the boundary of the feasible region. Value μ denotes the positive barrier parameter for each suboptimal problem and decreases gradually in the iterative procedure by a factor θ_μ.

$$\mu \leftarrow \theta_\mu \times \mu. \tag{9}$$

Hence, the effect of the barrier term diminishes because value μ converges to zero gradually.

The Lagrangian function $\boldsymbol{\mathcal{L}}$ is formed for solving the constrained optimization problem as follows.

$$\boldsymbol{\mathcal{L}}(\mathbf{x}, \mathbf{s}, \boldsymbol{\lambda}_g; \mu) = f_{\text{WMSE}}(\mathbf{x}) - \mu \sum_m \ln(s_m) + \boldsymbol{\lambda}_g^T(\boldsymbol{g}(\mathbf{x}) + \mathbf{s}) \tag{10}$$

where $\boldsymbol{\lambda}_g \in \mathbb{R}^M$ denotes the vector of Lagrange multipliers with nonnegative elements. Therefore, the minimum occurs when the partial derivatives of the Lagrangian function are zero. In other words, $\boldsymbol{\nabla \mathcal{L}} = \mathbf{0}$. Solution of the nonlinear equations can be found by the Newton's method. Let the step direction be $\mathbf{d} \equiv \left[\mathbf{d}_\mathbf{x}^T\ \mathbf{d}_\mathbf{s}^T\ \mathbf{d}_{\boldsymbol{\lambda}_g}^T\right]^T$ with $\mathbf{d}_\mathbf{x} \in \mathbb{R}^R$, $\mathbf{d}_\mathbf{s} \in \mathbb{R}^M$, and $\mathbf{d}_{\boldsymbol{\lambda}_g} \in \mathbb{R}^M$. Applying Newton's method, we have

$$(\boldsymbol{\nabla^2 \mathcal{L}})\mathbf{d} = -\boldsymbol{\nabla \mathcal{L}} \tag{11}$$

where $\boldsymbol{\nabla^2 \mathcal{L}}$ and $\boldsymbol{\nabla \mathcal{L}}$ denote the Hessian matrix and gradient of $\boldsymbol{\mathcal{L}}$.

To keep the variables in the feasible region, the maximal step size $\alpha_x^{\max}$, $\alpha_s^{\max}$, $\alpha_\lambda^{\max}$ must be constrained [9],

$$\alpha_x^{max} = \max\{\alpha \in (0, 1]: \tau(\mathbf{L} - \mathbf{x}) \le \alpha \mathbf{d}_\mathbf{x} \le \tau(\mathbf{U} - \mathbf{x})\} \tag{12}$$

$$\alpha_s^{max} = \max\{\alpha \in (0, 1]: -\tau \mathbf{s} \le \alpha \mathbf{d}_\mathbf{s}\} \tag{13}$$

$$\alpha_\lambda^{\max} = \max\left\{\alpha \in (0, 1]: -\tau \boldsymbol{\lambda}_g \le \alpha \mathbf{d}_{\lambda_g}\right\} \tag{14}$$

$$\alpha_z^{\max} = \min(\alpha_x^{max}, \alpha_s^{max}) \tag{15}$$

where $\mathbf{L}$ and $\mathbf{U}$ represent vectors of lower bounds and upper bounds of $\mathbf{x}$ in Table I; $\tau \in (0, 1)$ is a parameter defined for the fraction to the boundary. The variables are then updated along the step direction in the iterative procedure,

$$\mathbf{x} \leftarrow \mathbf{x} + \alpha_P \alpha_z^{\max} \mathbf{d}_\mathbf{x}, \tag{16}$$

$$\boldsymbol{\lambda}_g \leftarrow \boldsymbol{\lambda}_g + \alpha_P \alpha_\lambda^{\max} \mathbf{d}_{\boldsymbol{\lambda}_g}, \tag{17}$$

$$\mathbf{s} = -\boldsymbol{g}(\mathbf{x}). \tag{18}$$

In addition, we propose to include the ratio $\alpha_P \in \{1/P, 2/P, 3/P, \dots, P/P\}$ to refine the step size during line search, which is selected among the set according to the smallest value of $\varphi_\mu(\mathbf{x}, \mathbf{s})$ in (7).

As a result, the interior point method contains two loops. In the outer loop, the barrier parameter μ changes to define the suboptimal problem while in the inner loop, the solutions $\mathbf{x}$, $\mathbf{s}$ and $\boldsymbol{\lambda}_g$ for a given barrier parameter μ are calculated.

B. Methods for Solving Linear Equations

Algorithm 1 Interior Point Method for WPD

Input: initial values $\mathbf{x_0}$, lower bound $\mathbf{L}$ and upper bound $\mathbf{U}$
Output: variables of Gaussian component waves $\mathbf{x}$
Initialization: $\mu = \mu_0,\ i_1 = 1,\ \mathbf{x} = \mathbf{x}_0$

```
while i1 ≤ I1
    update μ by (9), i2 = 1
    while i2 ≤ I2
        Update s by (18)
        Form ∇ℒ and ∇²ℒ
        Solve linear equations (11) by LDL or CG
        Determine α_T, α_z^max, α_λ^max by (12)-(15)
        Update x by (16)
        Update λ_g by (17)
        if early termination is true
            return x
        endif
        i2 ← i2 + 1
    end while
    i1 ← i1 + 1
end while
return x
```

Let $\mathbf{A} = \nabla^2\mathcal{L}$ and $\mathbf{b} = -\nabla\mathcal{L}$. There are several methods that can be used for solving linear equation of $\mathbf{Ad} = \mathbf{b}$ in (11). The LU, LDL, and QR decomposition approaches are categorized as direct methods while conjugate gradient (CG) and preconditioned CG are iterative methods [10]. The direct method has the advantage of the time complexity while the iterative method such as CG consumes less memory. Note that the Hessian matrix $\mathbf{A}(= \nabla^2\mathcal{L})$ is symmetric. Factorizing matrix $\mathbf{A}$ into $\mathbf{LDL^T}$ can exploit the symmetry property and thus LDL decomposition has lower computation complexity than other direct methods [11].

Block LDL with Bunch Kaufman pivoting strategy [12] can decompose a square matrix that is symmetric indefinite into a lower triangular matrix $\mathbf{L}$, a block diagonal matrix $\mathbf{D}$, and a permutation matrix $\mathbf{P}$, namely

$$\mathbf{A} = \mathbf{P}^T\mathbf{L}\mathbf{D}\mathbf{L}^T\mathbf{P}. \tag{19}$$

Let

$$\mathbf{L}\mathbf{y_1} = \mathbf{P}\mathbf{b},\quad \mathbf{D}\mathbf{y_2} = \boldsymbol{y_1},\quad \mathbf{L}^T\mathbf{y_3} = \boldsymbol{y_2}, \tag{20}$$

$$\mathbf{P}\mathbf{d} = \mathbf{y_3}. \tag{21}$$

The desired vector $\mathbf{d}$ can be derived by forward and backward substitutions and permutation as described in (20) and (21), respectively.

The iterative CG method is suitable for solving large and sparse systems of linear equations. In the implementation, the maximum numbers of iterations are set as 5 and 25 times the dimension of Hessian matrix for double-precision and single-precision arithmetic, respectively. The residual will be updated and checked for every step. The iterative procedure will be terminated if the residual is smaller than or equal to $\Delta(= 10^{-10})$ to ensure that the result is a suitable solution for the equations. The complete algorithm is given in Algorithm 1 with $I_1 = 100,\ I_2 = 20$.

C. *Calculations for Exact Hessian and Gradient*

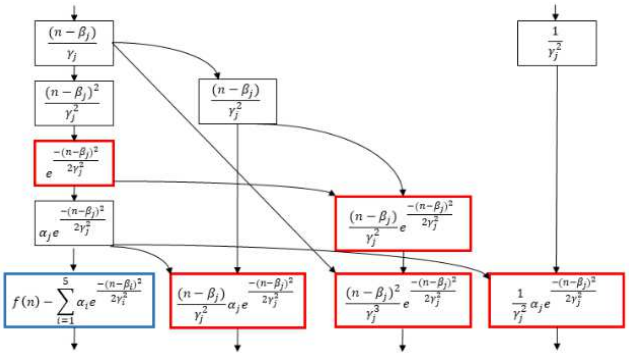


Fig. 2 Data flow graph of computations in Hessian matrix.

The Hessian matrix and gradient vector contain derivatives of the Lagrangian function. Although the finite difference can be used to approximate the derivatives in some software package, the exact derivatives are adopted in this work to acquire precise results during line search. However, the computation time of decomposing one PPG pulse is too long to satisfy real-time processing constraint if the derivatives are calculated intuitively in raster scan order. Since the derivative of an exponential function still consists of the original exponential function, the calculations for Hessian matrix and gradient vector can be simplified by exploiting dependency. Two types of dependency, intra-dependency and inter-dependency, are utilized. Consider two terms in Hessian matrix,

$$\frac{\partial}{\partial\beta_k}\left(\frac{\partial}{\partial\alpha_j}\right)\mathcal{L} = \frac{2}{N}\sum_n w(n)\left(e^{\frac{-(n-\beta_j)^2}{2\gamma_j{}^2}}\right)\left(\alpha_k e^{\frac{-(n-\beta_k)^2}{2\gamma_k{}^2}}\right)\left(\frac{n-\beta_k}{\gamma_k^2}\right), \tag{22}$$

$$\frac{\partial}{\partial\alpha_k}\left(\frac{\partial}{\partial\alpha_j}\right)\mathcal{L} = \frac{2}{N}\sum_n w(n)\left(e^{\frac{-(n-\beta_j)^2}{2\gamma_j{}^2}}\right)\left(e^{\frac{-(n-\beta_k)^2}{2\gamma_k{}^2}}\right). \tag{23}$$

The intra-dependency of $\frac{n-\beta_k}{\gamma_k^2}$ in (22) and the inter-dependency between (22) and (23) can be seen. The data flow graph is partly depicted to reveal the dependency as shown in Fig. 2. The result of the successor related with intra-dependency in the flow graph overwrites the result of the predecessor in the same storage to keep minimal storage size while the results associated with inter-dependency can be fetched from these storages to reduce computations. Therefore, for all j, k, a total of 62 ingredients are saved, indicated by the red and blue blocks, and the partial derivatives in Hessian matrix can be formed by their paired products. For each n value, only 245 flops and 5 exponential functions are executed as opposed to 3905 flops in raster scan order, 93.7% flops are eliminated in the multi-Gaussian fitting model.

IV. Simulation Results

The program of weighted pulse decomposition is realized by C without multi-threading. The simulations and measurements are first carried out with Intel CPU i7-6700, DDR4 32GB, under Ubuntu 18.04.6 for comparison with the Matlab solver.

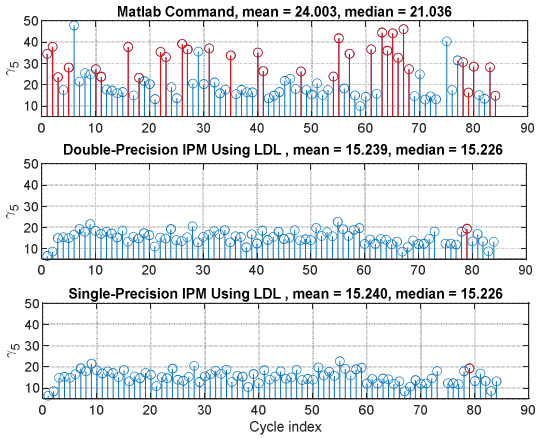


Fig. 3 Decomposition results of wave width of $G5$ from different approaches.

Table I Processing time when $\theta_\mu = 0.8$.

Case Setting	Case 1 CG $P = 16$	Case 2 LDL $P = 16$	Case 3 CG $P = 16$	Case 4 LDL $P = 16$	Case 5 LDL $P = 4$
SP	12.31s	1.00s	11.50s	0.24s	0.10s
DP	2.80s	1.87s	1.29s	0.36s	0.12s

The settings of the compiler flags -std=c99, -lm, -O3 are used with GCC 7.5.0.

In a one-minute PPG sequence, there are several tens of cardiac cycles. Usually, the morphology of PPG pulses and the component waves are slowly time-varying in stationary condition of the measurement conducted on a subject in a seated position. The signal quality index (SQI) of waveform decomposition is assessed by mean squared error, set to 2×10^{-3}, between $p(n)$ and $f(n, \mathbf{x})$. Fig. 3 shows the decomposition results of the width of $G5$, which plays an important role of the strongly-reflected diastolic wave, by Matlab solver (fmincon), double-precision and single-precision IPM using LDL from top to bottom respectively when $P = 16$. If the decomposed result fails to pass the SQI assessment, it is indicated by the red color. It is clear that our IPM program can generate stable decomposed component waves. In addition, both the single-precision arithmetic and double-precision arithmetic produce quite similar results. Only insignificant difference can be seen in their mean values.

Using different approaches to solve linear equations might generate different floating-point results. Usually, CG is sensitive to arithmetic precision and may need more iterations or compensations in order to suppress round-off errors for deriving small residual of the linear equations. Although the difference of the results obtained from CG and LDL is almost negligible when double-precision (DP) arithmetic is used, sometimes, apparent difference appears in some cases if single-precision (SP) arithmetic is employed. Table I lists the processing time, which is measured by the average of processing the weighted pulse decomposition problem using single and double precision arithmetic, respectively, for ten times. The intuitive computation in raster-scan order is adopted in Case 1 and Case 2. The improvement by exploiting intra-dependency and inter-dependency is shown in Case 3 and Case 4. In Case 5, the processing time is further reduced if operations such as schur complement and swap are implemented by taking advantage of the symmetric property for LDL. Besides, P is set to 4 for the refinement ratio α_P in (16) and (17). Thus, the effective of the proposed strategy is demonstrated. In addition, the processing time of the Matlab solver is about 0.15s. The program is then run on Raspberry Pi 4 Model B with ARM Cortex A72 processor. The processing times corresponding to DP and SP are listed in Table II with LDL for $P = 4$ and $P = 2$. Since one PPG cycle can be processed within 0.10s on Intel CPU or 0.24s on ARM for P=4, the real-time processing criterion can be satisfied. Furthermore, block LDL is a better approach for solving waveform decomposition owing to lower computation complexity, lower processing time, higher accuracy compared to the CG method if both single precision and double precision arithmetic operations are considered.

Table II Processing time on Raspberry Pi 4 Model B.

	DP		SP	
	mean	median	mean	median
P=4	0.31s	0.3s	0.24s	0.23s
P=2	0.256s	0.25s	0.198s	0.19s

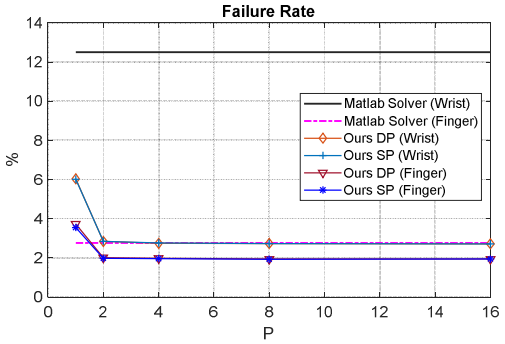


Fig. 4 Unqualified PPG cycles of different settings.

We then use the LDL-based IPM program to process the data set with 21441 finger PPG cycles and 20370 wrist PPG cycles from 60 persons. Fig. 4 shows the failure rate of unqualified decomposition results assessed by SQI versus various setting of P. Default setting is used for Matlab solver. The introduction of refined step size helps to achieve better adjustment of the variables to the barrier problem in (10). The effect of α_P in (16) and (17) during line search is verified. Usually the morphology of wrist PPG pulses is less recognizable than that of finger PPG pulses and thus more unqualified cycles are generated. However, when $P \geq 2$, our program shows a lower failure rate compared to the commercial solver, especially for wrist PPG. Table III provides the correlations of peak position and wave width from component waves $G3$, $G4$, $G5$. The correlations of $G1$ and $G2$ between writs PPG pulse and finger PPG pulse are lower than the values listed in the Table. There are 342 recordings from 60 subjects. The medians of the feature values from the synchronized and SQI-qualified finger and wrist PPG cycles of one-minute sequence are calculated and their correlation coefficients are examined. We can see that high correlation coefficients exist between the component waves of finger and

wrist PPG signals that are processed by the weighted waveform decomposition algorithm. Hence, the latent property of PPG pulse is acquired with consistency to a certain degree.

Table III Comparison of correlation coefficient (C.C.) of the decomposed component waves $G3$, $G4$, $G5$ between wrist and finger PPG cycles.

Feature	This work SP ($P = 4$)	This work DP ($P = 4$)	Matlab
Peak Position β_3	0.561	0.562	0.569
Peak Position β_4	0.679	0.679	0.658
Peak Position β_5	0.908	0.908	0.920
Wave Width γ_3	0.797	0.797	0.819
Wave Width γ_4	0.857	0.857	0.837
Wave Width γ_5	0.906	0.906	0.899

V. Conclusion

The weighted pulse decomposition of PPG signals solved by interior point method using exact derivatives is developed for wearable devices. The introduction of step refinement helps to improve the failure rate. The inter-dependency and intra-dependency in Hessian matrix and gradient are exploited to reduce 93.7% flops for partial derivatives of Gaussian-related function. Only 62 ingredients are reserved in the look-up table. The symmetry property of Hessian matrix is also utilized for schur complement and swap in LDL decomposition. From the experimental results, our program achieves better weighted pulse decomposition results of lower failure rate and less processing time compared to the commercial generic solver. In addition, both single-precision and double-precision arithmetic operations are supported for applications in wearable devices.

Acknowledgement

This work is supported by MediaTek Company under the Grant MTKC-2021-0477.

References

[1] F. Xu, F. Li and Y. Wang, “Modified Levenberg–Marquardt-Based Optimization Method for LiDAR Waveform Decomposition,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 4, pp. 530-534, April 2016.

[2] Z. Wang, L. Xu, D. Li, Z. Zhang and X. Li, “Online Multi-Target Laser Ranging Using Waveform Decomposition on FPGA,” *IEEE Sensors Journal*, vol. 21, no. 9, pp. 10879-10889, 1 May1, 2021.

[3] G. Zhou et al., “Comparison Analysis of Five Waveform Decomposition Algorithms for the Airborne LiDAR Echo Signal,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7869-7880, 2021.

[4] R. Couceiro, et al. “Assessment of Cardiovascular Function from Multi-Gaussian Fitting of a Finger Photoplethysmogram,” *Physiological measurement*, 36, pp, 1801-1825, 2015

[5] C. -H Huang *et al*., “Weighted Pulse Decomposition Analysis of Fingertip Photoplethysmogram Signals for Blood Pressure Assessment,” *IEEE International Symposium on Circuits and Systems (ISCAS),* 2020, pp. 1-5.

[6] S. R. Alty, N. Angarita-Jaimes, S. C. Millasseau and P. J. Chowienczyk, "Predicting Arterial Stiffness From the Digital Volume Pulse Waveform," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 12, pp. 2268-2275, Dec. 2007.

[7] H. Truong, A. Montanari and F. Kawsar, “Non-Invasive Blood Pressure Monitoring with Multi-Modal In-Ear Sensing,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6-10

[8] R. Waltz, J. Morales, J. Nocedal, *et al*.. “An Interior Algorithm for Nonlinear Optimization That Combines Line Search and Trust Region Steps,“ *Math. Program.* 107, 391-408 (2006).

[9] A. W¨achter, L. T. Biegler, “On the Implementation of an Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming,” *Math. Program.*106, 25–57 (2006).

[10] Gene H. Golub and Charles F. Van Loan, Matrix Computations 1996, page 192

[11] V. Barwell and J. A. George, “A Comparison of Algorithms for Solving Symmetric Indefinite Systems of Linear Equations,” *ACM Trans. Math. Softw.* 2, 242-251, Sep. 1976.

[12] J. R. Bunch and L. Kaufman, “Some Stable Methods for Calculating Inertia and Solving Symmetric Linear Systems.” *Mathematics of Computation*, vol. 31, no. 137, 1977, pp. 163–79.

QoS-Aware Downlink Beamforming for Joint Transmission in Multi-Cell Networks

Chen-Yen Lin
MediaTek Inc.
Hsinchu, Taiwan
freddie.lin@mediatek.com

Kuang-Hao (Stanley) Liu
Institute of Communications Engineering
National Tsing Hua University, Hsinchu, Taiwan
khliu@ee.nthu.edu.tw

***Abstract*—Multi-cell cooperation is an effective means to improve service quality to cellular users. Existing work primarily focuses on interference cancellation using all the degrees of freedom (DoF). This leads to low service quality for some users with poor channel quality to its serving base station. This work investigates the multi-cell beamforming design for simultaneously enhancing the downlink signal strength and mitigating interference. We first consider the ideal case when perfect channel state information (CSI) is available for determining the beamforming vectors and then extend to the case of imperfect CSI. For both cases, the beamforming optimization problems are non-convex. Assuming perfect CSI, we obtain the optimal joint transmit (JT) beamforming vectors based on the uplink-downlink duality. In the presence of unknown CSI errors, we use the semidefinite relaxation (SDR) with Bernstein-type inequality to derive the robust JT beamforming. Numerical results are presented to evaluate the performance of the proposed schemes.**

***Index Terms*—Imperfect CSI, joint transmission, multi-input and multi-output (MIMO), quality-of-service (QoS), robust beamforming, and semidefinite relaxation (SDR).**

I. Introduction

Multi-cell cooperation is a practical approach to enhance the service quality to subscribers in the cellular network. Through the coordination across multiple base stations (BSs) in different cells, a large virtual antenna array can be formed to increase the spatial degree of freedom (DoF), which may be used to enhance signal power by sending the same data to user equipment (UE), known as the joint transmission (JT). Alternatively, it can be used to mitigate interference by coordinated beamforming (CB). The increased DoF of multi-cell cooperation permits a flexible and robust design of multi-cell beamforming schemes that perform JT while simultaneously alleviating undesired interference, which is this work's objective.

In the literature, several works have explored multi-cell beamforming design with different objectives. As the single-cell counterpart, the summed throughput is a common design objective for multi-cell beamforming design [1]. Since it is challenging to solve the downlink (DL) throughput maximization problem directly, it is shown in [1] that reformulating the beamforming design as an uplink (UL) dual problem allows a more solvable form and the solution of the dual problem delivers the same summed throughput as the DL problem. Such a UL-DL dual relation has long been used to facilitate beamforming design in multiple-input multiple-output (MIMO) systems, such as the early work [2] for the single-cell MIMO system, and more recent work [3] for the cell-free MIMO system. The UL-DL duality has also been used to minimize the total transmit power required to meet a prescribed quality-of-service (QoS) constraint, namely a target signal-to-interference-plus-noise ratio (SINR) in [4] assuming perfect CSI and in [5], [6] considering CSI errors. Again, they assume a single serving BS per UE [4]–[7] and use all the DoF for interference mitigation. Another type of beamforming design considers fairness, as in [8] that jointly designs the transmit beamforming across multiple BSs to maximize the worst SINR among all the served UEs. Such a max-min formulation can also be tackled through the UL-DL duality. Unlike aforementioned works that use instantaneous CSI to design the beamforming, long-term CSI in terms of channel correlation matrices is adopted in [8] with the assumption of ideal CSI. Another popular method for solving the beamforming design problem is to linearize non-convex SINR expressions through the semi-definite relaxation (SDR) [7], [9], which provides a convex approximation to the non-convex optimization problem without closed-form solution. In contrast, the UL-DL duality is established based on the Lagrange duality that provides a closed-form solution for the optimal beamforming vectors.

To fully leverage the benefit of multi-cell cooperation, this work explores the optimal JT beamforming at DL, where several BSs are grouped into a cluster and transmit DL data to the same UE simultaneously. This generalizes the existing work with the cluster size of one and can be implemented using the control signaling defined in the standard, such as 3GPP Release 16 [10]. This way, multiple DL data streams can be transmitted simultaneously to different UEs to increase spectral efficiency. We aim to minimize the total transmit power while satisfying a target SINR as in [4]–[7] for CB. Unlike CB, our problem is more complicated because the transmit beamforming needs to maximize the summed signal power at the intended UE and suppress the interference to the other UEs, provided with the increased DoF via multi-cell cooperation. We show that the UL-DL duality can still be applied to obtain the optimal JT beamforming when global CSI is perfect. In practice, the backhaul links for CSI exchange have limited capacity, and thus the CSI exchanged across the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

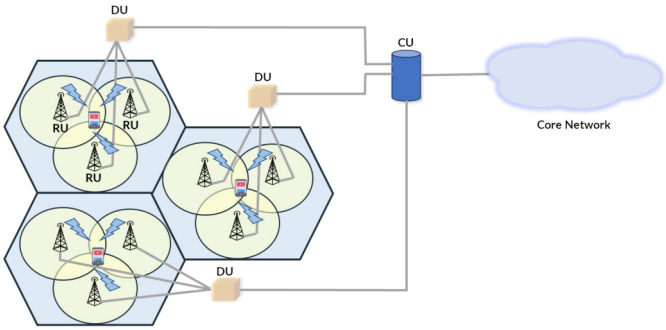


Fig. 1. Illustration of joint transmission in a multi-cell network.

cooperating BSs within the cluster is not perfect. The second part of this work addresses the optimal JT beamforming problem with imperfect CSI by extending [5] that considers the single-cell scenario. We demonstrate the significant gain of the optimal JT beamforming compared with the traditional non-JT beamforming and the zero-forcing (ZF) based beamforming through simulation results.

The remainder of this paper is organized as follows. Sec. II explains the system model. Sec. III presents the optimal JT beamforming design with perfect and imperfect CSI, respectively. Simulation results are shown in Sec. IV. Concluding remarks are given in Sec. V.

II. System Model

Consider a multi-cell network with multiple BSs deployed in a geographic area. Following the virtualized radio access network (vRAN) structure in 5G [11], a BS is split into different functional blocks, namely, the radio unit (RU) handling the RF transmission and reception, the distributed unit (DU) that carries out baseband signal processing, and the central unit (CU) performing the higher-layer protocols. Without loss of generality, one CU is deployed in the considered area, which is virtually partitioned into a set of cells denoted as $\mathcal{N}_c = \{\mathcal{C}_1, \cdots, \mathcal{C}_{N_c}\}$, each with N_b RUs connected to one DU via fronthaul links. The DU collects CSI from the RUs within the same cell and designs the DL beamforming. In this work, the term RU and BS will be used interchangeably. Each BS has N_t antennas, and each user equipment (UE) has a single antenna.

For the i-th cell, denote $\mathcal{C}_i = \{\text{BS}_{ib}\}_{b=1}^{N_b}$ the set of BSs connected to the same DU for $i = 1, \cdots, N_c$ and $\mathcal{U}_i = \{\text{UE}_{iu}\}_{u=1}^{|\mathcal{U}_i|}$ the set of UEs within the coverage area of the i-th cell. Considering the imperfect CSI, the true channel vector between BS_{ib} and UE_{iu} is modeled as

$$\mathbf{h}_{ib,iu} = \tilde{\mathbf{h}}_{ib,iu} + \mathbf{e}_{ib,iu} \tag{1}$$

where $\tilde{\mathbf{h}}_{ib,iu} \in \mathbb{C}^{N_t}$ is the estimated channel vector, and $\mathbf{e}_{ib,iu}$ represents the CSI error, which is assumed to be a complex Gaussian random vector with zero mean and the covariance matrix $\mathbf{Q}_{ib,iu} = \epsilon_{ib,iu}^2 \mathbf{I}_{N_t}$ where $\epsilon_{ib,iu}^2$ is referred to as the error radius. Then we can express the CSI error vector as

$$\mathbf{e}_{ib,iu} = \mathbf{Q}_{ib,iu}^{1/2} \mathbf{v}_{ib,iu} \tag{2}$$

where $\mathbf{v}_{ib,iu}$ is a normalized complex random vector with zero mean and the covariance matrix $\mathbf{I}_{N_t}$.

For simplicity, we assume that a cluster consists of all the BSs within the same cell, i.e., BSs in each cluster are coordinated by one dedicated DU. In the JT mode, all the BSs in the cluster send the same signal to the intended UE. The received signal at UE_{iu} is

$$y_{iu} = \sum_{b\in\mathcal{C}_i} \mathbf{h}_{ib,iu}^H \mathbf{w}_{ib,iu} s_{iu} + \sum_{b\in\mathcal{C}_i} \sum_{v\in\mathcal{U}_i\setminus u} \mathbf{h}_{ib,iu}^H \mathbf{w}_{ib,iv} s_{iv} + \sum_{\mathcal{C}_j\in\mathcal{N}_c\setminus\mathcal{C}_i} \sum_{b'\in\mathcal{C}_j} \sum_{v'\in\mathcal{U}_j} \mathbf{h}_{jb',iu}^H \mathbf{w}_{jb',jv'} s_{jv'} + z_{iu} \tag{3}$$

where s_{iu}, $\mathbf{w}_{iu} \in \mathbb{C}^{N_t}$, and $z_{iu} \sim N(0, \sigma^2)$ represent the DL signal, the transmit beamforming vector, and additive white Gaussian noise (AWGN) associated with UE_{iu}, respectively. The first term on the right-hand side is the collection of the intended signal for UE_{iu}, the second term is the intra-cell interference (Intra-CI), and the last term corresponds to the inter-cell interference (Inter-CI). Assuming the signal energy is normalized to one, the signal-to-interference-plus-noise ratio (SINR) of UE_{iu} based on true CSI is given by

$$\text{SINR}_{iu}(\mathbf{h}) = \frac{\sum_{b\in\mathcal{C}_i} |\mathbf{h}_{ib,iu}^H \mathbf{w}_{ib,iu}|^2}{I_{\text{Intra}} + I_{\text{Inter}} + \sigma^2} \tag{4}$$

where

$$I_{\text{Intra}} = \sum_{b\in\mathcal{C}_i} \sum_{v\in\mathcal{U}_i\setminus u} |\mathbf{h}_{ib,iu}^H \mathbf{w}_{ib,iv}|^2$$
$$I_{\text{Inter}} = \sum_{\mathcal{C}_j\in\mathcal{N}_c\setminus\mathcal{C}_i} \sum_{b'\in\mathcal{C}_j} \sum_{v'\in\mathcal{U}_j} |\mathbf{h}_{jb',iu}^H \mathbf{w}_{jb',jv'}|^2 \tag{5}$$

III. Optimal Beamforming Design for JT

Based on the considered system model, we formulate the optimal JT beamforming problem in this section, considering perfect and imperfect CSI, respectively.

A. Problem Formulation

For UE_{iu}, the DU collects the CSI from all the BSs in $\mathcal{C}_i$ and designs the beamforming vectors jointly to minimize the total transmit power consumption for serving all the UEs while satisfying the SINR requirement of individual UE. Denote γ_{iu} the SINR requirement of UE_{iu}, the JT beamforming design at DL can be formulated as

$$\text{(JT-DL)} \quad \max_{\{\mathbf{w}_{ib,iu}\}} \sum_{\mathcal{C}_i\in\mathcal{N}_c} \sum_{b\in\mathcal{C}_i} \sum_{u\in\mathcal{U}_i} \|\mathbf{w}_{ib,iu}\|^2$$
$$\text{s.t. } \text{SINR}_{iu}(\mathbf{h}) \geq \gamma_{iu}, \ \forall\, \mathcal{C}_i \in \mathcal{N}_c, u \in \mathcal{U}_i. \tag{6}$$

Due to the quadratic form of the beamforming vector $\mathbf{w}_{ib,iu}$ in $\text{SINR}_{iu}(\mathbf{h})$, (JT-DL) is a non-convex optimization problem. Besides, the beamforming vectors of different UEs are coupled, making it challenging to solve (JT-DL) directly. In

TABLE I
DUAL RELATIONSHIP

Variable	Downlink		Uplink
Channel vector	$\mathbf{h}_{ib,iu}^H$	$\Leftrightarrow$	$\mathbf{h}_{ib,iu}$
Beamforming vector	$\mathbf{w}_{ib,iu}$	$\Leftrightarrow$	$\hat{\mathbf{w}}_{ib,iu}$
Transmit Power	$\|\mathbf{w}_{ib,iu}\|^2$	$\Leftrightarrow$	$\lambda_{iu}\sigma^2$

this work, we employ the notion of *uplink-downlink duality* to solve (JT-DL) efficiently.

B. JT Beamforming with Perfect CSI

The uplink-downlink duality refers to the equivalence of the uplink beamforming and the downlink beamforming. It implies the minimum transmit power at the BS required to achieve the DL SINR requirement is identical to that for the UL. Based on this principle, one can transform the downlink beamforming design problem to the uplink, which is computationally easier to solve. This idea has been adopted in [4] for the non-JT case. This work extends the UL-DL duality to design the downlink beamforming for JT. Table I summarizes the UL-DL duality relation in our work.

The following theorem transforms (JT-DL) into the uplink counterpart.

Theorem 1. *The Lagrangian dual of (JT-DL) is the following uplink problem.*

$$(\textit{JT-UL}) \quad \min_{\{\lambda_{iu}\}} \sum_{i=1}^{N_c} \sum_{u\in\mathcal{U}_i} \lambda_{iu}\sigma^2 \quad \textit{s.t.} \Lambda_{iu} \geq \gamma_{iu} \tag{7}$$

where λ_{iu} is the dual variable of the u-th UE in the i-th cell and Λ_{iu} is the UL SINR given by

$$\Lambda_{iu} = \frac{\lambda_{iu}\sum_{b\in\mathcal{C}_i}|\hat{\mathbf{w}}_{ib,iu}^H\mathbf{h}_{ib,iu}|^2}{I_{\text{Intra}}^{\text{U}} + I_{\text{Inter}}^{\text{U}} + \sum_{b\in\mathcal{C}_i}\|\hat{\mathbf{w}}_{ib,iu}^H\|^2} \tag{8}$$

where $\hat{\mathbf{w}}_{ib,iu}$ is the UL receive beamforming vector given by

$$\hat{\mathbf{w}}_{ib,iu} = \left(\sum_{\mathcal{C}_j\in\mathcal{N}_c}\sum_{v\in\mathcal{U}_j}\lambda_{jv}\sigma^2\mathbf{h}_{ib,jv}\mathbf{h}_{ib,jv}^H + \sigma^2\mathbf{I}_{N_t}\right)^{-1}\mathbf{h}_{ib,iu}. \tag{9}$$

In addition, $I_{\text{Intra}}^{\text{U}}$ and $I_{\text{Inter}}^{\text{U}}$ are given by

$$I_{\text{Intra}}^{\text{U}} = \sum_{b\in\mathcal{C}_i}\sum_{v\in\mathcal{U}_i\setminus u}\lambda_{iv}|\hat{\mathbf{w}}_{ib,iu}^H\mathbf{h}_{ib,iv}|^2, \tag{10a}$$

$$I_{\text{Inter}}^{\text{U}} = \sum_{\mathcal{C}_j\in\mathcal{N}_c\setminus\mathcal{C}_i}\sum_{b'\in\mathcal{C}_j}\sum_{v'\in\mathcal{U}_j}\lambda_{jv'}\sigma^2|\hat{\mathbf{w}}_{ib,iu}^H\mathbf{h}_{ib,jv'}|^2. \tag{10b}$$

Proof. The proof is based on the fact that the non-convex SINR constraints in (JT-DL) can be transformed into the second-order cone constraints following [12]. Accordingly, the duality gap is zero, and strong duality holds for (JT-DL). This suggests that (JT-DL) can be solved by its Lagrangian dual problem. To this end, we first define the Lagrangian of (JT-DL) as given by

$$\begin{aligned} L(\mathbf{w}_{ib,iu}\lambda_{iu}) = &\sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_i}\sum_{u\in\mathcal{U}_i}\mathbf{w}_{ib,iu}^H\mathbf{w}_{ib,iu} - \sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{u\in\mathcal{U}_i}\lambda_{iu} \\ &\Big[\gamma_{iu}^{-1}\sum_{b\in\mathcal{C}_i}|\mathbf{h}_{ib,iu}^H\mathbf{w}_{ib,iu}|^2 - \sum_{b\in\mathcal{C}_i}\sum_{v\in\mathcal{U}_i\setminus u}|\mathbf{h}_{ib,iu}^H\mathbf{w}_{ib,iv}|^2 \\ &- \sum_{\mathcal{C}_j\in\mathcal{N}_c\setminus\mathcal{C}_i}\sum_{b'\in\mathcal{C}_j}\sum_{v'\in\mathcal{U}_j}|\mathbf{h}_{jb',iu}^H\mathbf{w}_{jb',jv'}|^2 - \sigma^2\Big] \end{aligned} \tag{11}$$

where $\lambda_{iu} > 0$ is the Lagrange multiplier associated with the SINR constraint of UE_{iu}. The dual function is

$$g(\lambda_{iu}) = \min_{\mathbf{w}_{ib,iu}} L(\mathbf{w}_{ib,iu}), \forall\ i, b, u. \tag{12}$$

Then the Lagrangian dual of (JT-DL), namely, the maximum of $g(\lambda_{iu})$ is

$$\begin{aligned} (\text{JT-DL Dual}) \quad &\max_{\lambda_{iu}} \sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{u\in\mathcal{U}_i}\lambda_{iu}\sigma^2 \\ &\text{s.t}\sum_{b\in\mathcal{C}_i}\mathbf{\Theta}_{ib} \succeq (1+\gamma_{iu}^{-1})\lambda_{iu}\mathbf{h}_{ib,iu}\mathbf{h}_{ib,iu}^H, \end{aligned} \tag{13}$$

where

$$\mathbf{\Theta}_{ib} = \sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{u\in\mathcal{U}_i}\lambda_{iu}\mathbf{h}_{ib,iu} + \mathbf{I}_{N_t} \tag{14}$$

To show (JT-DL Dual) is equivalent to (JT-DL), we first find the optimal receive beamforming vector $\mathbf{w}_{ib,iu}$ that maximizes Λ_{iu}, which is the minimum-mean-square-error (MMSE) receive beamforming as given by (9). By substituting (9) into (7), the uplink SINR constraint (7) is identical to (13). Hence, (Dual-JT-DL) is the same as (JT-UL) except that the minimization is replaced by maximization, and the inequality constraints are reversed. In addition, both problems have the optimal solution when the SINR constraints are satisfied with equality. Therefore, (JT-DL Dual) and (JT-UL) have the same solution. □

The optimal receive beamforming in (9) depends on the dual variable λ_{iu}, which must satisfy the first-order necessary condition of the optimality. To find λ_{iu}, we first combine all the terms related to $\mathbf{w}_{ib,iu}$ in (11), which yields

$$\begin{aligned} L(\mathbf{w}_{ib,iu},\lambda_{iu}) = &\sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{u\in\mathcal{U}_i}\lambda_{iu}\sigma^2 + \sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_i}\sum_{u\in\mathcal{U}_i}\mathbf{w}_{ib,iu}^H \\ &\left[\mathbf{I}_{N_t} - (1+\gamma_{iu}^{-1})\lambda_{iu}\mathbf{h}_{ib,iu}\mathbf{h}_{ib,iu}^H + \sum_{\mathcal{C}_j\in\mathcal{N}_c}\sum_{v\in\mathcal{U}_j}\lambda_{jv}\mathbf{h}_{ib,jv}\mathbf{h}_{ib,jv}^H\right] \\ &\times \mathbf{w}_{ib,iu}. \end{aligned} \tag{15}$$

Taking the gradient of $L(\mathbf{w}_{ib,iu},\lambda_{iu})$ with respective to $\mathbf{w}_{ib,iu}^H$ and setting the result to zero, we obtain the necessary condition for optimal λ_{iu} as

$$\lambda_{iu} = \frac{1}{\left(1+\gamma_{iu}^{-1}\right)\sum_{b\in\mathcal{C}_i}\mathbf{h}_{ib,iu}^H\mathbf{\Theta}_{ib}^{-1}\mathbf{h}_{ib,iu}}, \tag{16}$$

which is a function of $\mathbf{\Theta}_{ib}$ given in (14) that depends on λ_{iu}.

This suggests that one can iteratively optimize λ_{iu} and $\hat{\mathbf{w}}_{ib,iu}$ until convergence.

Based on the obtained receive beamformer $\hat{\mathbf{w}}_{ib,iu}$ and the UL-DL duality, the optimal DL JT beamforming can be derived from the following theorem.

Theorem 2. *The optimal DL JT beamforming $\mathbf{w}_{ib,iu}$ is a scaled version of the UL JT beamforming $\hat{\mathbf{w}}_{ib,iu}$ as given by*

$$\mathbf{w}_{ib,ij} = \sqrt{\delta_{ib,u}}\hat{\mathbf{w}}_{ib,iu}, \tag{17}$$

where $\delta_{ib,u}$ is an entry of the $N_b|\mathcal{U}_i| \times 1$ vector $\boldsymbol{\delta}_i = [\delta_{ib,u}]^T_{b\in\mathcal{C}_i,u\in\mathcal{U}_i}$ for $i = 1,\ldots,N_c$. Define the collection of $\boldsymbol{\delta}_i$ as a vector $\boldsymbol{\Delta}$ of dimension $N_cN_b|\mathcal{U}_i| \times 1$, i.e., $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_{N_c}]^T$, which can be obtained as

$$\boldsymbol{\Delta} = \mathbf{F}^{-1}\mathbf{1}\sigma^2, \tag{18}$$

where $\mathbf{1}$ is an all-one vector of dimension $N_c|\mathcal{U}_i| \times 1$ and $\mathbf{F}$ is an $N_c|\mathcal{U}_i| \times N_cN_b|\mathcal{U}_j|$ matrix as given by

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{1,11} & \ldots & \mathbf{F}_{1,1|\mathcal{C}_1|} & \mathbf{F}_{1,21} & \ldots & \mathbf{F}_{1,N_c|\mathcal{C}_1|} \\ \mathbf{F}_{2,11} & \ldots & \mathbf{F}_{2,1|\mathcal{C}_2|} & \mathbf{F}_{2,21} & \ldots & \mathbf{F}_{2,N_c|\mathcal{C}_2|} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{F}_{N_c,11} & \ldots & \mathbf{F}_{N_c,1|\mathcal{C}_{N_c}|} & \mathbf{F}_{N_c,21} & \ldots & \mathbf{F}_{N_c,N_c|\mathcal{C}_{N_c}|} \end{bmatrix} \tag{19}$$

where the (u,v)-th entry of each $|\mathcal{U}_i| \times |\mathcal{U}_j|$ sub-matrix $\mathbf{F}_{i,jb}$ is defined as

$$\mathbf{F}_{i,jb}(u,v) = \begin{cases} \gamma_{iu}^{-1}|\hat{\mathbf{h}}_{ib,iu}^H\hat{\mathbf{w}}_{ib,iu}|^2, & \text{if } j = i \text{ and } v = u, \\ -|\tilde{\mathbf{h}}_{ib,iu}^H\hat{\mathbf{w}}_{ib,iv}|^2, & \text{if } j = i \text{ and } v \neq u, \\ -|\tilde{\mathbf{h}}_{mb,iu}^H\hat{\mathbf{w}}_{jb,jv}|^2, & \text{if } j \neq i. \end{cases} \tag{20}$$

Proof. The proof can be found in [13]. □

A complete algorithm for iteratively obtaining the optimal DL JT beamforming vectors $\mathbf{w}_{ib,iu}$ is summarized in Algorithm 1, which is referred to as non-robust JT beamforming (NR-JTBF) in this work. The convergence of Algorithm 1 to the global optimal solution is guaranteed following the similar arguments in [4], which is omitted here due to space limit.

Algorithm 1 Non-Robust JT Beamforming (NR-JTBF)

Input: CSI $\mathbf{h}_{ib,iu}$, $\forall\ \mathcal{C}_i \in \mathcal{N}_c$, $b \in \mathcal{C}_i$, $u \in \mathcal{U}_i$.

1: Initialize $\lambda_{iu} = 1$, $\forall\ i, u$.
2: **repeat**
3: Compute $\boldsymbol{\Theta}_{ib}$ using (14).
4: Compute λ_{iu} using (16).
5: Compute $\hat{\mathbf{w}}_{ib,iu}$ using (9).
6: Compute the scaling factor $\delta_{i,bu}$ from (18).
7: Update $\mathbf{w}_{ib,iu}$ using (17).
8: **until** the predefined stopping criterion is met.

Output:

9: Optimal downlink JT beamformers $\mathbf{w}_{ib,iu}, \forall\ i, b, u$.

C. JT Beamforming with Imperfect CSI

Algorithm 1 uses the accurate CSI as the input. In practice, CSI acquired by each BS contains unknown estimation errors and thus is imperfect. Consequently, the beamforming vectors derived from imperfect CSI may not achieve the required SINR. This motivates us to revisit (JT-DL) and design robust beamforming for JT that minimizes the total transmit power while ensuring the target SINR with a certain probability. The resultant problem formulation based on estimated CSI is given below.

$$\begin{aligned} \text{(Robust JT-DL)} \quad & \min_{\mathbf{w}_{ib,iu}} \sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_i}\sum_{u\in\mathcal{U}_i} \|\mathbf{w}_{ib,iu}\|^2 \\ & \text{s.t } \Pr[\text{SINR}_{iu}(\tilde{\mathbf{h}}) \geq \gamma_{iu}] \geq 1 - \rho_{iu} \end{aligned} \tag{21}$$

where $\rho_{iu} \in (0,1)$. By imposing constraint (21), UE_{iu} is guaranteed to satisfy the target SINR with the probability no less than $1 - \rho_{iu}$. (Robust JT-DL) is non-convex since the beamforming vectors $\mathbf{w}_{ib,iu}$ are of quadratic forms and constraint (21) is a probability function without closed-form. Unfortunately, UL-DL duality can only handle deterministic constraints. To remedy this, we use the SDR technique to convert (Robust JT-DL) into a convex problem.

Firstly, we linearize the objective function of (Robust JT-DL) by introducing the rank-one positive semidefinite (PSD) matrix $\mathbf{W}_{ib,iu} = \mathbf{w}_{ib,iu}\mathbf{w}_{ib,iu}^H$. As a result, the quadratic term $\|\mathbf{w}_{ib,iu}\|^2$ is replaced by $\text{Tr}(\mathbf{W}_{ib,iu})$, which is a linear operation. Meanwhile, constraint (21) can be rewritten as

$$\begin{aligned} &\Pr\Bigg[\gamma_{iu}^{-1}\sum_{b\in\mathcal{C}_i}(\tilde{\mathbf{h}}_{ib,iu}+\mathbf{e}_{ib,iu})^H\mathbf{W}_{ib,iu}(\tilde{\mathbf{h}}_{ib,iu}+\mathbf{e}_{ib,iu}) \\ &-\sum_{b\in\mathcal{C}_i}\sum_{v\in\mathcal{U}_i\setminus u}(\tilde{\mathbf{h}}_{ib,iu}+\mathbf{e}_{ib,iu})^H\mathbf{W}_{ib,iv}(\tilde{\mathbf{h}}_{ib,iu}+\mathbf{e}_{ib,iu}) \\ &-\sum_{\mathcal{C}_j\in\mathcal{N}_c\setminus\mathcal{C}_i}\sum_{b'\in\mathcal{C}_j}\sum_{u'\in\mathcal{U}_j}(\tilde{\mathbf{h}}_{jb',iu}+\mathbf{e}_{jb',iu})^H\mathbf{W}_{jb',just} \\ &(\tilde{\mathbf{h}}_{jb',iu}+\mathbf{e}_{jb',iu}) \geq \sigma^2\Bigg] \geq 1-\rho_{iu}, \end{aligned} \tag{22}$$

which does not have a closed form and thus is difficult to handle. To simplify (22) into a more tractable form, we substitute (2) into (22) and define the following notations.

$$\begin{aligned} \boldsymbol{\Omega}_{ib,iu} &\triangleq \mathbf{Q}_{ib,iu}^{1/2}(\gamma_{iu}^{-1}\mathbf{W}_{ib,iu} - \sum_{v\in\mathcal{U}_i\setminus u}\mathbf{W}_{ib,iv})\mathbf{Q}_{ib,iu}^{1/2}, \\ \boldsymbol{\Omega}_{jb',iu} &\triangleq -\mathbf{Q}_{jb',iu}^{1/2}\sum_{u'\in\mathcal{U}_j}\mathbf{W}_{jb',ju'}\mathbf{Q}_{jb',iu}^{1/2}, \\ \boldsymbol{\omega}_{ib,iu} &\triangleq \mathbf{Q}_{ib,iu}^{1/2}(\gamma_{iu}^{-1}\mathbf{W}_{ib,iu} - \sum_{v\in\mathcal{U}_i\setminus u}\mathbf{W}_{ib,iv})\tilde{\mathbf{h}}_{ib,iu}, \\ \boldsymbol{\omega}_{jb',iu} &\triangleq -\mathbf{W}_{jb',iu}^{1/2}\sum_{u'\in\mathcal{U}_j}\mathbf{W}_{jb',ju'}\tilde{\mathbf{h}}_{jb',iu}, \\ d_{ib,iu} &\triangleq \tilde{\mathbf{h}}_{ib,iu}^H(\gamma_{iu}^{-1}\mathbf{W}_{ib,iu} - \sum_{v\in\mathcal{U}_i}\mathbf{W}_{ib,iv})\tilde{\mathbf{h}}_{ib,iv}, \\ d_{jb',iu} &\triangleq -\tilde{\mathbf{h}}_{jb',iu}^H\sum_{u'\in\mathcal{U}_j}\mathbf{W}_{jb',ju'})\tilde{\mathbf{h}}_{jb',iu}. \end{aligned} \tag{23}$$

Then (22) can be expressed as

$$\Pr\Bigg[\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\Big(\mathbf{v}_{lb,iu}^H\boldsymbol{\Omega}_{lb,iu}\mathbf{v}_{lb,iu}+\boldsymbol{\omega}_{lb,iu}^H\mathbf{v}_{lb,iu} + \mathbf{v}_{lb,iu}^H\boldsymbol{\omega}_{lb,iu}+d_{lb,iu}\Big)\geq\sigma^2\Bigg]\geq 1-\rho_{iu},\ \forall\,\mathcal{C}_i\in\mathcal{N}_c, u\in\mathcal{C}_i. \tag{24}$$

Note that $\boldsymbol{\omega}_{lb,iu}^H\mathbf{v}_{lb,iu}+\mathbf{v}_{l\beta,iu}^H\boldsymbol{\omega}_{lb,iu}=2\mathrm{Re}(\mathbf{v}_{lb,iu}^H\boldsymbol{\omega}_{lb,iu})$. Hence, (24) can be represented as

$$\Pr\Bigg[\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\Big(\mathbf{v}_{lb,iu}^H\boldsymbol{\Omega}_{lb,iu}\mathbf{v}_{lb,iu}+2\mathrm{Re}(\mathbf{v}_{lb,iu}^H\boldsymbol{\omega}_{lb,iu}\Big) \geq\sigma^2-\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{\beta\in\mathcal{C}_l}d_{lb,iu}\Bigg]\geq 1-\rho_{iu},\ \forall\,\mathcal{C}_i\in\mathcal{N}_c, u\in\mathcal{C}_i. \tag{25}$$

Let $K=N_cN_b$. Then we can rewrite (25) as

$$\Pr\Bigg[\sum_{k=1}^{K}\Big(\hat{\mathbf{v}}_{k,iu}^H\hat{\boldsymbol{\Omega}}_{k,iu}\hat{\mathbf{v}}_{k,iu}+2\mathrm{Re}(\hat{\mathbf{v}}_{k,iu}^H\hat{\boldsymbol{\omega}}_{k,iu}\Big) \geq\sigma^2-\sum_{k=1}^{K}\hat{d}_{k,iu}\Bigg]\geq 1-\rho_{iu}. \tag{26}$$

where $\hat{\mathbf{v}}_{k,iu}=\mathbf{v}_{lb,iu}$ for $k=(l-1)N_b+b$. Likewise, $\hat{\boldsymbol{\Omega}}_{k,iu}=\boldsymbol{\Omega}_{lb,iu}\in\mathbb{H}^{N_t}$, $\hat{\boldsymbol{\omega}}_{k,iu}=\boldsymbol{\omega}_{lb,iu}\in\mathbb{C}^{N_t}$ and $\hat{d}_{k,iu}=d_{lb,iu}$. Here $\mathbb{H}^{N_t}$ represents the N_t-dimensional Hermitian matrix. Although (26) is much simpler than (22), it remains a probabilistic inequality without a closed form, and thus a certain approximation is necessary. Since (26) contains the quadratic form of $\hat{\mathbf{v}}_{k,iu}^H$, we are motivated to apply the Bernstein-type inequality, which bounds the probability of the quadratic form of complex random variables deviating from its mean. To this end, define

$$g_{k,iu}=\hat{\mathbf{v}}_{k,iu}^H\hat{\boldsymbol{\Omega}}_{k,iu}\hat{\mathbf{v}}_{k,iu}+2\mathrm{Re}\left(\hat{\boldsymbol{\omega}}_{k,iu}^H\hat{\mathbf{v}}_{k,iu}\right), \tag{27}$$

$\forall\ k=1,\ldots,K$, $\mathcal{C}_i\in\mathcal{N}_c$, and $u\in\mathcal{U}_i$. Then the following Lemma is introduced.

Lemma 1. *For any $\tau_{iu}>0$,*

$$\Pr\Bigg[\sum_{k=1}^{K}g_{k,iu}\geq\sum_{k=1}^{K}\mathrm{Tr}(\hat{\boldsymbol{\Omega}}_{k,iu})-\sqrt{2\tau_{iu}}\Big(\sum_{k=1}^{K}\big(\|\hat{\boldsymbol{\Omega}}_{k,iu}\|^2 +2\|\hat{\boldsymbol{\omega}}_{k,iu}\|^2\big)\Big)^{1/2}-\tau_{iu}\lambda^-(\hat{\boldsymbol{\Omega}})\Bigg]\geq 1-e^{-\tau_{iu}}. \tag{28}$$

where $\lambda^-(\hat{\boldsymbol{\Omega}})=\max\{\lambda_{\max}(-\hat{\boldsymbol{\Omega}}_{1,iu}),\ldots,\lambda_{\max}(-\hat{\boldsymbol{\Omega}}_{K,iu}),0\}$ with $\lambda_{\max}(-\hat{\boldsymbol{\Omega}}_{k,iu})$ denoting the maximum eigenvalue of the matrix $(-\hat{\boldsymbol{\Omega}}_{k,iu})$.

Proof. The proof is motivated from [14], which considers robust hybrid beamforming without multi-cell cooperation. The idea is to find a conservative convex approximation to the probabilistic constraint (26) with the aid of the Bernstein-type inequality [15]. For our problem, a Bernstein-type inequality can be established as follows.

Let $S=K\cdot 2N_t$ for $K=N_cN_b$, $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_S]^T\in\mathbb{R}^S$ and $\boldsymbol{\beta}=[\beta_1,\ldots,\beta_S]^T\in\mathbb{R}^S$, where $\mathbb{R}^S$ denotes the S-dimensional real vector. Besides, $\boldsymbol{\nu}=[\nu_1,\ldots,\nu_S]^T\in\mathbb{R}^S$ be a random vector where $\nu_1,\ldots,\nu_S$ are i.i.d. real random variables following the standard normal distribution $\mathcal{N}(0,1)$. Define

$$G=\sum_{s=1}^{S}\Big(\alpha_s\nu_s^2+2\beta_s\nu_s\Big). \tag{29}$$

Then, given $\tau>0$, the following result holds.

$$\Pr\left[G\geq\sum_{s=1}^{S}\alpha_s-2\sqrt{\tau}\sqrt{\sum_{s=1}^{S}\left(\alpha_s{}^2+2\beta_s{}^2\right)}-2\tau\alpha^-\right] \geq 1-e^{-\tau}, \tag{30}$$

where $\alpha^-\triangleq\max\{\max\{-\alpha_s\mid 1\leq s\leq S\},0\}$. (30) is a real-valued constraint but can be applied to handle complex-valued variables. This is achieved with the aid of the following variables.

$$\bar{\boldsymbol{\Omega}}_{k,iu}=\frac{1}{2}\begin{bmatrix}\mathrm{Re}(\hat{\boldsymbol{\Omega}}_{k,iu}) & -\mathrm{Im}(\hat{\boldsymbol{\Omega}}_{k,iu})\\ \mathrm{Im}(\hat{\boldsymbol{\Omega}}_{k,iu}) & \mathrm{Re}(\hat{\boldsymbol{\Omega}}_{k,iu})\end{bmatrix}\in\mathbb{S}^{2N_t}, \tag{31a}$$

$$\bar{\boldsymbol{\omega}}_{k,iu}=\frac{1}{\sqrt{2}}\begin{bmatrix}\mathrm{Re}(\hat{\boldsymbol{\omega}}_{k,iu})\\ \mathrm{Im}(\hat{\boldsymbol{\omega}}_{k,iu})\end{bmatrix}\in\mathbb{R}^{2N_t}, \tag{31b}$$

$$\bar{\mathbf{v}}_{k,iu}=\sqrt{2}\begin{bmatrix}\mathrm{Re}(\mathbf{v}_{k,iu})\\ \mathrm{Im}(\mathbf{v}_{k,iu})\end{bmatrix}\sim\mathcal{CN}\left(\mathbf{0},\mathbf{I}_{2N_t}\right), \tag{31c}$$

for all $k\in\{1,\ldots,K\},\mathcal{C}_i\in\mathcal{N}_c$, and $u\in\mathcal{U}_i$. Denote $\bar{v}_{k,iu}[l]$ and $\bar{\omega}_{k,iu}[l]$ the lth element of $\bar{\mathbf{v}}_{k,iu}$ and $\bar{\boldsymbol{\omega}}_{k,iu}$, respectively. Here, $\mathbb{S}$ denotes the real symmetric matrix. Using (31a)-(31c), and decomposing $\bar{\boldsymbol{\Omega}}_{k,iu}$ as $\bar{\boldsymbol{\Omega}}_{k,iu}=\mathbf{U}_{k,iu}\boldsymbol{\Lambda}_{k,iu}\mathbf{U}_{k,iu}^T$ through eigenvalue decomposition, where $\mathbf{U}_{k,iu}$ is an orthogonal matrix and $\boldsymbol{\Lambda}_{k,iu}=\mathrm{diag}(\lambda_{k,iu}[l])$ is a diagonal matrix with $\lambda_{k,iu}[l]$ denoting the l-th eigenvalue of $\bar{\boldsymbol{\Omega}}_{k,iu}$, (27) can be rewritten as

$$\begin{aligned}g_{k,iu}&=\bar{\mathbf{v}}_{k,iu}^T\bar{\boldsymbol{\Omega}}_{k,iu}\bar{\mathbf{v}}_{k,iu}+2\bar{\boldsymbol{\omega}}_{k,iu}^T\bar{\mathbf{v}}_{k,iu}\\ &\overset{\text{(i)}}{=}\tilde{\mathbf{v}}_{k,iu}^T\boldsymbol{\Lambda}_{l,iu}\tilde{\mathbf{v}}_{k,iu}^T+2\tilde{\boldsymbol{\omega}}_{k,iu}^T\tilde{\mathbf{v}}_{k,iu}\\ &\overset{\text{(ii)}}{=}\sum_{l=1}^{2N_t}\Big(\lambda_{k,iu}[l]\tilde{v}_{k,iu}^2[l]+2\tilde{\omega}_{k,iu}[l]\tilde{v}_{k,iu}[l]\Big),\end{aligned} \tag{32}$$

where (i) is obtained by defining $\tilde{\mathbf{v}}_{k,iu}=\mathbf{U}_{k,iu}^T\bar{\mathbf{v}}_{k,iu}$ and $\tilde{\boldsymbol{\omega}}_{k,iu}=\mathbf{U}_{k,iu}^T\bar{\boldsymbol{\omega}}_{k,iu}$, and in (ii), $\tilde{v}_{k,iu}[l]$ and $\tilde{\omega}_{k,iu}[l]$ is the l-th element of $\tilde{\mathbf{v}}_{k,iu}$ and $\tilde{\boldsymbol{\omega}}_{k,iu}$, respectively. One can see that (32) follows the same form as (29). Similar to (29), denote G' the sum of $g_{k,iu},\ \forall\ k=1,\ldots,K$, i.e.,

$$\begin{aligned}G'&=\sum_{k=1}^{K}g_{k,iu}\\ &=\sum_{k=1}^{K}\sum_{l=1}^{2N_t}\Big(\lambda_{k,iu}[l]\tilde{v}_{k,iu}^2[l]+2\tilde{\omega}_{k,iu}[l]\tilde{v}_{k,iu}[l]\Big).\end{aligned} \tag{33}$$

According to (30), we obtain the following inequality.

$$\Pr\Big[G' \geq \sum_{k=1}^{K}\sum_{l=1}^{2N_t}\lambda_{k,iu}[l] - 2\sqrt{\tau_{iu}} \sqrt{\sum_{k=1}^{K}\sum_{l=1}^{2N_t}\lambda_{k,iu}^2[l] + 2\tilde{\omega}_{k,iu}^2[l]} - 2\tau_{iu}\tilde{\alpha}^-\Big] \geq 1 - e^{-\tau_{iu}}, \quad (34)$$

where

$$\tilde{\alpha}^- \triangleq \max\{\max_l\{-\lambda_{1,iu}[l]\},\ldots,\max\{\max_l\{-\lambda_{K,iu}[l],0\},$$

for all $\mathcal{C}_i \in \mathcal{N}_c$, $u \in \mathcal{U}_i$. (34) can be written more compactly by using the property of matrix trace, e.g., $\sum_{l=1}^{2N_t}\lambda_{k,iu}[l] = \mathrm{Tr}(\bar{\mathbf{\Omega}}_{k,iu})$ and also, $\bar{\mathbf{\Omega}}_{k,iu}$ has the same eigenvalues as $\hat{\mathbf{\Omega}}_{k,iu}$. Hence, (34) can be rewritten as

$$\Pr\Big[G' \geq \sum_{k=1}^{K}\mathrm{Tr}(\hat{\mathbf{\Omega}}_{k,iu}) - 2\sqrt{\tau_{iu}} \sqrt{\sum_{k=1}^{K}\frac{1}{2}\|\hat{\mathbf{\Omega}}_{k,iu}\|^2 + \|\hat{\boldsymbol{\omega}}_{k,iu}\|^2} - \tau_{iu}\lambda^-(\hat{\mathbf{\Omega}})\Big] \geq 1 - e^{-\tau_{iu}}, \quad (35)$$

which is obtained by noting that

$$\sum_{l=1}^{2N_t}\lambda_{k,iu}^2[l] = \|\bar{\mathbf{\Omega}}_{k,iu}\|^2 = \frac{1}{2}\|\hat{\mathbf{\Omega}}_{k,iu}\|^2, \quad (36a)$$

$$2\sum_{l=1}^{2N_t}\tilde{\omega}_{k,iu}^2[l] = 2\|\tilde{\boldsymbol{\omega}}_{k,iu}\|^2 = 2\|\bar{\boldsymbol{\omega}}_{k,iu}\|^2 = \|\hat{\boldsymbol{\omega}}_{k,iu}\|^2 \quad (36b)$$

$$\begin{aligned}\tilde{\alpha}^- &= \max\{\lambda_{\max}(-\bar{\mathbf{\Omega}}_{1,iu}),\cdots,\lambda_{\max}(-\bar{\mathbf{\Omega}}_{K,iu}),0\}\\ &= \frac{1}{2}\max\{\lambda_{\max}(-\hat{\mathbf{\Omega}}_{1,iu}),\cdots,\lambda_{\max}(-\hat{\mathbf{\Omega}}_{K,iu}),0\}\\ &\triangleq \frac{1}{2}\lambda^-(\hat{\mathbf{\Omega}}).\end{aligned} \quad (36c)$$

It is easy to check that (35) is identical to (28) that completes the proof. □

By letting $\rho_{iu} = e^{-\tau_{iu}}$ and comparing (28) with (26), we can see that the following inequality serves as a sufficient condition to satisfy (26).

$$\sum_{k=1}^{K}\mathrm{Tr}(\hat{\mathbf{\Omega}}_{k,iu}) - \sqrt{2\tau_{iu}}\Big(\sum_{k=1}^{K}\big(\|\hat{\mathbf{\Omega}}_{k,iu}\|^2 + 2\|\hat{\boldsymbol{\omega}}_{k,iu}\|^2\big)\Big)^{1/2} - \tau_{iu}\lambda^-(\hat{\mathbf{\Omega}}) \geq \sigma^2 - \sum_{k=1}^{K}\hat{d}_{k,iu}, \quad \forall\, \mathcal{C}_i \in \mathcal{N}_c, u \in \mathcal{U}_i. \quad (37)$$

Recall $K = N_cN_b$ and $k = (l-1)N_b+b$, (37) can be rewritten as

$$\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\Big(\mathrm{Tr}(\mathbf{\Omega}_{lb,iu}) + d_{lb,iu}\Big) - \tau_{iu}\lambda^-(\mathbf{\Omega}) - \sigma^2 \geq \sqrt{2\tau_{iu}}\Big(\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\|\mathbf{C}_{lb,iu}\|^2\Big)^{1/2}. \quad (38)$$

where $\mathbf{C}_{lb,iu} = [\mathbf{\Omega}_{lb,iu}\ \sqrt{2}\boldsymbol{\omega}_{lb,iu}]$. Since $\lambda^-(\mathbf{\Omega})$ in (38) is non-convex, we introduce the slack variable $x_{iu} \geq 0$ such that (38) is equivalent to

$$\sqrt{\sum_{l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\|\mathbf{C}_{lb,iu}\|^2} \leq \frac{1}{\sqrt{2\tau_{iu}}}\Big[\sum_{l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\Big(\mathrm{Tr}(\mathbf{\Omega}) + d_{l\beta,iu}\Big) - \tau_{iu}x_{iu} - \sigma^2\Big],$$

along with $x_{iu}\mathbf{I}_{N_t} + \mathbf{Q}_{lb,iu} \succeq \mathbf{0}$, $\forall\ \mathcal{C}_i,\mathcal{C}_l \in \mathcal{N}_c$, $b \in \mathcal{C}_l$, and $u \in \mathcal{U}_i$. As a result, we obtain a convex formulation for (Robust JT-DL) as follows.

$$\begin{aligned}(\text{Robust JT-DL}^2)\quad &\min_{\mathbf{W}_{ib,iu},x_{iu}\geq 0}\ \sum_{\mathcal{C}_i\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_i}\sum_{u\in\mathcal{U}_i}\mathrm{Tr}(\mathbf{W}_{ib,iu})\\ \text{s.t}\quad &\sqrt{\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\|\mathbf{C}_{lb,iu}\|^2} \leq \frac{1}{\sqrt{2\tau_{iu}}}\\ &\qquad\Big[\sum_{\mathcal{C}_l\in\mathcal{N}_c}\sum_{b\in\mathcal{C}_l}\Big(\mathrm{Tr}(\mathbf{\Omega}) + d_{lb,iu}\Big) - \tau_{iu}x_{iu} - \sigma^2\Big],\\ &x_{iu}\mathbf{I}_{N_t} + \mathbf{Q}_{lb,iu} \succeq \mathbf{0},\ \forall\ i,b,u.\end{aligned} \quad (39)$$

With the aid of SDR formulation and the Bernsteirn-type inequality, (Robust JT-DL^2) is convex that can be solved numerically. The obtained solution $\mathbf{W}_{ib,iu}$, which is referred to as robust-JT beamforming (R-JTBF), is not necessarily a rank-one matrix, in which case the Gaussian randomization technique can be used to find an approximated solution [9].

IV. Simulation Results

The performance of the proposed NR-JTBF and R-JTBF are evaluated by simulating the three-cell scenario in Fig. 1. In each cell, a UE is randomly distributed and 10,000 channel realizations are generated for each UE location. The channel vector between BS_{ib} and UE_{iu} is modeled as $\mathbf{h}_{ib,iu} = \sqrt{\beta_{ib,iu}}\psi_{ib,iu}\varphi_{ib,iu}(\tilde{\mathbf{h}}_{ib,iu} + \mathbf{e}_{ib,iu})$, where $\beta_{ib,iu}$ and $\psi_{ib,iu}$ capture the distance-dependent pathloss and lognormal shadowing, respectively, and $\varphi_{ib,iu}$ reflects the transmit-receive antenna gain. Besides, $\tilde{\mathbf{h}}_{ib,iu} \sim \mathcal{CN}(\mathbf{0},\mathbf{I}_{N_t})$ is the Rayleigh fading coefficient. Table IV lists the simulation parameters.

TABLE II
Simulation Parameters

Parameter	Value
Number of cells N_c	3
Number of BSs per cell N_b	3
Number of UEs per cell	1
Number of BS antennas N_t	4
Inter-cell distance	500 m
BS maximum power	24 dBm
Antenna gain $\varphi_{ib,iu}$	5 dBi
System bandwidth	10 MHz
Noise power spectral density	-174 dBm/Hz
Pathloss over a distance of d km	$145.4 + 37.5\log_{10} d$
CSI error radius ϵ	$\sqrt{0.1}$
SINR requirement γ_{iu}	12 dB
SINR violation probability ρ_{iu}	0.1

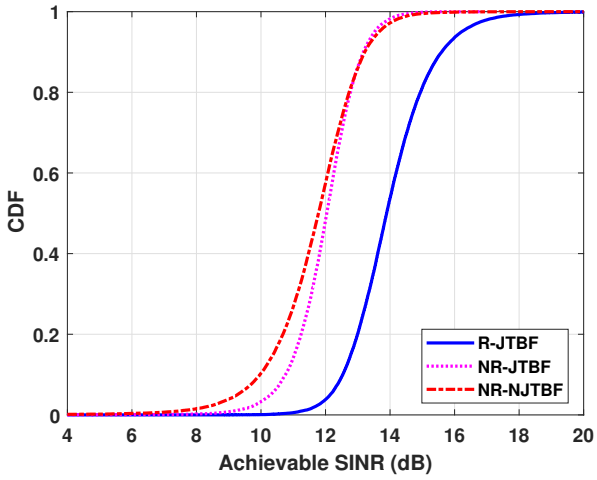


Fig. 2. The CDF of the achievable SINR.

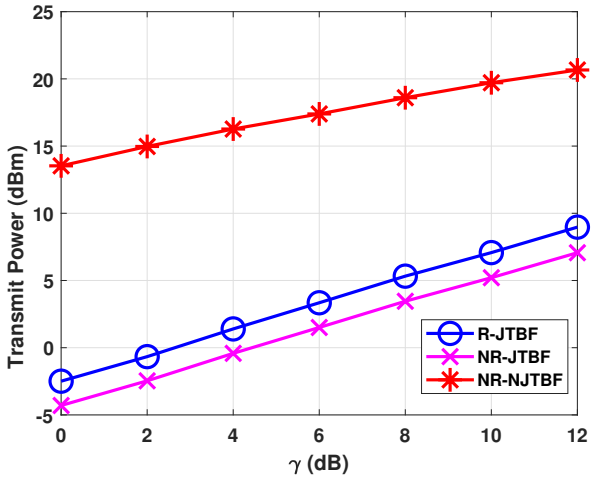


Fig. 3. Transmit power vs. SINR target γ.

We first show the distribution of the achievable SINR in terms of the cumulative distribution function (CDF) in Fig. 2. Here, the traditional multi-user MIMO beamforming based on imperfect CSI without JT referred to as non-robust non-JT beamforming (NR-NJTBF), is included for comparison. NR-NJTBF is a special case of the proposed NR-JTBF with one BS per cell, i.e., $N_b = 1$. One can see that R-JTBF significantly improves the achievable SINR over NR-JTBF and NR-NJTBF. The probability of violating the SINR target of 12 dB is only 0.04, which is much less than the prescribed probability of 0.1. As to NT-JTBF and NR-NJTBF, the likelihood of SINR target violation is 49% and 61%, respectively. The result justifies the importance of robust beamforming for practical systems with imperfect CSI.

Fig. 3 plots the transmit power used by different beamforming schemes for satisfying the SINR target γ, which is assumed to be identical for all UEs. As γ increases, a higher transmit power is required to meet the SINR target as expected. Given a specific SINR target, the proposed R-JTBF consumes extra power of 2 dBm compared with the proposed NR-JTBF. The additional power consumption of R-JTBF can be regarded as the cost for compensating the SINR loss due to imperfect CSI and the same price is maintained even for a more stringent SINR target. Comparing with NR-NJTBF, both the proposed R-JTBF and NR-JTBF dramatically reduce the required transmit power. For example, R-JTBF consumes 12 dBm less power than NR-NJTBF when $\gamma = 12$ dB. The power-saving gain attributes to JT, which uses the antennas across different BSs and, thus, it achieves a better beamforming gain than the beamforming scheme without JT.

We further examine the performance of R-JTBF under varied QoS requirements in terms of the designated SINR satisfaction probability $1 - \rho$. Here, the commonly considered multi-user beamforming, namely zero-forcing (ZF), is included in the comparison. The ZF beamforming vectors, as given by $\mathbf{w}_{ib,iu}^{\mathrm{ZF}} = \mathbf{h}_{ib,iu}(\mathbf{h}_{ib,iu}^H \mathbf{h}_{ib,iu})^{-1}$ for UE_{iu} served by BS_{ib}, aim to cancel all the interference, including intra-cell and inter-cell interference by jointly designing the beamforming using all the antennas in the network. This can be implemented at the CU, which collects global CSI. Fig. 4 shows the SINR satisfactory rate achieved by different schemes versus the designated SINR satisfactory probability. Except ZF without considering a specific QoS requirement, the same SINR target of 6 dB is applied to the rest schemes. In addition, the SINR satisfaction probability $(1-\rho)$ is varied from 0.1 to 0.9 for R-JTBF. The SINR satisfaction rate of R-JTBF consistently exceeds the designated satisfaction probability and is higher than the rest schemes. For the three non-robust beamforming schemes, the SINR satisfaction rate of ZF, NR-JTBF, and NR-NJTBF is about 63%, 51%, and 42%, respectively. It is worth pointing out that ZF outperforms NR-JTBF and NR-NJTBF because it utilizes all the available antennas in the network to determine the beamforming vector for each user, while the rest two use the antennas within the same cell only. Thus, the gain of ZF is the direct consequence of using N_c times more antennas than that of NR-JTBF and $N_c N_b$ times more antennas than that of NR-NJTBF with a single-BS operation.

Finally, we investigate how much spectral efficiency can be improved per unit of power. To this end, we define the power efficiency as the ratio between the spectral efficiency evaluated as $\sum_{\mathcal{C}_i \in \mathcal{N}_c} \sum_{u \in \mathcal{U}_i} \log_2(1 + \mathrm{SINR}_{iu})$ and the total transmit power. Fig. 5 illustrates the power efficiency as a function of the antenna per BS N_t. Increasing N_t leads to improved SINR due to a higher spatial DoF. Also, the required power to meet the SINR target is reduced by using a larger antenna array at the BS. As a result, each unit of energy can be used more efficiently, as seen from the increasing trend in Fig. 5. Both JT beamforming schemes, namely R-JTBF and NR-JTBF, have a much higher power efficiency than NR-NJTBF and ZF. The energy efficiency of R-JTBF is slightly worst than that of

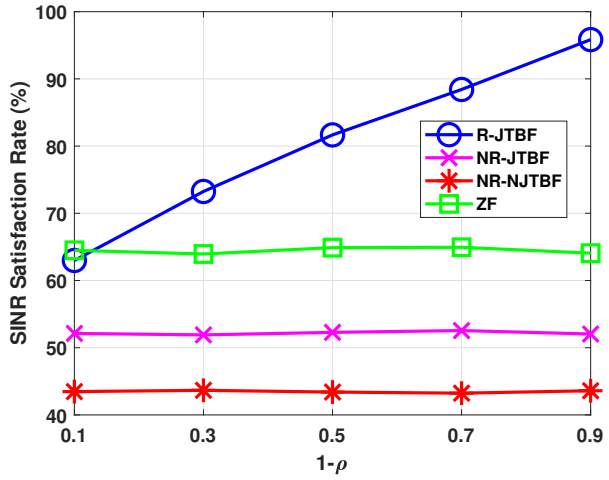


Fig. 4. SINR satisfaction rate vs. predefined SINR satisfaction probability.

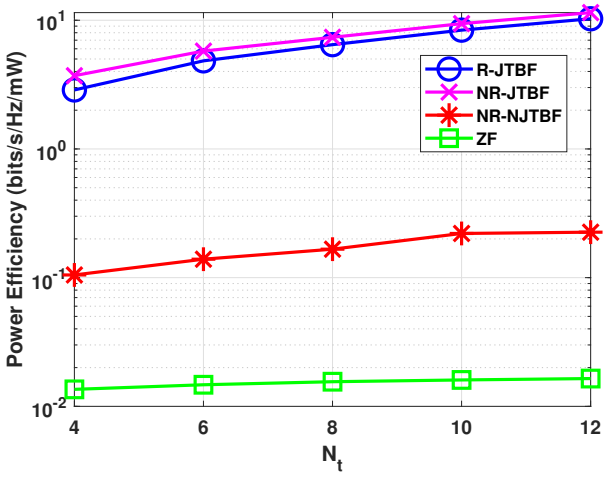


Fig. 5. Power efficiency vs. number of antennas per BS.

NR-JTBF because of a higher transmit power paid to achieve the desired SINR subject to CSI error as seen in Fig. 3. On the other hand, ZF has the worst energy efficiency because it uses all the available power at the BS. As a result, ZF is not energy efficient despite a certain spectral efficiency increase as N_t increases.

V. Conclusion

This work extends the previous work on the multi-cell coordinated beamforming design to the JT mode. It is shown that when perfect CSI is not available, the robust JT beamforming scheme significantly improves the achievable SINR, thereby increasing the SINR satisfaction rate over the non-robust JT beamforming scheme and ZF-based beamforming scheme. The improvement of the robust JT beamforming scheme comes at the cost of slightly increased transmit power compared to the non-robust JT beamforming scheme. Still, the cost is nearly negligible as the number of antennas per BS increases. On the other hand, the non-robust non-JT beamforming scheme incurs a dramatic increase in the total transmit power as the target SINR increases. Hence, its power efficiency is much lower than the two JT beamforming schemes. The results of this work demonstrate that signal enhancement and interference cancellation can be achieved simultaneously through multi-cell cooperation that yields promising gains even without perfect CSI.

Acknowledgment

This work was supported by the National Science and Technology Council, Taiwan, under Grant MOST 111-2221-E-007-076-MY3.

References

[1] J. Yang and D. K. Kim, "Multi-cell uplink-downlink beamforming throughput duality based on Lagrangian duality with per-base station power constraints," *IEEE Commun. Lett.*, vol. 12, no. 4, pp. 277–279, Apr. 2008.

[2] W. Yu, "Uplink–downlink duality via minimax duality," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361–374, Feb. 2006.

[3] F. Göttsch, N. Osawa, T. Ohseki, K. Yamazaki, and G. Caire, "Uplink-downlink duality and precoding strategies with partial CSI in cell-free wireless networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Austin, TX, Apr. 2022, pp. 614–619.

[4] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[5] K.-Y. Wang, T.-H. Chang, W.-K. Ma, A. M.-C. So, and C.-Y. Chi, "Probabilistic SINR constrained robust transmit beamforming: A Bernstein-type inequality based conservative approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 3080–3083.

[6] C. Shen, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed robust multicell coordinated beamforming with imperfect CSI: An ADMM approach," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2988–3003, Jun. 2012.

[7] C. Shen, K.-Y. Wang, T.-H. Chang, Z. Qiu, and C.-Y. Chi, "Worst-case SINR constrained robust coordinated beamforming for multicell wireless systems," in *Proc. IEEE International Conference on Communications (ICC)*, Kyoto, Japan, Jun. 2011, pp. 1–5.

[8] X. Gong, M. Jordan, G. Dartmann, and G. Ascheid, "Max-min beamforming for multicell downlink systems using long-term channel statistics," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Tokyo, Japan, Sep. 2009, pp. 803–807.

[9] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[10] 3GPP TS 38.214, "Physical layer procedures for data," 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR, Mar. 2023.

[11] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 1, pp. 146–172, 2019.

[12] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimizationfor fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

[13] C.-Y. Lin and K.-H. Liu, "QoS-aware downlink beamforming for joint transmission in multi-cell networks (with complementary materials)," arXiv:submit/4962970, 2023.

[14] G. Xu, C. H. Lin, W. Ma, S. Chen, and C. Y. Chi, "Outage constrained robust hybrid coordinated beamforming for massive MIMO enabled heterogeneous cellular networks," *IEEE Access*, vol. 5, pp. 13 601–13 616, 2017.

[15] I. Bechar, "Bernstein-type inequality for stochastic processes of quadratic forms of Gaussian variables," Apr. 2009. [Online]. Available: https://arxiv.org/abs/0909.3595

Deep-Learning-Based Lattice Reduction Preprocessing for Time-Correlated MIMO Systems

Yi-Mei Li, Jung-Chun Chi and Yuan-Hao Huang
Institute of Communications Engineering and Department of Electrical Engineering
National Tsing-Hua University, Hsinchu, Taiwan, R.O.C.
yhhuang@ee.nthu.edu.tw

***Abstract*—The lattice-reduction (LR) preprocessing can effectively improve the performance of multiple-input multiple-output (MIMO) systems especially in the highly correlated channel. This study designed a deep-learning model for the LR preprocessing in the MIMO system. Based on the AlphaGo Zero architecture, this paper proposes a deep learning-based lattice reduction (DLLR) algorithm by using reinforcement learning in the AlphaGo Zero. This work investigated the performances of the DLLR in the time-varying correlated MIMO channel environments. Compare with the traditional LR-aided MIMO algorithm, the simulation results show that the DLLR can greatly improve the orthogonality of the MIMO matrices for the detection and the DLLR-aided MIMO system has better bit-error-rate performance than the LR-aided MIMO system.**

I. Introduction

Multiple-input multiple-output (MIMO) [1] system can improve the spectral efficiency of a wireless system with limited bandwidth and the MIMO channel characteristics have strong impact on the spectral efficiency and detection performance of the MIMO system. Lattice reduction preprocessing [2] was first proposed to improve the detection performance because it can transform a correlated MIMO matrix into a more orthogonal form. An orthogonal channel matrix means that the transmission is free from the interference of other transmit antennas. Therefore, the detection performance is increased. Since the benefit of the lattice reduction, lattice reduction aided MIMO detector [3] was used to enhance the detection performance. Recently, various deep learning (DL) techniques have been adopted in communication systems to solve optimization problems because of their better performances, such as DL channel estimation [4], DL MIMO detection [5]. However, deep learning techniques have not been applied to lattice reduction preprocessing for the MIMO detection because the popular DL models are not suitable for lattice reduction. Therefore, this study proposed to use the DL architecture of AlphaGo Zero [6] to construct the discrete model of the QAM constellation. Then, we used correlated MIMO channel systems to generat the training data for training the AlphaGo Zero model. An initial neural network was used to perform the the Monte-Carlo tree search [7] for data generation. In the inference phase, we applied the reinforcement learning to the AlphaGo Zero model to implement the lattice reduction algorithm. In the experiments, the LR-aided MIMO system was simulated and the proposed LR preprocessing and MIMO detection were performed to evaluate the bit error rate performance.

The rest of this paper is organized as follows. Section II introduces the time-correlated MIMO system and the lattice reduction algorithm. Section III discusses the design issues of the lattice reduction and introduces the architecture of AlphaGo Zero. Section IV presents the proposed DLLR algorithm. Section V shows the simulation final results. Finally, Section VI concludes this paper.

II. MIMO System and Lattice Reduction

A. MIMO System

An $N_r \times N_t$ MIMO system can be expressed by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{y}$ is an $N_r \times 1$ received signal vector, $\mathbf{x}$ is an $N_t \times 1$ transmitted signal vector, $\mathbf{H}$ is an $N_r \times N_t$ channel matrix, and $\mathbf{n}$ is an $N_r \times 1$ vector of i.i.d $\mathcal{CN}(0, \sigma_n^2)$ additive Gaussian noise. The complex channel model can be equivalent to the real channel model as follows:

$$\mathbf{y}_r = \mathbf{H}_r\mathbf{x}_r + \mathbf{n}_r = \begin{bmatrix} \Re(\mathbf{H}) & -\Im(\mathbf{H}) \\ \Im(\mathbf{H}) & \Re(\mathbf{H}) \end{bmatrix} \begin{bmatrix} \Re(\mathbf{x}) \\ \Im(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} \Re(\mathbf{n}) \\ \Im(\mathbf{n}) \end{bmatrix}. \tag{2}$$

The dimension of $\mathbf{H}_r$ becomes $\mathfrak{n} \times \mathfrak{m} = 2N_t \times 2N_r$. This study adopted two channel models to evalute the proposed algorithm. The first is the time-correlated channel expressed by

$$\mathbf{H}_t = \sigma\mathbf{H}_{t-1} + (1-\sigma)\mathbf{H}_R, \tag{3}$$

where $\mathbf{H}_t$ and $\mathbf{H}_{t-1}$ are successive $N_r \times N_t$ channel matrices, σ is the time-correlated factor, $0 \leq \sigma \leq 1$, and $\mathbf{H}_R$ represent an $N_r \times N_t$ Rayleigh fading channel matrix of i.i.d $\mathcal{CN}(0,1)$. This model characterizes the correlation of successive channel matrices by the correlation value σ. The second is the practical system modelled by the QuaDRiGa [8], which is a geometry-based stochastic channel and can generate a 3D MIMO scenario as the mobile station (MS) moves along a track.

B. Lattice Reduction

Lattice is an infinite grid of discrete spaced points in a vector space. A complex lattice [9] is defined as

$$\mathcal{L} = \{t_1\mathbf{b}_1 + t_2\mathbf{b}_2 + ... + t_M\mathbf{b}_M \mid t_1...t_M \in \mathbb{G}\}, \tag{4}$$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Algorithm 1 LLL LR Algorithm [10]

Input: $\mathbf{Q}, \mathbf{R}$ such that $\mathbf{H} = \mathbf{QR}$
Initialize $\bar{\mathbf{Q}} = \mathbf{Q}, \bar{\mathbf{R}} = \mathbf{R}, \mathbf{T} = \mathbf{I}_N$
$t = 2$
while $t \leq M$ **do**
  for $p = t-1, ..., 1$ **do**
    $\mu = \lceil \bar{\mathbf{R}}_{p,t} / \bar{\mathbf{R}}_{p,p} \rfloor$
    if $\mu \neq 0$ **then**
      $\bar{\mathbf{R}}_{1:p,t} \leftarrow \bar{\mathbf{R}}_{1:p,t} - \mu \bar{\mathbf{R}}_{1:p,p}$
      $\mathbf{T}_{:,t} \leftarrow \mathbf{T}_{:,t} - \mu \mathbf{T}_{:,p}$
    end if
  end for
  if $\delta \times |\bar{\mathbf{R}}_{t-1,t-1}|^2 > |\bar{\mathbf{R}}_{t,t}|^2 + |\bar{\mathbf{R}}_{t-1,t}|^2$ **then**
    swap columns $t-1$ and t in $\bar{\mathbf{R}}$ and $\mathbf{T}$
    calculate GR matrix $\Theta = \mathbf{G}(t-1, t)$
    $\bar{\mathbf{R}}_{t-1:t,t-1:N} \leftarrow \theta \bar{\mathbf{R}}_{t-1:t,t-1:N}$
    $\bar{\mathbf{Q}}_{:,t-1:t} \leftarrow \bar{\mathbf{Q}}_{:,t-1:t} \Theta^H$
    $t \leftarrow max\{t-1, 2\}$
  else
    $t \leftarrow t+1$
  end if
end while
Output: $\bar{\mathbf{Q}}, \bar{\mathbf{R}}, \mathbf{T}$ such that $\mathbf{H} = \bar{\mathbf{Q}}\bar{\mathbf{R}}$

where $\{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_M\}$ are the basis vectors and $\mathbb{G}$ denotes the Gaussian integer set, $\mathbb{G} = \mathbb{Z} + \mathbb{Z}j$. If two matrix $\mathbf{B}$ and $\mathbf{BT}$ span the same lattice, it means that $\mathbf{B}\mathbb{Z}^N = \mathbf{BT}\mathbb{Z}^N$ and thus $\mathbb{Z}^N = \mathbf{T}\mathbb{Z}^N$. $\mathbf{T}$ is an unimodular matrix, that is, $\mathbf{T}$ is invertible and the elements of $\mathbf{T}$ are Gaussian integers and $|det(\mathbf{T})| = 1$. Thus, the MIMO system can be represented as a lattice form as follows:

$$\mathbf{y} = \mathbf{h}_1 x_1 + \mathbf{h}_2 x_2 + ... + \mathbf{h}_{N_t} x_{N_t} + \mathbf{n}, \tag{5}$$

where $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{N_t}$ are the columns of $\mathbf{H}$. The reduced channel matrix can be written as

$$\hat{\mathbf{H}} = \mathbf{HT}, \tag{6}$$

where $\mathbf{H}$ is the correlated channel matrix, $\mathbf{T}$ is the unimodular matrix derived by the LR algorithm, and $\hat{\mathbf{H}}$ is the lattice reduced channel matrix.

Lenstra-Lenstra-Lovsz (LLL) LR [10] is the generic lattice reduction algorithm and it can be reformulated as the QR decomposed form in **Algorithm 1**. The LLLLR algorithm uses two constraints to derve the unimodular matrix $\mathbf{T}$. The first is the size reduction

$$\frac{1}{2} \times |\bar{\mathbf{R}}_{k,k}| \geq |\bar{\mathbf{R}}_{k,t}| \quad for\ 1 \leq k < t \leq M. \tag{7}$$

and LLL reduction as

$$\begin{aligned} &\delta \times |\bar{\mathbf{R}}_{t-1,t-1}|^2 \leq |\bar{\mathbf{R}}_{t,t}|^2 + |\bar{\mathbf{R}}_{t-1,t}|^2 \\ &for\ 2 \leq t \leq M, 0 < \delta \leq 1. \end{aligned} \tag{8}$$

The processing of LLL LR algorithm starts from column $t = 2$ based on the size reduction and LLL reduction and every columns operation is recorded in the matrix $\mathbf{T}$.

To apply the LR preprocessing for MIMO detection, we can reformulate the MIMO channel after LR as

$$\mathbf{y} = \mathbf{Hx} + \mathbf{n} = \mathbf{HTT}^{-1}\mathbf{x} + \mathbf{n} = \hat{\mathbf{H}}\mathbf{z} + \mathbf{n}, \tag{9}$$

where $\hat{\mathbf{H}} = \mathbf{HT}$ represents the lattice-reduced matrix and $\mathbf{z} = \mathbf{T}^{-1}\mathbf{x}$ represents the symbol vector in the lattice domain. The signal $\mathbf{x}$ belongs to Ω^{N_t} because of the QAM mapping, so we convert $\mathbf{x}$ to the domain of $\mathbb{G}^{N_t}$ for the LR calculation in Eq. (9) by shifting and scaling operations. For a transmitted M-QAM symbol, the transformed signal becomes $\bar{\mathbf{x}} = \mathbf{x}/a - \frac{\mathbf{1}}{\mathbf{2}}(1+j)$, where $a = \sqrt{\frac{6}{(M-1)}}$ is the power normalization factor of M-QAM. Then, the signal model is reformulated by

$$\mathbf{y} = \mathbf{Hx} + \mathbf{n} = \mathbf{H}\{\bar{\mathbf{x}} + \frac{\mathbf{1}}{\mathbf{2}}(1+j)\}a + \mathbf{n}. \tag{10}$$

After dividing a (scaling) and subtracting $\mathbf{H}\frac{\mathbf{1}}{\mathbf{2}}(1+j)$ (shifting) on both sides, the signal model becomes

$$\hat{\mathbf{y}} = \frac{1}{a}\mathbf{y} - \mathbf{H}\frac{\mathbf{1}}{\mathbf{2}}(1+j) = \mathbf{H}\bar{\mathbf{x}} + \frac{1}{a}\mathbf{n}. \tag{11}$$

This signal model is equivalent to the original MIMO system in Eq. (1) except that the transmission symbol vector is in bounded continuous integers and the noise becomes $\frac{1}{a}$. After LR, the signal model for the LR-aided detector is written as

$$\hat{\mathbf{y}} = \hat{\mathbf{H}}\bar{\mathbf{z}} + \frac{1}{a}\mathbf{n}, \tag{12}$$

where $\hat{\mathbf{H}}$ is the reduced matrix and $\bar{\mathbf{z}} = \mathbf{T}^{-1}\bar{\mathbf{x}}$ is the transmit symbol vector in the lattice domain. Then, the lattice-reduction-aided maximum likelihood (ML) detection becomes

$$\hat{\mathbf{z}} = \underset{z \in \Omega^{N_t}}{\arg\min} |\hat{\mathbf{y}} - \hat{\mathbf{H}}\mathbf{z}|. \tag{13}$$

Afterwards, the function $round()$ must be used to slice $\mathbf{T}\hat{\mathbf{z}}$ to the closest point of $\{\bar{\mathbf{x}}\}$ as the detection result as follows:

$$\hat{\mathbf{x}} = \{round(\mathbf{T}\hat{\mathbf{z}}) + \frac{\mathbf{1}}{\mathbf{2}}(1+j)\}a. \tag{14}$$

III. Deep Learning Model for Lattice Reduction

A. Design Issues of Deep Learning for Lattice Reduction

This work aims to design the DL-based LR-aided MIMO preprocessing for the time-varying channel, in which the data association exists between channel $\mathbf{H}_t$ and $\mathbf{H}_{t-1}$. From Eq. (6), the lattice-reduced channel matrix is the product of the original channel matrix and the unimodular matrix. Therefore, if $\mathbf{H}_t$ and $\mathbf{H}_{t-1}$ have the data association, then so do $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{H}}_{t-1}$. Thus, $\mathbf{T}_t$ and $\mathbf{T}_{t-1}$ may also have the data association. The temporal relevancy can be used to predict the unimodular matrix $\mathbf{T}$ by the LR model. In this paper, the channel matrix $\mathbf{H}$ is represented in a real model with the same number of antennas in the transceiver and receiver, that is, $\mathbf{H} \in \mathbb{R}^{\mathcal{M} \times \mathcal{M}}, \mathbf{T} \in \mathbb{Z}^{\mathcal{M} \times \mathcal{M}}$ and $\mathcal{M} = 2N_t$. Because $\mathbf{T}$ is invertible and unimodular with all integer entries and $|det(\mathbf{T})| = 1$ and finding the unimodular matrix is an NP-hard problem [11], the DL model is incapable of accomplishing the LR algorithm directly. Therefore, we proposed a path matrix $\mathbf{P}$ by the DL model, thereby obtaining $\mathbf{T}$ indirectly. The path matrix $\mathbf{P}$ is represented as

$$\begin{aligned} &\mathbf{T}_t = \mathbf{T}_{t-1}\mathbf{P}, \\ &\mathbf{P} = (\mathbf{T}_{t-1})^{-1}\mathbf{T}_t = \mathbf{I}_{N_t} \cdot \mathbf{P}'_1\mathbf{P}'_2 \ldots \mathbf{P}'_I, \end{aligned} \tag{15}$$

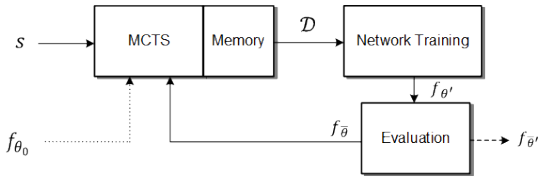


Fig. 1: Architecture of AlphaGo Zero.

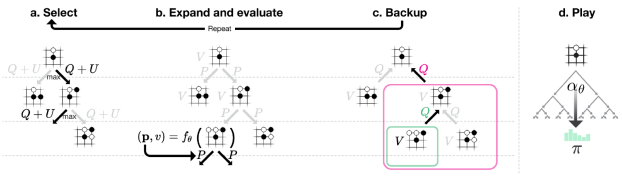


Fig. 2: Monte-Carlo Tree Search in AlphaGo Zero, from [6].

where $\mathbf{P}'_i$ is the step matrix for $i = 1, 2, ..., I$. Each step matrix is the function of (o, x, y), and a step matrix is expressed as

$$\mathbf{P}'_i = \mathbf{I}(o, x, y), \tag{16}$$

where o is the operator similar to μ in the LLL LR algorithm, and $o \in \mathbf{o}, \mathbf{o} = [\pm 1, \pm 2, \pm 4, \pm 8, \pm 16, swap, none]$. (x, y) are the column and row indices of the operator o in $\mathbf{P}'_i$. For example, if $\mathbf{I}(o, x, y) = \mathbf{I}(-2, 3, 4)$, then

$$\mathbf{P}'_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \tag{17}$$

Therefore, the unimodular matrix $\mathbf{T}_t$ can be obtained from $\mathbf{T}_{t-1}$ by the path matrix $\mathbf{P}$ consisting of step matrices $\mathbf{P}'_i$. Then, we can use step matrices $\mathbf{P}'_i$ in Eq. (16) to construct the DL model, which predicts integers o, x, and y for constructing path matrix $\mathbf{P}$ and $\mathbf{T}$ by Eq. (15). Because the defined lattice reduction DL model is similar to the Go game, this study proposed to adopt the DL architecture of the AlphaGo Zero to solve this complex decision problem.

B. AlphaGo Zero

AlphaGo Zero is a version of DeepMind's Go software AlphaGo [12] with the reinforcement learning technique [13] and it can be roughly divided into three stages of Monte-Carlo tree search (MCTS), neural network, and evaluation. Figure 1 shows the architecture of AlphaGo Zero, s is the state of the Go board, f_{θ_0} is the neural network with initial weights, $\mathcal{D}$ is the training dataset, $f_{\theta'}$ is the neural network with updated weights, $f_{\bar{\theta}}$ is the neural network with updated weights after evaluation, and $f_{\bar{\theta}'}$ is the final network after all iterations.

1) Monte-Carlo Tree Search

MCTS can be regarded as a form of self-play reinforcement learning. It focuses on the analysis of the most promising Go move, which is called policy. The goal of MCTS is to improve the selection policy of the Go move. MCTS in AlphaGo Zero is based on many "playouts", and a playout is a Go game. Each gain is played out to the very end by choosing Go moves through a selecting equation. Then, the final game result of each playout is used to weight the tree nodes, so that the better nodes are more likely to be chosen in the future. Additionally, MCTS is a tree-like structure consisting of many nodes, and each node stores several weights including a visit count $\mathcal{N}(s, a)$, a prior probability $\mathcal{P}(s, a)$, and an action-value $\mathcal{Q}(s, a)$, where s is the board state and a is the Go move. The process of MCTS in AlphaGo Zero is shown in Figure 2. The input of MCTS is the state of the game board, which is a 19×19 image with the position of Go stones. The output is the Go move policy, which is represented by a vector with the size $N \times 1$ with $N = 19 \times 19 + 1$ all possible Go moves on the board and "pass" moves. The neural network $f_{\bar{\theta}}$ guides the MCTS simulation and is composed of four steps. An iteration of running **step a** to **step c** is called a simulation, and the MCTS simulation is repeated K times in total ($K = 1600$). Then, **step d** will calculate the Go move policy.

a. Select

This step selects the child node down to the leaf node N_L continuously through the selecting equation starting from the root node. The root node is the current game state and the leaf node is the state after playing Go moves. The selecting equation is expressed as

$$a_i = \arg\max_{a_n} \{\mathcal{Q}(s_i, a_n) + \mathcal{U}(s_i, a_n)\}, i \leq N_L, n = 1, 2, ..., N, \tag{18}$$

where a_i is the selected node in the i-th layer of the tree, a_n are the nodes that represente all available Go moves. $\mathcal{Q}(s_i, a_n)$ is action-value with game state s_i and Go move a_n, $\mathcal{Q}(s_i, a_n) = \mathcal{W}(s_i, a_n) / \mathcal{N}(s_i, a_n)$, where $\mathcal{W}(s_i, a_n)$ is the accumulation of the evaluation value from the neural network and $\mathcal{N}(s_i, a_n)$ is the visit count. $\mathcal{U}(s_i, a_n)$ is the exploration weight with game state s_i and Go move a_n, $\mathcal{U}(s_i, a_n) = \mathcal{C} \cdot \mathcal{P}(s_i, a_n) / 1 + \mathcal{N}(s_i, a_n)$, where $\mathcal{C}$ is the exploration constant ($\mathcal{C} = 5$) and $\mathcal{P}(s_i, a_n)$ is the prior probability from the neural network.

b. Expand and Evaluate Afterwards, the search tree is expanded by the neural network. The leaf node s_{N_L} becomes the input of the neural network $f_{\bar{\theta}}$, and the network generates output move probabilities $\mathbf{p}$ and evaluation value v to expand the search tree. Then, the weights of the expanded node are initialized as

$$\begin{aligned} \mathcal{Q}(s_{N_L+1}, a_n) &= 0, \\ \mathcal{W}(s_{N_L+1}, a_n) &= 0, \\ \mathcal{N}(s_{N_L+1}, a_n) &= 0, \\ \mathcal{P}(s_{N_L+1}, a_n) &= p_{a_n}, \end{aligned} \tag{19}$$

where $p_{a_n} \in \mathbf{p}$ and p_{a_n} is the probability of move a_n. To ensure that all nodes be visited, the prior probability in the root node expansion is added with the Dirichlet noise as follows:

$$\mathcal{P}(s_1, a_n) = (1 - \epsilon) p_{a_n} + \epsilon \eta, \tag{20}$$

where ϵ is a constant ($\epsilon = 0.25$), and η is the Dirichlet noise, $\eta \sim \texttt{Dir}(0.03)$.

c. Backup

In this step, to make a better result in the next round selection, the parent nodes of the expanded node in **step b** are updated. The update function is expressed as

$$\begin{aligned}\mathcal{W}(s_i,a_n) &= \mathcal{W}(s_i,a_n)+v,\\ \mathcal{N}(s_i,a_n) &= \mathcal{N}(s_i,a_n)+1,\\ \mathcal{Q}(s_i,a_n) &= \mathcal{W}(s_i,a_n)\,/\,\mathcal{N}(s_i,a_n),\end{aligned} \tag{21}$$

where v is the evaluation value generated by the neural network $f_{\bar{\theta}}$ in **step b**.

d. Play

After K times simulations, the final output, the Go move policy, is calculated as

$$\boldsymbol{\pi} = \frac{\mathcal{N}(s_1,a_i)^{\frac{1}{\tau}}}{\sum_{n=1}^{N}\mathcal{N}(s_1,a_n)}, \forall i \in \{1,2,...,N\}, \tag{22}$$

where τ is the temperature parameter ($\tau = 1$). In the end, the results of MCTS are stored as training dataset $\mathcal{D}$ in the form of $(s_t, \boldsymbol{\pi}_t, z_t)$, where s_t is the state of the board at time t, $\boldsymbol{\pi}_t$ is the output policy for state s_t, and z_t is the winner from the current player's perspective at time t.

2) Neural Network

The input of the neural network is an image stack with size $19\times19\times17$, where 17 represents the number of feature planes. 8 of 17 feature planes, X_t, represent the current player's Go stones. $X_t = 1$ if the intersection i has a stone of the player at time t; $X_t = 0$ if the intersection has an opponent stone or is empty; if $t < 0$, X_t is also written as 0. Another 8 of 17 feature planes, Y_t, indicate the corresponding features for the opponent's Go stones. And the final feature plane, C, denotes the stone color to play, $C = 1$ if black is to play; $C = 0$ if white is to play. Thus, the input of the neural network $s_t = [X_t, Y_t, X_{t-1}, Y_{t-1}, ..., X_{t-7}, Y_{t-7}, C]$. The outputs to the neural network are an $N \times 1$ vector of the policy $\mathbf{p}_{\texttt{pred}}$ and a scalar of the evaluation value $v_{\texttt{pred}}$ corresponding to the MCTS results $\boldsymbol{\pi}$ and z, respectively.

IV. Proposed DL-Based LR-Aided MIMO System

A. Proposed Deep Learning Based Lattice Reduction

Figure 3 shows the architecture of the proposed deep learning-based lattice reduction. The proposed algorithm can be divided into offline and online schemes. In the offline scheme, the proposed algorithm architecture is referred to AlphaGo with similar neural network and MCTS model. Then, the trained neural network $f_{\bar{\theta}'}$ is used in the online scheme. In the online scheme, we adopted the MCTS model with the trained neural network to obtain step matrices $\mathbf{P}'_i$, which can be combined into the path matrix $\mathbf{P}$. Finally, we can obtain the unimodular matrix $\mathbf{T}_t$ through the path matrix. $\mathbf{S}$ is the product of the channel matrix $\mathbf{H}$ and the step matrix, and $\mathbf{B}$ is the unimodular matrix $\mathbf{T}$.

To apply AlphaGo Zero to the DL-based, the winning of a game is defined by channel orthogonality as follows:

$$O = \frac{\sum(|\widetilde{\mathbf{S}}|_{\text{diag}})}{\sum(|\widetilde{\mathbf{S}}|_{\text{off-diag}})}, \tag{23}$$

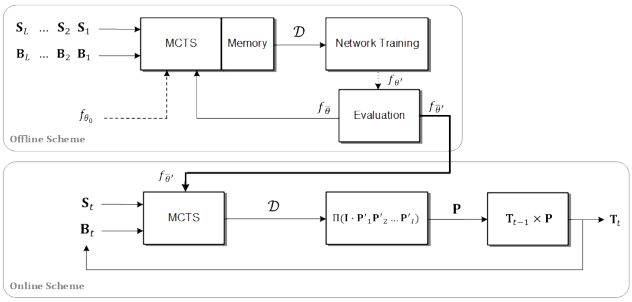


Fig. 3: The proposed architecture of DL-based LR.

where $\widetilde{\mathbf{S}}$ is the Gram matrix of $\mathbf{S}$. If the channel orthogonality is improved, it means the game wins. The following two inequalities are used to determine whether the orthogonality improves or not.

$$O_I > \lambda_O \text{ or } d > \lambda_d, \tag{24}$$

where O_I is the orthogonality after DLLR (orthogonality at step I), λ_O is the orthogonality threshold (λ_O = 10). d is the difference between O_I and O_0 (O_0 is the orthogonality of the original channel), and λ_d is the difference threshold (λ_d = 2). The action of the lattice reduction is defined as the step matrix associated with the operator and operating columns (o, x, y) in Eq. (16) and the number of actions N is

$$N = (10 \times \mathcal{M} \times (\mathcal{M}-1)) + (1 \times \mathcal{M} \times (\mathcal{M}-1)) + 1. \tag{25}$$

1) Offline Scheme

The offline scheme aims to train the DL model in Figure 3. The training process includes MCTS model, network training model, and evaluation model.

A. MCTS Model Figure 4 shows the architecture of the MCTS model in DLLR. The inputs of the MCTS are $\mathbf{S}$ and $\mathbf{B}$. $\mathbf{S}$ is the state matrix, which can be expressed by

$$\mathbf{S}_{l,i} = \mathbf{S}_{l,i-1} \times \mathbf{P}'_{l,i-1}, l = 1,2,...,L, i = 2,3,...,I, \tag{26}$$

where the initial $\mathbf{S}_{l,1} = \mathbf{H}_l$, $\mathbf{P}'_{l,1} = \mathbf{I}$. L denotes the channel number, and I is the step limit. The other input $\mathbf{B}$ is the bias matrix, which is the previous unimodular matrix

$$\mathbf{B}_l = \mathbf{T}_{l-1}, l = 1,2,...,L, \tag{27}$$

where the initial $\mathbf{T}_0$ is set to be $\mathbf{I}$. This bias matrix is used to find out the temporal features. The output of the MCTS simulation is a group of training data, which includes the state matrix $\mathbf{S}_{l,i}$, bias matrix $\mathbf{B}_l$, the policy $\boldsymbol{\pi}_{l,i}$ for the state $\mathbf{S}_{l,i}$, and the orthogonality improvement z_l. The former $\mathbf{S}_{l,i}$ and $\mathbf{B}_l$ are the aforementioned input matrices, and the latter policy $\boldsymbol{\pi}_{l,i}$ is the result of the tree search for state $\mathbf{S}_{l,i}$. z_l indicates whether the channel orthogonality is improved ($z_l = 1$) or not ($z_l = 0$), and it is only generated at the last step I. In addition, z_l of the training data from step 1 to step $I-1$ is equal to z_l at step I. Then, the data are stored as training dataset $\mathcal{D}$ for the network model. When the data is stored, the policy $\boldsymbol{\pi}_{l,i}$ is placed into the matrix operation at the same time. In

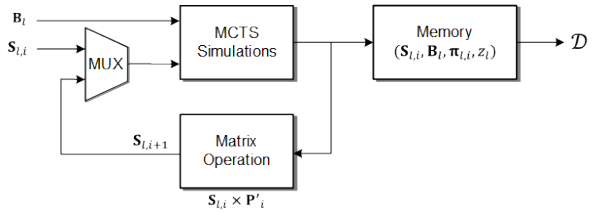


Fig. 4: Architecture of MCTS in DLLR.

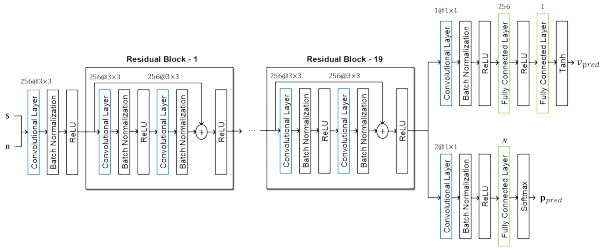


Fig. 5: Architecture of neural network in DLLR.

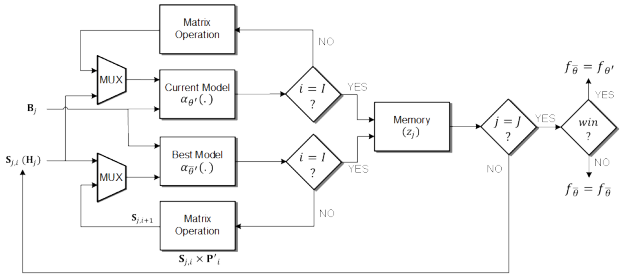


Fig. 6: Architecture of evaluation in DLLR.

matrix operation, we firstly transform $\boldsymbol{\pi}_{l,i}$ into step matrix $\mathbf{P}'_i$ as follows:

$$(o, x, y)_{\max} = a_{\max} = \max\{\boldsymbol{\pi}_{l,i}\}, \tag{28}$$

where $a_{\max}$ is the best action in $\boldsymbol{\pi}_{l,i}$ and $(o, x, y)_{\max}$ performs the column operations for the best action $a_{\max}$. Then, we can obtain the step matrix $\mathbf{P}'_i$ through Eq. (16). Finally, the input of the next step are generated as $\mathbf{S}_{l,i+1} = \mathbf{S}_{l,i} \times \mathbf{P}'_i$.

B. Network Training Model

Figure 5 shows the neural network architecture in DLLR. The DLLR network has two inputs, $\mathbf{S}$ and $\mathbf{B}$ in dataset $\mathcal{D}$. Then, $\boldsymbol{\pi}$ and z in dataset $\mathcal{D}$ are the labels for the neural network output $\mathbf{p}_{\text{pred}}$ and v_{pred}, respectively. $\mathbf{p}_{\text{pred}}$ is an $N \times 1$ vector that denotes the probability of actions, and v_{pred} is a scalar that denotes the probability of improving the orthogonality. We used the same loss function as that in the AlphaGo Zero. The Adam optimizer [14] was used to update all trainable parameters (weights) θ by

$$\begin{aligned} g_r &= \nabla_\theta \Phi \\ m_r &= \beta_1 m_{r-1} + (1-\beta_1) g_r \\ v_r &= \beta_2 v_{r-1} + (1-\beta_2) g_r^2 \\ \hat{m}_r &= \frac{m_r}{1-\beta_1^r} \\ \hat{v}_r &= \frac{v_r}{1-\beta_2^r} \\ \theta' &= \theta - \frac{\rho}{\sqrt{\hat{v}_r} + \epsilon} \hat{m}_r, \end{aligned} \tag{29}$$

where r is an iteration index, g_r is a gradient element, Φ denotes the loss function. m_r is the exponential moving averages of the gradient, v_r is the squared gradient, $\beta_1, \beta_2 \in [0, 1)$ are the hyper-parameters that control the exponential decay rates of these moving averages. ($\beta_1 = 0.9, \beta_2 = 0.999$). $\hat{m}_r$ and $\hat{v}_r$ are the bias corrections. θ' denotes the updated weights of the network, ρ is a learning rate ($\rho = 0.1$), and ϵ is a constant that avoids divisor by zero ($\epsilon = 10^{-7}$).

C. Evaluation Model

The evaluation is required to ensure the quality of training data for the DLLR. Figure 6 illustrates the architecture of the evaluation in the DLLR. We evaluated the current network f'_θ through the best network $f_{\bar{\theta}}$ at every checkpoint. $\mathbf{S}_{j,i}$ and $\mathbf{B}_j$ are inputs of the models of α'_θ and $\alpha_{\bar{\theta}}$, and the models operates for I steps to derive the results. Then, we determined the winner between the current network f'_θ and the best network $f_{\bar{\theta}}$, the winner z_j is represented as

$$z_j = \begin{cases} 1, & O_{I,\text{current}} > O_{I,\text{best}} \text{ or } d_{\text{current}} > d_{\text{best}}. \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

where $O_{I,\text{current}}$ and $O_{I,\text{best}}$ denote the orthogonality of the current network and the best network at step I, respectively, and d_{current} and d_{best} denote the differences of the current network and the best network, respectively. In addition, $d_{\text{current}} = O_{I,\text{current}} - O_{0,\text{current}}$ and $d_{\text{best}} = O_{I,\text{best}} - O_{0,\text{best}}$. After evaluating all input evaluation channels, we calculated the winning ratio Z

$$Z = \frac{\sum z_j|_{z_j=1}}{J}, \tag{31}$$

where J is the number of channels. The following inequality is used to determine whether the best network is updated from the current network or not

$$Z > \lambda_Z, \tag{32}$$

where λ_Z is the threshold of the winning rate ($\lambda_Z = 0.55$). If Z is greater than the threshold, we update the current network to the latest best network. Then, the best network is adopted in the next iteration of the MCTS model.

Algorithm 2 shows the offline algorithm. At the beginning, $\mathbf{T}$ is generated by the LLLLR algorithm (Line 1), as shown in **Algorithm 1**. Then, Line 2 to Line 5 prepare the input data for the MCTS model and T is the channel number. Line 6 splits the data into the training set and the evaluation set. Line 7 to Line 14 are the main process of DLLR. We initial $\bar{\theta}$ in Line 7, where θ_0 is generated randomly. R is the number of iterations. The MCTS model generates the training dataset $\mathcal{D}$ in Line 9, and the neural network is trained in Line 10. Then, we evaluate the networks at every checkpoint (Line 11

Algorithm 2 Deep Learning-based Lattice Reduction (Offline Scheme)

Input: $\mathbf{H}, T, R$
1: Generate $\mathbf{T} = LLL(\mathbf{H})$
2: **for** $t = 1, ..., T$ **do**
3: $\quad \mathbf{S}_t = \mathbf{H}_t$
4: $\quad \mathbf{B}_t = \mathbf{T}_{t-1}$
5: **end for**
6: Split $(\mathbf{S}, \mathbf{B})$ to $(\mathbf{S}_{\text{train}}, \mathbf{S}_{\text{eval}}, \mathbf{B}_{\text{train}}, \mathbf{B}_{\text{eval}})$
7: Initial $\bar{\theta} = \theta_0$
8: **for** $r = 1, ..., R$ **do**
9: $\quad$ Generate $\mathcal{D} = \alpha_\theta(\mathbf{S}_{\text{train}}, \mathbf{B}_{\text{train}})$
10: $\quad$ Calculate $(\mathbf{p}, v) = f_{\bar{\theta}}(\mathcal{D})$
11: $\quad$ **if** is checkpoint **then**
12: $\quad\quad$ Evaluate $\alpha'_\theta(\mathbf{S}_{\text{eval}}, \mathbf{B}_{\text{eval}})$ and $\alpha_{\bar{\theta}}(\mathbf{S}_{\text{eval}}, \mathbf{B}_{\text{eval}})$
13: $\quad$ **end if**
14: **end for**
15: $\bar{\theta}' = \bar{\theta}$
Output: $\bar{\theta}'$

to Line 13) of every 5 iterations. Finally, the final best network weights come out after running all iterations (Line 15).

2) Online Scheme After the offline scheme, the best weights of the neural network $\bar{\theta}'$ is used in the online scheme. Figure 3 shows the online scheme architecture and **Algorithm 3** shows the the online processing algorithm. Firstly, we initial $\mathbf{T}_0$ and set the input $\mathbf{S}_t$ and $\mathbf{B}_t$ (line 1 $\sim$ line 4). The MCTS model with the weights is denoted as $\alpha'_{\bar{\theta}}$, and it will output the step matrices $\mathbf{P}'_i$ (line 5 $\sim$ line 6). Then, the path matrix $\mathbf{P}$ is generated by multiplying those step matrices (line 7). Finally, multiplying $\mathbf{T}_{t-1}$ and path matrix $\mathbf{P}$ can get the expected unimodular matrix $\mathbf{T}_t$ (line 8), where $\mathbf{T}_{t-1}$ is $\mathbf{B}_t$.

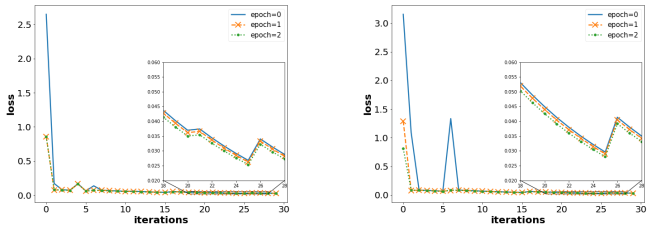
Algorithm 3 Deep Learning-based Lattice Reduction (Online Scheme)

Input: $\mathbf{H}, T, \bar{\theta}'$
1: Initial $\mathbf{T}_0 = \mathbf{I}$
2: **for** $t = 1, ..., T$ **do**
3: $\quad \mathbf{S}_t = \mathbf{H}_t$
4: $\quad \mathbf{B}_t = \mathbf{T}_{t-1}$
5: $\quad$ Generate $\mathcal{D} = \alpha_{\bar{\theta}'}(\mathbf{S}_t, \mathbf{B}_t)$
6: $\quad$ Transform $\boldsymbol{\pi}$ to $\mathbf{P}'$
7: $\quad$ Calculate $\mathbf{P} = \mathbf{P}'_1 \mathbf{P}'_2 \ ... \ \mathbf{P}'_I$
8: $\quad \mathbf{T}_t = \mathbf{T}_{t-1} \times \mathbf{P}$
9: **end for**
Output: $\mathbf{T}$

V. Simulation and Analysis Results

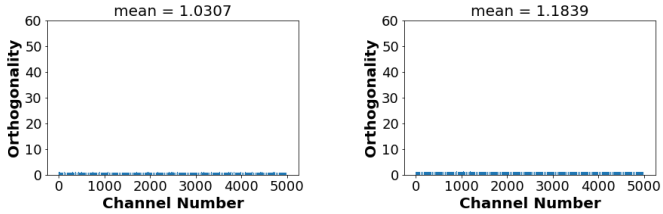
A. Simulation Environments

This study considered the MIMO system with $N_t = N_r = 4$. In the time-correlated channel, correlation factors of $\sigma = 0.2$ and $\sigma = 0.8$ were considered. For the QuaDRiGa channel, the mobile station moves along a 2000-meter random street track. The QuaDRiGa simulation scenario adopted mmMAGIC 3D urban micro-cell measured in Europe. This scenario is for typical terrestrial pico-base stations deployed below rooftops in densely populated urban areas. The number of cluster N_{cl} in the scenario is 3 and the number of subpaths N_{path} in each cluster is 20.



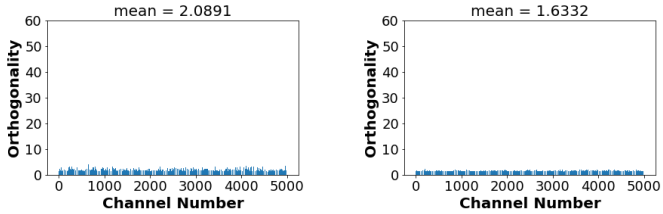
(a) (b)

Fig. 7: Convergence curves of training loss for the time-correlated factors of (a) $\sigma = 0.8$ and (b) $\sigma = 0.2$.



(a) (b)

Fig. 8: Orthogonality of the original channels with time-correlated factors of (a) $\sigma = 0.8$ and (b) $\sigma = 0.2$



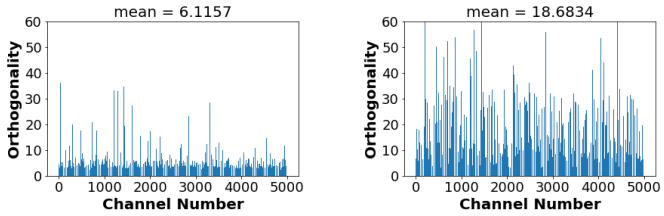
(a) (b)

Fig. 9: Orthogonality of the LLLLR-reduced channels with time-correlated factors of (a) $\sigma = 0.8$ and $\sigma = 0.2$.

In the offline scheme, we use the number of iterations $R = \{25, 30\}$ to train the system for $\sigma = 0.2$ and $\sigma = 0.8$, respectively. Thresholds of the orthogonality λ_O and difference λ_d are 10 and 2, respectively. The total channel number T is 20000, and the channel number per iteration L is 10. In the MCTS model, we adopt the simulation times $K = \{100, 150\}$ and step limit $I = \{15, 20\}$ for $\sigma = 0.2$ and $\sigma = 0.8$, respectively. In the network training, epoch $E = 3$, batch size $B = 32$, and learning rate $\rho = 0.1$ were used. In the evaluation model, the input channel number J is 20, the checkpoint is every 5 iterations, and the threshold of the winning rate $\lambda_Z = 0.55$.

B. Simulation Results of Time-Correlated Channel

Figure 7 shows the DLLR training loss in time-correlated factor $\sigma = 0.8$ and $\sigma = 0.2$ respectively. The step limit $I = 20$ and simulation times $K = 150$ were used in the simulation. We can see that the higher time-correlated channel model converges faster and the convergence results are better



(a) (b)

Fig. 10: Orthogonality of the DLLR-reduced channels with time-correlated factors of $\sigma = 0.8$ and $\sigma = 0.2$.

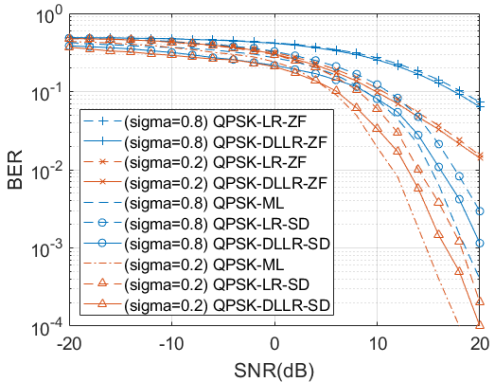
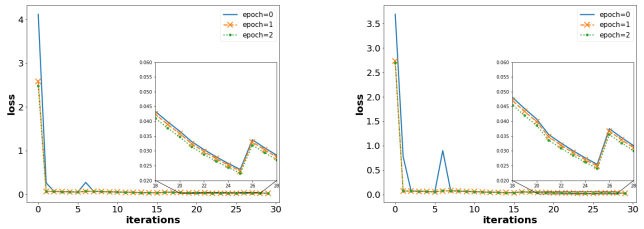


Fig. 11: BER performance curves of the LR-aided and DLLR-aided MIMO detectors with different time-correlated channels.

because the DLLR trains the data association in the time-varying channel. Thus, the higher temporal correlation resulted in better training results.

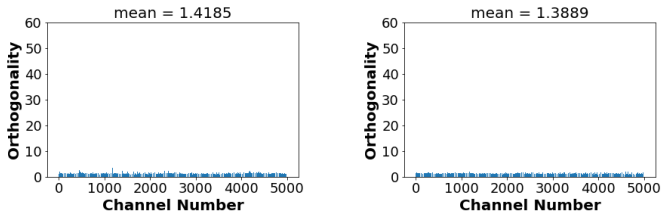
Figure 8 shows the channel orthogonality of the original channels with time-correlated factors of $\sigma = 0.8$ and $\sigma = 0.2$. Figure 9 plot the channel orthogonality of the reduced channels through the LLLLR algorithm. We can see that the lattice reduction algorithm (LLLLR) can effectively iincrease the channel orthogonalities under various correlation values. Figure 10 shows the channel orthogonality of the reduced channel through the DLLR algorithm. Compare with the channel orthogonalities of the original channel and the LLLLR-reduced channel, the DLLR can increase the channel orthogonality more effectively than the LLLLR algorithm. This result shows that the DLLR can figure out the data association between channels, so the model may be applied to the time-varying environment to increase the channel orthogonality. In Figure 10, the DLLR-reduced channel with $\sigma = 0.2$ has better orthogonality than that with $\sigma = 0.8$ because of the orthogonality $\lambda_O = 10$ and the threshold of the difference $\lambda_d = 2$. To speed up the DLLR calculations, the MCTS simulation can be stopped earlier if the channel orthogonality of the current step meets the winning condition. Therefore, the orthogonality of the channel with $\sigma = 0.8$ is lower than the one with $\sigma = 0.2$ because the MCTS simulations for $\sigma = 0.8$ was earlier stopped.

Figure 11 shows the BER performances of various lattice-



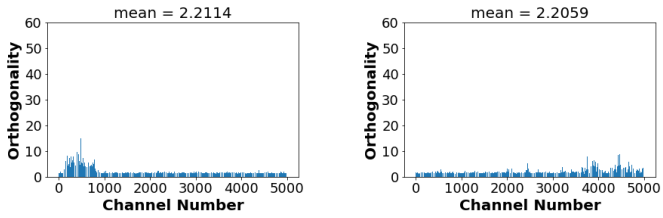
(a) (b)

Fig. 12: Convergence curves of training loss for the cluster numbers of (a) $N_{cl} = 3$ and (b) $N_{cl} = 6$.



(a) (b)

Fig. 13: Orthogonality of the original channels with (a) $N_{cl} = 3$ and (b) $N_{cl} = 6$ clusters.



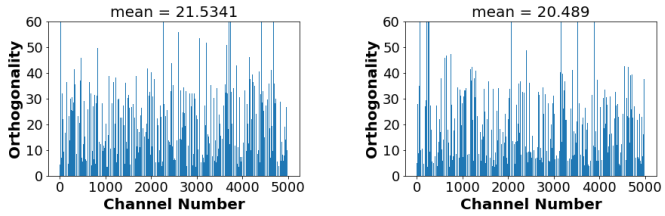
(a) (b)

Fig. 14: Orthogonality of the LLLLR-reduced channels with $N_{cl} = 3$ and $N_{cl} = 6$ clusters.

reduction-aided detectors in the time-correlated MIMO channels with $\sigma = 0.8$ and $\sigma = 0.2$. We used QPSK modulation in the simulations and compared the performances of the LLLLR and DLLR algorithms. Since the channel orthogonality of the DLLR algorithm is better than the LLLLR algorithm, we can see that the BER of the DLLR is lower than that of the LLLLR. The DLLR algorithm enhances the channel orthogonality in the time-correlated channel model and obtains a better performance in the simulation of the MIMO system.

C. Simulation Results of QuaDRiGa Channel

Figure 12 shows the convergence curves of the DLLR training loss in the QuaDRiGa channel model with the number of the clusters $N_{cl} = 3$ and $N_{cl} = 6$, respectively. The training results of two cases are almost the same and their loss values finally approach 0.025. Figure 13 shows the channel orthogonaltiy of the QuaDRiGa channel models with $N_{cl} = 3$ and $N_{cl} = 6$. Figure 14 shows the channel orthogonality of the LLLLR-reduced channels in two cases. The orthogonality



(a) (b)

Fig. 15: Orthogonality in the reduced channel (DLLR) under different cluster numbers $N_{cl} = 3$ $N_{cl} = 6$.

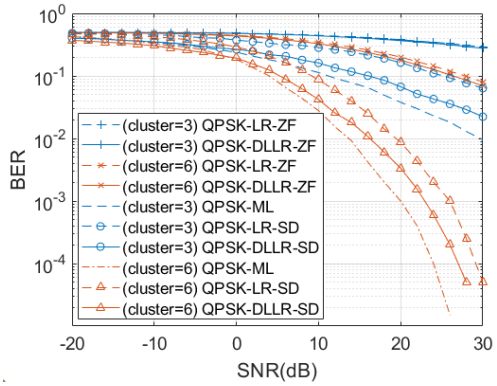


Fig. 16: BER performance curves of the LR-aided and DLLR-aided MIMO detectors with different cluster numbers in the QuaDRiGa channel.

of the LLLLR-reduced channel is better than that of the original channel similar to the time-correlated channel. Figure 15 shows the channel orthogonality of the DLLR-reduced channel in two cases. The DLLR can also improves the channel orthogonality significantly in the QuaDRiGa model. The data association between channels is trained by the DLLR algorithm and the number of clusters cause no difference in the training results of the DLLR algorithm. The primary reason is that the purpose of DLLR is to train the data association between channels. A larger cluster number represents the higher rank of the channel matrix, thus the DLLR algorithm causes no impact on the lattice reduced channels of different numbers of clusters. Figure 16 shows the BER performances of various lattice-reduction-aided detectors in the QuaDRiGa channels with $N_{cl} = 3$ and $N_{cl} = 6$ clusters. We also used QPSK modulation for the simulation and compared the performances of the LLLLR and DLLR algorithms with two cases. The orthogonality of the DLLR algorithm is better than that of the LLLLR algorithm, and the BER performances of the DLLR-aided detectors is also better than that of the LLLLR-aided detectors.

VI. Conclusion

This paper presents the lattice reduction preprocessing for time-correlated MIMO systems by the proposed AlphaGo-Based reinforcement learning. The offline and online trainining techniques were proposed to design the lattice reduction model that can adapt to the continous time-varying channel property. Final simulation results show that the proposed DLLR algorithm achieved much better performances than the original LR algorithm in terms of channel orthogonality and BER performance.

References

[1] K. Yu and B. Ottersten, "Models for mimo propagation channels: a review," *Wireless communications and mobile computing*, vol. 2, no. 7, pp. 653–666, 2002.

[2] D. Wbben, D. Seethaler, J. Jaldn, and G. Matz, "Lattice reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 70–91, 2011.

[3] H. Yao and G. Wornell, "Lattice-reduction-aided detectors for mimo communication systems," in *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE*, vol. 1, 2002, pp. 424–428 vol.1.

[4] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, "Deep learning-based channel estimation," *IEEE Communications Letters*, vol. 23, no. 4, pp. 652–655, 2019.

[5] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "A model-driven deep learning network for mimo detection," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 584–588.

[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[7] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of monte carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, 2012.

[8] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "Quadriga: A 3-d multi-cell channel model with time evolution for enabling virtual field trials," *IEEE transactions on antennas and propagation*, vol. 62, no. 6, pp. 3242–3256, 2014.

[9] Y. H. Gan, C. Ling, and W. H. Mow, "Complex lattice reduction algorithm for low-complexity full-diversity mimo detection," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2701–2710, 2009.

[10] J. Jalden, D. Seethaler, and G. Matz, "Worst- and average-case complexity of lll lattice reduction in mimo wireless systems," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2685–2688.

[11] A. Becker, N. Gama, and A. Joux, "Solving shortest and closest vector problems: The decomposition approach," *Cryptology ePrint Archive*, 2013.

[12] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[13] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] C. Ling and N. Howgrave-Graham, "Effective lll reduction for lattice decoding," in *2007 IEEE International Symposium on Information Theory*, 2007, pp. 196–200.

[18] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5353–5360.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Utilizing Unlabeled Data and Synthetic Data for Bird Sound Detection: Consistency Training, Mean Teacher, and Domain Adaptation Techniques

Fang-Ching Chen* and Yi-Wen Liu†
* National Tsing Hua University, Taiwan
E-mail: fumchin@gapp.nthu.edu.tw
† National Tsing Hua University, Taiwan
E-mail: ywliu@ee.nthu.edu.tw Tel/Fax: +886-35162205

***Abstract*—Labeling onset and offset time and class information for audio files has always been a laborious task for sound event detection (SED) research. Therefore, it would be appealing to consider using strongly-labeled synthetic data and a large amount of unlabeled data when training a SED model. To utilize these data, previous research showed that mean-teacher based semi-supervised consistency training strategies could work with adversarial domain adaptation. However, the effectiveness has only been demonstrated in the domestic environment. In this work, we expand the scope of application to bird sound event detection in the natural environment. To achieve this, we established a synthetic strongly-labeled bird sound dataset consisting of 20 species. Then, we tuned the consistency training strategies and domain adaptation techniques that have been proven effective in the domestic environment to this new dataset. Our evaluation on the Eastern North American bird sound dataset resulted in an F1 score of >30%, which approaches the upper-bound performance that could be obtained when strongly-labeled data are assumed to be available. Thus, the strategies exhibit potential for advancing research in both indoor and outdoor SED.**

I. Introduction

Sound event detection (SED) aims to recognize and determine the onset and offset of different events in an audio recordings. Over the years, different deep learning models and training strategies have been developed to address SED in domestic and urban environments. The annual challenge on domestic SED held by DCASE has accelerated research in this area. However, primarily due to the lack of strongly-labeled data in these domains, it is still challenging to apply SED techniques in natural soundscapes. Particularly, bioacoustic SED in the animals' natural habitats [1][2] remains relatively under-explored.

The study of bird vocalization covers various tasks, including bird audio detection [3], bird sound classification and bird sound event detection. Bird audio detection focuses on the presence or absence of bird sound in an audio recording without classifying the specific bird type. Bird audio detection is crucial to applications such as automatic wildlife monitoring. In contrast, bird sound classification focuses on accurately categorizing the species or types of birds present in an audio recording, without necessarily determining the exact starting and ending times of individual sound events. In recent years, the advance of signal processing [4] and the development of deep learning models [5][6] have revolutionized the field of bird audio detection and bird sound classification. Finally, bird sound event detection not only requires classification of bird species, but also aims to determine the starting time and ending time of the bird calls.

In this research, our objective is to extend several methods and strategies which were proven useful in domestic SED [7] to outdoor environments, especially for detection of bird sound events. A pioneering work utilized the hidden Markov model [8] to detect the onset and offset of Zebra finch calls that were recorded in an indoor aviary. Later, as deep learning became popular, convolutional recurrrent neural networks (CRNNs) were constructed and tested on a synthetic bird sound dataset that contained 5 bird species [9]; the CRNNs were shown to be capable of detecting overlapping bird sounds. More recently, there has been an investigation on whether CRNNs can handle highly dense bird acoustic scenarios where up to 10 birds can call at the same time [10], with a set of strongly-labeled bird sound recordings from the Eastern North America (ENA) region [11]. In this paper, while we also adopted the ENA dataset for it contains full annotation of bird species, onsets, and offsets, we focused on semi-supervised learning (SSL) and domain adaptation strategies. Overall, we aimed to utilize a large amount of out-of-domain synthetic data together a limited amount of in-domain data that were either unlabeled or weakly labeled. The question we are asking is – can unlabeled data and synthetic data help for SED when strongly-labeled in-domain data are scarce?

The techniques used in this paper are based upon our previous research [12], which includes feature-pyramid convolutional recurrent neural network (FP-CRNN), pseudo-labeling (PL), mean-teacher (MT) models, and adversarial domain adaptation (ADA) [13]. The rest of this paper is organized as follows. Section II describes the methods we adopt to detect bird sound events. Section III introduces the datasets for this research and the pre-processing steps thereof. Afterwards, experiment details, results and discussion, and conclusions will be given in Sec. IV, V, and VI, respectively.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

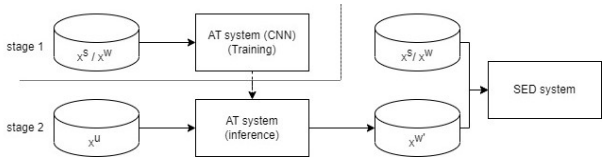


Fig. 1. Flow chart of pseudo-labeling techniques. x^{s}, x^{w}, x^{u} and $\mathrm{x}^{\mathrm{w}'}$ represent strongly-labeled, weakly-labeled, unlabeled, and pseudo weakly-labeled data, respectively. AT system is the model that needs to be trained to transform unlabeled data into pseudo weakly-labeled data.

II. Methods

In our previous work, several semi-supervised learning strategies and domain adaptation technique were employed to achieve outstanding performance in domestic SED using the DCASE 2020 Task4 dataset. In this section, we provide a thorough introduction to the model and the training strategies.

A. baseline deep learning model

FP-CRNN was chosen as our backbone model for this study. By combining the feature-pyramid component [14] with a basic CRNN model, the model is able to capture features at different scales. In detail, the FP-CRNN consists of seven CNN blocks, two layers of bidirectional gated recurrent units, a feature-pyramid component, and an attention layer which produces clip-level and frame-level predictions.

To utilize unlabeled and weakly-labeled data effectively, three semi-supervised learning strategies were adopted from our previous work — PL, mean teacher (MT) and shift-consistency mean teacher (SCMT) training, respectively. First, PL involves inferring weak labels for the originally unlabeled data by deploying an DNN-based audio-tagging (AT) system. The SED system is then trained with a combination of synthetic strongly-labeled and real weakly-labeled data. The dataflow of PL is shown in Fig. 1.

Secondly, the MT model consists of a student and a teacher model with the same structure. Parameters of the student model are updated through backpropagation, while parameters of the teacher model are set according to exponential moving average of the corresponding parameters of the student model, resulting in a slower but more stable way of updating the parameters. By calculating the mean square error (MSE) between the student and the teacher model, the consistency of the model can be enhanced.

Next, Fig. 2 illustrates an endeavor to incorporate shift consistency training (SCT) into the MT model, and the resulting system is referred to as the SCMT model. The integration can be achieved by jointly considering the MSE loss between the student model and the teacher model and a binary cross-entropy (BCE) loss between the frame-level predictions and the ground truth. The total loss function is expressed as follows,

$$\begin{aligned} L_{\mathrm{SCMT}} = & L_{wf,\mathrm{BCE}} + L_{sf,\mathrm{BCE}} + L_{st,\mathrm{BCE}} \\ & + r(t)(L'_{wt,\mathrm{MSE}} + L'_{wf,\mathrm{MSE}}) \\ & + r(t)(L'_{st,\mathrm{MSE}} + L'_{sf,\mathrm{MSE}}) \\ & + r(t)L_{st,\mathrm{MSE}} \end{aligned} \tag{1}$$

where $r(t) = e^{-5(1-t/T)^2}$ represents a ramp-up function of length T, subscripts w and s denote weakly-labeled outputs and strongly-labeled outputs, respectively, and f and t denote frequency shift and time shift, respectively.

Finally, we adopted two-stage ADA [15] technique to bridge the domain gap between synthetic data and real data. To cope with the ADA technique, the FP-CRNN was separated into a feature extractor and a label predictor. Besides, a CNN-based discriminator was trained to distinguish the input data as being synthetic or real. The following equation defines the learning objective of adversarial domain adaptation:

$$\begin{aligned} E(\theta_f, \theta_y, \theta_d) = & L_s(\theta_f, \theta_y) + L_w(\theta_f, \theta_y) \\ & - \lambda \cdot L_{D,\mathrm{BCE}}(\theta_f, \theta_y, \theta_d), \end{aligned} \tag{2}$$

where θ_f, θ_y and θ_d denote the parameters of feature extractor, label predictor and discriminator, respectively, λ denotes an adversarial loss weight, and $L_{D,\mathrm{BCE}}$ denotes the frame-level domain loss. The subscript w, s have the same meanings as in Eq. (1). Then, the following equations describe the adversarial training process,

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d), \tag{3}$$

and

$$\hat{\theta}_d = \arg\max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d). \tag{4}$$

III. Dataset

Experiments were conducted on multiple datasets. To begin with, a strongly-labeled dataset consisting of the sounds of Eastern North American (ENA) birds was retrieved. The dataset comprises 48 different bird species, totaling 385 minutes of audio recordings, and each recording is 5 minutes long. The dataset is marked with 16,052 annotations. In this research, we selected 20 bird species with more than 100 annotations as our identification targets.

Next, A strongly-labeled synthetic dataset comprising the same 20 bird species was created by mixing foreground bird calls with background environmental sounds. To obtain the foreground bird calls, audio clips of the 20 selected bird species were downloaded from xeno-canto, a website which allows bird observers to share wildlife sound recordings from around the world. Subsequently, bird calls contained in each audio clip were manually extracted in order to eliminate any silence between the calls. The duration of the resulting bird-sound segments ranges from 0.2 to 2 seconds.

Meanwhile, background environmental sound data were chosen from another bird sound dataset [16], which includes recordings captured in central and southern France and southern Spain. we specifically chose recordings without event labels to serve as our environmental sound clips. Finally, with the assistance of Scaper [17], a synthetic dataset with 3000 bird calls was generated, with fully-annotated event labels, and onset and offset times.

IV. Experiments

Selected details of the experiments are described in this section.

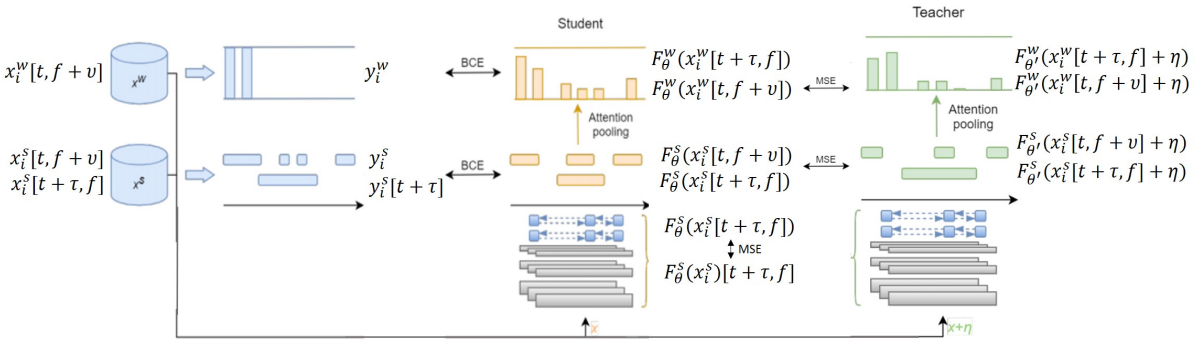


Fig. 2. the structure of shift consistency mean-teacher model. F_θ denotes a student model, $F_{\theta'}$ denotes a teacher model and the superscripts w and s denote clip-level and frame-level outputs from the model, respectively. f and t denote frequency and time, respectively; τ and ν denote the time and frequency shift, respectively. $\eta \sim \mathcal{N}(0, 0.5)$ denotes Gaussian noise that was added to the input of the teacher model.

A. Data Preprocessing and Feature Extraction

Each of the 5-minute recordings in the ENA dataset was partitioned into segments of 10 seconds with a hop length of 5 seconds. If the time interval between two events of the same class was shorter than 150 ms or if events of the same class overlapped in time, the events were merged into a single event. For each class, the shortest 10% of events were deleted unless they are 200 ms or longer. Table I summarizes the 20 bird species used in this research, along with the total number of annotations and the percentage of annotations with duration less than 200 ms after data pre-processing.

TABLE I
BIRD SPECIES IN THE ENA DATASET AFTER PRE-PROCESSING

Bird Species	Annotations	
	Total	Percentage (<0.2 sec)
AMCR	534	1.50
AMRE	274	0
BAWW	143	0
BCCH	1262	5.07
BGGN	320	1.25
BHCO	176	5.11
BHVI	174	1.72
BLJA	519	0.19
BTNW	1058	0
COYE	365	0
EATO	5238	1.18
HETH	162	0
KEWA	194	0
NOCA	884	0.23
OVEN	404	0.24
RBWO	145	4.23
REVI	592	1.18
SCTA	290	0.69
TUTI	885	1.12
WOTH	1826	0.43

Following the same settings as in our previous work [7], the audio files were first transformed into log-Mel spectrograms. The audio clips were first resampled to 32 kHz, and the window size, hop length, and the number of Mel bins were 2048, 255 and 128, respectively. Thus, the size of log-Mel spectrogram was fixed at 1255×128.

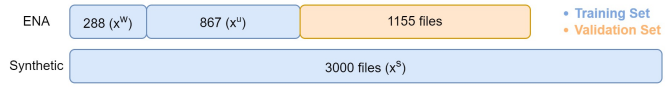


Fig. 3. Data splitting for training and validation. The superscripts w, u and s have the same meaning as in Fig. 2.

B. Other Configurations

The experiments were divided into several parts, and all of them used both the ENA dataset and the strongly-labeled synthetic dataset as training data. However, the role played by the ENA data could vary. In the first part, the ENA data was used as strongly labeled data, and the result will be reported in Sec. V.A to check the model's performance when trained on strongly-labeled and in-domain data. In subsequent parts of the experiments, we aimed to investigate whether the performance could be boosted via different training strategies when the out-of-domain synthetic data were strongly labeled while the in-domain ENA data were not. Thus, the ENA dataset was treated as unlabeled or weakly-labeled data; in other words, the class labels or the timing information were intentionally ignored. To achieve this, we split the ENA training data into two parts, 25% as weakly-labeled and 75% as unlabeled, and the data splitting scheme is illustrated in Figure 3.

To utilize the "unlabeled" ENA data, a pre-trained ResNet-based audio tagging system was fine-tuned for weak label generation [12]. This scheme is referred to as pseudo-labeling (PL) hereafter. As for implementing SCMT, the time shift and frequency shift were set to vary between -2.0 to 2.0 sec and -4 to 4 mel bins, respectively. During training, the Adam optimizer was adopted with the option of learning-rate adjustment.

V. RESULTS AND DISCUSSIONS

A. Evaluation of the baseline model on strongly-labeled ENA data and synthetic data

Table II shows the F1 scores (%) for the CRNN and FP-CRNN trained on strongly-labeled datasets. Both models were

effective to a certain extent, even when trained solely on out-of-domain synthetic data (SYN). It is worth noticing that basic CRNN model performed better than FP-CRNN when only trained on the ENA dataset. However, FP-CRNN performed better on synthetic-only and mixed dataset. The performance difference was probably due to data insufficiency in the ENA dataset for this part of the experiments. As FP-CRNN is a more complicated model than the basic CRNN, we speculate that it demands more data during the training phase to perform well.

TABLE II
F1 SCORE (%) COMPARISON OF BASELINE MODEL TRAINED WITH STRONGLY-LABELED ENA DATASET AND SYNTHETIC DATASET

Model	ENA	SYN	ENA+SYN
CRNN	30.81	13.49	39.30
FP-CRNN	22.75	14.01	43.44

B. Evaluation on models when using the ENA dataset as unlabeled or weakly-labeled data

In this part of the experiments, we simulated situations where strongly-labeled real training data were unavailable. Instead, we only had strongly-labeled synthetic data and unlabeled or weakly-labeled real data. Table III summarizes the F1 scores (%) for the situations in which all in-domain training data were unlabeled, or part of them were weakly labeled (see Fig. 3). Note that, the F1 score became higher when MT and SCMT strategies were adopted. Also, the performance further improved when accurate weak labels were provided. Note that 9 to 12% of net improvement in the three F1 scores could be obtained.

TABLE III
COMPARISON OF F1 SCORE (%) WITH DIFFERENT SEMI-SUPERVISED LEARNING (SSL) STRATEGIES

SSL Strategies	Unlabeled data	Weakly-labeled data
-	14.01	25.92
MT	16.92	26.39
SCMT	18.48	**27.84**

C. Evaluation on proposed semi-supervised learning strategies

Table IV indicates the F1(%) score before and after applying PL to the 867 unlabeled clips. Note that the performance consistently improved, and hence there seemed to be no conflict between MT or SCMT and PL.

TABLE IV
F1 SCORE(%) COMPARISON OF DIFFERENT MODELS BEFORE AND AFTER DOMAIN ADAPTATION

SSL Strategies	Before PL	After PL
-	25.92	26.34
MT	26.39	29.03
SCMT	27.84	**29.81**

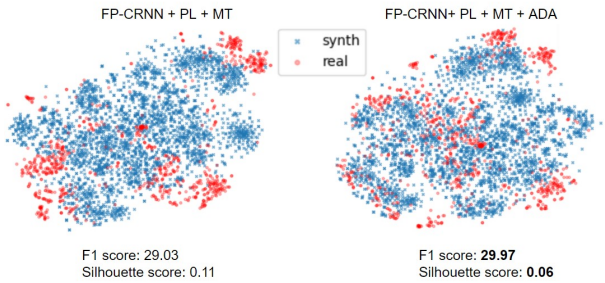


Fig. 4. Visualization of embedded features generated from a trained model before and after applying ADA .

D. Effectiveness of domain adaptation

TABLE V
F1 SCORE(%) COMPARISON BEFORE AND AFTER DOMAIN ADAPTATION

SSL Strategies	Before ADA	After ADA
-	25.92	26.97
PL	26.34	27.42
PL + MT	29.03	29.97
PL + SCMT	29.91	**30.72**

Finally, the performance of domain adaptation is summarized in Table V. Note that the F1 score consistently increased after applying the ADA technique, with the optimal score of 30.72% when the shift consistency, mean-teacher, pseudo-labeling, and adversarial adaptation strategies were all applied. Note that this F1 score approaches the CRNN score in Table II, 30.81%, that was obtained when in-domain strongly-labeled data were assumed to be available.

To quantify how well the ADA technique works, we calculated the silhouette score. Silhouette score ranges from -1 to 1, and a high score indicates high degree of separation between two clusters; in the present scenario, the two clusters are synthetic data and real data, respectively. Therefore, after we apply domain adaptation on our model, a low silhouette score is desired. Figure 4 shows a pair of t-SNE plots of the data before and after ADA was applied, and the corresponding silhouette scores are reported. Note that, in this particular combination of SSL strategies (PL+MT), the silhouette score was successfully decreased by ADA, which led to an increase of F1 score by approximately 1.0%. Thus, these results jointly show that domain adaptation can reduce the gap between synthetic data and real data, and the ADA technique could work with different combination of SSL strategies.

VI. CONCLUSIONS

In the present research, we created a synthetic dataset with 20 common bird species in the East Northern America region, and the synthetic data set was used with real data of the same bird species to test various sound event detection strategies. We demonstrate that all the strategies, including FP-CRNN, SCMT, PL, and ADA, could be applied jointly to improve the F1 score of sound event detection in this scenario, and

the improvement could be achieved whenever a new strategy is added. The present success shows that the models and training strategies previously designed for domestic sound event detection can now also be applied in an outdoor environment for bio-acoustic applications. In conclusion, through these experiments, we have shown that the model can be trained by supplying out-of-domain synthetic data and in-domain unlabeled and weakly-labeled data, and usage of these strategies is not limited to a single application scenario.

ACKNOWLEDGMENT

This research was supported by the National Science and Technology Council of Taiwan under grant No. 111-2634-F-002-023. Furthermore, we acknowledge the Xeno-canto community at www.xeno-canto.org for their valuable audio recordings used in our research.

REFERENCES

[1] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Visual object detector for cow sound event detection," *IEEE Access*, vol. 8, pp. 162 625–162 633, 2020. DOI: 10 . 1109 / ACCESS.2020.3022058.

[2] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, e0214168, 2019.

[3] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: A survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6. DOI: 10.1109/MLSP.2016.7738875.

[4] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, e488, 2014.

[5] C.-Y. Koh, J.-Y. Chang, C.-L. Tai, D.-Y. Huang, H.-H. Hsieh, and Y.-W. Liu, "Bird sound classification using convolutional neural networks.," in *CLEF (Working Notes)*, 2019.

[6] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *2017 25th European signal processing conference (EUSIPCO)*, IEEE, 2017, pp. 1744–1748.

[7] F.-C. Chen, K.-D. Chen, and Y.-W. Liu, "Domestic sound event detection by shift consistency mean-teacher training and adversarial domain adaptation," in *Proc. Int. Congress on Acoustics, Gyeongju, Korea.*, 2022.

[8] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2015, pp. 1–5.

[9] M. Crous, "Polyphonic bird sound event detection with convolutional recurrent neural networks," Ph.D. dissertation, Jul. 2019. DOI: 10.13140/RG.2.2.11943.09126.

[10] A. G. A. Parrilla and D. Stowell, *Polyphonic sound event detection for highly dense birdsong scenes*, 2022. arXiv: 2207.06349 [cs.SD].

[11] L. M. Chronister, T. A. Rhinehart, A. Place, and J. Kitzes, *An Annotated Set of Audio Recordings of Eastern North American Birds Containing Frequency, Time, and Species Information*. Wiley Online Library, 2021.

[12] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 376–380. DOI: 10.1109/ICASSP39728.2021.9414350.

[13] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[15] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2020.

[16] F. J. Bravo Sanchez, M. R. Hossain, N. B. English, and S. T. Moore, "Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[17] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2017, pp. 344–348.

A Comparative Evaluation of Video Codecs for rPPG-based Heart Rate Estimation

Muhammad H. Hyanda*[1], Nur Ahmadi*†‡§[2], Peter H. Charlton‡[3], Timothy G. Constandinou§¶, Ayu Purwarianti*†, Trio Adiono*[4]

*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, Indonesia
†Center for Artificial Intelligence (U-CoE AI-VLB), Bandung Institute of Technology, Bandung, 40132, Indonesia
‡Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, United Kingdom
§Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, UK
¶Care Research and Technology Centre, UK Dementia Research Institute at Imperial College London, London, UK
Email: [1]muhammadheronan@gmail.com, [2]nahmadi@itb.ac.id, [3]pc657@cam.ac.uk, [4]tadiono@itb.ac.id

***Abstract*—Remote photoplethysmography (rPPG) is a non-invasive technology that can be used to extract heart rate (HR) from video data. Video codecs are typically used for compressing video, which allows for efficient data storage, but could negatively impact the accuracy of HR estimation. To date, only limited number of studies investigate the impact of video codecs on rPPG-based HR estimation. In this paper, we present a comprehensive comparative evaluation of video codecs for rPPG-based HR estimation. Several factors being compared include video encoders, video containers, quality rate mode, and resolution. Experimental results using in-house dataset show that the lossy encoder like VP9 yield good performance nearly similar to the lossless counterparts. Additionally, the.avi container is preferable because it can contain all encoders tested except x265 (HEVC). Varying the resolution significantly affects the file size and accuracy of HR estimation. This study provides useful information for the creation of rPPG dataset and development of rPPG algorithm.**

I. Introduction

Heart rate (HR) monitoring is an important aspect of healthcare, sports, and fitness monitoring. Traditional HR monitoring methods involve direct contact with the skin, which can be uncomfortable and may cause skin irritation [1]. Remote photoplethysmography (rPPG) technology provides a noninvasive alternative to traditional methods by analyzing physiological signals from video data captured by cameras [2].

The use of rPPG-based HR estimation has gained significant attention in recent years [3]. To develop and evaluate rPPG algorithms, many researchers use publicly available datasets such as PURE [4], VIPL-HR [5], DEAP [6], COHFACE [7], MMSE-HR [8] and UBFC-Phys datasets [9]. These datasets differ in the way the data are stored. Some datasets are stored in sequences of PNG images with timestamps (e.g. PURE) while other datasets (e.g. VIPL-HR and UBFC-Phys) are stored in the form of videos. Large and diverse datasets are needed for fostering the development of accurate and reliable rPPG algorithms. However, creating datasets can be challenging due to the need for storing large amounts of video data and the lack of a guideline regarding the dataset storage format. In light of this, video codecs play a crucial role in reducing the size of video data and allowing efficient data storage and transmission [10]. Different video codecs use different algorithms and compression methods, resulting in varying levels of compression quality, which affects the accuracy of HR estimation. To date, however, there have been only few studies that investigate the impact of video codecs on rPPG-based heart rate estimation [10]–[12].

To address the above problem, in this paper, we present a comparative evaluation of different video codecs on rPPG-based HR estimation. Specifically, we compare the effects of video encoders, video containers, quality rate mode, and resolution to the accuracy of HR estimation. This paper aims to provide insights into the performance of different video codecs and their suitability for rPPG-based HR estimation.

The remainder of this paper is organized as follows. Section II describes the processing pipeline, video encodings, experimental setup, and performance metrics. Section III presents and discusses the impact of different video compressions, video containers, constant rate factor (CRF) values, resolution and scaling algorithms on the accuracy of rPPG-based HR estimation. Finally, conclusions are drawn in Section IV.

II. Methods

A. Processing Pipeline

The processing pipeline is illustrated in Fig. 1. It consists of three main blocks namely dataset recording, video making, and performance benchmarking. The dataset recording block is used to record in-house datasets from subjects and obtain the HR reference data using a pulse oximeter. These data will be used to make a video using several video encoders. In the performance benchmarking block, these videos will be evaluated using an rPPG algorithm to generate sequences of HR prediction and the error will be calculated using reference data recorded in the dataset recording block. The data processing pipeline is implemented in Python programming language.

The dataset recording block consists of a dataset acquisition program which receives some parameters such as camera de-

This work was supported by Bandung Institute of Technology (ITB) International Research Program 2022 (No. 4949/IT1.B07.1/TA.00/2022).



Fig. 1. The experiment processing pipeline and data flow.

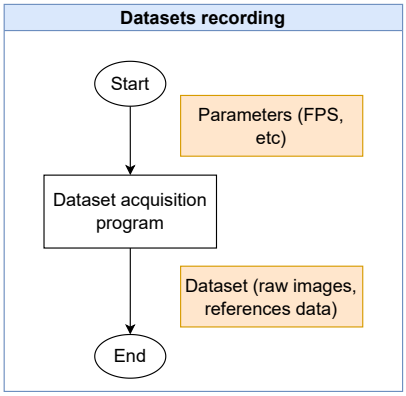


Fig. 2. Dataset recording data flow.

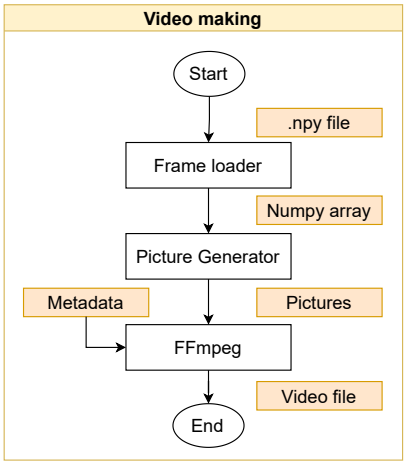


Fig. 3. Video-making data flow.

vice number, pulse oximeter address, frame rate, and duration of the dataset. After starting the program and inputting all the parameters, the program will start recording subject faces and save them as raw images contained in the NumPy array into a NumPy file format (NPY). This program also will save the reference data from the pulse oximeter in the same format. The process can be seen in Fig. 2.

The video making block consists of a frame loader and a video writing process. The frame loader unpacks the NumPy raw image file into a NumPy array which are then converted to PNG images with zero compression. After that, these images will be written into a video by using FFmpeg. The video encoder type used will be entered in this process. The process can be seen in Fig. 3.

The last block is performance benchmarking. Each video file is evaluated using an rPPG algorithm from our previous work [13]. This process produces an output in form of an

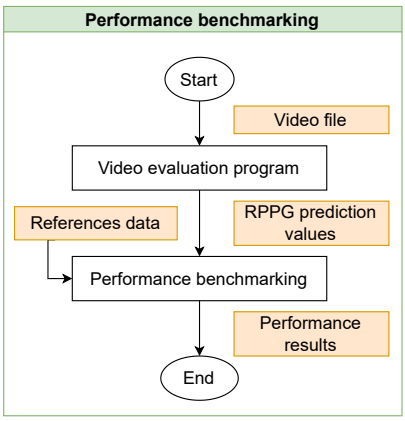


Fig. 4. Performance benchmarking data flow.

array of HR prediction values, which will proceed to the performance benchmarking sub-block. In this sub-block, mean absolute error (MAE) and root mean squared error (RMSE) will be calculated as performance metrics using the prediction and reference data array. . The process is shown in Fig. 4.

B. Video Encoding

Video encoding is a technique that allows video to be stored in much smaller size through compression because there is a lossy process involved. The process sweeps away some visual data in a video that is not really important so the visual data can be squeezed into fewer data bits. The video encoding process itself is determined by the video codecs or video compression standards.

*1) **Video Codecs**:* Video codecs are video compression standards carried out through software or hardware applications. They comprises two parts: an encoder and a decoder. The encoder is used to compress the video and the decoder is used to decode the encoded file and recreate the approximate original video. There are some publicly known video codecs such as x264, x265, FFV1 and etc.

Video codecs can be categorized into 2 categories: lossy and lossless. The key difference between these codecs is that lossless codecs try to preserve the video as it is while lossy codecs remove some part of the original data. Several popular lossless codecs used in this experiment are FFV1, HuffYUV, v410, and v210, whereas the used lossy codecs are x264, x265 (HEVC), VP9, and MPEG-4.

*2) **Video Containers**:* In this experiment, we also use some containers that comply with each codec used. Not to confuse with video codec, a video container is used to encapsulate everything such as encoded video and audio in one file. The different role between codec and container can be illustrated in Fig. 5. Some common containers were tested and used to provide complementary information in the final result.

*3) **Constant Rate Factor (CRF)**:* Constant Rate Factor (CRF) is one of the general rate control modes available in FFmpeg for x264, x265, and livbpx (VP9) encoder. By using this rate control, how many bits used in each frame

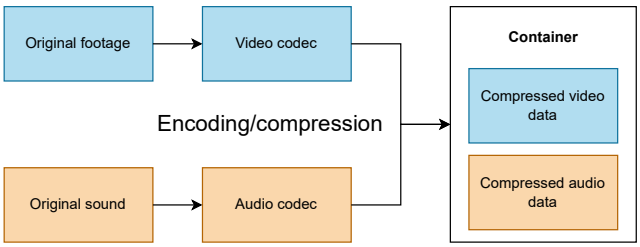


Fig. 5. Overview on video encoding process.

can be determined. By adjusting the so-called quantizer for each frame, it obtains the bitrate needed to keep the requested quality level [14]. In this experiment, the CRF range was 0 to 50 with an increment of 10.

4) ***Resolution****:* Resolution is considered an important factor to reduce the video file size. Each video was down-scaled using the FFmpeg scaler [10]. We tested different available resolutions and scaling algorithms.

C. Experimental Setup

The experimental setup consisted of a laptop equipped with a Logitech C920 camera [15] and connected to a pulse oximeter [16] to retrieve the ground truth (reference) data.

Data were recorded from 17 healthy subjects (9 males and 10 females) between the ages of 19 and 29. Each subject was sitting in a chair while facing his/her face to the camera that was 50 cm apart. Each recording was recorded for 90 seconds. All subjects gave their informed consent in accordance with the principles outlined in the Declaration of Helsinki. The ethical clearance was obtained from the Research Ethics Committee of the Bandung Institute of Technology, Indonesia (No. KEP/II/2022/X/M130723NA/PSPT).

For the downscaled video resolution data, the data from the highest resolution from each aspect ratio was downscaled to other lower resolutions. Downscaling was performed using several scaling algorithms provided by FFmpeg including fast bilinear, bilinear, bicubic, neighbor, area, gauss, lanczos, and sinc.

D. HR Estimation Algorithm

HR estimation algorithm used in this study was based on the rPPG algorithm described in [13] with modification to produce an HR estimate every 1 second. In [13], HR was estimated using all frames at the end, but in this new pipeline, the data was stored in a buffer until it was full, by which HR could then be estimated. After that, the new frames from the next 1 second were pushed into the buffer and the oldest 1 second frames in the buffer were popped out. By doing so, an HR estimate could be produced every second.

E. Performance Metrics

The performance for each video codec was measured using MAE and RMSE metrics. Performance results were displayed as a box plot, a bar plot, or a line plot. In the box plot, the horizontal line and the circle mark inside represent the median and mean, respectively. The colored solid box refers to the interquartile range (IQR) with the whisker extending to 1.5 times the IQR.

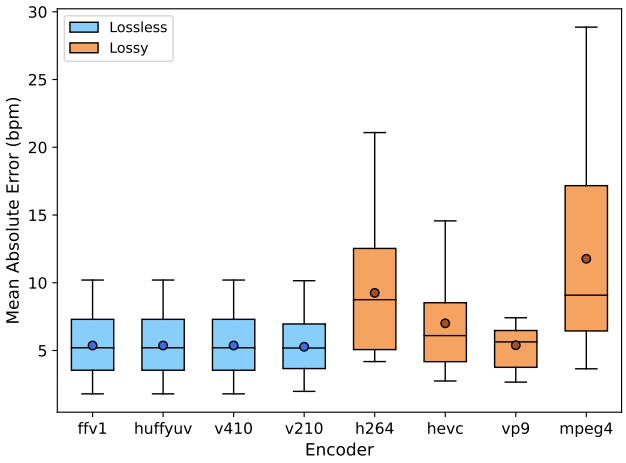


Fig. 6. Comparison between each encoder based on MAE metrics.

III. Results and Discussion

A. Lossless vs. Lossy Compression

In this experiment, for lossy compression, we used H264 or x264, H265 (HEVC) or x265, MPEG-4, and VP9. Whereas, for lossless compression, we used FFV1, HuffYUV, v410, and v210. All of these codecs are available in FFMPEG. The experimental results can be seen in Fig. 6.

According to the results, we can see that the lossless compression yields lower MAE than the lossy counterparts. FFV1, HuffYUV, v410, and v210 have better performance compared to the lossy compression encoders. This is due to the nature of lossless compression, which preserves the original bit data from the image. On the other hand, the lossy compression encoders have higher MAE because the output data is not the same as the original data. It can be caused by some factors such as inter-frame predictions, type of chroma sub-sampling, removed redundant frames, and motion compression.

Another important aspect is the reduction of file size. Lossy and lossless compressions have a big difference in the file size reduction. For this comparison, we take a sample to calculate the reduction in file size for each compression. Each file size is measured in a byte unit. The video resolution is 800 x 600 pixels. The comparison result can be seen in Fig. 7. Lossy compression can significantly reduce the file size to $0.1 - 0.3\%$ compared to the original file size. This reduction comes up with a trade-off, which is the decrease in the quality of image data due to the heavy compression algorithm used for producing output video. This trade-off heavily affects the quality of data for the rPPG algorithm which makes lossy compression having larger error as seen in Fig. 6. In this comparison, video codecs like H264, HEVC, and VP9 used the default CRF value.

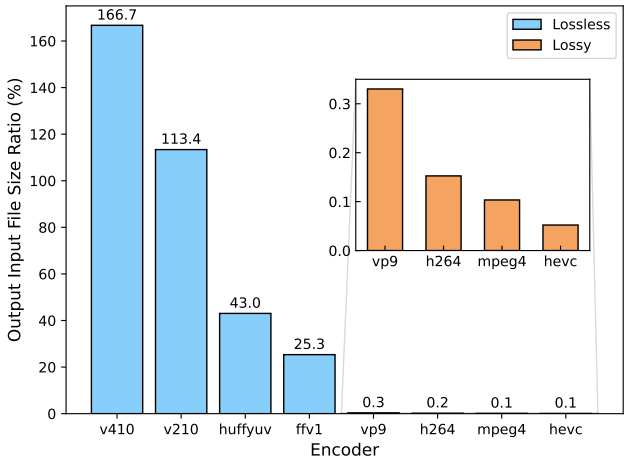


Fig. 7. File size reduction from the raw images.

On the other hand, the file size reduction using lossless was not as good as the lossy counterparts. It is due to the compression process in lossless compression is not as heavy as in lossy compression. It makes the output video file size still relatively large. We can also see that v410 and v210 resulted in larger file sizes than the raw file size. Therefore, v410 and v210 video codecs are not preferable for video compression. FFV1 produces the best result among the lossless by reducing the file size up to 25.30% and still keeps the performance metric similar to the other lossless compressions.

For the lossy compressions, the VP9 yields a relatively good file size reduction. It can compete with the lossless compressions while maintains lower variation on MAE over all subjects. The file size also could be compressed down to 0.3% size of the original file. It has been shown that lossy compression could be potentially used for video compression.

B. Effect on Different Container

As can be seen in the Table I, container does not affect the video quality. Different container is used for both lossless and lossy compression but none of them have differences. The data used for this experiment are just the top 3 subjects with lowest MAE over all codecs which causes the values slightly different from those shown in Subsection III-A.

Although there is no difference on the performance metrics on each video, different container has slightly different details. Using MediaInfo as video details checker, it is shown that .mkv has the smallest bitrate number among all these containers. For convenience, .avi can be selected to contain the video because it can be used to contain all types of encoder mentioned in this paper.

C. Effect of Constant Rate Factor (CRF)

Fig. 8(a) shows that the MAE tend to increase when the CRF value is increased. The default value for each encoder is different. For x264, 23 is the default value and for x265 (HEVC), 28 is the default CRF. Meanwhile, the default value for VP9 is 31. Near these default values, x265 and VP9 give

TABLE I
EFFECT OF DIFFERENT CONTAINERS

Encoder	Container	MAE	RMSE
FFV1	avi	2.32	2.97
	mkv	2.32	2.97
HuffYUV	avi	2.32	2.97
	mkv	2.32	2.97
v410	avi	2.32	2.97
	mkv	2.32	2.97
	mov	2.32	2.97
v210	avi	2.83	3.34
	mkv	2.83	3.34
	mov	2.83	3.34
H264	avi	5.02	6.26
	mkv	5.02	6.26
	mp4	5.02	6.26
HEVC	mkv	3.77	4.53
	mp4	3.77	4.53
MPEG4	avi	5.10	6.34
	mkv	5.10	6.34
	mp4	5.10	6.34
VP9	avi	4.60	5.54
	mkv	4.60	5.54
	mp4	4.60	5.54

better performance than x264. Overall, VP9 provides better performance than the other two encoders. It is also important to note that lowering CRF value will give higher bitrate and better quality, resulting in a lower MAE but at the same time will increase the file size, as shown in Fig. 8(b).

D. Effect of Resolution and Scaling Algorithm

Resolution tests were taken by using some common resolutions on 4:3 aspect ratio such as 640×480, 400×300, 320×240, 256×192, and 160×120. All of these resolutions were downscaled from the original video resolution of 800×600. In Fig. 9(a), it can be seen that increasing the resolution tends to reduce the MAE of HR estimation. This is because by increasing the resolution, the number of pixels contained by a video is also increased. Increasing the number of pixels means that the amount of information or the quality of the video also increases, leading to a decrease in the MAE of HR estimation. However, this comes at the expense of the increase in the file size as shown in Fig. 9(b).

IV. CONCLUSION

We have presented a comprehensive comparative evaluation of video codecs along with other relevant parameters such as video containers, quality rate mode, and resolution for rPPG based heart rate (HR) estimation. VP9 video codec provides balanced trade-off between the video quality, HR estimation accuracy, and file size. Even though video container does not affect HR estimation accuracy, .avi is preferable video container because it can contain all kinds of video encoders used in this study except x265 (HEVC). The decrease in the video resolution tends to increase the error in the HR

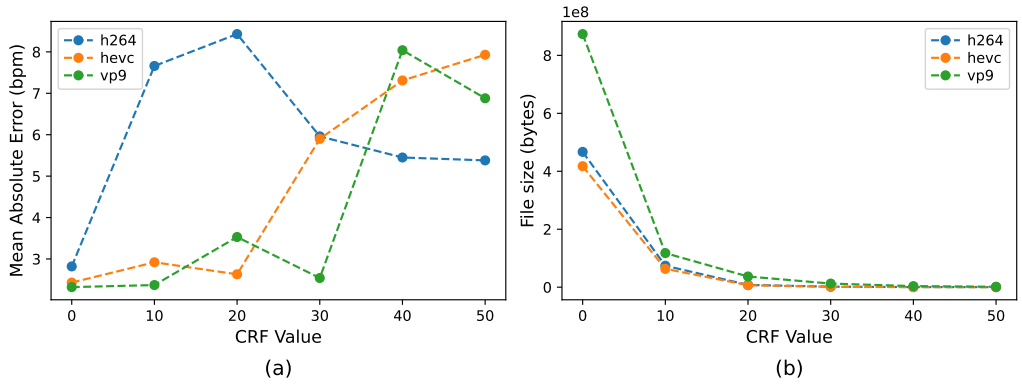


Fig. 8. Performance across different CRF value for each encoder. (a) CRF value vs. MAE. (b) CRF value vs. File size.

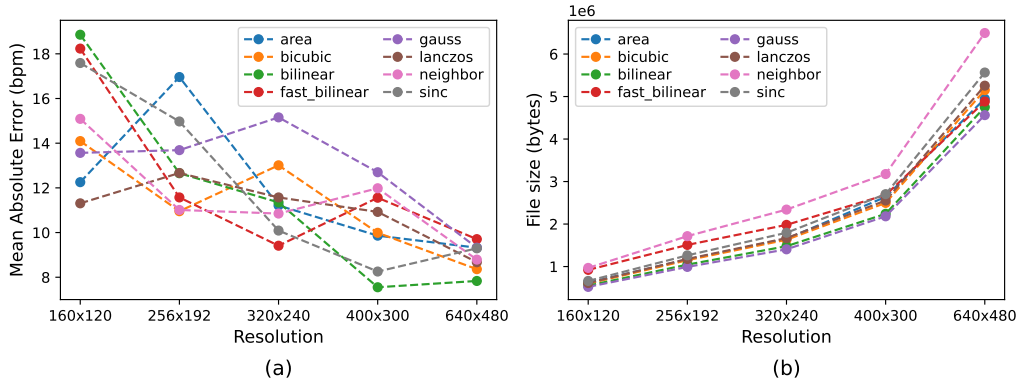


Fig. 9. Performance across different resolution with 4:3 aspect ratio. The data are down-sampled from native resolution 800×600. (a) Resolution vs. MAE. (b) Resolution vs. file size.

estimation, and vice versa. As a summary, this study could be useful for the creation of rPPG dataset and development of rPPG algorithm.

References

[1] D. McDuff, "Camera measurement of physiological vital signs," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–40, 2023.

[2] Z. Yu *et al.*, "Facial-video-based physiological signal measurement: Recent advances and affective applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 50–58, 2021.

[3] P. V. Rouast *et al.*, "Remote heart rate measurement using low-cost RGB face video: a technical literature review," *Front. Comput. Sci.*, vol. 12, no. 5, pp. 858–872, 2018.

[4] R. Stricker *et al.*, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 1056–1062.

[5] X. Niu *et al.*, "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*. Springer, 2019, pp. 562–576.

[6] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[7] G. Heusch *et al.*, "A reproducible study on remote heart rate measurement," *arXiv preprint arXiv:1709.00962*, 2017.

[8] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.

[9] R. M. Sabour *et al.*, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, 2021.

[10] M. Rapczynski *et al.*, "Effects of video encoding on camera-based heart rate estimation," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 12, pp. 3360–3370, 2019.

[11] D. J. McDuff *et al.*, "The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 63–70.

[12] E. B. Blackford and J. R. Estepp, "Effects of frame rate and image resolution on pulse rate measured using multiple camera imaging photoplethysmography," in *Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 9417. SPIE, 2015, pp. 639–652.

[13] N. Ahmadi *et al.*, "Development and evaluation of a contactless heart rate measurement device based on rppg," in *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. IEEE, 2022, pp. 1–4.

[14] FFmpeg, "Ffmpeg scaler documentation," accessed on March 30, 2023. [Online]: https://ffmpeg.org/ffmpeg-scaler.html#Scaler-Options

[15] Logitech, "C920 HD Pro Webcam," accessed on March 30, 2023. [Online]: https://www.logitech.com/en-ch/products/webcams/c920-pro-hd-webcam.960-001055.html

[16] Shanghai Berry Electronic Technology, "BM1000 Fingertip Pulse Oximeter," accessed on March 30, 2023. [Online]: https://www.shberrymed.com/products/fingertip-pulse-oximeter-bm1000b-60

Human Activity Recognition Based on FMCW Radar Using CNN and Transfer Learning

Listi Restu Triani*[1], Nur Ahmadi*†[2], Trio Adiono*[3]
*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, Indonesia
†Center for Artificial Intelligence (U-CoE AI-VLB), Bandung Institute of Technology, Bandung, 40132, Indonesia
Email: [1]listirestutriani@gmail.com, [2]nahmadi@itb.ac.id, [3]tadiono@itb.ac.id

***Abstract*—Radar-based human activity recognition (HAR) is a fascinating research topic as it can be applied in various fields such as healthcare, security, and smart homes. As a non-invasive or non-contact technology that uses electromagnetic waves to detect human activities, radar offers the advantage of protecting visual privacy and remaining resilient against environmental conditions. Common machine learning (ML) methods used for HAR, such as Support Vector Machine (SVM) and k-Nearest Neighbor (KNN), depend on complicated parameter tuning, and high-dimensional data can degrade their efficacy. We propose a unified approach based on a convolutional neural network (CNN), specifically the VGG19 architecture with transfer learning, to extract features and classify various human activities from radar data. The advantage of using CNN is the ability to integrate feature extraction and classification in one learning phase. Transfer learning accelerates learning by utilizing pre-trained models' knowledge, enabling swift adaptation to new problem domains in machine learning. We compare our proposed method against traditional ML classification methods (SVM, KNN) and the combination of CNN-based feature extraction and ML-based classification methods. Our experimental results show that our proposed method performs better than the others with an F1-score of 93%.**

I. Introduction

Human activity recognition (HAR) can be defined as the ability to recognize/detect current activity based on information received from various sensors. HAR aims to understand human activity in daily life and apply it to various fields such as health, sports, security, transportation, etc. Falling is critical to detect because it can cause severe injury and even death, especially in the elderly and people with weakened health conditions [1]. Injuries caused by falling can also affect a person's quality of life and require significant medical expenses [2]. It makes falling a significant global public health issue.

Methods for conducting HAR can be categorized into three types: based on wearable sensors [3], [4], based on cameras [5], [6], and based on radar [7]–[10]. HAR uses wearable sensors such as accelerometers, gyroscopes, and magnetometers to record human movements on the wrist, waist, or ankles. It has relatively high accuracy, but it can cause discomfort to the user. The primary constraint of wearable sensor-based HAR is that monitoring the user's activities becomes unattainable if they forget to wear the device. Another method is using a camera for HAR. It provides richer visual information than the sensor data generated by wearable devices, but it is susceptible to environmental conditions and can raise privacy concerns for the user. Radar-based HAR is a non-invasive or non-contact technology that utilizes electromagnetic waves to detect human movements, which can protect visual privacy. Additionally, radar is a simple device and is not affected by environmental conditions.

The principle of radar-based HAR involves the detection of Doppler frequency shifts caused by human movement. Doppler frequency shift occurs when there is a change in frequency or wavelength of the electromagnetic wave due to the relative motion of the target [11]. This frequency shift gives rise to the sidebands, which are additional frequency components that appear around the original transmitted frequency of the target. These sidebands further generate micro-Doppler (μD) signatures that can be used to determine dynamic features of the target [12]. In HAR, the identified target is human activity involving motion patterns of various body parts. The micro-Doppler (μD) signature can uniquely represent human actions, thus providing data that can be classified [13].

The common approach to HAR is an image-based method. The main concept involves converting the radar signal into a matrix or feature map to streamline further processing. From this feature map, the relevant information is extracted. This extraction stage usually includes two types of methods: traditional methods and deep learning methods. Traditional methods generally apply machine learning (ML) [14] such as the use of Support Vector Machined (SVM) and k-Nearest Neighbors (KNN). Deep learning (DL) methods in HAR are attracting attention due to their superiority in automated deep feature extraction and big data processing [8], [10], [11], [13]. The first application of a convolutional neural network (CNN) to interpret a time-Doppler map as a 2-D image was carried out by Kim and Moon [8].

ML algorithms like SVM and KNN have strengths but also limitations. SVM excels at complex data classification yet can be computationally intense and struggle with skewed data. KNN relies on local data distribution and is sensitive to outliers, losing efficiency as data grows. Both require careful tuning, and high-dimensional data can challenge their

This research was funded by the Indonesia Endowment Fund for Education (LPDP) graduate scholarship program (grant number KET-3118/LPDP.4/2021).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

effectiveness. Thus, considering constraints and alternatives is essential when dealing with diverse and demanding datasets. We propose a unified approach based on a convolutional neural network (CNN), especially VGG19 architecture with transfer learning to extract features and classify various human activities on radar data. We compare our proposed method against traditional ML methods (SVM, KNN) and the combination of CNN and ML methods. As one of the most successful deep learning algorithms, CNN is able to integrate feature extraction and classification in one learning phase, utilizing hierarchical convolutional layers to discern crucial visual features automatically. As a result, CNNs can extract specific local features for HAR that may differ based on the type of human activity being observed. The use of transfer learning (TL) also offers a notable advantage by accelerating the learning process through the utilization of the pre-trained model's acquired knowledge. This approach accelerates initial training stages using existing features, enabling faster model convergence and reducing the demand for extensive labeled data. This efficiency conserves resources and time and facilitates rapid adaptation of expertise to new problem domains, enhancing the development of efficient machine learning solutions.

The paper is organized as follows: In Section II, the methodology for radar-based HAR is explained, starting from data collection to the classification of human activities. Section III presents the findings of the experiments and engages in a discussion about them. Section IV, the final section, presents the conclusion.

II. Methods

A. Radar Data Collection

This study obtained data from a public repository provided by the University of Glasgow [14]. The data was collected using a Frequency Modulated Continuous Wave (FMCW) radar at the operating frequency of 5.8 GHz, a bandwidth of 400 MHz, and a 1ms chirp duration. There were six recorded activities: walking, sitting down, standing up, picking up an object, drinking water, and falling. The dataset consists of 1754 data as a vector of 1280000 complex values. The distribution data for each class activity is shown in Table I.

TABLE I: The dataset class distributions

Activity	Label	Number of data
Walking	L1	312
Sitting down	L2	312
Standing up	L3	311
Picking up an object	L4	311
Drinking water	L5	311
Falling	L6	197
Total		1754

During the radar data collection, the FMCW radar had two transmitting channels and four receiving channel antennas [15]. The radar transmitted a signal through the transmitting antenna, and then the signal reflected by an object was received by the receiving antennas. The mixer merged the pulse frequencies from the transmitter and receiver to generate the intermediate frequency (IF) signal. Subsequently, the analog-to-digital converter (ADC) in the radar equipment transformed the IF signal into a digital format for additional processing. Digital signal processing (DSP) was used to obtain the beat signal, the raw radar data matrix [16].

When an object moves towards or away from a radar system, it can cause a change in frequency in the backscattered signal due to the Doppler effect. This effect can occur when using a monostatic radar system that sends out a sinusoidal signal, as described in [17]:

$$S_T(t) = A_T \sin(2\pi f_c t) \tag{1}$$

where A_T is the amplitude of signal Tx and f_c represents the carrier frequency.

The radar signal captured from the moving target can be represented as:

$$S_R(t) = 2\rho \sin(2\pi f_c t - 2\pi f_c \tau(t)) = 2\rho \sin(2\pi f_c t + \phi(t)) \tag{2}$$

where ρ is the path loss, $\tau(t)$ is the delay due to the target's motion, and $\phi(t)$ denotes the phase due to the Doppler effect. This signal is both shifted in frequency and reduced in amplitude compared to the originally transmitted signal.

Once the signal is received, it is demodulated and converted to a baseband signal:

$$S(t) = \rho e^{-j2\pi f_c \tau(t)} = \rho e^{-\phi(t)} \tag{3}$$

The Doppler shift can be defined by the phase difference between the transmitted and received radar signals, as described in [17]:

$$f_D(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} = -f_c \frac{d\tau(t)}{dt} \tag{4}$$

B. Radar Data Processing

The radar-based HAR process algorithm is shown in Fig 1 and is implemented in Python. First, the raw data was processed into time-Doppler plots and spectrograms. The spectrogram was extracted to construct a matrix that was used as input for the machine learning (ML) classification method.

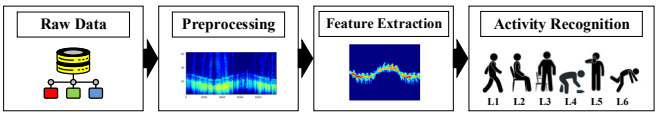


Fig. 1: The algorithm for radar-based HAR

The raw data was initially a 1D complex array and was processed through several steps. First, a fast Fourier transform (FFT) was performed on raw data to obtain the range profile in the time-range domain. Another FFT was then performed on the range profile to obtain the range-Doppler domain. Finally, a short-time Fourier transform (STFT) was performed on the range-Doppler domain to obtain the time-Doppler domain [15]. The time-Doppler domain provides information about the velocity and motion of various body parts of the target over time. When a person is in motion, the radar registers the radial

velocity of the object, and the change in the frequency of the reflected signal is referred to as the Doppler shift. This Doppler shift data contains speed details for distinct body parts, which vary as a person engages in different activities. Consequently, scrutinizing the Doppler shift enables the identification of the ongoing activity.

In the case of the 5.8 GHz FMCW radar, the wavelength can be calculated as $\lambda = \frac{c}{f} = 5.2\,\text{cm}$. Although the Doppler shift will not produce a perceivable change in frequency, there will be a considerable change in phase. The change in phase can be acquired by:

$$\Delta\phi = 2\pi f_c \Delta\tau = \frac{4\pi\Delta d}{\lambda} \tag{5}$$

As an illustration, when the scan period (T_c) is 1 ms and the velocity of the target is 10 m/s, the phase shift between each chirp will be approximately $13.8°$, whereas the range displacement during one chirp will only be 0.01 m. This indicates that the variation in phase between the peaks of two successive chirps holds the data on the target's velocity. If we substitute the relative displacement of the target as $\Delta d = vT_c$, we can obtain the velocity using the following formula:

$$\Delta\phi = \frac{4\pi v T_c}{\lambda} => v = \frac{\lambda\Delta\phi}{4\pi T_c} \tag{6}$$

C. Activity Recognition

This section involves the process of mapping the spectrum domain data obtained from radar signal processing to the features used for HAR.

1) K-Nearest Neighbors (KNN): The KNN is a simple and easy-to-implement classification algorithm, which assumes that data with the same label are normally near to each other. To classify a new data point or query, KNN calculates the distance between the query and all the training points in the dataset. The distance is usually the Euclidean distance, but other distance metrics can be used as well. Once the distances are calculated, the KNN algorithm sorts them and selects the k training points that are closest to the query [18]. The KNN algorithm was implemented using the Scikit-learn library in Python, with a specified value of $k = 5$.

2) Support Vector Machine (SVM): SVM is essentially a non-probabilistic binary linear classifier, meaning that an SVM training algorithm builds a model that assigns new examples to one category or the other [19]. SVM is a binary classification algorithm. In this case, a multiclass SVM algorithm was employed for classification using the "one vs all" strategy. This strategy involves training N different binary classifiers. Each classifier undergoes training to differentiate samples within a specific class from all other classes. When a new sample is encountered, all N classifiers are executed, and the classifier that produces the highest output is selected [20].

3) Convolutional Neural Network (CNN): CNN is a deep learning algorithm specifically designed for processing and analyzing visual data. CNN will automatically learn the hierarchical representation of visual data using multiple layers of interconnected neurons. A CNN typically comprises three primary building blocks: convolution, pooling, and fully connected layers. These layers are organized in a layered structure to extract higher-level information for classification purposes effectively [17]. Fig 2 illustrates how a simple CNN model can be visualized.

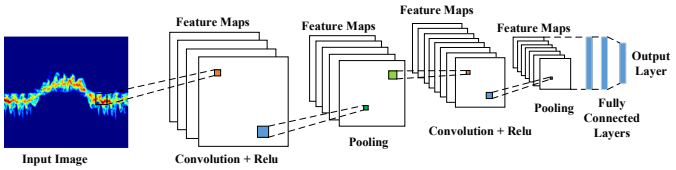


Fig. 2: An example of CNN architecture

In this study, CNN was used for pattern recognition based on micro-Doppler (μD) data that was translated into spectrogram data. The existing CNN were used such as ResNet50 [21], MobileNetV2 [22], VGG16 [23], and VGG19 [23]. To get the data into the right input format, the data was transformed to the time doppler-velocity spectrogram and rescaled to the desired 224x224 pixels.

III. Results and Discussion

This section presents and discusses our experimental results on the entire dataset. The dataset was divided randomly into three portions: 80% for training, 10% for validation, and 10% for testing. The dataset division utilized stratify feature to ensure that the proportions of each class remain balanced between the training and testing sets. We computed the weighted mean of classification accuracy, F1-score, and confusion matrix for every experiment.

A. Radar Data Processing

The dataset contains a raw data vector consisting of 1280000 complex values. This vector was converted into chirps, resulting in a 128×10000 complex-valued matrix. The matrix was then subjected to FFT to obtain range-time data. A moving target indication (MTI) filter removed unwanted signals. The filtered range-time data was then used to generate spectrograms. Fig 3 shows the spectrogram outcomes of each activity, indicating differences among all the activities.

B. Activity Recognition

In this study, we recognize human activity from radar data using the VGG19 architecture with transfer learning approach. The result of this method will be compared to traditional ML methods (SVM, KNN) and the combination of CNN and ML methods. In the early stages, a simple technique was used, which involved manual resizing and flattening of the spectrogram images to serve as input for the KNN and SVM classifiers. In addition, activity recognition using CNN through transfer learning (TL) was carried out early with different configurations such as ResNet50, MobileNetV2, VGG19, and VGG16. The feature representations were flattened and sequentially propagated through four dense layers. The final output was obtained using a softmax activation layer. In the

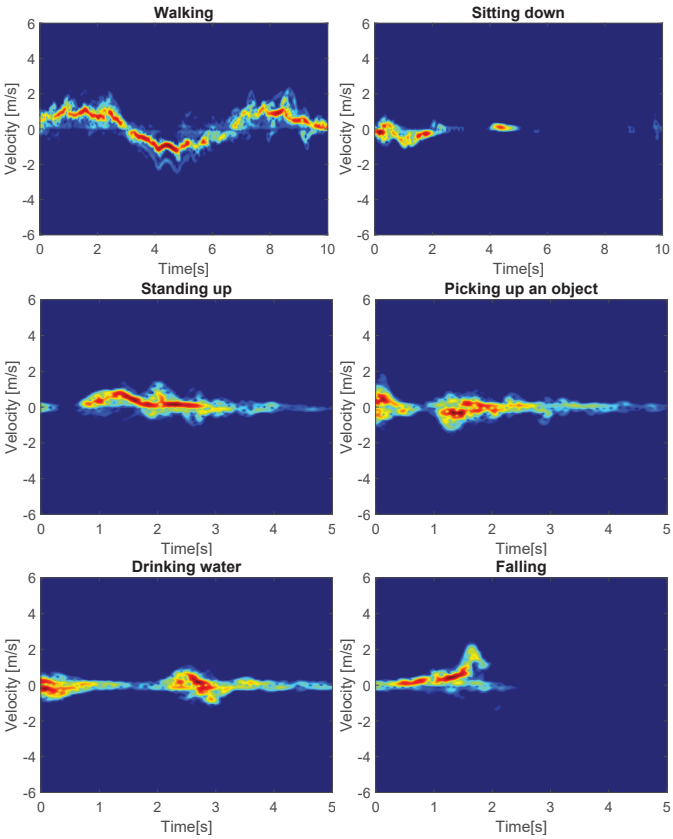


Fig. 3: Spectrogram of six human activities

final phase, activity recognition was performed using CNN as the feature extraction method and KNN and SVM as the classifiers. Fig 4 visualizes the activity recognition. The evaluation results for the different methods are presented in Table II, which shows the performance of each approach.

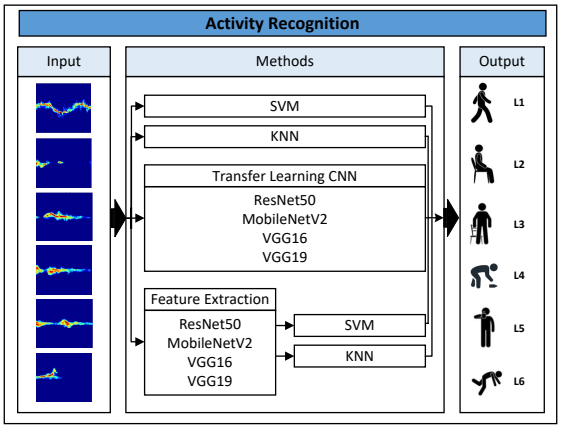


Fig. 4: Visualization of radar-based HAR

In the initial stage, the simple KNN configuration shows that the classification speed decreases with the increasing number of datasets used. The accuracy obtained with the parameter $k = 5$ is 89%. To estimate the accuracy of the KNN

TABLE II: Evaluation on different methods

Feature Extraction	Methods	Train Acc(%)	Test Acc(%)	F1 Score(%)
-	SVM	92	89	89
-	KNN	90	90	89
ResNet50-TL		80	86	86
MobileNetV2-TL		88	89	89
VGG16-TL		88	90	90
VGG19-TL		**90**	**93**	**93**
ResNet50	SVM	89	89	89
MobileNetV2	SVM	90	93	93
VGG16	SVM	91	92	92
VGG19	**SVM**	**90**	**92**	**93**
ResNet50	KNN	83	84	84
MobileNetV2	KNN	79	80	81
VGG16	KNN	88	85	85
VGG19	KNN	86	86	86

algorithm for the different configurations of features applied, the K-fold cross-validation approach was applied. The simple SVM configuration also provides the same accuracy value of 89%. Furthermore, activity recognition is carried out using deep learning called CNN with several configurations. The configurations used are ResNet50, MobileNetV2, VGG16, and VGG19. To train the CNN, the optimizer used was Adam, with a learning rate of 0.001. Sparse Categorical Cross Entropy was selected as the loss function since only one value at the output represents a classification label. The network was trained for 30 epochs with early stopping. At the end layer of the pre-trained model, we customize it by adding several dense layers. The accuracy results obtained sequentially were 86%, 89%, 90%, and 93%. Good results were obtained from the utilized configuration. The acquisition of activity recognition accuracy with KNN, SVM, and CNN shows that CNN is better at activity recognition. The confusion matrix in Fig 5 shows the best configuration performance of the CNN algorithm with VGG19 architecture.

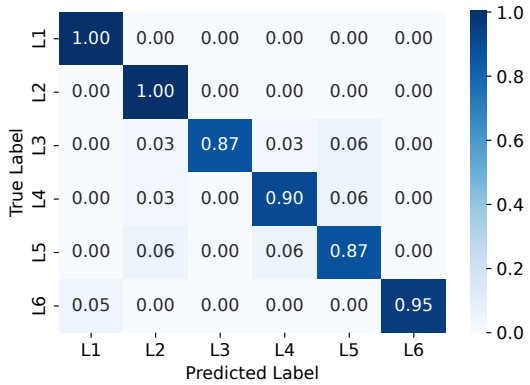


Fig. 5: Confusion Matrix VGG19-TL

Feature extraction uses CNN to speed up the model classification training time and get better results. Furthermore, the feature extraction results are input to the KNN and SVM

classifiers. Various configurations were used to evaluate the accuracy of the KNN algorithm, and the best accuracy of 86% was achieved by using VGG19 as feature extraction. In this case, the value of $k = 5$ is used to achieve optimal results. As previously stated, SVM is a binary classifier. However, HAR requires multiclass classification, so the "one versus all" strategy is used. Various configurations are used to assess the accuracy of the SVM algorithm. The confusion matrix in Fig. 6 shows the performance of the VGG19-SVM algorithm. Compared to the previous case, SVM gives superior results to KNN. Based on Table II, the comparison results of various methods indicate that both the VGG19-SVM and VGG19-TL models yield similar performance, with the highest accuracy of up to 93%. However, VGG19-TL holds an advantage, wherein feature extraction and classification are executed within a unified approach rather than separately. Another benefit of VGG19-TL is its potential for optimization to run on embedded devices or hardware accelerators through techniques like pruning and quantization. However, classifiers for L4 and L5 have the lowest accuracy.

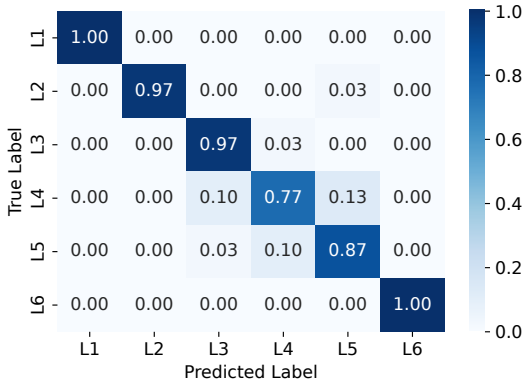


Fig. 6: Confusion Matrix VGG19-SVM

Additionally, a comparison is conducted between the proposed method and the reference method. The comparison reveals that the CNN (VGG19-TL) and the CNN-SVM methods with MobileNetV2 and VGG19 feature extraction exhibit superior accuracy compared to the other methods and study by Shah et al [14]. Shah et al. proposed GoogleNet, AlexNet+KNN, and AlexNet+SVM with performances of 75%, 77%, and 78%, respectively.

C. Future Works

Generally, radar-based HAR is implemented on the cloud or desktop. The future works will address the implementation of radar-based HAR on embedded systems because it offers several benefits, such as portability, data privacy, and low power consumption. The challenges for its implementation include real-time data processing, adaptation to changing environments, and using technology capable of processing radar-generated data. Therefore, the development of radar-based HAR on embedded systems still requires further research.

IV. Conclusion

This paper proposes radar-based human activity recognition (HAR) algorithm with a unified approach based on a convolutional neural network (CNN), especially VGG19 architecture with transfer learning. It presents its performance benchmark against other methods on a public dataset of people performing six daily activities. We compare our proposed method against traditional ML methods (SVM, KNN) and the combination of CNN and ML methods. The basic configuration for KNN and SVM achieves only about 89% accuracy. The CNN method achieves higher accuracy than KNN and SVM of 93%. Feature extraction is also needed to speed up the classification model training time and get better results. The KNN approach has the highest accuracy of 86%, with VGG19 as feature extraction. Furthermore, the SVM model using the "one vs. all" strategy was implemented for multi-class classification, achieving an accuracy of about 93% with MobileNetV2 and VGG19 as feature extraction. The VGG19-TL model produces the best accuracy of all the models performed. The VGG19-TL offers the advantage of unified feature extraction and classification alongside the potential for optimization to operate efficiently on embedded devices or hardware accelerators via techniques such as pruning and quantization.

References

[1] A. Chelli and M. Pätzold, "A machine learning approach for fall detection and daily living activity recognition," *IEEE Access*, vol. 7, pp. 38 670–38 687, 2019.

[2] M. Montero-Odasso, N. van der Velde, F. C. Martin, M. Petrovic, M. P. Tan, J. Ryg, S. Aguilar-Navarro, N. B. Alexander, C. Becker, H. Blain *et al.*, "World guidelines for falls prevention and management for older adults: a global initiative," *Age and ageing*, vol. 51, no. 9, p. afac205, 2022.

[3] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 029–13 040, 2021.

[4] G. Orengo, A. Lagati, and G. Saggio, "Modeling wearable bend sensor behavior for human motion capture," *IEEE Sensors Journal*, vol. 14, no. 7, pp. 2307–2316, 2014.

[5] Y.-G. Jiang, Q. Dai, W. Liu, X. Xue, and C.-W. Ngo, "Human action recognition in unconstrained videos by explicit motion modeling," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3781–3795, 2015.

[6] M.-C. Hu, C.-W. Chen, W.-H. Cheng, C.-H. Chang, J.-H. Lai, and J.-L. Wu, "Real-time human movement retrieval and assessment with kinect sensor," *IEEE transactions on cybernetics*, vol. 45, no. 4, pp. 742–753, 2014.

[7] H. T. Le, S. L. Phung, and A. Bouzerdoum, "Human gait recognition with micro-doppler radar and deep autoencoder," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3347–3352.

[8] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE geoscience and remote sensing letters*, vol. 13, no. 1, pp. 8–12, 2015.

[9] M. Wang, G. Cui, X. Yang, and L. Kong, "Human body and limb motion recognition via stacked gated recurrent units network," *IET Radar, Sonar & Navigation*, vol. 12, no. 9, pp. 1046–1051, 2018.

[10] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[11] Y. Lin, J. Le Kernec, S. Yang, F. Fioranelli, O. Romain, and Z. Zhao, "Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed with random forests," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9669–9681, 2018.

[12] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and electronic systems*, vol. 42, no. 1, pp. 2–21, 2006.
[13] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Personnel recognition and gait classification based on multistatic micro-doppler signatures using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 669–673, 2018.
[14] S. A. Shah and F. Fioranelli, "Human activity recognition: Preliminary results for dataset portability using fmcw radar," in *2019 international radar conference (RADAR)*. IEEE, 2019, pp. 1–4.
[15] D. F. Fioranelli, D. S. A. Shah, H. Li1, A. Shrestha, D. S. Yang, and D. J. L. Kernec, "Radar sensing for healthcare: Associate editor francesco fioranelli on the applications of radar in monitoring vital signs and recognising human activity patterns," *Electronics Letters*, vol. 55, no. 19, pp. 1022–1024, 2019.
[16] W. Ding, X. Guo, and G. Wang, "Radar-based human activity recognition using hybrid neural network model with multidomain fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 2889–2898, 2021.
[17] F. J. Abdu, Y. Zhang, and Z. Deng, "Activity classification based on feature fusion of fmcw radar human motion micro-doppler signatures," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 8648–8662, 2022.
[18] H. Son, J. W. Lim, S. Park, B. Park, J. Han, H. B. Kim, M. C. Lee, K.-J. Jang, G. Kim, and J. H. Chung, "A machine learning approach for the classification of falls and activities of daily living in agricultural workers," *IEEE Access*, vol. 10, pp. 77 418–77 431, 2022.
[19] T. D. Bufler and R. M. Narayanan, "Radar classification of indoor targets using support vector machines," *IET Radar, Sonar & Navigation*, vol. 10, no. 8, pp. 1468–1476, 2016.
[20] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *The Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

DQN Algorithm Design for Fast Efficient Shortest Path System

A Sumarudin*, Nana Sutisna*, Infall Syafalni*, Bambang Riyanto Trilaksono*, Trio Adiono*
* School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Indonesia
E-mail: 33222033@std.stei.itb.ac.id

***Abstract*—Reinforcement Learning is an algorithm with decision-making based on the Markov Decision Process (MDP). MDP uses actions that are assessed by rewards to achieve the goal depending on the environment. An action is taken by a policy. In this work, we consider a maze-based shortest path problem. The maze problem is solved based on the value of Q from the chosen action. The ordinary Q-learning algorithm uses these value-based states in two modes; explore and exploit. However, the continuous states require initialization. The problem appears in very large states. Thus, in this work, deep learning in the form of neural-network is embedded to determine the Q-value called by Deep Q-Network (DQN). The neural network architecture has a different number of layers and neurons for each implementation. Moreover, the implementation of agents on edges has limited resources. Thus, it requires an optimal architecture to achieve the goal. This study uses an environment maze with a size of 20×20 with four actions; up, down, left, and right to reach the destination state or goal. In this paper, we study how to get the optimum processing elements (PE) to be implemented in limited resources of edge computing. Experimental results show that the minimum number of layers of three layers with 14 hidden neurons, and 200 deep replay memory from DQN architecture have been able to complete the shortest path on a maze with a size of 20×20. The finding effectively minimizes the amount of computing and potential for edge computing applications.**

I. Introduction

Reinforcement Learning Algorithm is a sequential-based machine learning using the Markov Decision Process (MDP) to solve the problem without prior knowledge [1]. The algorithm requires a large computing resource to do exploration and exploitation while calculating the Q-Value (Quality). The Q-value will determine the action of the agent toward the goal. This becomes an issue when it is implemented with limited resources in edge computing [2].

The process starts by an agent takes action in the environment then the environment produces a state. Then, the state will be fed to a reward function, and the result is the reward (of the action). Moreover, based on the reward, the agent will take the next action to the environment. The agent needs to determine the best action using a reinforcement learning algorithm, which has value base and policy base algorithms. The best action is the one that gains the highest rewards.

The Deep Q-Network (DQN) algorithm implementations for finding optimal paths in robot navigation is resulted a better performances than A* and Dijkstra algorithms [3]. An application of DQN to finding the best path, utilizing guidance from A*, has been employed in an Automated Guided Vehicle (AGV) system for a Robotic Mobile Fulfillment System (RMFS). This application resulted a 71.51 % reduction in decision-making time for the complex model design of RMFS [4]. Thus, the DQN algorithm provides a promising solution for determining the best path in robot navigation. Edge computing has limited resources. In this specific case, the edge is designed to implement the Q-Leaning algorithm [2] despite the limitations it has. Thus, to increase the computational speed, the Reinforcement Learning (RL) algorithm is applied to an accelerator to provide a significant improvement in speed while reducing power consumption [5]. For instance, the implementation of deep Q-Learning in an Atari Game requires a relatively large amount of memory and a long period of consumption for training the network model. [6]. There are several approaches to expedite computational timing, including the utilization of Graphic Processing Units (GPUs), Tensor Processing Units (TPUs), and accelerators. [2]. The implementation of accelerators in this case has been proven to increase computational capacity while maintaining significantly lower power consumption compared to GPUs and TPUs. [7]–[11]. In order to ascertain the required calculation power for edge computing implementing DQN, it is imperative to initially outline the precise algorithm and artificial neural network (ANN) model. This process will involve minimizing the count of Processing Elements (PE) within the Computer System. The design will yield a more efficient model, considering the number of neurons and replay memory employed for problem-solving [12].

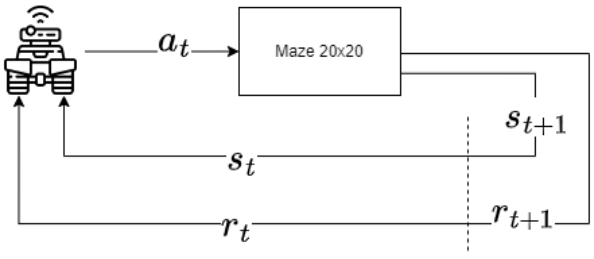


Fig. 1: Reinforcement Learning

The objective of this research is to identify the optimal

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

model by determining the appropriate number of neurons for the implementation of DQN in solving maze environments (Fig. 1). The experiments will investigate different options for the number of neurons, within 8, 10, 12, 14, 16, and 32, in order to achieve an optimal solution. Additionally, the study will encompass varying depths of replay memory options, specifically 50, 100, 150, and 200 for Deep Replay Memory (D). The outcomes of this study will play a crucial role in determining the required number of Processing Elements (PE) and on-chip memory necessary for the design of computing on Edge Computing.

Algorithm 1 Deep Q-Network Algorithm

Input: Hyper-Parameter
Output: θ

Initialize Hyper-Parameter
Initialisation Replay Memory Deep $[s_t, a_t, r_t, s_{t+1}]$
Initialisation action-Value function Q and T-Network
random weight and bias Network
while episode **do**
  Initialize T, T_n, cumulative reward, S_t
  Replay Memory **reset**
  while $t < T$ **do**
    Choose Action with Epsilon Greedy
    Observe Reward
    Calculate cumulative Reward equation (7)
    Convergence Check
    if $(s_t \neq goal)$ **then**
      $Y_i = r_t + \gamma \times \max Q(s_{t+1}, a_{t+1}|\theta)$
    else
      $Y_i = r_t$
      End Calculate Cumulative Reward
      Increment Success Rate
    end if
    Calculate Loss with Equation 9
    Store Replay Memory until full
    Calculate Stochastic Gradient Descent Randomly Replay Memory if Replay Memory Full
    Push Replay Memory
    update $\theta \leftarrow \theta^{'}$ every T_n
  end while
end while

II. DQN Algorithm

In this section, we explain our proposed method; the DQN for maze-based shortest path solver and the environment where the research is taking place.

A. Deep Q-Network Algorithm for Maze Solving

The algorithm developed in this study is the Deep Q-Network (DQN), which computes an Approximate Q-Value using a Neural Network (Alg. 1). This process is depicted in algorithm 1 based on [6], The replay memory is exclusively employed within each episode and is reset at the commencement of the subsequent episode. The algorithm is implemented for robot agents to ascertain the shortest paths in a 20 $\times$ 20 maze environment.

B. Environment of the Maze

The environment employed in this study features a maze with dimensions of 20 $\times$ 20, as illustrated in Figure 2. The neural network takes a state input of 400 neurons, and the objective of the agent is to navigate toward the destination state by identifying the shortest path. The maze used in a 20$\times$20 environment is of a complexity level suitable for maze solving. If a smaller maze is used, the computational time complexity decreases, and the memory requirements also decrease, and vice versa.

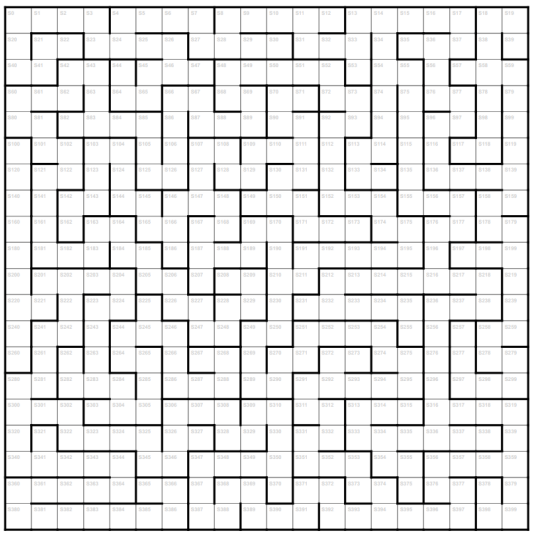


Fig. 2: Environment

III. Computational System Design for DQN Algorithm

A. Neural Network Design

The network design which is employed in the Deep Q-Network (DQN) utilizes a structure comprising three layers: the input layer, the hidden layer, and the output layer. The size of the weights for each of these layers is illustrated by Equation 1 and 2.

$$w^2_{i,j} = X_i \times Z2 \tag{1}$$

$$w^3_{i,j} = Z2 \times Z3 \tag{2}$$

The size of the bias value is determined based on the number of hidden layers and output layers. The quantities of weights and biases in the neural network model are presented in Table I.

In the computation of the Deep Q-Network (DQN), the input processing is based on Equation 3 utilizing one-hot encoding. The feed-forward equation is derived from Equation 4, it is involving a computational process of weight multiplication with neurons and the addition of input with the bias associated with each perception. The chosen activation function is the sigmoid

TABLE I: Size of Weight and Bias in DQN Neural Network

	Neuron					
Parameter	8	10	12	14	16	32
#Weight Layer 2	3,200	4,000	4,800	5,600	6,400	12,800
#Bias Layer 2	8	10	12	14	16	32
#Weight Layer 3	32	40	48	56	64	128
#Bias Layer 3	4	4	4	4	4	4
#Number of WB	3,244	4,054	4,864	5,674	6,484	12,964

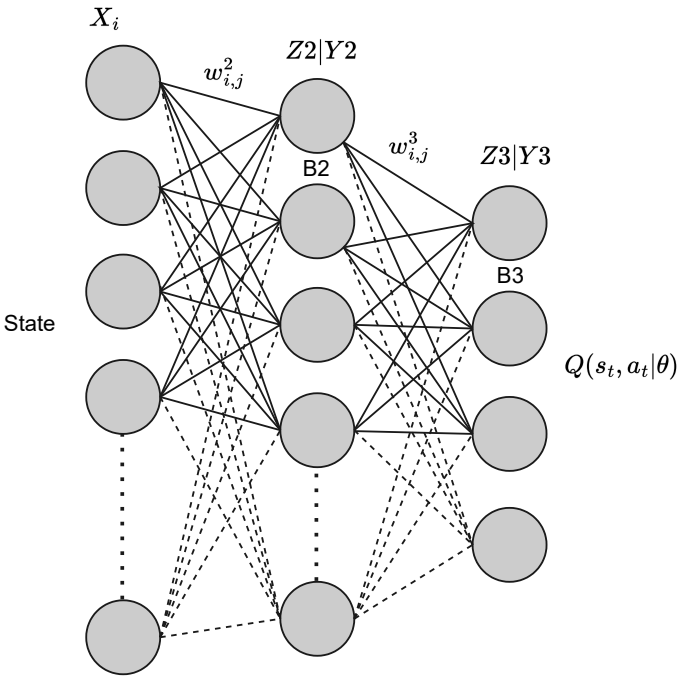


Fig. 3: Network Design for DQN

function, as depicted in Equation 5. The reward function is established according to Equation 6, where the agent receives a reward of 10 upon reaching the goal, -10 for being hit with a wall, and -1 for moving without encountering obstacles. The cumulative reward calculation is expressed in Equation 7. The algorithm's performance is assessed using the Success Rate parameter, which signifies the agent's achievement in reaching the goal within a specific time frame. In this experimental context, 1000 episodes are used, as shown in Equation 8.

$$a_j^1 = \begin{cases} 1, & \text{if } s_t = j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$y_j^m = f^m\left(\sum_{1=1}^{I_m}(w_{i,j}^m \times y_i^{m-1}) + b_j^m\right) \tag{4}$$

$$f(x) = \frac{1}{1+e^{-x}} \tag{5}$$

$$r_t = \begin{cases} 10, & \text{if Agent reaching the Goal} \\ -10, & \text{if Agent Hits the Wall} \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

$$R = \sum_{t=1}^{E} \times r_t \tag{7}$$

$$S_r = \frac{E_{goal}}{E_p} \times 100\% \tag{8}$$

After the Neural Network calculates the feed-forward process, the Deep Q-Network computes the Q-Value State's reward using Equation 10 to obtain the target value. The calculation of the Loss Function, utilizing MSE (Mean Squared Error), is depicted in Equation 9.

$$L(\theta) = \frac{[r_t^{'} - Q(s_t, a_t : \theta)]^2}{2} \tag{9}$$

$$r_t^{'} = r_t + \gamma \times \max Q(s_{t+1}, a_{t+1}|\theta) \tag{10}$$

To ascertain the weight and bias updates, a back-propagation process is employed, utilizing a chain rule (Eq. 11) in line with the stochastic gradient descent (SGD) approach (Eq. 12 and Eq. 13), incorporating replay memory $[s_t, a_t, r_t, s_{t+1}]$.

$$\frac{\delta L}{\delta w_{i,j}^j} = \frac{\delta L}{\delta z_{i,j}^j} \times \frac{\delta z_{i,j}}{\delta w_{i,j}^j} \tag{11}$$

$$w_{i,j}^{'m} = w_{i,j}^m - \alpha \Delta w_{i,j}^m \tag{12}$$

$$b_i^{'m} = w_{i,j}^m - \alpha \Delta b_{i,j}^m \tag{13}$$

The computation for stochastic gradient descent (SGD) involves calculating of derivatives through back-propagation, leading to weight and bias updates for layer 3, as demonstrated in Equations 14 and 15.

$$\Delta w_{i,j}^3 = Y_i^3 - r_t^{'} \times Y_{i,j}^2 \tag{14}$$

$$\Delta b_i^3 = Y_i^3 - r_t^{'} \tag{15}$$

The weight and bias updates for layer 2 are presented in Equations 16 and 17.

$$\Delta w_{i,j}^2 = \Delta w_{i,j}^3 \times w_{i,j}^3 \times (1 - Y_{i,j}^2) \times Y_j^2 \tag{16}$$

$$\Delta b_i^2 = \Delta w_{i,j}^3 \times w_{i,j}^3 \times (1 - Y_{i,j}^2) \times Y_{i,j}^2 \tag{17}$$

Following Equations 12 and 13, the subsequent step involves the back-propagation process, utilizing Equations 18 and 19.

$$w_{i,j}^{'3} = w_{i,j}^3 - \alpha \times Y_{i,j}^3 - (r_t + \gamma \times maxQ(s_{t+1,a_{t+1}|\theta}) \times Y_i^2 \tag{18}$$

$$b_i^{'3} = b_i^3 - \alpha \times Y_{i,j}^3 - (r_t + \gamma \times maxQ(s_{t+1,a_{t+1}|\theta}) \tag{19}$$

Furthermore, to update the weights and biases in layer 2, we use Equations 12 and 13. These equations are substitutions derived from Equations 16 and 17, and they are demonstrated in Equations 20 and 21.

$$w_{i,j}^{'2} = w_{i,j}^2 - \alpha \times \Delta w_{i,j}^3 \times w_{i,j}^3 \times (1 - Y_{i,j}^2) \times Y_{i,j}^2 \tag{20}$$

$$b_i^{'2} = b_i^2 - \alpha \times \Delta w_{i,j}^3 \times w_{i,j}^3 \times (1 - Y_{i,j}^2) \times Y_{i,j}^2 \tag{21}$$

B. DQN Hyper-Parameter Design

DQN training necessitates the initialization of parameters to ascertain the values shown in Table II for achieving an effective number of neurons in the experiment. Additionally, Table III is utilized to determine the size of the replay memory, which stores the agent's experiences $[s_t, a_t, r_t, s_{t+1}]$ in the environment for training the algorithm.

TABLE II: DQN Hyper-Parameter With Variable Number of Hidden Neurons

Parameters	Value
Number of Layer	3
Number of State Units	**400**
Number of hidden units	**8, 10, 12, 14, 16, and 32**
Number of Output Layer units	**4**
Activation Functions	Sigmoid/Linear
Back Propagation	Stochastic Gradient Descent
α	0.1625
γ	0.85
ϵ	0.9
Episode	1×10^6

TABLE III: DQN Hyper-Parameter with Variable Replay Memory

Parameters	Value
Number of Layer	3
Number of hidden units	14
Replay Memory	**50, 100, 150 and 200**
Activation Functions	Sigmoid/Linear
Back Propagation	Stochastic Gradient Descent
α	0.1625
γ	0.85
ϵ	0.9
Episode	1×10^6

C. Hardware for Training and Implmentation of DQN

This DQN algorithm model is executed on a computer equipped with an Intel Xeon Gold 5218 processor, featuring 16 cores operating at 2.294 GHz, and 128GB of RAM.

IV. Experiment Results

The training results are illustrated in Figure 4, depicting 50 paths from the initial state (s0) to the goal state (s399). Algorithm 1 successfully determines the shortest path for a 20x20 maze using the DQN approach.

Based on the experimental results, the success rate appears to converge around episode 250×10^3, as illustrated in Figure 6. The convergence of computation time requires approximately 1×10^4 to 2×10^4 seconds, as shown in Figure 7. The experimental findings reveal that a hidden neuron count of 14 is necessary to achieve a 100% convergence success rate, depicted in Figure 6.

When performing computational operations within the DQN algorithm, the determination of the number of Processing

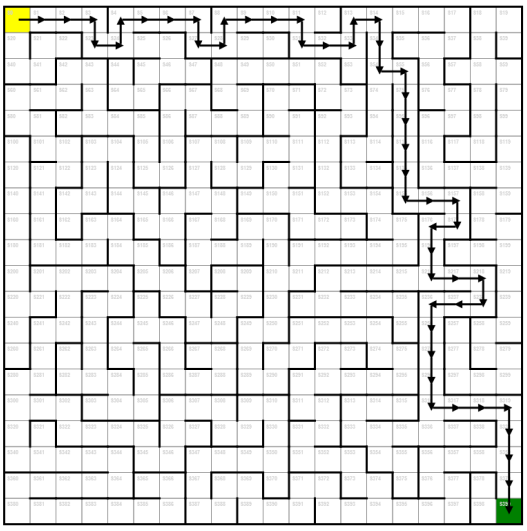


Fig. 4: Shortest Path of Agent in a Maze

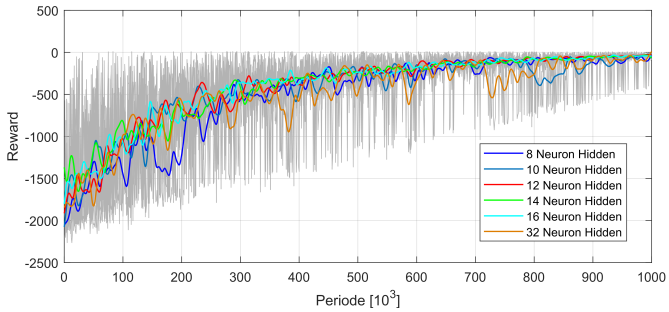


Fig. 5: Cumulative Reward of DQN against Neuron Size in Hidden Layer

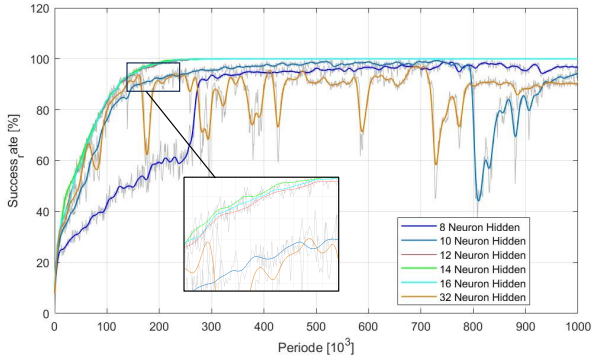


Fig. 6: Success Rate of DQN against Neuron Size in Hidden Layer

TABLE IV: Processing Elements Utilized by DQN for Feed Forward in Each Episode

	Neuron					
Parameter	8	10	12	14	16	32
#T per-episode	400	400	400	400	400	400
#Adder	1,369,600	1,712,000	2,054,400	2,396,800	2,739,200	5,478,400
#Multiplier	2,627,200	3,284,000	3,940,800	4,597,600	5,254,400	10,508,800
#Division	163,200	204,000	244,800	285,600	326,400	652,800
#Overall PE	4,160,000	5,200,000	6,240,000	7,280,000	8,320,000	16,640,000

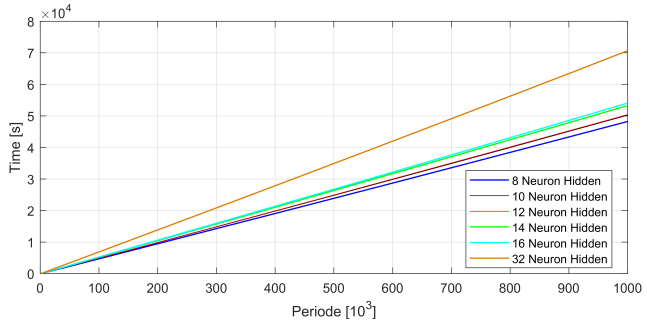


Fig. 7: Computation Time of DQN with neuron size in Hidden layer

Elements (PE) is also carried out during the feed-forward process. This calculation is based on equations 4 and 5, involving a count of the neurons without the Sigmoid activation function. In this context, **X** represents the input neurons, **Z2** denotes the hidden layer, and **Z3** corresponds to the output layer. The computation involving adders is employed in bias calculations for the hidden layer (**B2**) and the output layer (**B3**).

$$PE(MLP)_{mul} = (X \times Z2) + (Z2 \times Z3) \quad (22)$$

$$PE(MLP)_{add} = ((X-1)\times Z2)+((Z2-1)\times Z3)+B2+B3 \quad (23)$$

Based on the hyper-parameters in Table II, where each episode comprises 400 steps, the effective number of Processing Units (**PE**) required for 250×10^3 Episodes (**E**) and 400 steps per episode (**T**) determines the necessary computational resources. The computation of the activation function in Equation 5 involves employing the *numpy* exponential function [13], utilizing a computation process based on the exponential function [14] and employing a power series approach as outlined in Equation 24. This process requires n-1 adders, $2 \times (n-2)$ multipliers, and $N-2$ divisions. In addition to the Sigmoid Activation Function described in Equation 5, one extra division and one additional adder are necessary.

$$e^x = \sum_{n}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ..., \quad (24)$$

The computation involved in the backpropagation process is determined by equations 18, 19, 20, and 21. For updating W3, the number of Processing Elements (PE) requires is 2 subtractors, 1 adder, and 3 multipliers. For updating B3, the requirement is 2 subtractors, 1 adder, and 3 multipliers. Updating W2 and B2 requires 2 subtractors and 4 multipliers.

$$PE(sigmod)_{add} = (Z2 \times (N-1)) + 1 \quad (25)$$

$$PE(sigmod)_{mul} = (2 \times (N-2)) \times Z2 \quad (26)$$

$$PE(sigmod)_{div} = (Z2 \times (N-2)) + Z2 \quad (27)$$

The number of Processing Elements (PE) in the feed-forward process (Eq. 28) is the result of adding together the summation of weights and input neurons, along with all the bias in the Sigmoid function, within layer 2.

$$PE(FeedForward) = PE(MLP) + PE(sigmoid) \quad (28)$$

The number of Processing Elements (PE) used in the feed-forward process reached convergence by the episode indicated in Equation 29.

$$PE(FeedForward)_{conv} = PE(FeedForward) \times E \times T \quad (29)$$

The experimental results for various sizes of replay memory (50, 100, 150, and 200) in Deep Replay Memory are depicted in Figure 8 for cumulative reward and Figure 9 for success rate at each replay memory size.

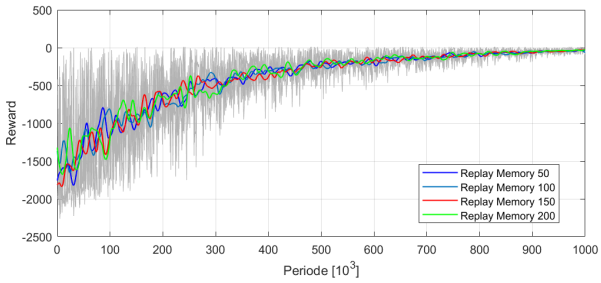


Fig. 8: Cumulative Reward of DQN with Different Replay Memory Sizes

For a replay memory size of 200, the details are provided in Table VI. Regarding the computation for 250×10^3 Episodes, the back-propagation computation time is illustrated in Equation 30, with the number of computations being $T-D$ per episode, where T represents the steps per episode and D signifies the depth of the replay memory.

$$PE(BP)_{convergen} = PE_{BP} \times (T-D) \times Episode \quad (30)$$

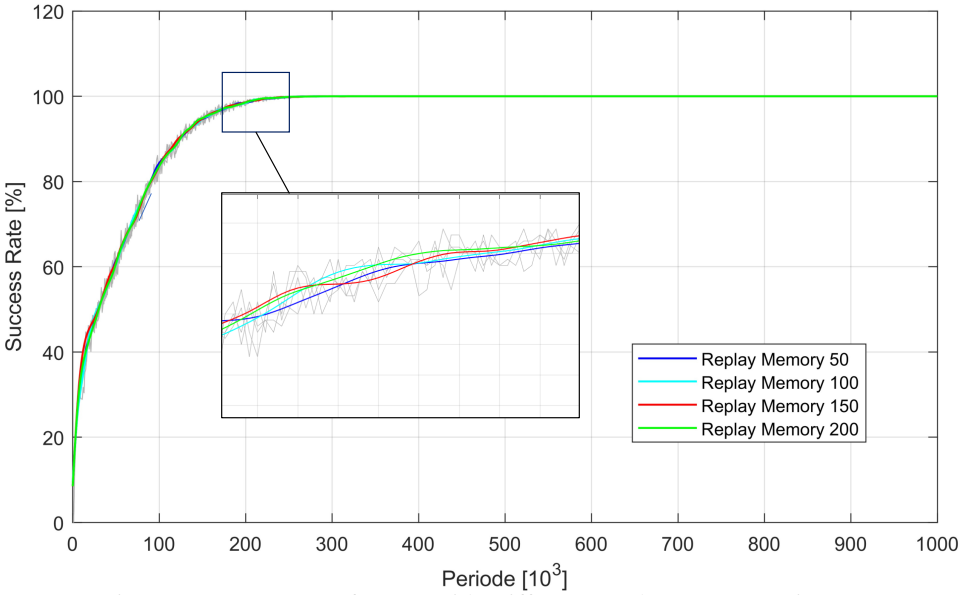


Fig. 9: Success Rate of DQN with Different Replay Memory Sizes

TABLE V: Processing Elements Utilized by DQN for Back-propagation with 100 Replay Memory per Episode

	Neuron					
Parameter	8	10	12	14	16	32
#T per-episode	400	400	400	400	400	400
#Replay Memory	100	100	100	100	100	100
#Adder	10,800	13,200	15,600	18,000	20,400	39,600
#Subtractor	1,946,400	2,432,400	2,918,400	3,404,400	3,890,400	7,778,400
#Multiplier	3,882,000	4,851,600	5,821,200	6,790,800	7,760,400	15,517,200
#Overall PE	5,839,200	7,297,200	8,755,200	10,213,200	11,671,200	23,335,200

TABLE VI: Processing Elements Utilized by DQN for Back-propagation with 200 Replay Memory per Episode

	Neuron					
Parameter	8	10	12	14	16	32
#T per-episode	400	400	400	400	400	400
#Replay Memory	200	200	200	200	200	200
#Adder	7,200	8,800	10,400	12,000	13,600	26,400
#Subtractor	1,297,600	1,621,600	1,945,600	2,269,600	2,593,600	5,185,600
#Multiplier	2,588,000	3,234,400	3,880,800	4,527,200	5,173,600	10,344,800
#Overall PE	3,892,800	4,864,800	5,836,800	6,808,800	7,780,800	15,556,800

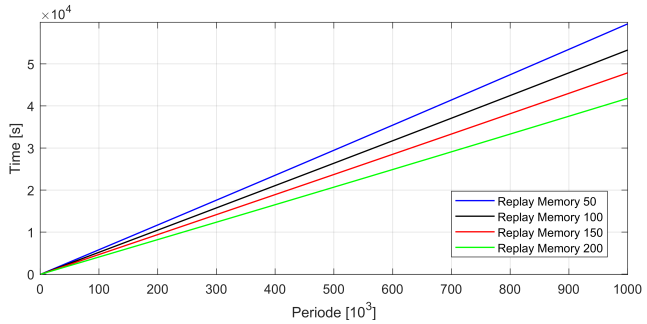


Fig. 10: Computation Time of DQN with Different Replay Memory Sizes

The computational requirement for a smaller replay memory is reduced because the back-propagation process is conducted over a shorter time period. However, the Success Rate and Reward Function exhibit nearly identical values across different replay memory sizes. Consequently, employing a replay memory with a depth of 200 proves to be more efficient by utilizing Processing Elements (PE) compared to utilizing a replay memory with a depth of 50. The edge device which will be use to solve the shortest path problem in a maze by a mobile robot will be implementing FPGA in the system. The number of processing elements (PEs) which are required on the edge device will affect the system's complexity, the goal is ensuring that the optimum value generated by the DQN

Algorithm model provides an efficient system comparable to full computing device. Based on the experimental results, an effective neuron count is determined to be 14 hidden layers with a replay memory depth of 200. Consequently, this necessitates a total of 7,280,000 PEs for feed-forward and 6,808,800 PEs for back-propagation per episode. For implementation on the edge device, parallelization and approximation techniques need to be employed to reduce resource usage on the FPGA that will be utilized.

V. Conclusion

The design of the DQN algorithm for determining the shortest paths in a maze of size 20x20 is adaptable to edge computing. The Neural Network model employs 14 neurons in the Hidden Layer and a Replay Memory of 200. The feed-forward process involves 2,396,800 adders, 4,597,600 multipliers, and 285,600 divisions as Processing Elements (PE). For back-propagation, using 14 neurons in the Hidden Layer and a Replay Memory depth of 200, the required PE amounts to 12,000 adders, 2,269,600 subtractors, and 4,527,200 multipliers. The algorithm attains convergence over 250×10^3 episodes, demanding a total of 3.522×10^{12} PE. In the context of accelerator DQN on edge computing, an approximation process and parallel computation are imperative for efficient execution.

Acknowledgment

This work is sponsored in part by the Education Funding Service Center (PUSLAPDIK) of the Ministry of Education and Culture, Research, and Technology, the Republic of Indonesia, and the Education Fund Management Institution (LPDP) of the Ministry of Finance, the Republic of Indonesia, with a contract No. 00016/J5.2.3./BPI.06/9/2022

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] M. Rothmann and M. Porrmann, "A survey of domain-specific architectures for reinforcement learning," *IEEE Access*, vol. 10, pp. 13 753–13 767, 2022. DOI: 10.1109/ACCESS.2022.3146518.

[3] S. Zhou, X. Liu, Y. Xu, and J. Guo, "A deep q-network (dqn) based path planning method for mobile robots," in *2018 IEEE International Conference on Information and Automation (ICIA)*, 2018, pp. 366–371. DOI: 10.1109/ICInfA.2018.8812452.

[4] L. Luo, N. Zhao, Y. Zhu, and Y. Sun, "A* guiding dqn algorithm for automated guided vehicle pathfinding problem of robotic mobile fulfillment systems," *Computers and Industrial Engineering*, vol. 178, p. 109 112, 2023, ISSN: 0360-8352. DOI: https://doi.org/10.1016/j.cie.2023.109112. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360835223001365.

[5] N. Sutisna, A. M. R. Ilmy, I. Syafalni, R. Mulyawan, and T. Adiono, "Farane-q: Fast parallel and pipeline q-learning accelerator for configurable reinforcement learning soc," *IEEE Access*, vol. 11, pp. 144–161, 2023. DOI: 10.1109/ACCESS.2022.3232853.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, ISSN: 00280836. [Online]. Available: http://dx.doi.org/10.1038/nature14236.

[7] J. Su, J. Liu, D. B. Thomas, and P. Y. Cheung, "Neural network based reinforcement learning acceleration on fpga platforms," *SIGARCH Comput. Archit. News*, vol. 44, no. 4, pp. 68–73, 2017, ISSN: 0163-5964. DOI: 10.1145/3039902.3039915. [Online]. Available: https://doi.org/10.1145/3039902.3039915.

[8] H. Watanabe, M. Tsukada, and H. Matsutani, "An fpga-based on-device reinforcement learning approach using online sequential learning," in *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2021, pp. 96–103. DOI: 10.1109/IPDPSW52791.2021.00022.

[9] Y. Meng, C. Zhang, and V. Prasanna, "Fpga acceleration of deep reinforcement learning using on-chip replay management," *ACM International Conference on Computing Frontiers*, 2022. DOI: 10.1145/3528416.3530227. [Online]. Available: https://par.nsf.gov/biblio/10350273.

[10] G. Gerogiannis, M. Birbas, A. Leftheriotis, E. Mylonas, N. Tzanis, and A. Birbas, "Deep reinforcement learning acceleration for real-time edge computing mixed integer programming problems," *IEEE Access*, vol. 10, pp. 18 526–18 543, 2022. DOI: 10.1109/ACCESS.2022.3147674.

[11] Q. Zhang, M. Lin, L. T. Yang, Z. Chen, and P. Li, "Energy-efficient scheduling for real-time systems based on deep q-learning model," *IEEE Transactions on Sustainable Computing*, vol. 4, no. 1, pp. 132–141, 2019. DOI: 10.1109/TSUSC.2017.2743704.

[12] S. Lee and H. Lee, "Musk-dqn: Multi-ubs selective-k deep q-network for maximizing energy-efficiency," in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, 2023, pp. 438–442. DOI: 10.1109/CCNC51644.2023.10059845.

[13] Numpy, *Numpy reference*, Available at htttps://numpy.org/doc/stable/reference/generated/numpy.exp.html (2023/07/22).

[14] M. Abramovitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1964.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Comparison of MPPT based on Deep Reinforcement Learning by DQN, DDPG and TD3

Jayandi Panggabean, Nana Sutisna, Infall Syafalni, and Trio Adiono
School of Electrical Engineering and Informatics
Bandung Institute of Technology
E-mail: jayandi@soriasi.net

***Abstract*—The renewable energy, especially photovoltaic (PV), is growing exponentially for being able to harvest solar power, which is abundant, especially in tropical countries, like Indonesia. There is the limit of the power harvested, first is PV efficiency and also how the generated power is transferred to the load. The power generated by PV (P_{PV}) should be able to be transferred to the load close to its maximum generated power. This task is done by an MPPT (Maximum Power Point Tracking) module in the PV System. MPPT has been widely implemented especially the conventional methods like Perturb and Observe (P&O). The review of conventional and even AI-based MPPT elucidates problems when MPPT transfers the power during the partial shading condition (PSC) and also at low generated power. The power fluctuates and is significantly inefficient in both conditions respectively. The latest studies showed that MPPT by Deep Reinforcement Learning (DRL) has provided a stable and more efficient P_{VP} compared to P&O and other conventional methods. The study in this research continues the previous one on DRL by reproducing MPPT using DQN (Deep Q-Network) and DDPG (Deep Deterministic Policy Gradient) and introducing TD3 (Twin Delayed Deep Deterministic Policy Gradient). The study shows that at irradiation 1000W/m^2 condition, TD3 produced output power 8.37% higher than DDPG and 1.61% higher than DQN. Then at low irradiation, 250W/m^2 TD3 is higher 13.93% than DDPG but lower 29.27% than DQN. On both conditions TD3 is the least fluctuated, the standard deviation is 0.02, while DQN is 4.04 at clear day and DDPG is 5.8 at cloudy day. This is a promising outcome for renewable energy harvesting, especially on a large scale.**

***Index Terms*—Photovoltaic, MPPT, Deep Reinforcement Learning, DQN, DDPG, TD3**

I. Introduction

Deep reinforcement learning has been applied to many applications including PV systems. The PV system required a controller to ensure the power generated by the PV module transferred to the load efficiently or as much as possible. [1] The PV system consists of PV, Converter, Controller of the Converter, and Load. MPPT is the controller, and it drives the converter to maximize the power transfer. MPPT works based on the information of PV specifications provided by the PV manufacturer. The specification has important information such as the maximum power PV could generate, maximum current and voltage (maximum power point), and the impact of temperature on the PV generated capacity. In uniform conditions, PV is expected to generate at its maximum capacity and less in non-uniform conditions. In uniform conditions, the temperature factor controls the maximum capacity of PV. On the contrary, in non-uniform conditions, the maximum capacity of PV is controlled not only by temperature factor but also by voltage level of PV, low irradiation will make PV cells provide less voltage than its rating. [2] Based on the conditions above, there is a requirement for MPPT to handle various conditions of PV systems as input and keep the maximum possible power output transferred to the load. This requirement has been handled by many conventional methods(non-AI), AI (particle swarm optimization), and a hybrid of AI and conventional. However, there are some drawbacks of these methods, which are: [2]:

1) Upon irradiation change, the output of the converter fluctuated and kept fluctuating.
2) At low irradiation, the MPPT is inefficient.

The initial study of DRL provides a promising result compared to the previous methods, as mentioned before. [3], [4], [5]. Even though the implementation of DRL is still under study since the outcome is not met with the model yet [3]. The study is reproducing and extending the MPPT by DRL by Bao et. all [3], Avila et. al. [4] and Gaviria et. al. [6]. The study result is a more stable output based on the TD3 algorithm compared to DDPG and DQN at uniform and non-uniform conditions. This study should be continued and implemented to ensure the benefits of DRL as an MPPT controller.

II. PV System Control Design

A. PV Model and its Components

The PV system for this research is a reproduction of Bao et. al. [3] and Avila et. al.[4] consisting of 3 PV of American Choice Solar, ACS-335-M; each has 80 cells per module, open circuit voltage (V_{oc}) is 49.9V, short circuit current (Isc) is 9A, the voltage at maximum power point (V_{mpp}) is 41.5V, current at maximum power point (I_{mpp}) is 8.07A, temperature coefficient at V_{oc} is -0.36%/°C, temperature coefficient at I_{sc} is -0.36%/°C, light-generated current I_L is 9.0297A, diode saturation current I_0 is 3.1312e-10A, diode ideality factor is 1.0111, shunt resistor R_{SH} is 77.6407 ohm and series resistance R_S is 0.25633 ohm. Each module has a bypass diode if the module has low or no irradiation.

Figure 1 shows the architecture of a PV system where power from PV is transferred to load through a boost converter which is driven by MPPT.

The frequency of PWM is 25kHz the L and C are 2.057mH and 12.9nF respectively, and R as load is 61.71003717 ohm.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

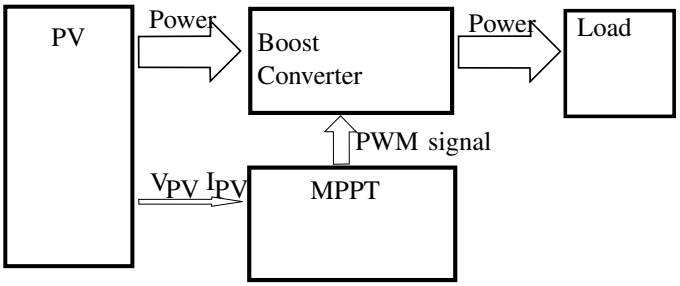


Fig. 1. PV System, PV is ACS-335-M, MPPT by DRL, Resistive Load, MPPT driving the converter by sending PWM signals.

There is a filter capacitor of 3mF. During training episodes, the irradiation and temperatures of PV is randomly initialized according to these conditions:

1) The irradiation values of each PV are randomly selected:
   a) The same value for all from 250W/m^2 to 1000W/m^2 steps 50W/m^2.
   b) Two same values but one is different, all selected from the same range as above.
   c) All three have different values from 150W/m^2 to 1000W/m^2 steps 50W/m^2.
2) The temperature value of all PV is randomly selected from 15°C to 30°C steps 5°C.

Therefore, in total, there are about 23,000 possibilities of initial conditions.

B. Converter Model

There are 3 converter models that are commonly in use for this purpose: Buck, Boost, and Buct-Boost. This study uses a boost converter and the switching is driven by a 25kHz Duty Cycle to the PWM signal module.

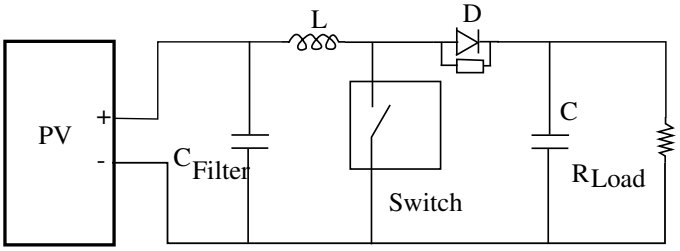


Fig. 2. Boost Converter, PV produces power and filtered by C_{Filter}, L will increase the current loop upon Switch on, D for ensuring current flow toward R_{load}, C for keeping the voltage level when Switch off. R_{load} is a resistive load [7]

The voltage output of PV will be converted by Boost Converter at continued conduction according to the equation 1 [8]:

$$E_c = \frac{T}{T - t_o n} E \tag{1}$$

E_C is the voltage at R_load and E is the V_{PV}, voltage at PV output. T is the frequency of PWM, and t_{on} is the pulse on duration. The converter will increase the PV output voltage based on the t_{on} duration.

C. Methodology of MPPT in PV Systems

At uniform conditions, the voltage and current of PV are at V_{mpp} and I_{mpp} respectively. At non-uniform conditions the PV may have several PV cells disconnected from PV output for safety reasons the current is not drawn from PV, instead, the cells are bypassed and go through a bypass diode for this purpose. This will make the overall PV voltage output decrease significantly.

MPPT will try to find the Global MPP instead of Local MPP, Global MPP is the highest possible power generated and sometimes it is not at the highest possible voltage. [9]

III. DRL as MPPT Controller

A simplified architecture of the PV System utilizing DRL as MPPT Controller is as figure 3 below [9].

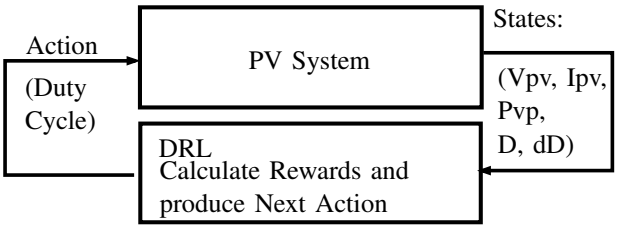


Fig. 3. The environment (PV System) and Actor (DRL system), Environment provides states to be calculated by DRL algorithm to produce rewards and next action. D is the duty cycle and dD is delta D compared to the previous D.

A. Deep Reinforcement Learning and Neural Network Composition

The DRL model free for solving the PV in this simulation is Deep Q-Network (DQN), Deep Deterministic Policy Gradient (DDPG), and Twin Delayed Deep Deterministic Policy Gradient (TD3). For having the sense of similarity, the neural network number of hidden layers and neurons is similar one to the other while keeping the architecture of each. The critic network used 200 neurons per layer and there are 4 hidden layers, the activation function is ReLU. The critic action network uses the *tanh* activation function. For DDPG and TD3, the action network used 200 neurons per layer and there are 3 layers. DQN, DDPG, and TD3 details are explained in the following subsections.

The simulation conditions for DQN algorithm are a revision of Gaviria et. al. [6] codes. The revision is required since there are functions used in Gaviria's code which is not recommended anymore for the simulation purpose based on Matlab recommendations [10]. Therefore the code is restructured, adjusted, and rewritten by new recommended functions.

The DDPG and TD3 Agents were created following the neural network structure of the DQN Agent. As stated previously, the DQN code is a revision of Gaviria's[6], however, DDPG and TD3 have no published code available. Bao et. al. [3] provide brief information on the DDPG algorithm structure but not the code. Therefore the code for DDPG Agent [11]

and TD3 Agent [12] were created according to the Mathworks recommendation structure while DDPG is following the recommendation from Bao et. al. [3].

Training episode for all methods is limited to 10000. Lastly, the GPU is used as a calculating device.

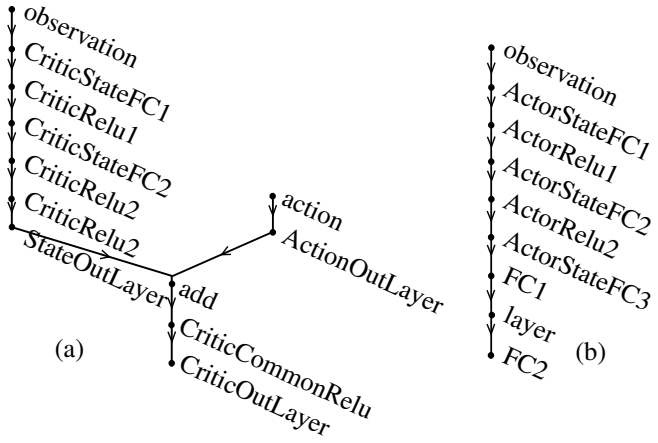


Fig. 4. (a) The Critic Network used in all DRL models (b) The action network used in DDPG and TD3

B. DQN as MPPT Controller

The methodology of DQN is describe in pseudo-code as Algorithm 1, [13] and modified by Bao et. al. [3]:

Algorithm 1 Algorithm for Deep Q-Network

Input: states
Output: action

Initialize replay memory D with a defined capacity N
Randomly initialize critic network Q(s, a|θ) with weight θ
Initialize target network $Q^{'}$(s,a|$\theta^{'}$) with weights $\theta^{'} \leftarrow \theta$
for episode = 1 to M **do**
  Initialize state s
  for t=1 to T **do**
    Following ϵ-greedy policy,
    select a_t= $\begin{cases} \text{a random action, if probability} \leq \epsilon \\ argmax_a \text{ Q}(s_t, a_t|\theta), \text{otherwise} \end{cases}$
    Execute action a_t to get reward r_t and observe the next states s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in D
    Sample a random mini-batch of (s_t, a_t, r_t, s_{t+1}) from D
    Set y_j = $\begin{cases} r_j, \text{ if episode terminates at step j+1} \\ r_j + \gamma \ max_a \ Q(s_{t+1}, a_{t+1}; \ |\theta^{'}), \text{otherwise} \end{cases}$
    Perform a gradient descent step to minimize the loss function for updating the critic network $L = \frac{1}{N}\sum_j (y_j - Q(s_j, a_j|\theta^Q))^2$, w.r.t. the network parameter θ
    Update the weights of the target network every C steps
  end for
end for

Actions are discrete values: [-0.03 -0.01 -0.005 -0.001 0 0.001 0.005 0.01 0.03], γ (discount factor) is 0.9, mini-batch size is 512, the learning rate is 0.0001 and adam optimizer applied. Double DQN is chosen as the method since it reduces the overestimate problem in DQN. [3]

C. DDPG as MPPT Controller

The algorithm is selected as the actor-critic representative for off-policy, model-free DRL [14] [3]. The pseudo-code for DDPG is Algorithm 2

Algorithm 2 Algorithm for DDPG

Randomly initialize critic network Q(s, a, $|\theta^Q$) and actor network $\mu(s|\theta^\mu)$ with weight parameters θ^Q and θ^μ, respectively.
Initialize target network Actor Q'(s, a, $|\theta^{'}_Q$) and Critic $\mu^{'}$(s $|\theta^{'}\mu$) with weight parameters $\theta^{Q'} \leftarrow \theta^Q$ and $\theta^{\mu^{'}} \leftarrow \theta^\mu$, respectively
Initialize experience replay memory D to a defined N
for episode = 1 to M **do**
  Initialize the random process $\mathbb{N}$ in the action exploration strategy.
  Observe initial states s_1
  for t=1 to T do **do**
    Select an action a_t= $\mu_t(s_t|\theta^\mu) + N$, based on the current policy $\mu(st)$ and exploring noise $\mathbb{N}$
    Execute action a_t in the environment to get reward r_t and the new state s_{t+1}
    Save transition (s_t, a_t, r_t, s_{t+1}) in D
    Sample random mini-batch of *Nth* transitions (s_t, a_t, r_t, s_{t+1}) from D
    Set $y_j = r_j + \gamma Q'(s_{j+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})|\theta^{Q'}$
    Update online critic network by minimizing loss: $L = \frac{1}{N}\sum_j (y_j - Q(s_j, a_j|\theta^Q))^2$, w.r.t. the network parameter θ
    Update the actor policy actor-network the sampled policy gradient.
    Update all target actor networks and target critic networks
  end for
end for

The critic and actor-network configurations are stated in Figure 4. The state element is limited as follows:

1) V_{PV}: -inf to inf
2) I_{PV}: -inf to inf
3) Duty Cycle: from 0 to 1.0
4) Delta of Duty Cycle: from -1.0 to 1.0

The action elements are limited from 0.0 to 1.0 (duty cycle values). The Target Smooth factor is 0.001, the discount factor is 1.0, the mini-batch size is 512, and the experience buffer length is 1e6. The Adam optimizer is applied to the network, the learning rate is 0.0001, and the gradient threshold is 1.

D. TD3 as MPPT Controller

The methodology of TD3 is describe in pseudo-code as Algorithm 3, modified from Fujimoto et. al. [15]:

Algorithm 3 Algorithm for TD3

Initialize critic network $Q_{\theta_1}, Q_{\theta_2}$, and actor network π_ϕ. with random parameters θ_1, θ_2, ϕ
Initialize target network $\theta_1' \leftarrow \theta_1, \theta_2' \leftarrow \theta_2, \phi' \leftarrow \phi$
Initialize replay memory D
for t = 1 to T do **do**
  Select action with exploration noise a $\sim \pi_\psi(s) + \epsilon, \epsilon \sim \mathbb{N}(0, \delta)$ and observe reward r and new state s_{t+1}
  Store transition tuple (s_t, a_t, r_r, s_{t+1}) in D
  Sample mini-batch of N transitions (s_t, a_t, r_r, s_{t+1}) from D
  $\tilde{a} \leftarrow \pi_{\phi'} + \epsilon, \epsilon \sim$ clip($\mathbb{N}(0,\tilde{\delta})$,-c,c)
  $y \leftarrow r + \gamma\ min_{i=1,2}\ Q_{\theta'}\ (s', \tilde{a})$
  Update critics $\theta_i \leftarrow argmin_{\theta_i} N^{-1} \Sigma (y - Q_{\theta_i}(s,a))^2$
  if t mod d then **then**
    Update ϕ by the deterministic policy gradient:
    $\nabla_\phi J(\phi) = N^{-1} \Sigma \nabla_a Q_{\theta_1}(s,a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$
    Update target networks
    $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$
    $\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
  end if
end for

The hyper-parameters are the same as DDPG.

IV. Simulation and Analysis

A. Simulation Conditions and Results

The training was on an i7 Intel computer with 16GB memory capacity (expandable by swap memory on hard drive up to 200GB), NVidia RTX 2060 using Matlab R2023a. DQN training took time around half of DDPG and TD3, the memory usage of DDPG is the highest compared to the two others, which is 120GB, and the TD3 and The DQN about 33GB. The agent was then tested against two extreme PV conditions: 1000W/m^2 and 250W//m^2 representing clear and mostly cloudy days, the results are shown in the figures 5 and 6 also in tables I and II.

There is still fluctuation on all power signals and they are all compared at 1000W/m2 irradiation and also at 250W/m2 one.

The duration compared between the results is the last 0.1s of the signals for it is considered all signals have been in steady state conditions. There are 41,436 samples during a duration of time 0.3 seconds to 0.4 seconds with time sampling is approximately 6.15µs.

B. Analysis of The Results

The training period for the TD3 is the longest while DQN is the fastest for TD3 structure has more neural networks compared to DDPG and DQN, therefore training time is not compared.

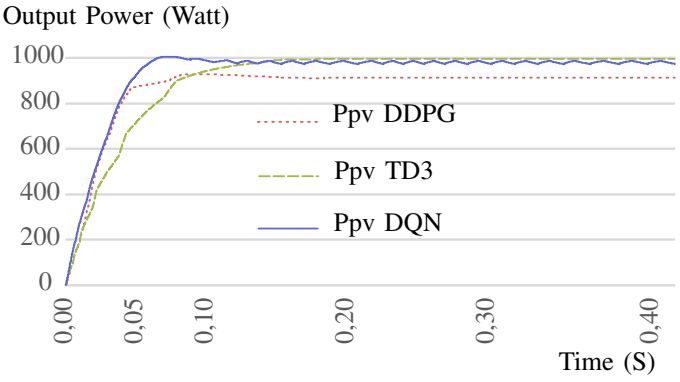


Fig. 5. Comparison of Output power of PV (P_{PV}) by DRL MPPT trained based on DDPG, TD3 and DQN algorithm. All PV is received irradiation 1000W/m^2

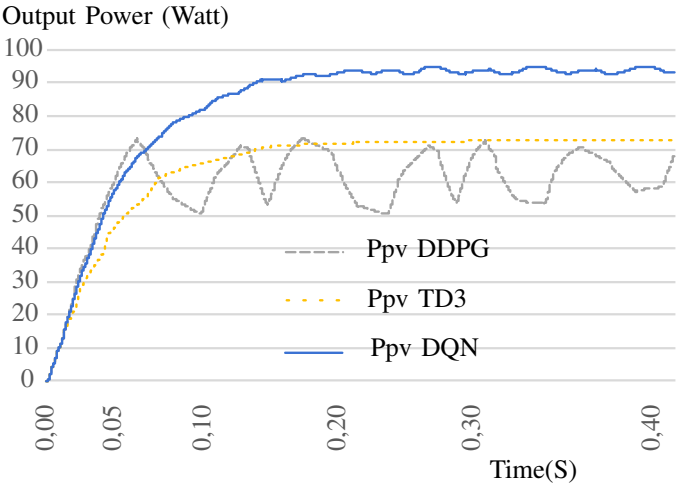


Fig. 6. Comparison of Output Power of PV (P_{PV}) by DRL MPPT trained based on DDPG, TD3 and DQN algorithm while all PV has irradiation 250W/m^2

TABLE I
Comparison of DQN, DDPG, and TD3 at Uniform Condition from 0.3s to 0.4s Period

	Irradiation at 1000W/m^2		
	DQN (W)	***DDPG (W)***	***TD3 (W)***
Mean	980.6801	913.3411	996.7506
SDev	4.043671	0.03989	0.029259

TABLE II
Comparison of DQN, DDPG, and TD3 at Uniform Condition from 0.3s to 0.4s Period

	Irradiation at 250W/m^2		
	DQN (W)	***DDPG (W)***	***TD3 (W)***
Mean	93.76509	62.42913	72.5331
SDev	0.738881	5.805285	0.021484

The main target of the research was comparing the power output produced by the converter circuit (P_{PV}) which is driven by DRL Agents, DQN Agent, DDPG Agent and TD3 Agent at given PV conditions (Irradiation on each PV and temperatures of the PV panels). The environmental conditions of PV were the same and each agent has same number of training episodes.

Based on the output power, at PV irradiation is 1000W/m^2, TD3 produced average maximum output power 8.37% higher than DDPG and 1.61% higher than DQN. At PV irradiation is 250W/m^2, average maximum output power of TD3 is higher 13.93% than DDPG but lower 29.27% than DQN.

Based on the power signal fluctuation at 0.1second period, TD3 has the least fluctuated power output, the standard deviation of TD3 power signal is 0.02 on high and low irradiation. While DDPG and DQN have higher fluctuations (standard deviation).

Standard deviation of DQN is 0.74 at cloudy day or low irradiation and DDPG is 0.04 at clear day or high irradiation and standard deviation of DQN's power signal is 4.04 at clear day and DDPG is 5.8 at cloudy day.

The response time of each curve to achieve its steady state condition is longer at 250W/m^2 compared to at 1000W/m^2. For DQN, it is about 2 times longer (0.05s to 0.1s), DDPG is nondeterministic since too fluctuated at 250W/m^2 and for TD3 it is about 1.5 time longer (0.1s to 0.15s). The response time is essential during a quick PSC day.

V. Discussion

The study found TD3 is the promising candidate as MPPT controller for the stability and efficiency in power transfer and also for the time response for PSC. The study had been carefully designed to be comparable to previous research and tested on 10000 episodes for training purpose of the agent. The study confirmed the previous research and extended to TD3 algorithm for TD3 is a revision of DDPG. The simulation done also found that TD3 is less memory consuming compared to DDPG.

At the moment a training consist of 100000 episodes is ongoing and may become a reference for other DRL or methods to achieve a stable and efficient MPPT controller or even as reference for implementation in the field.

Next, it is suggested to have further study covering control of each element of the PV System to gain optimum benefit of DRL in the system also effect of noise in the system.

VI. Conclusion

In this paper, DRL algorithm DQN and DDPG are implemented as for MPPT. All algorithm have almost similar neural network architecture and number of training episodes. TD3 provides a more stable power transfer compared to DQN and DDPG, either in 1000W/m^2 and 250W/m^2 irradiation and highest efficiency in high power transfer. At 1000W/m^2 condition, TD3 produced output power 8.37% higher than DDPG and 1.61% higher than DQN. At 250W/m^2 TD3 is higher 13.93% than DDPG but lower 29.27% than DQN.

On both condition TD3 has the least fluctuated power output, the standard deviation is 0.02 on both conditions. DQN is 4.04 at clear day and DDPG is 5.8 at cloudy day. Next, it is suggested to have further study covering control of each element of the PV System to gain optimum benefit of DRL in the system.

References

[1] J. Enslin, "Maximum power point tracking: A cost saving necessity in solar energy systems," in *[Proceedings] IECON'90: 16th Annual Conference of IEEE Industrial Electronics Society*, IEEE, 1990, pp. 1073–1077.

[2] N. Karami, N. Moubayed, and R. Outbib, "General review and classification of different mppt techniques," *Renewable and Sustainable Energy Reviews*, vol. 68, pp. 1–18, 2017.

[3] B. C. Phan, Y.-C. Lai, and C. E. Lin, "A deep reinforcement learning-based mppt control for pv systems under partial shading condition," *Sensors*, vol. 20, no. 11, p. 3039, 2020.

[4] L. Avila, M. De Paula, I. Carlucho, and C. S. Reinoso, "Mppt for pv systems using deep reinforcement learning algorithms," *IEEE Latin America Transactions*, vol. 17, no. 12, pp. 2020–2027, 2019.

[5] S. A. Raj and G. G. Samuel, "Survey of ai based mppt algorithms in pv systems," in *2022 4th international conference on smart systems and inventive technology (ICSSIT)*, IEEE, 2022, pp. 597–604.

[6] J. F. Gaviria, G. Narváez, C. Guillen, L. F. Giraldo, and M. Bressan, "Machine learning in photovoltaic systems: A review," *Renewable Energy*, Jul. 1, 2022, ISSN: 0960-1481. DOI: 10.1016/j.renene.2022.06.105. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148122009454 (visited on 07/03/2022).

[7] G. Petrone, C. A. Ramos-Paja, and G. Spagnuolo, *Photovoltaic sources modeling*. John Wiley & Sons, 2017.

[8] Y. Haroen, "Elektronika daya," *Hak Cipta Penerbit ITB, Bandung*, 2011.

[9] K.-Y. Chou, C.-S. Yang, and Y.-P. Chen, "Deep q-network based global maximum power point tracking for partially shaded pv system," in *2020 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*, IEEE, 2020, pp. 1–2.

[10] *(Not recommended) Value function critic representation for reinforcement learning agents - MATLAB — mathworks.com*, https://www.mathworks.com/help/reinforcement-learning/ref/rl.representation.rlvaluerepresentation.html, [Accessed 12-09-2023].

[11] *Deep Deterministic Policy Gradient (DDPG) Agents - MATLAB Simulink — mathworks.com*, https://www.mathworks.com/help/reinforcement-learning/ug/ddpg-agents.html, [Accessed 12-09-2023].

[12] *Twin-Delayed Deep Deterministic (TD3) Policy Gradient Agents - MATLAB Simulink — mathworks.com*, https://www.mathworks.com/help/reinforcement-learning/ug/td3-agents.html, [Accessed 12-09-2023].

[13] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[14] S. Guo, X. Zhang, Y. Zheng, and Y. Du, "An autonomous path planning model for unmanned ships based on deep reinforcement learning," *Sensors*, vol. 20, no. 2, p. 426, 2020.

[15] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*, PMLR, 2018, pp. 1587–1596.

Signal Quality Assessment for Wearable Multichannel Photoplethysmography Signals

Muhammad Dzaky Prihatmoko*[1], Nur Ahmadi*†‡[2], Peter H. Charlton‡[3], Trio Adiono*[4]
*School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, Indonesia
†Center for Artificial Intelligence (U-CoE AI-VLB), Bandung Institute of Technology, Bandung, 40132, Indonesia
‡Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, United Kingdom
Email: [1]dzakyprihatmoko@gmail.com, [2]nahmadi@itb.ac.id, [3]pc657@cam.ac.uk,[4]tadiono@itb.ac.id

***Abstract*—Photoplethysmography (PPG), a non-invasive method to measure health parameters, is highly dependent on signal quality to produce accurate and reliable measurements. Wearable PPG signals are susceptible to various factors that can degrade their signal quality. To address this issue, we present an algorithm for reducing errors by assessing signal quality and combining information from multichannel PPG to estimate heart rate (HR), interbeat interval (IBI) and heart rate variability (HRV) parameters using the time domain approach. Our method works by dividing into multiple 10 s nonoverlapping segments, assessing the quality of segments from multiple channels within the same period, selecting the best segments that are acceptable in terms of quality via template matching. Our proposed method significantly improves the number of acceptable samples (65.32%) and reduces the mean absolute error (MAE) for the estimation of HR and the average NN interval (AVNN) by 62.20% and 78.60%, respectively, compared to the single channel PPG. Overall results suggest that signal quality assessment from multiple channels can be used to improve the accuracy of the estimation of PPG-derived parameters from wearable devices.**

I. Introduction

Photoplethysmography (PPG) is implemented in various wearable devices such as wrist-worn smartwatches and finger-worn smart rings. Unlike fingertip PPG signals that are normally acquired in clinical settings where the subject is sitting still, wearable PPG devices are designed for daily activities in which many movements are involved. These movements can introduce noise and motion artifacts. Several approaches have been proposed to detect such artifacts [1], [2]. The motion artifact degrades the signal quality, in some cases obscuring pulse peaks and distorting the pulse wave shape. This can significantly reduce the accuracy of the derived health parameters.

To address the above issue, multiple preprocessing steps have been developed and widely used, such as detrending, bandpass filtering, and denoising using accelerometer sensors [3]. The preprocessed signals are then fed into the algorithms to estimate health parameters. Generally, the algorithms can be categorized into time domain methods and frequency domain methods. Time domain methods involve peak detection and calculation of the periods between the peaks in which health parameters can be derived. Frequency domain methods transform PPG signals into the frequency domain and typically identify the frequency with the highest peak to obtain the desired health parameters. Frequency domain methods are more robust than the time domain counterparts, since noises can be directly removed by excluding high-frequency noise in the frequency domain. However, frequency domain methods limit the capability of obtaining other parameters such as the interbeat intervals (IBIs).

Time domain methods measure the period between consecutive peaks in the PPG signal. Numerous peak detection methods have been investigated in the past [4], [5]. Signal quality usually varies throughout the signal and there are segments that have poor quality. When these segments are fed into peak detection algorithms, peaks can be missed or falsely detected, producing erroneous parameter estimations. Studies implementing these peak detection methods [6], opted to discard segments with flat lines due to sensor disconnection. There needs to be a way to determine whether or not the result of the time domain method is acceptable for further analysis.

In this paper, we propose a signal quality assessment (SQA) algorithm to determine whether a segment is usable or not by utilizing PPG signals from multiple channels. We select channels from different wavelengths that had been used in the past for PPG signals. The three channels that are widely used are green [7], [8], red [9], [10], and infrared [9]–[11]. Multichannel approaches have been previously developed, such as using data fusion to estimate respiratory rate, delivering promising results [12]. Each channel has its own advantages and disadvantages. For example, green PPG signals are more robust to motion artifacts but are affected by melanin, as green light only reaches the superficial part of the skin [13]. On the other hand, red and infrared light can penetrate deeper into the skin, meaning that they are less likely to be affected by melanin but are unfortunately more vulnerable to motion artifacts [13].

The remainder of this paper is organized as follows. Section II describes the methods which include data collection, proposed algorithm, signal quality assessment, and health parameters to be estimated. Section III presents and discusses

This research was funded by Bandung Institute of Technology (ITB) International Research Program 2022 (No. 4949/IT1.B07.1/TA.00/2022). PHC acknowledges funding from the British Heart Foundation grant [FS/20/20/34626].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

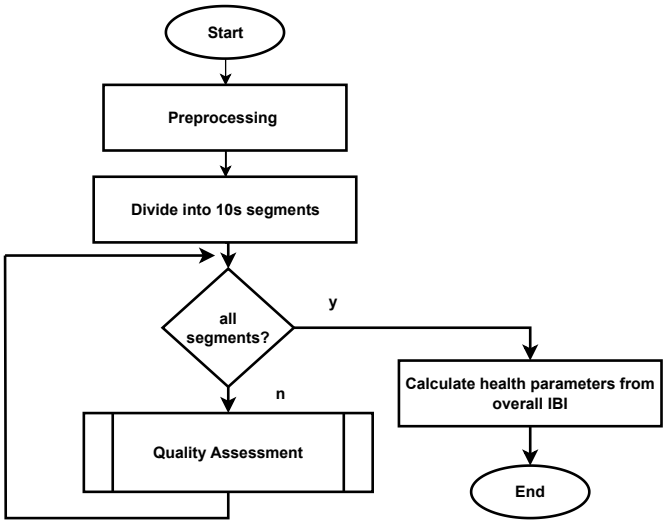


Fig. 1. The proposed signal quality asessment (SQA) algorithm

the performance evaluation of the proposed algorithm on the collected data. Finally, conclusions are drawn in Section IV.

II. Methods

A. Data Collection

Data were collected from 31 healthy subjects (17 males and 14 females) with age ranging from 20 to 24 years old. All subjects provided their informed consent following the guidelines established in the Declaration of Helsinki. The ethical approval was obtained from the Research Ethics Committee of the Bandung Institute of Technology, Indonesia (No. KEP/II/2022/X/M130723NA/PSPT). Electrocardiogram (ECG) signals were measured with the Polar H10 device at a sampling frequency of 100 Hz. PPG signals were acquired using our custom wrist-worn PPG device, equipped with the max30101 sensor with a sampling rate of 60 Hz. Three PPG channels collected green, red, and infrared wavelength PPG signals. The PPG and ECG signals were obtained simultaneously for 90 seconds. As a reference, we used IBI data obtained by the built-in function provided by Polar H10.

B. Proposed Algorithm

Fig. 1 shows the proposed SQA algorithm. All PPG channels were interpolated using cubic spline interpolation from 60 Hz to 100 Hz to ensure the same sampling frequency as the ECG. Interpolated signals were detrended to remove baseline wander as well as low-frequency noise. The signals were then filtered with a bandpass filter to keep frequency content between 1–8 Hz. Next, the signals were divided into multiple segments, each with a length of 10 s ($\sim$1000 samples per segment). IBI estimation and quality assessment were performed on each segment. Finally, the resulting IBIs were used to calculate health parameters. All signal processing was performed in Python with the help of Neurokit2 library [14].

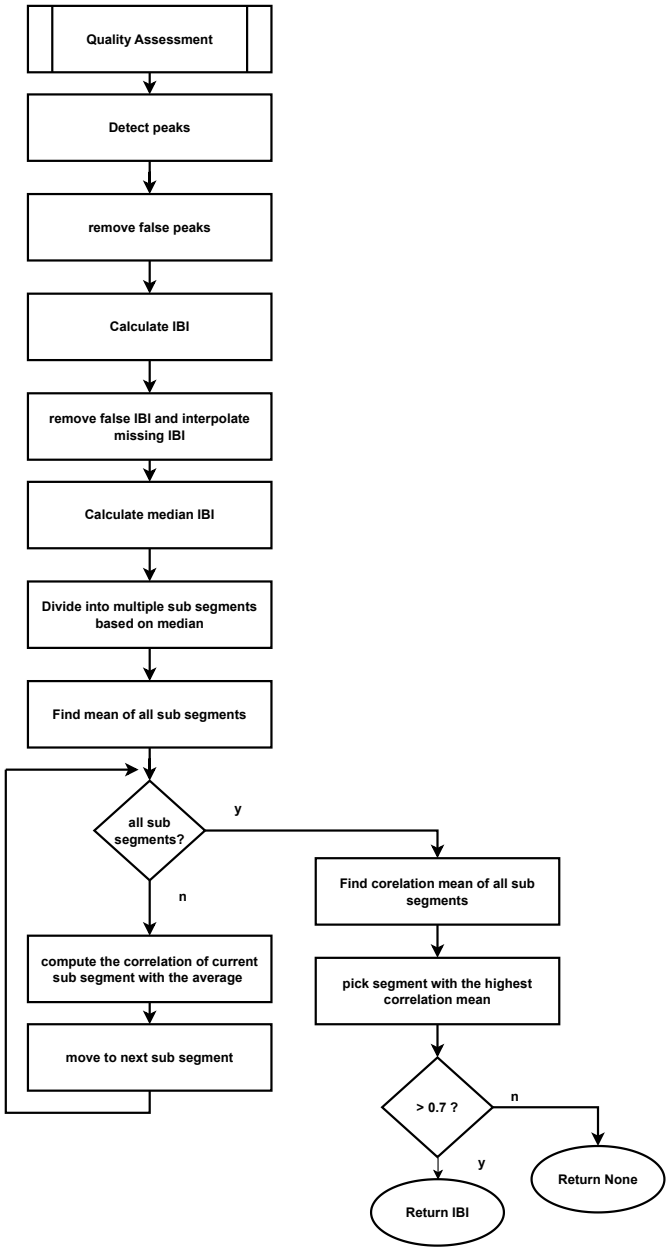


Fig. 2. Signal quality assessment (SQA) algorithm

C. Signal Quality Assessment

Fig. 2 illustrates the steps of the signal quality assessment (SQA) algorithm. IBIs were calculated using a time domain method which consists of detecting pulse wave peaks and then calculating IBIs as the time delays between successive peaks. We used qppg algorithm for peak detection because it was found to yield the highest accuracy particularly during movement [6], [15]. The qppg algorithm detects systolic upslopes that are identified with adaptive thresholding and with the help of a slope sum function, and calculates the total magnitude of upslopes in the previous 0.17 s. In our study, we observed that in some cases, it overestimates the number of peaks where additional peaks come from noise. Fig. 3 shows

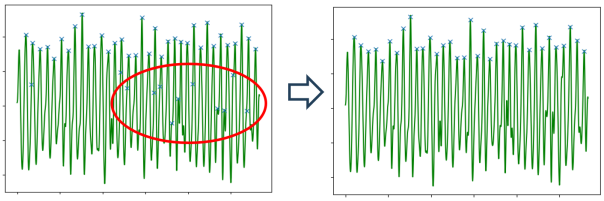


Fig. 3. Removal of false peak

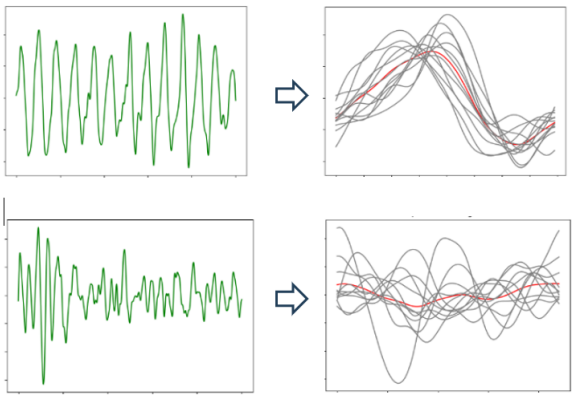


Fig. 4. Template matching

an example of peaks being falsely detected.

To solve the problem of false peak detection, we employed straightforward peak removal method. We first collected the data from 20 samples before and 20 samples after the current peak. If both the former and the latter were lower than the corresponding peak, the peak was kept. However, when the condition was not met, the peak was marked as false and rejected. Next, we calculated the IBI using the already confirmed peaks by using the difference in time between adjacent peaks. The resulting IBIs exhibited variations and sudden changes (spiky). While IBIs could vary over time [16], some of these variations are were too large. Falsely detected peaks rather than actual PPG systolic peaks contribute to this issue. Therefore, we smoothed the resulting IBIs by removing IBIs with differences greater than 0.15 s from the median of the overall IBI within this segment. In the cases where there were at least four IBIs left, we interpolated the number of IBIs to have the same number of IBIs as before the false IBI was removed.

Template matching [17] [18] searches for regularity within the PPG segment. It is based on the assumption that the shapes of pulse waves in high-quality PPG signals are almost consistent throughout the signal. In Fig. 4, we present the differences between high- and low-quality PPG segments. Grey lines represent individual pulse wave, while the red line is the average of all the pulse waves within the segment. High quality PPG segments have a regular shape, resulting in grey lines that overlap each other. On the contrary, low-quality segments have a variety of shapes that appear to be random and have low similarity. The median IBI was calculated for each 10 s PPG segment. The median was then used to divide the segment into multiple subsegments. We proceeded to estimate the mean IBI of all the sub-segments and used it as the reference shapes. We computed the correlation from each pulse wave with that of the reference shape and the average all the correlations. We found that the best threshold for accepting a PPG segment is 0.7 after experimenting with different threshold numbers, which was high enough to prevent low quality signals from passing in while still allowing imperfect signals with slight noise to be used as input. We selected one out of the three signal segments (green, red, and infrared) in the same period that exceeds the 0.7 minimum threshold and scores the highest. If all three segments had scores below the threshold, the segments were discarded.

TABLE I
HEALTH PARAMETER

Parameter Name	Definition
HR	Heart rate calculated from AVNN
AVNN	Average of NN intervals
SDNN	Standard deviation of NN intervals
RMSSD	Root Mean square of successive Differences
SDSD	Standard seviation of successive Differences
PNN50	Proportion of successive difference that are over 50 ms
PNN20	Proportion of successive difference that are over 20 ms

TABLE II
PERCENTAGE OF ACCEPTED SAMPLES

Signals	Accepted Samples	Percentage (%)
Green	163	38.72
Red	127	30.17
Infrared	221	52.49
Combined	275	65.32

III. RESULTS AND DISCUSSION

A. Low Quality Signal Identification

Health parameters that were estimated in this study along with their definition are shown in Table I. In total, 421 samples were gathered from 31 subjects. Some samples were in low-quality which could yield large errors. In our algorithm, if all the segments from the collected signals do not surpass the threshold in all periods, then the corresponding sample is discarded. For comparison, we also present the number of acceptable samples from three different channels computed individually using the proposed SQA algorithm. Based on Table II, we can see that the percentage of samples categorized as high quality increases when using multiple channels compared to single channel. The infrared channel produced the highest percentage of the accepted samples, but was still lower than when all three channels were combined with the percentage of 65.32%. Our results showed that multichannel

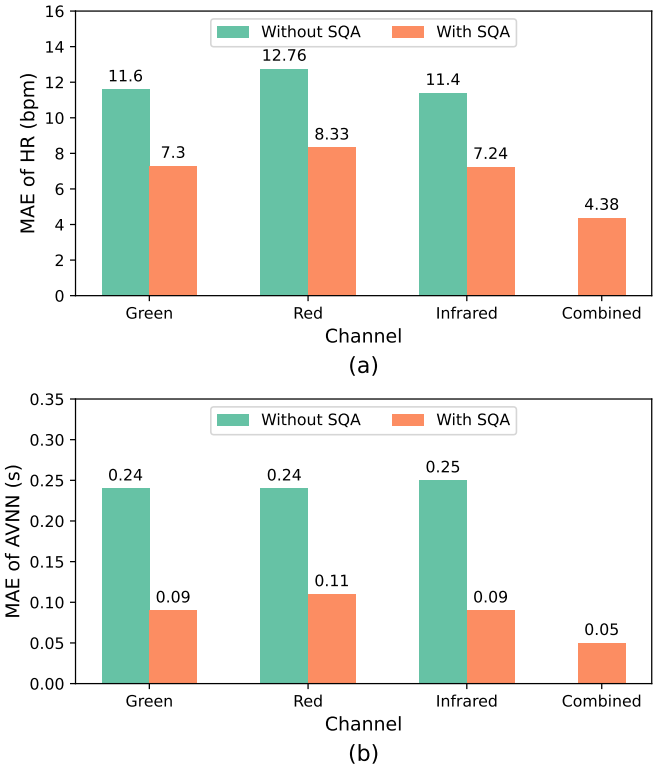


Fig. 5. MAE for (a) HR and (b) AVNN

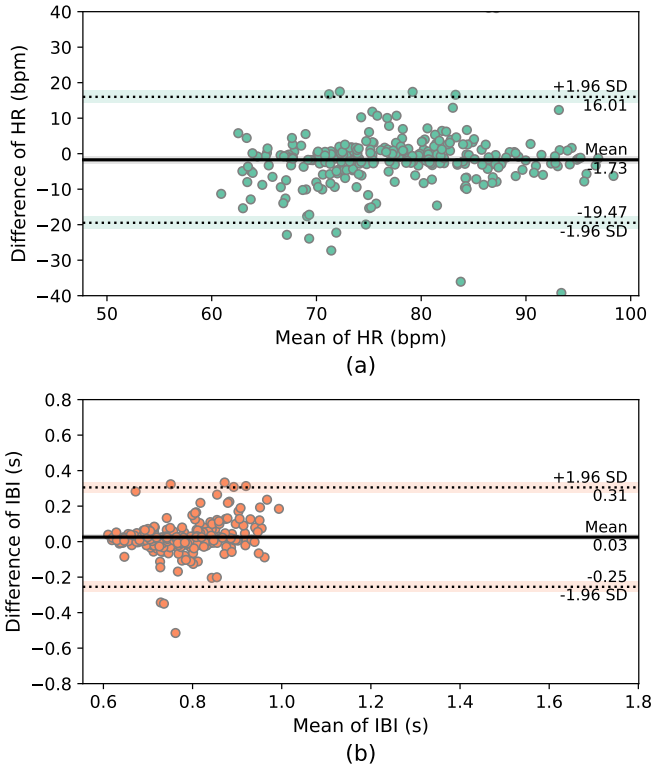


Fig. 6. Bland-Altman plot for (a) HR and (b) IBI from combined channels with signal quality assessment.

approach decreases the amount of data loss due to noise. This is beneficial in real-world applications where it is desired to keep as much data as possible for analysis.

B. Estimation Improvement

Fig. 5(a) shows mean absolute error (MAE) comparison from single channel (green, red, or infrared) and multichannel (combination of all the three signals) with and without SQA algorithm. It is shown that using the SQA on a single channel measurement alone was able to reduce the error by at least 34.72%. Overall, by combining all the PPG signals with SQA to select the best segment, we were able to reduce the MAE by 62.20%. AVNN is used to calculate HR with the following formula.

$$\mathrm{HR} = 60/\mathrm{AVNN} \tag{1}$$

Improvement of HR estimation was caused by the improvement of AVNN. From Fig. 5(b), we can see that the MAE of AVNN reduced with the use of SQA on a single channel source by at least 54.86%. The combined use of all three channels with SQA reduced the MAE by 78.60%. The MAEs were 0.24 s, 0.24 s, and 0.25 s for green, red, and infrared signal, respectively. The MAE obtained from SQA was as small as 0.05 s, which indicates that the proposed SQA algorithm successfully reduced the error by 44%. Looking into the details, the Bland-Altman plot in Fig. 6 illustrates the distribution of these errors. These outliers were low quality data that had not been marked as low quality by the algorithm. When the algorithm mistakenly identifies low quality signals as high ones, it could produce a large error .

C. Estimation of HRV Parameters

The good improvement shown in HR and AVNN encouraged us to apply the proposed SQA algorithm for the estimation of HRV parameters, which are typically calculated for extended periods such as 24 hours. The currently used dataset was only taken from 90 s measurements, which falls into ultra-short measurement category. Previous studies were conducted on ultra-short measurements [19] [20] and motivated us to evaluate our SQA algorithm on ultra-short term HRV measurements.

Fig. 7 illustrates the MAE obtained from the estimation of five HRV parameters: RMSSD, SDNN, SDSD, PNN20, and PNN50. From Fig. 7(a), the MAE of RMSSD, SDNN, and SDSD were above 0.1 s or 100 ms. The algorithm performed best (lowest MAE) when estimating SDNN. Fig. 7(b) illustrates the result in percentage form for PNN20 and PNN50. PNN50 and PNN20 measure the proportion of changes in NN interval that are above 50ms and 20ms, respectively. It is clear that for both PNN20 and PNN50, the error is large with more than 20%. These results indicate that despite the good improvement on HR estimation accuracy, it is still not enough to yield reliable HRV estimation with values typically less than 100 ms.

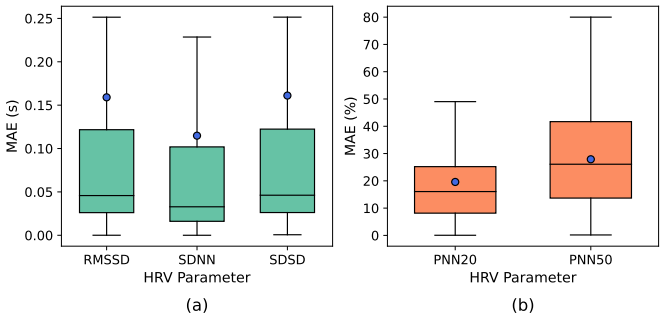


Fig. 7. Boxplot of MAE for HRV parameter estimation: (a) RMSSD, SDNN, and SDSD (b) PNN20 and PNN50

IV. Conclusion

We have presented a signal quality assessment (SQA) algorithm from wearable multichannel (combination of green, red, and infrared) PPG signals. We demonstrate that our proposed algorithm can improve the accuracy of health parameter estimation using our in-house dataset. It reduced the MAE by 62.2% for HR and 78.6% for AVNN compared to the result from individual signal without SQA. However, it was still not good enough to be used for measuring HRV in ultra-short duration. For future work, we suggest further development of the algorithm to allow for more accurate HRV measurement in low quality signals and in ultra-short duration.

References

[1] X. Liu, Q. Hu, H. Yuan, and C. Yang, "Motion artifact detection in ppg signals based on gramian angular field and 2-d-cnn," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 743–747.

[2] S. Hanyu and C. Xiaohui, "Motion artifact detection and reduction in ppg signals based on statistics analysis," in *2017 29th Chinese control and decision conference (CCDC)*. IEEE, 2017, pp. 3114–3119.

[3] L. B. Wood and H. H. Asada, "Noise cancellation model validation for reduced motion artifact wearable ppg sensors using mems accelerometers," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 3525–3528.

[4] A. E. Johnson, J. Behar, F. Andreotti, G. D. Clifford, and J. Oster, "R-peak estimation using multimodal lead switching," in *Computing in Cardiology 2014*. IEEE, 2014, pp. 281–284.

[5] S. M. Bishop and A. Ercole, "Multi-scale peak and trough detection optimised for periodic and quasi-periodic neuroscience data," in *Intracranial Pressure & Neuromonitoring XVI*. Springer, 2018, pp. 189–195.

[6] P. H. Charlton, K. Kotzen, E. Mejía-Mejía, P. J. Aston, K. Budidha, J. Mant, C. Pettit, J. A. Behar, and P. A. Kyriacou, "Detecting beats in the photoplethysmogram: benchmarking open-source algorithms," *Physiological Measurement*, vol. 43, no. 8, p. 085007, 2022.

[7] J. Lee, K. Matsumura, K.-i. Yamakoshi, P. Rolfe, S. Tanaka, and T. Yamakoshi, "Comparison between red, green and blue light reflection photoplethysmography for heart rate monitoring during motion," in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 1724–1727.

[8] A. A. Silverio, J. E. Asilo, S. L. D. Balagat, A. A. Pineda, J. A. Tangi, D. T. P. Tuazon, and B. M. B. Piedra, "Micromotion artefact reduction of a wrist worn ppg sensor using green light ppg and surface emg," in *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 2020, pp. 1–5.

[9] Y. Khan, D. Han, J. Ting, M. Ahmed, R. Nagisetty, and A. C. Arias, "Organic multi-channel optoelectronic sensors for wearable health monitoring," *IEEE Access*, vol. 7, pp. 128 114–128 124, 2019.

[10] M. Huotari, J. Röning, and K. Määttä, "Infrared and red ppg signals analysis of the healthy subjects and clinical patients," in *2016 15th Biennial Baltic Electronics Conference (BEC)*. IEEE, 2016, pp. 111–114.

[11] R. Shriram, M. Sundhararajan, and N. Daimiwal, "Effect of change in intensity of infrared led on a photoplethysmogram," in *2014 International Conference on Communication and Signal Processing*. IEEE, 2014, pp. 1064–1067.

[12] S. Nemati, A. Malhotra, and G. D. Clifford, "Data fusion for improved respiration rate estimation," *EURASIP journal on advances in signal processing*, vol. 2010, pp. 1–10, 2010.

[13] S. Ryals, A. Chiang, S. Schutte-Rodin, A. Chandrakantan, N. Verma, S. Holfinger, F. Abbasi-Feinberg, A. Bandyopadhyay, K. Baron, S. Bhargava *et al.*, "Photoplethysmography—new applications for an old technology: a sleep technology review," *Journal of Clinical Sleep Medicine*, vol. 19, no. 1, pp. 189–195, 2023.

[14] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021.

[15] A. N. Vest, G. Da Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, and G. D. Clifford, "An open source benchmarked toolbox for cardiovascular waveform and interval analysis," *Physiological measurement*, vol. 39, no. 10, p. 105004, 2018.

[16] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[17] H. Chan, G. Chen, M. Lin, and S. Fang, "Heartbeat detection using energy thresholding and template match," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 6668–6670.

[18] V. Krasteva and I. Jekova, "Qrs template matching for recognition of ventricular ectopic beats," *Annals of biomedical engineering*, vol. 35, pp. 2065–2076, 2007.

[19] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term hrv features as surrogates of short term hrv: A case study on mental stress detection in real life," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–13, 2019.

[20] A. Taoum, A. Bisiaux, F. Tilquin, Y. Le Guillou, and G. Carrault, "Validity of ultra-short-term hrv analysis using ppg—a preliminary study," *Sensors*, vol. 22, no. 20, p. 7995, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

After-Fatigue Condition: A Novel Analysis Based on Surface EMG Signals

Van-Hieu Nguyen[1], Gia Thien Luu[2,*], Thien Van Luong[1], Mai Xuan Trang[1]
Philippe RAVIER[3], Olivier BUTTELLI[3]
[1]Faculty of Computer Science, Phenikaa University, Hanoi 12116, Vietnam
[2]Biomedical Engineering Department, HUTECH Institute of Engineering, HUTECH University, Ho Chi Minh, Viet Nam
[3]University of Orléans, France PRISME Laboratoire, 12 rue de Blois, BP 6744, 45067 Orléans, France
E-mail: ngvan.hieu613@gmail.com, lg.thien@hutech.edu.vn, {thien.luongvan, trang.maixuan}@phenikaa-uni.edu.vn

***Abstract*—This study introduces a novel muscle activation analysis based on surface electromyography (sEMG) signals to assess the muscle's after-fatigue condition. Previous studies have mainly focused on the before-fatigue and fatigue conditions. However, a comprehensive analysis of the after-fatigue condition has been overlooked. The proposed method analyzes muscle fatigue indicators at various maximal voluntary contraction (MVC) levels to compare the before-fatigue, fatigue, and after-fatigue conditions using amplitude-based, spectral-based, and muscle fiber conduction velocity (CV) parameters. In addition, the contraction time of each MVC level is also analyzed with the same indicators. The results show that in the after-fatigue condition, the muscle activation changes significantly in the ways such as higher CV, power spectral density shifting to the right, and longer contraction time until exhaustion compared to the before-fatigue and fatigue conditions. The results can provide a comprehensive and objective evaluation of muscle fatigue and recovery, which can be helpful in clinical diagnosis, rehabilitation, and sports performance.**

***Index Terms*—Surface electromyography, maximal voluntary contraction, conduction velocity, EMG, power spectral density.**

I. Introduction

Surface electromyography (sEMG) is a non-invasive technique to measure the electrical activity of skeletal muscles. Surface EMG signals can provide valuable information about muscle fatigue, defined as the decrease of maximal force output to maintain or repeat tasks by muscles [1]. Muscle fatigue can affect the performance and health of muscles in various fields such as medicine, sports, rehabilitation, ergonomics, and human-machine interaction [2]–[4]. Therefore, assessing muscle fatigue based on sEMG signals is of great significance and especially interesting to researchers and health professionals.

Surface EMG signals are a valuable tool to evaluate muscle fatigue, and their applications in life and industrial fields are promising and diverse. However, assessing muscle fatigue based on sEMG is complex because it involves many complex factors, such as muscle type, contraction type, electrode placement, signal processing methods, and fatigue indices [5]. In addition, surface EMG signals are often affected by noise from various sources, such as motion artifacts, skin perspiration, skin impedance, and electromagnetic interference [6], [7]. Therefore, robust and reliable methods are in demand to extract meaningful features from sEMG signals and to classify the states of muscles before, during, and after muscle fatigue. Feature extraction of sEMG signals is a method of extracting useful information from signals under different muscle conditions. Muscle fatigue is usually identified by the decrease of frequency domain indices and time-frequency domain indices and the increase of time domain indices. The traditional approach to acquiring sEMG signals is still commonly used in physiological and clinical studies, based on a pair of electrodes placed on the skin over the muscle. However, the acquired signal depends on the location, the distance between electrodes, the size of the electrode pair, and the area along the muscle fiber, which can lead to very different spectral characteristics and amplification [5], [6].

Initially, sEMG signal features can be divided into three domains: time domain, frequency domain, and time-frequency domain. However, in the past 40 years, a large number of parameters extracted from sEMG signals to evaluate muscle fatigue have been developed. Currently, sEMG features are divided into amplitude-based parameters, spectral-based parameters, non-linear parameters, and muscle fiber conduction velocity (CV) estimation [8], [9]. For the classical bipolar measurement method, using methods based on more than two electrodes arranged in series allows collecting sEMG signals along the vertical or horizontal axis of the muscle [10]. A relatively recent approach includes using multiple electrode groups arranged in one- or two-dimensional arrays. In addition, the two-dimensional electrode array can determine the distribution of sEMG amplitude and describe the spectrum over the entire skin area covering the muscle [11]. Multi-channel sEMG signals also allow for more accurate and reliable estimation of CV [12]. Previous studies have investigated the changes in power spectral density (PSD) of sEMG signals during different muscle contraction and load levels. For example, [8], [13] examined the PSD of sEMG signals at various maximal voluntary contraction (MVC) levels and found that the PSD shifted to lower frequencies with increasing MVC. And [9] used a linear array sEMG to study the relationship between CV and the shift of PSD over time with different load levels. They observed that the CV decreased, and the PSD shifted to

*Corresponding Author

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

lower frequencies as the load increased.

Most studies mentioned above have mainly focused on the before-fatigue and fatigue conditions, but the after-fatigue condition has not been well explored. Moreover, previous studies exploited only several separate parameters, which did not fully interpret the objective and comprehensive evaluation of muscle in different conditions. The main contributions and findings in this work are summarized as follows:

- A novel analysis that analyzes all muscle fatigue indicators, such as Root mean square, Mean frequency, Power spectral density, and CV at various maximal voluntary contraction (MVC) levels. In contrast, the previous works analyzed only a few fatigue indicators at a few MVC levels and muscle conditions [14].
- The first attempt to analyze after-fatigue conditions compared to before-fatigue and fatigue conditions. Note that the previous studies only focused on before-fatigue and fatigue conditions as shown in [8], [9], [13]. Such analysis and comparison showed a significantly increasing contraction time of the after-fatigue condition until exhaustion, and this phenomenon was partially explained.

The rest of this work is constructed as follows. Section II presents the signal acquisition protocol and our analysis method, and then Section III discusses the experimental results. Finally, Section IV concludes the paper.

II. Materials and Methods

A. Signal acquisition protocol and data pre-processing

A monopolar sEMG signal was collected by the two-dimensional 64-electrode sEMG sensor, divided into thirteen rows and five columns, with an inter-electrode distance of eight millimeters and a sampling frequency of 2048 hertz. The sensor provides five groups of differential signals with thirteen electrodes along with the Biceps Brachii muscle group. A force signal over time is measured from the force gauge collected with each MVC level, as shown in Fig. 1 (A). In addition, the representation of the sensor matrix and the direction of signal propagation is described in Fig. 1 (B). The sEMG signals from all electrodes obtained from the MVC levels were digitally filtered (20 Hz-400 Hz) to reduce noise.

The data was collected on ten healthy subjects: three women and seven men (mean age 24, standard deviation 1.5 years). The MVC of an individual subject was defined as the average value of three times producing the maximum force, which separated every five minutes. After determining the MVC, the tests can be conducted at different force levels, expressed as a percentage of MVC. The subject should hold the force for at least ten seconds if possible. The experiment starts with a test at 10% of MVC so the subject can learn to maintain a certain force level. Furthermore, the process continues with the random order tests at 20%, 40%, 60%, and 90% of MVC, an exhaustion test at 70% of MVC, and a final test at 10% of MVC. Each subject takes a rest of five minutes between two consecutive tests. As such, the states before muscle fatigue include 10%MVC, 20%MVC, 40%MVC, 60%MVC, and 90%MVC, while the state during muscle fatigue is 70%MVC and the state after muscle fatigue is 10%MVC.

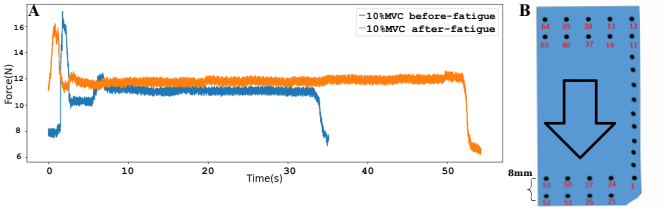


Fig. 1. (A) Representation of the longer force's contraction time until exhausted. (B) The matrix sensor is used in signal acquisition.

B. Amplitude-based parameters

Root mean square (RMS) is the main parameter used to investigate the amplitude of the sEMG signal. The formula for RMS is as follows:

$$RMS = \sqrt{\frac{1}{N}\sum_{n} x_n^2}, \tag{1}$$

where x_n are the values of the sEMG signal, and N is the number of data samples. RMS is believed that during maximal isometric contractions, amplitude falls progressively, in parallel with the decrease in force [15].

C. Spectral-based parameters

Two characteristic frequencies that have been used to quantify changes in spectral content based on the Fourier transform are the mean frequency (MNF) and the median frequency (MDF) of the power spectrum [16]. MDF is calculated as follows:

$$MDF = \int_{f_1}^{f_{median}} PS(f)df = \int_{f_{median}}^{f_2} PS(f)df, \tag{2}$$

while MNF is calculated as follows:

$$MNF = \frac{\int_{f_1}^{f_2} f.PS(f)df}{\int_{f_1}^{f_2} PS(f)df}, \tag{3}$$

where $PS(f)$ is the power spectrum calculated from the Fourier transform, f_1 and f_2 determine the lowest and highest frequency of the signal bandwidth, usually ranging from 20 hertz to 400 hertz. MDF and MNF are related to the change in CV and have been shown in isometric contractions that MNF will shift to lower frequencies during fatigue [17].

D. Conduction velocity estimation

The muscle's CV was initially calculated by placing electrodes along the muscle. However, this method can be biased in estimating CV when muscle fibers are not placed in a parallel plane. Moreover, electrodes placed on an undefined muscle domain can lead to misleading information [18]. Recently, multi-channel sEMG signals allowed more accurate estimation of CV both at the global level of the muscle and CV of individual motor units [15]. Therefore, this study conducted CV

estimation based on the multi-channels maximum likelihood estimation algorithm [19]. First, three single differential sEMG signals were selected to remove the mean value. Then, the algorithm calculates two double-differential sEMG signals and tries to maximize the likelihood delay function for observing the delay time between the two signals. CV was obtained after calculating the delay time by dividing the electrode distance by the delay time. The formula for calculating CV is as follows:

$$CV = \frac{d}{\theta}, \tag{4}$$

where d is the distance between two electrodes, θ is the delay time between two electrodes. The algorithm is described in more detail in [19].

CV parameter is related to the properties of the fiber membrane, fiber diameter, and contraction properties of the fiber. Therefore, measuring CV degradation is considered as the strongest indicator of muscle fatigue [20]. In addition, changes in the CV in muscle fatigue have a profound impact on the motor unit action potential waveform and, therefore, affect both amplitude parameters and spectral parameters extracted from sEMG, as analyzed in [21], [22]. Previous studies [23], [24] have also shown that CV is reduced due to the consequences of local metabolic changes in active muscles, mainly due to H+ and K+ distribution in the sarcolemma.

E. The proposed after-fatigue analysis methods

One of the novel aspects of this analysis method is that it calculates all indicators on the sEMG signal according to MVC levels and contraction time. While in [2] only shown a few indicators. The amplitude parameters (RMS) and spectral parameters (MNF) were obtained from the entire sEMG signal for each MVC level, excluding the noise parts at the beginning and end. The parameters were averaged over all signals. The electrodes' signals were visualized through a single differential filter to determine the innervation zone, as shown in Fig. 2.

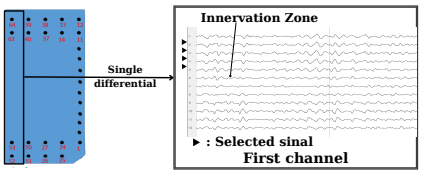


Fig. 2. Representation of Innervation zone and signals selection.

The multi-channel algorithm mentioned earlier estimated the CV with selected sEMG signals. The CV value will be accepted if the correlation coefficient between adjacent signals is more than 0.75. The estimated CV value of each MVC level was averaged over channels.

Furthermore, according to the analysis method of contraction time, each selected signal was divided into intervals, spaced 1000ms and 500ms long, which is believed to be stationarity. MNF and RMS obtained from each MVC level were averaged over all selected signals. The PSD representation of the signal was divided into three intervals, each with a length of 1000ms. "Onset" is taken at the beginning, "Middle" is taken in the middle, and "End" is taken at the end of the signal. These intervals allow capturing the changes in sEMG parameters over time and comparing them across MVC levels.

Finally, CV was calculated on each selected signal with segments 500ms long and spaced 1000ms apart. The CV value was accepted if the correlation coefficient between adjacent signals was more significant than 0.75. The estimated value of CV for each MVC level was averaged over channels. By using these techniques, the analysis method provides a comprehensive and accurate evaluation of muscle fatigue and recovery.

III. Experimental Results

A. Results from analysis of each MVC level

In this section, The analysis of the PSD, RMS, CV, and MMF of the sEMG signals recorded from the biceps brachii muscle at different levels of maximum voluntary contraction (MVC) was compared with previous studies. The findings' implications according to 10%MVC after-fatigue are also discussed for muscle fatigue assessment. The results interpretation will be constructed as follows.

The PSD in normalized units for each MVC level is shown in Fig. 3 (A). The PSD of 70%MVC fatigue is more concentrated in the lower frequency range, indicating a higher level of muscle activation. This is consistent with the report by [13]. On the other hand, the PSD of other MVC levels tends to shift to lower frequencies as the MVC level increases, suggesting a decrease in muscle activation.

Fig. 3 (B) compares the PSD of 10%MVC before-fatigue and 10%MVC after-fatigue. A significant difference between the two conditions was first found: the PSD of 10%MVC after-fatigue has a higher power spectrum in the higher frequency range. In comparison, the PSD of 10%MVC before-fatigue has a lower power spectrum at the same frequency range. This implies that 10%MVC after-fatigue signal has a higher muscle activation than 10%MVC before-fatigue. Note that this phenomenon has not been mentioned in any previous work and needs further analysis at higher MVC levels.

Fig. 3 (C) shows the general trend of RMS, which linearly increases to higher MVC levels. This coincides with previous studies such as [21], [23]. The RMS of 10%MVC after-fatigue and 10%MVC before-fatigue is almost unchanged. However, the RMS indicator needs to be combined with other indicators, such as MNF and CV, for a comprehensive understanding of muscle fatigue. In addition, 70%MVC fatigue has a higher RMS amplitude compared to other MVC levels because of higher muscle activation when the muscle is in the fatigue stage, which has been investigated in [17], [20].

As shown in Fig. 3 (D), the general trend of CV linearly decreases to higher MVC levels. This coincides with previous studies such as [9]. When the MNF of 10%MVC after-fatigue was compared with other MVC levels, we found that the CV of 10%MVC after-fatigue was the highest compared to other MVC levels. Note that the phenomenon has been overlooked

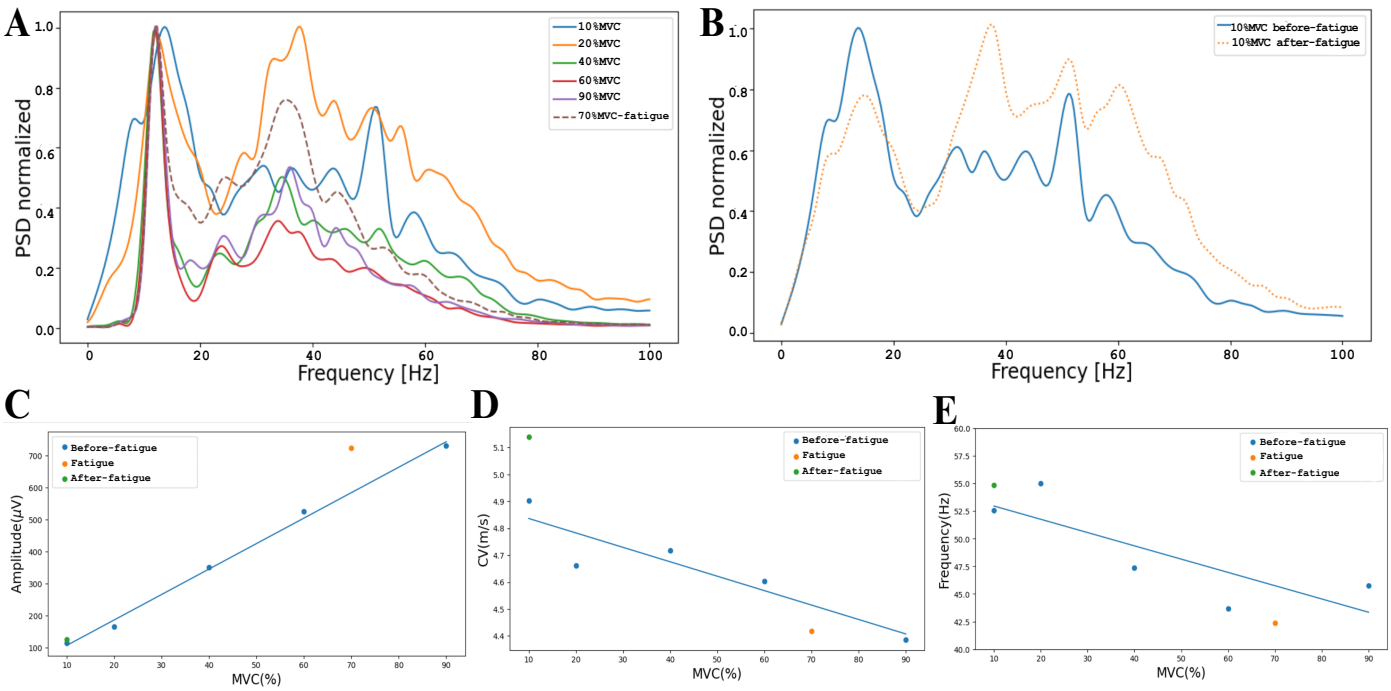


Fig. 3. (A) PSD normalized each MVC level. (B) PSD normalized of 10%MVC before-fatigue and after-fatigue. (C) RMS each MVC level. (D) CV each MVC level. (E) MNF each MVC level.

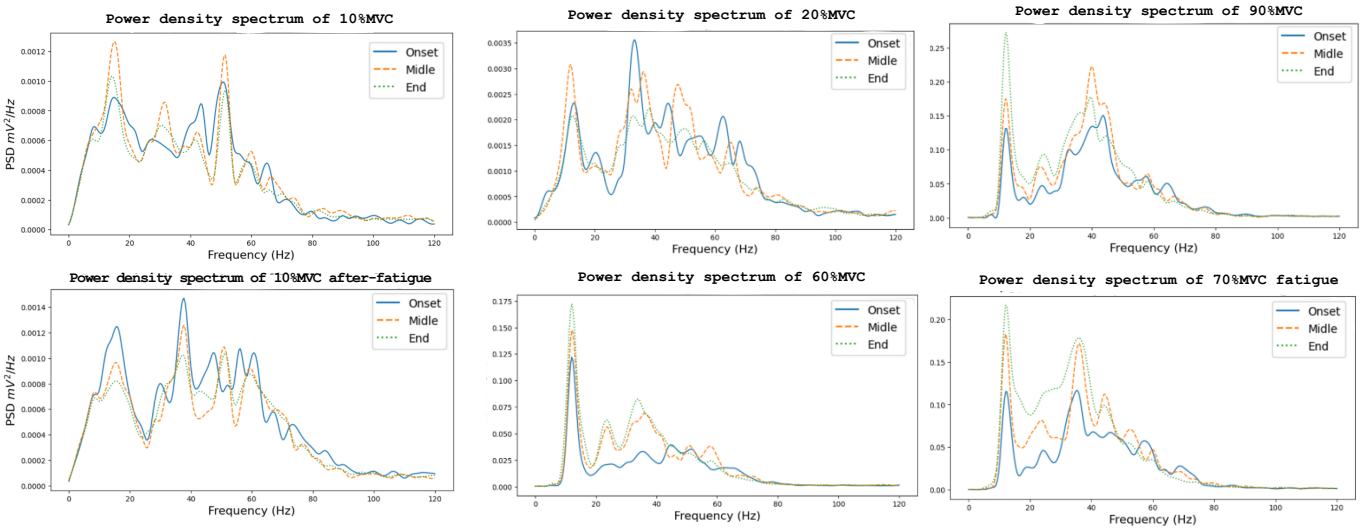


Fig. 4. Representation of the PSD shift during the contraction time.

in previous works. Therefore, we suggest that the robust increase in CV affects the PSD of 10%MVC after-fatigue, as mentioned above. In addition, 70%MVC fatigue has the lowest CV compared to other MVC because of the muscle fatigue stage. Finally, Fig. 3 (E) represents the general trend of MNF, which linearly increases to higher MVC levels. This also coincides with previous works such as [13], [16]. The MNF of 10%MVC after-fatigue was first observed, which has a higher frequency than 10%MVC before-fatigue; this phenomenon combines with indicators mentioned above that suggested a first evaluation of recovery conditions of muscle. In addition, 70%MVC has the lowest MNF compared to other MVC levels, which is the same as expected according to the above indicators.

In summary, novelty findings were found according to 10%MVC after-fatigue compared to other MVC levels. In addition, a shift of PSD to the right at 10%MVC after-fatigue was first found, compared to 10%MVC before-fatigue. This phenomenon comes with a robust increase in MNF and CV compared to other MVC levels. Based on the findings, We

TABLE I
PARAMETER'S SLOPE CHANGE OF EACH MVC LEVEL DURING CONTRACTION TIME

Parameters	MVC levels						
	10%MVC	20%MVC	40%MVC	60%MVC	90%MVC	70%MVC Fatigue	10%MVC (After-fatigue)
MNF	0.011	0.016	-0.321	-0.466	-1.019	-0.64	0.007
RMS	-0.357	-0.051	7.987	15.417	22.376	21.093	-0.004
CV	-0.001	-0.013	-0.019	-0.026	-0.045	-0.054	0.003

suggested a novelty evaluation of the after-fatigue conditions of muscle, where the shift to the right of PSD and increasing MNF denote higher muscle activation and a robust increase in CV denotes the reabsorption and distribution of ions such as H+ and K+ in the muscle membrane, leading to an increase in CV, as mentioned earlier. This further proved the more extended maintenance of contraction time at 10%MVC after-fatigue conditions.

B. Results from analysis contraction time

This section represents an analysis of the PSD indicator over time and evaluates the slope change of each parameter, such as MNF, RMS, and CV. Moreover, findings according to 10%MVC after-fatigue compared to other MVC levels and discuss their implications for muscle fatigue assessment. The results interpretation will be constructed as follows.

Fig. 4 represents the analysis of the PSD indicator over time. The compression trend of PSD was first generalized at various MVC levels and conditions over time, while [13] only exploited PSD on different MVC levels and some conditions. The PSD has a wide frequency, and the mean frequency is almost unchanged according to 10%MVC and 20%MVC before-fatigue. However, when MVC levels are increased, PSD tends to compress at a lower frequency apparent over time according to 60%MVC and 90%MVC. In addition, when the muscle is in the fatigue stage according to 70%MVC, the compression of PSD is not only at a lower frequency but also a robust increase in the power spectrum. Note that this finding has been overlooked in the previous studies.

In Table I, the slope change of all parameters such as MNF, RMS, and CV of each MVC level during contraction time were summarized. Firstly, we observed a general trend of MNF's slope at before-fatigue MVC levels. We first found that the higher the MVC levels, the more apparent MNF's slope decreases, which means the muscle tends to fatigue faster as higher force is required. In addition, MNF's slope of 70%MVC fatigue seems unaffected in fatigue conditions. MNF's slope of 10%MVC after-fatigue is nearly zero, which is almost unchanged over time and needs further analysis to understand fully.

In contrast to the general trend of MNF's slope, RMS' slope tends to increase as higher force is required according to before-fatigue conditions, as shown in Table I. In addition, RMS's slope of 70%MVC fatigue is nearly equal to 90%MVC, which means muscle needs more power to retain force than other MVC levels in fatigue conditions. In 10%MVC after-fatigue conditions, the slope nearly equals zero, requiring lower power to retain force. This phenomenon has been overlooked and needs further analysis. Finally, Table I shows the general CV's slope. CV's slope tends to decrease as higher force is required according to before-fatigue conditions. The CV's slope of 10%MVC after-fatigue was first found, which has a contrasted trend, slightly increasing compared to other MVC levels. Note that this phenomenon has not been reported in any previous works. We suggest that this phenomenon is related to the reabsorption and distribution of ions such as H+ and K+ in the muscle membrane. This leads to an increase in CV and the shift of PSD to the right, as we mentioned earlier.

In summary, a novelty analysis on the PSD indicators was proposed, which generalized the compression trend of PSD at various MVC levels and conditions over time and evaluated the slope change of each parameter, such as MNF, RMS, and CV. In addition, the findings were presented according to 10%MVC after-fatigue compared to other MVC levels.

IV. CONCLUSIONS

In this study, we proposed a method that analyzes muscle fatigue indicators at various MVC levels to fully evaluate muscle conditions according to the before-fatigue, fatigue, and after-fatigue conditions, and analysis of the contraction time of each MVC level is also analyzed with the same indicators to understand better how indicators change to time. The decline of CV, MNF, an increase of RMS, and a shift of power spectrum to lower frequency according to different MVC levels before and during muscle fatigue all provide results that we expect about the trend of muscle fatigue. Regarding the significant increase in muscle contraction maintenance time at 10%MVC level after muscle fatigue, we believe this increase in time is related to the reabsorption and redistribution of ions after muscle fatigue. As mentioned earlier, the change of CV has a solid relationship to the change of metabolism at the cell membrane. The CV also significantly impacts the amplitude and spectrum parameters, shown in the increased power spectrum density at higher frequencies and a slight increase in RMS and muscle force. However, this finding needs further investigation of the properties of the individual motor unit, such as firing rate, recruitment threshold, and synchronization.

References

[1] R. M. Enoka and J. Duchateau, "Muscle fatigue: What, why and how it influences muscle function," *The Journal of Physiology*, pp. 11–23, Jan. 2008.

[2] B. N. Cahyadi, W. Khairunizam, I. Zunaidi, *et al.*, "Muscle fatigue detections during arm movement using emg signal," *The Sports Medicine Resource Manual*, vol. 557, p. 012 004, Jun. 2019.

[3] J. Sun, G. Liu, Y. Sun, K. Lin, Z. Zhou, and J. Cai, "Application of surface electromyography in exercise fatigue: A review," *Frontiers in Systems Neuroscience*, vol. 16, p. 893 275, Aug. 2022.

[4] H. A. Yousif, A. Zakaria, N. A. Rahim, *et al.*, "Assessment of muscles fatigue based on surface emg signals using machine learning and statistical approaches: A review," *IOP Conference Series: Materials Science and Engineering*, vol. 705, p. 012 010, Dec. 2019.

[5] F. Bai, T. M. Lubecki, C.-M. Chew, and C. H. Teo, "Novel time-frequency approach for muscle fatigue detection based on semg," Nov. 2012.

[6] Z. Xiaoqi, L. Yun, and Z. Qianxiang, "Evaluation of muscle fatigue based on surface electromyography and subjective assessment," pp. 2003–2006, Jan. 2013.

[7] G. Y. Millet, M. F. Bertrand, T. Lapole, L. Féasson, V. Rozand, and D. Hupin, "Measuring objective fatigability and autonomic dysfunction in clinical populations: How and why?" *Frontiers in Sports and Active Living*, vol. 5, Mar. 2023.

[8] A. J. Lloyd, "Surface electromyography during sustained isometric contractions.," *Journal of Applied Physiology*, vol. 30, pp. 713–719, May 1971.

[9] T. Sadoyama, T. Masuda, and H. Miyano, "Relationships between muscle fibre conduction velocity and frequency parameters of surface emg during sustained contraction," *European Journal of Applied Physiology and Occupational Physiology*, vol. 51, pp. 247–256, Aug. 1983.

[10] R. Merletti, D. Farina, and M. Gazzoni, "The linear electrode array: A useful tool with many applications," *Journal of Electromyography and Kinesiology*, vol. 13, pp. 37–47, Feb. 2003.

[11] D. Falla and D. Farina, "Periodic increases in force during sustained contraction reduce fatigue and facilitate spatial redistribution of trapezius muscle activity," *Experimental Brain Research*, vol. 182, pp. 99–107, May 2007.

[12] M. Beretta-Piccoli, G. D'Antona, C. Zampella, M. Barbero, R. Clijsen, and C. Cescon, "Test-retest reliability of muscle fiber conduction velocity and fractal dimension of surface emg during isometric contractions," vol. 38, pp. 616–630, Mar. 2017.

[13] S. Nagata, A. B. Arsenault, D. Gagnon, G. Smyth, and P. A. Mathieu, "Emg power spectrum as a measure of muscular fatigue at different levels of contraction," *Medical & Biological Engineering & Computing*, vol. 28, pp. 374–378, Jul. 1990.

[14] A. Troiano, F. Naddeo, E. Sosso, G. Camarota, R. Merletti, and L. Mesin, "Assessment of force and fatigue in isometric contractions of the upper trapezius muscle by surface emg signal and perceived exertion scale," *Gait & Posture*, vol. 28, pp. 179–186, Aug. 2008.

[15] M. Beretta-Piccoli, C. Cescon, and G. D'Antona, "Evaluation of performance fatigability through surface emg in health and muscle disease: State of the art," *Arab Journal of Basic and Applied Sciences*, vol. 28, pp. 21–40, Dec. 2020.

[16] M. González-Izal, A. Malanda, E. Gorostiaga, and M. Izquierdo, "Electromyographic models to assess muscle fatigue," *Journal of Electromyography and Kinesiology*, pp. 501–512, Aug. 2012.

[17] R. Merletti and L. R. Lo Conte, "Surface emg signal processing during isometric contractions," *Journal of Electromyography and Kinesiology*, vol. 7, pp. 241–250, Dec. 1997.

[18] R. Merletti, *Surface Electromyography : Physiology, Engineering, and Applications*. Apr. 2016.

[19] K. C. McGill and L. J. Dorfman, "High-resolution alignment of sampled waveforms," vol. BME-31, pp. 462–468, Jun. 1984.

[20] A. J. Fuglevand, K. M. Zackowski, K. A. Huey, and R. M. Enoka, "Impairment of neuromuscular propagation during human fatiguing contractions at submaximal forces.," *The Journal of Physiology*, vol. 460, pp. 549–572, Jan. 1993.

[21] D. A. Gabriel and G. Kamen, "Experimental and modeling investigation of spectral compression of biceps brachii semg activity with increasing force levels," *Journal of Electromyography and Kinesiology*, vol. 19, pp. 437–448, Jun. 2009.

[22] L. R. Brody, M. Pollock, S. Roy, C. D. Luca, and B. R. Celli, "Ph-induced effects on median frequency and conduction velocity of the myoelectric signal," vol. 71, pp. 1878–1885, Nov. 1991.

[23] N. Dimitrova and G. Dimitrov, "Interpretation of emg changes with fatigue: Facts, pitfalls, and fallacies," *Journal of Electromyography and Kinesiology*, vol. 13, pp. 13–36, Feb. 2003.

[24] T. Masuda, H. Miyano, and T. Sadoyama, "The propagation of motor unit action potential and the location of neuromuscular junction investigated by surface electrode arrays," *Electroencephalography and Clinical Neurophysiology*, vol. 55, pp. 594–600, May 1983.

On the Semi-Blind Mutually Referenced Equalizers for MIMO Systems

Do Hai Son†, Karim Abed-Meraim‡, Tran Trong Duy†, Nguyen Linh Trung†, Tran Thi Thuy Quynh†
† AVITECH, VNU University of Engineering and Technology, Hanoi, Vietnam
‡ PRISME Laboratory (IUF member), University of Orléans, Orléans, France

***Abstract*—Minimizing training overhead in channel estimation is a crucial challenge in wireless communication systems. This paper presents an extension of the traditional blind algorithm, called "Mutually referenced equalizers" (MRE), specifically designed for MIMO systems. Additionally, we propose a novel semi-blind method, SB-MRE, which combines the benefits of pilot-based and MRE approaches to achieve enhanced performance while utilizing a reduced number of pilot symbols. Moreover, the SB-MRE algorithm helps to minimize complexity and training overhead and to remove the ambiguities inherent to blind processing. The simulation results demonstrated that SB-MRE outperforms other linear algorithms, i.e., MMSE, ZF, and MRE, in terms of training overhead symbols and complexity, thereby offering a promising solution to address the challenge of minimizing training overhead in channel estimation for wireless communication systems.**

***Index Terms*—Channel estimation, semi-blind, MRE, MIMO.**

I. Introduction

Multi-Input Multi-Output (MIMO) communication systems play a vital role in achieving reliable high data rate transmission while improving spectrum efficiency and channel capacity [1]. In order to mitigate channel distortions and separate source signals, MIMO systems employ an increased number of training sequences, usually referred to as pilots, which do not carry any useful data. However, this leads to a decrease in spectrum efficiency. In the 1990s, many studies proposed 'blind' (B) approach algorithms [2] to estimate the wireless channel without any pilot symbol. However, blind methods often have high complexity or require statistical information that is not ready for real-world systems. After all, pilot-based channel estimation techniques have dominated blind methods up to now.

In recent years, there has been a growing interest in the combination of pilot-based and blind methods, known as 'semi-blind' (SB) methods, which aim to reduce the number of pilots while maintaining stable performance [3]. In this work, we focus on the blind "Mutually referenced equalizers" (MRE) [4] algorithm, which is fast, global convergence, and flexible in implementation (i.e., Batch, LMS, RLS, ...). Since 1997, several studies have been proposed to develop the MRE algorithm. [5] firstly proposed recursive least-squares (RLS) implementation for the MRE method. In [6], Gesbert *et al.* introduced the MIMO version of the MRE method. In 2000, J. van der Veen *et al.* [7] proposed the combination of MRE with another blind algorithm, which is "Constant modulus" (CMA) for SIMO systems. [8] also used MRE method for SIMO models but reduced the complexity by computing only 2 instead of K equalizers as the original paper. Generally, most of the studies reviewed above and others tried to develop the MRE algorithm in the blind approach. As explained above, they often have one or more drawbacks, e.g., lack of support for MIMO; the need for intensive computational resources when the number of equalizers is large; and the requirement for several pieces of channel information, i.e., channel order and delay between receivers.

This motivates us to propose the SB-MRE method, a semi-blind approach for the MRE algorithm in MIMO systems. In detail, a few pilot symbols are used to improve the performance of the B-MRE component. To further enhance its effectiveness, we aim to reduce the overall cost of the SB-MRE method through two key improvements. Firstly, we decrease the complexity of the B-MRE component by reducing the number of equalizers to just 2. Secondly, we employ a straightforward adaptive algorithm that effectively minimizes the number of pilot symbols required in the SB-MRE method.

Our contribution in this paper is to propose an effective SB-MRE method for MIMO systems. The system model and B-MRE algorithm for MIMO systems are presented in section II. The novel SB-MRE algorithm is shown in section III. Two methods to reduce the cost of SB-MRE are presented in section IV. At last, we conduct numerical experiments to compare the performance of SB-MRE with other linear estimation methods.

Notations:

$\mathbf{X}^\top$	Transpose matrix of $\mathbf{X}$.
$\mathbf{X}^H$	Hermitian matrix of $\mathbf{X}$.
$\mathbf{I}_K$	Indentity matrix of shape K.
$\otimes$	Kronecker product operator.
$\mathbb{E}()$	Statistical expectation.
$\mathbf{0}$	Matrix of zeros.

II. MIMO-MRE

This section introduces the mathematical model of MIMO wireless communications used in this work. After that, the MRE algorithm for the MIMO model is briefly reviewed.

A. System model

The MIMO model, illustrated in Fig. 1, is composed of T transmitters and L receivers. Each channel between t-

This work was supported by the National Foundation for Science and Technology Development of Vietnam under Grant 01/2019/TN. Correspondence: Nguyen Linh Trung (linhtrung@vnu.edu.vn).

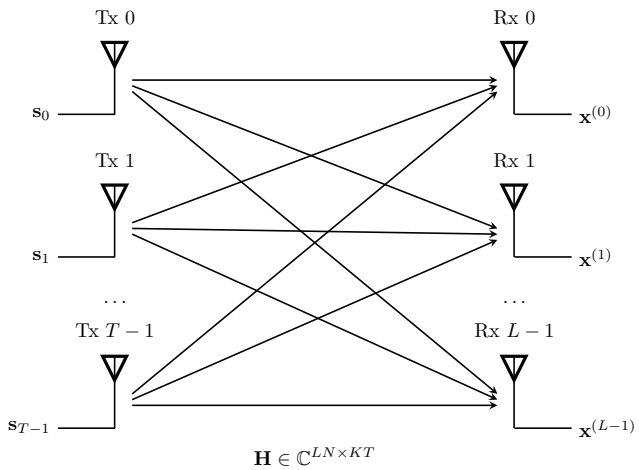


Fig. 1: Conventional MIMO communication model.

th transmitter and l-th receiver is formulated as a $M+1$ coefficients vector. At a time, N received symbols are simultaneously captured on each receiver. At time n, the following equation expresses the system model.

$$\mathbf{x}(n) = \sum_{t=0}^{T-1} \mathbf{H}_t \mathbf{s}_t(n) + \mathbf{w}_t, \tag{1}$$

where $\mathbf{s}_t(n) \in \mathbb{C}^{M+N\times 1}$ is the transmit symbols from the t-th transmitter. $\mathbf{H}_t$ is the channel convolution matrix between t-th transmitter and L receivers. $\mathbf{H}_t \in \mathbb{C}^{LN\times K}$ is assumed to be a full column rank ($K = M+N$) matrix. $\mathbf{x}(n) \in \mathbb{C}^{LN\times 1}$ denotes the observed signals and $\mathbf{w}_t \in \mathbb{C}^{LN\times 1}$ stands for additive white Gaussian noise matrix. Assume that channel and additive noise between each channel are i.i.d and have distributed $\mathcal{CN}(0, \sigma^2_{\mathbf{H}_t}\mathbf{I})$ and $\mathcal{CN}(0, \sigma^2\mathbf{I})$, respectively.

$$\mathbf{s}_t(n) = [s_t(n), s_t(n-1), \ldots, s_t(n-K+1)]^\top,$$

$$\mathbf{H}_t = \begin{pmatrix} h_{t,0}^{(0)} & \cdots & h_{t,M}^{(0)} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & h_{t,0}^{(0)} & \cdots & h_{t,M}^{(0)} \\ \vdots & \cdots & \vdots & \cdots & \cdots & \vdots \\ h_{t,0}^{(L-1)} & \cdots & h_{t,M}^{(L-1)} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & h_{t,0}^{(L-1)} & \cdots & h_{t,M}^{(L-1)} \end{pmatrix} \begin{matrix} \overset{K}{\longleftrightarrow} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \updownarrow LN,$$

$$\mathbf{x}(n) = \big[x^{(0)}(n), \cdots, x^{(0)}(n-N+1), \cdots, x^{(L-1)}(n), \cdots, x^{(L-1)}(n-N+1)\big]^\top.$$

B. Brief review of MRE method

Generally, MRE uses an N-taps linear equalizer to filter each channel. Let $\mathbf{g}_{t,i} \in \mathbb{C}^{LN\times 1}$ be an i-delay equalizer and t-th transmitter. For each transmitter, the number of equalizers equals the maximum delay, which is K. For $i = 0, \ldots, K-1$, at time n, we have

$$\mathbf{g}_{t,i}^H * \mathbf{x}(n) = \sum_{l=0}^{L-1}\sum_{k=0}^{N-1} g_{t,i}^H(k) x^{(l)}(n-k) \approx \mathbf{s}_t(n-i), \tag{2}$$

$$\mathbf{g}_{t,i} = \big[g_{t,i}^{(0)}(0), \ldots, g_{t,i}^{(0)}(N-1), \ldots, g_{t,i}^{(L-1)}(0), \ldots, g_{t,i}^{(L-1)}(N-1)\big]^\top. \tag{3}$$

The equalizers matrix for t-th transmitter is $\mathbf{G}_t \in \mathbb{C}^{LN\times K}$ as follows:

$$\mathbf{G}_t = [\mathbf{g}_{t,0}, \ldots, \mathbf{g}_{t,K-1}]. \tag{4}$$

In the noise-free case, the transmitted symbols can be perfectly recovered with $\bar{\mathbf{G}}$ is any left inverse of $\mathbf{H}$ since

$$\begin{aligned} [\mathbf{G}_0, \ldots, \mathbf{G}_{T-1}]^H \mathbf{x}(n) &= [\mathbf{s}_0^\top(n), \ldots, \mathbf{s}_{T-1}^\top(n)]^\top, \\ \bar{\mathbf{G}}^H \mathbf{x}(n) &= \bar{\mathbf{s}}(n). \end{aligned} \tag{5}$$

In noisy case, to estimate $\bar{\mathbf{G}}$, the MRE method exploits the delay diversity from multi-channel, $\mathbf{g}_i^H \mathbf{x}(n) = \mathbf{g}_{i+1}^H \mathbf{x}(n+1)$, to determine the full set of channel inverses. Where $\mathbf{g}$ is the vector form of $\bar{\mathbf{G}}$ equalizers matrix as shown in Eq. (12). The unconstrained MRE cost function of $\bar{\mathbf{G}}$ is given by:

$$\mathcal{J}(\bar{\mathbf{G}}) = \mathbf{g}^H \mathcal{R} \mathbf{g}, \tag{6}$$

where $\mathcal{R} \in \mathbb{C}^{LNKT\times LNKT}$ is the matrix of $\mathbf{x}(n)$ and $\mathbf{x}(n+1)$ observed signals, which is given by:

$$\mathcal{R} \overset{\text{def}}{=} \mathbb{E}\left(\mathbf{U}^H \mathbf{U}\right), \tag{7}$$

with

$$\mathbf{U} = \big(\mathbf{I}_{T(K-1)}, \mathbf{0}\big) \otimes \mathbf{x}^H(n) - \big(\mathbf{0}, \mathbf{I}_{T(K-1)}\big) \otimes \mathbf{x}^H(n+1). \tag{8}$$

Under the quadratic constraint [4], the unique stable minimum of $\mathbf{g}$ is estimated by selecting the smallest eigenvector of $\mathcal{R}$.

III. Propose SB-MRE

In t-th transmitter, a block data $\mathbf{s}_t$ is considered to send, including N_p pilot symbols and $N_s - N_p$ data symbols.

$$\mathbf{s}_t(n) = [s_t(n), \ldots, s_t(n-N_p+1), s_t(n-N_p), \ldots, s_t(n-N_s+1)]^\top. \tag{9}$$

Pilot signals estimate the full set of channel inverse by the least-square method.

$$\hat{\mathbf{G}} = \arg \min_{\bar{\mathbf{G}} \in \mathbb{C}^{LN\times KT}} \sum_{i=N-1}^{N_p-1} \|\bar{\mathbf{s}}(n) - \bar{\mathbf{G}}^H \mathbf{x}(n)\|_F^2. \tag{10}$$

The combining of pilot-based and blind MRE is a constrained optimization that can readily solve by the Lagrange multiplier method [9]. The total cost function of SB-MRE will be

$$\mathcal{J}(\bar{\mathbf{G}}) = \sum_{i=N-1}^{N_p-1} \|\bar{\mathbf{s}}(n) - \bar{\mathbf{G}}^H \mathbf{x}(n)\|_F^2 + \lambda \mathbf{g}^H \mathcal{R} \mathbf{g}, \tag{11}$$

with λ is a weighting factor, $\mathcal{R}$ in the quadratic form of the blind MRE criterion as shown in Eq. (7), and $\mathbf{g}$ is the vector form of $\bar{\mathbf{G}}$.

$$\begin{aligned} \mathbf{g} = \mathrm{vec}(\bar{\mathbf{G}}) &= \left[\vec{\mathbf{G}}_0^\top, \vec{\mathbf{G}}_1^\top, \ldots, \vec{\mathbf{G}}_{K-1}^\top\right]^\top, \\ \vec{\mathbf{G}}_i &= \left[\ \mathbf{g}_{0,i}^\top, \mathbf{g}_{1,i}^\top, \ldots, \mathbf{g}_{T-1,i}^\top\ \right]^\top. \end{aligned} \tag{12}$$

Without loss of generality, the least-square expression of Eq. (11) is conjugate transposed and the sum operator is turned into matrix forms of $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{X}}$. The cost function is expressed as follows:

$$\begin{aligned} \mathcal{J}(\bar{\mathbf{G}}) &= \sum_{i=N-1}^{N_p-1} \left\|\bar{\mathbf{s}}^H(n) - \mathbf{x}^H(n)\bar{\mathbf{G}}\right\|_F^2 + \lambda \mathbf{g}^H \mathcal{R} \mathbf{g} \\ &= \left\|\widetilde{\mathbf{S}}^H - \widetilde{\mathbf{X}}^H \bar{\mathbf{G}}\right\|_F^2 + \lambda \mathbf{g}^H \mathcal{R} \mathbf{g}. \end{aligned} \tag{13}$$

where $\widetilde{\mathbf{S}}, \widetilde{\mathbf{X}}$ are the matrices of shape $\mathbb{C}^{KT \times (N_p - N + 1)}$ and $\mathbb{C}^{LN \times (N_p - N + 1)}$, respectively.

$$\begin{aligned} \widetilde{\mathbf{S}} &= [\bar{\mathbf{s}}(N-1), \ldots, \bar{\mathbf{s}}\,(N_p - 1)], \\ \widetilde{\mathbf{X}} &= [\mathbf{x}(N-1), \ldots, \mathbf{x}\,(N_p - 1)]. \end{aligned}$$

The least-square expression is vectorized and thanks to the property for vector, i.e., $\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A}) * \mathrm{vec}(\mathbf{X})$. The SB-MRE cost function turned into

$$\begin{aligned} \mathcal{J}(g) &= \left\|\mathrm{vec}(\widetilde{\mathbf{S}}^H) - (\mathbf{I}_{KT} \otimes \widetilde{\mathbf{X}}^H)\,\mathrm{vec}(\bar{\mathbf{G}})\right\|_F^2 + \lambda \mathbf{g}^H \mathcal{R} \mathbf{g} \\ &= \|\bar{\mathbf{s}} - \mathbf{A}\mathbf{g}\|_F^2 + \lambda \mathbf{g}^H \mathcal{R} \mathbf{g} \\ &= \mathbf{g}^H \mathbf{A}^H \mathbf{A} \mathbf{g} + \|\bar{\mathbf{s}}\|_F^2 - 2\Re(\mathbf{g}^H \mathbf{A}^H \bar{\mathbf{s}}) + \lambda \mathbf{g}^H \mathcal{R} \mathbf{g}. \end{aligned} \tag{14}$$

In order to find minimum cost of Eq. (14), let derivative $\mathcal{J}(\mathbf{g})$ with respect to $\mathbf{g}$ as follows:

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{g}}(\mathbf{g}) &= 0, \\ \left(\mathbf{A}^H \mathbf{A} + \lambda \mathcal{R}\right) \mathbf{g} &= \mathbf{A}^H \bar{\mathbf{s}}. \end{aligned} \tag{15}$$

The final equalizers matrix in vector form of the proposed SB-MRE method is obtained through

$$\mathbf{g}_{SB} = \left(\mathbf{A}^H \mathbf{A} + \lambda \mathcal{R}\right)^{-1} \mathbf{A}^H \bar{\mathbf{s}}. \tag{16}$$

IV. Reduce the cost

In the ensuing, we aim to reduce the cost of the proposed SB-MRE algorithm by addressing two key factors, i.e., reducing the complexity of the B-MRE component and minimizing the training overhead.

A. Reducing the complexity of the B-MRE part

In the original study, the overall complexity of the blind MRE method is $\mathcal{O}(LNKT)$ [4]. Although all K equalizers are estimated for each transmitter, only one is ultimately utilized. However, this computational burden becomes unnecessary as N increases. Hence, in this section, we first considerably reduce the number of equalizers to 2, i.e., the 0-th and $(K-1)$-th equalizers. As a result, the overall complexity is reduced to $\mathcal{O}(LNT)$ and the equalizer matrix for the t-th transmitter can be represented as follows:

$$\mathbf{V}_t = [\mathbf{g}_{t,0},\ \mathbf{g}_{t,K-1}]. \tag{17}$$

Followed by the estimated signal source of t-th transmitter will be

$$\mathbf{V}_t^H \mathbf{x}(n) = [s_t(n), s_t(n-K+1)]^\top = \mathbf{s}_t(n). \tag{18}$$

Following that, we do not have to compute the full rank of $\mathcal{R}$ as the blind approach. Eq. (8) is modified to

$$\mathbf{U} = (\mathbf{I}_T, \mathbf{0}) \otimes \mathbf{x}^H(n) - (\mathbf{0}, \mathbf{I}_T) \otimes \mathbf{x}^H(n+K-1).$$

B. Reducing the training overhead for SB-MRE

In the least-square method, as indicated in Eq. (10), the performance of SB-MRE is primarily influenced by the number of pilot symbols (N_p). However, increasing the number of pilot symbols results in a decrease in spectral efficiency. To address this issue, we propose an adaptive algorithm that determines the minimum number of pilots required for the SB-MRE method.

First, we establish an assumption that after a few transmission sessions, the user equipment (UE) provides feedback to the base station (BTS) regarding the average symbol error rate (SER). This assumption is inspired by the feedback of the block error rate (BLER) in 5G standards [10]. Second, taking into account the channel state information (CSI), UE characteristics, and the specific service requirements, a target SER value ($\mathcal{T}$) is determined for a given transmission period. Subsequently, we define a straightforward loss function that measures the deviation from the target SER value as follows:

$$\mathcal{L}(\mathrm{SER}) = \log_{10}(\mathrm{SER}) - \log_{10}(\mathcal{T}). \tag{19}$$

The updating formula for the number of pilots N_p is

$$N_p = N_p + \delta * \mathcal{L}(\mathrm{SER}), \tag{20}$$

where δ is the learning rate.

Initially, the parameter δ is set to 1. If the feedback SER has not yet reached the target SER ($\mathcal{T}$), $\delta = \delta_1$. However, once the target SER is satisfied, the number of pilots is reduced to improve data efficiency, and during this phase, δ remains fixed at 1. The algorithm 1 shows our method to adapt the number of pilot symbols for SB-MRE. This approach can be seen as following a multiplicative-increase additive-decrease algorithm for adjusting the number of pilots. The reverse version, additive-increase multiplicative-decrease (AIMD), has been successfully adopted in networking, particularly the TCP congestion control algorithm [11]. Moreover, this scheme has been observed in nature, employed by biological systems, and even adapted for neural circuits [12].

Algorithm 1 Adaptive number of pilot symbols for SB-MRE.

Input: $\mathcal{T}, \delta_1, \text{SER}$
Output: N_p

$N_p \leftarrow N$
while true **do**
  $\text{SER} = \text{SB-MRE}(N_p)$
  $\mathcal{L}(\text{SER}) = \log_{10}(\text{SER}) - \log_{10}(\mathcal{T})$
  if $\mathcal{L}(\text{SER}) < 0$ **then**
    $\delta = 1$
  else
    $\delta = \delta_1$
  end if
  $N_p = N_p + \delta * \mathcal{L}(\text{SER})$
end while

V. Simulation Results

In this section, we present the experimental analysis of the proposed SB-MRE method using the simulation parameters outlined in Table I. The simulation results are based on an average of 500,000 runs. We begin by comparing the performance of the SB-MRE method against traditional channel estimation algorithms, namely Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) [13], in terms of SER.

Fig. 2 shows that ZF and MMSE outperform the proposed SB-MRE at lower signal-to-noise ratio (SNR) values. This can be attributed to the fact that the influence of the B-MRE component is negligible at low SNR levels. However, as the SNR increases, the proposed SB-MRE gradually catches up to the SER of ZF and MMSE, eventually surpassing them at SNR = 10dB. It is important to note that the proposed SB-MRE algorithm only utilizes 32/256 symbols for pilots, whereas ZF and MMSE require complete knowledge of the CSI. Furthermore, even after reducing the cost of the B-MRE component, the modified SB-MRE (SB-MRE_rc) still outperforms both the original B-MRE and B-MRE_rc in terms of SER. This highlights the effectiveness of the proposed modifications in enhancing the performance of the SB-MRE algorithm.

TABLE I: Simulation parameters

Parameters	Specifications
MIMO	$T = 2, L = 4$
Modulation	QPSK
Channel order	$M = 3$
Windows size	$N = 10$
Sample size	$N_s = 256$
Pilots	$N_p = 32$
Number of blind equalizers	2
Weighting factor	$\lambda = 0.1$
Increase rate of N_p	$\delta_1 = 2$
Target SER	$\mathcal{T} = 10^{-4}, \mathcal{T}_\text{rc} = 10^{-2}$

After that, we simulate to verify the performance of the proposed SB-MRE in different numbers of pilots (N_p) and SNR values. As shown in Fig. 3, N_p and SNR are turned in the range of $[10 \;\; 64]$ pilot symbols and $[5, 10, 15]$ dB, respectively. Overall, SER curves of both SB-MRE and SB-MRE_rc exhibit

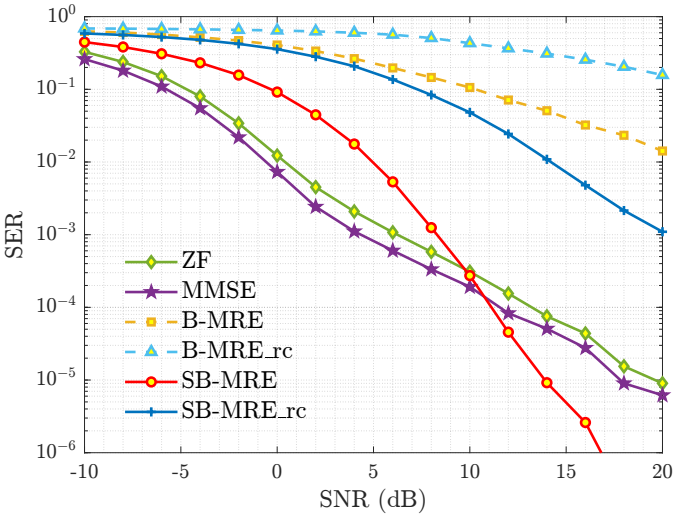


Fig. 2: Performance of proposed SB-MRE versus other channel estimation algorithms.

a gradual decrease as the number of pilot symbols (N_p) and SNR increase. This behavior represents a trade-off between spectrum efficiency and the accuracy of the channel estimation algorithm. At SNR = 15 dB, SB-MRE with $N_p > 40$ archives to SER $\approx 10^{-6}$. In the case of SB-MRE_rc, when the number of pilot symbols (N_p) is increased within the range of 10 to 40, there is a clear improvement in the SER curves. The SER decreases noticeably with an increasing number of pilots, indicating better performance. However, once N_p exceeds 40, the SER curves reach a stable state, and there is little to no further reduction in SER.

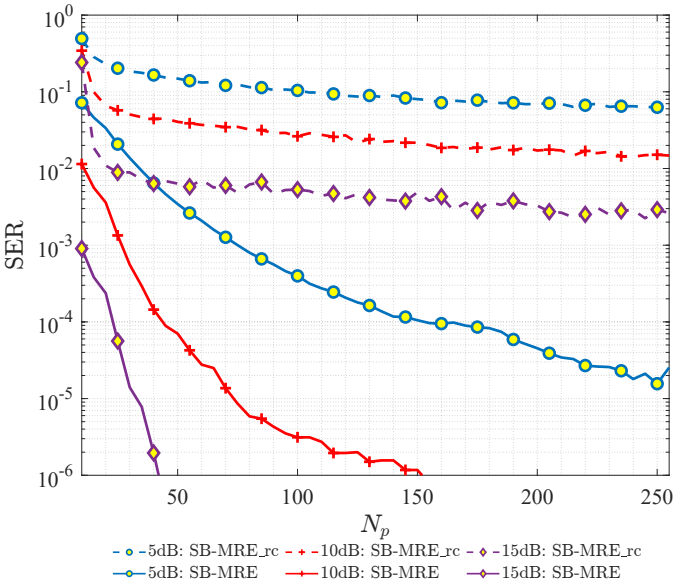


Fig. 3: Performance of proposed SB-MRE under various N_p and SNR values.

Fig. 4 demonstrates the impact of the weighting factor (λ) on the combination of pilot-based and B-MRE. The λ is varied

within the range of $[0.01 \;\; 0.2]$. For lower SNR levels (i.e., 5 dB and 10 dB), we observed a slight reduction in SER curves as λ values increased. However, at SNR $= 15$ dB, the performance of the B-MRE component became prominent, resulting in a significant decrease in the SER curves of SB-MRE. On the other hand, impacts of the weighting factor are negligible in the SB-MRE_rc version due to the small number of blind equalizers employed.

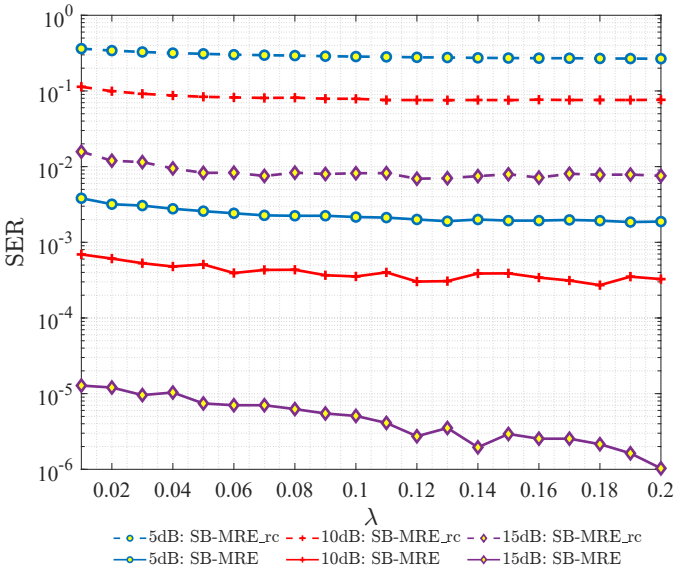


Fig. 4: Performance of proposed SB-MRE under various λ and SNR values.

Finally, the algorithm's performance to reduce the training overhead for SB-MRE is shown in Fig. 5. At SNR $= 12$ dB, the weighting factor (λ) is set at two levels, i.e., 0.1 and 0.15. According to simulation results in Fig. 3, the target SERs of SB-MRE, SB-MRE_rc are fixed at 10^{-4} and 10^{-2}, respectively. Overall, all scenarios' SER curves converge to their target SER after a certain number of iterations. However, the number of pilot symbols and iterations required for convergence vary significantly. For $\lambda = 0.1$, SB-MRE and SB-MRE_rc require 51 and 131 pilot symbols, along with 20 and 70 iterations, respectively. Conversely, for $\lambda = 0.15$, SB-MRE and SB-MRE_rc only need 13 and 15 pilot symbols, along with 1 and 5 iterations, respectively. These results highlight the effectiveness of the adaptive number of pilot symbols method for SB-MRE, with the λ value playing a crucial role in enhancing the performance of the SB-MRE algorithm, especially in high SNR scenarios.

VI. Conclusion

In this paper, we proposed a semi-blind version of the mutually referenced equalizers (MRE) algorithm designed for MIMO systems. Our algorithm uses a small number of pilot symbols to enhance the performance of the B-MRE. Furthermore, the SB-MRE algorithm also reduces the number of blind equalizers and pilot symbols required. Simulation results demonstrate that SB-MRE outperforms other B-MRE

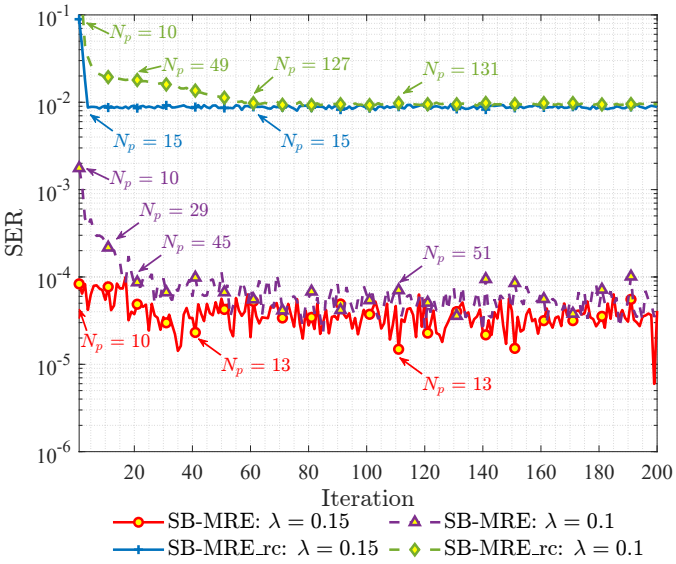


Fig. 5: Adapt the number of pilot symbols method for SB-MRE. The simulation parameters are $\mathcal{T} = 10^{-4}, \mathcal{T}_\text{rc} = 10^{-2}$, and SNR $= 12$ dB.

approaches and linear methods, particularly in high SNR scenarios. Additionally, we analyze the impact of the blind component and highlight the effectiveness of adapting the number of pilot symbols.

References

[1] G. George, R. K. Mungara, A. Lozano, and M. Haenggi, "Ergodic spectral efficiency in mimo cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2835–2849, May 2017.

[2] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug. 1997.

[3] A. Ladaycia, A. Belouchrani, K. Abed-Meraim, and A. Mokraoui, "Semi-blind mimo-ofdm channel estimation using expectation maximisation like techniques," *IET Communications*, vol. 13, no. 20, pp. 3452–3462, Dec. 2019.

[4] D. Gesbert, P. Duhamel, and S. Mayrargue, "On-line blind multichannel equalization based on mutually referenced filters," *IEEE Transactions on Signal Processing*, vol. 45, no. 9, pp. 2307–2317, Sept. 1997.

[5] D. Gesbert and P. Duhamel, "Unimodal blind adaptive channel equalization: an rls implementation of the mutually referenced equalizers," in *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, Paris, France, Apr. 1997, pp. 29–32.

[6] D. Gesbert, A. Paulraj, and P. Duhamel, "Blind joint multiuser detection using second-order statistics and structure information," in *Proceedings of 40th Midwest Symposium on Circuits and Systems*, vol. 2, Sacramento, CA, USA, Aug. 1997, pp. 1252–1255.

[7] A.-J. van der Veen and A. Trindade, "Combining blind equalization with constant modulus properties," in *34th Asilomar Conference on Signals, Systems and Computers*, vol. 2, Pacific Grove, CA, USA, Oct. 2000, pp. 1568–1572.

[8] C. Yu and L. Xie, "On recursive blind equalization in sensor networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 662–672, Feb. 2015.

[9] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[10] "Monitor of bler performance," ETSI, Standard 3GPP TS 32.425 version 16.2.0 Release 16, Mar. 2019.

[11] D.-M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1–14, Jun. 1989.

[12] J. Y. Suen and S. Navlakha, "A feedback control principle common to several biological and engineered systems," *Journal of The Royal Society Interface*, vol. 19, no. 188, p. 20210711, Mar. 2022.
[13] Y. Jiang, M. K. Varanasi, and J. Li, "Performance analysis of zf and mmse equalizers for mimo systems: An in-depth study of the high snr regime," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2008–2026, Apr. 2011.

Accurate continuous action and gesture recognition method based on skeleton and sliding windows techniques

Viet-Duc Le*, Thi-Lich Nghiem*†, Thi-Lan Le*
* School of Electrical and Electronic Engineering (SEEE),
Hanoi University of Science and Technology (HUST), Vietnam
Email: lan.lethi1@hust.edu.vn
† MIS and E-commerce faculty, Thuongmai University, Hanoi, Vietnam

***Abstract*—Vision-based action recognition with the main aim to determine the action class from image sequence has emerged recently thanks to its potential applications. Numerous methods have been proposed for action recognition. However, majority of them mainly focus on isolated action recognition with the assumption that videos are segmented in different clips or shots, each of them contains an image sequence of one sole action. Different from isolated action recognition, continuous action recognition involves recognizing human actions from a continuous stream of input data without requiring pre-segmentation into discrete action instances. In this paper, we propose a method for continuous action recognition. The proposed method incorporates sliding window technique and a light weight action classification model named DDNet. We evaluate the proposed approach on several benchmark datasets. Experimental results show that the proposed method the accuracy of 0.95, 0.97 and 0.71 for IPN, UOW and InHARD datasets respectively on isolated action recognition task. On continuous recognition task, it has achieved a performance of 0.57, 0.76 and 0.33 on IPN, UOW and InHARD datasets.**

***Index Terms*—skeleton-based action recognition, continuous action recognition**

I. Introduction

The ability to recognize and understand human actions in real-time has important applications in various domains, including surveillance, human-robot interaction, and healthcare. In surveillance, distracting moments of drivers, and illegal house intrusion can be detected [1], [2]. In human-robot interaction, gesture and action recognition can enable robots to respond to human commands and intentions, facilitating natural and intuitive communication [3] while in healthcare domain, action recognition can be used to monitor and assist patients through physical rehabilitation [4].

Recent advances in deep learning have enabled significant progress in action recognition, leveraging approaches, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and graph-based models. These methods have demonstrated encouraging outcomes in the identification of diverse human actions within video sequences.

However, majority of existing work focus on isolated action recognition with the assumption that videos are segmented in different clips or shots, each of them contains an image sequence of one sole action. Different from isolated action recognition, continuous action recognition involves recognizing human actions from a continuous stream of input data without requiring pre-segmentation into discrete action instances.

Compared with isolated recognition, recognizing continuous actions poses a notably greater challenge as it has to tackle to variation of actions' length, unclear boundary between actions.

In this paper, we introduce a novel approach designed for continuous action recognition. The proposed method incorporates sliding window technique and a light weight action classification model named DDNet [5]. We assess the performance of our approach on various benchmark datasets and conduct a comparative analysis with state-of-the-art approaches, demonstrating its effectiveness and robustness in continuous action recognition tasks.

The rest of this paper is organized as follows: Section 2 provides a review of the related work in human action recognition. Section 3 describes the detailed information of the proposed approach. Section 4 shows our experiments on three different datasets of continuous action and gesture recognition and analyze in detail the experimental results. Conclusion and future works are illustrated in Section 5.

II. Related Works

In this section, we provide an overview of the related work on action recognition, highlighting the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

key contributions and limitations of existing methods. The study on video-based action recognition can be categorized into two fundamental groups: isolated and continuous action recognition.

Isolated action recognition takes split videos that contain only a single instance of action as inputs and predicts the action label. Many approaches have been proposed for isolated action recognition. Some authors used CNNs to capture local spatial patterns and temporal dynamics, thereby effectively modeling the complex motion characteristics of actions such as [6]. Due to the limitation of the single-frame image data input, some authors proposed to use 3D CNN to learn the temporal information for action recognition including different fusion approaches, spatio-temporal attention mechanisms, and dual-stream 3D-CNN [7], [8]. In addition, the method based on Long Short-Term Memory (LSTM) and CNN that have been proposed to recognize actions quickly and accurately [9]. Nayan *et al.* [10] proposed a combined RNN that uses CNN to extract features for a bidirectional LSTM (BD-LSTM). While CNN, 3D CNN and LSTM based methods mainly work on RGB sequence, different GCN models have been proved to be robust for skeleton-based action recognition.

Recently, some methods have been proposed to address continuous action recognition from RGB [11] or skeleton sequences [12]. These methods built a detector to determine whether an images sequence contains an action or not and a classifier that aims to identify the action class. Building a detector is tedious task as it requires to prepare a training set that consists of image sequences for non action/non gesture. Moreover, some methods could not guarantee real time processing. In [13], the authors propose a model consisting of a Spatio-Temporal GCN and the sliding window method for continuous action recognition from skeleton sequence. A sliding window of a predetermined size was applied to a continuous stream of skeleton data, and it was shifted as time advanced. This method allows to accelerate the recognition speed, however, GCN is still complex model that requires computational resources. In [12], a light model has been introduced for continuous hand gesture recognition. However, the model is designed for IPN dataset and could not work on other datasets.

III. Proposed method

A. Sliding Window

The proposed strategy, which is shown in Fig. 1, is divided into two stages: training and inference. In the training step, a light weight model based on DDNet is trained on isolated sequences of different actions. Then, in the inference step, sliding window technique is incorporated to recognize the dynamic gestures from continuous stream of video. The window size denoted ws needs to be set to a fixed number same as the input of DDNet. The number of frames in the sliding window is will be chosen for each dataset depending on the frame rate and the length of gestures. This window strides every 1 frame and returns the predicted label for the middle frame of the window.

B. DDNet

DDNet (a Double-Feature Double-Motion network) is a light weight model proposed for skeleton-based human action recognition [5]. Although being a light weight model, DDNet and its variants provides state-of-the-art results on different datasets [2].

In Fig. 2, we conducted a performance comparison of recognition models using the MSRAction dataset [14] in term of both FLOPs and accuracy. We can observe that DD-Net results a good recognition rate while requires less computational time compared to GCN model.

DDNet consists of three branches that aims at extracting both spatial and temporal features from skeleton sequence. There branches are Slow motion, Fast motion and Joint Collection Distances (JCD). Each branch contains three CNN layers. Slow and Fast motions are designed to capture the movement of the actions while JCD is dedicated to spatial information.

Denote S^{t+1} and S^{t+2} are the skeleton of one frame and two frames behind the skeleton at frame t S^t, F^t_{slow} and F^t_{fast} that are denoted as the Slow Motion the Fast Motion and are computed as follows:

$$F^t_{slow} = S^{t+1} - S^t, t \in 1, 2, 3, ..., T-1; \quad (1)$$

$$F^t_{fast} = S^{t+2} - S^t, t \in 1, 3, ..., T-2; \quad (2)$$

Concerning JCD, it is extracted from a Euclidean distance matrix. Each element of the matrix represents the Euclidean distance between a pair of joints in the same frame. At frame t, JCD feature can be computed by using the following equation.

$$JCD^t = \begin{bmatrix} \left\|\overrightarrow{\mathbf{j_2^t j_1^t}}\right\| & & & \\ \vdots & \ddots & & \\ \vdots & \cdots & \ddots & \\ \left\|\overrightarrow{\mathbf{j_N^t j_1^t}}\right\| & \cdots & \cdots & \left\|\overrightarrow{\mathbf{j_N^t j_{N-1}^t}}\right\| \end{bmatrix} \quad (3)$$

where $\left\|\overrightarrow{\mathbf{j_k^t j_l^t}}\right\| (k \neq l)$ is the Euclidean distance between two joints $\mathbf{j_k^t}$ and $\mathbf{j_l^t}$ of the skeleton S^t.

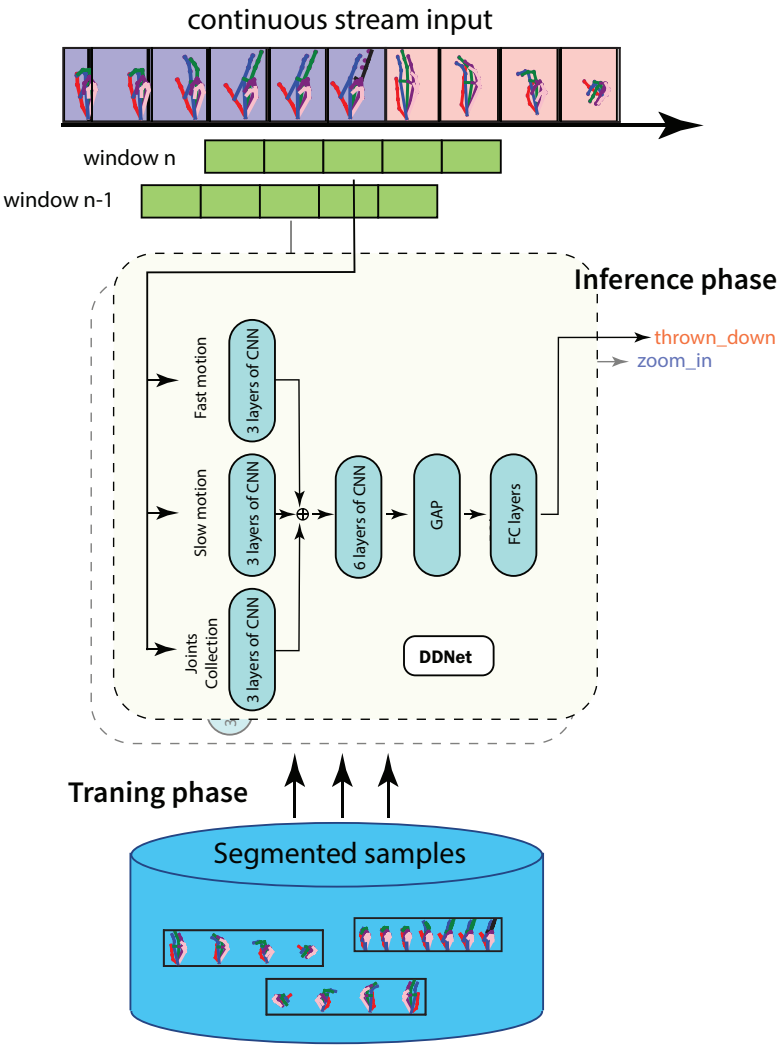


Fig. 1: Proposed method for skeleton-based continuous human action recognition

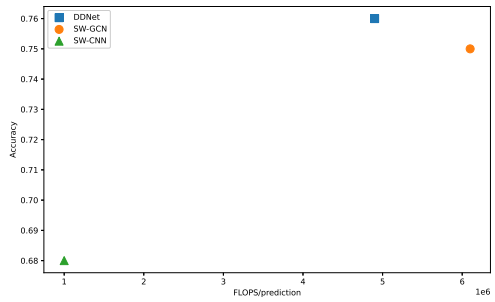


Fig. 2: Comparison of recognition models on MSRAction dataset [14]

TABLE I: Accuracy obtained by the proposed method on 3 datasets for isolated classification task.

	IPN	UOW	InHARD
Acc	0.95	0.97	0.71

TABLE II: Comparison of the proposed method with state-of-the-art methods on three benchmark datasets

	IPN	UOW	InHARD
SW-CNN [13]	-	0.68	-
SW-GCN [13]	-	0.75	-
Proposed method	**0.57**	**0.76**	**0.33**

IV. Experimental Results

A. Datasets

We conducted our experiments on the following continuous gesture datasets, which is widely used for evaluating continuous action recognition algorithms. **IPN** [11]: This dataset includes over 4,000 gesture samples and 800,000 RGB frames from 50 different individuals. 13 different static and dynamic gestures were created to focus on interactions with touchless screens. The main characteristic of the dataset is that the continuous gestures are performed without tran-

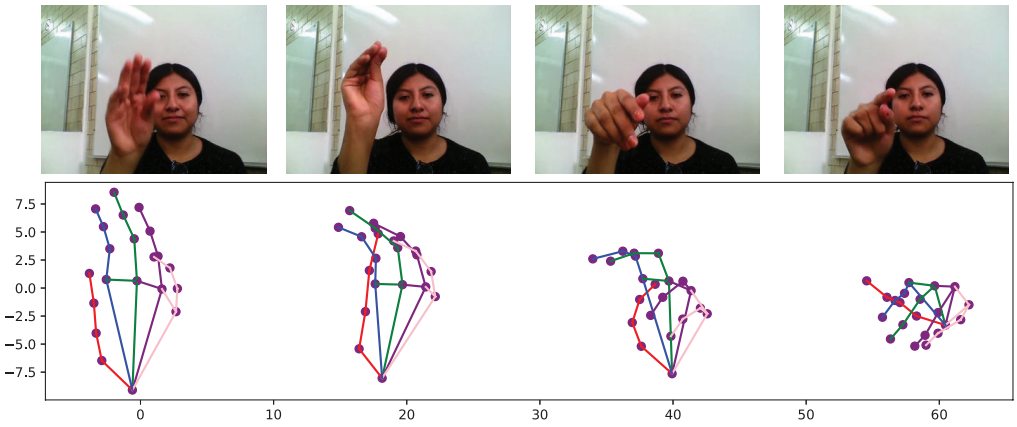


Fig. 3: RGB images sequence and its corresponding skeleton sequence of ”Throw down” action in IPN dataset

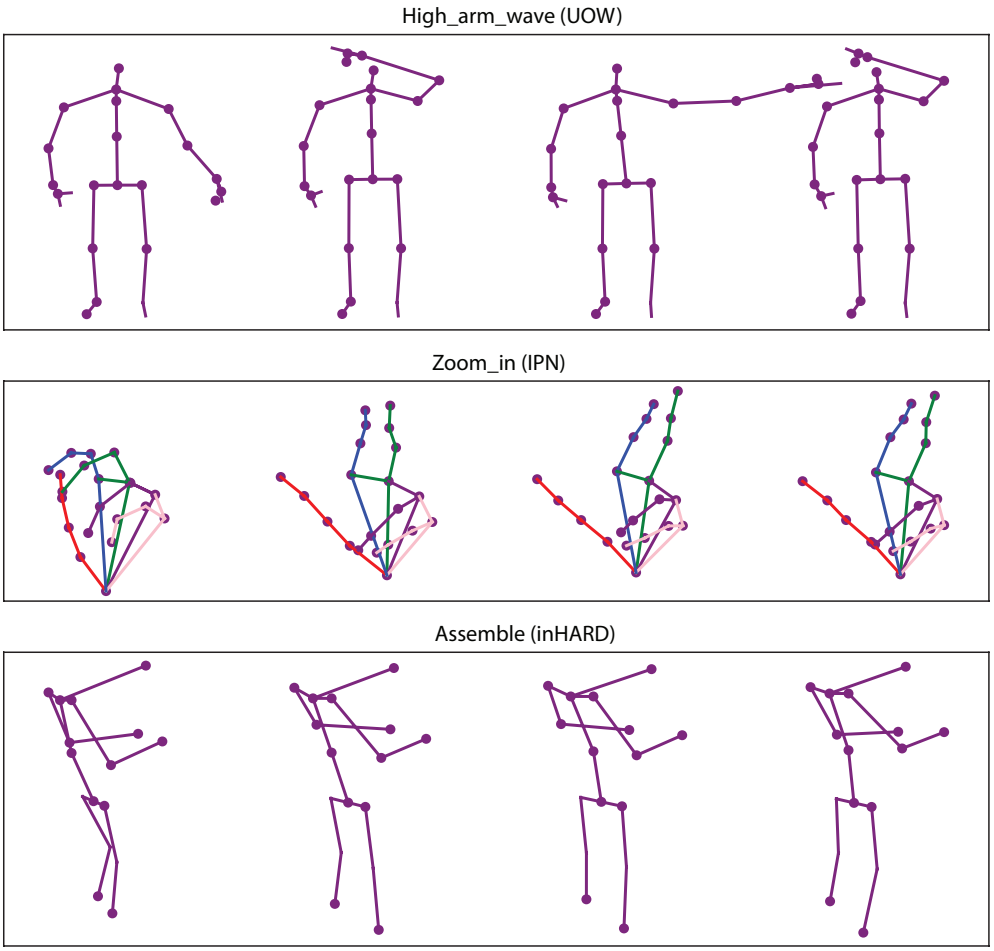


Fig. 4: Example of skeleton sequences of UOW, IPN and InHARD datasets

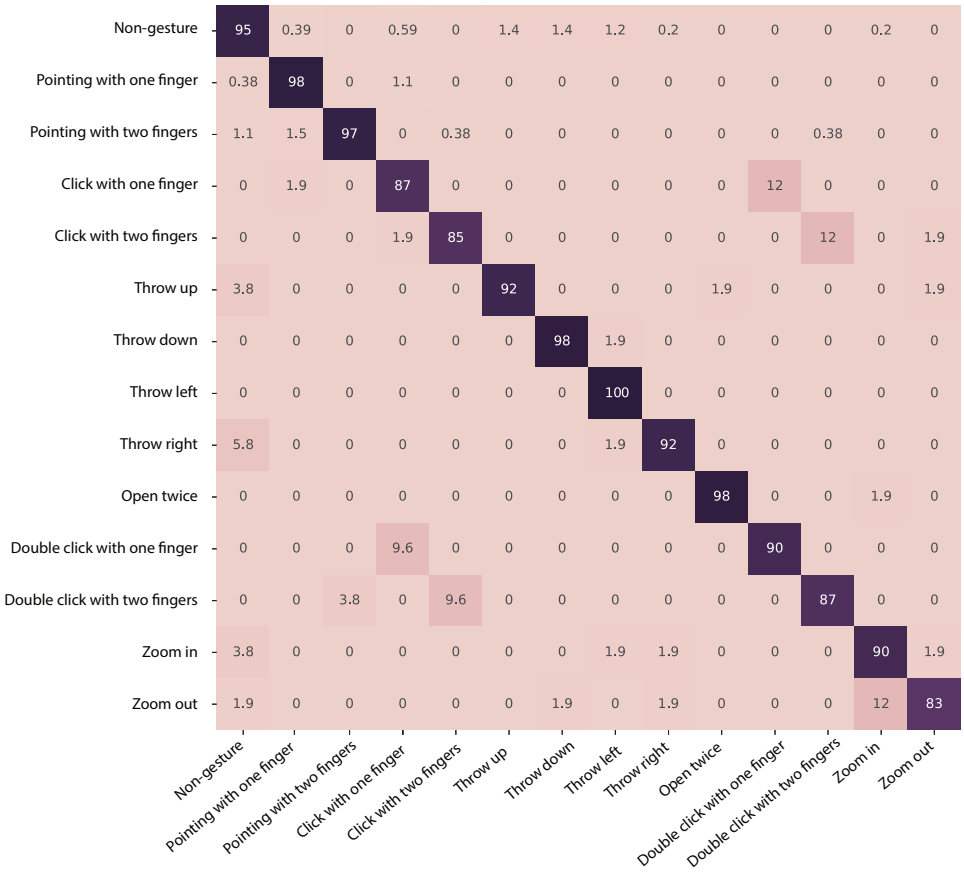


Fig. 5: Isolated action classification on IPN dataset

sition states and non-gesture actions are performed by subjects using natural hand movements. Gestures were captured in about 30 different settings with various backgrounds and lighting conditions to introduce real-world variation. Since the skeleton information is not available on this dataset, we estimate hand joint landmarks from provided video by using MediaPipe [15], an open-source cross-platform framework developed by Google that provides a pipeline for building real-time applications.

UOW: [16] The UOW Online Action3D Dataset contains action sequences of 20 actions. These actions are defined in MSR Action3D Dataset. The dataset were captured using Microsoft Kinect V.2 and comprises videos of 20 subjects. Each subject performed one action from 3 to 5 times before performing 20 actions continuously in a random order. The dataset has two subsets: a subset of continuous action sequences can be utilized for evaluation of continuous action recognition methods while the subset of repeated action sequences will be used for isolated action recognition or training purposes.

InHARD (Industrial Human Action Recognition Dataset): [17] has recently released for human-COBOT interaction context. This dataset contains over 4800 action samples that perform 13 different industrial action classes named meta-action. The authors use a motion sensor system to capture the human skeleton at 120 Hz. Figure 4 shows example of skeleton sequence of these datasets. It is worth to note that the number of joints for IPN, UOW and InHARD datasets are 21, 25 and 17 respectively.

B. Implementation details and evaluation metrics

The size of the window depends on the frame rate and gesture length of each dataset. We set the window size of UOW and IPN is 60 and of inHARD is 240. All experiments were conducted on a PC with an Intel Core i7 8750H processor.

To assess the performance action recognition system, we employ the video-wise accuracy for isolated recognition and frame-wise accuracy for continuous recognition. Frame-wise accuracy compares the predicted label vs the ground-truth label at frame level.

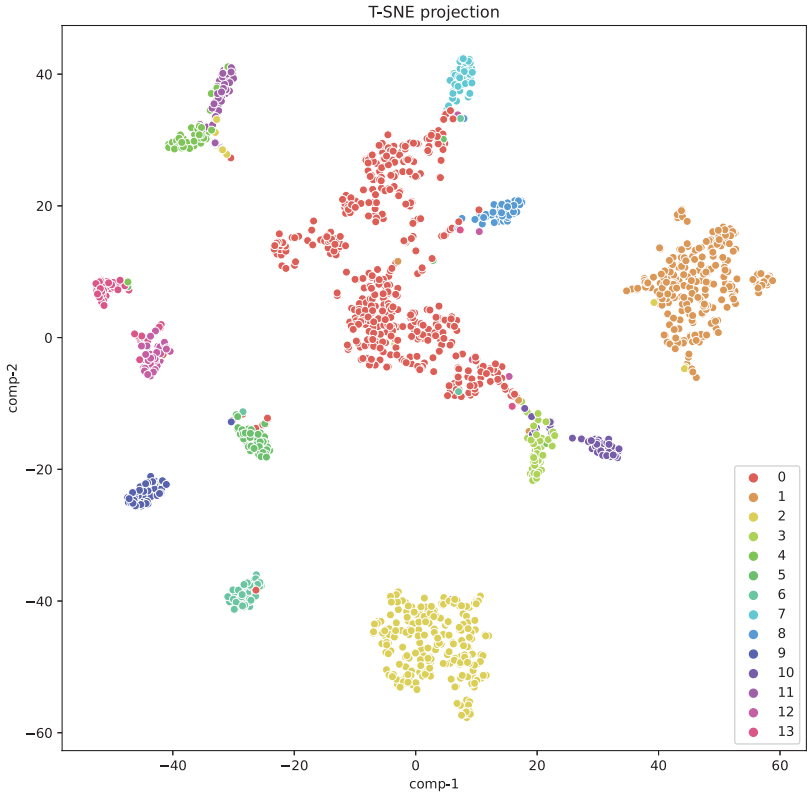


Fig. 6: Feature space learned from DDNet model for IPN dataset

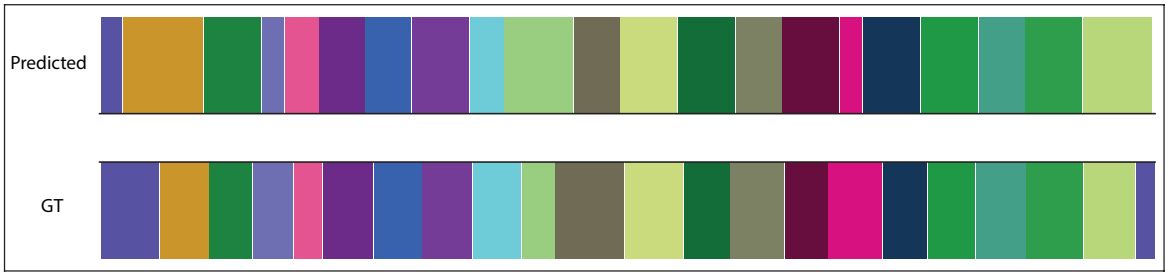


Fig. 7: Illustration of continuous action recognition on a sequence of UOW dataset. The first line represents the results of the proposed method while the second line shows the corresponding ground-truth. Each action label is represented by an identical color.

C. *Results*

As the proposed method consists of two main components: the sliding window technique and an action classifier based on DDNet. We evaluate first the performance of action classifier. Table I shows the accuracies obtained for three benchmark datasets. The accuracies obtained for IPN, UOW and InHARD are 0.95, 0.97 and 0.71 respectively. We can observe that the model based on DDNet could distinguish different actions with high accuracy. Among three datasets, the performance obtained for InHARD is lowest. The main reason is that in InHARD contains 74 actions belonging to 13 meta-actions, each type of meta-action has many variations of actions, therefore, it is hard to achieve the high accuracy in classification task. For IPN dataset, even with skeleton extracted by off the shelf method [15] and the high inter-similarity between action classes, the method still could recognize very well the actions. Figure 3 illustrates the RGB sequence and its corresponding skeleton sequence for ”Throw down” action of IPN dataset. Figure 5 shows the confusion matrix obtained for IPN while Fig. 6 shows the feature distribution of 13 actions in IPN dataset using t-SNE method. We can observe that the ”Click” and ”Double click” actions get a lot of confusions between 1 and 2

fingers. This can be explained by the similarity in the movement. In some cases, 1 of 2 fingers can be obscured leading Mediapipe only estimating 1 finger. Table II compares the performance of the proposed method with CNN and GCN based methods in [13] on three benchmark datasets: IPN, UOW, and InHARD. The two first row represents results of the SW-CNN, the SW-GCN method [13], which has achieved a performance of 0.68, 0.75 in term of frame-wise accuracy on the UOW dataset. The last row represents the proposed method, which has achieved a performance of 0.57 on the IPN dataset, 0.76 on the UOW dataset, and 0.33 on the InHARD dataset.

Figure 7 illustrates recognition for a continuous sequence in UOW dataset. The first line represents the proposal's results whereas the second line shows the corresponding ground-truth. Each action label is represented by an identical color. From this figure, we can observe that the capability of the proposed method to accurately localize and recognize different action instances from input sequence. However, for some actions such as the two first segments present the transition from non-gesture to "pick up & throw" gesture. Because the sliding window returns the label of the middle frame, in this case, skeletons in the window contains both non-gesture frame and the starting frame of "pickup" gesture. Therefore, it is recognize as pick up gesture.

V. Conclusion and Future Works

In this paper, a method for continuous action recognition from a skeleton sequence has been proposed. The proposed method incorporates sliding window technique and a light weight action classification model. Based on our experiments conducted on three benchmark datasets, the proposed method achieved remarkable accuracy scores of 0.95, 0.97, and 0.71 for the IPN, UOW, and InHARD datasets, respectively. These results pertain specifically to the isolated action recognition task. On continuous recognition task, it has achieved a performance of 0.57, 0.76 and 0.33 on IPN, UOW and InHARD datasets. Although the proposed method obtains promising results for isolated recognition task, there is still room for improvement for continuous recognition especially when working with challenging dataset such as InHARD.

Acknowledgment

This work was funded by the project "Research, design and manufacture of COBOT applied in industry and some other fields with human-robot interaction", grant number: KC- 4.0-35/19-35.

References

[1] H. Yan, B. Hu, G. Chen, and E. Zhengyuan, "Real-time continuous human rehabilitation action recognition using openpose and fcn," in *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 2020, pp. 239–242.

[2] T.-T. Nguyen, D.-T. Pham, H. Vu, and T.-L. Le, "A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation," *IET Computer Vision*, vol. 16, no. 8, pp. 709–726, 2022.

[3] Q. Gao, J. Liu, Z. Ju, and X. Zhang, "Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9663–9672, 2019.

[4] Q.-T. Pham, D.-A. Nguyen, T.-T. Nguyen, T. Nguyen, D.-T. Nguyen, D.-T. Pham, T. Tran, T. Le, and V. Hai, "A study on skeleton-based action recognition and its application to physical exercise recognition," in *Proceedings of the 11th International Symposium on Information and Communication Technology*, 12 2022, pp. 239–246.

[5] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM Multimedia Asia*. Association for Computing Machinery, 2020.

[6] S. Shinde, A. Kothari, and V. Gupta, "Yolo based human action recognition and localization," *Procedia Computer Science*, vol. 133, pp. 831–838, 2018, international Conference on Robotics and Smart Manufacturing (RoSMa2018).

[7] F. Liu and Z. Wu, "A kind of human action recognition algorithm based on Kinect and SVM," *Modern Computer*, vol. 18, pp. 55–58, 2019.

[8] N. M. and D. S., "Deep ensemble network using distance maps and body part features for skeleton based action recognition," *Pattern Recognition*, vol. 100, p. 107125, 2020.

[9] T.-H. Tan, C.-J. Huang, M. Gochoo, and Y.-F. Chen, "Activity recognition based on FR-CNN and attention-based LSTM network," in *2021 30th Wireless and Optical Communications Conference (WOCC)*, 2021, pp. 146–149.

[10] N. Nayan, D. Ghosh, and P. M. Pradhan, "A CNN bi-LSTM based multimodal continuous hand gesture recognition," in *2022 IEEE India Council International Subsections Conference (INDISCON)*, 2022, pp. 1–4.

[11] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai, "IPN hand: A video dataset and benchmark for real-time continuous hand gesture recognition," *CoRR*, vol. abs/2005.02134, 2020.

[12] T. Nguyen, N. Nguyen, D. Ngo, V. Phan, M. Pham, D. Nguyen, M. Doan, and T. Le, "A continuous real-time hand gesture recognition method based on skeleton," in *11th International Conference on Control, Automation and Information Sciences, ICCAIS, 2022*, 2022, pp. 273–278.

[13] M. Delamare, C. Laville, A. Cabani, and H. Chafouk, "Graph convolutional networks skeleton-based action recognition for continuous data stream: A sliding window approach," in *International Conference on Computer Vision Theory and Applications*, 2021.

[14] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 9–14.

[15] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," Jun. 2019.

[16] C. Tang, W. Li, P. Wang, and L. Wang, "Online human action recognition based on incremental learning of weighted covariance descriptors," *Information Sciences*, vol. 467, pp. 219–237, 2018.

[17] M. Dallel, V. Havard, D. Baudry, and X. Savatier, "InHARD-industrial human action recognition dataset in the context of industrial collaborative robotics," in *International Conference on Human-Machine Systems (ICHMS)*, 2020, pp. 1–6.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Transformer-Based Deep Learning Detector for Dual-Mode Index Modulation 3D-OFDM

Toan Gian[†], Tien-Hoa Nguyen[†]*, Trung Tan Nguyen[‡], Van-Cuong Pham[§] and Thien Van Luong[§]
[†] School of Electrical and Electronics Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
[‡] Faculty of Radio-Electronics, Le Quy Don Technical University, Ha Noi 11355, Vietnam
[§] Faculty of Computer Science, Phenikaa University, Hanoi 12116, Vietnam
Corresponding author*: Tien Hoa Nguyen (email: hoa.nguyentien@hust.edu.vn)

***Abstract*—In this paper, we propose a deep learning-based signal detector called TransD3D-IM, which employs the Transformer framework for signal detection in the Dual-mode index modulation-aided three-dimensional (3D) orthogonal frequency division multiplexing (DM-IM-3D-OFDM) system. In this system, the data bits are conveyed using dual-mode 3D constellation symbols and active subcarrier indices. As a result, this method exhibits significantly higher transmission reliability than current IM-based models with traditional maximum likelihood (ML) detection. Nevertheless, the ML detector suffers from high computational complexity, particularly when the parameters of the system are large. Even the complexity of the Log-Likelihood Ratio algorithm, known as a low-complexity detector for signal detection in the DM-IM-3D-OFDM system, is also not impressive enough. To overcome this limitation, our proposal applies a deep neural network at the receiver, utilizing the Transformer framework for signal detection of DM-IM-3D-OFDM system in Rayleigh fading channel. Simulation results demonstrate that our detector attains to approach performance compared to the model-based receiver. Furthermore, TransD3D-IM exhibits more robustness than the existing deep learning-based detector while considerably reducing runtime complexity in comparison with the benchmarks.**

***Index Terms*—TransD3D-IM, deep learning, BER, DNN, dual-mode, index modulation, DM-IM-3D-OFDM.**

I. Introduction

Combining orthogonal frequency division multiplexing (OFDM) with index modulation (IM) methods, namely OFDM-IM [1], [2] has gained dominant attention as a potential approach to substitute the traditional OFDM technique. In OFDM-IM, a subset of subcarriers is activated to transmit data through both subcarriers and their corresponding indices. As a result, the OFDM-IM approach outperforms for both reliability and energy efficient compared to traditional methods thanks to using indices of active subcarriers to convey data. Additionally, adjusting the number of activated subcarriers promises to offer an flexible trade-off between the error performance and spectral efficiency (SE).

In recent years, various advanced OFDM-based systems have been introduced [3]–[6], aiming at enhancing either the error performance or SE. Additionally, the dual-mode OFDM (DM-OFDM) [7] and three-dimensional OFDM (3D-OFDM) [8] are also known as the efficiency approach in reducing the bit error rate (BER) of the traditional OFDM. In particular, DM-OFDM utilizes multiple constellations to convey data using subcarriers that are not activated. In comparison with OFDM-IM, DM-OFDM has shown superior in higher achievable SE. A generalized variant of DM-OFDM which greatly improves the SE of the traditional version in [9]. In the process of 3D-OFDM, data bits are converted to the 3D constellation symbols, and the obtained set of symbols is divided into subcarriers, which are employed the 2D inverse fast Fourier transform for modulation. By employing OFDM with a 3D mapper instead of a 2D one, the minimum Euclidean distance among different constellations notably increases. These enhancement results in error performance thereby reduce the likelihood of transmission errors. In [10] and [11], authors presented the analysis for symbol error probability (SEP) and the solution to decrease the peak-to-average power ratio of 3D-OFDM. Nevertheless, its SE is only one-third compared to the classical OFDM when considering the constellation order [12]. Motivated by the IM idea, a novel scheme combined by 3D-OFDM [8] and DM-OFDM [7], namely dual-mode IM-aided 3D-OFDM (DM-IM-3D-OFDM), was introduced in [13] to enhance SE of 3D-OFDM. Within this model, data bits are conveyed via indices of activated subcarriers and the 3D mapper constellations based on the floor of the Poincare sphere [14]. In particular, subcarriers are divided into subblocks, and subcarriers in each subblock are split into two groups, and then mapped by dual-mode 3D constellations. In comparison with 3D-OFDM, DM-IM-3D-OFDM demonstrates superior to transmission reliability at large SNRs when employing the maximum likelihood (ML) detection under the fading channels. However, its computational complexity is high, especially as parameters of model are large. Even the complexity of the Log-Likelihood Ratio (LLR) algorithm, known as a low-complexity detector for signal detection in the DM-IM-3D-OFDM system, is also not impressive enough. Consequently, we propose a new detector to solve this fundamental issue.

Recently, Deep Learning (DL) [15] has been broadly applied to the physical layer of wireless communication systems [16]–[21]. Furthermore, the DL-based detector successfully employs the Transformer framework [22] for signal detection in the OFDM-IM system, especially when the model parameters are large [23]. To the best of our knowledge, the application of DL Transformer architecture to the DM-IM-3D-OFDM has been overlooked in the literature. In this contribution, our proposal focuses on a novel Transformer-based DL detector called TransD3D-IM for signal detection of the DM-IM-3D-OFDM

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

system. In particular, we build a new deep neural network (DNN) model of TransD3D-IM based on the Transformer framework. In our proposal, the Transformer block plays a crucial role as it calculates global dependencies and features with the 3D signal, which allows to obtain BER performance improvement. Aiming of minimizing BER, TransD3D-IM is trained offline by employing the simulation dataset. After that, it plays the role of an online detector to quickly estimate the conveyed bits. Our simulation results pointed out that TransD3D-IM can achieve an approach BER performance and much lower runtime complexity than the previous detectors.

The remaining parts of the paper are organized as follows. Section II describes the system model of DM-IM-3D-OFDM and conventional ML and LLR detectors. Section III presents the structure and model training of our proposed TransD3D-IM detector. Subsequently, Section IV shows the results achieved by our transformer-based detector and compares it with conventional detectors. Lastly, Section V is conclusion.

II. SYSTEM MODEL

The diagram in Fig. 1 illustrates the structure of a DM-IM-3D-OFDM simplified system, where data are conveyed by mapping 3D constellations and subcarrier indices. It is assumed that the message consisting of b bits are split into m subblocks of n subcarriers and p bits, i.e. $m = b/p = N/n$, where N is the number of subcarriers per subblock. Herein, the p bits are split into two components based on the dual 3D constellation-IM scheme. The first component, referred to as the index bit, consists of p_1 bits. The index bit serves as a selector to divide the subcarrier indices into two subsets. The second component, known as the information bits, consists of p_2 bits and is mapped to two distinct constellations: type A and type B, denoted by $\mathcal{S}_A$ and $\mathcal{S}_B$, respectively. Specifically, symbol sets of $\mathcal{S}_A$ and $\mathcal{S}_B$ are represented by $\mathcal{M}_A$ and $\mathcal{M}_B$ that associated with the size of s_A and s_B, respectively, note that $\mathcal{M}_A \cap \mathcal{M}_B = \varnothing$. In the frequency domain, each element of $\mathcal{M}_A$ can be expressed by $\mathcal{M}_A^j = (x, y, z)^T$, i.e., j-th symbol from the symbol set $\mathcal{M}_A$ where $1 \leq j \leq s_A$ and x, y, z are coefficients of base vectors in $x-$, $y-$, and $z-$axis. In addition, $(.)^T$ presents the transposition. Value of k stands for the number of subcarriers modulated by $\mathcal{S}_A$, thus the number of subcarriers mapped to $\mathcal{S}_B$ is $n - k$. Accordingly, we can determine the value of p as follows.

$$p = p_1 + p_2 = \lfloor \log_2 C_n^k \rfloor + \log_2((s_A)^k) + \log_2((s_B)^{n-k}), \quad (1)$$

where C_n^k is the binomial coefficient and $\lfloor . \rfloor$ denotes the integer floor operator. Additionally, the index selectors of the two 3D mapper constellations have an interdependent relationship. Specifically, once the index selector of $\mathcal{S}_A$ is determined, the corresponding index selector of $\mathcal{S}_B$ is also known. Therefore, it is sufficient to define the index selector of $\mathcal{S}_A$ in order to describe the principle of the index selector. For instance, it can be selected that $n = 4$, $k = 2$, and $s_A = s_B = 2$ for each OFDM subblock. The IM procedure is presented in Table I.

TABLE I
THE LOOK-UP TABLE FOR DM-IM-3D-OFDM SYSTEM

IM Bits	Indices for $\mathcal{S}_A$	Subblocks
$[0, 0]$	$[1, 2]$	$[\mathcal{S}_A^1, \mathcal{S}_A^2, \mathcal{S}_B^1, \mathcal{S}_B^2]$
$[0, 1]$	$[2, 3]$	$[\mathcal{S}_B^1, \mathcal{S}_A^1, \mathcal{S}_A^2, \mathcal{S}_B^2]$
$[1, 0]$	$[3, 4]$	$[\mathcal{S}_B^1, \mathcal{S}_B^2, \mathcal{S}_A^1, \mathcal{S}_A^2]$
$[1, 1]$	$[1, 4]$	$[\mathcal{S}_A^1, \mathcal{S}_B^1, \mathcal{S}_B^2, \mathcal{S}_A^2]$

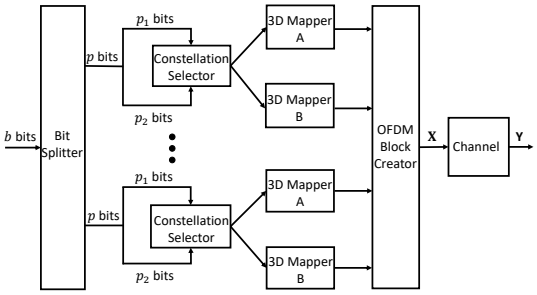


Fig. 1. Block diagram of the DM-IM-3D-OFDM system.

In the frequency domain, the transmitted signal $\mathbf{X} \in \mathbb{C}^{3\times n}$ in a block is constructed by $\mathbf{X} = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)]$, where $\mathbf{X}(i) \in \{\mathcal{S}_A, \mathcal{S}_B\}$ and $1 \leq i \leq n$. Particularly, the output of two mappers can be represented by $\mathcal{S}_A = [\mathcal{S}_A^1, \ldots, \mathcal{S}_A^k]$, where $\mathcal{S}_A^i = \mathcal{M}_A^j$ $(1 \leq i \leq n, 1 \leq j \leq s_A)$ and $\mathcal{S}_B = [\mathcal{S}_B^1, \ldots, \mathcal{S}_B^{n-k}]$, where $\mathcal{S}_B^i = \mathcal{M}_B^j$ $(1 \leq i \leq n-k, 1 \leq j \leq s_B)$. Noting that the process of designing 3D constellations to built-up the signal matrix $\mathbf{X}$ is presented in detail in [13], the index selector may be obtained by the look-up table as shown in Table I.

At the receiver, the obtained signal matrix $\mathbf{Y} \in \mathbb{C}^{3\times n}$ in the frequency domain can be represented as follows.

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{G}, \quad (2)$$

where $\mathbf{H} = [\mathbf{H}_1; \mathbf{H}_2; \mathbf{H}_3]^T \in \mathbb{C}^{3\times n}, \mathbf{H}_i \in \mathbb{C}^{1\times n}$ and $\mathbf{G} = [\mathbf{G}_1; \mathbf{G}_2; \mathbf{G}_3]^T \in \mathbb{C}^{3\times n}, \mathbf{G}_i \in \mathbb{C}^{1\times n}$ $(1 \leq i \leq 3)$, represent the fading channel coefficients and the additive white Gaussian noise (AWGN), respectively. Each entry of $\mathbf{H}$ and $\mathbf{G}$ are identically and independently distributed as a complex-valued random variable, with $\mathbf{H}_i(\alpha) \sim \mathcal{CN}(0, 1)$ and $\mathbf{G}_i(\alpha) \sim \mathcal{CN}(0, N_0)$ $(1 \leq \alpha \leq n)$, respectively, where N_0 is the noise variance.

At the received side, the ML detector is employed and the estimated signal denoted by $\hat{\mathbf{X}}$ is calculated by

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X}\in\delta} \| \mathbf{Y} - \mathbf{H}\mathbf{X} \|^2, \quad (3)$$

where $\| . \|^2$ is the squared Euclidean norm and δ stands for the set of all possible candidate delivery vectors modulated with $\mathcal{S}_A$ and $\mathcal{S}_B$. While ML can yield good performance, it becomes impractical as the values of p_1 and k increase due to its exponential growth. Moreover, $\hat{\mathbf{X}}$ can be recovered through the LLR detector by considering that the subcarriers are modulated by either $\mathcal{S}_A$ or $\mathcal{S}_B$, the LLR algorithm provides the ratio of the posterior probabilities, which can be expressed by

$$\delta_i = \ln\left(\frac{\sum_{j=1}^{s_A} P(\mathbf{X}(i)=\mathcal{M}_A^j) \mid \mathbf{Y}(i)}{\sum_{j=1}^{s_B} P(\mathbf{X}(i)=\mathcal{M}_B^j) \mid \mathbf{Y}(i)}\right), \quad (4)$$

where $1 \le i \le n$, $\mathcal{M}_A^j \in \mathcal{M}_A$ and $\mathcal{M}_B^j \in \mathcal{M}_B$. As observed in (4), the value of δ_i determines whether the i-th subcarrier is modulated by $\mathcal{S}_A$ or $\mathcal{S}_B$. For instance, if δ_i is larger, it is more probably that the i-th subcarrier is modulated by $\mathcal{S}_A$, whereas if δ_i is smaller, it is more probably to be modulated by $\mathcal{S}_B$. The LLR detector provides near-optimal performance by making hard decisions, while its complexity is significantly lower than the ML detector.

III. Proposal of TransD3D-IM Detector

This section first describes the operation of the TransD3D-IM detector and provides a comprehensive overview of its structure. Lastly, the offline training procedure is presented for our proposed detector.

A. Outline of TransD3D-IM

Firstly, we present a description of the TransD3D-IM detection process. As depicted in Fig. 2, the received signal and the channel $\mathbf{H}$ are fed to a pre-processing module. The resulting matrix from this module passed through a linear fully connected (FC) layer to produce a matrix with a higher row dimension. Following the linear FC layer, a Transformer block is employed, comprising a multi-head attention (MHA) layer, a multi-layer perceptron (MLP), and layer normalization operations. By performing a sequence of computations and projections within aforementioned DNN blocks, the TransD3D-IM input is transformed into the matrix of identical dimensions, which effectively represents the extracted dependencies and features from the input data. The neural networks have the output data, represented as intermediate matrix $\mathbf{O}$, providing soft probabilities for various symbols before next manipulations. Based on the intermediate matrix $\mathbf{O}$, the indices set of two mapper constellations are selected by the last column of this matrix. Then, the symbols corresponding to their mapped 3D constellation are recovered based on both matrix $\mathbf{O}$ and the previously recovered indices. By combining the process index and symbol parts, the transmitted data could be accurately recovered.

B. Structure of TransD3D-IM

Next, we present the core details of each DNN module in the proposed TransD3D-IM detector.

Pre-Processing: Pre-Processing module takes the received signal $\mathbf{Y}$ and the channel matrix $\mathbf{H}$ as inputs and implements zero-forcing (ZF) to build-up an equalized signal vector $\bar{\mathbf{Y}}$ at the receiver, which is presented by

$$\bar{\mathbf{Y}} = \mathbf{Y}\mathbf{H}^{-1}. \quad (5)$$

The received signal's energy matrix represented as $\mathbf{Y}_E$ is computed to realize various subcarriers clearly. We then concatenate $\bar{\mathbf{Y}}_R$, $\bar{\mathbf{Y}}_I$ and $\mathbf{Y}_E$ in parallel and get a matrix $\mathbf{I} = [\bar{\mathbf{Y}}_R, \bar{\mathbf{Y}}_I, \mathbf{Y}_E]$ as the input data of the Linear FC layer, where $\bar{\mathbf{Y}}_R$ and $\bar{\mathbf{Y}}_I$ represent the real and imaginary components of the vector $\bar{\mathbf{Y}}$, respectively.

Linear Fc layer: Next, the matrix $\mathbf{I}$ enters a linear FC layer, where employing the rectifier linear unit (ReLU), $f_{\text{ReLU}}(x) = \max(0, x)$ as the activation function. This module is implemented to get a matrix $\mathbf{C}$ with a higher row dimension. The resulting matrix $\mathbf{C}$ can be expressed as follows.

$$\mathbf{C} = f_{\text{ReLU}}(\mathbf{I} \times \mathbf{W}_{\text{C}} + \mathbf{b}_{\text{C}}), \quad (6)$$

where $\mathbf{W}_{\text{C}}$ and $\mathbf{b}_{\text{C}}$ represent weights and biases to be learned, respectively.

Transformer Block: Transformer Block plays a crucial role in our proposed detector. This block is employed to transform vector $\mathbf{C}$ into another matrix, effectively representing the extracted dependencies and features of the transmitted data bits. The details descriptions for every module are described along the following lines.

• *Multi-Head Attention Layer (MHAL):* MHA is derived by the single-head self-attention mechanism, which involves the calculation for queries of $\mathbf{Q}$, keys of $\mathbf{K}$, values of $\mathbf{V}$ [22] and the matrix $\mathbf{E}$. These components are computed as follows.

$$\begin{cases} \mathbf{Q} = \bar{\mathbf{C}} \times \mathbf{W}_{\text{Q}} + \mathbf{b}_{\text{Q}}, \\ \mathbf{K} = \bar{\mathbf{C}} \times \mathbf{W}_{\text{K}} + \mathbf{b}_{\text{K}}, \\ \mathbf{V} = \bar{\mathbf{C}} \times \mathbf{W}_{\text{V}} + \mathbf{b}_{\text{V}}, \end{cases} \quad (7)$$

Next, the matrix $\mathbf{E}$ of self-attention is computed by

$$\mathbf{E} = \text{Softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}, \quad (8)$$

where $\bar{\mathbf{C}}$ is the resulting matrix of Norm $(\mathbf{C})$ that stands for the function of layer normalization. $\mathbf{W}_{\text{Q}}$,$\mathbf{W}_{\text{K}}$,$\mathbf{W}_{\text{V}}$,$\mathbf{b}_{\text{Q}}$,$\mathbf{b}_{\text{K}}$, and $\mathbf{b}_{\text{V}}$ are weights and biases to be optimized likewise, respectively. Softmax $(\cdot)$ is employed as an activation function to implement normalization. In this paper, we employ a two-head attention layer. Therefore, the attention mechanism operated on two project versions of $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V}$, producing intermediate attention values of $\mathbf{E}_1$ and $\mathbf{E}_2$, as described in Fig. 3. The final attention value is obtained by cascading and projecting them once again, which is expressed by

$$\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2] \times \mathbf{W}_{\text{E}} + \mathbf{b}_{\text{E}}, \quad (9)$$

where $\mathbf{W}_{\text{E}}$ and $\mathbf{b}_{\text{E}}$ represent weights and bias to be optimized of projection, respectively.

• *Multilayer Perceptron (MLP):* This module comprises two FC sublayers, and ReLU is utilized as the activated function for the first FC sublayer. The output of MLP can be calculated by

$$\mathbf{A} = f_{\text{ReLU}}(\text{W}_{\text{A}_1} \times \bar{\mathbf{E}} + \text{W}_{\text{b}_1}) \times \text{W}_{\text{A}_2} + \text{b}_{\text{A}_2}, \quad (10)$$

where $\bar{\mathbf{E}}$ is the resulting matrix of Norm $(\mathbf{E})$, W_{A_1}, W_{A_2}, b_{A_1}, and b_{A_2} are params of the first and second FC sublayers, respectively.

• *Layer Normalization:* Layer normalization serves the purpose of stabilizing the distribution of data features, leading to faster model convergence as well as reduced training duration.

Output layer: In this layer, the Sigmoid function, $f_{\text{Sigmoid}}(x) = \frac{1}{1+e^{-x}}$ is adopted as the activation function.

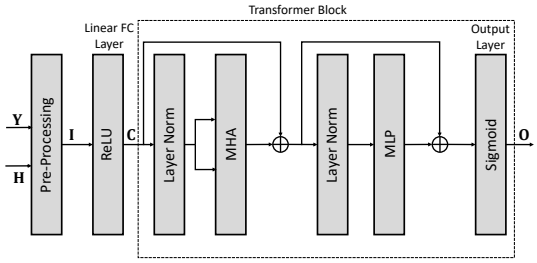


Fig. 2. Diagram of the proposed TransD3D-IM detector.

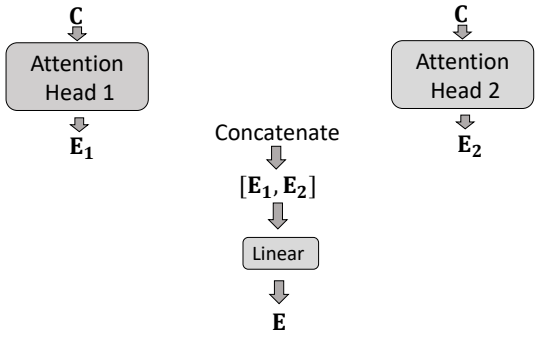


Fig. 3. Diagram of two-head attention.

Accordingly, the obtained matrix $\mathbf{O}$ of the output layer can be expressed by

$$\mathbf{O} = f_{\text{Sigmoid}}(\mathbf{W}_{\text{O}} \times \mathbf{A} + \mathbf{b}_{\text{O}}), \tag{11}$$

where $\mathbf{W}_{\text{O}}$ and $\mathbf{b}_{\text{O}}$ are weights and bias to be learned, respectively. The matrix of $\mathbf{O} = \{\mathbf{O}_1, ..., \mathbf{O}_{s_A+1}\}$ has the size of $[n \times (s_A+1)]$, where $\mathbf{O}_i \in \mathbb{R}^n$ with $1 \leq i \leq s_A+1$, provides soft probabilities for the corresponding symbols and the indices set of two mapper constellations based on the output layer of TransD3D-IM, which utilize Sigmoid as the active function that maps outputs in the interval (0,1). In fact, the proposed detector relies on the mechanism of the LLR-aided detection to recovered both indices and data separately corresponding to $\mathbf{O}$.

C. Model Training

Implementing into online TransD3D-IM detector is required sufficiently large offline training. In particular, various p-bit sequences are randomly generated and converted to the corresponding conveyed bits $\mathbf{X}$ according to the system model in Section II. Next, the processed signals are passed through the Rayleigh fading channel, where it is subject to interference from random AWGN noise. Channel and noise matrices are also randomly created and changed from one-bit sequence to another, according to their established statistical models. At the receiver, the obtained signal and CSI are pre-processed to obtain the training data. The conveyed signals $\mathbf{X}$ are transformed to the corresponding one-hot labels to enhance the accuracy of symbol recovery at the receiver side. For instance, suppose that sending a DM-IM-3D-OFDM subblock $[\mathcal{S}_A^1, \mathcal{S}_A^2, \mathcal{S}_B^1, \mathcal{S}_B^2]$ with via four subcarriers in Table I, the resulting one-hot label has the following form

$$\hat{\mathbf{O}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}. \tag{12}$$

Here, the row and column numbers of $\hat{\mathbf{O}}$ represent the number of subcarriers n and the constellation size $\mathcal{S}_A$ and $\mathcal{S}_B$, i.e., s_A and s_B, respectively. In particular, the first two columns indicate the order of 3D symbols in their corresponding constellations. Meanwhile, the last column shows the index selectors of the two mapper constellations. The index selector $\mathcal{S}_A$ corresponds to positions with a value of 0, while positions with a value of 1 indicate an index selector of $\mathcal{S}_B$.

The TransD3D-IM model performs a classification task to reduce the discrepancy among one-hot labels $\hat{\mathbf{O}}$ and the resulting matrix $\mathbf{O}$ of neural networks. Therefore, the binary cross-entropy (BCE) function is adopted as the loss function as follows.

$$\ell(\mathbf{O}, \hat{\mathbf{O}}; \theta) = -\hat{\mathbf{O}} \log_2(\mathbf{O}) - (1 - \hat{\mathbf{O}}) \log_2(1 - \mathbf{O}), \tag{13}$$

where θ is parameters to be learned of the model. The Adam optimizer [24] is employed and the learning rate is denoted by η. Aiming at achieving the best result, the training SNR level, presented as λ_{train}, must be carefully selected for training TransD3D-IM since it directly affects the detection performance.

IV. Simulation Results

In this section, we provide BER performance comparisons and runtime complexity of TransD3D-IM with its competitors, particularly ML, LLR, and the existing DL-based detector, term as DuaIM-3DNet [25]. As for TransD3D-IM, the linear FC layer has $q_{\text{linear FC}}$ nodes. We consider a two-head attention layer and implement MLP with two FC layers, which have q_{MLP} nodes. Meanwhile, the quantity of the hidden nodes of two subnets in the DuaIM-3DNet are denoted by q_{IndexNet} and $q_{\text{SymbolNet}}$, respectively. To ensure that DL-based detectors have fair performance, q_{IndexNet} and $q_{\text{SymbolNet}}$ have been used parameters as in [25], which are considered optimal parameters of DuaIM-3DNet model. These DL-aided detectors are trained for 100 epochs with every epoch consisting of 600 batches. The batch size used during training is set to 1000 epochs. The learning rate of $\eta = 0.001$ is applied for all parameters. In this paper, we consider the case with $(n, k) = 4, 2$ and use two types of params for all scenarios, namely Scenario 1 and 2, as shown in Table II. Our analyses are considered in scenarios with perfect CSI conducted in the Rayleigh fading channel where each entry is distributed by the complex Gaussian $\mathcal{CN}(0, 1)$.

A. BER Performance

In Fig. 4, we evaluate BER of TransD3D-IM in comparison with the current detectors versus SNR for Scenario 1 under perfect CSI channel condition. The DL-based detectors have been trained with $\lambda_{train} = 15$ dB. It can be seen that the

TABLE II
PARAMETERS FOR DNN-BASED DETECTORS

	s_A	s_B	DuaIM-3DNet		TransD3D-IM	
			q_{IndexNet}	$q_{\text{SymbolNet}}$	$q_{\text{Linear FC}}$	q_{MLP}
Scenario 1	2	2	768	768	32	128
Scenario 2	4	4	1000	1000	64	256

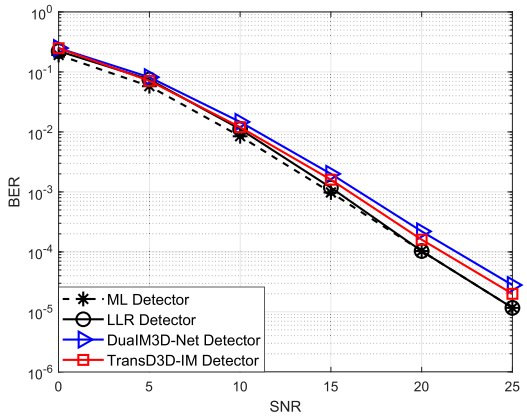


Fig. 4. BER performance of our proposed TransD3D-IM detector and the existing ones for Scenario 1.

proposed TransD3D-IM detector attains the BER performance approaching the model-based detectors and is slightly superior to the DuaIM-3DNet. For instance, the gap SNR between the Transformer-aided DL detector, LLR, and ML is negligible with less than 1 dB and 2 dB at BER of 10^{-3}, respectively.

Fig. 5 shows the BER performance of TransD3D-IM comparisons to its competitors versus SNR for Scenario 2 under certain CSI conditions. As shown in Fig. 5, we observed that TransD3D-IM achieves BER performance approximate to that of the model-aided detectors while significantly outperforming the existing DL-based detector. Considering at BER of 10^{-2}, the proposed detector is only inferior to ML and LLR by about 1 dB. Meanwhile, it has an edge over DuaIM-3DNet by about 5 dB. This indicates that differences in performance between various DL-aided detectors are increasingly evident. Our simulation results point out that TransD3D-IM can achieve nearly optimal BER performance in comparison with model-based receivers. Furthermore, our proposal exhibits more robustness than the current DL-aided detector while TransD3D-IM has significantly lower runtime complexity compared to its equivalents as detailed in the next section.

B. Runtime Complexity

We now compute the runtime complexity required for signal detection of the proposed detector and its competitors. All candidates are simulated using MATLAB on the same computer. The DNN model parameters are employed as presented in the previous subsection and its running time is calculated in online phase. We show the comparison of the resulting runtime of all models in Table III. It is clear that TransD3D-IM requires much less runtime than its counterpart under the same condition. Specifically, the runtime of TransD3D-IM is 460 times less than that of ML and 5 times less than that of LLR and DuaIM-3DNet detectors. This unequivocally validates the advantage of our proposal in terms of runtime.

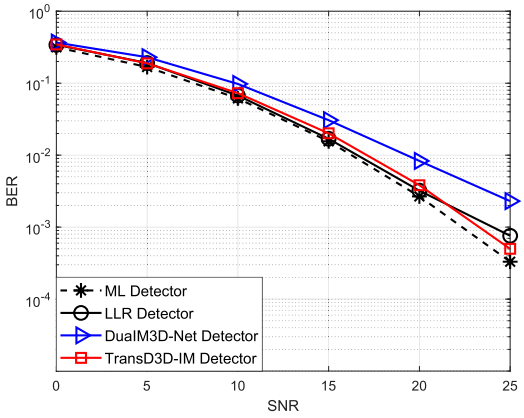


Fig. 5. BER performance of the proposed TransD3D-IM and existing detectors for Scenario 2.

TABLE III
RUNTIME OF PROPOSED TRANSD3D-IM AND ITS COMPETITORS IN MILLISECONDS

(n, k, s_A, s_B)	ML	LLR	DuaIM-3DNet	TransD3D-IM
(4,2,4,4)	12	0.134	0.136	0.026

V. CONCLUSIONS

In this work, we have put forward the TransD3D-IM detector for the DM-IM-3D-OFDM system. Specifically, we build a novel DNN model for TransD3D-IM based on the Transformer framework. Our proposal can comprehensively leverage the global dependencies and features of the obtained signal to produce nearly optimal BER performance. Under the Rayleigh fading channel with CSI certainty, TransD3D-IM can attain approximate BER performance in comparison with the model-based receiver. Besides, the proposed detector exhibits more robustness than the existing DL-aided method, especially when the model system parameters increase. Meanwhile, TransD3D-IM requires lower runtime complexity than its competitors. We believe that with such advantages, Transformer-based detectors may be an effective solution for signal detection in communication systems in the future.

REFERENCES

[1] E. Basar, U. Aygolu, E. Panayirci, and H. V. Poor, "Orthogonal frequency division multiplexing with index modulation," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5536–5549, 2013. DOI: 10.1109/TSP.2013.2279771.

[2] T. V. Luong and Y. Ko, "Impact of CSI uncertainty on MCIK-OFDM: tight, closed-form symbol error probability analysis," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1272–1279, Feb. 2018, ISSN: 0018-9545. DOI: 10.1109/TVT.2017.2691778.

[3] Y. Ko, "A tight upper bound on bit error rate of joint OFDM and multi-carrier index keying," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1763–1766, 2014. DOI: 10.1109/LCOMM.2014.2347280.

[4] T. V. Luong and Y. Ko, "Symbol error outage performance analysis of MCIK-OFDM over complex TWDP fading," in *Proc. Eur. Wireless*, May 2017, pp. 1–5.

[5] E. Basar, "On multiple-input multiple-output OFDM with index modulation for next generation wireless networks," *IEEE Trans. Signal Process.*, vol. 64, no. 15, pp. 3868–3878, 2016. DOI: 10.1109/TSP.2016.2551687.

[6] T. V. Luong and Y. Ko, "The BER analysis of MRC-aided greedy detection for OFDM-IM in presence of uncertain CSI," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 566–569, Aug. 2018, ISSN: 2162-2337. DOI: 10.1109/LWC.2018.2797066.

[7] T. Mao, Z. Wang, Q. Wang, S. Chen, and L. Hanzo, "Dual-mode index modulation aided OFDM," *IEEE Access*, vol. 5, pp. 50–60, 2017. DOI: 10.1109/ACCESS.2016.2601648.

[8] S. G. Kang, "An OFDM with 3-D signal mapper and 2-D IDFT modulator," *IEEE Commun. Lett.*, vol. 12, no. 12, pp. 871–873, 2008. DOI: 10.1109/LCOMM.2008.081197.

[9] M. Wen, E. Basar, Q. Li, B. Zheng, and M. Zhang, "Multiple-mode orthogonal frequency division multiplexing with index modulation," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3892–3906, 2017. DOI: 10.1109/TCOMM.2017.2710312.

[10] Z. Chen, E. C. Choi, and S. G. Kang, "Closed-form expressions for the symbol error probability of 3-D OFDM," *IEEE Commun. Lett.*, vol. 14, no. 2, pp. 112–114, 2010. DOI: 10.1109/LCOMM.2010.02.091347.

[11] H. Huang and L. Zhang, "3-D constellation extension-aided PAPR suppressing for OFDM systems," in *Proc 10th WCSP*, 2018, pp. 1–6. DOI: 10.1109/WCSP.2018.8555865.

[12] Z. Xiang-li, W. Yong, W. Dia-nhong, and C. Zhenx-ing, "Improvement of four-dimensional signal constellations and corresponding OFDM system model design," *ACTA ELECTONICA SINICA*, vol. 48, no. 8, 1486, pp. 1486–1492, 2020.

[13] M. Wang, Z. Chen, and Z. Chen, "Dual-mode index modulation aided 3D-OFDM," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 612–616, 2022. DOI: 10.1109/LCOMM.2021.3119252.

[14] Z. Chen and S. G. Kang, "Three-dimensional modulation formats with constant power for optical communications," *Opt. Express*, vol. 19, no. 23, pp. 22 358–22 363, Nov. 2011. DOI: 10.1364/OE.19.022358.

[15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015, ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2014.09.003.

[16] T. V. Luong, Y. Ko, N. A. Vien, D. H. N. Nguyen, and M. Matthaiou, "Deep learning-based detector for OFDM-IM," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1159–1162, 2019. DOI: 10.1109/LWC.2019.2909893.

[17] T. Van Luong, Y. Ko, N. A. Vien, M. Matthaiou, and H. Q. Ngo, "Deep energy autoencoder for noncoherent multicarrier MU-SIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3952–3962, 2020. DOI: 10.1109/TWC.2020.2979138.

[18] T. V. Luong, Y. Ko, M. Matthaiou, N. A. Vien, M. .-. Le, and V. .-. Ngo, "Deep learning-aided multicarrier systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2109–2119, 2021. DOI: 10.1109/TWC.2020.3039180.

[19] T. Van Luong, X. Zhang, L. Xiang, *et al.*, "Deep learning-aided optical IM/DD OFDM approaches the throughput of RF-OFDM," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 212–226, 2022. DOI: 10.1109/JSAC.2021.3126080.

[20] T. V. Luong, N. Shlezinger, C. Xu, T. M. Hoang, Y. C. Eldar, and L. Hanzo, "Deep learning based successive interference cancellation for the non-orthogonal downlink," *IEEE Trans. Veh. Technol.*, pp. 1–13, 2022. DOI: 10.1109/TVT.2022.3193201.

[21] T. Gian, V.-D. Ngo, T.-H. Nguyen, T. T. Nguyen, and T. Van Luong, "Deep neural network-based detector for single-carrier index modulation NOMA," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1805–1809. DOI: 10.23919/APSIPAASC55919.2022.9980150.

[22] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[23] D. Zhang, S. Wang, K. Niu, J. Dai, S. Wang, and Y. Yuan, "Transformer-based detector for OFDM with index modulation," *IEEE Communications Letters*, vol. 26, no. 6, pp. 1313–1317, 2022. DOI: 10.1109/LCOMM.2022.3158734.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] D.-Y. Hoang, T.-H. Nguyen, V.-D. Ngo, T. T. Nguyen, N. C. Luong, and T. Van Luong, "Deep learning-based signal detection for dual-mode index modulation 3D-OFDM," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1793–1797. DOI: 10.23919/APSIPAASC55919.2022.9979861.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

GAFormer: Wearable IMU-Based Human Activity Recognition with Gramian Angular Field and Transformer

Trung-Hieu Le*‡, Thai-Khanh Nguyen*, Trung-Kien Tran§, Thanh-Hai Tran*, Cuong Pham†

* School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
‡ Dainam University, Hanoi, Vietnam
§ Military Information Technology Institute, Hanoi, Vietnam
† Department of Computer Science, Posts and Telecom Institute of Technology, Hanoi, Vietnam
Email: hai.tranthithanh1@hust.edu.vn

***Abstract*—Recognizing human activities (HAR) from wearable motion sensors has widely practical applications due to its low cost, convenient use, and scalability. Plenty of works utilize 1D CNN or RNN to capture temporal information from time series, while others take advantage of 2D CNN architectures that can effectively handle spatial correlation, and distinctive recognition features, which are really useful for recognition tasks. This paper proposes GAFormer, a method, which exploits the latter, for human activity recognition from wearable motion sensors. GAFormer transforms the raw IMU data into a Gramian Angular Difference Field (GADF) image, which encodes the pair-wise angles between different sensor readings to capture the temporal dynamics and relationships among the sensor measurements. Next, a transformer model is employed to extract visual features from GADF images effectively. In addition, we adapted a state-of-the-art transformer CoAtNet as the backbone of GAFormer. GAFormer is evaluated over two published datasets C-MHAD and GesHome. With the accuracies of 98% on C-MHAD and 95.5% on GesHome's subset, GAFormer demonstrates that combining the GADF technique and a transformer model could be feasible and promising for motion sensor-based action recognition.**

***Index Terms*—Human activity recognition, inertial sensors, features extraction, Convolutional Neural Networks, Transformers, Gramian Angular Field.**

I. Introduction

Human activity recognition (HAR) has been an active research topic as it has diverse applications including, just a few to name, healthcare [1], security monitoring [2], and human-machine interaction [3]. Previously, cameras and motion sensors were commonly utilized to capture human activities. Although camera-based HAR has achieved a significant stride, it suffers several limitations such as privacy concerns and the recognition performance relies on the lighting conditions. The latter, however, has shown the feasibility of recognizing human activities everywhere while alleviating privacy invasion from the users. In addition, thanks to the proliferation of electronic devices and their low costs, motion sensors can be embedded in devices such as phones, watches, cars, and furniture. And they can be part of applications such as uninterruptedly and non-intrusively logging human motion information [4]. Therefore, human action recognition using wearable motion sensors (accelerometer and gyroscope sensors) inherits the advantages of mobility, compact size, low power requirements, and low cost.

In the literature, a number of methods for HAR from wearable motion sensors have been proposed. Traditional methods extracted hand-crafted statistical features from raw time series data and then input them to a classifier (e.g. SVM, Random Forest) [5]. Others proposed converting the raw signals by wavelet [6] or Fourier transforms [7] for better characterization of the signals. In recent years, with the increased development of deep learning, convolutional neural networks - CNNs [8], recurrent neural networks - RNNs (LTSM and Bi-LSTM and a combination of CNN with LSTM [9] have been widely exploited for motion sensor based HAR problems. Deep learning models have been shown to outperform the traditional methods [10], [11].

Sensor-based HAR often proposes a one-dimensional convolution kernel CNN that can easily handle time series, but it seems not effective to exploit the rich high-dimensional data features. To benefit from the advantages of 2D CNNs, several methods are used to convert the one-dimensional sensor data into two-dimensional data through matrix rearrangement. For instance, one-dimensional time series were converted to two-dimensional time-frequency images by Fourier transform in [12]. The Gramian Angular Difference Field was first introduced by [13] for converting time series into 2D images. In that work, GADF is inputted into a 2D CNN Fusion-Mdk-ResNet for activity classification [12], showing outstanding accuracy.

Motivated by this 2D representation, in this work, we employ the same Gramian Angular Difference Field technique to convert the time series data (accelerometer and gyroscope data) into a 2D image as [13]. However, we capture more of the dependency of features extracted from a 2D CNN by a CoAtNet-based transformer model [14]. CoAtNet was designed to recognize human activities from large datasets

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

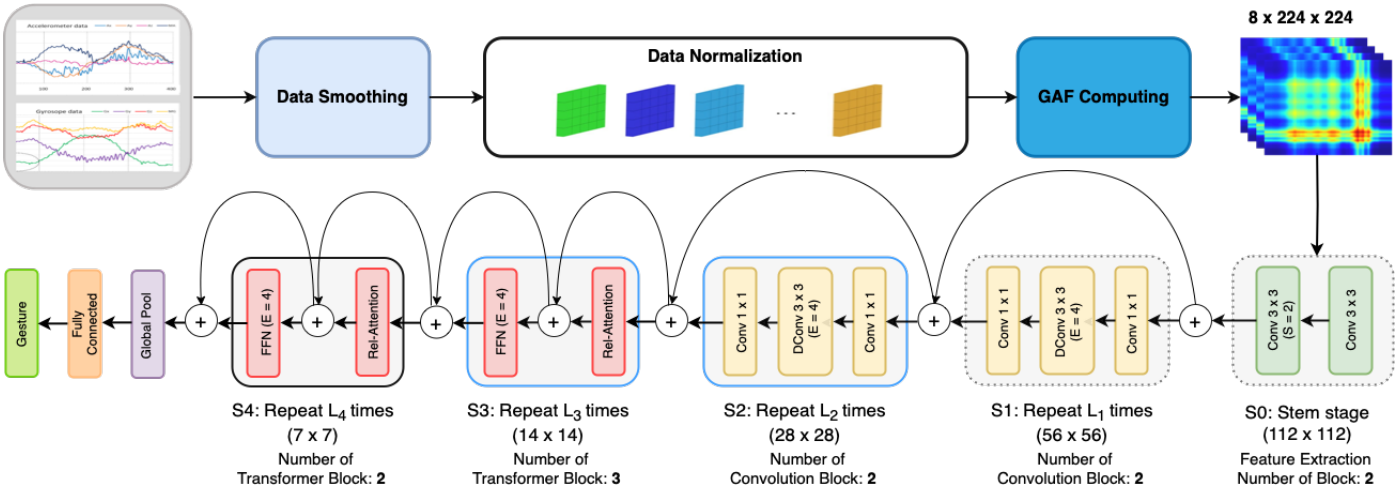


Fig. 1. Framework of GAFormer architecture for human activity recognition.

of RGB images that leverage the advantages of either CNNs or transformer models. In our work, GADF has different characteristics than conventional RGB images. In addition, the number of GADF samples is quite limited which is not suitable for transformer models. We then modify the original CoAtNet to fit better our problem. We name our proposed method GAFormer. The proposed method is validated on two datasets published in the research community C-HMAD [15] and a subset of GesHome [9] for the context of human-machine interaction. The results show that the proposed method achieves better results than the original methods.

In summary, the contributions of the paper are three-fold. Firstly, we propose a new method that combines GADF with a 2D CNN and a Transformer model. Second, we modify the transformer model to avoid overfitting when working with a small dataset in a new perspective (GADF) rather than conventional RGB space. Finally, we evaluate the performance of our proposed method on two public datasets showing promising results for activity recognition in human-machine interaction. The rest of the paper is structured as follows: We present the relevant works in section II. We introduce our framework in section III. Experiments and conclusions are presented in sections VI and V, respectively.

II. Related Work

In recent years, there have been many studies on user activity recognition using wearable devices for data collection. Two main approaches to feature learning for this problem are traditional machine learning methods and deep learning.

Regarding ML methods for extracting hand-crafted features using sensor data, these methods mainly extract features manually in the time or frequency domain to represent movements or activities. One of the recent studies [5] combined local and global features for action recognition from motion sensors. Some features, such as mean, energy, entropy, and frequency domain, are extracted through a sliding window, and then passed through different classifiers, such as SVM and Random Forest for gesture recognition. However, these methods all require complex and time-consuming feature engineering. The method based on deep learning avoids the step of feature engineering, enabling researchers to pay more attention to other aspects of HAR. In the work of E. Ramanujam et al. [16], the focus is on providing a profound and concise description of deep learning techniques used in smartphone and wearable sensor-based recognition systems. The proposed techniques are categorized into conventional and hybrid deep learning models, each with its own uniqueness, merits, and limitations.

Accordingly, several types of deep neural networks are commonly used to classify sensor signals in time series, such as Alemayoh et al. used an improved one-dimensional convolutional neural network (CNN) to classify the collected activity data [8]. Deng et al. [17] proposed two methods for activity image construction, transforming the sensor data into activity images by matrix rearrangement. Then, the CNN was applied to extract features and identify the types of activities. Uddin et al. proposed a multi-sensors data fusion network based on recurrent neural network (RNN) [18].

Nguyen et al. propose a 1D-CNN biLSTM architecture for 18 gestures using motion sensors [9] applied to home appliance control applications. BiLSTM allows time series to move in two directions: from past to future and from future to past. It works on the principle of two LSTM layers in one layer when the other layer is presented with a reverse copy of the original time series. Although the deep learning method has many advantages, it also has some unavoidable problems and disadvantages. The recognition methods based on RNN can only be carried out in sequence because the calculation of the next step in the training process depends on the results of the previous step, and the training process consumes a long time. Recognition methods based on CNN mostly use a one-dimensional convolution kernel which makes it difficult to fully exploit the rich high-dimensional data features.

Some signal-to-image conversion methods have also been

studied with positive results. Jiang et al. [19] use HAR inertia data to convert it into 2D active images, which they then feed into classifiers to improve performance. The active images were converted to the frequency domain and the spectral time domain using a discrete Fourier transform (DFT) and a 2D Wavelet transform, respectively. However, these images in different domains are used independently, and their union cannot be exploited. Xu et al. [12] used the method Gramian angular field (GAF) and deep CNN because of the weight sharing mechanism of CNN, the training speed of CNN is much faster than the other networks. At the same time, the GAF algorithm increases the interpretability of transforming from one-dimensional time series to two-dimensional images and also lays a foundation for the effectiveness of feature extraction. In 2021, Zeeshan Ahmad [20] also demonstrates the use of inertial sensor data using four types of spatial domain methods to convert inertial sensor data into active images. ResNet-18, a CNN model, is used to learn deep features from multimodality.

To combine the advantages of CNNs, RNNs, and self-attention networks, Vaswani et al.[21] designed a new architecture using centralization. This architecture, called Transformer, parallelizes by learning the response sequence with a centralized mechanism while also encoding the position of each element in the sequence. Dirgova et al. [22] also use the Transformer model for the direct application of deep neural networks to normalized time signal sequences from sensors using a different approach to time series processing. entirely based on the attention mechanism. The transformer model focuses directly on using the attention mechanism to find time series correlations between features and allows large parallelism of time series computations. However, this method is limited by the long computation time. To solve this problem, Zihang Dai [14] proposes a simple but effective network architecture called CoAtNet, which leverages strengths from both ConvNets and Transformers. In this study, it is shown that commonly used depth-first convolution can be efficiently merged into attention layers with relatively simple attention; stacking the convolution and attention layers in the appropriate manner can result in better efficiency, generalization, and capacity. CoAtNet not only enjoys the superior scalability of Transformer models but also achieves faster convergence and thus improved efficiency.

Our work in this paper will also convert the time series into an image using the Gramian Angle Field transform. Then, the generated image will be fed into the classification system, which is a combination of CNN and Transformer in image classification. Our methods and results will be presented in the following sections.

III. Method for action recognition from motion sensors

A. Proposed framework

Our proposed framework is illustrated in Fig.1. It is composed of four main steps:

- Pre-processing: This step denoises the raw accelerometer, gyroscope data, and their magnitude by using a smoothing filter. The smoothed signals are normalized to the range [0, 1].
- GADF computing: This step takes the normalized signals and converts them into a 2D image using Gramian Angular Different Field transform.
- Extracting 2D embedded features: A 2D CNN is deployed to extract discriminative features from GADF images.
- Self-attention computing: The embedded features are passed through a stack of self-attention blocks to capture better their correlation.

In the following, we will describe each step in detail.

B. Pre-processing

1) Data smoothing: The data needs to be smoothed before being processed in our proposed GAFormer model to maximize the effectiveness of the method. In this paper, we simply use the following smoothing function:

$$x_i = \frac{(x_{i-2} + x_{i-1} + x_i + x_{i+1} + x_{i+2})}{5} \tag{1}$$

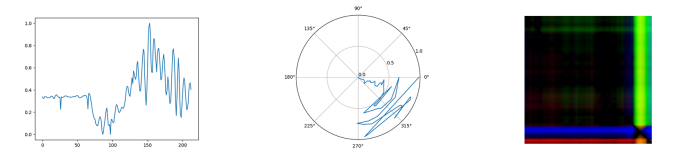
where x_i is the data signal at the time i. This formula is applied for all channels of accelerometer and gyroscope data and also the magnitudes of these two signals.

2) Data normalization: Let X be a real-valued time series of n samples $X = \{x_1, x_2, x_3, \ldots, x_l, \ldots, x_n\}$. We rescaled X into $\tilde{X}$ so that the value of each sample in $\tilde{X}$ falls between 0 and 1, which can be expressed as follows:

$$\tilde{x}^i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{2}$$

C. Gramian Angular Field

Wang et al. proposed Gramian Angular Field that encodes time series data into various types of images namely, Gramian Angular Summation/Difference Fields (GASF/GADF) and Markov Transition Fields (MTF) [13]. In this paper, we convert our time series from motion sensors by Gramian Angular Difference Field (GADF) to capture underlying dynamics while being easily processed by computer vision techniques.



(a) A rescaled one-dimentional time series in GesHome datasets (b) Transform time series into a polar coordinate system in GesHome datasets (c) GADF image in GesHome datasets

Fig. 2. Illustration of one-dimensional time series (left) converted to a 2D image (right) in the polar coordinate system (middle).

To this end, first, the normalized time series obtained from the pre-processing step is represented in the polar coordinates system. The value of the angular sine presents the angle while the time stamp presents the radius. The angular field is a new perspective of time series as time increases, the values

of angles will run along the circles of different radii. This encoding is formulated by the following equation.

$$\begin{cases} \phi_i = \arcsin(\tilde{x}_i), & 0 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ r_i = \frac{i}{N}, & i \in N \end{cases} \tag{3}$$

where $\tilde{x}_i$ is the rescaled sample of the time series, i is the time stamp and N is a constant factor to regularize the span of the polar coordinate system. Then the Gramian Angular Difference Field is utilized to create the corresponding 2D image:

$$GADF(i,j) = \sin(\phi_i - \phi_j) \tag{4}$$

where $i,\ j \in [1, N]$, N is the length of internal time series data and varies to the activity sample. In our work, we consider three acceleromerter channels (a_x, a_y, a_z), three gyroscope channels (g_x, g_y, g_z) and their corresponding magnitude m_a, m_g. As a result, a tensor $N \times N \times 8$ is generated. We then resize it to $224 \times 224 \times 8$ for further processing steps. Fig. 2 illustrates an example of one-dimensional time series (left) converted to a 2D image (right) in the polar coordinate system (middle).

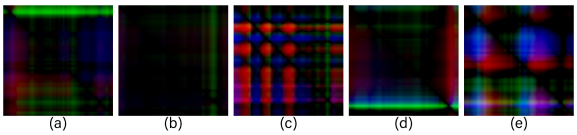


Fig. 3. GADF images generated from samples of five activities in C-MHAD dataset: (a) Swipe to left. (b) Swipe to the right. (c) Hand wave. (d) Draw a clockwise circle. (e) Draw a counter-clockwise circle.

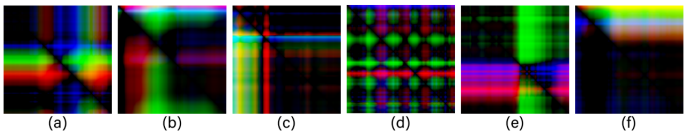


Fig. 4. GADF images generated from samples of five activities in GesHome dataset: (a) CCWCircle. (b) CWCircle. (c) Move down. (d) Move left. (e) Move right. (f) Move up.

Figure 3 and Figure 4 illustrate GADF images generated from time series of several activities in two experimented datasets. It is noticed that each activity generates a GADF which is different from others. That is a good characteristic for distinguishing the activities by the next steps.

D. Features extraction and classification

The GADF image generated from the previous step is inputted into a hybrid model that combines 2D CNN and Transformer. The idea to combine 2D CNN and Transformer is to boost the strengths of both models (generalization for CNN and capacity for transformer). In [14], Dai et al. have proposed a method, namely called CoAtNet, to efficiently hybrid the CNN and the Transformer.

CoAtNet combines the power of Convolutional Neural Networks (CNN) and Transformer networks. CNN aggregates local features from GAF images, while Transformer helps the model learn global relationships and extract important information from the data. This combination is demonstrated in the following form:

$$y_i^{\text{pre}} = \sum_{j \in \mathcal{G}} \left(\frac{\exp\left(x_i^T x_j + w_{i-j}\right)}{\sum_{j \in \mathcal{G}} \exp\left(x_i^T x_k + w_{i-k}\right)} \right) x_j \tag{5}$$

where $x_i, x_j, x_k \in R^D$ are the input at position i, j, k respectively, y_i^{pre} is the output at the position i before the Softmax normalization, w_{i-j} is the depthwise convolution kernel in CNN model and $\mathcal{G}$ indicates the global spatial space in transformer model. This formula combines convolution with self-attention to gather the information that shares the same receptive field $\mathcal{G}$. To enable the global convolution kernel without increasing the number of parameters, we consider w_i as a scalar as described in the original paper [14].

The architecture of CoAtNet is illustrated as a component in our GAFormer in Fig.1. It is composed of convolutions and transformers stacked together. As indicated in [14], the stack type C-C-T-T (C stands for Convolution and T stands for Transformer) has better generalization, model capacity, transferability, and efficiency then it is chosen to design CoAtNet in this work. Fig. 1 shows five stages S_0, S_1, S_2, S_3, S_4, the L_1 to L_4 may repeat several times. For convolution block, CoAtNet utilizes MBConv [23] because it was built upon depthwise convolutions so it can achieve both high accuracy and better efficiency. In the first block of each stage from S1 to S4, the down-sampling operator is performed independently for the residual branch and the identity branch. The global average pooling is applied to the last-stage output to get the representation.

CoAtNet has various versions (from CoAtNet-0 to CoAtNet-7) depending on the number of times the blocks in each stage are repeated. The higher version has more parameters but gives higher accuracy. However, some existing works have shown that Transformer scales better but its accuracy could be reduced compared to the CNN models. As a result, Transformer achieved only comparable performance when trained on a very large dataset. In this work, as we work on quite a limited dataset in a new perspective (GADF), we create a smaller CoAtNet, which we call CoAtNet-light by reducing the number of Attention Blocks. Table. I shows the configuration of our proposed CoAtNet-light compared with the smallest model CoAtNet-0. It is noticed that we keep the same number of channels but we change the number of repeat times at stages S_2 and S_3. Particularly, we reduce the number of Convolution Blocks (layer S2) from three to two, and the number of Transformer Blocks, in layer S3 from five to three. By doing so, we can avoid overfitting problems why lightening the size of models. As a result, the inference time is reduced (measurement by NVIDIA GeForce RTX 2080 Ti GPU) (Tab. II).

To sum up, GAFormer (GADF + CoAtNet-light) is a framework that combines the representation of motion data by GADF image and a CoAtNet-light network which is customized to fit better with the small-size datasets.

TABLE I
COMPARISON OF OUR COATNET-LIGHT WITH THE ORIGINAL COATNET-0

Stages	Our CoAtNet-light	CoAtNet-0
S0-Conv	L=2 D=64	L=2 D=64
S1-MBConv	L=2 D=96	L=2 D=96
S2-MBConv	**L=2** D=192	L=3 D=192
S3-TFM	**L=3** D=384	L=5 D=384
S4-TFM	L=2 D=768	L=2 D=768

TABLE II
COMPARISON NUMBER OF PARAMETER AND INFERENCE TIME BETWEEN GADF+COATNET-0 AND OUR PROPOSED GAFORMER

Method	Number of parameters	Inference time
GADF + CoAtNet-0	$19.03M$	$19.75ms$
GAFormer	$13.49M$	$15.42ms$

IV. EXPERIMENT

A. Datasets

We use two available datasets to evaluate the studied method, including C-MHAD [15] and GesHome [9]. The data of each gesture/activity used will include six signals: three signals in x, y, z axes of the accelerometer sensor and three signals in x, y, z axes of the gyroscope sensor. The details of each data set are described as follows:

- **C-MHAD[15]**: The continuous dataset contains five hand gestures for controlling smart TVs performed by 12 subjects (ten males and two females). The three-axis acceleration and three-axis angular velocity signals of the Shimmer3 wearable inertial sensor were fed in real time to a laptop via a Bluetooth link at a frequency of 50 Hz. It includes 41,410 samples split into a training set (41,310) and a test set (100). Each sample is then normalized and converted to GADF images and illustrated in Fig. 3.
- **GesHome*[9]**: The GesHome dataset includes 18 hand gestures from 20 non-professional subjects of various ages and occupations. Participants performed 50 reps of each movement for five days. Consequently, the number of collected data is 18,000 gesture samples, including three-axis linear acceleration and three-axis angular velocity with a frequency of 25 Hz. In our experiment, to demonstrate the effectiveness of the method and its application to the problem of controlling smart home devices, we choose six main gestures in the dataset focusing on using hand gestures to simulate device control for experimentation leading to only 9,689 samples available. We name this dataset GesHome*. Each sample in the dataset is normalized and then converted to the corresponding Gramian Angle Difference Field (GADF) image as shown in Fig. 4. We divided this dataset into two subsets, the training set with 7,843 samples and 1,846 samples of the test set.

B. Experimental results

The experiments are conducted on these two datasets. There are four combinations of GAF and CNN, namely GASF+CoAtNet-0, GADF+CoAtNet-0, GADF+ResNet50, and GAFormer. In addition to accuracy, we also provide some metrics, such as Precision, Recall, and F1-Score, to evaluate models. The results are illustrated on Tab. III, Tab. IV, respectively.

TABLE III
PERFORMANCE COMPARISON ON C-MHAD DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Baseline [15]	96.48	-	-	-
GASF + CoAtNet-0	95.00	94.20	94.80	94.40
GADF + ResNet50	94.00	93.96	94.00	93.80
GADF + CoAtNet-0	98.00	97.60	97.80	97.80
GAFormer	**98.00**	**97.80**	**98.20**	**98.00**

TABLE IV
PERFORMANCE COMPARISON ON GESHOME* DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1D-CNN-biLSTM [9]	91.77	92.33	91.66	91.66
GASF + CoAtNet-0	72.91	73.16	72.67	72.16
GADF + ResNet50	90.00	90.16	89.83	90.33
GADF + CoAtNet-0	93.31	93.42	93.21	93.23
GAFormer	**95.50**	**95.50**	**95.33**	**95.40**

- **C-MHAD dataset:**

In this dataset, the results of all four combinations are compared with the baseline method introduced in [15] as in Tab. III. The baseline method's accuracy is 96.48%, an average of five reported actions' accuracy. Firstly, the model with the lowest accuracy combines GADF and ResNet50, with only 94%. Secondly, the CoAtNet-0 model, combined with GASF and GADF transformations, could achieve 95% and 98% accuracy, respectively. These results indicate that GADF seems better than GASF in these tasks. Regarding our proposed GAFormer method, with an accuracy of 98.00 % compared to 96.48 %, our method could overcome the previous study. Although these results, including precision, recall, and f1-score, are slightly better than those of GADF+CoAtNet-0, GAFormer is still a preferred choice thanks to its number of parameters and inference time, as Tab. II indicates.

- **GesHome* dataset:**

Similar to the C-MHAD dataset, we also compare our proposed method with the original work, combining 1D-CNN and biLSTM. In this dataset, we re-implemented it with less action, leading to an accuracy of 91.77%. The first thing to notice is that the GASF + CoAtNet-0 method has the lowest performance compared to the other methods, with the accuracy of this method at only 72.91 %. Next is the

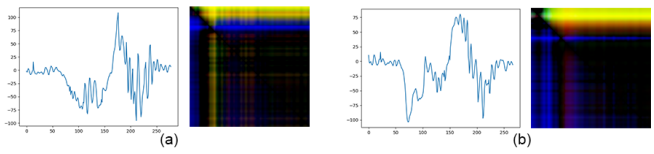


Fig. 5. Confusion image in GesHome datasets: (a) Image in CWCircle gesture. (b) Image in Move-left gesture.

GADF + ResNet50 method, with an accuracy of 90.00 %. The results of this method are relatively high but still lower than the baseline. Another combination, GADF + CoAtNet-0, has excellent performance with an accuracy of 93.31 %. The proposed GAFormer reaches 95.50 % yielding superior performance in terms of accuracy, precision, recall, f1-score, and the number of parameters. Tabs. IV shows the best performance of GAFormer compared to other methods and is highlighted with bold text.

V. Conclusion

In this work, we introduced a new method GAFormer for human activity recognition from wearable sensors. The method first transforms the motion data into Gramian Angle Difference Field (GADF) to create a 2D image. To extract the features from the image, we proposed a light version of CoAtNet that achieved better accuracy while reducing the inference time and the model size. Our method has been validated on two public datasets showing better accuracy than existing methods. In the future, we will deploy the method for continuous recognition from motion sensors in a practical application of human-machine interaction.

Acknowledgement

This research is funded by Hanoi University of Science and Technology (HUST) under project number T2021-SAHEP-003.

References

[1] S. Paraschiakos, R. Cachucho, M. Moed, *et al.*, "Activity recognition using wearable sensors for tracking the elderly," *User Modeling and User-Adapted Interaction*, vol. 30, no. 3, pp. 567–605, 2020.

[2] A. M. Buttar, M. Bano, M. A. Akbar, A. Alabrah, and A. H. Gumaei, "Toward trustworthy human suspicious activity detection from surveillance videos using deep learning," *Soft Computing*, pp. 1–13, 2023.

[3] H.-Q. Nguyen, T.-H. Le, T.-K. Tran, *et al.*, "Hand gesture recognition from wrist-worn camera for human-machine interaction," *IEEE Access*, pp. 1–1, 2023. DOI: 10.1109/ACCESS.2023.3279845.

[4] S. Zhang, Y. Li, S. Zhang, *et al.*, "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, no. 4, p. 1476, 2022.

[5] T. H. Le, Q. T. Nguyen, T.-H. Tran, and C. Pham, "Combined local and global features for action recognition from motion sensors," in *Intelligence of Things: Technologies and Applications: The First International Conference on Intelligence of Things (ICIot 2022), Hanoi, Vietnam, August 17–19, 2022, Proceedings*, Springer, 2022, pp. 146–155.

[6] Z. He, "Activity recognition from accelerometer signals based on wavelet-ar model," in *2010 IEEE International Conference on Progress in Informatics and Computing*, IEEE, vol. 1, 2010, pp. 499–502.

[7] W.-Y. Chung, A. Purwar, and A. Sharma, "Frequency domain approach for activity classification using accelerometer," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2008, pp. 1120–1123.

[8] T. T. Alemayoh, J. H. Lee, and S. Okamoto, "New sensor data structuring for deeper feature extraction in human activity recognition," *Sensors*, vol. 21, no. 8, p. 2814, 2021.

[9] K. Nguyen-Trong, H. N. Vu, N. N. Trung, and C. Pham, "Gesture recognition using wearable sensors with bi-long short-term memory convolutional neural networks," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 15 065–15 079, 2021.

[10] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Hand-crafted features vs residual networks for human activities recognition using accelerometer," in *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*, IEEE, 2019, pp. 153–156.

[11] C. Pham, S. Nguyen-Thai, H. Tran-Quang, *et al.*, "Scapsnet: Deep neural network for non-obtrusive sensing based human activity recognition," *IEEE Access*, vol. 8, pp. 86 934–86 946, 2020.

[12] H. Xu, J. Li, H. Yuan, *et al.*, "Human activity recognition based on gramian angular field and deep convolutional neural network," *IEEE Access*, vol. 8, pp. 199 393–199 405, 2020.

[13] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," *arXiv preprint arXiv:1506.00327*, 2015.

[14] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[15] P. C. Haoran Wei and N. Kehtarnavaz, "C-mhad: Continuous multimodal human action dataset of simultaneous video and inertial sensing," *Sensors Journal 2020*, vol. 20, 10 2020.

[16] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 029–13 040, 2021.

[17] Deng, "Convolutional neural networks for human activity recognition using multi-location wearable sensors," *Journal of Software*, vol. 30, no. 3, pp. 718–737, 2019.

[18] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, "A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare," *Information Fusion*, vol. 55, pp. 105–115, 2020.

[19] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1307–1310.

[20] Z. Ahmad and N. Khan, "Inertial sensor data to image encoding for human action recognition," *IEEE Sensors Journal*, vol. 21, no. 9, pp. 10 978–10 988, 2021.

[21] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] I. Dirgová Luptáková, M. Kubovčık, and J. Pospıchal, "Wearable sensor-based human activity recognition with transformer model," *Sensors*, vol. 22, no. 5, p. 1911, 2022.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Fatigue Classification and Onset estimation using Surface EMG Signals during Strength Training

Eswar Adapa, Anish C Turlapaty and Surya Naidu
Indian Institute of Information Technology Sri City, Sri City
E-mail:{geetasaieswar.a20, anish.turlapaty, sreesuryakiran.n20}@iiits.in Tel/Fax: +91-8179294256

***Abstract*—Muscle fatigue onset detection and estimation have applications in supporting athletes during strength training. In this paper, the problem of surface electromyography (sEMG) based binary classification of fatigue and non-fatigue of human muscle activity during strength training is addressed. Further, the learned model is used to determine the fatigue onset using sEMG signals corresponding to a training activity. A novelty of this work is the analysis of the proposed machine learning method under different measurement conditions, such as varying postures, experience levels and lifting loads. A novel sEMG dataset titled ElectroMyography Analysis of Human Activities - DataBase 6 (EMAHA-DB6) is developed for strength training activities. EMAHA-DB6 includes sEMG signals collected from the seven prominent muscle sites of the right arm of 11 subjects engaged in strength training. The participants performed five distinct training exercises using three load weights in two body postures. A set of wavelet and time domain features are extracted and the support vector machines are used to classify them. Based on the binary testing results from a given time series, the fatigue onset is estimated using the majority voting. The SVM model is compared with other classifiers in terms of testing accuracy in the mentioned measurement scenarios. The SVM has outperformed the other models with a test accuracy of 86.5% under all conditions scenario. In terms of the impact of measurement conditions, the SVM has 94% in the low load condition, 88.5% in the stand posture and nearly 88% in the case of subjects with intermediate training status. Finally, in the proposed approach, the estimation of the fatigue onset has an average relative error of 12%.**

I. Introduction

Background: Muscle fatigue is a common physiological phenomenon that occurs due to prolonged or intense muscle activity, including strength training [1]. Fatigue due to strength training primarily results in reduced nerve signal strength that regulates muscle movement, leading to a reduction in muscle output force and has the potential to cause muscle damage and increase the risk of joint-related issues [2]. So monitoring changes in muscle strength and endurance over time during strength training can provide important insights into fatigue onset and help identify strategies for preventing injury and promoting recovery. In applications such as usage of prosthetics and rehabilitation, an accurate prediction or atleast estimation of fatigue is essential for optimal outcomes. In the context of prosthetics, the estimation of fatigue is crucial because it helps prevent potential damage to the muscles and other tissues surrounding the prosthetic, as well as injuries and accidents that may occur due to decreased ability to control the prosthetic limb. By identifying and addressing muscle fatigue early on, such incidents can be minimized. In the context of rehabilitation, fatigue onset estimation is crucial for improving efficiency to minimize its adverse affects on the recovery process. Notably, predicting fatigue in the lower extremity has proven to significantly enhance the recovery of stroke patients by facilitating longer exercise duration and promoting active participation in the rehabilitation process [3].

There are various approaches available for direct or indirect onset detection and estimation of muscle fatigue, such as measuring lactate concentration [4], analyzing surface electromyography (sEMG) signals, and studying gait parameters [5]. Among these, sEMG is commonly used as a non-invasive estimation method because sEMG represent muscle contractions and can be utilized to characterize muscle fatigue [6][7][8]. In recent years, there were several studies with focus on differentiating between sEMG signals in non-fatigue and fatigue conditions during dynamic contraction. One such study utilized multivariate multifractal detrending moving average (MMFDMA) algorithm to extract features from the singular spectrum and differentiated between non-fatigue and fatigue conditions [9]. Another study investigated the detection of muscle state based on the working principle of synergistic muscles and combination of signals from different muscle groups [10]. Additionally, in another study different fatigue estimation algorithms and assessment methods are compared to understand intra and inter-subject separability of fatigue levels. Further comparative assessment was carried out at three different load levels, demonstrating the importance of considering various measurement conditions [11]. Most recently, a classification analysis of non-fatigue and fatigue data was carried out to determine the onset of fatigue using multi-modal data. These sensor modalities included EMG sensors, computer game data, and perceived data [3]. These studies demonstrate the importance of using ML techniques to detect and predict fatigue in various conditions, including rehabilitation exercises. The existing research does not address the problem of classification based fatigue onset estimation and there is no clear understanding of role of different measurement conditions during estimation.

Contributions: The key contributions of this study include 1) development of a dataset of multi-channel sEMG signals collected during weight training under various measurement conditions, 2) development of a machine learning framework for classification of segments of the sEMG signal into fatigue and non-fatigue categories, 3) a fatigue onset estimation framework based on the results from the classifier and 4) analysis

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of the classification scheme with respect to the measurement conditions.

II. Data Description

A. Data Collection

In this paper, a new sEMG signal dataset named: ElectroMyography Analysis of Human Activities - DataBase 6 (EMAHA-DB6) is developed. The EMAHA-DB6 consists of multi-channel sEMG signals acquired during various strength training activities under different conditions described as follows. The study involved 11 healthy strength training enthusiasts (subjects) comprising 10 males and 1 female with an average age of 20 years and an average body weight of 68kg. The selected subjects provided their voluntary written consent. The measurements and the research were approved by the Institutional Ethics Committee of IIIT Sri City in accordance with the "National Ethical Guidelines for Biomedical and Health Research involving human-participants" of India and in general accordance with the Declaration of Helsinki. As shown in Table I, based on the strength training status [12], the subjects were categorized into three levels and the number of subjects at each level are: 3 at advanced, 4 at intermediate, and 4 at beginner level. Again as shown in Table I, to investigate the effects of fatigue resulting from prolonged and intense physical activity, each subject was assigned three different loads based on their experience level and body weight. In this study, as shown in Table II, the subjects are asked to perform five different exercises with three different loads and two different postures. Note that in this work, the fatigue onset is defined as the moment the subject experienced a reduction in strength and they were instructed to halt the activity exactly at the end of second curl from the onset.

TABLE I
RECOMMENDED WEIGHTS FOR DIFFERENT LEVELS OF EXPERIENCE

Experience Level	Load Weights	Body Weight	Experience
Beginner	2.5 kg 5 kg 7.5 kg	50 - 60 kg	0 - 6 months
Intermediate	5 kg 7.5 kg 10 kg	60 - 70 kg	6 - 12 months
Advanced	7.5 kg 10 kg 12.5 kg	70 - 80 kg	12 - 18 months

TABLE II
EMAHA-DB6: STRENGTH TRAINING ACTIVITIES AT DIFFERENT LOADS

Activities	Load Weights	Postures
One Arm Overhead Tricep Extension One Arm Tricep Kickback Reverse Dumbbell Zottman Curl Dumbbell Hammer Curls Rear Front Rotations	Min Load Mid Load Max Load	Sit Stand

Each hand muscle activity is recorded with a 7-channel Noraxon Ultium wireless sEMG sensor setup at 4000 samples per second. Self-adhesive Ag/AgCl dual electrodes were placed at the centers of the seven representative muscle sites of the right arm. Prior to a measurement session, each subject is informed about the experimental procedures and instructed to avoid food consumption within two hours before the session and refrain from alcohol consumption for at least 24 hours leading to the experiments. During the experiment, the subjects are encouraged to consume water and soft drinks but are advised to avoid dairy products. For a given subject, the strength training experiment consisted of two sessions lasting up to one hour each. During the first session, in the *stand* posture, a subject performed the above mentioned training activities with three loads according to the training status and adhering to the fatigue based stopping protocol described above. In the second session the same set of activities and the fatigue protocol are repeated in the *sit* posture. Each activity involved two phases: a rest phase and an action phase. During the rest phase, subjects are allowed to relax. Once they felt ready, they transitioned to the action phase and actively engaged in the specified activity. The subjects discontinued the activity exactly after the second curl when they were in need of rest. Specifically, in the action phase, the maximum time recorded was 240s for the rear-front rotations with the min load. On the other hand, the minimum time observed was 15.25s for the one-arm tricep exercise with the max load. The average duration of all activities during the action phase is approximately 60.5 seconds.

III. Methodology

As shown in Fig. 1, the proposed methodology for fatigue classification and onset estimation consists of five stages: 1) data preparation consisting of preprocessing and signal labelling, where the data converted into patterns suitable for supervised learning, 2) segmentation and feature extraction, where the patterns are converted into feature vectors, 3) training: where a classification model is chosen and trained with the labelled data, 4) testing and analysis, where the trained model is used to test on unseen trials and optimized; and finally, 5) onset estimation, where the optimized model is used to determine the onset of fatigue based on majority voting on binary output from classification.

A. Data Preparation

The data preparation stage of this study involved two steps:Preprocessing, and signal labelling.

1) Preprocessing: The multi-channel sEMG signals are processed by a high pass filter with a cutoff frequency at 20Hz to remove low-frequency noise for example due to mechanical artefacts. To eliminate the power line noise, a band-stop filter at 50Hz is also applied.

2) Signal Labelling: In this study, a self-report based approach is adopted for signal partitioning into fatigue and non-fatigue. Specifically, the delineation is based on a subject's own report of inability to continue and need of rest leading to discontinuing the activity. Furthermore, sections of the sEMG time-series corresponding to the two closing curls of the activity are assigned the label "*fatigue*" and represent periods of fatigue as experienced by the subjects. On the other hand,

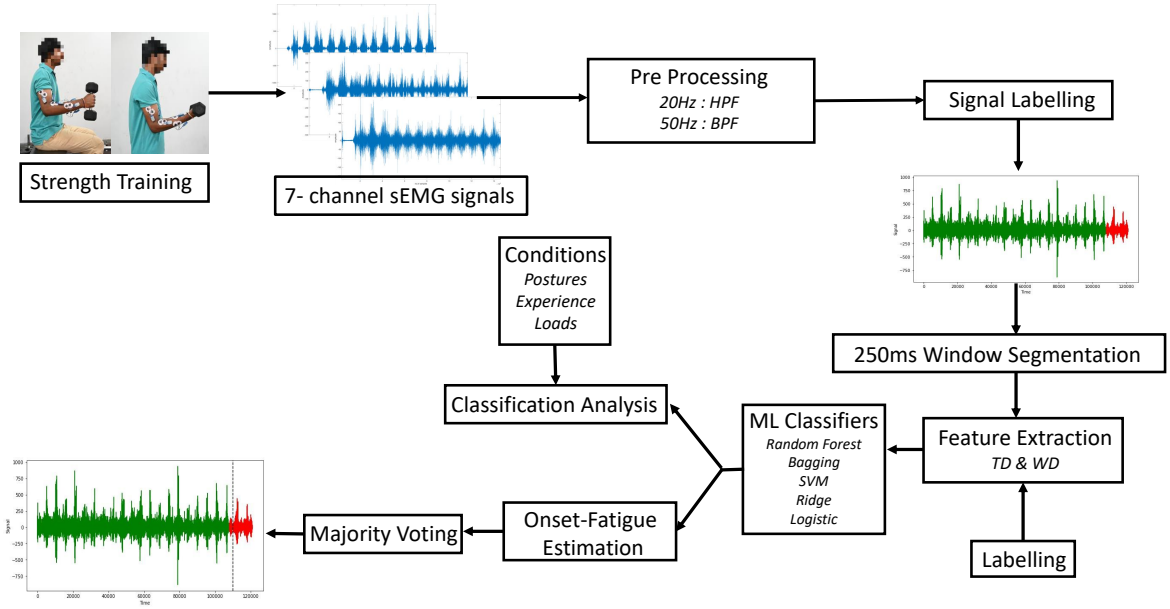


Fig. 1. Block Diagram for the Fatigue Classification and Onset estimation

the rest of the signal excluding the these two curls is labelled as "*Non Fatigue*".

B. Feature Generation

To facilitate classification, filtered signals are segmented into fixed length windows of 250ms duration. This window size is chosen to capture the short-term temporal dynamics of the muscle activity and provide sufficient resolution for further analysis. In the study, features are extracted for these 250ms segments. Feature extraction is performed in various domains, including time domain [13][14], frequency domain [15], wavelet domain [16], and eigenvalues, resulting in a total of 30 features. Class labels are then assigned accordingly.

C. Model Training

To balance the features from different modalities, the feature vectors are min-max scaled. This process standardized the data to enhance the performance of the classification models. During the training phase, to mitigate the significant class imbalance in the dataset, only 25% of the *non fatigue* segments in an activity are retained for further analyses. The scaled features are fed as input to the classification model for training. The classification algorithms analyzed in this study include the random forest (RF) classifier, bagging classifier, support vector machines (SVM), logistic regression (LR), and ridge classifier. These models were specifically selected due to their established effectiveness in binary classification tasks.The objective is to assess the classification performance from multiple perspectives and identify the most effective model for this task.

D. Testing and Analysis

For evaluating the performance of the trained models, in the testing phase, leave one activity out strategy is employed. Specifically, the feature vectors of segments of the left out activity are classified by the trained model and the classification performance is analysed as an average across the tested feature vectors. The performance metrics, include the confusion matrix and the classification accuracy. These metrics provided valuable insights into the effectiveness of the models in classifying the data into *Non Fatigue* and *Fatigue* segments.

E. Onset Estimation Framework

The onset estimation framework is described as follows. From Fig. 1, the classification model that provided the best performance during the testing phase is chosen as the onset estimator. This estimator is used to segregate the sEMG time series into *non fatigue* and *fatigue* partitions. To improve the quality of segregation a window based majority voting is employed. This window is equivalent to 10% of the signal length. In the majority voting approach, the predicted labels from multiple component segments of the time-series are aggregated. The label that occurs most frequently is considered the winner and an estimate of fatigue onset is obtained.

The performance of the onset estimation framework is analyzed as follows. For each activity set for estimation, the absolute error between the actual onset and the estimated onset through the previously described segregation process is determined. These errors are scaled by the length of the time series and statistics of the relative errors are reported. Furthermore, for visual inspection, the time series of a few

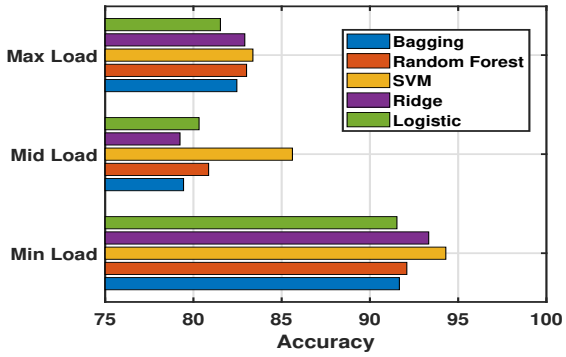


Fig. 2. Influence of Loads on Classification Performance.

sample signals with the segregated *non fatigue* and *fatigue* sections of the time series are plotted. Thus a combination of quantitative analysis and visual interpretation are used for performance assessment of the proposed onset estimation framework.

TABLE III
MODEL PERFORMANCE IN AGGREGATE DATA SCENARIO (EXPERIMENT 1)

Model	Accuracy (%)
SVM	86.51
Random Forest	85.08
Bagging	84.63
Ridge	84.4
Logistic	83.92

IV. IMPLEMENTATION AND RESULTS

A. Classification Experiments

In this study, various classification experiments are conducted to analyze the performance of the model under different measurement conditions. Note that in each of these experiments, the leave one activity out (LOAO) strategy is used for testing performance analysis.

1) Experiment 1: Aggregate Analysis: In this experiment, the activities from different body postures, lifting loads and training status levels are aggregated for classification performance analysis. The subsets of features mentioned are analyzed for optimal classification performance.

2) Experiment 2: Influence of Weights: In this experiment, the feature data is divided into three sets based on the training loads used by the subjects, namely, min, mid and max loads. The classification model is trained and tested separately on data from each group using the LOAO mechanism. The model performance is evaluated for each group and the average performance is reported.

3) Experiment 3: Influence of Postures: The experiment for understanding the impact of body postures has three different categories.

(i) Intra posture analysis: The classification model is trained and LOAO tested on the feature vectors from the same posture eg. *sit* posture. The experiment is repeated for the data from the *stand* posture.

(ii) Inter posture analysis: In the first phase of this analysis, the ML model is trained on feature vectors from the *sit* posture and tested on the data from the *stand* posture. Note that in this stage, the entire data from the other posture i.e., *stand* is tested. In the second phase, the order of postures is reversed.

(iii) Mixed posture analysis: In this category, in the first phase, the entire data from the *sit* posture and 30% of data from *stand* posture is used for training. For testing, the remaining 70% of the *stand* dataset is used. Note that the $30 : 70\%$ split is based on activities within the posture category. Again in the second phase, the process is vice-versa.

4) Experiment 4: Influence of Training Status: An experiment similar to the experiment 2 is conducted by dividing the data into sets based on subjects' training status advanced, intermediate, and the beginner. Again the training level performances are analyzed and the average is reported.

B. Classification Results and Analysis

1) Aggregate analysis: In this experiment, among the feature subsets tested, the combination of wavelet and time domain features demonstrated the best performance. Hence, these specific feature subsets are selected for further analysis. Table III shows the results of the testing in the aggregate data scenario from experiment 1. The SVM classifier outperforms the other classifiers with an average LOAO accuracy of 86.51%. The RF and the Bagging methods follow with 85% and 84.6% respectively.

2) Influence of Loads: Fig. 2 shows the model's test accuracy from the second experiment. The results show that there is clear dependence of performance on the activity load. For example, when training and testing on activities with the minimum load, highest accuracy of 94% for the SVM is observed. It indicates the model is effective in detecting gradual onset of fatigue, which is often associated with lower loads. On the other hand, when training and testing on the medium load, the model demonstrates 8% decline, suggesting its reduced ability to classify slightly intense fatigue situations. Finally, in the case of maximum load, the performance reduces to 83% a reduction of 11% in accuracy. This clearly indicates reduced classification capability in the case of fatigue related to maximum load.

3) Influence of Postures: Fig. 3 shows the results of the third experiment. It is observed that the SVM outperforms the other models by a decent margin in each of the categories of this experiment. For the SVM, the lowest performance is 83% and the highest is close to 89%. The performance of the models is clearly sensitive to different schemes in the third experiment. In the intra posture analysis, the performance in the case of *stand* posture is generally higher by 4% compared to the case of *sit* posture. In the inter posture analysis, when the model is trained on data from *sit* posture and tested on data from *stand*, a posture never trained, there is a slight decline in performance. In the mixed posture analysis, when only 30% of data from *stand* posture is introduced in training, the

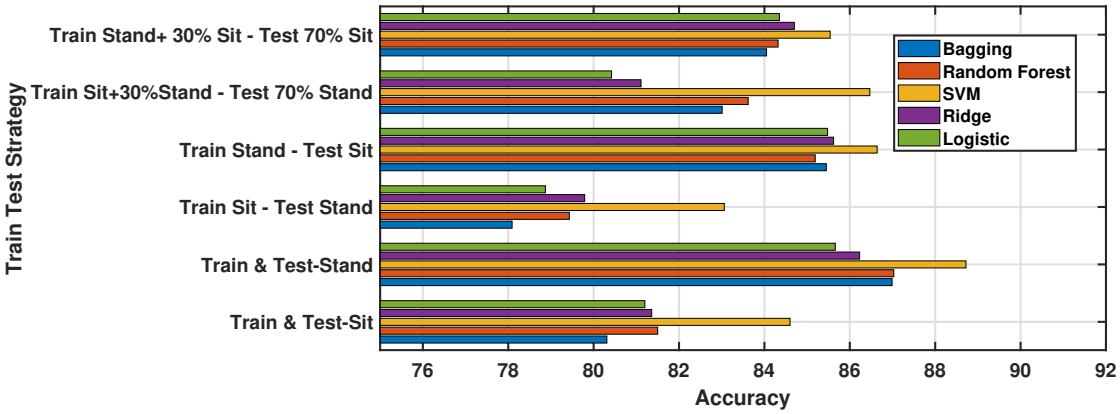


Fig. 3. Influence of Postures on Classification Performance.

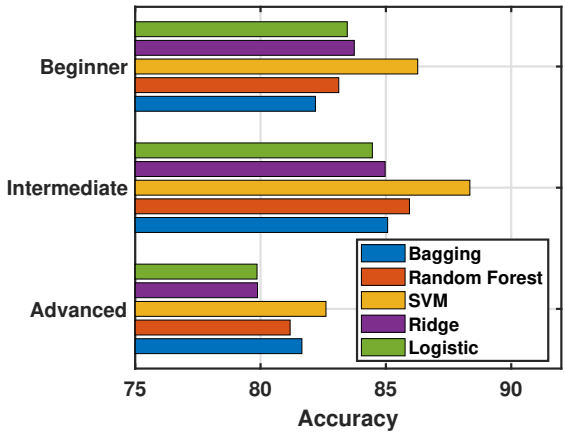


Fig. 4. Influence of Experience on Classification Performance.

model's performance improves by at-least 3%. This suggests that including a data from different postures in the training phase has a positive impact on testing accuracy. It should be noted that the performance of the model on the data from *stand* posture remains consistent across the different scenarios. In other words, the model seems more sensitive to data from *sit* posture. Overall, the model seems capable of generalization to data from new postures, for example see *Train Stand - Test Sit* results in Fig. 3.

4) Influence of Experience: Fig 4 shows the test accuracy results from experiment 4. Again the SVM outperforms the other classifiers. The average classification performance shows slight sensitivity to the strength training status of the subjects. For example, for the data from the *intermediate* training status, the model has higher performance of 87% compared to the other two levels.

C. Onset estimation analysis

Using the error analysis described in section III-E, the error percentages for the onset estimation are evaluated. The results revealed a minimum error of 0.43%, indicating a high level of accuracy in the onset estimation in some scenarios. On the other hand, the maximum error observed is 67.2%, suggesting instances where the model struggled to estimate the fatigue onset. The average error is 12.635% with a standard deviation of 13.14%. These findings show the possible range of performance outcomes, with some cases having excellent accuracy while others exhibiting more significant errors. These results underscores the need for further research. Fig. 5 illustrates the optimal onset estimation, revealing a 1-segment gap (250ms) between the actual and estimated onset location. In Fig. 6, the average onset estimation is depicted, exhibiting a 26-segment gap equivalent to 6.5 seconds. Both results demonstrate advanced reporting of fatigue onset.

V. Conclusion and Future Work

A. Conclusion

In this paper, a novel sEMG dataset, EMAHA-DB6, for strength training activities under different measurement conditions is developed. An ML based approach for fatigue vs. non-fatigue classification of sEMG signals followed by fatigue onset estimation is demonstrated. In the classification stage, several experiments are conducted to analyze the impact of measurement conditions on the classification performance. In the first experiment, the SVM model achieved an highest accuracy of 86% compared to the other models. In the second experiment, the best accuracy ($> 93\%$) is achieved by the SVM in the low load conditions. In the experiment three, the SVM achieved an accuracy of 88.5%. This suggests that the posture of the subjects plays a vital role in the accurate classification of fatigue. In addition to classification, fatigue onset is estimated using the sEMG signals for strength training activity. The estimation performance is evaluated by comparing it with the actual onset. An average error of 12% showcases the promise of our method.

B. Future Work

The future plans include further studies on the fatigue classification and onset estimation. Specifically, deep learning models such as the U-Net designed for time series analysis can be explored. One challenge is to determine if the model can classify and estimate the fatigue onset on data from entirely

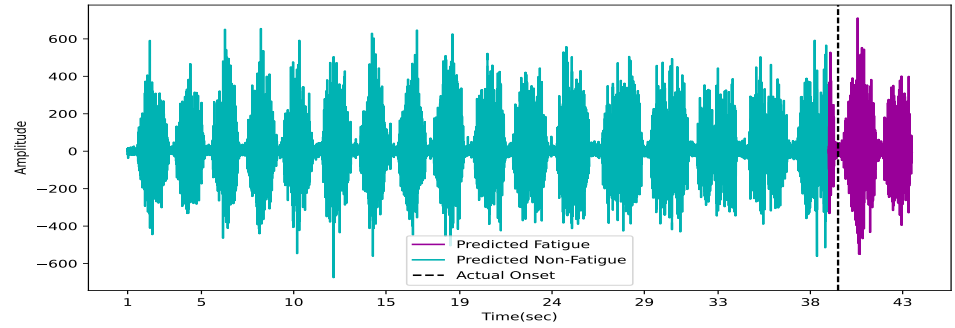


Fig. 5. An illustration of fatigue onset estimation: lowest error performance

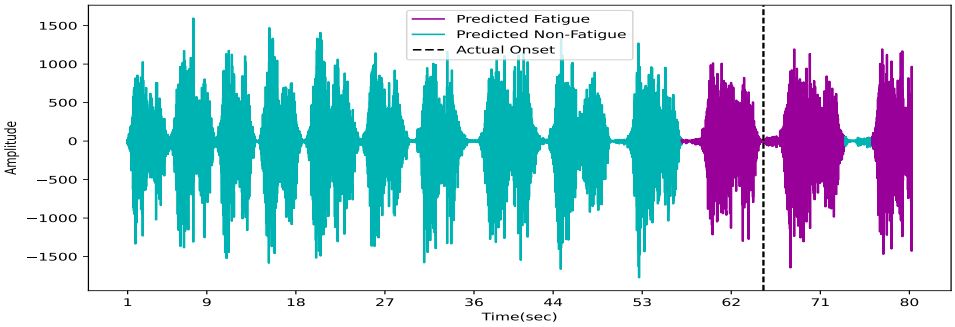


Fig. 6. An illustration of fatigue onset estimation: average error performance

new subjects without further training. In terms of measurement conditions the lifting load and the subject's training status can be further analyzed. Finally, another objective is to determine the amount of training data needed for optimal fatigue classification and onset estimation.

References

[1] R. Kataoka, E. Vasenina, W. B. Hammert, A. H. Ibrahim, S. J. Dankel, and S. L. Buckner, "Is there evidence for the suggestion that fatigue accumulates following resistance exercise?" *Sports Medicine*, vol. 52, no. 1, pp. 25–36, Jan. 2022, ISSN: 1179-2035. DOI: 10.1007/s40279-021-01572-0. [Online]. Available: https://doi.org/10.1007/s40279-021-01572-0.

[2] Q. Liu, Y. Liu, C. Zhang, *et al.*, "Semg-based dynamic muscle fatigue classification using svm with improved whale optimization algorithm," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 16 835–16 844, 2021. DOI: 10.1109/JIOT.2021.3056126.

[3] M. Golmohammadishouraki, *Prediction of fatigue in lower extremity using emg sensor and machine learning.* Thesis (M.App.Sc.) - Carleton University, 2022., Ottawa, 2022.

[4] K. Sahlin, "Muscle fatigue and lactic acid accumulation," *Acta physiologica Scandinavica. Supplementum*, vol. 556, pp. 83–91, 1986, ISSN: 0302-2994. [Online]. Available: http://europepmc.org/abstract/MED/3471061.

[5] M. Chardon, F. Barbieri, T. Penedo, P. C. Santos, and N. Vuillerme, "The effects of experimentally-induced fatigue on gait parameters during obstacle crossing: A systematic review," *Neuroscience Biobehavioral Reviews*, vol. 142, p. 104 854, 2022, ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2022.104854. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763422003438.

[6] B. Bigland-Ritchie, R. Johansson, and J. Woods, "Does a reduction in motor drive necessarily result in force loss during fatigue," *Biochemistry of exercise*, pp. 864–870, 1983.

[7] H. Kranz, A. M. Williams, J. Cassell, D. J. Caddy, and R. B. Silberstein, "Factors determining the frequency content of the electromyogram," *Journal of Applied Physiology*, vol. 55, no. 2, pp. 392–399, 1983, PMID: 6311777. DOI: 10.1152/jappl.1983.55.2.392. eprint: https://doi.org/10.1152/jappl.1983.55.2.392. [Online]. Available: https://doi.org/10.1152/jappl.1983.55.2.392.

[8] T. Moritani, M. Muro, and A. Kijima, "Electromechanical changes during electrically induced and maximal voluntary contractions: Electrophysiologic responses of different muscle fiber types during stimulated contractions," *Experimental Neurology*, vol. 88, no. 3, pp. 471–483, 1985, ISSN: 0014-4886. DOI: https://doi.org/10.1016/0014-4886(85)90064-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0014488685900640.

[9] K. Marri and R. Swaminathan, "Classification of muscle fatigue in dynamic contraction using surface electromyography signals and multifractal singularity spectral analysis," *Journal of Dynamic Systems, Measurement, and Control*, vol. 138, no. 11, 2016.

[10] M. Li, J. Li, and M. Shu, "Detection of muscle fatigue by fusion of agonist and synergistic muscle semg signals," in *2020 ieee 33rd international symposium on computer-based medical systems (cbms)*, IEEE, 2020, pp. 95–98.

[11] L. Kahl and U. G. Hofmann, "Comparison of algorithms to quantify muscle fatigue in upper limb muscles based on semg signals," *Medical Engineering & Physics*, vol. 38, no. 11, pp. 1260–1269, 2016.

[12] E. R. T. Santos Junior, B. F. de Salles, I. Dias, A. S. Ribeiro, R. Simão, and J. M. Willardson, "Classification and determination model of resistance training status," *Strength and Conditioning Journal*, vol. 43, no. 5, pp. 77–86, 2021.

[13] D. Yang, W. Yang, Q. Huang, and H. Liu, "Classification of multiple finger motions during dynamic upper limb movements," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 134–141, Jan. 2017. DOI: 10.1109/JBHI.2015.2490718.

[14] N. K. Karnam, S. R. Dubey, A. Turlapaty, and B. Gokaraju, "Emghandnet: A hybrid cnn and bi-lstm architecture for hand activity classification using surface emg signals," *Biocybernetics and Biomedical Engineering*, 2022.

[15] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for emg signal classification," *Expert Systems with Applications*, vol. 39, pp. 7420–7431, Jun. 2012. DOI: 10.1016/j.eswa.2012.01.102.

[16] R. Azhiri, M. Esmaeili, and M. Nourani, "Emg-based feature extraction and classification for prosthetic hand control," Jul. 2021.

P300 Event-Related Potential in Perception of Multiple Traffic Objects During Vehicle Driving

Yuki Yamamoto*, Sou Nobukawa **, Nobuhiko Wagatsuma ***, and Keiichiro Inagaki*
*Chubu University, Kasugai, Aichi, Japan
E-mail: tr22013-5881@sti.chubu.ac.jp, kay@isc.chubu.ac.jp Tel: +81-568-51-1111
**Chiba Institute of Technology, Tsudanuma, Chiba, Japan
E-mail: nobukawa@cs.it-chiba.ac.jp Tel: +81-47-478-0538
***Toho University, Funabashi, Chiba, Japan
E-mail: nwagatsuma@is.sci.toho-u.ac.jp Tel: +81-47-472-1176

Abstract — **Visual attention is a function of the brain that allocates its computational resources to enable us to attend to the most important visual information. In actual driving, there are many perceptual objects such as pedestrians, signals, and other vehicles, etc. To selectively recognize these traffic objects for safe driving, the driver's gaze is appropriately controlled under visual attention. It is known that the P300 event-related potential related to visual attention, and is an essential brain activity to understanding the visual-attention characteristics in visual search. Although the P300 response reflects visual attention, the response characteristics of the P300 especially regarding the multiple object search are largely unexplored. To understand the brain activity on visual attention for multiple objects, we measure the P300 response and examined especially its characteristics for the types of traffic objects in terms of the capacity of induction for visual attention.**

I. INTRODUCTION

Traffic accident is one of the crucial social problems. Recent highly developed various types of intelligent transport systems (ITS) contribute to reducing the occurrence of traffic accidents [1]. In spite of this development of ITS technology, human error-related traffic accidents are difficult to prevent [2].

A variety of human errors leading to traffic accidents such as miss perception, oversight, distracted driving, careless driving, etc. [2-7] is known, and most of those factors are related to vision. During vehicle driving, 90% of the information related to driving is based on vision. During driving, information on the driving scene is visually grasped and processed in the brain to determine appropriative driving behavior. In this processing, previously mentioned human visual processing errors lead to delays in the detection of traffic objects such as pedestrians, other vehicles, signals, etc., then these also make lose control of vehicle maneuvers. Given these factors, it is essential to understand brain signal processing regarding driving attention and perception in order to reduce road traffic accidents of visual origin.

Visual attention is a function of the brain that allocates its computational resources to enable us to attend to the most important current information [8]. In actual driving, there are many perceptual objects such as pedestrians, signals, and other vehicles, etc. Therefore, the driver should distribute the computational resource of visual attention to these objects due to that their recognition is closely related to safe driving behavior. The distribution of visual attention during driving was revealed by eye movement research—shifting attention to the side lane [9], fixating on the location where the vehicle will be in the next few seconds [10], varying the covered field of view (e.g. [11-13]). On this process, it is revealed that the attentional shift is related to the visual saliency that is determined by visual information of the scene such as color, intensity, orientation [14-16]. Despite large amounts of evidence for visual attention and visual search during driving, the brain activity for visual attention in visual search of multiple objects is largely unexplored.

In the present study, we focused on event-related potential P300, which is strongly correlated with visual attention [17,18], to explore the brain activity in the visual search of multiple objects during driving. To further investigate relationship between P300 and visual perception in driving, we examined the characteristics of P300 for the different types of traffic objects in terms of the capacity of induction for visual attention.

II. METHODS

A. Subjects

This study recruited 11 healthy male subjects aged 21-24 (22.18±1.03) years to participate in EEG measurement during visual perception of multiple traffic targets. All subjects had a

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

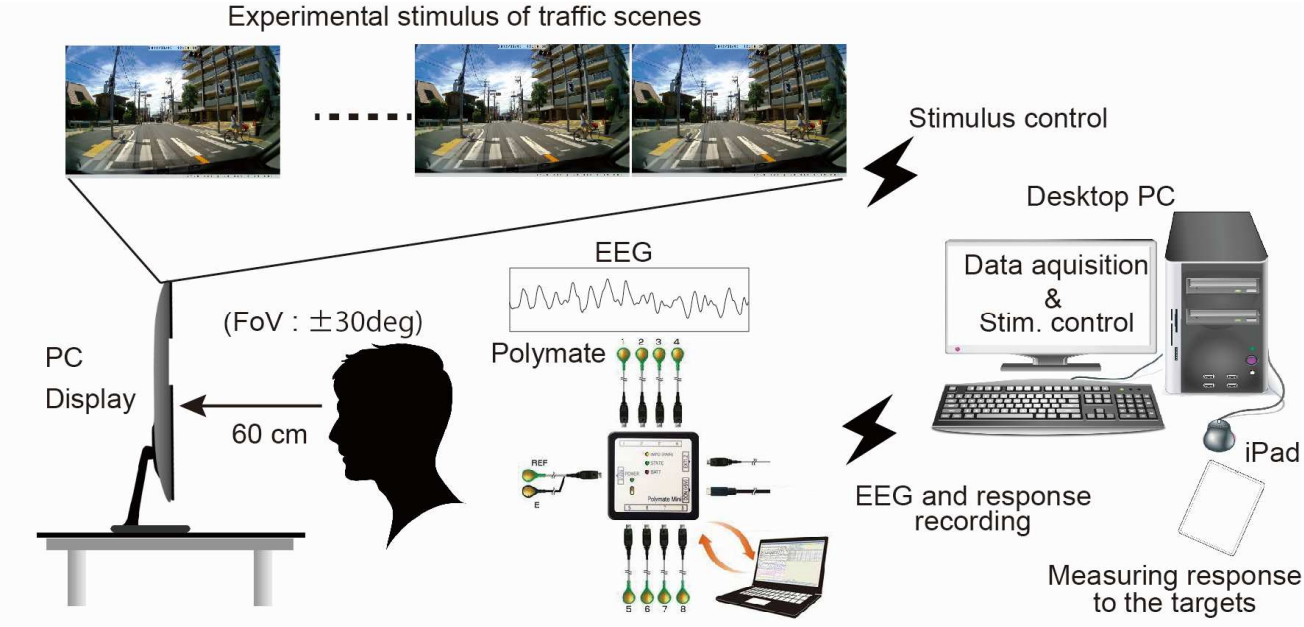


Fig. 1. Experimental setup for stimulus control, EEG and response recording.

valid driver's license, normal or corrected-to-normal vision using glasses or contact lenses which was specified in the license, and presented normal EEG. Here, we checked the arousal condition of the subjects using the Karolinska Sleepiness Scale [19] translated into Japanese, and subjects who reported more than level 3 (criteria of drowsy) did not participate in the experiments.

B. Experimental setup

Figure 1 illustrated the experimental setup for measuring EEGs during visual perception of the traffic objects. The experimental setup includes a chin rest, a driver's seat, , display (1920 × 1080), iPad utilized for measuring response to the traffic targets in perception, and a personal computer (PC) utilized for stimulus control and EEG acquisition. The 34 scenes of the vehicle driving were utilized as the visual stimulus. The distance between the projected PC display and the subject was maintained at 60 cm, which simulated the experience of watching traffic scenes from the driver's seat in a vehicle, and this condition imitates the field of view was ±30°, which encompassed the functional field of view during driving (approximately ±20°) [20].

C. Experimental Stimulus

We utilized 34 driving scenes captured by video camcorder in the experiments. In preceding the experiment, we carefully searched the traffic objects including road signs, signals, other vehicles, pedestrians, cross road, and intersections, which need perception for safe driving, and determined the timing of

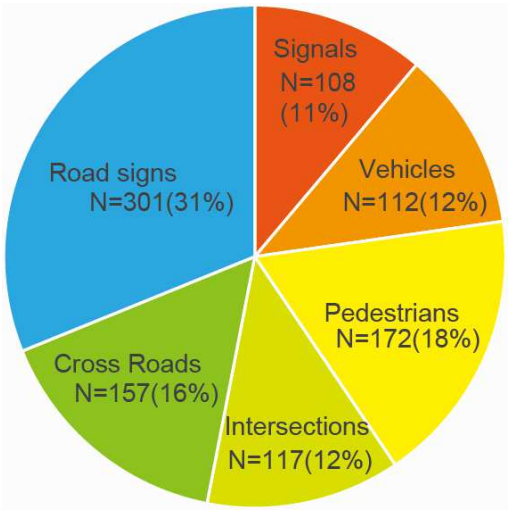


Fig. 2. Percentage of appearance of traffic objects.

appearance of the targets and the fraction of them. In this experiment, we defined the timing of appearance of the traffic objects when center of them was entered in 25% of the vertical area centered from the point at infinity. Figure 2 illustrated the percentage of the appearance of the traffic objects. In the case of the current driving scenes dataset, the fraction of each traffic target was 11%, 12%, 18%, 12%, 16%, and 31% for signals, vehicles, pedestrians, intersections, cross road, and road signs, respectively.

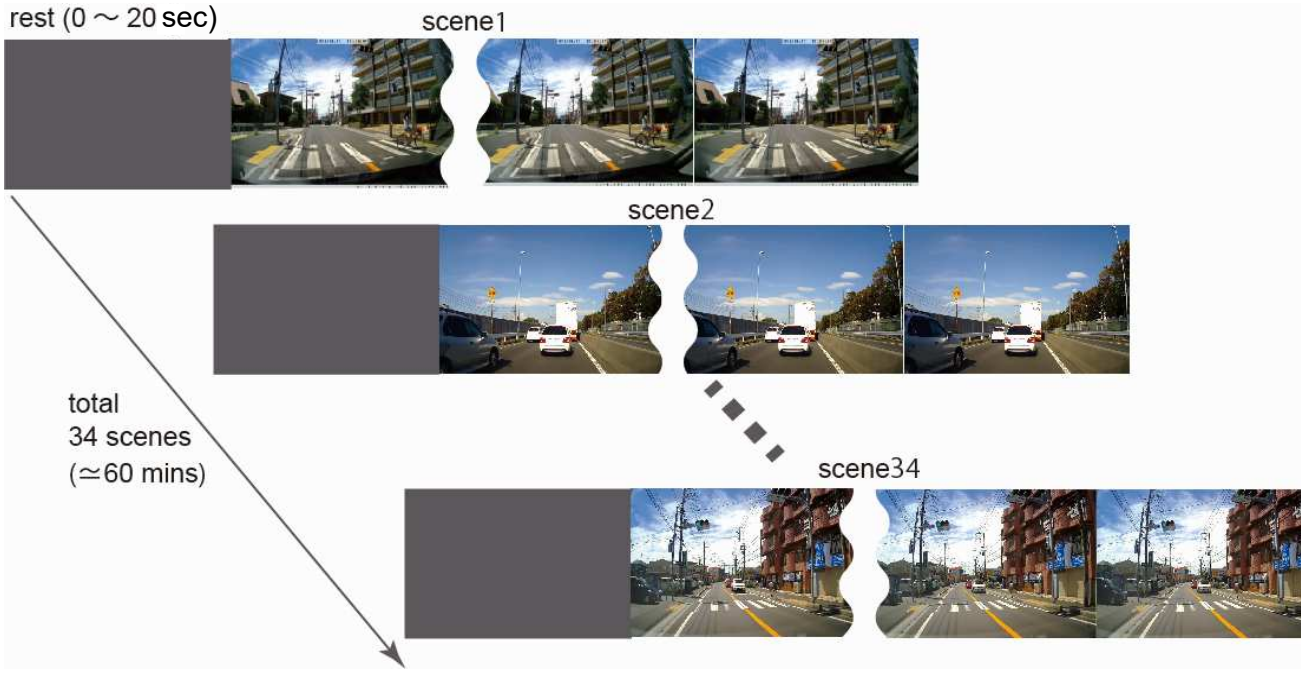


Fig. 3. Experimental stimulus and protocol. 34 traffic scenes were utilized in the experiment. Each scene consists of 20 seconds gray image for resting and traffic scenes. The duration of each traffic scene was approximately 2 minutes.

D. Experimental protocol

We first explained the experimental design to the subjects and obtained informed consent. Figure 3 summarized the stimulus of the driving scene and the experimental protocols. Each driving scene had 20 seconds of gray image for resting and the driving in the urban area with approximately 2 minutes length. These scenes were randomly projected to the PC display. We asked the subjects to watch the driving scenes and pushed the reaction button on the iPad as fast as possible when they could perceive the traffic objects. The Ethics Committee of Chubu University approved the study protocol (20210016). We performed all methods in accordance with the Declaration of Helsinki.

E. EEG recording

EEGs of all subjects were measured using Polymate Mini AP108 (Miyuki-Giken Co., Ltd.) with a 500-Hz sampling frequency. Standard scalp EEG electrodes were located in accordance with the International 10–20 System [21]. We employed 9-channel EEG at 7-electrode sites (AF3, AF4, T7, T8, Cz, O1, and O2 to cover entire brain area), and at 2-reference sites (A1 and A2).

F. Analysis

Topographic mapping of P300 event-related potential found that the P300 was dominantly distributed in the central to parietal derivations, without any left or right lateralization. Furthermore, P300 event-related potential was determined by averaging 20–30 EEG responses to the target paid attention [18]. In the present study, we focused on the EEG measured from Cz-channel.

At first, the EEG signal was filtered into the delta band (1 – 5 Hz) using a band-path filter due to that the P300 event-related potential most commonly observed in the delta band. To calculate the average response to determine P300 event-related potential, the filtered EEG between the pre- (-0.2sec) and post-appearance of the traffic target (0.8sec) was selected for each type of traffic object with adjustment of that DC level at the timing of appearance of traffic target become 0, and averaged.

Several definitions were reported to describe the P300 event-related response [18], which vary depending on stimulus modality, task conditions, and age. The EEG typically displays a peak response between 0.25 and 0.80 sec following stimulus presentation, defined as P300. In the present study, we defined the P300 event-related response as the peak response identified between 0.25 and 0.80 sec after the appearance of triggered traffic object. Note that if two peaks were observed those range, we chose second peak for the P300 event-related response to avoid choosing peak observed between N100 and N200.

III. RESULTS

We calculated the P300 event-related potential for each type of traffic object. Figure 4 summarized the P300 event-related

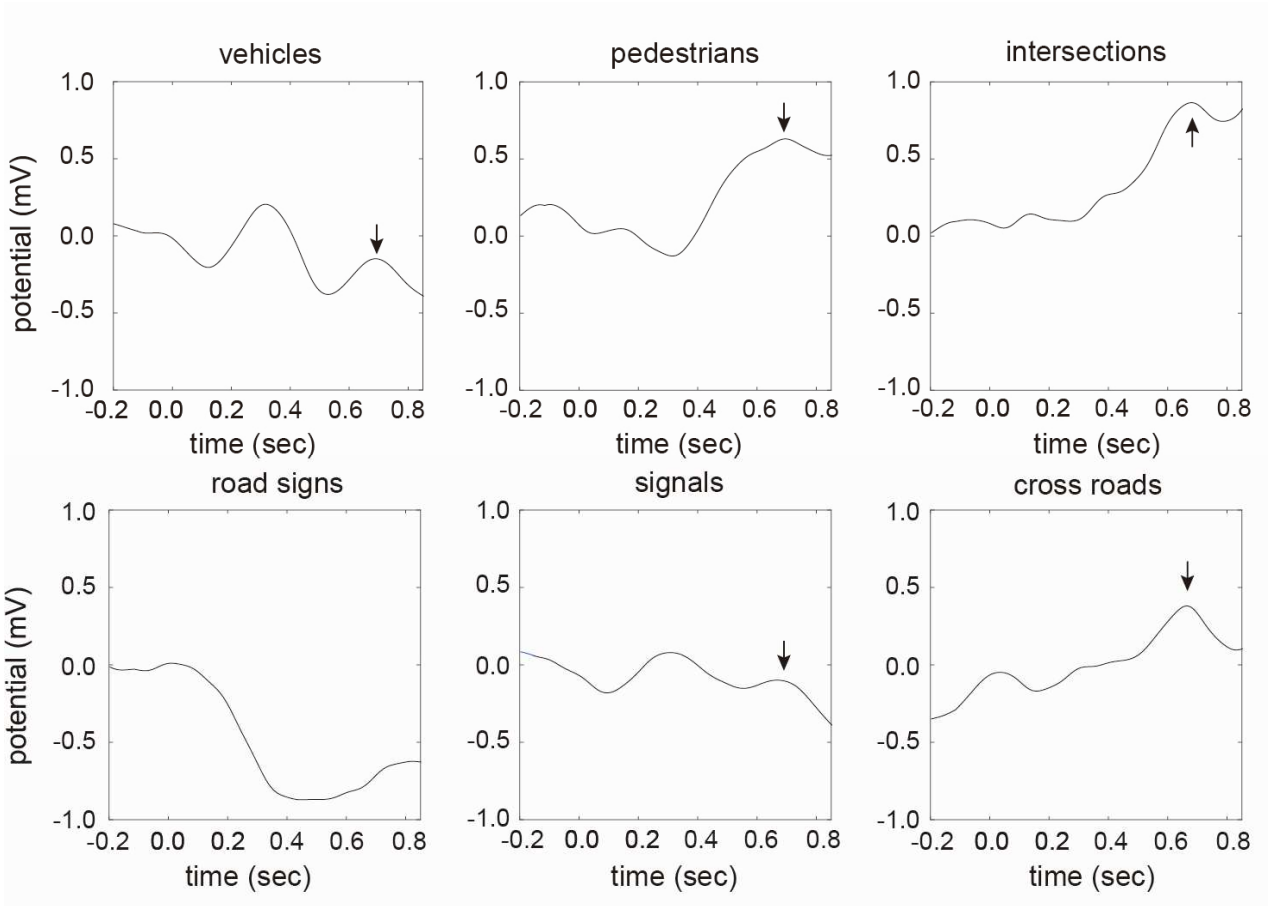


Fig. 4. Average response of P300 event-related potential for the different traffic objects. The arrow indicates P300 event-related potential.

potential for 6 different traffic object during perception of the driving scenes. For the intersection, and cross road, the P300 event-related potential with marked peak response was observed around 0.65-0.70 sec. For the vehicle, pedestrian, and signal, the P300 event-related potential response was also found around 0.65-0.70 sec; however, this response was dull (small). For the road sign, no P300 event-related potential was found.

The intensity of the attention for the traffic objects in the driving scene is different due to their characteristics such as color, luminosity, orientation, etc., and driver's experience and intention. Next, we investigated individual P300 event-related potential for the traffic objects. Table 1 showed the observation rate of P300 event-related potential for the different types of traffic objects. For the cross road and intersection, the observation rate of P300 event-related potential was, 81.8% and 81.8%, and those were relatively higher among the 6 different types of traffic objects. For the vehicle, pedestrian, and signal, the observation rate of P300 event-related potential was, respectively, 72.7%, 63.6%, and 54.5%. For the road sign, the observation rate of P300 event-related potential was 9.1%.

IV. Discussion

The present study focused on the P300 event-related potential, which is accompanied by attention, and investigated the P300 event-related potential response in visual search of multiple traffic targets. In our experimental protocol, we asked the subjects to pay attention and make a perception of traffic objects during watching driving scenes, and tap the reaction button when they perceive those traffic objects. As the results, we found that the occurrence of the P300 event-related potential in the multiple visual search depended on the types of traffic objects.

The occurrence of P300 event-related potential depended on the type of traffic object. At this point, the shapes of waveform for average response of P300 event-related potential (Figure 4) related to the observation rate of the P300 event-related potential for each type of traffic object. Namely, the clear response for P300 event-related potential appeared when the

Table. 1. Observation rate of P300 event-related potential for the different types of traffic objects

object type	P300 observation rate
vehicles	72.7% (8/11)
pedestrians	63.6% (7/11)
intersections	81.8% (9/11)
road signs	9.1% (1/11)
signals	54.5% (6/11)
cross roads	81.8% (9/11)

occurrence rate of the P300 event-related potential is high, and vice versa. Note that, even if a clear waveform of P300 event-related potential was not observed, a small number of P300 event-related potential occurred (Table1, the case of road sign).

The variability for observation of P300 event-related potential might indicate the essential ability of traffic objects to induce drivers' attention. Namely, when the ability of traffic objects to induce drivers' attention is originally high, observation of P300 event-related potential become high, and vice versa. Visual attention is a function of the brain that optimize computational resources for attention using the information from the visual field to enable us to attend to the most important current information [22] and enhance the visual perception of scenes [23]. By this process, it is revealed that the attentional shift is related to the visual saliency that is determined by visual information of the scene such as color, intensity, orientation [14-16]. In the future study, we will investigate relationship between the ability of traffic object to induce drivers' attention and characteristics of P300 event-related potential considering visual saliency of traffic objects.

V. Conclusion

We focused on the P300 event-related potential response on the traffic objects which need to make perception for the determination of appropriate driving behavior in safe driving. We found that the occurrence of P300 event-related potential response depends on the type of traffic objects. By this investigation, a part of the P300 event-related potential response for the multiple visual search could be quantified using the case of driving. In the future study, we further investigate the relationship between the ability of traffic objects to induce drivers' attention and P300 response considering the visual saliency of the traffic objects.

References

[1] Japanese National Police Agency. The Occurrence of Traffic Accidents and the Traffic Violation Control in 2017; Japanese National Police Agency: Tokyo, Japan, 2017.

[2] T.A. Dingus, F. Guo, S. Lee, J.F. Antin, M. Perez, M. Buchanan-King, J. Hankey, Driver crash risk factors and prevalence evaluation using naturalistic driving data. PNAS, vol.113, pp.2636–2641, 2016.

[3] J.R. Treat, N.S. Tumbas, S.T. McDonald, D. Shinar, R.D. Hume, R.E. Mayer, R.L. Stansifer, N.J. Castellan, N.J. Tri-Level Study of the Causes of Traffic Accidents: Final Report. Volume I: Causal Factor Tabulations and Assessments; Indiana University Institute for Research in Public Safety: Bloomington, IN, USA, 1979.

[4] K. Rumar, The role of perceptual and cognitive filters in observed behaviour. In Human Behaviour and Traffic Safety; Springer: Boston, MA, USA, pp. 151–167, 1985.

[5] T. Allahyari, G.N. Saraji, J. Adl, M. Hosseini, M. Iravani, M. Younesian, S.J. Kass, Cognitive failures, driving errors and driving accidents. Int. J. Occup. Saf. Ergon., vol.14, pp.149–158, 2008.

[6] E. Petridou, M. Moustakim, Human factors in the causation of road traffic crashes. Eur. J. Epidemiol., vol.16, pp.819–826, 2000.

[7] K.Bucsuházy, E.Matuchová, R.Zůvala, P.Moravcová, M.Kostíková, R.Mikulec, Human factors contributing to the road traffic accident occurrence. Transp. Res. Procedia, 45, 555–561, 2020.

[8] M.I. Posner, Orienting of attention. Quarterly Journal of Experimental Psychology, vol.32, pp.3–25, 1980.

[9] K. Sato, Visual search and peripheral vision performance by novice and experienced drivers. IATSS Rev., vol.19, pp.191–199, 1993.

[10] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, D. Crundall, Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. Ergonomics, vol.46, pp.629–646, 2003.

[11] D. Crundall, G. Underwood, Effects of experience and processing demands on visual information acquisition in drivers. Ergonomics, vol.41, pp.448–458, 1998.

[12] G. Underwood, P. Chapman, K. Bowden, D. Crundall, Visual search while driving: Skill and awareness during inspection of the scene. Transp. Res. Part F: Traffic Psychol. Behav. vol.5, pp.87–97, 2002.

[13] K. Inagaki, T. Maruno, K. Yamamoto, Evaluation of EEG activation pattern on the experience of visual perception in the driving. IEICE Trans. Info. Sys. , E103, 2032–2034, 2020.

[14] A.M. Treisman, G. Gelade, A feature-integration theory of attention. Cognitive Psychology, vol.12, no.1, pp.97–136, 1980.

[15] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. vol.4, no.4, pp.219-27, 1985.

[16] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions PAMI, vol.20, no.11, pp.1254-1259, 1998.

[17] D. Friedman, Y.M.Cycowicz, H. Gaeta, The novelty P3: An event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci. Biobehav. Rev. vol.25, pp.355–373, 2001.

[18] J. Polich, Updating P300: An integrative theory of P3a and P3b. Clin. Neurophysiol., 118, 2128–2148, 2007.

[19] T. Åkerstedt, M. Gillberg, Subjective and objective sleepiness in the active individual. Int. J. Neurosci., vol.51, no.1-2, pp.29-37, 1990.

[20] R. Tanishige, K. Doman, D. Deguchi, Y. Mekada, I. Ide, H. Murase, A study on the prediction of driver's pedestrian detectability considering characteristics of human fields-of-view while driving., IEICE PRMU2014-196. vol.114, pp.223-228, 2015.

[21] H.H.Jasper, The ten twenty electrode system of the international federation. Electroencephalography and Clinical Neurophysiology, 10, 371–375,1958.

[22] M.I. Posner, Orienting of attention. Quarterly Journal of Experimental Psychology, vol.32, pp.3–25, 1980.

[23] M. Carrasco, Visual attention: The past 25 years. Vis. Res. vol.51, pp.1484–1525, 2011.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Kernel Random Projection Depth for Outlier Detection

Akira Tamamori
Department of Information Science, Aichi Institute of Technology, Aichi, Japan
E-mail: akira-tamamori@aitech.ac.jp

***Abstract*—This paper proposes an extension of Random Projection Depth (RPD) to cope with multiple modalities and non-convexity on data clouds. In the framework of the proposed method, the RPD is computed in a reproducing kernel Hilbert space. With the help of kernel principal component analysis, we expect that the proposed method can cope with the above multiple modalities and non-convexity. The experimental results demonstrate that the proposed method outperforms RPD and is comparable to other existing detection models on benchmark datasets regarding Area Under the Curves (AUCs) of Receiver Operating Characteristic (ROC).**

I. Introduction

Outlier detection refers to "the problem of finding patterns in data that do not conform to expected normal behavior" [1]. Outlier detection has broad applicability and many application examples, such as fraud detection, intrusion detection, disease detection, medical diagnosis, etc. In supervised learning of classification problems, the authors have reported that utilizing outlier detection for data cleaning, i.e., removing anomalous samples as a preprocessing step, significantly improves the classification accuracy [2]. Accordingly, outlier detection is one of the important research topics in the field of data mining and machine learning.

Outlier detection algorithms often quantify the *outlyingness* (outlierness) of points on a data cloud in a multidimensional space. The opposite of outlyingness is the depth [3]. It measures the centrality of a data point, which defines how close the point is to the center of the data cloud. In other words, the depth introduces a center-outward ordering to the data cloud. Since we can construct analogs of quantiles and ranks in multidimensional datasets from the depth, many authors successfully have applied the depth function in related research fields, such as robust estimation [4], nonparametric tests [5], cluster analysis [6], discriminant analysis [7], multivariate regression [8], multivariate density estimation [9], functional data analysis [10], etc. Authors proposed various depth functions in the literature listed as follows: halfspace depth [11], simplicial volume depth [12], zonoid depth [13], spatial depth [14], projection depth [3, 15], Mahalanobis depth [3, 16], and so on. Among these depth functions, it is known that the halfspace depth and projection depth are favorable since they satisfy the desirable properties of the statistical depth [3]; affine invariance, maximality at the center, monotonicity on rays, null at infinity, etc. We also focus on the projection depth in this study because of these desirable properties. Furthermore, authors in the previous work studied an extension of the spatial depth-based outlier detection with kernel method [17]. The other study proposed an extension of Mahalanobis depth in the reproducing kernel Hilbert space [18]. Their studies are also related to ours because we consider an extension of the projection depth with the kernel method for outlier detection. In the following, we will describe our motivation in detail.

We have examined the performance of outlier detection methods based on Random Projection Outlyingness (RPO). The detection performance will degrade when data samples are not linearly separable, and the sample distribution exhibits multiple modalities. To cope with the such modalities, we have proposed the RPO-based detection method with a local score weighting [19]. However, the detection performance still needs to improve since it cannot cope with a nonlinearly separable data cloud. The RPO involves inner product computations, i.e., linear operations in the original feature space. In contrast, kernel-based detection methods such as One-Class Support Vector Machine (OCSVM) [20] can deal with such nonlinearly separable distributions using a feature map into Reproducing Kernel Hilbert Space (RKHS), and the kernel function induces the inner product in the RKHS. Combining kernel methods and projection depth will open the possibility of coping with nonlinear separability.

Given a large amount of training data available, we can use detection methods based on deep learning, such as Deep Support Vector Data Description (Deep SVDD) [21] and Deep AutoEncoding Gaussian Mixture Model [22]. These methods first perform a nonlinear mapping which maps the original features into a new feature space, and the detector is trained on the new space. Especially concerning RPO, Deep RPO [23] has been proposed; it combines Deep SVDD and RPO in the training objective. The authors demonstrated that it could improve detection performance than Deep SVDD and RPO. However, training such models on a small dataset will be challenging because sufficient data samples are only sometimes available.

This study proposes an outlier detection method based on Kernel Random Projection Depth (KRPD). A reproducing kernel function maps original feature vectors into the RKHS, and we can compute RPD on the RKHS. Although authors have proposed several methods to obtain a nonlinear mapping related to kernel approximation [24, 25], we adopt Kernel Principal Component Analysis (KPCA) to reduce the dimensionality of the feature vector. This dimensionality reduction helps make the algorithm feasible and avoids the curse of

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

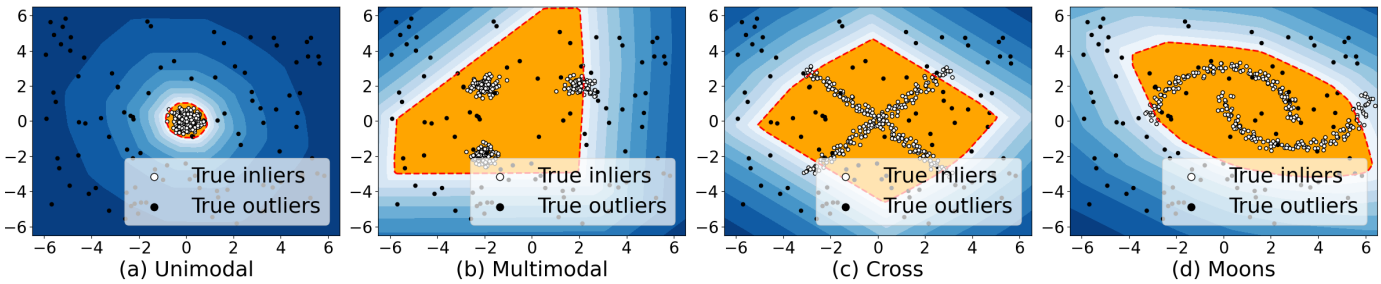


Fig. 1. Contour plots of RPD. The decision boundary is indicated by the red dot line and the central region is colored orange.

dimensionality related to the computations of an inner product in the RKHS. We conducted experiments on benchmark datasets for outlier detection. The results demonstrated that the proposed method outperformed RPD and performed better than other competitive detection methods, including OCSVM, Local Outlier Factor [26], Isolation Forest [27], and Gaussian Mixture Model [28], in terms of Area Under the Curves (AUCs) of Receiver Operating Characteristic (ROC).

II. Random Projection Depth

Suppose we are given a d-variate probability distribution P and a data cloud consisting of N data samples

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}, \quad \mathbf{x}_n \sim P, \quad \mathbf{x}_n \in \mathbb{R}^d. \tag{1}$$

Random Projection Outlyingness (RPO) $O_d(\mathbf{x} \mid X)$ on X is defined as

$$O_d(\mathbf{x} \mid X) := \sup_{\mathbf{u} \in \mathbb{S}^{d-1}} \frac{|\mathbf{u}^T\mathbf{x} - \mathrm{MED}(\mathbf{u}^T X)|}{\mathrm{MAD}(\mathbf{u}^T X)}, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a d-dimensional data point and $\mathbf{u}$ is a random unit vector drawn from the unit hypersphere $\mathbb{S}^{d-1}$. The one-dimensional, scalar-valued dataset $\mathbf{u}^T X$ is given by $\mathbf{u}^T X = \{\mathbf{u}^T\mathbf{x}_1, \mathbf{u}^T\mathbf{x}_2, \ldots, \mathbf{u}^T\mathbf{x}_N\}$. The abbreviations MED and MAD denote the median and the median absolute deviation, respectively. In accordance with the above definition, we can define Random Projection Depth (RPD) $D_d(\mathbf{x} \mid X)$ [29] as

$$D_d(\mathbf{x} \mid X) := \frac{1}{1 + O_d(\mathbf{x} \mid X)}. \tag{3}$$

Generally, the depth function is a statistic that gives a center-outward ordering to each data point $\mathbf{x}$. The point at which the depth function takes its maximum value is the median, a *deepest* point. RPD has the characteristic of not being able to deal with multimodal distributions. When we utilize RPD for outlier detection, this characteristic will likely decrease detection accuracy and robustness. To address this problem, we have proposed a method of weighting the outlyingness based on the cardinality of the clusters to which the samples belong [19]. However, this method requires the number of clusters to be determined in advance. Another characteristic of RPD is that the central region forms a convex hull, which also causes an increase in false positive errors when the data cloud is non-convex.

In order to demonstrate the above issues of RPD, we show contour plots and decision surfaces across synthetic datasets in Figure 1. In this figure, datasets are: (a) Unimodal, (b) Multimodal, (c) Cross, and (d) Moons. We generated each dataset with 400 samples (300 inliers, 100 outliers) in two-dimensional space. The true outlier samples were generated by a uniform distribution within $[-6, 6] \times [-6, 6]$. For Unimodal, the inlier samples were generated by a scaled standard normal distributions where the scale was set to 0.3. For Multimodal, the inlier samples were generated by three scaled standard normal distributions. The scale of each normal distribution was also the same value as 0.3. We generated inliers for Cross on the two vertical segments and Moons on the shifted semicircles. Small noises were added to those inliers; the strength of noise was set to 0.05 uniformly.

In this study, we utilize a negative projection depth as the outlier score; a high-depth value corresponds to an inlier, and a low-depth value corresponds to an outlier. We can identify a data point as an outlier if its outlier score is greater than a threshold [17]. The threshold was determined as 25 $(= 100 \times 100/400)$ percentile within whole predicted scores. We specify the central region as a set of points where its outlier score exceeds the threshold. The region is colored orange in the figure. We refer to the boundary of the central region as the decision boundary. The central region should include many inliers and few outliers in practice. On Unimodal, we can see that RPD successfully produces a tight boundary and the decision surface encloses almost all inliers. In this case, we can expect that outlier detection works well. However, on other datasets including Multimodal we can see that RPD fails to form tight decision boundaries to inliers and encloses inliers and many outliers. The false positive errors will increase in such cases.

III. Proposed method

We propose to compute RPD in an RKHS to resolve the issues described in the previous section, where RPD-based detection will fail on the multi-modal or not lineally separable data cloud. We refer to RPD in the RKHS as Kernel RPD (KRPD). The kernel method [30] makes the computation of KRPD feasible. In the following, we formulate the proposed method.

Suppose we are given a positive definite real-valued kernel

$$k : \Omega \times \Omega \to \mathbb{R}, \tag{4}$$

where Ω is a non-empty subset of $\mathbb{R}^d$. We refer the RKHS induced by k as $\mathcal{H}_k$ and define a feature map

$$\Phi : \Omega \to \mathcal{H}_k,\ \mathbf{x} \in \Omega \mapsto k(\cdot, \mathbf{x}). \tag{5}$$

The data cloud X is mapped to the data cloud $\mathcal{F}_k(X)$ on $\mathcal{H}_k$:

$$\mathcal{F}_k(X) = \{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_N)\} \subset \mathcal{H}_k. \tag{6}$$

We can define the outlyingness $O_k(\mathbf{x} \mid \mathcal{F}_k(X))$ and the projection depth $\mathcal{D}_k(\mathbf{x} \mid \mathcal{F}_k(X))$ on $\mathcal{H}_k$ sequentially as

$$O_k(\mathbf{x} \mid \mathcal{F}_k(X)) \coloneqq \sup_{\|f\|=1} \frac{|\langle f, \Phi(\mathbf{x})\rangle - \mathrm{MED}(\langle f, \mathcal{F}_k(X)\rangle)|}{\mathrm{MAD}(\langle f, \mathcal{F}_k(X)\rangle)}, \tag{7}$$

and

$$\mathcal{D}_k(\mathbf{x} \mid \mathcal{F}_k(X)) \coloneqq \frac{1}{1 + O_k(\mathbf{x} \mid \mathcal{F}_k(X))}, \tag{8}$$

where $\|\cdot\|$ denotes the norm in $\mathcal{H}_k$. The $\langle f, \mathcal{F}_k(X)\rangle$ is an abbreviation of the scalar-valued data cloud

$$\langle f, \mathcal{F}_k(X)\rangle = \{\langle f, \Phi(\mathbf{x}_1)\rangle, \langle f, \Phi(\mathbf{x}_2)\rangle, \ldots, \langle f, \Phi(\mathbf{x}_N)\rangle\}, \tag{9}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathcal{H}_k$. Through nonlinear kernel function k and its feature map, we expect that the projection depth is robust to the distribution with multi modality and non-convexity.

Since the dimensionality of $\mathcal{H}_k$ is not necessarily finite, the unit sphere $\mathcal{S}_k = \{f \in \mathcal{H}_k \mid \|f\| = 1\}$ is generally not a compact set. This leads to the case that the supremum of Eq.(7) cannot be approximated through a sequence of elements in $\mathcal{S}_k$ and the computation of the projection depth becomes infeasible. To derive the computationally feasible algorithm, we resort to the representer theorem [30]. We can assume that an element $f \in \mathcal{S}_k$ belongs to the linear span of $\mathcal{F}_k(X)$:

$$f = \sum_{n=1}^{N} \alpha_n \Phi(\mathbf{x}_n) = \sum_{n=1}^{N} \alpha_n k(\cdot, \mathbf{x}_n), \tag{10}$$

where $(\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathbb{R}^N$. The condition $\|f\| = 1$ can be rewritten as

$$\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} = 1, \tag{11}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ and $\mathbf{K}$ is the $N \times N$ Gram matrix:

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & k(\mathbf{x}_N, \mathbf{x}_2) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}. \tag{12}$$

We can indeed compute the projection depth approximately in $\mathcal{H}_k$ by sampling $\boldsymbol{\alpha}$ from the hyperellipsoid of $\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} = 1$. However, the inner product between high dimensional vectors suffers from the curse of dimensionality; almost all vectors in a high dimensional space are orthogonal to each other [31, 32]. This characteristic will significantly reduce the efficiency and accuracy of depth computation. To improve the efficiency and accuracy, we propose to perform a dimensionality reduction by KPCA [33] before computing projection depth. In this paper, we denote the principal dominant components as $\boldsymbol{\beta}$ for a data point $\mathbf{x}$ and its dimensionality as M. We also denote the new data cloud as $F(X) = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N\}$. By treating $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_n$ as new coordinates in $\mathbb{R}^M$, the projection depth can be computed in the same manner as in $\mathbb{R}^M$. If setting $M < N$, we can simultaneously achieve a nonlinear dimensionality reduction and a feasible computation of the projection depth. The overall computation algorithm is summarized in Algorithm 1. In the step 5), the projected data cloud $\mathbf{u}_l^T F(X)$ is defined as $\mathbf{u}_l^T F(X) \coloneqq \{\mathbf{u}_l^T \boldsymbol{\beta}_1, \mathbf{u}_l^T \boldsymbol{\beta}_2, \ldots, \mathbf{u}_l^T \boldsymbol{\beta}_N^T\}$. The projection depth $\mathcal{D}_k(\mathbf{x} \mid \mathcal{F}_k(X))$ is approximated as

$$\mathcal{D}_k(\mathbf{x} \mid \mathcal{F}_k(X)) \approx D_M(\boldsymbol{\beta} \mid F(X)) \tag{13}$$

$$\approx \left(1 + \max_{1 \le l \le L} \frac{|\mathbf{u}_l^T \boldsymbol{\beta} - \mathrm{MED}(\mathbf{u}_l^T F(X))|}{\mathrm{MAD}(\mathbf{u}_l^T F(X))}\right)^{-1}. \tag{14}$$

We also utilize a negative projection depth as the outlier score in the same way as RPD. The outlier detection algorithm is also the same.

Algorithm 1 Computation of outlier score based on KRPD.

Input: Data cloud $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \Omega$, kernel function k, dominant dimensionality M, the number of random projection vectors L, and a data point $\mathbf{x} \in \Omega$.

Output: Negative projection depth $-D_M(\boldsymbol{\beta} \mid F(X))$.

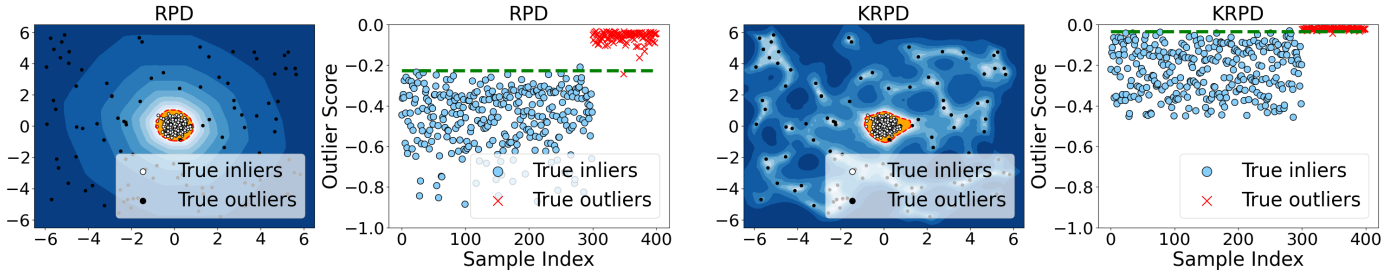
1: Compute the Gram matrix $\mathbf{K}$ between $\mathbf{x}$ and X, and obtain the centralized matrix $\mathbf{K}'$ from $\mathbf{K}$ [30].
2: Compute principal components $\boldsymbol{\beta}_n \in \mathbb{R}^M$ for each $\mathbf{x}_n$ by solving the eigenvalue problem of $\mathbf{K}'$, and obtain the new data cloud: $F(X) = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N\}$.
3: Draw a set of L random unit vectors from the unit hypersphere $\mathbb{S}^{M-1}$: $U = \{\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_L\} \subset \mathbb{S}^{M-1}$.
4: Compute the principal component $\boldsymbol{\beta} \in \mathbb{R}^M$ for $\mathbf{x}$.
5: Compute the projected data cloud $\mathbf{u}_l^T F(X)$ for $1 \le l \le L$.
6: **return** $-D_M(\boldsymbol{\beta} \mid F(X))$ in accordance with Eq. (14).

IV. Experimental Evaluation

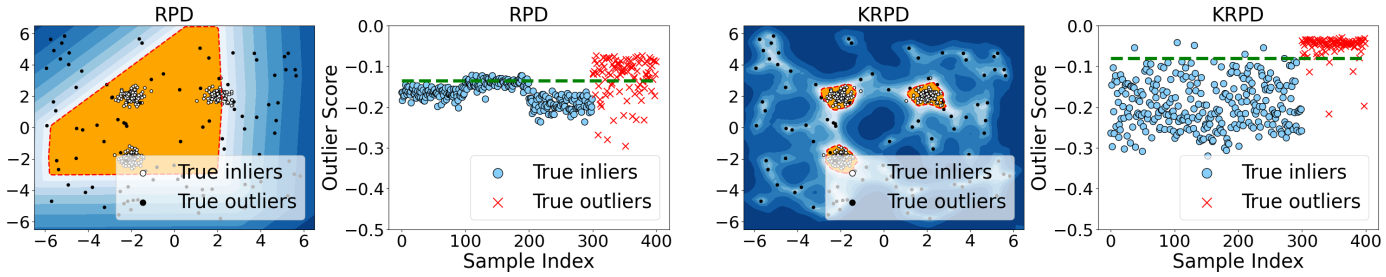
We conducted experiments on toy example data and benchmark datasets to evaluate the detection performance of proposed model. The following detection methods were compared with KRPD:

- k-Nearest Neighbors (kNN) [34]
- Local Outlier Factor (LOF) [26]
- Isolation Forest (IForest) [27]
- One-Class Support Vector Machine (OCSVM) [20]
- Kernel Principal Component Analysis (KPCA) [35]
- Random Projection Depth (RPD) [29]

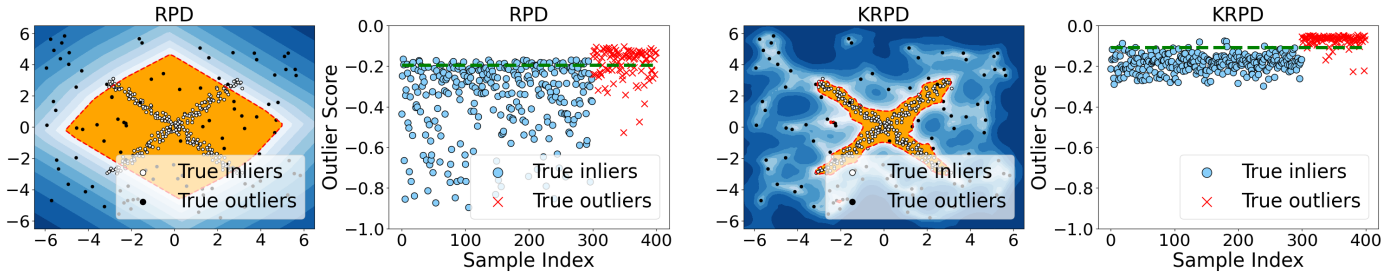
In this study, RBF kernel function $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma ||\mathbf{x} - \mathbf{y}||^2)$ was used for OCSVM, KPCA and KRPD.



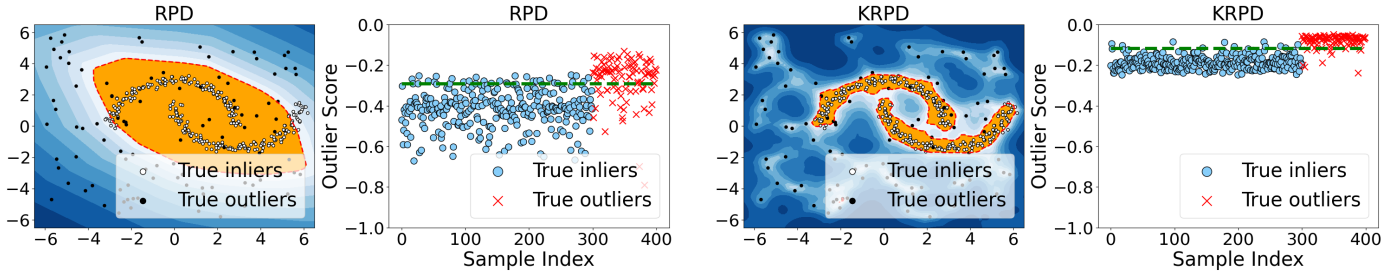
(a) Unimodal



(b) Multimodal



(c) Cross



(d) Moons

Fig. 2. Contour plots and outlier score comparison between RPD and KRPD. The threshold is indicated by the green dotted line in the outlier score plot. High outlier score corresponds to strong outlyingness.

A. Toy Example

Figure 2 shows contour plot comparisons between RPD and KRPD across toy datasets. We used the same datasets as in Figure 1. We fixed the kernel parameter γ and the dominant dimensionality M to 0.25 and 100, respectively. The threshold of the central region was determined according to the ratio of inliers to the total samples. Namely, the 75 ($= 100 \times 300/400$) percentile of the outlier scores. We also show the outlier scores on each toy data dataset. In these figures, the central region is colored orange and the threshold is represented as the green dotted line. First, we can see that the decision boundaries of both RPD and KRPD could enclose inliers on the unimodal. It is apparent that the corresponding scores are separated well, and consequently, we obtain few detection errors. Compared with RPD, we can see that the decision boundary of KRPD successfully encloses almost all inlier samples, and many outliers are outside the central region even when the data cloud exhibits multiple modalities and non-convexity. On the multimodal, scores of RPD from several inliers exceed the threshold. Such exceeded scores lead to an increase in false negative detection errors. Furthermore, many scores of RPD from outliers are lower than the threshold, which leads to an increase in false positive errors. In contrast, both false positive/negative detection errors decrease in KRPD, as we can

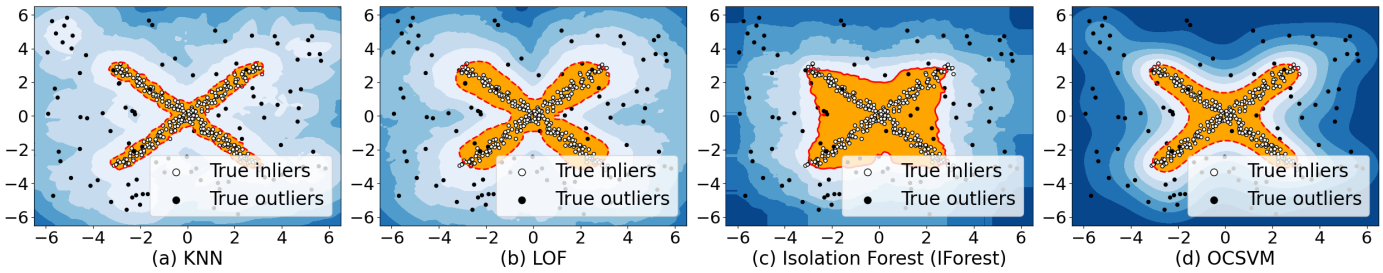


Fig. 3. Contour plots of detection methods on the Cross dataset.

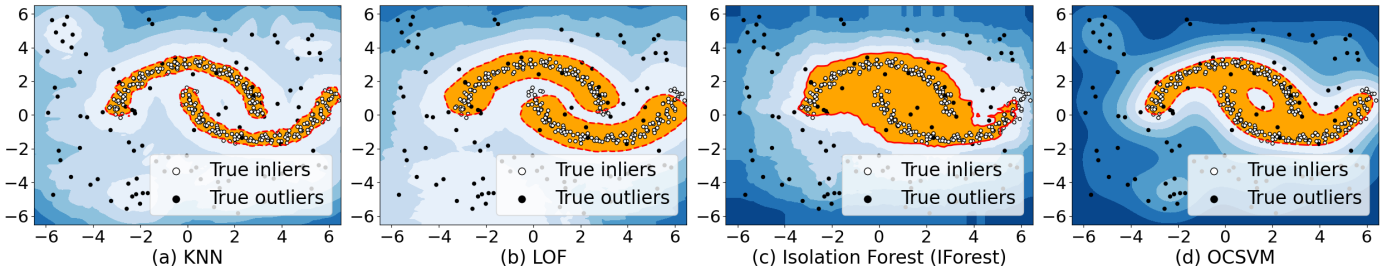


Fig. 4. Contour plots of detection methods on the Moon dataset.

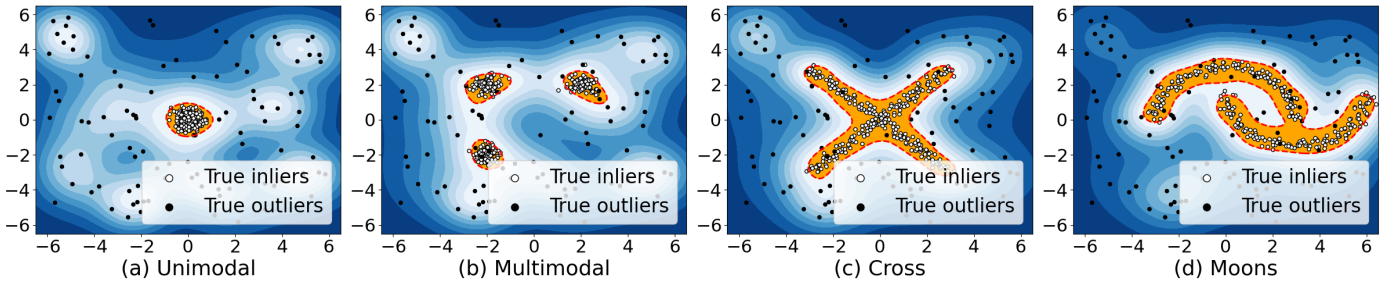


Fig. 5. Contour plots of KPCA-based detector on the toy datasets.

see from the right side of the figure. The trend is also generally true for the remaining Cross and Moons. From these results, we can expect that KRPD will demonstrate better detection performance than RPD.

Figure 3 and 4 show the contour plot comparisons on the Cross and Moons dataset, respectively. We can see that the decision boundary of IForest circumscribed the dataset and failed to capture the modalities, which led to an increase in false positive errors. kNN, LOF, and OCSVM achieved tighter boundaries to capture the modalities. Figure 5 shows the contour plots of KPCA on the toy datasets. We applied the same RBF kernel and the hyperparameter γ for both KPCA and KRPD. The dominant dimensionality M was set to 10 for KPCA. In the KPCA-based detection, we used the reconstruction error in the feature space as the outlier score [35]. From the figure, we can see that the boundary fitted the data cloud tightly, and even only KPCA obtained favorable feature spaces for outlier detection. Comparing with these figures and Figure 2, we consider that KRPD could achieve the competitive decision boundaries to other methods.

TABLE I
INFORMATION OF BENCHMARK DATASET [36]: DATASET NAME, NUMBER OF SAMPLES, NUMBER OF FEATURE DIMENSIONS, AND OUTLIER PERCENTAGES.

Dataset	#samples	#dims.	outliers (%)
Arrhythmia	452	274	14.6%
Cardio	1831	21	9.61%
Ionosphere	351	33	35.9%
Letter	1600	32	6.25%
MNIST	7603	100	9.21%
Musk	3062	166	3.17%
Optdigits	5216	64	2.88%
Pendigits	6870	16	2.27%
Satellite	6435	36	31.6%
Satimage-2	5803	36	1.22%
Vowels	1456	12	3.43%
Wbc	378	30	5.56%

B. Performance Evaluation on Benchmark Datasets

We conducted experiments on twelve standard benchmark datasets publicly available at ODDS library [36] for performance evaluations. The detailed information is listed in Table I.

TABLE II
ROC AUC SCORES OF OUTLIER DETECTION METHODS. AVERAGE OF 5 INDEPENDENT TRIALS IS SHOWN. THE BETTER METHOD IS INDICATED IN **BOLD**.

Dataset	kNN	LOF	IForest	OCSVM	KPCA	RPD	KRPD (ours)
Arrhythmia	0.783 ± 0.024	0.774 ± 0.028	**0.796 ± 0.040**	0.781 ± 0.022	0.768 ± 0.026	0.771 ± 0.021	0.791 ± 0.026
Cardio	0.841 ± 0.018	0.741 ± 0.017	0.928 ± 0.014	0.924 ± 0.012	0.852 ± 0.029	**0.936 ± 0.017**	0.926 ± 0.015
Ionosphere	0.924 ± 0.016	0.889 ± 0.030	0.839 ± 0.021	0.910 ± 0.017	0.918 ± 0.019	0.866 ± 0.014	**0.935 ± 0.012**
Letter	**0.897 ± 0.011**	0.865 ± 0.016	0.614 ± 0.029	0.792 ± 0.028	**0.897 ± 0.018**	0.617 ± 0.022	0.865 ± 0.018
MNIST	0.872 ± 0.004	0.779 ± 0.008	0.785 ± 0.008	0.859 ± 0.037	**0.875 ± 0.004**	0.844 ± 0.013	0.856 ± 0.006
Musk	0.972 ± 0.009	0.569 ± 0.028	**0.998 ± 0.000**	0.836 ± 0.085	0.991 ± 0.011	**0.998 ± 0.000**	0.997 ± 0.005
Optdigits	0.394 ± 0.028	0.605 ± 0.080	0.686 ± 0.043	0.528 ± 0.041	0.585 ± 0.053	0.552 ± 0.042	**0.755 ± 0.070**
Pendigits	0.842 ± 0.005	0.521 ± 0.025	0.950 ± 0.013	0.870 ± 0.038	**0.953 ± 0.010**	0.838 ± 0.029	0.920 ± 0.008
Satellite	0.717 ± 0.009	0.569 ± 0.007	0.714 ± 0.018	0.715 ± 0.039	0.716 ± 0.007	0.667 ± 0.004	**0.727 ± 0.005**
Satimage-2	0.992 ± 0.001	0.700 ± 0.057	0.989 ± 0.004	0.963 ± 0.033	0.993 ± 0.004	**0.996 ± 0.002**	0.994 ± 0.006
Vowels	**0.966 ± 0.009**	0.934 ± 0.009	0.757 ± 0.040	0.910 ± 0.049	0.940 ± 0.040	0.810 ± 0.049	0.905 ± 0.021
Wbc	0.812 ± 0.026	0.915 ± 0.021	**0.937 ± 0.028**	0.881 ± 0.062	0.884 ± 0.038	0.911 ± 0.032	0.906 ± 0.012
Average	0.834 ± 0.013	0.738 ± 0.027	0.832 ± 0.022	0.831 ± 0.039	0.864 ± 0.022	0.817 ± 0.020	**0.882 ± 0.017**

We randomly split each dataset into training and test data in a stratified fashion, specifying a seed; 60% for training and 40% for testing. The outlier percentage was kept the same on both training and test because of the stratified fashion. This split scheme intends to simulate data contamination and evaluate the robustness of the existence of outliers. Moreover, we performed a 5-fold stratified cross-validation on the training data for hyperparameter search. In each fold, 80% of training data were used for actual model training, and the rest of 20% were used for validation.

For both RPD and KRPD, the number of axes L for random projections was fixed to 1000. In order to optimize other hyperparameters, we utilized a framework using the Optuna package [37], which implements a Bayesian optimization. In this framework, we selected a strategy to maximize the ROC AUCs. For OCSVM, KPCA and KRPD, the kernel parameter γ was optimized between $[0.00001, 1.0]$. For both KPCA and KRPD, the number of principal components M was optimized between $[10, 500]$. We can obtain the label information as binary ones because the training data contains inliers and outliers. Hence we could also compute ROC AUCs on each fold of the training data, and the hyperparameters were determined semi-supervised. The Optuna optimized the hyperparameters according to the maximization criterion.

Table II lists ROC AUCs. We can see that KRPD outperformed RPD regarding ROC-AUCs on the nine datasets. Significantly, KRPD greatly improved the detection performance on Letter and Optdigits datasets. On the three datasets of Cardio, Musk, and Satimage-2, the performance of RPD was almost the same as KRPD. The averaged AUCs of KPCA was 0.864 ± 0.022, and it was comparable to other detection methods. Although KPCA outperformed KRPD on several datasets, KRPD outperformed KPCA on average, and especially, the improvement was significant on the Optdigits.

V. DISCUSSION

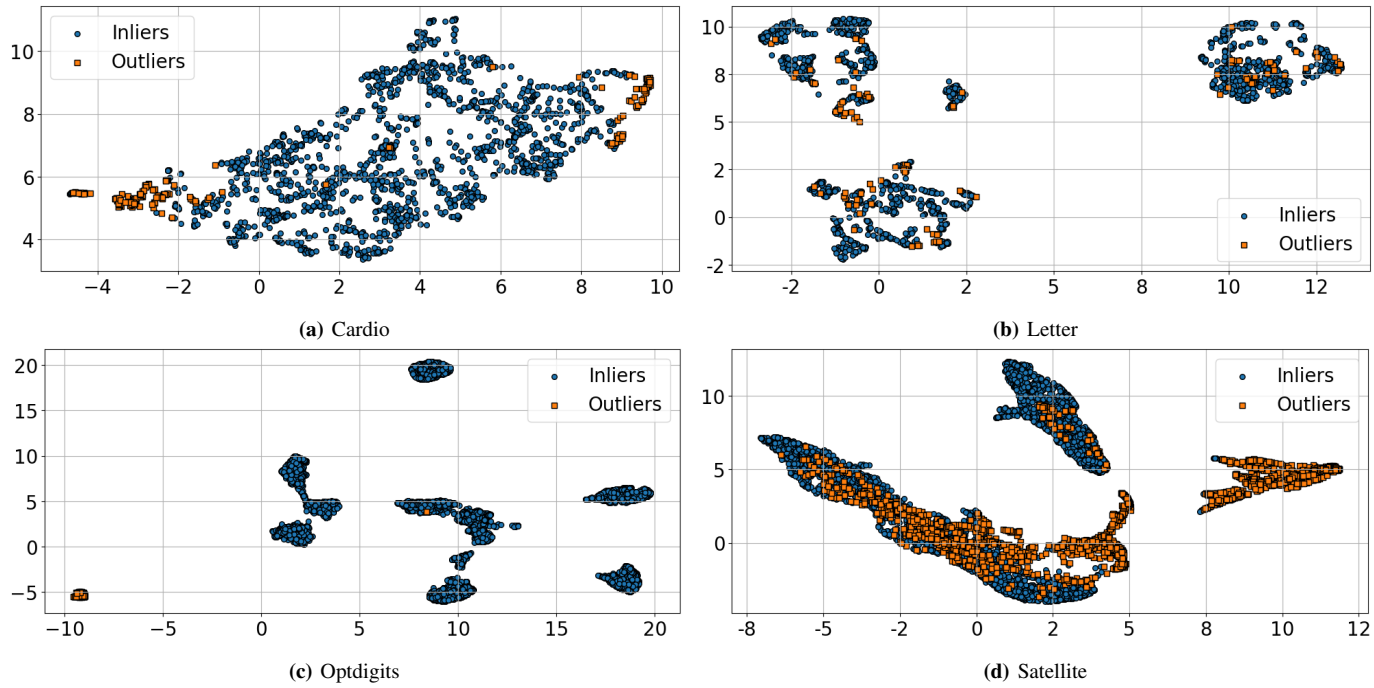
We have shown in the previous section that KRPD outperformed RPD and was comparable to other detection methods on average. These results demonstrate the effectiveness of KRPD in detecting outliers. However, KRPD did not deliver the best detection performance across all datasets. In this section, we will consider why we obtained variations across datasets in the detection performance of KRPD by using a visualization of datasets. We will also clarify the advantages against other methods and limitations of KRPD.

A. Variations of detection performance across datasets

Figure 6 shows visualizations of Cardio, Letter, Optdigits, and Satellite datasets by UMAP [38]. As shown in Figure 6a of Cardio, almost outliers locate around the edge of the inlier distribution, and the distribution looks roughly uni-modal. This observation explains why RPD outperformed KRPD slightly. Next, on Letter and Optdigits, KRPD outperformed RPD significantly. From the corresponding visualization of Letter as shown in Figure 6b, we can observe that roughly three clusters exist, corresponding to three capital letters in the English alphabet. Those clusters are further split into many small groups, and outliers are scattered. The detection with KRPD worked well even in such a challenging dataset. The Optdigits is a collection of instances of digits 0-9, and digit 0 is the outlier class. We can observe that its data cloud of Optdigits forms explicit clusters corresponding to each digit, and hence it shows multiple modalities. These observations also explain the superior performance of KRPD to RPD and other methods. As shown in Figure 6d of Satellite, the data cloud of outliers overlaps that of inliers, partly explaining the decrease of AUCs for KRPD. Namely, this decrease explains that KRPD is sensitive to the existence of outliers, that is, contamination in the training data, leading to an increase in the false alarm rate.

B. Detection performance without KPCA

As described in Section III, KPCA is applied to avoid reducing the efficiency and accuracy of depth computation. Hence, a natural question arises: How much will the performance degrade without KPCA? We conducted additional experiments to investigate the impact of using or not using KPCA on detection performance. We selected a kernel approximation method with random Fourier features (RFF) [24]. The RFF obtains a high dimensional vector for each $\mathbf{x}_n$ in the data



(a) Cardio **(b)** Letter

(c) Optdigits **(d)** Satellite

Fig. 6. Visualization of datasets by UMAP.

TABLE III
ROC AUC SCORES OF KRPD W/O AND W/ KPCA. AVERAGE OF 5 INDEPENDENT TRIALS IS SHOWN. THE BETTER METHOD IS INDICATED IN **BOLD**.

Dataset	w/o KPCA (RFF)	w/ KPCA
Arrhythmia	0.768 ± 0.033	**0.791 ± 0.026**
Cardio	**0.943 ± 0.015**	0.926 ± 0.015
Ionosphere	0.855 ± 0.039	**0.935 ± 0.012**
Letter	0.618 ± 0.018	**0.865 ± 0.018**
MNIST	0.820 ± 0.033	**0.856 ± 0.006**
Musk	**0.998 ± 0.002**	0.997 ± 0.005
Optdigits	0.543 ± 0.043	**0.755 ± 0.070**
Pendigits	0.896 ± 0.020	**0.920 ± 0.008**
Satellite	0.689 ± 0.020	**0.727 ± 0.005**
Satimage-2	**0.996 ± 0.003**	0.994 ± 0.006
Vowels	0.787 ± 0.035	**0.905 ± 0.021**
Wbc	0.896 ± 0.030	**0.906 ± 0.012**
Average	0.817 ± 0.025	**0.882 ± 0.017**

cloud X, and hence we can compute the projection depth. The number of Monte Carlo samples for RFF, corresponding to the dimensionality of the feature vector, was optimized between $[10, 500]$ with the Optuna package. The kernel parameter was also optimized between $[0.00001, 1.0]$.

Table III lists the ROC AUCs of KRPD without and with KPCA. We can see that the average performance significantly decreased without KPCA. We consider that this performance degradation was due to the curse of dimensionality because RFF involves computations of the inner product in a high dimensional space to approximate the kernel function. As described in Section III, almost all pairs of vectors on a unit hypersphere will become orthogonal. This fact significantly affects the computation of projection depth and partly explains the performance degradation without KPCA.

C. Limitations and future work

Since KRPD inherits advantages from KPCA, it also inherits its limitations. First, careful hyperparameter tuning is often required. In the experimental evaluation, the KRPD hyperparameters could be determined semi-supervised since we had labeled samples available. KRPD will be more effective in semi-supervised situations. Second, KPCA is not necessarily robust to outliers in the training data. Although KPCA has an application of denoising [39], its formulation does not assume the existence of outliers, and the outliers will distort the direction of eigenvectors. Authors in the previous study [40] have proposed several regularization techniques to obtain eigenvectors robustly against outliers. In the future, we will consider an extension of KPCA for RPD to cope with outliers and robust extraction of eigenvectors. Third, the computational complexity is still $O(N^3)$, where N is the number of training samples. Approximate kernel PCA [41] will be effective when N itself becomes a large number, e.g., $N > 10^4$. Kernel PCA with an online learning algorithm [42] will also reduce computational complexity. We will investigate correlations among the number of training samples, the dimensionality of the principal subspace, and its detection performance. Lastly, we have not given statistical support to explain the improved detection performance from the viewpoint of statistical depth. We are interested in how the desirable properties of statistical depth [3] behave when changing the dimensionality M and the number of training samples N.

VI. Conclusion

In this paper, we proposed KRPD, an extension of RPD for outlier detection. The original data cloud is mapped to RKHS by a reproducing kernel function, and we can compute RPD in the RKHS. With the help of KPCA, the decision boundary of KRPD fits the data cloud tightly even when the cloud shows complicated multiple modalities and non-convexity. We evaluated outlier detection performance for RPD, KRPD, and other popular methods on 12 benchmark datasets. The experimental results demonstrated that KRPD outperformed RPD and was comparable to other methods. We also discussed the detection performance between KPCA and KRPD and confirmed that KRPD was more effective than KPCA. We plan to extend KRPD further to utilize robust KPCA.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP21K17771.

References

[1] V. Chandola and V. Kumar, "Anomaly Detection : A Survey," *ACM Computing Surveys*, vol. 41, Jan. 2009.

[2] M. R. Smith and T. Martinez, "Improving classification accuracy by identifying and removing instances that should be misclassified," in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 2690–2697.

[3] Robert Serfling and Yijun Zuo, "General notions of statistical depth function," *The Annals of Statistics*, vol. 28, no. 2, pp. 461–482, 2000.

[4] Z. Chen and D. E. Tyler, "On the finite sample breakdown points of redescending m-estimates of location," *Statistics & Probability Letters*, vol. 69, no. 3, pp. 233–242, 2004.

[5] S. Chenouri and C. G. Small, "A nonparametric multivariate multisample test based on data depth," *Electronic Journal of Statistics*, vol. 6, pp. 760–782, 2012.

[6] R. Hoberg, "Cluster analysis based on data depth," in *Data Analysis, Classification, and Related Methods*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 17–22.

[7] A. K. Ghosh and P. Chaudhuri, "On data depth and distribution-free discriminant analysis using separating surfaces," *Bernoulli*, vol. 11, no. 1, pp. 1–27, 2005.

[8] M. Hallin, D. Paindaveine, and M. Šiman, "Multivariate quantiles and multiple-output regression quantiles: From l1 optimization to halfspace depth," *The Annals of Statistics*, vol. 38, no. 2, pp. 635–669, 2010.

[9] R. Fraiman, R. Y. Liu, and J. Meloche, "Multivariate density estimation by probing depth," *Lecture Notes-Monograph Series*, vol. 31, pp. 415–430, 1997.

[10] S. López-Pintado and J. Romo, "On the concept of depth for functional data," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 718–734, 2009.

[11] T. J. W., "Mathematics and the picturing of data," *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, vol. 2, pp. 523–531, 1975.

[12] H. Oja, "Descriptive statistics for multivariate distributions," *Statistics & Probability Letters*, vol. 1, no. 6, pp. 327–332, 1983.

[13] G. Koshevoy and K. Mosler, "Zonoid trimming for multivariate distributions," *The Annals of Statistics*, vol. 25, no. 5, pp. 1998–2017, 1997.

[14] Y. Vardi and C.-H. Zhang, "The multivariate L_1-median and associated data depth," *Proceedings of the National Academy of Sciences*, vol. 97, no. 4, pp. 1423–1426, 2000.

[15] R. Y. Liu, "Data depth and multivariate rank tests," *L1-statistical analysis and related methods*, pp. 279–294, 1992.

[16] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[17] Y. Chen, X. Dang, H. Peng, and H. Bart, "Outlier Detection with the Kernelized Spatial Depth Function," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 288–305, 2009.

[18] Y. Hu, Y. Wang, Y. Wu, Q. Li, and C. Hou, "Generalized mahalanobis depth in the reproducing kernel hilbert space," *Statistical Papers*, vol. 52, pp. 511–522, 2011.

[19] A. Tamamori, "Unsupervised outlier detection based on random projection outlyingness with local score weighting," *IEICE Transactions on Information and Systems*, vol. E106.D, no. 7, 2023, in press.

[20] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[21] L. Ruff *et al.*, "Deep one-class classification," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 4393–4402.

[22] B. Zong *et al.*, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.

[23] M. Bauw, S. Velasco-Forero, J. Angulo, C. Adnet, and O. Airiau, "Deep random projection outlyingness for unsupervised anomaly detection," *CoRR*, vol. abs/2106.15307, 2021. arXiv: 2106.15307.

[24] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, vol. 20, Curran Associates, Inc., 2007.

[25] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems*, vol. 13, MIT Press, 2000.

[26] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proc. of ICMD*, Dallas, Texas, USA, 2000, pp. 93–104.

[27] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[28] O. Barkan and A. Averbuch, "Robust mixture models for anomaly detection," in *Proc. of MLSP*, 2016, pp. 1–6.

[29] Y. Zuo, "Projection-based depth functions and associated medians," *The Annals of Statistics*, vol. 31, no. 5, pp. 1460–1490, 2003.

[30] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.

[31] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 3, pp. 427–444, 2005.

[32] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, 2019.

[33] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[34] F. Angiulli, T. Elomaa, H. Mannila, and H. Toivonen, "Fast outlier detection in high dimensional spaces," in *Proc. of PKDD*, Springer Berlin Heidelberg, 2002, pp. 15–27.

[35] H. Hoffmann, "Kernel pca for novelty detection," *Pattern Recognition*, vol. 40, no. 3, pp. 863–874, 2007.

[36] S. Rayana, *Odds library*, http://odds.cs.stonybrook.edu/, 2016.

[37] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. KDD*, Anchorage, AK, USA, 2019, pp. 2623–2631.

[38] L. McInnes and J. Healy, "UMAP: Uniform manifold approximation and projection for dimension reduction," *ArXiv e-prints*, Feb. 2018. arXiv: 1802.03426.

[39] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel pca and de-noising in feature spaces," in *Advances in Neural Information Processing Systems*, vol. 11, MIT Press, 1998.

[40] D. Wang and T. Tanaka, "Robust kernel principal component analysis with *l*2,1-regularized loss minimization," *IEEE Access*, vol. 8, pp. 81 864–81 875, 2020.

[41] B. K. Sriperumbudur and N. Sterge, "Approximate kernel pca: Computational versus statistical trade-off," *The Annals of Statistics*, vol. 50, no. 5, pp. 2713–2736, 2022.

[42] P. Honeine, "Online kernel principal component analysis: A reduced-order model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1814–1826, 2012.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Soft-Sensor Construction Method Based on Adaptive Modeling and Transfer Learning for Manufacturing Process Including Maintenance Periods

Kaito Katayama* and Koichi Fujiwara * and Kazuki Yamamoto †
* Nagoya University, Japan
E-mail: fujiwara.koichi@hps.material.nagoya-u.ac.jp
† AGC Inc., Japan

***Abstract*—Various soft-sensor construction methods based on transfer learning have been proposed to quickly adapt soft-sensors to changes in process characteristics caused by process maintenance. However, most of the conventional methods adopt transfer learning immediately after the maintenance. Since process characteristics may also change due to other than process maintenance, it is necessary for the conventional methods to switch to a soft-sensor that is applicable to the current process characteristics at the appropriate time; however, it is difficult to properly determine the switching timings of soft-sensors. This study proposes a new adaptive soft-sensor construction method called Latest Sample Targeting Frustratingly Easy Domain Adaptation (LST-FEDA) based on adaptive modeling and FEDA which is a kind of transfer learning method. In the proposed LST-FEDA, the application range of FEDA is updated and rebuilds a soft-sensor when a new query sample is measured. This approach enables soft-sensors to automatically adapt to changes in process characteristics without intentional soft-sensor switching.**

In order to demonstrate the effectiveness of the proposed LST-FEDA, it was applied to actual operation data of a distillation process including maintenance periods. The results demonstrated that LST-FEDA improved the RMSE by 19% on average in comparison with the conventional methods.

I. Introduction

In manufacturing processes, product quality or other important variables are often measured through offline analysis at regular intervals. Because offline analysis is burdensome and time-consuming, there is a risk of producing off-spec products between analysis intervals. Quality management using soft-sensors that predict product quality from process variables measured on-line has been widely used. However, there is a possibility that the prediction performance of soft-sensors deteriorates when the process characteristics change due to process maintenance or aging. Thus, it is essential to adapt soft-sensors to changes in process characteristics for stable and efficient process operation.

To deal with such changes and to update soft-sensors automatically, adaptive modeling methods, such as Moving Window (MW) modeling [1] and Just-in-Time (JIT) modeling [2], have been developed. In addition, soft-sensor construction methods based on transfer learning have been developed [3], [4], [5]. These methods enable soft-sensors to adapt quickly to changes in process characteristics due to process maintenance or aging. In transfer learning, the domain that transfers information is called the source domain (D^S) and the domain that receives information from the source domain is called the target domain (D^T) data.

Ye *et al.* proposed a method for training Online Sequential Extreme Learning Machines with Kernels (OS-ELMK) to utilize knowledge from old data for future forecasts using transition learning. The method was shown to provide highly accurate forecasts for various types of time series forecasting problems [4]. Lie *et al.* proposed a soft sensing method which transfers process information between different operating grades to construct soft-sensors for multi-grade chemical processes. Their proposed method, Domain Adaptation Extreme Learning Machines (DAELM), can construct high-performance soft-sensors even with limited supervised data [5]. As a simpler method of applying transition learning, Yamada *et al.* proposed a soft-sensor construction method combining an adaptive modeling method and a transfer learning method called Frustratingly Easy Domain Adaptation (FEDA) [6] for chemical processes that produce products with grade switching [7]. They used the method for transfering the process data from various grades to the process data of the target grade, and predicted the product quality shortly after grade switching. Their proposed method achieved a high prediction performance of product quality even after grade switching.

Transfer learning enables accurate predictions in processes with significant changes in operating conditions. However, when transfer learning is applied to adaptive soft-sensors, the similarity between the target domain and source domain may become low when an excessive amount of target domain samples are stored, which may lead to improper information transference [7]. If improper information is transferred, negative transfer may occur and degrade the prediction performance compared with even when transfer learning is not used [8]. In addition, when negative transfer occurs, the continued application of transfer learning has the potential to permanently deteriorate the prediction performance of the soft-sensors. Although most of the conventional methods using transfer learning are required to switch soft-sensors when a certain number of target domain samples were stored, it is often difficult to determine the appropriate timing for switching accurately.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

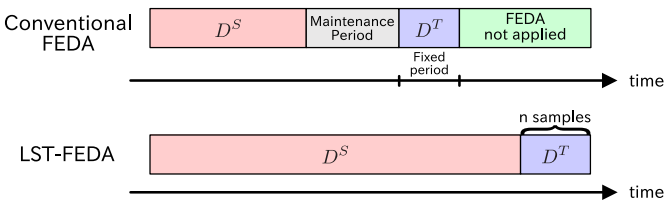


Fig. 1. Domain definitions of conventional FEDA and LST-FEDA

In this study, we propose a new adaptive soft-sensor construction method based on adaptive modeling and FEDA, called Latest Sample Targeting FEDA (LST-FEDA). In the proposed LST-FEDA, the specific number of samples in training data temporally close to a query sample are defined as the target domain. Whenever a new training sample is measured, the definition of the target domain is updated in a first-in-first-out manner and soft-sensors are rebuilt. In the proposed method, there is no need to switch soft-sensors to prevent negative transfer because it can keep the number of samples in the target domain constant.

We report application results of LST-FEDA to the actual operation data of a distillation process including process maintenance.

II. Methods

A. Frustratingly Easy Domain Adaptation

FEDA is a kind of transfer learning method that simply extends a feature space. When the source domain data $\boldsymbol{X}_S \in \mathbb{R}^{N_S \times M}$ and the target domain data $\boldsymbol{X}_T \in \mathbb{R}^{N_T \times M}$ are given, the new data $\boldsymbol{\Phi}(\boldsymbol{X}) \in \mathbb{R}^{(N_S+N_T)\times 3M}$ are constructed as follows:

$$\boldsymbol{\Phi}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{X}_S & \boldsymbol{X}_S & \boldsymbol{0} \\ \boldsymbol{X}_T & \boldsymbol{0} & \boldsymbol{X}_T \end{bmatrix} \tag{1}$$

Where N_S and N_T are the numbers of samples in the source and target domain, respectively; and M is the number of variables in $\boldsymbol{X}_S$ and $\boldsymbol{X}_T$ [6].

The common information between the source domain and the target domain can be modeled from $\boldsymbol{X}_S$ and $\boldsymbol{X}_T$. In addition, specific information in the source domain is modeled from $\boldsymbol{X}_S$ and $\boldsymbol{0}$, and specific information in the target domain is from $\boldsymbol{0}$ and $\boldsymbol{X}_T$ [7]. Thus, FEDA can construct models considering common and specific information of each domain by using (1).

B. Latest Sample Targeting FEDA

In this study, we propose a new adaptive soft-sensor construction method called Latest Sample Targeting FEDA (LST-FEDA), which defines only the latest n samples as the target domain, as shown in Fig. 1. The proposed method can prevent negative transfer caused by excessive sample storage in the target domain data and enables the continuous application of FEDA.

Among the input data $\boldsymbol{X} \in \mathbb{R}^{N\times M}$, the n temporally closest samples to the query sample $\boldsymbol{x}_q \in \mathbb{R}^M$ are defined as the target domain data $\boldsymbol{X}_{new} \in \mathbb{R}^{n\times M}$ and the past data before $\boldsymbol{X}_{new}$ are defined as the source domain data $\boldsymbol{X}_{past} \in \mathbb{R}^{(N-n)\times M}$. $\boldsymbol{X}$ and $\boldsymbol{x}_q$ are extended based on (1), and soft-sensors are constructed with the extended data $\boldsymbol{\Phi}(\boldsymbol{X}) \in \mathbb{R}^{N\times 3M}$ and $\boldsymbol{\Phi}(\boldsymbol{x}_q) \in \mathbb{R}^{3M}$ as follows:

$$\boldsymbol{\Phi}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{X}_{past} & \boldsymbol{X}_{past} & \boldsymbol{0} \\ \boldsymbol{X}_{new} & \boldsymbol{0} & \boldsymbol{X}_{new} \end{bmatrix} \tag{2}$$

$$\boldsymbol{\Phi}(\boldsymbol{x}_q) = \begin{bmatrix} \boldsymbol{x}_q^\top & \boldsymbol{0}^\top & \boldsymbol{x}_q^\top \end{bmatrix}^\top \tag{3}$$

When a new input sample is measured, the new sample is added to $\boldsymbol{X}_{new}$ and the oldest sample in $\boldsymbol{X}_{new}$ is moved to $\boldsymbol{X}_{past}$ to keep the number of samples in $\boldsymbol{X}_{new}$ constant in a first-in-first-out manner.

III. Case Study

A. Distillation process

In this study, we analyzed the actual operation data of the distillation process of a fluorinated telomer intermediate including maintenance periods, provided by AGC Inc, in Japan. The operation data consist of 21 process variables listed in Table I, and the objective quality variable of the telomer intermediate was analyzed using gas chromatography. Fig. 2 illustrates a flow diagram of the distillation process, and the numbers in Fig. 2 indicate the corresponding measurement sensors of the process variables in Table I. The process variables were measured every ten minutes, and the quality variable was analyzed approximately two or three times a day. The process is shut down due to maintenance from February to March every year. Thus, the data during the maintenance periods are lost.

In this study, we used the operation data measured from May 2020 to October 2022. The training data were from 2020 and 2021, and the test data were from 2022. The dataset for constructing soft-sensors used process variables 10, 20, and 30 minutes before the sampling time of the quality variable in addition to the process variables at the same time as the quality variable measurement in order to consider process dynamics. The purity of the telomer intermediate was the objective quality variable predicted by soft-sensors.

B. Comparison methods

To verify the effectiveness of the proposed LST-FEDA, we compared it with the following two methods:

1) Not Applied: Construct adaptive soft-sensors without FEDA.
2) Conventional FEDA: Constructs adaptive soft-sensors by applying FEDA over a period of one month after process maintenance. The process data before maintenance are the source domain and the data after maintenance are the target domain. In other periods, soft-sensors were constructed without FEDA.

Note that the period of application of FEDA in Conventional FEDA was set to the period when the influence of maintenance

TABLE I
PROCESS VARIABLES

No.	Variable Name	No.	Variable Name
1	TW-Feed Flow Rate [L/h]	12	Bottom Temperature of Raw Material Recovery Tower [°C]
2	TW-Steam Flow Rate [kg/h]	13	Bottom Temperature of Impurity Cut Tower [°C]
3	Reflux Flow Rate [L/h]	14	Bottom Temperature of Product Separation Tower [°C]
4	TW-Bottom Condensate Flow Rate [L/h]	15	Top Temperature [°C]
5	TW-Sidecut Distillate Flow Rate [L/h]	16	Middle Temperature 1 [°C]
6	Top Distillate Liquid Level [L]	17	Middle Temperature 2 [°C]
7	Bottom Pressure of Raw Material Recovery Tower [MPa]	18	Middle Temperature 3 [°C]
8	TW-Top Pressure [MPa]	19	Middle Temperature 4 [°C]
9	TW-Bottom Pressure [MPa]	20	Bottom Temperature [°C]
10	TW-Ejector Pressure [MPa]	21	Reflux Temperature [°C]
11	TW-Reboiler Pressure [MPa]		

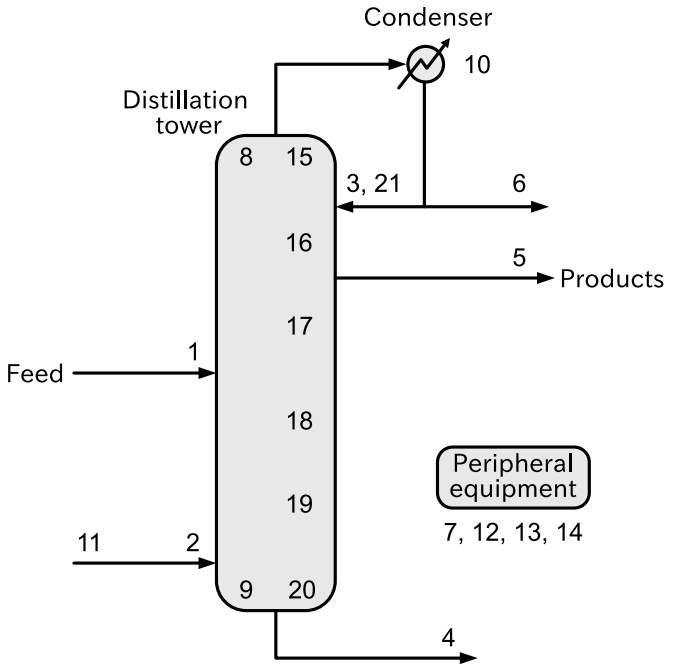


Fig. 2. Process flow diagram

was considered to be strongly present based on the opinions of the employees of AGC Inc.

C. Results

In the practical stage, the soft-sensors were reconstructed each objective variable measurement so that it adapts to the current process state. To evaluate the performance of the soft-sensors in the same situation, the soft-sensors of this study were constructed only for next sample prediction. After prediction, the sample was added to the training data and utilized for the next soft-sensor construction. we continued to update the soft-sensors in this way and calculated the prediction performance of each soft-sensor based on the prediction results for the samples considered as test data. In this study, we tried two types of soft-sensor construction methods: MWPLS and LWPLS. MWPLS is combined partial least squares (PLS) and MW modeling [9], and LWPLS is an extension of PLS based on locally weighted (LW) modeling, which is a type of JIT modeling [10]. We applied them to the datasets without and with FEDA, or with the proposed LST-FEDA.

Correlation coefficient (R) and root mean square error (RMSE) were used as evaluation metrics of their prediction performance. The number of target domain samples in LST-FEDA and hyper-parameters of each soft-sensor construction method were determined to minimize RMSE with three-fold time-series cross-validation [11]. In this study, the candidates of the number of target domain samples were $n = \{5, 15, 30, 60\}$. These numbers roughly correspond to a few days, a week, two weeks, and a month worth of data. We used cross-validation, and for both MWPLS and LWPLS, the number of samples was set to 5. The prediction performances during one month before and after process maintenance (January and April 2022) and a steady operation period (May to October 2022) and an entire period of each soft-sensor are summarized in Table II. The best performance for each evaluation metric is highlighted in bold.

In both MWPLS and LWPLS, the soft-sensors with LST-FEDA achieved higher prediction performance than the soft-sensors without and with the conventional FEDA. In comparison with the soft-sensor using the conventional FEDA, the soft-sensor using LST-FEDA and LWPLS improved R and RMSE by 46% and 13%, respectively; and the soft-sensor using LST-FEDA and MWPLS improved R and RMSE by 55% and 25%, respectively. The prediction results of soft-sensors constructed using LWPLS are shown in Figs. 3 - 5. The left figures show time-series trends, of which the horizontal axis represents the sampling time and the vertical axis represents the purity of the telomer intermediate. The red and blue lines represent the measured and predicted values, respectively. The right figures illustrate scatter plots, of which the horizontal and vertical axes

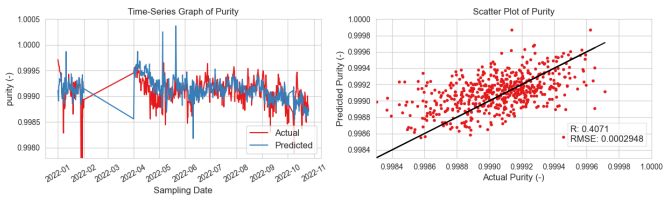


Fig. 3. Prediction result of soft-sensor using LWPLS without FEDA

TABLE II
PREDICTION PERFORMANCE OF SOFT-SENSORS

Soft-Sensor Construction Method	FEDA Application Method	One month before and after maintainance		Steady operation period		Entire period	
		R	RMSE ($\times 10^{-4}$)	R	RMSE ($\times 10^{-4}$)	R	RMSE ($\times 10^{-4}$)
MWPLS	Not Applied	0.5424	3.250	0.3669	3.556	0.3993	3.479
	Conventional FEDA	0.4459	3.451	0.3669	3.556	0.3772	3.529
	LST-FRDA	**0.5920**	**3.066**	0.5827	2.501	0.5829	2.660
LWPLS	Not Applied	0.3979	3.851	0.4294	2.554	0.4071	2.948
	Conventional FEDA	0.3595	3.660	0.4294	2.554	0.4065	2.884
	LST-FEDA	0.5734	3.084	**0.6048**	**2.266**	**0.5916**	**2.505**

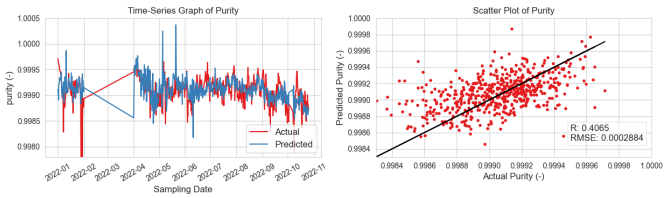


Fig. 4. Prediction result of soft-sensor with FEDA and LWPLS

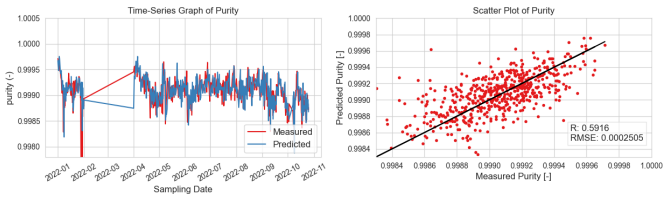


Fig. 5. Prediction result of soft-sensor with LST-FEDA and LWPLS

represent the measured and predicted values. These figures clearly show the usefulness of the proposed methods.

The performance improvement in one month before and after process maintenance suggests that the proposed LST-FEDA quickly and accurately adapted to changes in process characteristics caused by process maintenance. Moreover, the improvement of the prediction performance in the steady operation period suggests that LST-FEDA was able to deal with changes in process characteristics aging, or other factors because FEDA was not used in other methods during this period.

IV. Conclusion

In this study, we proposed a new adaptive soft-sensor construction method, LST-FEDA, which combines adaptive modeling and transfer learning. LST-FEDA continuously updates the definition of the target domain of transfer learning and constructs soft-sensors, which enables stable predictions without soft-sensor switching for preventing negative transfer.

We applied the soft-sensors based on the proposed LST-FEDA to the distillation process data including the maintenance periods and evaluated their prediction performance of the purity in the telomer intermediate. We confirmed that the proposed LST-FEDA improved the prediction performance during the steady operation period as well as the period before and after process maintenance in comparison with the conventional methods. It is indicated that the proposed method can construct soft-sensors applicable to both changes in process characteristics by process maintenance and changes in process characteristics by aging over time.

In future works, we will apply the proposed LST-FEDA to other real-process operational datasets to evaluate its performance. In addition, we will investigate whether dynamically adjusting the number of the target domain samples for each query sample, instead of using a fixed number to improve prediction performance.

References

[1] Burak Alakent. Soft sensor design using transductive moving window learner. *Computers & Chemical Engineering*, 140:106941, 2020.

[2] Koichi Fujiwara, Manabu Kano, Shinji Hasebe, and Akitoshi Takinami. Soft-sensor development using correlation-based just-in-time modeling. *AIChE Journal*, 55(7):1754–1765, 2009.

[3] Shohta Kobayashi, Masashi Miyakawa, Susumu Takemasa, Naoki Takahashi, Yoshio Watanabe, Toshiaki Satoh, and Manabu Kano. Transfer learning for quality prediction in a chemical toner manufacturing process. In Yoshiyuki Yamashita and Manabu Kano, editors, *14th International Symposium on Process Systems Engineering*, volume 49 of *Computer Aided Chemical Engineering*, pages 1663–1668. Elsevier, 2022.

[4] Rui Ye and Qun Dai. A novel transfer learning framework for time series forecasting. *Knowledge-Based Systems*, 156:74–99, 2018.

[5] Yi Liu, Chao Yang, Kaixin Liu, Bocheng Chen, and Yuan Yao. Domain adaptation transfer learning soft sensor for product quality prediction. *Chemometrics and Intelligent Laboratory Systems*, 192:103813, 2019.

[6] Hal Daumé III. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815, 2009.

[7] Nobuhito Yamada and Hiromasa Kaneko. Adaptive soft sensor based on transfer learning and ensemble learning for multiple process states. *Analytical Science Advances*, 3(5-6):205–211, 2022.

[8] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[9] Agnar Höskuldsson. Pls regression methods. *Journal of Chemometrics*, 2(3):211–228, 1988.

[10] Weiwei Bai, Junsheng Ren, and Tieshan Li. Modified genetic optimization-based locally weighted learning identification modeling of ship maneuvering with full scale trial. *Future Generation Computer Systems*, 93:1036–1045, 2019.

[11] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40 – 79, 2010.

Detecting Wire Bonding Defects in Point Clouds on Self-Generated Dataset

Shang Li Yuen*, Phooi Yee Lau*, Chin Wee Wong*, Muhammad Hafiz Samsuri, Zarina Hussin, Nur Afiqah Kamarudin, Muhammad Syukri Mohd Talib and Hock Woon Hon
Advanced Intelligence Lab, MIMOS Berhad, Kuala Lumpur, Malaysia
E-mail: { sl.yuen, py.lau, cw.wong, hafiz.samsuri, zarina.hussin, afiqah.kamarudin, syukri.talib, hockwoon.hon}@mimos.my

***Abstract* — Wire bonding involves bonding of thin wires precisely, typically made of gold or aluminum, often comes with intricate connections. Inspection systems often have to be placed in manufacturing lines to check the bonding process as defects could be introduced, thus decreasing the yield of IC chips. A lack of proper inspection system to ensure the overall quality and reliability of these bonds could cause premature device failures and costly recalls. The current 2D inspection systems have limitations in analyzing wire bonding 3D defects, as it may not able to identify defects that occur in the vertical dimension or are hidden from the top-down view. In this paper, we present a method to detect wire bonding defects in point cloud on our self-generated wire bonding mock-up datasets, named WireBondMockUp. In this work, we investigate the performance of different widely used 3D object detection methods to localize wire defect area in the WireBondMockUp, using different network architectures such as 3DETR, VoteNet, MLCVNet, and H3DNet. Our WireBondMockUp were annotated using LabelCloud and being randomly split into training and evaluation tailored to imitate wire bonding defects. Experimental results show that 3DETR, with learning rate 0.001, 1440 epoch and batch size of 4 could achieved 92.42 % in F1-score detection accuracy. The inference time was 0.08351 seconds. Experimental results also showed that the 3DETR could detect wire bonding defects in point cloud, highlighting the strengths in improving yield for IC manufacturing.**

I. INTRODUCTION

The demand for smaller size electronics products has kept engineer to shrink the sizes of electronic circuit elements from micrometer to nanometer, to ensure the capacity of integrated circuits (IC) can meet the emerging electronics products functions[reference]. Manufacturing ICs has thus become one of the fastest-expanding technologies due to the demand for higher end electronic products. As IC sizes decrease, wire bonding has become a popular choice for making electrical connections within the IC packaging. However, the bonding process can be complex, and there exist many precise controls of various parameters in order to achieve reliable and high-quality wire bonds. As such, placing a reliable inspection system to visually inspect and detect defects accurately is crucial especially as wire bond size decreases. However, the current inspection system focuses on 2D images, detecting defects such as missing bonds, short circuits, wire loops, or lifted bonds[reference]. On the other hand, there is a strong demand to detect defects occurring in the vertical plane, such as tilted bonds, lifted bonds, or bond deformation, which may not be apparent from a top-down image captured.

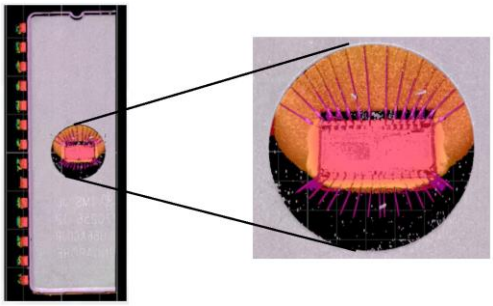


Fig. 1 Wire bonding

These days, the 3D object detection methods have shown remarkable performance on 3D public benchmark datasets [kitti, nuscene, waymo], but there is a lack of work reported for specific application domains. There exist limited reported implementation scenarios for specific industrial defects domain such as 3D defects detection, as it is relatively new and rapidly evolving. Nonetheless, as 3D imaging techniques could provide a detailed structure of the wires bonding structures, which permits the identification and analysis of defects, deploying 3D defect detection in manufacturing lines could prove to become a necessity, when manufacturing IC, to increase production yield.

In recent years, deep learning has gained significant popularity in the field of 3D object detection. Deep learning techniques have shown remarkable progress especially in detecting and localizing objects accurately in 3D space from different type of sensor inputs, such as LiDAR (Light Detection and Ranging) point clouds or RGB-D images [1-4]. In 2019, Qi et. al proposed VoteNet to detect objects in point clouds based

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

on a Hough voting strategy [1]. In 2020, Xie et. al extends Votenet with additional three-level useful contexts into the voting and classifying stages to encode contextual information [2]. In 2020, Zhang et. al proposed a 3D object detection method by using the hybrid predicted geometric primitives and then fitting the detected objects to these primitives and their associated features [3]. In 2021, Misra et. al proposed a transformer-based detection method with versatile attention-based backbone which could preserve the vanilla transformer blocks to reach comparable performance with CNN-based detectors [4]. However, these works fail to provide details on their performance for self-generated dataset, especially applying them in industrial applications, such as detecting wire bonding defects, which involves achieving a balance between inference speed and limited training datasets. Moreover, conventional 3D vision inspection systems often rely on height measurements to detect failures in wire bonds, and the complex nature of wire bond structures makes it challenging to reliably detect defects for each individual wire.

Based on the above, in this paper, we are proposing generating a mock-up dataset for imitating 3D wire bond structures and present a comprehensive analysis, using the 3D object detection methods discussed above, on self-generated wire bonding mock-up datasets. Our aim is to provide a more comprehensive and accurate assessment of wire bond quality, ensuring that even intricate structures can be thoroughly examined for defects. Our work emphasizes the practicability of these methods and their strength based on a common implementation scenario, i.e., applying these methods directly on our own datasets, named WireBondMockUp. We believe that evaluating the performance of these methods on custom datasets is crucial to understand their real-world applicability and to bridge the gap between theoretical advancements and practical applications.

The main objectives of this paper are:

- To propose a strategy to study and create a wire bonding point cloud dataset.
- To propose a unified strategy to detect wire bonding defects.

The rest of this paper is as follows. The materials are presented in Section II. Section III describes the proposed methodology. Section IV discusses the experiment results and lastly Section V concludes the paper with future works.

II. MATERIALS

In recent years, significant advancements have been made in the techniques used to collect and analyze 3D datasets, catering to diverse requirements. The common experimental set-up could include a camera associated to a recording in a specified grid for selectivity trials – see Fig. 2. In creating our mock-up dataset, we first set-up a wire bonding scenario, i.e. with/out failed bonding – see Fig. 3.



Fig. 2 Acquisition materials (left) Intel RealSense™ D435 (right) wire bonding mock-sup set-up

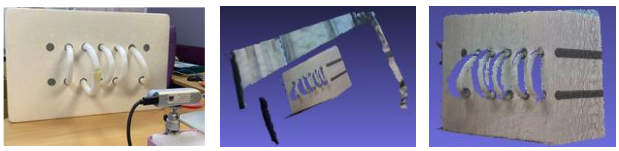


Fig. 3 Acquisition scenario (left) set-up (middle) uncropped (right) cropped

Then, an Intel RealSense™ D435 camera is mounted on a table top to record the scenario presented in the mock-up wire bonding set-up. Our acquisition scenario, with the above-mentioned set-up, notably produced: 1) some part of the point cloud which is required to be cropped, corresponding to the not-interesting objects around the wire bonding areas; 2) thin structure of the wire bond which may not work well with some 3D object detection method, and 3) artefacts appearing in the wire bonding area due to the camera and acquisition angle – see Fig. 3 for uncropped/cropped view. Our sample targets include detecting all wire bond and their associated defects such as tilted bonds, lifted bonds, or bond deformation – see Fig. 4.

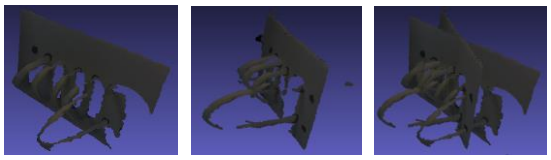


Fig. 4 Acquisition scenario (left) defect view1 (middle) defect view2 (right) comparison

The datasets include data collected from multiple viewpoints, ensuring a comprehensive coverage of the wire bonding defects. The resultant point cloud shows that in the experimental set-up, the RealSense camera is shifted to accommodate different

viewpoints highlight earlier. By offering wire bonding conditions from various viewpoints, we aim to enhance the robustness of our detection system.

TABLE I
DESCRIPTIONS OF TAG USED IN EXPERIMENTS

Tag	Description
accept	the wire bonding has no defects issue
reject	the wire bonding comes with defects such as tilted bonds, lifted bonds, or bond deformation.

In this work, we annotated the point cloud data using a graphical image annotation tool, named labelCloud, to outline and to label an object, in this instance as "reject" or "accept", with the bounding boxes drawn in order to extract label information used in the training process and export as .json file – see Fig. 5.

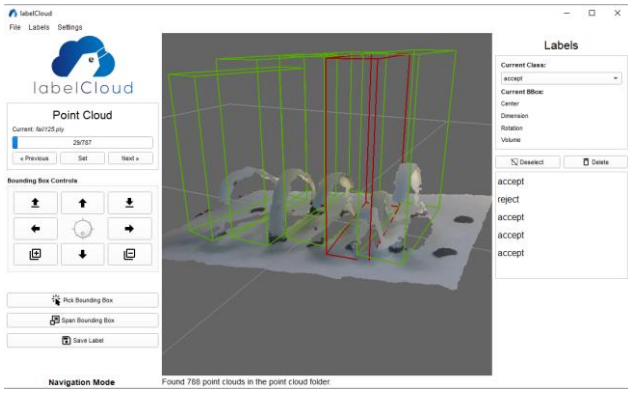


Fig. 5 labelCloud annotation

III. PROPOSED METHODOLOGY

In this section, we describe the methodology employed for evaluating the performance of the 3DETR, MLCVNet, H3DNet and VoteNet 3D object detection methods on our WireBondMockUp dataset – see Fig. 6. We highlight the key similarities and differences in their architectures, as well as the specific models used for our evaluation.

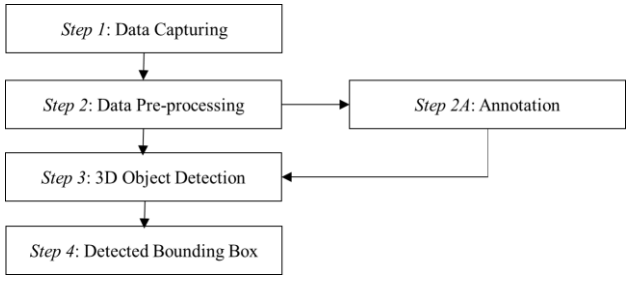


Fig. 6 3D Detection Pipeline

A. *Step1: Data Capturing*

The steps to obtain the WireBondMockUp was highlighted in Section II. A total of 800 3D point cloud data was collected. The annotation was completed for 170 3D point cloud data, used for training, while 34 is used for validation – see Fig. 7.



Fig. 7 WireBondMockUp train, test and validate dataset

B. *Step2: Data Pre-processing*

The 3DETR, MLCVNet, H3DNet and VoteNet 3D object detection methods accept point cloud data. The point cloud data is being converted based on the camera parameters, similar as used in VoteNet. The evaluation used the mean average precision(mAP) under different IoU thresholds, i.e., 0.25 and 0.5. The additional Step2A was detailed in Section II.

C. *Step3: 3D Object Detection*

We provide detailed descriptions of the specific models employed for our evaluation of the 3DETR, MLCVNet, H3DNet, and VoteNet methods – see Fig. 8. For each method, we outline the specific layers, modules, and parameter configurations for that model to accommodate the WireBondMockUp, in line with our evaluation objectives. We provide a comprehensive evaluation of the 3DETR, MLCVNet, and VoteNet methods on our self-generated datasets, highlighting both their shared architectural characteristics and distinct modeling choices in the discussion below.

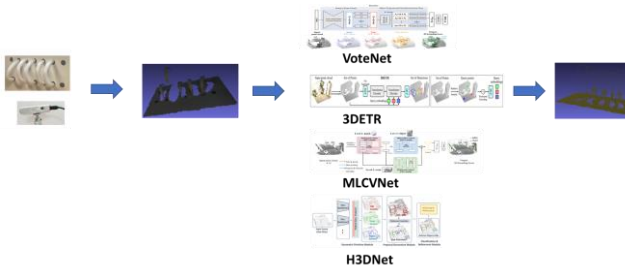


Fig. 8 Step3: 3D Object Detection [1-4]

VoteNet: VoteNet is presented by Qi et al. at 2019 and it is a point clouds based end-to-end 3D object detection algorithm using voting approach to estimate object's centre, bounding box and semantic classification. There are two major procedures has been highlighted in this VoteNet research: (1) point cloud feature learning through a backbone network and (2) learned Hough voting from seed points. For first procedure, they implement a deep network on point clouds for point

feature learning where the backbone of this deep network is constructed with multiple down-sampling (set-abstraction) layers and up-sampling (feature propagation) layers with skip connections and it outputs a group of seed points. The layers using this backbone network are from PointNet++ models [5]. Next, at the second procedure, each seed point generates a vote for centre coordinate of a bounding box through a shared multilayer perceptron (MLP). Then votes within a predefined radius distance are grouping and sampling to estimate the centroid of the object. Then bounding boxes and classification information are estimated around the predicted centroid of object through another shared MLP. All the estimated bounding boxes are filtering using 3D non-maximum suppression. The prediction outputs of this VoteNet including: an objectness score, bounding box parameters and semantic classification scores.

3DETR: 3DETR is presented by Misra et al. at 2021 and 3DETR is an end-to-end trainable Transformer that inputs 3D point clouds and outputs 3D bounding boxes. During training phase, this 3DETR uses a set-aggregation down-sampling operation from PointNet++ to reduce the number of points in point clouds [5]. Next the sampled points are set as input to Transformer encoder in order to produces a set of per-point features using self-attention layers. In the following, the output features from encoder and a set of 'query' embeddings are input to the Transformer decoder to predict bounding boxes using a parallel decoding scheme. For inference stage, the input point clouds are randomly sampled using Farthest Point Sampling [6] to generate a set of 'query' points and these that are embedded using Fourier positional embeddings [7] and pass to Transformer decoder. Then decoder takes this information and predicts centroid of object, bounding boxes and semantic classification, as well. There are five loss functions used in this 3DETR including: L1 regression losses for center and box dimensions (normalized in range [0,1]), Huber regression loss for angular residuals and cross-entropy losses for the angular and semantic classification. This 3DETR has been highlighted that it has the properties of simpler implementation and understanding by reducing many of the hard-coded design decisions in VoteNet and PointNet++, at the same time, it doesn't utilize ConvNet backbone in its architecture.

MLCVNet: MLCVNet is proposed by Xie et al. at 2020 and it is a 3D object detection network that alter the existing VoteNet algorithm by appending new self-attention and MLP modules to create a new framework which can capture and consider multi-level contextual information at patch, object and global scene levels. The contextual information at multiple levels in point clouds are captured using two self-attention modules and a multi-scale feature fusion module. As discussed by authors, it is important to employ multi-level contextual information for 3D object detection from colour-less and sparse point cloud data because it is difficult to identify the object when the point clouds appear independently, even for human naked eye. However, once there are extra information or knowledge from either surrounding environment or external domain, object recognition process will become relatively easier and this extra information is possible to be obtained from indoor scene, especially, due to indoor scene has specific man-made structural constraints or arrangement. The three important modules are highlighted in this research, which is patch-patch context (PPC) module, object-object context (OOC) module and global scene context (GSC) module. Their functionalities are explained as follows:

- Patch-patch context (PPC) module employs a self-attention layer to consolidate information from all the other point patches before sending it to the voting stage where the output point patch from this module contains not only its own local features, but also the information associated with all the other point patches
- Object-object context (OOC) module also employ a self-attention layer as well to capture relations between predicted objects where these predicted objects are in the form of vote clusters after voting stage as mentioned in VoteNet
- Global scene context (GSC) module extract the information from point patches and clusters using max pooling layers and concatenating them to form a global feature vector. Then this global feature vector is fed to an MLP to further aggregate the global information and this global information will be fused with the output from object-object context (OOC) module to generate final prediction result.

H3DNet: H3DNet is presented by Zhang et al. at 2020 as well. This H3DNet is an end-to-end 3D object detection network that inputs a dense set of 3D points (40000 points as proposed), predicts a group of geometric primitives (1 BB centers, 6 BB face centers, and 12 BB edge centers) with their latent features, then converts those information into object proposals and classifies each proposal as detected bounding box (BB center, BB size, and BB orientation, and a semantic label). There are three major components as discussed in this research:

- Geometric primitive module: A group of 3D points are fed into this module and this module predicts a set of geometric

primitives that includes locations of BB centers, BB face centers, and BB edge centers of the underlying objects

• Proposal generation module: This module inputs predicted geometric primitives and generates a set of object proposals where this object proposals are constructed as local minimums of a distance function, which is the innovate way as emphasized in this work.

• Classification and refinement module: This module classify each object proposal into a detected object and calculates offset vectors to refine the detected bounding box information (center, size and orientation) and a semantic label.The loss function of H3DNet uses L2 regression losses for geometric primitive location and shape, and uses cross-entropy loss for the rest, such as offset existence flag prediction, semantic labels.

D. Step4: Decision

The train model will be directly feed into the point cloud file for detection. For each detection, a counter is initialized with value 0 to keep track of how many objects detected. Objects detected can be viewed in sub folder, according to the configuration provided. The numbers of objects detected are displayed at the bottom, which corresponds to the number of the wire bond detected in the point cloud file – see Fig. 9.

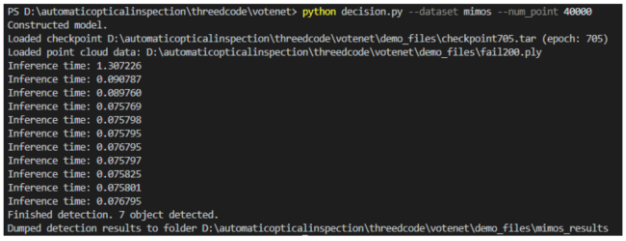


Fig. 9 Step4: Decision

IV. EXPERIMENTAL RESULTS

In this paper, we propose deploying different 3D object detection methods, such as 3DETR, MLCVNet, H3DNet and VoteNet, to study the WireBondMockUp datasets, which comes with different defects, in order to locate those defects accurately. Different methods could highlight the success or failure of the proposed architectures for the highly challenging dataset.

TABLE II
HYPERPARAMETER VALUE USED IN EXPERIMENTS

Hyperparameter	Description
Epoch	720
Batch Size	8 / 4 (3DETR)
Learning Rate	0.001
BN Decay Momentum	0.5
Platform (GPU)	NVIDIA GeForce RTX 2080

The performance of each method is determined by their precision and recall for the same testing set. All the experiments were performed on the same dataset for consistency, to provide a comprehensive evaluation of the methods on the WireBondMockUp, allowing for a fair and thorough comparison of their performance. – see Table II for more information.

A. Analysis1: General analysis on the model capability

In order to assess the performance of the 3DETR, MLCVNet, and VoteNet methods, we employ a set of hyperparameters used in the experiments – see Table II. A set of evaluation metrics shown in Table III, such as precision, recall, average precision, and Intersection over Union (IoU) are used to compare the performance of the method. We explain the significance of each metric in measuring the accuracy and robustness of the detection results.

During implementation, we encounter some deployment issues for each of the method. For VoteNet, generating own dataset proves to be a time-consuming task, as we have to familiarize with the format of the train and validation dataset. Though it is not superior in terms of performance, we found that VoteNet are able to support training, visualization and evaluation. For H3DNet, the main reason we select this method

TABLE III
PERFORMANCE COMPARISON ON DIFFERENT METHODS

Model	**Metrics**				**F1 score for 0.25**	**F1 score for 0.50**	**Inference (s)**
	mAP 0.25	**mAP 0.50**	**AR 0.25**	**AR 0.50**			
VoteNet[2019]	19.71	4.96	93.02	52.09	32.53%	9.06%	0.0788
MLCVNet[2020]	31.21	8.96	97.93	57.25	47.33%	15.50%	0.09175
H3DNet[2020]	0.0002	Not Available	0.8197	Not Available	0.03%	Not Available	Not Available
3DETR[2021]	64.35	17.76	88.31	40.55	74.44%	24.70 %	0.08351

is due to its ability to predict thin objects in 3D scenes. However, the performance on the WireBondMockUp may be influence by its computation limitation, having been able to execute only the single backbone due to the selected GPU (one GPU), and the wire bond structure 1) lacks of being a complete and solid object, and 2) poor planar structures. For MLCVNet, its performance is better than VoteNet due to its three sub-modules ability in 1) gathering data when data is missing, 2) gathering contextual information of object candidates, and 3) gathering global scene context. However, though the performance is superior than VoteNet, the performance for detecting "reject" and "accept" may be limited by this contextual information, as the structure of the "reject" part and "accept" part are quite similar. Through these extensive experiments, we show 3DETR outperforms VoteNet, MLCVNet and H3DNet. In terms of implementation, 3DETR is conceptually simple and easy to implement. The non-parametric queries and Fourier encodings techniques are able to detect the point clouds as they are irregular and sparse.

B. Analysis2: Further analysis on 3DETR loss and training performance

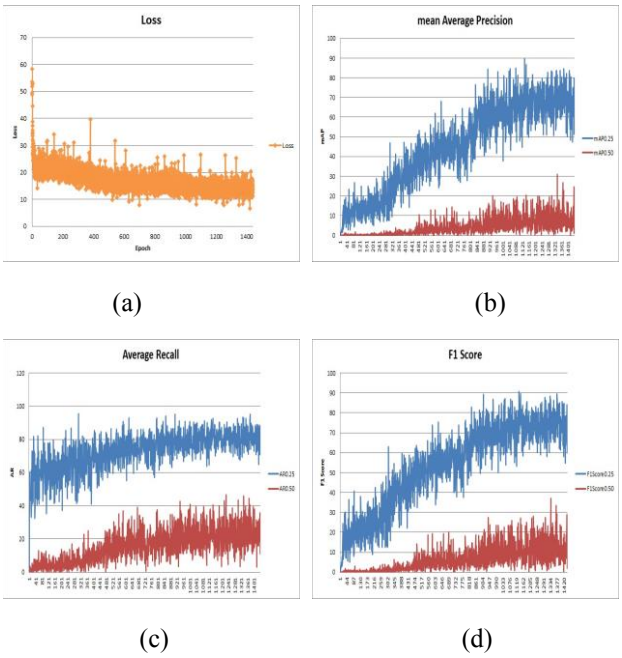


Fig. 9 Training performance of 3DETR on WireBondMockUp

In this further analysis, we explore the capability of 3DETR when we further trained the model – see Table IV.

TABLE IV
HYPERPARAMETER VALUE AND ACCURACY ACHIEVED IN THE EXPERIMENTS

Hyperparameter	Description
Epoch	1440
Batch Size	4
Learning Rate	0.001
BN Decay Momentum	0.5
Platform (GPU)	NVIDIA GeForce RTX 2080
mAP 0.25	90.85
mAP 0.5	22.11
AR 0.25	94.05
AR 0.5	41.23
F1 score for 0.25	92.42
F1 score for 0.5	28.78

The performance continues to increase when trained longer. The transformers-based 3D detection outperformed the traditional 3D object detection methods. As discussed in 3DETR publication, our experiment validated that 3DETR is capable to incorporate the inductive bias from irregular and incomplete points clouds from our WireBondMockUp dataset and perform accurate bounding box detection.

V. CONCLUSIONS

The recent rise in the popularity of deep learning has sparked interest among many in the field of machine learning. Not only does it outperform traditional machine learning techniques, but deep learning has also rapidly garnered a lot of attention and become a key technology behind many recent applications. These findings have motivated many products, especially in defect analysis, to embark on deploying deep learning tools in their product. Our self-generated datasets, named WireBondMockUp are able to capture component structures, capturing variations in object types, orientations, and occlusion levels that are relevant in the study. It is important to note that our work emphasizes the practical relevance by providing implementation scenarios and applying these methods directly on the WireBondMockUp. This work proposed to investigate the performance of the four 3D object detection methods which comes with different network architectures, on the WireBondMockUp. Experimental results indicate that extending the training epoch to 1440 for 3DETR shows that it could outperformed other architectures, achieving 92.42% in F1-score accuracy.

ACKNOWLEDGMENT

This work is supported by the RMK12 Project No. 120421/120422/120423 “Pembangunan Hyper Scale AI Analytics Untuk Kepintaran Operasi” from MOSTI, Malaysia.

REFERENCES

[1] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep Hough Voting for 3D Object Detection in Point Clouds," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9277-9286, Seoul, Korea, October 2019.

[2] Z. Zhang, B. Sun, H. Yang, and Q. Huang, "H3DNet: 3D Object Detection Using Hybrid Geometric Primitives," in Proceedings of the European Conference on Computer Vision, pp. 311-329, Springer, Glasgow, Scotland, UK, August 2020.

[3] Q. Xie, Y. K. Lai, J. Wu et al., “Mlcvnet: multi-level context votenet for 3d object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10447–10456, Seattle, WA, USA, June 2020.

[4] I. Misra, R. Girdhar, and A. Joulin, “An end-to-end transformer model for 3d object detection,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2906–2917, Montreal, Canada, October 2021.

[5] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in Advances in Neural Information Processing Systems 30, 2017.

[6] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi, "The Farthest Point Strategy for Progressive Image Sampling," IEEE Transactions on Image Processing, vol. 6, no. 9, pp. 1305-1315, 1997.

[7] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains," in Proceedings of the NeurIPS (Conference on Neural Information Processing Systems), 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Predicting Outcomes of Cognitive Behavioral Therapy for Depression Using Data Driven Approaches

Lily Tyszczuk*‡ and Liat Levita† and Jaime Delgadillo* and Zhang Haihong‡ and Mahnaz Arvaneh*
* The University of Sheffield, Sheffield, United Kingdom
† University of Sussex, Falmer, Brighton, United Kingdom
‡ Institute for Infocomm Research, A*STAR, Singapore

***Abstract*—Cognitive Behavioural Therapy (CBT) is a well established method for treating clinical depression. However, it is only considered successful for roughly half of those receiving it. Transitioning patients who are unlikely to benefit from CBT to alternative treatments, at an early stage, can significantly enhance their quality of life while alleviating the burden on mental health services. This study focuses on predicting the outcome of CBT using Linear Discriminant Analysis (LDA) models. These models are trained using demographic information, as well as data from depression (PHQ-9), anxiety (GAD-7), and functional impairment (WSAS) questionnaires administered during the initial four CBT sessions. The study aims to determine how the accuracy of the prediction models is influenced by different factors, such as the number of therapy sessions considered, the specific questionnaires used, and the analysis of individual symptoms versus the overall questionnaire scores. The data from over 8000 patients receiving high-intensity CBT (step 3 and 4) is analyzed. Prediction models are trained using different combinations of features and validated across multiple theray sites to ensure generalizability. The best performing model achieved a test accuracy of 69.7% (Odds Ratio OR = 5.20) in predicting whether Reliable and Clinically Significant Improvement (RCSI) was achieved. Interestingly, including baseline scores, demographic information, and WSAS/GAD7 data did not significantly improve the accuracy beyond using PHQ9 responses alone. However, incorporating data from successive therapy sessions consistently improved the prediction accuracy, up to a maximum of four sessions (the limit of the analysis). Surprisingly, analyzing individual symptom trajectories did not offer any advantages over analyzing the total scores of the questionnaires. In conclusion, the study demonstrates that monitoring responses to therapy during the initial sessions can be a valuable tool in predicting treatment outcome. Although there were significant correlations between RCSI and symptom level change, the study found no improvement in the model's prediction accuracy when considering: (1) individual symptom scores instead of the overall questionnaire score, (2) using GAD-7 and WSAS questionnaires in addition to PHQ-9, or (3) including demographic information.**

I. Introduction

Clinical depression, one of the leading contributors to global disability burden [1], can be debilitating, draining all motivation and pleasure from life, and in severe cases contributing to suicide. Co-morbidities are common - depression complicates care and worsens prognosis for conditions as varied as cancer [2], schizophrenia [3], and heart disease [4]. One of the most effective treatments for clinical depression, Cognitive Behavioural Therapy (CBT), is a talking therapy focused on interactions between thoughts, feelings, and behaviours. However, CBT still only appears to be successful in around half of patients - e.g. in the dataset analysed in this paper, 53.9% of patients achieved Reliable and Clinically Significant Improvement (RCSI).

This study was designed to predict which patients undergoing CBT are likely to achieve RCSI as measured using the Patient Health Questionnaire 9 (PHQ-9) [5], and determine which of the demographic, self-report questionnaire, and symptom information available were most important in making these predictions. Using such outcome prediction models, patients determined to be unlikely to achieve RCSI during CBT may be offered more individualised treatment plans, by way of alternate therapists, therapy styles, or treatments. This could lead to higher chances of recovery, enhancing their quality of life and reducing burden on mental health services.

Some studies have shown that prediction of psychological therapy outcomes can be made even before the first session. Finegan et al, in a review of socio-economic status and psychological therapy outcomes for depression and anxiety, found unemployment, low income, and relative neighbourhood deprivation to be indicators of poor therapy outcome, whereas social class and educational level were not found to be good predictors [6]. Huibers et al. used five predictors (gender, employment status, anxiety, personality disorder and quality of life) and six moderators (somatic complaints, cognitive problems, paranoid symptoms, interpersonal self-sacrificing, attributional style and number of life events) to make recommendations for depressed patients for choosing either cognitive therapy (CT) or interpersonal psychotherapy (IPT) [7]. Some features indicated good prognosis for both therapies (employment, low anxiety, female gender), while others were more common in those for whom a specific therapy style was more likely to be effective [7].

The therapist also affects outcome. The 'Common factors', including therapist empathy, alliance, cultural adaptation, and therapist effects have an impact on treatment [8]. In addition, the therapist's facilitative interpersonal skills [9], self-doubt

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

[10], and humour style [11] are important, as is the frequency and regularity of therapy, with weekly being better than less frequently [12].

Once therapy has commenced, there is consistent evidence that early symptom change is a reliable indicator of eventual remission. A recent systematic review and meta-analysis investigated the link between early responses to psychological therapy and treatment outcomes for depression and anxiety, finding that early responders had significantly better post treatment outcome, with a large overall effect size of early response on final outcome (g = 0.87, Odds Ratio OR = 4.84) [13]. While study designs and definitions varied across included trials, results were strong and consistent, and the authors stressed the importance of regularly monitoring a patient's symptoms.

We found no studies explicitly analysing changes in individual symptoms (rather than overall symptom score), but some symptoms of depression, such as insomnia, are studied in the context of being a co-morbidity. For example, Bei et al. used growth mixture modelling on overall levels of depression on 148 patients with insomnia and major depressive disorder (MDD) [14]. They linked symptom trajectories to insomnia, claiming early changes in insomnia may predict long-term depression outcomes.

Consequently, in this study we aim to investigate whether CBT outcome can usefully predicted in an early stage of therapy. Critically, we extend on previous work by investigating the utility of individual symptoms rather than overall level of depression, and which symptom questionnaires are most useful for outcome prediction models. As discussed, demographic information about the patient has demonstrable relationships to therapy outcome [6], but to what extent can this help a prediction model, and for how long is it still helpful when also tracking symptoms? This study has been designed to address these as yet unanswered questions by training Linear Discriminant Analysis (LDA) outcome prediction models on data from over 8800 patients receiving Cognitive Behavioural Therapy (CBT) at Improving Access to Psychological Therapy (IAPT) sites across England. This is a large and demographically diverse dataset, with the ability to cross-validate across IAPT sites across England, ensuring our models generalize well. These models were trained on different subsets of the available questionnaire and demographic data to answer the following questions. To the best of our knowledge, only the first of these five questions has been investigated explicitly by other researchers. Each has practical clinical use.

1) **How accurately can the proposed model predict success of CBT by tracking early symptom changes?** Calculating accuracy of the models trained will allow us to determine the utility of symptom tracking to mental health services.
2) **How does the proposed model's prediction accuracy change with each additional session of therapy analysed?** Participants filled in symptom questionnaires at each session of therapy, and we varied the number of therapy sessions included in training the prediction models from 0 (including only demographic information and baseline scores) up to 4.
3) **Which questionnaires and their derived features are key for outcome prediction by the proposed model?** We will restrict our prediction models to different questionnaires (out of PHQ-9, GAD-7, and WSAS) and combinations thereof, allowing us to assess the impact on the prediction accuracy.
4) **Does tracking individual symptoms offer any advantage to the proposed model over tracking the total symptom score?** We aim to expand on results of previous studies by comparing the use of total symptom level (i.e., PHQ-9 score) with individual symptoms (i.e. responses to each question item on the PHQ-9) in terms of achieved prediction accuracy.
5) **Does including demographic information improve the proposed model's prediction accuracy?** Demographic information alone has been shown to be useful in predicting therapy outcome [6]. Here, we aim to investigate to what extent its inclusion in prediction models can improve accuracy when tracking symptom scores.

II. Methodology

A. Participants and Treatment

Data were collected from 8865 (female = 5787, male = 3065, not available = 12) patients with a primary diagnosis of either 'depressive episode' or 'recurrent depressive disorder' who had attended at least 4 sessions of high intensity (Step 3 or 4) CBT. Patients were from 16 Improving Access to Psychological Therapy (IAPT) [15] Services across 7 Sites in the United Kingdom between 2010-2015. Sessions were 1-to-1, typically lasting an hour each. Patients had on average 12.7 (±5.0) sessions of CBT. Some patients had been on lower intensity group therapy sessions (Step 2) prior to step 3 or 4 - responses to questionnaires during step 2 were not included in this analysis.

Exclusion Criteria. Patients for whom RCSI was not calculable (beginning therapy with PHQ-9 score below 10, or no PHQ-9 data at either baseline or final therapy session) were removed, after which the dataset consisted of 8027 patients. Note that many of these removed patients still had calculable PHQ-9 change, and statistical tests involving PHQ-9 change were done on the larger sample of 8862 patients. Please see the section on *Outcomes*, for more details on how RCSI is calculated and why the baseline PHQ-9 score of less than 10 was an exclusion criteria.

B. Measures

Demographic Data. In this study, demographic data consists of gender, age, ethnicity, employment status, 2015 IMD (Index of Multiple Deprivation) decile of the region the patient lives in [16], whether the patient is taking medication, the presence of any long-term health conditions, and the presence of any disabilities. All these categories apart from age and IMD decile have been binarised for all proceeding analyses, i.e, female or male, White or Non-White, employed or unemployed,

taking medication or not taking medication, long term health condition or no long term health condition, disability or no disability. These variables are all taken at the first assessment by a general practitioner (GP), before therapy had started.

Symptom and Questionnaire Data. Baseline levels were included for the three following questionnaires at initial assessment with a GP. Additionally, responses to each of these questionnaires were recorded at each session of therapy. All were self-report Likert-type scales.

- **PHQ-9** The Patient Health Questionnaire 9 is a nine-item tool for screening, diagnosing, monitoring, and measuring the severity of depression. The PHQ-9 questions are based on diagnostic criteria of depression from DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) and ask about the patient's experience in the last two weeks [5]. Questions ask the patient "Over the last two weeks, how often have you been bothered by any of the following problems?", with four possible responses ranging from "Not at all" (score of 0) to "Nearly every day" (score of 3). A score of 10 or above has a sensitivity of 88% and a specificity of 88% for detecting major depression [5].
- **GAD-7** The Generalized Anxiety Disorder 7 is a seven-item that is used to measure or assess the severity of general anxiety disorder (GAD) [17]. Each item asks the individual to rate the severity of their symptoms over the past two weeks. Response options include "not at all", "several days", "more than half the days" and "nearly every day". GAD-7 shows excellent internal consistency and good test-retest reliability [17].
- **WSAS** The Work and Social Adjustment Scale is a five item scale assessing the impairment a patient feels as a result of a health condition [18]. This questionnaire includes five questions assessing impact of a condition on the patient's work, personal, and social life. Questions are of the form "Because of my [problem] my [. . .] is impaired. '0' means 'not at all impaired' and '8' means very severely impaired to the point I can't [...]."

Treatment Outcome. Full recovery from depression was defined as moving to below the clinical cutoff score of 10 points and more than 6 points of change on the PHQ-9 questionnaire. This standard is known as Reliable and Clinically Significant Improvement (RCSI), which is a widely accepted measure of recovery [19]. PHQ-9 scores used to determine RCSI were those taken at the initial assessment with the GP (i.e. before any session of therapy), compared with PHQ-9 score taken at the final session of the attended therapy.

Statistical Tests for Feature Comparison. Statistical analysis of the individual features were not the purpose of this paper and will not be commented on in depth. It was still informative to perform some statistical analysis on the features used. These tests in no way influenced the model training process. Tables I, II give some of these statistics, with the majority omitted for conciseness.

We considered ANOVA tests unsuitable, due to the large amount of dependency between features. We calculated rank biserial correlation (RBC) using the MES toolbox [20] in MATLAB between each of the symptom based features and the dichotomous outcome of RCSI or no RCSI. We also calculated Kendall correlation between each of the calculated features and overall PHQ-9 change during the analysed therapy sessions. The same was performed to compare the overall PHQ-9 change against age and IMD decile, the only nominal demographic data. For the categorical demographic data, we calculate Phi correlation with RCSI and Kendall correlation with PHQ-9 change. Appropriate p-values are calculated for each. Each of the tests used are non-parametric, allowing for fairer comparisons between features with different distributions.

C. Model Training

Prediction models were trained to determine whether RCSI was achieved or not by end of treatment, with cross-validation across different IAPT sites.

Feature Extraction. Responses to each question from PHQ-9, GAD-7 and WSAS from the first $0 \leq n \leq 4$ sessions of therapy were analysed. The mean, the slope, the root mean squared error (RMSE) from the slope, and the standard deviation (SD) of the first n sessions of each question as well as the summed total for each questionnaire were used as the features of the prediction model. RMSE and SD are similar but distinct variables - SD tells us the overall variability in a patient's symptom, while RMSE tells us how closely symptom change follows a linear slope. To calculate the mean, we needed at least one session of therapy. To calculate slope and SD, we needed at least two sessions, and RMSE requires at least three. For the sake of simplicity and ease in interpreting the results, we did not extract more complex non-linear features.

Splitting Data for Testing and Training. Our models were evaluated with cross-validation, whereby the data are split into distinct 'folds' for training and testing. Prediction models are trained on all but one fold of the data, and tested on this 'test fold'. This process is repeated until each fold has been the test fold exactly once, and the aggregated results from each test fold are reported.

The dataset includes data on patients from seven IAPT Sites across the UK, of unequal sizes. To get more evenly distributed cross-validation fold sizes, we merged the three smallest Sites into one, and divided the largest Site in two (this division still maintained distinction between Services within the Site). This resulted in 6 folds, of sizes 1263, 961, 1685, 1042, 1698, and 1378, respectively. From a machine learning perspective, the unequal split is not ideal, but we considered it important to demonstrate generalisation across IAPT sites, and regard this as a key strength of this paper.

Replacing Missing Data. With such a large dataset of this type missing values are inevitable. Missing continuous values in each test and train fold are replaced by the mean of each feature in the train fold. Missing categorical values are replaced by the mode in the train fold. To avoid biasing the results, we needed to make sure training data and model training do not have any dependency on the test data. Thus, replacing missing

values was done only based on the information from the train data. In a real-world application, the data used to train the model would also be used to estimate missing values in the testing data.

Training and Testing Prediction Models. Linear Discriminant Analysis (LDA) models were trained separately on a variety of feature subsets. LDA classifiers create a hyperplane in feature space separating the classes – RCSI or no RCSI. Feature subsets are characterised by

- Questionnaire data from how many sessions of therapy are included in analysis, from 0 to 4 (5 choices)
- Which subset of questionnaire data (including or excluding each of PHQ-9, GAD-7, and WSAS) are used in analysis, using at least one of the three questionnaires (7 choices)
- The use of summed total scores alone for each questionnaire, or individual symptom scores (2 choices)
- The inclusion or exclusion of demographic information (2 choices)

This results in $5 \times 7 \times 2 \times 2 = 140$ potential feature subsets for comparison, across 6 folds. LDAs, which are computationally cheap to train, were chosen so we could quickly compare results from this large number of feature subsets.

Permutation test. To guarantee the fairness of our methods we evaluated our training procedure by randomising the outcome variable (PHQ-9 RCSI) and then performed the exact same model training and evaluation method for each feature subset 100 times. The resulting models had mean accuracies (per feature subset) between 49.5% and 50.4%, with standard deviations of at most 2% - i.e, we are confident the models performed at random chance level (for binary classification). This is strong evidence that any prediction accuracies reported are a result of relationships within the data and not methodological biases.

III. Results

We present the results as answers to the research questions given in the introduction. For each, we consider the principle of selecting the most parsimonious model within one standard error (SE) of the highest performing model [21] - i.e., does including a subset of features improve prediction accuracy by at least one standard error? With many different feature subsets, it can be difficult to decide which is the simplest model – therefore we have selected two within one standard error of the best performing model. All classification accuracies given are the weighted mean accuracies of the test folds, with weighted standard error.

A. How accurately can the proposed model predict success of CBT by tracking early symptom changes?

The best performing outcome prediction model achieved 69.7% (±0.61) test accuracy, with Odds Ratio OR = 5.20. This model was trained on all three questionnaires, included demographic information, and did not use individual symptom scores. Models with no symptom tracking, based solely on demographic information and symptom level at initial assessment, (shown by those at 0 sessions of therapy in Figs. 1–3) achieved at most 60.8% (±0.72) accuracy (OR=2.38) when using all 3 questionnaires and demographic information.

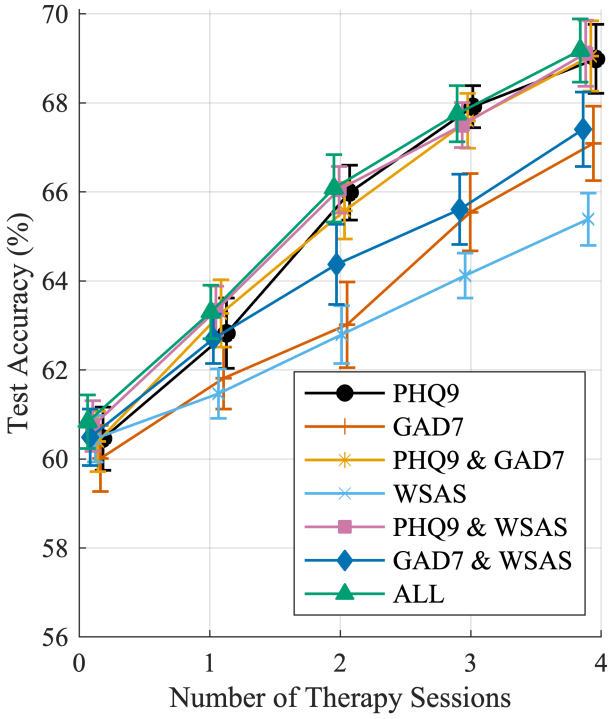


Fig. 1. Cross-validated mean test accuracy for prediction of RCSI with standard error bars against number of therapy sessions analysed, for a range of feature subsets. Mean and standard error are appropriately weighted by fold size. We can see the inclusion of questionnaires other than the PHQ-9 doesn't significantly improve prediction accuracy. For each symptom questionnaire combination in this figure, models include demographic information and use individual questionnaire responses.

B. How does the proposed model's prediction accuracy change with each additional session of therapy included in the model?

We found that, for each feature subset, increasing the number of analysed sessions improved prediction accuracy. For the best performing models, accuracy increased most during the first 2 sessions, from at most 60.78% (±0.72) to at most 66.08% (±0.77) and continued increasing to at most 69.7% (±0.61) by session 4. Inclusion of each subsequent therapy session in analysis had a mean accuracy greater than one SE from the previous mean accuracy. Figs. 1-3 show this effect.

C. Which questionnaires and symptoms are the most important for outcome prediction by the proposed model?

We found weak evidence for additional benefit of including WSAS or GAD-7 compared to PHQ-9 alone at either baseline or at any session monitored. This is evidenced by strong overlap in SE bars seen in Fig. 1. Given RCSI is determined by change in PHQ-9, it is no surprise that PHQ-9 was the most important for outcome prediction. Removal of PHQ-9 from session 2 onwards resulted in mean accuracy decreasing

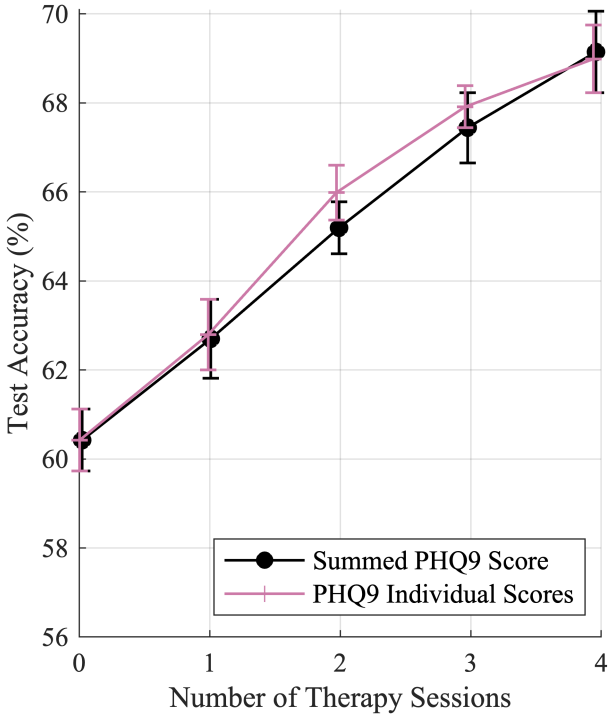


Fig. 2. Cross-validated mean test accuracy for prediction of RCSI with standard error bars against number of therapy sessions analysed. 'Summed PHQ9 Score' represents models trained only on total PHQ-9 score features, 'PHQ9 Individual Scores' represents models trained on features from all PHQ-9 questions. We see no significant benefit to the extra information from using individual questions rather than total score. Demographic information is included for the models in this figure.

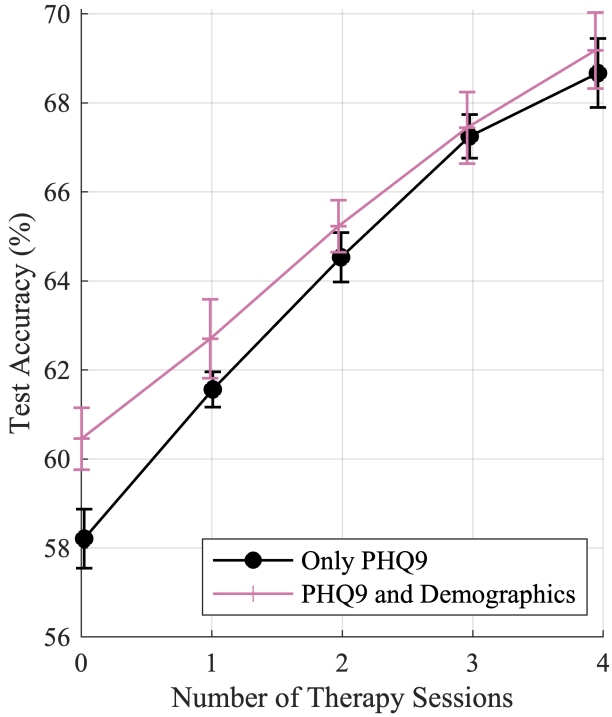


Fig. 3. Cross-validated mean test accuracy for prediction of RCSI with standard error bars against number of therapy sessions analysed. We see the benefit of including demographic information decays as more sessions are included. Individual PHQ-9 responses are used for the models in this figure.

by more than one SE as seen in Fig. 1. It is interesting that only small improvement to prediction accuracy was found by including either GAD-7 or WSAS, despite each individually achieving prediction accuracies of 67.41% (±0.75) and 65.39% (±0.52) respectively.

WSAS was used for one of the most parsimonious models within 1 SE of the best performing, achieving 69.05% (±0.77) with individual response scores and no demographic information. GAD-7 was used in the best performing model, but in neither of the most parsimonious models within 1 SE of the best performing. In Table I we can see that RBC against RCSI for WSAS baseline score (-0.170) is stronger than that of PHQ-9 baseline score (-0.141). Features calculated on 4 sessions of therapy [1] consistently show PHQ-9 features are the best performing for predicting both RCSI and PHQ-9 change, though many GAD-7 and WSAS derived features perform well. The features with strongest RBC against RCSI were consistently mean of the first four sessions, with higher mean making RCSI less likely. These were followed by slope features, with a steeper slope indicative of a higher chance of RCSI. A higher standard deviation and, to a lesser extent, higher RMSE from the slope, were indicators of achieving RCSI, with the most notable exceptions of PHQ-9 Q8 and Q9.

[1]At 96 features, we were unable to include this table in this paper

D. Does tracking individual symptoms offer any advantage to the proposed model over total symptom score?

Little evidence was found that including individual responses of the self-reported questionnaires added any benefit to the models over tracking total scores for each questionnaire. Fig. 2 shows consistent strong overlap in SE bars when we compare individual symptom analysis with total score. This is contrary to our expectations that a more detailed profile of a patients change in symptoms would help prediction accuracy. However, individual responses were used in one of the most parsimonious models within 1 SE of the best performing, achieving 69.05% (±0.77) when using PHQ-9 and WSAS responses.

E. Does including demographic information improve the proposed model's prediction accuracy?

Including demographic information in the analysis initially improves prediction accuracy by around 2%. As more sessions were included, we see the prediction accuracies converged, and by session 4 there was overlap in SE bars, visible in Fig. 3. Of the demographic features, only employment (φ= -0.193 RBC = 0.101) and disability (φ= 0.146) had effect sizes with absolute value > 0.1 (although the majority were statistically significant). In our dataset, disability had extremely high rates of missing values, so it's use in prediction is

TABLE I

RANK BISERIAL CORRELATION (RBC) AGAINST RCSI AND KENDAL CORRELATION (τ) AGAINST OVERALL PHQ-9 CHANGE FOR AGE, IMD DECILE (THE ORDINAL DEMOGRAPHIC FEATURES) AND BASELINE SCORES FOR THE THREE QUESTIONNAIRES. PHQ-9 RCSI AND CHANGE ARE CALCULATED BY THE DIFFERENCE BETWEEN FIRST AND FINAL SESSION OF THERAPY. NOTE THAT WHEN CALCULATING EFFECT SIZES FOR PHQ-9 CHANGE, SOME PATIENTS FOR WHOM RCSI WAS NOT CALCULABLE (E.G STARTING VALUE OF PHQ-9 TOO LOW) WERE INCLUDED. RCSI IS ENCODED AS 0 FOR NO RCSI AND 1 FOR RCSI.

Feature	RBC	p-value	τ	p-value
Age	-0.037	<0.001	-0.004	0.587
IMD decile	0.064	<0.001	-0.000	0.978
PHQ9 first	-0.141	<0.001	-0.289	<0.001
GAD7 first	-0.099	<0.001	-0.159	<0.001
WSAS first	-0.170	<0.001	-0.075	<0.001

likely less than suggested by the φvalue. One of the most parsimonious models within 1 SE of the best performing included demographic information and summed total PHQ-9 score, with accuracy of 69.18% (±0.71).

F. Summary

Tracking early symptom levels with the PHQ-9 questionnaire during CBT therapy is a promising tool in predicting outcome. We have demonstrated the use of adding each successive session of therapy, up to 4. The most parsimonious models with mean accuracies within 1 SE of the top performing model (69.67%±0.61) were trained on either (1) PHQ-9 total score and demographic information (69.18%±0.71) or (2) PHQ-9 and WSAS individual scores (69.05%±0.77). Models trained on just PHQ-9 score achieved 68.88% (±0.68) and 68.69% (±0.68) when using individual question responses or summed total respectively.

IV. DISCUSSION

A. Interpretation

Increasing accuracy with each successive session of therapy included in model training is as expected – the more sessions of therapy analysed, the clearer the prediction of the final outcome is. Contrary to what was predicted, analysing individual symptom trajectories gave little advantage when compared to summed symptom score, despite features derived from some individual questions, such as PHQ-9 Q2, having relatively strong effect sizes for both RCSI and symptom level change.

Similarly, while it was anticipated that models trained on PHQ-9 would outperform models trained on GAD-7 or WSAS, it was unexpected to find that adding GAD-7 and WSAS features to PHQ-9 features only resulted in a slight increase in outcome prediction accuracy. This is even more surprising because not only did many features derived from GAD-7 and WSAS exhibit similar or even better performance to those derived from PHQ-9 (in terms of correlation with CBT outcomes). For instance, GAD-7 total std (RBC = 0.212) demonstrated better correlation results compared to PHQ-9 total std (RBC = 0.198), and WSAS baseline score (RBC = -0.170) outperformed PHQ-9 baseline (RBC = -0.141).

TABLE II

PHI (φ) CORRELATION AGAINST PHQ-9 RCSI AND RANK BISERIAL CORRELATION (RBC) AGAINST PHQ-9 CHANGE FOR BINARY DEMOGRAPHIC FEATURES. RCSI AND PHQ-9 CHANGE ARE CALCULATED BY THE DIFFERENCE BETWEEN THE PHQ-9 SCORES OF THE FIRST AND FINAL SESSION OF THERAPY. NOTE THAT WHEN CALCULATING EFFECT SIZES FOR PHQ-9 CHANGE, SOME PATIENTS FOR WHOM RCSI WAS NOT CALCULABLE (E.G STARTING VALUE OF PHQ-9 TOO LOW) WERE INCLUDED. RCSI IS ENCODED AS 0 FOR NO RCSI AND 1 FOR RCSI. THE BINARY DEMOGRAPHIC FEATURES ARE ENCODED IN THE ORDER GIVEN IN THE TABLE AS 0 AND 1 RESPECTIVELY, WITH CLASS SIZES GIVEN (E.G, FOR GENDER, FEMALE (N=5787) IS ENCODED AS 0 AND MALE (N=3065) AS 1. φ=-0.025 SUGGESTS MALE PATIENTS ARE LESS LIKELY TO ACHIEVE RCSI AND RBC = 0.027 SUGGESTS MALE PATIENTS HAVE LESS OF A DECREASE IN PHQ-9 SCORE – NOTE HOWEVER, BOTH OF THESE EFFECT SIZES ARE VERY SMALL).

Feature	φ	p-value	RBC	p-value
Gender				
Female = 5787	-0.025	0.024	0.027	0.010
Male = 3065				
Ethnicity				
Non-white = 1614	0.029	0.010	0.002	0.832
White = 6872				
Unemployed				
Employed = 5986	-0.193	<0.001	0.101	<0.001
Unemployed = 2558				
Medication				
No Medication = 4204	-0.046	<0.001	-0.028	0.012
Medication = 3824				
Long Term Condition (LTC)				
LTC = 2587	0.089	<0.001	-0.037	0.001
No LTC = 4645				
Disability				
Disability = 355	0.146	<0.001	-0.074	0.002
No Disability = 1399				

The initial advantage of incorporating demographic information in the prediction model aligns with previous research findings, as expected. Notably, to our knowledge, our study is the first to demonstrate the diminishing impact of demographic information over time. Our study showed that relying solely on PHQ-9 is insufficient to achieve prediction accuracies within 1 standard error of the top-performing model. However, the inclusion of either (1) PHQ-9 total score and demographic information (69.18%±0.71) or (2) both PHQ-9 and WSAS individual scores (69.05%±0.77) were two of the most concise and effective models.

B. Strengths and Limitations

Modelling Assumptions. Statistical analysis of subjective, self-report, Likert scales is fraught with difficulty. Our feature extraction process assumes a linearity in score differences (e.g that the difference between '0' and '1' is the same as between '2' and '3'), which almost certainly oversimplifies and misses nuance.

Additionally, some of the assumptions required for LDA classifiers to function optimally were violated. The features included in the model did not follow gaussian distributions closely, and there is a large amount of correlation between features. However, these are not requirements to use an LDA classifier, and we are satisfied with the performance of our models.

In this study a simple model training procedure was used, with the emphasis on comparing the effects of different subsets of data rather than aiming to maximise accuracies. It remains possible that sophisticated models, which can more accurately capture interactions between features, may find benefit in including information our models did not require.

Generalisation. A key strength of this study was the use of a large dataset, with the ability to cross validate over different Sites and Services - there was never overlap in the IAPT Services on which models were trained and tested. This is an important, yet rare, element of our research which greatly strengthens our confidence that the results generalise well. Our best Odds Ratio was 5.20, larger than the mean of 4.84 found by Beard and Delgadillo in their 2019 review [13].

C. Implications for research, theory, and practice

We recommend more research and refinement of models before clinical use. This is a preliminary study demonstrating the potential of symptom tracking and investigating which subsets of information are most valuable to our models for predicting treatment response. Informed by this, we expect a more sophisticated model to achieve higher accuracies.

In this study, no additional benefit monitoring responses to the GAD-7 or WSAS (vs. PHQ-9 alone) was found. However, other questionnaires may prove more useful in outcome prediction. For example, the PHQ-9 has been found to inaccurately capture the presence and intensity of suicidal ideation and self-harm [22]. Another common criticism is, by asking only the frequency of symptoms in the last two weeks, the PHQ-9 is unable to determine accurately the intensity of any symptoms experienced.

Similarly, it may also be a useful research question to develop an 'optimally short' questionnaire that can accurately monitor whether therapy is on track. The fact that monitoring individual symptoms gives no benefit over total PHQ-9 score suggests there may be redundancy in the PHQ-9 for the purposes of therapy monitoring. The PHQ-2 may be a good candidate, as features derived from the questions it contains, Q1 and Q2, both had relatively strong effect sizes. A shorter questionnaire has two main benefits - reducing time spent filling in questionnaires for patient and therapist, and subsequently reducing the frequency of missing values, which could in turn increase prediction accuracies.

Synthesising results from studies of depression treatment outcome using varied predictors (such as biomarkers, demographic information, symptom profile, patient preference) for different treatment methods (such as therapies and medication) could greatly improve initial treatment selection. In a review on treatment efficacy for depression, Khan et al. highlight the need for better understanding of how different patient profiles may respond to different treatment methods, and how the heterogeneity of depression needs to be untangled to determine which treatment will work for whom [23].

Once treatment has commenced, we have demonstrated the value of treatment monitoring to determine whether it would be in the patient's interest to transition to an alternate treatment, perhaps with closer monitoring of those considered to be of high risk of not responding to therapy as suggested by Delgadillo et al. [24]. Given the devastating personal, societal, and financial costs of depression, an improvement in treatment assignment and outcome prediction would have enormous impact and should be considered a priority.

D. Conclusion

Tracking early symptom change during CBT is a promising method of predicting future treatment outcome, with each additional session of therapy included in analysis improving prediction accuracy by around 2%, achieving accuracies of 69.67%±0.61 (OR = 5.20) by session 4. The most parsimonious models within 1 SE were trained on either (1) PHQ-9 total score and demographic information (69.18%±0.71) or (2) PHQ-9 and WSAS individual scores (69.05%±0.77). Including the GAD-7 questionnaire gave negligible improvement on prediction accuracy, despite significant effect sizes of the individual features. Including demographic information had the strongest benefit to prediction accuracy before 2 sessions of therapy have been completed, and thereafter small. Outcome prediction could be used in the future to aid clinicians and patients in making the decision to personalise and alter treatment plans, increasing patient quality of life and reducing strain on mental health services.

REFERENCES

[1] S. L. James, D. Abate, K. H. Abate, *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, Nov. 2018, ISSN: 01406736. DOI: 10.1016/S0140-6736(18)32279-7.

[2] H. R. Smith, "Depression in cancer patients: Pathogenesis, implications and treatment (Review)," *Oncology Letters*, vol. 9, no. 4, pp. 1509–1514, Apr. 2015, ISSN: 1792-1074. DOI: 10.3892/ol.2015.2944. pmid: 25788991.

[3] R. Upthegrove, S. Marwaha, and M. Birchwood, "Depression and Schizophrenia: Cause, Consequence or Trans-diagnostic Issue?" *Schizophrenia Bulletin*, vol. 43, no. 2, sbw097, Jul. 2016, ISSN: 0586-7614. DOI: 10.1093/schbul/sbw097.

[4] I. Sokoreli, J. J. de Vries, S. C. Pauws, and E. W. Steyerberg, "Depression and anxiety as predictors of mortality among heart failure patients: Systematic review and meta-analysis," *Heart Failure Reviews*, vol. 21, no. 1, pp. 49–63, Jan. 2016, ISSN: 15737322. DOI: 10.1007/s10741-015-9517-4.

[5] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, Sep. 2001, ISSN: 0884-8734, 1525-1497.

[6] M. Finegan, N. Firth, C. Wojnarowski, and J. Delgadillo, "Associations between socioeconomic status and psychological therapy outcomes: A systematic review and meta-analysis," *Depression and Anxiety*, vol. 35, no. 6, pp. 560–573, Jun. 2018, ISSN: 1091-4269. DOI: 10.1002/da.22765.

[7] M. J. Huibers, Z. D. Cohen, L. H. Lemmens, *et al.*, "Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach," *PLoS ONE*, vol. 10, no. 11, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0140771.

[8] B. E. Wampold, "How important are the common factors in psychotherapy? An update," *World Psychiatry*, vol. 14, no. 3, pp. 270–277, 2015, ISSN: 20515545. DOI: 10.1002/wps.20238.

[9] T. Anderson, B. M. Ogles, C. L. Patterson, M. J. Lambert, and D. A. Vermeersch, "Therapist effects: Facilitative interpersonal skills as a predictor of therapist success," *Journal of Clinical Psychology*, vol. 65, no. 7, pp. 755–768, Jul. 2009, ISSN: 00219762. DOI: 10.1002/jclp.20583.

[10] P. Odyniec, T. Probst, J. Margraf, and U. Willutzki, "Psychotherapist trainees' professional self-doubt and negative personal reaction: Changes during cognitive behavioral therapy and association with patient progress," *Psychotherapy Research*, vol. 29, no. 1, pp. 123–138, 2019, ISSN: 14684381. DOI: 10.1080/10503307.2017.1315464.

[11] R. Yonatan-Leus, O. Tishby, G. Shefler, and H. Wiseman, "Therapists' honesty, humor styles, playfulness, and creativity as outcome predictors: A retrospective study of the therapist effect," *Psychotherapy Research*, vol. 28, no. 5, pp. 793–802, Sep. 2018, ISSN: 1050-3307. DOI: 10.1080/10503307.2017.1292067.

[12] L. Robinson, J. Delgadillo, and S. Kellett, "The dose-response effect in routinely delivered psychological therapies: A systematic review," *Psychotherapy Research*, vol. 30, no. 1, pp. 79–96, Jan. 2020, ISSN: 1050-3307. DOI: 10.1080/10503307.2019.1566676.

[13] J. I. L. Beard and J. Delgadillo, "Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: A systematic review and meta-analysis," *Depression and Anxiety*, vol. 36, no. 9, pp. 866–878, 2019, ISSN: 15206394. DOI: 10.1002/da.22931.

[14] B. Bei, L. D. Asarnow, A. Krystal, J. D. Edinger, D. J. Buysse, and R. Manber, "Treating insomnia in depression: Insomnia related factors predict long-term depression trajectories.," *Journal of Consulting and Clinical Psychology*, vol. 86, no. 3, pp. 282–293, 2018, ISSN: 19392117. DOI: 10.1037/ccp0000282. pmid: 29504795.

[15] D. M. Clark, "Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience," *International Review of Psychiatry*, vol. 23, no. 4, pp. 318–327, Aug. 2011, ISSN: 0954-0261, 1369-1627. DOI: 10.3109/09540261.2011.606803.

[16] T. Smith, M. Noble, S. Noble, G. Wright, D. McLennan, E. Plunkett, *et al.*, "The english indices of deprivation 2015," *London: Department for Communities and Local Government*, pp. 1–94, 2015.

[17] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, May 2006, ISSN: 00039926. DOI: 10.1001/archinte.166.10.1092. pmid: 16717171.

[18] J. C. Mundt, I. M. Marks, M. K. Shear, and J. H. Greist, "The Work and Social Adjustment Scale: A simple measure of impairment in functioning," *British Journal of Psychiatry*, vol. 180, pp. 461–464, MAY 2002, ISSN: 00071250. DOI: 10.1192/bjp.180.5.461. pmid: 11983645.

[19] N. S. Jacobson and P. Truax, "Clinical significance: A statistical approach to defining meaningful change in psychotherapy research.," *Journal of Consulting and Clinical Psychology*, vol. 59, no. 1, pp. 12–19, 1991, ISSN: 1939-2117, 0022-006X. DOI: 10.1037/0022-006X.59.1.12.

[20] H. Hentschke, *Measures of effect size toolbox*, Feb. 5, 2021. [Online]. Available: https://github.com/hhentschke/measures-of-effect-size-toolbox.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer, 2013, ISBN: 978-0-387-21606-5.

[22] A. Malpass, A. Shaw, D. Kessler, and D. Sharp, "Concordance between PHQ-9 scores and patients' experiences of depression: A mixed methods study," *British Journal of General Practice*, vol. 60, no. 575, e231–e238, Jun. 1, 2010, ISSN: 0960-1643, 1478-5242. DOI: 10.3399/bjgp10X502119.

[23] A. Khan, J. Faucett, P. Lichtenberg, I. Kirsch, and W. A. Brown, "A Systematic Review of Comparative Efficacy of Treatments and Controls for Depression," *PLoS ONE*, vol. 7, no. 7, C. Holscher, Ed., e41778, Jul. 30, 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0041778.

[24] J. Delgadillo, O. Moreea, and W. Lutz, "Different people respond differently to therapy: A demonstration using patient profiling and risk stratification," *Behaviour Research and Therapy*, vol. 79, pp. 15–22, Apr. 2016, ISSN: 00057967. DOI: 10.1016/j.brat.2016.02.003.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Learning Adapters for Code-Switching Speech Recognition

Chun-Yi He and Jen-Tzung Chien
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Multilingual code-switching speech recognition has been an emerging research direction in real-world applications since most of speakers are bilingual or multilingual. A code-switching sentence is the mixing of two or more languages especially within the same sentence. It is crucial to work out a multilingual speech recognition via a code-switching scheme based on a parameter-efficient learning by utilizing a pre-trained encoder. Using this scheme, it is essential to identify the languages within a single spoken utterance for code-switching speech recognition. However, collecting speech data from monolingual language is easier than collecting code-switching speech in multiple languages. This study develops a new code-switching Mandarin-English speech recognition by utilizing a large-scale pre-trained backbone model containing 53 single-code languages. The fine-tuning of backbone model is implemented by introducing the controllable language or task adapters and incorporating small number of Mandarin-English code-switching speech where the backbone model is frozen and only individual adapters for Mandarin and English are estimated. A limited amount of controllable parameters can be sufficiently calculated. Experiments on code-switching speech recognition for Taiwanese Mandarin and English shows the merit of the proposed method.**

I. Introduction

In the task of speech recognition, it is often faced with speech processing over two or more languages at the same time because most speakers are bilingual or multilingual. It has been demanding to develop multilingual automatic speech recognition (ASR) for a variety of applications. Generally speaking, an utterance or a sentence containing more than one language is seen as code-switching (CS) or code-mixing speech. There have been a number of works explored to handle the challenging task in a code-switching speech recognition system [1], [2]. A traditional ASR method was developed with three components consisting of acoustic model, word lexicon and language model [3], [4], [5], [6]. Acoustic model converts the input audio signal into a sequence of phonemes, word lexicon converts the phonemes generated by acoustic model into words, and language model combines the individual words into sentences. Modern ASR system is built via an end-to-end (E2E) model which requires only a single composite neural network. This network is exploited to optimize the probability of an output word string given an input speech sequence for sequence-to-sequence learning. E2E model has become a main stream method for ASR. E2E model can directly map an input speech signal to an output sequence of phones, characters or words. There are three categories of E2E models. The first one is the method of using the encoder and the connectionist temporal classification (CTC) [7], [8]. The second one is the solution of using an encoder-decoder framework with attention [9], [10]. The third one is the method based on the recurrent neural network (RNN) transducer which jointly estimates encoder and prediction networks [11]. These E2E ASR models have achieved remarkable performance in monolingual task for Mandarin [12], English [13], German. Recently, it was even more challenging to carry out multilingual multispeaker ASR tasks [14], [15] based on E2E setting which have obtained state-of-the-art performance.

In general, the challenge that code-switching ASR will encounter is the lack of code-switching speech data. A reasonable solution to handle insufficient code-switching data is to conduct the transfer learning [16], [17] based on the pre-trained models by using abundant speech data from monolingual Mandarin and English. In addition, a practical approach to handle multilingual ASR was to add the language-specific vectors to multilingual model to enhance the performance in individual languages [14], [18]. A similar strategy was also proposed in [19]. In these works, there was a dedicated layer for a specific language and a shared layer for different languages. The output ratio of two layers was determined through the gating decision process. The language-specific information was only captured in the encoder or from the output layer of a deep neural network. In addition to characterizing the language-specific information, another strategy was proposed to implement the multi-task learning by jointly training the ASR acoustic model and the language identifier [20]. Further, in code-switching ASR, it is difficult to identify target language corresponding to a speech segment [21] since different languages have similar pronunciations. To cope with this issue, language identification could be incorporated into a bilingual ASR model. In [22], the language identification (LID) tags was introduced in an augmented dictionary by adding the LID tokens "CH" and "EN" as the outputs of a model for Chinese and English, respectively. When a code-switching point in a speech signal was encountered, the decoder first predicted which language it was sensing, and subsequently recognized the word tokens corresponding to the selected language. Summarizing the two approaches mentioned above, the first one is to solve the problem of insufficient code-switching data by using monolingual language data training, and the second is to solve the problem of language identification errors by applying the multi-task learning or identifying the language specific information through LID tokens. This paper presents a parameter-efficient learning for bilingual Mandarin and English ASR under an E2E framework.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

The rest of the paper is organized as follows. In Section II, we present the issue of data sparseness in implementation of code-switching ASR system. In Section III, the method which incorporates LID into an E2E model using CTC with the proposed language-specific adapters is addressed. The experiments on bilingual Mandarin and English ASR are evaluated and analyzed in Section IV. The conclusions drawn from this study are given in Section V.

II. Code-Switching Speech Recognition

A traditional way to code-switching ASR [23], [24] was developed through language boundary detection and identification. In the procedure, the languages in various speech frames were first identified. The speech frames were then recognized by using the corresponding monolingual speech recognizer. However, this method was highly affected by the performance of boundary detection and language identification. In [25], an improved method based on deep neural network was proposed. More intuitively, the end-to-end methods were developed to relax the requirement of language boundary detection in a code-switching ASR system [26], [27]. In [28], it was proposed to use multiple encoders to capture the properties of individual languages. The acoustic representation of individual languages could be improved. The method in [29] presented a solution to accent-adaptive ASR by jointly learn the accent ID and the acoustic model so as to enhance the performance of accent speech recognition. This paper develops an E2E code-switching method where the parameter-efficient learning based on language-specific adapters is designed and implemented.

In particular, this study aims to handle the issue of data sparseness in a code-switching speech recognition system in presence of Mandarin and English speech with Taiwan accent. The issue is severe because the bilingual speech data with Taiwan accent are limited. A parameter-efficient learning is developed to build a low-resource bilingual ASR. A strong backbone model based on wav2vec architecture [30] is utilized in a transfer learning task where only a limited number of trainable parameters due to two language-dependent adapters are estimated for bilingual speech recognition. Language or task adapters could be pre-trained from monolingual languages with sufficient data. Some shared information in both languages is learned in a code-switching ASR. The requirement for fine-tuning the pre-trained model using code-switching speech data with target accent can be significantly reduced. Basically, using adapters allows efficient parameter tuning from the pre-trained models. Compared with fine-tuning a large-scaled pre-trained model, the lightweight adapters for both languages could be flexibly estimated and utilized to adjust the pre-trained model to a target task. In an E2E implementation, this paper employs the wav2vec2.0 [30] as a pre-trained encoder and incorporates the language adapters into a unified set of encoders and decoders in different layers for code-switching speech recognition. Experiments on the CS corpus ASCEND and NTUT-AB01 [31] show that the proposed model can distinguish the mixed languages while solving the training data scarcity problem.

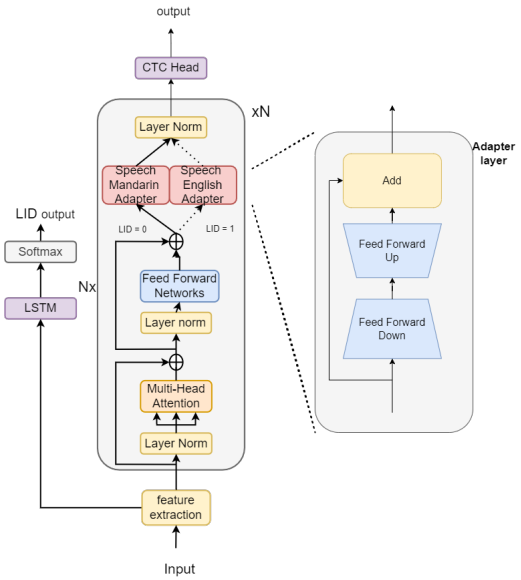


Fig. 1. An overview of the proposed method for code-switching speech recognition. The input speech frames are first passed through a long short-term memory (LSTM) and a softmax layer to find LID outputs for individual frames. The encoder accordingly switches between the corresponding adapters based on LID outputs.

III. Learning Adapters for Code-Switching ASR

Figure 1 overviews an architecture of adding language-specific adapters on a pre-trained backbone model. Language identifier for individual frames is implemented. LID outputs are used to perform language switch in encoder layers. The connectionist temporal classification (CTC) head is employed before finding output sequence. The detailed description of individual components in this work is provided.

A. Connectionist Temporal Classification

In this study, the connectionist temporal classification (CTC) is employed to handle time alignment between input speech signal and output word string in a sequence-to-sequence learning task. CTC network basically contains a linear layer and a softmax output layer with a dimension corresponding to the dictionary size. The dictionary is augmented with an additional "blank" token. In general, CTC calculates the probability distribution of the possible output sequence. The goal of CTC network is to maximize the conditional probability of finding the correct word labels by given the input speech frames. However, it is not always that each frame has a corresponding word label. The blank token is introduced to reflect those silence frames. Let $\mathbf{x} = \{\mathbf{x}_t\}$ denote the input speech frames, $l = \{l_t\}$ denote the output word labels, π denote the CTC path, Θ denote the network parameters, $B^{-1}(l)$ indicates all possible CTC paths from l extension. The CTC loss function is defined as the sum of the negative log probabilities of correct labels given by speech frames $\mathbf{x} = \{\mathbf{x}_t\}$ as

$$\mathcal{L}_{\text{CTC}} = -\log p_\Theta(\mathbf{l}|\mathbf{x}) = -\log \sum_{\pi \in B^{-1}(\mathbf{l})} p_\Theta(\pi|\mathbf{x}). \quad (1)$$

Based on the assumption of conditional independence for output units along T frames, the probability $p_\Theta(\pi|\mathbf{x})$ can be decomposed to a product of conditional probabilities from each time step t as follows

$$p_\Theta(\pi|\mathbf{x}) = \prod_{t=1}^{T} p_\Theta\left(\pi_t|\mathbf{x}_t\right) \tag{2}$$

where $\mathbf{x}_t$ is the input speech frame at time t, π_t is the output unit at time t, and T is the sequence length in frames.

B. Language Identification

This study implements a language identifier in sequence-to-sequence learning for bilingual ASR based on the connectionist temporal classification. In the proposed code-switching ASR, the problem encountered is that similar-sounding words in different languages are likely to be misrecognized. To tackle this issue, a language identification (LID) task is considered. In the implementation, the multilingual pre-trained models are utilized to provide language-related information through the pre-trained speech representations. Therefore, the code-switching ASR is handled by identifying LID for individual frames and then fulfilling a switching scheme between language-related adapters. In [32], [33], the utterance-level LID was proposed. But, in code-switching speech, there will be more than one language in a single utterance. It is necessary to carry out frame-level LID for code-switching ASR. However, the length of speech frames is usually short. It is still difficult to assure a reliable ASR performance based on frame-level LID, Nevertheless, this study integrates speech recognition and language classification where the frame-level LID is merged in a multi-task learning framework. The resulting LID accuracy is likely enhanced to improve learning representation in a joint optimization. For the task of LID, a recurrent neural network using bidirectional LSTM (BLSTM) [34], [35] is merged. Here, BLSTM and CTC models utilize the shared encoder. BLSTM is employed to compute the predicted LID outputs. Given the one-hot LID labels, the cross-entropy loss $\mathcal{L}_{\text{CE}}$ is calculated in an optimization procedure.

C. Adapter Learning

Importantly, this study incorporates individual adapter networks for Mandarin and English for code-switching speech recognition. Adapter network basically consists of a residual layer and two feedforward networks which are connected in a shape which is similar to extracting the bottleneck features. As shown by the shape in right-hand-side of Figure 1, the dimension of intermediate features between two feedforward networks is considerably reduced. A large number of parameters from a backbone model which were originally pre-trained are shared among different tasks, and only a small number of parameters from language-specific adapters are updated according to different objectives or tasks. Using the adapters as an alternative to fine-tuning all parameters of the pre-trained model has the advantage of retaining a large amount of knowledge of the pre-trained model and learning the schemes for specific tasks. When using adapters, only the parameters of the augmented module with adapters are changed, and all the other pre-trained parameters remain unchanged. Therefore, only a small amount of data are required to allow the efficiency in model convergence, and at the same time, avoid the weakness in model fitting due to the limited data. For the application of multilingual speech recognition [36], the adapters corresponding to different languages can be arranged. Adding adapters is also helpful for low-resource language modeling in multilingual machine translation [37]. A shared model could be learned from multiple language resources, and the adapters for low-resource languages are trained on top of the shared model. Different from the previous methods, this paper deals with the Mandarin-English code-switching speech recognition. A parameter-efficient learning based on adapters is proposed. To handle the spoken languages of Mandarin and English, a backbone pre-trained model from a number of different languages is utilized. The adapter-based code-switching speech recognition is built by training the adapters separately from individual languages. The multi-task learning for language identification and speech recognition is performed as detailed in what follows. First, each individual frame in an audio signal is classified as silence, Mandarin or English by using bidirectional LSTM with a linear classifier layer. In the encoder layer based on transformer [38], [39], [40], the audio signal is passed through the feature extraction layer to obtain latent features. These features are then passed through the multi-head attention layer and feedforward layer. Layer normalization is always performed. After that, depending on the LID label of an audio frame, the corresponding adapter is applied to find the word token based on CTC scheme using a CTC head. In case of detecting the input frame as silence, the outputs of transformer layer are not passed through any adapter and will be directly forwarding to CTC module for code-switching speech recognition. The multi-task learning is performed according to the following hybrid objectives

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{CTC}} + \lambda\mathcal{L}_{\text{CE}} \tag{3}$$

where λ is a regularization term to control the contribution of cross entropy loss due to language identification when compared with CTC loss due to speech recognition.

IV. Experiments

The proposed method was evaluated for a code-switching speech recognition under a low-resource setting where only a limited amount of code-switching Mandarin-English speech data with Taiwan accent was incorporated.

A. Experimental Data

In the following experiments, the Mandarin-English code-switching speech datasets including ASCEND [31] and NTUT-AB01 (a public-domain dataset obtained from the Association for Computational Linguistics and Chinese Language Processing (ACLCLP)) were used. ASCEND was a public-domain dataset containing Mandarin-English code-switching conversational speech with China accent. ASCEND consisted

of 10.62 hours of clean speech, collected from 23 bilingual speakers of Mandarin and English from Mainland China. The whole utterances in ASCEND were partitioned into training set, validation set and test set with the proportionals approximately 0.8, 0.1 and 0.1, respectively. In the partition, the total number of audio data for each gender were balanced. Table I summarizes the statistics of training, validation and test sets in ASCEND dataset.

NTUT-AB01 contained Mandarin, English, and Mandarin-English mixed speech which were collected as the reading speech in Taiwan. The Mandarin content was sampled from Chinese books, the English content was re-recorded from the ARCTIC corpus, and the code-switching speech was collected from common Mandarin-English mixed sentences collected from internet. This dataset could be also used to implement a code-switching speech synthesis system. There were 500, 105 and 376 minutes of speech in Taiwan accent for Mandarin, English and mixed Mandarin-English speech data.

TABLE I
SUMMARY OF TRAIN, VALIDATION, AND TEST DATA IN ASCEND.

Gender	Utterance				Duration(hr)			
	Train	Val	Test	Total	Train	Val	Test	Total
Female	4,591	484	861	5,936	4.04	0.46	0.48	4.98
Male	5,278	646	454	6,378	4.74	0.46	0.44	5.63

B. Experimental Settings

Before training the model, the dictionary based on the collected datasets was generated while some unnecessary characters and symbols were removed from the transcription data. This dictionary was used to expand the pre-trained tokenizer which was attached to the model. To enhance the model robustness, the data augmentation using SpecAugment [41] was applied by using the hyperparameter settings including 0.004 for the frequency masking probability, 0.065 for the time masking probability and 2 for the time masking length. The sampling rate of audio signals in ASCEND and NTUT-AB01 datasets was 16K Hz and 44K Hz, respectively, and the NTUT-AB01 dataset was accordingly downsampled as 16K Hz. Forced alignment was applied to automatically align the spoken recordings with the transcripts at the segment level. These schemes significantly reduced the time required to prepare data for speech analysis. The frame-level LID was performed to classify each frame as silent, Mandarin or English where each frame was windowed as 20ms. The frame labels of LID in training phase were obtained by forced alignment.

For the code-switching ASR task, this paper used the pre-trained XLSR-53 [42] which was known as a large-scaled pre-trained model from 53 languages definitely including Mandarin and English. Notably, those multilingual training utterances were *not* mixed or code-switching. The wav2vec 2.0 BASE and LARGE [30] had 95M and 317M parameters, respectively, which were pre-trained from 960 hours on the LibriSpeech [43]. For the language identification task, the module based on bidirectional LSTM was used. In the experiments, all of the models in the comparison were investigated through the HuggingFaces [44] framework.

C. Results on Language Identification

Language identification was configured by a bidirectional LSTM added by a single classifier layer and the fully connected (FC) feedforward classification network. FC network contained 4 hidden layers with the architecture 1024-2048-2048-512. Bidirectional LSTM (BLSTM) network had 2 hidden layers with 1024 nodes per direction in each layer. Table II reports the frame-level performance of LID model based on the extracted features using three backbone models wav2vec BASE, wav2vec LARGE and XLSR where FC and BLSTM are compared in case of using small amount of NTUT-AB01 training data. Test data in NTUT-AB01 are used. The results show that using BLSTM for different feature extractors achieves better accuracy than using FC. BLSTM is superior to FC when modeling time-series data. In addition, it is found that there are no significant gap in accuracy among three backbone models. Although the XLSR pre-trained model contains the knowledge of English and Mandarin, and the wav2vec 2.0 Base and Large pre-trained models contain only English, all these models can distinguish different languages in each sentence. It is because that these three feature extractors basically capture the characteristics of sound and have nothing to do with the labels of language ID. Table III further shows the same comparison but the training data from both ASCEND and NTUT-AB01 are used. LID is improved by considering increased amount of domain-related data.

TABLE II
COMPARISON OF FRAME-LEVEL ACCURACY (%) OF LID USING DIFFERENT METHODS. NTUT-AB01 IS USED AS TRAINING DATA.

Method	Accuracy%
Base + FC	66
Base + BLSTM	77
Large + FC	66
Large + BLSTM	77
XLSR + FC	67
XLSR + BLSTM	78

TABLE III
COMPARISON OF FRAME-LEVEL ACCURACY (%) OF LID USING DIFFERENT METHODS. BOTH ASCEND AND NTUT-AB01 ARE USED AS TRAINING DATA.

Method	Accuracy%
Base + FC	70
Base + BLSTM	81
Large + FC	70
Large + BLSTM	82
XLSR + FC	72
XLSR + BLSTM	82

D. Results on Code-Switching Speech Recognition

This study works on Mandarin-English Code-Switching ASR where the Mandarin and English utterances in ASEND and NTUT-AB01 are used. The character error rate (CER) and

word error rate (WER) were used as the metrics to evaluate ASR in presence of Chinese and English, respectively. For code-switching ASR, the mixed error rate (MER) is used in the evaluation where CER for Mandarin and WER for English are merged. In this set of experiments, XLSR trained from common voice [45] is used as the pre-trained backbone model since both ASCEND and NTUT-AB01 contained most of speech data in Mandarin rather than the backbone model wav2vec which was trained from English.

TABLE IV
ABLATION STUDY ON CODE-SWITCHING ASR WITH FINE-TUNING AND PARAMETER-EFFICIENT LEARNING USING LANGUAGE IDENTIFIER AND ADAPTER IN TERMS OF MERS AND PERCENTAGES OF CONTROLLED PARAMETERS. NTUT-AB01 IS USED AS TRAINING DATA.

Method	MER	Param
XLSR + FT	30.7%	100%
XLSR + adapter	30.0%	0.67%
XLSR + LID + adapter	28.7%	0.80%

TABLE V
ABLATION STUDY ON CODE-SWITCHING ASR WITH FINE-TUNING AND PARAMETER-EFFICIENT LEARNING USING LANGUAGE IDENTIFIER AND ADAPTER IN TERMS OF MERS AND PERCENTAGES OF CONTROLLED PARAMETERS. BOTH ASCEND AND NTUT-AB01 ARE USED AS TRAINING DATA.

Method	MER	Param
XLSR + FT	30.1%	100%
XLSR + adapter	29.2%	0.64%
XLSR + LID + adapter	28.1%	0.78%

Tables IV and V show the ablation studies on code-switching ASR using MER for different methods where the effects on parameter-fining (FT) and parameter-efficient learning with language identifier and adaptation [46], [47] are evaluated. The percentage of controlled parameters is investigated. The proposed hybrid method using language adapter combined with language identifier is compared with baseline fine-tuning method using XLSR and the baseline XLSR with adapter. This set of experiments would like to see how multi-task learning is beneficial to train an adaptive encoder via adapter to find correct text via correct language. The results show that the proposed method does improve speech recognition performance. Using LID as an auxiliary task [20] allows the model to likely catch the time frames to switch languages, From the analysis of results, it is found that the baseline system is prone to mis-identify language at the points where code-switching happens. The resulting prediction of an utterance become more likely an utterance with either monolingual Mandarin or English and less likely an code-switching utterance. Therefore, using LID is helpful to deal with this issue by coming out a distribution of output tokens which is more likely code-switched. Using multiple encoders to learn information from different languages is helpful for speech recognition. Different from [28] using two encoders for two languages, this study arranges two adapters in a single encoder where the number of controlled parameters is significantly reduced. Compared with the results using parameter-fining for all parameters, the proposed parameter-efficient learning with hybrid LID and adapter obtains lower mixed error rates in presence of very small percentage of controlled parameters.

V. CONCLUSIONS

This paper presented a parameter-efficient learning approach to handle Mandarin-English code-switching speech recognition by learning the language identifier for different languages and the language-specific adapters based on a pre-trained backbone model. The end-to-end speech recognition was implemented by using connectionist temporal classification. A multi-task learning with bilingual CTC loss and frame-level LID loss was implemented. The pre-trained model from a large amount of training data with a variety of languages was utilized to reduce the mis-prediction of code-switching points. A single encoder with augmentation of two small-scaled adapters was efficiently implemented. A shared backbone model was used to represent different languages. Only adapters were individual for languages. Experimental results on Mandarin-English code-switching speech recognition using ASCEND and NTUT-AB01 data demonstrated the effectiveness of the proposed method. Future study will be extended by improving the accuracy of language identification which considerably affect the performance of adapter-based code-switching speech recognition. The code-switching speech recognition over Mandarin, Taiwanese and English will be investigated.

REFERENCES

[1] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Chng Eng Siong, Tanja Schultz, and Haizhou Li, "A first speech recognition system for Mandarin-English code-switch conversational speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4889–4892.

[2] Ying Li and Pascale Fung, "Code switching language model with translation constraint for mixed language speech recognition," in *Proc. of International Conference on Computational Linguistics*, 2012, pp. 1671–1680.

[3] Xiaodong Cui, Jing Huang, and Jen-Tzung Chien, "Multi-view and multi-objective semi-supervised learning for hmm-based automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1923–1935, 2012.

[4] Dong Yu, Geoffrey Hinton, Nelson Morgan, Jen-Tzung Chien, and Shigeki Sagayama, "Introduction to the special section on deep learning for speech and language processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 4–6, 2012.

[5] George Saon and Jen-Tzung Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.

[6] Shinji Watanabe and Jen-Tzung Chien, *Bayesian Speech and Language Processing*, Cambridge University Press, 2015.

[7] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of International Conference on Machine Learning*, 2006, pp. 369–376.

[8] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. of International Conference on Machine Learning*, 2014, pp. 1764–1772.

[9] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960–4964, 2016.

[10] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[11] Alex Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton, "Speech recognition with deep recurrent neural networks," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

[12] Fu-Hao Yu, Kuan-Yu Chen, and Keda Lu, "Non-autoregressive ASR modeling using pre-trained language models for chinese speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1474–1482, 2022.

[13] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[14] Ngoc-Quan Pham, Alexander H. Waibel, and Jan Niehues, "Adaptive multilingual speech recognition with pretrained models," in *Proc. of Annual Conference of International Speech Communication Association*, 2022.

[15] Midia Yousefi and John H.L. Hanse, "Speaker conditioning of acoustic models using affine transformation for multi-speaker speech recognition," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 283–288.

[16] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7319–7323.

[17] Jui Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7304–7308.

[18] Bo Li, Yu Zhang, Tara N. Sainath, Yonghui Wu, and William Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5621–5625.

[19] Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat, "Share or not? learning to schedule language-specific capacity for multilingual translation," in *Proc. of International Conference on Learning Representations*, 2021.

[20] Zhiping Zeng, Yerbolat Khassanov, Haihua Xu Van Tung Pham, Eng Siong Chng, and Haizhou Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2019.

[21] Jen-Tzung Chien and Chuang-Hua Chueh, "Topic-based hierarchical segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 55–66, 2012.

[22] Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4919–4923.

[23] Marc A Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31, 1996.

[24] Koena Ronny Mabokela, Madimetja Jonas D. Manamela, and Mabu Manaileng, "Modeling code-switching speech on under-resourced languages for language identification," in *Proc. of Workshop on Spoken Language Technologies for Under-resourced Languages*, 2014.

[25] Hardik B. Sailor, Kiran Praveen, Vikas Agrawal, Abhinav Jain, and Abhishek Pandey, "SRI-B end-to-end system for multilingual and code-switching ASR challenges for low resource indian languages," in *Proc. of Annual Conference of International Speech Communication Association*, 2021.

[26] Ondrej Klejch, Electra Wallington, and Peter Bell, "The CSTR system for multilingual and code-switching ASR challenges for low resource indian languages.," in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2881–2885.

[27] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung, "Meta-transfer learning for code-switched speech recognition," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3770–3776.

[28] Xinyuan Zhou, Emre Yılmaz, Yanhua Long, Yijie Li, and Haizhou Li, "Multi-encoder-decoder transformer for code-switching speech recognition," *Proc. of Annual Conference of International Speech Communication Association*, pp. 1042–1046, 2020.

[29] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark A. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2018.

[30] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[31] Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, and Pascale Fung, "ASCEND: A spontaneous chinese-english dataset for code-switching in multi-turn conversation," in *Proc. of International Conference on Language Resources and Evaluation*, 2021.

[32] Weicheng Cai, Danwei Cai, Shen Huang, and Ming Li, "Utterance-level end-to-end language identification using attention-based CNN-BLSTM," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5991–5995, 2019.

[33] Ma Jin, Yan Song, Ian Mcloughlin, Wu Guo, and Lirong Dai, "End-to-end language identification using high-order utterance representation with bilinear pooling," in *Proc. of Annual Conference of International Speech Communication Association*, 2017.

[34] Jen-Tzung Chien and Yuan-Chu Ku, "Bayesian recurrent neural network for language modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 2, pp. 361–374, 2016.

[35] Jen-Tzung Chien and Chun-Wei Wang, "Hierarchical and self-attended sequence autoencoder," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4975–4986, 2022.

[36] Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," *Proc. of Annual Conference of International Speech Communication Association*, pp. 2130–2134, 2019.

[37] Ankur Bapna and Orhan Firat, "Simple, scalable adaptation for neural machine translation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 1538–1548.

[38] Chi-Hang Leong, Yu-Han Huang, and Jen-Tzung Chien, "Online Compressive Transformer for End-to-End Speech Recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2082–2086.

[39] Jen-Tzung Chien and Yu-Han Huang, "Bayesian Transformer Using Disentangled Mask Attention," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 1761–1765.

[40] Jen-Tzung Chien and Chia-Wei Tsao, "Graph evolving and embedding in transformer," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 538–545.

[41] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2019, pp. 2613–2617.

[42] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdel rahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2020.

[43] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[45] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. of Conference on Language Resources and Evaluation*, 2020, pp. 4211–4215.

[46] Jen-Tzung Chien and Chih-Hsien Huang, "Bayesian learning of speech duration models," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 558–567, 2003.

[47] Man-Wai Mak and Jen-Tzung Chien, *Machine Learning for Speaker Recognition*, Cambridge University Press, 2020.

FID-RPRGAN-VC: Fréchet Inception Distance Loss based Region-wise Position Normalized Relativistic GAN for Non-Parallel Voice Conversion

Sandipan Dhar*, MD. Tousin Akhter*, Padmanabha Banerjee†, Nanda Dulal Jana*, Swagatam Das‡
* National Institute of Technology Durgapur, India
E-mail: sd.19cs1101@phd.nitdgp.ac.in, mdtousinakhter@gmail.com, nandadulal@cse.nitdgp.ac.in
† Jalpaiguri Government Engineering College, India
E-mail: pbanerjee0801@gmail.com
‡ Indian Statistical Institute, Kolkata, India
E-mail: swagatamdas19@yahoo.co.in

***Abstract*—Voice conversion (VC) is the speech-to-speech (STS) synthesis process that converts the vocal identity of a source speaker to a target speaker by keeping the linguistic content unaltered. In recent years, VC research has been explored using generative adversarial network (GAN) models. However, a substantial difference exists between the real and the state-of-the-art (SOTA) VC model-generated speech samples as far as naturalness is concerned. This work proposes an improved GAN model for non-parallel VC to enhance the naturalness of the generated speech samples. The improved GAN model is integrated with a region-wise positional normalization technique in the generator, a relativistic mechanism-based discriminator, and a Fréchet inception distance (FID) based loss function. We tested the proposed model on VCC 2018, CMU Arctic, and a dysarthric speech dataset. The experimental results revealed the superiority of the proposed FID-RPRGAN-VC model over the SOTA MaskCycleGAN-VC model.**

***Index Terms*—Generative Adversarial Network (GAN), Voice Conversion (VC), Region Wise Positional Normalization Technique, Relativistic Mechanism, Fréchet Inception Distance (FID) loss.**

I. Introduction

Voice conversion (VC) [1] is a speech synthesis process that converts the vocal texture of a source speaker similar to a target speaker by keeping the linguistic content unchanged. Recently, the domain of VC has been majorly explored using deep learning (DL) models. Among the DL models, generative adversarial networks (GANs) [2] have been widely used for the VC task due to their capability of performing efficient style transfer of the target features on the source content.

Based on the linguistic contents of the speech datasets, a VC system can be categorized as a parallel or nonparallel VC [1]. In parallel VC, the speech content of both the source and target speakers are considered as same. Whereas for non-parallel VC, dissimilar speech contents are considered in the VC process [3]. On the other hand, based on the type of mapping from the source-to-target domain, it is divided into the following categories: one-to-one, one-to-many, many-to-many, etc. VC. In the early stage of the developments of VC, statistical models such as gaussian mixture models (GMMs) [4], phonetic posteriorgram (PPG) based models [5], non-negative matrix factorization based models [6], vector quantization (VQ) based models [7] etc., were widely used. Eventually, due to the progressive developments in the field of artificial intelligence (AI), the deep learning (DL) models appeared as a better alternative to the statistical models [8].

From the preceding works of GAN-based VC [1], it is observed that the successive models of CycleGAN-VC [8]–[10] have the potential ability to perform the VC task. Although, most of these models rely on multiple speech features for the VC process [11], [12]. This limits the ability of the VC models because of multi-feature dependency of a single VC model. In this context, MaskCycleGAN-VC [13] showed notable performance improvement by considering mel-spectrogram as the only input feature. Although, there is still a significant difference between the real and the existing VC model-generated speech samples in terms of speaker similarity and speech quality. To generate more natural-sounding speech, in this work, we put forth an improved GAN model with region-wise positional normalization technique in the generator, a relativistic mechanism based discriminator, and a Fréchet inception distance (FID) based loss function. The proposed model is named as FID loss-based region-wise position normalized relativistic GAN-VC (i.e., FID-RPRGAN-VC).

In our proposed FID-RPRGAN-VC model, the objective of using the region-wise positional normalization technique is to compute the normalization statistics at all the spatial positions independently by reducing the internal covariate shift [14], [15]. The internal covariate shift is the phenomenon of significant change in the distribution of the layers of a neural network model due to the change in network parameters during training. Most of the preceding works on GAN-based VC [8], [9], [13] have been explored using the instance normalization technique. However, based on the literature

of [14], [15], it is well observed that region normalization and positional normalization provide superior performance over instance normalization (IN) and other normalization techniques in image synthesis task [14], [15]. Moreover, an improvised version of gated linear unit (GLU) called gaussian error gated linear unit (GEGLU) [16], is incorporated into the proposed FID-RPRGAN-VC model to improve the performance of the model. Furthermore, the reason behind inclusion of a relativistic discriminator is to trace the similarity between the latent representation of real and generated mel-spectrogram. In addition to that, FID [17], [18] metric is used in the proposed model as a loss function for obtaining better quality generated speech samples. An Inception-v3 [19] architecture is incorporated to the proposed model to obtain the latent representation of real and generated mel-spectrogram for computing the FID loss.

The performance evaluation of the proposed FID-RPRGAN-VC model is conducted on VCC 2018 [20], and CMU Arctic dataset [21] using objective and subjective evaluation metrics. In addition to that the performance of the proposed model is also compared with MaskCycleGAN-VC for dysarthric-to-normal speech conversion, based on a dysarthric speech dataset called EasyCall corpus [22]. Based on the obtained results, it is observed that the proposed FID-RPRGAN-VC model performed superior to the state-of-the-art (SOTA) MaskCycleGAN-VC model. Our work is specifically compared with GAN-based VC model. The objective of considering MaskCycleGAN-VC for the performance comparison is that, among the other CycleGAN based VC models Mask-CycleGAN-VC has shown the most significant results [13].

The rest of the paper is organized as follows. Section II presents the proposed FID-RPRGAN-VC model in detail. Section III describes the details of the datasets and the experimental setups. In Section IV, the results obtained from the objective and subjective evaluations are discussed. Finally, Section V provides the conclusion and gives good scope for future research.

II. Proposed Model

We provide a detailed description of the framework of the proposed FID-RPRGAN-VC model and the loss functions described here in detail. As shown in Fig. 1, $\mathrm{x} \in \mathrm{X}$ is the mel-spectrogram of the source speaker and m is the mask such that $\mathrm{x}, \mathrm{m} \in \mathbb{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ (where C, H, W are channel, height, and width respectively) used for obtaining the masked mel-spectrogram $\mathrm{x}^{'}$. The mask m contains values zero and one, where zero indicates the masked region (the black region in Fig. 1) and one indicates the unmasked region (the white region in Fig. 1). The masked mel-spectrogram $\mathrm{x}^{'}$ of the source speaker is obtained by element-wise multiplication between x and m (i.e. $\mathrm{x}^{'} = \mathrm{x} \odot \mathrm{m}$). Similarly the masked mel-spectrogram of the target speaker $\mathrm{y}^{'}$ is obtained by element-wise multiplication between y and m (i.e. $\mathrm{y}^{'} = \mathrm{y} \odot \mathrm{m}$).

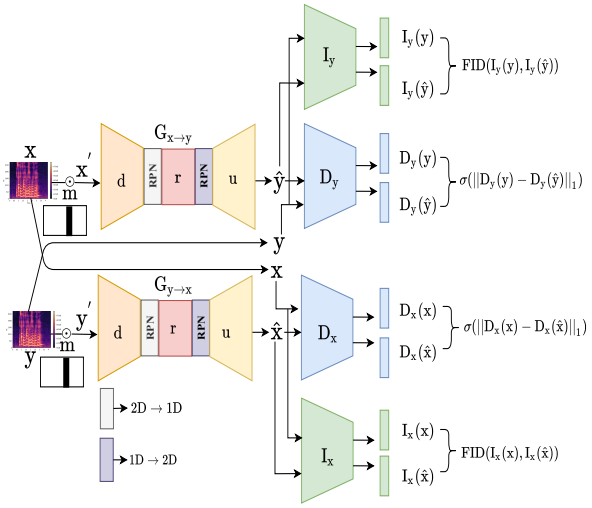


Fig. 1: The schematic overview of the proposed FID-RPRGAN-VC model.

As presented in Fig. 1, generator $\mathrm{G}_{\mathrm{X} \to \mathrm{Y}}$ maps $\mathrm{x}^{'}$ from source domain X to target domain Y, and generate the synthesized mel-spectrogram $\hat{\mathrm{y}}$. Likewise, $\mathrm{G}_{\mathrm{Y} \to \mathrm{X}}$ maps $\mathrm{y}^{'}$ from source domain Y to target domain X and generate $\hat{\mathrm{x}}$. Both $\mathrm{G}_{\mathrm{X} \to \mathrm{Y}}$ and $\mathrm{G}_{\mathrm{Y} \to \mathrm{X}}$ are region-wise positional normalization technique-based generators. And the discriminators D_{X} and D_{Y} of the proposed model are relativistic mechanism-based discriminators that estimate the similarity between the given real data and the synthesized data. In addition to that, Inception-v3 [19] architecture-based deep convolutional neural network (DCNN) model I_{X} and I_{Y} are incorporated in the FID-RPRGAN-VC model to compute the FID loss.

A. Region Wise Positional Normalization Technique-based Generator

The architectural framework of the proposed FID-RPRGAN-VC generators ($\mathrm{G}_{\mathrm{X} \to \mathrm{Y}}$ and $\mathrm{G}_{\mathrm{Y} \to \mathrm{X}}$) consists of downsample block $\mathrm{d}(.)$, residual block $\mathrm{r}(.)$ and upsample block $\mathrm{u}(.)$. Moreover, in between downsample and residual block, 2D $\to$ 1D CNN ($\mathrm{d}_{2 \to 1}(.)$) block is used. Similarly, in the intermediate stage of residual and downsample block, 1D $\to$ 2D CNN ($\mathrm{u}_{1 \to 2}(.)$) block is used. The region wise positional normalization technique is incorporated in both $\mathrm{d}_{2 \to 1}(.)$ and $\mathrm{u}_{1 \to 2}(.)$ block.

The components of $\mathrm{G}(.)$ consists of basic operations such as convolution operation ($\mathrm{Conv}(.)$), instance normalization ($\mathrm{IN}(.)$) [23], gaussian error gated linear unit ($\mathrm{GEGLU}(.)$) [16], [24], region wise positional normalization ($\mathrm{RPN}(.)$) [14], [15], and pixel shuffler ($\mathrm{PS}(.)$) [25].

The architectural framework of FID-RPRGAN-VC generator $\mathrm{G}(.)$ (i.e. both $\mathrm{G}_{\mathrm{X} \to \mathrm{Y}}$ and $\mathrm{G}_{\mathrm{Y} \to \mathrm{X}}$) is expressed in Eq. (1),

$$\mathrm{G}(.) \to \mathrm{u}(\mathrm{u}_{1 \to 2}(\mathrm{r}(\mathrm{d}_{2 \to 1}(\mathrm{d}(.))))). \tag{1}$$

The components of $G(.)$ are represented mathematically below,

$$\mathrm{d}(.) \rightarrow \mathrm{GEGLU}(\mathrm{IN}(\mathrm{Conv2D}(.))),$$
$$\mathrm{d}_{2\rightarrow1}(.) \rightarrow \mathrm{RPN}(\mathrm{Conv1D}(.)),$$
$$\mathrm{r}^{\mathrm{i}} \rightarrow \mathrm{r}^{\mathrm{i}-1} \oplus \mathrm{IN}(\mathrm{Conv1D}(\mathrm{GEGLU}(\mathrm{IN}(\mathrm{Conv1D}(\mathrm{r}^{\mathrm{i}-1}))))),$$

here, r^i indicates the i^{th} residual block.

$$\mathrm{u}_{1\rightarrow2}(.) \rightarrow \mathrm{RPN}(\mathrm{Conv1D}(.)),$$
$$\mathrm{u}(.) \rightarrow \mathrm{GEGLU}(\mathrm{IN}(\mathrm{PS}(\mathrm{Conv2D}(.)))).$$

1) Gaussian Error Gated Linear Unit (GEGLU): GLU [26] is a widely used gating mechanism in the field of language modeling and VC. However, one of the recent variants of GLU called GEGLU [24] has shown significant improvement in sequence-to-sequence modeling. Therefore, in the generator network of the proposed FID-RPRGAN-VC model, GEGLU is used to improve the model's performance for obtaining better quality generated speech samples. The mathematical expression of GEGLU is formulated below,

$$\mathrm{GEGLU}(\mathrm{x}^{'}, \mathrm{W}_1, \mathrm{W}_2, \mathrm{b}_1, \mathrm{b}_2) \rightarrow \mathrm{GELU}(\mathrm{x}^{'}\mathrm{W}_1 + \mathrm{b}_1) \otimes (\mathrm{x}^{'}\mathrm{W}_2 + \mathrm{b}_2),$$
$$\mathrm{GELU}(\mathrm{x}^{'}\mathrm{W}_1 + \mathrm{b}_1) \rightarrow (\mathrm{x}^{'}\mathrm{W}_1 + \mathrm{b}_1) \times \sigma(1.702 \times (\mathrm{x}^{'}\mathrm{W}_1 + \mathrm{b}_1)). \quad (2)$$

As shown in Eq. (2), GEGLU uses the activation function GELU [24]. And $\mathrm{x}^{'}$ is the input vector that passes through two learned linear transformations. Here W_1, W_2, are two different weight matrices and b_1, b_2 are two different bias vectors respectively.

2) Region Wise Positional Normalization (RPN): The RPN technique used in the reshaping block $\mathrm{d}_{2\rightarrow1}(.)$ and $\mathrm{u}_{1\rightarrow2}(.)$ utilize the positional normalization technique [14] in both the masked and unmasked regions of the input matrix. As shown in Fig. 1, the RPN technique is applied on the transformed output of the downsample block $\mathrm{d}(.)$ and the residual block $\mathrm{r}(.)$ respectively. The RPN(.) input matrix is divided into k regions as shown below.

$$\mathrm{x}^{'} = \mathrm{R}^1 \cup \mathrm{R}^2 \cup .... \cup \mathrm{R}^{\mathrm{k}}.$$

Here R^{k} is the k^{th} region of the input matrix $\mathrm{x}^{'}$. However, in this proposed work, k is set as 2 in order to consider masked and unmasked regions. Thereafter, the positional normalization technique is applied to each region (i.e., masked and unmasked). The computation of region wise positional normalization technique is shown through Eq. (3),

$$\hat{\mathrm{R}}^{\mathrm{k}}_{\mathrm{c,h,w}} = \frac{1}{\sigma^{\mathrm{k}}_{\mathrm{h,w}}}(\mathrm{R}^{\mathrm{k}}_{\mathrm{c,h,w}} - \mu^{\mathrm{k}}_{\mathrm{h,w}}),$$
$$\mu^{\mathrm{k}}_{\mathrm{h,w}} = \frac{1}{\mathrm{C}}\Sigma^{\mathrm{C}}_{\mathrm{c=1}}\mathrm{R}^{\mathrm{k}}_{\mathrm{c,h,w}}, \quad (3)$$
$$\sigma^{\mathrm{k}}_{\mathrm{h,w}} = \sqrt{\frac{1}{\mathrm{C}}\Sigma^{\mathrm{C}}_{\mathrm{c=1}}(\mathrm{R}^{\mathrm{k}}_{\mathrm{c,h,w}} - \mu^{\mathrm{k}}_{\mathrm{h,w}})^2}.$$

As per Eq. (3), $\mathrm{R}^{\mathrm{k}}_{\mathrm{c,h,w}}$ is an element of the matrix R^{k}. Whereas, $\hat{\mathrm{R}}^{\mathrm{k}}_{\mathrm{c,h,w}}$ is its corresponding normalized element. Likewise, the positional normalization technique is applied to each of the elements of R^{k} to obtain the normalized matrix $\hat{\mathrm{R}}^{\mathrm{k}}$. In the positional normalization technique, the mean (i.e., $\mu^{\mathrm{k}}_{\mathrm{h,w}}$) and standard deviation (i.e., $\sigma^{\mathrm{k}}_{\mathrm{h,w}}$) are calculated for each one of the elements along all channels for the given matrix. Finally, each normalized region is merged to obtain the final matrix. In this proposed work, the primary objective behind the fusion of positional normalization and region normalization techniques [14], [15] is to tackle the problem of covariate shift in a finer manner. The positional normalization technique computes the normalization statistics independently at all spatial positions and makes the statistics translation, scaling, and rational invariant. This results in better quality generated mel-spectrograms.

B. Relativistic Discriminator

The architectural framework of the relativistic discriminator $\mathrm{D}(.)$ (i.e. both $\mathrm{D_X}$ and $\mathrm{D_Y}$) can be written as follows,

$$\mathrm{D}(.) \rightarrow \mathrm{Conv2D}(\mathrm{d_l}(\mathrm{GEGLU}(\mathrm{Conv2D}(.)))),$$
$$\mathrm{d_l}(.) \rightarrow \mathrm{d_{l-1}}(.). \quad (4)$$

In Eq. (4), $\mathrm{d_l}(.)$ represents l numbers of $\mathrm{d}(.)$ blocks (in this work, l is considered as 4). The outputs of $\mathrm{D_X}$ and $\mathrm{D_Y}$ are the last layer flatten vector (i.e., the latent representation of the input feature) of the corresponding discriminators. The relativistic discriminator incorporated in the FID-RPRGAN-VC model works as a critic that measures the similarity between real and generated speech features. As shown in Eq. (5),

$$\mathcal{L}^{\mathrm{rel_X}}_{\mathrm{disc}} = \sigma(\|\mathrm{D_X}(\mathrm{x}) - \mathrm{D_X}(\hat{\mathrm{x}})\|_1), \quad (5)$$

the last layer flattened vector corresponding to real and synthesized mel-spectrograms are compared to compute the relativistic loss $\mathcal{L}^{\mathrm{rel_X}}_{\mathrm{disc}}$ [27]. Similarly, $\mathcal{L}^{\mathrm{rel_Y}}_{\mathrm{disc}}$ can be computed. The sigmoid activation function σ used in Eq. (5) normalizes the L_1 similarity between the latent representation of real and synthesized mel-spectrograms. Therefore, if the value of σ approaches 0.5, it indicates high similarity, whereas a value higher than 0.5 indicates high dissimilarity. The relativistic loss helps the discriminator to measure the similarity between the latent representations of the real and the synthesized mel-spectrograms.

C. Fréchet inception distance (FID) loss.

The FID metric is a widely used evaluation metric in the field of computer vision and it provides a similarity score between the real and the generated images. However, there is a scope to incorporate FID metric in VC models as a loss function in order to trace the similarity between original and generated samples, while training the generative model. In this work, mel-spectrograms are represented as a matrix of three-dimensional image format (i.e. $2 \times 80 \times 64$). Therefore, in the proposed FID-RPRGAN-VC model, FID metric [17], [18] is

incorporated as a loss function to trace the similarity between original and generated mel-spectrograms. The mathematical formulation of the FID metric is provided below:

$$\mathcal{L}_{\text{fid}}^{\text{X}} = \text{FID}(\text{I}_{\text{x}}(\text{x}), \text{I}_{\text{x}}(\hat{\text{x}})),$$
$$\text{FID}(\text{I}_{\text{x}}(\text{x}), \text{I}_{\text{x}}(\hat{\text{x}})) = ||\mu_{\text{I}_{\text{x}}(\text{x})} - \mu_{\text{I}_{\text{x}}(\hat{\text{x}})}||^2 + \text{T}_{\text{r}}(\Sigma_{\text{I}_{\text{x}}(\text{x})} + \Sigma_{\text{I}_{\text{x}}(\hat{\text{x}})} - 2(\Sigma_{\text{I}_{\text{x}}(\text{x})}\Sigma_{\text{I}_{\text{x}}(\hat{\text{x}})})^{1/2}). \quad (6)$$

Here, $\mathcal{L}_{\text{fid}}^{\text{X}}$ indicates the FID loss between the latent vectors $\text{I}_{\text{x}}(\text{x})$ and $\text{I}_{\text{x}}(\hat{\text{x}})$ that are obtained using Inception-v3 [19] architecture based DCNN model, as depicted in Fig. 1. In Eq. (6), the first term computes the mean difference, whereas the second term computes the covariance difference between $\text{I}_{\text{x}}(\text{x})$ and $\text{I}_{\text{x}}(\hat{\text{x}})$. The minimization of FID improves the similarity between the original and the generated mel-spectrograms, and helps in obtaining the formant information in a better manner. Though, both relativistic and FID loss compute the difference in latent space. But, the additional advantage of FID-loss is that it incentivizes the generator to generate mel-spectrograms by maintaining the feature correlation [17]. Apart from relativistic loss and FID loss, adversarial loss, cycle-consistency loss, and identity loss are also used in FID-RPRGAN-VC, keeping it same as MaskCycleGAN-VC [13]. However, in this work, the FID loss is utilized as an additional generator training loss.

III. Experimental Design

A. Dataset Description

The considered models are trained and tested on VCC 2018 [20], and CMU Arctic [21] dataset. For the VCC 2018 dataset, the considered speakers are VCC2TM1, VCC2SM3, VCC2TF1, and VCC2SF3. Whereas, for the CMU Arctic dataset, the regarded speakers are cmu-us-bld-Arctic, cmu-us-rms-Arctic, cmu-us-clb-Arctic, and cmu-us-slt-Arctic. For both the datasets, 81 speech samples and 35 speech samples were considered for training and testing, respectively. For CMU Arctic [21] dataset (out of 539 belonging to each speakers) we have considered 116 (i.e. 81 for training and 35 for testing) speech samples for each speakers such that the utterances are disjoint. The particular setting is considered for making the training process non-parallel. Additionally, we evaluated the performance of the proposed model for Easycall Dysarthric dataset [22]. The Easycall dataset contains parallel speech content for both normal and dysarthric speakers. Here, four speakers belonging to the male and female genders are considered. For each of the speakers, 264 samples were considered for training, and 66 samples considered for testing.

B. Network Architecture of FID-RPRGAN-VC

The FID-RPRGAN-VC model consists of region wise positional normalization technique based generator and relativistic discriminator. The baseline architecture of the generator is similar to MaskCycleGAN-VC. In the generator two $\text{d}(.)$ and $\text{u}(.)$ blocks are present. Moreover, the generator is comprise of one $\text{u}_{1\rightarrow 2}(.)$ and $\text{d}_{2\rightarrow 1}(.)$ block. In addition to that, six $\text{r}(.)$ blocks are incorporated in the generator. Whereas the discriminator consists of four $\text{d}(.)$ blocks. The architectural framework of the Inception-v3 model adapted to this work is based on [19] for the calculation of FID-loss.

C. Training Details

For training the FID-RPRGAN-VC model, Adam optimizer [28] is used with learning rate 1×10^{-4}. The proposed model is trained for 1×10^3 epochs (i.e. 8×10^5 iterations). The mini-batch size is considered as 1. In this work, pretrained MelGAN vocoder [29] is used for mel-spectrogram to audible speech synthesis. The size of the mel-spectrograms considered here is $2 \times 80 \times 64$, and the number of samples is 64. The mask size here is taken as 25% of the input size (along horizontal axis i.e. width). This value is considered based on the experimental results of MaskCycleGAN-VC [13].

D. Experimental Setup

The complete execution of the proposed FID-RPRGAN-VC model is carried out in the Dell precision 7820 workstation configured with ubuntu 18.04 64 bit Operating System, Intel Xeon Gold 5215 2.5GHz processor, 96GB RAM, and Nvidia 16GB Quadro RTX5000 graphics. All the experiments of this work are implemented in Python 3.6.9 using Pytorch 1.1.2 and Numpy 1.19.5. The audible speech data are preprocessed by using Librosa 0.9.1.

IV. Results and Discussion

A. Objective Evaluation

The average mel cepstral distortion (MCD) [30], modulation spectra distance (MSD) [31], and F_0 root mean square error (F_0 RMSE) [32], [33] values of the proposed FID-RPRGAN-VC and MaskCycleGAN-VC model-generated speech samples are provided in Table I,

TABLE I: MCD, MSD and F_0 RMSE values for VCC 2018 and CMU-Arctic dataset

Dataset	Models	M-M	F-F	M-F	F-M
MCD					
VCC 2018	**FID-RPRGAN-VC**	**6.40**	**6.45**	**6.53**	**6.73**
	MaskCycleGAN-VC	7.45	6.85	6.76	7.84
CMU-Arctic	**FID-RPRGAN-VC**	**6.97**	**7.48**	**8.09**	**7.97**
	MaskCycleGAN-VC	7.12	7.81	8.20	8.07
MSD					
VCC 2018	**FID-RPRGAN-VC**	**1.15**	**1.14**	**1.21**	**1.16**
	MaskCycleGAN-VC	1.17	1.18	1.50	1.24
CMU-Arctic	**FID-RPRGAN-VC**	**1.16**	**1.15**	**1.28**	**1.26**
	MaskCycleGAN-VC	1.18	1.19	1.32	1.29
F_0 RMSE					
VCC 2018	**FID-RPRGAN-VC**	**18.17**	**27.22**	**32.20**	**36.78**
	MaskCycleGAN-VC	18.77	28.37	34.20	38.43
CMU-Arctic	**FID-RPRGAN-VC**	**15.10**	**21.81**	**31.71**	**31.93**
	MaskCycleGAN-VC	15.25	23.62	35.12	34.98

for both VCC 2018 and CMU Arctic dataset[1]. In the case of the objective evaluation metrics, a lower value indicates high speaker similarity and vice versa. As shown in Table I,

[1]All the generated speech samples and the code implementation of our proposal can be downloaded from http://surl.li/fqgpz

for both intra-gender and inter-gender VC the obtained MCD values of FID-RPRGAN-VC is less than MaskCycleGAN-VC. Considering the average of MCD values from Table I, it can be observed that the performance improvement is 6.09%. Similarly, from Table I, it can also be found that FID-RPRGAN-VC performed superior to MaskCycleGAN-VC in terms of MSD and F_0 RMSE. Based on the average of MSD and F_0 RMSE values, the performance improvement is 2.58% and 5.31%, respectively.

We also report an ablation study to evaluate the performance of the proposed FID-RPRGAN-VC model for each contribution using the VCC 2018 dataset. In Table II, ABS(1)

TABLE II: MCD, MSD and F_0 RMSE values for ablation study

Models	M-M	F-F	M-F	F-M
MCD				
FID-RPRGAN-VC	**6.40**	**6.45**	**6.53**	**6.73**
ABS(1)	6.42	6.47	6.57	6.76
ABS(2)	6.56	6.72	6.75	7.09
ABS(3)	6.46	6.54	6.60	6.98
ABS(4)	6.50	6.61	6.69	7.11
MSD				
FID-RPRGAN-VC	**1.15**	**1.14**	**1.21**	**1.16**
ABS(1)	1.16	1.16	1.23	1.17
ABS(2)	1.23	1.27	1.30	1.28
ABS(3)	1.20	1.16	1.24	1.23
ABS(4)	1.21	1.23	1.24	1.21
F_0 RMSE				
FID-RPRGAN-VC	**18.17**	**27.22**	**32.20**	**36.78**
ABS(1)	18.23	27.30	32.28	36.87
ABS(2)	18.54	27.65	32.82	38.38
ABS(3)	18.35	28.02	32.61	38.46
ABS(4)	18.41	27.97	32.70	38.47

indicates the FID-RPRGAN-VC model without GEGLU (replaced by GLU of MaskCycleGAN-VC). ABS(2) indicates the FID-RPRGAN-VC model without RPN based generator (replaced by TFAN [34] of MaskCycleGAN-VC). ABS(3) denotes the FID-RPRGAN-VC model without relativistic discriminator (replaced by MaskCycleGAN-VC discriminator). And, ABS(4) denotes the FID-RPRGAN-VC model without FID loss. Here, the term ABS indicates the models for ablation study.

In Table III, based on the percentage ($\%$) of increment of MCD and MSD values w.r.t FID-RPRGAN-VC, the models ABS(1), ABS(2), ABS(3), and ABS(4) can be ranked as rank: 1, rank: 4, rank: 2, rank: 3, respectively. Since lower MCD and MSD values signify a greater degree of speaker similarity, the model achieving the lowest MCD and MSD is assigned the rank of 1, and so forth. It can be observed from Table III that for both MCD and MSD metrics, the models ABS(4) and ABS(2) obtained the rank: 3 and rank: 4 respectively, which implies the significance of FID loss and the RPN based generator in the proposed model. The removal of these two components significantly degraded the performance of the proposed model. It is also observable from Table III that in terms of F_0 RMSE ($\%$ of increment of F_0 RMSE) the models ABS(4) and ABS(2) obtained the rank: 4 and rank: 3 respectively. This indicates that the removal of these two components also degraded the results in terms of F_0 fundamental frequency or prosody.

TABLE III: MCD, MSD and F_0 RMSE values for ablation study

MCD							
ABS(1)	Rank	ABS(2)	Rank	ABS(3)	Rank	ABS(4)	Rank
0.41%	1	3.82%	4	1.77%	2	3.02%	3
MSD							
ABS(1)	Rank	ABS(2)	Rank	ABS(3)	Rank	ABS(4)	Rank
1.25%	1	9.03%	4	3.64%	2	4.95%	3
F_0 RMSE							
ABS(1)	Rank	ABS(2)	Rank	ABS(3)	Rank	ABS(4)	Rank
0.28%	1	2.46%	3	2.44%	2	2.55%	4

Apart from non-parallel uni-lingual VC, the performance of the proposed FID-RPRGAN-VC model is also evaluated for disordered speech to normal speech conversion considering a dysarthric speech dataset named as Easycall dataset [22], in a parallel setting. In Fig. (2), the schematic overview of the dysarthric to normal speech conversion process using FID-RPRGAN-VC model is shown. In Fig. (3), the visual

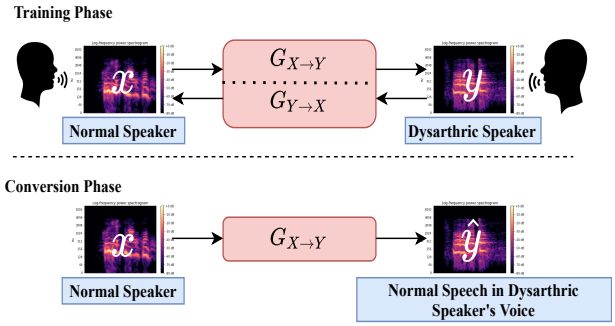


Fig. 2: Schematic overview of the dysarthric to normal speech conversion using FID-RPRGAN-VC model.

comparison of mel-spectrograms for MaskCycleGAN-VC and FID-RPRGAN-VC generated dysarthric to normal speech samples are provided. From the visual comparison (pointed with arrows and elliptical marker) of the mel-spectrograms it is well observed that the proposed FID-RPRGAN-VC generated mel-spectrograms are more similar to the normal speaker's mel-spectrograms as compared to MaskCycleGAN-VC generated samples. The FID-RPRGAN-VC generated mel-spectrograms captures well defined formant structures of the normal speech as compared to the MaskCycleGAN-VC as shown by the annotations 2, 3 and 4, in Fig. (3). We can also see that the generated spectrogram by MaskCycleGAN-VC is noisy (denoted by 1), whereas FID-RPRGAN-VC has obtained similar structures with respect to the target.

B. Subjective Evaluation

The subjective evaluations are conducted to measure the naturalness of the generated speech samples in terms of

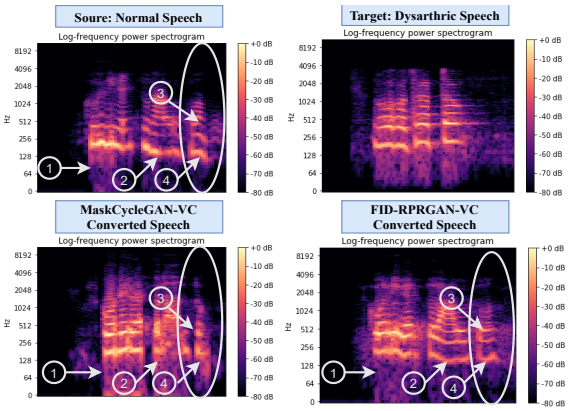


Fig. 3: Visual comparison of mel-spectrograms for MaskCycleGAN-VC and FID-RPRGAN-VC converted speech.

MOS [35]. To obtain the MOS values, 17 volunteers are considered. In order to collect the MOS values, 35 generated speech samples from both the datasets are taken. Each of the volunteers scored any randomly picked speech samples (for each categories i.e. M-M, F-F, M-F and F-M) in a range between 1-5 as per the standard MOS calculation method [35]. The MOS test is conducted for original speech samples, FID-RPRGAN-VC generated speech samples and MaskCycleGAN-VC-generated speech samples. The MOS values with 95% confidence intervals are provided in Table IV for both VCC 2018 and CMU-Arctic datasets.

TABLE IV: MOS for naturalness with 95% confidence intervals

Dataset	Models	M-M	F-F	M-F	F-M
VCC 2018	**FID-RPRGAN-VC**	**3.9±0.29**	**3.8±0.18**	**3.1±0.35**	**3.4±0.51**
	MaskCycleGAN-VC	3.1±0.32	2.9±0.42	2.6±0.23	2.9±0.24
CMU-Arctic	**FID-RPRGAN-VC**	**3.7±0.36**	**3.8±0.18**	**3.5±0.23**	**3.8±0.26**
	MaskCycleGAN-VC	2.9±0.15	3.1±0.41	3±0.04	3±0.05

Table IV shows that the MOS values for FID-RPRGAN-VC generated speech samples are higher than MaskCycleGAN-VC generated speech samples. This implies the superiority of the proposed FID-RPRGAN-VC model over MaskCycleGAN-VC in terms of subjective evaluation.

V. CONCLUSION

In this work, we proposed an improved GAN model for mel-spectrogram based VC and referred to it as the FID-RPRGAN-VC model that consists of a region-wise positional normalized generator, a relativistic discriminator, and a FID loss function. These modifications aim to generate mel-spectrograms that capture the target distribution better than the SOTA MaskCycleGAN-VC model. The proposed model is tested on VCC 2018, CMU Arctic, and Easycall speech datasets. The objective and subjective evaluation of the FID-RPRGAN-VC generated samples indicates the superiority of the proposed model. In the future, the GAN-based VC model can also be investigated for speech enhancement purposes. Moreover, there is also a scope to explore the model for multi-lingual VC.

REFERENCES

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in NIPS*, vol. 27. Curran Associates, Inc., 2014.

[3] S. Dhar, N. D. Jana, and S. Das, "An adaptive learning based generative adversarial network for one-to-one voice conversion," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2022.

[4] H. Uchida, D. Saito, N. Minematsu, and K. Hirose, "Statistical acoustic-to-articulatory mapping unified with speaker normalization based on voice conversion," in *Proc. Interspeech 2015*, 2015, pp. 588–592.

[5] L. Sun, K. Li, H. Wang, S. Kang, and H. M. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2016.

[6] R. Aihara, T. Takiguchi, and Y. Ariki, "Parallel dictionary learning for voice conversion using discriminative graph-embedded non-negative matrix factorization," in *Interspeech*, 2016.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 655–658 vol.1.

[8] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018, 26th EUSIPCO*, 2018, pp. 2100–2104.

[9] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," *ICASSP 2019*, pp. 6820–6824, 2019.

[10] S. Dhar, N. D. Jana, and S. Das, "An adaptive-learning-based generative adversarial network for one-to-one voice conversion," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 1, pp. 92–106, 2023.

[11] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Interspeech*, 2017.

[12] S. Seshadri, L. Juvela, P. Alku, and O. J. Räsänen, "Augmented cyclegans for continuous scale normal-to-lombard speaking style conversion," in *Interspeech*, 2019.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames," *ICASSP*, pp. 5919–5923, 2021.

[14] B. Li, F. Wu, K. Q. Weinberger, and S. J. Belongie, "Positional normalization," *ArXiv*, vol. abs/1907.04312, 2019.

[15] T. Yu, Z. Guo, X. Jin, S. Wu, Z. Chen, W. Li, Z. Zhang, and S. Liu, "Region normalization for image inpainting," in *AAAI*, 2020.

[16] N. M. Shazeer, "Glu variants improve transformer," *ArXiv*, vol. abs/2002.05202, 2020.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, 2017.

[18] A. Mathiasen and F. Hvilshøj, "Fast fréchet inception distance," *ArXiv*, vol. abs/2009.14075, 2020.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2015.

[20] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. H. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey*, 2018.

[21] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *SSW*, 2004.

[22] R. Turrisi, A. Braccia, M. Emanuele, S. Giulietti, M. Pugliatti, M. R. F. de Sensi, L. Fadiga, and L. Badino, "Easycall corpus: a dysarthric speech dataset," in *Interspeech*, 2021.

[23] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *ArXiv*, vol. abs/1607.08022, 2016.
[24] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *ArXiv*, vol. abs/1606.08415, 2016.
[25] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874–1883, 2016.
[26] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*, 2016.
[27] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *ArXiv*, vol. abs/1807.00734, 2019.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980
[29] K. Kumar, R. Kumar, T. de Boissière, L. Gestin, W. Z. Teoh, J. M. R. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, 2019.
[30] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing voice conversion network with cycle consistency loss of speaker identity," in *2021 IEEE SLT*, 2021, pp. 507–513.
[31] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.
[32] M. T. Akhter, P. Banerjee, S. Dhar, and N. D. Jana, "An analysis of performance evaluation metrics for voice conversion models," in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022, pp. 1–6.
[33] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer learning from speech synthesis to voice conversion with non-parallel training data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1290–1302, 2021.
[34] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," in *INTERSPEECH*, 2020.
[35] Y.-p. Li, Z. Ling-hua, and D. Hui, "Nonparallel voice conversion based on phoneme classification and eigenvoices," in *2010 IEEE 12th International Conference on Communication Technology*, 2010, pp. 662–665.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Deformable Aligned Fusion for Video Super Resolution

Sin-Hong Lee *, Chih-Hung Kuo † and Tsai-Chun Yu ‡
* Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
E-mail: leesinhong0312@gmail.com
† Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
E-mail: chkuo@ee.ncku.edu.tw
‡ Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
E-mail: n26100066@gs.ncku.edu.tw

***Abstract*— The key to the video super-resolution algorithm is to capture the information of adjacent frames to supplement the reconstruction of the current frame. It is necessary to apply the alignment operation to compensate the displacements of objects in different frames. We propose an innovative deformable aligned fusion block (DAFB) to exploit the complementary information from adjacent frames more efficiently. We also model the offsets of deformable alignments as probabilistic distributions to augment the robustness of a DAFB. Our proposed model outperforms state-of-the-art CNN-based models on benchmark testing datasets, including Vid4 and UDM10.**

I. Introduction

The super-resolution technique takes one or more input low-resolution images to recover the corresponding high-resolution image. The single image super resolution (SISR) restores the high-resolution image through a low-resolution image, and the advantage of the video super-resolution (VSR) over SISR is to utilize the adjacent frames to help restore the current frame.

Obtaining the complementary information from the adjacent frames is critical. There are different fusion strategies proposed for the VSR task. Direct fusion [1, 2, 3] exploits the convolutional layers to merge the features in multiple input images at the early stage of the network. Yi *et al.* [4, 5] propose the progressive fusion residual block (PFRB), which utilizes the convolutional layers to capture the temporal correlations between frames. By stacking multiple layers of PFRBs, their proposed network progressively fuses the adjacent feature maps. However, the aforementioned methods lack consideration of the motions between frames. When the motions are large, it is hard for the convolutional layers with small receptive fields to acquire the complementary information from the adjacent frames. Grabbing uncorrelated information results in poor reconstructions.

Dai *et al.* [6] first propose the deformable convolution, in which additional intra-frame offsets are learned to capture information beyond their local neighborhoods, enhancing the capability of convolutions. Tian *et al.* [7] utilize the deformable convolution to achieve the temporal alignment. They dynamically predict the inter-frame offsets of objects between different features maps, and leverage the deformable convolution on adjacent feature maps to align temporally by the predicted offsets. This method is referred to as the deformable alignment. Recent VSR methods [7, 8, 9, 10, 11, 12] utilize the deformable alignment to address the issue of object displacement between frames. Chan *et al.* [13] show that the performance of deformable alignment improves as the number of offset groups increases.

In this study, to mitigate the problem of misalignment, we design the deformable aligned fusion blocks (DAFB) by incorporating the deformable alignment module [7, 13] into the PFRB [4, 5] to better fuse the adjacent feature maps. Compared with prior fusion methods, the DAFB can utilize the alignment module to process object features with larger and more complex motion vectors.

We further model the offsets of deformable alignment as Gaussian distributions. Some prior image processing studies [14, 15, 16] utilize probability distributions to represent the uncertainty in predictions, helping neural networks to be more robust to noisy input data. However, there hasn't been a similar approach in the field of video super-resolution. The modified block, named DAFB-p, estimates the uncertainty of multiple groups of Gaussian distributed offsets, and accordingly find the most probable group as input to the deformable convolution. Through extensive experiments, we find that for the networks with fewer offset groups, the DAFB-p outperforms the DAFB.

Our main contributions are listed as follows.

1) We propose the deformable aligned fusion block, which progressively aligns and fuses neighboring feature maps. The alignment enhances the efficiency of capturing temporal information by the convolutional layers in the PFRB.
2) We measure the uncertainty of multiple offset groups using probability distributions, thereby increasing the tolerance for inaccurate offset estimations.

The rest of this paper is organized as follows. Section II reviews the related works. Section III details the proposed networks and the corresponding loss functions. Section IV

Acknowledgment: This work was sponsored by National Science and Technology Council, project number 111-2221-E-006-185.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

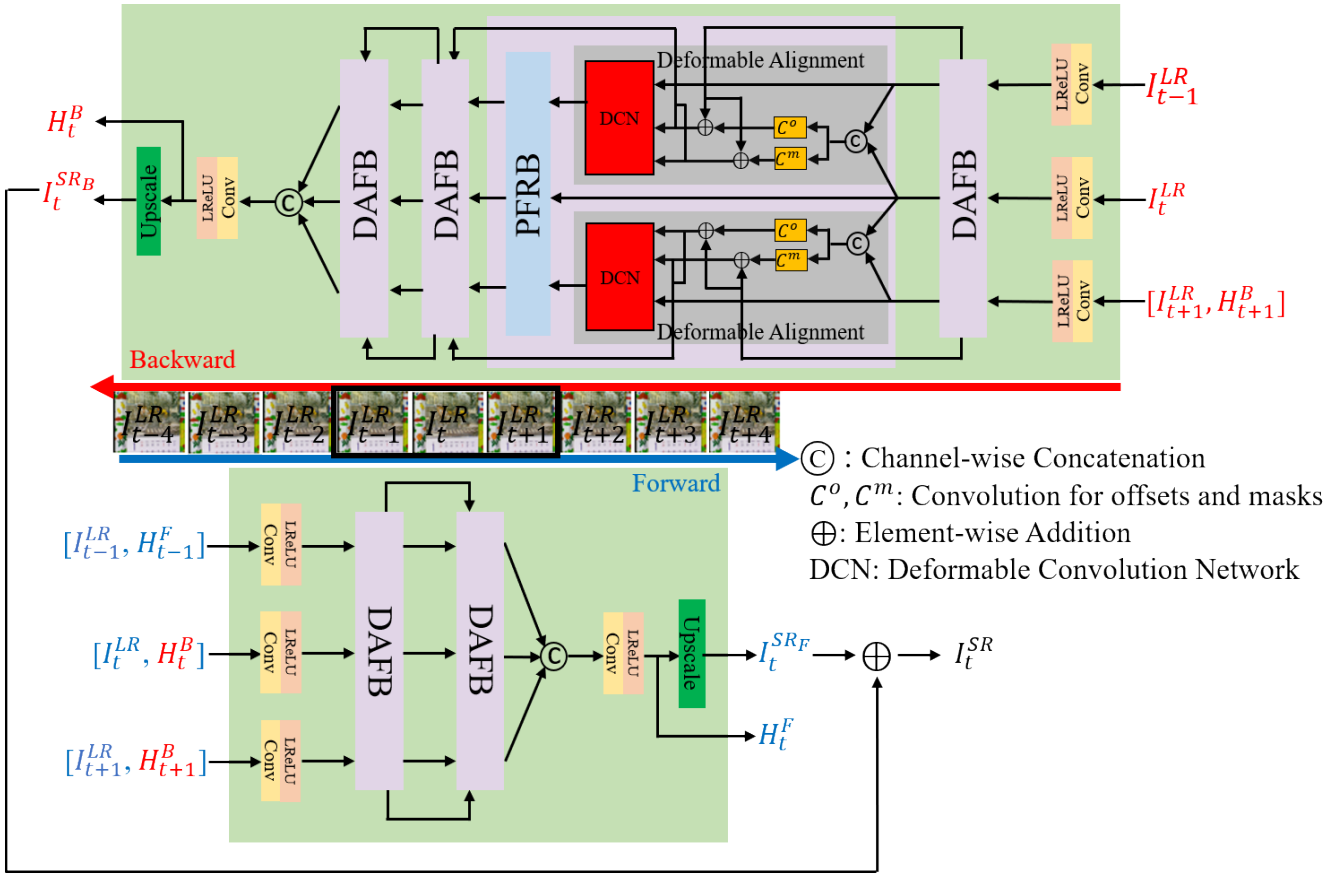


Fig. 1: Our proposed deformable aligned fusion network (DAFN).

presents our experimental results and ablation studies. Section V concludes this work.

II. Related works

Most VSR models [7, 8, 9, 10, 11, 17, 18, 19, 20] employ alignment operation to capture the temporal correlations between frames. The alignment methods of VSR tasks can be classified into three categories: image-level flow alignment [17], feature-level flow alignment [18, 19, 20] and deformable alignment [7, 8, 9, 10, 11].

The image-level flow alignment first estimates the optical flow between frames and then warps the neighboring frames. The BasicVSR [20] points out that warping frames at the image level inevitably destroys the original details of the images and results in blurring. The feature-level flow alignment warps the features from adjacent frames, and significantly improves the reconstruction performance.

The most commonly adopted alignment approach recently is the deformable alignment. Chan *et al.* [13] mention that the main advantage of the deformable alignment over the flow alignments comes from the offset diversity. In some areas that are difficult to recover, such as image boundaries or fast-moving objects, the deformable convolution improves alignment via multiple groups of offsets. The TDAN [7] is the fisrt method to incorporate the deformable alignment in VSR framework. The EDVR [9] designs a pyramid-shaped alignment module, which performs in a coarse-to-fine manner to process complex motions. Chan *et al.* [11] design a flow-guided alignment module that helps alleviate the burden of offsets learning in deformable alignment by using the optical flow as base offsets.

III. Methodology

A. deformable aligned fusion block

As shown in Fig. 1, the proposed VSR architecture, deformable aligned fusion network (DAFN) is based on the bidirectional recurrent neural network from GOVSR [5] and formed by stacking multiple layers of DAFB. Fig. 2 illustrates the DAFB in the n-th layer, including three input paths for the past feature map F_{t-1}^{n-1}, the present feature map F_t^{n-1} and the future feature map F_{t+1}^{n-1}, respectively. The deformable alignment on the future path utilizes a layer of convolution $(\mathrm{C^o}, \mathrm{C^m})$ to generate the residual offsets $\hat{o}_{t+1,k,g}^n$ and masks $\hat{m}_{t+1,k,g}^n$ from the concatenated feature maps,

$$\hat{o}_{t+1,k,g}^n = \mathrm{C^o}(\{F_{t+1}^{n-1}, F_t^{n-1}\}), \tag{1}$$

$$\hat{m}_{t+1,k,g}^n = \mathrm{C^m}(\{F_{t+1}^{n-1}, F_t^{n-1}\}), \tag{2}$$

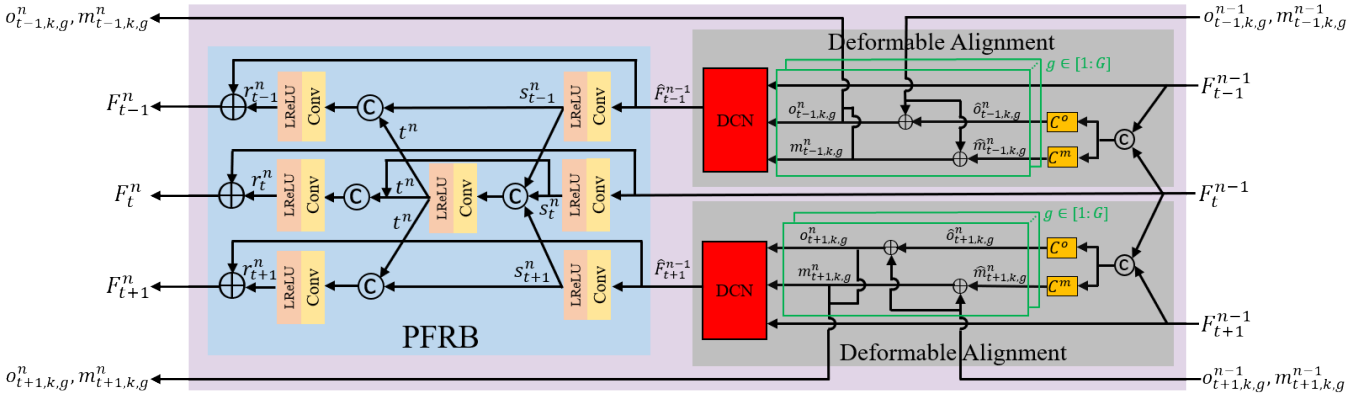


Fig. 2: Our proposed deformable aligned fusion block (DAFB) at the n^{th} layer.

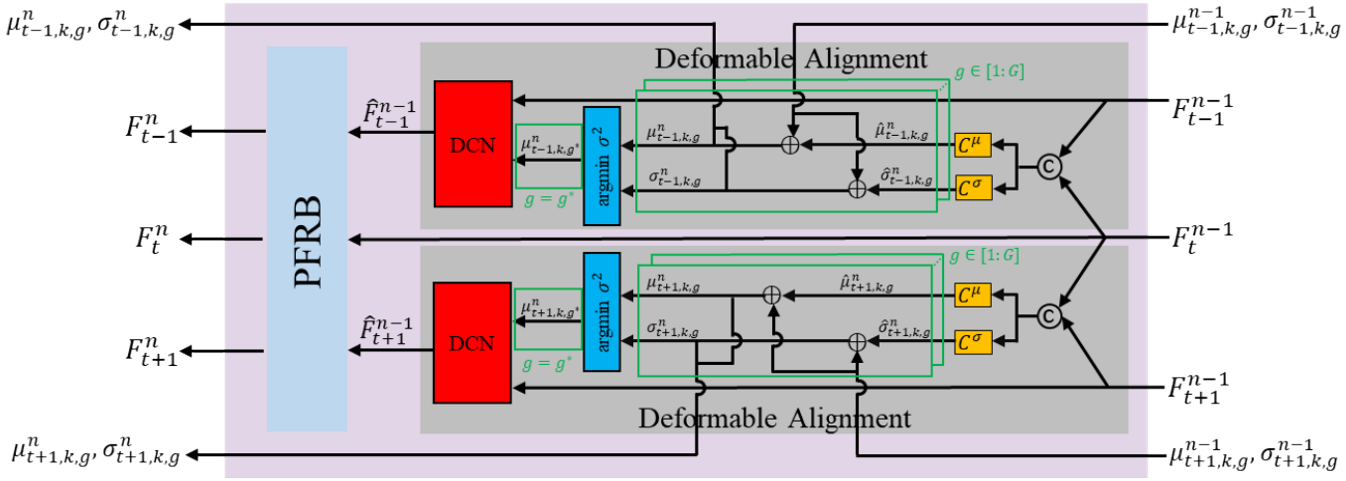


Fig. 3: Our proposed DAFB-p at the n^{th} layer.

for the k-th sampling point of the deformable convolution network (DCN) in the g-th group of offsets and masks.

The offsets and masks in the DAFB of different layers are similar since multi-layer deformable alignments are to align features in the same image. The residual offsets $\hat{o}^n_{t+1,k,g}$ and masks $\hat{m}^n_{t+1,k,g}$ of the current deformable alignment are element-wise added to the offsets $o^{n-1}_{t+1,k,g}$ and masks $m^{n-1}_{t+1,k,g}$ propagated from the previous deformable alignment,

$$o^n_{t+1,k,g} = o^{n-1}_{t+1,k,g} + \hat{o}^n_{t+1,k,g}, \tag{3}$$

$$m^n_{t+1,k,g} = m^{n-1}_{t+1,k,g} + \hat{m}^n_{t+1,k,g}. \tag{4}$$

The offsets $o^n_{t+1,k,g}$ and the masks $m^n_{t+1,k,g}$ are then fed into the DCN to generate the aligned feature $\hat{F}^{n-1}_{t+1}$. The DCN determines the sampling locations of the convolution kernel through the offsets, and the masks are regarded as attention maps to calculate the importance of sampling locations,

$$\hat{F}^{n-1}_{t+1}(p) = \sum_{k=1}^{s^2} w(q_k) \cdot F^{n-1}_{t+1}(p+q_k+o^n_{t+1,k,g}) \cdot m^n_{t+1,k,g}, \tag{5}$$

where s is the kernel size, p is the current position, and q_k is the k-th sampling offset of the convolution. For example, if kernel size $s = 3$, then $q_k \in \{(-1,-1),(-1,0),\ldots,(1,1)\}$. The offset $o^n_{t+1,k,g}$ and the mask $m^n_{t+1,k,g}$ are predicted for the position $p + q_k$.

As shown in Fig. 2, we adopt the PFRB [4, 5] to gradually fuse features. The PFRB extracts the intra-frame spatial information $s^n_{t-1}, s^n_t, s^n_{t+1}$ from the input feature maps $\hat{F}^{n-1}_{t-1}, F^{n-1}_t, \hat{F}^{n-1}_{t+1}$, respectively. These spatial information are then concatenated into a convolutional layer to capture the inter-frame temporal information t^n. Since the adjacent feature maps $\hat{F}^{n-1}_{t-1}, \hat{F}^{n-1}_{t+1}$ are pre-aligned to the current one F^{n-1}_t, the convolutional layer captures the temporal information t^n more efficiently. The temporal information t^n is sent back to each path, and individually fused with the spatial information $s^n_{t-1}, s^n_t, s^n_{t+1}$ through a convolutional layer to obtain the temporal-spatial information $r^n_{t-1}, r^n_t, r^n_{t+1}$. The PFRB outputs the summation of these temporal-spatial information and the input feature maps,

$$F^n_{t-1}, F^n_t, F^n_{t+1} = \mathrm{PFRB}(\hat{F}^{n-1}_{t-1}, F^{n-1}_t, \hat{F}^{n-1}_{t+1}). \tag{6}$$

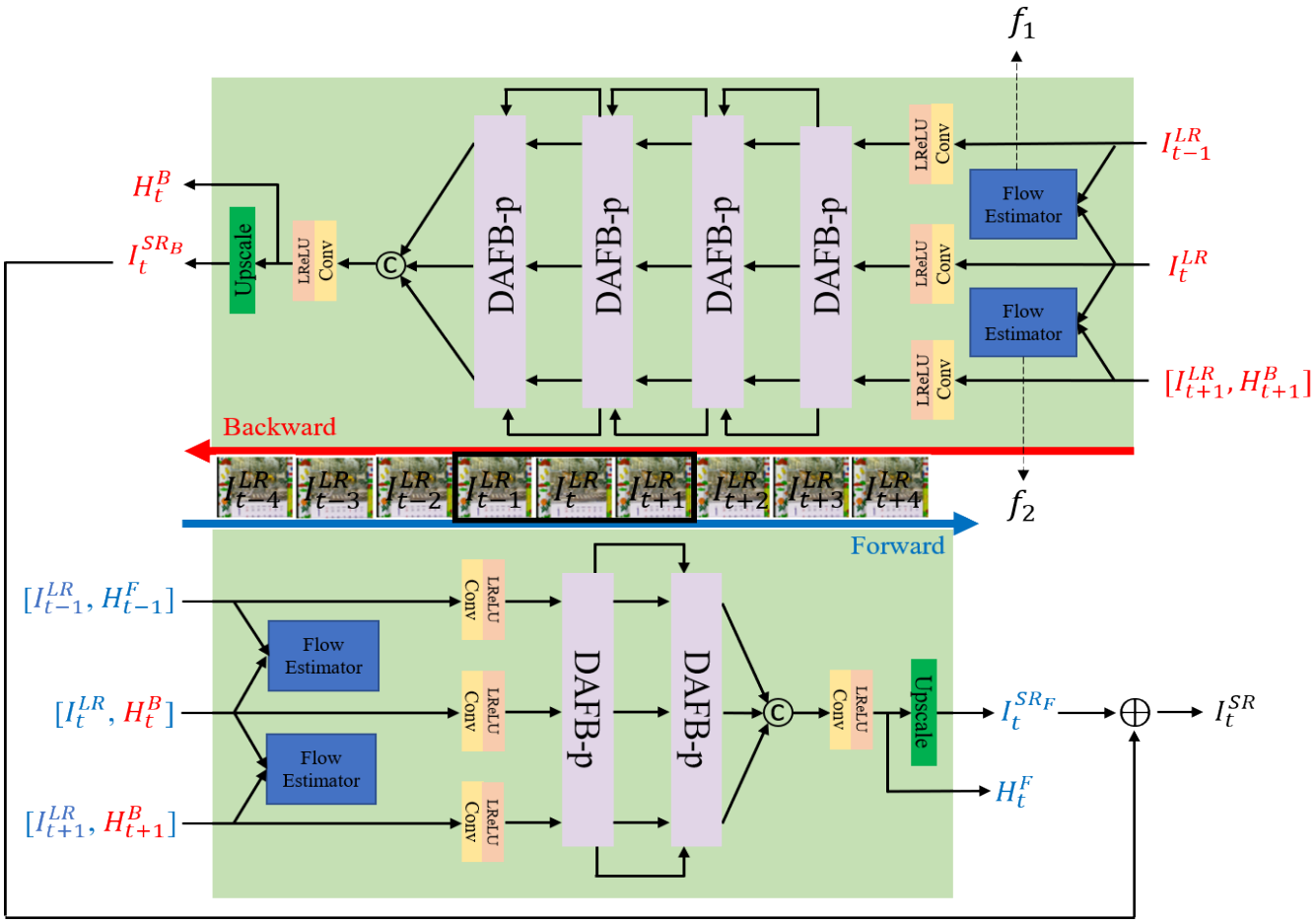


Fig. 4: Our proposed DAFN-p.

As shown in Fig. 1, the DAFN is stacked by a series of DAFB to progressively align and fuse the adjacent feature maps. The DAFN is a bidirectional recurrent neural network, and the final reconstructed image I_t^{SR} is the summation of the output of the backward network and that of the forward network,

$$I_t^{SR} = I_t^{SR_B} + I_t^{SR_F}. \tag{7}$$

We set the loss function by following the approach proposed by Yi *et al.* [5], who choose the Charbonnier loss function during training since it better handles outliers,

$$L = \sqrt{(I_t^{HR} - I_t^{SR})^2 + \varepsilon^2} + \alpha\sqrt{(I_t^{HR} - I_t^{SR_B})^2 + \varepsilon^2}, \tag{8}$$

where I_t^{HR} is the ground truth image, $\varepsilon = 10^{-3}$. We set the weighting factor $\alpha = 0.01$.

B. DAFB with probabilistic offsets

To measure the uncertainty of the offsets, the DAFB-p represents the offsets probabilistically rather than deterministically. Fig. 3 illustrates the DAFB-p in the n-th layer. For convenience, we simply model the distribution of the offsets $o_{t+1,k,g}^n$ as Gaussian distributions with mean $\mu_{t+1,k,g}^n$ and variance $\sigma_{t+1,k,g}^n$, and the deformable alignment utilizes a layer of convolution $(\mathrm{C}^\mu, \mathrm{C}^\sigma)$ to generate the residual mean and variance,

$$\hat{\mu}_{t+1,k,g}^n = \mathrm{C}^\mu(\{F_{t+1}^{n-1}, F_t^{n-1}\}), \tag{9}$$

$$\hat{\sigma}_{t+1,k,g}^n = \mathrm{C}^\sigma(\{F_{t+1}^{n-1}, F_t^{n-1}\}), \tag{10}$$

for the k-th sampling point of the deformable convolution network (DCN) in the g-th group. They are element-wise added to the mean $\mu_{t+1,k,g}^{n-1}$ and variance $\sigma_{t+1,k,g}^{n-1}$ propagated from the previous deformable alignment,

$$\mu_{t+1,k,g}^n = \mu_{t+1,k,g}^{n-1} + \hat{\mu}_{t+1,k,g}^n, \tag{11}$$

$$\sigma_{t+1,k,g}^n = \sigma_{t+1,k,g}^{n-1} + \hat{\sigma}_{t+1,k,g}^n. \tag{12}$$

The variance $\sigma_{t+1,k,g}^n$ may represent the uncertainty of the offset $o_{t+1,k,g}^n$. The smaller variance indicates the higher confidence. In Fig. 3, we select the offset with the minimum variance from G groups as the representative $N(\mu_{t+1,k,g^*}^n, \sigma_{t+1,k,g^*}^n)$,

$$g^* = \arg\min_g \ \sigma_{t+1,k,g}^n. \tag{13}$$

TABLE I: Quantitative comparison (PSNR (dB) / SSIM) on Vid4 [21] and UDM10 [4] for $4\times$ VSR. Red text indicates the best performance and blue indicates the second best.

Models	Parameters (M)	Runtime (ms)	BD Degradation	
			UDM10 [4]	Vid4 [21]
Bicubic	-	-	28.47 / 0.8253	21.80 / 0.5246
TOFlow [22]	-	-	36.26 / 0.9438	-
FRVSR [19]	5.1	137	37.17 / 0.9525	26.69 / 0.8220
MMCNN [23]	10.6	696.7	37.73 / 0.9561	26.28 / 0.7844
DUF_52L [2]	5.8	974	38.05 / 0.9586	27.34 / 0.8327
RBPN [24]	12.8	1507	39.00 / 0.9642	27.16 / 0.8190
EDVR [9]	20.7	378	39.05 / 0.9642	27.35 / 0.8264
PFNL [4]	3.0	295	39.00 / 0.9645	27.40 / 0.8384
TGA [25]	7.1	383.5	-	27.59 / 0.8419
RLSP [26]	5.6	-	39.02 / 0.9642	27.55 / -
RSDN9-128 [27]	6.2	94	39.30 / 0.9655	27.92 / 0.8505
RRN [28]	3.4	45	38.96 / 0.9644	27.69 / 0.8488
FFCVSR [29]	5.6	50.8	38.49 / 0.9599	26.97 / 0.8300
TDAN [7]	2.3	102.3	38.19 / 0.9586	26.86 / 0.8140
BasicVSR [20]	6.3	63	39.96 / 0.9694	27.96 / 0.8553
IconVSR [20]	8.7	70	40.03 / 0.9694	28.04 / 0.8570
GOVSR [5]	7.1	81.2	40.14 / 0.9713	28.41 / 0.8724
ETDM [30]	8.7	70	40.11 / 0.9707	28.81 / 0.8725
DAFN (Ours)	4.85	53	40.21 / 0.9695	28.77 / 0.8787
DAFN-p (Ours)	5.46	57	40.31 / 0.9699	28.51 / 0.8755
DAFN-l (Ours)	9.39	103	40.53 / 0.9713	29.13 / 0.8828

We treat the mean μ^n_{t+1,k,g^*} as the offset to be input into the DCN,

$$\hat{F}^{n-1}_{t+1} = \mathrm{DCN}(F^{n-1}_{t+1}; \mu^n_{t+1,k,g^*}). \tag{14}$$

Similar to the DAFB, we leverage the PFRB to gradually fuse features as Eq. (6).

The model DAFN-p shown in Fig. 4 replaces the DAFB with the DAFB-p. Note that we have no groundtruth of the offsets. Chan *et al.* [13] point out that the offsets and the optical flow are similar and can both be regarded as the motion vectors of objects between frames. They adopt optical flow to constrain the offsets so that they do not deviate much from the optical flow. In Fig. 4, to train for a better prediction of variances, we introduce a flow estimator to generate the optical flows,

$$\begin{aligned} f_1 &= \mathrm{FE}(I^{LR}_{t-1}, I^{LR}_t), \\ f_2 &= \mathrm{FE}(I^{LR}_{t+1}, I^{LR}_t). \end{aligned} \tag{15}$$

We set the target deviations as the differences,

$$\begin{aligned} d_1 &= |\mu^1_{t-1,k,g} - f_1|, \\ d_2 &= |\mu^1_{t+1,k,g} - f_2|, \end{aligned} \tag{16}$$

and include the second loss term L_{var},

$$\begin{aligned} L_{var} = &\sqrt{(\sigma^1_{t-1,k,g} - d_1)^2 + \varepsilon^2} \\ &+ \sqrt{(\sigma^1_{t+1,k,g} - d_2)^2 + \varepsilon^2}. \end{aligned} \tag{17}$$

The DAFN-p is trained by minimizing the overall loss function L_p,

$$L_p = L + \beta L_{var}, \tag{18}$$

where $\beta = 0.001$.

IV. Experimental Results

A. Experimental details

We select the same training dataset as previous studies [4, 5, 23], called MM522, which contains 522 groups of consecutive video frames of various scenes, including human, animal, natural and urban views. Diverse scenarios make the network robust when testing different datasets.

For BD (blur and downsampling) images, like existing approaches [2, 9, 27], we first blur the high-resolution video frames through a Gaussian filter with a standard deviation of 1.6 and convolution kernel of size 13×13, and then downsample frames through bicubic interpolation. The evaluation metrics, PSNR and SSIM values, are calculated only on the Y channel of YCbCr color space with skipping the first and last two frames.

The LR patches with the fixed size of 64×64 are randomly cropped from the BD images as the inputs, and the total number of iterations is 400 K. The initial learning rate is set to 10^{-3}, and gradually decays to 10^{-4} until the end. We set the batch size to 16 and the length of training video sequence

TABLE II: Ablation study on the DAFB and offsets propagation.

Models	PFRB	DAFB	Offsets propagation	Parameters (M)	Runtime (ms)	BD Degradation	
						UDM10 [4]	Vid4 [21]
(a)	✓			1.90	20	39.38 / 0.9647	27.87 / 0.8621
(b)		✓		4.85	53	40.07 / 0.9688	28.27 / 0.8705
(c)		✓	✓	4.85	53	40.21 / 0.9695	28.77 / 0.8787

TABLE III: Restoration performance between DAFN and DAFN-p with different groups of offsets.

Offset groups (G)	Models	Parameters (M)	Runtime (ms)	BD Degradation	
				UDM10 [4]	Vid4 [21]
2	DAFN	2.89	50	39.88 / 0.9679	28.22 / 0.8698
	DAFN-p	3.08	54	40.16 / 0.9694	28.34 / 0.8700
4	DAFN	3.54	52	39.99 / 0.9682	28.47 / 0.8746
	DAFN-p	3.86	56	40.22 / 0.9696	28.40 / 0.8722
8	DAFN	4.85	53	40.21 / 0.9695	28.77 / 0.8787
	DAFN-p	5.46	57	40.31 / 0.9699	28.51 / 0.8755

to 9 frames. All the experiments are implemented on Python 3.6.8, PyTorch 1.7.1 and V100 Tesla GPU.

B. Comparisons with other VSR models

Table I provides the quantitative evaluation for scale 4 to compare with several SOTA VSR methods. Since most methods train on different datasets, it would be unfair to solely compare results reported on their papers. This paper refers to performances evalutaed in OVSR [5], where those methods are re-trained on the MM522 dataset. The inference speed (Runtime) is measured on a V100 Tesla GPU by restoring images of HR size 1280×720.

To surpass the SOTA VSR methods, we double the number of DAFB in the DAFN, called DAFN-l. In Table I, the DAFN-l outperforms the ETDM [30], the latest VSR model, by up to 0.36 dB in PSNR. Even compared to OVSR [5], the baseline of our work, the DAFN improves PSNR by about 0.2 dB with 32% fewer parameters.

The qualitative comparisons with other SOTA methods are shown in Fig. 5. Our restoration images successfully recover more clear details than other images. For example, the city scene in the Vid4 dataset is restored particularly well. Areas that are difficult to restore, such as texture of the building, are well recovered.

C. Ablation study on DAFB and offsets propagation

We further provide some quantitative comparisons to demonstrate the effectiveness of our designs. As shown in Table II, model (a) solely stacks the PFRB for propagation. Although features from different frames are still gradually fused, the fusion performance is often disappointing since the different feature maps are not perfectly aligned.

In model (b), we replace the PFRB with our DAFB. Its performance (PSNR) is nearly 0.5 dB better than that of model (a). The improvement indicates that our proposed DAFB achieves superior restoration performance. Model (c) further passes the offsets predicted in the previous deformable alignment module to be the initial values of the current module. The performance is even improved by about 0.3dB in average, confirming the effectiveness of the offset propagations.

D. Comparisons between DAFN and DAFN-p

In Table III, we compare the restoration performance of the DAFN and DAFN-p with various offset groups $G = 2, 4, 8$. The DAFN-p outperforms the DAFN for the offset groups $G = 2$ and $G = 4$. But as the number of offset groups increases, the advantage of DAFN-p gradually becomes less significant. When the number of offset groups $G = 8$, the DAFN-p even performs slightly inferior to the DAFN. For more offset groups, even without the probability distribution, the DAFN exhibits robustness to imprecise predictions of offsets through the increase in diversity.

V. Conclusions

Most existing VSR methods rely on concatenating neighboring feature maps to obtain the complementary information from temporal correlations. In this work, we design a deformable aligned fusion block to align the neighboring feature maps and fuse the information. The proposed DAFN outperforms the SOTA models by up to 0.36 dB in PSNR on Vid4 and UDM10 datasets. We further increase the tolerance for inaccurate offsets by modeling them as probability distributions. By comparing two versions of deformable aligned fusion blocks, we observe that as the number of offset groups increases, a probabilistic model provides less advantage for robustness than a deterministic model with diversity. For future work, we need to explore more plausible combinations, including the type of probability distribution, and the number of offset groups

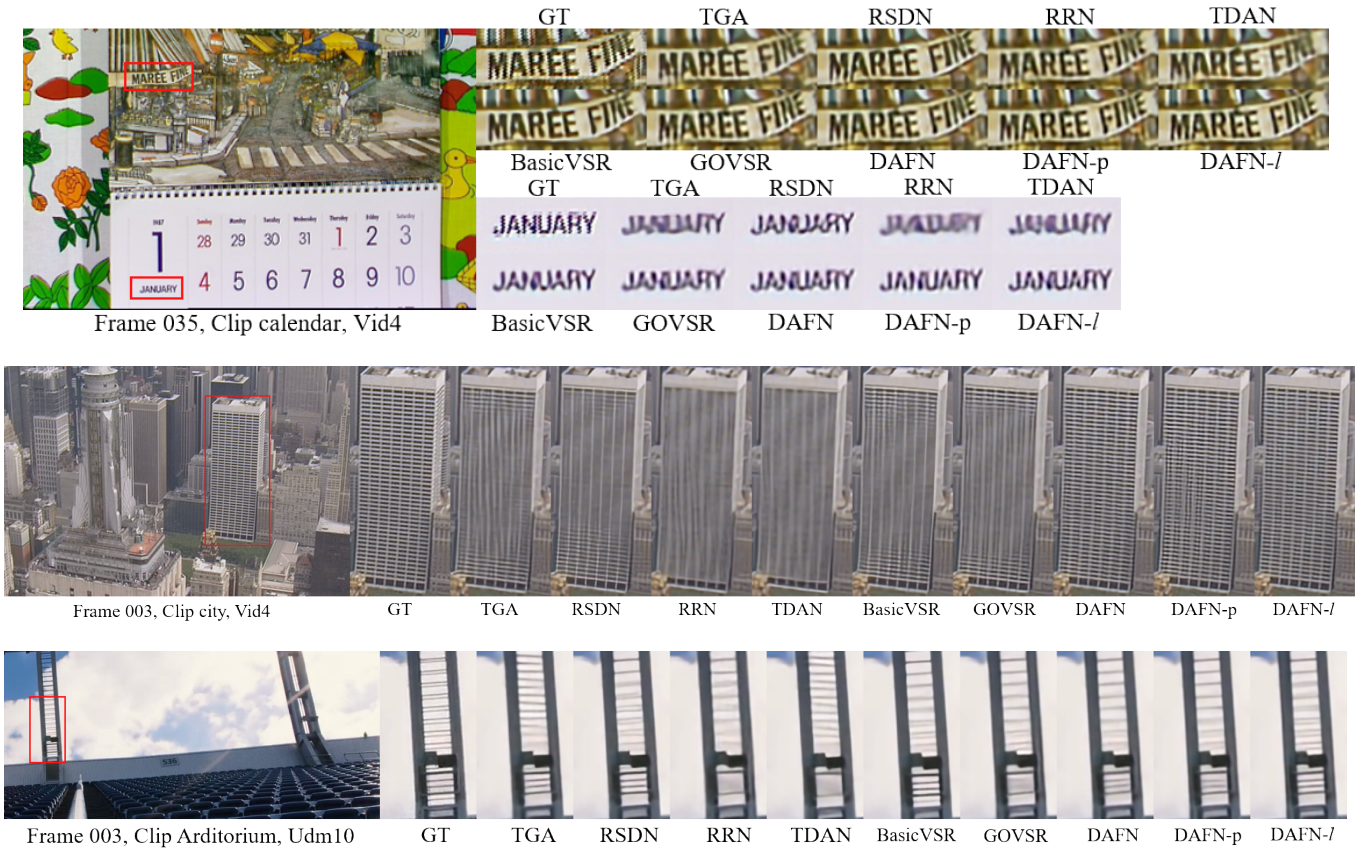


Fig. 5: Visual comparisons with other state-of-the-art VSR (X4) models.

generated through convolutional layers or selected for input into deformable convolutions.

REFERENCES

[1] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4778–4787, 2017.

[2] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232, 2018.

[3] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video Super-Resolution with Convolutional Neural Networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[4] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive Fusion Video Super-Resolution Network via Exploiting Non-Local Spatio-Temporal Correlations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3106–3115, 2019.

[5] P. Yi, Z. Wang, K. Jiang, J. Jiang, T. Lu, X. Tian, and J. Ma, "Omniscient Video Super-Resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4429–4438, 2021.

[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, 2017.

[7] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3360–3369, 2020.

[8] Z. Luo, L. Yu, X. Mo, Y. Li, L. Jia, H. Fan, J. Sun, and S. Liu, "EBSR: Feature Enhanced Burst Super-Resolution with Deformable Alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 471–478, 2021.

[9] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video Restoration with Enhanced Deformable Convolutional Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[10] Z. Luo, Y. Li, S. Cheng, L. Yu, Q. Wu, Z. Wen, H. Fan, J. Sun, and S. Liu, "BSRT: Improving Burst Super-Resolution with Swin Transformer and Flow-Guided De-

formable Alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 998–1008, 2022.

[11] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5972–5981, 2022.

[12] D. Fuoli, M. Danelljan, R. Timofte, and L. Van Gool, "Fast Online Video Super-Resolution with Deformable Attention Pyramid," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1735–1744, 2023.

[13] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding Deformable Alignment in Video Super-Resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 973–981, 2021.

[14] C. Zhang and B. Jin, "Probabilistic Residual Learning for Aleatoric Uncertainty in Image Restoration," *arXiv preprint arXiv:1908.01010*, 2019.

[15] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep Stereo Using Adaptive Thin Volume Representation with Uncertainty Awareness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2524–2534, 2020.

[16] Z. Zhang, C. Lan, W. Zeng, Z. Chen, and S.-F. Chang, "Uncertainty-Aware Few-Shot Image Classification," *arXiv preprint arXiv:2010.04525*, 2020.

[17] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for Video Super-Resolution Through HR Optical Flow Estimation," in *Asian Conference on Computer Vision*, pp. 514–529, Springer, 2018.

[18] J. Lin, Y. Huang, and L. Wang, "FDAN: Flow-Guided Deformable Alignment Network for Video Super-Resolution," *arXiv preprint arXiv:2105.05640*, 2021.

[19] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-Recurrent Video Super-Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6626–6634, 2018.

[20] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4947–4956, 2021.

[21] C. Liu and D. Sun, "On Bayesian Adaptive Video Super Resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2013.

[22] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video Enhancement with Task-Oriented Flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[23] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma, "Multi-Memory Convolutional Neural Network for Video Super-Resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2530–2544, 2018.

[24] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent Back-Projection Network for Video Super-Resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3897–3906, 2019.

[25] T. Isobe, S. Li, X. Jia, S. Yuan, G. Slabaugh, C. Xu, Y.-L. Li, S. Wang, and Q. Tian, "Video Super-Resolution with Temporal Group Attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8008–8017, 2020.

[26] D. Fuoli, S. Gu, and R. Timofte, "Efficient Video Super-Resolution through Recurrent Latent Space Propagation," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3476–3485, IEEE, 2019.

[27] T. Isobe, X. Jia, S. Gu, S. Li, S. Wang, and Q. Tian, "Video Super-Resolution with Recurrent Structure-Detail Network," in *European Conference on Computer Vision*, pp. 645–660, Springer, 2020.

[28] T. Isobe, F. Zhu, X. Jia, and S. Wang, "Revisiting Temporal Modeling for Video Super-Resolution," *arXiv preprint arXiv:2008.05765*, 2020.

[29] B. Yan, C. Lin, and W. Tan, "Frame and Feature-Context Video Super-Resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5597–5604, 2019.

[30] T. Isobe, X. Jia, X. Tao, C. Li, R. Li, Y. Shi, J. Mu, H. Lu, and Y.-W. Tai, "Look Back and Forth: Video Super-Resolution with Explicit Temporal Difference Modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17411–17420, 2022.

Learning Single Image Rain Streak Removal Based on Deep Attention Mechanism

Kuan-Hua Huang and Li-Wei Kang*
Department of Electrical Engineering
National Taiwan Normal University
Taipei, Taiwan
*E-mail: lwkang@ntnu.edu.tw

***Abstract* — Bad weather conditions (e.g., rain or hazy) may significantly degrade the visual quality of captured images/videos and the performances of related applications (e.g., outdoor visual surveillance). To solve this problem, this paper presents to learn rain steak removal from a single image. By using the ECNet (Embedding Consistency Network, by Li *et al*., 2022) as our basis network architecture, a deep encoder-decoder-based network with channel attention and the proposed multi-scale pixel attention module (MSPAM) is presented to single image rain streak removal, *i.e.*, deraining. Together with the "Rain Embedding Consistency" mechanism used in the ECNet, we have shown that the channel attention can be used to enhance the extracted features before being fed into the encoder, and our MSPAM can be embedded into the skip connection between the encoder and the decoder for further boosting the features to achieve better image reconstruction. Experimental results have demonstrated that the proposed framework outperforms the ECNet quantitatively and qualitatively.**

I. INTRODUCTION

Bad weather conditions, such as rain, haze, or snow, may significantly degrade the visual quality of outdoor-captured images/videos, and the performances of applications inspired by outdoor visual information [1], [2]. Therefore, single image-based deraining *i.e*., removing rain streaks from a single image, has received much attention recently [3], [4]. We have proposed the first research work for single image deraining in the literature, which is based on sparse signal decomposition without needing rain/non-rain image pairs [5], [6] (with extension to video-based rain streak removal in [7]). This work [5] was also further improved by introducing context-aware image segmentation [8], [9], and visual saliency detection [10], [11], respectively. Moreover, similar to sparse coding, the low-rank-based optimization mechanism was also popular to be used for single image deraining [12]. In addition, non-negative matrix factorization has also been shown to be effective to single image rain removal in [13].

As the advancement of deep learning techniques has achieved significant improvement in various perceptual tasks, such as image restoration [14], image classification [15], object detection [16], and several vision-based applications (e.g., autonomous driving [17]), and sufficient datasets of rain/non-rain image pairs have been available (e.g., [18]-[20]), several deep learning-based single image deraining frameworks have been presented [18]–[27]. The deep learning-based approaches have been shown to achieve state-of-the-art performances in single image deraining, which will be briefly reviewed in Sec. II.

To achieve better image deraining performance, an encoder-decoder-based single image rain removal framework with deep attention mechanism is proposed in this paper. By leveraging the rain embedding consistency mechanism presented in ECNet [27], the encoder can encode a higher quality rain embedding for better decoding performance. In the proposed framework, we have shown that applying channel attention to the feature map before being fed into the encoder would be beneficial for extracting rain streak features. Moreover, embedding the proposed MSPAM (multi-scale pixel attention module) into the skip connections between the encoder and the decoder would be useful for extracting the rain-attentive features for better reconstructing rain streaks $\hat{R}$. The finally rain streak-removed image can be obtained by $I - \hat{R}$, where I denotes the input rainy image.

The rest of this paper is organized as follows. Sec. II briefly reviews related works in deep learning-based single image rain streak removal. In Sec. III, we address the proposed deep encoder-decoder network with our MSPAM for rain streak removal from a single image. Experimental results are demonstrated in Sec. IV. Finally, Sec. V concludes this paper.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

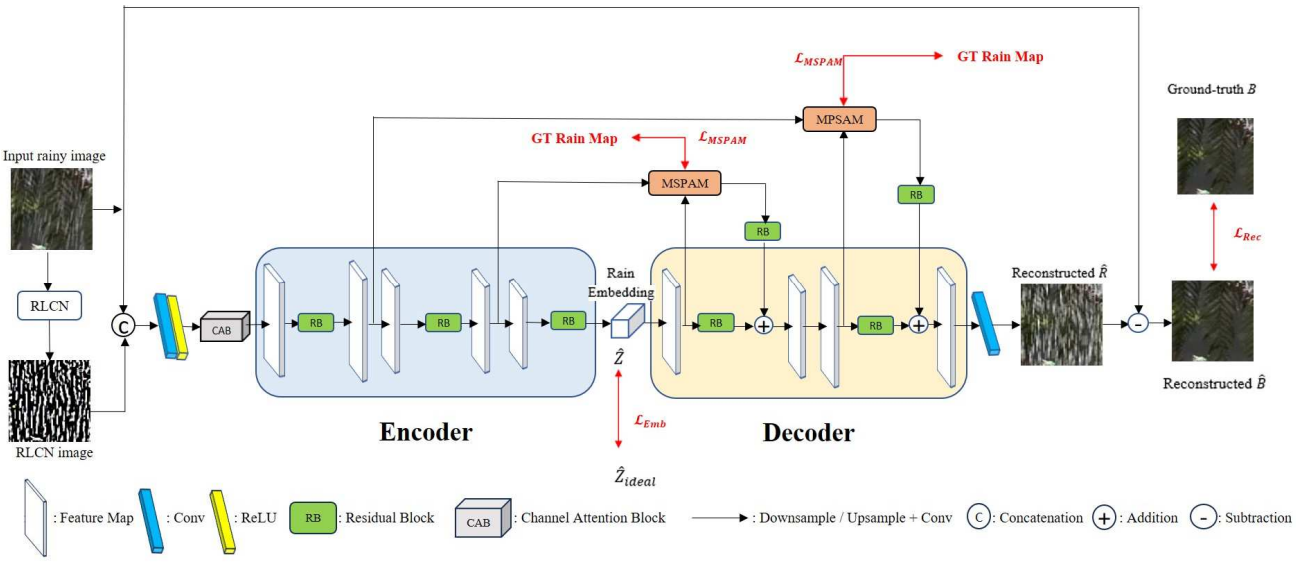


Fig. 1. Proposed deep encoder-decoder network for single image rain streak removal (modified from [27]).

II. Related Works

A. Deep Learning-based Single Image Deraining

A deep convolutional neural network (CNN) model, trained based on the mapping between paired images, embedded in an image decomposition framework (similar to [5]) was proposed for single image deraining in [21]. Moreover, a multi-task deep learning network was presented in [22] for detection and removal of rain streaks from a single rain image by learning the map of binary rain streaks, the appearance of rain streaks, and the clean background. Moreover, we have proposed a sequential dual attention-based single image deraining deep network (SSDRNet) [24] relying on a two-stage training strategy designed for capturing the rain streak distribution in a rainy image. On the other hand, a conditional generative adversarial networks (GAN)-based single image deraining framework was proposed in [23].

B. Embedding Consistency Network (ECNet)

In ECNet [27], an encoder-decoder architecture was presented to learn rain streaks, where the key is to include a "Rain Embedding Consistency" mechanism. The idea is to use the ideal encoded rain embedding obtained by an autoencoder and try to maintain the consistency between the ideal embedding and the real rain embedding derived by the network encoder. A rain embedding loss was used to supervise this encoding process, and a RLCN (rectified local contrast normalization) module was used to guide the extraction of candidate rain pixels. In this paper, by using ECNet as our basis, we propose to apply channel attention and the presented MSPAM (multi-scale pixel attention module) to further boost the feature extraction and the decoding process, described in Sec. III.

III. Proposed Single Image Deraining Framework

A. Overview of the Proposed Framework

As shown in Fig. 1, we adopt an encoder-decoder-based deep network architecture, where the residual block (RB) [28] is used for feature extraction, based on [27]. At the input side, each input rainy image I and its image I_L of candidate rain pixels obtained by the RLCN module presented in [27] are concatenated, denoted by $I_J = Concat(I, I_L)$. The map I_J is then fed into a convolution layer including the ReLU (rectified linear units) function [29], denoted by $J = \text{ReLU}\left(Conv(I_J)\right)$. Moreover, the input image I is also delivered through a skip connection to subtract the finally learned rain streak map $\hat{R}$ to get the finally rain streak-removed result, denoted by $\hat{B} = I - \hat{R}$. Different from [27], before feeding the feature map J into the encoder, we first apply the channel attention block (CAB) presented in [30] to obtain the channel-attentive feature map, denoted by $\text{CAB}(J)$. Then the encoder E encodes the feature map $\text{CAB}(J)$ to obtain the rain embedding, denoted by $\hat{Z} = E\left(\text{CAB}(J)\right)$. To learn the rain embedding, the autoencoder presented in [27] is used to generate the ideal rain embedding $\hat{Z}_{ideal}$. Then, the predicted rain streak map $\hat{R}$ can be obtained by the decoder D, denoted by $\hat{R} = D(\hat{Z})$. Moreover, for each skip connection between the encoder and the decoder, we

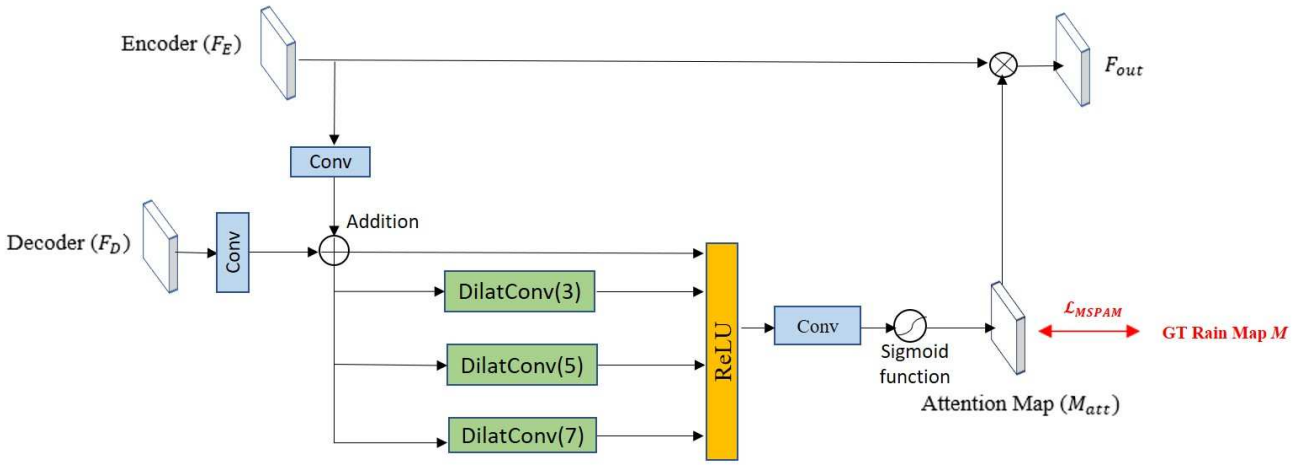


Fig. 2. Proposed multi-scale pixel attention module (MSPAM), where DilatConv(x) denotes the dilated convolution operation with the dilation rate of x.

embed the proposed MSPAM (described in Sec. III.B) to obtain the rain-attentive feature map for better enhancing the decoding process.

B. *Proposed Multi-Scale Pixel Attention Module*

The proposed multi-scale pixel attention module (MSPAM) is used to generate a pixel-level attention map to further refine the features, especially for suppressing possibly unexpected features extracted from the RLCN image which may include some non-rain pixels. As shown in Fig. 2, both of the feature maps F_E and F_D, respectively, from the encoder and the decoder first go through the convolution layers, and then are added. Then, we apply the dilated convolution operations (DilatConv) of the four different dilation rates to extract multi-scale features, followed by the ReLU and convolution operations. Then, the Sigmoid function is used to adjust the feature values to the range between 0 and 1 for generating the attention map M_{att}. The attention map is then used to generate the rain attentive-feature map, denoted by $F_{out} = F_E \otimes M_{att}$, where $\otimes$ is the element-wise multiplication operation.

C. *Leaning the Proposed Deep Model*

To learn the proposed deep model for single image deraining, the loss function is defined as follows. To supervise the reconstruction of background (rain-removed) image $\hat{B}$, the loss function is defined by

$$L_{Rec} = -SSIM(B, \hat{B}), \quad (1)$$

where $SSIM$ denotes the structural similarity (SSIM) index [31] between B and $\hat{B}$, and B is the ground truth (GT)/clean image. In addition, to guide our MSPAM to generate the rain-attentive feature map, the loss L_{MSPAM} between the ground truth of rain streak map/mask M and the predicted attention map M_{att} is defined accordingly. Moreover, the ideal rain embedding $\hat{Z}_{ideal}$ generated by the autoencoder is used to supervise the encoded rain embedding $\hat{Z}$, where the loss between them is denoted by L_{Emb}. As a result, the used total loss function is defined as

$$L = \lambda_{Rec} L_{Rec} + \lambda_{MSPAM} L_{MSPAM} + \lambda_{Emb} L_{Emb}, \quad (2)$$

where λ_{Rec}, λ_{MSPAM}, and λ_{Emb} are the respective weighting coefficients, empirically set to 1, 0.1, and 0.02, respectively.

IV. Experimental Results

To evaluate the proposed deep learning framework for single image deraining, the Rain100H dataset [18], Rain200H dataset [18], and Rain800 dataset [23] were used for training and testing our deep model. The Rain100H dataset contains 1800 training image pairs and 100 testing image pairs. The Rain200H dataset contains 1800 training image pairs and 200 testing image pairs. The Rain800 dataset contains 700 training image pairs and 98 testing image pairs. The patch size of 96×96 was used in the training process. The Adam optimizer [32] was used for model optimization (by 100 epochs) with the batch size set to 4, and the learning rate initialized to 0.001, which is decayed by multiplying 0.2 in every 25 epochs. We quantitatively (in terms of PSNR (in dB) and SSIM [31]) and qualitatively compared our method with the three state-of-the-art methods, PReNet [19], MPRNet [26], and ECNet [27]. The quantitative results were shown in Tables I-III, and the qualitative results were shown in Figs. 3-7. It has been shown that the proposed method recovers more image details while removing the rain streaks, and exhibits better quantitative results.

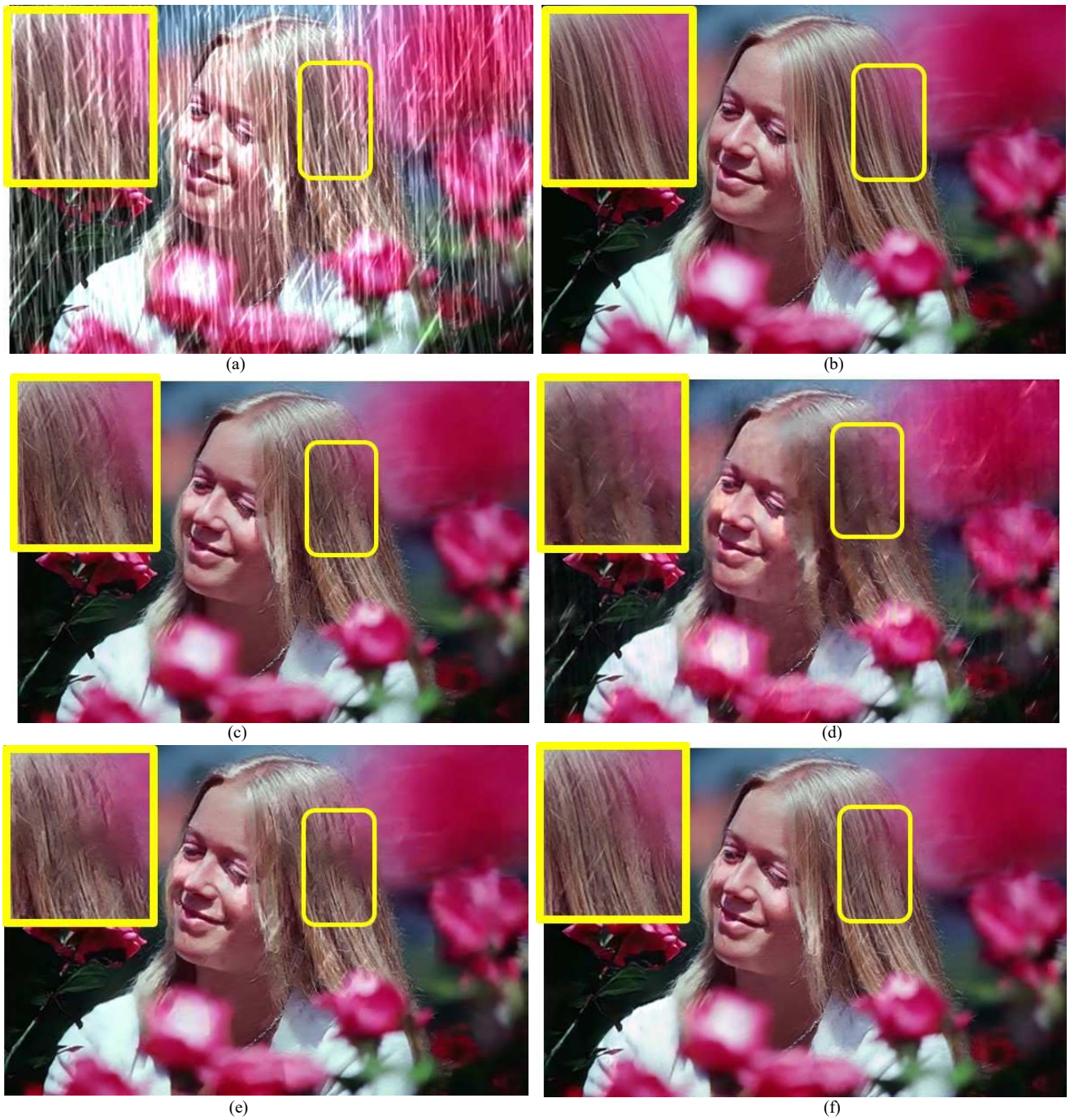


Fig. 3. Qualitative results on the Rain100H dataset: (a) rainy image; (b) ground truth; (c) PReNet [19]; (d) MPRNet [26], (e) ECNet [27]; and (f) proposed method.

Moreover, to evaluate the effectiveness of the embedded CAB (channel attention block) and the proposed MSPAM (multi-scale pixel attention module), we conducted the ablation studies by removing either CAB or MSPAM from the proposed deep rain streak removal framework, and evaluated the corresponding deraining performances, reported in Tables IV and V. Based on Tables IV and V, removing either CAB or MSPAM will degrade the deraining performances, and the proposed complete deep deraining network exhibits the best performance, which verifies the effectiveness of the two modules, CAB and MSPAM.

V. Conclusions

In this paper, we have proposed a deep learning-based encoder-decoder framework to learn rain steaks for single image-based deraining. By using ECNet [27] as our basis, an

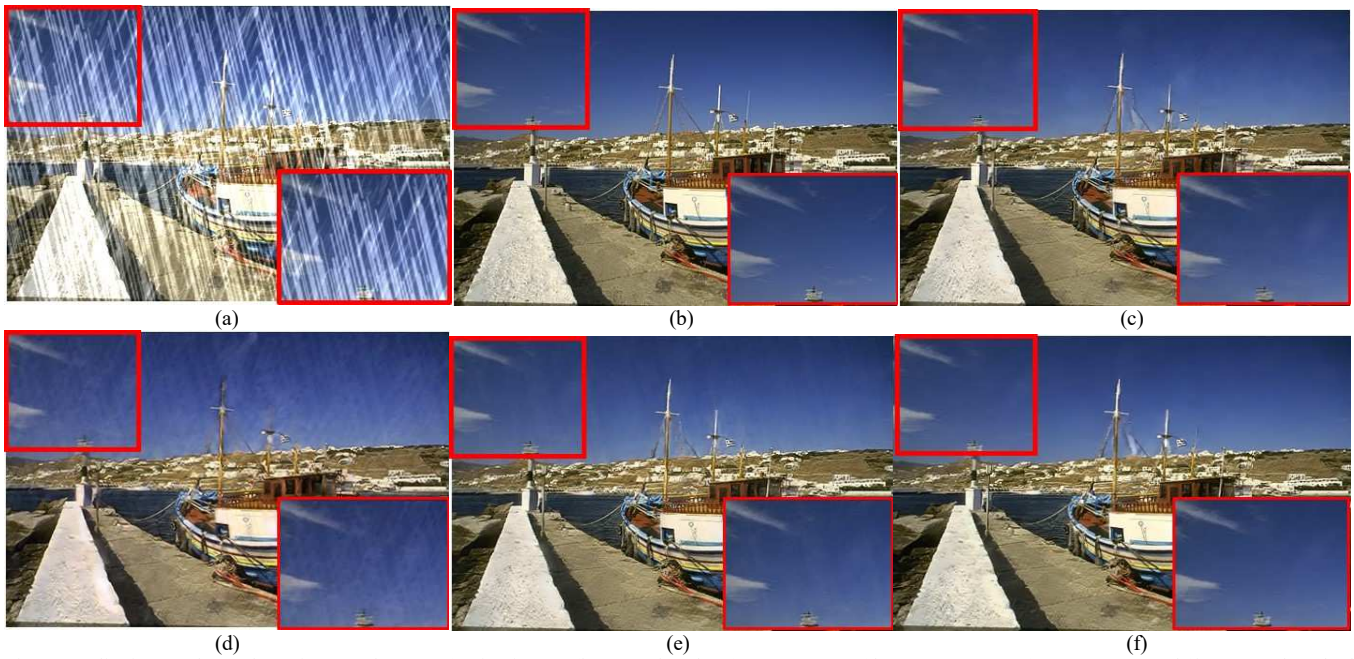


Fig. 4. Qualitative results on the Rain200H dataset: (a) rainy image; (b) ground truth; (c) PReNet [19]; (d) MPRNet [26], (e) ECNet [27]; and (f) proposed method.

encoder-decoder-based network with channel attention and the proposed multi-scale pixel attention module (MSPAM) was presented. Together with the "Rain Embedding Consistency" mechanism presented in the ECNet, we have shown that the channel attention can be used to enhance the extracted features before being fed into the encoder, and our MSPAM can be embedded into the skip connection between the encoder and the decoder for further boosting the features to achieve better image reconstruction. Experimental results have demonstrated that the proposed framework outperforms the three compared state-of-the-art deep learning-based deraining methods quantitatively and qualitatively.

ACKNOWLEDGMENT

This work was supported in part by National Science and Technology Council, Taiwan, under the Grant NSTC 111-2221-E-003-019-MY3.

REFERENCES

[1] K. Garg and S. K. Nayar, "Vision and rain," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 3–27, 2007.

[2] C.-H. Yeh, C.-H. Huang, and L.-W. Kang, "Multi-scale deep residual learning-based single image haze removal via image decomposition," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 3153–3167, Dec. 2020.

[3] S. Li, W. Ren, F. Wang, I. B. Araujo, E. K. Tokuda, R. H. Junior, R. M. Cesar-Jr., Z. Wang, and X. Cao, "A comprehensive benchmark analysis of single image deraining: Current challenges and future perspectives," *Int. J. Computer Vision*, vol. 129, pp. 1301–1322, 2021.

[4] Y. Yu, W. Yang, Y. Tan, and A. C. Kot, "Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 6003-6012.

[5] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1742–1755, Apr. 2012.

[6] Y.-H. Fu, L.-W. Kang, C.-W. Lin, and C.-T. Hsu, "Single-frame-based rain removal via image decomposition," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Prague, Czech Republic, 2011, pp. 1453-1456.

[7] L.-W. Kang, C.-W. Lin, C.-T. Lin, and Y.-C. Lin, "Self-learning-based rain streak removal for image/video," *Proc. IEEE Int. Sym. Circuits and Systems*, Seoul, Korea, 2012, pp. 1871-1874.

[8] D.-A. Huang, L.-W. Kang, Y.-C. F. Wang, and C.-W. Lin, "Self-learning based image decomposition with applications to single image denoising," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 83–93, Jan. 2014.

[9] D.-A. Huang, L.-W. Kang, M.-C. Yang, C.-W. Lin, and Y.-C. F. Wang, "Context-aware single image rain removal," *Proc. IEEE Int. Conf. Multimedia and Expo*, Melbourne, Australia, 2012, pp. 164-169.

[10] D.-Y. Chen, C.-C. Chen, and L.-W. Kang, "Visual depth guided color image rain streaks removal using sparse coding," *IEEE*

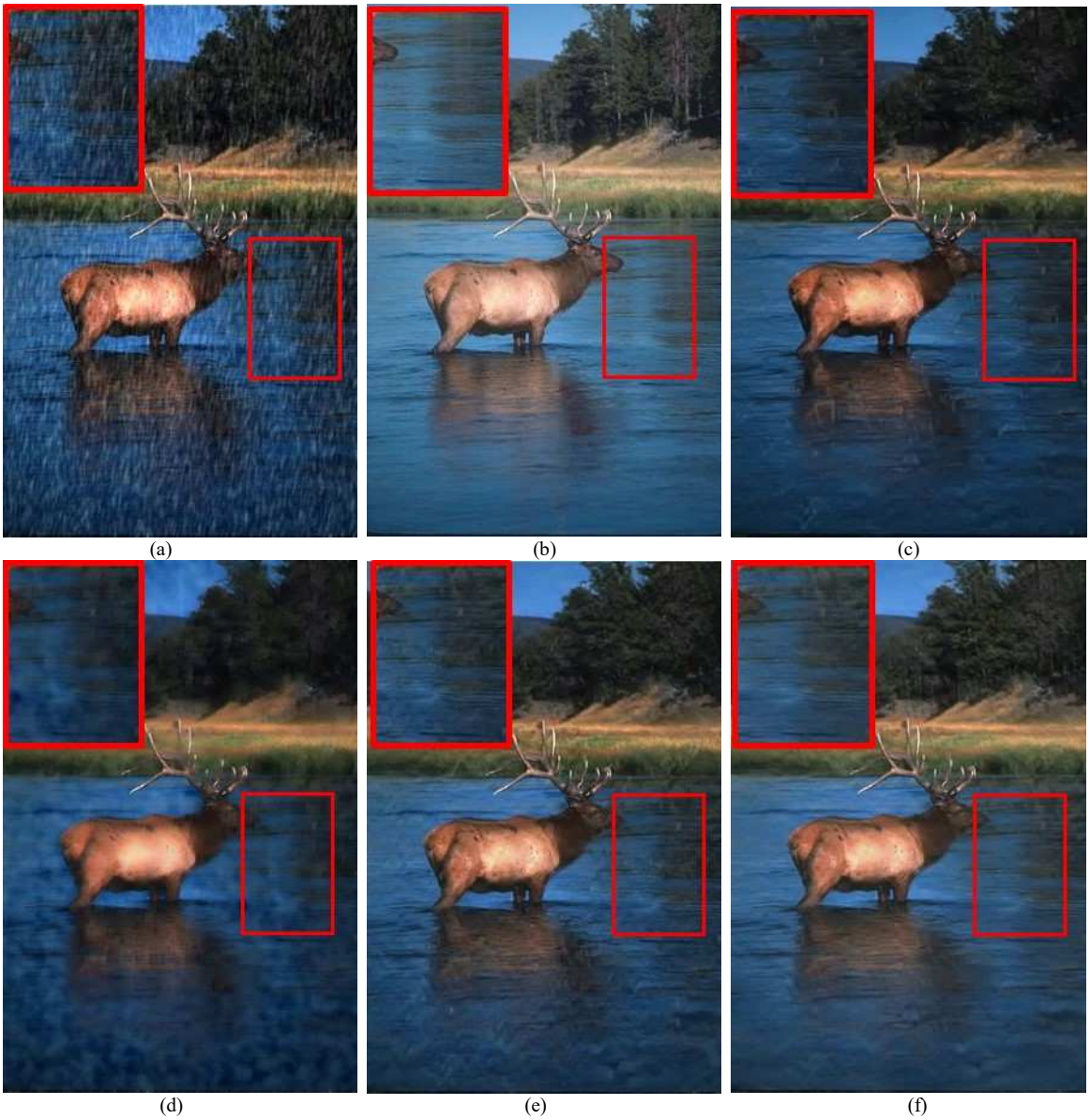


Fig. 5. Qualitative results on the Rain800 dataset: (a) rainy image; (b) ground truth; (c) PReNet [19]; (d) MPRNet [26], (e) ECNet [27]; and (f) proposed method.

Trans. Circuits Syst. Video Techn., vol. 24, no. 8, pp. 1430–1455, Aug. 2014.

[11] D.-Y. Chen, C.-C. Chen, and L.-W. Kang, "Visual depth guided image rain streaks removal via sparse coding," *Proc. IEEE Int. Sym. Intelligent Signal Processing and Communications Systems*, Tamsui, Taiwan, 2012, pp. 151-156.

[12] Y.-L. Chen and C.-T. Hsu, "A generalized low-rank appearance model for spatio-temporally correlated rain streaks," *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, Dec. 2013, pp. 1968–1975.

[13] P.-H. Liu, C.-Y. Lin, C.-H. Yeh, L.-W. Kang, K. S.-H. Lo, T.-H. Hwang, and C.-C. Kuo, "Rain removal using single image based on non-negative matrix factorization," Intelligent Systems and Applications, pp. 1137-1146, 2015.

[14] C.-H. Yeh, C.-H. Lin, M.-H. Lin, L.-W. Kang, C.-H. Huang, and M.-J. Chen, "Deep learning-based compressed image artifacts reduction based on multi-scale image fusion," *Information Fusion*, vol. 67, pp. 195-207, March 2021.

[15] C.-H. Yeh, M.-H. Lin, P.-C. Chang, and L.-W. Kang, "Enhanced visual attention-guided deep neural networks for image classification," *IEEE Access*, vol. 8, pp. 163447-163457, 2020.

[16] C.-H. Yeh, C.-H. Lin, L.-W. Kang, C.-H. Huang, M.-H. Lin, and C.-Y. Chang, and C.-C. Wang, "Lightweight deep neural network for joint learning of underwater object detection and color



(a) (b) (c) (d)

Fig. 6. Qualitative results on the Rain800 dataset: (a) rainy image; (b) ground truth; (c) ECNet [27]; and (d) proposed method.

TABLE I. Quantitative Results on Rain100H Dataset.

Method	PSNR (in dB)	SSIM
PReNet [19]	30.07	0.904
MPRNet [26]	27.59	0.834
ECNet [27]	29.53	0.900
Proposed	**30.11**	**0.908**

TABLE II. Quantitative Results on Rain200H Dataset.

Method	PSNR (in dB)	SSIM
PReNet [19]	28.94	0.900
MPRNet [26]	26.98	0.882
ECNet [27]	28.38	0.892
Proposed	**29.13**	**0.901**

TABLE III. Quantitative Results on Rain800 Dataset.

Method	PSNR (in dB)	SSIM
PReNet [19]	26.33	0.891
MPRNet [26]	23.84	0.804
ECNet [27]	26.67	0.879
Proposed	**27.78**	**0.897**

TABLE VI. Ablation studies on Rain100H Dataset.

CAB	MSPAM	PSNR (in dB)	SSIM
×	×	29.53	0.900
✓	×	29.65	0.903
✓	✓	**30.11**	**0.908**

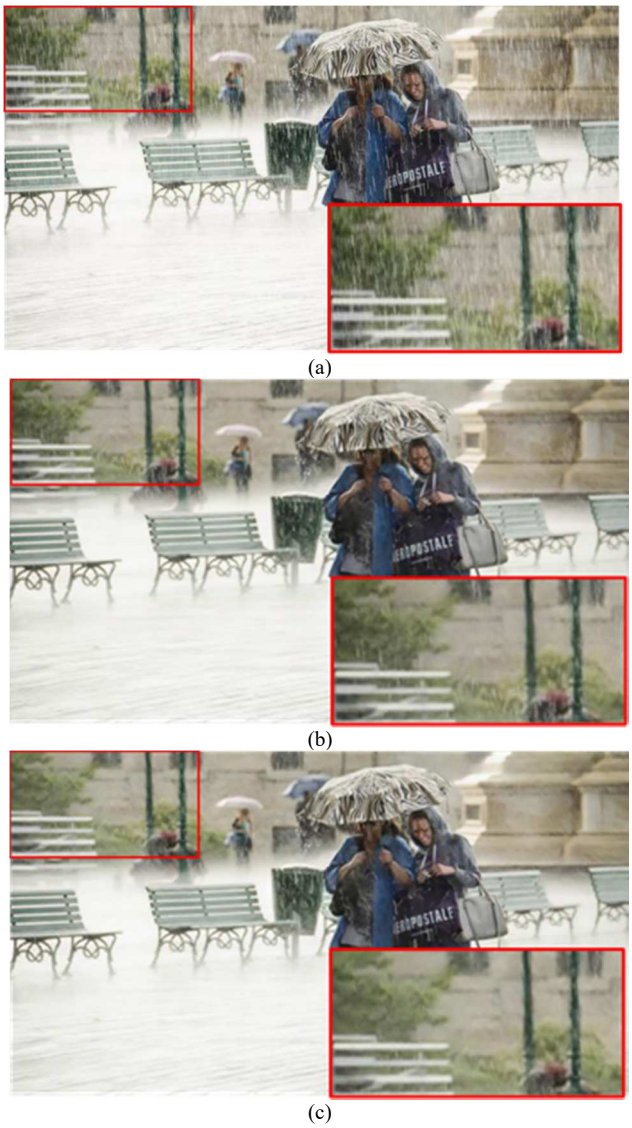
TABLE V. Ablation studies on Rain200H Dataset.

CAB	MSPAM	PSNR (in dB)	SSIM
×	×	28.38	0.892
✓	×	28.49	0.896
✓	✓	**29.13**	**0.901**

conversion," *IEEE Trans. Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6129-6143, Nov. 2022.

[17] C.-C. Hsu, L.-W. Kang, S.-Y. Chen, I-S. Wang, C.-H. Hong, and C.-Y. Chang, "Deep learning-based vehicle trajectory prediction based on generative adversarial network for autonomous driving applications," *Multimedia Tools and Applications*, vol. 82, pp. 10763–10780, 2023.

[18] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1685-1694.



(a) (b) (c)

Fig. 7. Qualitative results for natural image (no ground truth): (a) rainy image; (b) ECNet [27]; and (c) proposed method.

[19] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3937–3946.

[20] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. H. Lau, "Spatial attentive single-image deraining with a high quality real rain dataset," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2019, pp. 12270–12279.

[21] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2944–2956, June 2017.

[22] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Joint rain detection and removal from a single image with contextualized deep networks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 6, pp. 1377–1393, June 2020.

[23] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3943-3956, Nov. 2020.

[24] C.-Y. Lin, Z. Tao, A.-S. Xu, L.-W. Kang, and F. Akhyar, "Sequential dual attention network for rain streak removal in a single image," *IEEE Trans. Image Processing*, vol. 29, pp. 9250-9265, 2020.

[25] G. Wang, C. Sun and A. Sowmya, "Attentive feature refinement network for single rainy image restoration," *IEEE Trans. Image Processing*, vol. 30, pp. 3734-3747, 2021.

[26] S. W. Zamir et al., "Multi-stage progressive image restoration," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 14816-14826.

[27] Y. Li, Y. Monno, and M. Okutomi, "Single image deraining network with rain embedding consistency and layered LSTM," *Proc. IEEE Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2022, pp. 3957-3966.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proc. Int. Conf. Machine Learning*, Haifa, Israel, June 2010, pp. 807–814.

[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks", *Proc. European Conf. Computer Vision*, Munich, Germany, Sept. 2018.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process*., vol. 13, no. 4, pp. 600−612, 2004.

[32] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *Proc. Int. Conf. Learning Representations*, May 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Transformer-Based Framework for Tiny Object Detection

Yi-Kai Liao, Gong-Si Lin, Mei-Chen Yeh
Department of Computer Science and Information Engineering
National Taiwan Normal University, Taiwan
E-mail: {40947038s, myeh}@gapps.ntnu.edu.tw

***Abstract*—This paper proposes a fully transformer-based method for building an end-to-end model dedicated to tiny object detection. Our approach eliminates the components which are difficult to be designed in detecting tiny objects, such as anchor generation and non-maximum suppression. Additionally, we address the issue of receptive fields for tiny objects in convolutional neural networks through self-attention. The model named Swin-Deformable DEtection TRansformer (SD DETR) integrates Swin Transformer [1] and Deformable DETR [2]. Furthermore, we have introduced architectural enhancements and optimized the loss function to improve the model's ability in detecting tiny objects. Experimental results on the AI-TOD [3] dataset demonstrate that SD DETR achieves 10.9 AP for very tiny objects with only 2 to 4 pixels, showcasing a significant improvement of +1.2 AP compared to the current state-of-the-art model. The code is available at https://github.com/kai271828/SD-DERT**

I. Introduction

With the increasing maturity of drone aerial photography and satellite imaging technology, the demand for object detection in aerial images is rapidly growing. In recent years, significant progress has been made in the field of object detection, but the current main achievements are focused on detecting objects of general size. Successfully detecting tiny objects that often appear in aerial images is a challenging problem. Unlike detecting objects of general size, several difficulties exist in detecting tiny objects smaller than 64 pixels [3], including being easily affected by background noise, insufficient training samples, the limitation of using bounding boxes for tiny objects, and the downsampling process tpyically in convolutional neural network methods causing tiny objects to disappear in the final feature map. All of the factors mentioned above result in a significant decrease in the performance of detecting tiny objects.

Currently, the most representative solutions to the problem of detecting tiny objects include Normalized Gaussian Wasserstein Distance (NWD) [4] and Gaussian Receptive Field based Label Assignment (RFLA) [5]. Both methods utilize probabilistic modeling to alleviate the issue of imbalanced positive and negative samples caused by the IoU metric in the task of detecting tiny objects, enabling the model to successfully learn the features of tiny objects. While these methods have shown some success, overall, the task of object detection becomes more complex by incorporating additional probabilistic modeling in those hand-designed modules. Furthermore, there are still unresolved issues related to feature extraction, especially from images containing tiny objects.

We have observed that DETR [6] applied the Transformer model [7]—commonly used in natural language processing—to the field of object detection. Benefiting from the architecture of Transformer, DETR does not contain modules such as anchor generation, label assignment, and non-maximum suppression, which pose design challenges in the task of detecting tiny objects. Nevertheless, the DETR model performs poorly on small objects and has a slower convergence speed, as experimented and evidenced in [6]. To address these issues, many improved methods have been proposed, including Deformable DETR [2]. The multi-scale deformable attention mechanism proposed in [2] improves performance on small objects without using the Feature Pyramid Network (FPN) [8]. In addition, the Deformable DETR model modifies the self-attention mechanism by referring to Deformable Convolutional Networks [9] to accelerate model convergence. However, the performance on tiny objects still needs further investigation.

In addition, Transformers have recently been shown to be very effective backbone networks and outperform traditional convolutional neural networks in various visual tasks. For instance, Swin Transformer [1] uses a hierarchical design and shifted windows calculation to address the issues of Vision Transformer (ViT) [10] in object detection and image segmentation. Experimental results have shown that, under comparable model sizes, extensive pretraining allows the Swin Transformer to outperform commonly used backbone networks like ResNet [11]. We hypothesize that, in this study, the global self-attention mechanism of the Transformer contributes to mitigating the challenges associated with feature extraction for tiny objects.

Given the significant utility of detecting tiny objects and the potential of Transformers in object detection development, this paper aims to demonstrate that by using the elegant design of the DETR series models we can abandon hand-designed modules and achieve a concise and powerful end-to-end model for detecting tiny objects. Additionally, we investigate how the DETR series models can be improved for the detection of tiny objects.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Related work

A. Tiny Object Detection

Small objects typically occupy very few pixels in size. The current IoU-based metrics are very sensitive to variations in the position of small objects, which significantly affect the effectiveness of training. To alleviate this issue, Wang *et al.* [4] proposed a new metric based on Wasserstein distance. Specifically, bounding boxes (BBoxes) are modeled as Gaussian distributions, and a new measurement called Normalized Wasserstein Distance (NWD) is used to measure the similarity between the corresponding Gaussian distributions. NWD can be easily integrated into the label assignment, non-maximum suppression, and loss function of any anchor-based detector, replacing the commonly used IoU metrics.

Additionally, Xu *et al.* [5] made two observations. First, for tiny objects, existing label assignment methods heavily rely on overlapping areas. The ground truth hardly overlaps with any anchor boxes (i.e., IoU = 0), resulting in lack of positive samples for tiny objects. Therefore, heuristic methods are employed to ensure more positive samples are obtained for tiny objects. However, the allocator often struggles to compensate for positive samples based on zero-value IoU or centrality, leading to reduced attention to learning tiny objects. Second, the prior region follows a uniform distribution and treats each position within the prior region equally. However, prior information is essentially used to assist in label and feature point assignments. When mapping the feature points' receptive field back to the input image, the effective receptive field becomes a Gaussian distribution. The discrepancy between the uniform distribution of the prior and the Gaussian distribution of the receptive field causes a mismatch between the ground truth and the assigned receptive field for feature points.

To alleviate those issues, prior information based on Gaussian distribution is proposed and a label assignment strategy based on Gaussian Receptive Field is established [5], which is more beneficial for tiny objects.

B. Deformable DETR

DETR [6] has realized an end-to-end object detection model using transformers. Even after removing some modules that require lots of prior knowledge, DETR still performs remarkably well. However, this model also exhibits notable issues that need to be addressed, including poor detection performance for small objects and slow convergence speed. Therefore, Zhu *et al.* [2] proposed Deformable DETR, which effectively processes the features of images by only calculating a small number of key points through multi-scale deformable self-attention modules. Deformable DETR solves the problems of poor detection performance for small objects and slow convergence speed. It can surpass the performance of DETR with only one-tenth of the training times. Therefore, we developed our object detector based on this detection model.

C. Swin Transformer

Liu *et al.* [1] proposed shifted windows based on ViT [10]. By restricting the computation of self-attention mechanisms to non-overlapping regions while allowing cross-region connections, Swin Transformer can extract global features from images. The computational complexity is also reduced from being quadratic with respect to the input size to linearly related. In addition, the hierarchical structure enables Swin Transformer to extract multi-scale features, making it possible to use the transformer as a backbone network for various computer vision tasks. Experimental results also show that with sufficient pre-training, Swin Transformer performs well as a backbone network for object detection and semantic segmentation tasks. In this study, we investigate the effectiveness of this backbone for tiny object detection.

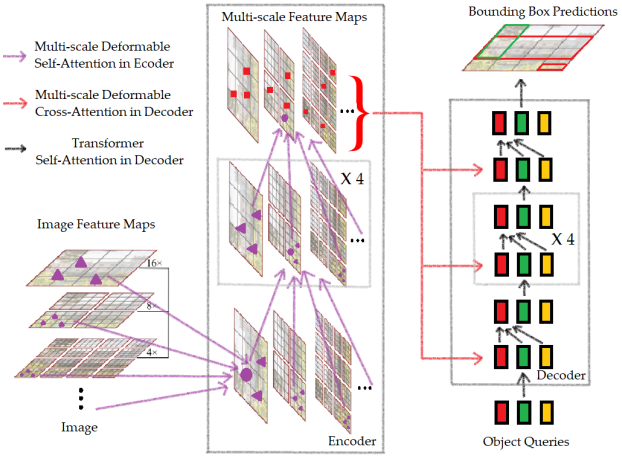


Fig. 1. Model architecture of the proposed tiny object detector. This model combines Swin Transformer [1] and Deformable DETR [2].

III. Method

The proposed tiny object detection model is developed upon Swin Transformer [1] and Deformable DETR [2], as displayed in Fig. 1. In the following we describe the technical details of our implementation.

A. Feature Extraction

As a backbone network, the Swin Transformer divides the input 800×800 image into 4×4 patches. Each patch is then transformed into a token through linear embedding. However, directly feeding these long tokens into the Transformer would result in excessively long computation time for self-attention, which is clearly unacceptable. Therefore, within a Swin Transformer Block, the tensor is divided into several 7×7 windows, and the self-attention computation is performed only within each window. As a result, the length of the tokens is fixed at 49. Subsequently, the windows are shifted, and another round of self-attention computation is performed to facilitate information exchange between windows. Finally, an output tensor of the same dimension is obtained, which serves as part of the feature map input for the subsequent Transformer model.

To incorporate features from other scales, the output tensor undergoes Patch Merging, which is essentially an interval

sampling technique. The output tensor is divided into four parts, and these parts are concatenated along the channel dimension. Subsequently, a 1×1 convolution is applied to adjust the dimension. This process is analogous to pooling in convolutional neural networks, and the resulting tensor can be used for further computations in the subsequent stages. This procedure is repeated several times in our model to obtain feature maps at four different levels.

B. Tiny Object Detection

We developed our tiny object detector based on Deformable DETR [2]. The model first applies positional encoding to the feature maps at four different scales. Then, the encoder utilizes the mechanism of deformable attention to determine the positions of keys that each query requires. This is achieved by learning the offset Δp_{mqk}. The computation of the deformable attention is:

$$\begin{aligned}&\mathrm{DeformAttn}\big(z_q, p_z, \mathbf{x}\big)\\&\quad=\sum_{m=1}^{M} W_m\left[\sum_{k=1}^{K} A_{mqk}\cdot {W'}_m x(p_q+\Delta p_{mqk})\right](1)\,,\end{aligned}$$

where x represents a feature map, z_q represents a query with index q, p_q represents a two-dimensional reference position, m is the index of attention heads, k is the index of sampling points, and A_{mqk} represents the weight of the self-attention. When applied to multi-scale features, the calculation becomes:

$$\begin{aligned}&\mathrm{MSDeformAttn}(z_q, p_z, \{x^l\}_{l=1}^{L})\\&=\sum_{m=1}^{M} W_m\left[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk}\cdot {W'}_m x^l(\phi_l(\hat{p}_q)+\Delta p_{mlqk})\right](2)\,,\end{aligned}$$

where l represents the index of the multi-scale feature map, $\hat{p}_q$ is normalized 2D coordinates, ϕ_l is used for re-scaling the normalized coordinates, and the remaining symbols are previously described. If the model is enabled with a two-stage approach, an additional encoder-only Deformable DETR is employed to generate region proposals. However, during this process, non-maximum suppression is not required.

Finally, 1,500 learnable object queries—together with the output of the encoder—are fed into the decoder for the standard self-attention computation. The feed-forward network then utilizes the decoder's output to perform predictions for class label and bounding box. To train an effective tiny object detector, we optimized the model by using the focal loss[12], together with the regression loss of the bounding boxes being a combination of the smooth L1 loss and the NWD loss. The NWD loss can address the issue of IoU-based metrics being sensitive to the location deviation of tiny objects.

IV. Experiments

A. Dataset

We evaluated the proposed method on the AI-TOD dataset [3], which contains 28,036 high-quality aerial/satellite images. AI-TOD consists of target objects that do not exceed 64 pixels in size. A single image in this dataset may contain thousands of objects, as shown in Fig. 2, making it more challenging compared to traditional object detection datasets. AI-TOD is divided into two versions—v1 and v2—, in both versions the images are identical, but the annotations in v2 have been improved based on v1[13]. Additionally, objects are categorized based on their sizes in this dataset. Objects ranging from 2 to 8 pixels are referred to as "very tiny" (vt), from 8 to 16 pixels as "tiny" (t), from 16 to 32 pixels as "small" (s), and from 32 to 64 pixels as "medium" (m). Objects are also classified based on their categories, including airplanes (AP), bridges (BR), storage tanks (ST), ships (SH), swimming pools (SP), vehicles (VE), people (PE), and windmill houses (WM).

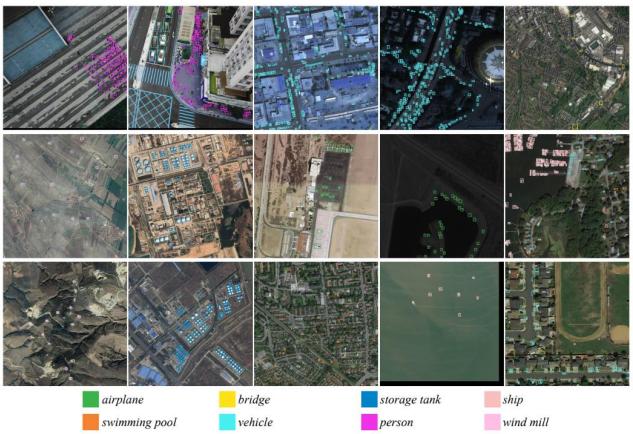


Fig. 2. A few samples in the AI-TOD dataset.

B. Implementation Details

We trained our models on a computer with eight NVIDIA Tesla V100 GPUs. The models were developed based on the MMdetection code library [14].We trained the models for 60 epochs on the AI-TOD-v1 and the AI-TOD-v2 train sets, respectively. In addition, the number of object queries was set to 1,500 to accommodate the dense nature of objects in AI-TOD. The number of feature levels was set to 4. Other hyper-parameter settings and training strategies mainly followed those in Deformable DETR [2].

C. Faster R-CNN and DETR Baselines

We first report the performance of the traditional object detection approach, Faster R-CNN [15], on AI-TOD v1. As shown in Table I, the performance of the Faster R-CNN baseline on AI-TOD is disappointing. While increasing the depth of the backbone network can enhance the overall AP slightly, it does not lead to an improvement in AP for detecting very tiny objects. In fact, the models fails to detect these tiny objects, resulting in an AP of zero. This limitation is a major problem for the Faster R-CNN baseline and other traditional object detection models.

On the other hand, the performance of the original DETR on AI-TOD is even worse. The original DETR does not have

TABLE I
RESULTS OF FASTER R-CNN AND DETR ON AI-TOD V1

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s	AP_m	Epochs
Faster R-CNN	ResNet34	8.9	22.0	5.6	0.0	5.8	18.9	27.8	12
Faster R-CNN	ResNet50	10.1	24.2	6.9	0.0	6.9	21.3	31.1	12
Faster R-CNN	ResNet101	13.1	30.7	9.2	0.0	9.7	24.6	35.5	12
DETR	ResNet50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	150

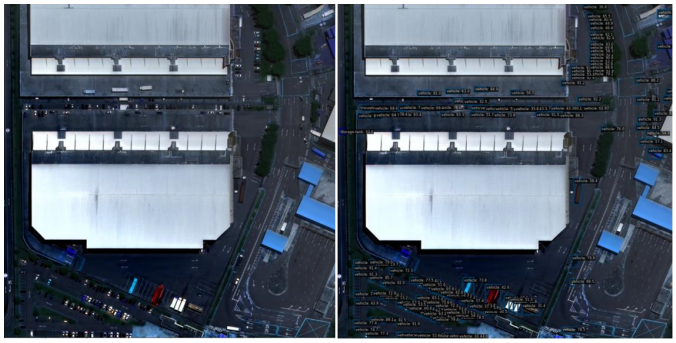


Fig. 3. Visualization of the detection result on AI-TOD v2. Left: the original image, right: the detection result obtained using SD DETR.

a network architecture for handling multi-scale features, and it generates only 100 object queries. Nevertheless, many images in AI-TOD contain a larger number of objects than this limit. It is expected that DETR would perform poorly in detecting tiny objects. However, experimental results show much more extreme outcomes, as the model fails to converge even after training for 150 epochs.

D. Comparison with State-of-the-Art Methods

Tables 2 and 3 compare the performance of our proposed method with state-of-the-art models on AI-TOD v1 and v2, respectively. After applying NWD-RKA [13] or RFLA, many models show significant improvements compared to the baselines. Our proposed models achieve APs that approach the level of state-of-the-art models. It is worth noting that the models outperform the state-of-the-art models on very tiny objects, with an improvement of 1 AP or more. An example of the visualization results of our models on the AI-TOD v2 is shown in Fig. 3.

E. Comparison with Other Designs

Table 4 presents the performance differences brought about by different model designs. Compared to the original deformable DETR, the proposed method performs better on AI-TOD v2 without using a two-stage approach. When incorporating the two-stage approach, the gap in AP widens even further. This improvement is particularly significant for detecting very tiny objects, with an increase of 4.8 AP, which is a notable enhancement.

V. CONCLUSION

We propose SD DETR, a Transformer-based efficient framework for tiny object detection. Addressing the challenges of detecting tiny objects, SD DETR introduces novel ideas and solutions that differ from previous related works. We achieve improved detection performance by directly discarding difficult-to-design modules in tiny object detection and enhancing feature extraction. Experimental results confirm the feasibility and effectiveness of our approach. On AI-TOD, SD DETR achieves comparable performance to state-of-the-art models and even surpasses them in detecting very tiny objects.

VI. ACKNOWLEDGEMENTS

This work was supported by the National Science and Technology Council of Taiwan (111-2221-E-003-016-MY2, 112-2813-C-003-018-E).

REFERENCES

[1] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *International Conference on Computer Vision (ICCV)*, 2021.

[2] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.

[3] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny object detection in aerial images," in *International Conference on Pattern Recognition (ICPR)*, 2020.

[4] J. Wang, C. Xu, W. W. Yang, and L. Yu, "A normalized gaussian wasserstein distance for tiny object detection," *arXiv preprint arXiv:2110.13389*, 2021.

[5] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Rfla: Gaussian receptive field based label assignment for tiny object detection," in *European Conference on Computer Vision (ECCV)*, 2022.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] J. Dai, H. Qi, Y. Xiong, *et al.*, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

TABLE II
COMPARISON OF SD DETR WITH STATE-OF-THE-ART METHODS ON AI-TOD V1. NOTE THAT THE EXPERIMENTAL RESULTS FOR THE RFLA METHODS ARE OBTAINED FROM [5].

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s	AP_m
RetinaNet [12]w/ RFLA	ResNet50	9.1	23.1	5.2	4.1	10.5	10.5	12.3
AutoAssign[16] w/ RFLA	ResNet50	14.2	37.8	6.9	6.4	14.9	17.4	21.8
FCOS[17] w/ RFLA	ResNet50	16.3	39.1	11.3	7.3	18.5	19.8	21.8
Faster R-CNN w/ RFLA	ResNet50	21.1	51.1	13.1	9.5	21.2	26.1	31.5
Cascade R-CNN[18] w/ RFLA	ResNet50	22.1	51.6	15.6	8.2	22.0	27.3	35.2
DetectoRS[19] w/ RFLA	ResNet50	**24.8**	**55.2**	**18.5**	9.3	**24.8**	**30.3**	**38.2**
SD DETR	Swin-t	21.5	53.9	12.2	**10.9**	21.6	27.0	35.1

TABLE III
COMPARISON OF SD DETR WITH STATE-OF-THE-ART METHODS ON AI-TOD V2. NOTE THAT THE EXPERIMENTAL RESULTS FOR THE NWD-RKA METHODS ARE OBTAINED FROM [13].

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s	AP_m
RetinaNet w/ NWD-RKA	ResNet50	10.5	28.5	5.2	3.5	11.0	15.2	24.2
Faster R-CNN w/ NWD-RKA	ResNet50	21.4	53.2	12.5	7.7	20.7	26.8	35.2
Faster R-CNN w/ NWD-RKA	ResNet101	20.8	52.4	12.3	8.5	20.4	24.9	35.1
Faster R-CNN w/ NWD-RKA	HRNet-w32[20]	22.6	55.3	14.2	8.4	21.8	27.6	36.0
Cascade R-CNN w/ NWD-RKA	ResNet50	22.2	52.5	15.1	7.8	21.8	28.0	37.2
DetectoRS w/ NWD-RKA	ResNet50	**24.7**	**57.4**	**17.1**	9.7	**24.2**	**29.8**	**39.3**
SD DETR	Swin-t	21.8	55.7	12.5	**10.9**	21.4	27.6	35.6

TABLE IV
COMPARISON OF OTHER DESIGNS ON AI-TOD V2. + MEANS WITH TWO-STAGE MECHANISM

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s	AP_m
Deformable DETR+	ResNet50	19.2	48.6	11.6	6.1	18.9	24.8	32.7
SD DETR	Swin-t	21.6	53.8	13.4	9.9	21.4	27.9	35.0
SD DETR+	Swin-t	21.8	55.7	12.5	10.9	21.4	27.6	35.6

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[11] K. He, X. Zhang, and S. R. J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark," in *Journal of Photogrammetry and Remote Sensing*, 2022.

[14] K. Chen, J. P. Jiaqi Wang and, Y. Cao, *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," in *CoRR abs/arXiv:1906.07155*, 2019.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.

[16] B. Zhu, J. Wang, Z. Jiang, *et al.*, "Autoassign: Differentiable label assignment for dense object detection," in *arXiv preprint arXiv:2007.03496*, 2020.

[17] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[18] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[20] K. Sun, B. Xiao, J. Dong Liu, and dong Wang, "Deep high-resolution representation learning for human pose estimation," in *In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Lightweight Models Distillation with Learnable Teaching Material: An Application for Smart Table Tennis System

Duan-Yu Chen and Yu-Hsuan Chen
Yuan Ze University, Taiwan
E-mail: dychen@saturn.yzu.edu.tw Tel: +886-34638800 ext. 7120
E-mail: s1094620@mail.yzu.edu.tw Tel: +886-34638800 ext. 636

Abstract — **The rapid advancement of technology has led to the increasing adoption of deep neural networks on edge computing and mobile devices. However, these devices face challenges in effectively handling the extensive parameters and operations associated with neural networks. This study aims to address this challenge by utilizing knowledge distillation methods and integrating them into a table tennis sports analysis system. Specifically, our focus is to enhance the accuracy of a lightweight model designed for edge devices, specifically for ball bouncing point detection in table tennis. Through comprehensive experimentation, we demonstrate the effectiveness of our proposed learnable teaching materials in improving the model's performance. On average, our approach achieves a 1% accuracy improvement compared to the baseline model before knowledge distillation, with the best performance achieving an accuracy of 91.3%.**

I. INTRODUCTION

In recent years, deep learning has witnessed remarkable advancements, leading to the proposal of diverse model architectures with high accuracy. However, these models often rely on computationally expensive parameters, posing challenges when deploying them on low-end hardware devices such as edge computing devices. To address this issue and achieve fast inference times without compromising accuracy, model compression techniques are crucial. Various approaches, including pruning [1-5], model quantization [6-8], and knowledge distillation [8-17], have been proposed for compressing neural networks.

Pruning [1-5] involves removing weights or channels from the model, making it smaller and faster, resulting in reduced storage requirements and accelerated inference. Model quantization [6-8], on the other hand, quantifies the model weights to achieve lightweight models, albeit with a potential reduction in accuracy. Knowledge distillation [8-21] is not a direct model compression method but aims to achieve similar goals and effects. It involves transferring knowledge from a large, accurate teacher model to a smaller, lightweight student model, enabling the latter to attain comparable accuracy levels.

In this work, we focus primarily on knowledge distillation techniques and practices for model compression. Knowledge distillation involves leveraging the knowledge learned by a high-precision teacher model with a large number of parameters and transferring this knowledge to a smaller, lightweight student model. Numerous knowledge distillation methods have been proposed, with many focusing on extracting appropriate lightweight student networks from complex teacher networks [8-17]. However, these methods often require substantial computational resources and entail prolonged training times.

Additionally, some methods explore loss functions that effectively utilize the transferred knowledge [18-19]. While these methods have lower training costs and time requirements, they are limited to specific tasks. For instance, loss functions designed for classification tasks may not be applicable to object detection tasks.

Through long-term knowledge distillation experiments, we have observed that augmenting the simple knowledge distillation approach with supplementary teaching materials can further enhance teaching quality, accuracy, and fine-grained details of the student model. In Section 3, we present the proposed methods, while Section 4 delves into the experimental results, providing a comprehensive explanation of these advancements.

II. RELATED WORK

Machine vision and deep learning have become popular areas of research in ball sports events, as the speed of the ball

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

makes it challenging for human vision to keep up. Intelligent systems that can gather more information to assist in judgments have been developed. One such system is TTNet [20], which focuses on table tennis and includes scene analysis, ball track detection, and ball bounce event detection.

Given our specific research conditions, which involve strict restrictions and the goal of real-time detection of the table tennis ball's drop point on an edge computing device, we employ lower resolution inputs and a lightweight model to tackle this challenge. The scene recognition in our study is subject to numerous limitations, and we are unable to obtain training data from any publicly available datasets. Therefore, we curated a private dataset to train our model.

Traditional object detection methods typically involve inputting an RGB image and using a network to determine the object's position and type within the image. However, detecting the precise moment when a ball bounces off a table presents a significant challenge for these traditional methods. To overcome this challenge, we curated a private dataset that collects trajectory maps generated by the ball's movement over short intervals. By analyzing these trajectory maps, our goal is to detect the presence of a bounce point. Among the various one-stage detectors available, anchor-free methods [21-30] offer the advantage of lower computational burden and faster inference. One branch of anchor-free detection methods is the keypoint-based approach [25, 27, 28], which involves detecting objects by identifying key points. The Centernet method [27] aligns closely with our objective (refer to Figure 3). Its process is straightforward, directly predicting the coordinates of the center point and subsequently estimating the length and width of the bounding box based on the center point's features. This notion of predicting the key center point of an object aligns with our task of detecting the ball's drop point. Consequently, we design our detection head to predict the center keypoint as part of our approach.

In the realm of knowledge transfer between different networks, Hinton et al. [31] introduced a novel concept of utilizing soft labels in conjunction with hard labels for simultaneous training. Soft labels represent the predicted output probability values from a larger model, while hard labels correspond to the original ground truth labels. This probability distribution provides the neural network with implicit information about the similarity between different categories, moving beyond a binary right-or-wrong approach. By incorporating soft labels, the smaller model can learn more effectively from the teacher model and exhibit improved generalization during training.

Since then, numerous studies have explored the idea of distillation from intermediate layers. Knowledge extraction is not limited to the output of the last layer in the teacher network; instead, it can be derived from multiple layers throughout the network. This enables the teacher to transfer knowledge to the student through feature maps from intermediate layers, facilitating faster and better initialization of the student network. Consequently, extending distillation from logit to intermediate layers has become a common approach in knowledge distillation. Examples include Fitnets proposed by Romero et al. [32] and AT (Attention Transfer) proposed by Zagoruyko et al. [33].

Fitnets suggest that directly fitting the output of a large model to a small model is a challenging task. Therefore, it is recommended to enhance the supervision signal in the intermediate layers, aligning the intermediate layers of the small model with those of the large model. This approach involves the small model emulating an intermediate learning step of the large model. On the other hand, the method proposed by Zagoruyko et al. [33] focuses on distilling knowledge about feature relationships between layers rather than solely focusing on the output. The student model is guided by extracting the attention map generated by the teacher model, ensuring that the student model's attention map closely resembles that of the teacher model. This method enables the student model to not only learn feature information but also learn how to refine feature information. Notably, the attention map differs from the feature map, but the number of channels remains consistent between the teacher and student models, facilitating a more intuitive transfer of knowledge.

III. Proposed Method

In this study, we propose a method based on knowledge distillation. The concept of knowledge distillation involves transferring the final predicted logit of the teacher model to the student model for the classification task, as proposed by Hinton et al. [38]. Our approach follows a similar principle, where the final output of our model takes the form of a heat map that predicts the central point of the ball trajectory upon bouncing. Figure 1 provides an illustration of our training process, which includes computing the loss function using the output heatmaps from both the teacher and student models, in which FCOS[26] architecture is employed and its lightweight model is constructed as shown in Fig. 2.

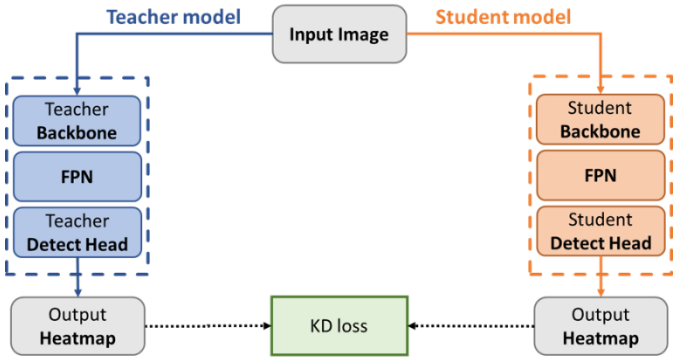


Fig. 1. Training structure of distillation from output layer.

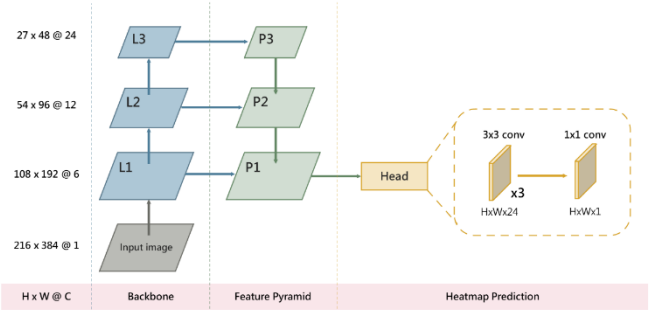


Fig.2. Our lightweight network architecture based on FCOS[26].

The loss functions for our proposed knowledge distillation mechanism are defined as follows:

$$\mathcal{L}_{out_{teach}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_t - \hat{y}_s)^2, \quad (1)$$

$$\mathcal{L}_{output\ layer-KD} = \mathcal{L}_{out_{teach}} + \mathcal{L}_{ori}, \quad (2)$$

where $\mathcal{L}_{output\ layer-KD}$ is the total loss of this part, where the distillation loss $\mathcal{L}_{out_{teach}}$ is the L2 norm, $\hat{y}_t$ and $\hat{y}_s$ are the prediction from teacher and student models, respectively. $\mathcal{L}_{ori}$ is the supervised loss calculated from ground truth. In addition, a loss function for the center response point is added, which is defined as

$$\mathcal{L}_{heatmap} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (3)$$

$$\mathcal{L}_{coordinate} = \frac{\left(C_x - \hat{C}_x\right)^2 + \left(C_y - \hat{C}_y\right)^2}{2}, \quad (4)$$

$$\mathcal{L}_{ori} = \mathcal{L}_{heatmap} + \mathcal{L}_{coordinate}, \quad (5)$$

where $\mathcal{L}_{heatmap}$ is the loss of heatmap, $\mathcal{L}_{coordinate}$ is the MSE loss of predicted coordinate of central response point, C_x, $\hat{C}_x$, C_y, $\hat{C}_y$ is the predicted x-axis coordinate, the ground truth x-axis coordinate, the predicted y-axis coordinate, and the ground truth y-axis coordinate, respectively.

To optimize the learning process of Knowledge Distillation (KD), we introduce the concept of Learnable Teaching Material (LTM) as illustrated in Fig. 3. The key characteristic of LTM lies in its capacity to accommodate an unlimited number of parameters for the learnable materials, depending on the layer of interaction with the teacher model. In Fig. 3, the pre-trained weights of the teacher model remain fixed, while the teaching materials are updated alongside the student model. The initial values of the materials correspond to the initialized weights, which vary according to different interaction patterns within the framework.

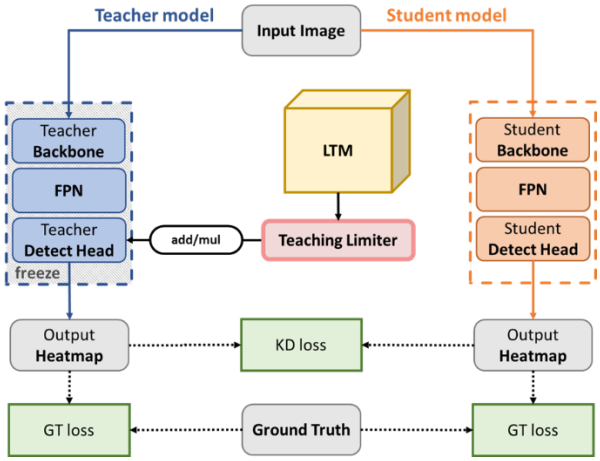


Fig. 3. Training structure of distillation with learnable teaching material.

The rationale behind freezing the pre-trained weights of the teacher model and solely updating the teaching material is to refine the teaching knowledge based on the students' learning progress, while preserving the foundational knowledge already acquired by the model. This approach allows for fine-tuning without directly modifying the existing knowledge. However, it is important to note that certain limitations are imposed on the teaching material as well by

$$\mathcal{T} = \mathcal{M}(LTM) = \gamma \cdot tan\,h(LTM)\,. \quad (6)$$

$\mathcal{M}(\cdot)$ is the teaching limiter, $\mathcal{T} \in [-\gamma \sim \gamma]$ is the output of the limiter, which can effectively fine-tune the underlying knowledge of the teacher model, and under the action of the limiter, LTM will not completely change the full structure of the knowledge. The related loss function is defined as

$$\mathcal{L}_{LTM-KD} = \mathcal{L}_{teacher_ori} + \mathcal{L}_{output\ layer-KD}\,. \quad (7)$$

In the whole training process, LTM will be trained and updated together. Therefore, the total loss $\mathcal{L}_{LTM-KD}$ also contains the loss $\mathcal{L}_{teacher_ori}$ from the output of the teacher model.

IV. Experiment Results

Our dataset is collected using the "Cumulative Frame Difference" (CFD) [34] method, which can detect bounce points by a single cumulative map. Examples of the original CFD and 8-bit CFD are demonstrated in Fig. 4 and Fig. 5, respectively.

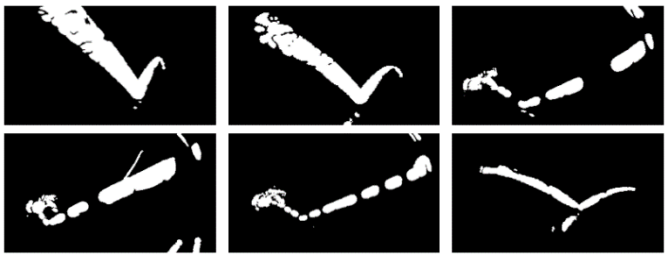
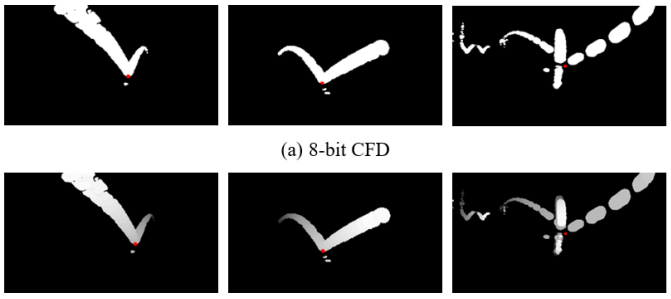


Fig. 4. Demonstration of the original CFD



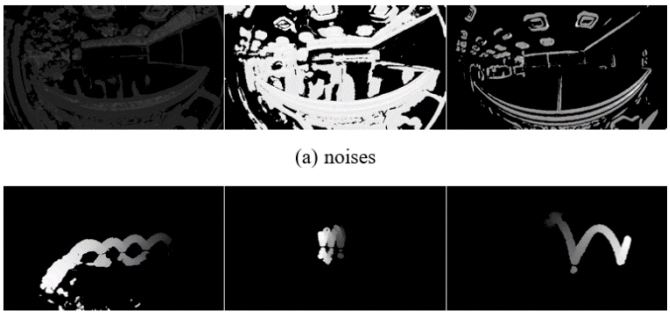
(a) 8-bit CFD

(b) 8-bit CFD w/ temporal

Fig. 5. Demonstration of 8-CFD

In our study, we obtained a comprehensive dataset from 23 private videos, which yielded a total of 7,467 original training images encompassing all employed methods. To enhance the quality of our results, we implemented diverse data augmentation techniques during the training process. These techniques involved common methods like "rotation" and "horizontal flip." Furthermore, we proactively considered potential challenges that may arise in real-time inference, such as camera shake and unexpected re-bouncing of the ball. To tackle these challenges, we employed the "copy-paste" method to create more intricate data augmentations, aiming to bolster the resilience of our training results. Remarkably, previous research [35] has demonstrated the significant performance improvement achievable through the simple act of copying data. Inspired by this insight, we developed a similar process to generate our augmented images.

To organize our training dataset, we divided it into two distinct categories: positive samples and negative samples. Positive samples consisted of images capturing bouncing points, whereas negative samples solely contained background and noise. We further categorized the negative samples into "noises" and "re-bounced ball" subcategories. Fig. 6 provides an example showcasing a negative sample.



(a) noises

(b) re-bounced balls

Fig. 6. Demonstration of negative samples.

We proceeded to develop three merging methods to combine positive and negative samples, resulting in new training data. The merging methods were as follows:

(a) Randomly selecting a "noises" negative sample and a chosen positive sample, and then summing the two samples. (b) Randomly selecting a "re-bounced ball" negative sample and a chosen positive sample, and then summing the two samples. (c) Randomly selecting a "noises" negative sample, a "re-bounced ball" negative sample, and a chosen positive sample, and then summing the three samples. To maintain diversity within the negative samples, we performed various transformations on them before combining the images. These transformations included random rotation, horizontal flip, random zooming, and random panning. Following the data augmentation process, the training images in the dataset expanded to four times the original size, resulting in a total of 29,868 images. This increase is due to each positive sample undergoing the aforementioned data augmentation techniques.

In our experiments, the system we use for training models is AMD Ryzen™ 9 3900X CPU @ 3.8 GHz, NVIDIA® TITAN RTX™ GPU with 24GB VRAM and 4608-core CUDA, and 128GB DDR4 system memory. In training the baseline of teacher (FCOS ResNet-19 lite) and student (FCOS ResNet-13 lite) models, the batch size is set to 64 and 128 respectively, and in knowledge distillation training, the batch size is set to 32. all the above cases trained 300 epochs. For LTM-KD method, we set the hyperparameter γ of the teaching limiter to 0.5, and use the "multiplying" method to let the teaching material assist the teaching.

In Table 1, the addition of Implicit Knowledge to the teacher model leads to noticeable improvements in the "16-bit" datasets. Specifically, compared to the Baseline, there is an increase of 0.9% and 1.3% in performance, respectively.

However, when it comes to the distilled student model in the "16-bit CFD w/temporal" dataset, the improvement is only 0.5%. On the other hand, LTM-KD demonstrates a 0.9% enhancement in the "16-bit CFD" dataset, achieving the highest accuracy of 91.3%. Despite the fact that the teacher model in LTM-KD has not undergone retraining, it is evident that it has been influenced by the LTM. This influence is demonstrated by the observed results. Interestingly, while the accuracy of the student model improves, the accuracy of the teacher model decreases under the influence of LTM. This observation indicates that the original knowledge of the teacher model is too complex to be directly extracted for effective teaching. However, with the assistance of LTM, this challenging knowledge can be transformed into more suitable and digestible forms for students.

Table 1. Performance comparison between KD using implicit knowledge(IK) and learnable teaching material(LTM)

Distillation Method	Teacher FCOS ResNet-19 lite				Student FCOS ResNet-13 lite			
	8-bit		16-bit		8-bit		16-bit	
	CFD	CFD w/ temporal	CFD	CFD w/ temporal	CFD	CFD w/ temporal	CFD	CFD w/ temporal
Baseline	0.896	0.913	0.917	0.906	0.879	0.862	0.894	0.893
IK-KD	0.890	0.909	0.926	0.919	0.878	0.879	0.892	**0.901**
LTM-KD[1]	0.889	0.897	0.905	0.903	0.880	0.881	**0.913**	0.894

[1] : Unlike other methods, LTM-KD does not retrain a pre-trained teacher model specifically for the corresponding method. Instead, it utilizes a baseline model that has been pre-trained with LTM. In the LTM-KD approach, only the parameters of the LTM are updated during the knowledge distillation process.

V. CONCLUSIONS

In this study, we have successfully tackled the challenge by incorporating knowledge distillation techniques and integrating an ultra-lightweight model into a smart table tennis sports system. Our focus was on enhancing the accuracy of a lightweight model designed for edge devices, specifically for ball bouncing point detection in table tennis. By employing LTM-KD, we have effectively improved the performance of the model. Through extensive experimentation, we have demonstrated the efficacy of our proposed knowledge distillation approach utilizing learnable teaching material (LTM). On average, our method achieves a significant 1% accuracy improvement compared to the baseline model prior to knowledge distillation. The best-performing model achieved an impressive accuracy of 91.3%, highlighting the effectiveness of LTM-KD in real-world environments.

REFERENCES

[1] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning Filters for Efficient ConvNets," *ArXiv,* vol. abs/1608.08710, 2017.

[2] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the Value of Network Pruning," p. arXiv:1810.05270.

[3] J.-H. Luo, J. Wu, and W. Lin, "ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression," *2017 IEEE International Conference on Computer Vision (ICCV),* pp. 5068-5076, 2017.

[4] S. Sen, N. Moha, B. Baudry, and J.-M. Jézéquel, "Meta-model Pruning," in *Model Driven Engineering Languages and Systems*, Berlin, Heidelberg, A. Schürr and B. Selic, Eds., 2009// 2009: Springer Berlin Heidelberg, pp. 32-46.

[5] M. Zhu and S. Gupta, "To prune, or not to prune: exploring the efficacy of pruning for model compression," *ArXiv,* vol. abs/1710.01878, 2018.

[6] A. J. Ahumada Jr and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," in *Human vision, visual processing, and digital display III*, 1992, vol. 1666: SPIE, pp. 365-374.

[7] H. Peterson, A. Ahumada, and A. Watson, "An Improved Detection Model for DCT Coefficient Quantization," *AIAA Computing in Aerospace 9: A Collection of Technical Papers,* vol. 1913, 09/08 1993, doi: 10.1117/12.152693.

[8] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *ArXiv,* vol. abs/1802.05668, 2018.

[9] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.

[10] Y. Kim and A. M. Rush, "Sequence-Level Knowledge Distillation," in *EMNLP*, 2016.

[11] S. Lee and B. C. Song, "Graph-based Knowledge Distillation by Multi-head Attention Network," p. arXiv:1907.02226.

[12] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured Knowledge Distillation for Semantic Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15-20 June 2019 2019, pp. 2599-2608, doi: 10.1109/CVPR.2019.00271.

[13] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved Knowledge Distillation via Teacher Assistant," in *AAAI*, 2020.

[14] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational Knowledge Distillation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3962-3971, 2019.

[15] H. Wang, S. Lohit, M. Jones, and Y. R. Fu, "Multi-head

Knowledge Distillation for Model Compression," *ArXiv,* vol. abs/2012.02911, 2020.

[16] J. Yim, D. Joo, J. Bae, and J. Kim, "A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 7130-7138, doi: 10.1109/CVPR.2017.754.

[17] L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng, "Revisit Knowledge Distillation: a Teacher-free Framework," *ArXiv,* vol. abs/1909.11723, 2019.

[18] X. Li, W. Wang, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized Focal Loss V2: Learning Reliable Localization Quality Estimation for Dense Object Detection," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 11627-11636, 2021.

[19] X. Li *et al.*, "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," *ArXiv,* vol. abs/2006.04388, 2020.

[20] R. Voeikov, N. Falaleev, and R. Baikulov, "TTNet: Real-time temporal and spatial video analysis of table tennis," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 3866-3874, 2020.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *ArXiv,* vol. abs/2004.10934, 2020.

[22] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint Triplets for Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 6568-6577, 2019.

[23] H. Law and J. Deng, "CornerNet: Detecting Objects as Paired Keypoints," *International Journal of Computer Vision,* vol. 128, pp. 642-656, 2019.

[24] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016.

[25] W. Liu, I. Hasan, and S. Liao, "Center and Scale Prediction: A Box-free Approach for Pedestrian and Face Detection," *arXiv: Computer Vision and Pattern Recognition,* 2019.

[26] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 9626-9635, 2019.

[27] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as Points," *ArXiv,* vol. abs/1904.07850, 2019.

[28] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-Up Object Detection by Grouping Extreme and Center Points," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 850-859, 2019.

[29] C. Zhu, Y. He, and M. Savvides, "Feature Selective Anchor-Free Module for Single-Shot Object Detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 840-849, 2019.

[30] C.-Y. Wang, I.-H. Yeh, and H. Liao, "You Only Learn One Representation: Unified Network for Multiple Tasks," *ArXiv,* vol. abs/2105.04206, 2021.

[31] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *ArXiv,* vol. abs/1503.02531, 2015.

[32] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for Thin Deep Nets," *CoRR,* vol. abs/1412.6550, 2015.

[33] S. Zagoruyko and N. Komodakis, "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer," *ArXiv,* vol. abs/1612.03928, 2017.

[34] Ching-Chung Kuan, “Ultra-Fast Deep Learning Based Bounce Point Detection for Smart Table Tennis System on Edge Computing Devices,” Master Thesis, July 2021. (https://hdl.handle.net/11296/3q9en7)

[35] G. Ghiasi *et al.*, "Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2917-2927, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Selecting Suitable Data Input for Deep-Learning Sign-Language Recognition with a Small Dataset

Yu-Jen Chen and Po-Chyi Su
Dept. of Computer Science and Information Engineering
National Central University, Taoyuan, Taiwan
E-mail: pochyisu@csie.ncu.edu.tw Tel/Fax: +886-4227151/+886-4222681

***Abstract*—Deep learning-based sign-language recognition usually requires abundant training videos. This research considers generating valid sign-language data for training the recognition models. MediaPipe is used to acquire the hand skeleton from the sign-language video. Then we analyze several hand skeleton adjustment policies with color-weighting strategies and generate hand masks to simulate different hands. Since miss detections of hands may happen due to motion blurring caused by rapid hand movements, we incorporate optical flows to ensure that the hand movement information is retained in each frame. We employ different spatial and temporal data augmentation strategies to simulate varying hand sizes, filming angles, and hand speeds. The experimental results show that the proposed method improves the accuracy of sign-language recognition in the American Sign Language dataset.**

I. Introduction

Sign language is an essential communication tool for people with hearing loss. Such visual languages are not universally understood as they vary depending on the customs of different regions, even in different cities of the same country. Sign language conveys messages through hand shapes, movements, and even facial expressions. Besides, sign languages usually express the meaning of words, instead of separated alphabets, making the recognition of sign languages require particular training. According to a study by the World Health Organization (WHO), more than 5% of the world's population, or 430 million, have a disability-related hearing loss. According to the Ministry of Health and Welfare of the Executive Yuan, $\%5$ of Taiwan's population, or $124,000$, have hearing impairment. Sign language serves as a bridge for people with hearing loss to receive information and communicate with other people. On the other hand, people who have not learned sign language cannot fully understand what the hearing-impaired person would like to express. Using video comprehension methods on terminal devices (e.g., smartphones, tablets, etc.) to help us analyze the gestures made by hearing-impaired persons will significantly reduce communication difficulty.

Traditional sign-language recognition research often requires a wearable device or a depth camera to help extract hand features, thus requiring considerable hardware costs. It also demands a series of image processing procedures to extract hand features, which is time-consuming and labor-intensive. In this research, we adopt deep-learning models to replace the traditional ways of using assistive devices and to effectively extract hand information like bounding boxes of hands or skeletons of hands. We employ hand masks, hand skeletons, and optical flows of sign language videos as the input to a deep learning network. By testing with different experiments, we analyze several hand skeleton adjustments and color weighting strategies and generate hand masks from the skeletons to simulate different hands. The experimental results show that the proposed methods improve recognition accuracy when testing on the American sign language dataset.

II. Related Work

For the related work of sign-language recognition, we first describe the available datasets and the potential solutions.

A. Datasets

There is a wide variety of sign languages; collecting data to form a feasible sign language dataset is not trivial. The publicly available sign-language datasets can be classified into isolated sign-language (ISL) datasets, as shown in Table I, and continuous sign–language (CSL) datasets, as shown in Table II. ISL datasets contain videos demonstrating the signing expression of single words, while CSL datasets consist of multiple words in each video. Each dataset may provide different data and labeling methods. The common data types are raw videos with RGB frames, with probably additional depth information. Such datasets as How2Sign [1] even contain sign-language videos of the same scene captured from different filming angles. Among the datasets, WLASL (Word-Level American Sign Language) [2] has quite complete data on American sign language. Some examples are shown in Fig. 1. There are four different subsets, WLASL100, WLASL300, WLASL1000, and WLASL2000, containing 100, 300, 1000, and 2000 sign language words respectively. We will use WLASL100 in this study.



Fig. 1. WLASL dataset [2]

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I
ISOLATED SIGN LANGUAGE DATASETS

Name	Language	#Gloss	#Video per gloss	Data type
DEVISIGN-G	Chinese	36	3	Video(RGB), Skeleton keypoints, Depth
DEVISIGN-D	Chinese	500	3	Videos(RGB), Skeleton keypoints, Depth
DEVISIGN-L	Chinese	2000	3	Videos(RGB), Skeleton keypoints, Depth
WLASL	English	2000	10.5	Videos(RGB), Skeleton keypoints, Frame startend
MS-ASL	English	1000	25.5	Videos(RGB)
RWTHBOSTON50	English	50	9.7	Videos(RGB)
Boston ASLLVD	English	2742	3.6	Videos(RGB), Multiple angles, Endstart frame
DGS Kinect-40	German	40	5	Videos(RGB), Depth
GSL	Greek	310	131.5	Videos(RGB), Depth
LSE-sign	Spanish	2400	2	Videos(RGB)
LSA64	Argentinian	64	50	Videos(RGB)
PSL Kinect 30	Polish	30	10	Videos(RGB), Depth
PSL ToF	Polish	84	20	Videos(RGB), Depth
AUTSL	Turkish	226	169	Videos(RGB), Depth, Skeleton keypoints

TABLE II
CONTINUOUS SIGN LANGUAGE DATASETS

Name	Language	#Gloss	Data type
Purdue ASL	English	14	Videos(RGB)
How2Sign	English	16000	Videos(RGB)
CUNY ASL	English	33000	Videos(RGB)
RWTH-PHEONIX-WEATHER 2014	German	1295	Videos(RGB), Start/End
SIGNUM	German	450	Videos(RGB)
GSL	Greece	331	Videos(RGB), Depth

B. Classifications of Sign Language Recognition

Sign language recognition is also divided into ISL recognition and CSL recognition. The methodologies are classified into traditional signal processing and deep learning methods. Early vision-based methods for sign-language recognition used manually crafted hand features [3] to represent gestures. For example, SIFT-based features [4], HOG-based features [5], and other features in the frequency domain may be employed. Auxiliary hardware such as Kinect was often installed to help capture hand gestures. The temporal relationships in the video sequences are then examined by the Hidden Markov Model (HMM) [6]. Finally, machine learning techniques such as SVM, template matching [7], instance-based learning, and decision tree [8] are used to classify the sign language videos. The effectiveness of traditional methods for sign language recognition is usually doubted due to high computational complexity, unpopularity of using additional hardware, and lack of generality of using a single setting for varying videos.

Deep learning methods for sign-language recognition attract much attention as more data are available, and the features can be extracted via training. Karen Simonyan proposed dual-stream convolutional neural networks [9], which could be the first scheme to leverage the feature extraction capability of deep learning techniques. The dual-stream convolutional neural network architecture includes both spatial and temporal networks to deal with the pixel-domain and temporal-domain data. The optical flows are extracted from the videos as the network input. The SVM is used as the classifier. No manual labeling is required. Reasonably good results are achieved on the task of UCF-101 [10] and HMDB-51 [11] datasets. To better establish the temporal relationship between visual features, Jeff Dona proposed LRCN [12], which extracts features from each video frame and models the hidden states by Conditional Random Field (CRF). LSTM is employed to learn the sequential dynamics. Another CNN+LSTM architecture was proposed by Ng et al. [13]. To explore different convolutional feature pooling architectures, feature pooling is used to maximize the collection of local messages. Feature aggregation is applied from frame to frame. A 2D CNN+LSTM-based sign-language video recognition was proposed by Li et al. [2]. The spatial features are extracted using VGG16 trained by ImageNet and then serve as the input to a stack of RNN-routed GRUs to process the temporal domain dependencies for recognition. Hara made use of Kinetics [14] to train a 3D ResNet network [15], replicating the success of 2D CNN on ImageNet. The model I3D [16] was also trained on ImageNet first and then fine-tuned on Kinetics-400 [14]. The results demonstrated that the fine-tuned I3D can better extract the features of sign-language videos. In addition to using RGB images or frames as the input, extracting skeletons is a common approach since many human behavior recognition applications use localized key points for pose analysis. The pose keypoint-based sign language recognition is often adopted in transformer-based single-word sign language recognition systems [17]. They

first estimated human pose key points and introduced a so-called pose normalization approach to process the skeletons. Several joint rotation strategies were proposed to apply the data augmentation on the hand skeletons to improve the model accuracy further.

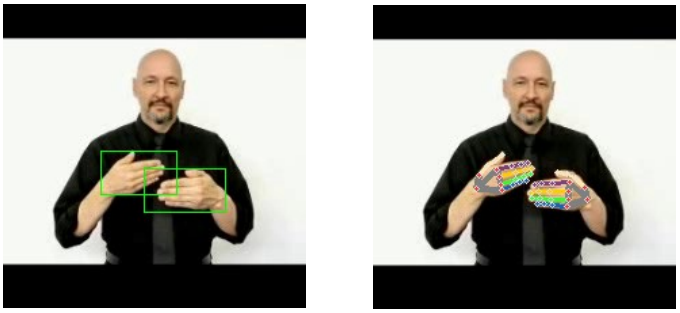
C. Feature Extraction

Traditional image processing methods to extract hands usually rely on a series of complex procedures. Edge detection can be applied first, followed by binarization and adjusting the thresholds to remove noises and further sharpen edge lines. The contours of the binarized images are then plotted, combined with the convex polygon detection and background removal. In addition to image processing methods, adopting special hardware assistance is also a feasible approach. RGB-D camera, Kinect, sensor gloves, etc. can be employed to obtain more information beyond color pixels. RGB-D cameras can use hand colors and depths to eliminate the background noises to better capture hands. Kinect is equipped with an RGB camera, an infrared emitter, and an infrared CMOS composed of a 3D structured light depth sensor. Kinect can scan the scene with an IR sensor emitting infrared. Kinect can also convert the detected depth images to the skeleton tracking system to facilitate the extraction of hand skeletons. Many open-source platforms integrate mainstream deep learning algorithms into a single library, allowing developers to call different modules using APIs, without markup or with only a small amount of markup, and using pre-trained models, which can significantly reduce development time. Detectron2 [18] is an object detection platform based on the Pytorch framework, developed by AI research in Facebook. As a combination of predecessor Detectron and MaskRCNN-benchmark [19], Detectron2 provides state-of-the-art object detection and object segmentation algorithms, such as Fast RCNN [20], Faster RCNN [21], Mask RCNN [22], which can be used to localize the hand positions. Mask RCNN can also be used to segment the hand and generate hand masks. Mediapipe [23] is a collection of frameworks that can be built on different platforms. Mediapipe supports a variety of different model inferences. It also provides solutions for vision tasks such as object detection, multiple hand tracking, face detection, and hair segmentation, etc. We use the hand detection/tracking module and a skeleton detection module to acquire the information of hands.

III. Proposed Method

It is well known that the regions of interest in frames of sign-language videos should be hands. We utilize Mediapipe to help capture all the hands in videos. Both the hand and skeleton detection modules in Mediapipe are considered to locate the bounding boxes of the hands and the joints of the skeleton of the hand, as shown in Fig. 2.

With the hand skeleton key points, we apply the dilation to expand the joints, draw contours surrounding all the hand key points, and then fill the contours to acquire the hand mask, as shown in Fig. 3. The masks are used to preserve or show the



(a) Bounding Box (b) Multi-Color Skeleton

Fig. 2. The examples of hand extraction by Mediapipe.

pixels of hands only and remove all the inferences from other objects, including other parts of a person or the background. In our opinion, hands look much similar for different persons, compared with other portions of bodies. When the dataset is not large, we have to keep hands only and then use data augmentation to increase the types of content.

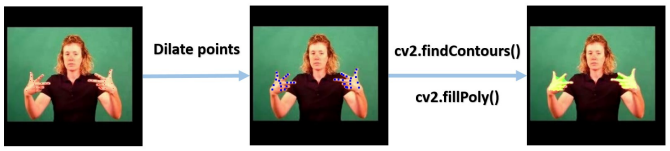


Fig. 3. The procedure of generating hand masks.

It should be noted that Mediapipe may fail to detect the hands when they move too fast or are too close to the face, which has similar skin colors. It is possible that both hands disappear and the whole frame becomes empty. We thus choose to calculate the optical flows to ensure that hand movement information is preserved in every frame, as shown in Fig. 4.

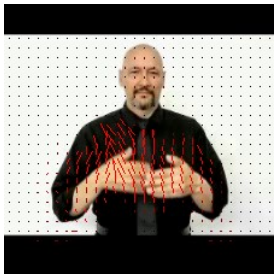


Fig. 4. Optical flows on one frame.

With a small sign-language training set, using a single feature extracted from the sign-language video may result in poor performance due to the limited data that the model can learn from. We examine different hand features in the three input channels of the model. The schematic diagram of the synthesized data is shown in Fig. 5. We have tried several combinations of possible input data, including different color information of hands combined with the optical flows, as

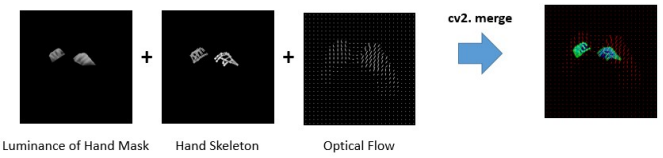
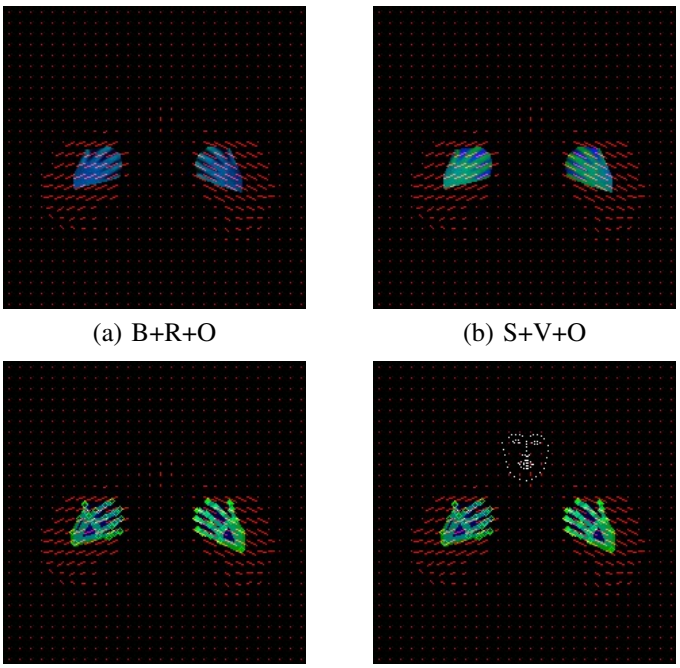


Fig. 5. The procedure of synthesizing the training data.

demonstrated in Fig. 6. The first attempt was to select two channels in the original RGB color space of videos and then integrate the optical flows as the input data. We also convert the videos from the RGB color space to HSV and replace the Hue channel with the optical flows as the input data. Then we train the hand skeletons and hand masks together, combined with the optical flows. We found that this combination yields better performance. Finally, we add the face landmarks into the input by extracting 68 features of the face and merging these features into the three channels, i.e., the hand mask, the hand skeleton, and the optical flows. However, the accuracy is not improved by adding the face landmarks.



(a) B+R+O (b) S+V+O

(c) L+Skeleton+O (d) L+Skeleton+O+F

Fig. 6. Examples of different combinations of input. **O** denotes Optical flow between two frames. **B** and **R** denote the Blue channel and the Red channel in RGB color space, respectively. **S** and **V** denote the Saturation channel and the Value channel in HSV color space respectively. **L** and **F** denotes Luminance of mask and Face Landmark respectively.

In order to simulate the situation where the hand is partially blurred due to fast movement and to let the model learn the different shapes of the hand, our idea is to reveal different pixels of the original hand. We implement this idea by changing the shape of the hand mask. We add noises to the hand skeleton's key points to change the shapes of hand masks. The steps are as follows:

1. Select a hand skeleton joint and choose a random value x in the range [0, X].
2. Randomly choose a direction y in the set {up, down, left, right, top-left, bottom-left, top-right, bottom-right}.
3. Move the joint to a point x unit away from the original joint in the direction y.

Through the steps mentioned above, different hand skeletons can be produced, and then hand masks are generated accordingly. We examine the following four hand skeleton adjustment strategies, as shown in Figure 7.

- **Random Noises**: The most intuitive step is to randomly adjust the data. We randomly pick 0 to 21 points from the 21 joints of the palm; each key point is randomly assigned with a distance value x, direction y, and then move the joint x unit away from the original joint in direction y.
- **Fingertip Noises**: Besides random adjustments, we start with applying minor adjustments on skeletons. Adjusting the fingertips should be the slightest change, compared with moving the whole fingers or retracting the entire palm. The strategy is randomly selecting 0 to 5 fingertips from the five fingers. The selected fingertip key points are randomly assigned with a value of distance x and a direction y. The key points are moved the joint x unit away from the original joint in direction y.
- **Move Fingers**: We randomly select 0 to 5 fingers, and the selected fingers are randomly assigned with a value of distance x and direction y. The key points of the same finger will move in the same direction.
- **Shrink Palm**: We fix our fingers and fingertips, but zoom in or zoom-out on the key points of the palm, and randomly select the distance value x and direction y.

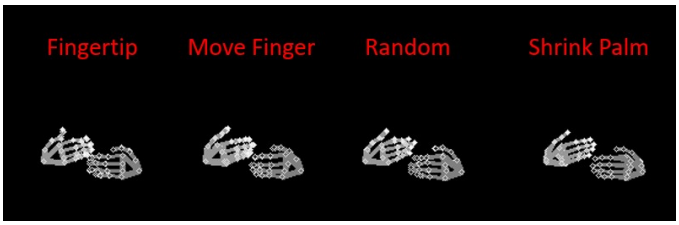


Fig. 7. Applying different skeleton adjustments.

ResNet is a well-known framework and has shown decent performance in image classification. Increasing the time dimension of all the convolutional kernels of ResNet into 3 dimensions, i.e., 3D ResNet, is considered in this research. Such 3D CNN can better combine the spatial and temporal domain features than 2D CNNs with dual-stream networks. In this study, we use a very basic 3D ResNet framework [15] to train a sign-language classification or recognition model. By doing so, the effectiveness of selecting suitable data inputs can be more obvious.

The data augmentation in the spatial domain is to simulate different positions of the hands, the brightness and darkness

of the scene, the size of the hands, and different filming angles, etc. We have applied the following spatial-domain data augmentation by introducing the brightness changes, displacement, scaling, slight rotation, contrast changes, and saturation, as shown in Fig. 8.

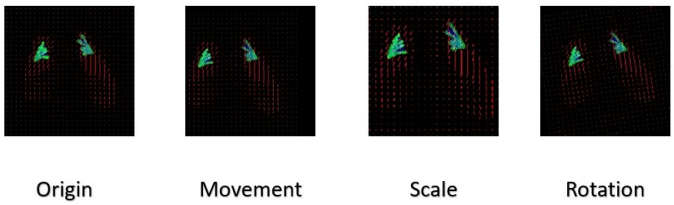


Fig. 8. Spatial augmentations.

The time-domain data augmentation simulates the varying speed of sign-language users. Selecting a few frames for processing reduces the memory burden of training the full video. We first calculated the average frame rate of the entire American Sign Language dataset, which is around 30 frames. We randomly selected 25 frames in American Sign Language videos and restricted the distance between the first and last frame with no more than the min_size set $\{15, 20, 25\}$ frames. If we select more than 25 frames, we randomly drop some frames until 25 frames are left; if the total number of selected frames is less than 25 frames, we randomly repeat some frames until 25 frames are reached. We use the following four different time domain sampling strategies:

1. Average Crop: Randomly select the start frame in the sign-language video interval $[1, min_size/2]$, then randomly select the step x from $\{2, 3\}$, and start sampling with step x from the start frame until 25 frames are selected.
2. Center Crop: First, take the middle frame of the sign-language video, then choose (index of the middle frame - min_size)/2 as the starting frame, and the last frame of the video as the ending frame. Randomly select 25 frames from this interval.
3. Random Crop: From the first to the last frames of the sign language video, we randomly pick 25 frames.
4. Begin Crop: In the sign language video interval $[1, min_size/2]$, pick a random number as the starting frame and sample 25 frames backward.

IV. Experimental Results

A. Implementation Details

This study was developed using Python version 3.6. The deep learning framework uses PyTorch 1.8.0 and Cuda 10.1, CuDNN 7. The operating system is Ubuntu 18.04 LTS. The system is equipped with Intel CoreTM i9-9900K CPU@3.6GHz, 64GB RAM, with Nvidia RTX Titan 24GB GPU.

B. Baseline

We use Mediapipe as the hand extractor to obtain the bounding boxes and skeletons of hands. Then, we use the hand skeleton to generate a hand mask. We train the hand bounding box, hand skeleton, and hand mask as benchmarks for improving sign language recognition accuracy.

According to the experimental results in Table III, the skeleton has the highest accuracy in the Top-1 ranking. Skeleton and Mask have the same accuracy in Top-3, and Mask has the highest accuracy in the Top-5 and Top-10 rankings. It can be found that Skeleton and Mask have their own advantages, while the accuracy of Bounding Box is relatively low. Therefore, we will use Skeleton and Mask for further improvement.

TABLE III
Single Feature Tests

Data type	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy	Top-10 Accuracy
Bbox	62.40%	79.07%	83.72%	87.21%
Mask	70.54%	**84.88%**	**89.92%**	**93.41%**
Skeleton	**72.87%**	**84.88%**	89.15%	91.86%

C. Comparison with different input data

To ensure that hand movement information is preserved in each frame and to integrate the optical flows into the training data, we tried different inputs as shown in Table IV. Mask+Skeleton+Optical flow has the best performance among all the accuracy rates. We believe that the decrease in accuracy with Face landmarks is due to the lack of effective data augmentation for facial features. We will then enhance the features of different channels of Luminance of Mask + Skeleton + Optical Flow to improve the recognition accuracy.

D. Mask+Skeleton+Optical flow

Optical flow color adjustment

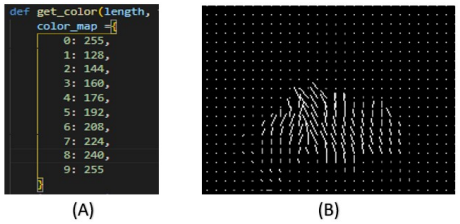
In sign-language videos, the more the hand moves, the longer the optical flow line is, as shown in Figure 9. We tried to differentiate the luminance value for different optical flow lines by giving a higher luminance value for longer optical flow lines and a lower luminance value for shorter optical flow lines. After examining the optical flows in the American Sign Language dataset, the lengths of most optical flows are lower than 10 pixels. In Experiment 1, we made minor adjustments. Selecting the luminance value from [128, 255] and assigning the values according to Figure 9.1 (A). The effect is shown in Figure 9.1(B). We wonder whether the adjustment is not large enough so similar results are observed. Therefore, Experiment 2 selects the luminance value from [16, 255] and assigns the values according to Figure 9.2 (A) The effect is shown in Figure 9.2 (B). Nevertheless, the accuracy dropped to 71.32%, which indicates that the difference in the optical flows was too large. We thought about keeping only the

TABLE IV
COMPARISON OF DIFFERENT DATA INPUTS

Data type	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy	Top-10 Accuracy
1.S+V+Optical flow	67.44%	82.95%	86.43%	90.31%
2.B+R+Optical flow	68.22%	85.66%	85.27%	91.47%
3.Mask+Skeleton+Optical flow	**74.81%**	**86.05%**	**87.21%**	**92.64%**
4.Mask+Skeleton+Optical flow+Face landmark	67.05%	79.84%	83.72%	88.37%

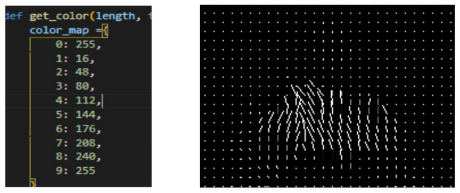
TABLE V
OPTICAL FLOW COLOR ADJUSTMENT

Experiment	Optical flow color	Top-1 Accuracy
Base	No Adjustment	**74.81%**
1	128-255	**74.81%**
2	16-255	71.32%
3	128-255 (focus on hand region)	72.48%



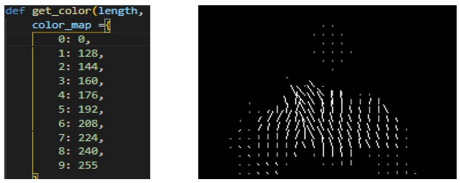
(A) (B)

(1) Experiment 1.



(A) (B)

(2) Experiment 2.

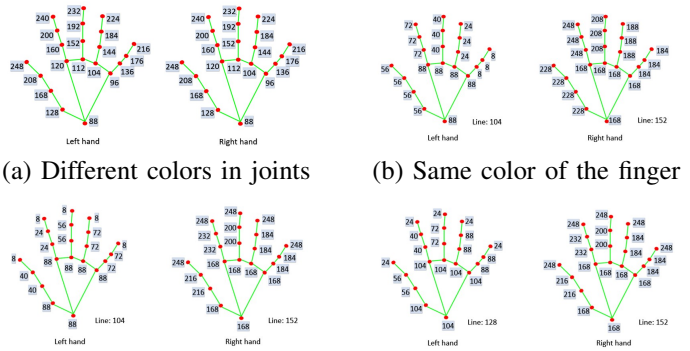


(A) (B)

(3) Experiment 3.

Fig. 9. Optical flow adjustment.

optical flow lines in the hand region and removing the points in the figure so that the model could focus on the hand region. The assignment of Experiment 3 is shown in Figure 9.3 (A), and the effect is shown in Figure 9.3 (B). The accuracy of Experiment 3 indicates that the dots were effective for sign language recognition. The results seem to show that adjusting the luminance value does not improve the accuracy. Therefore, the luminance value of the optical flow segment is maintained as a fixed value, 255, and the dots of the optical flow figure are retained in the following experiments.



(a) Different colors in joints (b) Same color of the finger

(c) Same fingertip of the hand (d) Brighter and darker hands

Fig. 10. Different skeleton adjustments.

Skeleton color adjustment

Based on the results of Table VI in different types of multi-color skeletons, we believe that adjusting the colors of the skeletons is quite helpful for sign language recognition. Therefore, as shown in Table VII, we conducted the experiments of color adjustment in skeletons for different hand regions, such as fingertips and fingers. The color configuration of Experiment 1 is shown in Figure 10(a). The left and right hands are in the same configuration. Each line of a skeleton and the joint are assigned a different color. The line's color is obtained by averaging the values in adjacent joints. The result of Experiment 1 shows that the accuracy is not always improved by assigning more colors. Figure 10(b) is the color configuration of Experiment 2. We reduced the number of colors. Giving different colors to the left and the right hands increases the recognizability. The model learns the finger-based features by assigning the same color to the same finger. The result of Experiment 2 indicates that the finger-based may be too coarse for the model to learn features. Figure 10(c) also shows that the model can learn the finger positions. Since the fingertips are the most common part of the palm, we dispatch the fingertips with the same color and configuration in Experiment 2. The result shows that our adjustment works correctly. The color configuration of Experiment 3 can improve the accuracy, so we further

TABLE VI
SKELETON COLOR ADJUSTMENT

Data type	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy	Top-10 Accuracy
1.Red line+Red joint	65.89%	82.56%	88.37%	90.31%
2.Red line+Green joint	69.38%	**85.27%**	88.37%	**92.64%**
3.Multi-Color skeleton	**72.87%**	84.88%	**89.15%**	91.86%

TABLE VII
SKELETON ADJUSTMENT EXPERIMENTS

Experiment	# of Colors	Adjustments	Left hand=Right hand	Line color	Top-1 accuracy
1	21	Different in every joint	v	(Sum of Adjacent Joint)/2	74.81%
2	L:7, R:7	2.1.Same color in the same finger 2.2.Right hand is brighter than left hand	x	Left:104, Right:152	74.03%
3	L:7, R:7	3.1.Same color in the same fingertip 3.2.Different colors between fingers	x	Left:104, Right:152	76.36%
4	L:7, R:7	Brighter left hand in Exp3.	x	Left:128, Right:152	**76.74%**

brightened the darker left hand in Experiment 4, as shown in Figure 10(d). Finally, we adjusted the skeleton color to achieve the best accuracy of 76.74%.

Mask shape adjustment

From the results shown in Table VIII, we select the color configuration with the highest accuracy and apply the skeleton adjustment strategy to change the shapes of the hand masks to reveal different pixels of the original hand to improve the robustness of the model. The results in Table VIII show that 79.84% accuracy can be achieved for the model trained with the hand mask generated by the skeleton adjustment strategy for the shrinking palm. In addition, we also validate our method on the larger dataset WLASL300. The results demonstrate the highest accuracy for the skeleton adjustment strategy of the shrinking palm.

TABLE VIII
MASK SHAPE ADJUSTMENT

Skeleton adjust policy	WLASL_100 Top-1 Accuracy	WLASL_300 Top-1 Accuracy
Move whole finger	74.42%	57.34%
Random noise	74.81%	60.48%
Finger noise	75.97%	60.63%
Shrink palm	**79.84%**	**61.08%**

E. Comparison with Other Models

Table IX shows that the accuracy of the proposed method is further improved in WLASL100 compared to the GCN-based, Transformer-based model. Similar improvement can be observed in WLASL300.

TABLE IX
COMPARISON WITH OTHER MODELS

Paper	Model	WLASL_100 Top-1 Accuracy	WLASL_300 Top-1 Accuracy
GCN-BERT	GCN	60.15%	42.18%
SPOTER	Transformer	63.18%	43.78%
WLASL	I3D	65.89%	56.14%
Ours	**3D ResNet**	**79.84%**	**61.08%**

V. CONCLUSION

We utilize Mediapipe to extract the hands and examine several data input combinations to see which combination demonstrates the best performance. We use the American sign language dataset to verify the usage of combining hand masks, hand skeletons, and optical flows, serving as the three channels of the input to the model. We have made different adjustments on each of the three channels, such as introducing the color weighting of the skeletons according to different parts of the hands and changing the shapes of the hand masks to reveal different hand pixels to increase the robustness of the model. Through careful data enhancement and training with a moderate 3D-ResNet model, we can achieve very good accuracy on WLASL100. For future work, we may consider combining multiple sign-language words to form a complete sign-language sentence by applying the proposed method and training with transformer-based models. GAN-related methods can simulate the lost skeletons when the hands move too fast. In addition, the virtual hands could be generated if the skeleton is used to generate more training data.

ACKNOWLEDGMENT

This research is supported by the National Science and Technology Council, Taiwan, under Grants NSTC 111-2221-

E-008-098 and 112-2221-E-008-077.

REFERENCES

[1] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, "How2sign: A large-scale multimodal dataset for continuous american sign language," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2735–2744.

[2] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459–1469.

[3] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 357–360.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[5] F. Yasir, P. C. Prasad, A. Alsadoon, and A. Elchouemi, "Sift based approach on bangla sign language recognition," in *2015 IEEE 8th international workshop on computational intelligence and applications (IWCIA)*. IEEE, 2015, pp. 35–39.

[6] M. Al-Rousan, K. Assaleh, and A. Tala'a, "Video-based signer-independent arabic sign language recognition using hidden markov models," *Applied Soft Computing*, vol. 9, no. 3, pp. 990–999, 2009.

[7] M. W. Kadous *et al.*, "Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language," in *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, vol. 165. DE Wilmington, 1996, pp. 165–174.

[8] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with kinect depth camera," *IEEE transactions on multimedia*, vol. 17, no. 1, pp. 29–39, 2014.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[10] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[13] ——, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[15] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[17] M. Boháček and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 182–191.

[18] Yuxin Wu and Alexander Kirillov and Francisco Massa and Wan-Yen Lo and Ross Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[19] Massa, Francisco and Girshick, Ross, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch," https://github.com/facebookresearch/maskrcnn-benchmark, 2018, accessed: [Insert date here].

[20] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[23] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Analysis of the Interaction Effect on Pruning and Transfer Learning in Model Training

Yu-Jen Wei, Jia-Hong Chen, and Tien-Ying Kuo
Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan
E-mail: t106319012@ntut.org.tw, asher.chen@sunplus.com, tykuo@ntut.edu.tw

***Abstract*— Pruning and transfer learning are extensively applied in developing convolutional neural network algorithms, and at times, they are concurrently employed as part of the training model strategy. However, the existing literature has only focused on studying these techniques separately, with no research focusing on the interaction effect between the two technologies. Therefore, we designed several scenarios to test and explore this issue in different target tasks. The experimental results showed significant performance differences among various strategies employed for the same task.**

I. INTRODUCTION

Convolutional neural networks (CNNs) have achieved remarkable success in computer vision. In the design process of these algorithms, increasing the number of convolutional layers is a common approach to improve model performance. However, such design choices can result in challenges related to training model convergence as well as an increase in the number of parameters and computational demands, thereby making it more difficult for algorithm development and deployment. In addition, while most tasks have sufficient training data for model training, certain specific tasks still have limited training data. Model compression [1-5] and transfer learning [6, 7] are the commonly used techniques for addressing these two issues, respectively.

The performance of the model architecture can be improved by increasing its depth, but it may result in redundant parameters [8]. To address this issue, we can employ model compression techniques to simplify the model, thus reducing the required memory space and computational burden. Common methods include low-rank approximation [1, 2], knowledge distillation [3], compact network design [4], and pruning [5], among others. The low-rank approximation [1, 2] reduces the overall number of parameters and computations by decomposing the multiplication operation of two large matrices into multiple smaller matrices. Knowledge distillation [3] refers to the process of using a pre-trained complex network (teacher network) to assist in training a lightweight network (student network), thus reducing the model capacity while maintaining the performance of the teacher network. The compact network design [4] applies the design of lightweight network architectures that enable the model to have extremely low parameters and computational requirements while maintaining high performance. Pruning [5] aims to measure the importance of parameters in a model by designing selection criteria, removing relatively redundant parameters, transforming a dense network into a sparse one, and achieving model simplification.

Transfer learning is derived from the human ability to apply previously learned knowledge to unfamiliar tasks, enabling easier problem-solving and achieving better outcomes. This concept can leverage CNNs to improve convergence speed during model training and achieve better performance [7]. We can transfer the knowledge and features learned from the pre-trained model on the source task as prior knowledge to facilitate the training of the model for the target task.

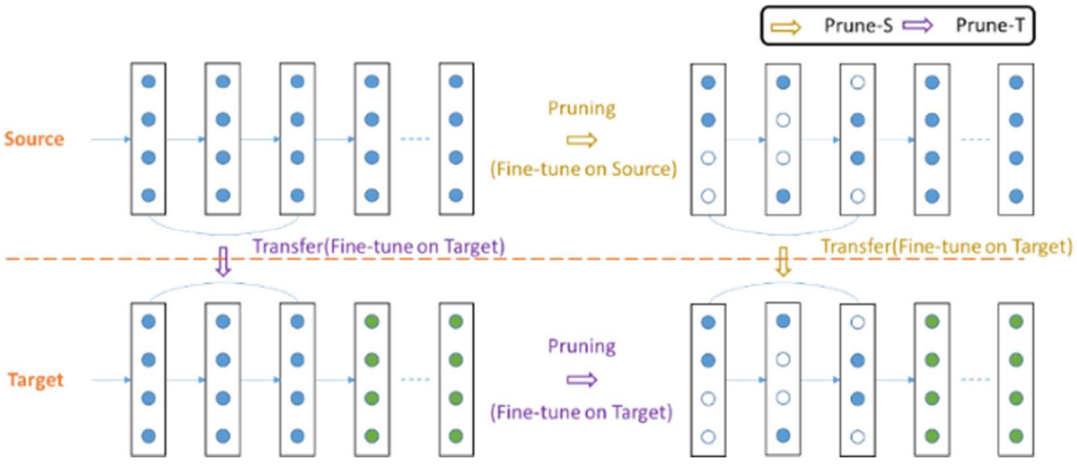


Fig. 1 The operational process of the Prune-S and Prune-T strategies.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Table 1 Integration of information from the source task and the target task.

Transfer Learning	Computer Vision Task	Abbrev.	Image (-centric)	Dataset	Architecture	# of Class	# of Training Data
Source Task	Classification	C	Object	ImageNet	VGG16 [9]	1,000	1.28 M
Target Task	Classification	C1	Object	Flowers-102	VGG16 [9]	102	6.5 k
	Classification	C2	Scene	Indoor-67	VGG16 [9]	67	5.5 k
	Detection	D	Object	Pascal VOC 2007	Faster R-CNN [10]	20	5 k
	Semantic Segmentation	S	Object	Pascal VOC 2012 + SBD	FCN [11]	20	10 k

Typically, we select the shallow convolutional layers of the model for transfer learning since they are responsible for learning distinct types of image features at different levels. Among them, the shallow convolutional layers can learn more general features, while the deep convolutional layers tend to focus on learning task-specific features. In addition to the mentioned advantages, transfer learning can also overcome the problem of insufficient training data [12]. The model can learn to analyze image content from a source task with large data. Then, this model is guided by a smaller amount of data from the target task to learn how to solve the problem of the target task.

Currently, in the development of CNN-based algorithms, it is common to employ not only model compression or transfer learning individually but also both techniques simultaneously. However, to the best of our knowledge, the existing literature predominantly focuses on the investigation of individual techniques, lacking an exploration of the simultaneous application of these two techniques. Therefore, the experiment was conducted with relevant experiments to serve as a reference for researchers in designing training model strategies.

We analyzed the impact of different model simplification strategies on model performance by adjusting three factors: the order of utilizing model compression and transfer learning techniques, the quantity of the target task dataset, and varying compression ratios of the model. During our analysis, we observed that not all model compression techniques could be effectively combined with transfer learning, and such as low-rank approximation, knowledge distillation, and compact network design may not be compatible with transfer learning. These methods require a redesign of the model architecture to simplify the model. As a result, the architectures used in these methods differ significantly from the model architecture of the source task, making the transfer infeasible. Therefore, to work with transfer learning, our experiments use the L1-norm filter structured pruning [5] to compress the model.

II. PROPOSED METHOD

We define the names of two pruning strategies: Prune-S (Prune-on-Source) and Prune-T (Prune-on-Target). Prune-S means pruning the model trained on the source task and fine-tuning it on the target task. Prune-T involves first transferring the model trained on the source task to the target task and then pruning it using the training data of the target task. The operational flow of both strategies is shown in Fig. 1, where the middle orange dashed line serves as a boundary separating the upper and lower sections, representing the source task and the target task individually.

For the selection of the source task, we opted for the commonly used image classification task (C) in transfer learning, along with the widely used VGG16 backbone [9] in classification task architectures. In the target task, we not only selected image classification tasks (C1, C2) similar to the source task but also used object detection (D) and semantic segmentation (S) tasks for evaluation. In selecting the datasets, we considered the image relevance between the source and target tasks to avoid negative transfer during transfer learning. This ensures that the chosen datasets are aligned to minimize any potential negative effects during transfer learning. We have presented detailed information on the source and target tasks in Table 1. Among them, Faster R-CNN [10] and FCN [11] are designed based on the VGG16 architecture. Therefore, we employ these two architectures for tasks D and S, respectively.

Table 2 The result of pruning the source task model

Dataset	Model (VGG-16)		Top-1	Top-5	# of Parameter			FLOPs		
					w/ FC	w/o FC	Times (↓)	w/ FC	w/o FC	Times (↓)
ImageNet	w/o Pruning		71.59	90.38	138.36 M	14.71 M	-	15.61 G	15.36 G	-
	Prune Ratio	25%	70.36	89.68	134.43 M	10.79 M	1.36	9.40 G	9.15 G	1.68
		50%	69.22	89.14	131.44 M	7.8 M	1.89	4.92 G	4.67 G	3.29

Table 3 Hyperparameter configuration for training the model

Hyper-parameters	Source Task	Target Task			
	C	C1	C2	D	S
Optimizer	SGD + momentum				
Weight decay	1e-4	1e-4	1e-4	5e-4	5e-4
Image resolution	224x224	224x224	224x224	Depend on data	Depend on data
Training epoch	-	45	75	22	40
Initial learning rate	-	1e-3	1e-3	1e-3	1e-10

Next, we will explain the data allocation for the target task. Since the limited availability of data for specific tasks compared to others, we simulate this situation by adjusting the amount of data used for the target task. In our experiment, we evaluate the impact of the amount of training data on the Prune-S and Prune-T strategies by using two different configurations, as shown in Fig. 2. First, we fine-tune the model parameters using the entire dataset (100%) of the target task. Second, we randomly select 10% of the data from the dataset and use it for fine-tuning. This allows us to assess the influence of the quantity of training data on the Prune-S and Prune-T strategies.

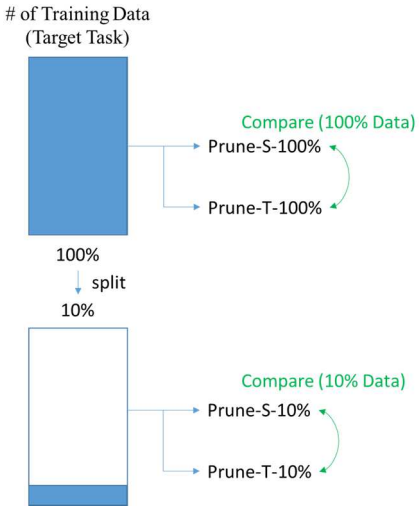


Fig. 2 Illustration of training data splitting.

The model pruning ratio also has a significant effect on the performance of the model. Hence, we will observe the situation when different pruning ratios are used. We applied 25% and 50% pruning ratios to compress the model for the source task, leading to retentions in model parameters by 75% and 50%, respectively. In the following sections, we refer to these as the low pruning ratio (25%) and high pruning ratio (50%). We specifically targeted the first 10 layers of the model for pruning and transfer. This decision was based on the following reasons: 1) As mentioned earlier, shallower convolutional layers tend to learn more generic features. 2) Zhang et al. [1] mention that the parameters in the shallower layers of VGG16 are relatively redundant, making pruning in these layers have a smaller impact on model performance. 3) More than 90% of the computational burden in VGG16 comes from the first 10 convolutional layers, indicating that pruning these layers would be more beneficial. Table 2 presents the results of the pruning on the source task. In Table 2, "w/ FC" indicates that the computed result includes the fully connected layer, while "w/o FC" indicates the exclusion of this layer.

III. Experiment Results

We conducted the experiments on hardware with a CPU Intel i7-7700K and a GPU NVIDIA 2080Ti. Table 3 shows the hyperparameters used during the training of our models, where the model parameters for the source task were obtained from PyTorch [13]. Both the Prune-S and Prune-T strategies were evaluated for the target tasks C1, C2, D, and S. The models were fine-tuned using 100% and 10% of the training data from the target task. Then, their performance was analyzed under different pruning ratios. In the evaluation of model performance, we used Top-n, mAP, and mIoU as the evaluation metrics for image classification (C), object detection (D), and semantic segmentation (S) tasks, respectively. The experimental results are shown in Table 4, which presents the accuracy differences obtained by subtracting the performance of the Prune-T from the Prune-S strategies.

The experimental results confirm that the Prune-S strategy outperforms Prune-T in most conditions, especially in the case of C2 and S, where it demonstrates a significant advantage over Prune-T. The experimental results confirm that the Prune-S strategy outperforms Prune-T in most conditions, especially in the case of C2 and S, where it demonstrates a significant advantage over Prune-T. We can notice that only target task D has better results for Prune-T at low pruning ratios.

We attribute this to the high pruning ratio in the Prune-S strategy, which results in the convolutional layers extracting more generic features which apply to all tasks. Therefore, Prune-S outperforms Prune-T in all cases of high pruning ratios. In contrast, the convolutional layer of the low pruning ratio model in the Prune-S strategy captures not only the general features but also the redundant ones for the target task. These redundant features have a negative impact on the training of the target task, resulting in better performance of Prune-T. When there is low object relevance between the source task and the target task, and a model with a low pruning ratio is employed, it is difficult to determine which strategy is superior for training the target task.

Table 4 The results of the accuracy difference between Prune-S and Prune-T on the target task.

# of Training Data	Prune Ratio	Low pruning ratio (25%)				High pruning ratio (50%)			
	Target Task	C1	C2	D	S	C1	C2	D	S
100%	Prune-S − Prune-T	0.09	0.1	−0.48	0.28	0.33	3.63	0.69	1.31
10%	Prune-S − Prune-T	0.73	0.8	−0.64	0.6	0.71	4.86	0.46	4.94

Finally, we will discuss the time required for model compression under the conditions of 100% data volume and a high pruning ratio. In our experiments, it takes 15 days to prune the model for the source task (C), 1.5 hours for the target tasks C1 and C2, and 1.5 days for the target tasks D and S. The significant difference in time cost is due to the iterative process involved in the pruning and compressing of the model. The algorithm iteratively identifies and removes the unimportant parameters in the model until the assigned pruning ratio is achieved. As a result, the size of the dataset used during the compression process will impact the time required to compress the model. While the compression time for the source task model is several times longer than target tasks, the compressed model can be transferred to other target tasks, reducing the development time for algorithms.

IV. CONCLUSIONS

This study explores the simultaneous application of transfer learning and pruning techniques in convolutional neural networks, aiming to investigate their combined impact on model simplification strategies. Following the analysis of experimental results, we recommend the combination of the Prune-S strategy and a high pruning ratio model because it is more suitable for application across various target tasks. Despite the model compression process for the source task in Prune-S requiring much more time compared to Prune-T, the advantage is that it only needs to be performed once and can be applied to other tasks. Furthermore, the compressed model allows for the development of algorithms with lower hardware requirements, effectively reducing the development time.

V. ACKNOWLEDGMENT

This research was funded by National Science and Technology Council grant number 111-2221-E-027-065

REFERENCES

[1] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 1943-1955, 2015.

[2] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing neural networks," Advances in neural information processing systems, vol. 28, 2015.

[3] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4320-4328.

[4] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4373-4382.

[5] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," arXiv preprint arXiv:1608.08710, 2016.

[6] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27, 2018: Springer, pp. 270-279.

[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1717-1724.

[8] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," Advances in neural information processing systems, vol. 27, 2014.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440.

[12] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4918-4927.

[13] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Old Damaged Photo Recovery with Style Transfer-Based Data Augmentation

Chih-Hao Wang, Yu-Jen Wei, Ching Hsiang Chang, and Tien-Ying Kuo
Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan
E-mail: hank19960918@gmail.com, {t106319012, t110318081}@ntut.org.tw, tykuo@ntut.edu.tw

***Abstract*— Restoring damaged old photos to their original state can enhance the viewing and comprehension of the historical context and time period captured by these photos. Recent studies attempted to restore these photos using proposed methods, but the results often suffer from blurring and loss of details. Therefore, we propose a restoration model designed explicitly for damaged old photos in this paper. Furthermore, due to the limited quantity of damaged old photos compared to other task datasets, we employ data augmentation techniques to increase the quantity and diversity of training data for the restoration method. The experimental results demonstrate that our restoration method surpasses existing methods across various evaluation metrics and provides a satisfying viewing experience.**

I. INTRODUCTION

Old photos preserve scenes from the past, allowing future generations to reminisce about the lives and events of that time. Unfortunately, many photos preserved until today often suffer from damage such as creases, mold, and insect infestation due to poor storage. Removing these damages from the photos would significantly enhance their visual quality and make viewing more enjoyable.

The most straightforward restoration method is to infer the damaged content based on the intact portions and use appropriate tools to restore the photos accordingly. In the restoration process, besides directly repairing physical photographs, we can scan them into digital files and use digital editing software to modify damaged areas. Although this practice improves the viewing quality of the photos, the restoration results heavily rely on the prior knowledge of the operator, and the entire process can be time-consuming.

The advancement of deep learning has led to improvements in addressing this issue, as demonstrated in recent inpainting methods [1, 2]. These methods can generate missing image content by analyzing the manually annotated regions where content is missing and the surrounding image context. Liu et al. [1] proposed the partial convolution structure, which uses a mask of the input region to be restored, allowing the model to avoid referencing incorrect information during the image analysis. Yu et al. [2] designed gated convolution to analyze the features of each layer to generate the corresponding masks, allowing the model to preserve the semantic structure and details of the image while filling in the missing regions. Inpainting methods can produce results more quickly than manual restoration, but the process of manual annotation remains laborious and prone to errors.

To solve this problem, [3, 4] proposed restoration methods that only require the input of damaged images to obtain restored results. Wan et al. [3] designed a module to remove damaged features of the image in the latent space and obtained the results by converting restored features back to the image space. Although [3] can save significant time and labor costs, we also note that this method has limitations in identifying damages. It may mistakenly consider objects in the image as damaged areas and remove them.

In our previous research [4], we developed an automatic recovery system to address the shortcomings of existing methods. This system consists of a damage detection model and a restoration model. The damage detection model would identify the damaged areas in old damaged photos and then utilizes existing restoration methods to obtain the restored photo. Despite the effective damage detection achieved in [4], the restored regions using existing restoration models still suffer from blurriness and loss of detail. The reason is that these restoration methods are designed for modern images and thus cannot effectively analyze the content of old photos.

To obtain results with more details, we proposed a restoration method designed explicitly for damaged old photos and combined it with our previous research [4]. In the development process, we noticed that the limited quantity and diversity of old photos constrained the performance of the model. Therefore, we apply style transfer technology to transfer modern images into old-style images, thus increasing the amount of training data.

II. PROPOSED METHOD

The flowchart of the old photo recovery system is shown in Fig. 1. In the damage detection step, we use the detection model proposed in [4] to generate the damaged area mask, then combine it with the damaged image as inputs for restoration model, which predicts and generates the content of the damaged regions. Additionally, before training the restoration model, we employ our designed data augmentation model to generate many old-style images, expanding the training data for the restoration model.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

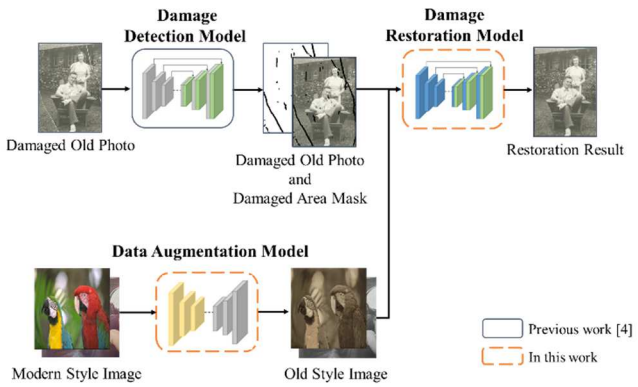


Fig. 1 Flowchart of our old photo recovery system

A. *Damaged Restoration Model*

Our restoration model is based on generative adversarial networks (GANs), as shown in Fig. 2. The adversarial learning between the generator and the discriminator is employed to enhance the realism of the restoration results. We adopt U-Net [5] as the generator backbone architecture because it effectively handles the details of the image. Instead of the standard convolutional blocks in U-Net, we replace them with a combination of Residual Dense Blocks (RDB) [6] and Partial Convolution (PConv) [1] to improve the capability of analyzing damages. RDB allows the model to retain more information from shallow features and accurately predicts the content of the damaged regions by fusing these features. In addition, dense connections and residual learning improve the convergence speed and stability of the deep network. The formula for the partial convolution is written in Eq.(1). The concept is to exploit masks to aid in model restoration, where x' means output feature, W and b represent the weight and bias of the convolution filter, X represents pixel values or feature values at the location of the current convolution filter, M denotes the corresponding binary mask, $sum(1)$ and $sum(M)$ indicate the number of valid pixels and the number of the total pixels in the convolution window, and $\odot$ denotes the element-wise product. The mask needs to be updated while restoring, and the method for the updated mask m' is shown in Eq. (2) which means the center of the convolution window sets to a valid pixel as long as the within region of the convolution window has a valid pixel.

$$x' = \begin{cases} W^T(X \odot M)\dfrac{sum(1)}{sum(M)} + b, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \quad (1)$$

$$m' = \begin{cases} 1, if\ sum(M) > 0 \\ 0, \quad otherwise \end{cases} \quad (2)$$

We adopt the Patch-GAN [7] as discriminator. Unlike typical discriminators of GAN that judge the realism of an entire image, Patch-GAN divides the image into multiple N×N patches and evaluates the realism of each patch. This design allows for effectively identifying inconsistencies in different regions of the generated images, improving the image quality produced by the generator.

B. *Data Augmentation Model*

To expand the training data for our model, we designed a data augmentation model based on the architecture of Cycle-GAN [8], as shown in Fig. 3. We used it to learn how to transform images in a modern style into the style of old photos, thus generating old-style images with diverse contents.

Compared to other style transfer network [7], Cycle-GAN does not require paired image datasets for training, which reduces the difficulty of data collection and model training in this case. To enhance the ability of model to perform style transfer, we replaced the transposed convolution used in the original Cycle-GAN with sub-pixel convolution [9]. This substitution helps to avoid generating images with noise or checkerboard artifacts. Sub-pixel convolution rearranges deep features to generate high-resolution features, improving the level of detail in the generated images.

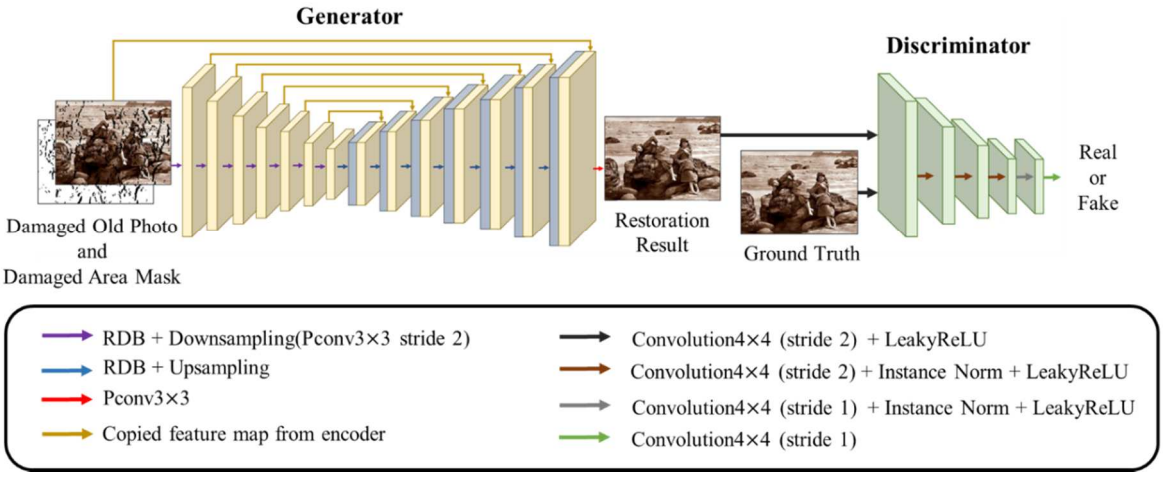


Fig. 2 Architecture of the image restoration model

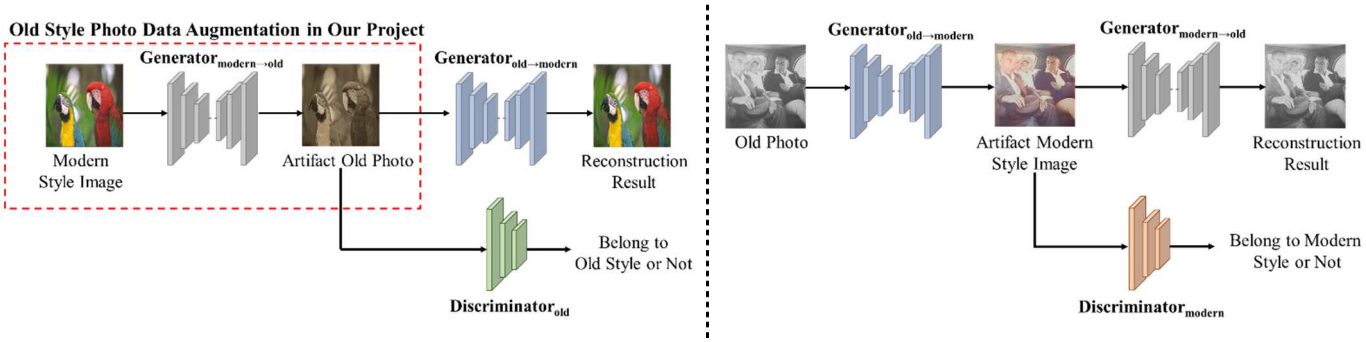


Fig. 3 Architecture of the data augmentation model

We collected modern-style images as our training dataset for the training style transfer model. These images were used separately as the input and target style datasets for the data augmentation model. Specifically, within the target style dataset, we divided the images into two subsets: yellowish style and monochromatic style. We trained the corresponding data augmentation models using these two subsets. This approach was adopted to avoid interference between the two styles and to prevent the generated results from containing both color tones simultaneously.

The Eq. (3) presents the loss function $L_{restoration}$ for our restoration model, which consists of six terms. The valid loss L_{valid} and hole loss L_{hole} are employed to measure the differences between the restored image and the ground truth image in the non-damaged and damaged regions, respectively. The perceptual loss $L_{perceptual}$ compares the feature differences between the transformed images in the feature domain. The style loss. L_{style} evaluates the similarity of styles between images in the feature space. The variation loss L_{tv} is used to assess the smoothness of the boundaries between the damaged and non-damaged regions to ensure seamless blending. The adversarial loss L_{GAN} is employed for the discriminator to determine the authenticity of the generated content. Parameters for each loss term in $L_{restoration}$ are set based on the reference [1].

$$L_{restoration} = L_{valid} + 6L_{hole} + 0.05L_{perceptual} + 120L_{style} + 0.1L_{tv} + 0.01L_{GAN} \quad (3)$$

We utilize Eq. (4) as the loss function, denoted as $L_{transfer}$, for training the two data augmentation models. This loss function consists of an adversarial loss L_{GAN} and a cycle consistency loss L_{cyc}. The cycle consistency loss L_{cyc} ensures that the stylized images preserve the content from the input images. The adversarial loss L_{GAN} helps improve the realism of the stylized images after the style transfer process.

$$L_{transfer} = 10L_{cyc} + L_{GAN} \quad (4)$$

III. Experiment Result

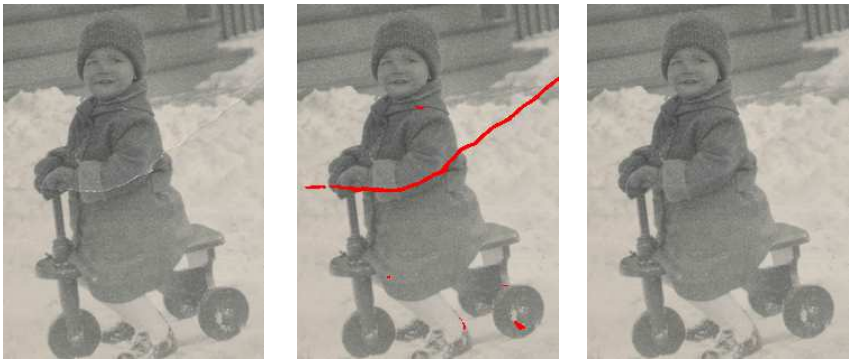
Our experiments were conducted on hardware consisting of a CPU E5-2698 and a GPU TESLA V100. We used four widely used metrics, namely PSNR, SSIM, LPIPS [10], and FID [11], to evaluate the performance of restoration system. Higher values for the first two metrics indicate better results, while lower values for the latter two metrics indicate better results. We obtained synthesized damaged old photos in the same manner as the training process to create a testing dataset for evaluating the restoration method.

Table I presents the results of our method compared to [1-3]. It can be seen that our restoration method outperforms other methods in all metrics, and the performance is further improved with the addition of data augmentation. Furthermore, it is notable that Wan et al. [3] show significant differences in the evaluation results compared to other methods. This is because their approach modifies the undamaged regions during the restoration process, resulting in a substantial discrepancy between the generated images and the ground truth.

Table I Comparison of restoration results

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
Liu et al. [1]	40.35	0.98	0.0102	6.97
Yu et al. [2]	38.25	0.98	0.0678	19.24
Wan et al. [3]	22.30	0.72	0.173	80.22
Our method w/o DA	41.66	0.99	0.0066	4.54
Our method w/ DA	**41.78**	**0.99**	**0.0063**	**4.40**

We show the restored results of different models applied to real damaged old photos in Fig. 4 and Fig. 5. Since there is no corresponding ground truth available for real damaged old photos, we relied on subjective evaluations to evaluate restoration results. In Fig. 4, our method can effectively restore the damage on the sleeve, while restored content of other methods appears more blurry.



(a) Damaged old photo (b) Damage detection (c) Our method



(d) Liu et al. [1] (e) Yu et al. [2] (f) Wan et al. [3]

Fig. 4 The restored result of different models applied to real monochromatic style damaged old photo



(a) Damaged old photo (b) Damage detection (c) Our method



(d) Liu et al. [1] (e) Yu et al. [2] (f) Wan et al. [3]

Fig. 5 The restored result of different models applied to real yellowish style damaged old photo

Fig. 5 shows the results of restoration methods applied to a yellowish-style photo. The content restoration applied to the shoulder area in [2] is significantly inconsistent with the surrounding regions. Wan et al. [3] causes the hand of the baby to appear blurry and indistinguishable, and adjusted the overall tone of the image. In contrast, our method does not have these problems. Overall, our method can distinguish between the damaged patterns and the objects, allowing the restoration method to predict more accurate restoring content that closely resembles the original image texture.

IV. CONCLUSIONS

We proposed a damaged old photo restoration model based on generative adversarial networks (GANs). We combine residual dense blocks and partial convolution to enhance the restoration effect on the details of the damaged areas. In the data augmentation method, we optimize the architecture of Cycle-GAN to generate various old-style images to increase the training data and then improve the generalization capability of the model. In the experiments, we demonstrated that our restoration method produces better image quality and viewing quality than other methods when applied to damaged old photos.

V. ACKNOWLEDGMENT

This research was funded by National Science and Technology Council grant number 111-2221-E-027-065

REFERENCES

[1] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 85-100.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4471-4480.

[3] Z. Wan et al., "Bringing old photos back to life," in proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2747-2757.

[4] T.-Y. Kuo, Y.-J. Wei, P.-C. Su, and T.-H. Lin, "Learning-Based Image Damage Area Detection for Old Photo Recovery," Sensors, vol. 22, no. 21, p. 8580, 2022.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 2015: Springer, pp. 234-241.

[6] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2472-2481.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125-1134.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223-2232.

[9] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874-1883.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586-595.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Advances in neural information processing systems, vol. 30, 2017.

A Deep Learning based Sustainable Energy Scheduling System

Kun-Lin Tsai*, Yan-Hao Chen, Choa-Ting Huang, Guo-Wei Huang, and Shih-Ting Tseng
Department of Electrical Engineering, Tunghai University, Taichung, Taiwan
E-mail: kltsai@thu.edu.tw Tel: +886-4-23590121

Abstract — **Climate change greatly affects human life. According to previous studies, the biggest cause of climate change is the increase of greenhouse gases in the atmosphere. In order to effectively reduce carbon emissions, many countries not only formulate many carbon reduction policies, but also focus on developing green energy. Green energy, e.g., solar energy and wind energy, is a solution to reduce carbon emission; however, it is difficult for green electricity to provide stable energy supply due to climate factors. In this paper, we proposed a long short-term memory (LSTM) deep learning model based sustainable energy scheduling system to control the energy usage between green power and utility power. The proposed system analyzes the data of green power generation and weather conditions, and based on the weather forecast and the results calculated by this system, predicts the power generation status of solar and wind energy in the next few days. According to the experimental results, by using the proposed system, the goals of stable power supply and minimization of carbon emissions from energy consumption can be achieved.**

I. INTRODUCTION

In recent years, the issue of global warming has led to a surge in severe weather conditions, consequently driving up electricity consumption. To minimize the environmental repercussions of energy usage, many countries tried to provide various sustainable solutions, with a particular focus on green energy sources, including hydropower, solar energy, wind energy, biomass energy, geothermal energy, etc. According to the estimations of the International Energy Agency (IEA), by the year 2050, green energy can potentially contribute to more than 70% of the world's electricity supply [1]. Several countries have already pledged to achieve zero carbon emissions, emphasizing their commitment to reducing carbon footprints and transitioning away from fossil fuels towards 100% renewable energy.

Even though solar energy and wind power are two main sources of green energy, the intermittent nature of solar and wind power poses a challenge in maintaining a consistent power supply due to climate variability. To maximize the generation of green electricity, numerous wind-solar hybrid systems have been proposed [2-4]. Several wind-solar hybrid products are also available in the market that incorporate energy storage batteries. These innovative solutions, e.g., street lamps and small power supplies, have been designed and implemented to harness the combined potential of wind and solar energy. By integrating different technologies, the hybrid systems can optimize energy production and storage, providing a reliable and sustainable source of electricity.

The objective of this study is to reduce the impact of climate-related variables on the stability of green power supply. To address this issue, in this paper, a deep learning based sustainable energy scheduling method is proposed. The proposed method involves the integration and analysis of solar and wind power generation data in conjunction with diverse weather patterns. The collected data is then visualized and utilized in a long short-term memory (LSTM) forecasting model. By using deep learning techniques, the model compares various weather patterns, the predicted electricity outputs, and actual monitored electricity data to predict the power generation status of solar and wind energy in the next few days.

The experimental results demonstrate the effectiveness of the proposed method in controlling the storage capacity of green power. By accurately forecasting power generation, the method enables the optimized utilization of green power resources, even during adverse weather conditions, so as to enhance the overall efficiency and reliability of green power systems.

II. RELATED STUDIES

Various studies [5-10] have been conducted to enhance the effective utilization of green energy. Chen *et al*. [5] conducted a comprehensive review focusing on the application of electric power in wind energy systems. They examined different types

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of wind turbines, highlighting variations in performance and controllability. Among these, the SCIG (Squirrel Cage Induction Generator) incorporating the VSC-HVDC (Voltage Source Converter-High Voltage Direct Current) concept demonstrated remarkable adaptability to grid faults. Wang *et al*. [6] addressed the fan configuration of wind farms in their research. As wind passes through the turbines, the blades absorb the wind's kinetic energy, inducing turbulence and reducing wind speed, which is known as the wake effect. Wang's paper introduced a novel non-linear wake expansion model suitable for flat or slightly undulating terrains and offshore wind farms. Additionally, the model incorporated cost-benefit calculations to determine the most economically efficient arrangement of wind turbines in a wind power field.

Bukala *et al*. [7] addressed the issue of small wind turbines, primarily catering to individuals and small-scale users. They recognized that wind turbine sellers often fail to provide customers with crucial performance data, including environmental parameters, technical specifications, and power output. Consequently, their objective was to develop a concise framework capable of describing the factors influencing wind turbines. Touafek *et al*. [8] focused on enhancing solar radiation power through solar energy concentration. The efficiency of the concentrator relied on the radiation concentration and the temperature difference between the absorber and the surrounding air, where the latter contributes to inevitable heat loss. This heat loss directly impacted the overall efficiency, necessitating further investigation into materials that can minimize it. Their study introduced a novel approach by employing two reflectors on a hybrid PhotoVoltaic Thermal (PVT) collector to concentrate solar energy and improve production efficiency.

Bedreddine *et al*. [9] conducted research on parabolic solar panel concentrators, aiming to achieve a focal point temperature of 400 degrees Celsius. They identified various parameters and validated their findings by assessing the performance of zinc oxide thin film batteries at different temperatures. Compared to traditional methods, their approach demonstrated higher safety and efficiency in converting green energy. Reza [10] developed a solar tracking system utilizing Arduino as a development board. This system continuously monitored the sun's position and adjusted the solar panel accordingly to optimal conversion efficiency. Implementing a single-axis solar tracking system on a 10W solar panel resulted in an average power increase of approximately 1W per hour. Furthermore, when applied to a 100W battery, the power output improved by up to 80W.

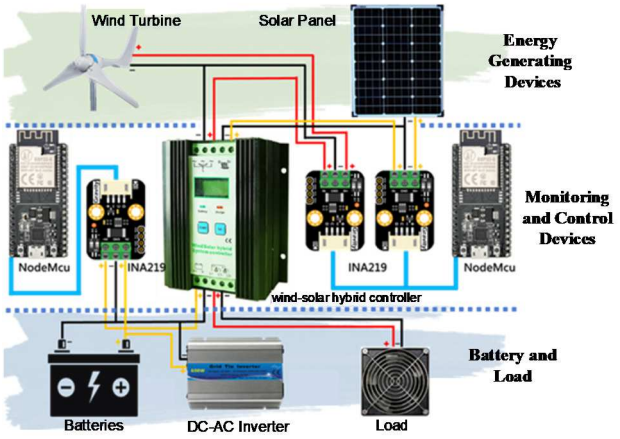


Fig. 1 System Architecture of proposed wind-solar hybrid system.

While the aforementioned papers presented promising solutions, there is still room for improvement in maximizing the efficiency of green energy utilization, particularly in light of climate-related challenges. Therefore, the primary focus of this paper is to introduce a method aimed at optimizing the efficiency of green energy usage.

III. WIND-SOLAR HYBRID SYSTEM ARCHITECTURE

The system architecture, shown in Fig. 1, comprises three main levels: the energy generating devices, the monitoring and control devices, and the storage and load end. At the production side, solar panels and wind turbines are employed to generate electricity. The monitoring and control side facilitates the combination of wind and solar power and ensures the battery is charged without the risk of overcharging. This is achieved through the utilization of voltage and current sensors, which collect data. The collected information is then transmitted to a computer via Wi-Fi using the NodeMCU platform. The data, including voltage, current, and power, is transmitted using the MQTT protocol and Python is employed to connect to the MQTT Broker for data acquisition. Simultaneously, a database is generated to facilitate research and analysis. The storage and load end encompass key components such as batteries, DC-AC inverters, and fan loads. Monitoring between the battery and the wind-solar hybrid is conducted using INA219 voltage and current sensors. These sensors provide insights into the voltage and power status of the battery. The subsequent sections provide a detailed description of each system component.

A. Energy Generating Device

In terms of energy collection, this paper utilizes two components.

(1) Wind Turbine: The energy collection system incorporates a set of 100W five-blade wind generators. These

generators operate based on the principle of converting wind power into electrical energy. The rotation of the five blades is driven by the force of the wind, and the wind turbine head facilitates the conversion of wind energy into electrical energy. To optimize energy generation, the empennage is designed to adjust the wind turbine's direction according to the prevailing wind direction.

(2) Small Monocrystalline Silicon Solar Panel: The energy collection system incorporates a 50W solar panel, specifically utilizing monocrystalline silicon technology. Solar panels can be classified into different types based on the materials and manufacturing methods employed, including silicon, compound, and organic semiconductor types. Among these, silicon-based panels are the most commonly used. Polysilicon panels are cost-effective due to their simpler manufacturing process, but they exhibit lower conversion efficiency. On the other hand, monocrystalline silicon panels offer higher power generation efficiency (>20% module efficiency) and longer lifespan due to the higher purity of the silicon element used.

It is worth noting that the temperature has a notable influence on the efficiency of solar power generation. Lower temperatures lead to slower current drop rates, resulting in more stable current output compared to high temperatures. With increasing temperature, the output voltage and maximum power of the solar cell decrease. Consequently, it becomes necessary to monitor the actual power generated by the solar cells under varying conditions. Weather data is recorded as sample data, allowing for a comparison between weather types and their respective impact on power generation. This analysis aims to determine the degree of correlation between weather conditions and the power generated by the solar cells in different environments.

B. Monitoring and Control Devices

The monitoring and control end of the system encompasses a wind-solar hybrid controller and power sensors. The wind-solar hybrid controller performs the crucial tasks of transmitting and storing power generated by both wind turbines and solar energy sources. It effectively manages and balances the power generation status of both sources to achieve a synergistic and efficient power storage effect. This controller is designed to automatically adjust the power generation operations of the wind turbines and solar energy systems.

During daylight hours, power is generated simultaneously by both solar energy and wind turbines. Conversely, during nighttime hours, only wind turbines generate electricity, as solar energy is limited to daytime generation. The wind-solar hybrid controller optimizes the utilization of both energy sources, ensuring a continuous and balanced power supply while maximizing energy storage capabilities.

C. Battery and Load

One of the batteries used in the proposed system is the Yuasa REC22-12, which is a 12V/22Ah battery. This battery is specifically designed for uninterruptible power supply (UPS) systems. It exhibits exceptional recovery performance after deep discharge and demonstrates an ultra-low self-discharge rate. The maximum charging current for this battery is 5.5A.

The charge-discharge rate of a battery is defined as the ratio of the current during charging or discharging to the rated capacity of the battery. It represents the speed at which the battery can be charged or discharged. This rate is commonly denoted by "C" and is expressed in units of time. For example, a 1C discharge signifies that the battery is discharged in one hour, while a discharge rate of 0.2C indicates that the battery is discharged over a period of five hours. The charge-discharge rate is calculated using formula (1), where I represents the charge-discharge current, Cap represents the rated capacity of the battery, and C_{Rate} represents the charge-discharge rate.

$$C_{Rate} = \frac{I}{Cap} \tag{1}$$

The load side of the system comprises a DC-AC converter and a fan. The DC-AC converter is responsible for converting the DC power generated by the battery into AC power, producing an analog sine wave. This converter is designed to ensure compatibility with connected devices, such as motors, without causing any damage to their components. By directly connecting the devices to the battery, uninterrupted power supply requirements are met, guaranteeing a continuous power source. This requirement is also extended to the NodeMCU, which requires a stable voltage supply for real-time data transmission and updates.

D. Energy Collection and Monitoring Methods

In order to maximize the role of green electricity and ensure a stable power supply, we have developed a sustainable energy utilization and deployment system based on a hybrid of wind and solar power. Fig. 2 provides an actual layout diagram. The system's production capacity encompasses both solar energy and wind energy, with an energy storage battery being charged through a wind-solar hybrid controller.

The wind-solar hybridization offers several advantages. During periods when there is no wind but ample sunlight during the day, the solar panels take charge of power supply and also recharge the energy storage battery. Conversely, the solar panels are inactive at night. In the presence of wind, the power generation is facilitated by wind energy. However, if



Fig. 2 Implement of the proposed system.

there is insufficient wind or no wind at all, the energy storage batteries serve as a backup power source. Furthermore, our system is capable of monitoring weather conditions during the deployment of wind and solar energy. By integrating weather data with algorithms, it can assess weather patterns and optimize power generation accordingly.

The key functionalities of the proposed system include:

(a) Energy Use and Conversion: Our system optimizes the deployment of solar energy by utilizing wind and solar complementarity. This ensures efficient conversion of solar energy into electrical energy, which is then stored for later use. By improving the energy conversion efficiency, the energy of the system is optimized.

(b) Research and Analysis of Electricity Storage and Discharge: To facilitate effective management of electricity storage, we record the hourly electricity storage status and conduct detailed analysis. This analysis incorporates various weather factors such as weather conditions and wind strength. By considering these factors, we can accurately predict the usage of electric energy and make informed decisions regarding its discharge.

(c) Data Monitoring and Transmission: For comprehensive monitoring, we employ the NodeMCU system equipped with the INA219 sensor. This combination allows us to monitor voltage, current, and power parameters. Additionally, we incorporate the MQTT transmission protocol to facilitate efficient and organized data transmission. The monitored data, including voltage, current, and power information, is transmitted every 10 seconds via wireless transmission to a server for storage. The server then generates a data file containing essential information such as voltage and current, enabling further analysis, calculations, and recording of power storage data and weather conditions. Table 1 lists the monitored battery voltage, current, and power.

Table. 1 Energy data of 2022-04-22.

Time	Voltage (V)	Current (mA)	Power (mW)
2022-04-22 12:00:23	13.52	238.16568	3219.999994
2022-04-22 12:00:33	13.51	460.39970	6219.999947
2022-04-22 12:00:43	13.78	960.81277	13239.99997
2022-04-22 12:00:54	13.56	843.65782	11440.00004
2022-04-22 12:01:04	13.65	646.84982	8829.500043
2022-04-22 12:01:14	13.83	809.83369	11199.99993
2022-04-22 12:01:24	13.85	1088.8087	15080.00050
2022-04-22 12:01:34	13.58	494.84536	6719.999989
2022-04-22 12:01:45	13.51	387.86084	5239.999948

E. Weather Pattern

Web crawling, an automated process of extracting data from webpages, was employed in this study using the Spyder (Anaconda2) Python programming language. The weather data utilized in this research was obtained from the weather.com website. The crawler program was developed to retrieve weather-related information from the Weather.com webpage and store it in an Excel file format. The extracted data encompasses various parameters including forecast time, temperature, rainfall, UV index, cloud cover, and more. Subsequently, this data was analyzed and visualized to enhance the weather forecasting program.

F. Long Short-Term Memory Model

The LSTM (Long Short-Term Memory) model which is a recurrent neural network (RNN), aimed to deal with the vanishing gradient problem, is a valuable approach for data analysis and forecasting, particularly when examining the correlation between historical data and predicted future data. It is applicable to classification, processing and predicting data based on time series. It leverages the concept of time series analysis to address the irregularity inherent in monitoring systems influenced by varying weather conditions on a daily basis. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

With different weather patterns corresponding to each time point, the complexity of the problem requires a more advanced approach like time series analysis. By considering the sequential nature of the data, the LSTM model used in the proposed system captures temporal dependencies and patterns that may exist within the dataset, including the sensed voltage data, current data, power data, and weather data. This enables more accurate predictions and a deeper understanding of the relationship between past observations and future outcomes.

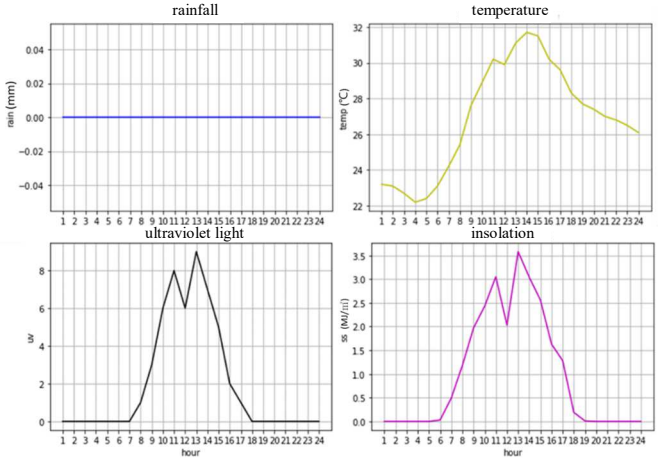


Fig. 3 Weather patterns observed on 2022/04/22.

IV. Experimental Results

A. Weather Patterns

To investigate the impact of weather patterns on the system, this study utilized web crawlers to extract weather pattern data. Fig. 3 displays the weather patterns observed on 2022/04/22. The figure provides the following insights:

(a) The weather conditions were sunny, with no recorded rainfall.

(b) The average temperature exhibited its highest values between 13:00 and 16:00, peaking at 14:00.

(c) Ultraviolet (UV) radiation levels were highest around 13:00 and 14:00, reaching approximately UVI 8.7.

(d) Insolation, which represents the intensity of sunlight, reached its peak at 13:00.

These observations contribute to understanding the specific weather conditions on the mentioned day and their potential impact on the system under study.

B. Solar and Wind Power Generation

Based on the monitoring data, the solar generator's power generation is limited to the time frame from 6:00 am to 6:00 pm. Outside of this period, no power is generated. During unfavorable weather conditions such as cloudy or rainy weather, the voltage, power, and battery voltage decrease noticeably. Using the data from April 22, 2022, as an example, Fig. 4 illustrates the voltage, current, and power curves generated by the solar power system on 2022/04/22.

Upon observing Fig. 4, it is apparent that the voltage begins to rise as it approaches 5 o'clock, indicating the nearing of sunrise. At 6 o'clock, the current and power start to gradually increase, signifying the initiation of the charging process. By around 10 o'clock in the morning, the current and power reach their peak, indicating the highest level of power generation during that day.

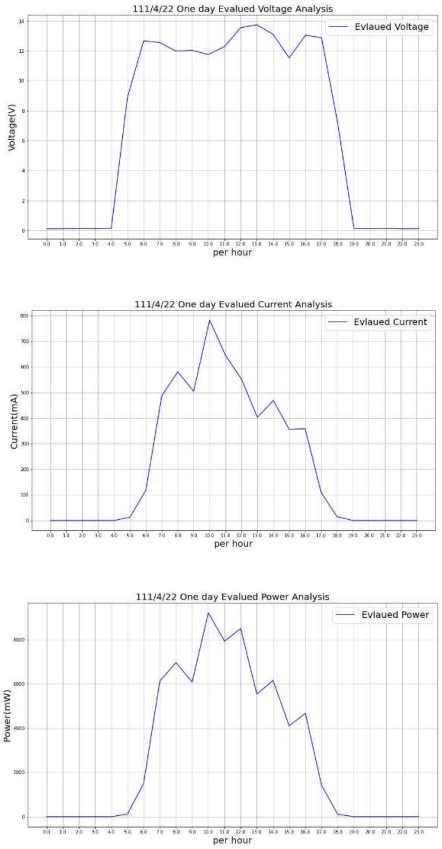


Fig. 4 Voltage, current, and power curves generated by the solar power system on 2022/04/22.

Typically, the voltage tends to stabilize at a certain value, primarily due to limitations imposed by the battery voltage. The current and power, on the other hand, are influenced by the wind-solar hybrid controller to prevent battery overcharging. Consequently, the protection system is activated to maintain optimal operation.

Due to the seasonal and weather-related variability, the power generation status of wind power is relatively unstable. Fig. 5 provides an overview of the wind speed recorded from 2022/04/22 to 2022/04/26. From the figure, it is evident that the wind speed remained relatively low during these five days. Consequently, the corresponding power generation was insufficient to meet the battery charging requirements. This highlights the dependence of wind power generation on favorable wind conditions, which can impact its reliability as a consistent power source.

C. LSTM Prediction

In order to provide comprehensive predictions, training, and comparison of monitored values, we have employed the LSTM prediction model. This model allows us to make accurate forecasts and assess the performance of the system. In

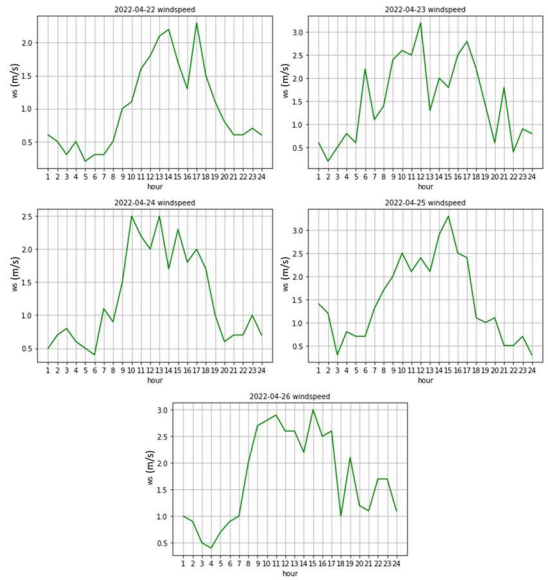


Fig. 5 Wind speed recorded from 2022/04/22 to 2022/04/26.

the generated figures, the blue line represents the actual monitored values, while the red line represents the predicted values.

Fig. 6 illustrates the monitoring data and corresponding forecast for a single day, whereas Fig. 7 displays the monitoring and forecast results for a period of six days. Upon observation, it is evident that the multi-day forecast results align more closely with the actual power generation values. This indicates the effectiveness of the LSTM prediction model in capturing the underlying patterns and trends within the data, leading to improved predictive accuracy and reliability.

D. Discussion

To summarize the above, we utilized the LSTM forecasting model algorithm by incorporating the monitored voltage, current, power, and weather patterns. By feeding the corresponding weather patterns into the LSTM forecasting model algorithm, we observed a certain degree of correlation between the predicted and actual values. Analyzing the resulting curves, we have identified the following findings:

(a) Integral data analysis: We conducted an analysis of the area under the two curves and the time axis. By averaging the differences between the predicted values at each time point, we obtained a value A. Similarly, by averaging the predicted values in the same manner, we obtained a value B. Taking the absolute difference between these two values, we obtained a value of 0.007337. A smaller value indicates a smaller data error and higher reliability.

(b) Analysis of correlation coefficient: The correlation coefficient is used to measure the degree of correlation between two variables, X and Y. Its value ranges from -1 to 1. This coefficient is commonly employed to assess the linear dependence between variables. In our analysis, we calculated the correlation coefficients for voltage, current, and power. The voltage correlation coefficient was found to be 0.7644, the current correlation coefficient was 0.8057, and the power correlation coefficient was 0.8341. These results indicate a high degree of correlation between the predicted and monitored values.

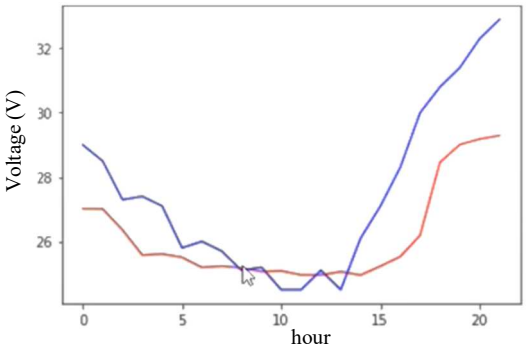


Fig. 6 Monitoring data and corresponding forecast for a single day.

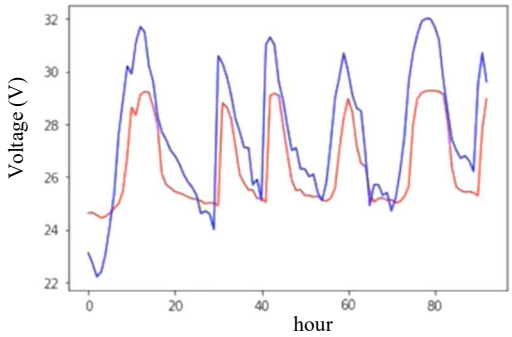


Fig. 7 Monitoring and forecast results for a period of six days.

(c) Analysis of mean square error (MSE): The MSE is a commonly used loss function for regression tasks. It quantifies the deviation between observed and true values. The MSE is calculated as the sum of squared differences between predicted and monitored values. In our analysis, the MSE value between the predicted and monitored values was calculated as 13.267, reflecting the overall deviation between the two sets of values.

Based on the data analysis, it is evident that the power generation of solar energy is significantly influenced by various weather patterns, including cloud cover, ultraviolet rays, temperature, and sunshine hours. By incorporating these impact factors into our linear prediction algorithm, we can establish a strong relationship between weather patterns and power generation. The aforementioned analysis also highlights the credibility of our prediction values.

Similarly, for wind energy, we consider two key weather parameters: wind direction and wind speed. The wind speed

must meet or exceed the minimum startup wind speed of the wind turbine (typically around 5 m/s) to ensure effective power generation. By considering these factors, we can assess the feasibility and potential output of wind energy generation.

Overall, the analysis underscores the importance of considering weather conditions and their impact on power generation when forecasting and optimizing renewable energy systems.

V. CONCLUSION

In this paper, we propose a Deep Learning-based Sustainable Energy Scheduling System that aims to optimize the utilization of green power by leveraging the relationship between solar and wind power generation and weather patterns. Our research focuses on establishing the correlation between green power generation and weather patterns through the collection and analysis of power generation data. To achieve this, we employ a data integration and visualization approach, followed by the utilization of the LSTM forecasting model. By comparing weather patterns, we obtain predicted electricity values and assess their correlation with the actual monitored electricity. Additionally, we utilize weather forecasts to estimate power generation in the coming days, further enhancing the forecasting capabilities of our model.

The experimental results demonstrate the clarity and feasibility of the proposed method. Moving forward, our future work will involve incorporating remote monitoring data and incorporating additional factors that influence green power generation. By developing a more comprehensive model, we aim to enhance the value and effectiveness of sustainable energy optimization. Overall, this paper presents a novel approach to sustainable energy scheduling, leveraging deep learning techniques and the analysis of weather patterns. The findings highlight the potential for maximizing the utilization of green power and pave the way for future research in this area.

REFERENCES

[1] https://www.iea.org/

[2] P. Roy, J. He, T. Zhao, and Y. V. Singh, "Recent advances of wind-solar hybrid renewable energy systems for power generation: A review," *IEEE Open Journal of the Industrial Electronics Society*, vol. 3, pp. 81-104, Jan. 2022.

[3] P. S. Kumar, R. P. S. Chandrasena, V. Ramu, G. N. Srinivas, and K. V. S. M. Babu, "Energy management system for small scale hybrid wind solar battery based microgrid," *IEEE Access*, vol. 8, pp. 8336-8345, Jan. 2020.

[4] R. Wang, D. Ma, M. J. Li, Q. Sun, H. Zhang, and P. Wang, "Accurate current sharing and voltage regulation in hybrid wind/solar systems: an adaptive dynamic programming approach," *IEEE Transactions on Consumer Electronics*, vol. 68, no. 3, pp. 261-272, Jun. 2022

[5] Z. Chen, J. M. Guerrero, and F. Blaabjerg. "A review of the state of the art of power electronics for wind turbines," *IEEE Transactions on Power Electronics*, vol. 24, no. 8, 1859-1875, Aug. 2009.

[6] F. Wang, D. Liu, L. Zeng. "Modeling and simulation of optimal wind turbine configurations in wind farms," in *Proceedings of 2009 World Non-Grid-Connected Wind Power and Energy Conference*, pp. 1-5, September 2009.

[7] J. Bukala, K. Damaziak, K. Kroszczynski, M. Krzeszowiec, and J. Malachowski. "Investigation of parameters influencing the efficiency of small wind turbines," *Journal of Wind Engineering and Industrial Aerodynamics*, Vol. 146, pp. 29-38, November 2015.

[8] K. Touafek, A. Khelifa, H. Haloui, M. T. Baissi, L. Boutina, S. Haddad, I. Tabet and A. Malek, "Improvement of Energy Efficiency of Solar Hybrid Water Collectors," in *Proceedings of 2017 International Renewable and Sustainable Energy Conference*, pp. 1-4, December 2017.

[9] M. Bedreddine, A. Yacine, and N. Abdelbaki, "Exploitation of solar energy for the elaboration of zinc oxide thin films according to a new method of solar spray rolysis," in *Proceedings of 2020 6th International Symposium on New and Renewable Energy*, pp. 1-5, October 2021.

[10] N. Reza, S. Hossain, N. Mondol, and A. Kabir, "Design and Implementation of an Automatic Single Axis Solar Tracking System to Enhance the Performance of a Solar Photovoltaic Panel" in *Proceedings of International Conference on Science & Contemporary Technologies*, pp. 1-6, August 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Computational Efficient Direct Position Determination Approach of Narrow-band Emitter

Yuan Zhao*, Hanmin Sheng † * ‡, Jinliang Shao † * ‡
* Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
† University of Electronic Science and Technology of China, Chengdu, China
‡ Laboratory of Electromagnetic Space Cognition and Intelligent Control, Beijing, China
E-mail: gabrizhao@cuhk.edu.cn

***Abstract*—Direct position determination (DPD) technique has been shown its appealing to emitter localization thanks to high estimation accuracy. Motivated by the practical applications, this paper proposed a computational efficient DPD approach using the multiple signal classification (MUSIC) formulation. Whereas, the proposed approach calculates the signal and the noise subspace using the multi-stage Wiener filter (MSWF) without employing the hardware-consuming eigen-decomposition. Hence, the proposed MSWF-DPD can be cheap equipment friendly. And we verify the performance loss is small through the simulations in the end of this paper.**

I. Introduction

Passive localization of emitter has found its significant in various applications, including wireless communication, unmanned aerial vehicle surveillance, Internet of things (IoT) networks [1], [2]. Now days, these techniques can be easily achieved by software defined radio (SDR) equipment. And this may accelerate the development of other possible IoT applications thanks to the cheap price.

The localization approaches are two-folder formulated. The classical approach relies on the well studied two-step notations. At the first step, some intermediate parameters, such as angle of arrival [3], time of arrival [4], Doppler frequency shift [5] and/or signal strength [6], are estimated from the emitter. While at the second step, these intermediate parameters are used to formulate the non-linear equations according to the geometric. The locations are then estimated through solving these non-linear equations. However, these methods are sub-optimal, since the formulation of non-linear equation to a transform linear equation requires some approximation and will introduce a bias to the estimation [4]. Meanwhile, the intermediate parameters are estimated independently without considering the constraint that all the observations relate to the same emitter location [7].

Fortunately, the notation of Direct Position Determination (DPD) approach was firstly proposed in [8], and enables one to achieve the location estimation goal in a single step. Different from the two-step approaches, DPD based approach can directly obtain the location estimation according to the cost function formulated by the radio frequency received signals (sometimes down-converted to the intermediate frequency). By this mean, the estimation error cased by intermediate measurements can be avoided. It is readily shown in [8] that the DPD approaches have a better estimation performance comparing classical two-step counterpart, especially in low signal to noise ratio (SNR) regime.

In fact, the DPD approaches can be directly determined by the maximum likelihood (ML) criteria, which is equivalent to relate the location estimation to the minimized negative log-likelihood function (LLF) While this naive DPD approach has been extensively studied, it fails to work for the multiple emitter regime. On the contrary, spectral based DPD approaches are derived using the eigen-structure of the second-order statistic. So that the eigenvalues with greater values corresponds to the signal subspace while the remaining corresponds to the noise subspace. We can also calculate the expanded steering vector considering the topology of the receivers. Hence, it allows us to estimate the location of the emitter the same manner as the multiple signal classification (MUSIC) used for AOA estimation [9], [10].

Unfortunately, the computational burden of DPD approaches, especially demanded by eigen-decomposition, prohibits us from using this high-precious method. Besides, the eigen-decomposition also consumes vast of hardware resources. Therefore, a low computational complexity approach with slight performance degradation will be more preferable.

In the paper, we develop a new localization approach inherent the improved MUSIC formulation [11]. While the proposed approach can be both computational simple and robust. To achieve this goal, we employ the well-studied multistage orthogonal projection employed in the multi-stage Wiener filter (MSWF) [12] to calculate signal and noise subspace, rather than employing the eigenvectors of the covariance matrix. Finally, a grid searching is employed to yield the localization of the emitter. The computer simulations are conducted to show the feasibility of proposed MSWF-DPD approach.

$Notations$: Throughout this paper, functions/scalars, vectors and matrices are written in lightfaced lowercase, boldfaced lowercase and boldfaced uppercase letters. $\mathrm{j} = \sqrt{-1}$ represents the unit complex. c is a constant representing the speed of light. The $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^\dagger$ stand for the conjugate, transpose and conjugate transpose operators. $\mathbb{E}\{a\}$ and $\hat{a}$ are the expected value and estimation of a. Given a vector $\boldsymbol{a}$, $|\boldsymbol{a}|_2$ returns its

This paper is sponsored by the National Natural Science Foundation of China (No. 62201363), and the funding from Shenzhen Institute of Artificial Intelligence and Robotics for Society.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

2-norm. Given a square matrix $\boldsymbol{A}$, $|\boldsymbol{A}|$ and $\mathrm{tr}(\boldsymbol{A})$ are the determinant and trace operator. The $\odot$ represents the Hadamard (element wise) product, while $\otimes$ calculates the Kronecker product. $\boldsymbol{I}_M$ represents the $M \times M$ identity matrix. $\boldsymbol{a}[i]$ stands for the ith entry of the vector $\boldsymbol{a}$, and $[\boldsymbol{A}]_{i,j}$ represents the (i,j)th entry of the matrix $\boldsymbol{A}$. Finally, $\delta_{l,k} = 1$ for $l = k$ and 0, otherwise.

II. Problem Formulation

Consider without loss of generality that L spatially distributed receivers, indexed by l, equipping with M_l omni-directional antenna. For derivation simplicity, we assume $M_l = M$ for all receivers. We then suppose there exist a emitter transmitting the narrow-band unknown signal $s(t)$. The goal of this paper is to determine the location of this emitter.

Denote the unknown source and the known lth receiver location by $\boldsymbol{p}_0$ and $\boldsymbol{p}_l$, respectively. It has been well investigated that according to the geometry, the time of arrival (TOA) and the direction of arrival (DOA) are respectively calculated by

$$\begin{aligned} \tau_l &= \|\boldsymbol{p}_l - \boldsymbol{p}_0\| / c \\ \phi_l &= \frac{\pi}{2} - \arctan\left(\frac{\boldsymbol{p}_l[2] - \boldsymbol{p}_0[2]}{\boldsymbol{p}_l[1] - \boldsymbol{p}_0[1]}\right), \end{aligned} \tag{1}$$

with index $[1]$ indicates the first element of the location and $[2]$ refers to the second element. Hence, the received snapshot of the lth receiver can be formulated by

$$\begin{aligned} \boldsymbol{y}_l(t) &= b_l \boldsymbol{a}_l(\phi_l)\, s(t-\tau_l) + \boldsymbol{\omega}_l(t) \\ &\approx b_l \boldsymbol{a}_l(\phi_l)\, s(t)\, e^{\mathrm{j}2\pi f_0 \tau_l} + \boldsymbol{\omega}_l(t), \end{aligned} \tag{2}$$

where b_l stands for the intensity of the observation, and $\boldsymbol{a}_l(\phi_l)$ is the steering vector corresponds to the array manifold allocated in arbitrary manner. Furthermore, we make the assumption that the sources are located in the far-field regime so that the array manifold is a function of the DOAs only. Therefore, one can denote the total observation from all the sensors by stacking the individual observation into a $L \times M$ vector, so that the total received vectors can be expressed as

$$\begin{aligned} \boldsymbol{y}(t) &= \left[\boldsymbol{y}_1^T(t), \boldsymbol{y}_2^T(t), ..., \boldsymbol{y}_L^T(t)\right]^T \\ &= \mathrm{diag}\{\boldsymbol{b}\} \otimes \boldsymbol{I}_M \boldsymbol{\alpha}(\boldsymbol{p})\, s(t) + \boldsymbol{\omega}, \end{aligned} \tag{3}$$

with noise vector and the extended steering vector being defined as

$$\begin{aligned} \boldsymbol{\omega} &= \left[\boldsymbol{\omega}_1^T(t), \boldsymbol{\omega}_2^T(t), ..., \boldsymbol{\omega}_L^T(t)\right]^T \\ \boldsymbol{\alpha}(\boldsymbol{p}) &= \left[\bar{\boldsymbol{a}}_1^T(\boldsymbol{p}), \bar{\boldsymbol{a}}_2^T(\boldsymbol{p}), ..., \bar{\boldsymbol{a}}_L^T(\boldsymbol{p})\right]^T, \end{aligned} \tag{4}$$

where $\bar{\boldsymbol{a}}_l(\boldsymbol{p}) = \boldsymbol{a}_l(\phi_l)\, e^{-\mathrm{j}2\pi f_0 \tau_l}$.

Meanwhile, we assume the additive noise satisfies the zero-mean circulant Gaussian distribution with variance σ_0^2, i.e. $\omega_l(t) \sim \mathcal{CN}\left(0, \sigma_0^2 \boldsymbol{I}_M\right)$. For derivation convenience, we further assume the additive noise are equal power[1]. The transmitted signal is supposed to have the normalized power $\sum_{t=1}^{N} |s(t)|^2 / N = 1$, while the amplitudes b_l's are jointly stationary zero-mean Gaussian distributed with variance σ_l^2. However, the sensors may spatially coherence so that b_l's are mutually correlated by factor $\rho_{l,k} = \sqrt{\mathbb{E}\{b_l b_k^*\}}$

Under these assumptions, the total observation vector is a complex Gaussian random process with zero mean and the (l,k)th block of the covariance matrix yields

$$\begin{aligned} [\boldsymbol{R}_{\boldsymbol{y}}]_{l,k} &= \mathbb{E}\left\{\boldsymbol{y}_l(t)\, \boldsymbol{y}_k^{\dagger}(t)\right\} \\ &= \rho_{l,k} \boldsymbol{a}_l(\phi_l)\, \boldsymbol{a}_k^{\dagger}(\phi_k)\, e^{\mathrm{j}2\pi f_0(\tau_l - \tau_k)} + \delta_{l,k}\sigma_0^2 \boldsymbol{I}_M, \end{aligned}$$

The likelihood function is then given by

$$f_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{p}_0) \propto |\boldsymbol{R}_{\boldsymbol{y}}|^{-1} \exp\left(-\boldsymbol{y}^H \boldsymbol{R}_{\boldsymbol{y}}^{-1} \boldsymbol{y}\right).$$

The principle of DPD is to estimate the underlying source location by minimizing the negative LLF. However, to the best of our knowledge, the non-linearity equation (1) embedded in LLF under statistic model admits no closed-form representation regarding the location estimation $\hat{\boldsymbol{p}}_0$. Thus, in the sequel, we resort to the grid-search approach and the optimized estimation is in accordance with the grid that minimize (or maximize) the cost function.

Disscussion 1: Under the deterministic model where b_l's are determined yet unknown, one can formulate the cost function by the first-order statistic. The DPD approach is then derivative by nonlinear least square (NLS) estimation criteria regardless the distribution [7], [13]. Whereas, under the statistic model where b_l's are random distributed with zero mean, the covariance matrix is sufficient for computing the cost function. Under this circumstance, one may also access the cost function in the same manner as NLS. However, subspace approaches such as MUSIC are capable to achieve a high resolution estimation. Hence, we formulate the cost function by MUSIC proposed in [9] but recast to a more computational efficient manner.

III. The Proposed DPD without Eigen-Decomposition

In this section, we will present a reduced-rank DPD approach based on MSWF. For comparison purpose, we firstly revisit a recently proposed DPD approach using MUSIC [9], which is under narrow band assumption. We will then present our proposed DPD approach using MSWF to avoid calculating time-consuming eigen-decomposition. In the end of this section, we will summarize our proposed approach.

A. MUSIC DPD

Regarding our received data model, it is non-trivial to decompose the population covariance matrix (PCM) of the received signal by

$$\boldsymbol{R}_{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\dagger}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_{ML}]$ spans the signal and noise subspace. To be concrete, in our *single* source regime, $\boldsymbol{u}_1$ corresponds to the signal vector, while $\boldsymbol{U}_N = [\boldsymbol{u}_2, ..., \boldsymbol{u}_{ML}]$ corresponds the noise subspace.

[1]This assumption is somehow unrealistic, but one can estimate the noise level and "whiten" the received signal. Hence, throughout this paper, we assume the noise level are equal for all receivers.

The principle concept of MUSIC-based approach is that the steering vector lies in the same direction with signal vector, therefore orthogonal with the noise subspace. Thus, the cost function is formulated by

$$\xi_1(\boldsymbol{p}) = \boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{U}_N}\hat{\boldsymbol{U}_N}^{\dagger}\boldsymbol{\alpha}(\boldsymbol{p}) \tag{5}$$

where $\hat{\boldsymbol{U}_N}$ is noise subspace of sample covariance matrix (SCM). Here, the nuisance parameter $\boldsymbol{b}$ is substituted by all-one vector since it has no influence on the cost function.

Further, by adopting the asymptotic distribution given by [9], [11], we arrive at

$$\epsilon_i \triangleq \boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{u}}_{i+1} \sim \mathcal{N}(0, \gamma_1(\boldsymbol{p})), \forall i = 1, ...ML-1, \tag{6}$$

where $\gamma_1(\boldsymbol{p}) = \boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\boldsymbol{W}\boldsymbol{\alpha}(\boldsymbol{p})$ with $\boldsymbol{W} \propto \boldsymbol{u}_1\boldsymbol{u}_1^{\dagger}$ by a constant scalar $\sigma_0^2\lambda_1/N\left(\sigma_0^2-\lambda_1\right)$. In fact, this scalar can be treated as a weight on the signal subspace for multiple sources. In our single source situation, however, we discard this scalar in $\boldsymbol{W}$ without influencing the global searching for the optimize value. And according to [11], ϵ_i's are asymptotically independent $\mathbb{E}\left\{\epsilon_i\epsilon_j^*\right\} = 0$. Therefore, an improved MUSIC-DPD based cost function is proposed by

$$\xi_2(\boldsymbol{p}) = \frac{\xi_1(\boldsymbol{p})}{\gamma_1(\boldsymbol{p})} = \frac{\boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{U}_N}\hat{\boldsymbol{U}_N}^{\dagger}\boldsymbol{\alpha}(\boldsymbol{p})}{\boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\boldsymbol{u}_1\boldsymbol{u}_1^{\dagger}\boldsymbol{\alpha}(\boldsymbol{p})}. \tag{7}$$

Finally, one can calculate the cost function for all possible locations through grid-search. The optimal grid with maximized cost function corresponds to the estimation of source. Whereas, either $\xi_1(\boldsymbol{p})$ or $\xi_2(\boldsymbol{p})$ demands the calculation of eigen-decomposition, thereby requires heavy computational efforts. To cure this problem, we propose a MSWF-based DPD approach to estimate the signal and the noise subspace iteratively, in the following part of this section.

B. Multi-Stage Winier Filter based MUSIC DPD

To begin with, we assume an extra fusion center with single omni-directional antenna. The received signal of the fusion center can be denoted by

$$y_0(t) \approx b_0 s(t)\,\mathrm{e}^{-\mathrm{j}2\pi f_0\tau_0} + \omega_0(t). \tag{8}$$

We define $d_0(t) = y_0(t)$ as the reference, the cross correlation in the fusion center is given by

$$\boldsymbol{r}_0(t) = \mathbb{E}\{\boldsymbol{y}(t)\,d_0^*(t)\} = \mathrm{diag}\{\boldsymbol{c}\} \otimes \boldsymbol{I}_M\boldsymbol{\beta}(\boldsymbol{p}). \tag{9}$$

where $\boldsymbol{c}[l] = b_l b_0^*$, $\boldsymbol{\beta}(\boldsymbol{p}) = \boldsymbol{\alpha}(\boldsymbol{p})\,\mathrm{e}^{\mathrm{j}2\pi f_0\tau_0}$ and the additive noise is eliminated due to the assumption that the noise of each receivers are independent.

[14] indicates this correlation vector can be regarded as matched filter to capture the source information $\boldsymbol{\alpha}(\boldsymbol{p})$. Therefore, we can define a normalized filter by

$$\boldsymbol{h}_1 = \frac{\boldsymbol{r_0}}{\|\boldsymbol{r}_0\|},$$

which partitions the observation by the desired component $d_1(t) = \boldsymbol{h}_1^{\dagger}\boldsymbol{y}(t)$ and orthogonal component $\boldsymbol{y}_1(t) = \boldsymbol{y}(t) - \boldsymbol{h}_1 d(t)$. Similarly, if we partition the signal for the ith stage, which yields

$$\begin{aligned} d_i(t) &= \boldsymbol{h}_i^{\dagger}\boldsymbol{y}_{i-1}(t), \\ \boldsymbol{y}_i(t) &= \boldsymbol{y}_{i-1}(t) - \boldsymbol{h}_i d_i(t). \end{aligned}$$

where the matched filter at ith stage is computed by

$$\boldsymbol{h}_i = \frac{\mathbb{E}\left\{\boldsymbol{y}_{i-1}(t)\,d_{i-1}^*(t)\right\}}{\left\|\mathbb{E}\left\{\boldsymbol{y}_{i-1}(t)\,d_{i-1}^*(t)\right\}\right\|} \tag{10}$$

We can repeat this recursions by ML stages, so that the desired signal forms a vector defined by

$$\boldsymbol{d}(t) \triangleq [d_1(t), d_2(t), ..., d_M(t)]^T = \boldsymbol{H}^{\dagger}\boldsymbol{y}(t),$$

where the unity of matrix $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_{ML}]$ has readily proved in [14]. Moreover, [12] has also shown that $\boldsymbol{H}_N = [\boldsymbol{h}_2, ..., \boldsymbol{h}_{ML}]$ spans the noise subspace as $\boldsymbol{U}_N$, for our single source regime.

Following this conclusion, we define the statistic

$$\zeta_i = \boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{h}}_{i+1}, \forall i = 1, 2, ..., ML-1,$$

where $\hat{\boldsymbol{h}}_i$ is the data-driven match filter at ith stage and $\boldsymbol{\alpha}(\boldsymbol{p})$ is given by (4). The following proposition provides the statistic property of ζ_i.

Proposition 1: The statistic $\boldsymbol{\zeta} = [\zeta_1, \zeta_2, ..., \zeta_{ML-1}]^T$ are asymptotically joint Gaussian distributed with zero mean, whose i, jth element of covariance matrix is given by

$$\begin{aligned} [\boldsymbol{R}_{\boldsymbol{\zeta}}]_{i,j} &= \mathbb{E}\left\{\zeta_i\zeta_j^*\right\} \\ &= \mathbb{E}\left\{\boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{h}}_{i+1}\hat{\boldsymbol{h}}_{j+1}^{\dagger}\boldsymbol{\alpha}(\boldsymbol{p})\right\}, \end{aligned}$$

which is equal to 0 for $i \neq j$ and ML for diagonal elements.

Proof: A detailed proof for the asymptotic Gaussian property in $Proposition1$ will be presented in our future extended paper. Meanwhile, one can easily prove the diagonal property of $\boldsymbol{R}_{\boldsymbol{\zeta}}$ following the orthogonal of $\hat{\boldsymbol{h}}_i$ and $\hat{\boldsymbol{h}}_j$ for $i \neq j$. And the diagonal element can be calculated by noticing $\left\|\hat{\boldsymbol{h}}_i\right\|_2^2 = 1$ and $\|\boldsymbol{\alpha}(\boldsymbol{p})\|_2^2 = ML$. □

Thanks to Proposition 1, we can reformulate the cost function the same manner as (7) by

$$\xi_3(\boldsymbol{p}) = \frac{\boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{h}}_1\hat{\boldsymbol{h}}_1^{\dagger}\boldsymbol{\alpha}(\boldsymbol{p})}{\boldsymbol{\alpha}^{\dagger}(\boldsymbol{p})\,\hat{\boldsymbol{H}}_N\hat{\boldsymbol{H}}_N^{\dagger}\boldsymbol{\alpha}(\boldsymbol{p})}. \tag{11}$$

This result is straightforward since $\hat{\boldsymbol{h}}_1$ spans the signal subspace and $\hat{\boldsymbol{H}}_N$ spans the noise subspace. Therefore, we can estimate the location of the emitter by maximize the cost function as one did in (7). Whereas, one can estimate the subspace by iteratively calculate the MSWF stage by stage, instead.

C. A Overview of the Proposed MSWF-DPD Approach

In this paper, we resort to the grid search method by maximize the cost function in (11). Concretely speaking, the section of interest (SOI) will be divided into several square grid with a user-defined resolution. The centers of each girds $\boldsymbol{p}$ are plugged into the cost function, so that one can estimate the location by searching for the maximum value of the cost function, i.e.

$$\hat{\boldsymbol{p}}_0 = \arg\max \xi_3 (\boldsymbol{p}),$$

which corresponds to the optimal location estimation. However, if one search the emitter by a very fine grid at the first stage (say 1×1 grid in 1000×1000 area), the computational load caused by searching (1×10^6) will be very heavy. Hence, one can search the existence of the emitter by a rough resolution and then refine the resolution if the emitter appears in the specific grid.

In summarize, the proposed approach can achieve a refined localization by the algorithm 1.

Algorithm 1 MSWF-MUSIC-DPD

Input:

The SOI, the location of the receivers $\{\boldsymbol{p}_l\}$ and the number of receivers L.

The total received data $\boldsymbol{y}(t)$, number of antenna for each receivers M_l and the number of snapshots N

Output:

The location estimation of emitter $\hat{\boldsymbol{p}}_0$

Initialize: grid size Δ, searching center $\boldsymbol{p}_0^{(1)}$ for the SOI, iterative index $i = 1$, $\boldsymbol{p}_0^{(0)} \leftarrow \boldsymbol{p}_1$

while $\left|\boldsymbol{p}_0^{(i)} - \boldsymbol{p}_0^{(i-1)}\right|_2 > \epsilon_0$ **do**

  $j \leftarrow 0$, $\boldsymbol{y}_0(t) = \boldsymbol{y}(t)$

  Calculate the cross correlation $\boldsymbol{r}_0$ with respect to the fusion center given by (9)

  for $j = 1 : \sum_{l=1}^{L} M_l$ **do**

    $\boldsymbol{h}_j = \boldsymbol{r}_{j-1} / |\boldsymbol{r}_{j-1}|_2$

    $d_j(t) = \boldsymbol{h}_j^{\dagger} \boldsymbol{y}_{j-1}(t)$

    $\boldsymbol{y}_j(t) = \boldsymbol{y}_{j-1}(t) - \boldsymbol{h}_j d_j(t)$

    $\boldsymbol{r}_j = \sum_t \boldsymbol{y}_j(t) d_j^*(t)$

  end for

  $\boldsymbol{p}_0^{(i)} = \arg\max \xi_3 (\boldsymbol{p})$, where ξ_3 is defined in (11)

  $i \leftarrow i + 1$, $\Delta \leftarrow \Delta/2$

end while

IV. Simulation Results

In this section, we report some representative numerical results to investigate the localization performance. The MUSIC-DPD formulated in (7) is adopted as the competitive.

Additionally, each result represents an average of 1×10^3 independent trials. At each run, the gird size is initiated at $10m$. The SOI is a $100m \times 100m$ square centered at the original point. The receivers and the emitter are randomly distributed in this area, as we show in Fig.1 for one run. We assume the emitter transmits the simple pulse modulated by cosine function for illustrative purpose. The duration of the pulse is $50\mu s$. The carrier frequency is assumed to be 2.4GHz. At each run, we assume the receives signals are down convert to the intermediate frequency at 1MHz. The sample rate is twice of the intermediate frequency. So that each receiver can collect 100 snapshots. We evaluate the estimation performance by root mean-squared-error (rMSE) calculated by

$$\text{rMSE} = |\boldsymbol{p}_0 - \hat{\boldsymbol{p}}_0|_2.$$

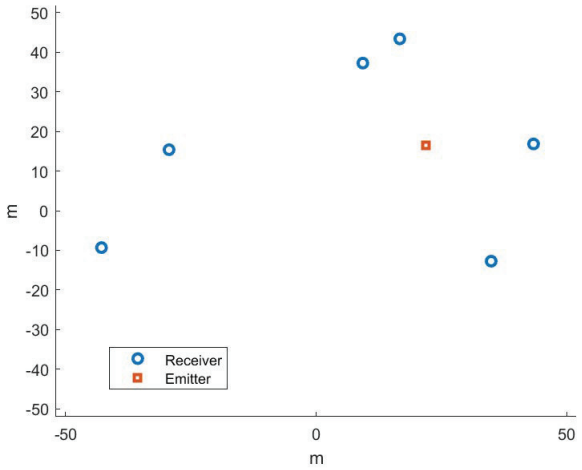


Fig. 1: The randomly generated locations of emitter and receivers for $L = 6$.

We firstly investigate the estimation performance as the function of the SNR. The simulation results is shown in Fig.2, where we assume 6 receivers are randomly distributed in the SOI and each receiver equips 2 antenna except for the fusion center (single channel). It can be observed that the rMSE of both MUSIC-DPD and proposed MSWF-DPD decreases as the SNR increases, and the performance of MSWF-DPD is coverage to the MUSIC-DPD. This is reasonable since the signal component dominates the matched filter in the first stage, i.e. $\boldsymbol{h}_1$, when the SNR is high. Hence, one can have a better estimation of signal subspace using normalized cross correlation (9). In this experiment, we assume the receivers are equipped with 2 antenna because we will extend our result to Universal Software Radio Peripheral (USRP) B210[2] to collect field data in future works. And USRP B210 is designed with 2 receiving channels using AD9361[3].

It can be observe that both approaches have a estimation rMSE for around $5m$ when SNR is equal to 6dB. This result is somehow unsatisfactory, but we can refine this estimation performance by increase either the number of antenna or the number of receiver. Fig.3 and Fig.4 show the rMSE versus number of antenna and receivers, respectively. All the curves decrease as the dimension of total observation, $\boldsymbol{y}(t)$, increases.

[2]visit https://www.ettus.com/all-products/ub210-kit/ for detailed information

[3]visit https://www.analog.com/en/products/ad9361.html for detailed information

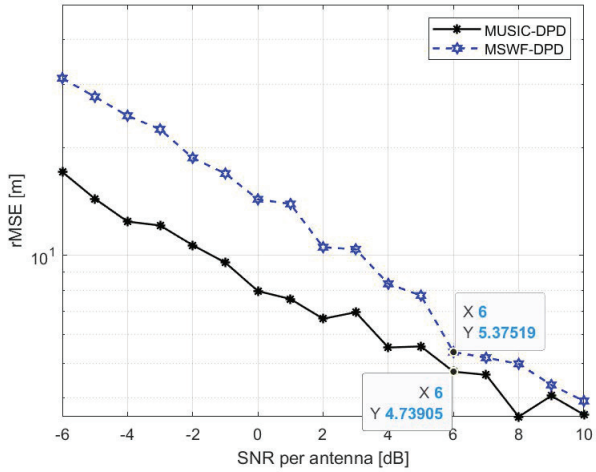


Fig. 2: Localization performance versus SNR with $M = 2$, $L = 6$.

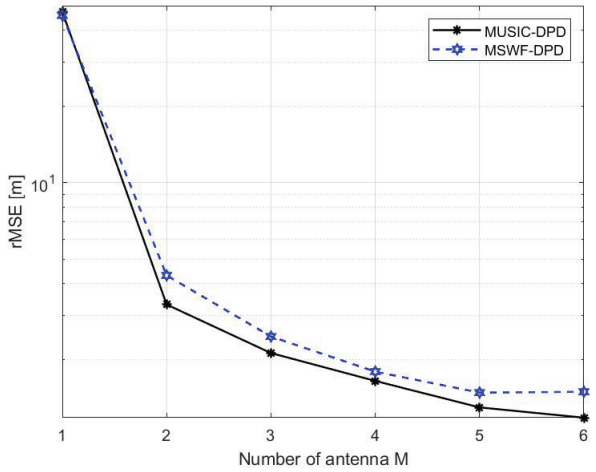


Fig. 3: Localization performance versus M with SNR= 6dB, $L = 6$.

This is because one can have a better estimation of noise subspace as the dimension gets greater. These results are also in accordance with our conclusions in Fig.2.

Although the performance of MSWF-DPD is overwhelmed by MUSIC-DPD, but the computational complexity is significantly decreased [14]. To be specific, the MSWF-DPD spends the most computational effort in computing the covariance matrix. While the MUSIC-DPD requires some extra computational effort in eigen-decomposition, which is $\mathcal{O}\left(M^3L^3\right)$. Hence, the proposed MSWF-DPD can achieve a better estimation efficiency for the centralized location technique.

V. Conclusion and Future Works

In this paper, we proposed a computational efficient method to estimate the localization of a narrow-band emitter. We inherent the grid-search scheme as the MUSIC-DPD.

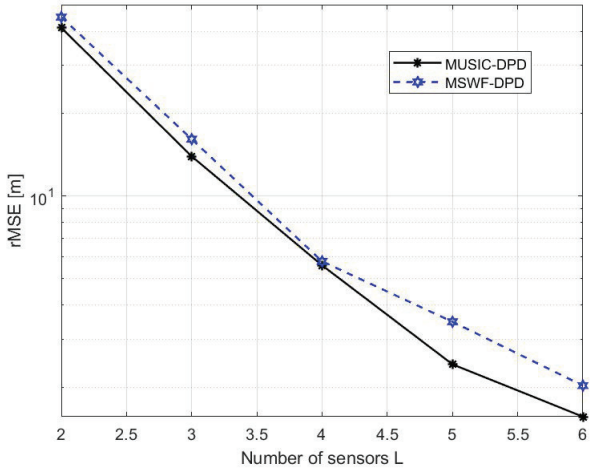


Fig. 4: Localization performance versus L with SNR= 6dB, $M = 6$.

But we achieve the partition without calculating the eigen-decomposition. In the simulation, we present some typical results due to the length constraint. And our method can be easily extended to the multiple emitter regime. For better illustrate our proposed approach, the experiment results using USRP will be presented in our future extended work. Besides, we can further reduce the computational complexity through the selection on the order of MSWF, which require a sophistic trade-off. And the extension to the situation when each emitters is not well calibrated would also left in our future works.

References

[1] Y. Zhao, R. Lv, S. He, L. Wu, B. Wang, and L. Zhao, "Cooperative positioning of single target based on TDOA using UAV cluster," pp. 111–115, 2022.

[2] F. Khelifi, A. Bradai, A. Benslimane, P. Rawat, and M. Atri, "A survey of localization systems in internet of things," *Mobile Networks and Applications*, vol. 24, no. 3, pp. 761–785.

[3] Y. Wang and K. C. Ho, "An asymptotically efficient estimator in closed-form for 3-d aoa localization using a sensor network," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6524–6535, 2015.

[4] K. C. Ho, "Bias reduction for an explicit solution of source localization using TDOA," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2101–2114, 2012.

[5] G. Wang, S. Cai, Y. Li, and N. Ansari, "A bias-reduced nonlinear WLS method for TDOA/FDOA-based source localization," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8603–8615, 2016.

[6] J. He, Y. J. Chun, and H. C. So, "A unified analytical framework for RSS-based localization systems," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6506–6519, 2022.

[7] T. Tirer and A. J. Weiss, "High resolution direct position determination of radio frequency sources," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 192–196, 2016.

[8] A. Weiss, "Direct position determination of narrowband radio frequency transmitters," *IEEE Signal Processing Letters*, vol. 11, no. 5, pp. 513–516, 2004.

[9] L. Wang, Y. Yang, and X. Liu, "A direct position determination approach for underwater acoustic sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 033–13 044, 2020.

[10] M. Oispuu and U. Nickel, "Direct detection and position determination of multiple sources with intermittent emission," *Signal Processing*, vol. 90, no. 12, pp. 3056–3064, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168410002124

[11] F. Wen, Q. Wan, R. Fan, and H. Wei, "Improved MUSIC algorithm for multiple noncoherent subarrays," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 527–530, 2014.
[12] L. L. Scharf, E. K. P. Chong, M. D. Zoltowski, J. S. Goldstein, and I. S. Reed, "Subspace expansion and the equivalence of conjugate direction and multistage Wiener filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5013–5019, 2008.
[13] A. J. Weiss, "Direct geolocation of wideband emitters based on delay and doppler," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2513–2521, 2011.
[14] H. Lei, L. Teng, and W. Shunjun, "Source enumeration for high-resolution array processing using improved Gerschgorin radii without eigendecomposition," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 5916–5925, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Modeling and Analysis of the Epidemic-Behavior Co-evolution Dynamics with User Irrationality

Wenxiang Dong and H. Vicky Zhao
Department of Automation, BNRist, Tsinghua University, Beijing, P. R. China 100084
E-mail: dongwx19@mails.tsinghua.edu.cn; vzhao@tsinghua.edu.cn

***Abstract*—During a public health crisis like COVID-19, individuals' adoption of protective behaviors, such as self-isolation and wearing masks, can significantly impact the spread of the disease. In the meanwhile, the spread of the disease can also influence individuals' behavioral choices. Moreover, when facing uncertain losses, individuals' decisions tend to be irrational. Therefore, it is critical to study individuals' irrational behavior choices in the context of a pandemic. In this paper, we propose an epidemic-behavior co-evolution model that captures the dynamic interplay between individual decision-making and disease spread. To account for irrational decision-making, we incorporate the Prospect Theory in our individual behavior modeling. We conduct a theoretical analysis of the model, examining the steady states that emerge from the co-evolutionary process. We use simulations to validate our theoretical findings and gain further insights. This investigation aims to enhance our understanding of the complex dynamics between individual behavior and disease spread during a pandemic.**

***Index Terms*—Irrationality, Epidemic, Individual behavior**

I. Introduction

The COVID-19 pandemic has triggered a profound public health crisis and inflicted substantial economic damage. Throughout such public health crisis, individuals' behavior and the pandemic influence each other in a complex way. For instance, individuals' behaviors, such as wearing masks, social distancing, and self-quarantine, can help limit the spread of the disease, while people tend to adopt protective behaviors more readily when the severity of the pandemic becomes apparent. Furthermore, individuals often display irrational behavior during the pandemic, including heightened panic at the early stage of an outbreak or a tendency to underestimate the severity of the disease as it spreads further. These irrational responses can significantly impact people's decision-making processes and ultimately influence the pandemic. Therefore, it is crucial to model individuals' irrational behavior and to analyze its impact on the pandemic.

A. Literature Review

There have been extensive works on user behavior modeling during an epidemic, which we will briefly introduce in the following.

1) Behavior Modeling in Disease Spread: Several previous studies focus on modeling individual behavior during disease outbreaks. The authors of [1]–[3] believe that as the number of infected individuals in a given environment increases, people are more inclined to adopt precautionary measures. Zhang et al. propose that individuals are more inclined to engage in protective behaviors when they have a higher number of infected neighbors [1]. Wu et al. posit that an individual's adoption of protective behaviors is influenced not only by the proportion of infected neighbors but also by the regional and global infection rates [2]. The works in [4]–[8] also propose that information dissemination plays a crucial role in shaping people's protective behavior during disease outbreaks, and individuals who receive relevant information are more likely to adopt precautionary measures in response to the pandemic. Funk et al. believe that infected individuals play a significant role in disseminating information about the disease across networks. This information dissemination can effectively encourage others to adopt protective measures in response to the ongoing spread of the disease [4]. Granell et al. develop a dual network model that incorporates both disease spread and information spread. In their model, one network represents the transmission of the disease, while the other network represents the dissemination of information. Their findings suggest that individuals who receive informational messages are more vigilant and, as a result, exhibit a lower infection rate [5]. However, these studies often overlook a common yet crucial factor: the influence of individual irrationality on decision-making during a pandemic. Understanding and accounting for the impact of irrational behavior are essential as they can greatly shape individuals' responses and actions during an ongoing outbreak.

2) Irrational Behavior Modeling: Individuals often make irrational decisions when faced with risks, such as the risk of being infected during a disease outbreak. The Prospect Theory [9], [10], an influential theory put forth by Nobel laureates Kahneman and Tversky, sheds light on this phenomenon. It says that individuals tend to overestimate the likelihood of low-probability events and underestimate the likelihood of high-probability events. Moreover, individuals demonstrate a heightened sensitivity to potential losses compared to gains [11]. To effectively guide people's behavior during an epidemic and control the spread of the disease, it is of great importance to model such irrational behavior and analyze its impact on the pandemic. The works in [12]–[14] examine the influence of people's irrationality on their vaccination decisions. Vaccination presents a fundamental behavioral choice problem where individuals must make a single binary decision on whether or not to receive a vaccine during an epidemic, wherein opting for vaccination eliminates the risk of infection. However, during an epidemic, individuals also have the option to engage in

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

various protective behaviors, such as wearing masks, practicing hand hygiene, and self-isolation at home. These behaviors are discretionary and may change from time to time depending on the severity of the pandemic. Note that during an epidemic, individuals only decide once on whether to receive the vaccination. But for protective behaviors such as wearing masks, practicing hand hygiene, and self-isolation at home, they have to constantly decide whether to adopt them until the pandemic ends. The dynamic interplay between disease spread and individual behavioral choices occurs continuously throughout the outbreak, highlighting the complex relationship between the two. To our knowledge, no prior work modeled and analyzed the complex interplay between disease spread and behavioral choice with consideration of user irrationality.

B. Our Contribution

The main contributions of this paper are:

- We develop an epidemic-behavior co-evolution model that provides a framework for analyzing the co-evolution of the pandemic and individuals' behavior. This model allows us to study the dynamic interplay between disease spread dynamics and human behavior.
- In our model, we incorporate the concept of individuals' irrationality and specifically model irrational behavior using the Prospect Theory. By considering irrational decision-making processes, we enhance the realism of the model and capture the complex human decision-making process during a pandemic.
- We conduct theoretical analyses of the co-evolution of the pandemic and individuals' behavior. By exploring the dynamics of the model, we identify and characterize the steady states, which provide insights into the long-term behavior patterns and their interplay with the pandemic.

II. The epidemic-behavior co-evolution model

In this section, we model the co-evolution of individual behavioral choices and disease spread during the pandemic. In our model, extending the model in [15], we assume that individuals can choose between 2 possible behaviors: risky behavior a_1 (like going out) and conservative behavior a_2 (like self-isolation). Individuals would choose the behavior according to the severity of the epidemic, and these behavior choices will, in turn, affect the epidemic. We use an undirected graph to model the interpersonal connections among individuals. This graph captures the spread of the disease through the network, and in addition, neighbors in this network influence each other's behavior. To simplify the analysis, in this work, we assume a regular network consisting of N nodes, with degree $\bar{k}$.

Our model contains two parts, the disease spread model and the behavior change model. Given the current actions of all users, the disease spread model quantifies how the pandemic will spread in the network; while given the current number of infected people and their neighbors' current actions, the behavior change model describes how users choose their protective behaviors. We will introduce them one by one in the following.

A. The Disease Spread Model

We adopt the classical Susceptible-Infected-Susceptible (SIS) model to characterize the spread of the disease. In this model, each individual can be in one of two health states: susceptible or infected. In each time slot, a susceptible individual has the potential to become infected by an infected individual, with a given infection rate. Conversely, infected individuals have a chance to recover at a certain recovery rate. To mitigate the risk of infection, susceptible individuals can choose to adopt various protective measures, such as home quarantine, wearing masks, or taking no action. In this study, our focus is primarily on the protective behavior of susceptible individuals. We make the assumption that individuals who are already infected will not engage in such protective behaviors. This assumption is based on the understanding that infected individuals may not have the same motivation or need to adopt protective measures since they are already infected.

Assume that the susceptible individuals have two possible behavioral options a_1 and a_2. Furthermore, we assume that all individuals who choose a_1 have the same infection rate, denoted as β_1, and similarly, all individuals who choose a_2 have the same infection rate β_2. The recovery rate is assumed to be the same for all infected people and is denoted as γ. At time slot t, let $s(t)$ and $i(t)$ represent the fractions of susceptible and infected individuals, respectively. Additionally, let $x_1(t)$ be the fraction of susceptible individuals adopting behavior a_1, and $x_2(t)$ represent the fraction adopting behavior a_2. Consequently, the dynamics of the disease spread can be described by the following differential equations:

$$\begin{aligned} \frac{ds(t)}{dt} &= \gamma i(t) - s(t)i(t)\overline{\beta}\bar{k}, \\ \text{and} \quad \frac{di(t)}{dt} &= s(t)i(t)\overline{\beta}\bar{k} - \gamma i(t), \end{aligned} \tag{1}$$

where $\overline{\beta} = \beta_1 x_1(t) + \beta_2 x_2(t)$. Note that $s(t) + i(t) = 1$, (1) can be rewritten as

$$\frac{di(t)}{dt} = i(t)(1 - i(t))\overline{\beta}\bar{k} - \gamma i(t). \tag{2}$$

B. The Behavior Change Model

The behavior change model quantifies the change in $\{x_1(t), x_2(t)\}$, the proportion of susceptible individuals taking actions $\{a_1, a_2\}$ at time t. The change in $\{x_1(t), x_2(t)\}$ may be due to two possible reasons. First, due to the change in the proportion of the infected individuals, or influenced by their neighbors, susceptible individuals select different actions at time t. We use $E_i(t)$ to denote the change in $x_i(t)$ due to such users' active behavior change, where $i = 1, 2$. Additionally, users' health states may change at time t, causing the change in $x_i(t)$. For example, a susceptible user taking action a_i at time $t-1$ becomes infected at time t, or vice versa. We use $B_i(t)$ to denote the change in $x_i(t)$ due to the change in users'

health states. Thus, we have

$$\frac{dx_i}{dt} = E_i(t) + B_i(t),\ i = 1, 2. \tag{3}$$

In the following, we will discuss $E_i(t)$ and $B_i(t)$ one by one.

1) Analysis of $E_i(t)$: To model users' active behavior change due to their neighbors' influence and the percentage of individuals currently being infected, we use evolutionary game theory to model individuals' decision-making processes. Evolutionary game theory offers a valuable tool for understanding how individuals make choices in response to environmental influences. The fundamental components of the evolutionary game theory include strategy, payoff, and strategy update rules, each of which will be introduced in detail in the following.

Strategy: Each behavior a_i is a strategy. In each time slot, m percent of susceptible individuals are randomly selected as focal individuals. They observe their neighbors' behaviors and choose one possible strategy. Other susceptible individuals keep their actions unchanged.

Payoff: Each individual would get a payoff based on their chosen action and their interactions with their neighbors. To model users' irrationality under conditions of uncertainty, we define the payoffs of different behaviors using the Prospect Theory (PT). The Prospect Theory seeks to explain how individuals make decisions under conditions of uncertainty. PT suggests that people do not always make rational choices, and their decisions can be influenced by various cognitive biases and subjective evaluations. According to PT, if an individual selects a behavior a_i that leads to L potential outcomes $o_1, o_2, ..., o_L$ with corresponding probabilities $p_1, p_2, ..., p_L$, respectively, then the payoff U_i associated with choosing a_i can be defined as follows:

$$U_i = \sum_{i=1...L} u(o_i)\boldsymbol{\omega}(p_i), \tag{4}$$

where $u(x)$ is the value function and $\boldsymbol{\omega}(p)$ is the weighting function, which we will introduce one by one.

For the value function $u(x)$, according to the Prospect Theory, the perceived payoff differs from the actual payoff due to the presence of cognitive biases and subjective evaluations that individuals make when faced with uncertain outcomes. These biases and evaluations influence how individuals perceive and interpret information, leading to deviations from rational decision-making. The value function $u(x)$ models the relationship between the actual payoff x and the perceived payoff $u(x)$. A widely used value function is the power function

$$u(x) = \begin{cases} x^{\sigma}, & if\ x \geq 0 \\ -\lambda(-x)^{\sigma}, & if\ x < 0 \end{cases} \tag{5}$$

where λ reflects the individual's different sensitivity to gain and loss. The sensitivity coefficient $\sigma \in (0, 1]$ determines the shape and curvature of the value function [16].

For the weighting function $\boldsymbol{\omega}(p)$, when facing uncertainty, individuals' perceived probability is different from real probability. Individuals have a tendency to overestimate the probability of small risks, while underestimating the probability of larger risks [17]. According to the work in [18], the weighting function describing the relationship between the perceived probability and the real probability is:

$$\boldsymbol{\omega}(p) = e^{(-(-lnp)^{\alpha})},\quad p \in [0, 1],\ \alpha \in (0, 1], \tag{6}$$

where p is the real probability of a certain outcome, $\boldsymbol{\omega}(p)$ is the perceived probability, and α is the rationality coefficient. A smaller α indicates a higher degree of individuals' irrationality.

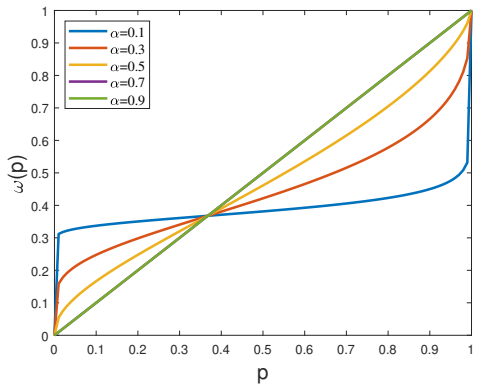


Fig. 1. The weighting function of different rational coefficients α.

Fig.1 plots the weighting function $\boldsymbol{\omega}(p)$ with different α. When the actual probability p is very small, the perceived probability $\boldsymbol{\omega}(p)$ is much larger than the actual probability p, showing the phenomenon that individuals tend to overestimate events/risks with small probabilities. On the contrary, when the actual probability is large, the perceived probability $\boldsymbol{\omega}(p)$ is smaller than the actual probability p, modeling the phenomenon that individuals tend to underestimate events/risks with large probabilities. Also, from Fig.1, a smaller α (that is, a higher degree of individuals' irrationality) intensifies individuals' tendency to overestimate the probability of small risks while simultaneously underestimating the probability of larger risks.

In our model, each susceptible individual who adopts behavior a_i receives a payoff of c_i with probability 1, which is determined by the behavior itself. Also note that an individual taking behavior a_i has an approximate probability of $\bar{k}\beta_i i(t)$ to be infected at time t [19], which incurs an additional loss of c_n (we consider those epidemics with $\beta_i^t \ll 1$, such as SARS, MERS and common influenza[20], and we assume that $\bar{k}\beta_i < 1$). To simplify the analysis, we assume that all individuals have the same form of the utility function $u(x)$ and the weighting function $\boldsymbol{\omega}(p)$. Thus, from (4), the perceived payoff when choosing behavior a_i is

$$U_i = u(c_n) \cdot \boldsymbol{\omega}[\bar{k}\beta_i i(t)] + u(c_i),\ i = 1, 2. \tag{7}$$

Strategy Update Rule: At each time step, m percent of the susceptible individuals are randomly selected as focal individuals and update their actions, while the rest individuals' actions

remain unchanged. These focal individuals are more likely to imitate the behaviors of individuals who have higher payoffs. Following the work in [21], a focal individual with strategy a_i compares its payoff U_i to the payoff U_j of another randomly chosen individual with strategy a_j. The focal individual then updates its strategy to a_j with probability:

$$p(a_i \to a_j) = \frac{1}{2} + \frac{\omega}{2}\frac{1}{U_{max}}(U_j - U_i), \tag{8}$$

where $\omega \in (0, 1]$ measures the strength of selection, and U_{max} is the normalization term.

For the two behavior a_1 and a_2, the probability that $x_1(t)$ increases by $\frac{1}{N_0(t)}$ is:

$$p\left(\Delta x_1 = \frac{1}{N_0(t)}\right) = x_1(t)x_2(t)p(a_2 \to a_1), \tag{9}$$

where $N_0(t)$ is the number of susceptible individuals at t and $N_0(t) = Ns(t)$. Similarly, the probability that $x_1(t)$ decreases by $\frac{1}{N_0(t)}$ is:

$$p\left(\Delta x_1 = -\frac{1}{N_0(t)}\right) = x_1(t)x_2(t)p(a_1 \to a_2). \tag{10}$$

Combining (8), (9) and (10), we have:

$$\begin{aligned} E_1(t) =& mN_0(t)\left\{p\left(\Delta x_1 = \frac{1}{N_0(t)}\right) \times \frac{1}{N_0(t)}\right. \\ & \left. -p\left(\Delta x_1 = -\frac{1}{N_0(t)}\right) \times \frac{1}{N_0(t)}\right\} \\ =& \frac{m\omega}{U_{max}} x_1(t)x_2(t)\left\{u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 i(t)]\right. \\ & \left. +u(c_1) - u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_2 i(t)] - u(c_2)\right\}, \end{aligned} \tag{11}$$

where m is the fraction of individuals who are chosen as the focal individuals at each time. We can get $E_2(t)$ in the same way, which is omitted here.

2) Analysis of $B_i(t)$: Even if individuals do not change their behavior, the proportion of different behaviors among susceptible individuals will change over time due to the change of their health states. Let $s_1(t)$ be the fraction of individuals who are both susceptible and adopt behavior a_1 at time t among the entire population, that is, $s_1(t) = s(t)x_1(t)$ where $s(t)$ is the fraction of susceptible individuals and $x_1(t)$ is the fraction of individuals adopting a_1 among all the susceptible individuals. As $s(t) = s_1(t) + s_2(t)$, we have:

$$\begin{aligned} B_1(t) =& \frac{d}{dt}\left(\frac{s_1(t)}{s(t)}\right) = \frac{d}{dt}\left(\frac{s_1(t)}{s_1(t) + s_2(t)}\right) \\ =& \frac{s_1'(t)s_2(t) - s_2'(t)s_1(t)}{s^2(t)}. \end{aligned} \tag{13}$$

Note that $s_1'(t)$, the first order derivative of $s_1(t)$, contains two parts. The first part is the change caused by the infection of susceptible users, while the second part is the change caused by the recovery of infected individuals. For the first part, we have $s_{1a}'(t) = -s_1(t)\beta_1\bar{k}i(t)$, which represents the decrease in the fraction of individuals adopting behavior a_1 due to infection. For the second part, we assume that the recovered individuals would choose their behaviors based on the current ratio of different behaviors in the environment, as defined in [22]. Therefore, we have $s_{1b}'(t) = \gamma x_1(t)i(t)$, which represents the increase in the fraction of individuals adopting behavior a_1 due to the recovery of infected individuals. Combining these two parts, we obtain the overall change in the fraction of individuals adopting behavior a_1 as $s_1'(t) = s_{1a}'(t) + s_{1b}'(t)$. Then we have:

$$\begin{aligned} & s_1(t) = s(t)x_1(t), \\ & s_1'(t) = -s(t)x_1(t)\beta_1\bar{k}i(t) + \gamma x_1(t)i(t). \end{aligned} \tag{14}$$

We can get $s_2(t)$ and $s_2'(t)$ in the same way. Then based on (13) and (14), we have

$$B_1(t) = x_1(t)x_2(t)\bar{k}i(t)(\beta_2 - \beta_1). \tag{15}$$

We can get $B_2(t)$ in the same way, which is omitted here.

Combining (3), (15) and (11), the complete differential equation describing the dynamics of individual behavior change is:

$$\begin{aligned} \frac{dx_1}{dt} =& x_1(t)x_2(t)\bar{k}i(t)(\beta_2 - \beta_1) + \frac{m\omega}{U_{max}}x_1(t)x_2(t) \\ & \cdot\left\{u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 i(t)] + u(c_1) - u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_2 i(t)] - u(c_2)\right\}. \end{aligned} \tag{16}$$

Note that $x_1(t) + x_2(t) = 1$, we have $\frac{dx_2}{dt} = -\frac{dx_1}{dt}$.

Combining (2) and (16), we get the epidemic-behavior co-evolution model in (12). The first differential equation represents the dynamics of individuals' health states. The second equation models the changes in the proportions of individuals adopting the risky behavior.

C. Steady State

At the steady states of the above epidemic-behavior co-evolution model, both $i(t)$ and $\{x_1(t), x_2(t)\}$ evolve to a stable state where there are no further changes in the proportions of infected individuals and the choices of different behaviors. As $x_2 = 1 - x_1$, we omit the explicit mention of x_2 in the representation of the steady state, and use (i^*, x_1^*) to represent the steady state. Similar to the work in [22] and the Lyapunov's first method [23], we define the *steady-state* of the epidemic-behavior co-evolution model as follows.

Definition 1: The steady state (i^*, x_1^*) satisfies:

$$\left.\frac{di}{dt}\right|_{i=i^*} = 0, \ \left.\frac{dx}{dt}\right|_{x_1=x_1^*} = 0, \ Re(\lambda_1) < 0, \ Re(\lambda_2) < 0, \tag{17}$$

where λ_1 and λ_2 are the eigenvalues of the Jacobian matrix

$$\left.\begin{pmatrix} \frac{\partial i'}{\partial i} & \frac{\partial i'}{\partial x_1} \\ \frac{\partial x_1'}{\partial i} & \frac{\partial x_1'}{\partial x_1} \end{pmatrix}\right|_{i=i^*, x_1=x_1^*}, \tag{18}$$

and $Re(x)$ means the real part of x.

As (17) are often difficult to solve, we often use numerical methods to find the solution of (17) and the steady states of the co-evolution of disease spread and behavioral choice.

$$\frac{di}{dt} = i(t)(1-i(t))\overline{\beta}\bar{k} - \gamma i(t), \tag{12}$$
$$\frac{dx_1}{dt} = x_1(t)x_2(t)\bar{k}i(t)(\beta_2-\beta_1) + \frac{m\omega}{U_{max}}x_1(t)x_2(t)(u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 i(t)] + u(c_1) - u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_2 i(t)] - u(c_2)).$$

III. The steady state Analysis

During the COVID-19 crisis, governments have been promoting home isolation as a crucial measure to contain the spread of the disease. Home isolation helps significantly reduces the risk of being infected and controls the pandemic, while it also leads to substantial economic loss as well as impacts people's physical and mental well-being. During the pandemic, individuals need to choose between high-cost low-risk conservative behavior and low-cost high-risk risky behavior. Their decisions are often influenced by the severity of the epidemic and the potential loss due to home isolation. People tend to choose self-isolation when the pandemic poses a greater threat to their health; while they may be inclined to choose to go out when the loss due to home isolation is too high (e.g., losing their jobs and income).

Therefore, in this section, based on our model in the previous section, we analyze the dynamics of people's choices between two behaviors: risky behavior (going out) represented by a_1, and conservative behavior (home isolation) represented by a_2, and the evolution of the epidemic. We assume that the infection rate for conservative behavior is $\beta_2 = 0$ (as isolation ensures no infection risk).

Then we analyze the steady state of our model. To find the steady state, we first introduce Lemma 1.

Lemma 1: The necessary and sufficient conditions that (i^*, x_1^*) is a steady state is:

$$\begin{aligned} &\frac{di}{dt}\Big|_{i=i^*} = 0,\ \frac{dx_1}{dt}\Big|_{x_1=x_1^*} = 0,\\ &(\frac{\partial i'}{\partial i} + \frac{\partial x_1'}{\partial x_1})\Big|_{i=i^*,x_1=x_1^*} < 0,\\ \text{and}\quad &(\frac{\partial i'}{\partial i}\frac{\partial x_1'}{\partial x_1} - \frac{\partial i'}{\partial x_1}\frac{\partial x_1'}{\partial i})\Big|_{i=i^*,x_1=x_1^*} > 0. \end{aligned} \tag{19}$$

Proof: By the definition of the steady state, we first have $\frac{di}{dt} = 0$ and $\frac{dx_1}{dt} = 0$. For simplicity, we let

$$\begin{aligned} &P(i^*, x_1^*) = (\frac{\partial i'}{\partial i} + \frac{\partial x_1'}{\partial x_1})\Big|_{i=i^*,x_1=x_1^*},\\ \text{and}\quad &Q(i^*, x_1^*) = (\frac{\partial i'}{\partial i}\frac{\partial x_1'}{\partial x_1} - \frac{\partial i'}{\partial x_1}\frac{\partial x_1'}{\partial i})\Big|_{i=i^*,x_1=x_1^*}. \end{aligned} \tag{20}$$

Then we have:

$$\lambda_1 = \frac{P+\sqrt{P^2-4Q}}{2},\ \text{and}\ \lambda_2 = \frac{P-\sqrt{P^2-4Q}}{2}. \tag{21}$$

From (17), the steady state requires $Re(\lambda_1) < 0$ and $Re(\lambda_2) < 0$. If $Q \le 0$, then it implies $Re(\lambda_1) \ge 0$, which contradicts the steady state conditions. Therefore, we must have $Q > 0$. Additionally, since $Q > 0$ and $Re(\lambda_1) < 0$, it follows that $P < 0$. Therefore, we have $P < 0$ and $Q > 0$, which means $\frac{\partial i'}{\partial i} + \frac{\partial x_1'}{\partial x_1} < 0$ and $\frac{\partial i'}{\partial i}\frac{\partial x_1'}{\partial x_1} - \frac{\partial i'}{\partial x_1}\frac{\partial x_1'}{\partial i} > 0$. ■

From Lemma 1, we can have Theorem 1 about the steady state (i^*, x_1^*).

Theorem 1: The steady state (i^*, x_1^*) of (12) satisfies:

$$(i^*, x_1^*) = \begin{cases} (0,1), & \text{if } \bar{k} < \frac{\gamma}{\beta_1},\\ \left(1-\frac{\gamma}{\bar{k}\beta_1}, 1\right), & \text{if } \bar{k} > \frac{\gamma}{\beta_1}, \Phi < 0,\\ \left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right), & \text{if } \bar{k} > \frac{\gamma}{\beta_1}, \Phi \ge 0, \end{cases} \tag{22}$$

where $\Phi \triangleq -k_0(u(c_1)-u(c_2)) - (\gamma - \bar{k}\beta_1) - k_0 u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 - \gamma]$, $k_0 = \frac{m\omega}{U_{max}}$, and $i^{(2)}$ is the solution of the equation

$$k_0 u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 i^{(2)}] - \bar{k}\beta_1 i^{(2)} + k_0(u(c_1)-u(c_2)) = 0. \tag{23}$$

Proof: By setting $\frac{di}{dt} = 0, \frac{dx_1}{dt} = 0$, we can get four steady state candidates: $(0,0)$, $(0,1)$, $\left(1-\frac{\gamma}{\bar{k}\beta_1}, 1\right)$ and $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$. Then, we check whether they are stable. Note that in reality, the payoff of going out should be larger than quarantine, we have $c_1 - c_2 > 0$ (that means $u(c_1) - u(c_2) > 0$). Since c_n is the loss of being infected, so $c_n < 0$ (that means $u(c_n) < 0$). Based on Lemma 1, the steady state should satisfy $P(i^*, x_1^*) < 0$ and $Q(i^*, x_1^*) > 0$, where $P(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ are defined in (20). Then we have:

- For the state $(0,0)$, we have $P = k_0(u(c_1)-u(c_2)) - \gamma$ and $Q = -\gamma k_0(u(c_1)-u(c_2))$. Since $u(c_1)-u(c_2) > 0$, $\gamma > 0$ and $k_0 > 0$, we have $Q < 0$. So $(0,0)$ is unstable.
- For the state $(0,1)$, we have $P = -(\gamma - \bar{k}\beta_1) - k_0(u(c_1) - u(c_2))$ and $Q = k_0(\gamma - \bar{k}\beta_1)(u(c_1)-u(c_2))$. As $u(c_1) - u(c_2) > 0$ and $k_0 > 0$, if $Q > 0$, we have $\gamma - \bar{k}\beta_1 > 0$, which is equivalent to $\bar{k} < \frac{\gamma}{\beta_1}$. When $\gamma - \bar{k}\beta_1 > 0$, we also have $P < 0$. Therefore, when $\bar{k} < \frac{\gamma}{\beta_1}$, $(0,1)$ is a steady state.
- For the state $\left(1-\frac{\gamma}{\bar{k}\beta_1}, 1\right)$, we have:

$$\begin{aligned} P =& -k_0(u(c_1)-u(c_2)) - k_0 u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 - \gamma],\\ Q =& (\gamma - \bar{k}\beta_1)\{-k_0(u(c_1)-u(c_2))\\ &- (\gamma - \bar{k}\beta_1) - k_0 u(c_n)\cdot\boldsymbol{\omega}[\bar{k}\beta_1 - \gamma]\}. \end{aligned} \tag{24}$$

Since $1 - \frac{\gamma}{\bar{k}\beta_1}$ is the proportion of infected individuals, we have $1 - \frac{\gamma}{\bar{k}\beta_1} > 0$, which is equivalent to $\bar{k} > \frac{\gamma}{\beta_1}$. (If $1 - \frac{\gamma}{\bar{k}\beta_1} = 0$, then the state $\left(1-\frac{\gamma}{\bar{k}\beta_1}, 1\right)$ becomes $(0,1)$,

which we have discussed). To satisfy $P < 0$ and $Q > 0$, we first define

$$\begin{aligned} h_1 &\triangleq -k_0(u(c_1)-u(c_2)) - k_0 u(c_n)\cdot \boldsymbol{\omega}[\bar{k}\beta_1 - \gamma], \\ \text{and } h_2 &\triangleq -k_0(u(c_1)-u(c_2)) - (\gamma - \bar{k}\beta_1) \\ &\quad - k_0 u(c_n)\cdot \boldsymbol{\omega}[\bar{k}\beta_1 - \gamma]. \end{aligned} \tag{25}$$

Note that $P < 0$ and $Q > 0$ are equivalent to $h_1 < 0$ and $h_2 < 0$. It is obvious that $h_1 = h_2 - (\bar{k}\beta_1 - \gamma)$. Since $(\bar{k}\beta_1 - \gamma) > 0$, if $h_2 < 0$, then $h_1 < 0$, and $P < 0$, $Q > 0$. Note that $\Phi = h_2$. Therefore, $\left(1 - \frac{\gamma}{\bar{k}\beta_1}, 1\right)$ is a steady state when $\bar{k} > \frac{\gamma}{\beta_1}$ and $\Phi < 0$.

- First, as $i^{(2)}$ and $\frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}$ are the proportion of infected individuals and the proportion of individuals who choose risky behavior, by definition, we have $0 \le i^{(2)} \le 1$ and $0 \le \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1} \le 1$, which is equivalent to:

$$0 \le i^{(2)} \le 1 - \frac{\gamma}{\bar{k}\beta_1}. \tag{26}$$

From (26), it is obvious that we should have $0 < 1 - \frac{\gamma}{\bar{k}\beta_1}$, which means $\bar{k} > \frac{\gamma}{\beta_1}$ (If $0 = 1 - \frac{\gamma}{\bar{k}\beta_1}$, then the state $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$ becomes $(0,1)$, which we have discussed). Then, we define $f_2(x) \triangleq k_0 u(c_n)\cdot\boldsymbol{\omega}(\bar{k}\beta_1 x) - \bar{k}\beta_1 x + k_0(u(c_1)-u(c_2))$. From (23), $i^{(2)}$ is the solution of $f_2(x)$ so we have $f_2(i^{(2)}) = 0$. It is obvious that $f_2(x)$ is a monotonically decreasing function of $i^{(2)}$. Then $0 \le i^{(2)} \le 1 - \frac{\gamma}{\bar{k}\beta_1}$ is equivalent to:

$$f_2(0) \ge 0, \quad f_2(1 - \frac{\gamma}{\bar{k}\beta_1}) \le 0. \tag{27}$$

Note that $u(c_1) - u(c_2) > 0$ and $k_0 > 0$. Therefore we have $f_2(0) = k_0(u(c_1)-u(c_2)) > 0$. So when $\bar{k} > \frac{\gamma}{\beta_1}$ and $f_2(1 - \frac{\gamma}{\bar{k}\beta_1}) \le 0$, we have $0 \le i^{(2)} \le 1$ and $0 \le \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1} \le 1$.

Then, if $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$ is a steady state, it should satisfies $P < 0$ and $Q > 0$, so we have:

$$\begin{aligned} P &= -\gamma + (1-2i^{(2)})\bar{k}\beta_1 x_1^{(2)} = -\bar{k}\beta_1 i^{(2)} x_1^{(2)}, \\ Q &= -\bar{k}\beta_1 i^{(2)} x_1^{(2)}(1-i^{(2)})(1-x_1^{(2)}) \\ &\quad \cdot (k_0 u(c_n)\boldsymbol{\omega}'[\bar{k}\beta_1 i^{(2)}]\bar{k}\beta_1 - \bar{k}\beta_1). \end{aligned} \tag{28}$$

It is obvious that $P < 0$. Since $\boldsymbol{\omega}'[\bar{k}\beta_1 i^{(2)}] > 0$ and $k_0 u(c_n) < 0$, we have $k_0 u(c_n)\boldsymbol{\omega}'[\bar{k}\beta_1 i^{(2)}]\bar{k}\beta_1 - \bar{k}\beta_1 < 0$. So $Q > 0$.

Therefore, when $\bar{k} > \frac{\gamma}{\beta_1}$ and $f_2(1 - \frac{\gamma}{\bar{k}\beta_1}) \le 0$, $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$ is steady state. Note that $f_2(1 - \frac{\gamma}{\bar{k}\beta_1}) = -\Phi$, therefore, when $\bar{k} > \frac{\gamma}{\beta_1}$ and $\Phi \ge 0$, $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$ is steady state.

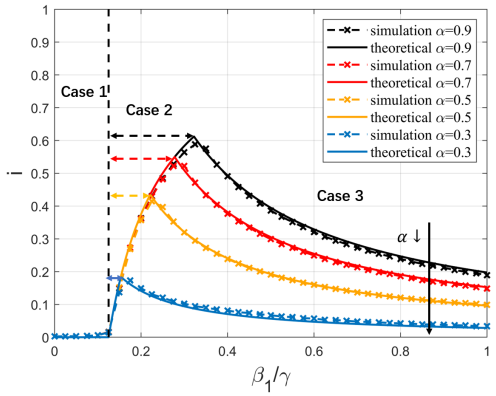
■

From Theorem 1, our epidemic-behavior co-evolution model has possible three steady states, and each corresponds to a particular situation in reality:

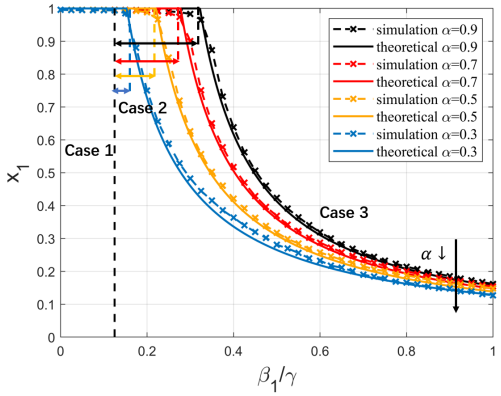
- Case 1: The steady state $(0,1)$ represents an extreme situation where the infection rate is too low, and the disease would eventually die out even without any intervention or change of behavior. In this case, all individuals choose risky behavior. For the steady state $(0,1)$, the stability condition is $\bar{k} < \frac{\gamma}{\beta_1}$, which is equivalent to $\frac{\beta_1}{\gamma} < \frac{1}{\bar{k}}$. Here, $\frac{1}{\bar{k}}$ represents the epidemic threshold of a regular network with degree $\bar{k}$ [24]. If $\frac{\overline{\beta}}{\gamma} < \frac{1}{\bar{k}}$, the epidemic would die out; otherwise, it would spread out [24]. Since $\overline{\beta} \in [0, \beta_1]$, the disease would die out regardless of how people choose their behavior. In this situation, all individuals would choose the risky behavior in the steady state. Therefore, when $\bar{k} < \frac{\gamma}{\beta_1}$, the stable point is $i = 0$ and $x_1 = 1$.
- Case 2: For the steady state $\left(1 - \frac{\gamma}{\bar{k}\beta_1}, 1\right)$, the first constraint is $\bar{k} > \frac{\gamma}{\beta_1}$, which means if all individuals choose the risky behavior, the disease will spread out. The second constraint is $\Phi < 0$. When all individuals choose to go out with $x_1 = 1$, the percentage of infected individuals will reach the stable state $\bar{i} = 1 - \frac{\gamma}{\bar{k}\beta_1}$, which is the maximum extent to which the disease can spread (please refer to Appendix A for a proof). If for all possible i in the range $[0, \bar{i}]$, we have $\frac{dx_1}{dt}|_{x_1 \ne 0,1} > 0$ (if $x_1 = 0$ or 1, $\frac{dx_1}{dt} = 0$), then more people will choose risky behavior as time goes on, and $x_1 = 1$ with all individuals choosing the risky behavior at the steady state. This may happen when the risky behavior's payoff is much higher than the conservative behavior's with $c_1 \gg c_2$, or when the cost of being infected c_n is very low. If $\frac{dx_1}{dt}|_{i=\bar{i}, x_1 \ne 0,1} > 0$, which is equivalent to $\Phi < 0$, then for all $i \in [0, \bar{i}]$, we have $\frac{dx_1}{dt}|_{x_1 \ne 0,1} > 0$. Therefore, when $\bar{k} > \frac{\gamma}{\beta_1}$ and $\Phi < 0$, $\left(1 - \frac{\gamma}{\bar{k}\beta_1}, 1\right)$ is the steady state.
- Case 3: The steady state $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$ represents the general situation other than the two extreme cases. In this scenario, the disease does not extinct, nor does it spread to the maximum extent, and at the steady state, $0 < i^{(2)} < \bar{i}$ percent of individuals will be infected. Meanwhile, $0 < \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1} < 1$ of susceptible individuals will choose the risky behavior. When $\bar{k} > \frac{\gamma}{\beta_1}$, and $\Phi \ge 0$, the state $\left(i^{(2)}, \frac{\gamma}{(1-i^{(2)})\bar{k}\beta_1}\right)$ is stable.

IV. Simulation Results

In this section, we run simulations to verify our analysis in the above sections. The network we used is a regular network with 500 nodes and a fixed degree of 8. Since the risky behavior of going out is the default behavior that most people take in their daily lives, we set the payoff of this behavior to 0, while the conservative behavior of self-isolation can be regarded as behavior with losses and a negative payoff. As an example, we set the payoff of the risky behavior as $c_1 = 0$, the payoff for the conservative behavior as $c_2 = -1$, and the disease loss as $c_n = -10$. For the value function, we set $\sigma = 0.65$ according to [16]. In our simulations, we set $\lambda = 1$, and observe the same trend for other values of λ.

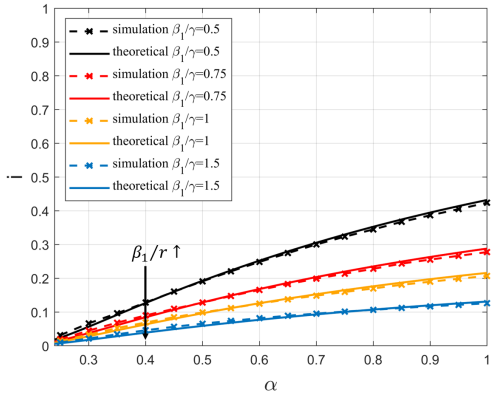


(a) The proportion of infected individuals

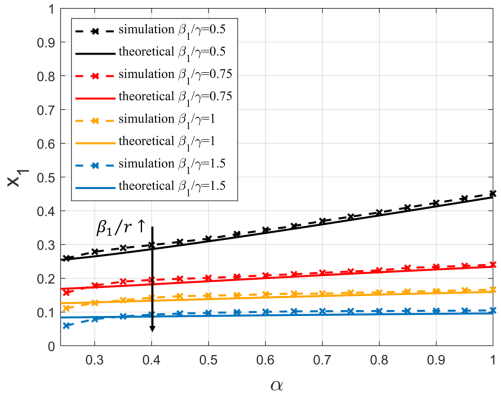


(b) The proportion of individuals choosing risky behavior

Fig. 2. The simulation and theoretical results of steady states with different disease spread ability



(a) The proportion of infected individuals



(b) The proportion of individuals choosing risky behavior

Fig. 3. The simulation and theoretical results of steady states with different disease rationality coefficients α.

Fig. 2 shows the results of steady state with changing $\frac{\beta_1}{\gamma}$ which reflects the spread ability of the disease. A higher value of $\frac{\beta_1}{\gamma}$ indicates a greater ability of the disease to spread within the population. First, from Fig. 2, the theoretical results fit well with the simulation results. The three cases in Theorem 1 can be clearly seen in Fig. 2. When $\frac{\beta_1}{\gamma}$ is low, even if all individuals choose risky behavior, the disease will die out, which corresponds to Case 1 in the figure. When $\frac{\beta_1}{\gamma}$ is high enough so that the disease would spread out, it enters Case 2. As $\frac{\beta_1}{\gamma}$ increases, the proportion of infected individuals in the steady state gradually increases, and at this time, all individuals would still choose risky behaviors since it has a higher payoff. As $\frac{\beta_1}{\gamma}$ continues increasing, it enters Case 3, where more and more people choose self-isolation due to the high risk of infection. In this case, although $\frac{\beta_1}{\gamma}$ become higher, more people choose to self-isolate due to the increased risk of infection, which makes the proportion of infected individuals in the steady state gradually decrease.

Moreover, from Fig. 2, given $\frac{\beta_1}{\gamma}$, individuals with a lower rationality coefficient α (meaning that such individuals are more irrational) have a smaller probability to be in Case 2, which means that irrational individuals are more inclined to choose conservative behaviors during a pandemic. This may be because from the Prospect Theory, irrational users tend to overestimate the small probability of being infected, and thus they tend to be more conservative. We also run simulations on the steady state of Case 3 under different rationality coefficients α and the results are shown in Fig. 3. It can be seen that fewer individuals choose risky behavior when the rational coefficient α is smaller, confirming our previous conclusion about their conservative behavior.

V. Conclusion

In this paper, we propose the epidemic-behavior co-evolution method to capture individuals' behavior during a pandemic and analyze the co-evolution of disease spread and behavior change. We incorporate irrational behavior based on Prospect Theory and examine how it influences people's decision-making processes and, in turn, the spread of the disease. Through rigorous theoretical analysis, we identify three steady states under different scenarios and investigate

their practical implications. To validate our theoretical findings and gain further insights, we conduct simulation experiments. The experimental results align closely with our theoretical predictions, confirming the correctness of our model. Notably, our findings consistently indicate that irrationality tends to steer individuals towards more conservative behavioral choices during a pandemic, effectively mitigating the spread of the disease.

APPENDIX A
PROOF FOR THE MAXIMUM SPREAD EXTENT

For the steady state (i^*, x_1^*), it satisfies $\frac{di}{dt} = 0$. If $i^* \neq 0$, then based on (12), we have:

$$i^* = 1 - \frac{\gamma}{\bar{k}\beta_1 x_1^*}, \tag{29}$$

and i^* is an increasing function of x_1^*, where $x_1^* \in [0, 1]$. Therefore, i^* reaches its maximum value when $x_1^* = 1$. Substituting this value into the equation, we obtain $\bar{i} = 1 - \frac{\gamma}{\bar{k}\beta_1}$.

REFERENCES

[1] H.-F. Zhang, J.-R. Xie, M. Tang, and Y.-C. Lai, "Suppression of epidemic spreading in complex networks by local information based behavioral responses," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 24, no. 4, p. 043 106, 2014.

[2] Q. Wu, X. Fu, M. Small, and X.-J. Xu, "The impact of awareness on epidemic spreading in networks," *Chaos: an interdisciplinary journal of nonlinear science*, vol. 22, no. 1, p. 013 101, 2012.

[3] F. Bagnoli, P. Lio, and L. Sguanci, "Risk perception in epidemic modeling," *Physical Review E*, vol. 76, no. 6, p. 061 904, 2007.

[4] S. Funka, E. Gilada, C. Watkinsb, and V. A. Jansena, "The spread of awareness and its impact on epidemic outbreaks," *PNAS*, vol. 106, no. 16, 2009.

[5] C. Granell, S. Gómez, and A. Arenas, "Dynamical interplay between awareness and epidemic spreading in multiplex networks," *Physical review letters*, vol. 111, no. 12, p. 128 701, 2013.

[6] C. Granell, S. Gómez, and A. Arenas, "Competing spreading processes on multiplex networks: Awareness and epidemics," *Physical review E*, vol. 90, no. 1, p. 012 808, 2014.

[7] Z. Wang, Q. Guo, S. Sun, and C. Xia, "The impact of awareness diffusion on sir-like epidemics in multiplex networks," *Applied Mathematics and Computation*, vol. 349, pp. 134–147, 2019.

[8] C. Zheng, C. Xia, Q. Guo, and M. Dehmer, "Interplay between sir-based disease spreading and awareness diffusion on multiplex networks," *Journal of Parallel and Distributed Computing*, vol. 115, pp. 20–28, 2018.

[9] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," in *Handbook of the fundamentals of financial decision making: Part I*, World Scientific, 2013, pp. 99–127.

[10] D. Kahneman and A. Tversky, "Choices, values, and frames," in *Handbook of the fundamentals of financial decision making: Part I*, World Scientific, 2013, pp. 269–278.

[11] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and uncertainty*, vol. 5, no. 4, pp. 297–323, 1992.

[12] T. Oraby and C. T. Bauch, "Bounded rationality alters the dynamics of paediatric immunization acceptance," *Scientific reports*, vol. 5, no. 1, pp. 1–12, 2015.

[13] A. R. Hota and S. Sundaram, "Game-theoretic vaccination against networked sis epidemics and impacts of human decision-making," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 4, pp. 1461–1472, 2019.

[14] X.-J. Li and X. Li, "Perception effect in evolutionary vaccination game under prospect-theoretic approach," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 329–338, 2020.

[15] P. Poletti, B. Caprile, M. Ajelli, A. Pugliese, and S. Merler, "Spontaneous behavioural changes in response to epidemics," *Journal of theoretical biology*, vol. 260, no. 1, pp. 31–40, 2009.

[16] D. Prelec, "Compound invariant weighting functions in prospect theory," *Choices, values, and frames*, pp. 67–92, 2000.

[17] K. A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979.

[18] D. Prelec, "The probability weighting function," *Econometrica*, pp. 497–527, 1998.

[19] J. O. Kephart and S. R. White, *DIRECTED-GRAPH EPIDEMIOLOGICAL MODELS OF COMPUTER VIRUSES*. Computation: The Micro and The Macro View.

[20] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: A modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[21] A. Traulsen, J. C. Claussen, and C. Hauert, "Coevolutionary dynamics: From finite to infinite populations," 2004.

[22] J. Rowlett and C. J. Karlsson, "Decisions and disease: The evolution of cooperation in a pandemic," 2020.

[23] A. M. Lyapunov, "The general problem of the stability of motion," *International journal of control*, vol. 55, no. 3, pp. 531–534, 1992.

[24] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, "Epidemic thresholds in real networks," *Acm Transaction on Information & System Security*, vol. 10, no. 4, p.1–26, 2007.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Noise-robust Pitch Detection Based on Super-Resolution Harmonics

Dongjie Zhu*, Weibin Zhu*, Tianrui Wang*, Yingying Gao†, Junlan Feng† and Shilei Zhang†
* Institute of Information Science, Beijing Jiaotong University, Beijing, China
E-mail: 21120323@bjtu.edu.cn Tel: +886-18847171485
† China Mobile Research Institute, Beijing, China
E-mail: gaoyingying@chinamobile.com Tel: +886-15210565528

***Abstract*—In order to improve the performance of pitch detection algorithms in a noisy scenario, we propose a noise-robust pitch detection model based on super-resolution harmonics. The model consists of denoise, integration, and super-resolution modules. The Denoisy module highlights the harmonic structure of speech in the frequency domain to reduce noise interference. The enhanced spectrum is multiplied by an integral matrix to obtain a pitch-harmonic integration spectrum. The super-resolution module adjusts the low-resolution pitch-harmonic integration spectrum and outputs fine results. In addition, dynamic programming is employed for pitch tracking to eliminate those singular spots in the preliminary results. The experimental results show that the proposed model significantly outperforms the referenced approaches, including SWIPE, CREPE, HarmoF0, and HGCN+. The results also show that the model can work well even under low SNR.**

I. Introduction

Pitch often represented by f_0 is an important characteristic of speech. It is widely used in speech applications, such as speech enhancement [1], speech recognition [2], voice conversation [3], and speech synthesis [4], etc.

The traditional approaches can be divided into the time-domain methods (RAPT [5], YIN [6], and pYIN [7]) and the frequency-domain methods (SWIPE [8]). Generally, pitch detection is implemented in two steps. First, the pitch candidates are detected based on the periodicity of voiced segments [5] or the instant of uttering events [9]. Then the pitch tracking adopts dynamic programming [10] or hidden Markov [11] to model temporal continuity constraint. In recent years, pitch detection has been formulated as a supervised learning task. In [12], a supervised learning algorithm to estimate pitch using a convolutional neural network (CNN) and CREPE [13] operates on the waveform to estimate the pitch. [14] proposed a network that jointly detects the voiced segments and pitch. These approaches work well on clean speech, but their performance will decline for noisy speech. can be divided into the time-domain methods (RAPT [5], YIN [6], and pYIN [7]) and the frequency-domain methods (SWIPE [8]). Generally, pitch detection is implemented in two steps. First, the pitch candidates are detected based on the periodicity of voiced segments [5] or the instant of uttering events [9]. Then the pitch tracking adopts dynamic programming [10] or hidden Markov [11] to model temporal continuity constraint. With the introduction of deep learning, pitch detection is formulated as a supervised learning task, and its performance is significantly improved. In [12], a supervised learning algorithm to estimate pitch using a convolutional neural network (CNN) and CREPE [13] operates on the waveform to estimate the pitch. [14] proposed a network that jointly detects the voiced segments and pitch. These approaches work well on clean speech, but their performance will decline for noisy speech. In the speech spectrum, harmonic frequencies are integer multiples of the pitch, where these harmonics show energy peaks locally and present a comb-like structure. Even at low SNR, the harmonic structure is seldom completely masked. HarmoF0 [15] introduced multiple rates dilated causal convolution network [16] to capture the harmonic structure in the logarithmic scale spectrum and then estimate the pitch. Since there is no dedicated module for harmonic enhancement, this approach does not seem to work well at low SNR.

High-resolution fundamental frequencies are essential for speech enhancement and timbre conversion. The high-resolution pitch is verified to be necessary for the use of high-order harmonics. Still, the double error and/or half error are occasionally ignored in our former system (HGCN+ [17]). In HGCN+, a high-resolution harmonic-based pitch detection was proposed and adopted to guide the model for speech enhancement. HGCN+ first performs a coarse spectrum enhancement, and then harmonics integration is utilized to predict the pitch. Finally, the harmonic gating mechanism is used to guide the compensation model to adjust the coarse results from CRN to obtain the refinedly enhanced results. HGCN+ adopts pitch as an additional feature, and the enhanced results would be better with a higher resolution fundamental frequency. The fundamental frequency is often used for prosody modeling in voice conversion. VQMIVC [3] employs vector quantization (VQ) for content encoding and introduces mutual information (MI) as the correlation metric during training to achieve proper disentanglement of content, speaker and pitch representations by reducing their inter-dependencies in an unsupervised manner. High-resolution fundamental frequency enables model prosody to be more precise, resulting in better disentanglement. And the resolution of the estimated pitch can also be improved by a larger frame-size of Fourier transform, in addition to the complexity problem, the over-smoothing problem is also inevitable.

For a practical pitch detection algorithm to run in the real

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

world, there are two issues need to be solved, which include noise robustness and frequency resolution. In this study, starting from the pitch detection module of HGCN+, we improve the integral matrix. Then the U-Net [18] based harmonic enhancement, integral matrix based pitch detection, and super-resolution module are integrated into our network, resulting in a noise-robust high-resolution pitch detection method. We compared our methods with four other advanced ones on both noisy and clean datasets. The outstanding performances with low complexity of the method are verified experimentally.

II. Proposed Method

The overall architecture of the proposed model is shown in Fig. 1, which consists of the harmonic enhancement module (HEM), integral module (IM), and confidence super-resolution module (CSM). The HEM enhances the harmonics of noisy speech. Then IM calculates the pitch-harmonic integration spectrum based on the enhanced magnitude spectrum. Finally, the CSM is used to generate the confidence spectrum. Each module is described as follows.

A. Harmonic Enhancement Module

The harmonic enhancement module (HEM) is designed to highlight the harmonics while suppressing the noise on the magnitude spectrum for more accurate pitch detection. It estimates a mask $\boldsymbol{M}$ from a given noisy magnitude spectrum. And then, the mask is applied to the noisy signal,

$$\boldsymbol{S'} = \boldsymbol{S} \odot \boldsymbol{M} \tag{1}$$

where $\boldsymbol{S'}$ and $\boldsymbol{S}$ are the enhanced and noisy magnitude spectra, respectively. And the $\odot$ denotes the element-wise multiplication. A U-Net based enhancement network is adopted, which is an encoder-decoder architecture. Both encoder and decoder consist of Batch normalization (BN) [19], 2D convolution layer (Conv2d), and Parametric Rectified Linear Unit (PReLU) [20]. Between the encoder and decoder, long short-term memory (LSTM) [21] is inserted to model the temporal dependencies. Additionally, skip connections are utilized to concatenate the output of each encoder layer to the input of the corresponding decoder layer. The channel number of encoder and decoder is $(16, 32, 64, 128, 128, 128)$. And a 512units FC layer after a 128-units LSTM is adopted.

To highlight the harmonics in magnitude spectrum, the training label for HEM is a harmonic enhanced magnitude spectrum, which will be detailed in the section II-E.

B. Integral Module

Harmonics are the key to spectrum-based pitch detection, which appear as comb-like structures in the magnitude spectrum, as peaks are located at integer multiples of the pitch with the locally highest energy. An artificial comb-structure integral matrix $\boldsymbol{K}$ is proposed to mimic the distribution of harmonics, which is adopted as a checklist for different pitches, as shown in Fig. 1.

The integral matrix $\boldsymbol{K}$ is designed as shown in Algorithm 1, where F is the frequency dimension of the magnitude spectrum, and sr is the sample rate. R is the resolution of pitch candidates, which is set artificially. Because the matrix dimension and computational cost increase with higher resolution, R is set to 1 to compromise the computational costs. A total of 360 pitches in 60∼420 Hz are taken as candidates. For special pitch candidate f_c, the energy above $sr/(4 \cdot R \cdot f_c)$ order harmonic has no reference with lower energy. So only the harmonics from 1 to $sr/(4 \cdot R \cdot f_c)$ are allowed to participate in the integration. And to alleviate the double-pitch and half-pitch errors, decaying weighting factors are applied to the integral matrix, and the weight of $(R \cdot f_c \cdot p)^{0.23}$ is applied for the p order harmonic. The weights between $(p-1)$ order and p order harmonic are filled with cosine curves.

Algorithm 1 Integral matrix

$\boldsymbol{K} \leftarrow \boldsymbol{0} \in \mathbb{R}^{[420/R] \times F}$
for $f_c \leftarrow [60/R] \rightarrow [420/R]$ **do**
  $\text{loc}_{\text{last}} \leftarrow 0$; $\text{peak}_{\text{last}} \leftarrow 1$; $j \leftarrow f_c$
  for $p \leftarrow 1 \rightarrow [sr/(4 \cdot R \cdot f_c)]$ **do**
    $\text{loc} \leftarrow [R \cdot f_c \cdot p \cdot F/\,(sr/2)]$
    $\text{peak} \leftarrow 1/(R \cdot f_c \cdot p)^{0.23}$; $\boldsymbol{K}_{j,\text{loc}} \leftarrow \text{peak}$
    if $\text{loc}_{\text{last}} \neq 0$ **then**
      if $\text{loc} - \text{loc}_{\text{last}} > 1$ **then**
        $\text{num}_{\text{iner}} = \text{loc} - \text{loc}_{\text{last}}$
        $\boldsymbol{F}^{\cos} \leftarrow \cos\left(\text{linspace}(0, 2\pi, \text{num}_{\text{iner}})\right)$
        $\boldsymbol{F} \leftarrow \text{linspace}(\text{peak}_{\text{last}}, \text{peak}, \text{num}_{\text{iner}})$
        for $i \leftarrow 1 \rightarrow \text{num}_{\text{iner}}$ **do**
          $\boldsymbol{K}_{j,i+\text{loc}_{\text{last}}} = \boldsymbol{F}_i^{\cos} \cdot \boldsymbol{F}_i$
      else
        $\boldsymbol{K}_{j,\text{loc}} \leftarrow \boldsymbol{K}_{j,\text{loc}} - (\text{peak}_{\text{last}} + \text{peak})/2$
        $\boldsymbol{K}_{j,\text{loc}_{\text{last}}} \leftarrow \boldsymbol{K}_{j,\text{loc}_{\text{last}}} - (\text{peak}_{\text{last}} + \text{peak})/2$
    $\text{loc}_{\text{last}} \leftarrow \text{loc}$; $\text{peak}_{\text{last}} \leftarrow \text{peak}$

As shown in Fig. 1, the enhanced $\boldsymbol{S'}$ is fed into IM to calculate the pitch-harmonic integration spectrum $\boldsymbol{Q}$,

$$\boldsymbol{Q} = \boldsymbol{K} \cdot \boldsymbol{S'} \tag{2}$$

where $\boldsymbol{Q} \in \mathbb{R}^{420 \times 500}$ represents the similarity of the magnitude to the harmonics distribution of different pitch candidates (row of $\boldsymbol{K}$ in Fig. 1).

C. Super-Resolution Module

The super-resolution is adopted to refine the resolution of the given pitch-harmonic integration spectrum with 1 Hz resolution from the IM. The pitch resolution can be increased to 0.2 Hz by the CSM, which is composed of four Conv2d layers and one fully-connected layer (FC). The feature size of the input is $\boldsymbol{Q} \in \mathbb{R}^{420 \times 500}$. The stride of all Conv2d is 1, the kernel size is $(5, 3, 3, 1)$, and the channel number is $(32, 16, 5, 1)$. The first three Conv2d are used to expand the channels up to 5 times. The fourth Conv2d gathers channel features, and the output is flattened into a vector. The vector

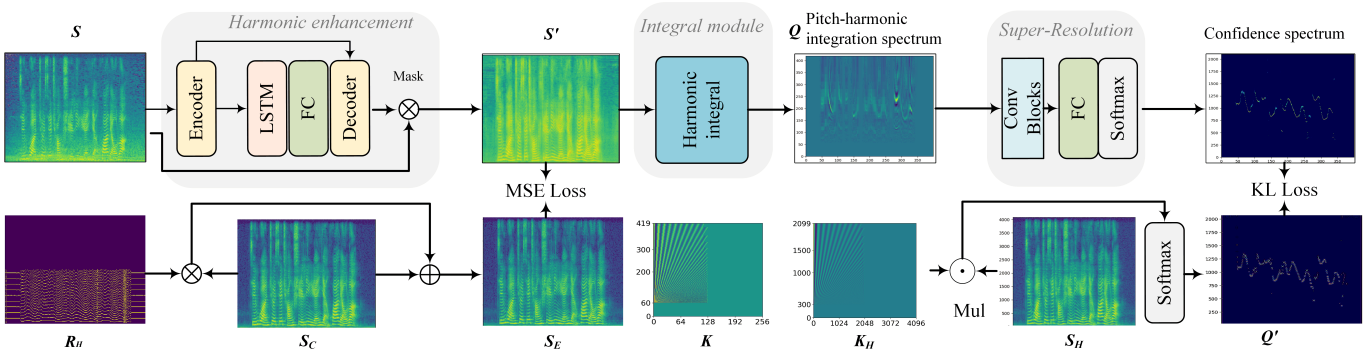


Fig. 1. Architecture of the proposed model.

is fed into FC, resulting in the boosted feature. The activation function of the last FC is softmax. The output of CSM is referred to as the confidence spectrum, which represents the probability of the pitch candidates.

D. Pitch Tracking

The preliminarily detected result may have singular spots without post-processing. Pitch tracking can smooth the expected result by the dynamic programming algorithm. The state probability at each stage is the confidence spectrum. The transfer probability is modeling the temporal continuity constraint, which can be modeled by a normal distribution.

$$p_t(\Delta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(\Delta-\mu)^2}{2\sigma^2}\right) \tag{3}$$

where Δ represents the pitch change of adjacent frames. Here, we set $\mu = 0$ and $\sigma = 0.1$. The transfer probability reduces as Δ increases, which limits the pitch changing significantly. The dynamic programming algorithm will maximize both state probability at each stage and the transfer probability. Then the final continuous pitch contour is generated by backtracking.

E. Label Generation and Loss Function

There is no pitch annotation on the DNS Challenge dataset [22]. According to Algorithm 1, a high-resolution integra matrix $\boldsymbol{K_H}$ is generated by setting R to 0.2, F to 4097. $\boldsymbol{K_H}$ is shown in Fig 1. A total of 2100 frequencies in 60∼420 Hz are taken as candidates. Then, the matrix $\boldsymbol{K_H}$ is applied to the 4097-dimension magnitude spectrum transformed by the 1024 Hanning window and a hop size of 160. $\boldsymbol{Q'}$ is the target of CSM.

$$\boldsymbol{Q'} = \boldsymbol{K_H} \cdot \boldsymbol{S_H} \tag{4}$$

The KL divergence is used as the loss function to optimize the model. The candidate corresponding to the maximum value in $\boldsymbol{Q'}$ is selected as the pitch, and the corresponding harmonic locations are used as the result $\boldsymbol{R_H} \in \mathbb{R}^{T\times F}$, where the harmonic locations are 1, and the non-harmonic locations are 0.

$$\boldsymbol{S_E} = \boldsymbol{S_C} \odot (\boldsymbol{1} + \boldsymbol{R}_H) \tag{5}$$

$\boldsymbol{S_E}$ denotes the target of the HEM and $\boldsymbol{S_C}$ denotes the spectrum of the clean speech. The MSE loss function is used for HEM.

III. EXPERIMENTAL SETUP

A. Datasets

The proposed model is trained on DNS Challenge (INTERSPEECH 2020) dataset. The DNS Challenge dataset includes 500 hours of clean speech from 2150 speakers and over 180 hours of noise data from 150 classes. For training, we generate 200 hours of noisy speech with DNS noise. The SNR range is -5 dB to 15 dB. And data is divided into training and validation sets at 4: 1. For testing, noisy speech is generated based on PTDB-TUG dataset[23]. The PTDB-TUG dataset comprises 4720 speech and laryngograph recordings of 20 native English speakers (10 males and 10 females) with a total length of 576 minutes. The noises of test data are from the ESC-50, and the SNRs are -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. A total of 2880 minutes of test audio are generated.

B. Model and Training Setup

The Hamming window of size 512 and a hop size of 160 are used, resulting in the 257 dimension spectral features. The model is trained using Adam [24] with default hyperparameters. And the initial learning rate is 0.001, which will decay to 50% when the validation loss plateau for 5 epochs, and the training is stopped if the loss plateau for 15 epochs. We trained the model with 4 TITAN X GPUs for one day.

C. Evaluation Metrics

Raw Ritch Accuracy (RPA) [25] is used to evaluate and compare the performance of the proposed model with the compared ones. RPA measures the percentage of audio frames where the frequency estimate is accurate within the threshold value. We select 10 cents, 30 cents, and 50 cents as threshold values. The octave is divided into 12 equal parts, which is called a semitone. The semitone is divided into 100 equal parts, and each part is a cent. The higher the RPA is, the better the model performs. Note that RPA ignores voicing errors.

TABLE I
RPA OF DIFFERENT MODELS ON NOISY PTDB-TUG DATASET, TESTED WITH THE 10, 30 AND 50 CENTS THRESHOLD.

Model	10 cents					AVG	30 cents					AVG	50 cents					AVG
	-5db	0db	5db	10db	15db		-5db	0db	5db	10db	15db		-5db	0db	5db	10db	15db	
SWIPE	2.7	5.4	8.4	12.9	17.6	9.4	15.9	11.8	18.6	28.2	37.4	22.4	7.2	14.3	22.4	33.7	44.8	24.5
CREPE	15.5	19.7	21.9	22.9	24.5	20.9	39.1	48.4	52.4	55.7	59.3	51.0	52.6	63.7	68.0	72.1	76.5	66.6
HGCN+	17.7	19.3	20.7	21.5	22.5	20.3	43.3	46.9	50.5	52.7	54.5	49.6	60.6	65.3	69.0	71.6	73.8	68.1
HarmoF0	21.2	27.2	29.4	30.1	31.6	27.9	46.2	57.0	61.2	62.7	65.3	58.5	55.9	68.0	72.3	74.0	76.4	69.3
Proposed	**31.1**	**35.2**	**36.5**	**36.5**	**36.6**	**35.2**	**60.4**	**66.6**	**68.3**	**68.5**	**68.3**	**66.4**	**72.6**	**78.8**	**80.2**	**80.3**	**81.1**	**78.6**

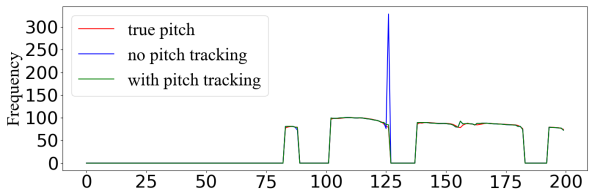


Fig. 2. Example output of the proposed method. The example mixture is a male utterance which is mixed with noise at 0 dB.

IV. RESULTS AND ANALYSIS

A. Performance in Noisy Conditions

The performance in noisy environments is related to the robustness and the application in real situations. We conduct experiments to validate the performance of the proposed framework on the PTDB-TUG dataset with various noises.

As shown in Table I, our model achieved state-of-the-art accuracy. The RPA of our proposed framework shows $(54.1\%, 12\%, 10.5\%, 9.3\%)$ of relative improvement compared to SWIPE, CREPE, HGCN+, and HarmoF0 with 50 cents threshold. The HGCN+ model enhances the speech coarsely, so it predicts well in the noisy environment. This indicates that the enhancement module is important for the proposed model to predict the fundamental frequency in a noisy environment. A similar trend can be seen in other thresholds. Since the HEM recovers the harmonic structure, the proposed framework is noise-robust.

For the average RPA, the ratios of 10 cent accuracy to 50 cent accuracy of SWIPE, CREPE, HGCN, HarmoF0, and the proposed model are $(38.4\%, 31.4\%, 30\%, 40.3\%, 44.8\%)$, respectively. The detected results of proposed model have a higher proportion within 10 cents, which indicates that the proposed method performs more precisely. This owes to CSM. The detected result may have a rough spot, which can be smoothed by dynamic programming (Fig. 2). Overall, our proposed method achieves better performance under noise scenarios.

B. Pitch Accuracy

For clean speech, the pitch detection performance of the proposed model is depicted in Table II. Our proposed model outperforms other models in different thresholds, which is around 5 million as compared with the CREPE model, which uses 22.2 million parameters. The RTF is second only to the HarmoF0. The RPA of HGCN+, HarmoF0, and the proposed model are similar with 50 cents threshold. The difference of performance becomes more obvious as the threshold decreases, which indicates our method predicts more precisely. To summarize, our proposed model outperforms the compared ones in terms of raw pitch accuracy on PTDB-TUG datasets for clean speech.

TABLE II
RPA ON CLEAN PTDB-TUG DATASET, TESTED WITH THE 10, 30 AND 50 CENTS THRESHOLD

Threshold	Params	RTF	10 cents	30 cents	50 cents
SWIPE	-	0.078	23.1	50.0	60.4
CREPE	22.2M	0.231	25.0	60.0	77.3
HGCN+	3.92M	0.066	18.1	46.4	60.1
HarmoF0	**0.38M**	**0.012**	32.9	64.0	73.4
Proposed	4.45M	0.051	**36.4**	**68.5**	**80.2**

C. Model Analysis

In addition, we further analyze the role of the HEM and CSM. We reduce the HEM and the CSM of the proposed model, respectively, and the results are presented in Table III. The proposed framework significantly outperforms the model without the HEM. Due to the absence of the HEM, the degraded harmonic structure will lead to failing to detect correctly. Without the CSM, the performance is similar with the 50 cents threshold while a gap of 7% with the 10 cents threshold. It indicates that the CSM makes the detected result more precise.

TABLE III
EVALUATION RESULTS OF THE ABLATION STUDY OF PROPOSED MODEL ON 0 SNR ENVIRONMENT.

Threshold	10 cents	30 cents	50 cents
Proposed	35.2	66.6	78.8
w/o HEM	10.8	30.1	45.7
w/o CSM	27.9	50.0	78.5

V. CONCLUSION

This paper proposes a noise-robust pitch detection network, which is based on explicitly modeling harmonics and can estimate pitches with a high resolution. The proposed model outperforms referenced ones like SWIPE, CREPE, HarmoF0, and HGCN+ on the clean PTDU-TUG dataset. Due to the special module (HEM) being set to enhance the harmonic structure from noisy speech, the proposed model is more robust to noise. The performance gap between the proposed model and the referenced counterparts is more significant in noisy

environments. Furthermore, the experimental results show that the model remains high accuracy at a strict evaluation threshold of 10 cents, which mainly owes to the super-resolution module (CSM).

REFERENCES

[1] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "Hgcn: Harmonic gated compensation network for speech enhancement," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 371–375.

[2] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 2494–2498.

[3] D. Wang, L. Deng, Y. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," *Interspeech*, 2021.

[4] M. Reddy and K. Rao, "Excitation modelling using epoch features for statistical parametric speech synthesis," *Computer Speech & Language*, vol. 60, p. 101 029, 2020.

[5] D. Talkin and W. Kleijn, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[6] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[7] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 659–663.

[8] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *JASA*, vol. 124, no. 3, pp. 1638–1652, 2008.

[9] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *TASLP*, vol. 17, no. 4, pp. 614–624, 2009.

[10] W. Chu and A. Alwan, "Safe: A statistical approach to f0 estimation under clean and noisy conditions," *TASLP*, vol. 20, no. 3, pp. 933–944, 2011.

[11] Z. Jin and D. Wang, "Hmm-based multipitch tracking for noisy and reverberant speech," *TASLP*, vol. 19, no. 5, pp. 1091–1102, 2010.

[12] H. Su, H. Zhang, X. Zhang, and G. Gao, "Convolutional neural network for robust pitch determination," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 579–583.

[13] J. Kim, J. Salamon, P. Li, and J. Bello, "Crepe: A convolutional representation for pitch estimation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 161–165.

[14] D. Tran, U. Batricevic, and K. Koishida, "Robust pitch regression with voiced/unvoiced classification in nonstationary noise environments.," in *INTERSPEECH*, 2020, pp. 175–179.

[15] W. Wei, P. Li, Y. Yu, and W. Li, "Harmof0: Logarithmic scale dilated convolution for pitch estimation," *arXiv preprint arXiv:2205.01019*, 2022.

[16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *in Proc. Open Rev. Int. Conf. Learn. Representations, pp. 1–4*, 2016.

[17] T. Wang, W. Zhu, Y. Gao, Y. Chen, J. Feng, and S. Zhang, "Harmonic gated compensation network plus for icassp 2022 dns challenge," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9286–9290.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, Springer, 2015, pp. 234–241.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, PMLR, 2015, pp. 448–456.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015, pp. 1026–1034.

[21] A. Gf, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," in *IDSIA*, 1999.

[22] C. Reddy, V. Gopal, R. Cutler, *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *Interspeech*, pp. 2492–2496, 2020.

[23] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *INTERSPEECH*, 2011.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[25] C. Raffel, B. McFee, J. Humphrey E.and Salamon, *et al.*, "Mir_eval: A transparent implementation of common mir metrics," in *ISMIR*, Citeseer, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Subband Approach to Personal Sound Zone with Joint Optimization of Sound Pressure and Particle Velocity

Yingke Zhao*, Wen Zhang†, and Jingdong Chen†
* Center of Intelligent Acoustics and Immersive Communications, School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
† Center of Intelligent Acoustics and Immersive Communications, School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi, China

***Abstract*—A joint optimization of sound pressure and particle velocity method has been recently proposed for personal sound zone control, which demonstrates promising performance in terms of reproducing perceptually relevant sound field intensity in the bright zone and reducing the sound pressure in the dark zone. However, this joint optimization method requires a long control filter for the fullband control in reverberant room environments. To make the method suitable for real-time applications, we present in this paper a subband approach to the generation of personal sound zones by joint control of the sound pressure and particle velocity. We show that with the subband filtering strategy, the control filter is shortened so that the computational complexity is reduced. Numerical results demonstrate the superiority and effectiveness of the proposed subband method in comparison with the traditional fullband approach.**

***Index Terms*—personal sound zone, subband filtering, sound pressure, particle velocity**

I. Introduction

Generating personal sound zones (PSZ) is one of the most important topics in sound field control. The objective is to utilize a loudspeaker array to generate spatially separated sound zones for different listeners [1]–[10]. The early work in PSZ is based on pressure matching (PM), which attempts to generate the desired sound pressure at the control points [11], [12]. In order to reproduce a desired sound field in a spatial region, the PM approach requires a large number of control points to achieve satisfying performance.

Sound pressure is only a scalar that describes the sound field at a control point. There are some other important parameters that can be utilized to better characterize the desired sound field, e.g., the particle velocity and sound intensity. The application of such information leads to two new PSZ control approaches, i.e., intensity-based reproduction, and joint optimization of pressure and particle velocity. For example, an intensity-based reproduction method was developed in [13] to reproduce the three-dimensional multizone sound field, in which the amplitude of the loudspeaker control filter is obtained by solving an intensity matching optimization problem. An improved version of this method was developed in [14] where a minimum-phase filter was derived by applying the Hilbert transform, but the resulting control filter is not optimal in terms of phase control.

The particle velocity information has also been explored for PSZ to reproduce the directional information of the sound field in the bright zone [6]. A benefit of using the particle velocity information is that the resulting PSZ system can work with irregularly spaced loudspeaker arrays [14]–[16]. In [6], [17], a multizone sound rendering method was proposed based on the joint control of the sound pressure and particle velocity at the control area contour. It demonstrated that by controlling the radial component of the particle velocity can help improve the reproduction performance. In [18], a particle velocity assisted three-dimensional sound field reproduction method in the modal domain was proposed. Recently, a method with joint control of particle velocity and sound pressure was developed [15], which was shown to work with irregular loudspeaker arrays and can achieve similar performance as the aforementioned methods but with less loudspeakers and less control points. However, this joint optimization method requires a long control filter, particularly in reverberant room environments, which makes it challenging to use in real-time applications.

This paper studies the problem of PSZ production. We present a subband approach to the generation of sound zones by joint optimization of the sound pressure and particle velocity. The control filter is achieved by solving a joint optimization problem in the frequency domain at every subband. The frequency-domain control filter is then transformed to the time domain using the inverse Fourier transform. With the subband filtering strategy, the control filter can be much shorter, making the method computationally more efficient than the fullband control method. As a result, this method is better suited to real-time applications.

II. System Model and Problem Formulation

Consider a loudspeaker array with L loudspeakers to generate a bright zone and a dark zone in a room. In the frequency domain, the sound pressure generated by the loudspeaker array

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

at the control point $\mathbf{x}_m$ is written as

$$p(\mathbf{x}_m,\omega)=\sum_{\ell=1}^{L}S(\omega)w_\ell(\omega)H_\ell^{(p)}(\mathbf{x}_m,\omega)$$
$$=S(\omega)\left[\mathbf{h}^{(p)}(\mathbf{x}_m,\omega)\right]^T\mathbf{w}(\omega), \quad (1)$$

where $S(\omega)$ is the source signal, $w_\ell(\omega)$ is the ℓth loudspeaker control filter, $H_\ell^{(p)}(\mathbf{x}_m,\omega)$ denotes the sound pressure acoustic transfer function (ATF) from the ℓth loudspeaker to the control point $\mathbf{x}_m$, $\omega=2\pi f$ is the angular frequency, and vector

$$\mathbf{h}^{(p)}(\mathbf{x}_m,\omega)=\left[H_1^{(p)}(\mathbf{x}_m,\omega)\ \cdots\ H_L^{(p)}(\mathbf{x}_m,\omega)\right]^T \quad (2)$$

consists of all the loudspeaker sound pressure ATFs.

The sound pressure at all the control points is written as

$$\mathbf{p}(\omega)=\left[\mathbf{p}_{\mathrm{B}}^T(\omega)\quad \mathbf{p}_{\mathrm{D}}^T(\omega)\right]^T$$
$$=S(\omega)\mathbf{H}^{(p)}(\mathbf{x})\mathbf{w}(\omega), \quad (3)$$

where the two vectors $\mathbf{p}_{\mathrm{B}}(\omega)$ and $\mathbf{p}_{\mathrm{D}}(\omega)$ denote the reconstructed sound pressure at the control points in the bright and dark zones respectively, the matrix $\mathbf{H}^{(p)}(\omega)$ denotes the sound pressure ATF matrix, and M_{B} and M_{D} are, respectively, the numbers of control points in the bright and dark zones.

The particle velocity reflects the direction information of the sound field. In the Cartesian coordinate system, the particle velocity generated by the loudspeaker array at the control point $\mathbf{x}_m$ is represented as $\mathbf{v}(\mathbf{x}_m,\omega)=[v^{(x)}(\mathbf{x}_m,\omega)\quad v^{(y)}(\mathbf{x}_m,\omega)\quad v^{(z)}(\mathbf{x}_m,\omega)]^T$. The particle velocity is related to the sound pressure through the following form [19]

$$\mathbf{v}(\mathbf{x}_m,\omega)=-\frac{i}{\omega\rho_0 c}\overrightarrow{\bigtriangledown}p(\mathbf{x}_m,\omega), \quad (4)$$

where ρ_0 is the air density, c is the speed of sound, $p(\mathbf{x}_m,\omega)$ is sound pressure at position $\mathbf{x}_m$, and $\overrightarrow{\bigtriangledown}$ denotes gradient. In practice, the components of $\mathbf{v}(\mathbf{x}_m,\omega)$ can be obtained by using the spatial difference to approximate its derivative in (4). For example, along the x axis, one has

$$v^{(x)}(\mathbf{x}_m,\omega)\approx-\frac{i}{\omega\rho_0 c}\frac{p(\mathbf{x}_m,\omega)-p(\mathbf{x}_{m,x},\omega)}{\Delta R}, \quad (5)$$

where $p(\mathbf{x}_{m,x},\omega)$ is the pressure of the microphone near $\mathbf{x}_m$ along the x axis, and $\Delta R=\|\mathbf{x}_m-\mathbf{x}_{m,x}\|_2$ is the distance between the control point and a closely-spaced microphone in the x axis. In this work, we assume that, at every control point, there is a pressure microphone and three additional microphones located along the $x-$, $y-$ and $z-$axis, which are very close to the control point with a distance of ΔR to estimate the particle velocity.

Similar to (1), (5) can been rewritten as

$$v^{(x)}(\mathbf{x}_m,\omega)=\sum_{\ell=1}^{L}S(\omega)w_\ell(\omega)H_\ell^{(v_x)}(\mathbf{x}_m,\omega)$$
$$=S(\omega)\left[\mathbf{h}^{(v_x)}(\mathbf{x}_m,\omega)\right]^T\mathbf{w}(\omega), \quad (6)$$

where

$$H_\ell^{(v_x)}(\mathbf{x}_m,\omega)=-\frac{i}{\omega\rho_0 c}\frac{H_\ell(\mathbf{x}_m,\omega)-H_\ell(\mathbf{x}_{m,x},\omega)}{\Delta R} \quad (7)$$

is the particle velocity ATF in the x axis, and

$$\mathbf{h}^{(v_x)}(\mathbf{x}_m,\omega)=\left[H_1^{(v_x)}(\mathbf{x}_m,\omega)\quad\cdots\quad H_L^{(v_x)}(\mathbf{x}_m,\omega)\right]^T \quad (8)$$

is the control filter vector. The other particle velocity components $v^{(y)}(\mathbf{x}_m,\omega)$ and $v^{(z)}(\mathbf{x}_m,\omega)$ have similar form to (6).

We write the reproduced particle velocity at the bright zone control points in a vector form as

$$\mathbf{v}(\omega)=S(\omega)\mathbf{H}^{(v)}(\omega)\mathbf{w}(\omega), \quad (9)$$

where

$$\mathbf{H}^{(v)}(\omega)=\Big[\mathbf{h}^{(v_x)}(\mathbf{x}_{1_{\mathrm{B}}})\quad\mathbf{h}^{(v_y)}(\mathbf{x}_{1_{\mathrm{B}}})\quad\mathbf{h}^{(v_z)}(\mathbf{x}_{1_{\mathrm{B}}})\quad\cdots$$
$$\mathbf{h}^{(v_x)}(\mathbf{x}_{M_{\mathrm{B}}})\quad\mathbf{h}^{(v_y)}(\mathbf{x}_{M_{\mathrm{B}}})\quad\mathbf{h}^{(v_z)}(\mathbf{x}_{M_{\mathrm{B}}})\Big]^T. \quad (10)$$

The multi-zone reproduction can be achieved in different ways, which are summarized as follows.

A. Intensity-based Reproduction

Let us first review the frequency-domain sound intensity matching method. The reconstructed intensity at a control point $\mathbf{x}_m$ is $\mathbf{I}(\mathbf{x}_m)=\left[I^{(x)}(\mathbf{x}_m)\ I^{(y)}(\mathbf{x}_m)\ I^{(z)}(\mathbf{x}_m)\right]^T$, where the x component of sound intensity is

$$I^{(x)}(\mathbf{x}_m)=\left(\mathbf{I}^{(x)}\right)^T\overline{\mathbf{w}}, \quad (11)$$

where

$$\mathbf{I}^{(x)}=\left[I_1^{(x)}(\mathbf{x}_m)\ \cdots\ I_L^{(x)}(\mathbf{x}_m)\right]^T, \quad (12)$$
$$\overline{\mathbf{w}}=\left[|\omega_1|^2\quad\cdots\quad|\omega_L|^2\right]^T. \quad (13)$$

We have

$$I_\ell^{(x)}(\mathbf{x}_m)=p_\ell^*(\mathbf{x}_m)v_\ell^{(x)}(\mathbf{x}_m)$$
$$=|S(\omega)|^2\left[H_\ell^{(p)}(\mathbf{x}_m)\right]^*H_\ell^{(v_x)}(\mathbf{x}_m). \quad (14)$$

The intensity can be written in a matrix form as

$$\bar{\mathbf{I}}=\Big[\mathcal{R}\left\{\mathbf{I}^{(x)}(\mathbf{x}_1)\right\}\quad\mathcal{I}\left\{\mathbf{I}^{(x)}(\mathbf{x}_1)\right\}\quad\mathcal{R}\left\{\mathbf{I}^{(y)}(\mathbf{x}_1)\right\}\quad\cdots$$
$$\mathcal{I}\left\{\mathbf{I}^{(y)}(\mathbf{x}_M)\right\}\quad\mathcal{R}\left\{\mathbf{I}^{(z)}(\mathbf{x}_M)\right\}\quad\mathcal{I}\left\{\mathbf{I}^{(z)}(\mathbf{x}_M)\right\}\Big]^T, \quad (15)$$

where $\mathcal{I}\{\cdot\}$ and $\mathcal{R}\{\cdot\}$ denote the imaginary and real parts of a complex vector, respectively, and $M=M_{\mathrm{B}}+M_{\mathrm{D}}$ is the total number of control points within the bright and dark zones. The desired sound intensity at all the control points can also be represented in a vector form as

$$\bar{\mathbf{I}}_{\mathrm{d}}=\Big[\mathcal{R}\left\{I^{(x)}(\mathbf{x}_1)\right\}\quad\mathcal{I}\left\{I^{(x)}(\mathbf{x}_1)\right\}\quad\cdots$$
$$\mathcal{R}\left\{I^{(z)}(\mathbf{x}_M)\right\}\quad\mathcal{I}\left\{I^{(z)}(\mathbf{x}_M)\right\}\Big]^T. \quad (16)$$

Note that here the desired intensity in the dark zone is set to 0 in (16). The amplitude of the control filter can be obtained by minimizing the mean square error between the reconstructed intensity and the desired intensity, which is formulated as

$$\min \|\bar{\mathbf{I}}\overline{\mathbf{w}} - \bar{\mathbf{I}}_{\mathrm{d}}\|^2 \;\; \text{s.t.} \;\; \overline{\mathbf{w}} \geq 0, \tag{17}$$

where the constraint ensures that the elements in the vector $\overline{\mathbf{w}}$ are non-negative. The problem in (17) is a non-negative least squares (NNLS) problem, which can be solved by the Lawson and Hansons method [20]. To make sure that the PSZ control system is causal and stable, a minimum-phase filter is adopted by utilizing the Hilbert transform on the calculated amplitude [14]. As will become clear in Section V, the intensity matching approach is not able to achieve the optimal phase control, leading to insufficient acoustic contrast (AC) between the bright and dark zones.

B. Joint Control of Sound Pressure and Particle Velocity

The desired sound pressure at all the control points is written in a vector form as

$$\mathbf{p}_{\mathrm{d}}(\omega) = [p_{\mathrm{d}}(\mathbf{x}_{1_{\mathrm{B}}}, \omega) \quad \cdots \quad p_{\mathrm{d}}(\mathbf{x}_{M_{\mathrm{B}}}, \omega) \quad \mathbf{0}_{M_{\mathrm{D}}}]^T. \tag{18}$$

The desired particle velocity at the bright-zone control points is also written in a vector form as

$$\mathbf{v}_{\mathrm{d}}(\omega) = \left[\mathbf{v}_{\mathrm{d}}^T(\mathbf{x}_{1_{\mathrm{B}}}, \omega) \quad \mathbf{v}_{\mathrm{d}}^T(\mathbf{x}_{2_{\mathrm{B}}}, \omega) \;\cdots\; \mathbf{v}_{\mathrm{d}}^T(\mathbf{x}_{M_{\mathrm{B}}}, \omega)\right]^T. \tag{19}$$

A frequency-domain method for joint optimization of sound pressure and particle velocity to generate PSZs was developed in [6]. The principle is to reproduce the desired particle velocity and sound pressure within the bright zone and meanwhile keep the dark zone sound pressure as small as possible. The loudspeaker control filter is obtained by solving the following optimization problem:

$$\min_{\mathbf{w}(\omega)} \tau\|\mathbf{p}(\omega) - \mathbf{p}_{\mathrm{d}}(\omega)\|^2 + (1-\tau)\rho(\omega)\|\mathbf{v}(\omega) - \mathbf{v}_{\mathrm{d}}(\omega)\|^2. \tag{20}$$

Let us introduce a frequency dependent normalization term, i.e., $\rho(\omega) = |\omega\rho_0 c\Delta R|^2$, to the particle velocity term in (20). Substituting (3) and (9) into (20) gives the following equivalent optimization problem:

$$\begin{aligned}\min_{\mathbf{w}(\omega)} &\tau\|\mathbf{H}^{(p)}(\omega)\mathbf{w}(\omega) - \mathbf{h}_{\mathrm{d}}^{(p)}(\omega)\|^2 + \\ &(1-\tau)\rho(\omega)\|\mathbf{H}^{(v)}(\omega)\mathbf{w}(\omega) - \mathbf{h}_{\mathrm{d}}^{(v)}(\omega)\|^2,\end{aligned} \tag{21}$$

whose solution is

$$\begin{aligned}&\mathbf{w}(\omega) = \\ &\left\{\tau\left[\mathbf{H}^{(p)}\right]^H \mathbf{H}^{(p)} + (1-\tau)\rho(\omega)\left[\mathbf{H}^{(v)}\right]^H \mathbf{H}^{(v)}\right\}^{-1} \\ &\left\{\tau\left[\mathbf{H}^{(p)}\right]^H \mathbf{h}_{\mathrm{d}}^{(p)} + (1-\tau)\rho(\omega)\left[\mathbf{H}^{(v)}\right]^H \mathbf{h}_{\mathrm{d}}^{(v)}\right\}.\end{aligned} \tag{22}$$

Note that when $\tau = 1$, the filter in (22) becomes the conventional PM method.

The aforementioned joint control method was evaluated in free field environments [6]. When used in real reverberant environments, particularly when reverberation is strong, the control filter has to be long, making the algorithm computationally very expensive. To reduce complexity, we introduce a subband filtering strategy to reduce the length of the control filters in the next section.

III. Subband design for PSZ systems

A. Subband Filtering

We use the quadrature mirror filter (QMF) to fulfil the subband filtering structure [21]. QMF provides a two-band division and achieves an alias-free reconstruction of signals. The QMF subband division method consists of two steps: 1) the analysis step, where the original signal is filtered with the low pass analysis filter f_0 (half-band FIR filter) and the filtered signal is then decimated by two; and 2) the reconstructive step, in which the down-sampled signal is first up-sampled then filtered with the reconstructive filter g_0. The higher frequency band shares a similar process except that the analysis filter is a high-pass filter f_1, and the synthesis filter is g_1. The reconstructed signal is obtained by adding up the synthesized subband signals [21]. Multiband can be obtained by further applying the two-band analysis and synthesis to each subband in a cascaded way.

B. Subband PSZ Method

In the subband method, the ATFs are first converted to subband ATFs. The loudspeaker control filter in the ith subband is then obtained by applying the PSZ control method within this subband where the ith subband loudspeaker control filter is obtained as follows

$$\begin{aligned}\mathbf{w}_i(\omega) = &\left\{\tau\left[\mathbf{H}_i^{(p)}\right]^H \mathbf{H}_i^{(p)} + (1-\tau)\rho\left[\mathbf{H}_i^{(v)}\right]^H \mathbf{H}_i^{(v)}\right\}^{-1} \cdot \\ &\left\{\tau\left[\mathbf{H}_i^{(p)}\right]^H \mathbf{h}_{\mathrm{d},i}^{(p)} + (1-\tau)\rho\left[\overline{\mathbf{H}}_i^{(v)}\right]^H \mathbf{h}_{\mathrm{d},i}^{(v)}\right\},\end{aligned} \tag{23}$$

where $\mathbf{H}_i^{(p)}$, $\mathbf{H}_i^{(v)}$, $\mathbf{h}_{\mathrm{d},i}^{(p)}$ and $\mathbf{h}_{\mathrm{d},i}^{(v)}$ are subband ATFs. To generate the loudspeaker driving signal, the frequency-domain subband control filter $\mathbf{w}_i(\omega)$ is transformed to the time domain using the inverse Fourier transform, thereby giving $\mathbf{w}_i(t)$. The time-domain filter can be truncated to further reduce the filter length. With the subband source audio signal being $\mathbf{s}_i(t)$, the subband loudspeaker driving signal is obtained as $\mathbf{y}_i(t) = \mathbf{s}_i(t) * \mathbf{w}_i(t)$. An example of the subband PSZ control method with 4 subbands is illustrated in Fig. 1.

IV. Performance Measures

In order to evaluate the performance of different PSZ methods, we use acoustic contrast (AC) and reproduced intensity error as performance measures. AC is defined as the ratio of the reproduced signal power at the bright zone to that at the

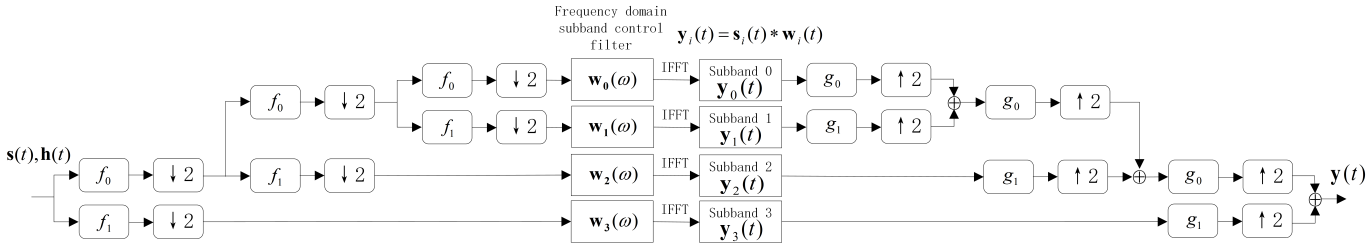


Fig. 1. Illustration of a subband PSZ method.

dark zone, i.e.,

$$\mathrm{AC}(\omega)=\frac{\mathbf{w}^H(\omega)\left[\mathbf{H}_{\mathrm{B}}^{(p)}(\omega)\right]^H\mathbf{H}_{\mathrm{B}}^{(p)}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^H(\omega)\left[\mathbf{H}_{\mathrm{D}}^{(p)}(\omega)\right]^H\mathbf{H}_{\mathrm{D}}^{(p)}(\omega)\mathbf{w}(\omega)}. \tag{24}$$

The normalized mean square error (NMSE) between the reproduced sound intensity and the desired sound intensity in the bright zone is defined as

$$\epsilon(\omega)=\frac{\sum_{m=1}^{M_{\mathrm{B}}}\|\mathbf{I}(\mathbf{x}_m)-\mathbf{I}_{\mathrm{d}}(\mathbf{x}_m)\|^2}{\sum_{m=1}^{M_{\mathrm{B}}}\|\mathbf{I}_{\mathrm{d}}(\mathbf{x}_m)\|^2}, \tag{25}$$

where $\mathbf{I}(\mathbf{x}_m)=p^*(\mathbf{x}_m,\omega)\mathbf{v}(\mathbf{x}_m,\omega)$ and $\mathbf{I}_{\mathrm{d}}(\mathbf{x}_m)=p_{\mathrm{d}}^*(\mathbf{x}_m,\omega)\mathbf{v}_{\mathrm{d}}(\mathbf{x}_m,\omega)$ are, respectively, the reproduced sound intensity and desired sound intensity.

V. Simulation Results

The performance of different PSZ methods are evaluated through numerical simulations. An irregular circle array of 30 loudspeakers are placed in a square room of size $6\ \mathrm{m}\times 6\ \mathrm{m}\times 3\ \mathrm{m}$. The loudspeakers and microphone control points are at the same horizontal plane with a height of $1.5\mathrm{m}$ above the floor. There is a dark zone and a bright zone inside the circular array, and in each zone, we placed 8 control points on the contour of the control area. For each control point, there are extra microphones along the x-, y- and z-axis to obtain the particle velocity estimates. These microphones are $0.01\ \mathrm{m}$ away from the control point. The position of the loudspeakers and microphones are illustrated in Fig. 2. We apply the image source model to simulate the ATFs [22]. The sampling frequency is $8\ \mathrm{kHz}$. The signals are separated into four subbands with the cutoff frequencies being, respectively, $0.5\ \mathrm{kHz}$, $1\ \mathrm{kHz}$, and $2\ \mathrm{kHz}$. For comparison, the reproduction performance based on pressure matching, intensity matching, and the corresponding full-band control are also examined.

In the first simulation, we compare the performance of the joint control PSZ method with the PM and intensity matching method. In order to apply the intensity matching for multizone sound field reproduction, we set the desired sound intensity in the dark zone as zero. A room environment with reverberation time $T_{60}\approx 130\ \mathrm{ms}$ is simulated. From the results shown in Fig. 3, one can see that the joint control method has the smallest reproduced intensity error as compared to both the PM ($\tau=1$) and intensity matching methods at both the control and uncontrolled points. The intensity matching method can obtain acceptable reproduction error at the control point in the bright zone, but almost no AC difference between the bright and dark zones. The underlying reason is that with the intensity matching method, the phase of each control filter is fixed (i.e., a minimum phase as a Hilbert transform of the control filter amplitude), so the required phase control for AC is not guaranteed. The intensity matching method also fails to fulfill the control at the uncontrolled points.

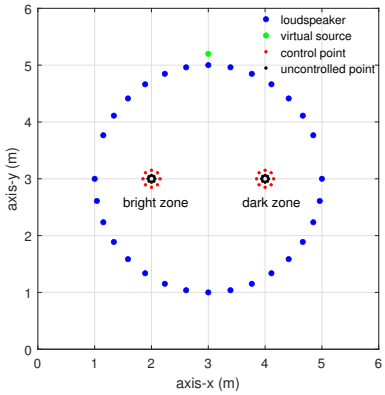


Fig. 2. The simulation setup.

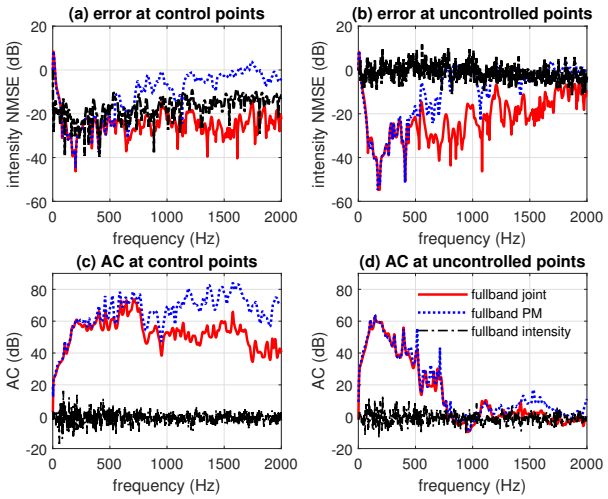


Fig. 3. Performance of the PM method, intensity matching method, and the joint controlled method with $\tau=0.6$.

In the second simulation, the performance of the subband joint control method is studied and compared to the original full-band control method. The results in Fig. 4 show that the subband processing leads to slightly degraded performance compared to its full-band counterpart. However, the subband method has a much shorter filter length. The full-band method requires a filter length of 1800 to obtain good PSZ control

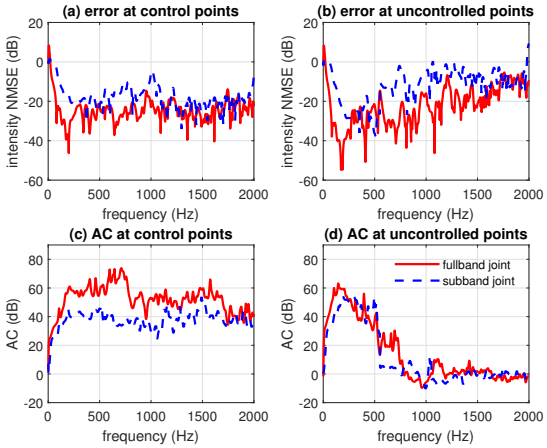


Fig. 4. Performance of the joint control method ($\tau = 0.6$) using fullband and subband filtering strategy.

TABLE I
SUBBAND FILTER LENGTH.

subband (kHz)	0-0.5	0.5-1	1-2	2-4
filter length (samples)	180	180	350	800

performance. In this case, the computational complexity of the full-band method is $(1800)^3$. The filter length of each subband filter is shown in Table. 1. We further shorten the subband filter length by cutting off the points in the tail of the filter which is close to zero. One can see that the subband filter has a much shorter filter length. With the filter length being illustrated in Table 1, the computational complexity of the subband method is $(800)^3 + (350)^3 + 2(180)^3$ which is much less than the full-band method. The less computational complexity of subband PSZ method makes it more appropriate to implement in real applications.

VI. CONCLUSIONS

This paper studied the problem of generating PSZ based on optimization of sound pressure and particle velocity. We first reviewed two state-of-the-art methods, i.e., the frequency-domain intensity matching method and the method with pressure and particle velocity being jointly controlled. Both methods can work with irregular loudspeaker array setups. To reduce the complexity so the joint optimization method can be implemented and applied in reverberant environments, we proposed a subband PSZ control method based on the joint optimization of sound pressure and particle velocity. Simulations show that the intensity matching method, which has no control of reproduced phase, cannot achieve the desired AC. In comparison, the proposed subband method can achieve the desired performance and meanwhile is computationally more efficient since it has a much shorter filter length.

REFERENCES

[1] L. Shi, G. Ping, X. Shen, and M. G. Christensen, "Generation of personal sound fields in reverberant environments using interframe correlation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 1066–1070.

[2] V. Molés-Cases, G. Piñero, M. de Diego, and A. Gonzalez, "Personal sound zones by subband filtering and time domain optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2684–2696, 2020.

[3] L. Vindrola, M. Melon, J.-C. Chamard, and B. Gazengel, "Use of the filtered-x least-mean-squares algorithm to adapt personal sound zones in a car cabin," *J. Acoust. Soc. Amer.*, vol. 150, no. 3, pp. 1779–1793, 2021.

[4] V. Molés-Cases, S. J. Elliott, J. Cheer, G. Piñero, and A. Gonzalez, "Weighted pressure matching with windowed targets for personal sound zones," *J. Acoust. Soc. Amer.*, vol. 151, no. 1, pp. 334–345, 2022.

[5] T. Lee, L. Shi, J. K. Nielsen, and M. G. Christensen, "Fast generation of sound zones using variable span trade-off filters in the DFT-domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 363–378, 2020.

[6] M. Buerger, C. Hofmann, and W. Kellermann, "Broadband multizone sound rendering by jointly optimizing the sound pressure and particle velocity," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 1477–1490, 2018.

[7] Y. Cai, M. Wu, L. Liu, and J. Yang, "Time-domain acoustic contrast control design with response differential constraint in personal audio systems," *J. Acoust. Soc. Amer.*, vol. 135, no. 6, pp. 252–257, 2014.

[8] M. Zhu and S. Zhao, "An iterative approach to optimize loudspeaker placement for multi-zone sound field reproduction," *J. Acoust. Soc. Amer.*, vol. 149, no. 5, pp. 3462–3468, 2021.

[9] J. Zhang, L. Shi, M. G. Christensen, W. Zhang, L. Zhang, and J. Chen, "Robust pressure matching with ATF perturbation constraints for sound field control," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 8712–8716.

[10] T. Lee, J. K. Nielsen, and M. G. Christensen, "Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2412–2426, 2020.

[11] F. Olivieri, F. M. Fazi, S. Fontana, D. Menzies, and P. A. Nelson, "Generation of private sound with a circular loudspeaker array and the weighted pressure matching method," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1579–1591, 2017.

[12] T. Betlehem and P. D. Teal, "A constrained optimization approach for multi-zone surround sound," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 437–440.

[13] H. Zuo, T. D. Abhayapala, and P. N. Samarasinghe, "3D multizone soundfield reproduction in a reverberant environment using intensity matching method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 416–420.

[14] H. Zuo, P. N. Samarasinghe, and T. D. Abhayapala, "Intensity based spatial soundfield reproduction using an irregular loudspeaker array," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1356–1369, 2020.

[15] X. Hu, J. Wang, W. Zhang, and L. Zhang, "Time-domain sound field reproduction with pressure and particle velocity jointly controlled," *Appl. Sci.*, vol. 11, no. 22, p. 10880, 2021.

[16] H. Zuo, P. N. Samarasinghe, and T. D. Abhayapala, "Intensity based soundfield reproduction over multiple sweet spots using an irregular loudspeaker array," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 486–490.

[17] M. Buerger, R. Maas, H. W. Löllmann, and W. Kellermann, "Multizone sound field synthesis based on the joint optimization of the sound pressure and particle velocity vector on closed contours," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2015, pp. 1–5.

[18] H. Zuo, T. D. Abhayapala, and P. N. Samarasinghe, "Particle velocity assisted three dimensional sound field reproduction using a modal-domain approach," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2119–2133, 2020.

[19] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[20] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 6, pp. 672–680, 1980.

[21] H. So and J.-W. Choi, "Subband optimization and filtering technique for practical personal audio systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 8494–8498.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

An Multi-evidence Fusion Based on C-Distance with Uncertain Reasoning for Classification

Cuiping Cheng, Pengcheng Yue and Taihao Li*
Artificial Intelligence Research Institute, Zhejiang Laboratory, Hangzhou, China
E-mail: chengcuipingswu@163.com, pcyue@zhejianglab.com, lith@zhejianglab.com

***Abstract*—Dempster-Shafer (D-S) evidence theory has been extensively used to handle uncertain information. However, evidence fusion remains a challenging problem because counter-intuitive results would occur when combining highly conflicting evidence. To address this issue, we propose a weighted fusion model with multi-evidence uncertainty reasoning based on a novel similarity quantification function. Specifically, we first define a dissimilarity representation function C-distance. This function considers two factors, the angle and the norm of the evidence vector, so that the similarity between evidences can be described more comprehensively. In particular, unlike previous weights defined as the ratio between distances, we use a nonlinear function to fit the weight and the C-distance, which expands the learning space of weights. Based on the above work, we propose a weighted fusion framework. Experiments on real datasets illustrate the excellent performance of the proposed method in information fusion.**

I. Introduction

With the rapid development of science and technology, the types of information are gradually diversified. Information from different sources often helps to comprehensively describe the nature of things, hence, how to integrate multi-source information has attracted much attention [1], [2]. The linear combination formulation works only when there is no high degree of conflict between all tasks, however, such cases are rare. Multiple-evidence learning with conflicts requires a trade-off of weights between tasks that goes beyond what is achieved by linear combinations. Many fusion models have been constructed, such as Dempster-Shafer evidence theory [3], [4] and evidential reasoning [5].

D-S evidence theory [6], [7] has been widely used in many fields such as decision making [8], target recognition and so on [9], [10], [11], [12]. However, counterintuitive results may obtained when using evidence theory to combine highly conflicting evidence [13]. In order to solve this problem, an effective method is to determine the contribution of different evidences in the combined operation by measuring the conflict between the evidences[14]. Many methods to measure the similarity between evidences have been proposed to model the conflict degree [15], [16], [17], among which, the Jousselme $et\ al.$'s distance is relatively simple to calculate and the geometric meaning is clear, it has been widely used to manage conflicts of evidence. Although effective, Jousselme $et\ al.$ diatence is a global similarity measure that cannot capture local differences, which degrades the performance of classification tasks.

To cope with evidence uncertainty and fusion conflicts, we present a novel distance measure to quantify the uncertainty of the belief distribution and the similarity of two evidences, then use the nonlinear constraint relationship to learn the optimal distribution of weights, and finally use the weighted discounted classification results for D-S combination to make the final decision making. In specifically, we first define the C-distance, which estimates the angle and norm difference between evidence vectors, so that it can not only learn local belief structures more sensitively, but also effectively discern non-heterogeneous evidence. Compared with linear weights, dynamic weights [18] can flexibly adjust the sensitivity of conflict, thereby increasing the robustness of the fusion framework and providing decision makers with more space for expression of preferences. We then dynamically learn the optimal weights based on the proposed similarity function. Finally, we design a multi-evidence fusion framework to deal with classification tasks. Specifically, the obtained weight matrix will guide the fusion process to pay more attention to the single evidence with high reliability. We evaluate our model on the challenging dataset, and the empirical results demonstrate the effectiveness of our method. In addition, this paradigm is generalizable and can be extended to more tasks, which is of great significance for improving the performance of fusion decision-making.

In summary, our contributions in this paper are summarized as follows:

- A new similarity measure function is proposed to quantify the angle and modulus length differences of evidence vectors. The proposed measure takes into account the uncertainty of belief distribution and the inclusion relationship between elements, which allows the measure able to handle data where the conflicts between evidences are not significant enough.
- This paper designs a weighted fusion strategy based on C-distance and dynamic weight generation, which gives reliable evidence a higher weight to deal with high-conflict scenarios.
- We conduct extensive experiments on iris dataset and sentiment analysis datasets, the results validate the robustness and effectiveness of our method on the classification task.

II. Related Work

A. Evidence theory

A substantial number of work has been done to deal with the conflict issue from evidence dissimilarity perspectives. One of the classic schemes, Deng $et\ al.$ [19] provide an improved

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

combination averaging method of confidence function based on evidence distance measurement. This method first uses distance to construct a similarity matrix, and then determines the importance of the evidence according to the credibility degree. Fuyuan Xiao [20] presented a new reinforced belief divergence measure in D-S evidence theory for data fusion. The proposed divergence measurement is the first such measure that takes into account the correlations between mass functions and subsets of the sets of belief functions. The credibility of each piece of evidence is obtained by modeling the reinforced belief divergence measure matrix, which is treated as the final weight for each evidence. Kaiyi Zhao $et\ al.$ [21] developed an improved evidence fusion algorithm in multi-sensor systems. First, a new distribution distance measurement method is proposed to obtain the credibility of the evidence. Next, an improved calculating method was introduced to generate the amount of information of the evidence. Finally, modify the credibility of each evidence according to the amount of information to determine the weight of the evidence. There are also many studies focus on the use of entropy and distance to combine conflicting evidence. Chaosheng Zhu $et\ al.$ [14] defined a Hellinger distance for evidence theory. Based on the presented distance, a multisource information fusion method is devised, the proposed method takes both distance between evidences and the uncertainty of the evidences into account. In the proposed belief fusion method, the newly defined distance is used as a tool to measure the distance between two sources of evidence, and the information volume weight of each piece of evidence is calculated by belief entropy.

B. Dynamic weight generation

In [18], an efficient dynamic weight generation scheme was proposed. Given n pieces of evidence, let m_i and m_j be two BPAs on the same frame of discernment Θ.

Step 1: Calculate the dissimilarity representation between all the evidences:

Let

$$d_{sum}(m_i) = \sum_{j=1, j\neq i}^{N} D_C(m_i, m_j) \tag{1}$$

Step 2: Let ω_i be the weight of evidence m_i, the weight ω_i is obtained by minimizing the objective function:

$$minimize \sum_{i=1}^{N} \omega_i^{\tau} d_{sum}(m_i), \tag{2}$$

where $\tau > 1$ is the elasticity coefficient. In order to achieve the result of assigning significant dissimilar evidence with a small weight, optimization is performed by minimizing the objective function. Here, the optimal weight w_i is given by the following equation:

$$argmin_{w_i} \sum_{i=1}^{N} \omega_i^{\tau}(m_i) d_{sum}(m_i) \tag{3}$$

Then, we use the Lagrangian multiplier method to solve this extreme value problem, the Eq.3 has a generalized Lagrangian:

$$\zeta(d_{sum}(m_i), \omega_i(m_i)) = \sum_{i=1}^{N} \omega_i^{\tau}(m_i) d_{sum}(m_i) + \lambda(\sum_{i=1}^{N} \omega_i(m_i) - 1)$$

$$s.t. \sum_{i=1}^{N} \omega_i(m_i) = 1$$

III. The proposed method

A. C-distance

Let us consider an example, m_1, m_2 and m_3 are three BPAs in the $\Theta = \{A, B, C\}$.

$$\begin{aligned} &m_1(A) = 0.5, m_1(B) = 0.2, m_1(C) = 0.3; \\ &m_2(A) = 0.7, m_2(B) = 0.1, m_2(C) = 0.2; \\ &m_3(A) = 0.3, m_3(B) = 0.3, m_3(C) = 0.4. \end{aligned}$$

According to the Jousselme's distance, we have $d_J(m_1, m_2) = d_J(m_3, m_2)$. In this example, m_1 and m_2 have a consistent preference, both support the element A. However, Jousselme's distance fails to capture this information. It is worth noting that, according to information entropy, we can know evidence m_2 has the lowest uncertainty, while m_3 is the most uncertain. More importantly, the norm between the three vectors satisfies the following relationship: $|m_2| > |m_1| > |m_3|$. We therefore propose the following theorem:

Theorem 3.1: The norm of the vector is inversely proportional to its information uncertainty.

Since the Euclidean distance quantifies the global difference, the direction information of the vector is ignored. Especially for high-dimensional vectors, the sparsity of the data is more likely to damage the sensitivity As shown in Fig.1, it can be observed the angle $\theta(\vec{oa}, \vec{ob}) > \theta(\vec{oe}, \vec{of})$ according to Euclidean distance, we have $d_E(\vec{oa}, \vec{ob}) = d_E(\vec{oe}, \vec{of})$. This shows that two sets of vectors with equal Euclidean distances do not mean that their element distributions are also the same. It is necessary to combine the factor of the difference in angle to improve the sensitivity. If the angles and norms of the two sets of vectors are equal, it can be shown that the distance between the two vectors is the same.

In fact, the distance function is a general function of the distribution of random variables, which does not depend on the actual value of the random variable, but only on its belief distribution. Therefore, if the direction difference and value difference can be considered at the same time, the measurement of information similarity will be more reliable. Here, we propose a new conflict measure called C-distance, which overlaps with Euclidean distance and $et\ al.$'s distance, but also extends cosine distance.

Definition 3.1: Given two BPAs $\mathbf{P} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{Q} = (q_1, q_2, \ldots, q_n)$, The C-distance between $\mathbf{P}$ and $\mathbf{Q}$ is

defined as:

$$D_C(\mathbf{P},\mathbf{Q}) = (1 - \frac{\sum_{i=1}^{i=n} p_i \times q_i}{|\mathbf{PS}^P| \times |\mathbf{QS}^Q|})|\mathbf{P}| \times |\mathbf{Q}| \tag{4}$$

where $\mathbf{S}$ is an $1 \times n$ matrix whose elements are

$$\mathbf{S}^P = (\frac{1}{|p_1|}, \frac{1}{|p_2|}, \cdots, \frac{1}{|p_n|})^T$$

where $\mathbf{S}$ is the likelihood ratio matrix whose elements measure the similarity between classified elements, matrix $\mathbf{S}$ reflects the association between each element in the vector, which plays the role of changing the angle between the vectors. Intuitively, the C-distance compares the distance between two vectors in the decision space whose angle is less than or equal to 90 degrees. this measure puts improved robustness and improved similarity conflicts under a common umbrella. This normalization provides effective guidance on which distance metric to use for a given case, depending on the details of the decision task.

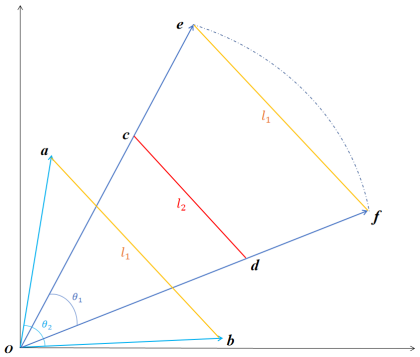


Fig. 1.

B. Evidence weighted fusion

To address the problem of belief theory failing when combined with highly contradictory evidence, in this section, we design a weighted fusion framework, from the aspect of evidence preprocessing. The architecture of presented method is as shown in Fig.2. The framework consists of three modules: formulation of evidence, dynamic weight generation and fusion of weighted beliefs. Our approach provides a new paradigm for fusion learning by combining dynamic evidence outputs from different sources at the decision level.

After obtaining dynamic weights w_i, a weighted average is performed on all the belief functions:

$$\tilde{m}(A) = \sum_i w_i \cdot m_i \tag{5}$$

Suppose two evidences m_1 and m_2, which can be combined into a new evidential body m through Dempster combination rule

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - k} \tag{6}$$

TABLE I
THE BPAS OF IRIS SAMPLE (ORIGINAL).

BPA	Attribute			
	$SL(m_1)$	$SW(m_2)$	$PL(m_3)$	$PW(m_4)$
$m(SE)$	0.3337	0.3164	0.6699	0.6996
$m(VC)$	0.3165	0.2501	0.2374	0.2120
$m(VI)$	0.2816	0.2732	0.0884	0.0658
$m(SE,VC)$	0.0307	0.0304	0.0000	0.0000
$m(SE,VI)$	0.0052	0.0481	0.0000	0.0000
$m(VC,VI)$	0.0272	0.0515	0.0043	0.0226
$m(SE,VC,VI)$	0.0052	0.0304	0.0000	0.0000

TABLE II
THE BPAS OF IRIS SAMPLE WITH NOISE (NOISY).

BPA	Attribute			
	$SL(m_1)$	$SW(m_2)$	$PL(m_3)$	$PW(m_4)$
$m(SE)$	0.3337	0.00	0.6699	0.6996
$m(VC)$	0.3165	0.99	0.2374	0.2120
$m(VI)$	0.2816	0.01	0.0884	0.0658
$m(SE,VC)$	0.0307	0.00	0.0000	0.0000
$m(SE,VI)$	0.0052	0.00	0.0000	0.0000
$m(VC,VI)$	0.0272	0.00	0.0043	0.0226
$m(SE,VC,VI)$	0.0052	0.00	0.0000	0.0000

with $k = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$. Here, k represents the degree of conflict between evidences and also called normalization constant.

Finally, fuse the evidence $\tilde{m}(A)$ $N-1$ times through Dempster's combination rule:

$$m_f(A) = \tilde{m}(A) \oplus \tilde{m}(A) \oplus \cdots \oplus \tilde{m}(A) \tag{7}$$

The output $m_f(A)$ is used for decision making.

IV. EXPERIMENTS

In the following, let us study the properties of our model through numerical examples.

A. Iris dataset

The initial Iris data set comes from the UCI repository of the database. For the credibility and fairness of comparison, we quote the BPAs in [22] for experiments. In the Iris dataset, there are three kinds of Iris flower Setosa, Versicolor and Virginica. and each type of iris has the following four attributes $SepalLength(SL)$, $SepalWidth(SW)$, $PetalLength(PL)$ and $PetalWidth(PW)$. Therefore, the frame of discernment is denoted as $\Psi = \{SE, VC, VI\}$, the BPAs generated based on the Iris dataset is shown in Tab.I. For the purpose of verifying and comparing the performance of models in handling conflicts of evidence, the test data set obtained by adding noise samples is given as Tab.II.

The distance and weight results on the two datasets, as shown in Tab.III, shows the feasibility of our distance measure and weight generation strategy. Our distance is more discriminative for conflicts between conflicting evidence, which contributes to the reliability of weights generated based on distance.

From Tab.IV, for BPA that does not have high conflicts, all methods including our model can assign the highest reliability to the correct sample after fusion. Since our dissimilarity

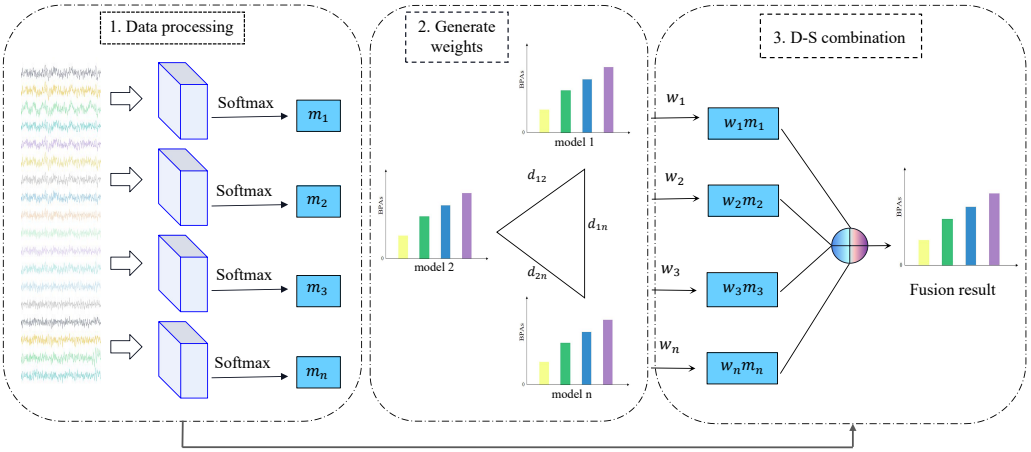


Fig. 2. Overview of our proposed framework.

TABLE IV
THE ORIGINAL IRIS FUSION RESULTS

Measures	$m(SE)$	$m(VC)$	$m(VI)$	$m(SE,VC)$	$m(SE,VI)$	$m(VC,VI)$	$m(SE,VC,VI)$
Dempster	0.8940	0.0933	0.0127	0.00	0.00	0.00	0.00
Murphy's method [23]	0.8854	0.0880	0.0266	0.00	0.00	0.00	0.00
Deng's	0.8803	0.0911	0.0286	0.00	0.00	0.00	0.00
Zhu's	0.7189	0.1726	0.1076	0.00	0.00	0.00	0.00
Proposed	**0.9400**	0.0480	0.0120	0.00	0.00	0.00	0.00

TABLE V
THE FUSION RESULTS OF NOISY DATASET

Measures	Step	$m(SE)$	$m(VC)$	$m(VI)$	$m(SE,VC)$	$m(SE,VI)$	$m(VC,VI)$	$m(SE,VC,VI)$
Dempster	(1, 2)	0.0000	0.9916	0.0084	0.00	0.00	0.00	0.00
	(1, 2, 3)	0.0000	0.9968	0.0032	0.00	0.00	0.00	0.00
	(1, 2, 3, 4)	0.0000	0.9988	0.0012	0.00	0.00	0.00	0.00
Murphy's method [23]	(1, 2)	0.0655	0.8828	0.0505	6×10^{-6}	4×10^{-5}	5×10^{-4}	5×10^{-4}
	(1, 2, 3)	0.2112	0.7749	0.0139	8×10^{-6}	2×10^{-7}	9×10^{-6}	3×10^{-8}
	(1, 2, 3, 4)	0.4422	0.5546	0.0032	8×10^{-8}	5×10^{-10}	6×10^{-7}	3×10^{-11}
Deng's	(1, 2)	0.0655	0.8828	0.0505	6×10^{-4}	4×10^{-5}	5×10^{-4}	1×10^{-5}
	(1, 2, 3)	0.3219	0.6534	0.0247	2×10^{-5}	4×10^{-7}	2×10^{-5}	5×10^{-8}
	(1, 2, 3, 4)	0.7301	0.2652	0.0047	1×10^{-7}	7×10^{-10}	9×10^{-7}	5×10^{-11}
Zhu's	(1, 2)	0.2740	0.5210	0.2003	0.0025	2×10^{-4}	0.0020	5×10^{-5}
	(1, 2, 3)	0.5353	0.3717	0.0928	9×10^{-5}	2×10^{-6}	8×10^{-5}	3×10^{-7}
	(1, 2, 3, 4)	0.8255	0.1556	0.0189	1×10^{-6}	7×10^{-9}	3×10^{-6}	5×10^{-10}
Proposed's	(1, 2)	0.0655	0.8828	0.0505	6×10^{-4}	4×10^{-5}	5×10^{-4}	1×10^{-5}
	(1, 2, 3)	0.3545	0.6095	0.0359	3×10^{-5}	6×10^{-7}	3×10^{-5}	9×10^{-8}
	(1, 2, 3, 4)	**0.8780**	0.1120	0.010	0	0	0	0

measurement is more comprehensive, the difference between m_2 and other BPAs is deepened by supplementing preference information. In this example, the reason why we set $\tau = 1.5$ is that when there are highly conflicting data, let $\tau < 2$, it will magnify the difference, the difference would be amplified and the weight of unreliable evidence would be reduced. The comparison results of our strategy and other methods in the noisy dataset are shown in Tab.V. It can be seen that the convergence speed of the proposed method is the fastest. The results obtained by the weighted fusion method are generally better than the Dempster fusion rule, and the weight tendency determined by the proposed method is more obvious. After combining the four pieces of evidence, the presented framework is superior to other methods.

B. EMOCAP dataset

1) Experimental setting: Recent advances in deep neural networks (DNNs) have improved various machine learning problem, but how to effectively fuse massive labeled training data is still the key to improving the accuracy of DNNs [24]. It is worth noting that due to the internal architectures of neural networks are different, for the same data set, the predictive ability of the models learned by different networks is not consistent. Therefore, it is unreasonable to share unified weight parameters in multi-model fusion decision-making. Accordingly, this allows us to conduct multi-model fusion experiment to

TABLE III
DISTANCE VALUES AND WEIGHT DISTRIBUTIONS OF DIFFERENT DISTANCES IN TWO CASES. ω_1 REPRESENTS THE WEIGHT OBTAINED BY THE DENG'S METHOD, ω_2 REPRESENTS THE WEIGHT OBTAINED BY OUR APPROACH

Dissimilarity	$d(SL)$	$d(SW)$	$d(PL)$	$d(PW)$
		Original		
$Jousselme$	0.2117	0.2140	0.1972	0.2135
$proposed$	0.1177	0.1115	0.1065	0.1255
		Weight		
ω_1	0.2492	0.2485	0.2537	0.2486
$\omega_2(m=1.5)$	0.2822	0.1874	0.2718	0.2586
		Noisy		
$Jousselme's$	0.3816	0.6714	0.3408	0.3574
$proposed's$	0.2546	0.5953	0.2771	0.3016
		Weight		
ω_1	0.2492	0.2485	0.2537	0.2486
$\omega_2(m=1.5)$	0.2822	0.1874	0.2718	0.2586

TABLE VI
THE ACCURACY OF EACH NEURAL NETWORK MODEL

Models	Accuracy result
V-FT	67.333
CMP	45.178
GoogleNet	47.437
ResNet	62.815
VGG	50.565
DenseNet	50.565

illustrate that our approach can effectively mitigate the impact of conflicting data in information fusion, thereby improving the reliability of decision-making. In order to simulate on real datasets, the first step is to build a reasonable belief generation framework to process the outputs of DNNs. DNNs use the softmax operator to convert the continuous activations of the output layer into class probabilities, these probabilities are then compressed into simplex by softmax according to the maximization of the DNNs parameters. Therefore, in this paper, we train the classification probability output by the softmax layer of the neural network as the belief distribution, so that different network models output different evidence to form decision opinions. The output of the softmax operation of the neural network N_i be the predictive distribution m_i:

$$m_i = softmax(N_i)$$

For the evidence given by the ith network model, we can find that the focal element of each evidence is a single subset. It is worth noting that different transformation strategies will have different belief generation results. We chose a one that is simple to calculate while ensuring that the predictions given by the model are not corrupted to the greatest extent.

The evaluation form for EMOCAP contains 10 items (neutral, happy, sad, angry, surprised, fearful, disgusted, depressed, excited, other). We chose 4 of these: angry, happy, neutral and sad, to be consistent with previous research.

2) Results: The result in Tab.VI is the classification accuracy of each neural network model, one may notice that there is a clear gap in accuracy between them. $V-FT$ has the highest accuracy 67.333%, and the gap between the worst-performing G model is greater than 20%. The accuracy of the six models has a certain difference, and the inherent properties may cause the networks to give conflicting decision-making opinions for the same sample. To verify the effectiveness of our model in integrating different and various information sources, we conduct experimental comparisons with existing evidence fusion frameworks on different numbers of models.

TABLE VII
COMPARISON RESULTS OF DIFFERENT METHODS, TO VERIFY THE PROPOSED DISSIMILARITY QUANTIZATION FUNCTION,

Measures	Accuracy result
Dempster [7]	70.287
Murphy's method [23]	67.941
Deng's [19]	70.721
Zhu's [25]	64.292
Proposed	**71.329**

In Fig.VII, we report the results of different methods on the dataset of EMOCAP. We can seen that the accuracy of all model fusion strategies is greatly improved compared with the single model, the performance of ours is better than the comparison models. As can be observed that the performance of multiple models has generally been improved, it may be due to the data enhancement effect of multiple models. Models embedded in evidence theory help characterize the uncertainty of images that are difficult to identify, and the comparison results also demonstrate this advantage. This comparison result verifies the excellent robustness of the proposed model, the reason is that our model is more sensitive to the reliability of the evidence, which can effectively reduce the degree to which conflicting information is considered. Therefore, our model is less dependent on data quality.

V. CONCLUSION

Fusion of conflicting evidence is an open issue of belief function. In this paper, a distance function between evidence vectors is proposed, which not only calculates the angle of the vector, but also considers the norm of the vector. Specifically, in order to make the difference between the evidence more comprehensive, we referred to the cosine function as a prototype to characterize the angle difference, and the norm of the vector was used to solve the shortcoming of the Jousseme $et\ al.$'s. Furthermore, we used the obtained dissimilarity and the constraint condition that the sum of weights is equal to 1 to construct a multivariate function, the solution obtained by calculating this function is used as the optimal fusion weight of each evidence. A set of application simulations has been conducted to demonstrate the effectiveness of the proposed algorithms. We intend to further study the proposed weight determination algorithm in our future work and to extend the proposed method to more applications.

ACKNOWLEDGMENT

The work is supported by Grant No. 2020KB0AC01 from the Scientific Project of Zhejiang Lab, and the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2021ZD0114303)

REFERENCES

[1] B. Li, D. Pi, Y. Lin, I. A. Khan, and L. Cui, "Multi-source information fusion based heterogeneous network embedding," 2020.

[2] R. R. Yager, "Set Measure Directed Multi-Source Information Fusion," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 6, pp. 1031–1039, 2011.

[3] F. Xiao, "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy," *Information Fusion*, vol. 46, no. 2019, pp. 23–32, 2019.

[4] J. Qian, X. Guo, and Y. Deng, "A novel method for combining conflicting evidences based on information entropy," *Applied Intelligence*, vol. 46, no. 4, pp. 1–13, 2016.

[5] W. Jiang, Z. Zhang, and X. Deng, "A novel failure mode and effects analysis method based on fuzzy evidential reasoning rules," *IEEE Access*, vol. 7, pp. 113605–113615, 2019.

[6] A. P. Dempster, "Upper and Lower Probabilities Induced by a Multi-valued Mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325 – 339, 1967.

[7] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[8] L. Fei, Y. Feng, and H. Wang, "Modeling heterogeneous multi-attribute emergency decision-making with Dempster-Shafer theory," *Computers & Industrial Engineering*, vol. 161, 2021.

[9] T. Zhang, J. Guo, C. Xu, X. Zhang, C. Wang, and B. Li, "A new oil spill detection algorithm based on Dempster-Shafer evidence theory," *Journal of Oceanology and Limnology*.

[10] C. Zhu, B. Qin, F. Xiao, Z. Cao, and H. M. Pandey, "A fuzzy preference-based Dempster-Shafer evidence theory for decision fusion," *Information Sciences*, vol. 570, pp. 306–322, 2021.

[11] K. Guo and L. Zhang, "Multi-source information fusion for safety risk assessment in underground tunnels," *Knowledge-based System*, vol. 227, 2021.

[12] C. Qiang and Y. Deng, "A new correlation coefficient of mass function in evidence theory and its application in fault diagnosis," *Application Intelligence*, 2021.

[13] L. A. Zadeh, "A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination," *AI Magazine*, vol. 7, no. 2, pp. 85–90, 1986.

[14] C. Zhu and F. Xiao, "A belief Hellinger distance for D-S evidence theory and its application in pattern recognition," *Engineering Applications of Artificial Intelligence*, vol. 106, no. 104452, 2021.

[15] A. Sarabi-Jamab, B. N. Araabi, and T. Augustin, "Information-based dissimilarity assessment in Dempster-Shafer theory," *Knowledge-based SystemS*, vol. 54, no. SI, pp. 114–127, 2013.

[16] L. Xiong, X. Su, and H. Qian, "Conflicting evidence combination from the perspective of networks," *Information Sciences*, vol. 580, pp. 408–418, 2021.

[17] F. Xiao, "CED: A Distance for Complex Mass Functions," *IEEE Transactions on Neural Network and Learning Systems*, vol. 32, no. 4, pp. 1525–1535, 2021.

[18] C. Cheng, X. Zhang, and T. Li, "Distance-based dynamic weight: A novel framework for multi-source information fusion," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 378–385, IEEE, 2022.

[19] Y. Deng, W. Shi, Z. Zhu, and Q. Liu, "Combining belief functions based on distance of evidence," *Decision Support Systems*, vol. 38, pp. 489–493, 2004.

[20] "A new divergence measure for belief functions in ds evidence theory for multisensor data fusion," *Information Sciences*, vol. 514, pp. 462–483, 2020.

[21] K. Zhao, R. Sun, L. Li, M. Hou, G. Yuan, and R. Sun, "An improved evidence fusion algorithm in multi-sensor systems," *Applied Intelligence*, pp. 1–11, 2021.

[22] J. Qian, X. Guo, and Y. Deng, "A novel method for combining conflicting evidences based on information entropy," *Applied Intelligence*, vol. 46, no. 4, pp. 876–888, 2017.

[23] C. K. Murphy, "Combining belief functions when evidence conflicts," *Decision Support Systems*, vol. 29, no. 1, pp. 1–9, 2000.

[24] M. Yüzkat, H. O. Ilhan, and N. Aydin, "Multi-model cnn fusion for sperm morphology analysis," *Computers in Biology and Medicine*, vol. 137, p. 104790, 2021.

[25] C. Zhu and F. Xiao, "A belief hellinger distance for d–s evidence theory and its application in pattern recognition," *Engineering Applications of Artificial Intelligence*, vol. 106, p. 104452, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

On Uncertainty Principles for Lowband Graph Signals

Na Li*, Linbo Shang†, Zhichao Zhang‡
* School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
E-mail: 202212150017@nuist.edu.cn Tel/Fax: +86-17805103957
† School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
E-mail: ashanglinbo@163.com Tel/Fax: +86-18555331185
‡ School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
E-mail: zzc910731@163.com Tel/Fax: +86-13376073017

***Abstract*—In recent years, the research of graph signal processing has attracted extensive attention of scholars, especially graph uncertainty principle. The lower bound of uncertainty is closely related to time-frequency resolution, therefore it is of great application value to explore the minimum lower bound of uncertainty in time-frequency analysis, biological data processing and radar imaging. In this paper, we obtain two uncertainty principles of lowband graph signals, basing on matrix theory and operator norm theory. By using the extreme value property of quadratic matrix eigenvalue and the related definition of matrix norm, the basic sparse uncertainty principle and the sparse lower bound of graph uncertainty principle based on multiplicative uncertainty are compared, which shows the superiority of the latter in signal detection performance. Then, examples and numerical comparison results are provided to verify the previous theoretical analysis, showing that the graph uncertainty principle based on multiplicative uncertainty is closer to the time-frequency resolution than the basic sparse uncertainty principle. The application of uncertainty principle involving signal sparsity expansion in compressed sampling is also given.**

I. Introduction

In the field of signal processing, the classical uncertainty principle means that a signal can't be too concentrated in time and space domain and frequency domain at the same time. This means that the resolution of the signal in time domain and frequency domain can't be determined at the same time. The uncertainty principle is not a single theorem, but a group of theorems. The most famous one is Heisenberg's uncertainty principle. Heisenberg inequality [1], [2],

$$\triangle_t^2 \triangle_\omega^2 \geq \frac{1}{4}, \tag{1}$$

as a theoretical tool to measure the resolution in time domain and frequency domain, plays an important role in time-frequency analysis and it is always a research focus. Signals are co-located in time and frequency, and the uncertainty principle represents the resolution trade-off between these two domains. Up to now, in the field of discrete signal processing, the research work on Heisenberg-like uncertainty principle has been very mature. The information characteristics of various stationary signals can be analyzed in detail, and preliminary applications have been made in compressed sensing, image denoising and signal recovery [3], [4]. Among many research results, the Heisenberg inequality is especially widely popularized, mainly involving the spread around a reference point. The related concepts of time spread and frequency spread are introduced below. For a signal $\mathbf{x}(t) \in l^2(R)$,$\langle \mathbf{x}(t), \mathbf{x}(t) \rangle_R = \int_{-\infty}^{\infty} \mathbf{x}^T(t)\mathbf{x}(t)\,dt$, its time spread about a point t_0 is defined by [5]

$$\triangle_t^2 = \frac{\langle (t-t_0)\mathbf{x}(t), (t-t_0)\mathbf{x}(t) \rangle_R}{\langle \mathbf{x}(t), \mathbf{x}(t) \rangle_R}. \tag{2}$$

Similarly, in the classical domain, for a signal $\mathbf{x}(t) \in l^2(R)$, its frequency spread is defined by

$$\triangle_\omega^2 = \frac{\langle \omega\widetilde{\mathbf{x}}(\omega), \omega\widetilde{\mathbf{x}}(\omega) \rangle_R}{2\pi \langle \mathbf{x}(t), \mathbf{x}(t) \rangle_R}. \tag{3}$$

For the classical uncertainty principle in time-frequency domain, see [6], [7], [8].

With the development of digital signal processing theory and the continuous improvement of information analysis ability, more non-stationary and nonlinear signals are encountered in practical application scenarios, such as infrastructure networks, social networks, economic networks and so on. These signals are no longer evenly sampled in time, but exist in the form of graphs. Graph signal includes vector signal, matrix signal and tensor signal. In this paper, we mainly describe vector graph signal. In order to further process and analyze irregular graph signals, it is necessary to extend the basic concepts of Fourier transform, filtering and multi-resolution analysis in classical signals to graph signals. In classical Euclidean space, the underlying domain is homogeneous, and the uncertainty principle applies to all signals, no matter how they are distributed on the coordinate axis. However, in the graph space, the underlying domain is irregular, and its topological structure, the relationship between nodes, needs to be considered. A little change in the structure of each small area on the graph will greatly affect the uncertainty boundary. According to the

This work was supported by the National Natural Science Foundation of China under Grant No 61901223 and the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant No 2021K205B.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

characteristics of the above graph signal, a large number of scholars have successively entered the research field of graph signal uncertainty principle, and put forward the uncertainty principle in different applications. In the following, we mainly introduce two types of common expressions of uncertainty principle on graph signal.

The measurement of the first kind of uncertainty principle mainly involves the spread around a reference point, usually the average position of the energy contained in the signal. For graph signals, it is difficult to extend the definitions of time spread and frequency spread in classical domain to graphs. Because of the structural complexity of graph signal, the extended definition of signal in vertex domain and graph domain is no longer a simple form about the mean or variance of signal. For a signal $\mathbf{x}(t) \in l^2(R)$,$\langle \mathbf{x}, \mathbf{x} \rangle_G = \mathbf{x}^T \mathbf{x}$, its graph spread about a vertex u_0 is defined by [5]

$$\triangle_g^2(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{P}_{u_0}^2 \mathbf{x}}{\langle \mathbf{x}, \mathbf{x} \rangle_G}, \tag{4}$$

where is $\mathbf{P}_{u_0}$ a diagonal matrix defined as

$$\mathbf{P}_{u_0} = diag\left\{d(u_0, v_1), \ldots, d(u_0, v_N)\right\}, \tag{5}$$

where v_i represents the ith vertex on the graph, $d(u, v)$ represents the shortest distance connecting two vertices. Similarly, for a signal $\mathbf{x}(t) \in l^2(R)$, its spectral spread is defined by

$$\triangle_s^2(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{\langle \mathbf{x}, \mathbf{x} \rangle_G}, \tag{6}$$

where $\mathbf{L}$ is the Laplace matrix corresponding to Fourier transform on the graph.

Ameya Agaskar and Yue M. Lu [5], [9] introduced the uncertainty principle involving these two concepts and gave the definition of uncertainty curve, which intuitively verified the conditions for realizing the lower bound in the uncertain feasible region. Based on the extension of the extended definition, in this paper, we introduce the first graph uncertainty principle. This indicates that the graph uncertainty principle is only applicable to some weighted graph families, which shows that it has certain limitations. Later, Bastien Pasdeloup [10] and others continued their in-depth research in this field and found a method to extend Ameya Agaskar's work [5] on unweighted graphs. Daniel Lenz [11] and others also studied on weighted graphs, and proved Laplace's principle of low-energy quantitative uncertainty on general weighted graphs. These all belong to the first kind of uncertainty principle. However, the previous work has its limitations. It shows that the extension of the classical uncertainty principle in time-frequency domain on the graph is not simple, which is undoubtedly a major focus of follow-up research.

The second kind of uncertainty principle mainly involves the absolute sparsity or concentration of signals. The research of Oguzhan Teke and Palghat P. Vaidyanathan [12] plays an outstanding role in this category, and its most important contribution is the graph uncertainty principle based on multiplicative uncertainty proposed according to the nonlocal measure of sparsity concept. Specifically, l_0 norm or l_1 norm is the related concept of signal expansion. This method is important because it quantifies the size of signal support. That is, the lower bound of the number of non-zero elements considering the graph signal and its Fourier transform. In order to find the signal that reaches the lower limit of derivation, sparse feature vectors are especially considered. Before this research Michael Elad and Alfred M. Bruckstein first introduced and proved the basic sparse uncertainty principle in early research, which has been widely used and popularized in many subsequent studies. Michael Elad and others also proved the uniqueness of sparse signal representation based on this important result. The above two research is a important achievement in the field of uncertainty principle on graph signals. Although the lower bound of uncertainty is derived from different research angles, they all use graph Laplacian operator to define the properties of graph signals when analyzing graph signals. Because of the importance of basic sparse uncertainty principle and graph uncertainty principle based on multiplicative uncertainty, we want to make a horizontal study and comparison between them.

Specific extension related concepts correspond to specific uncertainty principles, and they all have different applications. In order to further analyze the uncertainty of irregular graph signals, it is necessary to compare different uncertainty principles. The main reason is that if the lower bound is a constant with a parametric variable instead of a constant, it is possible to maximize the time-frequency resolution, thus maximizing the signal detection performance. Therefore, we focus on the basic sparse uncertainty principle and graph uncertainty principle based on multiplicative uncertainty, and deeply studies the minimum lower bound problem by comparing the extreme properties of quadratic matrix eigenvalues and the related definitions of matrix norms. A lowband signal indicates that the adjacent nodes on the graph are similar, and the node characteristics on the graph domain are smooth. In order to better compare the two kinds of uncertainty principle results, we determine to compare the sparse uncertainty boundaries based on l_0 norm on the same kind of lowband graph signals. Finally, the correctness of the previous theoretical derivation is verified by examples and experiments, which shows that the graph uncertainty principle based on multiplicative uncertainty has been improved in signal processing performance. The application of uncertainty principle involving signal sparsity expansion in compressed sampling is given.

II. Comparison of uncertainty principles for lowband graph signals

Graph signal processing is an extension of Fourier analysis. At present, scholars pay attention to the theoretical derivation of various forms of inequalities about Fourier transform of graph signals. In this section, we analyze and compare the uncertainty principle of lowband graph signals.

A. Notation

Given a graph $G(U, E)$, where U represents the set of N nodes in the graph, E is the set of edges. In this paper,

the fundamental structure of a graph G be captured by its adjacency matrix $\mathbf{A}$ of the graph, and each data element corresponds to one edges. Specifically, for any two different nodes $u_i, u_j \in U$ has $\mathbf{A}(i,j) = \omega_{ij}$, where ω_{ij} represents the connection weight value between node u_i and u_j. Given a graph G, the data set constitutes a graph signal, which is defined as a map, $\mathbf{x} : u_n \longrightarrow x_n$. Then the graph signal can be expressed as a vector $\mathbf{x} = [x_1, \ldots, x_N]^T$, which is not only a list, but also a graph.

The Laplacian matrix of a graph is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the graph degree matrix and $\mathbf{A}$ is the graph adjacency matrix. Let $\mathbf{L}$ be a real symmetric matrix. According to the theorem, $\mathbf{L}$ can be decomposed into features, $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{V}$ is an orthogonal matrix, $\mathbf{\Lambda}$ is a diagonal matrix and has non-negative real eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_N$. Let $\mathbf{V} = (\mathbf{v}_1, \ldots \mathbf{v}_N)$ be the Fourier basis of the graph, then for any signal $\mathbf{x}$ on the graph, the graph Fourier transform is

$$\hat{x}_m = \sum_{i=1}^{N} \mathbf{V}_{mi}^T x_i = \langle \mathbf{v}_m, \mathbf{x} \rangle_G . \tag{7}$$

The Graph Fourier Transform (GFT) $\widehat{\mathbf{x}}$ of a graph signal $\mathbf{x}$ is defined as $\hat{\mathbf{x}} = \mathbf{F}\mathbf{x} = \mathbf{V}^T\mathbf{x}$. Taking the characteristic vector of $\mathbf{L}$ as the Fourier basis and $\hat{x}_m$ as the Fourier coefficient of the signal $\mathbf{x}$ on the m-th Fourier basis. Its essence is the projection on the m-th Fourier basis, which measures the similarity between the graph signal and the Fourier basis, and then connects the graph domain with the spectrum domain.

In order to facilitate the later research, the definition of k-lowband signal is introduced below. If only the first k frequencies in the Fourier transform of a signal are meaningful, where $k \neq 0$, then the signal $\mathbf{x}$ is called a k-lowband signal, that is

$$\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_k, 0, \ldots, 0)^T . \tag{8}$$

The space formed by k-lowband signals is represented as $PW_G(\lambda_k)$ [13]. On the other hand, in the following, $\|\mathbf{x}\|_0$ represents the number of non-zero elements in signal $\mathbf{x}$, $\|\mathbf{A}\|_2$ represents the maximum singular value of matrix $\mathbf{A}$, and $\|\mathbf{A}\|_{max}$ represents the maximum absolute value of elements in matrix $\mathbf{A}$, that is

$$\|\mathbf{A}\|_2 = \max_{x \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}, \|\mathbf{A}\|_{max} = \max_{i,j} |a_{i,j}| . \tag{9}$$

B. Uncertainty Principles

Theorem 1 (Basic Sparse Uncertainty Principle): Given a vector $\mathbf{S} \in R^N$, for two orthogonal bases $\mathbf{\Phi}$ and $\mathbf{\Psi}$, the vector $\mathbf{S}$ can be expressed as $\mathbf{S} = \mathbf{\Phi}\boldsymbol{\alpha}, \mathbf{S} = \mathbf{\Psi}\boldsymbol{\beta}$. For each pair of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, there is:

$$\frac{\|\boldsymbol{\alpha}\|_0 + \|\boldsymbol{\beta}\|_0}{2} \geq \sqrt{\|\boldsymbol{\alpha}\|_0 \|\boldsymbol{\beta}\|_0} \geq \frac{1}{M}, \tag{10}$$

where M is a scalar value, indicating the mutual coherence between bases $\mathbf{\Phi}$ and $\mathbf{\Psi}$ [4].

Applying the above results to the signal $\mathbf{x} \in PW_G(\lambda_k)$, the first coefficient uncertainty boundary is obtained:

$$\|\mathbf{x}\|_0 \|\widehat{\mathbf{x}}\|_0 \geq \frac{1}{\mu_k^2}, \tag{11}$$

where μ_k represents the mutual coherence of the gauge basis. Let $\left(\mathbf{u}^1, \ldots, \mathbf{u}^k\right)$ represent the first k eigenvectors of Laplacian matrix $\mathbf{L}$, then μ_k has the following expression

$$\mu_k = \max_{\substack{1 \leq i \leq N \\ 1 \leq j \leq k}} \left| u_i^j \right| . \tag{12}$$

Definition 2 (Multiplicative Uncertainty): For a graph with GFT $\mathbf{F}$, the multiplicative uncertainty of a nonzero signal $\mathbf{x}$ is:

$$p_0(\mathbf{x}) = \sqrt{\|\mathbf{x}\|_0 \|\mathbf{F}\mathbf{x}\|_0}. \tag{13}$$

And $p_0(\mathbf{x})$ is called multiplicative uncertainty of signal $\mathbf{x}$ [12].

Theorem 2 (Multiplicative Uncertainty Principle): For a graph with (GFT) $\mathbf{F}$, the multiplication uncertainty of a nonzero signal $\mathbf{x}$ is lower bounded as follows [12]:

$$p_0(\mathbf{x}) \geq \frac{1}{\|\mathbf{F}^{-1}\|_2 \|\mathbf{F}\|_{\max}}. \tag{14}$$

C. Comparison of sparse uncertainty principles

In order to compare the two uncertainty principles conveniently, this paper limits the signal. It specifies the graph signal $\mathbf{x} \in PW_G(\lambda_k)$, so $\|\widehat{\mathbf{x}}\|_0 \leq k$. Then there is a lower sparsity bound for (11)

$$\|\mathbf{x}\|_0 \geq \frac{1}{k\mu_k^2}. \tag{15}$$

Remember that this uncertainty boundary is $A(k)$, that is $A(k) = \frac{1}{k\mu_k^2}$. Similarly, for a signal $\mathbf{x} \in PW_G(\lambda_k)$, the multiplicative uncertainty principle (14) also has the corresponding sparsity lower bound

$$\begin{aligned} \|\mathbf{x}\|_0 &\geq \frac{1}{k\left(\|\mathbf{F}^{-1}\|_2 \|\mathbf{F}\|_{\max}\right)^2} \\ &= \frac{1}{k\left(\|\mathbf{V}\|_2 \|\mathbf{V}^{-1}\|_{\max}\right)^2} \end{aligned} . \tag{16}$$

Remember that basic sparse uncertainty boundary is $B(k)$, that is $B(k) = \frac{1}{k\left(\|\mathbf{F}^{-1}\|_2 \|\mathbf{F}\|_{max}\right)^2}$. The main conclusion of this paper is given below.

Theorem 3: For a graph with (GFT) $\mathbf{F}$ and a nonzero signal $\mathbf{x}$, the following inequality holds:

$$A(k) \geq B(k) . \tag{17}$$

Proof: Proving (17) is equivalent to proving that the following inequality holds for any k

$$\left\|\mathbf{F}^{-1}\right\|_2 \|\mathbf{F}\|_{max} \geq \mu_k. \tag{18}$$

For any $\mathbf{F} \in R^{N \times N}$ and $\mathbf{x} \in R^{N \times 1}$, there is

$$max \frac{\mathbf{x}^T \left(\left(\mathbf{F}^{-1}\right)^T \left(\mathbf{F}^{-1}\right)\right) \mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_{max}\left(\left(\mathbf{F}^{-1}\right)^T \left(\mathbf{F}^{-1}\right)\right). \tag{19}$$

Let the matrix $\mathbf{F}^{-1} = (a_{ij})_{N\times N}$, identity matrix $\mathbf{E}_N = (\mathbf{e}_1, \ldots, \mathbf{e}_N)$, and $\mathbf{x} = \mathbf{e}_i$, we have

$$\lambda_{max}\left(\left(\mathbf{F}^{-1}\right)^T\left(\mathbf{F}^{-1}\right)\right) \geq a_{1i}^2 + \cdots + a_{Ni}^2. \tag{20}$$

Further limit the smaller boundary

$$\sqrt{\lambda_{max}\left(\left(\mathbf{F}^{-1}\right)^T\left(\mathbf{F}^{-1}\right)\right)} \geq \max_{i,j}\left|a_{ij}\right|, \tag{21}$$

hence

$$\left\|\mathbf{F}^{-1}\right\|_2 \geq \left\|\mathbf{F}^{-1}\right\|_{max}. \tag{22}$$

Since $\|\mathbf{F}\|_{max} > 0$, there is

$$\left\|\mathbf{F}^{-1}\right\|_2 \|\mathbf{F}\|_{max} \geq \left\|\mathbf{F}^{-1}\right\|_{max}. \tag{23}$$

For any k, the following inequality holds

$$\left\|\mathbf{F}^{-1}\right\|_2 \|\mathbf{F}\|_{max} \geq \mu_k, \tag{24}$$

namely $A(k) \geq B(k)$.

Remark: In the above proof process, (23) shows that Theorem 3 is still valid for non-lowband signals.

III. Experiments

In this section, the basic sparse uncertainty lower bound $A(k)$ and multiplicative uncertainty principle $B(k)$ are compared in the following four different data graphs in order to verify the correctness of the above conclusions. It shows the superiority of multiplicative uncertainty principle in signal detection performance.

(1) $G1$: Uniformly sampled graph 500 positions are generated uniformly and randomly in the unit square, and $G1$ is generated by the 4- nearest neighbor process.

(2) $G2$: Road-minnesota graph $G2$ is a Minnesota road network from the road network class in MatlabBGL library, which consists of 2,642 nodes and 3,304 vertices [14].

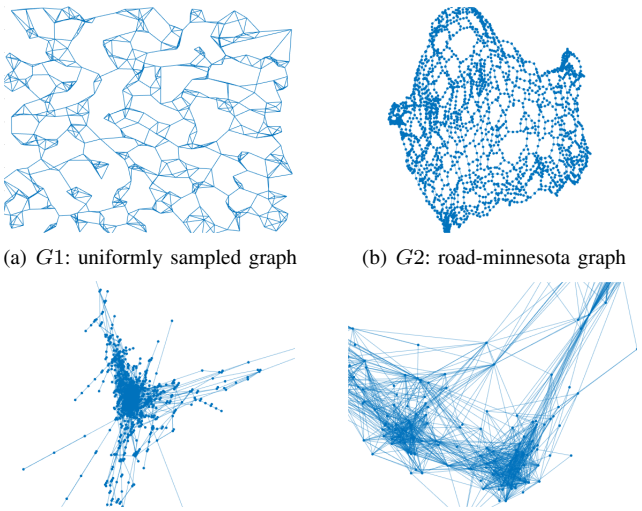
(3) $G3$: Bio-yeast graph $G3$ is a biological yeast graph from the biological network class in MatlabBGL library, which consists of 1458 nodes and 1948 vertices [14].

(4) $G4$:Infect-dublin graph $G4$ is a Dublin infection graph from adjacent network classes in MatlabBGL library, which consists of 410 nodes and 2765 vertices [14].

Fig. 1 shows four experimental graphs. Fig. 2 shows the function change images of two sparse boundaries about k on these graphs, where the orange line represents the lower bound of basic sparse uncertainty $A(k)$ and the blue line represents the lower bound of multiplicative uncertainty principle $B(k)$. As shown in the figure, for each graph studied, the lower bound $A(k)$ is higher than $B(k)$, which reflects the correctness of the conclusion in the previous section. On the one hand, the smaller the lower bound, the higher the time-frequency resolution and the stronger the information processing ability. Obviously, the lower bound $B(k)$ is closer to the limit of time-frequency resolution. Using this relationship, the time-frequency resolution can be maximized and the signal detection performance can be maximized. At the same time, it shows that the multiplicative uncertainty principle improves the signal processing ability on $PW_G(\lambda_k)$ compared with the basic sparse uncertainty principle. On the other hand, the larger the lower bound, the higher the accuracy of time-frequency resolution, so the lower bound $B(k)$ also has certain application value. In addition, as can be seen from the figure, when k tends to N, $A(k) = B(k)$, since $\mu_N = \|\mathbf{V}\|_{max}$ and $\|\mathbf{V}\|_2 = 1$. It shows that these two sparse lower bounds tend to be equal when k approaches N.

IV. Potential Application

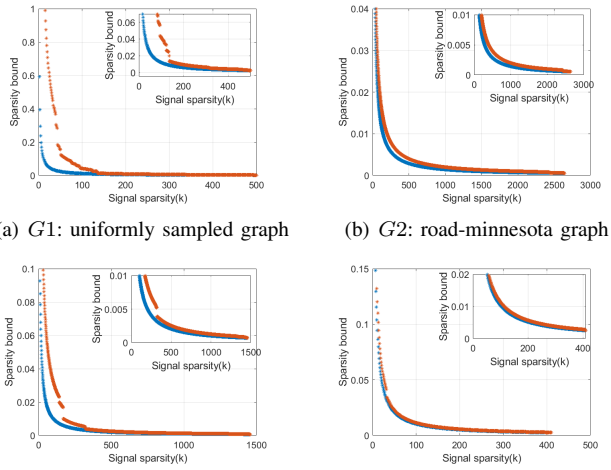
Compressive sensing is a technology for signal sampling, which specifically refers to the process of data compression in the sampling process. It breaks through the Nyquist Theorem in the field of signal processing, and achieves the same effect as full sampling with few sampling points in the process of signal sampling. Compressive sensing technology has a wide application prospect. Taking magnetic resonance imaging as an example, the long imaging time is a major defect of MRI. The application of compressed sampling technology not only improves the imaging speed, but also avoids wasting a lot of sampling resources and reduces the cost of signal measurement. Compressive sensing theory mainly includes three parts, which are sparse representation of signal, construction of measurement matrix and design of signal recovery algorithm. The sparse representation of the signal requires that the signal is sparse in a certain change domain. For the signal that meets the sparsity, the noise can be removed by setting the threshold. In [3], David L. Donoho and Xiao M. Huo applied the basic sparse uncertainty principle to compressive sensing and sparse signal processing to define reconstruction conditions. In addition, the uniqueness of sparse representation is proved, and how sparse the representation should be depends on the boundary provided by uncertainty principle. It can be seen from the above results that the multiplicative uncertainty principle improves the sparse boundary. This means that the improved lower bound can be used for compressive sensing



(a) $G1$: uniformly sampled graph (b) $G2$: road-minnesota graph

(c) $G3$: bio-yeast graph (d) $G4$: infect-dublin graph

Fig. 1. Graphs $G1$ to $G4$



(a) $G1$: uniformly sampled graph (b) $G2$: road-minnesota graph

(c) $G3$: bio-yeast graph (d) $G4$: infect-dublin graph

Fig. 2. Function diagram of uncertainty boundaries $A(k)$ and $B(k)$ from $G1$ to $G4$ with respect to k.

to reduce the sparse boundary of reconstruction based on coherence index. It further shows that the uncertainty principle related to signal sparsity expansion has direct application in the model construction of compressed sampling.

The specific application process is given below. Let the graph signal $\mathbf{x} \in PW_G(\lambda_k)$, the sparse matrix $\mathbf{H} \in R^{N\times N}$, then the sparse coefficient is $\mathbf{x} = \mathbf{Hs}$. The measurement matrix $\mathbf{H} \in R^{M\times N}$ is a universal independent gaussian random measurement matrix with the same distribution , which corresponds to the sub-sampling process. The observation signal $\mathbf{y} \in R^M$ obtained by $\mathbf{y} = \mathbf{Ax}$, which is the result of sub-sampling. The problem of reconstructing the graph signal is transformed into the optimization problem of solving linear equations $\mathbf{y} = \mathbf{AHs}$. In general, the number of equations is far less than the number of unknowns. The equation has no definite solution, so the signal cannot be reconstructed. However, if the graph signal is sparse, the unknown quantity can be accurately obtained from the M observed values and the signal reconstruction can be realized. The approximation of signal is obtained generally. According to the reduction of the sparse lower bound, the original signal can find a more sparse and effective expression to improve the reconstruction rate.

V. Conclusions

The lower bound of uncertainty is closely related to the time-frequency resolution, and studying the size of the lower bound is an important way to optimize the performance of signal analysis and processing. In this paper, we mainly study and compare two lower bounds of sparse uncertainty on lowband graph signals. By using the extreme value property of quadratic matrix eigenvalue and the related definition of matrix norm, we compare the basic sparse uncertainty principle of lowband graph signal with the sparse lower bound of graph multiplicative uncertainty principle, which shows that the latter is stricter and closer to time-frequency resolution. The correctness of the previous theoretical derivation is verified by examples and experiments, which shows that the graph uncertainty principle based on multiplicative uncertainty is optimized in signal processing compared with the basic sparse uncertainty principle. Finally, the application of uncertainty principle involving signal sparsity expansion in compressed sampling is given.

References

[1] M. A. Eastwood, "Heisenberg's uncertainty principle," *QJM : Monthly Journal of the Association of Physicians*, vol. 110 5, pp. 335–336, 2016.

[2] G. B. Folland and A. Sitaram, "The uncertainty principle: A mathematical survey," *Journal of Fourier Analysis and Applications*, vol. 3, pp. 207–238, 1997.

[3] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, pp. 2845–2862, 2001.

[4] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 48, pp. 2558–2567, 2002.

[5] A. Agaskar and Y. M. Lu, "A spectral graph uncertainty principle," *IEEE Transactions on Information Theory*, vol. 59, pp. 4338–4356, 2012.

[6] R. A. Somaraju and L. Hanlen, "Uncertainty principles for signal concentrations," *2006 Australian Communications Theory Workshop*, pp. 38–42, 2005.

[7] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM Journal on Mathematical Analysis*, vol. 49, pp. 906–931, 1989.

[8] Z. C. Zhang, X. Y. Shi, A. Y. Wu, and D. Li, "Sharper N-D Heisenberg's uncertainty principle," *IEEE Signal Processing Letters*, vol. 28, pp. 1665–1669, 2021.

[9] A. Agaskar and Y. M. Lu, "Uncertainty principles for signals defined on graphs: Bounds and characterizations," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3493–3496, 2012.

[10] B. Pasdeloup, R. Alami, V. Gripon, and M. G. Rabbat, "Toward an uncertainty principle for weighted graphs," *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1496–1500, 2015.

[11] D. Lenz, P. Stollmann, and G. Stolz, "An uncertainty principle and lower bounds for the dirichlet laplacian on graphs," *Journal of Spectral Theory*, vol. 10, pp. 115–145, 2016.

[12] O. Teke and P. P. Vaidyanathan, "Uncertainty principles and sparse eigenvectors of graphs," *IEEE Transactions on Signal Processing*, vol. 65, pp. 5406–5420, 2017.

[13] A. Mazarguil, L. Oudre, and N. Vayatis, "An uncertainty principle for lowband graph signals," *IEEE Signal Processing Letters*, vol. 29, pp. 727–731, 2022.

[14] R. A. Rossi and N. Ahmed, "The network data repository with interactive graph analytics and visualization," *AAAI Conference on Artificial Intelligence*, 2015.

CoA-DLinkNet: Connectivity-Enhanced Dual-Branch Road Extraction Network Based on D-LinkNet

Linghan Li, Heliu Chen, Renjie He*, Yuchao Dai, and Mingyi He
Northwestern Polytechnical University, China
E-mails: {lilinghan,chenheliu}@mail.nwpu.edu.cn, {davidhrj, daiyuchao,myhe}@nwpu.edu.cn

***Abstract* - Extracting roads from high-resolution remote sensing image data presents a challenging task in the field of remote sensing image processing, which is of great significance for urban planning, vehicle navigation, and geographic information system updates. In 2018, the champion of the DeepGlobe Road Extraction Challenge proposed a creative solution named as D-LinkNet, which employed a dilated convolution cascade module to expanded the network receptive field and achieved impressive results with a concise network structure. However, D-LinkNet still suffers from low connectivity caused by road breaks. Therefore, based on D-LinkNet, we have made three improvements in this paper: 1) incorporating a strip pooling module to capture long-range anisotropic contextual information, 2) employing a parallel upsampling structure in the decoder to supervise the shallow layers, and 3) introducing a connectivity branch with a connectivity attention module to model local connectivity of roads. We have named the upgraded network as CoA-DLinkNet. Our experimental results have demonstrated that the proposed network significantly improves the prediction accuracy and connectivity of the extracted road.**

I. INTRODUCTION

Roads are the backbone of modern transportation system, and an integral component of modern urban infrastructure and geographic information systems. Directly extracting roads from high-resolution remote sensing (RS) images is an important and efficient approach for updating road databases. However, challenges in extracting road breaks and missing parts arise due to large intra-class variations, small inter-class differences, large spans with small proportions, and shadow occlusions in RS images [1]. With the development of artificial intelligence in recent years, deep learning (DL) algorithms have been widely utilized in the field of RS image processing, and have been initially applied in road extraction tasks with certain breakthroughs and achievements.

Currently, the automatic road extraction tasks based on DL, with CNN as the representative, includes two sub-tasks, namely road segmentation and road centerline extraction. Road segmentation is a narrow road extraction task, which is typically regarded as a two-category semantic segmentation problem, where the roads are considered as the foregrounds and the others as the backgrounds. According to the different models used, methods can be classified into four categories: CNN based models [2], FCN based models [3-5], Encoder-Decoder based models [2, 6-9] and GAN based models [10]. On the other hand, road centerline extraction is to obtain the vector representation of roads, which is a key step in road network extraction and modeling. There are many application prospects in practical applications, such as autonomous driving, urban planning, vehicle navigation, etc. According to the different DL methods used, existing road centerline extraction algorithms can be divided into two categories: methods based on segmentation post-processing [11-13] and methods based on road iterative tracking [14-16]. The former heavily rely on the quality of road segmentation, and segmentation disruptions usually result in low road connectivity. The latter constructs the road network by iteratively predicting road points, offering better connectivity. However, such methods are complicate to the implement, and suffer from missed detection and misalignment due to local receptive field.

D-LinkNet [7] is an important network in the field of road extraction. It adopts an encoder-decoder architecture and incorporates a cascade of dilated convolutions in the middle of the network to expand the receptive field. However, it still exhibits three major problems. Firstly, it has limitations to expand the receptive field by stacking convolution. Secondly, the usage of square convolution kernels would inevitably introduce unwanted background interference. Lastly, the method based on segmentation post-processing predicts that the road is overly refined, resulting in poor road connectivity. To address these limitations, this paper proposes three improvements to the original D-LinkNet, taking the characteristics of narrow and long connected roads into account: 1) A strip pooling module [17] is integrated with the utilization of a long and narrow strip pooling kernel, which can

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

obtain anisotropic long-range context information in the horizontal and vertical directions, effectively expand the receptive field and reduce the background interference. 2) The decoder designs a multi-stage parallel upsampling structure, allowing the fusion of multi-scale features and enabling layer-by-layer depth supervision to refine network predictions. 3) A connectivity branch is introduced to demonstrate the connectivity between the modeled center pixel and the surrounding eight neighboring pixels, thereby improving the connectivity of predicted roads. The improved network is named CoA-DLinkNet. Experimental results shown that the CoA-DLinkNet achieves higher accuracy and improves connectivity in road extraction tasks.

II. PROPOSED METHOD

A. *Network Architecture*

The overall process of the proposed method is shown in Fig. 1. The CoA-DLinkNet is a multi-task learning network, which has SB (segmentation branch) and CB (connectivity branch) in parallel. The SB introduces a strip pooling module based on D-LinkNet, which directly outputs a predicted road extraction result shown in Fig. 2. The CB shown in Fig. 3 introduces several parallel connectivity attention modules [17] to predict the connectivity of among the pixels in eight directions of the road. The two branches improve learning performance through information exchange during the learning process.

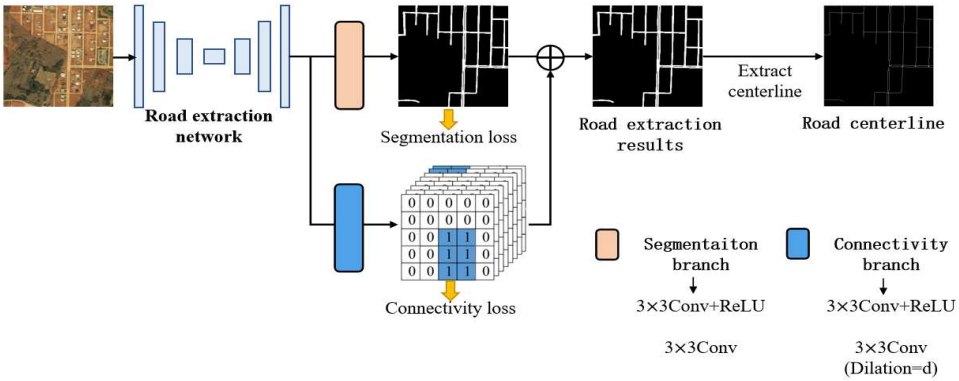


Fig. 1: An overview of SP-DLinkNet-CoA Network

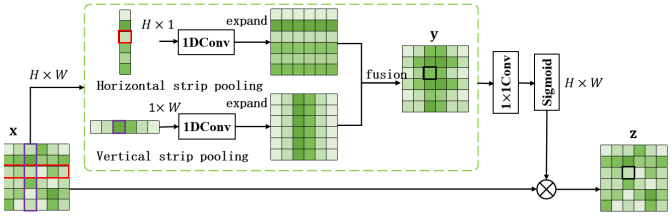


Fig. 2: Strip Pooling Module (SPM)

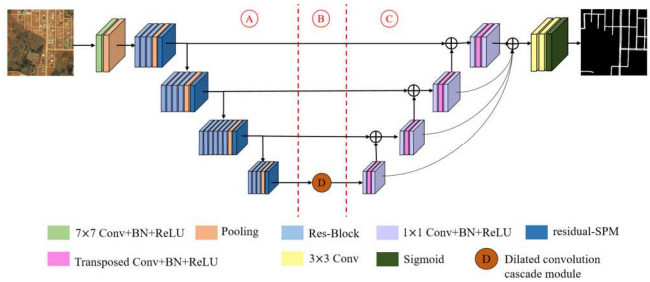


Fig. 3: Segmentation branch of CoA-DLinkNet

In the prediction stage, the final road extraction result can be obtained by the weighted summation of the two branch predictions. Finally, through the post-processing oof the centerline extraction method, the road centerline can be obtained for a wider range of real-world applications.

B. *Strip Pooling Module*

The structure of the Strip Pooling Module (SPM) [16] is shown in Fig. 2, given an input feature $\boldsymbol{x} \in \boldsymbol{R}^{C\times H\times W}$, feed it to two parallel horizontal and vertical strip pooling layers that average the row and column features. Then, it goes through a 1D convolutional layer with a kernel size of three, which is used to obtain the neighborhood features of the current location. In order to obtain the output $\boldsymbol{z} \in \boldsymbol{R}^{C\times H\times W}$ containing the global context information, finally we repeatedly expand the obtained features and combine them together. After the Sigmoid function is multiplied with the original feature $\boldsymbol{x}$, the output feature map $\boldsymbol{z}$ is obtained.

SPM combines horizontal and vertical strip pooling layers, conforming to the general shape of roads in RS images, and it is easy to build long and narrow long-range dependencies. At the same time, since the pooling kernel shape is narrower along the other dimension, it also focuses on capturing local details. SPM can be regarded as a spatial attention mechanism. Through $1\times N$ and $N\times 1$ pooling, each pixel can obtain the average of all pixel features in its row and column, that is, the global context information in the horizontal and vertical directions. Afterwards, close-range local context information can be obtained in the orthogonal direction of pooling through

1D convolution. Multiplying the fused features of the two pools with the original feature map is equivalent to using the correlation between pixels to enhance or weaken the features of each position in the original feature map. Compared with attention-based modules, SPM modules are lightweight and can be easily embedded into any building block to improve the ability to capture long-range spatial dependencies.

The segmentation branch of CoA-DLinkNet is based on D-LinkNet and integrates SPM to obtain long-range context information in a specific direction. To use pretrained ResNet weights, we add residual SPM after each stage. For the irregularly shaped country roads in RS images, their directions are changeable, and it is difficult to use horizontal or vertical pooling to obtain the contextual information of the road. In order to ensure that the pixels of these roads can also obtain sufficient contextual information, we retain the dilated convolution cascade module in the central part of D-LinkNet to expand the receptive field and to perform multi-scale information fusion. In addition, after each stage of the decoder, a multi-stage upsampling (MPU) is designed to restore the feature maps to 1024×1024 through different multiples. At the same time, in order to reduce the memory consumption, the dimension is reduced to a unified dimension and added. After that, it is fused by 3×3 convolution to get the final road prediction. The advantage of parallel upsampling is that it can aggregate multi-scale features, and at the same time, the loss can be directly back-propagated to each stage of the network during training, thereby generating an implicit deep supervision, which can enable the shallow layers of the network to be fully trained and reduce The loss of detailed features, to obtain more accurate road extraction results, also help to train deeper networks.

C. Connectivity Attention Module

The connectivity branch (CB) of CoA-DLinkNet introduces the Connectivity Attention Module (CoA)[18] with local relation perception. The CB consists of several parallel CoA modules at different intervals d . We experiment with CoA modules of $d=1, d=3$ and $d=5$, each supervised by the corresponding ground-truth connectivity cube, respectively. The final predicted connectivity cubes are summed according to the channel direction, and then added to the prediction map of the split branch according to the weight.

The CoA module consists of convolutions and a channel attentions. First, the input feature map is dimensionally reduced by a 3×3 convolution, and the number of channels after dimensionality reduction is 8. Second, the connectivity between the center pixel and the eight neighboring pixels is modeled by a 3×3 dilated convolution with a dilation rate $r=d+1$. Finally, we utilize the channel attention module to recalibrate the connectivity predictions after dilated convolutions. These steps are shown in Fig. 4.

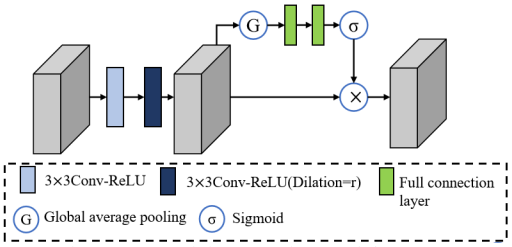


Fig. 4: Connectivity Attention Module (CoA)

The ground-truth connectivity cube $\boldsymbol{O} \in R^{H \times W \times C_0}$ can be generated by using binary ground-truth road labels, where C_0 represents the number of adjacent pixels sampled for a given pixel, and we predict the connectivity of the square neighborhood of a pixel, and set $C_0=8$, the interval between adjacent pixels is d. By traversing to check whether each pixel is connected to the eight neighborhood pixels with interval d, the connectivity of C_1-C_8 position can be obtained, and then the truth connectedness cube $\boldsymbol{O}$ can be obtained. In the connectivity cube $\boldsymbol{O}$, $\boldsymbol{O}_{i,j,c}=1$ represents the connection relation between the central pixel and adjacent pixels, where (i, j) represent the spatial position of pixels, and C represents the orientation $C_1 \rightarrow C_8$, if $\boldsymbol{O}_{i,j,c}=1$, it means that the center pixel (i, j) is connected to the pixel at position c, that is, both pixels are road pixels. We do not predict the connectivity of background pixels, and set the connectivity of pixels with pixel value of 0 in all directions to 0 to reduce irrelevant background noise. After obtaining the connectivity cube of truth value, it can supervise us to predict the connectivity between adjacent pixels through the network.

The predicted connectivity cube can be viewed as a subtask of the road segmentation task, which divides each pixel into foreground and background while ensuring overall semantic rationality and regional smoothness. In the connectivity prediction subtask, each task only classifies pixels in one direction. Therefore, the subtask must not only judge whether the pixel is the foreground or the background, but also judge whether the adjacent pixels in a specific direction are also the foreground, so the network is forced to model the local connectivity of the image and learn more robust features, so that the network predicts Road connectivity is stronger. In addition, the introduction of the connectivity task can explicitly integrate the prediction of the relationship between pixels into the optimization problem, which can be seen as a process of simulating graph inference in the model, which can enable

better preservation of road contours and prevent rough segmentation results. Since connectivity is a symmetric metric, it can be seen as a simple ensemble method that combines the predictions of two corresponding adjacent pixels to determine the connectivity of a pair of pixels.

III. EXPERIMENTS

A. Datasets

Our experiments are performed on the DeepGlobe Road Extraction dataset [19], which includes images from three different cities with pixel-level annotations. The spatial resolution is 0.5 meters, and the image size is 1024×1024. We divide the 6226 images of the official training set into 4226 training images, 1000 validation images, and 1000 test images. To evaluate the connectivity of roads, we skeletonized the road extraction results and used the DP algorithm to simplify the curve to obtain the ground truth of the road centerline.

B. Evaluation metrics

1) Pixel-level metrics: F1-score and IoU are two common metrics to evaluate pixel-level segmentation performance.

2) Graph-Based Metrics: Average Path Length Similarity (APLS)[20] measures the similarity of the average shortest path lengths between all pairs of nodes in the ground truth graph G and the predicted graph $\hat{G}$, and the topological correctness and connectivity of roads can be evaluated in (1).

$$APLS = \frac{1}{N}\sum\left(2\Big/\left(\frac{1}{S_{\hat{G}\to G}}+\frac{1}{S_{G\to\hat{G}}}\right)\right), \tag{1}$$

where N is the total number of images in the dataset, and $S_{\hat{G}\to G}$ measures the similarity of all corresponding node pairs in the ground−truth graph G to the predicted graph $\hat{G}$, defined as (2). $S_{G\to\hat{G}}$ is a symmetric term of $S_{\hat{G}\to G}$ in order to prevent false positive(FP).

$$S_{\hat{G}\to G} = 1-\frac{1}{M}\sum \min\left(1,\frac{\left|L(a,b)-L\left(\hat{a},\hat{b}\right)\right|}{L(a,b)}\right) \tag{2}$$

where M is the number of unique paths in the ground truth map G; a,b are all possible starting and ending points in G; $\hat{a},\hat{b}$ are the points in the predicted map $\hat{G}$ that are closest to the positions of a and b in the ground truth map G; $L(a,b)$ and $L(\hat{a},\hat{b})$ are the path lengths between corresponding nodes in G and $\hat{G}$, respectively.

C. Implementation Details

We train the network using the Adam optimizer[21] with a batch size 8, a maximum epoch of 300 and an initial learning rate of 2e-4. When the loss of the validation set does not decrease for three consecutive epochs, the learning rate is divided by 5, and the training is terminated early when the loss of six consecutive epochs no longer decreases, and the training is also stopped when the learning rate is less than 5e-7.

We perform large-scale data augmentation at training time, including color and spatial transformation. In the testing phase, we performed test-time augmentation by flipping horizontally, vertically, and diagonally, making eight predictions for each image, and averaging the test results of eight images as the final result.

Different from the loss function used by D-LinkNet, the segmentation branch of CoA-DLinkNet adds the SSIM loss [22], which can measure the similarity of the local structure of the image, and the loss function is shown in (3). where the BCE loss evaluates the difference between the predicted map and the ground-truth map at the pixel level, the SSIM loss evaluates the road structure difference at the local level, and the Dice coefficient loss evaluates the performance of the model road extraction at the image global level.

$$L_{seg} = L_{BCE} + \lambda_1 L_{Dice} + \lambda_2 L_{SSIM} \tag{3}$$

where λ_1, λ_2 are the weight parameters, which are set to 1 in the experiment.

To address the imbalance of positive and negative samples in the connectivity branch, we use the loss function Focal loss [22] for class imbalance. The prediction map of each direction of each connectivity cube uses the Focal loss to calculate the loss separately, so the loss L_{d1} of the CoA module with interval $d=1$ is defined as (4).

$$\begin{aligned} L_{d1} = -\frac{1}{C_o\times N}\sum_{c=1}^{C_o}\sum_{i=1}^{N} \Big[& y_i^c \alpha\left(1-\hat{y}_i^c\right)^{\gamma} y_i^c \log\left(\hat{y}_i^c\right) \\ & +\left(1-y_i^c\right)(1-\alpha)\hat{y}_i^{c\gamma} log\left(1-\hat{y}_i^c\right) \Big] \end{aligned} \tag{4}$$

where $C_o = 8$ and N is the total number of pixels in the image.

Since the task of predicting connectivity cubes can be viewed as a subtask of segmentation, we can recover road segmentation maps from the connectivity branch. Specifically, the predicted connectivity cube $\boldsymbol{O}$ is normalized by Sigmoid, and the threshold t is set for binarization. At the same time, we require that the predictions of the corresponding adjacent pixels (i', j') should be consistent, for example, if and only if $\sigma(\boldsymbol{O})_{i,j,5} > t$ and $\sigma(\boldsymbol{O})_{i,j+1,4} > t$, the two Adjacent pixels are connected.

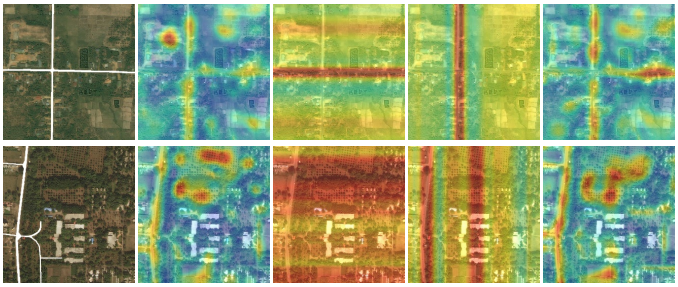
D. Results and Analysis

We use the original D-LinkNet34 as the baseline model to test each improved method, and the results are shown in Table 1. We first modified the encoder part, and by adding SPM at the end of each stage, F1-score, IoU, and APLS were improved by 0.71%, 0.85%, and 1.33%, respectively. In order to illustrate the role of the SPM module more intuitively, we performed the visualization of the feature map to observe the

distribution of features in the image, as shown in Fig. 5. It can be seen that the two pooling can accurately locate the horizontal and vertical roads, and the characteristics of the road area are significantly enhanced, indicating that the SPM module can effectively identify large-span roads.

Table 1 Ablation study (%) for the proposed strip pooling module (SPM), multi-stage parallel upsampling (MPU), SSIM loss and connected attention module (CoA)

Model	loss	F1-score	IoU	APLS
D-LinkNet	BCE+Dice	77.85	63.74	66.51
D-LinkNet+SPM	BCE+Dice	78.56	64.69	67.92
D-LinkNet+SPM+MPU	BCE+Dice	78.75	64.95	68.33
D-LinkNet+SPM+MPU	BCE+Dice+SSIM	79.56	66.05	70.34
D-LinkNet+SPM+MPU+CoA (CoA-DLinkNet)	BCE+Dice+SSIM	80.41	67.24	73.54



(a) (b) (c) (d) (e)

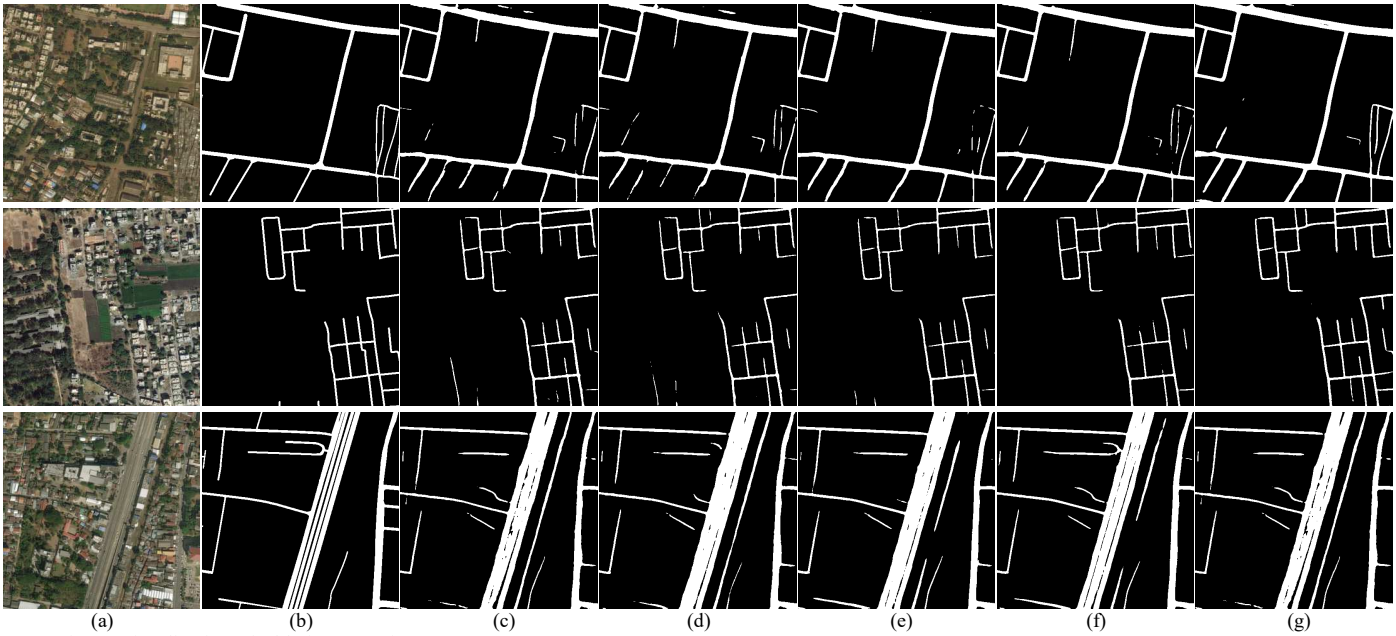
Fig. 5: Feature map visualization after adding SPM module. (a) to (e) correspond to: Image with GT, D-LinkNet, Vertical pooling, Horizontal pooling, and DLinkNet + SPM.

We continued to add the MPU structure at each stage of the decoder, and the F1-score, IoU, and APLS increased by 0.19%, 0.26%, and 0.41%, respectively. When we added the loss function SSIM for network training, the F1-score increased by 0.81%, the IoU increased by 1.10%, and the APLS increased by 2.01%. All metrics have been significantly improved, indicating that the supervision of SSIM loss has been significantly improved. The network's ability to identify road pixels makes the identified road more complete. Finally, the best performance obtained by our CoA-DLinkNet with the added connectivity branch is 80.41% for F1-score, 67.24% for IoU, and 73.54% for APLS, showing that the addition of CoA module to the network makes the predicted road connectivity better, more complete.

The visualization results of the above methods are shown in Fig. 6. The D-LinkNet can obtain a larger receptive field by dilated convolution, but it still has limitations when it contains the contamination information of irrelevant regions brought by the squared kernel. As shown in Fig. 6(c), when the road is occluded, the entire road segment or several parts of the road are lost. The limitations of D-LinkNet can be alleviated by employing the strip pooling to capture anisotropic long-range contextual information.

After adding SPM in the encoder, D-LinkNet can identify some occluded roads, shown in Fig. 6(d). We add parallel upsampling at each stage of decoder of the D-LinkNet+SPM, as shown in Fig. 6(e). After adding the multi-stage parallel upsampling, the network prediction is more accurate, and the gap recognition ability for two-way lane lines is stronger.

After adding SSIM loss, the network can identify more unidentified road segments, so the overall structure of the



(a) (b) (c) (d) (e) (f) (g)

Fig. 6: Visualization of Ablation Experiments. (a) Input Image, (b) Ground-Truth Mask, (c) D-LinkNet, (d) D-LinkNet+ SPM, (e) D-LinkNet + SPM + MPU, (f) D-LinkNet + SPM + MPU + SSIM Loss, (g) D-LinkNet+SPM+MPU+SSIM Loss + CoA.

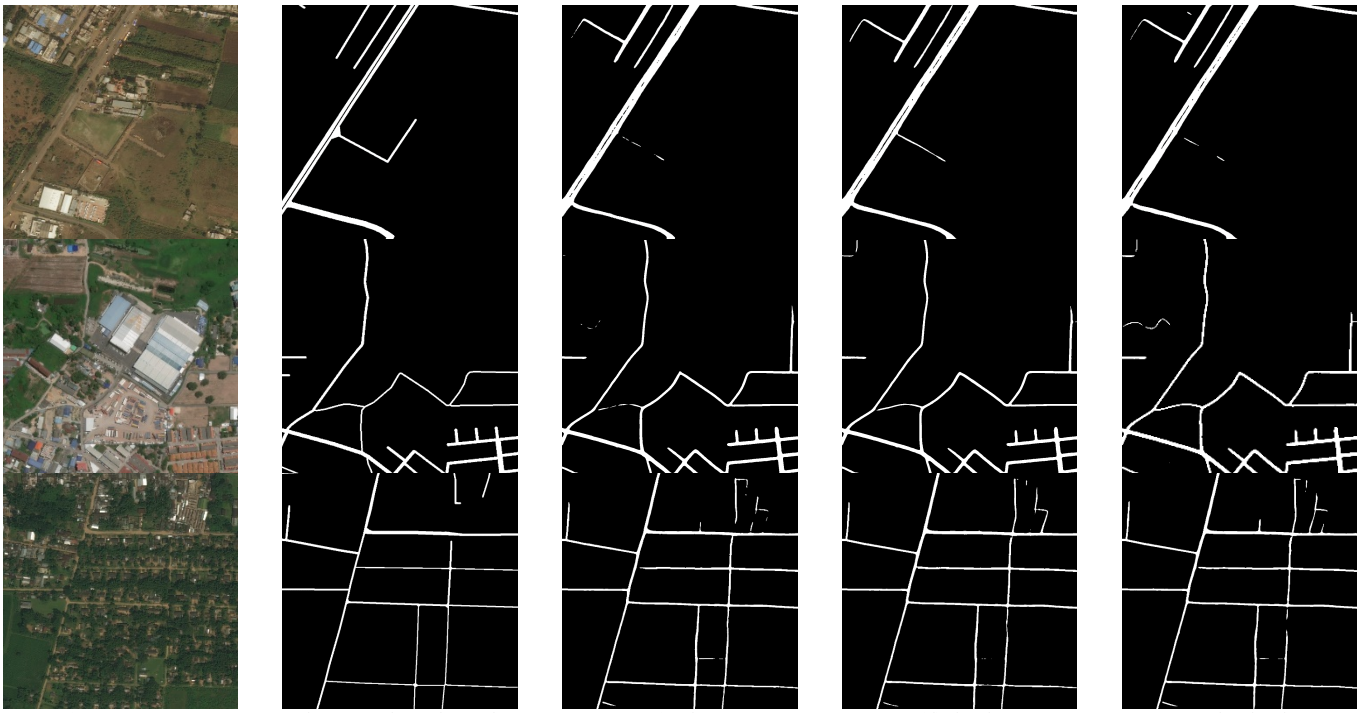
road is more complete, as shown in Fig. 6(f). Fig. 6(g) shows the visualization of the connectivity branch road extraction effect after adding CoA modules with d=1 and d=3. It can be seen that after adding the connectivity branch, the road fracture problem has been significantly improved, some obviously broken roads are connected, and the integrity and connectivity of the road have been significantly improved.

In order to explore the impact of connectivity branches of different CoA module configurations on network performance, we experimentally compared the connectivity branches of three different CoA module configurations, as shown in Table 2, where the second column represents the interval d of parallel CoA modules. The loss function used by the default connectivity branch is Focal loss.

When using the connectivity branch containing only the CoA module with $d=1$, the F1-score is improved by 0.76%, the IoU is improved by 1.06%, and the road centerline APLS is improved by 3.01%. When using the connectivity branch including the CoA module with $d=1$ and $d=3$, compared with $d=1$, F1-score, IoU, and APLS are improved by 0.09%, 0.13%, and 0.19%, respectively. This shows that when the multi-scale CoA module is used, the extracted roads are more complete and better road extraction results can be obtained. When using the connectivity branch containing CoA modules with d=1, d=3 and d=5, the F1-score, IoU, and APLS decreased by 0.05%, 0.07%, and 0.27%, respectively, compared to using the two scales. We can analyze the reason through Fig. 7.In Fig. 7(c), when the connectivity branch only contains CoA modules with d=1, the network can improve the situation of some road breaks. However, due to the small interval between adjacent pixels of CoA modules, there are still some broken roads with large intervals that do not be connected. When the connectivity branch contains CoA modules with d=1 and d=3, the number of broken roads in Fig. 7(d) is greatly reduced, part of the broken roads are reconnected, and the connectivity of the roads is improved. When the connectivity branch includes CoA modules with d=1, d=3 and d=5, due to the increase of the distance between pixels, more background information is modeled in the CoA module, resulting in a decrease in the reliability of prediction. In Fig. 7(e), roads with narrower widths will break again due to more surrounding backgrounds, and the network will detect redundant pixels that do not belong to the road.

Table 2 Ablation study of the proposed Connectivity Attention Module (CoA) with different configurations (%). d1, d3, and d5 represent CoA modules with d=1, d=3, and d=5, respectively.

Model	CoA	F1-score	IoU	APLS
CoA-DLinkNet	-	79.56	66.05	70.34
CoA-DLinkNet	1	80.32	67.11	73.35
CoA-DLinkNet	1, 3	80.41	67.24	73.54
CoA-DLinkNet	1, 3, 5	80.36	67.17	73.26



(a) RS Image (b) Ground-Truth Mask (c) d1 (d) d1, d3 (e) d1, d3, d5

Fig. 7: Visualizing the effect of different configurations of connectivity branches on road extraction

We compare the performance of CoA-DLinkNet with several state-of-the-art road extraction methods on the DeepGlobe road extraction dataset, including Improved Connectivity [13], ResUNet [6], LinkNet34 [7], D-LinkNet[8] and Deep Road Mapper [11] as shown in Table 3. Among them, Deep Road Mapper and Improved Connectivity utilize complex designed post-processing steps to improve road connectivity, while other methods directly generate extraction results through road skeleton extraction.

Table 3. Performance comparison of different models on DeepGlobe dataset (%)

Model	F1-score	IoU	APLS
DeepRoadMapper[11]	78.06	63.98	65.56
ImprovedConnectivity[13]	79.96	66.58	71.69
ResUNet[6]	75.34	60.96	61.06
LinkNet[7]	77.09	62.32	65.33
D-LinkNet[8]	77.85	63.74	66.51
CoA-DLinkNet	80.41	67.24	73.54

In Table 3, the CoA-DLinkNet algorithm achieves state-of-the-art results in F1-score, IoU, and APLS metrics, achieving 80.41%, 67.24%, and 73.54%, which are 0.68%, 0.66%, and 1.85% higher than the best-performing ImprovedConnectivity algorithm, respectively. Since our proposed CoA-DLinkNet integrates road segmentation and modeling of connectivity between adjacent pixels into one framework, it can more accurately extract roads from RS images. More importantly, the topological connectivity of roads can be further improved by using the CoA module.

IV. CONCLUSION

The CoA-DLinkNet proposed in this paper is a dual branch network based on D-LinkNet, the segmentation branch and the connectivity branch. In the segmentation branch of the network, a strip pooling module is added to the D-LinkNet encoder, enabling the network to capture anisotropic long- range contextual information. Parallel upsampling at each stage in the decoder can deeply supervise the network and make the network prediction more accurate. The connectivity branch models the connectivity between local pixels through the connectivity attention module, which is beneficial to improve the problem of predicting road breaks. Our experiments show that, compared with D-LinkNet, the extracted road by our CoA-DLinkNet network is more accurate, and the connectivity of road centerline is greatly improved, subject to little additional computational cost.

V REFERENCES

[1] Lian R, Wang W, Mustafa N, et al. Road extraction methods in high-resolution remote sensing images: A comprehensive review[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 5489-5507.

[2] Mnih V, Hinton G E. Learning to detect roads in high- resolution aerial images[C]. European Conference on Computer Vision, 2010: 210-223.

[3] Zhong Z, Li J, Cui W, et al. Fully convolutional networks for building and road extraction: Preliminary results[C]. IEEE International Geoscience and Remote Sensing Symposium, 2016: 1591-1594.

[4] Varia N, Dokania A, Senthilnath J. DeepExt: A convolution neural network for road extraction using RGB images captured by UAV[C]. IEEE Symposium Series on Computational Intelligence, 2018: 1890-1895.

[5] Kestur R, Farooq S, Abdal R, et al. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle[J]. J App RS, 2018, 12(1): 016020.

[6] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual u-net[J]. IEEE GRSL, 2018, 15(5): 749-753.

[7] Chaurasia A, Culurciello E. Linknet: Exploiting encoder representations for efficient semantic segmentation[C]. 2017 IEEE Visual Comm. and Image Processing. IEEE, 2017: 1-4.

[8] Zhou L, Zhang C, Wu M. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction[C]. Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2018: 182-186.

[9] Wang Y, Seo J, Jeon T. NL-LinkNet: Toward lighter but more accurate road extraction with nonlocal operations[J]. IEEE GRSL, 2021, PP(99): 1-5.

[10] Shi Q, Liu X, Li X. Road detection from remote sensing images by generative adversarial networks[J]. IEEE access, 2017, 6: 25486-25494.

[11] Máttyus G, Luo W, Urtasun R. Deeproadmapper: Extracting road topology from aerial images[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 3438-3446.

[12] Cheng G, Wang Y, Xu S, et al. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6): 3322-3337.

[13] Batra A, Singh S, Pang G, et al. Improved road connectivity by joint learning of orientation and segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 10385-10393.

[14] Ventura C, Pont-Tuset J, Caelles S, et al. Iterative deep learning for road topology extraction[J]. arXiv preprint arXiv:1808.09814, 2018.

[15] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. European conference on computer vision, 2016: 483-499.

[16] Tan Y Q, Gao S H, Li X Y, et al. Vecroad: Point-based iterative graph exploration for road graphs extraction[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 8910-8918.

[17] Hou Q, Zhang L, Cheng M M, et al. Strip pooling: Rethinking spatial pooling for scene parsing[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4003-4012.

[18] Mei J, Li R J, Gao W, et al. CoANet: Connectivity Attention Network for Road Extraction From Satellite Imagery[J]. IEEE Transactions on Image Processing, 2021, 30: 8540-8552.

[19] Demir I, Koperski K, Lindenbaum D, et al. Deepglobe 2018: A challenge to parse the earth through satellite images[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018:172-175..

[20] Van E A, Lindenbaum D, Bacastow T M. Spacenet: A remote sensing dataset and challenge series[J]. arXiv preprint arXiv:1807.01232, 2018.

[21] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[22] Qin X, Zhang Z, Huang C, et al. Basnet: Boundary-aware salient object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7479-7489.

[23] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Black-box Lossless Fragile Watermarking Based on Hidden Space Search for DNN Integrity Authentication

Gejian Zhao and Chuan Qin*
University of Shanghai for Science and Technology, Shanghai, PR China
E-mail: 202430387@st.usst.edu.cn, qin@usst.edu.cn

***Abstract*—A lossless and fragile watermarking scheme based on black-box with hidden space search is proposed for DNN. This is a publicized remote verification scheme, any authorized party can generate and verify the watermark of the target model without knowing the internal situation of the model and only by the model ground-truth input and output results. Specifically, we generate some sensitive samples near the decision boundary of the target model and exploit these samples as the fingerprints of the model. Once these target models are attacked or tampered to a certain extent, the labels corresponding to these sensitive samples will be flipped, thus accomplishing the purpose of authentication. Experimental results show that the proposed scheme is sensitive to the some current common model attack schemes, and only a small number of sensitive samples are needed to efficiently identify whether the watermarked model is intact or not.**

***Index Terms*—Artificial intelligence security, Fragile watermarking, Model integrity, Model attacks**

I. Introduction

Today, the theory and technology of deep learning have been gradually improved, the application scenarios has been expanded, and the commercialization of DNN has become more and more mature. Many technology companies choose to provide off-the-shelf models to customers on a subscription or rental basis, mainly through MLaaS (Cloud-enabled Machine Learning as a Service) [1], where the customers only need to request the API (Application programming interface) provided by the cloud service vendors from a remote location. However, potential risks and threats of MLaaS are emerging, such as hackers who can maliciously tamper models with backdoor attacks [2] or Trojan attack [3], and dishonest cloud server providers who degrade models to save storage costs [4], or even embed backdoor through model compression [5].

Model fragile watermarking technology [6], [7] is commonly used as an effective mean of identifying whether a watermarked DNN model has been tampered. An ideal fragile watermarking scheme should be sensitive in the face of various attacks against the model, and any modification will change the watermark information in the target model.

This work was supported in part by the National Natural Science Foundation of China under Grants 62172280, U20B2051 and 62172281, in part by the Natural Science Foundation of Shanghai under Grants 21ZR1444600 and 20ZR1421600, and in part by the STCSM Capability Construction Project for Shanghai Municipal Universities under Grant 20060502300.

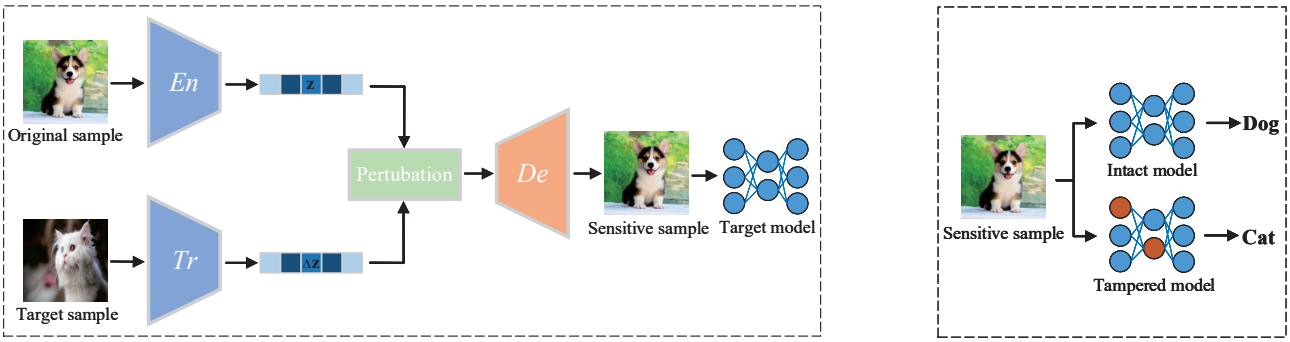
There are two mainstream fragile watermarking schemes, one is similar to the image-oriented fragile scheme, called white-box watermarking, where a specified string is embedded as watermark information in the target model by modifying the model parameters. For example, Botta *et al*. [8] proposed a fragile watermarking scheme with tampering localization. This scheme extracts the parameters of the model in order, and uses cryptographic hashing function, linear transformations and genetic algorithms to embed the watermark. Guan *et al.* [9] proposed a model integrity authentication scheme based on reversible watermarking scheme, which first obtained the hashes of model parameters by utilizing SHA-256 cryptographic hashing function, and then embedded the hashes into the host sequence by histogram shifting.

Another mainstream is generation of sensitive samples, called black-box watermarking, where the authentication task is achieved by observing the input-output relationships of the compromised models. He *et al.* [10] first proposed a black-box fragile watermarking scheme, which is able to detect attacks, such as data poisoning attack and backdoor attack, without knowing the priori knowledge of the model.

Black-box fragile watermarking is convenient, efficient and concealed compared to white-box watermarking, and has been gradually gaining attention from researchers in recent years. In this work, a black-box lossless fragile watermarking scheme based on hidden space search is proposed, and the main contributions of the proposed scheme are summarized as follows.

- The scheme does not require any internal information about the target model, or even access to train the model, and any authorized party can generate sensitive samples using the label information returned by the model.
- Design a high-performance VAE-GAN-based mapping and reconstruction model that maps the original samples into hidden space to form the corresponding hidden vectors, which can then be reconstructed into visually satisfactory sensitive samples by the decoder.
- A strategy is designed to search for perturbations in the hidden space on the hidden vector of the original sample, which can keep the search in the direction of the decision boundary of the target model, and in turn generates efficient sensitive samples that can be attached to the vicinity of the decision boundary.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



(a) Generation of sensitive samples

(b) Integrity authentication of the target model

Fig. 1. Flowchart of the proposed black box lossless fragile watermarking scheme.

II. The Proposed Fragile Watermarking Scheme

This part explains in detail the process from the generation of sensitive sample to model integrity authentication, and gives the corresponding mathematical optimization.

A. The Overall Flow of The Proposed Scheme

The proposed fragile watermarking scheme based on image hidden space is a black-box watermarking type. For an image classification model f to be protected, referred as target model, the licensees adopt MLaaS approach to generate some sensitive samples locally for the target model by querying it at a remote location. These sensitive samples and associated labels are retained by the licensees as fingerprints of the target model and are exploited as a criterion for determining whether the target model has been tampered. The specific flow of the proposed watermarking scheme is shown in Fig. 1(a).

The framework consists of three main parts, namely encoder En, decoder De and converter Tr. A high-performance encoder En is first trained on the dataset $\mathbf{X}$ with a corresponding decoder De. The purpose of the encoder En is to map an original sample x from dataset to hidden space, thus obtaining the corresponding hidden vector $\mathbf{z}$. Then an image with a different label from the original sample, called target sample x', is fed into the converter Tr to obtain its vector $\Delta\mathbf{z}$. Finally, the output of the target model f is used as the guidance to search perturbation on $\mathbf{z}+\Delta\mathbf{z}$ in the direction of the target sample in the low-dimensional hidden space. After the perturbation search, $\mathbf{z}+\Delta\mathbf{z}$ is fed to the decoder De, which reconstructs the input vector to generate the sensitive sample v. After that, we use the generated sensitive samples and the corresponding labels as the fingerprints $\mathbb{FG}=\{(v_i, f(v_i))\}_{i=1}^{n}$ of the target model, and hand them over to the authorized party for retention.

In the verification session, in the face of a suspicious model f' deployed on a cloud server, the authorized party can access the model f' remotely by using API interface. If there exists $f'(v_i)\neq f(v_i), i\in\{1,\cdots,n\}$, then it can be assumed that the parameters or structure of the model f' has been tampered and its integrity has been compromised, and the verification session is shown in Fig. 1(b).

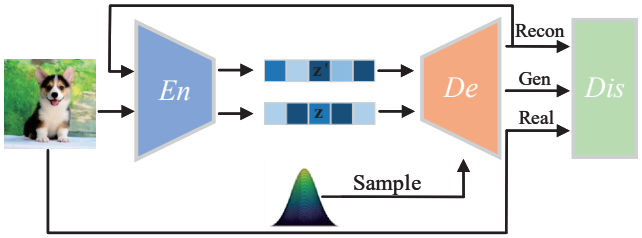


Fig. 2. Structure of VAE-GAN-based mapping and reconstruction model.

B. Generation of sensitive samples

The encoder module consists of several convolutional layers, in the process of the convolutional kernel downsampling the input image, the main features of the image are learned, which maps the input image to a low-dimensional vector, constituting a structure called hidden space $\mathcal{Z}$, and these compressed vectors are called hidden vectors $\mathbf{z}\in\mathcal{Z}$. The decoder consists of a number of transposed convolutional layers, and the effect of the decoder is to reconstruct the low-dimensional hidden vector into the original image.

The encoder-decoder coupling ensures a fixed mapping between the hidden vectors and the image samples, and means that the vectors in the hidden space can be manipulated to generate the new images needed for our task, i.e., the sensitive samples in proposed scheme. However, the samples generated by standard VAE are often blurred and visually poor [11]. To address this issue, we design a high-performance VAE-GAN-based mapping and reconstruction model, which can extract more critical features from the input images and generate images with better visual quality, the specific structure of the model is shown in Fig. 2.

Specifically, for the input original sample x, it is first transformed by the encoder into a low-dimensional vector $\mathbf{z}$, i.e., $\mathbf{z}\sim En(x)=q(\mathbf{z}\mid x)$, after which it is reconstructed by the decoder into reconstruction sample $\widetilde{x}$, i.e., $\widetilde{x}\sim De(\mathbf{z})=p(x\mid\mathbf{z})$. Then VAE model imposes a priori constraint term on the hidden vector $\mathbf{z}$ such that it conforms to $\mathbf{z}\sim\mathcal{N}(0,I)$.

Thus the loss function of VAE can be shown in Eq. (1):

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q(\mathbf{z}|x)}\left[\log \frac{p\left(x \mid \mathbf{z}\right) p\left(\mathbf{z}\right)}{q\left(\mathbf{z} \mid x\right)}\right] = \mathcal{L}_{\text{like}}^{\text{pixel}} + \mathcal{L}_{\text{prior}} \tag{1}$$

$$\mathcal{L}_{\text{like}}^{\text{pixel}} = -\mathbb{E}_{q(\mathbf{z}|x)}\left[\log p\left(x \mid \mathbf{z}\right)\right] \tag{2}$$

$$\mathcal{L}_{\text{prior}} = D_{KL}\left(q\left(\mathbf{z} \mid x\right) \parallel p\left(\mathbf{z}\right)\right) \tag{3}$$

where $\mathcal{L}_{\text{like}}^{\text{pixel}}$ is the reconstruction loss used to keep the reconstructed sample as similar as possible to the original sample, and $\mathcal{L}_{\text{prior}}$ is the KL scatter loss, which measures the similarity between two distributions.

In order to obtain visually better sensitive samples, we adopt the idea of GAN [12] by using the decoder De as the generative network in GAN, which is followed by a discriminator Dis. Detailedly, we sample the gaussian noise $\mathbf{z}_p$ as the input to the decoder De to obtain generated samples $\bar{x}$. Since the distribution $p(\mathbf{z})$ and the distribution $q(\mathbf{z} \mid x)$ tend to differ to some extent, in order to minimize this difference, we feed the real sample x, the reconstructed sample $\widetilde{x}$ and the generated sample $\bar{x}$ together into the discriminator. In the ideal case, the discriminator can not recognize the difference between the three. Besides, in order to generate more accurate sensitive samples after adding perturbations in the hidden space, we need to reduce the difference between the hidden vector $\mathbf{z}'$ obtained from the reconstructed sample $\widetilde{x}$ by the encoder De and the hidden vector $\mathbf{z}$ of the real sample. Therefore, we design the $\mathcal{L}_{\text{GAN}}$ loss and the $\mathcal{L}_{\mathbf{z}}$ loss, as shown in Eq. (4) and Eq. (5), respectively:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} =& log(Dis(x)) + log(1 - Dis(De(\mathbf{z}_p))) + \\ & log(1 - Dis(De(En(x))))\end{aligned} \tag{4}$$

$$\mathcal{L}_{\mathbf{z}} = \left\|\mathbf{z} - \mathbf{z}'\right\|_2 \tag{5}$$

where the $Dis(\cdot)$ in Eq. (4) is the probability that the discriminator can recognize the input true sample as the natural image from the training set.

The difference between the reconstructed sample $\widetilde{x}$ and the generated sample $\bar{x}$ can be captured by the discriminator, so we can obtain a decoder with better reconstruction capability. Therefore, we can optimize the parameters in the encoder, decoder and discriminator simultaneously by Eq. (6).

$$\mathcal{L}_{\text{En-De-Dis}} = \mathcal{L}_{\text{like}}^{\text{pixel}} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\mathbf{z}} \tag{6}$$

Through the joint training of encoder, decoder and discriminator, the conditions for the generation of sensitive samples are provided.

C. Strategies for Perturbation Search

After completing training on encoder En and decoder De, both of which have acquired the ability of mapping and reconstructing. Considering the visual quality of sensitive samples, we should impose as little as possible perturbation on original samples, so the selection of target samples is needed. Specifically, we perform Euclidean distance calculation for each available sample that differs from the label of original sample, then select the samples with the closest distance to the original sample as the target samples, see Eq. (7):

$$\min_{l \neq y} \min_{f(x')=l} \left\|x' - x\right\|_2 \tag{7}$$

where $f(\cdot)$ is the target model, x is the original sample, y is the label of the original sample, x' is the target sample, and l is the label of the target sample.

After selecting a suitable target sample x', we fed it into the converter Tr such that the converter extract the key semantic information in the target sample and converts these information into a hidden vector $\Delta\mathbf{z}$. To facilitate the addition of suitable perturbations to the hidden vector, the converter Tr is constructed in the same structure as the decoder En. The input samples are mapped by the encoder to the corresponding hidden vector $\mathbf{z}$, then $\Delta\mathbf{z}$ is superimposed onto the hidden vector $\mathbf{z}$ as a perturbation, and finally $\Delta\mathbf{z} + \mathbf{z}$ is used as the input to the decoder De. The decoder can reconstruct the input from the hidden space to the original sample space to generate the sensitive sample. The core idea of the scheme is to add the perturbation to the hidden vector of the input samples, which makes the samples move closer to the decision boundary of the target model. We can identify the task as a mathematical optimization problem, see Eq. (8):

$$\begin{aligned}\mathcal{L}_{\text{target}} =& \max\left(f(De(\Delta\mathbf{z} + \mathbf{z}))_y - \max f(De(\Delta\mathbf{z} + \mathbf{z}))_{y \neq l}, 0\right) \\ & s.t. \|\Delta\mathbf{z} - \mathbf{z}\|_2 \leq \varepsilon\end{aligned} \tag{8}$$

where $f(\cdot)_y$ denotes the confidence that the output label is y, $\|\cdot\|_2$ is the Euclidean distance, and ε denotes the maximum threshold of perturbation to constrain the visual quality of sensitive samples.

Besides, we should control the direction of the search perturbations in the hidden space towards the location of the target sample. Specifically, the original sample x, the sensitive sample v and the target sample x' are mapped through the encoder En to obtain their corresponding hidden vectors $\mathbf{z}$, $\mathbf{z}_{wmk}$ and $\mathbf{z}_{tgt}$. Then the differential vector $\mathbf{z}_{tgt-ori}$ between the target sample and the original sample and the differential vector $\mathbf{z}_{wmk-ori}$ between the sensitive sample and the original sample can be obtained by Eq. (9) and Eq. (10), respectively.

$$\mathbf{z}_{tgt-ori} = \mathbf{z}_{tgt} - \mathbf{z} \tag{9}$$

$$\mathbf{z}_{wmk-ori} = \mathbf{z}_{wmk} - \mathbf{z} \tag{10}$$

Since the differential vectors $\mathbf{z}_{tgt-ori}$ and $\mathbf{z}_{wmk-ori}$ contain information about the direction in the hidden space, the angle between the two vectors can be measured by calculating the cosine similarity, thus we impose a certain constraint on the search direction of the perturbation, see Eq. (11).

$$\begin{aligned}&\text{Cos}\left(\mathbf{z}_{wmk-ori}, \mathbf{z}_{tgt-ori}\right) = \\ &\frac{\sum_{i=0}^{n}\left(\mathbf{z}_{wmk-ori}^{i} \times \mathbf{z}_{tgt-ori}^{i}\right)}{\sqrt{\sum_{i=0}^{n}\left(\mathbf{z}_{wmk-ori}^{i}\right)^2} \times \sqrt{\sum_{i=0}^{n}\left(\mathbf{z}_{tgt-ori}^{i}\right)^2}}\end{aligned} \tag{11}$$

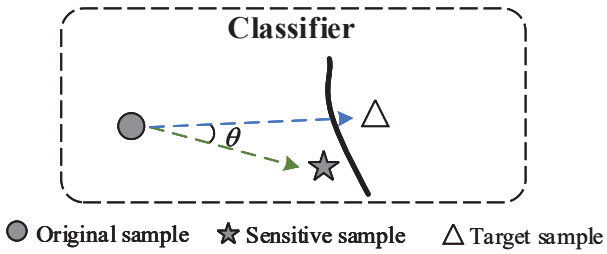


Fig. 3. Diagram of perturbation search direction.

The utilization of cosine similarity allows the search direction of the perturbation to coincide with the direction of $\mathbf{z}_{tgt-ori}$. In the sample space this is expressed as the direction of the perturbation remaining on the line from the original sample to target sample, as shown in Fig. 3.

The visual quality of the sensitive samples need to be maintained while imposing constraints on the direction of perturbation search, so the overall loss function for the perturbation search can be expressed as:

$$\mathcal{L}_{\text{perturbation}} = \mathcal{L}_{\text{target}} + (1 - \text{Cos}\left(\mathbf{z}_{wmk-ori}, \mathbf{z}_{tgt-ori}\right)) + \|v - x\|_2 \tag{12}$$

After the classification information of the target samples are obtained by accessing the model, according to the designed perturbation search strategy, the sensitive samples can be made to reach near the decision boundary of the target model with the loss function $\mathcal{L}_{\text{perturbation}}$.

III. Experimental Results and Analysis

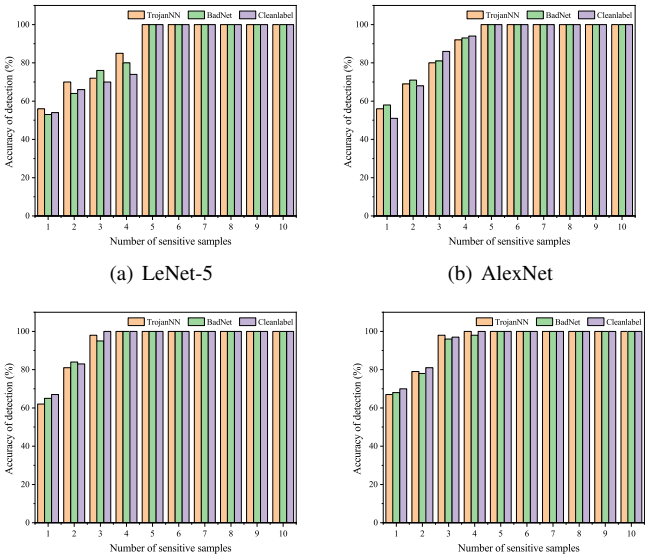
In the following experiments, we selected four typical DNN models, namely LeNet-5, AlexNet, VGGNet19 and ResNet50, with different performance and structural complexity, to demonstrate the fitness of our scheme. To further prove the generalization ability of the schemes, these models were trained and manipulated on two different datasets, namely MNIST and Cifar-10. All experiments were conducted on a computer equipped with NVIDIA TITAN RTX GPU, Intel Xeon W-2123 CPU and 64 GB RAM.

A. Setup

We simulated two different types of attacks against the integrity of the model. For the backdoor attack type, three typical attack schemes, BadNet [2], TrojanNN [3] and Cleanlabel [13], were chosen, all resulting in incorrect output of the model during inference. In this scenario, we assume that the attacker attempts to minimize the prediction accuracy of the target model, as well as to maximize the likelihood of misclassifying samples into the target category, so we adjusted the success rate of three attack schemes to the highest. We use attack intensity to measure the strength of the attack, i.e. the accuracy of successfully outputting a toxic sample as a predefined mislabel. The second type is that the dishonest cloud server provider compresses the target model to save storage costs, and used the compression rate to measure the attack intensity, with the higher the compression rate the higher the attack intensity.

TABLE I
Experimental Models and Attack Settings

Target Model	Attack Type	Dataset	Attack scheme	Attack Intensity
LeNet-5	Backdoor attack	MNIST	BadNet	100%
			TrojanNN	96%
			Cleanlabel	95%
AlexNet	Backdoor attack	Cifar-10	BadNet	100%
			TrojanNN	98%
			Cleanlabel	97%
VGGNet19	Backdoor attack	Cifar-10	BadNet	100%
			TrojanNN	98%
			Cleanlabel	98%
ResNet50	Backdoor attack	Cifar-10	BadNet	100%
			TrojanNN	99%
			Cleanlabel	98%
AlexNet	Compress attack	Cifar-10	Pruning	Compress 15%
VGGNet19	Compress attack	Cifar-10	Pruning	Compress 30%



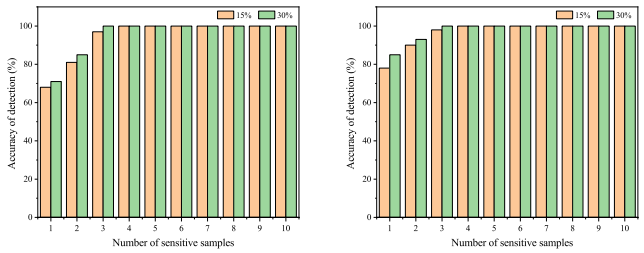
(a) LeNet-5 (b) AlexNet

(c) VggNet19 (d) ResNet50

Fig. 4. Detection accuracy rate of backdoor attacks for sensitive samples in different models.

The specific attack settings are shown in TABLE I. We conducted backdoor attacks on four different watermarked models with BadNet, TrojanNN and Cleanlabel. 100 attack operations were performed on the watermarked models for each attack scheme, and all the average success rate of each attack are over 95%. In addition, we performed compression attacks on two watermarked models, namely AlexNet and VGGNet19, trained on Cifar-10, by means of pruning attack.

The length of the hidden vector was set to 68, and we trained 300 epochs with Adam optimizer, the learning rate was 0.08, the weight decay was set to 0.005, and the threshold ε was set to 3.5. 20 images of each category from each dataset were randomly selected as the original and target samples for perturbation search.



(a) AlexNet (b) VGGNet19

Fig. 5. Detection accuracy rate of compression attacks for sensitive samples in different models.

TABLE II
DIFFERENCES IN THE PROPERTIES OF DIFFERENT FRAGILE WATERMARKING SCHEMES

Scheme	Watermarking type	Performance	Sample generation
RWDCNN [9]	White-box	↓	N/A
SSFDNN [10]	Black-box	=	White-box
NeuNac [8]	White-box	↓	N/A
FNNW [14]	Black-box	↓	Black-box
Proposed	Black-box	=	Black-box



Fig. 6. Visualization of the training process of reconstructed samples.

B. Tampering Detection

To assess the effectiveness of the watermarking scheme in detecting attacks on the integrity of the model, i.e. the fragility of the watermark, we conducted statistics on how the labels associated with the sensitive samples changed after the model was compromised. The experimental results are shown in Fig. 4, where the horizontal coordinates indicate the number of sensitive samples and the vertical coordinates indicate the success rate of identifying tampering attacks. It should be noted that the proposed scheme does not identify the intact models as tampered models, i.e. the false positive rate is 0.

As can be seen from Fig. 4, the correct detection rate for the three backdoor attacks gradually increases with the number of sensitive samples increases. When the number of sensitive samples is set to 5, 100% correct detection rate is achieved on all models, which indicates that the verifier only needs 5 API accesses to be informed of the integrity, thus proving that the proposed scheme has a very high verification efficiency.

In order to verify the effectiveness of the scheme in detecting model compression attacks, AlexNet and VGGNet19 after compression attacks were also tested, the experimental results are shown in Fig. 5. When the number of sensitive samples was set to 3, both AlexNet and VGGNet19 achieved 100% correct detection for 30% pruning attack, and 96% and 98% accuracy for 15% pruning attack, respectively. When the number was set to 4, both models achieved 100% accuracy in detecting different levels of pruning attacks.

C. Comparisons with State-of-the-art Schemes

To verify the superior performance of the proposed scheme, we compared it with some other model fragile watermarking schemes. According to the type of watermarking, NeuNac [8] and RWDCNN [9] can be classified as white-box watermarking, both of which embed a specific string into the model, and inevitably modify some parameters and cause performance degradation. SSFDNNN [10], FNNW [14] and the proposed scheme are black-box watermarking schemes, the FNNW requires the target model for fine-tuning, which also distorts the decision boundaries of the model to some extent. The state-of-the-art scheme SSFDNN does not impact on the performance of the target model, but it requires information of the model during the generation of sensitive samples. TABLE II lists the differences in the properties of these fragile watermarking schemes.

To further demonstrate the effectiveness, we conducted more detailed comparison experiments and analysis between SSFDNN and the proposed scheme. We selected VGGNet19 trained on Cifar-10 as the experimental subject, and generated three different groups of experiments with sensitive sample numbers of 1, 2 and 3 using the SSFDNN and the proposed scheme, respectively. Then we attacked the experimental model 100 times with each of the five attacks in TABLE I. At the authentication end, changes of associated labels were used to determine whether the experimental model had been tampered, as illustrated by the results in TABLE III.

With the number of sensitive samples set to 1, the proposed scheme achieved a higher detection accuracy than SSFDNN. As the number of sensitive samples increases, SSFDNN's effectiveness increased but is still slightly worse than the proposed scheme. Overall, the proposed scheme is able to detect a wide range of attacks with excellent results and does not require any internal information of the target model, and only acquires the output of the model to generate several efficient sensitive samples.

D. Imperceptibility

In addition to the effectiveness, imperceptibility is also a very important evaluation criterion in watermarking schemes. The sensitive sample should have good visual quality, so that the human eye cannot distinguish the difference between the sensitive samples and the original samples. The experiment adopted imperceptibility to measure the visual difference, and imperceptibility can be obtained by Eq. (13):

$$Imperceptibility = \sqrt{\frac{1}{n} \parallel v - x \parallel_2} \tag{13}$$

where n is the number of pixels in the image, v is the sensitive sample and x is the original sample.

A lower imperceptibility indicates a better visual quality for the sensitive sample. We first conducted imperceptibility experiments on the reconstruct effect of VAE-GAN, and we show the reconstructed samples obtained every 30 epochs, as shown in Fig. 6. It can be seen that, the reconstruction process

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED SCHEME WITH SSFDNN ON THE VGGNET19

	SSFDNN [10]			Proposed		
	1 Sample	2 Samples	3 Samples	1 Sample	2 Samples	3 Samples
BadNet	65%	90%	100%	64%	83%	96%
TrojanNN	68%	87%	98%	62%	80%	98%
Cleanlabel	70%	89%	99%	68%	82%	100%
15% pruning	51%	80%	94%	78%	90%	97%
30% pruning	55%	82%	98%	83%	93%	100%



Fig. 7. Visual quality of sensitive samples on MNIST dataset.

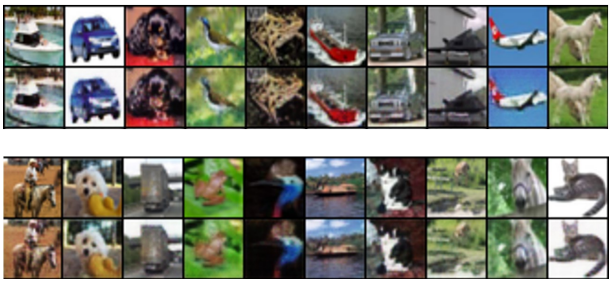


Fig. 8. Visual quality of sensitive samples on Cifar-10 dataset.

gradually enriches from smooth regions at low frequencies to complex details at high frequencies, which is in line with the reconstruction law of deep learning models. After training the model for 300 epochs, the reconstructed samples achieved high visual quality, and the imperceptibility between the reconstructed sample and its original sample is 23.6.

To further verify the imperceptibility of the sensitive samples, AlexNet was selected to trained on MNIST and Cifar-10, respectively. The data in MNIST are simple single-channel images, which allows the model to learn the core features more easily, while the images in Cifar-10 are colourful and more complex, which makes the perturbation search more difficult. In this experiment, we randomly selected 50 original samples from each of the two datasets, which can generate 50 sensitive samples, and then calculated the imperceptibility in pairs, followed by the average mean. After calculation, the mean values of imperceptibility on MNIST and Cifar-10 are 51.25 and 98.83, respectively. Some representative sensitive samples generated from the corresponding datasets are shown in Fig. 7 and Fig. 8.

IV. CONCLUSION

In this paper, a black-box fragile watermarking scheme is proposed, which can generate sensitive samples only by the output results returned by the model. Firstly, a high-performance mapping and reconstruction model is designed, which superimposes the semantic features of the target sample on the hidden vector corresponding to the original sample by searching for perturbations, so that the original samples keep moving towards the decision boundary. After that, the decoder reconstructs the perturbed vector to obtain sensitive samples with good visual quality. Extensive experiments demonstrate that the sensitive samples can effectively detect model backdoor attack and compression attack. In addition, the user only needs a low cost for samples generation and model integrity verification, which has practical application value for image classification models.

REFERENCES

[1] M. Ribeiro, K. Grolinger, and M. A. Capretz, "Mlaas: Machine learning as a service," in *Proceedings of 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. 2015, pp. 896–902.

[2] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[3] R. Tang, M. Du, N. Liu, F. Yang, and X. Hu, "An embarrassingly simple approach for Trojan attack in deep neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 218–228.

[4] S. Wiedemann, H. Kirchhoffer, S. Matlage, P. Haase, A. Marban, T. Marinč, D. Neumann, T. Nguyen, H. Schwarz, T. Wiegand *et al.*, "Deepcabac: A universal compression algorithm for deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 700–714, 2020.

[5] D. Wu and Y. Wang, "Adversarial neuron pruning purifies backdoored deep models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16913–16925, 2021.

[6] G. Zhao, C. Qin, H. Yao, and Y. Han, "Dnn self-embedding watermarking: Towards tampering detection and parameter recovery for deep neural network," *Pattern Recognition Letters*, vol. 164, pp. 16–22, 2022.

[7] Z. Yin, H. Yin, and X. Zhang, "Neural network fragile watermarking with no model performance degradation," in *Proceedings of 2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 3958–3962.

[8] M. Botta, D. Cavagnino, and R. Esposito, "NeuNac: A novel fragile watermarking algorithm for integrity protection of neural networks," *Information Sciences*, vol. 576, pp. 228–241, 2021.

[9] X. Guan, H. Feng, W. Zhang, H. Zhou, J. Zhang, and N. Yu, "Reversible watermarking in deep convolutional neural networks for integrity authentication," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2273–2280.

[10] Z. He, T. Zhang, and R. Lee, "Sensitive-sample fingerprinting of deep neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4729–4737.

[11] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 1133–1141.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[13] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[14] R. Zhu, P. Wei, S. Li, Z. Yin, X. Zhang, and Z. Qian, "Fragile neural network watermarking with trigger image set," in *Proceedings of Knowledge Science, Engineering and Management: 14th International Conference*. 2021, pp. 280–293.

Hiding patient information in medical images:A high-capacity and reversible hiding algorithm for E-healthcare

Xiaoyi Zhou*, Shuai Lee†,
*Hainan University, Hainan, Haikou 570228 ,China
Email: xy.zhou@hainan.edu.cn
†Hainan University, Hainan
Email: 21220854000067@hainanu.edu.cn

***Abstract*—The rise of online medical-related technologies has brought great convenience to doctors and patients. However, electronic patient information may be subject to attacks and tampering during medical transmission, leading to compromised information. Therefore, it is necessary to protect the security of data in medical transmission by authenticated electronic patient information. This paper proposes a high-capacity reversible data hiding technology based on interpolated images, which can successfully protect patients' confidential information and verify received content. By using predictors to generate high-quality cover images, and then using data foding strategy to hide electronic patient information into medical images, the entire scheme can be divided into image preprocessing, data processing, and data embedding processes. Image preprocessing includes generating predicted images and performing boundary value on them to generate high-quality covered images. Data processing encrypts the secret data, divides the encrypted data into n blocks of 3 bits, and converts each block into a decimal number. Before embedding the secret data, these numbers are subjected to data foding strategy operation, and data labeling is carried out to ensure the recovery of secret data. The embedding process divides the cover image into 2x2 blocks and embeds secret data in the non-seed pixel diagonal of each block. Last but not least, a fragile watermark is also embedded in the cover image to determine whether the transmitted information has been tampered with. Experimental results show that the method can achieve a high embedding capacity of an average of 2.25 bits per pixel, while the average peak signal-to-noise ratio of the steganographic image reaches 42dB, which is good compared to existing technologies.**

***Keywords*—Electronic patient information (EPI), Interpolation, Data folding strategy (DFS), Predictor, Reversible data hiding**

I. Introduction

E-healthcare has improved healthcare services during pandemics, minimizing disease transmission and reducing morbidity and mortality rates [1]. In order to address the problem of uneven distribution of medical resources, remote healthcare has received extensive attention within the medical industry of the Internet of Things [2]. The security of remote medical data and the protection of patient privacy have always been a key concern [3]–[5]. In the entire remote healthcare system, EPI serves as a medium between doctors and patients, and if there are minor changes it will interfere with the doctor's diagnosis. In this case, researchers have been looking for effective methods to ensure the security of EPI.

Data hiding algorithms are found to be applicable to the remote healthcare system for protecting patient information. The direction of data hiding research can be roughly divided into digital watermarking, steganography and reversible data hiding (RDH) [6], [14]. Due to the special mechanism of RDH, it is impossible to cause any damage to the diagnostic medical images. Therefore, its work and research play an important role in protecting EPI.

II. Literature review

Reversible data hiding algorithms based on image interpolation (IT) has become a new research hot topic. Image interpolation technology is widely used in the field of medical digital image processing [7], [8].A large amount of patient diagnostic information needs to be collected so that doctors can make better diagnoses. Compared with traditional RDH algorithms [15]–[18], IT-based RDH algorithms [13], [19], [20] have the advantages of large single-layer embedding capacity and high visual quality.

Jung et al. [9] first generate interpolated pixels by calculating the average value of neighboring pixel values, this method is called Neighborhood Mean Interpolation (NMI) method, and then embeds secret data of corresponding bit numbers by comparing the relationship between the original pixels and the interpolated pixels. In order to pursue better visual quality, Lee et al. [10] implemented the RDH method through an interpolation of neighboring pixels (INP). A RDH method based on enhanced neighborhood mean interpolation (ENMI) was proposed by Chang et al. [11], which further improves image quality by replacing the diagonal interpolated pixels . At the same time,An improved neighbor mean interpolation (MNMI) method was proposed by Malik et al. [12]. Overall, in terms of visual quality, ENMI and MNMI methods can achieve higher image quality than NMI method. Compared with traditional RDH techniques, IT-based RDH techniques can achieve a high embedding rate of 2.0 or higher because multiple bits of secret data can be embedded in each interpolated pixel block, achieving truly large-capacity secret data embedding. In practical applications, if you need to embed a large amount

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of secret data in the cover image, the IT-based scheme can be said to be an ideal choice.

We can see that most of the reversible data hiding techniques based on interpolation images have disadvantages such as high computational complexity, low efficiency in EPI transmission, and lack of tampering detection capabilities. As a result, we propose an improved interpolation approach to optimize the EPI transmission efficiency while minimizing the impact on image quality. Encoding the data can reduce differences between pixel values, providing better imperceptibility. Tampering detection capabilities are obtained by inserting fragile watermarks in the spatial domain.The main contributions of this paper are:

- An improved predictor is designed for three types of interpolation pixels. The predictor based on the classification idea of this paper effectively retains the edge texture information of the original image.
- A data double-folding mechanism was developed to map the corresponding data to a fixed sequence of shifted data, resulting in less degradation for a given set of hidden data. The main purpose of data mapping is to reduce the difference between pixel values and achieve high imperceptibility

III. Proposed method

The entire embedding process of this method is shown in Figure 1.

First, the original image generates a cover image through interpolation. As in previous research schemes, the cover image was divided into each block of 3×3, and the seed pixels are kept while the non-seed pixels are predicted using the improved predictor method to ensure reversibility. At the same time, the encrypted bitstream to be embedded is divided into 3-bit blocks, and each block is first composed of decimal equivalents of 0, 1, 2, 3, 4, 5, 6, 7, then converted into -1, 0, 1, 2 representations through data folding and tagged. As later sections show, the encoding mapping system can improve the imperceptibility of the steganographic image by further reducing the difference between pixel values. Border conditions need to be imposed on each pixel block to handle overflow pixel values. The proposed method also has tampering detection capability.

A. Predictor

To generate a cover image for hiding secret information, the predictor generates relevant interpolated pixels by referencing the relationships between neighboring pixels. On the one hand, this ensures reversibility, and on the other hand, it also performs well in terms of image fidelity. We re-considered the reference position of the interpolated image and made it an adaptive process.

B. Predictor computation mechanism

In each block, the pixels within each block are divided into two types of pixels due to different uses, namely reference pixels and interpolation pixels. The classification rule is: the original pixels in each block are classified as reference pixels, and other pixels are classified as virtual pixels. The interpolated pixels are calculated from weighted reference pixels and given parameters. The original pixels located in each block are used to calculate the interpolated pixels because they have a strong local correlation with the predicted interpolated pixels.During the calculation process of weight parameters, different labels are assigned to the calculated pixels, as shown in Figure 2. "C" represents the pixel to be interpolated at this time, "D" represents the reference pixel, and "×" represents the pixels that do not affect the generation of the interpolated pixel.

Types 1 and Type 2: As shown in figure 2, if the interpolated pixel has neighboring pixels on its left and right, or above and below, that are original pixels, we classify this as Type 1 and Type 2 respectively.

The formula for predicting pixel C is: C = PxD_1 + (1-P)xD_2

$$P = \begin{cases} \frac{\text{overnum}_1}{\text{num}}, & \text{if } D_1 > D_2 \\ \frac{\text{overnum}_2}{\text{num}}, & \text{if } D_2 > D_1 \\ \frac{1}{2}, & \text{if } D_1 = D_2 \end{cases} \tag{1}$$

In this formula, P is an adaptive parameter whose value is determined by the difference between the interpolated pixel and its neighboring pixels.D_1 refers to the value of the pixel to the left or above the predicted pixel value, and D_2 refers to the value of the pixel to the right or below the predicted pixel value. num represents the number of non-seed pixels around the predicted pixel. $Overnum_1$ and $Overnum_2$ represent the number of non-seed pixels whose values are greater than the D1 and D2 pixel values, respectively.

For Type 3, as shown in Figure , where the interpolated pixel has no neighboring original pixels on its left, right, above, and below, we classify it as Type 3.

The formula for predicting pixel C is:

$$C = \begin{cases} \frac{D_1+D_4}{2}, \text{ if } |D_1 - D_4| < |D_2 - D_3| \\ \frac{D_2+D_3}{2}, \text{ if } |D_1 - D_4| > |D_2 - D_3| \\ \frac{D_1+D_2+D_3+D_4}{4}, \text{ if } |D_1 - D_4| = |D_2 - D_3| \end{cases} \tag{2}$$

A working example of a simple predictor is presented here. The process of interpolating prediction on the original pixel points with grayscale values [53 42 64 31] is shown in Figure 3. The two pixels with values of 53 and 31 are treated as labels D_1 and D_4, respectively, and the two pixels with values of 42 and 64 are treated as labels D_2 and D_3, respectively. Then, the predicted interpolated pixel "C" is calculated based on different scenarios.

C. Prediction of boundary interpolated pixels

Due to insufficient original pixels, the proposed method cannot predict interpolated pixels at edges. Therefore, in this paper, the following special cases will be handled as follows:

Case 1: Interpolate the first and second-to-last rows of the cover image, or interpolate the pixels in the last row of the cover image, where the boundary pixels at this time are the average of the adjacent left and right pixels.

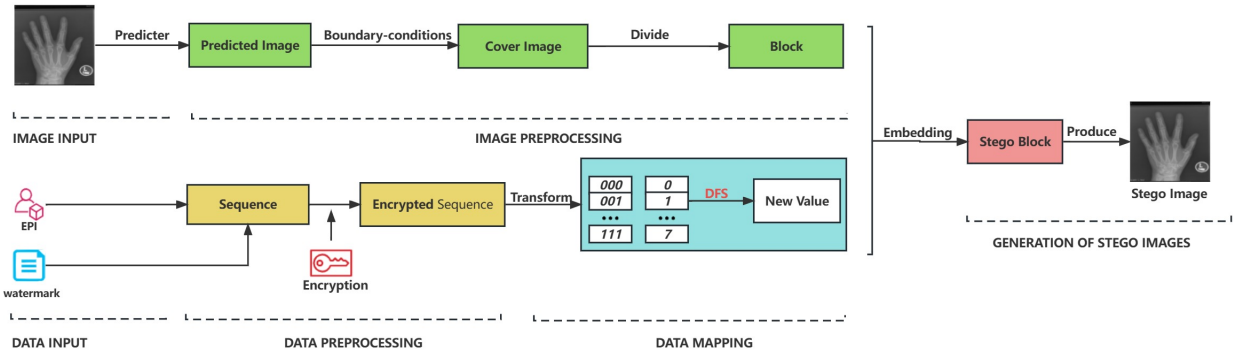


Fig. 1. The flowchart of the proposed scheme.

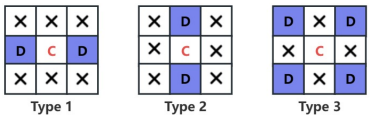


Fig. 2. Different predictor types.

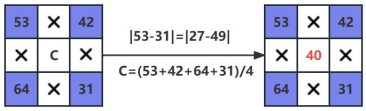


Fig. 3. Pixel prediction using the predictor.

Case 2: Interpolate the pixels in the first and second-to-last columns of the cover image, or interpolate the pixels in the last column of the cover image, where the boundary pixel is the average of the adjacent top and bottom pixels.

After passing the method, we get the cover image that hides the secret data.

D. Encryption and data folding schemes

Security of confidential data is the top priority of telemedicine systems.Various data breaches occur in the medical industry every year, so new mechanisms need to be developed to protect the security and privacy of data. In this scheme, this paper uses the AES-128 encryption scheme to encrypt confidential data to enhance its security. In our scenario, we are using 128-bit key length because it provides faster encryption. After encrypting the data, the previously used method, namely, data left-shift method (LDM), is typically used to embed data into a cover image. Instead, in this paper, we use a data folding scheme (DFS) to optimize the data in the cover image. The scheme enables us to further hide more data while reducing the image degradation level. First, the data D to be embedded is converted into 3-bit blocks, all of which are converted into decimal value data from 0 to 7. Then, these numbers are folded according to a corresponding mapping (such as a formula), and the range of the folded data L is between -4 and 3. Therefore, the data is divided into two parts: the positive part and the negative part. However, we found that the values of each part are still very large, so we need to repeatedly fold each part to obtain new data (New D) with a range of [-2,1] to reduce the difference between pixel values. Embedding data in this way will result in stego images with lower degradation levels compared to the traditional method.

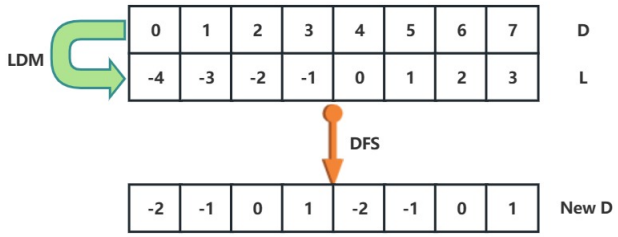


Fig. 4. Proposed implementation of DFS

E. Data Embedding Process

This section will demonstrate embedding confidential data and fragile watermarks in the generated cover image.After being combined with the watermark, the secret data is embedded in the pixels designated on the reverse diagonal while keeping the seed pixels unchanged. For the purpose of authentication,a 64×64 fragile watermark is concatenated with the data to be embedded, with one watermark bit embedded after every 47 bits of confidential data. The purpose of the scheme is to ensure high susceptibility to attacks. See Algorithm 1 for details.

F. Data Extraction Process

This section briefly describes the data extraction process. The data extraction process operates exactly opposite to the data embedding process. The extraction process consists of two parts: extraction of original images and extraction of secret data.The process of extracting the original image, which is done by retaining the seed pixels of the cover image. For data extraction, the algorithm used returns the original values, and

Algorithm 1 Embedding Algorithm.

Require:

secret data SD and Cover image C;

Ensure:

Stego image S;

1: Take n-bits secret data SD and the cover image C.;
2: Divide the cover image of size M×N into 2×2 blocks, each block contains seed pixels and 3 interpolated pixels,where the number of blocks is M×N/4 ;
3: Concatenate watermark bits and the EPI into a sequence ”S” with a size of 6MN. At the same time, select the key K and encrypt the sequence ”S” to get ”E”. The encryption sequence process can be expressed as E=ENC(S,K);
4: Divide the encrypted sequence ”E” into one part every 3 bits, and convert each part from binary to decimal (i.e, 000 through 111 is converted to 0 through 7)
5: Convert these decimal numbers to new value ”X” by data folding strategy;
6: The two decimal-converted secrets are sequentially embedded into the anti-diagonal of each block while preserving the original pixels. The embedding process is: P_i=P+X,where P_i is the pixel value after embedding and P is the pixel value before embedding;
7: Follow steps 1-6 above until you obtain the stego image containing the secret data;

the decrypted information is then obtained by applying the corresponding algorithm.

IV. Results and Discussion

In this section, to evaluate the performance of our proposed algorithm, we randomly tested eight images with a size of 512 × 512 (4 regular images and 4 medical images). The test images shown in Figure 8 are 4 medical images acquired from the Open Access Biomedical Image Search Engine (OPENI) and 4 general images acquired from USC-SIPI. The proposed method extended the images before embedding using a predictor to generate predicted images with a size of 512 × 512. To embed, we used data generated by a pseudorandom function. The experiment was conducted using MATLAB R2014a.

A. The embedding capacity

The embedding capacity is one of the important metrics for reversible data hiding schemes.

The embedding capacity (EC) is defined as follows:

$$EC(bpp) = \frac{\text{The number of bits embedded}}{\text{Total number of pixels}}. \quad (3)$$

Therefore, the method in this paper can hide 2.25bpp after calculation. It should be noted that when fragile watermarks are added, the maximum embeddable EPI is 1.5 bpp.

The payload of the method proposed in this paper has been compared with some related and existing state-of-the-art methods. The results shown in Figure 6 show that our scheme outperforms most techniques in terms of payload.

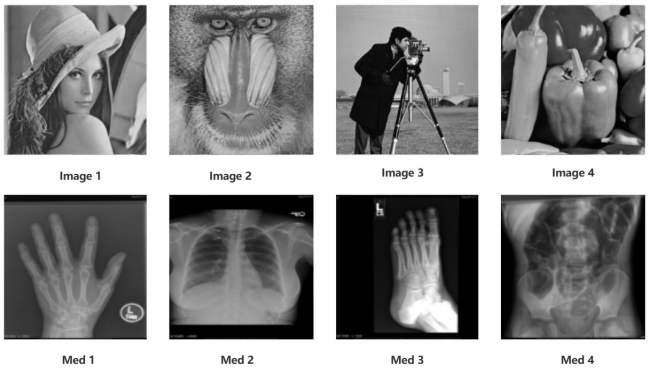


Fig. 5. Test images.

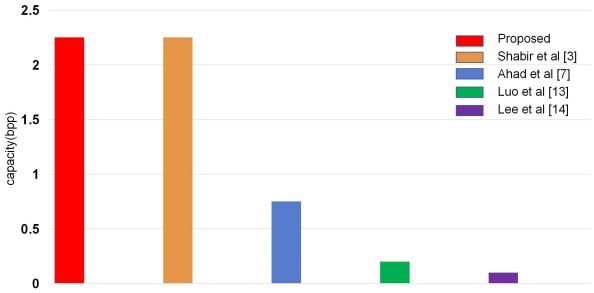


Fig. 6. Payload comparison of images.

TABLE I
Payload(bits), PSNR (dB),SSIM, BER

Images	Payload(bits)	PSNR(dB)	SSIM	BER(%)
Img1	589824	41.96	0.9731	0
Img2	589824	41.88	0.9635	0
Img3	589824	41.75	0.9802	0
Img4	589824	41.94	0.9737	0
Med1	589824	41.99	0.9466	0
Med2	589824	41.92	0.9583	0
Med3	589824	41.96	0.9861	0
Med4	589824	41.99	0.9553	0
Average	**589824**	**41.92**	**0.9671**	**0**

B. imperceptibility analysis

An important criterion for an information embedding system is that even after high-capacity data embedding, the perceptual quality of the image must have a satisfactory result. For imperceptibility, commonly used measurement methods are similarity index (SSIM) and peak signal-to-noise ratio (PSNR).

For a stego image S and an original image C, PSNR is defined as:

$$PSNR \triangleq 10\log_{10}\frac{255^2}{MSE} \quad (4)$$

SSIM is calculated as follows:

$$\text{SSIM} = \frac{(2\mu_{Cl}\mu_S + C_1)(2\sigma_{CS} + C_2)}{(\mu_C^2 + \mu_S^2 + C_1)(\sigma_C^2 + \sigma_S^2 + C_2)} \quad (5)$$

Where C_1 andC_2 are two defined constants. The value of SSIM ranges from 0 to 1. An SSIM value of 0 means that the two images are completely different. At the same time, an SSIM value of 1 means that the two images are identical.

Table 1 shows the average PSNR value of the images to be 41.92 dB, which means the changes in the image are imperceptible to humans through their senses.Table 1 also shows the SSIM values of sample test images, which ranges from 0.9466 to 0.9861. The average SSIM value is 0.9671. The results show that the stego image is almost identical to the cover image.

C. Reversibility analysis

This section discusses the reversibility analysis of the proposed method. The original image and the restored image are shown in Figure 7. The differences between the two images are almost imperceptible to the human eye, but the fact that the pixel values of the original image and the recovered image are exactly the same indicates that our method is completely reversible and the original image can be completely recovered without any loss. We randomly tested two typical images to confirm this, as shown in the figure.

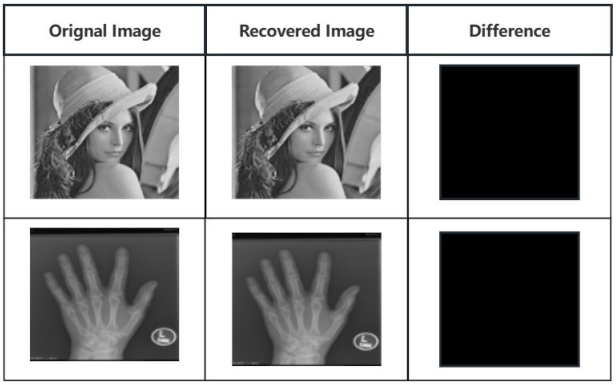


Fig. 7. Reversibility analysis of the method.

D. Authentication Analysis

The ability to extract the same watermark embedded by the sender at the receiver end is the standard for determining whether the real stego image has been intentionally or unintentionally damaged. If a successful extraction is possible at the receiver end, it proves that the real stego image has not been damaged during transmission. Otherwise, the stego image has been damaged during transmission. In addition, If the steganographic image undergoes signal interference or malicious attacks, the recipient cannot successfully obtain the information, and notify the recipient that the received content is invalid. The table shows the extracted information and its associated quality parameters. Bit Error Rate (BER) is generally used to measure the vulnerability of hidden algorithms. The formula is defined as follows.:

$$BER = \frac{\left[\sum_{i=1}^{W}\sum_{j=1}^{H} WM_O(i,j) \oplus WM_R(i,j)\right]}{W \times H} \times 100\% \tag{6}$$

Where H represents the height of the embedded watermark, W represents the width of the embedded watermark; in addition, WM_O is represented by the original watermark, and the restored watermark is represented by WM_R.

This paper uses a 64 $\times$ 64 digital watermark for authentication and vulnerability analysis testing. In this experiment, we randomly selected Img13 from general natural images and Med1 from medical images. Figure 8 shows the extracted digital watermarks from different stego images that have undergone various attacks. The bit error rate is mainly concentrated around 80%, indicating that our scheme is very vulnerable when the image is subjected to various attacks. The experiments show that the proposed method can effectively perform identity authentication and tampering detection.

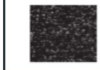
Attacks	Image 1		Med 1	
	Extracted logo	BER (%)	Extracted logo	BER (%)
No Attack		0		0
White Nosie		79.316		80.416
Salt and Pepper		18.981		18.312
JPEG 90		85.641		86.216
Low Pass Filtering		89.213		89.781

Fig. 8. Authentication and fragility analysis for propound scheme.

V. Conclusion

The remote medical patient data hiding scheme needs to consider three key factors: high capacity, low distortion, and reversibility. This paper proposes a reversible data hiding scheme that embeds EPI into medical images to solve the issues of patient data privacy and integrity. By hiding the patient's EPI in the medical image, the scheme provides dual security protection for embedded electronic medical records:The first protection is achieved using AES-128 encryption and then using an information embedding process. A DFS-based scheme is proposed and used to increase the effective payload. Compared with existing methods, the computational cost of spatial domain embedding is lower. The inclusion of fragile watermarking enables our proposed scheme to authenticate the EPI hidden in medical images. Experimental results show that

our scheme has better performance than many state-of-the-art schemes in terms of imperceptibility, effective payload, computational complexity, and tampering detection capability.

ACKNOWLEDGMENT

This work was supported by Hainan Province Key R&D plan project (No. ZDYF2022GXJS224)

REFERENCES

[1] Islam S M R, Kwak D, Kabir M D H, et al. The internet of things for health care: a comprehensive survey[J]. IEEE access, 2015, 3: 678-708.
[2] Chen M, Li W, Hao Y, et al. Edge cognitive computing based smart healthcare system[J]. Future Generation Computer Systems, 2018, 86: 403-411.
[3] Parah S A, Kaw J A, Bellavista P, et al. Efficient security and authentication for edge-based internet of medical things[J]. IEEE Internet of Things Journal, 2020, 8(21): 15652-15662.
[4] Khan S, Abbas N, Nasir M, et al. Steganography-assisted secure localization of smart devices in internet of multimedia things (IoMT)[J]. Multimedia tools and applications, 2021, 80: 17045-17065.
[5] Mansour R F, Parah S A. Reversible data hiding for electronic patient information security for telemedicine applications[J]. Arabian Journal for Science and Engineering, 2021, 46(9): 9129-9144.
[6] Ni Z, Shi Y Q, Ansari N, et al. Reversible data hiding[J]. IEEE Transactions on circuits and systems for video technology, 2006, 16(3): 354-362.
[7] Parah S A, Ahad F, Sheikh J A, et al. A new reversible and high capacity data hiding technique for E-healthcare applications[J]. Multimedia Tools and Applications, 2017, 76: 3943-3975.
[8] Loan N A, Parah S A, Sheikh J A, et al. Hiding electronic patient record (EPR) in medical images: a high capacity and computationally efficient technique for e-healthcare applications[J]. Journal of biomedical informatics, 2017, 73: 125-136.
[9] Jung K H, Yoo K Y. Data hiding method using image interpolation[J]. Computer Standards & Interfaces, 2009, 31(2): 465-470.
[10] Lee C F, Huang Y L. An efficient image interpolation increasing payload in reversible data hiding[J]. Expert systems with applications, 2012, 39(8): 6712-6719.
[11] hang Y T, Huang C T, Lee C F, et al. Image interpolating based data hiding in conjunction with pixel-shifting of histogram[J]. The Journal of Supercomputing, 2013, 66: 1093-1110.
[12] Malik A, Sikka G, Verma H K. A reversible data hiding scheme for interpolated images based on pixel intensity range[J]. Multimedia Tools and Applications, 2020, 79: 18005-18031.
[13] Luo L, Chen Z, Chen M, et al. Reversible image watermarking using interpolation technique[J]. IEEE Transactions on information forensics and security, 2009, 5(1): 187-193.
[14] Lee S, Yoo C D, Kalker T. Reversible image watermarking based on integer-to-integer wavelet transform[J]. IEEE Transactions on information forensics and security, 2007, 2(3): 321-330.
[15] Yu C, Zhang X, Wang D, et al. Reversible data hiding with pairwise PEE and 2D-PEH decomposition[J]. Signal Processing, 2022, 196: 108527.
[16] Mao N, Chen F, He H, et al. Reversible data hiding based on adaptive IPVO and two-segment pairwise PEE[J]. Signal Processing, 2022, 198: 108577.
[17] Kamil Khudhair S, Sahu M, KR R, et al. Secure Reversible Data Hiding Using Block-Wise Histogram Shifting[J]. Electronics, 2023, 12(5): 1222.
[18] Yu C, Zhang X, Li G, et al. Reversible data hiding with adaptive difference recovery for encrypted images[J]. Information Sciences, 2022, 584: 89-110.
[19] Xiong X, Wang L, Li Z, et al. An adaptive high capacity reversible data hiding algorithm in interpolation domain[J]. Signal Processing, 2022, 194: 108458.
[20] Jana M, Joardar S, Jana B. A New Reversible Data Hiding Scheme by Altering Interpolated Pixels Exploiting Neighbor Mean Interpolation (NMI)[M]//Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2022. Singapore: Springer Nature Singapore, 2022: 393-402.

A Visually Meaningful Image Encryption Algorithm with Attention Mechanism and Artificial Bee Colony Optimization

Jiarong Mao, Yuting An and Xiaoyi Zhou*
* Hainan University, Haikou
E-mail: xy.zhou@hainanu.edu.cn Tel: +86-13648616622

Abstract — **The existing encryption algorithms that generate visually meaningful images usually focus on the recovery of secret information, while neglecting the visual quality degradation of the carrier image, thereby increasing the possibility of being intercepted. This paper proposes a visually meaningful image encryption algorithm based on the attention mechanism and artificial bee colony optimization. It combines the attention mechanism in deep learning and the artificial bee colony optimization in optimization algorithms, aiming to maximize the visual effect of visually meaningful images. Firstly, the plaintext image is compressed and scrambled to obtain the secret information. Then the attention mechanism algorithm is applied to divide the cover image into visually salient regions, and the non-salient regions blocks are prioritized for embedding the secret information. By introducing the artificial bee colony optimization algorithm, the optimal values of noise visibility function (NVF), information entropy, and contrast weight are iteratively obtained. On this basis, select the positions and order of the sub-blocks to be embedded and perform IWT and LSB embedding to obtain the visually meaningful cipher images. Experimental results demonstrate that the proposed scheme effectively improves the quality of visually meaningful images.**

Keywords — **Image encryption; Visually meaningful cipher image; Attention Mechanism; Artificial bee Colony.**

I. INTRODUCTION

The visually meaningful image encryption algorithm is an effective method for protecting private images. It enhances certain aspects of visual quality while ensuring the security of the images. The encrypted images produced by this technology have a similar visual perception to the original images, making the ciphered images less suspicious and reducing the risk of interception in communication channels. In visually meaningful images, it is important not only to ensure the security of the secret information but also to minimize the degradation of the carrier image, thereby reducing the risk of channel interception and enhancing the overall security of the encryption algorithm.

To minimize the degradation of the carrier image, the selection of embedding positions for the ciphertext in the carrier image is crucial. Chai et al. [1] utilized the least significant bit (LSB) to minimize changes in the pixel values after embedding. Huang et al. [2] employed integer wavelet transform to embed information in high-frequency regions. Zhu et al. [3] used SVD embedding, modifying smaller singular values to reduce the impact on the carrier image. References [4-5] propose using the information entropy and the noise visibility function as evaluation metrics for each embedding block to determine the priority of embedding. These approaches can effectively enhance the overall visual quality of the carrier image. However, they overlook the prioritization of important aspects in image analysis by the human visual system.

In recent years, attention mechanisms have been effectively used to capture visually salient regions in images by simulating human visual perception. Gan et al. [6] applied attention mechanisms to divide the plaintext image into different regions with varying compression ratios, thereby enhancing the visual quality of the decrypted image. Inspired by this, the carrier image can be partitioned into salient and non-salient regions using the attention mechanism, and the ciphertext can be prioritized for embedding in the non-salient regions. Building upon the work of [4-5], this paper introduces the calculation of contrast [7], which is highly sensitive to the human visual system. The Artificial Bee Colony algorithm (ABC) [8] is then employed to compute the optimal weights for information entropy [9], noise visibility function [10], and contrast. Based on these weights, the selection of sub-blocks for embedding is performed to improve the visual quality of the final generated

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

image. Our main contributions are as follows:

- A visually meaningful image encryption algorithm is proposed, which utilizes the attention mechanism to divide the carrier image into critical and non-salient regions. The ciphertext is prioritized to be embedded in the non-salient regions, enhancing the visual security quality of the encrypted image.
- A method is proposed to iteratively optimize the weights of NVF, information entropy, and contrast using the Artificial Bee Colony algorithm in order to determine the priority of sub-blocks to be embedded and improve the quality of visually secure images.

In the next section, we will introduce the relevant background knowledge, followed by a detailed explanation of the proposed algorithm in Section 3. The simulation results and analysis will be presented in Section 4, and the work will be summarized in Section 5.

II. PRELIMINARIES

A. Attention Mechanism

Attention mechanism was initially introduced as a visual attention model and has since been widely applied and developed in various fields. Reference [11] proposed a fast visual saliency method based on Multi-Scale Difference-of-Gaussians (MDF) in the frequency domain. This method approximates the scale coefficients of image foregrounds one by one by observing the phenomenon that foreground energy in certain bands is prominently displayed and densely distributed in the image. Then, using the linear integral property of Fourier transform, the feature spectrum can be obtained by weighted infinite integration of Gaussian feature maps relative to object scale differences. Finally, the saliency of each channel is obtained from the feature spectrum through inverse Fourier transform and scale filtering method, which is used to segment the focus regions of the image. The effect of the attention mechanism is shown in Figure 1.

B. Artificial Bee Colony algorithm

The Artificial Bee Colony algorithm (ABC) [8] simulates the process of honey bees in selecting the best food source during honey collection. It consists of three groups of bees: employed bees, onlookers, and scouts. They go through three stages iteratively until a predefined maximum number of iterations is reached.

In the initialization stage, all employed bees start as scouts and randomly discover initial nectar sources, representing candidate solutions x_{ij} (decision variables), using Eq.(1), where i = 1; 2,..., NP is the number of population (NP) and j = 1; 2; ..., D is the dimension of decision vector x_{min} and x_{max} are the minimum and maximum bounds of the search space.

$$x_{ij} = x_{min} + rand(0,1) \times (x_{max} - x_{min}) \quad (1)$$

In the employed bee stage, the employed bees generate new solutions mij near their previous solutions based on Eq.(2) and evaluate the fitness value, equation as follows

$$m_{ij} = x_{ij} + \emptyset_{ij} \times (x_{ij} - x_{kj}) \quad (2)$$

where $(k \neq i) \in \{1,2,..., NP\}$ is random index different from i and $j \in \{1,2,..., D\}$; $\varnothing_{ij} \in [-1,1]$ is a random number.

In the onlooker bee stage, the onlooker bees decide their solutions by using roulette wheel selection and then start consuming the selected nectar sources, meaning they generate new m_{ij} around the selected solution using the operators in Eq.(1). The fitness value of the modified m_{ij} is then computed to evaluate the quality of the new nectar source. The better solution m_{ij} replaces the old solution. Meanwhile, if an employed bee exhausts a nectar source, it becomes a scout and is assigned to a brand-new nectar source.

These operations are repeated until the predefined maximum number of iterations is reached, leading to termination of iteration and obtaining the best solution.



(a) (b)

Figure 1: (a) Plaintext image; (b)Attention Mechanism Visualization.

III. PROPOSED ALGORITHM

The algorithm process is shown in Figure 2, and the steps are as follows:

Step 1: Perform compression perception, scrambling, and diffusion on the plaintext image to obtain secret information.

Step 2: Preprocess the carrier image and use the attention mechanism to divide the image into salient regions and non-salient regions.

Step 3: Divide the non-salient regions into blocks, calculate the fitness value, and use ABC iteration to determine the position and order of the embedding block.

Step 4: Embed the secret information obtained in Step 1 into the sub-block obtained in Step 3 using skew transform, and obtain the final visually meaningful image.

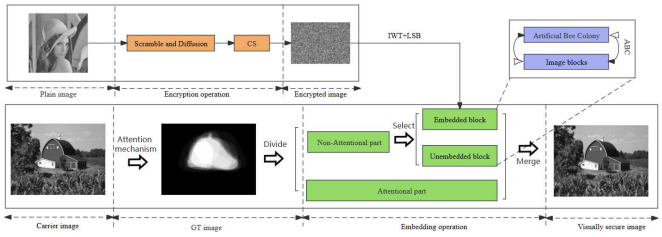


Figure 2: Overall encryption diagram.

A. Image Encryption

In this paper, we use the fully chaotic mapping proposed by Reference [12] to generate chaotic sequences and measurement matrices, whose mathematical expressions are shown in Eq.(3).

$$\begin{aligned} v_{i+1} &= mod((-10 - 9.1uv_i)(e^{9.3} - 6.1w_i),1) \\ w_{i+1} &= mod((-7.9 - 8.0uw_i)(e^{10} - 9.8v_{i+1}),1) \end{aligned} \tag{3}$$

Firstly, the plaintext image is scrambled using Arnold scrambling, and Eq.(4) is used to rearrange and exchange the positions of pixels, which disrupts the relationships between the original pixels,

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 1 & a \\ b & ab+1 \end{bmatrix} \begin{bmatrix} 1 \\ j \end{bmatrix} \tag{4}$$

where

$$p = 1 + aj \tag{5}$$

$$q = b + (ab + 1)j \tag{6}$$

Next, cyclic additive modular diffusion is employed, and Eq.(7) is used to transform the pixel values, spreading the pixel values of the plaintext image throughout the entire image,and increasing the complexity of the encryption.

$$C_i = (C_{i-1} + S_i + P_i) mod\ 256 <<< LSB_3(C_{i-1}) \tag{7}$$

Regarding compression perception, we use chaotic sequences to generate Gaussian random matrices, which are used to measure the encrypted image. The entire compressive sampling process can be described by Eq.(8),whereΦ is the measurement matrix, Φ needs to satisfy the Restricted Isometry Property (RIP)., x is the signal, x=Ψs, Ψ is the sparse basis, and s is the sparse coefficients.

$$\text{y}=\Phi\text{x} = \Phi\Psi \tag{8}$$

In the subsequent reconstruction stage, we use the OMP reconstruction algorithm to reconstruct and restore the encrypted image using the measurement results and Gaussian random matrices, in order to recover the original plaintext image.

B. Image embedding

Attention mechanism preprocessing

Before performing image embedding, the carrier image was preprocessed using the attention mechanism proposed in [11]. Based on the GT map, a threshold is set to divide the plaintext image into salient and non-salient regions, while preserving the positional coordinate information for future restoration. The division of salient regions can also be adjusted according to different needs, and other processing can be performed for different regions to balance the visual quality of the decrypted image with the compression rate. The region division in this paper is illustrated in Figure 3. By setting a threshold, the GT map is traversed from the top, bottom, left, and right directions until the first pixel that exceeds the threshold is encountered. The connected pixels are then linked to form a rectangle. Based on the coordinates, the plaintext image is divided into salient regions and non-salient regions.



(a) (b)

Figure 3 Region Division: (a) Plaintext image region division; (b) GT map region division.

ABC optimization and embedding.

We use the ABC algorithm to select the blocks of the carrier image for embedding, to pursue optimal visual security and image quality. We set up the fitness function, required data, and optimization objectives. The optimization process flow chart is shown in Figure 4.

First, we use PSNR[13-14] and SSIM[15] as target values. PSNR considers the error between corresponding pixel points, but does not consider factors such as the visual sensitivity of the human eye, which may result in inconsistency with subjective perception. In contrast, SSIM measures image similarity from three aspects: brightness, contrast, and structure, and can more comprehensively evaluate image quality. Then, based on the ABC algorithm, we generate optimal weights w1, w2, and w3 for input data of information entropy, noise visibility function, and contrast. The fitness value calculation is performed according to Eq. (9).

$$Fitness = w_1 \times f(NVF) + w_2 \times f(IE) + w_3 \times (D) \tag{9}$$

The formulas for the three performance metrics are as follows:

$$NVF_k = \frac{1}{1 + \sigma_k} \tag{10}$$

$$H = -\sum_{i=0}^{L} p(i) log_2 p(i) \tag{11}$$

$$D = \sum_{\delta} \delta(i,j)^2 P_{\delta}(i,j) \tag{12}$$

where σ_k represents the variance of the k-th sub-image, $p(i)$ represents the probability of occurrence of pixel value x in the image, $\delta(i.j) = |i - j|$represents the difference in gray-scale between adjacent pixels, and $P_{\delta}(i.j)$ represents the probability distribution of pixels with a gray-level difference of δ between adjacent pixels.

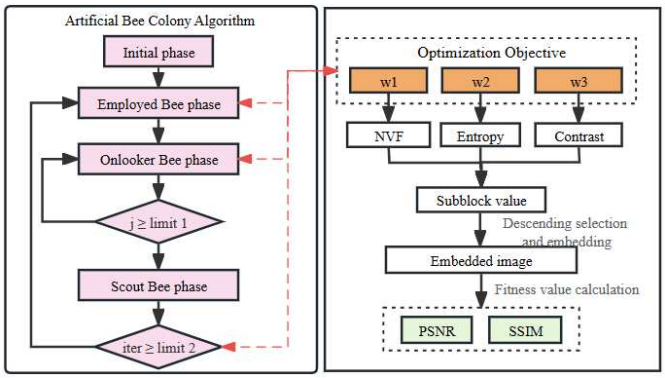


Figure 4 Optimization Process of ABC Algorithm.

Finally, we select the sub-blocks to be embedded in descending order based on their fitness values, and perform IWT (Inverse Wavelet Transform) and LSB (Least Significant Bit) embedding until all secret information has been embedded. The embedding method involves first decomposing the sub-block into four coefficient matrices, namely CA, CH, CV, and CD, through IWT. Then, each pixel value of the secret information is converted to an 8-bit binary representation, and the high two bits, middle three bits, and low three bits are respectively embedded into CH, CV, and CD.

C. Image decryption

The decryption process is reversible and consists of two main steps. Firstly, the secret information is extracted from the stego-image, and secondly, reconstruction is performed using the OMP (Orthogonal Matching Pursuit) algorithm. Then, inverse permutation and diffusion operations are applied to obtain the final decrypted image, D. The detailed steps are as follows:

Step 1: Extract the embedded encrypted information from the carrier image.

Step 2: Reconstruct and restore the extracted encrypted information using the OMP reconstruction algorithm.

Step 3: Perform reverse scrambling and diffusion operations on the image to obtain the decrypted image.

IV. EXPERIMENT

To verify the effectiveness of the proposed image encryption algorithm, simulation experiments were carried out on a desktop computer running Windows 11 (CPU 2.6GHz i7-10750H, RAM 16G) using Matlab 2021b as the simulation platform. In the experiment, we selected the images in the Li-Jian dataset used in Reference [3] as the carrier images, and the standard database Lena image as the plaintext image for testing. Our parameters were selected as: CR=0.7, w1=0.43577, w2=0.40204, w3=0.013402. Figure 5 shows the encryption and decryption effect of the plaintext image, and Figure 6 shows the carrier image and visually meaningful image with region division coordinates at (49, 209) and (304, 464).



Figure 5 : (a) Plain image; (b) Decrypted image without embedding;(c)Decrypted image.



Figure 6 : (a) Carrier image; (b) Visual meaningful image.

To demonstrate the improvement of our algorithm in the quality of visually meaningful encrypted images, we conducted comparative experiments with the sequential embedding method. Firstly, we applied different compression ratios to the Lena image for carrier image embedding, and the results are shown in Table 1. It can be observed that our method effectively enhances the PSNR and SSIM of visually meaningful images. Secondly, we embedded the Lena, Cameraman, Baboon, and Barbara images into the carrier image separately and compared the results with IE-based sorting from Reference [9] and NVF-based sorting from Reference [10]. The outcomes are presented in Table 2. It can be seen that our algorithm provides better decryption quality be seen that our algorithm provides better decryption quality for the private image and improved visual quality for the carrier image compared to other methods.

Tab.1 PSNR and SSIM values of simulation results

CR	Ours		Sequential Embedding		IE-based		NVF-based	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0.3	49.3998	0.9979	49.1598	0.9868	49.2317	0.9975	48.3927	0.8928
0.5	46.9841	0.9937	46.8712	0.9488	46.9680	0.9921	46.8325	0.8911
0.7	45.7533	0.9629	45.4165	0.9310	45.5045	0.9479	45.1753	0.9401

Tab.2 PSNR values between our algorithm and Unoptimized

Plain image	Ours		Sequential Embedding	
	PSNRste	PSNRdec	PSNRste	PSNRdec
Lena	45.7533	18.7115	45.4165	18.4125
Cameraman	46.3920	16.5321	46.0159	15.8011
Baboon	45.9931	17.4352	45.4582	17.2435
Barbara	46.0120	17.2345	45.7364	16.9032

V. CONCLUSIONS

This paper proposes a visually meaningful image encryption algorithm based on attention mechanism and artificial bee colony optimization. Experiments demonstrate that the proposed algorithm can effectively improve the visual effect and robustness of the overall visually secure images. In the future, we will further study the division of carrier images of the same type, so that they can share the same or similar optimized weights to reduce the consumption of optimization time and increase the universality.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 62362025) and the Hainan Province Key R&D plan project (No.ZDYF2022GXJS224).

REFERENCES

[1] X. Chai, H. Wu, Z. Gan, Y. Zhang, Y. Chen, K.W. Nixon, An efficient visually meaningful image compression and encryption scheme based on compressive sensing and dynamic LSB embedding, Opt. Lasers Eng. 124 (2020) 105837.

[2] X Huang, Y Dong , G Ye, et al. Meaningful image encryption algorithm based on compressive sensing and integer wavelet transform[J]. Frontiers of Computer Science, 2023, 17(3):173804

[3] L Zhu, et al. "A robust meaningful image encryption scheme based on block compressive sensing and SVD embedding." Signal Processing 175 (2020): 107629

[4] X Wang, L Cheng, and D Jiang. "A novel visually meaningful image encryption algorithm based on parallel compressive sensing and adaptive embedding." Expert Systems with Applications 209 (2022): 118426.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] D Jiang, et al. "Adaptive embedding: A novel meaningful image encryption scheme based on parallel compressive sensing and slant transform." Signal Processing 188 (2021): 108220.

[6] Z Gan, et al. "Exploiting compressed sensing and polynomial-based progressive secret image sharing for visually secure image selection encryption with authentication." Journal of King Saud University-Computer and Information Sciences 34.10 (2022): 9252-9272.

[7] A. K Tripathi, M Sudipta, and K. D Ashis. "Performance metrics for image contrast." 2011 International Conference on Image Information Processing. IEEE, 2011.

[8] D Karaboga, and B Bahriye. "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm." Journal of global optimization 39 (2007): 459-471.

[9] D. Y Tsai, Y Lee, and E Matsuyama. "Information entropy measure for evaluation of image quality." Journal of digital imaging 21 (2008): 338-347.

[10] Barni, Mauro, B Franco, and P Alessandro. "Improved wavelet-based watermarking through pixel-wise masking." IEEE transactions on image processing 10.5 (2001): 783-791.

[11] W Li, X Yang, C Li, R Lu, and X Xie. Fast visual saliency based on multi-scale difference of gaussians fusion in frequency domain. IET Image Processing, 14(16):4039–4048, 2020.

[12] Toktas, Abdurrahim, and E Uğur. "2D fully chaotic map for image encryption constructed through a quadruple-objective optimization via artificial bee colony algorithm." Neural Computing and Applications (2022): 1-25.

[13] Musanna, Farhan, D Deepak, and K Sanjeev. "A novel chaos-based approach in conjunction with MR-SVD and pairing function for generating visually meaningful cipher images." Multimedia Tools and Applications 79 (2020): 25115-25142.

[14] Shannon, E Claude. "Communication theory of secrecy systems." The Bell system technical journal 28.4 (1949): 656-715.

[15] Hore, Alain, and Z Djemel. "Image quality metrics: PSNR vs. SSIM." 2010 20th international conference on pattern recognition. IEEE, 2010.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

High-Quality Triggers Based Fragile Watermarking for Optical Character Recognition Model

Yujie Yin* and Heng Yin* and Zhaoxia Yin† and Wanli Lyu* and Sha Wei‡
* Anhui Province Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei, China
† Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China
‡ China Academy of Information and Communications Technology
E-mail: *{e21201003, e21301335}@stu.ahu.edu.cn, †zxyin@cee.ecnu.edu.cn, *lwl@ahu.edu.cn, ‡weisha@caict.ac.cn

***Abstract*—Deep neural networks have demonstrated exceptional performance in numerous domains, yet they also face significant security issues. To safeguard the integrity of deep neural networks and prevent unauthorized parameter manipulation, researchers have proposed fragile model watermarking techniques. However, current methods concentrate on image classification models and are not suitable for text image recognition models. Consequently, this paper proposes a remote black-box integrity authentication method for text image recognition models. Specifically, we embed a trigger set into the model to be protected as its watermark. This trigger set is carefully constructed by adding tiny noise to meticulously selected training set samples, with each trigger set sample labeled using model-specific sentence. Utilizing the principle of information entropy theory, we employ an optimal approach to fine-tuning the model, aiming to ensure that the outputs of the trigger set exhibit a pronounced sensitivity to variations within the model. Experiments indicate that with the structural similarity between the trigger set samples and the original samples exceeding 0.9, the embedded watermark has a negligible impact on model performance. The minimum decrease in accuracy on the test set is less than 0.5%. Moreover, even after undergoing modifications limited to fine-tuning only the last two layers and altering a mere one ten-thousandth of the parameters, the model watermarking is effective in identifying such changes.**

I. Introduction

Deep neural networks (DNNs) have made significant contributions to various fields, including computer vision [1]–[3], natural language processing [4]–[6] and automatic driving [7]–[9]. Therefore the security challenges associated with DNNs have also become increasingly critical. One notable concern is backdoor attack [10], wherein hidden patterns are injected into DNNs by manipulating their parameters during the training process. These backdoor patterns do not affect the model performance on regular inputs but trigger unexpected behaviors when specific inputs are provided. Consequently, preventing tampering from attackers and protecting the integrity of DNNs are imperative to ensure their security, reliability, and successful deployment in practical applications.

Over the past few years, researchers have made substantial progress in protecting model integrity by drawing inspiration from the traditional digital watermarking framework [11]. To address the need for preserving the performance of original model while remaining sensitive to parameter modifications, the concept of fragile model watermarking has emerged. Current fragile watermarks can be classified into two categories, one of which is known as white-box watermarking [12]–[15]. This approach needs access to the model and its parameters in order to extract the watermark, presenting challenges in practical scenarios where such access may be limited. The other type of watermark is referred to as black-box watermarking [16]–[20], which differs from the former by enabling extraction solely through querying the model. Although the aforementioned methods adequately fulfill the requirements for protecting DNNs in image classification task, they always disregard other types of models, such as those specifically designed for optical character recognition (OCR) that maps text images to texts.

Generating fragile watermarks for OCR models using existing watermarking methods designed for classification models is a straightforward approach. However, it is challenging to achieve this goal using either white-box or black-box fragile watermarking techniques, owing to the inherent differences between the DNNs employed in these two types of tasks. In contrast to convolutional neural networks (CNNs) commonly employed in classification tasks, OCR models also rely on recurrent neural networks to capture features. If we were to only embed a watermark within the parameters of the convolutional or fully connected layers of an OCR model, similar to the approach employed in classification tasks, it becomes evident that these specific parameters constitute only a small fraction of the overall parameter space. Consequently, the effectiveness of such localized watermarking methods is likely to diminish or even result in failure. In black-box watermarking techniques, the watermark embedding problem usually encompasses two crucial aspects. The first aspect involves constructing an appropriate trigger set that aligns with the output format of OCR task. It is important to emphasize that the OCR task differs significantly from image classification tasks, as it involves handling variable-length text information instead of a single category. Consequently, it becomes essential to carefully adapt the design of the trigger set to accommodate this unique characteristic. Additionally, ensuring that the trigger set exhibits sensitivity to changes in the model is of utmost importance. While fine-tuning

‡ Corresponding author
This research work is partly supported by National Natural Science Foundation of China No. 62172001 and Development of Intelligent Manufacturing Data Resource Public Service Platform(2021-0174-1-1).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the model to minimize the variance between the predicted probabilities of each category in the trigger set is a common approach [17] for classification models to watermark fragile trigger sets, it is not suitable for OCR models. The ability to calculate variance arises from the fact that CNNs produce one-dimensional output vectors. However, in the case of OCR models, the model generates two-dimensional output vectors that include a temporal dimension.

To address the issue of fragile watermarking for OCR models, we propose a novel black-box remote integrity authentication method using a trigger set based watermarking mechanism. Different from existing methods [17][20] that use noisy images as trigger sets, our method needs to add tiny noises to clean samples and design trigger set labels as sentences specified by the model owner. The addition of tiny noise to clean samples aims to prevent the model from denying remote access due to abnormal input. Randomly generated noisy images as trigger sets can easily be distinguished from clean samples by human eyes or machines. In contrast, adding random noise to clean samples can ensure normal query of the remote model while providing enough information to guide the target model to output specific sentences without affecting its original performance. Additionally, to ensure the sensitivity of the trigger set to model changes, we explore the intuition behind zero variance in existing methods. Specifically, we aim to maximize the information entropy of the selected trigger samples, placing them on the decision boundary and achieving the desired model sensitivity. To this end, we extend the 1D variance loss to a generalized information entropy loss, which enables the embedding of fragile watermarks even when the OCR model output is 2D data.

In order to prove the effectiveness of the proposed method, we conduct extensive experiments on the popular OCR model, and present the integrity verification results of the watermarking model in the face of parameter tampering. The main contributions of this paper are as follows:

- To the best of our knowledge, this paper is the first time to focus on the integrity protection of OCR models and proposes a black-box fragile watermarking method based on trigger set.
- The constructed trigger set samples exhibit a high degree of similarity to the clean samples, making them challenging to detect both by machines and human observers.
- Experiments demonstrate the effectiveness of the proposed method, which can effectively distinguish different degrees of tampering.

II. Proposed method

The OCR model we are interested in is the convolutional recurrent neural network (CRNN) [21]. It is a popular framework that forms the basis of subsequent state-of-the-art works on OCR task [22]–[24].

During the recognition process, the CRNN utilizes a combination of CNNs and recurrent neural networks to extract and process features from a given text image. The process can be described as follows: First the CNNs component of the CRNN is responsible for extracting a sequential feature representation $\boldsymbol{X} = x_1, \ldots, x_T$ from an input text image, where T is the length of frames. This feature sequence captures important visual patterns and characteristics of the text. And then $\boldsymbol{X}$ is fed into bi-directional long short-term memory (BiLSTM) layers. The BiLSTM effectively fuses $\boldsymbol{X}$ and learns to capture the contextual information of the character sequence. It takes into account both past and future context of the characters to enhance the understanding of the text. The output of BiLSTM network is a sequence of context features, as a label distribution y_t for each frame x_t in the feature sequence $\boldsymbol{X}$. This sequence is then generated a probability distribution by connectionist temporal classification (CTC) layer [25]. The probability is defined for label sequence $\boldsymbol{L}$ conditioned on the per-frame predictions $\boldsymbol{Y} = y_1, \ldots, y_T$.

The formulation of the CTC can be described as follows: The input is a sequence $\boldsymbol{Y}$. Here, each y_t is a probability distribution over the set where contains all labels in the task and 'blank'. A sequence-to-sequence mapping function B is defined on sequence $\boldsymbol{\pi}$ with length T. B maps $\boldsymbol{\pi}$ onto $\boldsymbol{L}$ by removing the repeated labels and 'blank's. Then, the CTC is defined as the sum of probabilities of all $\boldsymbol{\pi}$ that are mapped by B onto $\boldsymbol{L}$:

$$P(\boldsymbol{L}|\boldsymbol{Y}) = \sum_{\boldsymbol{\pi}:B(\boldsymbol{\pi})=\boldsymbol{L}} p(\boldsymbol{\pi}|\boldsymbol{Y}), \tag{1}$$

where the probability of $\boldsymbol{\pi}$ is defined as $P(\boldsymbol{\pi}|\boldsymbol{Y}) = \prod_{t=1}^{T} y_{\pi_t}^t$, $y_{\pi_t}^t$ is the probability of having label π_t at time stamp t.

Ultimately transcription is the process of converting the per-frame predictions into a label sequence. It selects the character with the highest probability on per-frame to construct the final recognized text.

Inspired by existing integrity-preserving methods for image classification models, we introduce an integrity-preserving method specifically designed for OCR tasks. As shown in Fig. 1, our approach comprises two key steps: watermarking construction and watermarking embedding. A trigger set consists of a set of samples featuring a specific identifier, denoted as the watermark. These samples are accompanied by labels that can either be meaningful or nonsensical sentences. To construct a trigger set containing M samples, we input a batch of clean samples into the original model, meticulously documenting the samples that the model inaccurately recognizes. Subsequently, we introduce tiny perturbation to this batch of samples to yield the candidate trigger set. The information entropy value of each sample within the candidate trigger set is computed and sorted, culminating in the selection of M samples featuring the largest information entropy values. During the phase of watermarking embedding, a synthesis of clean samples and trigger set samples forms the training set. Guided by the alternating application of $Loss_{acc}$ and $Loss_{fra}$, the original model undertakes training with the training set until the watermark is comprehensively embedded. To detect potential tampering of the watermarking model, the

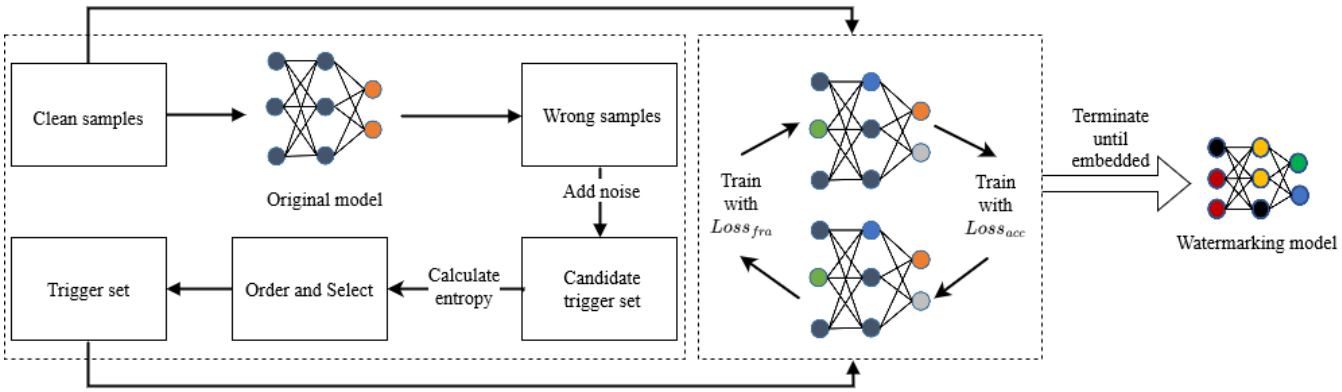


Fig. 1: Overview of the proposed method.

stored trigger set is exclusively required to be inputted into the watermark model. Any model tampering is evident as a distinct drop in recognition accuracy for the trigger set. In the subsequent sections, we provide a comprehensive description of these two steps.

A. Construct trigger set

We wish to generate M trigger set samples with consistent label. Specifically, suppose the target OCR model with parameters θ be denoted as $\mathcal{R}_\theta$, and its training set be represented as $Tra = \{(x_i, y_i) : i = 1, \ldots, N\}$, where x_i, y_i correspond to the input and ground truth, respectively. Let $Wr = \{(wx_i, wy_i) : wx_i, wy_i \in Tra, \mathcal{R}_\theta(wx_i) \neq wy_i, i = 1, \ldots, K\}$ be the set of samples misidentified by $\mathcal{R}_\theta$ in Tra, where $\mathcal{R}_\theta(\cdot)$ represents the output of the model. In order to minimize the impact of the trigger set on the performance of the model, we generate random noise δ by the gaussian distribution F:

$$F(k) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(k-\mu)^2}{2\sigma^2}), \tag{2}$$

where k is the random key, μ is the mean, and σ is the standard deviation. The noise δ has the same number and size as the images in Wr. We then add δ to the corresponding samples in Wr to obtain the candidate trigger set $C = \{(cx_i, cy_i) : cx_i = wx_i + \delta, cy_i = wy_i, i = 1, \ldots, K\}$. Adding triggers to misidentified samples does not result in significant degradation of model performance, as misidentification is an inherent property of the model. However, directly modifying the recognition results of misidentified samples without the guidance of global information can lead to inefficient embedding of trigger sets. In addition, the reason why we do not directly use noise samples as input is that they have a significantly different distribution from training samples and can be easily identified as invalid inputs by humans or machines.

Next, we select M samples from the candidate sample set C as the trigger set, denoted as $Tri = \{(tx_i, ty_i) : i = 1, \ldots, M\}$. Here, ty_i represents the target output for tx_i. Our criterion for Tri selection is to minimize the distance between the trigger set samples and the model decision boundary. In other words, we aim to position the trigger samples near the boundary. By embedding such samples into the model, we ensure that they are sensitive to changes in the boundary and can effectively trigger corresponding responses. Inspired by [26], we can measure the distance between a sample and the model decision boundary by information entropy

$$E(I) = -\sum_{i=1}^{n}\sum_{j=1}^{T} p_i^j(I) \log p_i^j(I), \tag{3}$$

where I and p_i^j indicate input image and the probability of recognizing I as category i at time step j respectively. According to information entropy theory, information entropy is a measure of the uncertainty associated with various possible events occurring in an information source. In the context of OCR models, characters can be considered as information sources, and the information entropy of each character is defined as the uncertainty of being recognized as various possible characters. The information entropy of an image can be defined as the sum of the information entropy of its constituent information sources. It is evident that as the entropy value of a sample increases, the likelihood of being recognized as a fixed label decreases, indicating its proximity to the decision boundary. Therefore, our methodology for selecting the trigger set involves computing the entropy of each sample in C, sorting the results in descending order, and subsequently choosing the top M samples as the trigger set images. The labels assigned to these samples are set as wm, which can be either a randomly generated sentence or a meaningful sentence.

B. Watermarking Embedding

During this phase, we primarily pursue two objectives. The first one is to encourage the target model to change its behavior on the trigger set and output a predefined watermark, while keeping its behavior on the test set unchanged. Secondly, it is necessary to improve the sensitivity of the trigger set to the model, that is, once the model parameters change, the result

of the trigger set should deviate from the expectation. Then we can formalize the above task as minimize the negative log-likelihood of conditional probability of ground truth and maximize information entropy:

$$-\sum_{i=1}^{N} logP(\mathcal{R}_\theta(x_i)|y_i) - \sum_{i=1}^{M} logP(\mathcal{R}_\theta(tx_i)|ty_i), \quad (4)$$

$$\sum_{i=1}^{M} E(tx_i). \quad (5)$$

As Tri is obtained through a certain transformation of Tra, it can be considered as a sample set resulting from data augmentation. Consequently, enhancing the recognition accuracy on both Tri and Tra is not contradictory. We can integrate these two tasks into a unified formula (4). However (4) and (5) are opposite for Tri. Equation (4) aims to maximize the accuracy of each sample in Tri, ensuring that they are correctly recognized as the preset label. On the contrary, Equation (5) requires to maximize the entropy value E of the samples in Tri, as discussed earlier, samples with higher information entropy tend to have less accurate recognition results.

In order to address the conflicts in the optimization process, we alternately train the models with two loss functions respectively:

$$\begin{aligned} Loss_{acc} = -(&\sum_{i=1}^{N} logP(\mathcal{R}_\theta(x_i)|y_i) \\ &+ \sum_{i=1}^{M} logP(\mathcal{R}_\theta(tx_i)|ty_i)), \end{aligned} \quad (6)$$

$$Loss_{fra} = Loss_{acc} + \lambda(K - E(tx_i)), \quad (7)$$

where λ is a hyperparameter that controls the magnitude of sensitivity and K is a constant used to restrict the entropy value within a certain range. During training, we iteratively update the two loss functions in an sequential manner. After each update of the (7) for embedding the watermark, which may result in a decrease in accuracy, it is necessary to update the (6) for accuracy restoration and improvement. We continue this process until the watermark is fully embedded in the model. In other words, the model achieves a predetermined accuracy level on both the trigger set and the test set.

III. Experiments

In this section we use experiments to illustrate the contribution of the proposed method and compare it with other methods in Table I. The following content will begin with an introduction to experimental setup, followed by a series of ablation experiments. Subsequently, we demonstrate the impact of our proposed method on model fidelity, showcasing its effectiveness in verifying integrity. Finally, we present the visual impact of the trigger set.

TABLE I: Comparison among various fragile watermarking methods. Attributes for comparison include: (1) the type of verification method; (2) available for OCR models; (3) imperceptible trigger set samples(if a trigger set is used), emphasizing that our method is a trigger set based black-box watermarking method applicable to OCR models, and the trigger set is agnostic.

Method	Verification method	Available for OCR models	Imperceptible trigger set
ACMMM2020-[12]	White-box	✓	N/A
ACMMM2022-[15]	White-box	✓	N/A
KSEM2021-[17]	Black-box	×	×
KNOSYS2022-[18]	Black-box	×	✓
ICIP2022-[20]	Black-box	×	×
Ours	Black-box	✓	✓

A. Settings

To validate the effectiveness of proposed method, extensive experiments were conducted on Caffe-ocr Chinese synthetic dataset, using the widely adopted CRNN model for text image recognition as a case study. The provided dataset comprises of 3.6 million training set samples, 360,000 test set samples, and has the ability to recognize 5990 characters. Each sample in the dataset is a picture with black or gray characters on a gray background, with a size of $280 \times 32 \times 3$. We fine-tune a pre-trained CRNN model to embed a specific trigger set as its watermark. Since the data set is relatively large and our training resources are limited, we select 4500 samples from the training set for fine-tuning and 5000 samples from the test set for verification. We set the number of trigger sets M to be 36, and add Gaussian noise with mean of 0 and variance of 5 to each sample. During the embedding process, we employ the Adam optimizer with a learning rate of 1e-4 and set λ in (7) to 0.5. Both model and watermarking performance are measured by accuracy. The implementation is based on Python3.8 and the PyTorch deep learning platform. Each experiment is repeated five times to obtain the average value in tables.

B. Ablation Experiment

In this section, we demonstrate the effectiveness of our proposed method through an ablation experiment. We compare three different watermarked models, including: (1) NL model, which utilizes randomly selected normal samples from the training set to add noise as the trigger set and embeds it with only $Loss_{acc}$. (2) N2L model, which utilizes the same trigger set as NL but embeds it by with $Loss_{acc}$ and $Loss_{fra}$. (3) The complete model watermarking, constructed through our proposed method.

To ensure fair experimentation, we have modified the training termination conditions for the three models used in the watermark embedding process. Unlike the method mentioned earlier, we no longer rely on the accuracy of the test set and trigger set to meet a preset value as the end of training. Instead,

TABLE II: Recognition accuracies (%) on test set and trigger set for different watermark models after 10 epochs of watermark embedding. The values in parentheses represent the decreased accuracy compared to clean model.

Model	Recognition Accuracy (%)	
	Test Set	Trigger Set
Clean Model	88.40	0.00
NL Model	86.23(2.17)	100.00
N2L Model	86.21 (2.19)	33.33
Proposed Model	88.07(0.33)	100.00

we judge the end based on whether the specified training epochs have been reached. It is important to note that all other experimental settings remain unchanged, except for the differences in trigger set selection and training process. Table II illustrates the recognition accuracies on test set and trigger set for different watermark models after 10 epochs of watermark embedding.

Based on the results in Table II, we can observe the NL model which without fragile training can identify trigger sets with 100% accuracy after 10 rounds. This is because the constructed trigger set samples can be regarded as data augmentation of training set samples, so the model can quickly learn new samples. However, when NL model is subjected to tampering, regardless of the severity of the tampering, the accuracy of the NL model rapidly decreases to 0. We will illustrate this situation in detail later Fig. 2. Next we found the N2L model achieves a lower trigger set accuracy of 33.3% compared to the NL model after the same number of training epochs. This decrease in the accuracy of trigger set can be attributed to the inherent conflict between $Loss_{fra}$ and $Loss_{acc}$ for the trigger set as discussed in II-B. We can vividly understand that in the same length of time the NL model primarily emphasizes enhancing the recognition accuracy of trigger set, while the N2L model not only aims to improve the accuracy of trigger set but also strives to move it closer to the decision boundary, lagging behind in the accuracy of trigger set. However, our proposed watermarking model uses samples that are naturally located near the decision boundary to construct trigger sets. As a result, proposed model can reduce the required training time for $Loss_{fra}$ and quickly memorize trigger set samples.

Fig. 2 depicts the decline in trigger set accuracy across various watermarking models as the epochs of fine-tuning progressively increases. Fine-tuning is a common practice in model transform. In this section, we focus on fine-tuning embedding layer of the model to minimize any impact on the accuracy of test set. Proposed watermarking model demonstrates high accuracy in identifying trigger sets with limited extent of parameter modification, while the watermarking accuracy of the other two models decreases significantly after fine-tuning. From the perspective of the magnitude of decline, the NL model exhibits a significantly larger decrease compared to the N2L model. The reason behind this can be attributed to the lack

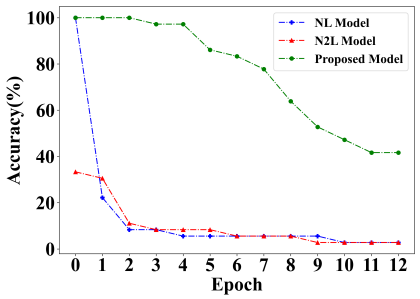


Fig. 2: Recognition accuracies (%) on trigger set across various watermarking models as the epochs of fine-tuning progressively increases.

TABLE III: Maximum recognition accuracies (%) achieved on test set and number of training epochs required for embedding watermark.

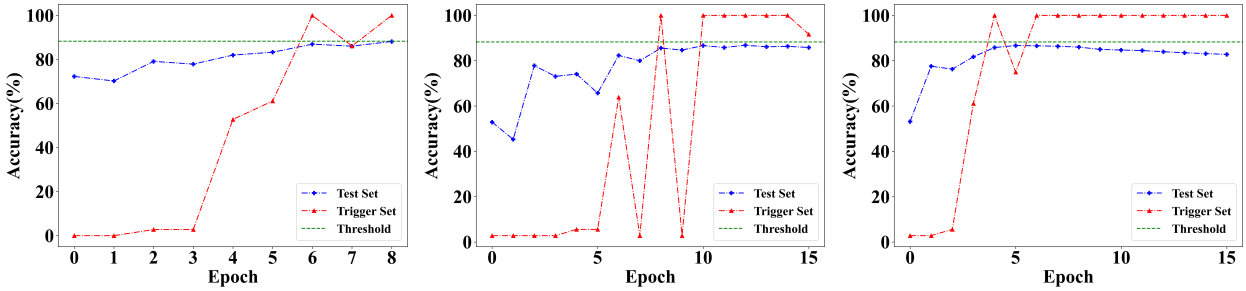
Model	Training Epoch	Recognition Accuracy (%)
NL Model	9	85.71
N2L Model	15	86.39
Proposed Model	9	87.84

of $Loss_{fra}$ training in the NL model. The sensitivity observed in the NL model is primarily a consequence of overfitting to the trigger set, which hinders its ability to effectively differentiate model tampering.

C. Fidelity Evaluation

Fidelity refers to the amount of performance degradation that occurs in the clean model as a result of watermark embedding. In this section, we demonstrate that our proposed method has a limited impact on the performance of clean model. We set the accuracy of test set to drop no more than 0.3% during the watermark embedding process, and the accuracy of trigger set is greater than 99.9%. Fig. 3 respectively shows the changes in recognition accuracy of proposed watermark model, NL model and N2L model on test set and trigger set with the increase of embedding epochs.

It is evident that proposed watermark model only needs 8 epochs to effectively integrate a watermark into the model, while the model accuracy remains at 88.4%, which is only about 0.2% lower than the clean model accuracy. While for the other two baseline models NL and N2L, we found that even if the training epochs are increased to 16, they could not achieve the desired accuracy on the test set. We also show in Table III the maximum accuracy achieved on the test set and the number of training epochs required for embedding watermark under multiple trials. Compared to the other two models, the proposed watermark model achieves higher recognition accuracy with fewer training epochs. This shows that the proposed method is efficient and high-fidelity.



(a) Proposed Model (b) N2L Model (c) NL Model

Fig. 3: Accuracy of test set and trigger set recognized by different models with the increase of embedding epochs. The green dotted line in indicates the test set accuracy value we set.

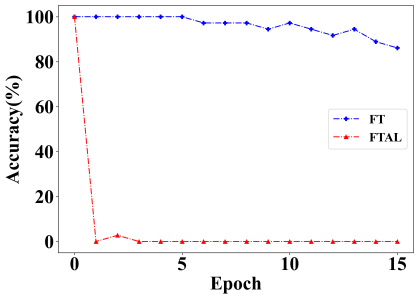


Fig. 4: Recognition accuracies (%) on trigger set for models with different types of fine-tuning as the training epochs increases.

TABLE IV: Recognition accuracies (%) on test set and trigger set for models tampered with different proportions (‰) parameters.

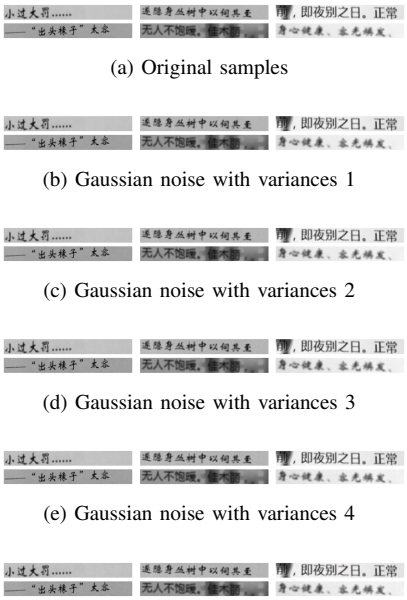
Proportions(‰)	0.1	1	10
Test Set	86.78	86.64	67.14
Trigger Set	19.44	11.11	2.78

D. Sensitivity Comparison

In this paper, sensitivity refers to the changes in the accuracy of trigger set when watermarked model is subjected to model tampering such as fine-tuning and least significant bit replacement (LSBR) [10].

We considered two extreme scenarios for fine-tuning. The first is to freeze all parameters, but only update the last layer parameters (FT) during back-propagation; In the second scenario, we conducted fine-tuning with changing all parameters (FTAL). We show the experimental results of these two different methods in Fig. 4. It can be seen that when limited parameters are modified, the watermark still maintains a high accuracy due to the small magnitude of changes. But when the parameters are significantly tampered, the watermark accuracy is close to zero. In practice, we usually freeze most of the parameters and only adjust a small number of weights.

LSBR is a backdoor embedding method, where the least



(a) Original samples

(b) Gaussian noise with variances 1

(c) Gaussian noise with variances 2

(d) Gaussian noise with variances 3

(e) Gaussian noise with variances 4

(f) Gaussian noise with variances 5

Fig. 5: Original samples and trigger set samples constructed by adding gaussian noise with different variances respectively.

TABLE V: Similarity metrics between trigger set samples and original samples as the amount of noise increases.

Metrics\σ^2	1	2	3	4	5
PSNR	45.113	39.923	35.375	31.921	29.020
SSIM	0.996	0.985	0.962	0.930	0.901
LPIPS	0.001	0.008	0.021	0.033	0.048

significant bit of parameters is replaced by a secret string. In our study, we manipulate parameters at varying proportions and evaluate the recognition accuracy of the trigger set for the manipulated model, as presented in Table IV. It is worth noting that even with a minimal modification of only 0.1‰to the original parameters, we are able to reliably detect it, as the accuracy of trigger set drops significantly to less than 20% in this case.

TABLE VI: Training epochs required for embedding trigger set and corresponding recognition accuracies (%) of test set under various trigger sets with different noises.

σ^2	1	2	3	4	5
Training Epoch	12	12	10	10	10
Recognition Accuracy (%)	85.42	86.04	86.90	87.30	88.17

E. Perceptual Evaluation

This section evaluates the visual quality of the trigger set samples constructed on Caffe-ocr Chinese synthetic dataset from the perspectives of human eyes and machines, and also studies the impact of the trigger sets obtained after adding different sizes of noise on the performance and effect of watermark embedding. Fig. 5 shows original samples and trigger set samples constructed by adding Gaussian noise with a variance of 1-5 using the proposed method, respectively. It can be seen that the generated trigger set samples are very similar to the original samples.

In Table V, we quantitatively evaluate the similarity between trigger set samples and original samples using three metrics: peak signal-to-noise ratio (PSNR) [27], structural similarity (SSIM) [28] and learned perceptual image patch similarity (LPIPS) [29]. We consider a PSNR value greater than 30, an SSIM value closer to 1, and an LPIPS value closer to 0 as indications of higher similarity between the samples. The data in the Table V all meet the conditions. In addition, we also show in Table VI the epochs required to fully embed different trigger sets into the model and maximize the test set accuracy, and the corresponding test set accuracy. We found that the changes in the noise will affect the fidelity. Trigger sets with larger noise exhibit a greater difference from the test set distribution, resulting in less impact on test set performance. Regardless of the visual quality, all trigger sets are successfully and efficiently embedded into the model.

Therefore, we can conclude that the trigger set samples utilized in this paper exhibit a comparable visual effect to original samples when embedding watermarks, in contrast to the previous method of embedding watermarks using trigger sets. Additionally, even in trigger pattern with small noise added, these samples demonstrate high embedding efficiency.

IV. Conclusions

This paper presents a trigger-set based fragile model watermarking approach for OCR tasks. The trigger sets, constructed using misclassified samples, are naturally situated near the decision boundary. We utilize two-dimensional information entropy to guarantee their susceptibility to tampering. Our experiments demonstrate that the proposed method can embed fragile watermarks with negligible loss while ensuring the generation of high-quality trigger sets. Furthermore, the method can detect various tampering techniques. Moreover, in the future, we can consider optimizing the trigger set during the model fine-tuning process to enhance the invisibility and pertinence of disturbances.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[4] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[5] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1422–1432.

[6] Z. Jiang, M. Yang, M. Tsirlin, R. Tang, Y. Dai, and J. Lin, ""low-resource" text classification: A parameter-free classification method with compressors," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 6810–6828.

[7] A. Casanova, G. Cucurull, M. Drozdzal, A. Romero, and Y. Bengio, "On the iterative refinement of densely connected representation levels for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 978–987.

[8] Y. Wang, Z. Xu, H. Shen, B. Cheng, and L. Yang, "Centermask: Single shot instance segmentation with point representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9313–9321.

[9] C. Badue, R. Guidolini, R. V. Carneiro, *et al.*, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113 816, 2021.

[10] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2017. arXiv: 1708.06733.

[11] X. Zhang and S. Wang, "Fragile watermarking with error-free restoration capability," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1490–1499, 2008.

[12] X. Guan, H. Feng, W. Zhang, H. Zhou, J. Zhang, and N. Yu, "Reversible watermarking in deep convolutional neural networks for integrity authentication," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2273–2280.

[13] M. Botta, D. Cavagnino, and R. Esposito, "Neunac: A novel fragile watermarking algorithm for integrity

protection of neural networks," *Information Sciences*, vol. 576, pp. 228–241, 2021.

[14] A. Abuadbba, H. Kim, and S. Nepal, "Deepisign: Invisible fragile watermark to protect the integrity and authenticity of cnn," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021, pp. 952–959.

[15] C. Xiong, G. Feng, X. Li, X. Zhang, and C. Qin, "Neural Network Model Protection with Piracy Identification and Tampering Localization Capability," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2881–2889.

[16] Z. He, T. Zhang, and R. Lee, "Sensitive-Sample Fingerprinting of Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2019, pp. 4724–4732.

[17] R. Zhu, P. Wei, S. Li, Z. Yin, X. Zhang, and Z. Qian, "Fragile Neural Network Watermarking with Trigger Image Set," in *International Conference on Knowledge Science, Engineering and Management*, Springer, 2021, pp. 280–293.

[18] D. P. Kuttichira, S. Gupta, D. Nguyen, S. Rana, and S. Venkatesh, "Verification of integrity of deployed deep learning models using Bayesian Optimization," *Knowledge-Based Systems*, vol. 241, p. 108 238, 2022.

[19] Y. Lao, W. Zhao, P. Yang, and P. Li, "Deepauth: A dnn authentication framework by model-unique and fragile signature embedding," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 9, pp. 9595–9603, 2022.

[20] Z. Yin, H. Yin, and X. Zhang, "Neural network fragile watermarking with no model performance degradation," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 3958–3962.

[21] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[22] J. Baek, G. Kim, J. Lee, *et al.*, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4715–4723.

[23] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9809–9818.

[24] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7098–7107.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[26] J. Shlens, M. B. Kennel, H. D. Abarbanel, and E. J. Chichilnisky, "Estimating information rates with confidence intervals in neural spike trains," *Neural computation*, vol. 19, no. 7, pp. 1683–1719, 2007.

[27] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*, IEEE, 2010, pp. 2366–2369.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Coupled Transformed Induced Tensor Nuclear Norm for Robust Tensor Completion

Mengjie Qin*†, Zheyuan Lin*†, Minhong Wan*†, Chunlong Zhang*†, Jason Gu*†‡, Te Li*†
* Research Center for Intelligent Robotics, Research Institute of Interdisciplinary Innovation,
Zhejiang Lab, Hangzhou 311100, China.
† Zhejiang Engineering Research Center for Intelligent Robotics, Hangzhou 311100, China.
‡ Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS Canada B3H 4R2.

***Abstract*—This study focuses on exploring the problem of low-rank tensor completion (LRTC), where the main challenge lies in accurately representing the inherent structures of the underlying data. Typical approaches exhibit either suboptimal performance due to the lack of suitable priors or combine many constraints, resulting in increased model complexity. In this paper, we propose a robust transformed t-SVD by utilizing the discrete cosine transform in conjunction with a unitary transformation matrix. Nonetheless, the t-SVD-based framework exhibits a deficiency in its capacity to flexibly explore diverse correlations across tensor modes. To address this issue, we employ the representation of the tensor as a series of multi-dimensional unfolding tensors, which fully capture the embedded structure in the original data. Moreover, on the basis of this tensor representation, we construct a transformed t-SVD method within the LRTC model and solve it by the alternating-direction method of multiplier (ADMM) approach. Extensive experiments demonstrate that the proposed model outperforms existing state-of-the-art methods.**

I. Introduction

Over the past decade, machine vision has undergone significant advancements, transitioning from "visible" to "clearly visible", and currently reaching an "understanding" intelligent stage. However, challenges persist in scenarios characterized by fragmented visual perception, such as adverse weather conditions and limited incident energy, which still impose limitations on the generation of high-quality images. These limitations greatly diminish the visual quality and directly affect the performance of subsequent processing steps. This motivation drives us to develop more efficient tensor completion (TC) [1], [2] models specifically designed to complete unknown elements based on observational element information. The overarching concept entails the characterization and utilization of meaningful correlations among distinct elements.

Recently, many prior-based TC techniques have been extensively explored in the realm of visual inpainting. These main methods can be categorized into two principal groups: data-driven based methods[3], [4] and conventional properties-representation-based methods. The first group of methods is based on the intrinsic non-linear representation capacities of deep learning, which exhibit outstanding performance in image representation. However, these approaches fundamentally rely on the diversity and quantity of the training dataset, necessitating substantial time investment in fine-tuning a multitude of network parameters. Furthermore, the effectiveness of convolution neural network-based approaches is particularly susceptible to the specific missing rates present in the relevant data. As a result of this limited adaptability, these methods cannot be directly extended to address the overarching TC challenges encountered across diverse samples.

Another kind of algorithm, properties representation-based methods, follows a principled strategy to demonstrate the potential structure implicitly hidden in incomplete data. In this work, we focus on this kind of algorithm due to its interpretability and direct applicability in singleton scenarios. Focusing on the tensor representation, this technique aims to approximate higher-order tensors by utilizing a set of underlying factors, enabling the transformation of original large-scale data into a lower-dimensional space. Researchers have extensively studied and proposed various tensor factorization methods by investigating the diverse structures of potential factors and their multiple linear representations. Among these, Tucker decomposition [5] and CANDECOMP/PARAFAC (CP) decomposition [6] have garnered significant attention and are considered the most classical models for tensor decomposition in recent decades. Recently, new tensor decomposition techniques such as Tensor Train (TT) decomposition [7] and Tensor Ring (TR) decomposition [8] have emerged. These methods have demonstrated strong compression capabilities for higher-order large-scale tensors. Furthermore, it is challenging to pre-specify the optimal rank in the TT-rank or the TR-rank. Particularly within tensor decomposition models, where their ranks are defined as vectors, identifying the optimal rank becomes a complex endeavor due to the potential for extensive cross-validation.

Indeed, more effective definitions of tensor rank, including the tensor tubal rank and multi-rank, have been garnering an increasing amount of research attention. They can be computed and validated through tensor singular value decomposition (t-SVD). The t-SVD model was originally introduced by Kilmer et al.[9]. It emerged as a method for decomposing three-way tensors within the framework of the tensor-tensor product (t-product). In contrast to alternative tensor decomposition meth-

This work is supported in part by the Key Research Project of Zhejiang Lab (No. G2021NB0AL03), and in part by the Youth Foundation Project of Zhejiang Lab (No. K2023NB0AA03).
Corresponding author: Te Li, lite@zhejianglab.com

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

ods, t-SVD has the advantage of exploring spatially shifted correlations. With t-SVD, a three-way tensor is regarded as a cohesive entity, mitigating the inherent information loss associated with matrices or the flattening of tensors.

Inspired by the above results, Hu et al. [10] introduced a twist tensor nuclear norm that relaxed the multi-rank tensor within the Fourier domain. Similarly, Lu et al.[11] consider TNN based on T-product and t-SVD, and theoretically analyze the convergence of the model. Nevertheless, owing to the limitations of Fourier representation, previous TNN-based approaches have shown inadequacies in characterizing the attributes of the data. Other existing works have focused on exploring transformations more suitable than the Fourier transform, aiming for a more effective depiction of data correlations, *e.g.*, Fast Fourier Transform (FFT)[11], discrete cosine transform (DCT)[12], [13], and invertible linear transform (ILT)[14]. Applying t-SVD in other domains is highly practical and interesting, as it not only enhances the low-rank prior but also has the potential to reduce computational costs.

To cope with the mentioned limitations and leverage the benefits of transformation-based operations, we propose utilizing the DCT and unitary transform as a unified transformation techniques within the t-SVD framework. Precisely, this transformation redistributes the tubal components of the tensor into matrix slices, thereby enabling the application of SVD on each transformed matrix slice. Therefore, we propose a coupled transformed induced TNN, termed CTTNN, founded on the preference for DCT-based unitary transformations. This innovation yields superior approximation of tensor rank. Importantly, in this paper, we focus on the application of TNN-based method to higher-order tensors. To improve the applicability of TNN in high-dimensional data, we employ mode-k_1k_2 unfolding operator to fully represent the different modes along higher dimensional tensors. The contributions of this article can be highlighted as follows.

- With the comparative analysis of different transformation techniques, we argue that the low-rank property within unitary representation preferred over others is relatively stricter. On this basis, we propose a unified transformation-based t-SVD approach and introduce the CTTNN model that combines the DCT and unitary transformation within a unified framework.
- We consider the correlations among different modes and transform them into corresponding mode-unfolded tensors, which are superior in preserving the intrinsic structure of tensors.
- To address the proposed approach, we introduce the guaranteed convergent alternate direction method of multipliers (ADMM) for model optimization. Extensive experiments indicate that our proposed model outperforms the state-of-the-art methods.

II. Tensor Notations and Definitions

Generally, an N-order tensor is denoted as $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$, and its $(i_1, i_2, \ldots, i_N)$-th component is denoted as $x_{i_1 i_2 \ldots i_N}$. For a better reading, we summarize the basic notations of tensor in Table. I. Following, we introduce some necessary definitions of tensor representations.

TABLE I
TENSOR NOTATIONS.

Nation	Description
x, $\boldsymbol{X}, \boldsymbol{\mathcal{X}}$	scalar, matrix, tensor
$\boldsymbol{\mathcal{X}}_{(i_1,i_2,\ldots,i_N)}$	The $(i_1, i_2, \ldots, i_N)$th element of $\boldsymbol{\mathcal{X}}$
$\text{unfold}_k(\boldsymbol{\mathcal{X}})$, $\boldsymbol{\mathcal{X}}_{(k)}$	the mode-k unfolding of $\boldsymbol{\mathcal{X}}$
$\text{fold}_k(\boldsymbol{X}_{(k)})$	the fold operation along the mode-k
$\|\boldsymbol{\mathcal{X}}\|_F$	the Frobenius norm of $\boldsymbol{\mathcal{X}}$, $\|\boldsymbol{\mathcal{X}}\|_F = \sqrt{\sum_{i_1,\ldots,i_N} (\boldsymbol{\mathcal{X}}_{i_1 \ldots i_N})^2}$
$*$	tensor-tensor product
$\hat{\boldsymbol{\mathcal{X}}}$, $\bar{\boldsymbol{\mathcal{X}}}$	the DFT and DCT results of $\boldsymbol{\mathcal{X}}$
$\boldsymbol{\mathcal{X}}^H$	the conjugate transpose of $\boldsymbol{\mathcal{X}}$
$\boldsymbol{X}^{(i)}$	the i-th frontal slice of $\boldsymbol{\mathcal{X}}$(three-order)

Definition 1. *(tensor mode-k unfolding and folding[15]): Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ be an N-order tensor, its mode-k unfolding matrix is denoted as $\boldsymbol{X}_k \in \mathbb{R}^{I_k \times \prod_{s \neq k} I_s}$, where $(i_1, i_2, \ldots, i_N)$-th element of $\boldsymbol{\mathcal{X}}$ maps to (i_k, j)-th element of $\boldsymbol{X}_k$,*

$$j = 1 + \sum_{s=1, s \neq k}^{N} (i_s - 1) J_s, J_s = \prod_{n=1, n \neq k}^{s-1} I_n. \tag{1}$$

The corresponding mode-k unfolding operator and its inverse operator are respectively denoted as unfold_k *and* fold_k*, i.e.,* $\boldsymbol{\mathcal{X}} = \text{fold}_k(\text{unfold}_k(\boldsymbol{\mathcal{X}})) = \text{fold}_k(\boldsymbol{X}_{(k)})$.

Definition 2. *(identity tensor[16]): For a identity tensor $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{I_1 \times I_1 \times I_3}$ whose first frontal slice is $I_1 \times I_1$ identity matrix ($\boldsymbol{I}_{I_1}$) and others are all zeros.*

Definition 3. *(orthogonal tensor[16]): For a orthogonal tensor $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{I_1 \times I_1 \times I_3}$ which satisfies*

$$\boldsymbol{\mathcal{Q}}^H * \boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{Q}} * \boldsymbol{\mathcal{Q}}^H = \boldsymbol{\mathcal{I}}, \tag{2}$$

where $\boldsymbol{\mathcal{I}}$ is the identity tensor.

Definition 4. *(t-product[17]): Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_2 \times I_4 \times I_3}$, the t-product $\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}}$ is defined as*

$$\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}} = \text{fold}(\text{bcirc}(\boldsymbol{\mathcal{X}}) \cdot \text{unfold}(\boldsymbol{\mathcal{Y}})) \in \mathbb{R}^{I_1 \times I_4 \times I_3}, \tag{3}$$

where $\cdot$ is the matrix product. When $I_3 = 1$, t-product simplifies to matrix multiplication.

III. Proposed Model

A. Coupled Transformed t-SVD Framework

For a three-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, owing to the circular convolution in t-product, its t-SVD can be efficiently computed with DFT. Using the DFT matrix $\boldsymbol{F}_{I_3}$, its Fourier transformed tensor can be formulated as

$$\hat{\boldsymbol{\mathcal{X}}} = \text{fold}_3(\boldsymbol{F}_{I_3} \boldsymbol{X}_{(3)}) \in \mathbb{R}^{I_1 \times I_2 \times I_3}, \tag{4}$$

where $\boldsymbol{X}_{(3)}$ is mode-3 unfolding matrix of $\boldsymbol{\mathcal{X}}$.

Next, by further analysis, we adopt the DCT as a substitute for the DFT. DCT is one of the most standard and widely used transformations, its transform matrix $\boldsymbol{F}_{dct} \in \mathbb{R}^{I_3 \times I_3}$ is defined as:

$$\boldsymbol{F}_{dct} = \alpha_k \sqrt{\frac{2}{I_3}} \cos(\theta_{k,n}), \theta_{k,n} = \frac{\pi}{I_3} k(n + \frac{1}{2}) \tag{5}$$

where the ranges of the indexes k and n are $0 \leq n, k \leq I_3 - 1$, $\alpha_0 = \frac{1}{\sqrt{2}}$, and $\alpha_k = 1, k \neq 0$.

Similar to the DFT, the transformed tensor can be formulated as $\bar{\boldsymbol{\mathcal{X}}} = \text{fold}_3(\boldsymbol{F}_{\text{dct}} \boldsymbol{X}_{(3)})$. Let $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} *_{\text{dct}} \boldsymbol{\mathcal{S}} *_{\text{dct}} \boldsymbol{\mathcal{V}}^T$ be the t-SVD based DCT with $*_{\text{dct}}$, we denote the tensor nuclear norm of $\boldsymbol{\mathcal{X}}$ as $\|\boldsymbol{\mathcal{X}}\|_{\text{TNN}}$,

$$\|\boldsymbol{\mathcal{X}}\|_{\text{TNN}} = \|\bar{\boldsymbol{X}}\|_*, \tag{6}$$

where $\|\cdot\|_*$ represents the matrix nuclear norm. For a $I_1 \times I_2$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_* = \Sigma_i^{min(I_1, I_2)} \sigma_i$, σ_i is the i-th singular value of matrix $\boldsymbol{A}$.

In the above discussion, we define t-SVD based, where $\boldsymbol{F}_{dct}$ is an arbitrary invertible matrix. In the following, we introduce the t-product based on unitary transformation for property augmentation.

Let $\boldsymbol{\Phi} \in \mathbb{R}^{I_3 \times I_3}$ be a unitary transform matrix with $\boldsymbol{\Phi}\boldsymbol{\Phi}^H = \boldsymbol{\Phi}^H\boldsymbol{\Phi} = \boldsymbol{I}_{I_3}$. $\bar{\boldsymbol{\mathcal{X}}}_{\Phi}$ denotes a three-order tensor obtained by multiplying via $\boldsymbol{\Phi}$ on all slices along three-mode of $\boldsymbol{\mathcal{X}}$, we have

$$\bar{\boldsymbol{\mathcal{X}}}_{\Phi}(i, j, :) = \boldsymbol{\Phi}(\boldsymbol{\mathcal{X}}(i, j, :)). \tag{7}$$

For simplicity, we denote $\bar{\boldsymbol{\mathcal{X}}}_{\Phi} = \boldsymbol{\Phi}[\boldsymbol{\mathcal{X}}]$, the corresponding inverse transform is $\boldsymbol{\mathcal{X}} = \boldsymbol{\Phi}^H[\bar{\boldsymbol{\mathcal{X}}}_{\Phi}]$.

The t-product based on $*_{\boldsymbol{\Phi}}$ is consistent to t-product based on DCT with $*_{\text{dct}}$. Therefore, the definition about t-product with $*_{\boldsymbol{\Phi}}$ is omitted. In addition, we have the following t-SVD induced by unitary transform, that is,

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} *_{\boldsymbol{\Phi}} \boldsymbol{\mathcal{S}} *_{\boldsymbol{\Phi}} \boldsymbol{\mathcal{V}}^H. \tag{8}$$

Based on t-SVD framework with DCT, we propose a more suitable unitary matrix $\boldsymbol{\Phi}$. The main motivation is that the data represented by DCT can be better approximated with lower rank than other tranforms. Since, the unitary matrix $\boldsymbol{\Phi}$ can be computed by $\boldsymbol{\Phi} = \boldsymbol{U}_{\text{dct}}$, where $\boldsymbol{U}_{\text{dct}}$ is SVD factor of matrix $\bar{\boldsymbol{X}}_{(3)}$. The above details are summarized in Algorithm 1.

Recalling Eq.(6), the tensor nuclear norm of $\boldsymbol{\mathcal{X}}$ equals to the matrix nuclear norm of its block diagonal form in the transformed domain, which motivates us to ropose a coupled transformed induced TNN (CTTNN) based on the unitary transformation preferred by DCT, that is,

$$\|\boldsymbol{\mathcal{X}}\|_{\text{CTTNN}} = \sum_{i=1}^{I_3} \|\bar{\boldsymbol{X}}_{\Phi}^{(i)}\|_*. \tag{9}$$

For an in-depth understanding of Algorithm 1, its advantages can be analyzed by comparing various transform representations. Illustrating this with the instance of an MSI "fake and real beers," where 70% of the pixels are missing. The original multispectral image (MSI) data can be represented as $\boldsymbol{\mathcal{X}}$, its DCT [18], FFT, Hankelization[19], ILT[14], and our tensor are represented by distinct variables. In Figure 1, we depict the distribution of singular values for the frontal slices of diverse transformation tensors. Additionally, we delineate the proportional allocation of singular values within each amplitude interval to offer further differentiation. Interestingly, a notable concentration of singular values derived from our proposed transformed tensor lies within the intervals $[10^{-3}, 10^{-2})$, $[10^{-4}, 10^{-3})$, and $[0, 10^{-4})$. On the contrary, the majority of singular values resulting from alternative transformations predominantly span the intervals $[10, \infty)$, $[1, 10)$, and $[10^{-1}, 1)$. This observation implies that the data representation afforded by Algorithm 1 can be more precisely approximated through a diminished rank.

Algorithm 1 Robust Transformed t-SVD.

Input: $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$.

Output: t-SVD components $\boldsymbol{\mathcal{U}}$, $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{V}}$.

$\bar{\boldsymbol{\mathcal{X}}} = \text{fold}_3(\boldsymbol{F}_{\text{dct}} \boldsymbol{X}_{(3)})$
$[\boldsymbol{U}_{\text{dct}}, \boldsymbol{S}_{\text{dct}}, \boldsymbol{V}_{\text{dct}}] = \text{SVD}(\bar{\boldsymbol{X}}_{(3)})$
$\boldsymbol{U}_{\text{dct}} \rightarrow \boldsymbol{\Phi}$
$\bar{\boldsymbol{\mathcal{X}}}_{\Phi} = \boldsymbol{\Phi}[\boldsymbol{\mathcal{X}}]$
for $i = 1, \cdots, I_3$
$[\bar{\boldsymbol{U}}_{\Phi}^{(i)}, \bar{\boldsymbol{S}}_{\Phi}^{(i)}, \bar{\boldsymbol{V}}_{\Phi}^{(i)}] = \text{SVD}(\bar{\boldsymbol{X}}_{\Phi}^{(i)})$
end for
$\boldsymbol{\mathcal{U}} = \boldsymbol{\Phi}^H[\bar{\boldsymbol{U}}_{\Phi}^{(i)}]$,$\boldsymbol{\mathcal{S}} = \boldsymbol{\Phi}^H[\bar{\boldsymbol{S}}_{\Phi}^{(i)}]$,$\boldsymbol{\mathcal{V}} = \boldsymbol{\Phi}^H[\bar{\boldsymbol{V}}_{\Phi}^{(i)}]$

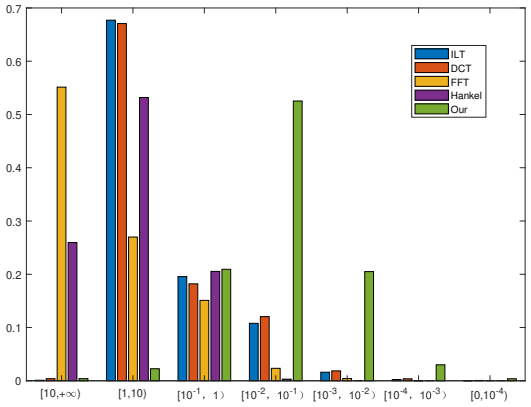


Fig. 1. The distribution of singular values. Here, the singular values are derived from various transformation tensors.

B. Completion using CTTNN and Unfolding Tensors

Based on the above analysis, t-SVD can effectively preserve the underlying structure of tensors. However, the framework of t-SVD is only appropriate to three-order tensors and lacks the flexibility necessary to handle different correlations along different modes. From this perspective, we consider the t-SVD in multi-mode unfolding tensors, which are obtained by

mode-k_1k_2 unfolding operator. Mathematically, the operation of mode-k_1k_2 tensor unfolding is a process of stacking all mode-k_1k_2 slices of the k_1k_2 order tensor sequentially into 3D tensors, whose definition is given in [15].

For simplicity, we define the operation of mode-k_1k_2 tensor unfolding as $\boldsymbol{\mathcal{X}}_{(k_1k_2)} = \text{unfold}_{(k_1k_2)}(\boldsymbol{\mathcal{X}}, (k_1, k_2))$, and its inverse operator is $\boldsymbol{\mathcal{X}} = \text{fold}_{(k_1k_2)}(\boldsymbol{\mathcal{X}}_{(k_1k_2)}, (k_1k_2))$. Take a three-order as an example, the mode-k_1k_2 tensor unfolding operator rearranges the tensor slices, i.e.,

$$\begin{aligned}\boldsymbol{\mathcal{X}}(I_1, I_2, I_3) &= \boldsymbol{\mathcal{X}}_{(12)}(I_1, I_2, I_3) = \boldsymbol{\mathcal{X}}_{(13)}(I_1, I_3, I_2) \\ &= \boldsymbol{\mathcal{X}}_{(23)}(I_2, I_3, I_1)\end{aligned} \tag{10}$$

Taking a 3D tensor as an example, a graphical diagram is shown in Fig. 2. More intuitively, it can also be seen as different perspectives of the data cube. In other words, the modek_1k_2 unfolding tensors can preserve the essential feature of each material, because the modek_1k_2 tensor unfolding operator can learn multiple bundles of signature related to each dimension by reshaped tensors.

Based on the above superiority of the representation tensors, we formulate a novel LRTC model with CTTNN under unfolding tensors, which is as follows:

$$\begin{aligned}&\min_{\boldsymbol{\mathcal{X}}} \sum_{1\le k_1<k_2\le I_N} \|\boldsymbol{\mathcal{X}}_{(k_1k_2)}\|_{\text{CTTNN}} \\ &s.t. \boldsymbol{\mathcal{P}}_\Omega(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}) = 0,\end{aligned} \tag{11}$$

where $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_1\times I_2,\ldots,\times I_N}$ is the observed tensor, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1\times I_2,\ldots,\times I_N}$ is the underlying tensor, $\boldsymbol{\mathcal{P}}_\Omega$ is the binary observation index tensor. Synthesizing the powerful data representation capability of unfolding tensors and property augmentation with transformed t-SVD, the proposed model can achieve better performance in the LRTC problem.

C. Optimization Scheme

In this subsection, we develop the well-known alternate direction multiplier (ADMM) [20] algorithm to solve the model (11). By introducing an auxiliary variable $\boldsymbol{\mathcal{M}}_{(k_1k_2)} = \boldsymbol{\mathcal{X}}$, the problem (11) can be formulated as the equivalent constrained optimization problem

$$\begin{aligned}&\min_{\boldsymbol{\mathcal{X}}} \sum_{1\le k_1<k_2\le I_N} \|\boldsymbol{\mathcal{M}}_{(k_1k_2)}\|_{\text{CTTNN}} + \iota_{\mathbb{S}}(\boldsymbol{\mathcal{X}}) \\ &s.t. \boldsymbol{\mathcal{M}}_{(k_1k_2)} = \boldsymbol{\mathcal{X}},\end{aligned} \tag{12}$$

where

$$\iota_{\mathbb{S}}(\boldsymbol{\mathcal{X}}) = \begin{cases} 0 & if\ \boldsymbol{\mathcal{X}} \in \mathbb{S} \\ \infty & if\ \text{otherwise}, \end{cases} \tag{13}$$

and $\mathbb{S} := \{\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{P}}_\Omega(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}) = 0\}$. Then the corresponding augmented Lagrangian function can be expressed as

$$\begin{aligned}&\mathcal{L}_{\beta_{k_1k_2}}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}_{(k_1k_2)}, \Lambda_{k_1k_2}) = \\ &\sum_{1\le k_1<k_2\le I_N} \|\boldsymbol{\mathcal{M}}_{(k_1k_2)}\|_{\text{CTTNN}} + \iota_{\mathbb{S}}(\boldsymbol{\mathcal{X}}) + \\ &\frac{\beta_{k_1k_2}}{2}\|\boldsymbol{\mathcal{M}}_{(k_1k_2)} - \boldsymbol{\mathcal{X}} + \frac{\Lambda_{k_1k_2}}{\beta_{k_1k_2}}\|_F^2,\end{aligned} \tag{14}$$

where $\Lambda_{k_1k_2}(1 \le k_1 < k_2 \le I_N)$ are Lagrange multipliers and $\beta_{k_1k_2}$ are positive penalty parameters. Within the ADMM iteration system, problem (12) can be solved by iterating the target variable with other fixed variables. Specifically, in the t-th iteration, subproblems $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{M}}_{(k_1k_2)}$ and $\Lambda_{k_1k_2}$ are alternately updated as follows:

$$\begin{cases} \boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)} = \arg\min_{\boldsymbol{\mathcal{M}}_{(k_1k_2)}} \mathcal{L}_{\beta_{k_1k_2}}(\boldsymbol{\mathcal{X}}^t, \boldsymbol{\mathcal{M}}_{(k_1k_2)}, \Lambda^t_{k_1k_2}) \\ \boldsymbol{\mathcal{X}}^{t+1} = \arg\min_{\boldsymbol{\mathcal{X}}} \mathcal{L}_{\beta_{k_1k_2}}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)}, \Lambda^t_{k_1k_2}) \\ \Lambda^{t+1}_{k_1k_2} = \arg\min_{\Lambda_{k_1k_2}} \mathcal{L}_{\beta_{k_1k_2}}(\boldsymbol{\mathcal{X}}^{t+1}, \boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)}, \Lambda_{k_1k_2}) \\ \beta_{k_1k_2} = \min(\rho\beta_{k_1k_2}, \beta_{k_1k_2\max}) \end{cases} \tag{15}$$

where $\rho > 1$ denotes a predetermined constant of the penalty.
1) $\boldsymbol{\mathcal{M}}_{(k_1k_2)}$ **sub-problem**: The $\boldsymbol{\mathcal{M}}_{(k_1k_2)}$ at the t-th step is

$$\begin{aligned}\boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)} =& \arg\min_{\boldsymbol{\mathcal{M}}_{(k_1k_2)}} \|\boldsymbol{\mathcal{M}}_{(k_1k_2)}\|_{\text{CTTNN}} \\ &+ \frac{\beta_{k_1k_2}}{2}\|\boldsymbol{\mathcal{M}}_{(k_1k_2)} - \boldsymbol{\mathcal{X}} + \frac{\Lambda_{k_1k_2}}{\beta_{k_1k_2}}\|_F^2.\end{aligned} \tag{16}$$

Similarly to the proximal mapping of the nuclear norm of the matrix, following the tensor singular value thresholding (t-SVT)[21] theorem. The problem (16) can obtain the closed-form solution as,

$$\boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)} = \text{t-SVT}(\mathcal{D}_{\frac{1}{\beta_{k_1k_2}}}(\boldsymbol{\mathcal{X}} - \frac{\Lambda_{k_1k_2}}{\beta_{k_1k_2}}), k_1, k_2), \tag{17}$$

where $\mathcal{D}(\cdot)$is a t-SVT operator.

2) $\boldsymbol{\mathcal{X}}$ **sub-problem**: By fixing $\boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)}$, $\Lambda^t_{k_1k_2}$, the $\boldsymbol{\mathcal{X}}^{t+1}$ can be given by

$$\begin{aligned}\boldsymbol{\mathcal{X}}^{t+1} =& \arg\min_{\boldsymbol{\mathcal{X}}} \iota_{\mathbb{S}}(\boldsymbol{\mathcal{X}}) + \\ &\frac{\beta_{k_1k_2}}{2}\|\boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)} - \boldsymbol{\mathcal{X}} + \frac{\Lambda_{k_1k_2}}{\beta_{k_1k_2}}\|_F^2,\end{aligned} \tag{18}$$

which is differentiable and its closed-form solution can be obtained by

$$\begin{cases} \boldsymbol{\mathcal{P}}_\Omega(\boldsymbol{\mathcal{X}}^{t+1}) = \boldsymbol{\mathcal{P}}_\Omega(\boldsymbol{\mathcal{Y}}) \\ \boldsymbol{\mathcal{P}}_{\Omega^c}(\boldsymbol{\mathcal{X}}^{t+1}) = \boldsymbol{\mathcal{P}}_{\Omega^c}\left(\frac{\sum_{1\le k_1<k_2\le I_N} \beta_{k_1k_2}\left(\boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)} - \frac{\Lambda^t_{k_1k_2}}{\beta_{k_1k_2}}\right)}{\sum_{1\le k_1<k_2\le I_N} \beta_{k_1k_2}}\right), \end{cases} \tag{19}$$

where Ω^c denotes the complement of the index set Ω.

3) $\Lambda_{k_1k_2}$ **sub-problem**: According to the ADMM iteration standard, we update the multiplier $\Lambda_{k_1k_2}$ by

$$\Lambda^{t+1}_{k_1k_2} = \Lambda^t_{k_1k_2} + \beta_{k_1k_2}(\boldsymbol{\mathcal{M}}^{t+1}_{(k_1k_2)} - \boldsymbol{\mathcal{X}}^{t+1}). \tag{20}$$

The whole pseudocode of the ADMM iteration strategy to solve problem (12) is illustrated in Algorithm 2.

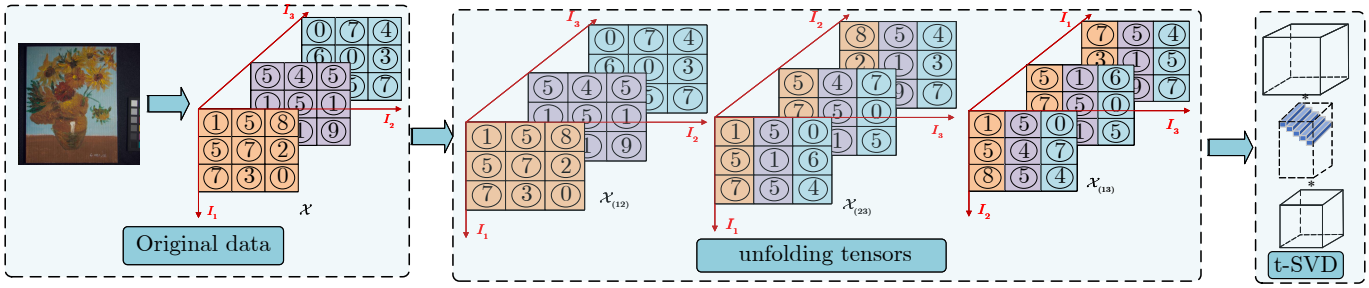


Fig. 2. Illustration of CTTNN, which contains two components: unfolding tensors (middle) and t-SVD (right).

Algorithm 2 ADMM for solving (12).

Input: The incomplete tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2, \ldots, \times I_N}$, the binary observation index tensor $\mathcal{P}_\Omega$, the positive penalty parameters $\beta_{k_1 k_2}$, the predetermined constant ρ, the maximum number of iterations T.

Output: the completed tensor $\mathcal{X}_\Omega$.

1: **Initialization**: $\mathcal{X}_\Omega = \mathcal{Y}_\Omega$, $\mathcal{X}_{\Omega^c} = 0$, $\mathcal{X}^0_{(k_1 k_2)} = 0$, $\mathcal{M}^0_{(k_1 k_2)} = 0$, $\rho = 1.1$, $T = 500$;
2: **while** not converged and $t \leq T$ **do**:
3: update $\mathcal{M}_{(k_1 k_2)}$ by solving (17);
4: update $\mathcal{X}$ by solving (19);
5: update $\Lambda_{k_1 k_2}$ by solving (20);
6: update $\beta_{k_1 k_2} = \min(\rho \beta_{k_1 k_2}, \beta_{k_1 k_2 \max})$;
7: **end while**

IV. Experimental Results

To illustrate the effectiveness of the proposed CTTNN model, extensive tensor completion experiments on images under two types of tasks, missing-pixel restoration. In these experiments, all Multispectral Images (MSIs) are resized to dimensions of 256×256 pixels, encompassing 31 spectral bands. The entire array of experimental images is depicted in Figure 3. Meanwhile, several state-of-the-art and representative methodologies are enlisted: TRLRF [22], FCTN-TC [23], TNN [21], HFTNN [19], and WSTNN[15]. Among these, TNN follows a scheme grounded in tSVD, while HFTNN and WSTNN are refinements of TNN that employ distinct tensor representation strategies. TRLRF and FCTN-TC are based on tensor decomposition techniques, utilizing the tensor ring and fully-connected tensor ring for low-rank approximation, respectively. To ensure equitable evaluation, the parameters of each competing method are meticulously fine-tuned within the bounds suggested by the original authors, aiming for optimal outcomes. All experimental results are produced in Matlab2020a, and the hardware environment is AMD Ryzen 7 4800H CPU and 16GB RAM.

In this subsection, we apply all methods to six MSIs from the Columbia MultiSpectral Image database[1], i.e., flowers, thread spools, fake and real beers, oil painting, stuffed toys,

[1]http://www.cs.columbia.edu/CAVE/databases/multispectral/.



(a) (b) (c) (d) (e) (f)

Fig. 3. Samples of ground truth, from left to right are (a) flowers, (b) thread spools, (c) fake and real beers, (d) oil painting, (e) stuffed toys, (f) fake and real strawberries.

and fake and real strawberries. To assess the completion results of all methods, we perform visual inpainting by randomly selecting positions at various ratios, spanning from 10% to 30%. Table II reports the MPSNR and MSSIM quality metrics of each method on all MSIs. The best results in the table are bolded. As can be seen from Table II, our proposed CTTNN method outperforms most of the competitors across nearly all scenarios, indicating that the recovered MSI by our model exhibits spatial and spectral structures that are closer to the original data. The non-bold outcomes of CTTNN are reflected in the MPSNR and MSSIM metrics for the oil painting and tuffed toys. Despite this, our results maintain a strong second position and exhibit a marginal difference from the top-performing result.

For visual comparison, the recovered MSI on of thread spools at a 10% sampling rate is presented in Figure 4. This illustration depicts the reconstructed MSIs at band 10, along with the corresponding error images at band 10 from various competing methods. From this depiction, the initial observation highlights that TRLRF yields a subpar visual impact, characterized by a blurry and desaturated recovered image lacking in semantic image details. In contrast, other competing techniques such as FCTN-TC and TNN struggle to recover all the elements, displaying noticeable errors. The outcomes of WSTNN and HFTNN present marginally better acceptability, while our novel CTTNN approach accurately estimates nearly all pixels. In the magnified thumbnail showcasing pseudo-color images, the proposed model distinctly delineates target contours of the mountain. Concurrently, the error images confirm that the discrepancy between the recovered outcome by CTTNN and the ground truth is minimal. Evidently, in terms of visual quality, the results attained by the proposed CTTNN model underscore the advantages of our holistic approach, which

TABLE II

COMPARISON OF THE RECOVERED RESULTS ON MSIS DATA SETS WITH DIFFERENT MISSING RATES. THE OPTIMAL VALUES OF MPSNR AND MSSIM, FOR EACH TEST, ARE HIGHLIGHTED IN BOLD.

Test image	SR	Evaluation metrics	Observed	TRLRF	FCTN-TC	TNN	WSTNN	HFTNN	CTTNN
fake and real beers	10%	MPSNR	9.889	34.306	42.827	37.762	41.133	39.709	**42.452**
		MSSIM	0.035	0.933	0.982	0.957	0.990	0.976	**0.991**
	20%	MPSNR	10.400	36.570	45.853	43.522	48.138	45.288	**47.788**
		MSSIM	0.051	0.957	0.990	0.987	0.997	0.991	**0.996**
	30%	MPSNR	10.983	37.621	47.001	47.386	52.331	49.281	**52.403**
		MSSIM	0.065	0.962	0.992	0.994	0.998	0.995	**0.997**
oil painting	10%	MPSNR	13.913	31.749	32.634	31.747	**35.937**	35.627	35.658
		MSSIM	0.126	0.828	0.868	0.864	**0.951**	0.948	0.945
	20%	MPSNR	14.426	36.243	35.081	36.400	40.215	38.560	**41.659**
		MSSIM	0.184	0.958	0.916	0.941	0.977	0.932	**0.979**
	30%	MPSNR	15.007	37.540	36.245	40.193	43.260	40.231	**45.759**
		MSSIM	0.238	0.948	0.933	0.971	0.986	0.970	**0.994**
stuffed toys	10%	MPSNR	13.589	30.704	**38.721**	32.155	37.748	36.500	38.455
		MSSIM	0.302	0.801	0.953	0.882	0.967	0.949	**0.969**
	20%	MPSNR	14.106	32.812	42.278	38.490	43.237	42.595	**44.343**
		MSSIM	0.365	0.831	0.976	0.961	0.986	0.983	**0.993**
	30%	MPSNR	14.682	39.654	43.796	43.248	45.53	46.622	**48.569**
		MSSIM	0.425	0.958	0.424	0.985	0.989	0.991	**0.997**
fake and real strawberries	10%	MPSNR	15.256	35.352	39.582	36.724	40.092	39.18	**40.723**
		MSSIM	0.276	0.925	0.947	0.923	0.970	0.951	**0.977**
	20%	MPSNR	15.776	36.706	42.949	41.441	44.582	42.267	**45.998**
		MSSIM	0.357	0.922	0.974	0.968	0.985	0.980	**0.992**
	30%	MPSNR	16.359	37.397	44.453	45.080	46.751	46.301	**48.643**
		MSSIM	0.430	0.932	0.981	0.984	0.989	0.990	**0.994**
thread spools	10%	MPSNR	15.257	35.352	39.582	35.506	39.527	39.180	**39.680**
		MSSIM	0.276	0.925	0.947	0.923	0.968	0.951	**0.971**
	20%	MPSNR	15.772	36.706	40.074	42.949	43.364	42.267	**43.874**
		MSSIM	0.357	0.922	0.974	0.967	0.988	0.980	**0.990**
	30%	MPSNR	16.352	37.397	44.453	43.674	46.872	46.301	**47.214**
		MSSIM	0.430	0.932	0.981	0.984	0.992	0.990	**0.995**
flowers	10%	MPSNR	17.075	32.578	36.045	32.072	37.305	38.133	**38.533**
		MSSIM	0.428	0.787	0.885	0.838	0.945	0.955	**0.966**
	20%	MPSNR	17.599	38.991	38.852	36.89	42.066	43.183	**44.276**
		MSSIM	0.486	0.939	0.933	0.926	0.974	0.986	**0.989**
	30%	MPSNR	18.169	40.471	40.959	40.567	44.369	47.354	**48.753**
		MSSIM	0.540	0.962	0.954	0.961	0.981	0.994	**0.966**

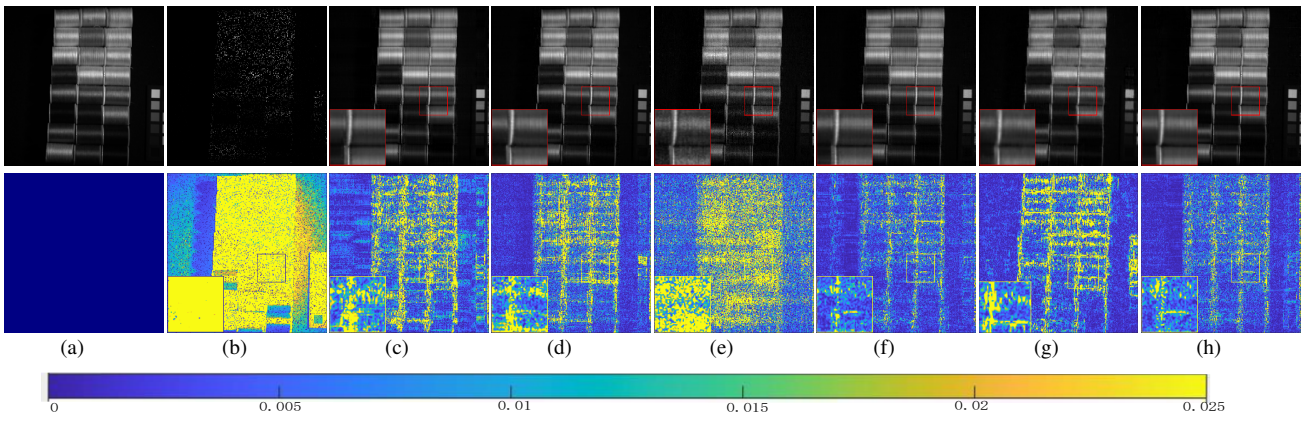


Fig. 4. Visual results of all comparative algorithms on thread spools at band 10. (a) Original image. (b) Observed data. (c)-(h) Recovered images from TRLRF, FCTN-TC, TNN, WSTNN, HFTNN and CTTNN, respectively. The second row presents the corresponding error images of the first row.

intelligently incorporates enhanced low-rank properties and the utilization of unfolding tensors.

Efficiency of execution stands as another crucial indicator for real-world applicability. In this context, Figure 5 provides a visual representation of the average execution times of all competing models across varying random missing pixel rates applied to the six recovered MSIs. Notably, the efficiency of TRLRF and TNN outperforms that of the other algorithms considerably. As computational requirements escalate, FCTN-TC, WSTNN, HFTNN, and CTTNN demonstrate comparatively diminished efficiency due to their unique data expansion strategies aimed at enhancing data properties. This outcome conforms to the reasoning that such strategies inherently introduce greater computational intricacies. However, when compared to HFTNN, CTTNN manages to sustain an acceptable level of execution efficiency. Moreover, as a counterbalancing facet, CTTNN accentuates the reinforcement of low-rank attributes and the resilience exhibited across diverse bands of MSIs.

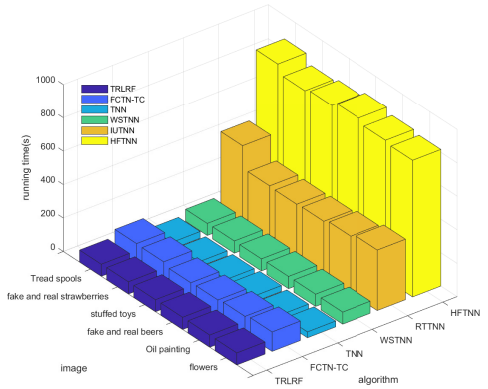


Fig. 5. The execution time of all algorithms for MSIs recovery.

V. CONCLUSION

We propose a novel tensor completion model, CTTNN, which amalgamates the enhanced low-rank property of t-SVD with the multi-dimensional tensor unfolding technique. In this model, the constituents of higher-order tensors undergo a reconstruction process into a sequence of three-way tensors. Moreover, we extend the t-SVD framework and devise the LRTC model through the incorporation of a conjoint transform-induced TNN. Subsequently, the ADMM algorithm optimizes the proposed model, yielding a systematically derived solution at each iterative step. Extensive experimental results validate that the CTTNN method surpasses other cutting-edge approaches in terms of both qualitative and quantitative measures.

REFERENCES

[1] H. Wu, K. Fu, Y. Zhao, H. Song, and J. Li, "Joint self-supervised and reference-guided learning for depth inpainting," *Computational Visual Media*, pp. 1–16, 2022.

[2] T. K. Sinha, J. Naram, and P. Kumar, "Nonnegative low-rank tensor completion via dual formulation with applications to image and video completion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3732–3740.

[3] X. Deng, Y. Zhang, M. Xu, S. Gu, and Y. Duan, "Deep coupled feedback network for joint exposure fusion and image super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 3098–3112, 2021.

[4] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[5] Y.-F. Li, K. Shang, and Z.-H. Huang, "Low tucker rank tensor recovery via admm based on exact and inexact iteratively reweighted algorithms," *Journal of Computational and Applied Mathematics*, vol. 331, pp. 64–81, 2018.

[6] P. Tichavskỳ, A.-H. Phan, and A. Cichocki, "Numerical cp decomposition of some difficult tensors," *Journal of Computational and Applied Mathematics*, vol. 317, pp. 362–370, 2017.

[7] H. Zhang, X.-L. Zhao, T.-X. Jiang, M. K. Ng, and T.-Z. Huang, "Multiscale feature tensor train rank minimization for multidimensional image recovery," *IEEE Transactions on Cybernetics*, 2021.

[8] F. Sedighin, A. Cichocki, and A.-H. Phan, "Adaptive rank selection for tensor ring decomposition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 454–463, 2021.

[9] M. E. Kilmer and C. D. Martin, "Factorization strategies for third-order tensors," *Linear Algebra and its Applications*, vol. 435, no. 3, pp. 641–658, 2011.

[10] W. Hu, D. Tao, W. Zhang, Y. Xie, and Y. Yang, "The twist tensor nuclear norm for video completion," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 12, pp. 2961–2973, 2016.

[11] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis with a new tensor nuclear norm," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 925–938, 2019.

[12] C.-Y. Pang, R.-G. Zhou, B.-Q. Hu, W. Hu, and A. El-Rafei, "Signal and image compression using quantum discrete cosine transform," *Information Sciences*, vol. 473, pp. 121–141, 2019.

[13] X. Shen, J. Yang, C. Wei, *et al.*, "Dct-mask: Discrete cosine transform mask representation for instance segmentation," in *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, 2021, pp. 8720–8729.

[14] C. Lu, X. Peng, and Y. Wei, "Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5996–6004.

[15] Y. Zheng, T. Huang, X. Zhao, T. Jiang, T. Ji, and T. Ma, "Tensor n-tubal rank and its convex relaxation for low-rank tensor recovery," *Information Sciences*, vol. 532, pp. 170–189, 2020.

[16] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 148–172, 2013.

[17] C. Lu, X. Peng, and Y. Wei, "Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5996–6004.

[18] X. Zhang, F. Peng, and M. Long, "Robust coverless image steganography based on dct and lda topic classification," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3223–3238, 2018.

[19] H. Xu, J. Zheng, X. Yao, Y. Feng, and S. Chen, "Fast tensor nuclear norm for structured low-rank visual inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 538–552, 2022.

[20] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[21] Z. Zhang and S. Aeron, "Exact tensor completion using t-svd," *IEEE Transactions on Signal Processing*, vol. 65, no. 6, pp. 1511–1526, 2016.

[22] L. Yuan, C. Li, D. Mandic, J. Cao, and Q. Zhao, "Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9151–9158.

[23] Y. Zheng, T. Huang, X. Zhao, Q. Zhao, and T. Jiang, "Fully-connected tensor network decomposition and its application to higher-order tensor completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 11 071–11 078.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multi-Frequency Feature Enhancement for Multi-Granularity Visual Classification

Meijiang Fu[1], Yixiao Zheng[1], Dongliang Chang[2*], Wenpan Li[1], and Zhanyu Ma[1]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Tsinghua University, Beijing, China
[1]{fmj,zhengyixiao,liwenpan,mazhanyu}@bupt.edu.cn
[2]{changdongliang}@tsinghua.edu.cn

***Abstract*—Multi-granularity visual classification is a challenging task derived from traditional image recognition. Previous methods commonly use the features from the final convolutional layer to perform multi-granularity visual classification. However, the features required for different granularity label classification are not consistent. The finer the granularity of the label, the more detailed the features are needed. So, the key to multi-granularity visual classification is to extract effective features for different granularity levels. Generally, the high-frequency parts of natural images usually encode detailed information, while low-frequency parts often encode global structures. Therefore, mapping the output features of convolutional layers into high-frequency and low-frequency parts may enhance feature learning with multi-granularity. In this paper, we decompose the output features from convolutional layers into high-frequency and low-frequency counterparts, and use the combinations of the high-frequency and low-frequency features from various convolutional layers to learn different granularity labels. We evaluate the proposed method on the three public available datasets, including CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets. The experimental results demonstrate the effectiveness of the proposed method.**

I. Introduction

Fine-grained visual classification (FGVC) [1]–[10] is a research field in computer vision, with the goal of identifying sub-classes with high similarity under the same super-class. Many researchers [11]–[17] are currently striving to improve the performance of FGVC models, but there is little research on the definition of fine-grain itself. In fact, the definition of fine-grain is subjective, and people at different professional levels have different definitions of fine-grain. In order to meet people's demands for different granularities, some researchers [18]–[22] have constructed a hierarchical label classification system to provide multi-granularity labels from coarse to fine granularities.

By analyzing existing multi-granularity visual classification methods [18]–[22], a common and challenging point is how to extract effective features representing different granularities to learn multi-granularity labels. Chang et al. [18] extracted features for multi-granularity labels by decomposing the final convolutional layer's features into granularity-specific parts, and each part was used for one granularity label learning. Then, they proposed fine-grained features to participate in coarse-grained label prediction, which promoted the learning of coarse-grained labels. This method disentangles the decision spaces of different granularities and improves the performance of the model. Chen et al. [19] extracted features for multi-granularity labels by constructing a feature pyramid structure using the output features at different convolutional layers. Then, they proposed upsampling deep features and combining them with shallow features to make decisions. This method uses the relationships between labels at different levels to learn features with different granularities, and has achieved better results. Meanwhile, the high-frequency parts of natural images usually encode detailed information, while low-frequency parts often encode global structures [23]–[25], as shown in Fig. 1. Therefore, mapping the output features of convolutional layers into high-frequency and low-frequency parts may enhance feature learning with multi-granularity. In this paper, we use octave convolution [26] to decompose the output features of convolutional layers into high-frequency and low-frequency features. Then, we combine cross-layer frequency features to extract effective features representing different granularities.

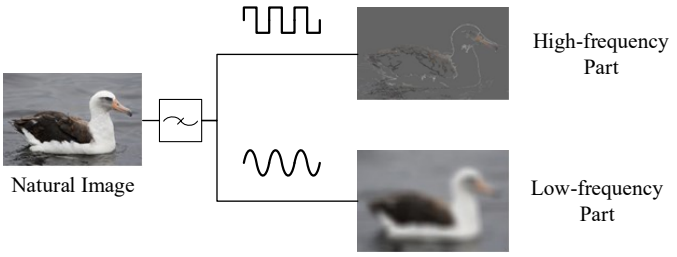


Fig. 1: Natural image information can be decomposed into high-frequency and low-frequency information.

As shown in Fig. 2, we propose a multi-frequency feature enhancement model for multi-granularity visual recognition. Specifically, we take the three-granularity labels as an example. Firstly, in order to disentangle the learning of detailed information and global structural information, we decompose the convolutional layers' features into high-frequency and low-frequency features through octave convolution [26]. Secondly, by using the relationships between labels at different levels to learn features with different granularities, we decompose the final convolutional layers' features into three feature parts, and combine the low-frequency features of shallower convolutional layers with these feature parts. Lastly, we set a gradient controller, to ensure that the gradient flow of each granularity-

*indicates the corresponding author.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

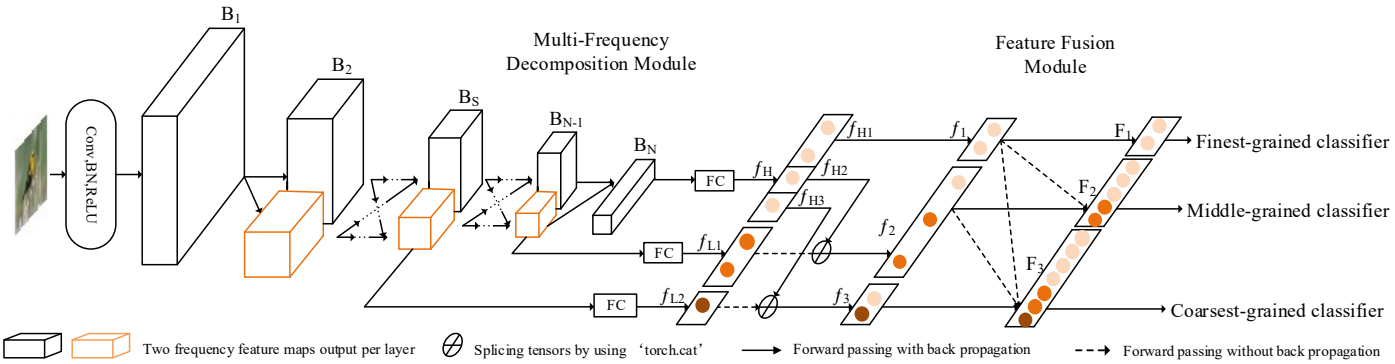


Fig. 2: Schematic diagram of our multi-granularity visual classification model, with the multi-frequency decomposition module on the left and the feature fusion module module on the right.

specific classifier don't affect others during the reverse propagation phase. Following Chang et al. [18], we allow fine-grained features to participate in coarse-grained label learning. Through the above method, we have achieved multi-frequency feature enhancement for multi-granularity visual classification. Our contributions are as follows:

1)We decompose the features into high-frequency and low-frequency parts, to enhance the learning of detailed features and global structural features respectively.

2)We propose a new method based on combining the high-frequency and low-frequency features of different convolutional layers, to achieve multi-frequency feature enhancement for multi-granularity visual recognition.

3)We conduct extensive experiments on three publicly available datasets: CUB-200-2011 [27], Stanford Cars [28], and FGVC-Aircraft [29]. The experimental results demonstrate the effectiveness of the proposed method.

II. RELATED WORK

A. Fine-Grained Visual Classification

Fine-grained visual classification (FGVC) [1]–[10] is a highly challenging research field in computer vision, with the goal of identifying sub-classes with high similarity under the same super-class. In contrast to traditional image recognition, the difficulty of FGVC is that there are significant intra-class differences and subtle inter-class differences in images. Currently, researchers mostly use deep learning methods to solve FGVC task, and have achieved good results. Specifically, these methods can be roughly divided into three categories. The first category is based on localization-recognition methods [4]–[7], which locate distinctive feature regions and then learn fine-grained features in these regions. In the early days, researchers [4], [5] used strong supervision to locate distinctive feature regions, such as adding additional manual annotation information. In recent years, researchers [6], [7] have gradually proposed using weak supervision methods to locate distinctive feature regions, such as introducing attention mechanisms. The second category is based on network ensemble methods [8], [9], which divide fine-grained datasets into several similar subsets for classification, or directly use multiple neural networks to improve the performance of fine-grained classification. The third category is based on high-order feature encoding methods [2], [10], which improve the expression ability of CNN features through high-order synthesis. In this paper, we study a different task based on FGVC, which generates images into multi-granularity label outputs.

B. Multi-Granularity Visual Classification

Multi-granularity visual classification (MGVC) task aims to provide a sequence of labels from coarse to fine for single label images. In contrast to FGVC, MGVC not only focuses on improving model recognition performance, but also meets the demands of people with different professional levels for different granularities. Now, many researchers [18]–[20], [30], [31] have used different methods to solve MGVC tasks. For example, Chang et al. [18] decomposed the features of the final convolutional layer into several specific-granularity parts, and proposed fine-grained features to participate in coarse-grained label prediction. Chen et al. [19] constructed a feature pyramid structure using the output features of different convolutional layers, and then proposed upsampling deep features and combining them with shallow features to make decisions. Wang et al. [20] proposed a unified probability framework based on deep learning for label hierarchy transformation, to solve hierarchical classification problems by using the correlation between various categories at different levels. The above models all extracted effective features of different granularities through different methods. In this paper, we decompose the convolutional layer into multi-frequency features, and combine the frequency features of different layers to extract more effective features of different granularities.

III. METHODOLOGY

In this section, we will introduce our model for multi-granularity visual classification. As shown in Fig. 2. We divide our model into two parts, namely the multi-frequency decomposition module and the feature fusion module respectively. We process the input image with octave convolution [26] to obtain multi-frequency features. Then, we combine the high-frequency and low-frequency features of different convolutional layers to enhance the learning of multi-granularity labels.

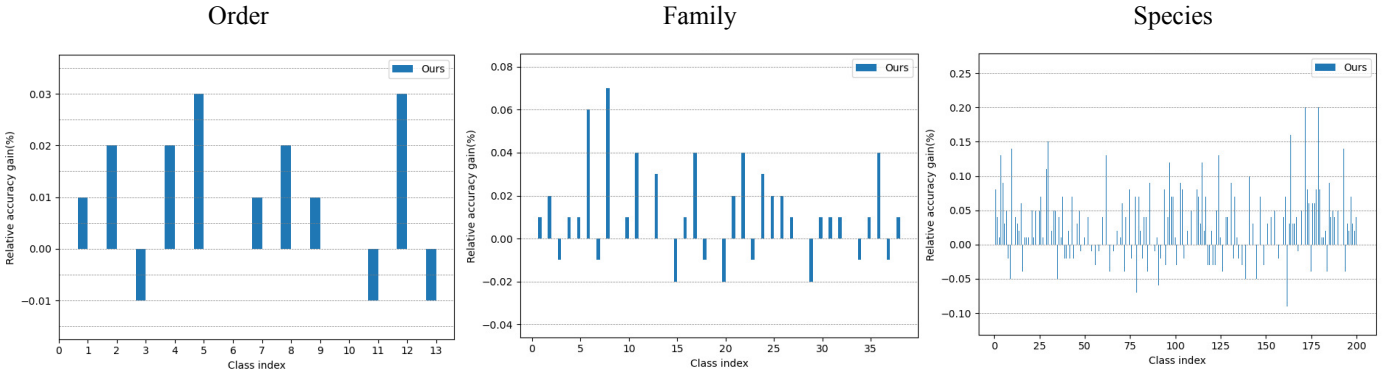


Fig. 3: Relative accuracy gain of Oct-FGoN-based, Ours compared with Oct-FGoN for each class recognition on the CUB-200-2011 dataset. X-axis represents the class index. Y-axis represents the relative accuracy gain for each class.

A. Multi-Frequency Decomposition Module

The goal of this module is using multi-frequency features to disentangle the learning of detailed information and global structural information. We decompose convolutional layers' output features into high-frequency and low-frequency features through octave convolution [26]. We replace all traditional convolutions of backbone network with octave convolutions and form our multi-frequency decomposition module.

In multi-granularity visual classification task, an image x has a multi-granularity label $\{y_1,y_2,...,y_K\}$ (K represents the number of granularities). Using ResNet50 as an example, we denote the output features of each block to $\{B_1,B_2,...,B_S,...,B_N\}$, where N is the total number of convolutional blocks. In our method, we decompose each block's output features $B_S(C_S \times W_S \times H_S)$ into high-frequency $B_{SH}(C_S/2 \times W_S \times H_S)$ and low-frequency $B_{SL}(C_S/2 \times W_S/2 \times H_S/2)$ through octave convolution [26]. We take the final convolutional layer's features and other blocks' low-frequency features as the multi-frequency features $\{B_N, B_{(N-1)L}, B_{(N-2)L}, ..., B_{(N-K+1)L}\}$. Then, we pass the multi-frequency features through the fully connected layer FC to obtain the input $\{f_H, f_{L1}, f_{L2}, ..., f_{L(K-1)}\}$ of the feature fusion module.

B. Feature Fusion Module

In this module, we use two feature fusion operations. We have the features $\{f_H, f_{L1}, f_{L2}, ..., f_{L(K-1)}\}$ output from the multi-frequency decomposition module. For the convenience of explanation, we use the three-granularity labels as an example in Fig. 2. Firstly, following Chang et al. [18], we decompose the f_H into three feature parts $\{f_{H1}, f_{H2}, f_{H3}\}$. To be more specific, we divide half of f_H into f_{H1}, and divide the other half equally into f_{H2} and f_{H3}. Because fusing cross-layer features can enhance the learning of multi-granularity labels [19], so we fuse f_{H2} with f_{L1} to denote as f_2, and fuse f_{H3} with f_{L2} to denote as f_3. We transfer the first feature regions f_{H1} to f_1, thus obtain new three feature parts $\{f_1, f_2, f_3\}$. The above feature fusion operations are set with gradient controller functions δ to stop the propagation of gradient flow, and $\oplus$ is denoted concatenation operation. The specific methods are as follows:

$$f_1 = f_{H1},$$
$$f_2 = f_{H2} \oplus \delta(f_{L1}), f_3 = f_{H3} \oplus \delta(f_{L2}). \quad (1)$$

Secondly, we have mentioned above that fine-grained features can enhance the prediction of coarse-granularity classifier. So, we fuse the fine-grain feature part into the coarse-grained feature part. To be more specific, we fuse f_1 and f_2 into f_3 to obtain F_3 that works for prediction of coarse-grained label, and fuse f_1 into f_2 to obtain F_2 that works for prediction of middle-grained label. We transfer f_1 to F_1 that works for prediction of fine-grained label. We still set gradient controller functions δ for these fusions, and $\oplus$ is denoted concatenation operation. The specific methods are as follows:

$$F_1 = f_1, F_2 = f_2 \oplus \delta(f_1),$$
$$F_3 = f_3 \oplus \delta(f_2) \oplus \delta(f_1). \quad (2)$$

Through the two feature fusion operations mentioned above, we have achieved multi-frequency features enhancement for multi-granularity visual classification task. Our optimisation objective is three independent cross-entropy loss $\sum_{k=1}^{3} L_{CE}(\hat{y_k}, y_k)$, where $\hat{y_k}$ is the label predicted by the model, and y_k is the real label. Similarly, for K granularity labels, our optimisation objective is K independent cross-entropy loss $\sum_{k=1}^{K} L_{CE}(\hat{y_k}, y_k)$.

IV. EXPERIMENTAL SETTINGS

A. Datasets and Evaluation Criterion

We validate our method on the three public available datasets. Chang et al. [18] have constructed a multi-granularity label structure for each dataset. Details are as follows. **CUB-200-2011** [27] is a bird dataset consisting of 11778 bird images from 200 categories, officially divided into 5994 training images and 5794 test images. This dataset is divided into 13 orders, 38 families, and 200 species. **FGVC Aircraft** [29] is an aircraft dataset consisting of 10000 aircraft photos from 100 categories, officially divided into 6667 training images and 3333 test images. This dataset is divided into 30 manufacturers, 70 families, and 100 models. **Stanford Cars** [28] is a vehicle dataset consisting of 16185 vehicle images from

TABLE I: Accuracy Comparisons of Different Multi-granularity Visual Classification Methods
(The highest accuracy of each granularity is marked in red, the second accuracy is marked in blue)

Method	CUB-200-2011				FGVC-Aircraft				Stanford Cars		
	order_acc	family_acc	species_acc	avg_acc	maker_acc	family_acc	model_acc	avg_acc	maker_acc	model_acc	avg_acc
Vanilla_single	95.38 ± 0.47	87.70 ± 0.79	74.24 ± 0.86	85.77	90.82 ± 1.02	88.73 ± 1.17	86.26 ± 1.37	88.60	95.30 ± 0.11	88.66 ± 0.45	91.98
Vanilla_multi	95.13 ± 0.53	89.70 ± 0.13	78.31 ± 0.35	87.71	90.69 ± 0.48	89.23 ± 0.53	88.10 ± 0.10	98.34	95.24 ± 0.20	89.14 ± 0.16	92.19
FGoN [18]	96.37 ± 0.16	90.39 ± 0.15	77.95 ± 0.04	88.24	93.04 ± 0.25	90.73 ± 1.19	88.35 ± 0.18	90.71	95.58 ± 0.06	89.66 ± 0.16	92.62
Oct_FGoN	97.12 ± 0.10	92.21 ± 0.16	80.41 ± 0.13	89.91	94.41 ± 0.18	92.11 ± 0.12	88.66 ± 0.14	91.73	95.81 ± 0.12	90.07 ± 0.14	92.94
CGVC [19]	97.51 ± 0.12	92.22 ± 0.10	80.54 ± 0.13	90.09	95.40 ± 0.07	92.72 ± 0.10	89.00 ± 0.15	92.37	96.43 ± 0.11	90.70 ± 0.13	93.57
Oct_CGVC	97.21 ± 0.14	91.92 ± 0.13	80.51 ± 0.11	89.88	94.72 ± 0.13	92.51 ± 0.11	88.79 ± 0.16	92.01	96.13 ± 0.17	90.16 ± 0.18	93.15
Ours	97.62 ± 0.15	92.65 ± 0.12	81.61 ± 0.16	90.63	95.45 ± 0.08	92.80 ± 0.13	89.06 ± 0.17	92.44	96.45 ± 0.15	90.74 ± 0.12	93.60

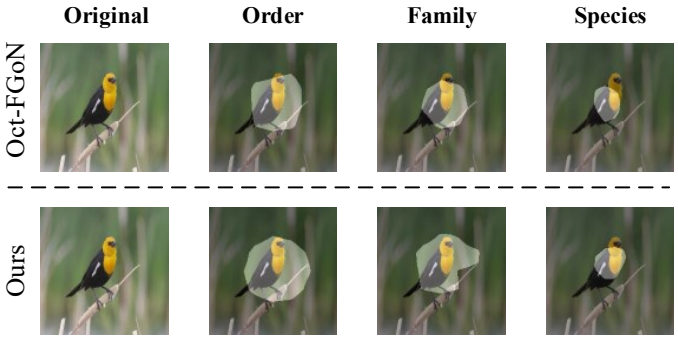


Fig. 4: Visualizations of Oct-FGoN and Ours at Order, Family, and Species granularities on CUB-200-2011 dataset.

196 categories, officially divided into 8144 training images and 8041 test images. This dataset is divided into 9 car types and 196 models.

We use the same evaluation metrics as [18]. We evaluate the network performance through accuracy, which means percentage of images whose labels correctly classified. Then calculate the average accuracy of multi-granularity labels, denoted by "avg_acc". Each experiment is run three times, and the accuracy results are shown with the mean and standard deviation.

B. Implementation Details

We use Oct-ResNet50 (replacing the traditional convolution of ResNet50 with octave convolution) [26], [32], pre-trained on ImageNet as the backbone network. On the CUB-200-2011 and FGVC-Aircraft datasets with labels of three granularities, we use the final features of residual block4 in Oct-ResNet50, and the low-frequency features of residual block3 and block2 as the multi-frequency features, denoted as f_H, f_{L1}, and f_{L2} respectively. On the Stanford Cars dataset with labels of two granularities, we use the final features of residual block4 in Oct-ResNet50, and the low-frequency features of residual block3 as the multi-frequency features, denoted as f_H and f_{L1} respectively.

For a fair comparison with other methods, we scale the input size of all images to 224×224, and use common data augmentation training methods, including horizontal flipping, random cropping, and color dithering. We set the initial learning rate as 0.001 for the pre-trained convolutional layers, and 0.1 for the newly added fully connected layers in the classifiers. We decay the learning rate with cosine annealing, and train the networks for 100 epochs per experiment. We use Momentum Optimizer with momentum value as 0.9, and weight decay value as 5×10^4.

C. Baselines

To verify the effectiveness of our method, we have conducted extensive comparative experiments with existing models: (i) **Vanilla_single**: the method uses only one network backbone (Resnet50), followed by multiple classification headers. (ii) **Vanilla_multi**: it has multiple independent backbone networks (Resnet50), each of which handles one granularity. (iii) **FGoN** [18]: it decomposes the final features into several parts, and uses fine-grained features to promote the learning of coarse-grained labels. (iv) **Oct-FGoN** [18]: replacing the convolutions in the FGoN with octave convolutions. (v) **CGVC** [19]: it constructs the features of different blocks into a feature pyramid structure, and then deep block's features participate in shallow block's feature learning. (vi) **Oct-CGVC** [19]: replacing the convolutions in the CGVC with the octave convolutions. (vii) **Ours**: decomposing the convolutional layers' features into high-frequency and low-frequency features, combining the high-frequency and low-frequency features of different convolutional layers to learn labels with different granularities.

D. Results and Analysis

Comparing with existing models, whether in coarse-grained labels or fine-grained labels, our method has achieved satisfactory results. The experimental results are shown in Table 1. From the table, it can be observed that:

Comparing to Vanilla_single and Vanilla_multi, FGoN [18] obtains better results on each granularity label. The difference from Vanilla_single is that FGoN [18] disentangles decision spaces of different granularities. And the difference from Vanilla_multi is that FGoN [18] uses fine-grained features to participate in the learning of multi-granularity labels. So, we have obtained two inspirations: (i) disentangling decision spaces of different granularities, and (ii) using fine-grained features to participate in coarse-grained label learning. They are beneficial for promoting the recognition of multi-granularity labels.

Compared to FGoN [18] and Oct-FGoN, our model has exceeded the accuracy of Oct-FGoN and FGoN [18] at each granularity. So, it is effective to combine the different convolutional layers' low-frequency features with the final convolutional layer' features. Because the relationships of the different convolutional layers' features is similar to the relationship of

the different granularity's labels. So, utilizing the relationships between different granularity labels to combine cross-layer features can enhance the learning of multi-granularity labels.

Compared to CGVC [19] and Oct-CGVC, our model has a better performance. The difference with CGVC [19] and Oct-CGVC is that we only take the low-frequency parts of the convolutional layers, and we combine these cross-layer frequency features after passing through the fully connected layer. The experimental results prove that our method is effective, because our method reduces the impact of inconsistent feature scales between different convolutional layers.

In order to more directly demonstrate the performance of our method. In Fig. 3, we compare the classification results of each class at three granularities in the CUB-200-2011 dataset with Oct-FGoN [18].

E. Visualization

In Fig. 4, we use Grad-CAM [33] to visually compare Oct-FGoN [18] with our method. Experiments show that our method has a larger receptive field in coarse-grained areas and a more accurate receptive field in fine-grained areas, which means that our method achieves disentangling the learning of detailed information and global structural information.

V. CONCLUSION

We propose a multi-frequency feature enhancement for multi-granularity visual recognition. We realize decomposing the features of convolutional layers into multi-frequency features can disentangle the learning of detailed information and global structural information. Meanwhile, the relationship of the different convolutional layers' features is similar to the relationship of the different granularity's labels, so we combine the final convolutional layer's features with low-frequency features of different convolutional layers, to obtain effective features representing different granularities. Compared with existing methods, we achieve satisfactory performance on CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets.

ACKNOWLEDGMENT

This work was supported in part by Beijing Natural Science Foundation Project No. Z200002, in part by National Natural Science Foundation of China (NSFC) No. U19B2036, 62225601 and in part by Youth Innovative Research Team of BUPT No. 2023QNTD02.

REFERENCES

[1] T. Berg and P. N. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[2] S. Cai, W. Zuo, and L. Zhang, "Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] H. Zheng, J. Fu, M. Tao, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognition*, 2018.

[5] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked cnn for fine-grained visual categorization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] C.-L. Zhang, Y.-H. Cao, and J. Wu, "Rethinking the route towards weakly supervised object localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Z. Ge, C. McCool, C. Sanderson, and P. Corke, "Subset feature learning for fine-grained category classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] Z. Ge, A. Bewley, C. McCool, P. Corke, B. Upcroft, and C. Sanderson, "Fine-grained classification via mixture of deep convolutional neural networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[10] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," *arXiv preprint arXiv:1406.2952*, 2014.

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[13] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *in IEEE Transactions on Image Processing*, 2017.

[14] T. Chen, W. Wu, Y. Gao, D. Le, and L. Liang, "Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding," *in ACM International Conference on Multimedia*, 2018.

[15] D. Chang, Y. Ding, J. Xie, *et al.*, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *in IEEE Transactions on Image Processing*, 2020.

[16] D. Chang, K. Pang, R. Du, *et al.*, "Making a bird ai expert work for you and me," *in IEEE Transactions*

on Pattern Analysis and Machine Intelligence (TPAMI), 2023.

[17] D. Chang, Y. Tong, R. Du, T. Hospedales, Y.-Z. Song, and Z. Ma, "An erudite fine-grained visual classification model," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your" flamingo" is my" bird": Fine-grained, or not," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] J. Chen, D. Chang, J. Xie, R. Du, and Z. Ma, "Cross-layer feature based multi-granularity visual classification," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2022.

[20] J. Chen and Y. Qian, "Hierarchical multilabel ship classification in remote sensing images using label relation graphs," *in IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[21] J. Chen, P. Wang, J. Liu, and Y. Qian, "Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] R. Du, W. Yu, H. Wang, *et al.*, "Multi-view active fine-grained recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[23] F. W. Campbell and J. G. Robson, "Application of fourier analysis to the visibility of gratings," *The Journal of physiology*, 1968.

[24] R. L. Devalois and K. K. Devalois, "Spatial vision (oxford psychology)," *Politics*, 2003.

[25] D. G. Low, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[26] Y. Chen, H. Fan, B. Xu, *et al.*, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] K. Duan, D. Parikh, D. Crandall, and K. Grauman, "Discovering localized attributes for fine-grained recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[28] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[29] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *HAL - INRIA*, 2013.

[30] X. Shu, L. Zhang, Z. Wang, L. Wang, and Z. Yi, "Fine-grained recognition: Multi-granularity labels and category similarity matrix," *Knowledge-Based Systems*, 2023.

[31] A. Garg, S. Bagga, Y. Singh, and S. Anand, "Hiermatch: Leveraging label hierarchies for improving semi-supervised learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Improving Aspect Sentiment Classification via Retrieving from Training Data

Tongtao Ling[1,2,†], Lei Chen[1], Chen Liao[3], Shilei Huang[3], Zhipeng Yu[2] and Yi Liu[2,*]
[1] Guangdong University of Technology, Guangzhou, China
[2] IMSL Shenzhen Key Lab, PKU-HKUST Shenzhen Hong Kong Institute, Shenzhen, China
[3] Shenzhen Raisound Technologies, Co., Ltd, Shenzhen, China
E-mail: tongtao.ling@imsl.org.cn, chenlei3@gdut.edu.cn, {chen.liao, shilei.huang}@raisound.com, {zhipeng.yu, yi.liu}@imsl.org.cn

***Abstract*—Aspect sentiment classification (ASC) is an essential subtask of aspect-based sentiment analysis. Recently, pre-trained language models (PLMs) have gradually become the mainstream building block for the ASC task and retrieval-based methods are shown to be effective in helping PLMs better understand various downstream tasks. However, retrieval-based methods need to introduce external knowledge, and building an index of large-scale corpus leads to huge cost. To this end, we propose an effective method that utilizes training data to construct a retrieval corpus and retrieve instances most similar to current input to enhance semantic representations. In the proposed method, three kinds of retrieval metrics are applied, allowing us to quickly adapt to ASC tasks. Experimental results show that this simple retrieval augmented method can achieve significantly better performance on three benchmark ASC datasets (Twitter, Laptop, and Restaurant). Our code is available at https://github.com/rickltt/re-bert.**

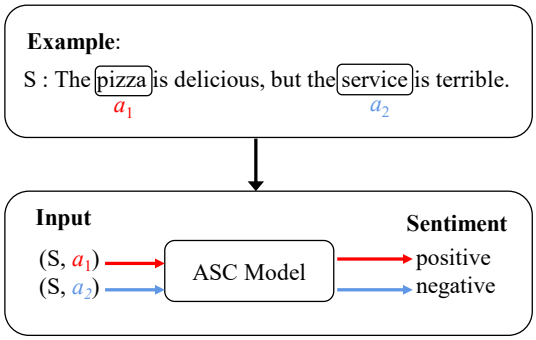


Fig. 1. An example of aspect sentiment classification process.

I. Introduction

Sentiment analysis (SA) is an important task in natural language processing (NLP) that aims to classify the sentiment polarity of a given sentence [1]. Traditional sentiment analysis studies have focused on sentence-level or document-level predictions, yet there may be multiple aspects in a given text. For example, As shown in Fig. 1, a plain sentence contains different aspects of "pizza" and "service". The polarity in "pizza" is positive, and the polarity regarding "service" is negative. In this context, the need to identify more fine-grained aspect-level perspectives and emotions, also known as aspect-based sentiment analysis (ABSA) [2], has received increasing attention over the past decade.

Aspect sentiment classification (ASC), also called aspect-level sentiment classification, is a subtask of ABSA, which aims to classify the sentiment polarity (positive, negative, or neutral) of aspects extracted from the original sentences. Fig. 1 illustrates the flow of ASC tasks, there are two aspect terms a_1 and a_2 extracted from sentence s and formed into two pairs (s,a_1) and (s,a_2), subsequently inputted into the ASC model to obtain the sentiment polarity of aspects.

The earlier work on ASC relied heavily on feature engineering [3] and subsequently ASC based on deep learning has attracted a lot of interest. After that, a variety of neural network-based models have been proposed including the CNN-based network [4, 5], memory network [6, 7], and gated network [8], and significant performance improvements have been made. In recent years, fine-tuning pre-trained language models (PLMs) such as BERT [9] and RoBERTa [10] have been reported with high performance in ASC tasks [11]. PLMs can learn generic text representations from large amounts of unlabeled corpus to enhance downstream tasks, which attracts considerable attention to improve the semantic representations of sentences for the ASC task. In natural language processing, retrieval-based methods usually work by selecting extra information related to input and concatenating it with retrieval results as a new input [12]. This way can remarkably improve the performance of models for various downstream tasks. As a result, a series of retrieval-based approaches have been applied with great success to many NLP tasks such as language modeling [13, 14], open-domain question answering [15] and machine translation [16]. However, these methods rely on a large-scale corpus and building an index requires a huge computational burden. From the perspective of the supervised learning paradigm, the most similar distribution to the test data is the training data. Moreover, the size of training data is usually much smaller than the external large-scale corpus. To this end, we present a **RE**trieval-based **BERT** (**RE-BERT**) model, which aims to utilize training data as the retrieval corpus to enhance the semantic representations for the ASC task. Specifically, as shown in Fig. 2, we first build an index

† Work was done when Tongtao Ling interned at IMSL.
* Corresponding author.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of training data as a retrieval corpus. Then, for each input instance, we retrieve the top k most similar instances to the current input. Finally, we concatenate retrieval instances with the original input as a new input and feed it into the ASC model to predict the polarity of the aspect.

In our experiments, we use three different retrieval metrics to measure the similarity between instances and evaluate our approach on three ASC benchmark datasets. Compared to various state-of-the-art baselines, we find that: (1) after integrating our method, pre-trained models (i.e., BERT) for the ASC task can achieve better performance than other models that use different network architectures; (2) for different retrieval metrics, dense-vector retrieval (dense vectors encoded using pre-trained language models) outperforms spare-vector retrieval (i.e., TFIDF and BM25); (3) the results on ablation study shows that BERT-base with retrieval method can achieve similar performance with Bert-large without retrieval method, the parameter of BERT-large is nearly three times than BERT-base (340M vs. 110M). This phenomenon illustrates that even though models have hundreds of millions of parameters, a model cannot memorize all the patterns in the training data. Therefore, retrieving the relevant training data to enhance the semantic information is an effective method to enhance the inference of the model. Besides, we conduct extensive ablation studies on three widely-used datasets and perform insight analysis. The results demonstrate the effectiveness and interpretability of our proposed model.

The remainder of this paper is organized as follows. Related work is described in Section II. Our proposed method is described in Section III, including the pipeline of our model and the kinds of retrieval metrics. In Section IV, the experimental results indicate that the competitiveness with other state-of-the-art ASC models. In section V, further analysis of the ablation experiments is performed. Finally, Section VI concludes this paper.

II. Related Work

A. Aspect-based Sentiment Analysis

Aspect-based sentiment analysis aims to identify the polarity of a specific aspect term in the given sentence. Early ASC models focused on manually designed features such as dependency parse [17] and term frequency [3]. With the rapid development of deep learning, most research studies make use of neural networks to learn the word's semantic representation [5, 7, 18, 19]. Among them, Ma et al. [19] proposed a model that learns interactively representations between aspect and context by using two LSTM layers and attention mechanisms. Chen et al. [7] used bidirectional long short term memory (BiLSTM) and gated recurrent unit (GRU) network to combine multiple attention outputs. Wang et al. [18] appends the target embedding with each word embedding and gets the final representation by attention mechanisms that combine with LSTM. Li et al. [5] introduced a target-specific network to make use of the aspect information.

More recently, pre-trained language models have been achieved remarkably successful in the aspect-based sentiment analysis task. Song et al. [11] proposed to use manual auxiliary sentences to convert ABSA into a sentence-pair classification task. Xu et al. [20] proposed the Review Reading Comprehension (RRC) task and utilized the post-training method for RRC and ABSA, which can let the model acquire domain knowledge and task-aware knowledge. To address the problem of implicit sentiment in ASC (sentiments with no explicit opinion word), Li et al. [21] proposed a pre-training method for the ASC model. Xiao et al. [22] utilized inter-layer's features and position information of BERT to enhance graph convolutional network (GCN) encoding for ASC. Tian et al. [23] proposed Type-aware GCN that uses multiple layers of GCN information instead of just the last layer. Qin et al. [24] proposed a two-stage method to complete the ASC task: the first stage adds a decoder to recover the original sentence and uses a discriminator to weigh the recovered result, and the second stage removes the part and only classifies sentiment. Zhang et al. [25] design a lightweight dynamic reweighting Adapter to learn dynamic aspect-oriented semantics for aspect-based sentiment analysis. Dependency trees are widely used in GCN, but Chen et al. [26] thinks that dependency trees are not optimized and proposes an aspect-specific and language-agnostic discrete latent opinion tree model as an alternative structure for ABSA.

B. Retrieval-based Methods

Retrieval-based methods fuse recent deep learning technology and traditional retrieval technology, which has achieved remarkably successful performance in many natural language processing tasks and attracted more attention from NLP researchers [27, 28]. The parameters of pre-trained language models are very huge, but even GPT-3 [29] is unable to remember everything and it is also important to collect retrieval information from external knowledge to enhance the performance of various NLP tasks. Specifically, template retrieval is an effective intermediate step for the dialogue response generation task. On the task of machine translation, retrieved sentences are used to improve statistical machine translation and neural machine translation models. And other applications of text generation such as code generation [30], abstractive summarization [31] and knowledge-intensive generation [14].

The most important key of the retrieval-based method is the retrieval metric. There are two kinds of retrieval metrics: sparse representation and dense representation. For sparse representation, the most commonly used are bag-of-words and rule-based keyword matching (i.e., TFIDF and BM25 [32]) method. These methods have been widely applied to a general large-scale search, such as open domain question answering and machine translation. For dense representation, it requires an encoder (i.e., BERT) to transform the context into low-dimensional dense vectors and compute the similarity between two vectors. Dense representation can outperform sparse representation on different tasks, such as knowledge-grounded generation, open domain question answering and machine translation. The drawback of dense representation is the cost of time and it can not process long sequences, usually less than 512 tokens.

Considering the consequence of different retrieval metrics for retrieval-based methods, in this paper, we use TFIDF, BM25 and Dense to score the similarity and retrieve from the training data.

III. Methodology

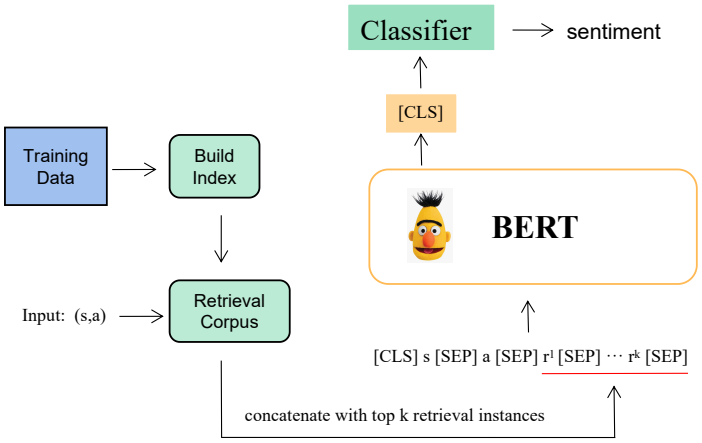


Fig. 2. The pipeline of aspect sentiment classification with retrieval from training data.

In this section, we first introduce the description of aspect sentiment classification, then we elaborate our proposed **RE-BERT** and the detail of retrieval metrics.

A. ASC Task

Given a natural language sentence s containing a set of aspect terms: $\{a_1, ..., a_n\}$, which can be wrapped into several pairs (s,a_1),...,(s,a_n) as inputs. The aspect term could be a single word or a long phrase and the goal of ASC task is to find the polarity $y \in Y$ of each aspect term, $Y = \{$ *positive*, *negative*, *neutral*$\}$. As a result, ASC is typically cast as a sentence pair classification task.

As shown in Figure 2, the ASC model mainly consists of an encoder and a classifier. We utilize pre-trained language model BERT as the base encoder to generate word vectors of sequence. For an input (s,a), we insert special tokens to construct a new input:

$$x = [CLS]\ s\ [SEP]\ a\ [SEP] \tag{1}$$

[CLS] and [SEP] are special tokens in BERT. [CLS] is a symbol in front of each sentence and [SEP] used to distinguish different sentences.

B. RE-BERT

In this paper, we aim to improve the performance of the ASC task by using retrieval results as input to generate better results. In both training and testing, for each input-label (s,a), we retrieve top k sentences $\{r^1, \cdots, r^k\}$ from training data most similar to the current input. Then, we concatenate them with the input sentence:

$$x_{concat} = [CLS]\ s\ [SEP]\ a\ [SEP]\ r^1\ [SEP] \cdots r^k\ [SEP] \tag{2}$$

After that, we input x_{concat} to the encoder BERT and obtain the contextual representation:

$$h = BERT(x_{concat}) \tag{3}$$

We leverage the representations of [CLS] $h_{[CLS]}$ as the input of the classifier, because $h_{[CLS]}$ is the aspect-aware representation of whole concatenated sentences. The classifier is a linear layer. To conduct aspect sentiment classification, a weight matrix $W \in \mathbb{R}^{D\times|Y|}$ and a bias term $b \in \mathbb{R}^{|Y|}$ are introduced to classify each aspect into a corresponding polarity, where D is the hidden size of BERT output and $|Y|$ is the number of polarities. The classification logit is defined as:

$$p = W \cdot h_{[CLS]} + b \tag{4}$$

The final predicted polarity of aspect:

$$\hat{y} = softmax(p) \tag{5}$$

The optimization loss function is a cross entropy loss between classification logits and true polarity, which is defined as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j \log(\hat{y}_i^j) + \lambda \sum_{\theta\in\Theta} \theta^2 \tag{6}$$

where y_i^j is the ground truth of polarity and $\hat{y}_i^j$ is the predicted polarity. C is the number of labels and N is the number of training instances. Θ is the parameter set corresponding to all the trainable parameters and λ is coefficient.

C. Retrieval Metrics

Given an input (s,a) and a retrieval corpus C, we aim to retrieve a set of relevant instances $z = \{(s^r, a^r)\}$ from C. How to measure the score between (s,a) and (s^r,a^r) is a considerable problem in retrieval-based methods. The widely used retrieval measurements such as BM25 and TF-IDF [32] could match keywords efficiently with an inverted index. However, these methods generate sparse-vector that is incapable to capture deep-level semantics. To alleviate the above problem, some recent works attempt to retrieve similar instances by using PLMs to generate dense-vector.

a) BM25: BM25 algorithm, usually used for search relevant score. The general formula for the BM25 algorithm is as follows:

$$Score_{bm25}(Q,d) = \sum_{i}^{n} W_i \cdot R(q_i,d) \tag{7}$$

where query $Q = \{q_1, \cdots, q_n\}$ is a tokenized sequence and we need to compute the correlation with a document $d \in D$, and D contains many documents. The correlation is calculated as:

$$R(q_i,d) = \frac{f_i \cdot (k_1+1)}{f_i + K} \tag{8}$$

$$K = k_1 \cdot (1 - b + b \cdot \frac{dl}{avgdl}) \tag{9}$$

where k_1 and b are regulating factors that usually could be set to 1.5 and 0.75 respectively. dl is the length of D and $avgdl$ is

the average length of D. fi is the frequency of occurrence of qi in Q. In addition to correlation, the other key factor is the weight W_i of q_i that is calculated to obtain final score. There are various ways to determine the correlated weight of q_i to d, the most common being IDF (Inverse Document Frequency). IDF is defined as:

$$W_i = IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{10}$$

b) TF-IDF: TF-IDF (Term Frequency Inverse Document Frequency) is a weighting technique commonly used in information processing and data mining. The technique uses a statistical method to calculate the importance of a word in the overall corpus based on the number of occurrences in the text and the frequency of documents in the overall corpus. It has the advantage of filtering out common but irrelevant words while retaining the important words that affect the whole text. The correlation is calculated as:

$$Score_{tfidf}(Q, d) = \sum_{i}^{n} f_i \cdot IDF(qi) \tag{11}$$

c) Dense: How to encode text to low-dimension dense vectors is an issue that has been studied in the past decades [33]. PLMs can learn deep-level semantic representations under large amount of unlabeled corpus and encode text into low-dimensional vectors. To make use of PLMs, we try to use BERT to encode document d and query Q. Then, we compute relevant scores via inner products between d and Q. The score is calculated as:

$$Score_{dense}(Q, d) = BERT(Q) \cdot BERT(d) \tag{12}$$

IV. Experiment

In this section, we present our experimental comparison and the corresponding analysis of three widely-used ASC datasets.

A. Datasets

We conduct experiments on three ASC benchmark datasets: Restaurant, Laptop, and Twitter. Restaurant reviews and Laptop reviews are from SemEval-2014 Task 4 [34], and Twitter reviews are from the social platform Twitter [35]. These datasets are labeled with three sentiment polarities: positive, neutral, and negative. We use *Accuracy* and *Macro-F1* as metrics for all experiments. Table I shows the number of training and testing instances in each category.

TABLE I
Statistics of the benchmark datasets.

Dataset	Train			Test		
	Positive	Neutral	Negative	Positive	Neutral	Negative
Twitter	1561	3127	1560	173	346	173
Restaurant	2164	637	807	728	196	196
Laptop	994	464	870	341	169	128

B. Baseline

To investigate the effectiveness of the proposed method, we compare it with a range of baselines that can be categorized into three classes: non-BERT-based models, BERT-based models, and our retrieval-based models. For non-BERT-based models, we compare with (1) **ATAE-LSTM** [18], using attention-based long short-term memory network for aspect-level sentiment classification; (2) **AOA** [36], introduced an attention-over-attention neural network for aspect-level sentiment classification; (3) **MemNet** [6] considers contexts as external memories and benefits from a multi-hop architecture; (4) **ASGCN** [37], building a Graph Convolutional Network (GCN) over the dependency tree of a sentence to exploit syntactical information and word dependencies; (5) **RAM** [7] adopts multiple-attention mechanism to capture sentiment features separated by a long distance. For BERT-based models, we compare with (6) **BERT** [9], fine-tuning for text classification; (7) **BERT-PT** [20], post-training for BERT's weight using a joint post-training algorithm; (8) **BERT-SPC** [11] treats ASC task as a sentence pair classification by using BERT; (9) **AEN-BERT** [11] eschew recurrence and employs attention-based encoders for the modeling between context and target; (10) **BERT-ADA** is proposed by [38], fine-tuning BERT on the cross-domain corpus.

C. Implementation Details

The proposed model is implemented based on the Transformers toolkit[1] and PyTorch Framework. We employ BERT-BASE-UNCASED[2] as the base encoder. The maximum sentence length is set to 512 and the dropout rate is set to 0.3. Our model is trained using AdamW optimizer with the learning rate amongst [2e-5,3e-5,5e-5] and the batch size is set to 16. We set the ratio of validation to 0.1 and randomly split the original training set into a new training set and a validation set. We run all experiments three times with different seeds and report the average performance using PyTorch 1.7.0 on the Nvidia RTX 3090 GPU. The accuracy and macro-F1 score are used as the main evaluation metrics in the experimental studies.

D. Main Results

The main experimental results for three datasets are shown in Table II, where we have the following observations.

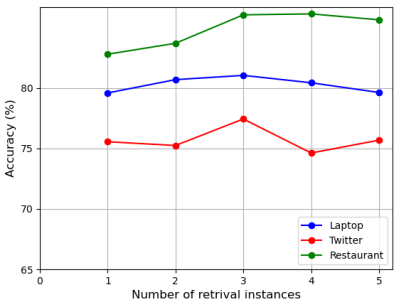
First, compared to large-scale pre-trained language models, the performance of conventional neural network models is unpromising, including the RNN-based model (ATAE-LSTM), attention-based model (AOA), GCN-based model (ASGCN) and memory network model (MemNet, RAM). On the contrary, BERT-based models have achieved effective results, indicating that the rich knowledge of PLMs captured from large-scale unlabeled data is important for downstream tasks. That is the reason why we choose BERT as base encoder to learn overall semantic representations of whole sentences.

[1] https://github.com/huggingface/transformers
[2] https://huggingface.co/bert-base-uncased

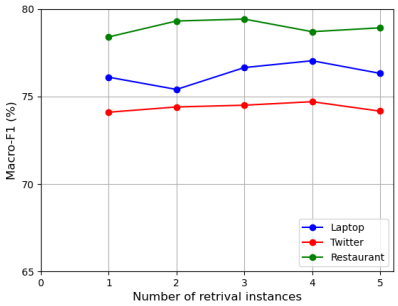
TABLE II

MAIN RESULTS. WE REPORT THE MEAN ACCURACY AND MACRO-F1 SCORE OVER 3 TIMES WITH 3 DIFFERENT SEEDS. "-" MEANS NOT REPORTED IN ORIGINAL PAPER. THE TOP 3 SCORES ARE IN **BOLD**

Category	Models	Twitter		Restaurant		Laptop	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Non-BERT-based baselines	ATAE-LSTM [18]	-	-	77.20	-	68.70	-
	AOA [36]	-	-	79.70	-	72.60	-
	MemNet [6]	-	-	80.95	-	72.37	-
	ASGCN [37]	72.15	70.40	80.86	72.19	75.55	71.05
	RAM [7]	69.36	67.30	80.23	70.80	74.49	71.35
BERT-based baselines	BERT [9]	73.42	72.17	82.40	73.17	77.29	73.36
	BERT-PT [20]	-	-	84.95	76.96	78.07	75.08
	BERT-SPC [11]	73.55	72.14	84.46	76.98	78.99	75.03
	AEN-BERT [11]	74.71	73.13	83.12	73.76	79.93	**76.31**
	BERT-ADA [38]	-	-	**85.51**	78.09	79.19	74.18
Retrieval-based (Ours)	RE-BERT (TFIDF)	**75.64**	**74.31**	84.62	77.94	**79.94**	75.89
	RE-BERT (BM25)	**76.41**	**74.68**	**86.11**	**79.92**	**81.03**	**77.00**
	RE-BERT (Dense)	**76.64**	**75.14**	**86.84**	**80.94**	**81.42**	**77.35**



(a) The performance of Accuracy in three datasets.



(b) The performance of Macro-F1 in three datasets.

Fig. 3. The ablation study on different numbers of retrieval instances which conducted on the test set of three datasets.

Second, compared the non-specific BERT models (i.e., BERT and BERT-PT) with task-specific BERT models (i.e., ANE-BERT), we find that task-specific models work better than non-specific models for ASC tasks. In BERT-based baselines, we can see the trend that AEN-BERT >BERT-SPC >BERT-PT >BERT, which illustrates that the information of aspect term is the core influence factor for the performance of the aspect sentiment classification model.

Finally, our retrieval-based method outperforms the most advanced baselines on three datasets, no matter in terms of Accuracy or Macro-F1 score. The results demonstrate that the effectiveness of the retrieval-based method could improve the performance of the aspect sentiment task. Meanwhile, it also indicates that our RE-BERT can better enhance sentence semantics by retrieving similar instances from the retrieval corpus. In addition, unlike the keyword matching method, Dense outperforms the other two methods (TFIDF and BM25), which is consistent with the previous assumption that PLMs can retrieve semantically deep-level sentences instead of surface lexical features. Meantime, it also indicates that our proposed RE-BERT can improve semantics by concatenating with retrieved sentences, which is better than other BERT plus-in components in previous methods.

V. ANALYSIS

In this section, we conduct several ablation studies and discuss a case study to further analyze the effectiveness of our method. For simplicity, we conduct experiments on the best performance retrieval metrics **Dense**.

A. Ablation Study

a) Ablations on the number of retrieval instances: In Figure 3, we study the influence of different retrieval instances from training data in our model. We experimented with the number of retrieval sentences in a range (i.e.,1 to 5). Specifically, the accuracy and macro-f1 of RE-BERT in the three datasets become higher as the number of retrieval sentences increases and reaches the optimal result around 3. As the number increases, performance tends to decline. This phenomenon indicates that retrieval sentences have limited semantic improvement, and too many similar sentences may cause deviations in the semantic understanding of the model. In addition, we can also find that RE-BERT always achieves

TABLE III
ABLATION STUDY ON DIFFERENT PRE-TRAINED LANGUAGE MODELS.

Retrieval	PLMs	Twitter		Restaurant		Laptop	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
✖	BERT-base	73.42	72.17	82.40	73.17	77.29	73.36
	BERT-large	76.84	74.91	85.74	81.03	82.02	78.15
✔	BERT-base	76.64	75.14	86.84	80.94	81.05	77.35
	BERT-large	77.43	76.23	87.04	81.42	82.52	79.23

TABLE IV
ABLATION STUDY ON THE POSITION OF CONCATENATING.

Position	Twitter		Restaurant		Laptop	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Post	76.64	75.14	86.84	80.94	81.05	77.35
Prefix	76.78	75.25	86.45	80.67	80.94	76.97
Derange	73.45	72.93	83.52	79.05	79.83	75.46

TABLE V
CASE STUDY. EXAMPLE: "EASY TO START UP AND DOES NOT OVERHEAT AS MUCH AS OTHER LAPTOPS."

Metrics	Retrieval Sentences
TFIDF	It 's fast , it 's easy easy easy to set up , easy to hook to my wireless network . The price premium is a little much , but when you start looking at the features it is worth the added cash . He loves it and it is easy to use and well the schools start teaching the kids early about computers so it was easy for him to get started .
BM25	It is easy to operate and i have already ordered more software and gadgets for my new rolls royce of laptops . Everything is so easy to use, mac software is just so much simpler than microsoft software. You can do what you want in just few seconds, even to start up your computer takes few seconds.
Dense	Easy to carry , can be taken anywhere , can be hooked up to printers , headsets . It is easy to use , fast and has great graphics for the money . The mac takes about the same amount of starting-up time as the average pc , but keeps itself cleaned up and ready to use .

a higher performance compared to the BERT-based models. All these results suggest that RE-BERT can better enhance the semantics of the whole sentence in aspect sentiment classification.

b) Ablations on the backbone of pre-trained language models: To evaluate the potential of our method, we try to use different parameter sizes of PLMs (BERT-base and BERT-large) as the encoder of **RE-BERT**. As shown in Table III, the ablation result of BERT-base with retrieval-based method rivals BERT-large without the retrieval-based method, which contains nearly three times the parameters. Compared with BERT-large without retrieval, BERT-base with retrieval has little difference of Macro-F1 on Twitter (74.91 % v.s.75.14%), Macro-F1 on Restaurant (81.03% v.s. 80.94%), Macro-F1 on Laptop (78.15% v.s. 77.35%). In addition, whether using the BERT-base or BERT-large model, the retrieval-based approach outperformed the approach without retrieval. Compared to BERT-base without retrieval, BERT-base with retrieval achieves improvements of 3.0% Macro-F1 on Twitter (72.17 % vs.75.14%), 7.7% Macro-F1 on Restaurant (73.17% vs. 80.94%) and 4.0% Macro-F1 on Laptop (73.36% vs. 77.35%).

c) Ablations on the position of concatenating: We tried to use different ways of concatenating with the original inputs and the retrieval instances. We use three types of concatenating methods: (1) **Post**, concatenating after the original input; (2) **Prefix**, concatenating before the original input; (3) **Derange**, concatenating with deranged retrieval instances. As shown in Table V, we find that **Post** and **Prefix** performed similarly, but **Derange** performed poorly. The main reason is that BERT can learn the contextual semantic representation of the whole sentence, and since the retrieval sentences have high semantic similarity with original inputs, the semantic representations learned by BERT do not differ too much, regardless of the **Post** and **Prefix** method chosen. However, the disordering may cause BERT's semantic understanding of context to be wrong, so the classification performance will be degraded.

B. Case Study

As shown in Table V, we display a case study on the Laptop dataset retrieved by three different retrieval metrics. From the example "Easy to start up and does not overheat as much as other laptops", we show the top 3 most similar sentences for each metric. We can intuitively find that the retrieved sentences show the same topic about "easy to use laptop" and the same polarity: positive. It proves that our retrieval metrics can successfully retrieve similar sentences from training data. Although we can not visually see how the neural network works with the retrieved sentences, we have shown that the

retrieved sentences have a very strong correlation with the original sentences.

VI. Conclusion

In this paper, we propose a simple and effective retrieval augmented method to fully make use of training data to enhance the semantic representations for aspect sentiment classification. Our proposed method is easily implemented with different language models and retrieval metrics on ASC task. The experimental results and analysis on three ASC benchmark datasets have shown the effectiveness and potential of the proposed method.

VII. Acknowledgment

This work was in part supported by Shenzhen Science and Technology Program (No:JCYJ20210324135809025).

References

[1] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, 2020.

[2] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2022.

[3] C. Brun, D. N. Popa, and C. Roux, "XRCE: Hybrid classification for aspect-based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 838–842.

[4] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2514–2523.

[5] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 946–956.

[6] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 214–224.

[7] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 452–461.

[8] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Attentional encoder network for targeted sentiment classification," *arXiv preprint arXiv:1902.09314*, 2019.

[12] S. Wang, Y. Xu, Y. Fang, Y. Liu, S. Sun, R. Xu, C. Zhu, and M. Zeng, "Training data is more valuable than you think: A simple and effective method by retrieving from training data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3170–3179.

[13] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang, "Generating sentences by editing prototypes," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 437–450, 2018.

[14] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," *arXiv preprint arXiv:1911.00172*, 2019.

[15] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879.

[16] U. Khandelwal, A. Fan, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Nearest neighbor machine translation," *arXiv preprint arXiv:2010.00710*, 2020.

[17] X. Ding and B. Liu, "The utility of linguistic rules in opinion mining," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 811–812.

[18] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[19] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 4068–4074.

[20] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2324–2335.

[21] Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, "Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 246–256.

[22] Z. Xiao, J. Wu, Q. Chen, and C. Deng, "BERT4GCN: Using BERT intermediate layers to augment GCN for aspect-based sentiment classification," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9193–9200.

[23] Y. Tian, G. Chen, and Y. Song, "Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2021, pp. 2910–2922.

[24] H. Qin, Y. Tian, F. Xia, and Y. Song, "Complementary learning of aspect terms for aspect-based sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 7029–7039.

[25] K. Zhang, K. Zhang, M. Zhang, H. Zhao, Q. Liu, W. Wu, and E. Chen, "Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3599–3610.

[26] C. Chen, Z. Teng, Z. Wang, and Y. Zhang, "Discrete opinion tree induction for aspect-based sentiment analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2051–2064.

[27] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *arXiv preprint arXiv:1811.01241*, 2018.

[28] J. Weston, E. Dinan, and A. Miller, "Retrieve and refine: Improved sequence generation models for dialogue," in *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 87–92.

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[30] T. B. Hashimoto, K. Guu, Y. Oren, and P. S. Liang, "A retrieve-and-edit framework for predicting structured outputs," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[31] H. Peng, A. Parikh, M. Faruqui, B. Dhingra, and D. Das, "Text generation with exemplar-based adaptive decoding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2555–2565.

[32] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[33] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A survey on retrieval-augmented text generation," *arXiv preprint arXiv:2202.01110*, 2022.

[34] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *International workshop on semantic evaluation*, 2016, pp. 19–30.

[35] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent Twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 49–54.

[36] B. Huang, Y. Ou, and K. M. Carley, "Aspect level sentiment classification with attention-over-attention neural networks," in *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer, 2018, pp. 197–206.

[37] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[38] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4933–4941.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

CH-MEAD: A Chinese Multimodal Conversational Emotion Analysis Dataset with Fine-Grained Emotion Taxonomy

Yu-Ping Ruan, Shu-Kai Zheng, Jiantao Huang, Xiaoning Zhang, Yulong Liu, and Taihao Li*
Zhejiang Lab, Hangzhou, P.R.China
{ypruan, zhengsk, jthuang, zhangxn, liuyl2020, lith}@zhejianglab.com
* Corresponding Author

***Abstract*—Emotion recognition in conversations (ERC) is important for building human-like chatbots. Much research work has been devoted to the English community, however, there is still missing large-scale multimodal ERC dataset for the Chinese community. On the other hand, emotion states in the dynamic flow of conversations are typically subtle and complex. But currently most adopted emotion taxonomy for emotion recognition in conversations is based on Ekman's six basic emotion categories [1], which cannot cover the diverse set of emotions occurring in multi-turn dialogues, such as the emerging states of anxiety, puzzled and so on. In this paper, we present a Chinese multimodal conversational emotion analysis dataset (CH-MEAD), which contains $25,292$ video segments (utterances) labeled for 26 emotion categories (including *neutral*) and are collected from more than 400 Chinese TV series, movies and show. To our best knowledge, the CH-MEAD is the first large-scale Chinese multimodal conversational dataset with fine-grained emotion taxonomy, which is suitable for studying the speaker's subtle and complex emotion states in dialogues and will also promote the development of multimodal ERC in Chinese research community. Based on the CH-MEAD dataset, we propose a speaker-aware contextual multimodal fusion (SCMF) network and demonstrate its efficiency over the current SOTA models in our experiments[1].**

I. Introduction

With the popularity of conversational AI research, there has been a growing interest in the topic of multimodal emotion recognition in conversations (multimodal ERC) [2], [3]. The multimodal ERC task aims to identify the emotion of each utterance in a conversation based on the speaker's audio, visual, and textual modality information, which is crucial for building human-like chatbots [4] and has potential applications in many scenarios, such as education, health-care [5], and opinion mining in social media [6].

In the past decades, significant research work has been devoted to the multimodal ERC task. Among these, several manually annotated English multimodal emotional conversational datasets have been released, including IEMOCAP [7], MELD [8], and SEMAINE [9]. Though above public datasets have greatly promoted the research development in the field of multimodal ERC [2], [3], the vast majority of them adopt the Ekman's six basic emotion categories (joy, anger, fear, sadness, disgust, and surprise) [1], which covers a limited emotion taxonomy. Indeed, accompanied with the moving forward of conversations, some subtle and complex emotion states (i.e., curiosity, anxiety, etc.) which are not included in the Ekman's six basic emotion categories, frequently appear in dialogues. In view of this, a limited work has been devoted to building datasets with more emotion categories. [10] proposed a multimodal ERC dataset (CHEAVD) with 26 emotion categories. However, there is no textual modality in CHEAVD dataset, which is essential for multimodal emotion recognition. Moreover, the CHEAVD dataset has only $2,322$ samples, which is limited for the ERC task.

TABLE I
THE EMOTION TAXONOMY OF THE CH-MEAD DATASET, WHICH INCLUDES 8 PRIMARY EMOTIONS, 17 COMPOUND EMOTIONS, AND THE *neutral* CATEGORY.

Type	Categories
primary emotions	joy (高兴), sadness (悲伤), anger (愤怒), fear (恐惧), trust (信任), disgust (厌恶), surprise (惊讶), anticipation (期待)
compound emotions	love (爱), alarm (警觉), remorse (懊悔), contempt (蔑视), optimism (乐观), curiosity (好奇), pride (得意), guilt (内疚), envy (妒忌), disappointment (失望), pessimism (悲观), anxiety (焦虑), shy (害羞), puzzled (困惑), hesitant (犹豫), sympathetic (同情), schadenfreude (幸灾乐祸)
others	neutral (中立)

Compared with Ekman's basic emotion taxonomy, the Plutchik's wheel of emotions [11] is much more diverse and comprehensive, in which eight primary bipolar emotions are suggested, i.e., joy v.s. sadness, anger v.s. fear, trust v.s. disgust, and surprise v.s. anticipation. Also, [11] claimed that the primary emotion categories can be combined to represent the diverse set of human's subtle and complex emotions. Specifically, there are more than 20 binary compound emotions, such as love (joy + trust) and curiosity (trust + surprise) based on the primary emotions. Therefore, the Plutchik's wheel of emotions provides us a better tool for the modelling of human subtle and complex emotion states in dialogues.

[1]Our dataset will be released upon accepted.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Considering that the existing research work are most devoted to the English community, and there is still missing large-scale Chinese multimodal ERC dataset, this paper presents a Chinese multimodal conversational emotion analysis dataset (CH-MEAD) which contains audio, visual, and textual modality information, to lay foundations for the further development of multimodal ERC in the Chinese research community. We build the CH-MEAD's emotion taxonomy mainly based on the Plutchik's wheel of emotions [11] to alleviate the deficiency of the Ekman's six basic emotion categories. Specifically, We first include all eight basic emotions in our emotion taxonomy. For the compound categories in Plutchik's wheel of emotions, we filtered the ones which are rare in our daily life and cannot be clearly defined in psychology (e.g., *fatalism*) based on the votes of 5 recruited psychological professionals, and finally 12 compound emotions are kept. Considering the specificity of different cultures, we added another 5 compound emotions which are common in Eastern culture [12], i.e., shy, puzzled, hesitant, sympathetic, and schadenfreude, according to the advice of psychological professionals. The final 26 emotion categories (including *neutral*) are present in Table I. To ensure the diversity of samples in the annotated dataset, such as the variety of topics and speakers, we collected more than 400 Chinese TV series, movies and show from a large TV resource pool (> 1000) for the construction of our CH-MEAD dataset. As known, the generalization ability of model is highly correlated with the diversity of training samples. To ensure the quality of CH-MEAD dataset, we selected high quality video segments from the raw large TV resource pool by the recruited experts over two months. Then the annotation on above selected valid data are conducted by trained annotators over three months.

Based on the CH-MEAD dataset, we propose a speaker-aware contextual multimodal fusion (SCMF) network as a baseline, which first utilizes three long short-term recurrent neural networks (LSTM-RNNs) [13] to model the contextual information along the audio, visual, and textual modality respectively in a speaker-aware manner, and then performs the modality fusion based on several Transformer encoder networks [14]. Experimental results demonstrate the effectiveness of our SCMF model, and show that multimodal information is important for the emotion recognition in conversations.

II. Related Work

A. Multimodal Emotion Datasets

There has been much research work devoted to the task of multimodal emotion analysis. *The current available multimodal datasets for emotion or sentiment analysis are most non-conversational*, such as CMU-MOSI [15], CMU-MOSEI [16], MOUD [17], ICT-MMMO [18], YouTube [19], and CH-SIMS [20]. For the multimodal emotion analysis in conversations, several manually annotated datasets have been released, including the IEMOCAP [7], MELD [8]. However, all the above datasets built their emotion taxonomy based on the Ekman's six basic emotion categories, which are insufficient for the research of human dynamic and subtle emotion transition process in dialogues. [10] present a Chinese multimodal dataset with fine-grained emotion taxonomy (more than 20 emotion categories), i.e., CHEAVD, for the emotion analysis in conversations, but it only has the audio and visual modality and do not provides the textual information, which are important for the multimodal ERC task. And its emotion taxonomy are completely defined by annotators, which is lack the support of the psychological theory.

B. Emotion Taxonomy

Most existing emotion analysis datasets are constructed based on the six basic emotion categories proposed in [1] or along the affective dimensions (valence and arousal) that underpin the circumplex model of affect [21], [22]. Generally, the emotion taxonomy based on discrete categories are more intuitively interpretable than those based on the continuous affective dimensions. So this paper focused on the discrete emotion analysis.

Indeed, in the scenarios of multi-turn dialogue, the speaker's emotion expression is a dynamic process accompanied with the conversations moving forward, in which human emotion states are usually subtle and complex, such as the emerging of curiosity and anxiety, which are not included in the Ekman's six basic emotion categories but appears frequently in our daily life. Recent advances in psychology research have offered new conceptual and methodological approaches to capture the subtle human emotion expressions by studying the distribution of emotion responses to a diverse array of stimuli [23]. Studies guided by these principles have demonstrated that the emotion states of human are complex and delicate [24], [25], [26], which cannot be covered by the simple Ekman's six basic emotion categories. Plutchik's wheel of emotions [11] defines 8 primary bipolar emotions from four different dimensions and claims that the primary emotion categories can be combined to represent the diverse set of human subtle and complex emotions. Specifically, there are more than 20 compound emotion categories by combining the primary emotions. Therefore, the Plutchik's wheel of emotions offers researchers a better tool of diving deep into the human emotion analysis in conversations.

III. CH-MEAD Dataset

A. Data Collection

To ensure the diversity of samples in CH-MEAD dataset, we collected the raw data from a variety of Chinese television series, movies, and shows (a large pool of more than $1,000$ TV series/movies/shows, which are crawled from the Internet)[2]. Compared with the unimodal datasets, the requirements of multimodal datasets are relatively high. The principles for the data selection are listed as below:

[2]We consulted a legal office to verify that the academic usage and distribution of very short length videos fall under the fair use category.

1) The themes of candidate television series, movies, and show must come from real life scenarios. Fantasy novels and so forth are not accepted.
2) Mandarin is preferred and materials with regional accent are excluded. To ensure the clarity of the speaker's voice, the video segments with background music (sound) are also excluded.
3) The speaker's face and voice must appear in the frames at the same time, and there should be only the speaker appears in the pictures (no other person).

To ensure the quality of each single modality, we set individual standards for the audio, visual, and textual content respectively.

- *Audio & Textual*: The voice of the speaker must be clear, and if there's background noise, the SNR of the audio should $\geq 15db$. The textual transcription, including punctuations, must be accurate.
- *Visual*: The resolution of the video segments should $\geq$ 480p. The video pictures should be bright enough. In addition, only the video segments with camera angle $\leq 30°$ are selected.

We recruited 10 experts for the data selection. To ensure the judges are qualified, we showed them acceptable and unacceptable examples for each selection criterion and then asked them to finish a test. And only those passed the test can join the data selection process. We used the Adobe Premiere Pro[3] to crop the raw videos into utterance segments. Finally, $10,000$ video segments (from more than 400 Chinese television programs) are selected from the original video database by the qualified experts over two months.

B. Taxonomy of Emotions

As described in Section I, we build a fine-grained emotion taxonomy based on the Plutchik's wheel of emotions [11]. Specifically, we first include the primary eight emotion, i.e., joy, sadness, anger, fear, trust, disgust, surprise, and anticipation, into our emotion taxonomy. For the original compound emotion categories in Plutchik's wheel of emotions, we recruited 5 psychological professionals and asked them to vote for the ones which are common in our daily life and can be clearly identified by ordinary annotators. Finally, 12 compound emotion categories, i.e., love, alarm, remorse, contempt, optimism, curiosity, pride, guilt, envy, disappointment, pessimism, and anxiety are reserved. Considering the specificity of different cultures [12], the recruited 5 psychological professionals are asked to discuss other compound emotions which are common in Eastern culture but missing in Plutchik's wheel of emotions, and finally 5 compound emotions are included, i.e., shy, puzzled, hesitant, sympathetic, and schadenfreude. The final emotion taxonomy adopted in our dataset is shown in Table I.

C. Annotation

The data annotation in our work is performed by the recruited graduate annotators. We trained the annotators to minimize subjectivity and inconsistency as much as possible. Specifically, we first asked the professionals in psychology to label a subset of the valid data (300 utterances) collected in Section III-A to build a ground truth set. Then we show annotators the definition of each emotion category as in Table II, and half of the samples in the ground truth set are used to further illustrate different emotions and train the annotators, and the other half are used for testing. Only those who passed the test are qualified to participate in the data annotation.



Fig. 1. The online annotation interface used for building our dataset.

With reference to the number of annotators in existing multimodal ERC datasets (mostly 3-5 annotators) [7], [8], [10] and considering the cost of labeling (i.e., time and money), we assigned three annotators for each sample. The annotators were asked to label the emotion category for each utterance segment under the taxonomy defined in Table I. Except for these, the gender and age of the speaker are also required to mark. Specifically, the age is labeled as one of the child, teenagers, youth, middle-aged, and elderly. The online annotation interface present to the annotators is shown in Figure 1.

For each utterance segment, the majority annotated emotion was decided as the final label. And those without agreement (less than 2 consistent annotations) were discarded. Finally, we collected $8,787$ dialogs by three annotators over three months, which contain $25,292$ utterances totally from the $10,000$ valid video segments collected in Section III-A. The Fleiss's kappa score [27] on the final CH-MEAD dataset is 0.79, which demonstrates a substantial agreement among annotators.

Statistics & Discuss: The summary statistics of our labeled CH-MEAD dataset are present in Table III. We find that the speaker's gender distribution is very balanced, and most of the speakers are youth or middle-aged. Also, we find that about half of the dialogs collected in the CH-MEAD dataset has only 2 utterances, actually, this is because we adopted a very rigid standard for the data selection in Section III-A to ensure the quality of data. Table IV presents the number of utterance segments for each emotion category, we find that there are few examples for some emotion categories, such as fear and envy. We suspected that these emotions are rare in TV series, movies, and shows. We kept those emotion categories with few examples in our CH-MEAD dataset to provide an entrance to study the few-shot learning for the multimodal ERC task.

[3]https://www.adobe.com/products/premiere.html

TABLE II
THE DEFINITION OF EACH EMOTION PRESENT FOR THE ANNOTATORS.

Emotion	Definition
joy	Feel happy and comfortable
sadness	the feeling of being sad
anger	strong feeling of displeasure and hostility
fear	an emotion experienced in anticipation of some specific pain or danger
trust	the belief that somebody/something is good, sincere, honest, etc.
disgust	a strong feeling of dislike for somebody/something that you feel is unacceptable
surprise	an event, a piece of news, etc. that is unexpected or that happens suddenly
anticipation	a feeling of excitement about something (usually something good) that is going to happen
love	a very strong feeling of liking and caring for somebody/something
alarm	fear and worry that somebody feels when something dangerous or unpleasant might happen
remorse	the feeling of being extremely sorry for something wrong or bad that you have done
contempt	the feeling that somebody/something is without value and deserves no respect at all
optimism	a feeling that good things will happen and that something will be successful
curiosity	a strong desire to know about something
pride	a feeling of being pleased or satisfied that you get when you own something that other people admire
envy	feeling of discontent caused by sb else's good fortune or success
disappointment	a feeling of dissatisfaction that results when your expectations are not realized
pessimism	the feeling that things will turn out badly
anxiety	the state of feeling nervous or worried that something bad is going to happen
guilt	the unhappy feelings caused by knowing or thinking that you have done something wrong
schadenfreude	a feeling of pleasure at the bad things that happen to other people
shy	nervous or embarrassed about meeting and speaking to other people
puzzled	unable to understand something or the reason for something
hesitant	slow to speak or act because you feel uncertain, embarrassed or unwilling
sympathetic	kind to somebody who is hurt or sad

TABLE III
THE SUMMARY STATISTICS OF OUR LABELED CH-MEAD DATASET.

Number of dialogs	8,787
Number of utteraces	25,292
Average tokens per utterance	21.9
Number of emotions	25 + neutral
Dialog length	1: 9.7% 2: 44.2% 3: 22.0% 4: 11.5% ≥5: 12.6%
Speaker gender	male: 50.7% female: 49.3%
Speaker age	child: 1.9% teenager: 1.4% youth: 63.5% middle-aged: 30.0% elderly: 3.2%

TABLE IV
NUMBER OF UTTERANCE SEGMENTS FOR EACH EMOTION CATEGORY IN OUR LABELED CH-MEAD DATASET.

Emotion	#Num.	Emotion	#Num.
neutral (中立)	3606	joy (高兴)	2008
sadness(悲伤)	667	anger (愤怒)	2070
fear (恐惧)	153	trust (信任)	269
disgust (厌恶)	823	surprise (惊讶)	760
anticipation (期待)	1821	love (爱)	207
alarm (警觉)	228	remorse (懊悔)	217
contempt (蔑视)	1635	optimism (乐观)	582
curiosity (好奇)	2982	pride (得意)	1004
envy (妒忌)	120	disappointment (失望)	340
pessimism (悲观)	455	anxiety (焦虑)	1910
guilt (内疚)	399	schadenfreude (幸灾乐祸)	187
shy (害羞)	334	puzzled (困惑)	1594
hesitant (犹豫)	615	sympathetic (同情)	306

D. Feature Extraction

We trained three classifiers based on the utterance level audio, visual, and textual modality information respectively, and use these classifiers to extract the feature vector for each utterance from different modality information. Followings are the details:

1) Audio feature: For the input audio of each utterance, we first extract the 128-dim fbank acoustic features with the frame size of 50ms and frame shift of 12.5ms. Then the sequence of fbank acoustic features are input to a 1-layer long short-term memory recurrent neural network (LSTM-RNN), the max-pooling of the hidden states are used as the final audio feature representation for corresponding utterance, which has a dimension of 768.

2) Visual feature: Frames are extracted from the original video segments at 30Hz. The area of speaker's face for each image frame are cropped by the bounding box, which are derived by the MTCNN face detection algorithm [28]. Then

the cropped face images are all resized to 240×240 and input to the EfficientNet-B1 [29]. The max-pooling of the last feed-forward (FC) layer in EfficientNet-B1 is used as the final vector representation for corresponding utterance. The final extracted visual feature vector for each utterance has a dimension of 1280.

3) Textual feature: We fine-tune the pretrained BERT-base [30] using the text of each utterance as input, and the emotion label as the target. Specifically, the PyTorch implementation of BERT with the pretrained model file `bert-base-chinese` provided by Google was employed. The vector representation for token "[CLS]" of last Transformer layer is used as the final texutal feature representation for corresponding utterance, which has a dimension of 768.

IV. SCMF Network

Based on the CH-MEAD dataset, we propose a speaker-aware contextual multimodal fusion (SCMF) network for the emotion recognition in conversations. Figure 2 presents the overview architecture of our proposed SCMF model. Followings are the details:

For an input conversation with n utterances, its extracted audio, visual, and textual features are $[\mathbf{u}_1^a, \mathbf{u}_2^a, ..., \mathbf{u}_n^a]$, $[\mathbf{u}_1^v, \mathbf{u}_2^v, ..., \mathbf{u}_n^v]$, and $[\mathbf{u}_1^t, \mathbf{u}_2^t, ..., \mathbf{u}_n^t]$ respectively (see Section III-D). To better model the dependency among different speakers, the unimodal utterance representations are first concatenated with corresponding speaker embedding and then input to corresponding LSTM-RNN networks separately. The derived contextual unimodal representations are $[\mathbf{h}_1^a, \mathbf{h}_2^a, ..., \mathbf{h}_n^a]$, $[\mathbf{h}_1^v, \mathbf{h}_2^v, ..., \mathbf{h}_n^v]$, and $[\mathbf{h}_1^t, \mathbf{h}_2^t, ..., \mathbf{h}_n^t]$ respectively, which are the hidden states of corresponding LSTM-RNN cells.

The multimodal fusion module (MFM) in SCMF is built based on Transformer [14] encoder, which has been proved to be effective for the fusion of multimodal information [31]. As illustrated in Figure 2, a MFM network contains two unique Transformer encoders, which process the input two modality representation matrices in opposite directions. Specifically, for two different modality representation $H^A = [\mathbf{h}_1^A, \mathbf{h}_2^A, ..., \mathbf{h}_n^A]$ and $H^B = [\mathbf{h}_1^B, \mathbf{h}_2^B, ..., \mathbf{h}_n^B]$, the MFM module process them as follows,

$$\begin{aligned} \tilde{\mathbf{h}}_1^A, ..., \tilde{\mathbf{h}}_n^A &= Trans.(Q = H^A, (K, V) = H^B), \\ \tilde{\mathbf{h}}_1^B, ..., \tilde{\mathbf{h}}_n^B &= Trans.(Q = H^B, (K, V) = H^A), \\ \mathbf{u}_i^{AB} &= Concat([\tilde{\mathbf{h}}_i^A, \tilde{\mathbf{h}}_i^B]), i \in [1, n], \\ & A \in [a, v, t], B \in [a, v, t], A \neq B \end{aligned} \tag{1}$$

It's worth mentioning that the causal attention mask is applied for all Transformer encoders to ensure the subsequent utterances are unseen for current step. The final multimodal representation $\tilde{\mathbf{u}}_i$ for $i-$th utterance is the concatenation of $\mathbf{u}_i^{av}, \mathbf{u}_i^{at}, \mathbf{u}_i^{vt}$,

$$\tilde{\mathbf{u}}_i = Concat([\mathbf{u}_i^{av}, \mathbf{u}_i^{at}, \mathbf{u}_i^{vt}]), \tag{2}$$

The classification module in the SCMF model is a fully connected network with one linear hidden layer. It accepts the $\tilde{\mathbf{u}}_i$ as input and makes the final emotion prediction,

$$\begin{aligned} probs_i &= softmax(W\tilde{\mathbf{u}}_i + \mathbf{b}), \\ \hat{y}_i &= \arg\max_k(probs_i[k]). \end{aligned} \tag{3}$$

in which k is number of emotion categories.

V. Experiments

A. Baseline Models

Previous studies have shown that the utterance-level contextual modelling is essential for the emotion recognition in conversations [3], [32]. So we compare our SCMF model with current state-of-the-art models which consider the modelling of multi-turn dialogue context. The baseline models used in our experiments are as follows:

- **ICON**: [3] proposed the interactive conversational memory network (ICON), in which several gated recurrent units (GRUs) [33] are utilized to model the dialog context from self- and inter-speaker view respectively, and a memory block is also introduced to summary the global contextual information. *Different modality representations for each utterance are directly concatenated together and input to downstream networks in original ICON model.*
- **DialogueRNN**: The DialogueRNN model present in [34] treats parties in a conversation individually, and applied different GRUs for the contextual information modelling for different speakers. Also, it introduces another GRUs for the dynamic modelling of emotion states. *The audio, visual, and textual utterance representations are concatenated together and input to GRUs in DialogueRNN model.*
- **CIM**: [35] proposed the contextual inter-modal (CIM) attention model, in which the utterance-level unimodal representations are input to three different bi-directional GRUs respectively to perform the dialogue context modelling. Then an inter-modal attention module is utilized to perform the multimodal information fusion.
- **DAG-ERC**: [36] proposed the DAG-ERC model for emotion recognition in conversations, which combined the directed acyclic graph network with recurrent neural network to model the conversational context, and achieved impressive results in the ERC task. Similar with ICON and DialogueRNN, *three different modality representations for each utterance are concatenated together for downstream processing in DAG-ERC.*
- **MMGCN**: [37] proposed the MMGCN model to utilize both multimodal and long-distance contextual information based on multimodal fused graph convolutional network for the multimodal ERC task. The MMGCN model can not only make use of multimodal dependencies effectively, but also leverage speaker information to model inter-speaker and intra-speaker dependency.

To make the comparison experiments between different models more fair, **the extracted unimodal features in Section**

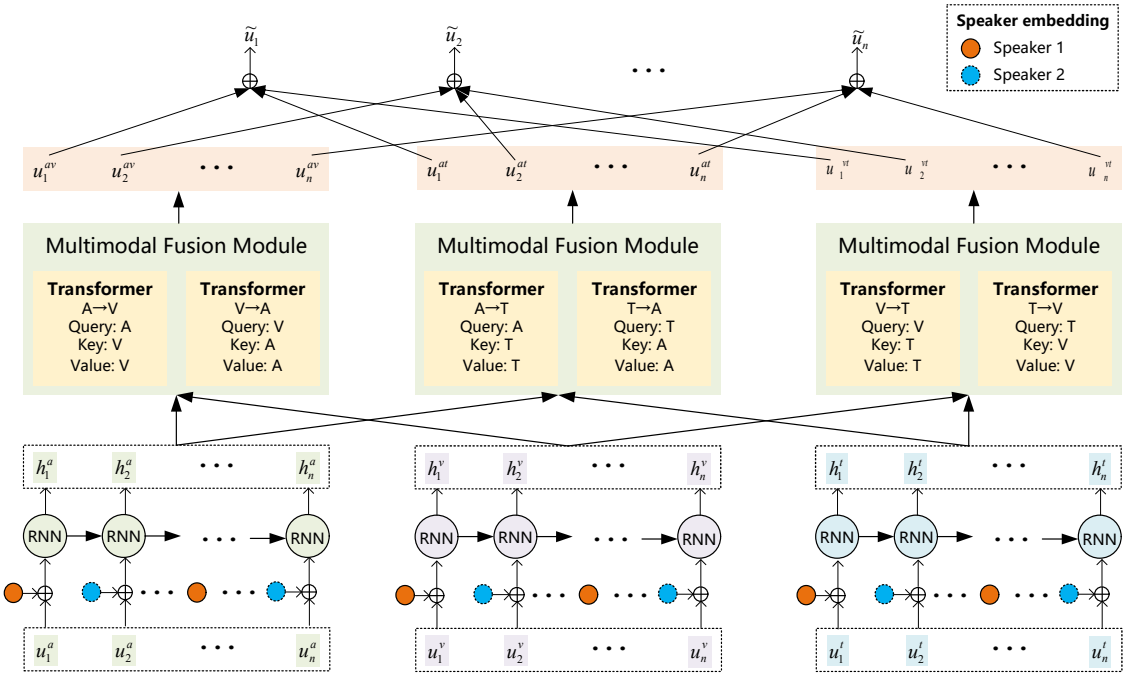


Fig. 2. The architecture of our proposed SCMF model.

TABLE V
THE OVERALL PERFORMANCE OF DIFFERENT MODELS ON THE 9-CLASS'S SUBSET OF CH-MEAD DATASET.

	Acc.	W.- F1	F1								
			joy	sad.	anger	fear	trust	disgust	surprise	anticipation	neutral
ICON	49.0	48.5	56.0	32.0	48.0	21.0	12.0	**29.0**	43.0	50.0	59.0
DialogueRNN	52.9	51.6	58.0	34.0	59.0	28.0	18.0	19.0	50.0	51.0	61.0
CIM	53.0	49.5	58.1	25.9	**59.3**	8.7	7.6	6.1	43.2	51.3	61.4
DAG-ERC	52.3	52.0	59.0	31.2	55.3	38.1	10.5	25.3	50.6	**56.1**	58.5
MMGCN	51.8	50.8	50.9	25.0	58.6	**58.9**	**40.4**	3.3	2.3	24.0	59.5
SCMF	**54.9**	**54.2**	**59.9**	**38.5**	59.1	37.8	18.3	28.4	**52.8**	51.9	**63.1**

III-D are used as input for all baseline models and our proposed SCMF model.

B. Parameter Settings

The number of layers of LSTM-RNNs in our SCMF model are all set to 1. The dimension of the speaker embedding and hidden state of LSTM cell are set to 16 and 768 respectively. The number of layers, number of heads, dimension of hidden states for each Transformer encoder are set to 1, 16, and 768 respectively. The method of Adam [38] is adopted for optimization. The initial learning rate is 0.0001. Gradient clipping is set to 1, and the batch size is 32. Dropout with rate of 0.1 is used in the training process.

VI. RESULTS

We use the accuracy and weighted-average F1 score (denoted as Acc. and W.- F1 respectively) as the evaluation metrics in our experiments. We perform the experiments on 9-class's subset (8 primary emotion categories and the *neutral*) and 26-class's fullset respectively. The dataset are randomly split into train/dev/test set at a ratio of 3:1:1. The number of train/dev/test utterances in 9-class's subset and 26-class's fullset are $7,316/2,450/2,411$ and $15,161/5,074/5,057$ respectively.

A. Overall Performance

1) 9-class's subset: The overall evaluation results of different models on the 9-class's subset of CH-MEAD dataset are present in Table V. From Table V, we find that most models obtained better prediction scores on joy, anger, and neutral categories than other emotion categories with few samples. Compared with the baseline models which all utilized the same extracted unimodal features, we find that our proposed SCMF model achieved the best performance on both accuracy and weighted-average F1 score metrics, which obtains an improvement of 1.9% (over CIM) and 2.2% (over DAG-ERC) respectively .

2) 26-class's fullset: Table VI shows the performance of different models on the 26-class's fullset of CH-MEAD dataset. It can be found that our SCMF model outperformed other baseline models with obvious improvements, i.e., 2.3% (over CIM) on accuracy and 4.6% (over DAG-ERC) on

TABLE VI
THE OVERALL PERFORMANCE OF DIFFERENT MODELS ON THE 26-CLASS'S FULLSET OF CH-MEAD DATASET.

	Acc.	W.- F1
ICON	28.3	23.2
DialogueRNN	35.2	31.0
CIM	35.7	32.1
DAG-ERC	32.1	32.1
MMGCN	35.2	30.9
SCMF	**38.0**	**36.7**

TABLE VII
THE DETAILED F1 SCORE OF SCMF MODEL ON THE 26-CLASS'S FULLSET OF CH-MEAD DATASET.

Emotion	F1	Emotion	F1
neutral (中立)	46.2	joy (高兴)	48.3
sadness(悲伤)	19.6	anger (愤怒)	43.9
fear (恐惧)	35.3	trust (信任)	13.3
disgust (厌恶)	14.8	surprise (惊讶)	26.6
anticipation (期待)	39.6	love (爱)	41.8
alarm (警觉)	15.1	remorse (懊悔)	55.4
contempt (蔑视)	23.6	optimism (乐观)	20.9
curiosity (好奇)	58.2	pride (得意)	22.3
envy (妒忌)	32.4	disappointment (失望)	2.7
pessimism (悲观)	20.9	anxiety (焦虑)	33.4
guilt (内疚)	42.5	schadenfreude (幸灾乐祸)	8.3
shy (害羞)	35.4	puzzled (困惑)	38.5
hesitant (犹豫)	22.6	sympathetic (同情)	20.4

weighted-average F1 score metrics respectively. Table VII presents the detailed F1 score of SCMF model achieved on each emotion category, it can be found that there's a lot of room for improvement on a bunch of emotion categories, such as sympathetic, alarm, and trust, which has few samples, we believe these provide an entrance to study the few-shot learning for multimodal ERC task. Also, we find that the model's performance on different emotion categories is not proportional to the number of samples in each category, such as pride vs. remorse (F1: 24.1 vs. 62.3, #number: 1004 vs. 217), which indicates that some emotion categories are easy to distinguish, while others are difficult.

B. Ablation Studies

To inspect the effects of speaker information and fusion of different modalities in the SCMF model, we conducted several ablation studies. Specifically, we remove all speaker embeddings in the SCMF model and directly input the original extracted unimodal features of each utterance into the LSTM-RNNs to leave out the speaker information. To examine the importance of unimodal information, we use the contextualized unimodal representation output from the LSTM-RNNs, i.e., $\mathbf{h}_i^a$, $\mathbf{h}_i^v$, and $\mathbf{h}_i^t$, for the final emotion classification to evaluate the importance of audio, visual, and textual modality respectively. For the fusion of different modalities, we eliminate three unimodal features respectively and directly use the utterance encodings output from the MFM module, i.e., $\mathbf{u}_i^{av}$, $\mathbf{u}_i^{at}$, and $\mathbf{u}_i^{vt}$, for the emotion classification respectively.

TABLE VIII
THE ABLATION STUDIES FOR THE SCMF MODEL WITH CONSIDERING THE SPEAKER EMBEDDING AND FUSION OF DIFFERENT MODALITIES.

	9-class		26-class	
	Acc.	W.- F1	Acc.	W.- F1
A	33.9	25.4	18.1	9.4
V	41.1	35.4	22.5	15.5
T	48.2	47.7	33.2	31.0
A+V	46.2	42.1	25.4	20.9
A+T	49.6	48.8	34.6	33.2
V+T	52.0	51.2	36.5	35.1
A+V+T	54.9	54.2	38.0	36.7
w/o spk.	54.3	53.5	37.8	36.4

The evaluation results are present in Table VIII. We find that the performance of SCMF model slightly declined when without using speaker embeddings, which indicates the speaker identity information is useful for the multimodal emotion recognition in conversations.

For the effects of different unimodal information, we find that the textual input occupies the most important position, which owns obvious performance advantages over the audio and visual modality. And the audio inputs obtains the worst performance on both accuracy and weighted-averaged F1 score metrics. For the fusion of different two modalities, it can be found that the combination of visual and textual modalities, i.e., "V+T", outperformed the "A+V" and "A+T". Above comparison results imply the primary effects of textual information for the emotion recognition in conversations. Also, the results in Table VIII confirm the necessity of comprehensive utilization of audio, visual, and textual modality information for the speaker's emotion recognition.

VII. CONCLUSION

In this paper, we have presented a Chinese multimodal conversation emotion analysis (CH-MEAD) dataset, which contains $25,292$ video segments labeled for 26 emotion categories and collected from more than 400 Chinese TV series, movies, and shows. To our best knowledge, the CH-MEAD is the first large-scale Chinese multimodal conversation dataset with fine-grained emotion taxonomy, which allows researchers to dive deep into the modeling of speaker's subtle and complex emotional states in conversations, and provides an entrance to study the few-shot learning on the multimodal ERC task. Also, we proposed a speaker-aware contextual multimodal fusion (SCMF) network, which first models the contextual information along the audio, visual, and textual modality respectively in a speaker-aware manner, and then performs the modality fusion based on Transformer encoders. Experimental results demonstrate the effectiveness of our SCMF model and the importance of multimodal information for emotion recognition in conversations. To study the few-shot learning on the CH-MEAD dataset will be our future work.

ACKNOWLEDGEMENTS

This work was supported in part by the Youth Foundation Project of Zhejiang Province (Grant No. LQ22F020035), the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2021ZD0114303), and the Key Research Project of Zhejiang Lab (No. 2020KB0AC01).

REFERENCES

[1] Paul Ekman, "Are there basic emotions?," 1992.

[2] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, 2017, pp. 873–883.

[3] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *EMNLP*, 2018, pp. 2594–2604.

[4] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Mime: Mimicking emotions for empathetic response generation," in *EMNLP*, 2020, pp. 8968–8979.

[5] Tim Althoff, Kevin Clark, and Jure Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463–476, 2016.

[6] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal, "Semeval-2019 task 3: Emocontext contextual emotion detection in text," in *SemEval*, 2019, pp. 39–48.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *ACL*, 2019, pp. 527–536.

[9] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

[10] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu, "Cheavd: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 913–924, 2017.

[11] Robert Plutchik, "A psychoevolutionary theory of emotions," 1982.

[12] Alvin C Dueck, Yin Yang, et al., "Understanding emotion in chinese culture: Thinking through psychology," 2017.

[13] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[16] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018, pp. 2236–2246.

[17] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Utterance-level multimodal sentiment analysis," in *ACL*, 2013, pp. 973–982.

[18] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[19] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[20] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *ACL*, 2020, pp. 3718–3727.

[21] James A Russell, "Core affect and the psychological construction of emotion.," *Psychological review*, vol. 110, no. 1, pp. 145, 2003.

[22] Sven Buechel and Udo Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *EACL*, 2017, pp. 578–585.

[23] Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner, "Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 69–90, 2019.

[24] Alan S Cowen and Dacher Keltner, "What the face displays: Mapping 28 emotions conveyed by naturalistic expression.," *American Psychologist*, vol. 75, no. 3, pp. 349, 2020.

[25] Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nature human behaviour*, vol. 3, no. 4, pp. 369–382, 2019.

[26] Alan S Cowen and Dacher Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[27] Joseph L Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, pp. 378, 1971.

[28] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[29] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019.

[31] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL*. NIH Public Access, 2019, vol. 2019, p. 6558.

[32] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 2470–2481.

[33] Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP 2014*, 2014.

[34] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *AAAI*, 2019, vol. 33, pp. 6818–6825.

[35] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *NAACL*, 2019, pp. 370–379.

[36] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan, "Directed acyclic graph network for conversational emotion recognition," in *ACL-IJCNLP*. 2021, pp. 1551–1560, Association for Computational Linguistics.

[37] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin, "Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *ACL)*, 2021, pp. 5666–5675.

[38] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Evolutionary Analysis and Cultural Transmission Models of Color Style Distributions in Painting Arts

Eita Nakamura* and Yasuyuki Saito†
* Kyoto University, Kyoto, Japan
E-mail: eita.nakamura@i.kyoto-u.ac.jp Tel/Fax: +81-75-753-5396
† National Institute of Technology, Kisarazu College, Chiba, Japan
E-mail: saito@j.kisarazu.ac.jp Tel/Fax: +81-438-30-4137

***Abstract*—This paper studies the evolutionary dynamics and cultural transmission of color styles in painting arts. Creative cultures, such as visual arts, develop through repeated processes of knowledge transmission and modification, and characteristics of these processes can lead to macroscopic evolutionary patterns. While recent studies have found intriguing trends and dynamic patterns in the evolution of visual arts, the relationship between these trends and cultural transmission processes remains poorly understood. We first analyze the evolutionary dynamics of color distributions in oil paintings by applying a clustering method. The results reveal that some of the frequencies of the color style clusters exhibit synchronous changes that are related to creators' biases in producing paintings belonging to specific clusters. We subsequently construct models of cultural evolution to represent how these biases are transmitted between creators, based on two hypothetical transmission modes: one guided by influencer-to-descendant connections and another guided by community-to-descendant connections. Through statistical inference, we show that the influencer-guided transmission model better fits the historical data. Furthermore, highly influential creators, as inferred by the model purely from content data, remarkably align with those commonly mentioned in the literature on art history.**

I. Introduction

The recent advancement of generative AI [1]–[3] has drawn attention to the application of image processing in the field of arts, such as painting. The automatic creation technology based on machine learning enables the generation of images in diverse styles. While machine learning relies heavily on a large accumulation of image data historically created by humans, the generation of paintings in new creation styles remains a big challenge [4]. In fact, there is still limited quantitative and scientific understanding of the process in which humans develop new creation styles. If we can understand this human's creative intelligence, the knowledge can be utilized to realize automatically generating paintings in a much wider variety of styles than current systems.

Creative cultures have evolved through repeated processes of transmission and modification of creative knowledge. Cultural evolution including such cultural transmission processes has been studied based on theoretical analogies with biological evolution, and two factors, environmental and dynamical, in the evolution process are generally considered to shape evolutionary processes [5]. Environmental factors refer to various factors that generate selective pressure, and determine the direction and constraints of evolution. In the case of painting culture, these include people's preferences, the supply of painting materials, and social backgrounds such as economics and politics. On the other hand, dynamical factors refer to the effects caused by the processes of knowledge transmission, mutation, and selection, and are generally relevant for understanding the speed and regularities of evolution.

Research on the evolutionary analysis of painting art data, mainly focusing on Western art, has recently become an active field [6], thanks to the availability of large-scale digital data and the development of image analysis techniques including deep learning. Previous studies have found trends in certain time periods, in the entropy-complexity features used in physical system analysis [7] and in the features extracted by convolutional neural networks (CNNs) [8], [9]. It has also been shown that the characteristic temporal changes found in these analyses generally agree with the analysis by art experts, and the creators who represent the style of each era coincide with those widely mentioned in art history [7], [8], [10]. Another study suggested that there were significant changes in the average features of paintings, particularly during the 15th and 16th centuries, and in the late 19th century, from the analysis of color contrast, which can be associated with the popularization of oil paints and the increase in individuality in painting styles [11]. The existence of trends spanning several centuries and the presence of revolutionary periods with rapid changes have also been observed in music data, where similar evolutionary analysis has been conducted [12]–[15]. These quantitative results have revealed the relationship between the temporal changes in painting styles and social/environmental factors. However, the impact of dynamic factors due to cultural transmission on art evolution still remains poorly understood.

In this study, focusing on the fact that creators usually produce artworks in several creation styles, we investigate the impact of cultural transmission of these creative style distributions on macroscopic evolution of the painting culture. First, we analyze the color styles of oil paintings by using a clustering method and show that there are synchronous changes in the frequencies of certain color style clusters. We also confirm, through an analysis of the network of clusters obtained from creators belonging to multiple style clusters, that the creators' biased distribution of color styles had significant influence on the aforementioned synchrony.

To understand the process by which creators select the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

distribution of color styles, we construct cultural evolution models incorporating a transmission process of this distribution between creators and compare them with a random selection model. In cultural transmission, various modes of learning are generally possible, such as learning from a specific influencer, as in a master-descendant relationship, or learning from numerous creators or artworks. We here consider two extreme possibilities: *influencer-guided* transmission, where one influencer is mainly involved in the transmission, and *community-guided* transmission, where the average features from a community of creators are transmitted. We formulate the mathematical models to represent these hypothetical cases. Through statistical inference based on these models, we show that the influencer-guided transmission model better explains the historical data. Furthermore, we validate the potential of this model, which allows for the estimation of the creator's fitness (influence) from image data by unsupervised learning, by measuring the degree of match between creators with high estimated fitness and those mentioned in an influential textbook on art history.

II. Cluster Analysis of Color Style

A. Materials and method

We used image data of mostly Western art paintings obtained from a public website WikiArt.org. Some meta data are associated with the obtained image data, including the year of production, creator's name, and media (materials). For our analysis, we extracted artworks with media labeled as "oil", resulting in 32 401 artworks. We then extracted color frequency statistics from each of the images as follows. First, all images were converted to 100×100 pixels by downsampling (Fig. 1). Second, we applied color reduction to $I = 40$ representative colors, which form a set of equally spaced points in the cone-shaped HSV color space. Third, the relative frequencies $\boldsymbol{\theta}_n = (\theta_{ni})_{i=1}^{I}$ of the representative colors i for each image n were computed. The tuple $(a_n, t_n, \boldsymbol{\theta}_n)$ of creator a_n, year of production t_n, and color statistics $\boldsymbol{\theta}_n$ was used for the following analysis. The years of production in the data ranged from 1270 to 2022. The number of distinct creators was 1128.

To quantitatively study the evolution of creation styles, we analyze the temporal changes in the distribution of color statistics. To deal with the high-dimensional space of the statistics, we first apply clustering in this space to identify several concentrated regions, which we call color style clusters. Specifically, we use for clustering the discrete distribution mixture model defined as

$$P(\boldsymbol{\theta}_n) = \sum_{k=1}^{K} \pi_k P(\boldsymbol{\theta}_n | \bar{\boldsymbol{\theta}}_k),$$

where K is the number of clusters, π_k denote the probabilities of clusters (called the mixture probabilities), and $\bar{\boldsymbol{\theta}}_k$ is the mean statistics of cluster k. The k-th component discrete distribution model $P(\boldsymbol{\theta}_n | \bar{\boldsymbol{\theta}}_k)$ is defined as

$$P(\boldsymbol{\theta}_n | \bar{\boldsymbol{\theta}}_k) = \prod_{i=1}^{I} \bar{\theta}_{ki}^{L\theta_{ni}},$$

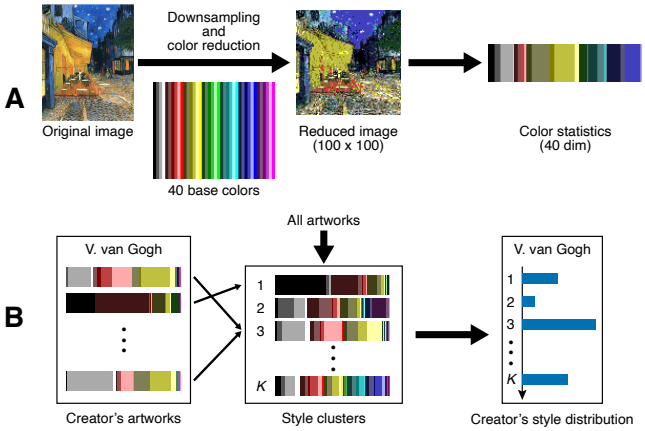


Fig. 1. Processes for analyzing color styles. A: Preprocessing for extracting color statistics. B: Color style clusters and creator's style distribution.

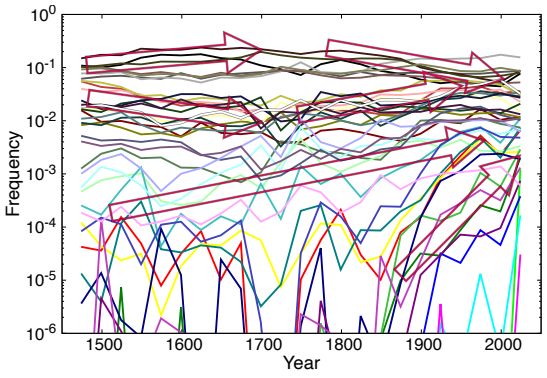


Fig. 2. Evolution of color frequencies. The plotted lines are colored in the indicating base colors (while line is shaded). Arrows indicate notable trends.

where L is an auxiliary parameter corresponding to a "sample size" to compute the probability. The clustering algorithm can be derived based on the expectation-maximization (EM) algorithm. In the E-step, the posterior probability is computed as $P(k|n) \propto \pi_k P(\boldsymbol{\theta}_n | \bar{\boldsymbol{\theta}}_k)$, where L is set as $L = 10\,t$ (t denotes the iteration time) and plays a similar role as the inverse temperature in the simulated annealing method. In the M-step, the mixture probability is updated as $\pi_k \leftarrow \sum_n P(k|n)$ and the mean statistics as $\bar{\theta}_{ki} \leftarrow \sum_n P(k|n)\theta_{ni}$. Note that we apply the clustering without using the temporal information t_n. After the convergence of the EM algorithm, we assign each artwork n to the most probable cluster $k_n = \operatorname{argmax}_k P(k|n)$, for subsequent analyses.

To get an overview of the temporal changes of the distribution of data samples, we also project the color statistics on a two-dimensional visualization space. For this purpose, we use the t-SNE method [16] with the Jensen-Shannon divergence used as the distance measure. This dimensional reduction process is conducted independently of the clustering and is used only for visualization.

B. Results

We used images whose year of production is 1450 or later because there were only a small number of images produced

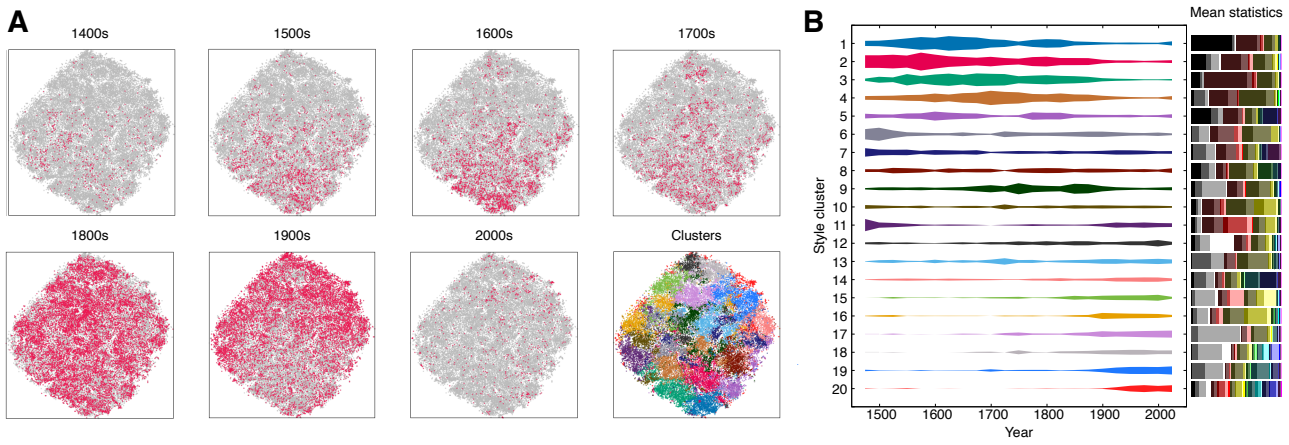


Fig. 3. A: Distribution of artworks in the visualization space. Red dots indicate artworks in each period and gray dots represent all the artworks. In the 'cluster' panel, artworks are illustrated in the colors of the clusters they belong to (cluster colors are arbitrary). B: Evolution of style cluster frequencies.

before the year. Fig. 2 shows the temporal evolution of the relative frequencies of colors, where the frequencies were calculated for each 25-years window. The most frequent colors have low brightness and low saturation, and bright and high-saturation colors are generally rare. Several trends can be observed in this figure, as indicated by arrows. First, from the 15th century to the 17th century, some of the most frequent colors increased their frequencies in the data, whereas some relatively frequent bright colors decreased their frequencies. The frequencies of these colors changed in opposite directions from the mid-18th century to the 20th century. Second, several rare bright colors had an overall gradual increase of the frequency from the 16th century to the 20th century. Last, the frequencies of vivid colors that were rarely used in early centuries increased rapidly from the mid-19th century onwards. While these general trends provide some insights, the periodwise average frequencies cannot represent variation and biases of color combinations in individual artworks. The analysis method based on style clusters solves this problem, as presented in the results below, thus providing a useful framework for studying underlying evolutionary processes.

The results of clustering and visualization of color statistics are shown in Fig. 3, where we used $K = 20$ clusters to approximate the distribution. In the visualization space, most artworks in early centuries are located in the lower half region, which represents color combinations containing mostly dark colors. Several important observations can be made in the evolution of the relative frequencies of color style clusters. First, we can observe concurrent and transient cluster structure [13], [17], that is, several clusters are active in each time period and active clusters gradually change over time. That there always exist several active clusters reflects the fact that there are usually multiple art styles and genres in any society. Second, there are two notable periods where multiple style clusters arose: from the 15th to the 16th centuries, and from the mid-19th to the mid-20th centuries. These time periods coincide with the previous result [11] showing periods with significant changes in color contrast. Third, we can observe some style clusters with synchronous rises and declines. For example, clusters 1 and 3, and clusters 17 and 18 respectively have synchronous rises in different centuries. These results indicate the presence of regularities in the evolution of color styles. In the next section, we analyze this synchrony in more detail and study its relationships with creators' behavior.

III. Synchronous Changes of Style Cluster Frequencies and Creators' Style Distributions

A. Analysis method

The relative frequencies of style clusters, as depicted in Fig. 3, can be expressed as a time-dependent mixture probability $\pi_k(t)$. The local synchrony $S_{kk'}(t)$ between the rises and declines of clusters k and k' can be formulated as

$$S_{kk'}(t) = \frac{1}{\sqrt{C_k C_{k'}}} \frac{\mathrm{d}\pi_k(t)}{\mathrm{d}t} \frac{\mathrm{d}\pi_{k'}(t)}{\mathrm{d}t},$$

where $C_k = \int \mathrm{d}t \, (\mathrm{d}\pi_k(t)/\mathrm{d}t)^2$ is a normalization constant. The (overall) synchrony can be defined as $S_{kk'} = \int \mathrm{d}t \, S_{kk'}(t)$. In the numerical analysis, the mixture probabilities are observed at discrete time steps, and the differentials and integration in these equations should be replaced with differences and summation w.r.t. time. It is noted that since the relative frequencies must sum up to unity, a rise of a cluster naturally induces a decline of another, and $S_{kk'}$ generally has a negative value if the relative frequencies of clusters k and k' fluctuate independently. Therefore, a positive value of $S_{kk'}$ indicates synchronous changes between the two clusters.

There are at least two possible factors that cause synchronous changes of cluster frequencies. First, some style clusters with similar mean color statistics, such as those located closely in the visualization space in Fig. 3, can represent separate parts of the same color style. In such a case, a change in the frequency of a color style can result in coherent changes in the frequencies of the relevant clusters. Second, the behavior of creators who usually produce artworks in several creation styles can induce the synchrony between clusters that do not necessarily represent similar color statistics. For

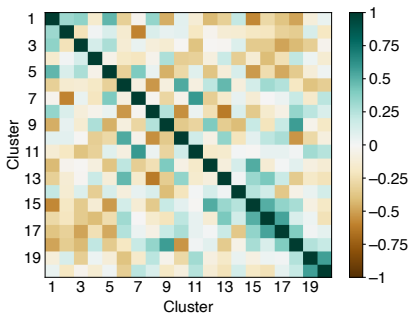


Fig. 4. Pairwise synchrony $S_{kk'}$ between clusters.

example, there are many creators who produced both portraits and landscapes, each of which are usually painted in different color combinations, and a change of artistic taste shared in a community can result in a synchronous change of color styles of both genres. To study this aspect, we calculate the *color style distribution* $\boldsymbol{\pi}_a = (\pi_{ak})_{k=1}^K$ of creator a, by counting the number of artworks by the creator belonging to cluster k and normalize the frequencies.

In both cases, some relations, feature similarity or association through creators, induce connections between clusters that lead to synchronous changes in the cluster frequencies. As in the case of association through creators, such connections can temporally change in general. These connections can be expressed as a weighted network $A(t) = \{A_{kk'}(t)\}$ of clusters, where edges $A_{kk'}(t)$ represent the degree of connection between clusters k and k' at time t. The contribution of a network on the synchrony can be quantified as $S(A) = \int \mathrm{d}t \sum_{kk'} A_{kk'}(t) S_{kk'}(t)$. To compare different networks, we impose conditions that self-connections are zero ($A_{kk}(t) = 0$) and that a normalization condition $\sum_{kk'} A_{kk'}(t) = 1$ holds.

To examine possible factors of the synchrony of style cluster frequencies, we consider the following four cluster networks. (i) Uniform network: $A_{kk'} = 1/(K-1)$ ($k \neq k'$). This is a null hypothesis. (ii) Feature similarity network: $A_{kk'} \propto \sum_i \bar{\theta}_{ki} \bar{\theta}_{k'i}$. This represents the cosine similarity between the mean statistics of the clusters. (iii) Creator-mediated association network: $A_{kk'}(t) \propto \sum_a D(t,a) \pi_{ak} \pi_{ak'}$, where $D(t,a) = 1$ if creator a is relevant at time t and 0 otherwise. This represents the co-occurrence of style clusters in the artworks produced by individual creators. (iv) Frequency-based association network: $A_{kk'}(t) \propto \pi_k(t) \pi_{k'}(t)$. This represents a hypothetical situation where creators choose the style clusters of their artworks by randomly sampling from the cluster frequencies of the time, and consequently their style distributions π_{ak} coincide with $\pi_k(t)$ of time t when creators a are active. We compare these networks in terms of the derived synchrony $S(A)$; if a network $A_{kk'}(t)$ is relevant to the macroscopic evolution of the cluster frequencies, then $S(A)$ is expected to be large.

B. Results

Fig. 4 shows the pairwise synchrony $S_{kk'}$ between style clusters, where we have used discrete time steps of 25-year width. We see some off-diagonal elements with positive values,

TABLE I
DERIVED SYNCHRONIES FOR DIFFERENT CLUSTER NETWORKS.

Network A	Derived synchrony $S(A)$
Uniform	-0.042
Feature similarity	-0.007
Creator-mediated association	0.033
Frequency-based association	-0.034

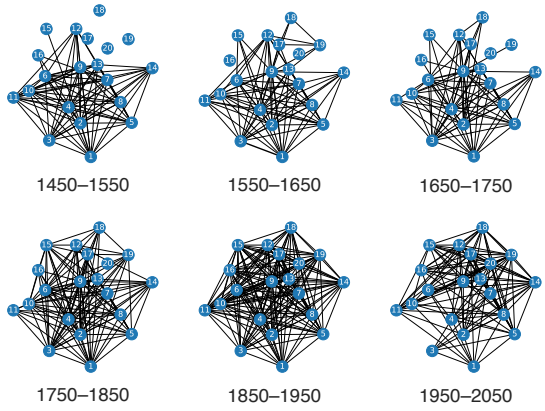


Fig. 5. Creator-mediated cluster networks in different centuries. The nodes are located similarly as the cluster centers in the visualization space in Fig. 3.

while a majority of the elements have negative values. Table I lists the derived synchronies for the four cluster networks. Only the creator-mediated association network yielded a positive synchrony, indicating that the creators' style distributions had an impact on the dynamics of style cluster frequencies and induced the synchronous changes. The result also suggests that the feature similarity of clusters also induced the synchrony. On the other hand, the small derived synchrony by the frequency-based network indicates that creators' style distributions deviated from the random selection case, in a biased and coherent (i.e. shared by many creators of the same time) way.

The creator-mediated cluster networks visualized in Fig. 5 clearly show that the association of clusters changed considerably over centuries. The connection weights between some clusters are much larger than others, and we can observe some edges between distant clusters with distinct styles also have large weights. Overall, these results show that creators' color style distributions are biased in a coherent way that cannot be simply explained by the feature similarity or random selection of style clusters, and as they vary over time these biased style distributions markedly affected the macroscopic evolution of color styles in painting arts.

IV. Cultural transmission of color style distributions

A. Cultural evolution models for creators' style distributions

The biased and coherent color style distributions of creators discussed in the previous section pose an important question: How did the creators' style distributions determine? While certain social factors such as the demand from the church or patrons are conceivable, we here focus on a simple possibility

of cultural transmission. Since complex knowledge about the art of paining is usually learned from other creators or their artworks, it is possible that the color style distribution is transmitted in this learning process.

We construct two models to examine what particular mode of cultural transmission is plausible to explain the data. In cultural phenomena, unlike in genetic evolution, the number of cultural parents is not limited to one (asexual) or two (sexual), and transmission from a varying number of cultural parents are possible. We thus consider two hypothetical situations that represent extreme cases within a broad range of possibilities. In the first *influencer-guided* transmission model, we suppose that a creator learns his/her creation style mainly from a specific influencer's artworks, as in a master-descendant relationship. In the second *community-guided* transmission model, we suppose that a creator learns the creation style from numerous creators' artworks belonging to some stylistic community. In the following, we mathematically formulate these models.

We formulate the cultural evolution models based on a stochastic description of the transmission-mutation-selection process [5]. This is a formulation similar to the models of biological evolution used in population genetics [18], except that the potential of a past creator to produce cultural offspring can in principle continue for long after his/her death. We denote by $t_a^{\rm s}$ the year that creator a started creation, which is defined as the earliest year of the creator's artworks in the data. Let $\tilde{\boldsymbol{\pi}}_a = (\tilde{\pi}_{ak})_{k=1}^K$ denote the frequency counts of style clusters in artworks of creator a. By normalizing these frequencies, we obtain the style distribution $\boldsymbol{\pi}_a$. We also denote by $\boldsymbol{\pi}_a^{<t}$ the style distribution obtained from the artworks of creator a produced before year t. Our cultural evolution models are formulated as explicit realizations of the probability distribution $P(\tilde{\boldsymbol{\pi}}_a)$ for all creators a.

B. Influencer-guided transmission model

In the influencer-guided transmission model, each learner creator selects an influencer creator among previous creators and learns the influencer's style distribution. This model can be formulated as

$$P(\tilde{\boldsymbol{\pi}}_a) = \sum_{a':t_{a'}^{\rm s}<t_a^{\rm s}} P_{\rm sel}^{t_a^{\rm s}}(a') P_{\rm dis}(\tilde{\boldsymbol{\pi}}_a|\boldsymbol{\pi}_{a'\to a}),$$

where the summation is taken over all potential creators a' with $t_{a'}^{\rm s} < t_a^{\rm s}$, $P_{\rm sel}^{t_a^{\rm s}}(a')$ is the selection probability of the influencer a' of creator a, $\boldsymbol{\pi}_{a'\to a}$ represents the style distribution transmitted from a' to a, and the last factor is given as a discrete probability distribution

$$P_{\rm dis}(\tilde{\boldsymbol{\pi}}_a|\boldsymbol{\pi}_{a'\to a}) = \prod_k [\pi_{(a'\to a)k}]^{\tilde{\pi}_{ak}}. \tag{1}$$

We incorporate two biases in the selection probability to represent potential tendencies of learner creators, which are similar to those used for modeling citation dynamics of scientific papers [19]. The first is the recency bias, which represents the creators' tendency to more likely choose a reference influencer that lived more recently. This bias can be represented by a weighting factor $\mathrm{e}^{-(t_a^{\rm s}-t_{a'}^{\rm s})/\tau_{\rm c}}$, where the time constant $\tau_{\rm c}$ represents the time scale for the bias. The second bias is the (intrinsic) fitness of an influencer, which represents a time-independent likelihood of being learned by successors due to possible factors such as prestige and visibility. To represent this bias, we introduce a fitness parameter $v_{a'}$ for each creator. We also require that a potential influencer a' must have produced at least $N_{\rm bound}$ artworks (specifically, we set $N_{\rm bound} = 7$). The selection probability can then be given as

$$P_{\rm sel}^{t_a^{\rm s}}(a') \propto \mathbb{I}(t_{a'}^{\rm s} < t_a^{\rm s})\, \mathbb{I}(N_{a'}^{<t_a^{\rm s}} \geq N_{\rm bound})\, \mathrm{e}^{-(t_a^{\rm s}-t_{a'}^{\rm s})/\tau_{\rm c}+v_{a'}}, \tag{2}$$

where $\mathbb{I}(\mathcal{C}) = 1$ if condition $\mathcal{C}$ is true and otherwise $\mathbb{I}(\mathcal{C}) = 0$, and $N_{a'}^{<t}$ denotes the number of artworks produced by creator a' before year t.

The transmitted style distribution represents the influencer's style distribution with modifications including mutations. Here, we include the influence of creators other than the influencer since it is certainly possible that a learner creator observes many others' artworks and is influenced by the average style distribution of the time. This situation can be formulated as

$$\boldsymbol{\pi}_{a'\to a} = (1-\eta-\epsilon)\boldsymbol{\pi}_{a'}^{<t_a^{\rm s}} + \eta\, \boldsymbol{\pi}^{<t_a^{\rm s}} + \epsilon\, \mathbf{1}/K,$$

where $\boldsymbol{\pi}^{<t_a^{\rm s}}$ represents the average style distribution at (just before) time $t_a^{\rm s}$, $\mathbf{1}/K = (1/K, \ldots, 1/K)$ denotes a uniform distribution, and η and ϵ are nonnegative constants. The first, second, and third terms on the right-hand side represent the transmission from the influencer, transmission from the average style distribution of the time, and random mutation, respectively. The average style distribution $\boldsymbol{\pi}^{<t}$ also incorporates the recency effect and is defined as

$$\pi_k^{<t} \propto \sum_{n:t_n<t} \mathbb{I}(t_n < t)\mathbb{I}(k_n = k)\, \mathrm{e}^{-(t-t_n)/\tau_{\rm p}}, \tag{3}$$

where the summation is taken over all artworks n produced before t, k_n denotes the cluster assigned to n, and $\tau_{\rm p}$ denotes the time constant for artworks. The parameters η and ϵ represent the strength of the second and third terms and are called the oblique transmission coefficient and the random mutation rate, respectively. It is noted that the time-decaying factors in Eqs. (2) and (3) reduce the influence of past creators and artworks as time proceeds. Additionally, the restrictions on time in these equations ensure the causal constraint that future information is not used in the definition of probability $P(\tilde{\boldsymbol{\pi}}_a)$.

C. Community-guided transmission model

In the community-guided transmission model, we consider that a learner creator first analyzes previous artworks and finds communities of creators sharing similar style distributions. The learner creator then selects a community and learns the mean style distribution of the community. This model is described as

$$P(\tilde{\boldsymbol{\pi}}_a) = \sum_{c=1}^{C} \psi_c^{<t_a^{\rm s}} P_{\rm dis}(\tilde{\boldsymbol{\pi}}_a|\boldsymbol{\pi}_{c\to a}), \tag{4}$$

where c indexes communities of creators, C is the number of communities, $\psi_c^{<t}$ represents the selection probability of

communities at time t, and $\boldsymbol{\pi}_{c\to a}$ represents the style distribution transmitted from community c to creator a. The discrete probability distribution $P_{\rm dis}(\tilde{\boldsymbol{\pi}}_a|\boldsymbol{\pi}_{c\to a})$ is the same as in Eq. (1). The creator communities are formalized as a dynamic mixture model with the following log-likelihood:

$$\mathcal{L}(t) = \sum_{\substack{a:t_a^{\rm s}<t,\\ N_a^{<t}\geq N_{\rm bound}}} \mathrm{e}^{-(t-t_a^{\rm s})/\tau_{\rm c}} \ln\Bigg[\sum_c \psi_c^{<t} P_{\rm dis}(\tilde{\boldsymbol{\pi}}_a^{<t}|\boldsymbol{\pi}_c^{<t})\Bigg], \tag{5}$$

where $\tau_{\rm c}$ is the creator time constant, $\psi_c^{<t}$ is the mixture probability, and $\boldsymbol{\pi}_c^{<t}$ is the mean style distribution of community c. Assuming that how creators analyze the community structure is time-dependent and recent creators are given more attention in this process, the time-decaying factor $\mathrm{e}^{-(t-t_a^{\rm s})/\tau_{\rm c}}$ is introduced to weigh the creators' contributions on the likelihood. We use in Eq. (4) the parameter values of $\psi_c^{<t}$ and $\boldsymbol{\pi}_c^{<t}$ that maximize Eq. (5). See Sec. IV-D for the inference method.

The transmitted style distribution in Eq. (4) is formulated similarly as in the influencer-guided transmission model: we suppose that the mean style distribution of the selected community has the dominant influence, but the mean average style distribution of the time also has some influence and there is a small chance of random mutation. This situation can be formalized as

$$\boldsymbol{\pi}_{c\to a} = (1-\eta-\epsilon)\boldsymbol{\pi}_c^{<t_a^{\rm s}} + \eta\,\boldsymbol{\pi}^{<t_a^{\rm s}} + \epsilon\,\mathbf{1}/K,$$

where η represents the oblique transmission coefficient also in this model.

D. Inference method

The main purpose of the evolution models presented in the previous sections is to compare the two hypotheses on the cultural transmission process. Both models are formulated as realizations of the same probability distribution $P(\tilde{\boldsymbol{\pi}}_a)$, and we can compare them by their total log-likelihoods normalized by the total number of artworks

$$\mathcal{L}_{\rm tot} = \Bigg(\sum_a \ln P(\tilde{\boldsymbol{\pi}}_a)\Bigg)\Bigg/\Bigg(\sum_{a,k}\tilde{\pi}_{ak}\Bigg),$$

or the exponentiated quantities $\kappa = \exp(\mathcal{L}_{\rm tot})$ called the perplexity. The perplexity can be interpreted as the effective number of style clusters to choose. The smaller this quantity is, the more predictive a model is. The log-likelihood value $\mathcal{L}_{\rm tot}$ depends on the model parameters in both evolution models. Following the maximum likelihood principle, we estimate the model parameters to maximize $\mathcal{L}_{\rm tot}$, that is, we compare the two models in their optimal conditions. Since the values of the mutation rate ϵ and artwork time constant $\tau_{\rm p}$ did not affect the likelihood so much, as they are effectively smoothing constants, we fixed them as $\epsilon = 10^{-6}$ and $\tau_{\rm p} = 20$ yrs. In the following, due to the lack of space, we briefly explain specific methods for the inference of model parameters.

In the influencer-guided model, the model parameters to be estimated are $\{v_a\}$, $\tau_{\rm c}$, and η. We use an iterative method for optimizing these parameters; when one parameter (set) is optimized the other parameters are fixed. For creators' fitness v_a, we use the expectation-maximization (EM) algorithm with the influencer variable a' treated as a latent variable for each learner creator a, which can be derived in a manner similar to the Gaussian mixture model. For parameters $\tau_{\rm c}$ and η, we use the grid search method.

In the community-guided model, the model parameters to be estimated are $\{\psi_c^{<t}, \boldsymbol{\pi}_c^{<t}\}$, $\tau_{\rm c}$, and η. We again use an iterative method for optimizing these parameters, and the parameters $\tau_{\rm c}$ and η are optimized by grid search. We use the EM algorithm for estimating $\psi_c^{<t}$ and $\boldsymbol{\pi}_c^{<t}$ with the likelihood function in Eq. (5). To ensure the temporal consistency of the community estimation, we first perform community estimated on a coarse time scale of 25-years width and use this result to initialize the parameters $\psi_c^{<t}$ and $\boldsymbol{\pi}_c^{<t}$ for individual years.

TABLE II
PERPLEXITIES AND ESTIMATED PARAMETERS OF THE EVOLUTION MODELS. RESULT FOR THE MODEL USED FOR ANALYZING CREATOR FITNESS (SEE TEXT) IS ALSO SHOWN AT THE BOTTOM.

Model	Perplexity	$\tau_{\rm c}$ (yrs)	η
Random	19.1	NA	NA
Influencer-guided (optimal)	15.6	117	0.37
Community-guided ($C = 10$)	16.4	160	0.38
Community-guided ($C = 20$)	16.2	110	0.52
Community-guided ($C = 30$)	16.1	220	0.58
Influencer-guided (for analysis)	16.2	50	0.1

E. Results

Table II shows the model comparison result. As a reference, we also compared the random model where a creator generates artworks with cluster frequencies $\tilde{\boldsymbol{\pi}}_a$ according to the temporal average style distribution $\boldsymbol{\pi}^{<t_a^{\rm s}}$: $P(\tilde{\boldsymbol{\pi}}_a) = P_{\rm dis}(\tilde{\boldsymbol{\pi}}_a|\boldsymbol{\pi}^{<t_a^{\rm s}})$. The creator time constant $\tau_{\rm c}$ was estimated in the resolution of approx. 1 year for the influencer-guided model, and 10 years for the community-guided model, due to the large computational cost of the latter model. We tested $C = 10$, 20, and 30 for the community-guided model. Compared to the random model, the influencer-guided and community-guided models both had significantly lower perplexities, indicating a clear effect of cultural transmission. We see that the number of communities C does not have much impact on the perplexity, while large values of the oblique transmission coefficient were estimated for $C = 20$ and 30, suggesting that these models overfitted the community structure. By comparison, the influencer-guided model had a smaller perplexity, which shows that influencer-guided transmission better fits our data. On the other hand, the oblique transmission coefficient was also relatively large for this model, implying that the color style distribution is learned not only from one influencer but also from others. The estimated value $\tau_{\rm c} = 117$ yrs suggests that the influence of past creators sustains for a relatively long time.

An interesting possibility of the influencer-guided model is that it can estimate the fitness (influence) of creators in an unsupervised manner. To examine the model's potential,

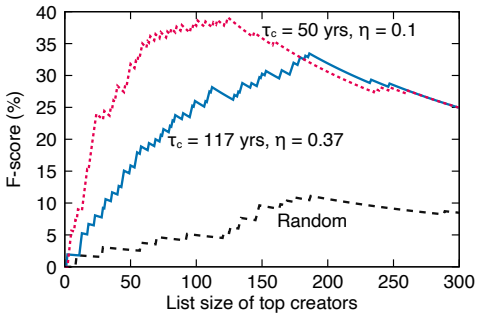


Fig. 6. Comparison of F-score measuring the accuracy of estimating influential creators in comparison with Gombrich's book [20].

we compared the list of creators with high estimated fitness with the list of creators appearing in Gombrich's classical textbook on art history [20]. Among the creators listed in [20], we removed a few who did not have major activities of painting and obtained a list of 101 creators. Using the model estimation, we made lists of varying sizes containing creators with the highest fitness, and we calculated the F-score to measure how much the results align with the Gombrich's list. We also conducted a similar calculation for a randomly sorted list of creators. The result in Fig. 6 shows that the F-score curve by the influencer-guided model with maximum likelihood parameters had significantly high values than that for the random list. While this result is already remarkable, we found that with alternative parameterization with a shorter time constant $\tau_c = 50$ yrs and a smaller oblique transmission constant $\eta = 0.1$, the F-score curve had even higher values (Fig. 6), even though the perplexity becomes higher (Table II). We discuss this mismatch between the optimized models for the likelihood measure and the accuracy of estimating "well-known" creators in the discussion section.

The 25 creators with the highest estimated fitness by the influencer-guided model with $\tau_c = 50$ and $\eta = 0.1$ are shown in Table III. We can see that the list includes creators active in various centuries. Importantly, the list contains creators with a relatively small number of artworks in the dataset, including such well-known creators as S. de Vlieger, R. Lichtenstein, and H. Hoffman. This result shows that the model predicts the fitness not merely from the number of artworks in the data, which although highly correlates with the creator's prestige. That this result was obtained purely from the image data and the meta data about the year of production for ensuring causal constraints demonstrates the potential of the present model for quantitatively and objectively analyzing painting art data and cultural evolution.

TABLE III
HIGHEST FITNESS CREATORS ESTIMATED BY THE INFLUENCER-GUIDED MODEL ($\tau_c = 50$ YRS, $\eta = 0.1$). NAMES IN BOLD FONTS INDICATE THOSE APPEARING IN GOMBRICH'S BOOK [20].

Fitness e^{v_a}	Creator a	Year t_a^s	No. of works
1.81	T. van Rysselberghe	1880	125
1.77	I. Aivazovsky	1835	238
1.75	**S. de Vlieger**	1629	29
1.70	**Titian**	1503	200
1.63	**C. Lorrain**	1630	62
1.62	C. Hernandez	2001	17
1.60	D. Teniers II	1633	54
1.58	**G. Bellini**	1455	56
1.57	M. Hafftka	1986	8
1.54	**P.-A. Renoir**	1858	913
1.53	H. Hofmann	1902	16
1.48	F. Boucher	1734	51
1.45	K. Malevich	1904	113
1.44	**P. Picasso**	1893	499
1.44	**D. Velazquez**	1616	127
1.43	**T. Gainsborough**	1745	125
1.43	**J. van Eyck**	1426	63
1.42	R. Lichtenstein	1961	26
1.42	**C. Monet**	1861	191
1.41	**H. Matisse**	1890	188
1.40	**V. van Gogh**	1881	747
1.38	**C. Corot**	1825	409
1.38	**P. Gauguin**	1873	460
1.38	**A. Altdorfer**	1507	25
1.37	R. Bujnowski	2001	9

V. CONCLUSIONS AND DISCUSSION

We have analyzed the evolution of color styles in painting arts using statistical machine learning and evolutionary modeling methods. By analyzing the dynamics of color style clusters, we have found synchronous changes in the cluster frequencies. We also found that creators' color style distributions are biased in a coherent way and this property is a major factor of this macroscopic phenomenon. We then constructed cultural evolution models incorporating transmission and selection processes of creators' color style distributions and found that the influencer-guided transmission model better fits the data than the community-guided transmission model. We also showed that the model can estimate the creators' fitness by unsupervised learning and automatically discover important figures in the history of painting art only from the image data. All together, these results indicate that in the cultural transmission of knowledge about painting styles, not only the average color combinations but also the distribution of different sets of color combinations is transmitted from creators to creators, and this microscopic process of individual creators can lead to the macroscopic patterns in the evolution of painting style.

It should be emphasized that the present models represent a simplified description of the real evolutionary process. For example, the model can be extended to include more than one cultural parent, or dynamic changes in the parameters such as the creator time constant and the oblique transmission coefficient. The mismatch of the preferred parameterizations by the maximum likelihood estimation and the optimal accuracy of reproducing the human expert's account, found in Sec. IV-E, may be an indication of such incompleteness of the current model. It may also be possible that the present model, which "analyzes" a much larger amount of data than a human can, can find objectively correct results that humans may not find easily. Therefore, future work should study more elaborated models and examine how these models will (or will not) alter the results, and additionally, conduct a benchmark test, for

example, using synthetic data, to quantitatively measure the ability of the present model. It is also important to examine the influence of the number and values of representative colors, which were somewhat arbitrarily chosen in this study.

Our result has prospective applications for enhancing the generative AI technology. Using the evolution model with influencer-guided transmission, we can estimate the transmission paths of knowledge about creation style. In scientometrics, for example, such knowledge reference data (i.e. citation data) are used as fruitful sources for studying the dynamics of creative activities including the quantification of the scientific contribution [21], the formation of communities [22], and the nature of creativity [23], [24]. Since complete knowledge transmission paths are rarely provided in self-reports or historians' accounts especially in art fields [25], a data-driven method would provide useful complemental data for revealing the mechanisms by which new creation styles are developed. The found mechanisms can be integrated with machine learning techniques to realize "creative" generative AIs, for example, using the framework of evolutionary computation.

Finally, the present framework of applying a model integrating a probabilistic generative process and cultural evolution process can be applied to a wide domain of data. While we focused on the color style in this study, there are other important stylistic features such as spatial composition, brushstroke pattern, and painted objects. We can apply DNNs for analyzing these various features [8], [9] and with a modification of the data generation process, we can conduct a similar analysis on more general features. Other possible cultural domains of application include music [26]–[28] and literary art [29], and culinary art [30], where cultural transmission of complex knowledge is considered equally important to understand the creative aspect of human intelligence.

ACKNOWLEDGMENT

We thank Takayuki Itoh, Daichi Mochihashi, Hitomi Kaneko, and Gumpu Nakamura for useful discussions. This study was supported by JSPS KAKENHI Grant Numbers 21K12187, 21K02846, and 22H03661, and JST FOREST Program Grant Number JPMJPR226X.

REFERENCES

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[2] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[3] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 14 745–14 758, 2021.

[4] A. Elgammal *et al.*, "Can: Creative adversarial networks: Generating "art" by learning about styles and deviating from style norms," in *Proc. International Conference on Computational Creativity (ICCC)*, 2017, pp. 96–103.

[5] L. L. Cavalli-Sforza and M. W. Feldman, *Cultural Transmission and Evolution*. Princeton University Press, 1981.

[6] M. Perc, "Beauty in artistic expressions through the eyes of networks and physics," *Journal of Royal Society Interface*, vol. 17, p. 20190686, 2018.

[7] H. Y. D. Sigaki, M. Perc, and H. V. Ribeiro, "History of art paintings through the lens of entropy and complexity," *Proceedings of the National Academy of Sciences of the USA*, vol. 115, no. 37, pp. E8585–E8594, 2018.

[8] A. Elgammal *et al.*, "The shape of art history in the eyes of the machine," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 2183–2191.

[9] E. Cetinica, T. Lipica, and S. Grgic, "Learning the principles of art history with convolutional neural networks," *Pattern Recognition Letters*, vol. 129, pp. 56–62, 2020.

[10] H. Wölfflin, *Principles of Art History: The Problem of the Development of Style in Later Art*. Dover Publications Inc., 1950.

[11] B. Lee *et al.*, "Heterogeneity in chromatic distance in images and characterization of massive painting data set," *PLOS One*, vol. 13(9), no. e0204430, 2018.

[12] P. H. R. Zivic, F. Shifres, and G. A. Cecchi, "Perceptual basis of evolving Western musical styles," *Proceedings of the National Academy of Sciences of the USA*, vol. 110, no. 24, pp. 10 034–10 038, 2013.

[13] M. Mauch *et al.*, "The evolution of popular music: USA 1960–2010," *Royal Society Open Science*, vol. 2, no. 150081, 2015.

[14] C. Weiß *et al.*, "Investigating style evolution of Western classical music: A computational approach," *Musicae Scientiae*, vol. 23, no. 4, pp. 486–507, 2019.

[15] E. Nakamura and K. Kaneko, "Statistical evolutionary laws in music styles," *Scientific Reports*, vol. 9, no. 15993, 2019.

[16] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[17] R. Singh and E. Nakamura, "Dynamic cluster structure and predictive modelling of music creation style distributions," *Royal Society Open Science*, vol. 9, no. 220516, 2022.

[18] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory*. Harper and Row Publishers, New York, NY, USA, 1970.

[19] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[20] E. H. Gombrich, *The Story of Art (16th ed.)*. Phaidon Press, 1995.

[21] O. Penner *et al.*, "On the predictability of future impact in science," *Scientific Reports*, vol. 3, no. 1, pp. 1–8, 2013.

[22] R. Klavans and K. W. Boyack, "Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?" *Journal of the Association for Information Science and Technology*, vol. 68, no. 4, pp. 984–998, 2017.

[23] V. Larivière, S. Haustein, and K. Börner, "Long-distance interdisciplinarity leads to higher scientific impact," *PLOS One*, vol. 10, no. 3, p. e0122565, 2015.

[24] L. Wu, D. Wang, and J. A. Evans, "Large teams develop and small teams disrupt science and technology," *Nature*, vol. 566, no. 7744, pp. 378–382, 2019.

[25] D. Park, J. Nam, and J. Park, "Novelty and influence of creative works, and quantifying patterns of advances based on probabilistic references networks," *EPJ Data Science*, vol. 9, no. 2, pp. 1–15, 2020.

[26] D. Park *et al.*, "Topology and evolution of the network of Western classical music composers," *EPJ Data Science*, vol. 4, no. 2, 2015.

[27] M. Interiano *et al.*, "Musical trends and predictability of success in contemporary songs in and out of the top charts," *Royal Society Open Science*, vol. 5, p. 171274, 2018.

[28] E. Nakamura, "Computational analysis of selection and mutation probabilities in the evolution of chord progressions," in *Proc. International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2023, to appear.

[29] C. O. Brand, A. Acerbi, and A. Mesoudi, "Cultural evolution of emotional expression in 50 years of song lyrics," *Evolutionary Human Sciences*, vol. 1, no. e11, 2019.

[30] O. Kinouchi *et al.*, "The non-equilibrium nature of culinary evolution," *New Journal of Physics*, vol. 10, no. 073020, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Ultimatelink Between Characters Having a Certain Meaning in Physical Space to URL in Cyberspace with Robust Print and Scan

Keiji Yamadera* and Michiharu Niimi†
* Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: yamadera.keiji334@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—We propose a method that links character string having a certain semantic meaning on papers in real world to the digital information in cyberspace, and call it Ultimatelink. Ultimatelink adds additional information to characters by superimposing a circular color mark on the characters without changing the shape of the characters. A URL (Universal Resource Locator) can be generated by connecting additional information of semantically organized characters which are come from captured images by digital devices. In the experiment, we investigated the behavior of Ultimatelink for the five type of Japanese characters. The target characters are multiple katakana characters, multiple kanji characters only, multiple kanji and hiragana characters, kanji with a few strokes, and kanji with many strokes. As a result of experiments, it was confirmed that information superimposition and information extraction are possible in most cases, but there are problems such as failure to extract a single character or the ability to extract color markers appropriately depending on the character shape.**

I. Introduction

In today's modern society, many people own smartphones, and the use of their cameras is not limited to photographing people and backgrounds, but has become one of the ways to connect to the Internet, for example by reading QR codes[1]. However, QR codes must be printed on paper separately from textual information, which would disrupt the layout of printed materials. In addition to QR codes, there is a technology called screen codes[2], which read information in the computer world from information on paper. Screen code is a method that prints and reads invisible microscopic patterns on paper, but because it uses special color printing technology, it is not intended to be used for printing information specified by general users using common computer devices. Therefore, it would be very convenient for us if we could easily link "textual information on paper" with "digital contents" in the computer space using common printing devices and paper without disrupting the layout of the printed matter. For example, we assume that each rendered character of "Mt.Fuji" has additional information and is printed on papers, then the printed characters are captured by digital camera. From the captured image, a URL reconstructed by the additional information extracted from each rendered characters may be assumed to a web site of explaining Mt.Fuji.

There are several methods that hide an information within characters in watermarking or information hiding research fields. For example, additional information is represented by transforming very small parts of characters to another shape[3][4] or removing a small part of characters[5][6]. However, embedded information is quite very low in those methods.

In this paper, we propose Ultimatelink to embed URL for multiple characters having a certain semantic meaning printed on papers. This method is realized by putting small color circles on the body of characters. The position of small color circles is decided by skeletonization of rendered characters and computing pseudo-stroke order which is a scanning order of skeletonized pixels.

In section 2, we describe how Ultimatelink embeds additional information to characters and extract it. In section 3, we show some experimental results, and discuss the results in section4. Finally, in section 5, we summarize the results and discuss future issues.

II. Ultimatelink

Ultimatelink is a method that links a certain semantic cohesion in the real world to accompanying/supplementary information in cyberspace. Specifically, it is a technology for superimposing and extracting binary data on characters. The process can be divided into two main categories: embedding information in a binary image and extracting information from a printed image that has been embedded and photographed by a camera such as a smartphone. Each of these processes is described below.

A. Embedding Process

The embedding process follows the flow below.

(1) Character recognition
(2) Character segmentation
(3) Prepossessing thinning (erosion)

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

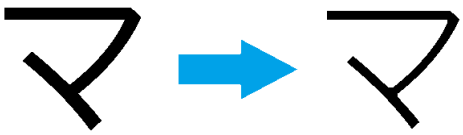


Fig. 1: Example of erosion



Fig. 2: Example of thinned image

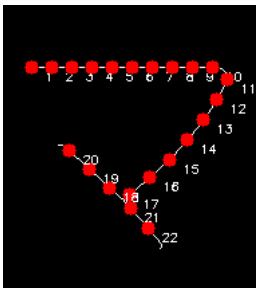


Fig. 3: Example of grid points

(4) Character image thinning
(5) Pseudo-stroke order calculation
(6) Grid points assignment
(7) Colored circles draw
(8) Final processing

Each procedure is described below in detail.

(1) Character Recognition
Images with multiple characters rendered in which you want to embed information (referred to as "original image") are prepared. We assume that the original images consist of two colors: black and white pixels which are corresponding to foreground pixels and background pixels respectively, in other words, this is equivalent to that we write characters by pencils on white color papers. The characters in the original image are recognized using an optical character recognition engine.

(2) Character Segmentation
The character string to be embedded is segmented from the original image into one character. Assuming horizontal writing, this segmentation is performed by obtaining the width and height of each character in the image. The width is obtained by extracting the consecutive rows of black pixels in the image. In this case, there are some characters, such as "リ (ri)" and "ハ (ha),"a that have an empty space within a single character. This means that these characters often have two small widths of consecutive black pixels within a character. The two small widths are able to be categorized as outliers by thresholding. Therefore, the width of those characters can be obtained by considering adjacent characters with smaller widths to be one. Similarly, the height is obtained by extracting consecutive lines of black pixels in the image. The character string can be segmented by cropping the image with a margin between the width and height obtained by this method.

(3) Prepossessing thinning (erosion)
Before performing the thinning of each segmented character image, the erosion is performed for the characters whose example is shown in Fig.1.

(4) Character image thinning
After erosion of each character, we perform thinning to it. A binary image with a line width of 1 is obtained from this process. This information is used for the position of drawing circles. Fig.2 shows an example of a thinned image.

(5) Pseudo-stroke order calculation
The pseudo-stroke order is calculated by tracing the order of foreground pixels from an image of a thinned character image. This order is based on the contour tracking method and is calculated by the following procedure.

Step1 : The first pixel found in the foreground by raster scanning is the pixel of interest.
Step2 : Find the pixel adjacent to the pixel of interest in a counterclockwise direction. The found pixel is the next pixel of interest.
Step3 : Repeat Step 2, and return to Step 1 when there are no more adjacent pixels.
Step4 : Repeat Steps 1 through 3 until all foreground pixels are found.

The pseudo-stroke order is calculated by the above procedure. The obtained pseudo-stroke order is saved for each character.

(6) Grid points assignment
A circle of radius r is drawn in pseudo-stroke order with d as the distance between circles. We call the position of circles "grid point." In some case, the circles may overlap depending on the character shape and scanning order when we draw on grid points. In such cases, the pseudo-stroke order calculated above is not used. Then the new pseudo-stroke order is searched so the the circles do not overlap. Fig.3 shows an example of drawing grid points.

(7) Colored circles draw
For the obtained grid points, circles colored according to the embedding information are drawn. The position at which the circle is drawn is randomly selected from among the grid points with a fixed random seed according to the length of the embedding information. One circle should represent two bits (four patterns), and five colors should be set in total, since colors are also needed for projection markers. In HSV color space, we fixed the value for two



Fig. 4: Colors for grid points/projection marker

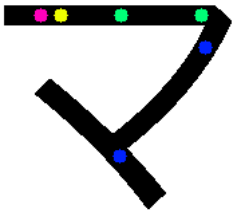


Fig. 5: Example of information embedding

axis: S = 255 and V = 255, and assign five values for one axis : H. As a result of this, five colors are generated in HSV space shown in Fig.4. The colors in Fig.4 are, from top to bottom, projection markers, embedded two bit binary information:"00," "01," "10," and "11."

(8) Final Processing
As colored circles are superimposed on rendered character images, we replace original characters with it. In addition, four projection markers are drawn to normalize the image during extraction. Fig 5 shows an example of a character image with embedded information. The character with embedded information is shown in Fig.5 and multiple colored circles are drawn on the character according to the embedded information.

B. Extraction Process

The extraction process follows the flow below.

(1) Projection markers detection
(2) Character recognition
(3) String extraction
(4) Circles detection by colors
(5) Circles detection by grid points
(6) Bit information extraction

Each procedure is described in detail below.

(1) Projection markers detection
A projective transformation is performed to transform the captured image to the same position and size as when embedding. To detect the projection markers as the basis for the transformation, mask images are created with the colors for the projection markers assigned in embedding process. We apply dilation processing to the mask images. Then circles are detected from the dilation mask image by Hough transform.

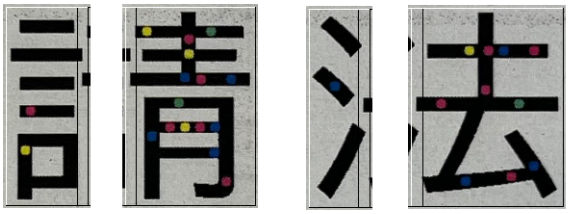


Fig. 6: Example of character segmentation failure

(2) Character Recognition
Character recognition is performed in the same way as in the embedding process.

(3) String Segmentation
We perform character string segmentation on the image after the projective transformation. The splitting of multiple characters procedure is basically the same as in the embedding process, but, in some cases, a failure to split one character into two characters by mistake may occur, as shown in Fig.6. When such a failure to split a string occurs, the string is split again according to the procedures shown below.
First, we suppose that n characters are recognized as a result of character recognition. At this time, if the length of the width of the entire character string is X, we can get the width by (X/n).

(4) Circles detection by colors
We detect colored circles from segmented images. The detection method is the same as the detection of the projection markers for each predefined color in HSV color space.

(5) Circles detection by grid points
We find grid points from the stored pseudo-stroke order information for each character. Using the same random seed determined in the embedding process, we can identify the center positions of colored circles allocated in the embedding process. Therefore, we estimate the colored circles allocated in the embedding process by finding circles with the closest distance to the identified center position from detected circles by colors in (4).

(6) Bit information extraction
The bit sequence is extracted from the estimated circles in (5) by arranging the bit sequence in specific order.

III. Experiments

In order to evaluate Ultimatelink, we conducted five experiments, which are called type A, B, C, D and E, with varying type of Japanese characters. We prepared two types of embedded information which is called type E_1 and type E_2 respectively. In the type E_1 and type E_1, the shortened URL "bit.ly/3oXpEIm" and the simple string "abcdefghijklmn" were used for the embedded information and each string consists of 14 characters which became 98 bits in ASCII code form.

TABLE I: Type of experiments

Type	Type of characters to be processed (Japanese)	Actual string	Embedded information
A	Katakana	マイナンバーカード	E_1
B	Kanji	申請方法	E_2
C	Kanji and Hiragana	お問合わせ	E_2
D	Kanji with a few strokes	一上一下	E_2
E	Kanji with many strokes	諸艶大艦	E_2



Fig. 7: Rendered target image with Gothic font (Type B)

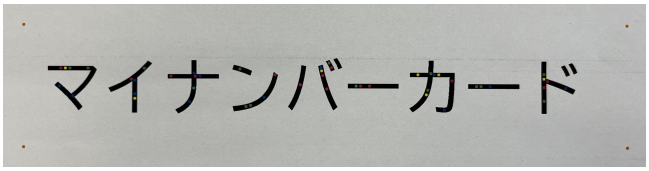
Experimental conditions are summarized in TABLE I.

We assume that A4 paper size is 2339×1654 pixels in a digital binary image, a character is rendered within 200×200 and those characters are placed around center of A4 paper. Actual string is rendered by using Gothic font, and an example of this is shown in Fig.7.

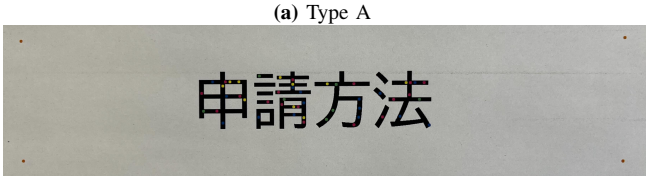
For these rendered characters, the embedding procedure of Ultimatelink is performed. Then, the characters with additional information embedded are printed out on A4 size of paper. In the extraction process, characters on papers are captured. Fig.8 shows clipped images which were captured 4032×3024 in size by digital camera.

For all type of experiments, we confirmed that the embedded information was extracted correctly. Although, in type E of kanji characters with many strokes, the reproducibility was not guaranteed. In order to be successfully achieved, we needed to change parameters adequately in erosion step during the embedding process and in character recognition during the extraction process. These parameter adjustments were made manually based on try and error.

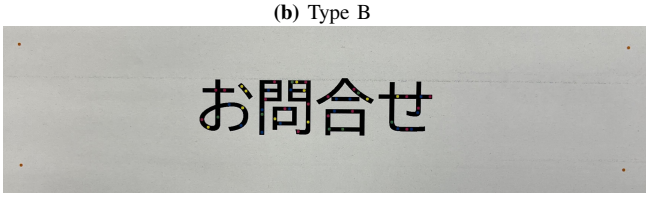
When splitting character strings in information extraction by using the method of splitting character strings in information embedding, the second character from the left and the fourth character are split incorrectly as shown in Fig.6, but the character string is split again using the number of characters during character recognition and the width of the entire character string as described in (3) of II-B in II. As a result, we confirmed that the string was correctly divided as shown in Fig 9.



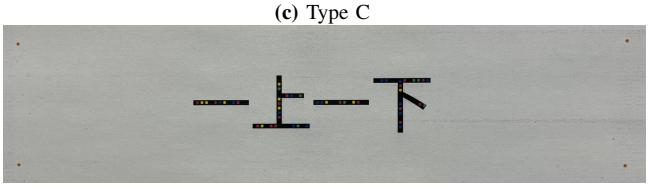
(a) Type A



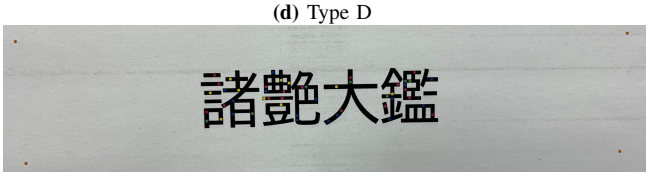
(b) Type B



(c) Type C



(d) Type D



(e) Type E

Fig. 8: Clipped captured image

IV. Discussion

We confirmed that Ultimatelink was able to accurately extract information for katakana only, kanji and hiragana, and kanji only with a standard number of strokes or a small number of strokes. This is the reason why the fact that the pseudo-stroke order is preserved for each character by performing character recognition and dividing the character string into individual characters. If a pseudo-stroke order is obtained from the captured image, the thinned image of each character may change depending on the position of the drawn circle, resulting in differences in the bit sequence extracted instead of being the same as when embedding. However, by using the same pseudo-stroke order for embedding and extraction, the reference grid points can be obtained in the correct order and the information can be extracted.

In addition, the character is eroded before the character is thinned when the information is embedded. This is to prevent the circles from being too close to each other and causing failure in circle detection. The comparison of the result of drawing a circle on an image that has not undergone erosion before the thinning process and the result of drawing a circle after the erosion process is shown in Fig.10.

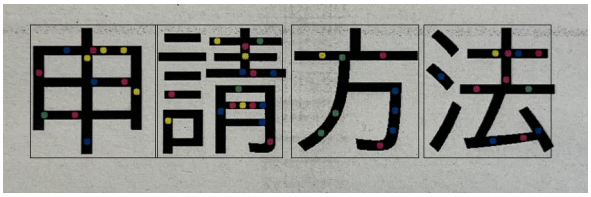
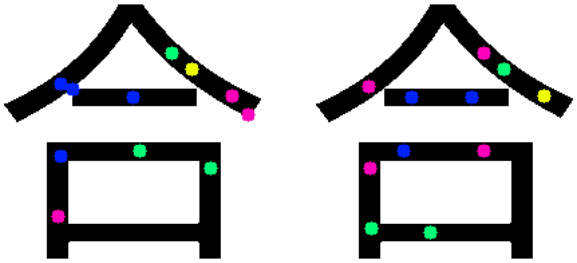


Fig. 9: Example of split success



(a) processing without erosion **(b)** eroded

Fig. 10: Comparison of circle drawing results

Fig.10(a) shows that the blue circles are too close to each other. On the other hand, Fig.10(b) shows that the distance between the circles is appropriately maintained. It is thought that this process ensures that circles are not undetected and that the information is extracted correctly.

In the case of multiple kanji characters with a large number of strokes, the results correctly extracted the information, but the reproducibility to other character images was not ensured because various parameters had to be adjusted. In the future, it will be necessary to search for parameters that can guarantee reproducibility in the case of multiple kanji characters with many strokes.

V. Conclusion

In this paper, we propose Ultimatelink for multiple characters with robust for print and scan. While the system was able to extract information without problems for katakana-only, kanji and hiragana, and kanji-only character images with a standard number of strokes and a small number of strokes, it was unable to reproduce kanji character images with a large number of strokes. Therefore, it is necessary to construct the system that is robust even when it contains kanji characters with a large number of strokes.

In addition, since this experiment was conducted on text printed on paper with a white background, it is necessary to conduct the experiment on posters and other displays with a colored background. Furthermore, since the lighting conditions in this experiment were a single lighting condition, under fluorescent light only, it is necessary to conduct the experiment under multiple lighting conditions, such as under sunlight or in a dark room.

In the method used in this experiment, images taken with a smartphone camera were imported once to a PC and processed. In consideration of future practicality, we should construct a system that can display Web pages and complete the information extraction process within a smartphone, and evaluate its practicality.

Acknowledgment

This research was supported by JSPS Grant-in-Aid for Scientific Research JP 19K12879.

References

[1] J. I. Standards, *Two-dimensional code symbols-qr code -basic specifications*, JIS, X0510, 1999.

[2] *Screen code*, https://www.apollo-japan.ne.jp/technology/.

[3] S. Rane, A. Vetro, and A. L. Varna, "Data hiding in hardcopy text documents robust to print, scan and photocopy operations," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2009, pp. 1397–1400. DOI: 10.1109/ICASSP.2009.4959854. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICASSP.2009.4959854.

[4] S. Das, S. Rane, and A. Vetro, "Hiding information inside structured shapes," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 1782–1785. DOI: 10.1109/ICASSP.2010.5495425.

[5] K. Hayashi, Y. Ueda, M. Iwata, and H. Kimura, "A proposal for information hiding to hardcopy by using cutting processed characters (in japanese)," *IPSJ Journal*, vol. 44, no. 11, pp. 2817–2825, Nov. 2003, ISSN: 0387-5806. [Online]. Available: https://cir.nii.ac.jp/crid/1050564285889926528.

[6] A. Takehiko, "Information hiding for hardcopy by using cutting processed characters and morse code," *Proceedings of the 7th Asia Pacific Industrial Engineering and Management Systems Conference, Bangkok, Thailand*, pp. 1747–1753, 2006. [Online]. Available: https://cir.nii.ac.jp/crid/1010282256913643787.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Human Flow Measurement System Using Floor Estimation of Depth Images for Low-End IoT Devices

Takuya NAGATOSHI* and Michiharu NIIMI†
* Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka-shi, Fukuoka, 820-8502 Japan
E-mail: nagatoshi.takuya769@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology,
680-4, Kawazu, Iizuka-shi, Fukuoka, 820-8502 Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—We propose a system that uses depth information, which represents the distance from the sensor, instead of color information to do both measure human flow and protect privacy for low-end IoT devices. The system is designed to detect the position and number of persons from depth information. Since administrative organizations or educational institutions have been reducing their budgets, this system should be implemented at as low a cost as possible. In order to realize human flow detection system on low-end Iot devices, we use low performance depth cameras as data acquisition device controlled by single board computers, such as Raspberry Pi. As one of our goals is all data processing are performed on single board computers, we adopt computational methods for the detection as possible as simple. The proposed method is based on the background subtraction method, which prepares a reference depth image and extracts moving regions from one frame extracted from the video depth image. Furthermore, we aim to achieve high-precision people flow measurement by combining the following three elements: segmentation of moving objects using edge information, identification of human areas using floor information, and human tracking using areas where people overlap in the direction of the time axis. Experiments were also conducted and evaluated in a real space using a program that implements the presented method.**

***Index Terms*—privacy protection, depth image, human flow detection, V disparity map, low-end IoT devices**

I. Introduction

Due to the corona disaster of the past few years, we have become concerned about spatial density. If information about human flow could be automatically obtained, it would certainly help our live, not only for the problem of the Corona disaster, but also from the perspective of crowd detection during events.

All of the previous studies have attempted to measure human flow using video and image processing from existing surveillance cameras and other fixed cameras which is able to provide RGB images. Angelova et al.[1] combine a cascade classifier and a small DNN (Deep Neural Network) to accelerate processing and achieve real-time detection of the location and number of people for multiple pedestrians in suburban and urban areas taken from a vehicle. Yuhong et al.[2] for counting the number of people at the level of hundreds, called crowd measurement. CSRNet (Congested Scene Recognition Network) is a supervised deep learning model that uses density maps. Weizhe et al.[3] have used a density map that includes information on the flow of people from video images in consecutive frames and the correlation of people tracking by optical flow to more accurately estimate the number of people in a crowd of several hundred or more people and the direction of the flow of people.

However, such methods using RGB cameras, including existing surveillance cameras, simultaneously acquire information on the face and clothing that can identify a person, which raises privacy issues.

To that concern, it is one solution using depth images. In a related study, Thi Thi et al.proposed a monitoring system for elderly people using depth information[4]. In this system, one elderly person living in a room is monitored. Detection is not performed for multiple persons, which is necessary for human flow measurement.

This paper proposes a system that uses depth information to both measure human flow and protect privacy. In this system, we assume that our proposed system is utilized in administrative organizations or educational institutions. As you know, these organization have been reducing their budgets, therefore this system should be implemented at as low a cost as possible. In order to realize human flow detection system on low-end Iot devices, we use low performance depth cameras as data acquisition device controlled by single board computers, such as Raspberry Pi. As one of our goals is all data processing are performed on single board computers, we adopt computational methods for the detection as possible as simple. The proposed method is based on the background subtraction method, which prepares a reference depth image and extracts moving regions from one frame extracted from the video depth image. Furthermore, we aim to achieve highly accurate human flow measurement by combining three methods: segmentation of moving objects using edge information, identification of person areas using floor information, and person tracking using overlapping person areas in the time axis direction.Experiments were also conducted and evaluated in an

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

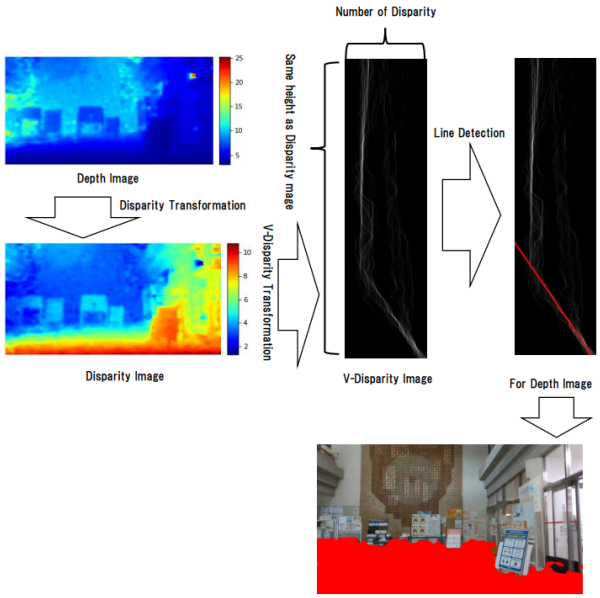


Fig. 1: Outline of floor detection from depth image

actual space using a program that implements the presented method.

II. PROPOSED SYSTEM

A. PERSON DETECTION SUMMARY

The proposed method is based on the background subtraction method of depth images. Moving objects will appear as the difference of the subtraction. To construct our method, we need two technical issues: depth images often contain noise, and a blob on binary image which is generated after the subtraction may represent overlapped persons. For the former issue, we adopt the thresholding of the size of blob and the judgment whether person is on the floor or not. For the latter issue, we use the image processing based on edge detector. The detailed method is described as follows.

1) BACKGROUND SEPARATION BY BACKGROUND DIFFERENCE METHOD: Firstly, for the depth information of the acquired F-frames, the background difference is calculated with the previously acquired background depth information, which is called a reference depth image, in order to separate the background and foreground information.

Since the depth difference image does not take into account errors in depth information by the sensor, there is a possibility that the difference between the reference depth image and frames may be generated even in areas where no person actually exists. Therefore, binarization processing is performed based on the parameter:T_1, and the area where the difference exists is designated as a tentative person area. From this region, the one with sufficient area which is selected by the parameter:T_2 is regarded as a candidate of person region. After this step, a binary image which foreground corresponds to a candidate of person region is generated.

2) EXTRACTION OF FLOOR AREA: The flow of extraction of the floor region is shown in Fig.1. It is known that flat regions such as the floor region appear as straight lines in the histogram of disparity values, called the V disparity map[5][6]. In Fig. 1, the disparity image is color-coded according to the magnitude of the value (the closer to blue, the smaller the value; the closer to red, the larger the value). Since the disparity value is the reciprocal of the depth value, the coloring is reversed compared to the original depth image.

The V disparity image is obtained by taking a histogram of the disparity values for each row. Therefore, the vertical axis equals the height of the disparity image and the horizontal axis equals the number of disparity values. Since the disparity values of flat areas such as floors and walls are constant or change step wise, they appear as straight lines in the V disparity image with a histogram taken for each row. In the V disparity image of Fig.1, the vertical lines correspond to the walls and the diagonal lines to the floor. This time, diagonal straight lines are detected by the Hough transform and mapped to the depth image to extract the floor area.

3) EDGE DETECTION FROM DEPTH INFORMATION: Within a candidate of person region, a region may contain two or more person if persons are overlapped. In order to separate those overlapped persons, we use raw depth information of a candidate of person region. There is a difference in depth between overlapping persons, which appears as a large gradient when observing the change in depth at a pixel. Therefore, we use edge detector for depth information. The contour information is detected from the depth information and subtracted from the candidate of person region This process can divide the overlapping person areas and enables us the detection of multiple persons. Because small region may not be person region, we remove those region using the area criterion T_3.

4) DETECTION BY INVERSE ORDER TRACKING CONSIDERING FLOOR AREA: If person detection by floor area is performed for each frame, it is not possible to detect persons who are not in the floor area even though they are persons. These persons may invade the floor area over time. Therefore, if the detected area is tracked and the same area exists in the floor area in a certain frame, the area is considered to be a person even if it does not exist in the floor area in the frames before and after the frame, and the frame is not deleted retroactively but remains as a person. Specifically, tracking is performed in reverse order from frame n to frame 0. If there is an object area in the floor area at frame n+1 or n-1, the same area at frame n is detected as a person even if it is not present in the floor area. The proposed method uses a measure of how much common area there is between regions called the IoU (Intersection over Union). If this IoU value is high, which means there is overlap between frames, the regions are considered to be the same moving object.

III. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT

In this experiment, a commercially available Realsense series D435i from Intel was used to acquire depth images.

TABLE I: Evaluation Result

Scene	condition	CC	MAE	PDA
1	Canny	**0.878**	0.235	0.806
1	Sobel	0.831	**0.195**	**0.930**
2	Canny	**0.817**	0.660	**0.900**
2	Sobel	0.659	**0.280**	0.885
3	Canny	0.758	**0.11**	0.852
3	Sobel	**0.818**	1.33	**0.930**

Person detection is performed from the 200 frames of depth images obtained by capturing at three locations, which is called Scene 1, Scene 2, and Scene 3, within the our campus.

The parameters used in this experiment are then described. The threshold T_1 for binarizing the background differential depth image was set to 600 mm. The threshold values T_2 and T_3 for the micro-area area were both set to 5500, and we used the values that were good while changing the parameters. The threshold value of the Houghline transform for floor area extraction was set to the value that best covers the actual floor area by experimenting with each scene. For edge detection, we used two detectors, one is the Canny detector and the other is Sobel detector.

B. EVALUATION METHODS

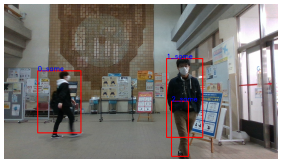
We check the RGB image Visually and determine the actual number of persons in the i-th frame as Pt_i. The number of persons output by the proposed method is denoted as Po_i. Based on this number, the correlation coefficient (CC) and mean absolute error (MAE) between Pt_i and Po_i are calculated for all frames. To evaluate whether the same person is detected without duplication, or whether the detected frame is a person, the number of detected areas that correctly capture a person (hereafter Pc_i) is also visually checked to evaluate these as person detection accuracy. Pc_i/Po_i is calculated for the frame where $Po_i \neq 0$, and the average value over all frames to be processed is the person detection accuracy (PDA). The average value over all the frames to be processed is used as the human detection accuracy.

C. EXPERIMENTAL RESULTS

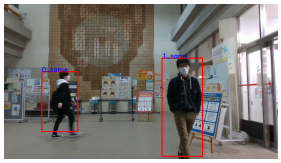
An example of the person detection results by the two edge detectors is shown in Fig. 2, respectively. To visually check the person detection results of the system, a bounding box is created by taking the bounding rectangle for each region. The same person tracked is assigned the same number.An example of how errors in large areas due to measurement errors can be suppressed is shown in Fig. 3. Fig. 4 shows an example where the same person is detected even if he/she is not in the floor area by tracking. In addition, the evaluation in each scene is shown in Table I.

D. Studies

The Canny method extracts edges in detail, resulting in more cases of duplicate detection of one person, as in Fig. 2(a), Fig. 2(c), and Fig. 2(e). On the other hand, the Sobel filter takes the edges roughly, which suppresses overlap and leads



(a) Scene 1 : Canny Detector



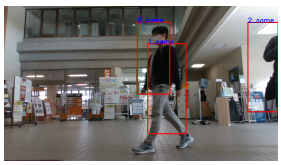
(b) Scene 1 : Sobel Detector



(c) Scene 2 : Canny Detector



(d) Scene 2 : Sobel Detector

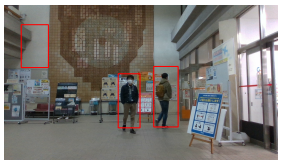


(e) Scene 3 : Canny Detector

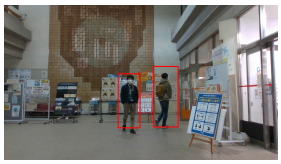


(f) Scene 3 : Sobel Detector

Fig. 2: Detection results in Canny Detector and Sobel Detector for each scene



(a) An example of large measurement error



(b) After taking account of floor detection

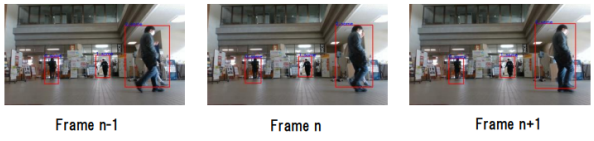
Fig. 3: Effects of floor detection to reduce measurement error

to highly accurate evaluation results. In Fig. 3(b), the miss detection region due to the large measurement error in the upper left corner of Fig. 3(a), which existed in the process up to the person segmentation by edge detection, can be removed because it is not in the floor area.

In addition, we confirmed that if only the area that exists



(a) floor detection performed in each frame



(b) floor detection performed over frame using tracking

Fig. 4: Effects of consecutive frame traking to reduce missing detection

in the floor area is considered as a person, a person who is not in the floor area may be mistakenly deleted as an error. By tracking between consecutive frames and retroactively detecting the person outside the floor area when the person enters the floor area over time, our method is able to detect the person without omissions as shown in Fig. 4(b).

The evaluation results show that the Canny method has a higher correlation coefficient in Scenes 1 and 2. This is because the number of persons is sometimes the same as the actual number of persons because of duplicate counting. For scene 1, this is confirmed by the fact that the processing with the Sobel filter has better person detection accuracy, which indicates whether the person is actually correctly present in the area being detected. The Canny method also shows better results in terms of person detection accuracy in scene 2. In Scene 2, it may be likey to occur because the background changes as the elevator door opens and closes. Canny method does not detect the elevator portion, so the person detection accuracy is better than the Sobel one. However, since the absolute mean error is better with the Sobel filter, it can be said that the number of detected frames that actually capture the person is higher with the Sobel filter.

IV. CONCLUSION.

In this paper, we conducted experiments to verify the effectiveness of a method for detecting persons in space from depth images using an algorithm that combines the background subtraction method, removal of background change regions by variance, edge detection using a Sobel filter, and determination of tracking objects using floor regions extracted from V disparity images. As a result, we were able to segment multiple overlapping persons if sufficient depth differences were generated.The problem of false detection of people due to shooting errors was improved by the presence or absence of tracking objects in the floor area.

Future issues include how to account for measurement errors in depth images and how to reject information that is detected as human motion even though it is not human motion.

REFERENCES

[1] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proceedings of the British Machine Vision Conference (BMVC)*, X. Xie, M. W. Jones, and G. K. L. Tam, Eds., BMVA Press, Sep. 2015, pp. 32.1–32.12, ISBN: 1-901725-53-7. DOI: 10.5244/C.29.32. [Online]. Available: https://dx.doi.org/10.5244/C.29.32.

[2] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1091–1100. DOI: 10.1109/CVPR.2018.00120.

[3] W. Liu, M. Salzmann, and P. Fua, "Counting People by Estimating People Flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8151–8166, 2022. DOI: 10.1109/TPAMI.2021.3102690.

[4] T. T. Zin, Y. Htet, Y. Akagi, *et al.*, "Real-Time Action Recognition System for Elderly People Using Stereo Depth Camera," *Sensors*, vol. 21, no. 17, 2021, ISSN: 1424-8220. DOI: 10.3390/s21175895. [Online]. Available: https://www.mdpi.com/1424-8220/21/17/5895.

[5] A. Iloie, I. Giosan, and S. Nedevschi, "UV disparity based obstacle detection and pedestrian classification in urban traffic scenarios," in *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2014, pp. 119–125. DOI: 10.1109/ICCP.2014.6936963.

[6] Z. Hu and K. Uchimura, "U-V-disparity: An efficient algorithm for stereovision based scene analysis," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, 2005, pp. 48–54. DOI: 10.1109/IVS.2005.1505076.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Holo-QoI: A Human Factor-Based Dataset and Prediction Framework for Assessing Quality of Interaction in Augmented Reality

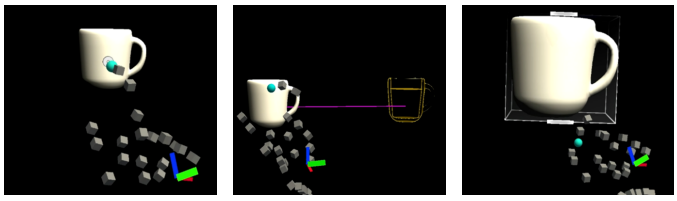
Seongjean Kim*, Seonghwa Choi* and Sanghoon Lee*†
* Department of Electrical and Electronic Engineering, Yonsei University, South Korea
† Department of Radiology, College of Medicine, Yonsei University, South Korea
E-mail: {jin.k, csh0772, slee}@yonsei.ac.kr Tel/Fax: +82-2-2123-7734

***Abstract*—With the revolution of Augmented Reality (AR) Head Mounted Display (HMD), AR technology enables the provision of immersive virtual content using various interaction methods. To bridge the real and virtual world, hand-based interactions have been utilized for their simplicity and effectiveness in interacting with AR applications. Thus, assessing the Quality of Interaction (QoI) is vital for enhancing the experience of AR applications, but it has been hindered by the absence of a comprehensive database that includes human factors. In this paper, we construct an interaction dataset, Holo-QoI, which includes 49 sets of human factor metadata with associated human subjective scores. To analyze the contribution of various factors to QoI, we process HMD metadata into human factor-based features (e.g. head, gaze, and hand motion). We then design a QoI prediction framework based on a Transformer architecture to model the relationship between human factor-based features and deep features, which is used to predict the QoI score of an interaction. Through rigorous experiments, we demonstrate that our proposed model achieves an 86% accuracy using the constructed dataset.**

I. Introduction

The development of Augmented Reality (AR) technology has been driven by the desire for more immersive communication with the digital world. Using AR Head Mounted Displays (AR-HMD), AR applications have been developed in various areas such as entertainment, education, healthcare, and engineering [1]–[3]. Existing applications utilize interaction [4] through gaze movement, voice commands, controllers, and hands to bridge the gap between the virtual and real world. Among these, hand-based interactions are particularly prominent as they are intuitive, efficient, and widely used. Hand-based interactions involve complex interactions with human factors such as eye gaze and head movements. For instance, in user-friendly interactions, users tend to focus their gaze on the target, resulting in minimal movement of the eyes and head. On the contrary, in less immersive interactions, the dispersion of eye gaze and head movements leads to a decrease in user engagement and results in lower-quality interactions.

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)(No.2020-0-00537, Development of 5G based low latency device – edge cloud interaction technology)



(a) Select (b) Move (c) Rotate

Fig. 1. Content scene example for the 3 interaction types, select, move, and rotate. The original video is recorded with each data paired with the real world. We separate the virtual parts from the full video capture, and paint the background of each data black for uniformity of further process

Therefore, high-quality hand-based interactions are necessary for an improved immersive experience in AR.

To analyze the relationship between human factors and hand-based interaction quality, there is a need for a quantitative measure to assess the interaction quality. We define this as the Quality of Interaction (QoI). To measure QoI, subjective assessment is commonly used because it can provide valuable insights into qualitative aspects such as usefulness, satisfaction, and naturalness. However, this method is time-consuming and requires a large number of participants to achieve statistically significant results [5]. It is also sensitive to various factors such as mood and prior experiences, leading to inconsistent and unreliable results. In this regard, it is crucial to have a method that can quantitatively measure the QoI and a comprehensive dataset that can provide rich data for analysis.

Recent studies have explored alternative approaches to suggest guidelines for improving the naturalness and intuitiveness of hand interactions. Kang et al. [6] present design suggestions for 3D interaction designers by analyzing the importance of the visual appearance of virtual objects for natural hand interactions. Du et al. [7] propose a new interaction method by combining Leap Motion with the HMD and assess its effectiveness through game experiments and subjective evaluation. The aforementioned methods aim to improve the quality of interaction through different approaches, but they do not provide a specific quantitative method for measuring interaction quality. Therefore, they have a limitation in terms of generalizability.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I
AR GESTURE INTERACTION CONTENT TABLE

Scene Index	Interaction type	Distance	Order	Mean Opinion Score
S101	Select	Near	Fixed	4.531
S102		Exact	Fixed	4.286
S103		Far	Fixed	4.020
S104		Near	Random	4.449
S105		Exact	Random	4.184
S106		Far	Random	3.980
S201	Move	Near	Fixed	3.816
S202		Exact	Fixed	3.571
S203		Far	Fixed	3.286
S204		Near	Random	3.796
S205		Exact	Random	3.510
S206		Far	Random	3.245
S301	Rotate	Near	Fixed	3.286
S302		Exact	Fixed	2.633
S303		Far	Fixed	1.020
S304		Near	Random	3.306
S305		Exact	Random	2.592
S306		Far	Random	1.898

To investigate the perceptual QoI in AR HMDs during object interaction tasks, our previous research [8] utilized an AR content dataset paired with subjective scores collected from a total of 32 subjects. These subjective scores served as the ground-truth labels for supervising the QoI prediction model. To quantify the level of QoI, we extracted various features from the recorded video data and metadata obtained from the HMD. Subsequently, we trained an SVR-based model using these features to predict the QoI score of a given content. However, such an approach had limitations in both constructed dataset and predictive performance. Given the significant importance of human factors in evaluating virtual interactions [9], [10], our previous dataset had limitations in terms of both the limited quantity of data and the lack of human factor information beyond hand position data. Due to the limitations, the proposed method lacked in modeling various human factors that influence interaction quality, resulting in unsatisfying prediction performance.

In this paper, to overcome the limitations of our previous research, we analyze the contribution of various factors to QoI. We construct a large-scale QoI dataset, Holo-QoI, that includes HMD metadata sets consisting of gaze, head, hand, and object, with associated human subjective scores. Using this dataset, we overcome the limitations of our previous method. First, we design a method for processing the metadata into human factor-based features to model the influence of each human factor on interaction quality. Then, we design a Transformer-based QoI prediction framework that uses human factor-based features along with deep features to predict the QoI of the content. Experimental results show that our framework achieves a prediction accuracy of 86%, significantly outperforming the results of our previous method. Such results indicate that the applied features are appropriate and that our framework provides a more efficient way to predict QoI in AR.

II. DATASET CONSTRUCTION

To analyze the contribution of various factors to QoI, we first construct a new large-scale interaction dataset, Holo-QoI, which consists of in-device metadata paired with subjective scores for Unity AR contents. Metadata refers to the time-continuous gaze, head, hand, and virtual object 3-D position coordinates and rotation angle data obtained using the internal API of the Microsoft HoloLens 2. It is important to note that there are no existing datasets available that contain this type of data, making our dataset particularly valuable for this research.

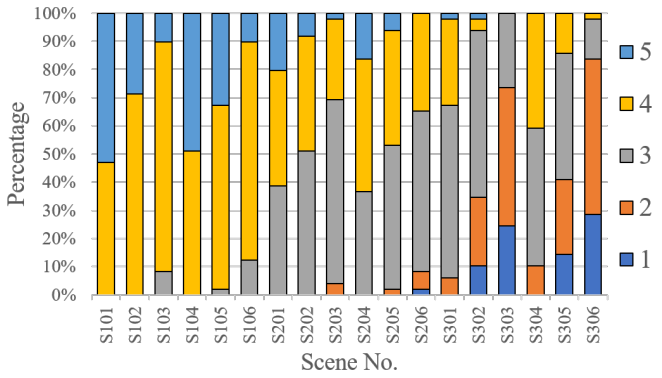


Fig. 2. Score distribution for each content scene. The bars of each column represents the score percentage of the 49 subjects.

A. Holo-QoI Dataset

To achieve our objective, we constructed 18 AR gesture interaction content scenes by varying several conditions, as summarized in Table I. Fig. 1 provides a visualization example of the representative AR gesture interaction content. Each scene is constructed and run using the Unity3D engine, with the background set to empty space to minimize visual disturbance during the assessment.

Interaction Type The interaction type condition specifies the interaction task that the subject must perform with the virtual object. The Hololens offers various built-in interaction methods, which enable users to select, move, and rotate virtual objects. In the select case, the subject is required to touch the virtual object with their bare hand. For both move and rotate cases, the subject is asked to grab the object and drag it to the target location or target pose.

Distance The distance condition specifies the distance between the subject and the virtual object in each content. Before evaluation, the arm length of each subject is measured, and the position of the virtual object is adjusted accordingly to either be closer, further or at an exact arm length.

Order The order condition determines the positioning order of the virtual objects that appear on the screen for interaction in each scene. A total of 5 objects continuously appear, and this condition decides whether the objects appear in a fixed or random order.

We aimed to obtain a dataset of HMD metadata paired with subjective scores to describe the interaction quality of content scenes, using a Microsoft Hololens 2 [11] system for the assessment. The system contains a head-mounted display with 2k 3:2 light engines, which generate holograms on see-through holographic lenses at a framerate of 30 FPS.

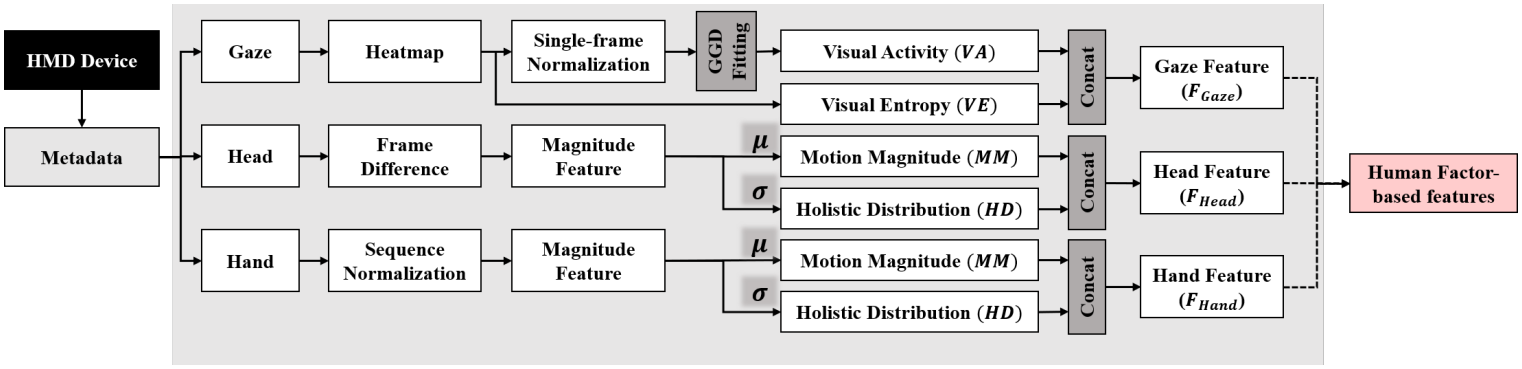


Fig. 3. Block diagram of the human factor-based feature extraction method

The metadata is obtained as a set of head, gaze, object, and hand data for each frame. The head data contains 7 dimensions, the gaze data contains 6 dimensions, the object data contains 7 dimensions, and the hand data contains 182 dimensions (7 dimensions per joint * 26 joints). The position and direction of each dimension are expressed as (x,y,z), while the rotation is expressed as (x, y, z, w). The head data captures the movement and rotation of the HMD in global coordinates, while the gaze data captures the starting point and direction of the eye gaze. Despite having two eyes, only one vector is returned because the device averages the position and direction of the eyes. The object data captures the position and rotation of the virtual object used for the interaction task. The hand data captures the movement and rotation of the dominant hand through the hand joint position and hand joint rotation.

B. Subjective Assessment

Following the subjective assessment protocol of standard IEEE P3333.1.3 [12], we design an evaluation process to measure the interaction quality of each constructed content. A total of 49 subjects participated in the evaluation, consisting of 40 males and 9 females with an age range from 24 to 35 years. All subjects were undergraduate and graduate students majoring in electronic engineering or software engineering, who were all right-handed and had little or no experience in AR interaction before.

The assessment protocol for this study was as follows: prior to beginning, the calibration settings of the HMD were adjusted, and the arm length of each subject was measured. Subjects were instructed to use their right hand to interact with each content scene. After each scene, subjects were asked to rate their interaction experience on a 5-point Likert scale, which was labeled as Very Comfortable (5), Comfortable (4), Mildly Comfortable (3), Uncomfortable (2), and Extremely Uncomfortable (1). Following each scene, subjects were given a short break to reduce visual fatigue and minimize the effects of AR sickness on subsequent scenes. This process was repeated for a total of 18 content scenes.

C. Data analysis

Table I presents the Mean Opinion Score (MOS) averaged for 49 participants for each content scene, while Fig. 2 shows the score distribution. The MOS was compared across different groupings, including interaction type (select, move, and rotate), distance (near, exact, and far), and order (fixed and random). The average MOS for each grouping was $4.269, 3.537, 2.622$ for interaction type, $3.870, 3.483, 3.074$ for distance, and $3.494, 3.458$ for order. The results indicate that the interaction type condition had the most significant impact on the score. From the results of [6], the difficulty of a hand-based interaction in AR is determined by factors such as the complexity of the movement, accomplishment time, and hand naturalness. Considering this, from the interaction types, rotate had the highest difficulty, followed by move, then select. Therefore, we derive that the QoI score is highly related to the interaction difficulty, and to improve user experience, one should consider the factors previously mentioned. The distance condition also had a minor contribution to the score, with closer objects being associated with higher scores due to their visibility from size.

III. Method

To find the relationship between the metadata and the paired scores, we process the metadata into human factor-based features. We design a QoI Predictor based on a Transformer architecture [13] to consider object and hand coordinate information along with the human factor-based features. At the output of the Transformer, this score token is then passed through an MLP (Multi-Layer Perceptron) head, resulting in the QoI prediction score of the input metadata.

A. Human Factor-based Features

Fig. 3 illustrates the human factor-based feature extraction process. In addition to hand data, which is the main interaction method, we also incorporate gaze and head data to reflect additional human features. We use appropriate feature extraction methods to extract features from these data and model the relationship between the metadata and QoI score.

Gaze For gaze (F_{Gaze}), we first process the gaze distribution of each scene per subject and convert it into a heatmap, as shown in Fig. 5. The heatmap is just an image generated from metadata, so we adopt visual activity (F_{va})[14] and visual entropy (F_{ve})[15] to extract representative features that

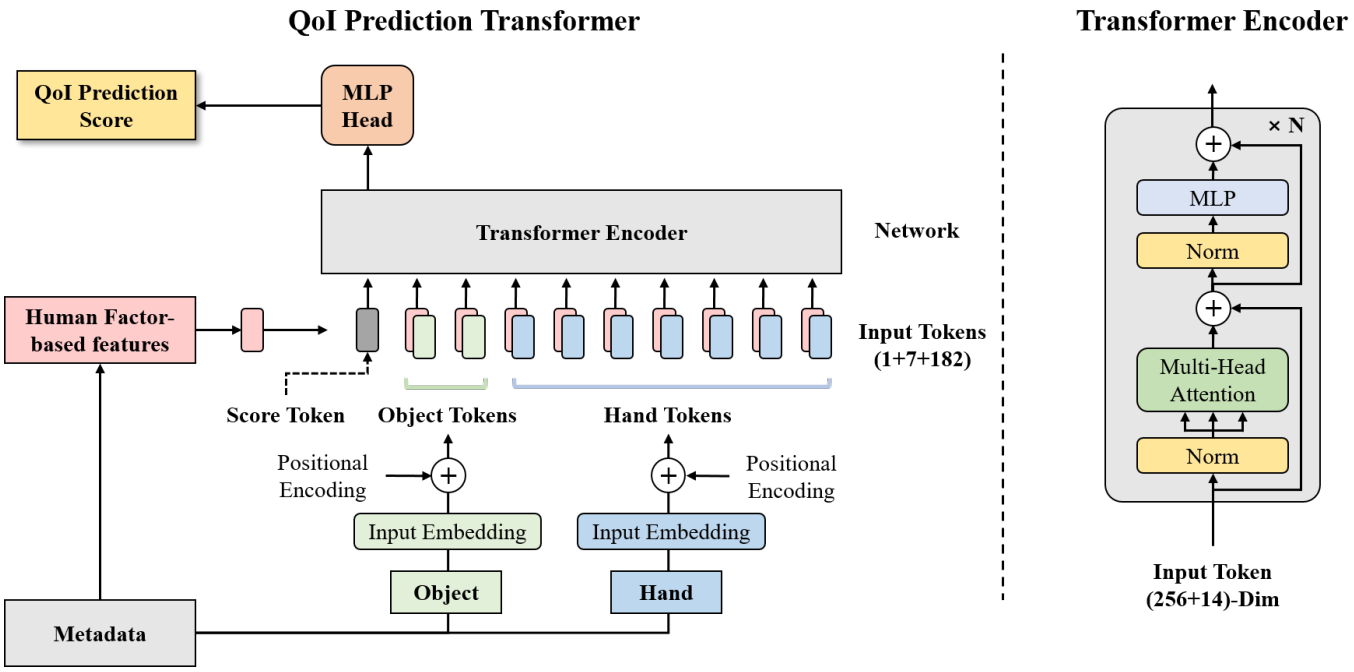


Fig. 4. Overview of the proposed QoI prediction framework

reflect human perception. These features are commonly used to evaluate the quality of virtual reality content [16]. We concatenate the two extracted features to create the gaze feature.

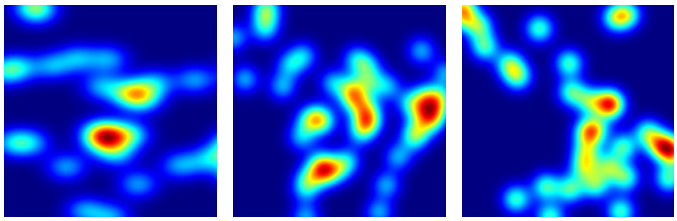
The visual activity is obtained by first normalizing the obtained heatmap, followed by a discrete wavelet transformation. This process quantifies gaze distribution and movement in the frequency domain. Finally, the histogram of the wavelet coefficients of each sub-band is fitted with a generalized Gaussian distribution (GGD) [17]. Let β_i be the i^{th} sub-band shape parameter of the wavelet transform histogram. Using the shape parameter, the F_{va} of the i^{th} sub-band of the heatmap is defined as:

$$F_{va} = \begin{cases} \sum_i (1 - \frac{1}{1+\exp(-2(\frac{\beta_i - C_t}{C_{tu}}))}), & \text{if } \beta_i \geq C_t \\ \sum_i (1 - \frac{1}{1+\exp(-2(\frac{\beta_i - C_t}{C_{td}}))}), & \text{otherwise} \end{cases} \quad (1)$$

where $C_t, C_{tu} C_{td}$ are normalization factors on the feature map. From experimental results of [18], normalization factors are set as $C_t = 1.1322, C_{tu} = 0.1338, C_{td} = 0.2904$.

The visual entropy score quantifies the degree of visual spread in the heatmap, where interactions with a high Mean Opinion Score (MOS) tend to show a low visual entropy score. Let (x, y) be the pixel coordinate of the heatmap H_{map}, with the center of the heatmap fixed at the coordinate $(0, 0)$. The formula for calculating the visual entropy score F_{ve} is the negative sum of the distance from each pixel (x, y) to the center, multiplied by its log value. The equation is defined as:

$$F_{ve} = -\sum_{x,y} H_{map}(x, y) \log \left(H_{map}(x, y) \right) \quad (2)$$



(a) QoI score: 5 (b) QoI score: 3 (c) QoI score: 1

Fig. 5. Heatmaps generated from different users on the same task. Each heatmap is labeled with its paired QoI score. Color represents the distribution density of gaze data(red to blue denotes high to low values).

Head and Hand For head (F_{Head}) and hand (F_{Hand}), we apply two processing techniques. First, we compute the frame difference of the head data to consider the change in the head movement. This involves subtracting the next frame's head data from the current frame's head data. Then, we normalize the hand data sequence, which smooths the hand joints' position in time to minimize noise. Specifically, we compute the magnitude feature function, which returns the mean (μ) and standard deviation (σ). We use these values to derive the motion magnitude (F_{mm}) and holistic distribution (F_{hd}) as follows:

$$F_{mm} = [\mu, \mu_h, \mu_l], \quad (3)$$

$$F_{hd} = [\sigma, \sigma_h, \sigma_l] \quad (4)$$

where $[\cdot]$ is the concatenation of the values, μ_h, σ_h is the mean and standard deviation of the top 10% data, and μ_l, σ_l is of the lower 10% data. The values are then concatenated and

TABLE II
PLCC AND SRCC PERFORMANCE COMPARISON ON HOLO-QOI DATASET

Method	PLCC(↑)	SRCC(↑)
Kim et al. [8]	0.6752	0.6914
Ours(F_{Gaze})	0.4775	0.5282
Ours(F_{Head})	0.5868	0.5936
Ours(F_{Hand})	0.8123	0.8148
Ours(F_{Gaze}, F_{Head})	0.6586	0.6698
Ours(F_{Gaze}, F_{Hand})	0.8279	0.8307
Ours(F_{Head}, F_{Hand})	0.8391	0.8447
Ours($F_{Gaze}, F_{Head}, F_{Hand}$)	**0.8598**	**0.8663**

used as the feature of each head and hand data. The motion magnitude represents the absolute difference of the head and hand position coordinates and creates a value proportional to the size of the motion. The holistic distribution is a measure of how spread out the velocity distribution is in each AR content, and it is calculated based on the size distribution of the head and hand movement.

B. QoI Predictor

Fig. 4 depicts an overview of our proposed framework. It takes in an interaction data sequence as input and predicts the score of the interaction. Our proposed framework follows a Transformer architecture to regress the sequence data to a single score. For each sequence of metadata, we apply our proposed feature extraction method to process the gaze, head, and hand data. In order to capture the relationship between the object and hand data, we utilize positional encoding to generate object and hand tokens. With a set of a score token followed by object tokens and hand tokens, we employ self-attention mechanisms and simultaneously regress the tokens. The state of the score token at the output of the Transformer encoder is then passed through an MLP head to retrieve a single value as the QoI prediction score.

C. Training

To train the Transformer encoder, we apply an L_1 regularization function to the output of the MLP head as follows;

$$\mathcal{L}_1 = |y - s| \tag{5}$$

where s is the predicted score and y is the ground truth score rated by the subject. $|x|$ denotes the absolute value of x. The encoder is trained to reduce the error between the prediction and the ground truth. For each trial, the dataset was randomly divided into $80\% : 20\%$ by scene, so that the training set would consist of 80% of the constructed dataset's data-score pairs, while the remaining 20% forms the test set.

D. Experimental results

We follow the model validation strategy of [19]–[22] and evaluate the models' predictive performance using Pearson's linear correlation coefficient (PLCC) and Spearman's rank-order correlation coefficient (SRCC) relative to the QoI score by computing the median correlation over 100 iterations.

We conducted experiments to evaluate the impact of individual features on our proposed method's prediction performance by comparing the results with our previous study [8]. Table II presents the evaluation results in terms of the PLCC and SRCC scores, with scores closer to 1 indicating better performance for each evaluation method. Our newly proposed method outperforms our previous method by a large margin. We also note that both PLCC and SRCC scores of our previous method showed an improvement in prediction by using our new dataset, due to an increase in the number of subjects participated.

E. Ablation study

To evaluate the impact of individual features on our proposed method's prediction, we compare the performance of all combinations of the extracted features. From the results shown in Table II, we observed that despite the simplicity of the features, the hand feature showed the highest correlation, followed by the head and gaze features. This is reasonable, as the overall process of AR interaction is mainly processed with the hands. However, the head and gaze features are also meaningful as auxiliary means of predicting scores, since the correlation increases when combined with the hand features. The best predictive performance was achieved when all features were used, demonstrating that the applied feature extraction methods effectively capture the relationship with QoI scores. These findings underscore the effectiveness of our analysis in establishing the metadata-QoI score relationship and our proposed method in predicting interaction quality from metadata.

IV. CONCLUSIONS

Interaction plays a vital role in AR, as better interactions lead to a more immersive user experience. Therefore, quantifying the QoI is essential for enhancing the AR experience. However, the existing dataset for this purpose has limitations in the quantity and diversity of human factor information, prompting us to develop our own large-scale AR interaction dataset. Hand-based interaction is the most commonly used method in AR, and its quality is influenced by the gaze and head movements. To model this relationship, we extract human factor-based features and design a model to estimate the interaction quality based on these features. The experimental results demonstrate that our prediction model has a strong correlation with human cognitive factors, and highlights the importance of considering gaze and head movements to enhance immersion in virtual environments through interaction in AR.

REFERENCES

[1] C. Moro, C. Phelps, P. Redmond, and Z. Stromberga, "HoloLens and mobile augmented reality in medical and health science education: A randomised controlled trial," *British Journal of Educational Technology*, vol. 52, no. 2, pp. 680–694, 2021, ISSN: 1467-8535. DOI: 10.1111/bjet.13049.

[2] S. Park, S. Bokijonov, and Y. Choi, "Review of Microsoft HoloLens Applications over the Past Five Years," *Applied Sciences*, vol. 11, no. 16, p. 7259, Jan. 2021, ISSN: 2076-3417. DOI: 10.3390/app11167259.

[3] A. Palumbo, "Microsoft HoloLens 2 in Medical and Healthcare Context: State of the Art and Future Prospects," *Sensors*, vol. 22, no. 20, p. 7709, Jan. 2022, ISSN: 1424-8220. DOI: 10.3390/s22207709.

[4] S. Rokhsaritalemi, A. Sadeghi-Niaraki, and S.-M. Choi, "A Review on Mixed Reality: Current Trends, Challenges and Prospects," *Applied Sciences*, vol. 10, no. 2, p. 636, Jan. 2020, ISSN: 2076-3417. DOI: 10.3390/app10020636.

[5] W. F. Bassett and R. L. Lumsdaine, "Probability Limits: Are Subjective Assessments Adequately Accurate?" *The Journal of Human Resources*, vol. 36, no. 2, pp. 327–363, 2001, ISSN: 0022-166X. DOI: 10.2307/3069662.

[6] H. J. Kang, J.-h. Shin, and K. Ponto, "A Comparative Analysis of 3D User Interaction: How to Move Virtual Objects in Mixed Reality," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2020, pp. 275–284. DOI: 10.1109/VR46266.2020.00047.

[7] G. Du, D. Guo, K. Su, *et al.*, "A Mobile Gesture Interaction Method for Augmented Reality Games Using Hybrid Filters," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022, ISSN: 1557-9662. DOI: 10.1109/TIM.2022.3181305.

[8] S. Kim, J. Kim, and S. Lee, "Quality of interaction arising from augmented reality content: A comprehensive study," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, pp. 1470–1474.

[9] N. T. Banerjee, A. J. Baughman, S.-Y. Lin, Z. A. Witte, D. M. Klaus, and A. P. Anderson, "Side-by-side comparison of human perception and performance using augmented, hybrid, and virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 4787–4796, 2021.

[10] Y. Bao, J. Wang, Z. Wang, and F. Lu, "Exploring 3d interaction with gaze guidance in augmented reality," in *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, IEEE, 2023, pp. 22–32.

[11] D. Ungureanu, F. Bogo, S. Galliani, *et al.*, *HoloLens 2 Research Mode as a Tool for Computer Vision Research*, Aug. 2020. arXiv: 2008.11239 [cs].

[12] "IEEE Standard for the Deep Learning-Based Assessment of Visual Experience Based on Human Factors," IEEE, Tech. Rep. DOI: 10.1109/IEEESTD.2022.9781357.

[13] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1954–1963.

[14] Kwanghyun Lee, A. K. Moorthy, Sanghoon Lee, and A. C. Bovik, "3D Visual Activity Assessment Based on Natural Scene Statistics," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 450–465, Jan. 2014, ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2013.2290592.

[15] K. Lee and S. Lee, "A New Framework for Measuring 2D and 3D Visual Information in Terms of Entropy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2015–2027, Nov. 2016, ISSN: 1051-8215, 1558-2205. DOI: 10.1109/TCSVT.2015.2477915.

[16] W. Kim, S. Lee, and A. C. Bovik, "VR Sickness Versus VR Presence: A Statistical Prediction Model," *IEEE Transactions on Image Processing*, vol. 30, pp. 559–571, 2021, ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2020.3036782.

[17] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, Feb. 1995, ISSN: 1558-2205. DOI: 10.1109/76.350779.

[18] R. Cappelli, D. Maio, and D. Maltoni, "Combining Fingerprint Classifiers," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2000, pp. 351–361, ISBN: 978-3-540-45014-6. DOI: 10.1007/3-540-45014-9_34.

[19] J. Kim and S. Lee, "Fully Deep Blind Image Quality Predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, Feb. 2017, ISSN: 1941-0484. DOI: 10.1109/JSTSP.2016.2639328.

[20] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep Video Quality Assessor: From Spatio-temporal Visual Sensitivity to A Convolutional Neural Aggregation Network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 219–234.

[21] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-Based Blind Image Quality Predictor," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 11–24, Jan. 2019, ISSN: 2162-2388. DOI: 10.1109/TNNLS.2018.2829819.

[22] W. Kim, A.-D. Nguyen, S. Lee, and A. C. Bovik, "Dynamic Receptive Field Generation for Full-Reference Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4219–4231, 2020, ISSN: 1941-0042. DOI: 10.1109/TIP.2020.2968283.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Supervised Single-channel EEG Decomposition using Detector-kernel Networks for Noise Reduction

Hiroshi Higashi
Osaka University, Japan
E-mail: higashi@comm.eng.osaka-u.ac.jp

***Abstract*—Signal decomposition techniques are ubiquitously utilized for the analysis, noise reduction, and classification of electroencephalogram (EEG) signals. Predominantly, these techniques harness the spatial features of EEG signals, thereby targeting multi-channel signals. However, decompositions employing spatial features often prove inefficient when applied to scenarios involving limited channel recording, such as with consumer-use equipment. This paper presents a novel approach for single-channel EEG signal decomposition. Rather than relying on a multitude of electrodes, we leverage training samples to cultivate a decomposition model based on convolutional neural networks. Our model hypothesizes that an observed EEG signal has essential components called kernels and comprises the time-shifted and amplitude-modulated kernels. Accordingly, the convolutional neural network-based signal decomposition model estimates these kernels and identifies their occurrence positions, thereby pinpointing time-shift and amplitude modulation. We validated our proposed decomposition method in the context of a noise reduction problem within a BCI application. Our experiment substantiated that our method effectively decomposes noise-related and non-noise-related components.**

I. Introduction

Signal decomposition is a pivotal technique in signal and data processing [1]. It is employed extensively not only for audio and speech signals [2] and images [3], but also for electroencephalogram (EEG) signal. There exists a wealth of techniques specifically designed to dissect EEG signals into various components. The application of EEG signal decomposition spans a broad range of uses, such as noise reduction [4], source separation/localization [5], and brain–computer interfaces (BCI) [6].

While EEG is predominantly measured with multiple electrodes (multi-channel) in most scenarios, signal decomposition techniques utilizing spatial features of signals, such as beamforming [7] and spatial filtering, have been thoroughly developed. For instance, independent component analysis (ICA) [8] predicates on the statistical independence among decomposed signals and estimates how these signals are spatially mixed. Apart from multi-channel measurement devices employed for research and medical purposes, there is an emergence of devices with fewer channels designed for consumer use. However, for these devices equipped with fewer channels, wherein the electrodes may be considerably distant from one another, it poses a challenge to leverage the spatial features for signal decomposition.

For measurement devices observing signals with spatially-sparse electrodes, a signal decomposition technique typically applied is single-channel signal decomposition. Classical time-frequency-based methods such as Fourier and (continuous) wavelet transforms [9] can be employed in this case. In these decomposition approaches, the decomposed signals are anchored to certain theoretical signals such as sinusoidal signals and brief waveforms, commonly referred to as mother wavelets.

Conversely, methods that tailor the decomposed signal for the target signal to be decomposed are termed data-driven or heuristic methods. When it comes to EEG decomposition, components may not always be assignable to specific frequency bands and can overlap across frequency bands. In such scenarios, data-driven methods—known for their flexibility in decomposing signals—prove to be highly effective [10]. Empirical mode decomposition (EMD) [11], a well-established data-driven decomposition method, is often applied to single-channel EEG signals.

The majority of existing data-driven decomposition techniques aim for adaptive application, meaning they rely solely on the signal-to-be-decomposed for the design of the decomposition. Given that the signal-to-be-decomposed may be relatively short, its decomposition can be significantly influenced by noise. Furthermore, the decomposition lacks consistency across multiple observed signals, making it challenging to determine whether a decomposed signal is desired or not.

In addition to the signal intended for decomposition, there are instances where other observed signals may be utilized to design the decomposition. For instance, in neuroscience studies, EEG signal samples gathered from multiple trials are typically batched and analyzed collectively. Similarly, for applications in brain–computer interfaces, training samples may exist that aid in the supervised training of classifiers. Despite the availability of an abundance of EEG samples, data-driven decomposition is typically conducted individually for each signal, with the decomposition being designed solely with the signal-to-be-decomposed.

Our research is motivated by the goal of leveraging

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

these EEG samples. We operate under the assumption that training samples are available and apply a (semi) supervised method to design short-length components, referred to as kernels. Similar to continuous wavelet transforms, we reconstruct an observed signal using a combination of time-shifted and amplitude-modified kernels. These adjustments in time-shift and amplitude are estimated by neural network modules, termed detectors. The optimization of these detectors occurs in tandem with the kernels. If the training samples originate from a specific person or task, the designed kernels and detectors could potentially be person-specific or task-specific.

To illustrate the efficacy of our proposed method, we conducted an experiment involving noise reduction for a BCI. In our experimental setup, the training samples were tagged with labels indicating whether they were free from noise events or had noise events at each sampling time index. Leveraging a supervised optimization strategy based on these labels, our method designs two modules that respectively output a noise-related signal and a non-noise-related signal.

II. Proposed noise reduction

We present a supervised signal decomposition technique leveraging an artificial neural network. This section commences by introducing the basic structure of the network. Subsequently, we elucidate a noise reduction method based on the proposed network.

A. Basic structure: A detector-kernel network

The proposed decomposition is founded on the assumption that an EEG signal can be reconstructed through a linear combination of short signals, denoted as kernels. This concept bears similarity to the Fourier transform, with kernels represented as complex sinusoidal signals, and the wavelet transform, with kernels depicted as mother wavelets. However, our proposed method distinguishes itself by designing these kernels in a supervised manner. Our proposed method employs a model comprising of detector and kernel networks. The detector network outputs coefficients for the linear combinations, while the kernel network produces the linear combinations using the detector output and a kernel.

To illustrate, let us consider a straightforward instance of a single reconstruction model that decomposes a signal into several components and reconstructs the original signal by summing the decomposed components. We posit that EEG signals can be reconstructed by N kernels, that is, an EEG signal is decomposed into N components. This model can be expressed as

$$x[i] = \sum_{n=1}^{N} \sum_{j=1}^{M_n} k_n[j] z_n[i-j], \tag{1}$$

where $x[i]$ is an observed signal with time index i, $\{k_n[j]\}_{j=1}^{M_n}$ is the nth kernel, M_n is the length of the nth kernel, and $z_n[i]$ is the magnitude of the nth kernel at time index i. Given that $z_n[i]$ represents the magnitudes, $z_n[i]$ must be a non-negative value ($z_n[i] \geq 0$). The objective of our proposed networks is to estimate the kernels k_n and their magnitudes $z_n[i]$.

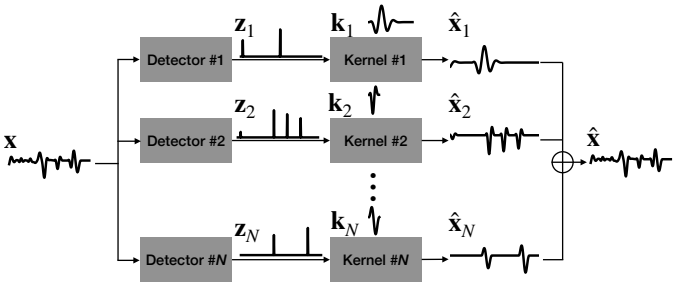


Fig. 1: Basic structure of the proposed detector-kernel network.

The proposed detector-kernel network estimates these variables in a supervised manner. The basic structure of the proposed network is depicted in Figure 1. A detector's role is to estimate $z_n[i]$. Any network that can output a non-negative signal of the same signal length as the input signal $x[i]$ may serve as a detector. For simplicity, we use a single convolutional layer in this instance. Then, the output of the nth detector is given as

$$z_n[i] = \mathrm{ReLU}\left(\sum_{j=1}^{P_n} d_n[j] x[i-j] \right), \tag{2}$$

where $x[i]$ is the observed signal and input for the detector, and $\{d[j]\}_{j=1}^{P_n}$ represents coefficients for the convolutional layer. This can be described in vector form with signal length N_t as:

$$\boldsymbol{z}_n = D_n(\boldsymbol{x}) \tag{3}$$

where $\boldsymbol{x}, \boldsymbol{z}_n \in \mathbb{R}^{N_t}$ are vector forms of the observed signal and the nth detector output defined as $\boldsymbol{x} = [x[0], x[1], \ldots, x[N_t-1]]$ and $\boldsymbol{z}_n = [z_n[0], z_n[1], \ldots, z_n[N_t-1]]$, respectively, and $D_n(\cdot)$ is an operator as described in (2). Both $\boldsymbol{x}$ and $\boldsymbol{z}_n$ have the same signal lengths. Due to the non-negative activate function, the detector output yields non-negative values. The detector output is expected to identify how time-shift and amplitude-modulate the kernel, represented by the preceding kernel network K, within the input signal $x[i]$.

The convolution with the kernels subsequently produces a decomposed signal:

$$\hat{x}_n[i] = \sum_{j=1}^{M_n} k_n[j] z_n[i-j]. \tag{4}$$

In vector form, the signal reconstruction by the N kernel network outputs can be described as:

$$\hat{\boldsymbol{x}}_n = K_n(\boldsymbol{z}_n) \tag{5}$$

where $\hat{\boldsymbol{x}}_n \in \mathbb{R}^{N_t}$ is a vector of the reconstructed signal defined as $\hat{\boldsymbol{x}} = [\hat{x}_n[0], \hat{x}_n[1], \ldots, \hat{x}_n[N_t - 1]]$ and $K_n(\cdot)$ is an operator described in (4).

The parameters to be optimized are the set of coefficients for each detector and each kernel:$\{\{d_1[i]\}_{i=1}^{P_1}, \{k_1[i]\}_{i=1}^{M_1}\}, \ldots, \{\{d_N[i]\}_{i=1}^{P_N}, \{k_N[i]\}_{i=1}^{M_N}\}$. The objective of the optimization is to minimize the discrepancy between the observed signal $x[i]$ and the reconstructed signal $\sum_{n=1}^{N} \hat{x}_n[i]$. This discrepancy is measured by

$$L = \sum_i \left\| x[i] - \sum_{n=1}^{N} \hat{x}_n[i] \right\|. \tag{6}$$

This loss function L is minimized during the training process. In this way, the detector-kernel network is trained to accurately decompose the input signal into components, each represented by a kernel.

B. Noise reduction using noisy-event occurrence periods

In this section, a method for noise reduction is described for a situation where one has a single EEG channel as input and the occurrence timings of events that might generate noise in EEG are known, but the nature of those noise components is not known. An example of this scenario could be when both an EEG channel and electrooculogram (EOG) channel are observed. The EOG signal can provide the timing of eye-related events such as blinks and saccades, which can cause noise in the EEG signal. Traditional methods for noise reduction in this situation often involve the use of multi-channel signal decomposition techniques like principal component analysis (PCA) or ICA to reduce noise components, assuming that noise components in the EEG signal are similar to the EOG signal [12]. In contrast, the method described in this section does not make such an assumption and is thus more general. The method can be used with any noise reference, as long as the timings of events that generate noise in the EEG can be detected. Other examples of noise references could include an eye tracker to detect the timings of eye blinks and saccades, or a motion capture system to reduce noise components due to body movements.

The network structure for the noise reduction is based on a signal observation model where the EEG signal, $x[i], i = 0, 1, \ldots, N_t$, contains both brain-related signals, $s_l[i]$, and noise-event-related signals, $n_l[i]$ formulated as

$$x[i] = s[i] + n[i] = \sum_{l=1}^{N_s} s_l[i] + \sum_{l=1}^{N_n} n_l[i], \tag{7}$$

where N_s is the number of the brain-related signals and N_n is the number of the brain-related signals. The purpose of noise reduction is to estimate $\{s_l[i]\}_{l=1}^{N_s}$ and $\{n_l[i]\}_{l=1}^{N_n}$.

The network consists of two detector-kernel networks: the first one is designed to extract brain-related signals and the second one to extract noise-event-related signals. Figure 2 provides a schematic representation of the network

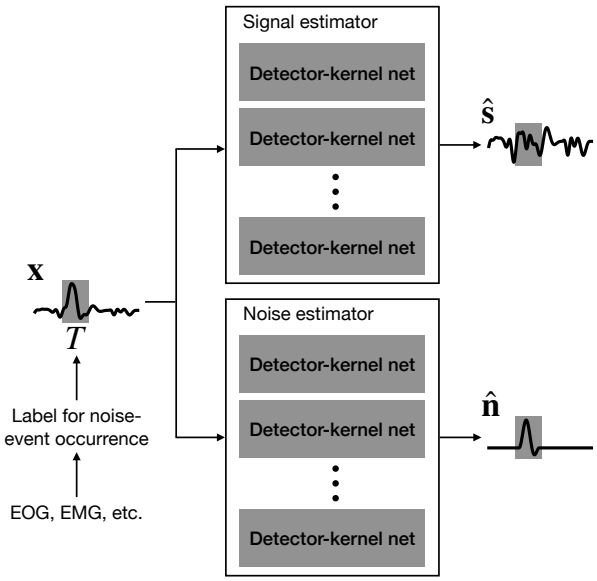


Fig. 2: Basic structure of the network for noise reduction.

structure. Here $\{D_{1,l}\}_{l=1}^{N_s}$ and $\{K_{1,l}\}_{l=1}^{N_s}$ are the detectors and kernels for the first network (signal-estimator), which is designed to extract the brain-related signals. Similarly, $\{D_{2,l}\}_{l=1}^{N_n}$ and $\{K_{2,l}\}_{l=1}^{N_n}$ are the detectors and kernels for the second network (noise-estimator), which is designed to extract the noise-related signals. The estimated noise-event-related signal $\hat{\boldsymbol{n}} \in \mathbb{R}^{N_t}$ is computed as the sum of the results of passing the input signal $\boldsymbol{x} \in \mathbb{R}^{N_t}$ through each detector-kernel pair in the second network:

$$\hat{\boldsymbol{n}} = \sum_{l=1}^{N_m} K_{2,l}(D_{2,l}(\boldsymbol{x})) = N_{\text{noise}}(\boldsymbol{x}). \tag{8}$$

where N_{noise} is the combined detector-kernel pairs for the noise-estimator. After estimating $\hat{\boldsymbol{n}}$, it is subtracted from the input signal $\boldsymbol{x}$ to obtain the signal without noise components. The result is then passed through each detector-kernel pair in the first network to estimate the brain-related signal $\hat{\boldsymbol{s}} \in \mathbb{R}^{N_t}$:

$$\hat{\boldsymbol{s}} = \sum_{l=1}^{N_s} K_{1,l}(D_{1,l}(\boldsymbol{x} - \hat{\boldsymbol{n}})) = N_{\text{signal}}(\boldsymbol{x} - \hat{\boldsymbol{n}}), \tag{9}$$

where N_{signal} is the combined detector-kernel pairs for the signal-estimator.

By minimizing loss functions, the parameters for the detector-kernels pairs are optimized. The loss function for the signal-estimator, N_{signal}, quantifies the discrepancy between the input signal and the sum of the outputs of the signal- and noise-estimators. This is calculated as

$$L_{\text{signal}} = \sum_{i=0}^{N_t - 1} |x[i] - (\hat{s}[i] + \hat{n}[i])|_2^2 = \|\boldsymbol{x} - (\hat{\boldsymbol{s}} + \hat{\boldsymbol{n}})\|, \tag{10}$$

where $\hat{s}[i]$ is the ith component of $\hat{\boldsymbol{s}}$ and $\hat{n}[k]$ is the kth component of $\hat{\boldsymbol{n}}$. On the other hand, the loss function

Algorithm 1 Learning a decomposition model for noise reduction

INPUT A set of training samples $\{\boldsymbol{x}\}$ and set of time indexes $\{T\}$ at which noise-events have occurred.
OUTPUT A signal estimator N_{signal} and noise-estimator N_{noise}
repeat
  Optimize N_{signal} by minimizing L_{signal}
  if Iterations is smaller than N_{iter} **then**
    Optimize N_{noise} by minimizing L_{noise}
  else
    Optimize N_{noise} by minimizing L'_{noise}
  end if
until Maximum iterations reach

L_{noise} for the noise-estimator, N_{noise}, evaluates the discrepancy between the input signal and the output of the noise-estimator during periods when noise events have occurred, as well as the power of the estimated signal during non-noise-event periods. With T which denotes the set of time indexes where the noise events have occurred in the epoch, the loss L_{noise} is

$$L_{\text{noise}} = \sum_{i \in T} |x[i] - \hat{n}[i]|_2^2 + \sum_{i \notin T} |\hat{n}[i]|^2. \tag{11}$$

This ensures that the noise-event-related signal is accurately estimated during noise-event-occurrence periods while giving no effect to the estimation of brain-related signals during noise-event-free periods. However, since this loss function does not take into account the output of the signal-estimator, $\hat{\boldsymbol{s}}$, a modification to L_{noise} is proposed after L_{signal} has converged or after a certain number of iterations in the optimization. The modified loss function L'_{noise} is defined as

$$L'_{\text{noise}} = \sum_{i \in T} |x[i] - (\hat{n}[i] + \hat{s}[i])|^2 + \sum_{t \notin T} |\hat{n}[i]|^2. \tag{12}$$

The pseudocode is provided in Algorithm 1 succinctly encapsulates the training process for the noise reduction model.

III. Experiment

A. Dataset

An open dataset for a motor-imagery-based BCI [13] used in this experiment. The dataset includes data from 15 participants, with each one performing seven different motor-imagery task. The dataset proposed the classification problem in which a signal is classified into a class out of C_1 (elbow flexion), C_2 (elbow extension), C_3 (supination), C_4 (pronation), C_5 (hand close), C_6 (hand open), and C_7 (rest). The dataset also included three EOG channels, with the electrodes placed on the left, right, and top of eyes.

As the focus of this study is on single-channel decomposition, only data from the F1 channel was used. The F1 channel is located on the frontal head area of the head, an area is often affected to noise-events such as eye blinks and saccades.

B. Classification procedures

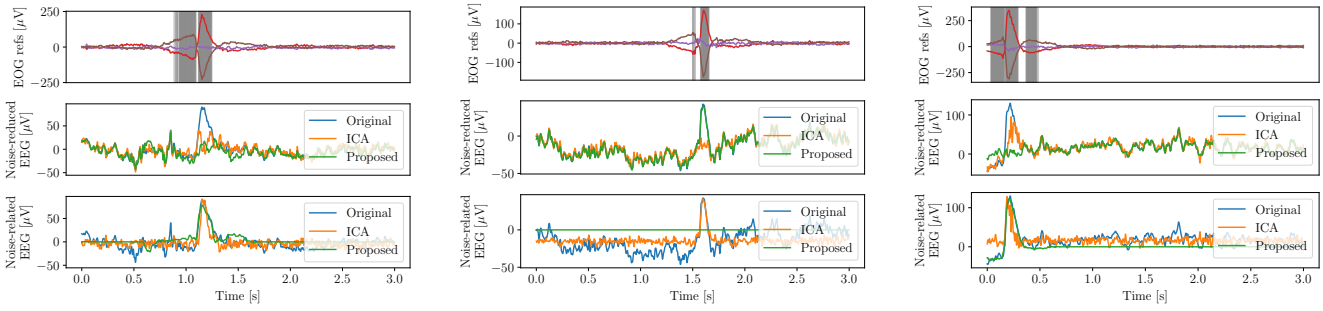
The available samples were divided into three distinct sets: samples for signal decomposition training ($\mathcal{R}_1$), samples for classifier training ($\mathcal{R}_2$), and a testing sample set ($\mathcal{E}$). For $\mathcal{R}_1$, samples were from C_1, C_2, C_4, C_5, and C_6. The samples from C_3 and C_7 were used for the classifier training and testing sets ($\mathcal{R}_2$ and $\mathcal{E}$). These were split into training and testing samples using a leave-one-out cross validation (CV) approach.

The proposed method trained a signal decomposition model on the available samples $\mathcal{R}_1$. The signal- and noise-estimators employed 32 pairs of detector-kernel networks ($N_s = N_m = 32$). These detectors were constructed using four layers of convolutional networks, with a kernel size of 60 ms and four channels for the intermediate layers. As for the kernel networks, a kernel size was 1 s. The noise-estimators were optimized with the noise-event-free and noise-event-occurrence labels at each time index for each sample which were obtained from the EOG signals. From the three EOG channels, three bipolar EOG signals were derived. Time indexes where the EOG signals exceeded ± 50 μV were labeled as noise-event-occurrence. For the optimization of the decomposer network, we used an Adam optimizer (0.001 of learning rate, 0.5 of β_1, 0.999 of β_2, and 10^{-5} of weight decay). The optimization was carried out over 1,000 iterations with a batch size of 50. Following optimization, the signal decomposition model produced a noise-related signal and a brain-related signal for each sample. The brain-related signal was then input into a support vector machine (SVM) classifier that employs a radial basis function kernel. This SVM classifier (kernel coefficient γ ="scale", regularization parameter $C = 1$) was tasked with classifying each signal into either C_3 or C_7.

In addition to the proposed method, two alternative conditions were also evaluated for performance comparison: *Raw* and *Reject*. In the condition Raw, the signal was not subjected to any form of decomposition or noise reduction. The observed signals in $\mathcal{R}_2$ were input directly into the classifier. In the condition Reject, a simple form of noise rejection was implemented. Any samples having time indexes that were labelled as containing noise events were entirely excluded from classification.

C. Results

Figure 3 presents some examples of signal by noise reduction. The top panel exhibits the reference EOG signals, which were used to label each time index as either noise-event-free or noise-event-occurrence. The gray-shaded areas indicate time indexes labeled as noise-event-occurrence. It seems that large amplitude fluctuations in these areas on the EOG signals are caused by noisy events,



(a) (b) (c)

Fig. 3: Examples of signal samples from the sample set $\mathcal{R}_2$. The top panel shows the EOG reference signals. The middle panel shows the observed EEG signals (Original), the noise-reduced signals by ICA, and the noise-reduced signal by the proposed method. The bottom panel shows the observed EEG signals (Original), the noise-related signals by ICA, and the noise-related signal by the proposed method.

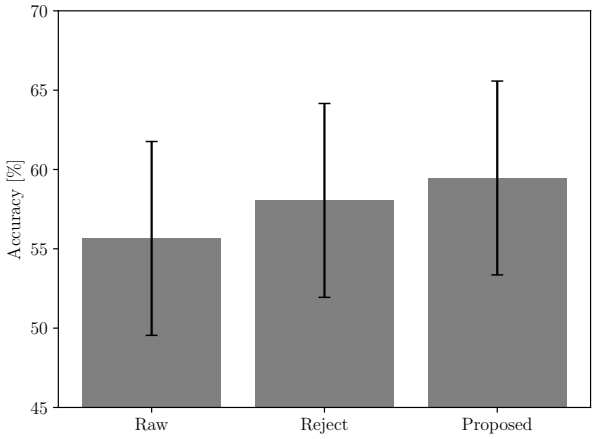


Fig. 4: Classification accuracy rates.

such as eye blinks or eye movements. The middle panel displays the original EEG signals from the F1 channel alongside their noise-reduced version. Noise reduction was performed using both the proposed method and ICA for comparison. In the ICA noise reduction, signals observed from all 64 EEG electrodes were used. Any independent components that had a correlation coefficient greater than 0.7 and less than −0.7 with any EOG signals were removed from the reconstruction for the noise-reduced signals. As can be seen in Figure 3a, both the proposed method and ICA successfully removed large fluctuations that were supposedly caused by noisy events. However, both methods often failed to completely remove these fluctuations, as demonstrated in Figure 3b for the proposed method and Figure 3c for ICA. These results show that while both methods have the potential to reduce noise-related fluctuations in the signal, neither is completely infallible.

The classification accuracy rates for each condition are presented in Figure 4. The proposed method exhibited superior accuracy rates compared to the condition Raw, indicating a significant improvement with a p-value of 0.007 given by a paired t-test. Moreover, the proposed method achieved comparable accuracy rates to those of the condition Reject, with a non-significant difference reflected by a p-value of 0.298. This highlights that the effectiveness of the proposed method in reducing noise-related components and its subsequent contribution to a robust classification.

IV. Conclusions

This paper presents a novel approach to single-channel EEG signal decomposition, grounded in the premise that a signal consist of several time-shifted and amplitude-modulated kernels. By leveraging a neural network-based model, we are able to estimate these kernels and accurately determine their time shifts and amplitude modulations. Our method was validated using a noise reduction problem in the context of BCI applications. Experimental results showed that our proposed method effectively separates noise-related and brain-related signals in a supervised manner. This presents promising prospects for the enhancement of noise reduction techniques and, consequently, for the performance and reliability of BCIs.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science KAKENHI, grant numbers: 19K16894 and 22H05163.

References

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley, 2002.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007. DOI: 10.1109/TASL.2006.885253.

[3] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022. DOI: 10.1109/TPAMI.2021.3059968.

[4] S. Vorobyov and A. Cichocki, "Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis," *Biological Cybernetics*, vol. 86, no. 4, pp. 293–303, 2002. DOI: 10.1007/s00422-001-0298-6.

[5] S. Asadzadeh, T. Yousefi Rezaii, S. Beheshti, A. Delpak, and S. Meshgini, "A systematic review of EEG source localization techniques and their applications on diagnosis of brain abnormalities," *Journal of Neuroscience Methods*, vol. 339, no. February, 2020, ISSN: 1872678X. DOI: 10.1016/j.jneumeth.2020.108740.

[6] F. Lotte, L. Bougrain, A. Cichocki, *et al.*, "A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, aab2f2, 2018. DOI: 10.1088/1741-2552/aab2f2.

[7] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988. DOI: 10.1109/53.665.

[8] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, Jun. 2000. DOI: 10.1016/S0893-6080(00)00026-5.

[9] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[10] V. Bajaj and R. B. Pachori, "Classification of seizure and nonseizure EEG signals using empirical mode decomposition," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1135–1142, 2012. DOI: 10.1109/TITB.2011.2181403.

[11] N. E. Huang, Z. Shen, S. R. Long, *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998. DOI: 10.1098/rspa.1998.0193.

[12] H. Nolan, R. Whelan, and R. B. Reilly, "FASTER: Fully automated statistical thresholding for EEG artifact rejection," *Journal of Neuroscience Methods*, vol. 192, no. 1, pp. 152–162, Sep. 2010. DOI: 10.1016/j.jneumeth.2010.07.015.

[13] P. Ofner, A. Schwarz, J. Pereira, and G. R. Müller-Putz, "Upper limb movements can be decoded from the time-domain of low-frequency EEG," *PLoS ONE*, vol. 12, no. 8, pp. 1–24, 2017. DOI: 10.1371/journal.pone.0182578.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Cross-Subject Classification of Spoken Mandarin Vowels and Tones with EEG Signals: A Study of End-to-End CNN with Fine-Tuning

Xinyu Wang*, Mingtao Li*†, Hao Li*, Sio Hang Pun† and Fei Chen*
* Southern University of Science and Technology, Shenzhen, China
E-mail: 12232141@mail.sustech.edu.cn, limt@mail.sustech.edu.cn, lih9@sustech.edu.cn,
fchen@sustech.edu.cn Tel/Fax: +86-755-88018554
† University of Macau, Macau, China
E-mail: YB97473@umac.mo, lodgepun@um.edu.mo Tel/Fax: +853-88224467

***Abstract*—Direct speech brain-computer interface (DS-BCI) is an ideal way for speech communication by decoding signals collected from the brain. Electroencephalogram (EEG) has gained widespread use in DS-BCI studies due to its simplicity of operation and high temporal resolution. However, as human brain exhibits considerable inter-individual variability, classification models trained on the basis of data from one subject may not generalise well to other individuals, which is a major challenge in existing EEG signal classification studies. In this paper, the cross-subject classification performance of spoken Mandarin speech with EEG signals was investigated by using an end-to-end convolutional neural network (CNN) model pre-trained on the source data and fine-tuned on the target data. The raw EEG signals were directly used as the input to the model without using extracted features. In addition, adding Gaussian noise was used as the data augmentation method in order to deal with the unbalanced dataset. The proposed method was tested on a collected EEG dataset of spoken Mandarin speech, including vowel classification and tone classification tasks. The average classification accuracies of four vowels and four tones were 63.1% and 51.7% respectively. The average accuracy of tone classification was significantly improved compared with the machine learning and subject-dependent methods. The results of this work showed the potential of the fine-tuning based CNN model in the cross-subject studies of EEG decoding.**

I. Introduction

Brain-computer interface (BCI) provides a direct access between the brain and external devices [1]. Direct speech brain-computer interfaces (DS-BCIs) are proposed as an ideal way of communication, because speech serves as a fundamental form of human communication and plays a key role in conveying thoughts and ideas. DS-BCIs decode human thoughts by reading and analysing signals from the brain [2]. They are useful for individuals with neurological disorders, such as aphasia, enabling them to effectively communicate with the world beyond their immediate limitations [3].

In the studies of DS-BCIs, speech includes three distinct modalities: overt speech, intended speech, and imagined speech [4]. Imagined speech refers to the internal dialogue or thoughts that occur in the mind of an individual without any actual form of verbal expression. Overt speech, i.e., spoken speech, refers to speech that is explicitly produced with articulation and vocalisation. While imagined speech is the most ideal form of speech BCIs, it is hard to decode due to its low signal-to-noise ratio and difficulty of temporal alignment. Therefore, this work concentrated on spoken speech, which has stronger intensity and detectable motor activities.

Many current DS-BCI studies use machine learning methods and require feature extraction [5]. In the previous studies, electroencephalogram (EEG) signals of spoken speech were analyzed using common spatial patterns and Riemannian manifolds as feature extraction methods and a linear discriminant analysis as the classifier [6], [7]. Four classes of Mandarin vowels and four classes of Mandarin tones were classified, and the average accuracies are 63.1% and 42.9% respectively. In another EEG study on Mandarin, Yang et al. classified tonal and nontonal sentences across subjects using rational asymmetry and binomial variable algorithm with 98.82% accuracy [8]. With the development of deep learning techniques, they have been applied to the studies of many EEG classification problems. For instance, some studies converted EEG signals into spectrograms as input to convolutional neural network (CNN) models [9], [10].

Extracting speech features carries the risk of losing potentially valuable information. Deep learning based models can avoid this problem by learning features directly from raw data. End-to-end neural networks have been used in different EEG classification tasks, such as motor imagery [11], attentive mental state detective [12], etc. Sereshkeh et al. employed a multilayer perceptron artificial neural network in a ternary classification task, i.e., distinguishing between 'yes', 'no' and the remaining trials, and achieved an average accuracy of 54.1 ± 9.7% [13]. By utilizing an extreme learning machine, Tan et al. classified EEG signals in BCI system and achieved a higher classification accuracy compared with support vector machines [14].

However, deep learning methods require a large amount of training data. Most current works relied on data obtained from the same subject [15]. Collecting a large dataset is difficult due to the limitation of human attention and the health status of patients. Therefore, it is necessary to explore cross-subject

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

classification methods that can improve the generality performance of classification models [16]. For instance, Cooney et al. used deep CNN and transfer learning methods to classify EEG-based imagined vowels, with the highest overall accuracy of 35.7% [17].

This work presented a deep learning method for decoding spoken Mandarin with EEG signals and investigated its cross-subject classification performance. The experimental design consists of two tasks using the collected dataset to classify four vowels and four tones of Mandarin respectively. A CNN model was pre-trained on the source data and then fine-tuned on the target data. The raw EEG data was used as input to the model. In addition, the EEG dataset was augmented by adding Gaussian noise to avoid the unbalanced dataset.

II. Methods

A. Dataset

The dataset employed in this experiment was obtained from the previous studies [6], [7], involving EEG signals of four speech modalities (i.e., listening, imagined speech, intended speech, and overt speech). The participants consisted of eleven individuals, including four female and seven male subjects (22.6 ± 3.2 years). The dataset collected EEG signals containing 70 Mandarin monosyllabic words as stimuli. It contained a total of 54 words consisting of different combinations of 4 vowels (i.e., /a/, /i/, /u/ and /ü/), 5 consonants (i.e., /b_/, /f_/, /j_/, /l_/ and /m_/) and 4 tones (i.e., Mandarin tones 1, 2, 3 and 4). It also contained 16 single vowels, which were combinations of four vowels and four tones in Mandarin. All words used in the dataset were displayed in Table I.

Scalp EEG signals were recorded using a 64-channel electrode cap manufactured by Neuroscan Inc. The stimuli and audio recordings were acquired at a sampling rate of 500 Hz. The ground channel was positioned on the forehead. The reference channels were placed at the tip of the nose, bilateral mastoids, and above and below the left eye to remove oculomotor signal interference. Data from one subject (i.e., the 6th subject) was excluded from the analysis due to the presence of numerous artifacts. After manual elimination of the less valid data, a total of 3297 trials were retained. More details of the experimental paradigm can be found in [6], [7].

TABLE I
List of Monosyllabic Stimuli

	b_	**f_**	**j_**	**l_**	**m_**	**_**
/a/	ba	fa		la	ma	a
/i/	bi		ji	li	mi	i
/u/	bu	fu		lu	mu	u
/ü/			ju	lü		ü

B. Data Preprocessing

All the EEG data were preprocessed using EEGLAB [18], which involved re-referencing and filtering. Filtering included a finite impulse response band-pass filter with cut off frequency of 0.5 $\sim$ 70 Hz and a notch filter with cut off frequency of 49 $\sim$ 51 Hz. The data were then subjected to independent component analysis and manual removal of artifacts, including those from electrocardiography, electrooculography, electromyography, etc [19].

This experiment extracted a total duration of 2500 milliseconds of spoken speech EEG data across all channels. The size of each trial was (1250, 62), corresponding to the number of samples and the number of channels. The start point and the duration of each trial were identified using a voice activity detection algorithm. To handle varying length of data, the input matrices were processed with zero-padding. The final size of each input matrix was (900, 62). All data were normalised.

The experiment was divided into two tasks, which were tone classification and vowel classification. For each trial, there were two labels that indicated its tone and vowel respectively.

C. Network Structure

The CNN architecture employed in this study was designed for EEG decoding based on EEGNet [20]. The structure of the CNN architecture was shown in Table II. The three dimensions of the size of feature map corresponded to channel, width and height respectively.

The feature extraction block of the network comprised three convolutional layers. The raw EEG data was used as input of the network. The first convolutional layer was composed of two individual layers, namely 1a and 1b. Layer 1a performed temporal convolution, while layer 1b performed spatial filtering. Subsequently, batch normalization was applied, and the results were fed into a non-linear activation function. Similar structures were employed for the second and third convolutional layers, accompanied by max pooling operations. The non-linear activation functions for the first three layers were all ReLU. The output of the network was generated using a fully connected network and a softmax function, yielding the final classification results. A dropout layer with a scale of 0.3 was used.

D. Data Augmentation

In both vowel and tone classification tasks, an imbalance in sample size between the four classes was observed, which may lead to a decrease in classification accuracy. To solve this problem, the dataset was augmented by adding Gaussian noise to the time-domain data before the pre-training process [21]. For the class with a small sample size (e.g., vowel /a/ and tone 4), trails were randomly selected and Gaussian noise with a mean of 0 and a standard deviation of 0.01 was added. This method allowed for a balanced sample size for all four classes in both tasks. The maximum difference in sample size between any two classes was not greater than 10.

E. Model Training Procedure

The training process for the pre-trained model used data from all 9 individuals except the target subject. The dataset was divided into training set and validation set. 80% of the

TABLE II
NETWORK PARAMETERS

Layer Type	Size of Feature Map	Kernel Size	Stride	Padding
Input EEG	1×900×62			
Conv1a	40×895×66	10×1	1	2
Conv1b	40×899×62	1×9	1	2
Batch normalization	40* 899×62			
ReLU	40×899×62			
Conv2	100×899×62	5×5	1	2
Batch normalization	100×899×62			
ReLU	100×899×62			
Max pooling	100×149×10	6×6	6	0
Conv3	200×149×10	5×5	1	2
Batch normalizaiton	200×149×10			
ReLU	200×149×10			
Max pooling	200×14×1	10×10	10	0
Flatten	1×2800			
Linear	1×100			
ReLU	1×100			
Dropout layer	1×100			
Linear	1×4			
Softmax	1×4			

source data were used for training and the remaning 20% were used for validation. During the training process, a batch size of 16 was employed. The optimiser was Adam [22]. The loss function used to measure the difference between the predicted outputs and the true labels was the cross-entropy loss function. Learning rate for the pre-training phase was set to 0.001. The training process for the pre-trained model continued for 100 epochs. At the end of the training process, the model with the highest accuracy on the validation set was selected as the pre-trained model. This model was then used as the starting point for further fine-tuning.

For the fine-tuning stage, data of the target subject was used, allowing the model to learn and adapt to the specific features of the target subject. The data of the target subject was divided into training set, validation set, and testing set. 50% of the trials were divided into 25% for training and validation respectively, and the remaining 50% were used for testing. This split ratio was set to evaluate the performance of the model and to ensure that it generalized well to unseen data. The process did not use any data augmentation techniques. However, when splitting the target dataset, the numbers of trials for each class of data balanced in the training and validation sets, which prevented bias in the model learning process and ensured that all classes of data were fully utilised. The learning rate was set to 0.002, while keeping the other parameters consistent with the pre-training step. At the end of this process, the model with the highest accuracy on the validation set was selected as the final model.

III. RESULTS

Table III and Table IV show the confusion matrices for tone classification and vowel classification, respectively, including average accuracy and standard error of mean value for all 10 subjects as the target subject. The column labels in the confusion matrix correspond to the correct labels and the row labels correspond to the predicted labels. The highest accuracy values are marked in bold.

The average accuracy of tone classification is 51.7%. Tone 4 has the highest accuracy of 62.4%. In the previous study of tone classification task [7], i.e., the subject-dependent method using Riemannian manifold, the classification accuracies for the four classes are 37.4%, 40.5%, 48.3% and 44.7%, respectively. The average accuracy is 42.9%. The results in Table III show that the classification accuracy of tone 1 is comparable to that in the previous study. The accuracies of tone 2, tone 3 and tone 4 improve by 3.8%, 10.6%, 17.7% points respectively. The overall accuracy is 8.8% point higher than that in the previous study.

The average accuracy of vowel classification is 63.1%. Vowel /a/ has the highest accuracy of 79.2%. In the previous study [6], the average accuracy of vowel classification is 68.5%. The accuracy in this study is slightly lower than that of the subject-dependent experiment. It is noted that the classification accuracy of vowel /ü/ reaches 42.5%. It is 13.3% higher than that of the previous method, which is 29.2% [6].

The average precision, recall, and F1 scores for the two tasks are summarized in Table V. In both tasks, the three evaluation metrics yield similar results. The precision, recall, and F1 scores of the vowel classification task consistently outperform those of the tone classification task, aligning with the accuracy findings.

TABLE III
CONFUSION MATRIX OF SPOKEN TONE CLASSIFICATION

Accuracy (%)	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	37.1±3.6	31.8±4.3	14.7±4.1	16.4±4.8
Tone 2	21.5±3.5	44.3±5.4	18.9±5.1	15.2±4.5
Tone 3	7.8±2.5	14.7±2.9	58.9±7.7	18.6±6.7
Tone 4	15.2±3.6	12.4±2.9	10.2±5.4	**62.4±7.0**

TABLE IV
CONFUSION MATRIX OF SPOKEN VOWEL CLASSIFICATION

Accuracy (%)	/a/	/i/	/u/	/ü/
/a/	**79.2±5.3**	8.28±3.2	9.26±2.2	3.22±1.3
/i/	8. 6±2.8	51.7±3.7	11.78±2.5	27.89±5.1
/u/	8.5±2.8	18.17±4.2	62.0±6.6	11.17±2.8
/ü/	7.4±3.1	30.71±3.5	19.55±6.5	42.5±9.2

IV. DISCUSSION AND CONCLUSIONS

This work presented a cross-subject study for classifying spoken speech with EEG signals. It was tested on a Mandarin spoken speech EEG dataset, including tone classification and vowel classification tasks. A CNN network was employed, with the model pre-trained on data from non-target subjects

TABLE V
PRECISIION, RECALL AND F1 SCORE OF VOWEL CLASSIFICATION AND TONE CLASSIFICATION

	Precision (%)	**Recall** (%)	**F1 Score** (%)
Vowel	58.6	58.9	58.7
Tone	50.3	50.7	50.5

and subsequently fine-tuned using data from the target subject. The model was cross-validated with data from 10 subjects. The average accuracies of vowel classification and tone classification were 63.1% and 51.7% respectively, achieving comparable performance to the previous subject-dependent studies. The tone classification task exhibits an improvement of 8.8% point in average accuracy.

Cross-subject classification studies can be valuable in making full use of existing datasets to train models used on target subjects. The average accuracy achieved in the tone classification in this study surpassed that of the previous subject-dependent experiment. In the case of vowel classification, while the average accuracy fell slightly below the subject-dependent experiment, it still yielded comparable results. This can be attributed to the inherent challenge posed by subject-independent experiments, where data variances among different subjects introduce additional complexity. The outcomes of this study exhibit the ability of pre-trained models to extract generalizable features or patterns from the source data and effectively apply them to the target subject after fine-tuning. However, this study only focuses on spoken speech EEG signals. Future extension of this research could explore imagined speech, considering the correlation between the envelope of imagined speech and that of spoken speech [23].

Vowel /ü/ has a much lower classification accuracy than the other three classes in the previous study, because of its similar formant structure to other vowels [6]. However, this experiment showed a remarkable improvement in classification accuracy for vowel /ü/, indicating the advantage of the deep learning method over the machine learning approach in distinguishing vowels with similar formant structures.

In addition, the subject-independent study revealed improved results in the tone classification compared with the subject-dependent method, and the average accuracies of three classes demonstrated improvement, suggesting the presence of shared features from Mandarin tones across different subjects. This finding can be useful for future studies in Mandarin DC-BCIs, particularly in enhancing the adaptability of cross-user classification models by extracting appropriate features from tones.

In conclusion, the proposed method demonstrated its potential to decode spoken Mandarin speech with EEG signals across different subjects. The vowel classification obtained higher average accuracy and the accuracy of the previously challenging class was effectively improved. On the other hand, the tone classification had a better cross-subject performance.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61971212), the Basic Research Foundation of Shenzhen (Grant No. JCYJ20220818101217037), and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022B1515120056).

REFERENCES

[1] S. Vaid, P. Singh, and C. Kaur, "EEG signal analysis for bci interface: A review," *Fifth International Conference on Advanced Computing & Communication Technologies*, pp. 143–147, 2015.

[2] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, 2010.

[3] S. Sarasso, S. Määttä, F. Ferrarelli, R. Poryazova, G. Tononi, and S. L. Small, "Plastic changes following imitation-based speech and language therapy for aphasia: A high-density sleep EEG study," *Neurorehabilitation and Neural Repair*, vol. 28, no. 2, pp. 129–138, 2014.

[4] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *IScience*, vol. 8, pp. 103–125, 2018.

[5] N. Hashim, A. Ali, and W.-N. Mohd-Isa, "Word-based classification of imagined speech using EEG," *Computational Science and Technology*, pp. 195–204, 2018.

[6] M. Li, S. H. Pun, and F. Chen, "A preliminary study of classifying spoken vowels with EEG signals," *International IEEE/EMBS Conference on Neural Engineering*, pp. 13–16, 2021.

[7] M. Li, S. H. Pun, and F. Chen, "Mandarin tone classification in spoken speech with EEG signals," *European Signal Processing Conference*, pp. 1163–1166, 2021.

[8] S. Yang, T. Jung, C. Lin, *et al.*, "Recognizing tonal and nontonal mandarin sentences for EEG-based brain–computer interface," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1666–1677, 2021.

[9] G. Vrbancic and V. Podgorelec, "Automatic classification of motor impairment neural disorders from EEG signals using deep convolutional neural networks," *Elektronika ir Elektrotechnika*, vol. 24, no. 4, pp. 3–7, 2018.

[10] A. Vilamala, K. H. Madsen, and L. K. Hansen, "Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring," *IEEE 27th International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2017.

[11] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Systems with Applications*, vol. 114, pp. 532–542, 2018.

[12] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *Journal of Neural Engineering*, vol. 16, no. 2, 026007, 2019.

[13] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG classification of covert speech using regularized neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2292–2300, 2017.

[14] P. Tan, W. Sa, and L. Yu, "Applying extreme learning machine to classification of EEG BCI," *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, pp. 228–232, 2016.

[15] A. Craik, Y. He, and J. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *Journal of Neural Engineering*, vol. 16, no. 3, 031001, 2019.

[16] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring EEG features in cross-subject emotion recognition," *Frontiers in Neuroscience*, vol. 12, pp. 162–162, 2018.

[17] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG," *IEEE International Conference on Systems, Man and Cybernetics*, pp. 1311–1316, 2019.

[18] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[19] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech synthesis using EEG," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1235–1238, 2020.

[20] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 056013, 2018.

[21] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *Journal of Neuroscience Methods*, vol. 346, 108885, 2020.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *San Diego: The International Conference on Learning Representations*, 2015.

[23] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura, "Synchronization between overt speech envelope and EEG oscillations during imagined speech," *Neuroscience Research*, vol. 153, pp. 48–55, 2020.

Decoding time-course of saliency network of fMRI signals by EEG signals using optimized forward variable selection: a concurrent EEG-fMRI study

Tung Dang[1,2], Kentaro Ono[1], Takafumi Sasaoka[1], Shigeto Yamawaki[1], and Maro G Machizawa[1,3*]
[1] Hiroshima University, Hiroshima, Japan
[*]Email: machizawa@hiroshima-u.ac.jp Tel: +81–3–6897–3568
[2] The University of Tokyo, Tokyo, Japan
[3] Xiberlinc Inc., Tokyo, Japan

Abstract — **Coinciding with the rapid growth of neural engineering technologies, the social demand for digital mental health is rapidly rising to date. While neuroimaging techniques, such as functional MRI or MEG, allow us to monitor neural activities in depth brain regions and sparse networks, direct application of those tools is limited in real life to assess neural well-being in real life. Literature suggests human interoceptive awareness is associated with activities and connections from the insular cortex as a part of the saliency network. Accumulating evidence suggests monitoring and modulation of these functional connectivities may reflect or improve our mental states. The neurophysiological link between signals from fMRI and EEG is isolated due to varieties of reasons; however, here we propose a novel signal processing approach to decode the time-course of functional connectivity obtained from fMRI by the concurrently recorded EEG signals (n = 28). To exploit the feasibility of this fusion network decoding, we present a novel machine learning method, "optimized forward variable selection decoder," that achieves higher decoding accuracy over classical methods. We propose a unified EEG model for brain-computer-interface applications in real life that can be applied for monitoring mental well-being in daily life.**

I. Introduction

In the last decade, the applications of machine learning (ML) approaches have attracted increasing interest in the vast majority of modalities of fields from biology to psychology, including neuroimaging studies. Previously, the forward variable selection approach was proposed in the domain of genetic data analysis. Here, we present a novel application of an optimized decoding algorithm for neuroimaging data. The decoding of neural information has dramatically increased and improved in neuroscience to date; however, the curse of its high dimensionality and non-stationarity of neurophysiological data variability yields a number of challenges.

For instance, generalization performance is one of the most important focuses of ML studies, such as ensemble learning methods. The ensemble learning methods improve their generalization ability by fusing the prediction results of basic models (e.g., the random forest is an ensemble of decision trees). However, the high dimensionality of data typically impedes the generalization performance of decoding approaches, causing estimation instability, model overfitting, local convergence, or critical estimation errors. Specifically, it has been a challenge to successfully decode with basic ML models because the decoding accuracy notably decreases in the high dimensional computations [1], [2]. Moreover, the high dimensionality is also linked to longer computational time. Therefore, it is pivotal to efficiently identify optimal sets of features for ML models that achieve successful generalization.

The recent development of the brain-computer-interface (BCI) has gained hope for solutions for mental health issues in real-life scenarios. Various BCI models have been proposed; functional connectivity (FC) is thought to reflect remotely interconnected neural activities across brain regions, and EEG correlates of FC are thought to reflect mental states [3]. However, the direct link between fMRI and electrophysiological responses is reportedly low [4], indicating potential risks to the fidelity of neural markers connecting the two distinct neural responses.

A successful integration of decoded fMRI-driven functional networks using EEG signals can have several potential benefits. It empowers us to investigate both the spatial and temporal dimensions of brain activity, reflecting remotely

This research was supported by JST Moonshot R&D Goal 9 (JPMJMS2296) and JST COI (JPMJCE1311 and JPMJCA2208).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

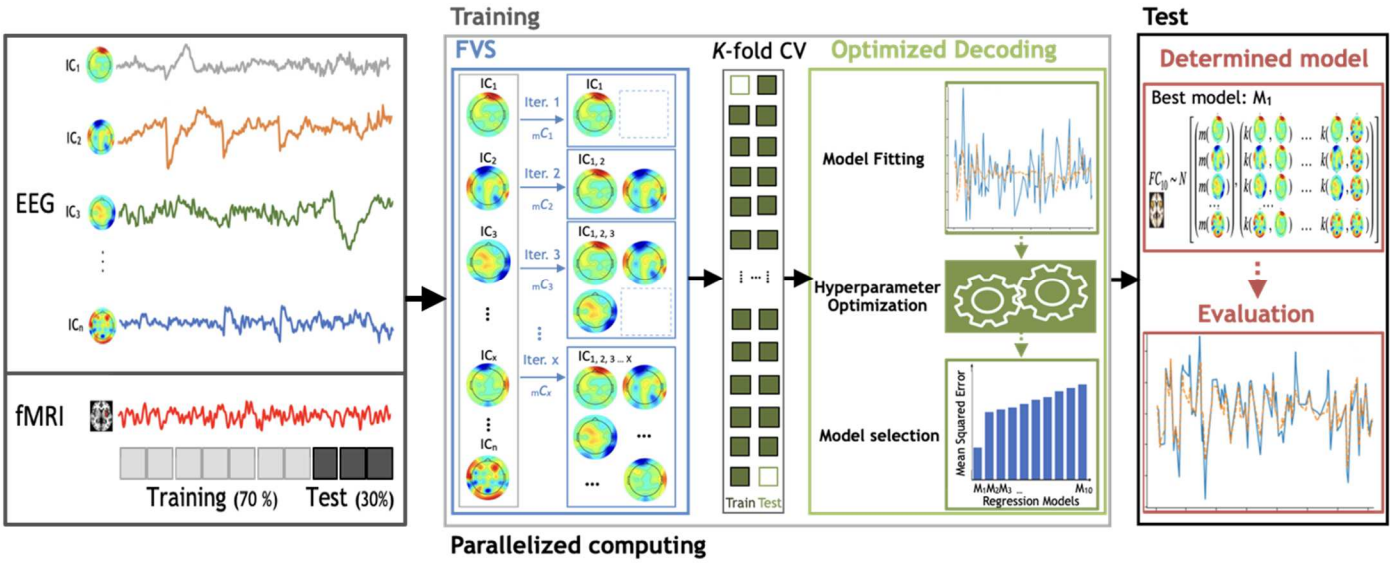


Fig. 1. Computational workflow of the optimized forward variable selection decoding toolbox (oFVSD) on the multimodal neuroimaging dataset. All features, i.e., the EEG ICAs, underwent the iterative FVS steps ('Iter.'), followed by regression-based machine learning with K-fold cross-validation ('CV'). At the decoding step, the random and grid search strategies with cross-validation were adopted to optimize the hyperparameters of model(s) at each iteration of the FVS algorithm. The training data was used to guide the refinement of decoding outcomes by minimizing the mean squared error (MSE) value, and the performance of the selected model was subsequently assessed on a separate test dataset to ensure its predictive accuracy. n: the total number of features; x: a user-defined maximum number of features. A determined computational model (a case of Gaussian process regression) is shown in the top right corner.

interconnected brain regions. However, linking the brain-to-brain activities obtained from concurrent EEG-fMRI datasets has been challenging. Yet, it can potentially connect putatively distant neural activities between indirect and direct measures of brain activities obtained from fMRI and EEGs, respectively. Here, we examined the feasibility of the enhanced forward selection decoding toolbox, which was developed in our previous study using structural MRI datasets [5] on the cross-modal neuroimaging data. Moreover, a saliency network was selected as a use case sample for this feasibility as one of the major FCs.

II. Methods and Material

We developed a novel decoding pipeline to overcome these challenges by incorporating two frameworks. Firstly, we developed an automatic ML framework that can automatically select the best model out of multiple regression models as well as classifier models with the forward variable selection (FVS) for each model [3,4]. To examine the feasibility of our approach, we examined regression models to decode the time course of FC of the saliency network obtained from functional magnetic resonance imaging (fMRI) by the concurrently recorded cortical electroencephalography (EEG) signals recorded from healthy young adults (n = 28). On ten selected regression models (see details in [5]), the FVS was followed by a hyperparameter optimization. The FVS reduces the dimension of this input, including regularization models for feature selection, as well as models that perform a random sampling of the available features to generate different subspaces of features in order to achieve a bias-variance trade-off. Note kernel models were considered to let interactions between features and nonlinear functions, therefore driving the prediction. Secondly, we employed a 'parallelized' pipeline that was coupled with the FVS algorithms to maximize the predictive performance.

A. Concurrent EEG-fMRI Data Acquisition

A 15-min eyes-open resting-state fMRI was recorded in the Siemens Skyra scanner (T2*-weighted echo-planar imaging: TR = 2,500ms; TE = 29ms; flip angle 90°; 3.0x3.0x3.5mm; slice thickness 3.5mm; FOV 192mm). The structural images of each participant were acquired using the following parameters: TR = 2,300 ms, TE = 2.98 ms, voxel size of 1.0 × 1.0 × 1.0 mm, flip angle = 9°, and FOV = 256 mm. Concurrent EEG signals were continuously recorded at 1,000 Hz with an MRI-compatible 32 channels system (BrainAmp MR plus, Brain Products).

B. fMRI Data Preprocessing

To extract the FC from the resting-state fMRI data, we performed a group independent component analysis (group-ICA) using CONN Toolbox (version 21.a; [7]). Data were

preprocessed by the default preprocessing pipeline in the CONN Toolbox, including realignment, slice timing correction, outlier detection, direct segmentation, normalization to Montreal Neurological Institute (MNI) space, and smoothing. Then, motion artifact from the blood-oxygen-level-dependent (BOLD) signal was denoised. A group-ICA was performed to separate fMRI time series into spatial source maps and loadings. The number of components was set to 20, resulting in 20 independent components (fMRI-IC). Of those fMRI-ICs, a saliency network was identified based on contributing brain regions. For each participant, we extracted the time series of each IC that represents a measure of FC in different intrinsic networks.

C. EEG Data Preprocessing

EEG data were preprocessed with the FMRIB plug-in (https://fsl.fmrib.ox.ac.uk/eeglab/fmribplugin/) for EEGLAB (https://sccn.ucsd.edu/eeglab/) followed by conventional artifact rejection procedures implement on the EEGLAB and down-sampled to 500Hz and FIR bandpass filtered at 0.01–40 Hz and artifact rejection was performed with MARA plug-in and additional visual inspections. Cleaned EEG signals underwent a logistic infomax independent component (IC) analysis, resulting in twenty-some ICs per subject (note, the number of ICs is not equal across subjects due to the variability across subjects). The time course was then convolved with a hemodynamic function and down-sampled to 0.5Hz.

D. Decoding with oFVSD.

We used the FVS algorithm [5] to identify a small number of features (EEG ICs). At each learning step, the FVS selects a feature that provides the largest increase in predictive performance and the largest reduction of MSE and then adds it to the set of 'selected' features. The FVS algorithm continued until no feature improved performance or a maximum number of selected features had been reached. At the end of FVS iterations, feature identification of selected features. Hyperparameter optimization was applied for both cases *with* and *without* the FVS. Cross-validation was applied at each forward iteration. Given the selected features, samples were randomly split into a training dataset comprising the first 70% of the samples and a test dataset with the remaining 30%. Additional fivefold cross-validation was used on the training data to select an appropriate parameter configuration for a specific model to optimize the respective hyperparameters. At this step, mean squared error (MSE) was used to determine the best-performing hyperparameters for each approach on the training dataset. Moreover, the optimization algorithm was implemented to identify the best hyperparameter values of the ML model. Depending on the number of hyperparameters of each model, either an 'exhaustive grid search' that considers all parameter combinations or the 'random search' that randomly selects a given number of values from a parameter space was considered for parameter optimization in the cross-validation step [8]. MSE and Spearman correlation coefficients between the true and predicted values were used as a criterion for decoding performance [5] so that the outcome can be comparable across nonparametric regression models.

A computer-friendly procedure was proposed for dealing with very high dimensional microbiome data [9]. This is true for neuroimaging data [5]. The number of these subsets (or comparisons) is mostly larger than the number of processors in a single computer system. Parallel computing was applied as implemented in the oFVSD. For each forward iteration, the input variables were separated randomly into a number of subsets. The computation time for each model was assessed.

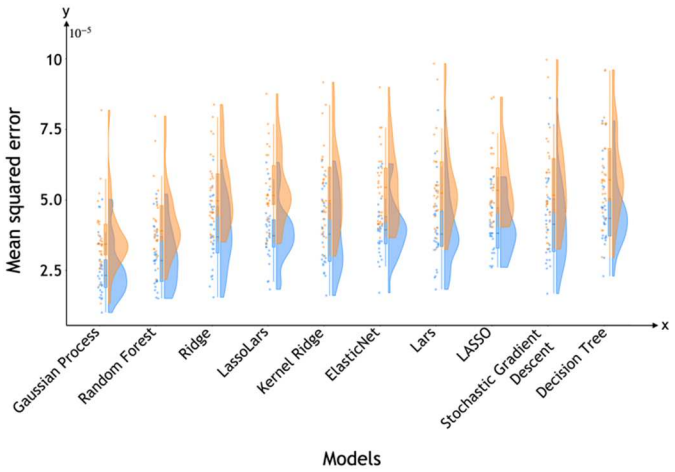


Fig. 2. Violine plots of mean squared errors of regression models *with* forward variable selection (FVS: blue color) and *without* FVS (orange color). The time-courses of the saliency network from fMRI were decoded by that of EEG ICs in a concurrent EEG-fMRI recording at resting-state. Entries are sorted in order of ascending mean squared error (MSE) values.

III. Results

A. Decoding Performance

We systematically evaluated regressor performances using the EEG and fMRI databases. As for the regression model, we predicted the time course of FC obtained from fMRI as a function of selected ICAs from the concurrently recorded EEG signals. Fig. 2 shows the MSE values for each ML model *with* and *without* the FVS algorithm to predict the time course of the

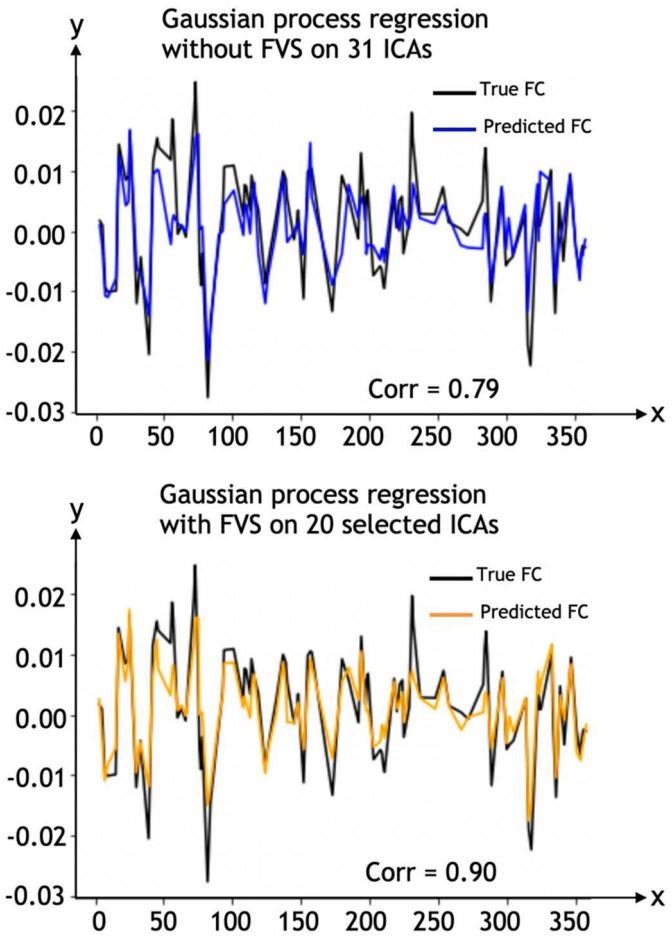


Fig. 3. Activity amplitudes of the saliency network as a function of time (in seconds) for true time-course from fMRI (in black color) and for predicted time-course decoded from EEG signals from a subject. A case of Gaussian process regression without FVS algorithm (in blue color) with all 31 EEG ICAs (overall MSE = 3.37 x 10^{-5}, Spearman ρ ('Corr') = .79, p = .058), and that with the FVS algorithm (in orange) with a subset of EEG ICAs selected with the FVS algorithm (MSE = 2.08 x 10^{-5}, Spearman ρ = .90, p < .0001).

FC from all subjects. Across all ten models, the use of the FVS algorithm strongly outperformed the conventional ML models *without* the FVS algorithm (repeated-measure ANOVA: $F_{1,28}$ = 132.517; p < .0001; η_p^2 = .826). The ML models also showed a strong effect ($F_{9,127.76}$ = 40.059; p < .001; η_p^2 = .589, corrected by Greenhouse-Geisser). Across all models, 'Gaussian Process' regression was among the best (an average MSE = 3.63 x 10^{-5}; p < .001, Bonferroni corrected) followed by 'Random Forest' (p = .002). There were little differences across models (p > .05, not significant) except that 'Decision Tree' was the worst than almost all models (not significant against 'Stochastic Gradient Descent').

Fig. 3 depicts a single case example of the time course of the saliency network activity obtained from fMRI and its predicted time course from EEG signals for both *without* and *with* FVS algorithms decoded with the Gaussian process regression. For a subject, among all model comparisons, 20 out of 31 EEG ICs were identified using the FVS-supported Gaussian process regression model.

B. Computational Cost.

Notably, it is essential to consider the computational cost (see Table 1 for the computational cost of each model and algorithm). In terms of decoding accuracy, 'Random Forest (second best)' models (MSE = 4.19) without variable selection comparatively were as good as the Gaussian Process model (MSE = 3.33, see Table 1 and Fig. 2). However, the computational costs of the Random Forest model (CPU time = 80 seconds without FVS) model were more expensive because the parameters were more complex than those of the Gaussian Process model (CPU time = 37 seconds without FVS).

Table 1 Computational time (CPU time) of the ML models to predict age. FVS: forward variable selection algorithm; Entries are sorted in order of ascending mean squared error (MSE) values.

Model	MSE (10^{-5})		CPU time (secs)	
	with FVS	without FVS	with FVS	without FVS
Gaussian Process	2.19	3.33	582	37
Random Forest	2.96	4.19	988	80
Ridge	3.80	5.34	255	5
LassoLars	3.84	5.38	376	6
KernelRidge	3.90	5.38	609	21
ElasticNet	3.97	5.45	393	22
LASSO	4.06	5.53	261	5
Lars	4.07	5.51	271	6
Stochastic Gradient Descent	4.27	5.58	777	42
DecisionTree	4.66	6.00	626	38

Additionally, the ridge regression model achieved a fairly comparable performance (MSE = 5.34) to the Random Forest (MSE = 4.19), while the computational time of ridge regression was considerably shorter (CPU time = 5 seconds without FVS) than that of Random Forest model (CPU time = 80 seconds without FVS). Conversely, the stochastic gradient descent and

decision tree models had poor performances with low values (MSE = 5.58 and MSE = 6.00, respectively).

C. Application for BCI

The decoded model can then be applied for a BCI as an estimate of the target FC time course for monitoring or neurofeedback purposes. Fig. 4 depicts a sample decoded model using the Gaussian Process. The resultant algorithm can be implemented using IC-based BCI [10]. While practicality, such as computational time, needs to be assessed for real-time BCI, this model can be surely applicable for offline assessment without using fMRI outside the laboratory setting. While we demonstrated the fusion of EEG-fMRI at the individual level in this study, future studies could elaborate on building a standardized or generalized model for practical solutions.

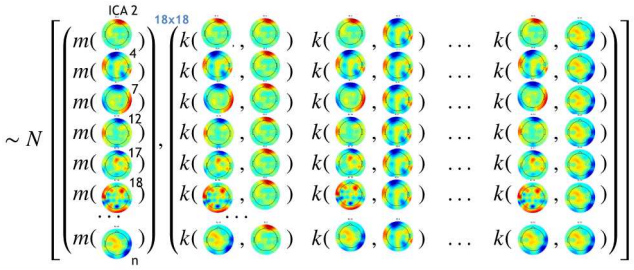


Fig. 4. A schematic sample case of decoded Gaussian Process model for a FC. Out of 32 EEG-ICs, 18 ICs were selected for this case. These model can be implemented for IC-based BCI model.

IV. Conclusions

The proposed approach presents the following advantages compared with the previous methods. Firstly, neuroscientists could benefit from the automatic selection of the most appropriate (or best-performing) model for their own data sets. In our proposed approach, 1) the benefit of feature selection was apparent; 2) Gaussian process regression appeared to perform better than the other ML models to link the EEG to fMRI FC signals; and 3) a unified BCI model has been proposed (see Fig. 1 top right panel). Further interpretation is needed to fully understand why we observed such decoding patterns and why there is a large variability in decoding performance across individuals. Although we applied parallel computational techniques for overcoming large-scale problems in the FVS algorithm, a high-performance machine is still necessary for completing our proposed method.

Our toolbox [5] can identify a minimal-size core set of features that can provide insights into the neurophysiological link between signals from fMRI and EEG. Having shared the toolbox in the Python environment, we hope our package, originally built for genetic analysis, would also be a great tool for studies on the brain-computer interface and neuroimaging.

References

[1] G. D. C. Cavalcanti, L. S. Oliveira, T. J. M. Moura, and G. V. Carvalho, "Combining diversity measures for ensemble pruning," *Pattern Recognit. Lett.*, vol. 74, pp. 38–45, Apr. 2016, doi: 10.1016/j.patrec.2016.01.029.

[2] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, no. 1–2, pp. 239–263, May 2002, doi: 10.1016/S0004-3702(02)00190-X.

[3] J. Wirsich, A.-L. Giraud, and S. Sadaghiani, "Concurrent EEG- and fMRI-derived functional connectomes exhibit linked dynamics," *Neuroimage*, p. 116998, 2020, doi: 10.1016/j.neuroimage.2020.116998.

[4] A. T. Winder, C. Echagarruga, Q. Zhang, and P. J. Drew, "Weak correlations between hemodynamic signals and ongoing neural activity during the resting state," *Nature Neuroscience*, vol. 20, no. 12, pp. 1761–1769, Dec. 2017, doi: 10.1038/s41593-017-0007-y.

[5] T. Dang, A. S. R. Fermin, and M. G. Machizawa, "oFVSD: A Python package of optimized forward variable selection decoder for high-dimensional neuroimaging data," Neuroscience, preprint, Dec. 2022. doi: 10.1101/2022.12.25.521906.

[6] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, 2012.

[7] S. Whitfield-Gabrieli and A. Nieto-Castanon, "Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks," *Brain Connect.*, vol. 2, no. 3, pp. 125–141, 2012, doi: 10.1089/brain.2012.0073.

[8] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and*

prediction, 2nd ed. in Springer series in statistics. New York, NY: Springer, 2009.

[9] T. Dang and H. Kishino, "Forward Variable Selection Improves the Power of Random Forest for High-Dimensional Micro Biome Data," *J. Cancer Sci. Clin. Ther.*, vol. 06, no. 01, 2022, doi: 10.26502/jcsct.5079147.

[10] M. G. Machizawa *et al.*, "Quantification of anticipation of excitement with a three-axial model of emotion with EEG," *J Neural Eng*, vol. 17, no. 3, p. 036011, 2020, doi: 10.1088/1741-2552/ab93b4.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multimodal recognition of speech and electrocorticogram

Mitali Ahuja*, Shuji Komeiji†, Takumi Mitsuhashi‡, Yasushi Iimura‡,
Hiroharu Suzuki‡, Hidenori Sugano‡, Koichi Shinoda* and Toshihisa Tanaka†
* Department of Computer Science, Tokyo Institute of Technology
† Department of Electronic and Information Engineering, Tokyo University of Agriculture and Technology
‡ Department of Neurosurgery, Juntendo University School of Medicine

***Abstract*—Brain-Computer Interface (BCI) provides a novel way of communicating with computers but their poor performance in terms of speed and accuracy in comparison to the other modes of communication has been the biggest obstacle in their usage for practical applications. In this work, we aim to enhance the performance of BCI by utilizing speech data collected along with the electrocorticogram (ECoG) recordings when the person is speaking. While some BCI users may have difficulty in speaking at all, many of them can speak, even though their speech may be unclear. We propose that information from this speech data can help in improving accuracy and to employ such speech context to assist the decoding process, we apply a multimodal recognition method. We use speech data contaminated by noise in our evaluation to simulate the cases where the available speech quality is low. Our experiments using data from five subjects suffering from Epilepsy show that our method of using multimodal input has a significant improvement, an absolute reduction in phrase error rate by 1.1 points from recognition using speech alone and by 51.3 points from recognition using ECoG alone.**

I. Introduction

Brain-Computer Interfaces (BCIs) allow humans to communicate with computers directly through their brains by translating their brain activity into computer-readable and implementable formats. Their application areas are rehabilitation of disabled people along with non-medical fields such as Gaming, Internet of Things (IoT), and autonomous driving [1]–[3].

BCIs are not yet widely adopted in our society because of their sub-par performance due to challenges such as the trade-off between quantity and quality of brain data that can be recorded, complex and non-stationary brain signals, and difficult feature extraction process [1]. Past studies achieved acceptable results for several tasks such as controlling a cursor or prosthesis by translating the brain activity produced while imagining motor movements [4], [5]. However, such BCIs are still limited in terms of their applicability.

Lately, the focus has been on the decoding of speech from the brain as it further extends the scope of usability. Amongst different methods of brain data acquisition, Electrocorticography (ECoG), which measures the electrical activity of the brain from sensors placed directly on the cortical surface, has been frequently used due to its high temporal and spatial resolution which is required for fast-paced continuous speech recognition [6]. However, its invasive setup restricts its usage only under strict medical requirements resulting in the availability of limited amounts of data.

Most speech-based BCI studies perform spoken speech recognition as the first step towards implementing BCIs based on imagined speech since it is easy to monitor changes in brain activity when corresponding speech data is available. Ref. [7] was the first to decode continuously spoken speech from ECoG using the Gaussian mixture model.

Recently, advancements in the field of Automatic Speech Recognition (ASR) are being employed to BCIs even though the ECoG datasets are fairly smaller in size compared to frequently used speech datasets. For example, [8] used a Bi-directional Long Short Term Memory (BLSTM) based sequence-to-sequence encoder-decoder architecture to translate brain signals into sentences. Their approach was based on multi-task learning which required training the encoder with features obtained from the simultaneously produced speech and they achieved Word Error Rate (WER) as low as 3% for a subject. Ref. [9] extended this approach by switching the encoder architecture from BLSTM to Transformer as Transformers are better at learning long-range dependencies thereby achieving improved performance. Ref. [10] proposed a novel Brain2Char architecture that employed a Connectionist Temporal Classification (CTC) Decoder along with an external language model to avoid dealing with the alignment between speech and brain and they also achieved low WER of around 10% for a much larger vocabulary.

While these studies achieved good performance, their accuracy degrades with less training data or if the electrode arrays used to measure brain activity are less dense. To mitigate this problem, they assist their model with simultaneously recorded speech data by guiding their network with speech features so that it learns to make better predictions, but the improvement in performance is still not as significant. Another limitation is that none of them consider the scenario where clean spoken speech data is not available such as where the subject has difficulty in speaking.

Taking inspiration from studies employing multimodal speech recognition [11], we propose to utilize the speech information by using it as input along with ECoG. Different from previous works where speech features were used as a target for intermediate layers, we combine information from

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

speech and ECoG features in a multimodal manner through a late fusion approach so that we can utilize the speech context efficiently. Additionally, to simulate the case when available speech quality is low, we propose to use noise-added speech in our experiments. Our aim with this approach is to show that if speech modality is available, irrespective of the quality of speech, it can be employed in BCIs to improve their performance.

II. Method

A. Dataset

Five subjects (js4, js5, js6, js7, and js8) undergoing brain monitoring through the intracranial placement of ECoG arrays for treatment of epilepsy at Juntendo University Hospital volunteered to be a part of this study. A written, informed consent approved by the Ethics Committee of Tokyo University of Agriculture and Technology, Tokyo Institute of Technology, and Juntendo University Hospital was obtained beforehand. The details about this dataset are summarised below but more information can be referred to from [9].

During the experiment, the subjects were asked to speak out loud the sentence being displayed on a screen in front of them while their corresponding speech and ECoG data were collected. All subjects were conscious during the experiments and had no trouble speaking.

The sentences used for the experiments consisted of 8 different Japanese language sentences. Each sentence is made up of three phrases. The first phrase is either 「わたしは」(I) or 「きみと」(with you), second is 「がっこうへ」(to school) or 「しょくばに」(to office) and the last one is 「いった」(went) or 「むかう」(go) .

Speaking one sentence constituted one trial, and 80 trials were recorded for each subject. The speech and ECoG data were digitized at a sampling frequency of 9,600 Hz. The number and placement of electrodes varied between subjects because it depended on their medical condition, so we adopted a subject-dependent approach by performing experiments for each subject separately.

B. Preprocessing

Simultaneously recorded speech and ECoG data for each subject were trimmed to a length equal to the maximum duration of the sentence spoken by that subject so that the model does not learn to predict the output based only on the length of the sentence.

To be inclusive of conditions where clean speech is not available or feasible, we added Signal-to-Noise Ratio (SNR) equal to -10 white noise to our raw speech data. Since it's extremely difficult to collect data from speech-disabled subjects with implanted ECoG arrays who also volunteer to participate in research, preparing such data by manually adding noise helped in simulating conditions close to the real-world scenario.

C. Network Architecture

Taking inspiration from recent works which showed end-to-end architectures can decode speech from ECoG efficiently [8]–[10], we decided on employing a similar model consisting of an encoder and a decoder with CTC and a Recurrent Neural Network (RNN) based language model for our task.

As shown in Fig. 1, our network consists of the following components:

1) **Encoder:** It contains stacked bi-directional LSTM layers which take input features $\mathbf{X}$ of dimension $[T, f]$ where T and f represent time frames and features respectively and outputs a hidden vector $\mathbf{h}$ of dimension $[T', H]$, where T' is subsampled time frames, and H is the number of projection units in the final layer of the encoder.
2) **Decoder:** Our decoder is composed of CTC block and an RNN-based language model.
   CTC: It consists of LSTM layers that take the hidden vector from the encoder as input and output probabilities over each token c_{t} in the dictionary for each time frame. The output probability for tokens in the dictionary is given as
   $$p(c_t|X) = \mathrm{Softmax}(\mathrm{LinB}(\mathbf{h}_t)), \tag{1}$$
   where LinB and $\mathrm{Softmax}$ denote the linear layer and the softmax function, respectively.
   RNN language model: To assist CTC, we use an external RNN-based model that calculates the probability of a sequence of tokens as follows:
   $$p(c_1, c_2, \ldots, c_n) = p(c_1)p(c_2|c_1)\ldots p(c_n|c_{1:n-1}) \tag{2}$$

During training, the encoder and CTC are trained to maximize the model's probability of predicting correct output by minimizing the corresponding CTC loss. RNN language model is trained separately using the text corpus.

During inference, the hidden vector from the trained encoder is fed to the decoder. For each time step t, the probabilities from the CTC block and RNN language model are fused, and the token with the maximum probability is selected. The obtained output for the current time step is then fed to the language model to predict the probabilities for the next token. Finally, when all time frames are exhausted, or <EOS> label is predicted, beam search decoding is performed to get the final predicted output.

D. Decoding Approach

To test our proposed idea of using speech features as input along with ECoG will help in making the model learn better from speech modality, we performed speech recognition experiments in the following three steps. First, we used only noisy speech features as input and employed the architecture shown in Fig. 1 while discarding the ECoG branch to perform speech recognition. Next, we used only ECoG features as input while dropping the speech branch and performed the same experiment again. Finally, using noisy speech and ECoG features

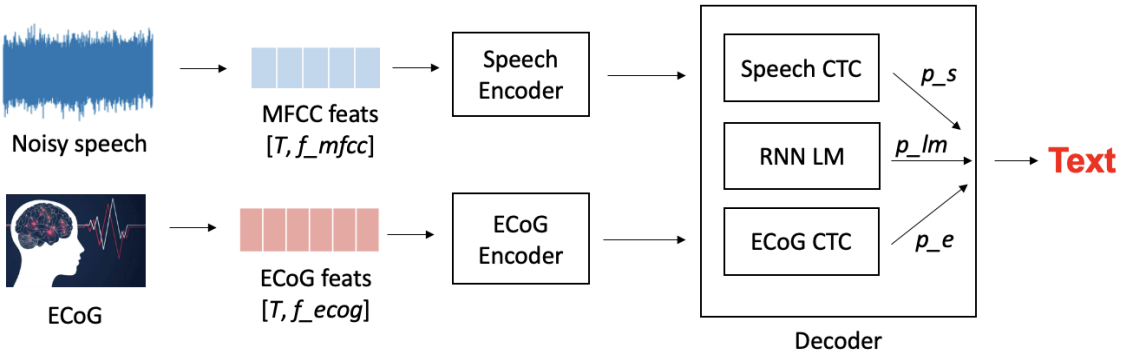


Fig. 1. Network architecture for multimodal recognition of speech and ECoG

together as input, we performed multimodal recognition and compared performance under these three scenarios.

For combining information from speech and ECoG, we used a late fusion method. We first retrieved the models trained on speech and ECoG modality individually. Then, in the decoder, we used their respective CTC scores and combined them with the language model score as follows:

$$\hat{c}_{\mathrm{t}} = \underset{c_{\mathrm{t}}}{\operatorname{argmax}} \{ w_{\mathrm{s}} \times \log p_{\mathrm{ctc}}(c_{\mathrm{t}}|X_{\mathrm{s}}) + w_{\mathrm{e}} \times \log p_{\mathrm{ctc}}(c_{\mathrm{t}}|X_{\mathrm{e}}) + w_{\mathrm{lm}} \times \log p_{\mathrm{lm}}(c_{\mathrm{t}}) \} \tag{3}$$

to get the most likely set of tokens for each frame. In (3), w_{s} and w_{e} are the weights assigned to the speech and ECoG CTC blocks, respectively, and w_{lm} is the weight for the language model. We believe that fusing information this way should prevent the model from becoming biased towards any particular modality while helping it in learning from each of these components effectively.

E. Feature Extraction from Speech

We extracted Mel Frequency Cepstral Coefficients (MFCCs) as acoustic features from speech data as they have been successfully used in many ASR as well as speech-based BCI-related studies [8]–[11]. For the calculation of MFCCs, a window of 20 msec with a frame period of 5 msec was used. We obtained MFCC features of dimension $[T, f_{\mathrm{mfcc}}]$ where T is the number of frames and f_{mfcc} is the dimension for features which was set to 13, similar to previous studies [8], [9]. The values for window length and frame period were set so that the output dimension of MFCC features across the time axis coincides with that of ECoG features.

F. Feature Extraction from ECoG

Our feature extraction process from ECoG has been inspired by past studies [8], [9] as they achieved state-of-the-art results in decoding speech from neural signals.

We began with removing channels that showed obvious signs of contamination by noise resulting in 54 channels for subjects js4, js5, js6 and js8 and 53 channels for subject js7. Next, a lowpass filter at 200 Hz followed by downsampling at 400 Hz was performed. A notch filter was applied at 50 Hz and 100 Hz to attenuate power line noise. To extract brain activity in the high gamma range (70–150 Hz), which is shown to be associated with speech [8], we used eight bandpass filters with centers between 70 and 150 Hz and averaged them across bands. ECoG data was further downsampled at 200 Hz and normalized using the z-score method to avoid bias due to the presence of any outliers.

The resultant ECoG features had dimension $[T, f_{\mathrm{ecog}}]$ where T is the number of time frames which is the same as that obtained for MFCC features and f_{ecog} is equal to the number of channels or electrodes used for that subject.

G. Implementation

We performed our experiments using ESPnet [12], an end-to-end speech processing toolkit. Our speech recognition model consisted of an encoder with four BLSTM layers; each layer had 320 units followed by a projection layer with 320 units. The subsampling factor for the encoder was set to 1/8 to reduce computational complexity. The CTC block consisted of one LSTM layer with 300 units. For the RNN language model, we trained a phrase-level model using one BLSTM layer with 1,000 units.

Since the focus of our work was to test how speech and ECoG features impact the recognition performance, we kept the language model weight constant at 0.5 across all types of input while the remaining 0.5 came from the weights for speech and/or ECoG CTC blocks. For recognition using only noisy speech as input, $w_{\mathrm{s}} = 0.5$ and $w_{\mathrm{e}} = 0.0$, for ECoG only input, $w_{\mathrm{e}} = 0.5$, and $w_{\mathrm{s}} = 0.0$, and lastly, for the case of multimodal features, w_{lm} was kept fixed at 0.5, while we tried different combinations for w_{s} and w_{e}.

H. Evaluation

As discussed in Section II-A, we performed speaker-dependent experiments, which restricted us from pooling data across subjects resulting in around five minutes of data per subject. To mitigate the effects of data insufficiency, we performed a five-fold cross-validation where for each fold, we split our data (80 trials) into the train (56 trials), validation (eight trials) and test set (16 trials).

TABLE I
AVERAGE PER(%) FOR SPEECH RECOGNITION USING DIFFERENT INPUT FEATURES

Features	PER(%)
Speech	6.7
ECoG	56.9
Multimodal	**5.6**

We evaluated the performance of our model in terms of Phrase Error Rate (PER), which is given as

$$\text{PER} = \frac{S + I + D}{N}, \tag{4}$$

where S, I, and D are, respectively, the minimum number of substitutions, insertions, and deletions required to convert the predicted output into ground truth, and N is the total number of phrases in the actual output. Since this is an error rate, a lower PER is desirable.

III. RESULTS

The average PERs obtained across five subjects for speech recognition using different features as input (Speech, ECoG, and Multimodal) have been presented in Table I. Here, the results presented for multimodal features are when using the following combination of weights, $w_{\text{lm}} = 0.5$, $w_{\text{s}} = 0.3$ and $w_{\text{e}} = 0.2$.

As can be verified from Table I, recognizing speech using multimodal features has the lowest PER. Our late fusion method significantly reduced PER by 51.3 points from the case of using only ECoG features and by 1.1 points from the case of using only speech features. This proves that through proper utilization of information from speech, we can majorly enhance the accuracy of BCI systems. In particular, PER obtained using multimodal features was even lower than that obtained for using only speech features, which has been the goal of this research, i.e., to make the performance of speech recognition from ECoG comparable to that of ASR systems to make it possible to employ them for different applications.

We also tested our method for different combinations of Speech and ECoG CTC weights while keeping the language model weight fixed at 0.5, and the results have been presented in Fig. 2. When $w_{\text{e}} = 0.0$, w_{s} becomes 0.5, which correponds to the case of using only speech as input, whereas when $w_{\text{e}} = 0.5$, w_{s} becomes 0.0 corresponding to the case of using only ECoG features as input. For all subjects, multimodal features have the lowest PER and, therefore, better performance than using only speech or only ECoG features for different combinations of speech and ECoG CTC weights, thereby proving the effectiveness of our method.

IV. DISCUSSION

Our proposed method was able to achieve competing PERs for all subjects even with less training data and a lesser number of electrodes as compared to previous studies [8], [10]. Usually, it is difficult to directly compare these error rates because of the large gap in the experiment conditions,

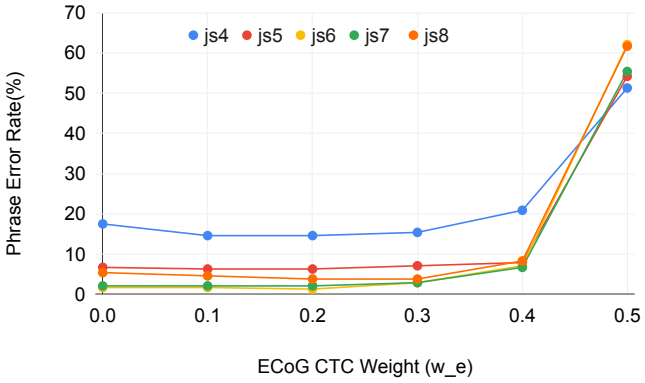


Fig. 2. PER(%) for different fusion weights for speech and ECoG

TABLE II
COMPARISON OF PER(%) FOR SPOKEN SPEECH RECOGNITION

Method	js4	js5	js6	js7	js8	Avg.
Komeiji [9]	18.7	35.8	33.1	30.9	41.9	32.1
Ours	**14.6**	**6.3**	**1.3**	**2.1**	**3.8**	**5.6**

such as vocabulary size, variability across subjects as well as the implementation architecture. However, just to evaluate the improvement in performance using our proposed approach, we compared our results with Komeiji et al. [9] since they have also used the same dataset and evaluation metric as ours. As presented in Table II, our method of using speech along with ECoG as input exceeds their performance where they used speech features to train the encoder by a large margin across all five subjects showing a promising application of our approach in future.

V. CONCLUSION

In this paper, we proposed multimodal recognition of ECoG for improved BCI communication. To our knowledge, we are the first to employ speech features as input and use noise-added speech data in our experiments. We achieved an average PER of 5.6% across subjects which outperforms the previous state-of-the-art results.

Currently, our method requires spoken speech data from subjects to perform multimodal recognition. However, we should also cater to scenarios where it is not available such as when the subject has completely lost their ability to speak or for decoding imaginary speech from ECoG. We would like to extend our method so that it does not rely on the necessity of having spoken speech data from the same subject during inference. We would also like to evaluate our method with non-invasive brain data acquisition techniques such as EEG or fMRI where we can explore a bigger dataset with many subjects and a larger vocabulary size in the future.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI 20H00235.

REFERENCES

[1] Abdulkader S N, Atia A, Mostafa MS M, "Brain computer interfacing: Applications and challenges", in *Egyptian Informatics Journal*, Volume 16, Issue 2, 2015, Pages 213-230, URL: http://www.sciencedirect.com/science/article/pii/S1110866515000237

[2] Ajmeria R et al., "A Critical Survey of EEG-based BCI Systems for Applications in Industrial Internet of Things," in *IEEE Communications Surveys Tutorials*, URL: https://ieeexplore.ieee.org/abstract/document/10000406

[3] Blankertz B, Tangermann M, Vidaurre C et al., "The Berlin Brain–Computer Interface: Non-Medical Uses of BCI Technology", in *Frontiers in Neuroscience*, Volume=4, 2010, URL: https://www.frontiersin.org/article/10.3389/fnins.2010.00198

[4] Pandarinath C et al., "High performance communication by people with paralysis using an intracortical brain–computer interface" in *eLife*, 6, e18554 (2017), URL: https://doi.org/10.7554/eLife.18554

[5] Attallah O, Abougharbia J, Tamazin M et al., "A BCI System Based on Motor Imagery for Assisting People with Motor Deficiencies in the Limbs" in *Brain Sciences*, 2020, 10, 864, URL: https://doi.org/10.3390/brainsci10110864

[6] Schalk G, Leuthardt E C, "Brain-Computer Interfaces Using Electrocorticographic Signals," in *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140-154, 2011, URL: https://ieeexplore.ieee.org/abstract/document/6047564

[7] Herff C, Heger D, De Pesters A, "Brain-to-text: Decoding spoken phrases from phone representations in the brain" in *Frontiers in Neuroscience*, Volume=9, Year=2015, Pages=217, URL: https://doi.org/10.3389/fnins.2015.00217

[8] Makin J G, Moses D A, Chang E F, "Machine translation of cortical activity to text with an encoder-decoder framework", *Nature Neuroscience*, 23, 575–582 (2020), URL: https://doi.org/10.1038/s41593-020-0608-8

[9] Komeiji S et al., "Transformer-Based Estimation of Spoken Sentences Using Electrocorticography", in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP) 2022*, pp. 1311-1315, URL: https://doi.org/10.1109/ICASSP43922.2022.9747443

[10] Sun P, Anumanchipalli G K and Chang E F, "Brain2Char: A Deep Architecture for Decoding Text from Brain Recordings", in *Journal of Neural Engineering*, 2019, volume=17, URL: https://doi.org/10.48550/arXiv.1909.01401

[11] Shruti Palaskar, Ramon Sanabria and Florian Metze, "End-to-end multimodal speech recognition", in *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 5774-5778, 2018, URL: https://doi.org/10.48550/arXiv.1804.09713

[12] Watanabe S, Takaaki H, Shigeki K, Tomoki H et al., "Espnet: End-to-end speech processing toolkit" in *arXiv e-prints*, 2018, URL: http://arxiv.org/abs/1804.00015

[13] Martin S, Brunner P, Iturrate I et al., "Word pair classification during imagined speech using direct brain recordings" in *Sci Rep* 6, 25803 (2016), URL: https://doi.org/10.1038/srep25803

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhancing Real-Time Semantic Segmentation with Textual Knowledge of Pre-Trained Vision-Language Model: A Lightweight Approach

Chia-Yi Lin*, Jun-Cheng Chen† and Ja-Ling Wu*
* National Taiwan University, Taiwan
E-mail: linchiayi@cmlab.csie.ntu.edu.tw
† Academia Sinica, Taiwan

***Abstract*—In this paper, we present a lightweight method for real-time semantic segmentation models by leveraging the power of pre-trained vision-language models. Our approach incorporates the CLIP text encoder, which provides rich semantic embeddings for text labels, and effectively distills its rich textual knowledge to the segmentation model. The proposed framework integrates the image and text embeddings, enabling visual and textual information alignment. Besides, we introduce learnable prompt embeddings to capture class-specific information and enhance the semantic understanding of the model. To ensure efficient learning, we devise a two-stage training procedure that allows the segmentation backbone to learn from fixed text embeddings in the first stage and optimize the prompt embeddings in the second stage. Extensive experiments and ablation studies demonstrate the effectiveness of our method in significantly improving the performance of the real-time semantic segmentation model.**

I. Introduction

Semantic segmentation is one of important computer vision tasks that aims to assign semantic labels to every pixel in an image, effectively dividing it into meaningful parts. The primary purpose is to enable machines to understand the content of an image at a granular level, allowing them to distinguish between different objects, regions, or parts in the image. Numerous real-world applications, such as autonomous driving and augmented reality, require high efficiency in inference speed, as fast interaction and response are often essential. Therefore, developing real-time semantic segmentation models has become a critical topic of interest in recent research.

Several models [1]–[5] have been proposed to meet the demand of low-latency algorithms for real-time semantic segmentation. Despite these advances, there are still challenges that existing methods cannot fully resolve, especially the trade-off between the inference speed and the accuracy of the model. In addition, designing more sophisticated architectures to further improve the feature representation of the models while keeping low latency at the same time is challenging to strike a balance.

To this end, we propose to leverage the power of pre-trained vision-language models [6], [7] to enhance the real-time semantic segmentation models. Specifically, our designed framework and training procedure aims to transfer the rich textual knowledge from the Contrastive Language-Image Pre-training (CLIP) [6] text encoder to the real-time semantic segmentation model STDC-Seg [4]. The complete framework is shown in Fig. 1. First, we integrate the image and text embeddings, facilitating the alignment of information from both visual and textual domains. By calculating the cosine similarity between these embeddings, we generate a score map that captures the semantic similarity between different image regions and text labels. The score map is then concatenated with the image feature maps, enabling the fusion of visual and textual information. To further enhance the semantic understanding of the model, we introduce class-specific learnable prompt embeddings, which are added to the label token embeddings. This augmented input is subsequently passed through the CLIP text encoder, resulting in enriched embeddings for each class with enhanced semantic meanings. By incorporating learnable prompts, the model gains the flexibility to optimize the prompt embeddings and capture class-specific discriminative information effectively.

Moreover, to ensure efficient learning, we devise a two-stage training procedure. In the first stage, we allow the segmentation backbone to learn from fixed text embeddings, enabling it to align image features with the text embeddings of the label from the CLIP text encoder. The purpose of this stage is to distill the backbone with the rich textual knowledge from the CLIP model into its feature representation. In the second stage, we focus on optimizing the learnable prompt embeddings while keeping the backbone fixed. This allows the model to refine the prompt embeddings to capture class-specific information, further enhancing the segmentation performance.

By integrating these components and training stages, our proposed framework significantly improves the performance of the real-time semantic segmentation model STDC-Seg. It is important to note that our method incurs only negligible additional latency during inference, making it a lightweight and practical solution for real-world applications.

Our main contributions are summarized as follows:

- We propose a method to enhance the real-time semantic segmentation model STDC-Seg by effectively transferring the CLIP textual knowledge. This results in significant

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

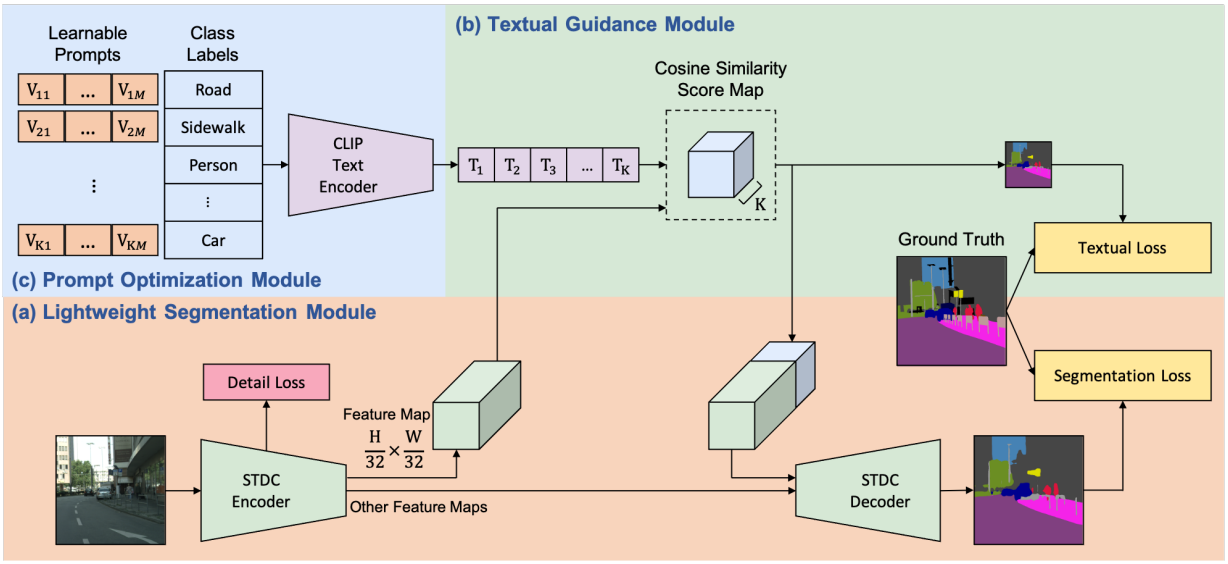


Fig. 1. Block Diagram and Information Flow of the Proposed Framework. (a) Lightweight Segmentation Module; (b) Textual Guidance Module; (c) Prompt Optimization Module.

performance improvements with minimal additional latency during inference.

- We introduce a two-stage training procedure that effectively trains both the lightweight segmentation backbone and class-specific learnable prompts, leading to enhanced performance in real-time semantic segmentation tasks.
- We conduct extensive experiments on two standard datasets for semantic segmentation, including ablation studies, to validate the effectiveness of our method and provide valuable insights.

II. Related Work

A. Real-Time Semantic Segmentation

Real-time semantic segmentation is an active research area that aims to enable efficient and accurate segmentation of images in real-time scenarios. BiSeNetV1 [1] proposed a compact and efficient dual-path network architecture to deal with spatial information and global context. DFANet [2] utilized multi-scale feature propagation and feature aggregation to balance computation and accuracy. BiSeNetV2 [3] further improved the performance of BiSeNetV1 with a refined architecture design. STDC-Seg [4] employed spatio-temporal decomposition and context aggregation mechanisms to enhance accuracy and efficiency simultaneously, and the proposed STDC backbone is proved to be more efficient and robust than previous methods in their experiments. Recently, PP-LiteSeg [5] redesigned the decoder of STDC-Seg, resulting in improvements in both speed and accuracy. This study focuses on STDC-Seg as the baseline method since it proposed the STDC backbone with state-of-the-art performance.

B. Pre-Trained Vision-Language Model

Vision-language pre-training [6]–[9] has become a hot research topic in recent years, as it enables models to learn joint representations of images and natural language. CLIP [6] and Align [7] are two of the most well-known methods, especially CLIP, which has gained much attention since its release. CLIP leverages a vision-language contrastive training process to learn a shared embedding space for images and texts, enabling zero-shot inference on various downstream tasks. CLIP's success is attributed to its ability to learn from diverse data sources and utilize contrastive learning, which encourages the model to learn discriminative representations for positive and negative pairs. In this study, we utilize the pre-trained CLIP text encoder as the textual knowledge source in our framework.

C. Prompt Tuning

Prompt tuning [10]–[12] is a method used to fine-tune pre-trained language models to improve the performance of language models on downstream tasks by adapting the model to the specific task at hand. Many methods have been proposed in recent years, including hand-craft-based prompt template generation [10] and automated prompt template generation [11] with sequence-to-sequence large language models like T5 [13]. Besides, there are also prompt tuning methods that directly optimize the prompt in the embedding space [12] instead of generating a prompt template in the text space. Especially in vision-language prompt tuning, CoOp [14] introduced learnable contexts to CLIP text encoders to improve performance on few-shot image classification. CLIP-Adapter [15] proposed to train the lightweight feature adapters attached to CLIP using a small amount of labeled data. These advancements in the prompt tuning for vision-language models have led to significant improvements in the ability of models to understand and generate natural language descriptions of visual content. In our study, since we understand that the prompt tuning method of CoOp best fits the real-time scenario, we employ it as part of our framework.

D. Semantic Segmentation with Pre-Trained Vision-Language Model

Pre-trained vision-language models have revolutionized the field of semantic segmentation, with several methods exploring their potential to enhance performance. Language-Driven Semantic Segmentation [16] combined the CLIP text encoder with a Vision Transformer-based [17] segmentation model and aligned the output feature maps of the decoder with text embeddings, allowing the segmentation model to achieve zero-shot semantic segmentation [18] capabilities. DenseCLIP [19] aligned image features from the output of the image encoder with text embeddings. It also utilized unified context optimization methods as in CoOp, and incorporated an additional Transformer [20] for visual context-aware prompt tuning, resulting in state-of-the-art performance across various semantic segmentation datasets.

Despite their impressive results, both methods do not acquire real-time capabilities due to their design. In our study, we specifically address the task of real-time semantic segmentation. We leverage the textual knowledge provided by the CLIP text encoder to enhance the accuracy of STDC-Seg while minimizing the impact on inference speed. By incorporating this textual knowledge, we aim to strike a balance between accuracy and inference speed in real-time semantic segmentation tasks.

III. Methodology

A. Model Architectures

This section presents a comprehensive description of our proposed framework for real-time semantic segmentation. Our framework consists of three main modules: the Lightweight Segmentation Module (LSM), Textual Guidance Module (TGM), and Prompt Optimization Module (POM). We aim to integrate these modules to enhance the performance of STDC-Seg, a real-time semantic segmentation model. This chapter will delve into each module, providing a comprehensive explanation, followed by our designed training procedure.

1) Lightweight Segmentation Module (LSM): Our Lightweight Segmentation Module (LSM) follows an encoder-decoder structure for semantic segmentation. Specifically, the encoder is responsible for extracting image features at different levels, ranging from shallow to deep layers. This results in a set of feature maps at various stages. On the other hand, the decoder integrates these feature maps to produce the final segmentation output, as shown in Fig. 1(a).

The typical decoding process starts by handling the high-level features, which are upsampled and combined with the low-level features using pixel-wise addition. This process is repeated from deep to shallow features, ensuring that the integrated features contain the necessary information for pixel-wise classification, which is essential for accurate segmentation. Let X represent the input image, and Y represent the ground-truth semantic mask. We aim to have the predicted mask $Y' = D(E(X))$ closely approximate Y, where E and D represent the encoder and the decoder, respectively. In this paper, we employ STDC-Seg as our LSM.

2) Textual Guidance Module (TGM): Inspired by DenseCLIP [19], the Textual Guidance Module (TGM) is designed to enhance the power of the STDC backbone in the LSM by leveraging textual knowledge from the CLIP text encoder. This module takes two inputs: the feature map from the last stage of the STDC backbone and the text embeddings of the label from the CLIP text encoder, as shown in Fig. 1(b).

Let $F^{(1)}, F^{(2)}, F^{(3)}, F^{(4)} = E(X)$ represent the output feature maps of the STDC backbone from stage one to stage four, with $F^{(4)} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times d}$, where H and W are the height and width of the input image, and d is the channel dimension. Additionally, let $L_1, L_2, ..., L_K$ denote the label texts corresponding to K classes (e.g., road, person, car). We pass each label text through the CLIP text encoder to obtain their respective embeddings: $T_1, T_2, ..., T_K \in \mathbb{R}^d$.

To transfer the textual knowledge from the CLIP text encoder to the STDC backbone, we compute the cosine similarity between each pixel of $F^{(4)}$ and the text embeddings $T_1, T_2, ..., T_K$. This results in a score map $S \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times K}$, which is computed as follows:

$$S_{i,j,k} = \frac{F^{(4)}_{i,j} \cdot T_k}{\|F^{(4)}_{i,j}\| \|T_k\|}, i = 1...\frac{H}{32}, j = 1...\frac{W}{32}, k = 1...K \quad (1)$$

The score map generated by the TGM can be treated as a predicted semantic map at a lower resolution. We can then supervise it by comparing it with the ground-truth semantic map as an auxiliary task. By computing the cosine similarity, we align the image features with the corresponding label text embeddings provided by the CLIP text encoder. This alignment allows the STDC backbone to incorporate the textual knowledge from the CLIP text encoder, thereby enhancing the representation power.

3) Prompt Optimization Module (POM): Inspired by CoOp [14], the Prompt Optimization Module (POM) aims at handling the class-specific learnable prompt embeddings, $V_{ij} \in \mathbb{R}^{d_t}$, $i = 1, 2, ...K$, $j = 1, 2, ..., M$, where K represents the number of classes, M represents the number of learnable embeddings for each class, and d_t is the dimension of the CLIP token embeddings. Additionally, we have $E_{ik} \in \mathbb{R}^{d_t}$, $i = 1, 2, ..., K$, $k = 1, 2, ..., l$, which represents the token embeddings of the class label text, where l denotes the sequence length of the token embeddings. The purpose of having learnable prompt embeddings is to have trainable text embeddings provided by the CLIP text encoder, thereby enhancing the discriminative qualities of the text embeddings associated with each class label, as shown in Fig. 1(c).

To obtain the trainable text embeddings, we concatenate $V_{i1}, V_{i2}, ..., V_{iM}$, and $E_{i1}, E_{i2}, ..., E_{il}$ for each class i, and feed this concatenated input to the CLIP text encoder, resulting in the output text embedding $T_i \in \mathbb{R}^d$, which can be expressed as:

$$T_i = POM([V_{i1}\ V_{i2}\ ...\ V_{iM}\ E_{i1}\ E_{i2}\ ...\ E_{il}]) \quad (2)$$

Throughout the training procedure, the parameters of the CLIP text encoder remain fixed. As a result, the trainable text embeddings T_i are always in the CLIP pre-trained domain, providing valuable textual knowledge. We rely on this property to enhance the performance of our model.

B. Two-Stage Training Procedure

In our framework, we design a two-stage training procedure to effectively transmit textual knowledge from the CLIP text encoder to the lightweight STDC backbone. In the following sections, we will first describe the loss functions employed during training and then delve into the separate training processes for each stage.

1) Loss Functions: The primary task of semantic segmentation is carried out by the LSM, which is responsible for generating the predicted semantic mask. To optimize this task, we employ the cross-entropy loss with Online Hard Example Mining (OHEM) [21], a commonly used technique in existing methods. The cross-entropy loss (CE) is calculated as follows:

$$CE(x, y) = \frac{1}{N} \sum_{n=1}^{N} -\log \left(\frac{\exp(x_{n,y_n})}{\sum_{k=1}^{K} \exp(x_{n,y_k})} \right) \quad (3)$$

Here, x represents the logit output by the model, y denotes the ground-truth class, and N is the total number of the pixels. OHEM enhances the optimization process by emphasizing those challenging examples contributing significantly to the loss.

In our framework, we refer to the loss of the primary task as L_{seg}. Additionally, as mentioned earlier, the TGM generates a score map that can be considered a low-resolution semantic mask, suitable for training as an auxiliary task using the same loss function as the main task. In our implementation, the low-resolution semantic mask is first bilinearly upsampled to match the size of the ground-truth semantic mask. Then, we compute the textual loss, denoted as L_{text}, between them using cross-entropy. The textual loss helps the method to leverage the rich semantic information provided by the pre-trained CLIP model. The alignment with the textual space of the CLIP model enforces a good regularization and enables us to learn an efficient segmentation model. Furthermore, following the approach of STDC-Seg, we also incorporate the detail loss, denoted as L_{detail}, into our training process. Finally, the total loss, denoted as L, is calculated as follows:

$$L = \lambda_1 * L_{seg} + \lambda_2 * L_{text} + \lambda_3 * L_{detail} \quad (4)$$

Here, we set λ_1, λ_2, and λ_3 all to 1.

2) First Training Stage: In the first stage, as depicted in Fig. 2, we do not use learnable prompts. Instead, we only input the label text's token embeddings into the CLIP text encoder, keeping the text embeddings fixed. This stage allows the STDC backbone to learn and align with the text embeddings provided by the CLIP text encoder. This alignment enables the STDC backbone to leverage textual knowledge as additional information for semantic segmentation.

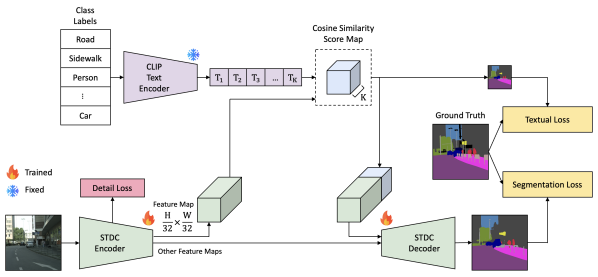


Fig. 2. The Setups of our First Training Stage.

3) Second Training Stage: In the second stage, as depicted in Fig. 3, we introduce the utilization of learnable prompt embeddings in the POM while keeping the parameters of the STDC backbone fixed. This choice is based on the understanding that in the first stage, the STDC backbone already assimilated textual knowledge from the CLIP text encoder, allowing it to adapt to changing text embeddings. Consequently, the POM focuses on optimizing the prompts to provide precise descriptions for each class label, thereby augmenting the discriminative characteristics of the optimized text embeddings for the STDC backbone.

Our framework achieves significant improvements after the two training stages compared to the original model. The LSM can simultaneously handle semantic segmentation tasks with visual and textual knowledge. The POM provides optimized text embeddings with discriminative information. The TGM connects these two modules and integrates the two embedding spaces, successfully transmitting textual knowledge from the CLIP text encoder to the STDC backbone.

C. Inference

After completing the two-stage training procedure, we can achieve real-time inference speed with minimal additional time compared to the original STDC-Seg model during the inference phase. This achievement is accomplished by retaining only the final trained text embeddings and discarding the learnable prompt embeddings and the CLIP text encoder, as illustrated in Fig. 4. By adopting this approach, the computational cost is limited to calculating cosine similarity in the TGM and including a few additional channels to the encoded feature maps. Consequently, the overall increase in Floating Point Operations (FLOPs) is negligible, enabling efficient real-time inference while preserving the accuracy and performance gains of our training methodology.

IV. Experiments

A. Experimental Settings

1) Datasets: To evaluate the performance of our proposed methods, we conduct experiments on two real-time semantic segmentation datasets, including the Cityscapes [22] dataset and the CamVid [23] dataset. The Cityscapes dataset is a large-scale dataset for urban scene understanding with pixel-level annotations of 30 different classes, where 19 of them are generally used for semantic segmentation tasks. There are

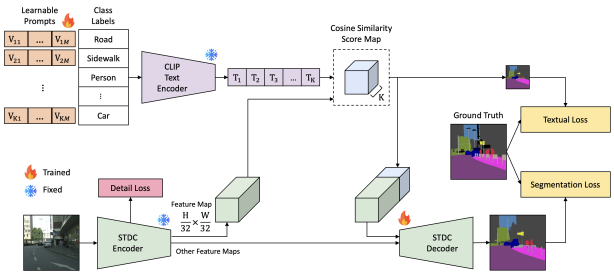


Fig. 3. The Setups of our Second Training Stage.

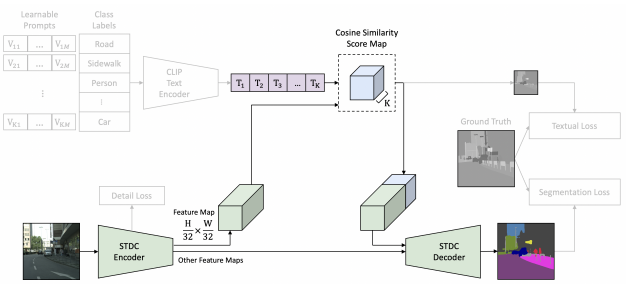


Fig. 4. The Colored Area Indicates the Inference Process of our Framework.

5,000 finely annotated images with 2048 x 1024 pixels in size, split into three subsets: training, validation, and testing, with 2,975, 500, and 1,525 images, respectively. Besides, the CamVid dataset, collected from vehicle-mounted cameras, has 11 semantic classes. It comprises 701 finely annotated images with 960 x 720 pixels in size, also split into three subsets: training, validation, and testing, with 367, 101, and 233 images, respectively. We train our models with a training set and evaluate them on a validation set of each dataset separately.

2) Implementation Details: In our experiments, we utilize stochastic gradient descent (SGD) [24] as our optimizer with a momentum of 0.9 and weight decay of 0.0005. For the Cityscapes dataset, the initial learning rate is set to 0.005 for the backbone and learnable prompts and 0.05 for other parameters. The batch size is set to 24, and we employ the "poly" learning rate policy, where the initial rate is multiplied by $(1 - \frac{\text{current_iteration}}{\text{max_iterations}})^{\gamma}$ with a total of 160,000 iterations and γ equals 0.9. Additionally, we use a linear warm-up strategy for the first 1,000 steps.

For the CamVid dataset, the initial learning rate is set to 0.01 for the backbone and learnable prompts and 0.1 for other parameters. The batch size is set to 16, and we also use the "poly" learning rate policy with the same parameters as before but with a total of 10,000 iterations. The warm-up strategy is applied for the first 200 steps.

We apply data augmentation techniques in both datasets, including color jittering, random horizontal flip, random crop, and random resize, to enhance the training data. For the Cityscapes dataset, we train the model using a scale range of [0.125, 1.5] stepping by 0.125, and the cropped resolution is set to 1024 x 512. On the other hand, for the CamVid dataset, the scale range is [0.5, 2.5] stepping by 0.25, and the cropped resolution is set to 960 $\times$ 720.

During training and evaluating performance, we use the mmsegmentation [25] library version 0.30.0 along with PyTorch [26] 1.12.1. The training is performed on a system equipped with two NVIDIA GTX 3090 GPUs, utilizing CUDA 11.3 and cuDNN 8.2. For testing the inference speed, we conduct experiments on a different setup. We modify the code provided by STDC-Seg [4] and use PyTorch 1.2.0, CUDA 10.0, and cuDNN 7.5. Additionally, we employ TensorRT 5.1.5.0 to optimize the inference process as STDC-Seg did. These experiments are carried out on an NVIDIA GTX 1080Ti GPU to benchmark the inference speed of our method.

B. Main Results

In our evaluation, we compare the performance of our proposed methods with the baseline STDC-Seg model on both the Cityscapes dataset and the CamVid dataset. By doing so, we aim to demonstrate the effectiveness of our approach in improving the performance of STDC-Seg using textual knowledge.

1) Results on Cityscapes: In Table I, we present our proposed method's segmentation accuracy and inference speed on the Cityscapes validation set. Notably, compared to the baseline STDC-Seg50 model, which achieves a validation mean Intersection over Union (mIoU) of 72.2, our proposed methods incorporating textual knowledge through a two-stage training procedure can enhance the performance to 72.9. This improvement comes with only a slight sacrifice in inference speed.

2) Results on CamVid: Table II presents the comparison results of our method and STDC-Seg on the CamVid dataset. We train the STDC-Seg model on the training set and evaluate its performance on the validation set. By applying our two-stage training procedure with text knowledge guidance, we are able to increase the mIoU of STDC-Seg from 75.3 to 78.0. This significant performance improvement further demonstrates the effectiveness of our method in enhancing segmentation accuracy by leveraging textual knowledge.

3) Qualitative Results: Here, we present qualitative results on the Cityscapes validation set, as depicted in Fig. 5. Specifically, our method achieves more precise segmentation of poles (colored in gray) compared to STDC-Seg. Poles are challenging to segment accurately due to their thin shape, which can often be overlooked by the model. However, our results demonstrate that the transfer of textual knowledge significantly enhances performance. In (a), (b), and (c), the segment masks of the poles exhibit significantly improved accuracy, addressing the issue of missing or incomplete segmentation in STDC-Seg. Furthermore, we observe improvements in segmenting fences (colored in Tuscan) as well. Fences are challenging to segment as they frequently overlap with other objects, such as buildings or walls. Our method demonstrates better results, suggesting that the transmitted textual knowledge aids in discriminating the fence from other objects. In (c) and (d), we can observe

TABLE I
QUANTITATIVE COMPARISONS OF THE BASELINE AND THE TWO STAGES OF OUR MODELS ON THE CITYSCAPES VALIDATION SET.

Model	#Params ↓	FLOPs ↓	mIoU (%) ↑	Inference Speed (FPS) ↑
STDC1-Seg50 (baseline)	8.57M	16.95G	72.2	250.4
Ours (stage 1)	8.60M	16.96G	72.7	245.3
Ours (stage 2)	8.60M	16.96G	**72.9**	245.3

TABLE II
QUANTITATIVE COMPARISONS OF THE BASELINE AND THE TWO STAGES OF OUR MODELS ON THE CAMVID VALIDATION SET

Model	#Params ↓	FLOPs ↓	mIoU (%) ↑	Inference Speed (FPS) ↑
STDC1-Seg (baseline)	8.57M	22.42G	75.3	197.6
Ours (stage 1)	8.59M	22.43G	76.6	196.8
Ours (stage 2)	8.59M	22.43G	**78.0**	196.8

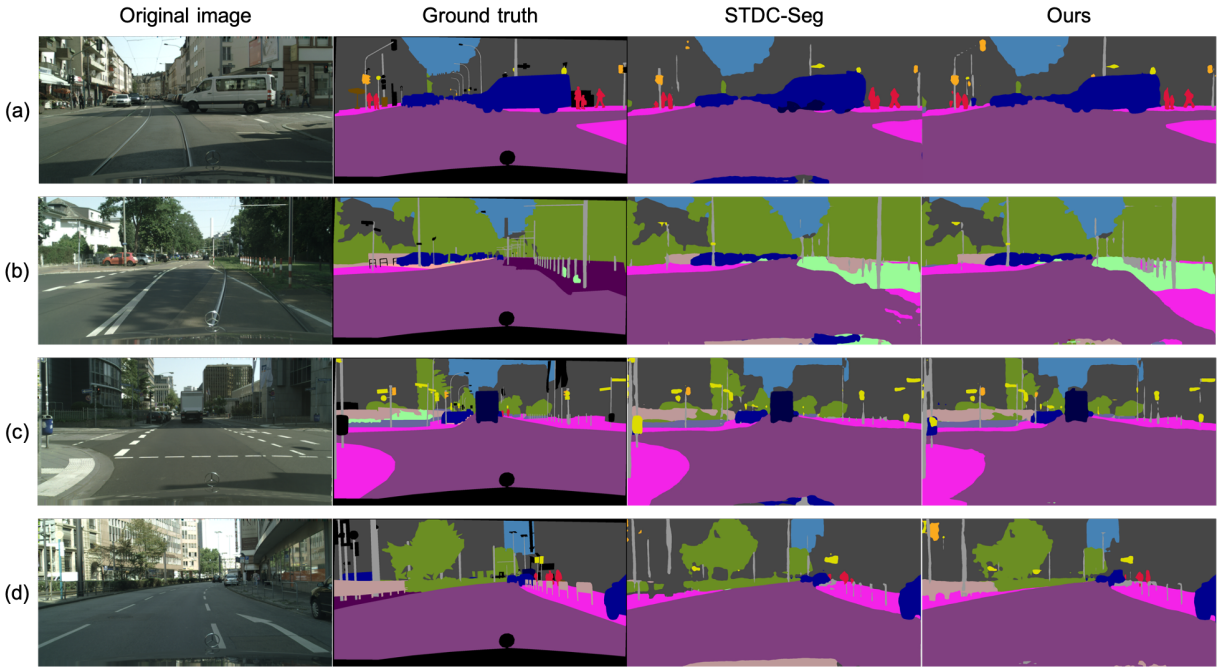


Fig. 5. Comparisons of Qualitative Results between the Proposed Method and the SOTA (i.e., the STDC-Seg). Note that the region colored in black is marked as "ignore" in the standard evaluation. Best viewed in color and zoom in.

substantial improvements in fence segmentation compared to STDC-Seg.

These qualitative results highlight the effectiveness of our approach in enhancing the segmentation of challenging objects such as poles and fences. By leveraging the textual knowledge transmitted to the model, we address the limitations of the base STDC-Seg method and achieve more accurate and detailed segmentations.

C. Ablation Studies

1) Effectiveness of Two-Stage Training Procedure: In the early stages of our experiments, we encountered challenges in directly transfer textual knowledge from the CLIP text encoder to the STDC backbone with learnable prompts. Training the learnable prompts and the STDC backbone at the same time results in worse performance than our two-stage settings. We believe this is because the STDC backbone lacks textual knowledge before our training since it is not pre-trained with a vision-language training procedure. Consequently, when trained simultaneously with learnable prompts, the text embeddings from the CLIP text encoder constantly change, making it difficult for the STDC backbone to keep up and leading to unstable performance.

To address this issue, we adopt a two-stage training procedure. In the first stage, we fix the text embeddings and allow the STDC backbone to align itself with the textual knowledge. In the second stage, we fix the STDC backbone and adjust the learnable prompts to optimize the "prompt sentence" from the embedding space.

As shown in Table III, on the Cityscapes dataset, training the model without the two-stage procedure yield an mIoU of 72.7, whereas, with the two-stage training, the mIoU is improved to

TABLE III
ACCURACY COMPARISON BETWEEN TRAINING PROCEDURES WITHOUT AND WITH MULTIPLE STAGES (CITYSCAPES VALIDATION SET).

Procedure	mIoU (%) ↑
end-to-end 1 stage	72.7
2 stages (1^{st} stage)	72.7
2 stages (2^{nd} stage)	**72.9**

TABLE IV
ACCURACY COMPARISON BETWEEN TRAINING PROCEDURES WITHOUT AND WITH MULTIPLE STAGES (CAMVID VALIDATION SET).

Procedure	mIoU (%) ↑
end-to-end 1 stage	76.4
2 stages (1^{st} stage)	76.6
2 stages (2^{nd} stage)	**78.0**

TABLE V
COMPARISONS OF THE EFFECTS OF DIFFERENT NUMBERS OF LEARNABLE PROMPT EMBEDDINGS ON THE CAMVID VALIDATION SET.

M	mIoU (%) ↑
8	76.3
11	**76.4**
14	75.3
17	76.0

TABLE VI
COMPARISONS OF THE EFFECTS OF WEIGHT DECAY IN LEARNABLE PROMPT EMBEDDINGS ON THE CAMVID VALIDATION SET.

Weight decay (M=8)	mIoU (%) ↑
0.0005	75.9
No weight decay	**76.3**

72.9. Similarly, as shown in Table IV, on the CamVid dataset, training without the two-stage procedure results in an mIoU of 76.4, while training with the two-stage procedure achieves an mIoU of 78.0. These results indicate a significant improvement in adopting of the two-stage training procedure.

Based on these findings, we strongly suggest that the two-stage training procedure is crucial for effectively transmitting textual knowledge from the CLIP text encoder to the STDC backbone. It allows for better alignment and optimization, ultimately improving the segmentation performance on both the Cityscapes and CamVid datasets.

2) Number of Learnable Prompt Embeddings: During the second stage of training, we conduct experiments to determine the optimal number of learnable prompt embeddings, denoted as M. We explore different values of M, specifically $M = 8,\ 11,\ 14,$ and 17, and analyze their impact on performance. As shown in Table V, we observe that the performance gain reaches a saturation point around $M = 11$. This suggests that increasing the number of embeddings beyond a certain threshold does not yield significant improvements in performance.

We conjecture that there is a limitation to the number of embeddings because excessively long prompts may contain unrelated or extraneous information for the task of classifying between labels in segmentation. Therefore, there is a balance to be struck between having enough prompt embeddings to capture relevant information and avoiding an excessive number that includes noise or irrelevant details.

3) Regularization of Prompt Embedding Space: We make an interesting observation regarding the regularization of the learnable prompt embeddings and their impact on performance. Specifically, in the CamVid dataset, we notice that applying an L2 weight decay regularization of 0.0005 to the prompt embeddings results in a significant drop in performance, with the mIoU decreasing from 76.3 to 75.9, as shown in Table VI. This trend is consistent across most of our experiments.

We believe the reason behind this phenomenon lies within the embedding space. If we imagine the embeddings space of all tokens as a normal distribution, with the centers representing frequently occurring words and the outer regions containing less frequent words, applying L2 weight decay regularization to the learnable prompt embeddings tends to push these embeddings towards the zero vector, which corresponds to the center of the distribution. Consequently, the prompts would predominantly consist of frequent words.

However, we argue that a powerful prompt should incorporate specific words that help the model distinguish between different labels. Relying solely on frequent words may not provide the necessary discriminative power. From a higher-level perspective, let's consider the well-known prompt "A photo of a [CLASS]." In this example, words like "a" and "of" are likely close to the distribution's center, representing frequent words. On the other hand, the word "photo" is more specific and likely to be further away from the center. Restricting the prompt embeddings to only frequent words might hinder the ability of the prompts to convey specific meanings.

Based on these observations, we suggest that: to equip the learnable prompts with the ability to capture the specific meanings from the embedding space, one has better not to apply L2 weight decay regularization on them. By doing so, we provide the prompts with the opportunity to reach beyond frequent words and incorporate more specific and contextually relevant terms, ultimately improving the performance of the segmentation model.

V. CONCLUSION

In conclusion, this research has contributed to the field of real-time semantic segmentation by leveraging the pre-trained vision-language model CLIP. The proposed framework enhances the STDC-Seg model by effectively incorporating textual knowledge from the CLIP text encoder. This integration enables the fusion of visual and textual information, leading to substantial performance improvements while maintaining minimal additional latency during inference. To further improve the model's semantic understanding and capture class-specific details, learnable prompt embeddings are introduced. The two-stage training procedure effectively trains both the lightweight

segmentation backbone and the learnable prompts, optimizing the model's performance in real-time semantic segmentation tasks.

Acknowledgment

The project is supported by National Science and Technology Council (NSTC), Taiwan (R.O.C) under grant numbers of 111-2634-F-002-022 and 111-2221-E-002-134-MY3.

References

[1] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[2] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9522–9531.

[3] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051–3068, 2021.

[4] M. Fan, S. Lai, J. Huang, *et al.*, "Rethinking bisenet for real-time semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9716–9725.

[5] J. Peng, Y. Liu, S. Tang, *et al.*, "Pp-liteseg: A superior real-time semantic segmentation model," *arXiv preprint arXiv:2204.02681*, 2022.

[6] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[7] C. Jia, Y. Yang, Y. Xia, *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 4904–4916.

[8] Y.-C. Chen, L. Li, L. Yu, *et al.*, "Uniter: Universal image-text representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, Springer, 2020, pp. 104–120.

[9] P. Zhang, X. Li, X. Hu, *et al.*, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[10] F. Petroni, T. Rocktäschel, P. Lewis, *et al.*, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[11] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[12] X. Liu, Y. Zheng, Z. Du, *et al.*, "Gpt understands, too," *arXiv preprint arXiv:2103.10385*, 2021.

[13] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[14] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[15] P. Gao, S. Geng, R. Zhang, *et al.*, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[16] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv:2201.03546*, 2022.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] Y. Rao, W. Zhao, G. Chen, *et al.*, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 082–18 091.

[20] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

[22] M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[23] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, Springer, 2008, pp. 44–57.

[24] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[25] M. Contributors, *MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark*, https://github.com/open-mmlab/mmsegmentation, 2020.

[26] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

EEG study on anticipation of difficulty for upcoming auditory task

Zichen Song*, Hiroshi Higashi† and Shin Ishii ‡
* Graduate School of Informatics, Kyoto University, Japan
E-mail: song.zichen.65v@st.kyoto-u.ac.jp
† Graduate School of Engineering, Osaka University, Japan
E-mail: higashi@comm.eng.osaka-u.ac.jp
‡ Graduate School of Informatics, Kyoto University, Japan
E-mail: ishii.shin.3u@kyoto-u.ac.jp Tel/Fax: +81-75-753-4908

***Abstract*—The intricate workings of the human brain, which utilizes acquired information from the environment to anticipate future challenges, still lack comprehensive understanding. To understand the mechanism, this study investigated the effects on electroencephalographic (EEG) signals of the process of anticipation of difficulty in auditory tasks. We designed a similarity problem in which human participants were first presented with two tones, followed by a questioning tone, and were asked to answer which of the first two tones was similar to the questioning tone. We analyzed EEG signals recorded during the presentation of the second of the first two tones, a period during which we assumed the participants were anticipating the upcoming problem. By manipulating the similarity of the first two tones in terms of two attributes (position and frequency), we eliminated other effects on the EEG signals caused by the tone difference and attributes, and extracted the brain activity related to the anticipation of the task difficulty. We found event-related potentials (P2, P3, and N4) and frequency bands (theta, alpha, and beta) in the prefrontal regions, alpha band in the EEG signals were significantly correlated with task difficulty. These results demonstrated the reliability of EEG to evaluate the anticipation of difficulty.**

I. Introduction

The human brain has the remarkable ability to anticipate and make anticipations about future events [1]. This capacity allows us to prepare and adapt our behavior based on our anticipations [2] and greatly impacts decision-making [3]. These anticipations can be related to various aspects, including the difficulty of future tasks or events. For example, if we are familiar with a particular task or activity, such as solving a complex math problem or playing a challenging musical piece, our brain can draw upon past experiences and knowledge to anticipate the level of difficulty we might encounter. When it comes to anticipating the difficulty of future events, our brain relies on a combination of past experiences, knowledge, and contextual cues. Through a process called mental simulation, our brain generates internal representations of possible future scenarios and evaluates their anticipatory difficulty based on existing information [4]. This anticipation allows us to allocate appropriate cognitive resources and prepare ourselves mentally and physically [5]. Understanding how anticipations are formed and modulated is a complex and intriguing area of research. Additionally, understanding the neural mechanisms and responses involved in anticipation could lead to numerous engineering applications. For instance, techniques that predict individual anticipatory responses to future events could aid in the development of alert systems for autonomous driving, or stress monitoring systems for mental health assessments.

Previous studies have primarily relied on behavioral measures to investigate these anticipations, such as reaction times and self-reported ratings [6]. However, these measures provide limited insight into the underlying neural processes that give rise to anticipations. Advancements in neuroimaging techniques, specifically electroencephalography (EEG), have revolutionized our ability to measure the electrical activity in the brain with remarkable precision at the millisecond level. This unprecedented level of temporal resolution offers a distinctive and unparalleled opportunity to delve into the realm of perception-based anticipations and explore the underlying neural processes. By analyzing the patterns of neural activity associated with anticipations processing, EEG can reveal neural markers of perceived anticipation. For instance, Simonet et al. used EEG to investigate the neural marks underlying movement anticipation and found an enhanced alpha band in response to movement anticipation [7]. Additionally, Kverage et al. employed EEG and examined oscillatory activity in the low-frequency bands in vision anticipation, revealing that low-frequency bands reflected vision anticipation [8]. Vanhollebeke et al. found that the anticipatory phase of stress and challenges is also associated with the alpha band [9]. These findings underscore the significance of oscillatory activity in mediating the integration of anticipation.

This research aims to investigate brain activities related to the anticipation for the difficulty of an upcoming auditory task. We designed an experiment where human participants were required to listen to two tones, followed by a questioning tone. They were then prompted to determine which of the first two tones was more similar to the question tone. The difficulty level of the question was manipulated by controlling the similarity of the initial two tones: if the tones were similar, the question was considered difficult, whereas if they were dissimilar, the question was relatively easy. Our hypothesis

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

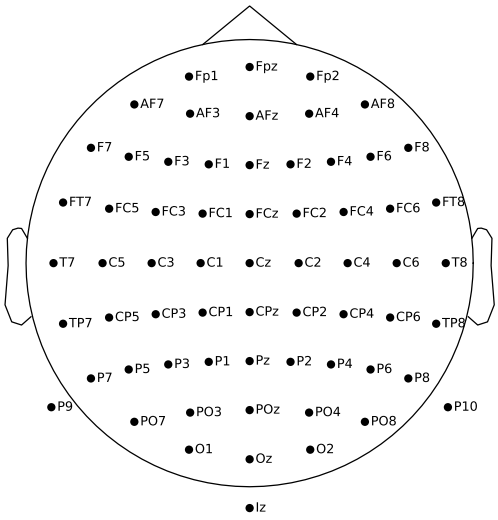


Fig. 1. The layout of the channels

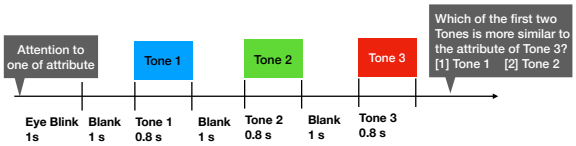


Fig. 2. The procedure of a single trial

posited that participants would anticipate the difficulty level of the upcoming problem while the first two tones were being presented.

To explore perception anticipation in auditory tasks, this research utilizes EEG to analyze event-related potentials and oscillatory activity. By capitalizing on the temporal precision of EEG, the study aims to uncover neural markers associated with anticipation in the face of partial information about upcoming problems i.e., the presentation period for the first two tones. Since EEG signals can be influenced by factors other than the anticipation of difficulty—such as the differences between the two tones—we employed two tone attributes, frequency and position. The participants evaluated the similarity between the tones in terms of either of these attributes. We then designed a normalized feature by combining the observed signals of the attributes to mitigate the impact of these other factors. By deciphering the neural makers underlying auditory perception and anticipation, we can advance our understanding of human cognition and develop innovative solutions that enhance safety, user experience, and mental well-being in various domains.

II. Experiments and Methods

A. Participants

All 26 participants (13 female; age range 22–28) had normal hearing and reported no psychological or neurological disorders. All studies were approved by the Committee for Human Research at Graduate School of Informatics, in Kyoto University. All participants signed a written informed consent prior to participating in their respective studies and received monetary reimbursement afterward.

B. Experimental environment

The participants were seated comfortably in front of an LCD screen (24-inch) on a chair in a dark room. On a table between the participants and the display, a keyboard was provided to input actions by pressing its keys. Auditory stimuli were sent using Psychopy3. The pure tones for the stimuli are spatial sounds that were generated based on the head-related transfer function [10] and presented through in-ear headphones (Etymotic ER3). The volume for stimulus sounds was tuned for each participant such that the sound did not make them uncomfortable and was kept constant during the experiment.

EEG signals were recorded using an ActiveTwo system (BioSemi) from 64 active scalp electrodes and digitized at 512 Hz. The 64 electrodes were positioned to cover the head as shown in Figure 1 according to the extended International 10/10 system. Additional signals were measured in the extra active electrodes placed on the left and right earlobes, on the left and right temples, and upper and lower sides of the left eye.

C. Task

The procedure of a trial is illustrated in Figure 2. At the beginning of each trial, the participants were provided with clear instructions regarding the specific attribute of the tones they needed to focus on. We utilized two attributes of the tones: their frequency and location. By integrating these attributes, we designed normalized EEG features to isolate brain activity specifically associated with anticipating difficulty (refer to Section II-G for details). Following the instructions, participants were given a 1.0-s duration to blink and relax their eyes. This brief period aimed to minimize blinking during the EEG recording phase. Then three sound stimuli (Tone 1, Tone 2, and Tone 3), which were pure tones with various positions and frequencies, were presented in sequence. Each sound stimulus lasted 0.8 s with a 1.0 s interval period. At the end of the trial, the participants were prompted to indicate which of Tone 1 and Tone 2 was more similar in the instructed attribute to Tone 3 and they answered within 3 s by the keyboard.

The experiment had two parts: the behavior and the EEG experiments. To account for variations in individual sensitivity to tone frequency and position, as well as the potential differences in perceived difficulty among participants for the same tones, we conducted a preliminary behavioral experiment prior to the

EEG study. This preliminary experiment aimed to fine-tune the difficulty level of the task for each participant, ensuring a tailored and optimized experience based on their specific abilities and perceptions. This difficulty level was subsequently tested in the EEG experiment.

In the behavior experiment, five similarity levels for each frequency/position attribute (see Section II-D) were tested. Each participant performed 10 trials for each similarity level of the attributes, 100 trials in total. Out of the five similarity levels for each attribute, the two similarity levels were assigned as the high and low difficulty levels for each participant based on the task performance in the behavior experiment. This assignment for the similarity level was performed such that the accuracy closer to 50% was assigned to the high difficulty level, and closer to 100% was assigned to the low difficulty level. In case there were two or more with 100% accuracy, the highest similarity level among them was assigned as the low difficulty level. For a special case where the participants performed the task perfectly for all similarity levels, the highest similarity level was assigned as the high difficulty level.

EEG signals were then recorded using these two difficulty levels. Two attributes were randomly selected out of the candidates (see Section II-D) to achieve the high and low difficulty levels, and then these two values for the attributes were randomly combined. Each participant performed 48 trials for each difficulty level of the attributes, for a total of 192 trials.

D. Auditory stimuli with similarity level

We manipulated the difficulty of the task by controlling the similarity between Tone 1 and Tone 2. To extract the neural activity related to the difficulty anticipation, we used two attributes, frequency and position, for controlling the similarity. This section describes the definition of the similarity levels for the attributes. Each attribute has five similarity levels. Level 1 represents the highest similarity and Level 5 is the lowest.

The first attribute is frequency, which defines the pitch of a tone [11]. In our experimental setting, 880 Hz (high A) was the lowest frequency. Based on the principle of equal temperament [12], we defined 24 frequencies:

$$\phi_n = 880 \times (\sqrt[24]{2})^n, \ (n = 0, 1, ..., 23). \tag{1}$$

The frequencies for Tone 1 and Tone 2, f_1 and f_2, were randomly selected out of these 24 candidates. Then the similarity level for f_1 and f_2 was calculated as

$$D_f = \left| \log_2 \left(\frac{f_1}{f_2} / 880 \right) * 24 \right|. \tag{2}$$

In order to set the frequency for Tone 3 (f_3), the higher and lower frequencies in f_1 and f_2, were randomly selected from $\{\max(f_1, f_2)/(\sqrt[96]{2})^n, \min(f_1, f_2) \cdot (\sqrt[96]{2})^n\}$. Table I shows some examples of the combinations of the tone frequencies and their similarity levels in the frequency attribute.

The second attribute was position, which defines the arrival direction of the tones. As shown in Figure 3, a semicircle surrounding a participant's head was defined, and 36 positions were set at $5°$ intervals from $-90°$ to $90°$ as:

$$\psi_n = -90 + 5n, (n = 0, ..., 35). \tag{3}$$

The positions for Tone 1 and Tone 2, p_1 and p_2, were randomly selected out of these 36 candidates. Then the similarity level for p_1 and p_2 was calculated as

$$D_p = |p_1 - p_2| / 30°. \tag{4}$$

In order to set the position for Tone 3 (p_3), the more rightward and leftward positions in p_1 and p_2, p_r and p_l, were randomly selected from $\{\max(p_1, p_2) - 5, \min(p_1, p_2) + 5\}$. Table II shows some examples of trials about the position attribute.

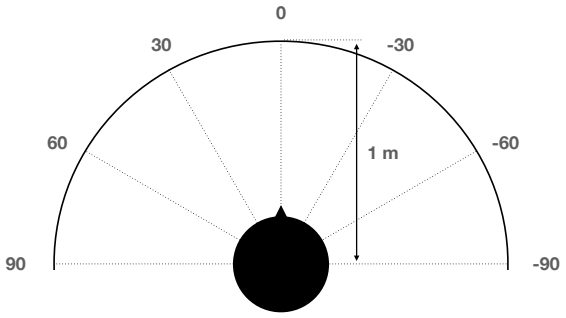


Fig. 3. The definition of the position attribute for the tones. The position attribute represents the arrival direction of the tone that is defined with the semicircle surrounding a head.

TABLE I

EXAMPLES OF TRIALS ABOUT FREQUENCY ATTRIBUTE

A: 880 Hz; $A^{\#}$: 932.328 Hz; B: 987.767 Hz				
Trial Number	Tone 1	Tone 2	Question Tone	Similarity Level
1	A	B	$A \times (\sqrt[96]{2})^n$	4
2	A	B	$B/(\sqrt[96]{2})^n$	4
3	A	$A^{\#}$	$A^{\#}/(\sqrt[96]{2})^n$	2
4	$A^{\#}$	A	$A \times (\sqrt[96]{2})^n$	2

E. Data Preprocessing

The EEG signals were preprocessed and analyzed by using Python 3.9.7 and an open-source Python package MNE 1.2 [13]. We reconstructed the equivalent sources from scalp EEG recordings with average reference [14] and the data were band-pass filtered from 0.2 Hz to 40 Hz. Continuous data was segmented into epochs ranging from -0.2 s to 0.8 s at the onset of tone stimulus. An automatic routine was applied to exclude data epochs contaminated by eye movements and muscle artifacts [15]. Then we used spherical spline interpolation to repair bad channels [16].

The trials had conditions that are presented with three letters:

TABLE II

EXAMPLES OF TRIALS ABOUT POSITION ATTRIBUTE

Trial Number	Tone 1	Tone 2	Question Tone	Similarity Level
1	$60°$	$-30°$	$-25°$	3
2	$-30°$	$60°$	$55°$	3
3	$60°$	$30°$	$55°$	1
4	$30°$	$50°$	$25°$	1

TABLE III

DESCRIPTION OF LABELS IN DIFFERENT EXPERIMENTAL CONDITIONS

	Frequency attention	Position attention
High difficulty level	FHH	PHH
	FHL	PHL
Low difficulty level	FLH	PLH
	FLL	PLL

- The first letter: The attribute on which the participant evaluated the similarity. "F" represented the attribute of frequency and "P" represented the attribute of position.

- The second letter: The difficulty level for the frequency attribute. "H" represented the high difficulty level and "L" represented the low difficulty level.

- The third letter: The difficulty level for the attribute on which the participant did not evaluate the similarity. "H" represented the high difficulty level and "L" represented the low difficulty level. For example, the label "FHH" represents a trial condition in which the target is frequency, the frequency similarity is at the high difficulty level, and the position similarity is at the high difficulty level.

F. EEG features

In order to find the neural effects of the anticipation of the difficulty level, we extracted some EEG features described in this section. Here, an EEG signal of participant p for trial k at channel c was represented as $\text{EEG}_k[p, c, n_{(t_{\text{start}}, t_{\text{end}})}]$, where $n_{(t_{\text{start}}, t_{\text{end}})}$ are all points from t_{start} to t_{end}.

1) Event-related potential (ERP): ERPs are small voltages generated in response to a specific stimulus. They reflect the summed activity of postsynaptic potentials generated when a large number of similarly oriented cortical pyramidal neurons discharge synchronously while processing information [17].

We obtained the ERPs of the participant p for the trial k at the channel c by normalizing to the baseline of 200 ms pre-stimulus activity (baseline correction) as:

$$\begin{aligned}&\text{ERP}_k[p, c, n_{(-0.2,0.8)}] = \\ &\text{EEG}_k[p, c, n_{(-0.2,0.8)}] - \overline{\text{EEG}_k[p, c, n_{(-0.2,0)}]}.\end{aligned} \tag{5}$$

where the bar in the equations represented the average of the time series $n_{(t_{\text{start}}, t_{\text{end}})}$ of EEG signals from t_{start} to t_{end}.

2) Time-frequency representation (TFR): TFR visualizes the frequency content of brain activity as it varies over time. It can provide insight into the brain's dynamic activity.

We used the Wavelet transform [18] with 300 ms of sliding window, 50 ms of sliding step, and 2 Hz of spectral smoothing. Then, we obtained the TFR of the participant p for the trial k at the channel c by normalizing to the baseline of 200 ms pre-stimulus activity (baseline correction) as:

$$\begin{aligned}&\text{TFR}_k[p, c, n_{(-0.2,0.8)}] = \\ &Wavelet(\text{EEG}_k[p, c, n_{(-0.2,0.8)}]) - \\ &\overline{Wavelet(\text{EEG}_k[p, c, n_{(-0.2,0)}])}.\end{aligned} \tag{6}$$

where the bar in the equations represented the average of the time series $n_{(t_{\text{start}}, t_{\text{end}})}$ of TFR from t_{start} to t_{end}.

G. EEG analysis

Let $X_k[p, c]$ be an EEG feature, which is one of the features of the participant p for the trial k at the channel c defined in Section II-F and y_k is the label for trial k.

$Original$ EEG features were obtained by averaging the EEG signals with the same label of difficulty level as:

$$o_{\text{H}}[p, c, n] = \overline{X_{k|(y_k \in \{\text{FHH,FHL,PHH,PHL}\})}[p, c, n]}, \tag{7}$$

$$o_{\text{L}}[p, c, n] = \overline{X_{k|(y_k \in \{\text{FLH,FLL,PLH,PLL}\})}[p, c, n]}, \tag{8}$$

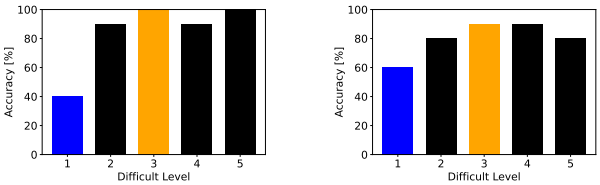
where the bars in the equation represent the average of the EEG features for all trials k that satisfied condition y_k.

The original EEG features, which are typically averaged based on the same label, might not exclusively capture the anticipation of difficulty. They could also be influenced by other responses, such as the perceived differences between Tone 1 and Tone 2 [19], [20]. Therefore, we proposed EEG features by combining the two attributes. To isolate the brain activities associated with difficulty anticipation under trial conditions d, a subtraction can be performed by subtracting the EEG feature of the trial conditions that the participant did not evaluate from the EEG feature of the trial conditions that the participant evaluated.

We defined the proposed EEG feature (the normalized feature) as follows. Initially, we calculated features that captured the difference between the attended and unattended trials of a specific attribute, while maintaining the difficulty level of this attribute constant. For instance, we computed the difference between trials where the participant was asked to assess the similarity of tones based on frequency, and trials focused on position. Four variations of this feature were formulated, as dictated by:

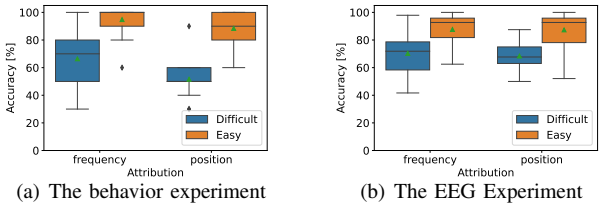
$$\begin{aligned}&s_{\text{FH}}[p, c, n] = \\ &\overline{X_{k|(y_k \in \{\text{FHH,FHL}\})}[p, c, n]} - \overline{X_{k|(y_k \in \{\text{PHH,PLH}\})}[p, c, n]},\end{aligned} \tag{9}$$

$$\begin{aligned}&s_{\text{FL}}[p, c, n] = \\ &\overline{X_{k|(y_k \in \{\text{FLH,FLL}\})}[p, c, n]} - \overline{X_{k|(y_k \in \{\text{PHL,PLL}\})}[p, c, n]},\end{aligned} \tag{10}$$



(a) Frequency selection (b) Position selection

Fig. 4. An example of the accuracy of one of the participants for the five difficulty levels, as well as for the high and low difficulty levels. The blue and orange boxes represent the distributions of accuracy for high and low difficulty levels, respectively.



(a) The behavior experiment (b) The EEG Experiment

Fig. 5. The task performance (accuracy) in behavior and EEG experiments. The blue and orange boxes represent the distributions for middle 50% range of accuracy for high and low difficulty levels, respectively. The green triangles indicate mean values and the black horizontal lines indicate median values, while black dots are outliers.

$$s_{\mathrm{PH}}[p,c,n] = \overline{X_{k|(y_k\in\{\mathrm{PHH,PHL}\})}[p,c,n]} - \overline{X_{k|(y_k\in\{\mathrm{FLL,FLH}\})}[p,c,n]}, \tag{11}$$

$$s_{\mathrm{PL}}[p,c,n] = \overline{X_{k|(y_k\in\{\mathrm{PLH,PLL}\})}[p,c,n]} - \overline{X_{k|(y_k\in\{\mathrm{FHL,FLH}\})}[p,c,n]}, \tag{12}$$

where the bars in the equation represent the average of the EEG features for all trials k that satisfied condition y_k.

Next, to derive EEG features that are independent of the focused tone attribute, we averaged the previously mentioned features across the same difficulty level. The normalized feature was ultimately defined as follows.

$$s_{\mathrm{H}}[p,c,n] = (s_{\mathrm{FH}}[p,c,n] + s_{\mathrm{PH}}[p,c,n])/2, \tag{13}$$

$$s_{\mathrm{L}}[p,c,n] = (s_{\mathrm{FL}}[p,c,n] + s_{\mathrm{PL}}[p,c,n])/2. \tag{14}$$

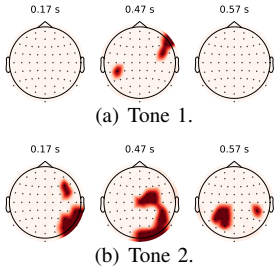
III. Result

A. Behavior

The participants completed the behavior experiment with the five similarity levels to determine their own high and low difficulty levels for the EEG experiment. The difficulty levels were selected based on the task performance for the five similarity levels in the behavior experiment. An example of the accuracy of one of the participants for the five similarity levels, as well as the selection of high and low difficulty levels, is shown in Figure 4. Furthermore, the accuracy for the two difficulty levels are shown in Figure 5. The task performance between the high and low difficulty levels showed significant differences ($p < 0.01$) for the frequency and position attributes in both the behavior and EEG experiments.

B. EEG

1) Event-related potential (ERP): Figure 6 illustrated topographic maps of Normalized ERPs highlighting regions with significant differences between the two difficulty levels. Specifically, the differences are observed at 0.17 s, 0.47 s, and 0.57 s. These topographic maps provide a visual representation of the neural activity variations associated with different levels of difficulty during the specified time intervals. In the period for Tone 2, significant differences were found in the right central vertex, right parieto-occipital, and left parieto-occipital regions, while there were almost no differences in Tone 1. Because of this result, all results below are only for Tone 2. We selected channels, C2, P5, and T8, as the representatives for the areas with the significant differences, respectively.



(a) Tone 1.

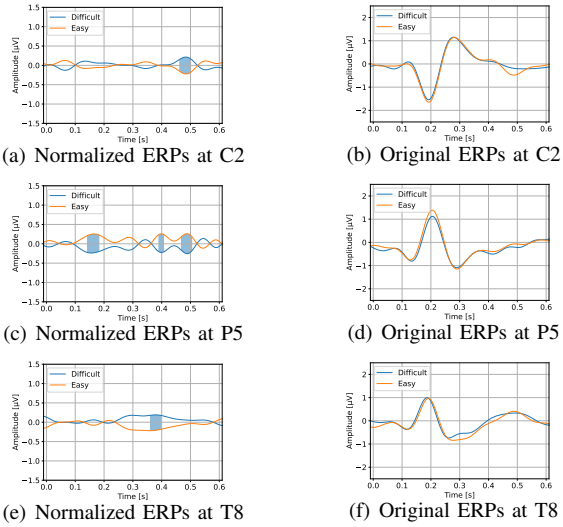
(b) Tone 2.

Fig. 6. Topographic maps of the significant differences in the normalized between the two difficulty levels of the sound stimulus at 0.14 s, 0.23 s and 0.32 s. The red parts indicate significant differences ($p < 0.05$).

The ERP results of these channels are depicted in Figure 7. The significant difference in the ERPs in the right central vertex region can be attributed to the variation to N4, as shown in Figure 7(a) and 7(b). Additionally, in the left parieto-occipital region, the ERPs can be attributed to the variation to P2, N4, and P3, as shown in Figure 7(c) and 7(d). Moreover, in the right temporal region, the ERPs can be attributed to the variation to P3, as shown in Figure 7(e) and 7(f).

2) Time-frequency representation (TFR): Figure 8 illustrates topographical maps showing areas with significant differences in Tone 2 between the two difficulty levels in the time-frequency representation at 0.07 s, 0.50 s, 0.58 s, and 0.64 s. In the theta (4–7 Hz), alpha (8–11 Hz), and beta (12–20 Hz) bands, significant differences were all found in the prefrontal and occipital regions at 0.07 s, while at 0.50 s, 0.58 s, and 0.64 s in the alpha and beta bands, significant differences were found in the left central vertex and occipital regions. The gamma band showed significant differences in the right and left parieto-occipital regions. Because of these results, we selected channels, Fz, C3, and Pz, as the representatives for the prefrontal, left central vertex, and occipital areas, respectively. The TFR results of these channels are depicted in Figure 9.

The significant difference in the TFR in the prefrontal area

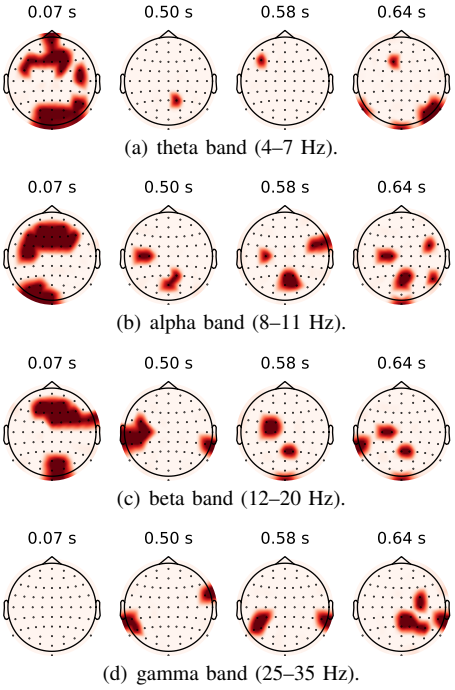


(a) Normalized ERPs at C2 (b) Original ERPs at C2

(c) Normalized ERPs at P5 (d) Original ERPs at P5

(e) Normalized ERPs at T8 (f) Original ERPs at T8

Fig. 7. The normalized (left) and original (right) ERPs in Tone 2. The blue and orange lines are ERPs for the high and low difficulty levels, respectively. The shadows represent time ranges showing significant differences ($p < 0.05$).



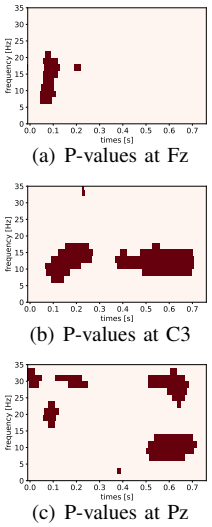
(a) theta band (4–7 Hz).

(b) alpha band (8–11 Hz).

(c) beta band (12–20 Hz).

(d) gamma band (25–35 Hz).

Fig. 8. Topographic maps for the normalized TFR with significant differences between two difficulty levels at different frequency bands in Tone 2. The red parts indicate significant differences ($p < 0.05$).



(a) P-values at Fz

(b) P-values at C3

(c) P-values at Pz

Fig. 9. The normalized TFR in Tone 2. The parts of red indicate significant differences ($p < 0.05$).

can be attributed to the variation to the theta, alpha, and beta bands at 50–100 ms, as shown in Figure 9(a). Additionally, in the left central vertex region, the TFR can be attributed to the variation in alpha band and beta1 band (12–16 Hz) at 80–280 ms and 380–700 ms, as shown in Figure 9(b). Moreover, in the occipital region, the TFR can be attributed to the variation to alpha band and gamma band (25–35 Hz) at 500–700 ms, as shown in Figure 9(c).

IV. Discussions

In this study, we investigated EEG features that are involved in anticipation of the difficulty of upcoming auditory tasks. By manipulating sound similarity, the participants anticipated how much difficult the upcoming task would be. The behavior experiment revealed significant differences in task performance between high and low difficulty levels. This suggests that the chosen level of difficulty was appropriate. In the EEG experiment, the analysis of ERPs revealed significant differences in several brain regions. The right central vertex region showed notable variations primarily in the N4 component. In contrast, the left parieto-occipital region displayed variations in the P2, N4, and P3 components, while the right temporal region exhibited variations specifically in the P3 component. Furthermore, the analysis of TFR uncovered additional significant differences across various brain regions. During the earlier time interval, the prefrontal and occipital regions demonstrated significant variations across multiple frequency bands. Moving to later time intervals, the left central vertex and occipital regions displayed significant differences, particularly in the alpha and beta bands. Additionally, the right and left parieto-occipital regions exhibited distinctive variations specifically in the gamma band.

In the prefrontal area, according to earlier studies, synchronization of low and medium frequency bands was involved in attention transformer [21]. In the present study, there were significant differences in the low and medium frequency bands (5–20 Hz) at the beginning of the event (50–100 ms). This indicates that the participants assigned more attention to the task of the high difficult level than that of the low difficult level. Furthermore, previous studies have indicated that the evocation of N4 components occurs when there is semantic

inconsistency [22]. In the present study, since similarity was used as a basis for judging the level of difficulty, participants can clearly perceive this inconsistency when the task was easy, so N4 was clearly evoked. In contrast, the N4 was not evoked when the task was difficult. We also found a P3 variation in the left parieto-occipital region, which, according to earlier studies, was usually associated with self-perception, that is, metacognition [23]. This suggests that after perceiving information the participants usually compared the perceived environmental information with their self-competence, to anticipate the difficulty of the problem. In addition, following time-frequency analysis, significant differences were observed in the alpha band between 400–700 ms in both the left central vertex region and the parieto-occipital region. However, the nature of these differences varied: in the left central vertex region, alpha band synchronization was observed, whereas, in the parieto-occipital region, desynchronization was observed in both the alpha and gamma bands. We believe these were important neural markers to prepare resources in response to upcoming problems of varying difficulty.

In summary, this study explores the neural mechanisms involved in anticipation during auditory tasks. The findings highlight the role of attention and predictive behavior in responding to environmental information. The results also reveal distinct patterns of neural activity associated with different levels of task difficulty, as measured by ERPs and TFR. These patterns suggest the involvement of specific brain regions and frequency bands, indicating variations in cognitive processing. The study enhances our understanding of how the brain adapts to different levels of cognitive demand and opens possibilities for applications in cognitive load monitoring and adaptive task design. Further research can expand upon these findings to investigate the specific cognitive processes associated with task difficulty levels.

REFERENCES

[1] M. Bar, "The proactive brain: Using analogies and associations to generate predictions," *Trends in Cognitive Sciences*, vol. 11, no. 7, pp. 280–289, 2007.

[2] C. Brunia, "Neural aspects of anticipatory behavior," *Acta Psychologica*, vol. 101, no. 2-3, pp. 213–242, 1999.

[3] B. Knutson and S. M. Greer, "Anticipatory affect: Neural correlates and consequences for choice," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1511, pp. 3771–3786, 2008.

[4] M. Bar, "The proactive brain: Memory for predictions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1235–1243, 2009.

[5] V. Bianco, M. Berchicci, E. Mussini, R. L. Perri, F. Quinzi, and F. Di Russo, "Electrophysiological evidence of anticipatory cognitive control in the stroop task," *Brain Sciences*, vol. 11, no. 6, p. 783, 2021.

[6] W. C. McSherry and J. E. Holm, "Sense of coherence: Its effects on psychological and physiological processes prior to, during, and after a stressful situation," *Journal of Clinical Psychology*, vol. 50, no. 4, pp. 476–487, 1994.

[7] M. Simonet, H. B. Meziane, O. R. Runswick, *et al.*, "The modulation of event-related alpha rhythm during the time course of anticipation," *Scientific Reports*, vol. 9, no. 1, p. 18 226, 2019.

[8] K. Kveraga, A. S. Ghuman, and M. Bar, "Top-down predictions in the cognitive brain," *Brain and Cognition*, vol. 65, no. 2, pp. 145–168, 2007.

[9] G. Vanhollebeke, S. De Smet, R. De Raedt, C. Baeken, P. van Mierlo, and M.-A. Vanderhasselt, "The neural correlates of psychosocial stress: A systematic review and meta-analysis of spectral analysis EEG studies," *Neurobiology of Stress*, p. 100 452, 2022.

[10] T. Nishino, S. Kajita, K. Takeda, and F. Itakura, "Interpolating head related transfer functions in the median plane," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*, IEEE, 1999, pp. 167–170.

[11] J. H. McDermott and A. J. Oxenham, "Music perception, pitch, and the auditory system," *Current Opinion in Neurobiology*, vol. 18, no. 4, pp. 452–463, 2008.

[12] D. De Klerk, "Equal temperament," *Acta Musicologica*, vol. 51, no. Fasc. 1, pp. 140–150, 1979.

[13] A. Gramfort, M. Luessi, E. Larson, *et al.*, "MEG and EEG data analysis with mne-python," *Frontiers in Neuroscience*, p. 267, 2013.

[14] D. Yao, "A method to standardize a reference of scalp EEG recordings to a point at infinity," *Physiological Measurement*, vol. 22, no. 4, p. 693, 2001.

[15] F. C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener, "Semi-automatic identification of independent components representing EEG artifact," *Clinical Neurophysiology*, vol. 120, no. 5, pp. 868–877, 2009.

[16] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and Clinical Neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[17] S. Sur and V. K. Sinha, "Event-related potential: An overview," *Industrial Psychiatry Journal*, vol. 18, no. 1, p. 70, 2009.

[18] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[19] I. Anourova, V. V. Nikouline, R. J. Ilmoniemi, J. Hotta, H. J. Aronen, and S. Carlson, "Evidence for dissociation of spatial and nonspatial auditory information processing," *Neuroimage*, vol. 14, no. 6, pp. 1268–1277, 2001.

[20] I. Anourova, P. Rämä, K. Alho, S. Koivusalo, J. Kahnari, and S. Carlson, "Selective interference reveals dissociation between auditory memory for location and pitch," *Neuroreport*, vol. 10, no. 17, pp. 3543–3547, 1999.

[21] M. D. Sacchet, R. A. LaPlante, Q. Wan, *et al.*, "Attention drives synchronization of alpha and beta rhythms between right inferior frontal and primary sensory neocortex," *Journal of Neuroscience*, vol. 35, no. 5, pp. 2074–2082, 2015.

[22] K. K. Toffolo, E. G. Freedman, and J. J. Foxe, "Evoking the N400 event-related potential (ERP) component using a publicly available novel set of sentences with semantically incongruent or congruent eggplants (endings)," *Neuroscience*, vol. 501, pp. 143–158, 2022.

[23] K. Desender, F. Van Opstal, G. Hughes, and E. Van den Bussche, "The temporal dynamics of metacognition: Dissociating task-related activity from later metacognitive processes," *Neuropsychologia*, vol. 82, pp. 54–64, 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Event-Related Potential in Rapid Serial Visual Presentation-based Partial Face Cognition Depends on Visible Face Components

Ingon Chanpornpakdi* and Toshihisa Tanaka*
* Graduate School of Engineering, Tokyo University of Agriculture and Technology, Japan
E-mail: mint@sip.tuat.ac.jp, tanakat@cc.tuat.ac.jp Tel/Fax: +81-42-388-7123

***Abstract*—Under the SARS-CoV-2 pandemic situation, wearing masks became a must-do action in our daily life. Although some parts of the face are covered, humans can still recognize people's faces precisely. In this study, we aimed to identify how the visibility of different parts of the face affects face recognition using an electroencephalogram (EEG). We also investigated how physical response affects the cognition mechanism. In the experiment, we performed a face cognition task using a rapid serial visual presentation (RSVP) stimulus. The experiment consisted of two cognition tasks, images displayed at 1000 msec with a button press when the target is shown and images displayed at 200 msec without any physical responses. Six faces under seven visibility conditions (full-face shown, eyes covered, nose covered, mouth covered, eyes and nose covered, eyes and mouth covered, and nose and mouth covered) were presented as the stimulus to identify the effect of each part being visible or not visible. As a result, we observed two components of P300 event-related potential (ERP) evoked, P3a and P3b, that related to partial face cognition. For both physical response tasks, we found that the P300 peak value was significantly lower, and the P3a and P3b latency were significantly delayed when the eyes were not visible. These results suggest that the eye could be the most crucial feature in face cognition, and there is a possibility of substituting the no physical response task with the physical response task.**

I. Introduction

During the SARS-CoV-2 pandemic, wearing masks has become a health precaution that everyone needs to follow. By wearing masks, part of the face became hidden, but people could still recognize faces regardless of the small visible face area. Previous studies claimed that holistic face processing enhances the identification judgment [1]–[3]. Covering part of the face decreases the ability to recognize faces [1], [4], [5] However, whether holistic face processing applies to partial face cognition and how people can recognize masked faces are still in question. To clarify the question on the partial face mechanism, we utilized the rapid serial visual presentation (RSVP) paradigm, which is a widely-used stimulus for visual perception study [6].

RSVP is a temporal sequence of target and non-target images with the probability of the target image being considerably smaller than the non-target image. Generally, the rate of presentation is expeditious. The electroencephalogram (EEG) response was focused on in this study due to its high temporal resolution [7], [8]. During the RSVP task, a time-lock response related to the perceived target image called event-related potential (ERP), is induced with the particular pattern [9], [10]. The pattern of ERP response consists of many components, such as P300 which is related to the target recognition found at 250-700 msec after the stimulus invoked [6], [11]. Each component was typically named according to its polarity and latency after the stimulus onset.

Among many ERP components, N170 and vertex positive potentials (VPP) are considered as a face responsive ERP [12]–[14]. For target and non-target face cognition, the relevance of N200 and P300 potentials outweighs that of the N170 and VPP components [13], [14]. In partial face cognition, some studies claimed that lacking the eye component results in N170 latency extension, but the N170 amplitude remained unchanged [15]–[19]. In contrast, other studies reported that a face with eyes invisible affected both N170 amplitude and latency [20], [21]. With these arguments, how the face feature affects the ERP response needs to be clarified.

Based on the aforementioned studies, ERP components were utilized as a biomarker for partial face and whole face cognition to study how each component of the face features plays a role in face cognition. We hypothesized that people could recognize an uncovered face easier and more precisely than a partially covered face due to the holistic face processing and more components of the face are provided in the whole face cognition. Additionally, many studies in the EEG-based RSVP paradigm usually utilized button press tasks [6] which can cause mental fatigue and the results can be affected by decreasing attention during the task [22]. Therefore, we also investigated how ERP changes during 'button press' and 'no button press' tasks in order to minimize workload during ERP tasks in future studies.

II. Materials and Methods

A. Dataset

The data recorded from 18 young and healthy participants (average age: 26.05 ± 2.61 years; minimum age: 22; maximum age: 31) during the partial face cognition task in our previous study [23] was used. The experiment consists of two tasks; button press and no-button press task. Both tasks have the same flow of experiment, starting with the instruction and the target face presentation, followed by 500 msec of fixation presentation, and the presentation of the face images. The

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

stimulus presentation time was set to 1,000 msec for the button press task and 200 msec for the non-button press task with an inter-stimulus interval (ISI) of 500 msec. Each task consists of seven blocks, showing seven face image conditions; full face, face with eyes covered, face with nose covered, face with mouth covered, face with eyes and nose covered, face with eyes and mouth covered, and face with nose and mouth covered, respectively. Six Asian face images were used in each block; one target face image and five non-target face images. Each face was presented 20 times per block, resulting in 120 trials in total. The EEG data was recorded together with the eye movements using a 72-channel Refa amplifier (electrode alignment: 10-10 international system; reference: average of all EEG input channels; ground: left wrist). (see [23] for details)

B. Data Analysis

The acquired signals were pre-processed in the same manner as done in Chanpornpakdi and Tanaka [23] using MNE package 1.3.0 [24] in Python 3.5.8. The data were first re-referenced with the average of earlobe electrodes (M1 and M2). To obtain as less-contaminated EEG data as possible, they were filtered with a 1–30 Hz bandpass filter and submitted to extended Infomax independent component analysis (ICA). The clean EEG data were segmented from -100 msec to 1000 msec from the target onset, resulting in 120 epochs. The epoched data were grouped into 12 electrodes based on the brain regions; frontal, parietal, occipital, and temporal regions, and averaged among the electrodes in the same group. (see [23] for details)

To acquire the ERP, we concatenated the preprocessed EEG epochs of each participant in each condition together and took the average of the concatenated data, called the grand average, across the participant and a topographic map of the grand average was plotted. The grand average ERP was calculated on 20 trials per participant across 18 participants, making the total number of trials 360. In the statistical test, the maximum amplitude and latency of the corresponding amplitude of the two ERP components, N170 (negative peak at 100–200 msec) and P300 (positive peak at 300–400 msec for early P300, or P3a, and 400–600 msec for later P300, or P3b), were extracted from each participant. An upper-tailed dependent t-test was applied for each amplitude and a lower-tailed dependent t-test was used for each latency of the ERP components (N170, P3a, and P3b with peaks of 18 participants for each component as the dependent variable) to test our hypothesis that the amplitude of ERPs during full-face cognition is larger than the amplitude of ERPs during partial face cognition. The ERP peaks evoked and latencies in the full-face condition were compared with those evoked in each partial face condition (six conditions: covered eyes, covered nose, covered mouth, covered eyes and nose, covered eyes and mouth, and covering nose and mouth).

Furthermore, we also applied a two-way repeated measures analysis of variance (ANOVA) test to the amplitude of the ERP peaks; N170, P3a, and P3b, to test our hypotheses that: 1) all the face condition groups have equal mean ERP peaks, 2) all the physical task condition groups have equal mean ERP peaks, and 3) the factors are independent, or the interaction effect of the face cognition and task conditions does not exist. The two factors in our ANOVA test were seven face cognition conditions (full-face and six partial face conditions) and two button press conditions (button press and no button press condition) and Greenhouse–Geisser correction was used in multiple comparisons. The dependent variable was the same as that used in the t-test and the sample size was 18. After identifying the factors (button, face condition, and the interaction of both factors) that have significantly different in ERP among themself, we performed pairwise t-tests to explore the relationships that showed significant differences using Benjamini/Hochberg (non-negative) correction.

III. Results

In this study, we compared the ERP elicited during full-face and partial face cognition in button press and no button press tasks. Figure 1 shows the examples of the grand average ERP and topographic map of the full-face, partial face with eyes visible, and partial face with eyes invisible conditions during the button press task and no button press task, respectively. The face conditions shown were the full-face, face with nose covered, and face with eyes and nose covered image cognition condition. The face image with nose covered stimulus presentation was to show as an example of a part of the face covered and the P3b of this condition was found to be the largest among all the conditions. The face with the nose and eyes covered was shown as an example of the ERP when the two parts of the face were covered and how the ERP was elicited differently when the eyes were covered in addition to the nose.

A. Amplitud of Evoked ERP

Four ERP components were observed: N170, P200, N250, and P300. We have highlighted the two essential ERP components for face cognition in Figure 1; N170 (a characteristic ERP for face cognition) in gray and P300 (an attention ERP for target and nontarget cognition) in yellow and blue color. N170, P200, and N250 of the perceived target were observed with larger amplitudes in the frontal and central electrodes that became smaller as the position moved toward the posterior region, regardless of the type of image condition. The other component observed was the P300 component, which contained two subcomponents: early P300 (P3a) observed at a latency of approximately 300 msec, predominant at the frontal region of the brain, and later P300 (P3b) observed at a later latency around 400–600 msec, noticeably at the parietal region of the brain. We could also see that the N170 of the target and non-target face images were almost at the same level whereas the P300 component of the target image has a larger amplitude than that of the nontarget image. Moreover, both tasks shared a common result: there were drops in the P300 peaks evoked when the eye component of the target was invisible.

To test our hypothesis, we have performed the upper-tailed dependent t-test on the amplitude of N170, P3a, and P3b. The results were plotted and shown in Figure 2. Figure 2 shows the peak of N170 in the frontal region, P3a in the frontal region,

and P3b in the parietal region in each face condition during the button press and no button press tasks. When comparing the ERP peaks evoked in each condition, peaks of N170 in the frontal electrode region from both the button press and no button press tasks were at a similar level, and there was no significant difference observed between the peaks evoked in full-face and partial face cognition. However, we found a significantly larger peak in N170 in the comparison of the full-face cognition condition and face with eyes and nose covered cognition in the button press condition ($p < 0.05, d = 0.449$).

For P300, the peaks of both P3a and P3b subcomponents dropped when the eye component was invisible during both button press tasks when compared to the full-face cognition condition. However, the significant differences of P3a amplitude in the frontal area between the full-face and partial face cognition were found only when the eyes and mouth were covered ($p < 0.05, d = 0.423$) in the button press task and when the eyes and nose were covered ($p < 0.05, d = 0.478$) in the no button press task. Furthermore, the peak of P3b in the parietal region during full-face cognition was significantly larger than the partial face condition when the eyes ($p < 0.001, d = 0.882$), and eyes and mouth were covered ($p < 0.01, d = 0.648$) in the button press task and when the eyes and nose ($p < 0.01, d = 0.647$), and eyes and mouth were covered ($p < 0.001, d = 1.015$) in the no button press task.

B. Latency of Evoked ERP

Figure 1 also shows the topographic map around 0.150 sec, 0.300 sec, 0.450 sec, 0.500 sec, and 0.550 sec that represents the N170, P3a, and P3b, respectively. The N170 peak, in which either the nose or mouth was covered, has slightly larger peaks in the central area in both button press tasks. We could also clearly see from the topographic map in Figure 1 that there are some delays in the latencies of the ERPs.

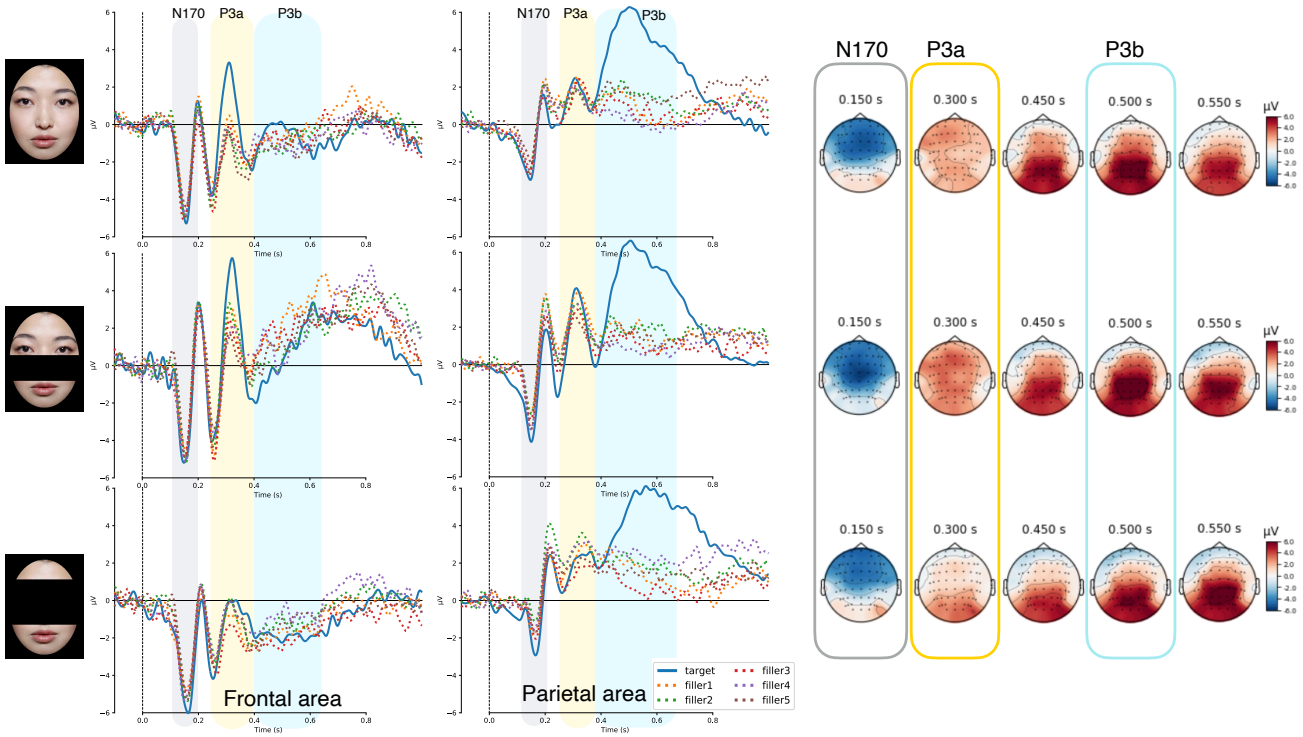
We have performed the lower-tailed dependent t-test to observe the delay of elicited ERPs caused when removing parts of the face during the cognition task. As a result, we found that there were significant delays in N170 elicited when the eyes and nose were covered ($p < 0.01, d = -0.644$) around 0.010 sec and when the eyes and mouth were covered ($p < 0.05, d = -0.557$) around 0.011 sec in the button press task. However, we did not observe any significant delay of N170 elicited between the full-face and partial face cognition in no button press task. Similarly, we also found that the P3a elicited in both the button press task was significantly delayed when the eyes, eyes and nose, and eyes and mouth were covered. In the button press task, when the eyes ($p < 0.05, d = -0.475$), eyes and nose ($p < 0.05, d = -0.549$), and eyes and mouth were covered ($p < 0.05, d = -.478$), P3a got extended by 0.015 sec, 0.024 sec and 0.021 sec, respectively. In the no button press task, when the eyes ($p < 0.05, d = -0.517$), eyes and nose ($p < 0.05, d = -0.570$), and eyes and mouth were covered ($p < 0.005, d = -0.699$), P3a got extended by 0.017 sec, 0.020 sec and 0.026 sec, respectively.

For the latency of the evoked P3b, we found that P3b got extended in all the conditions except when the nose and mouth were covered in the button press task. When the eyes ($p < 0.001, d = -1.368$), nose ($p < 0.05, d = -0.420$), mouth ($p < 0.05, d = -.519$), eyes and nose ($p < 0.001, d = -1.106$), and eyes and mouth were covered ($p < 0.001, d = -1.099$), the P3b subcomponent was delayed by 0.062 sec, 0.019 sec, 0.020 sec, 0.055 sec, and 0.046 sec, respectively. Contradictory, we only found a significant delay in P3b latency only when the eyes ($p < 0.001, d = -1.021$), eyes and nose ($p < 0.001, d = -0.977$), and eyes and mouth ($p < 0.05, d = -0.463$) were covered in the no button press task. The delays were around 0.041 sec, 0.045 sec, and 0.027 sec, respectively.

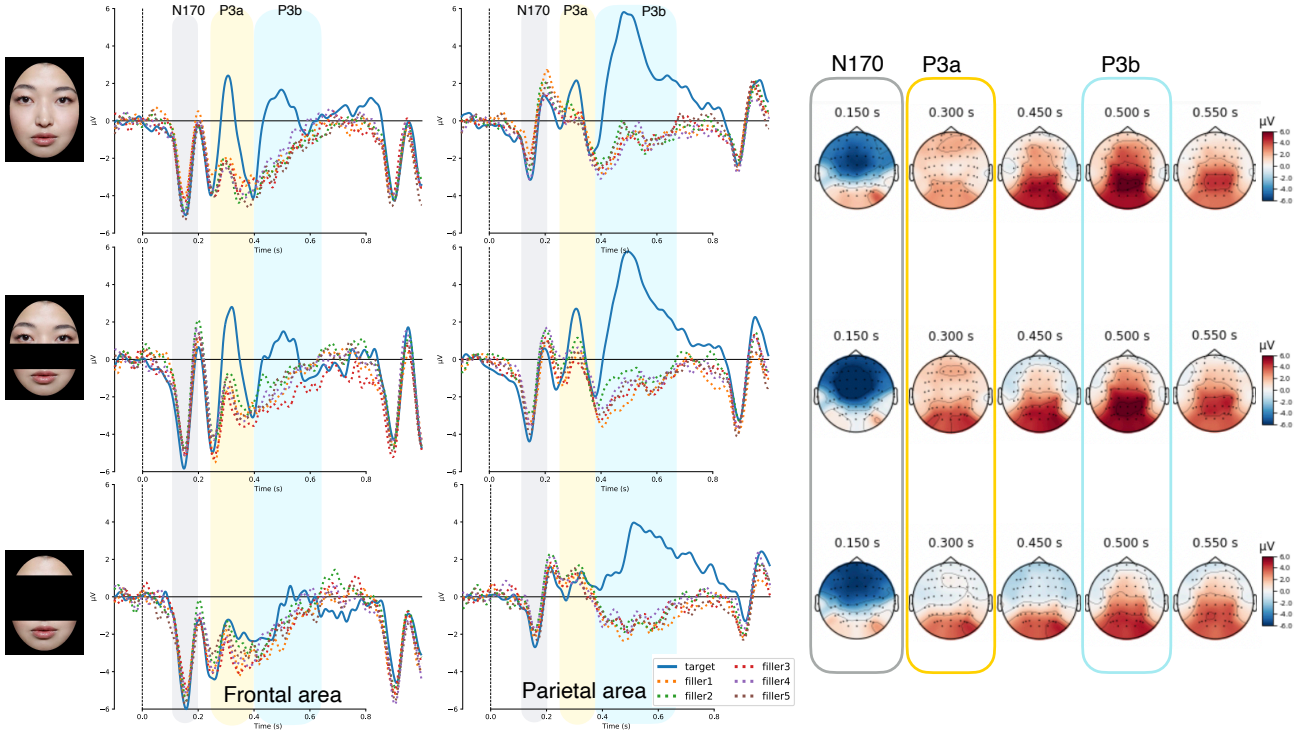
C. Interaction of Face Conditions with Physical Response

The results of the two-way repeated-measures ANOVA test for the amplitude of N170, P3a, and P3b peaks around the frontal and parietal regions are presented in Tables I, II, and III, respectively. From Table I, when comparing the full and partial face cognition conditions, the amplitude of the N170 component associated with partial face cognition has similar peaks in both the button press and no button press conditions, $F(1, 17) = 0.692, p = 0.010.417, {\eta_p}^2 = 0.003$. However, we could not observe the significant differences in conditions, $F(6, 102) = 0.779, p = 0.552, {\eta_p}^2 = 0.725$, and the interaction of button press tasks face cognition conditions in the elicited N170, $F(6, 102) = 0.218, p = 0.945, {\eta_p}^2 = 0.765$.

From Tables II and III, we found that the button press task significantly affected both P3a, $F(1, 17) = 11.082, p < 0.01, {\eta_p}^2 = 0.028$, and P3b, $F(1, 17) = 9.708, p < 0.01, {\eta_p}^2 = 0.042$. In addition, the face cognition condition had a significant main effect on P3a, $F(6, 102) = 5.893, p < 0.001, {\eta_p}^2 = 0.062$, and P3b, $F(6, 102) = 9.305, p < 0.001, {\eta_p}^2 = 0.075$. Pairwise t-tests revealed that the P3a peak evoked in partial face cognition was significantly different when: covered eyes were compared with covered nose ($p < 0.05, d = -0.592$); covered eyes and mouth were compared with covered nose ($p < 0.05, d = -0.601$); covered eyes and nose compared with mouth ($p < 0.05, d = -0.467$), mouth ($p < 0.01, d = -0.751$), nose and mouth ($p < 0.05, d = -0.677$); and covered mouth were compared with covered nose ($p < 0.05, d = -0.338$). The P3b peaks evoked in partial face cognition were also significantly different from each other when comparing covered eyes with covered eyes and mouth ($p < 0.01, d = 0.464$), full-face ($p < 0.05, d = -0.256$), mouth ($p < 0.001, d = -0.503$), and nose ($p < 0.05, d = -0.330$). The comparison of the P3b peak between covered eyes and mouth and covered eyes and nose ($p < 0.05, d = -0.400$), full-face ($p < 0.005, d = -0.759$), mouth ($p < 0.001, d = -0.997$), nose ($p < 0.005, d = -0.817$), and nose and mouth ($p < 0.005, d = -0.702$) were also significantly different. In addition, the P3b peak evoked when the eyes and nose were covered was significantly different from the peak evoked when the mouth was covered ($p < 0.005, d = -0.610$), and the P3b peak elicited during

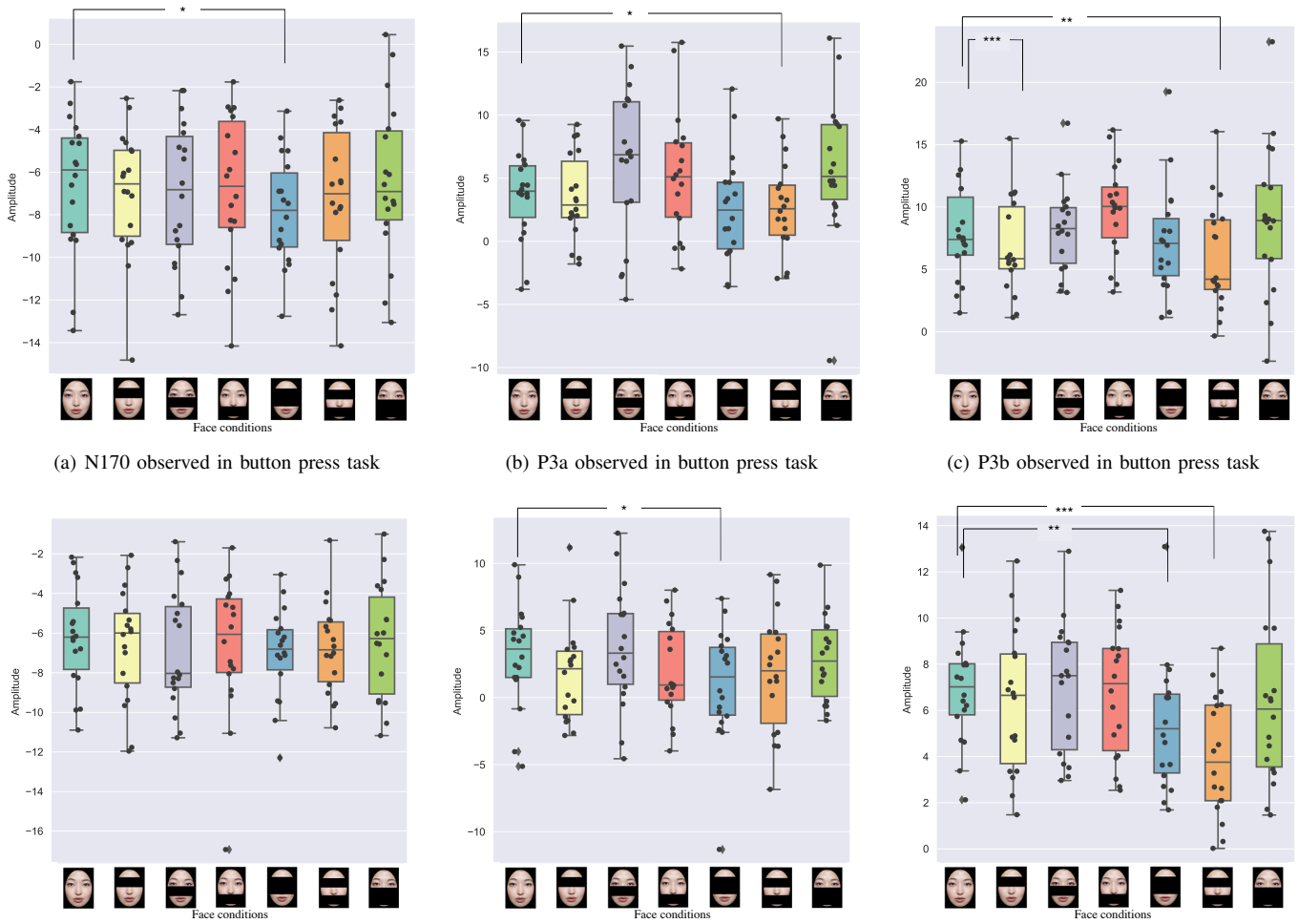


(a) button press task



(b) no button press task

Fig. 1. Examples of grand average ERP and topographic map of full-face, partial face with eyes visible and partial face with eyes invisible condition during the button press and no button press task. The grey, yellow, and blue highlighted areas represent the latency of N170, P3a, and P3b. The thick blue line represents the ERP of the target, while the dotted line represents non-target 1, 2, 3, 4, and 5 respectively.



(a) N170 observed in button press task (b) P3a observed in button press task (c) P3b observed in button press task

(d) N170 observed in no button press task (e) P3a observed in no button press task (f) P3b observed in no button press task

Fig. 2. ERP amplitudes (N170 at frontal region, P3a at frontal region, and P3b at parietal region) comparison of each condition in the button press (a, b, and c) and no button press task (d, e, and f), where $n = 18$, $p < 0.05 : *$, $p < 0.01 : **$, and $p < 0.001 : ***$.

TABLE I
THE TWO-WAY REPEATED MEASURES ANOVA RESULT OF N170 AROUND THE FRONTAL AREA.

Source	SS	df1	df2	MS	F	p-unc	p-GG-corr	ng2	eps
Button	7.199	1.000	17.000	7.199	0.692	0.417	0.417	0.003	1.000
Condition	27.259	6.000	102.000	4.543	0.779	0.588	0.552	0.012	0.725
Button * Condition	3.914	6.000	102.000	0.652	0.218	0.970	0.945	0.002	0.765

TABLE II
THE TWO-WAY REPEATED MEASURES ANOVA RESULT OF P3A AROUND THE FRONTAL AREA.

Source	SS	df1	df2	MS	F	p-unc	p-GG-corr	ng2	eps
Button	116.003	1	17	116.003	11.082	0.004	0.004	0.028	1.000
Condition	269.421	6	102	44.903	5.893	0.00002	0.0004	0.062	0.657
Button * Condition	55.728	6	102	9.288	1.551	0.169	0.206	0.014	0.578

TABLE III
THE TWO-WAY REPEATED MEASURES ANOVA RESULT OF P3B AROUND THE PARIETAL AREA.

Source	SS	df1	df2	MS	F	p-unc	p-GG-corr	ng2	eps
Button	119.751	1	17	119.751	9.708	0.006	0.006	0.042	1.000
Condition	219.175	6	102	36.529	9.305	3.965E-08	0.00003	0.075	0.537
Button * Condition	48.287	6	102	8.048	2.632	0.021	0.067	0.017	0.438

full-face cognition when compared with partial face cognition with the mouth covered was also significantly different ($p < 0.05, d = -0.275$). We found no significant interaction between face cognition condition and button press condition for neither P3a, $F(6, 102) = 1.551, p = 0.206, {\eta_p}^2 = 0.014$ (Table II), nor P3b, $F(6, 102) = 2.632, p = 0.067, {\eta_p}^2 = 0.01$ (Table III). However, pairwise t-tests revealed that some of the face cognition condition results varied significantly between the button conditions.

IV. DISCUSSION

A. Evoked ERP Components

As mentioned in the introduction, the N170 and VPP are typical face-characteristic ERPs associated with face cognition. The observations relating to N170 and P300 match the finding that these ERP components are characteristic of face stimuli,

whereas the amplitude of the P300 depends on the target and non-target conditions of the face stimulus [13], [14]. In addition, we found significant differences between the N170 peaks of the full-face and partial face cognition conditions only in the button press task. This result emphasizes that the N170 component is the face characteristic ERP component and is sensitive to facial features only either when performing behavior response or participants had enough time to perceive the face image.

We also found a positive peak at approximately 200 msec, with a greater amplitude around the frontal region. This positive peak latency is consistent with the VPP and P200 latencies [25]–[27]. Hence, there are two options that this peak could belong to. However, N170 and VPP have been reported to be face-sensitive responses from the same neuron generator in the occipito-temporal region [15], [25]. On the other hand, P200 reflects the feedback of the visual cortex, which is necessary for perception, and the response is more sensitive to configural face processing or spacing processing rather than featural face processing [28]. On this account, the observed peak is more likely to be P200, and the subsequent negative peak found around 250 msec could be named N250. N250 indicates the categorization process for individual faces [26].

Regarding the P300 component, the P3a subcomponent is related to focal attention and represents the response to a new event or stimulus that causes an attention shift, whereas P3b is related to the memory stage and represents the response to cognition [29]–[31]. These findings indicate that the early P3a subcomponent found in this experiment might reflect the attention to the facial image, and the later P3b subcomponent reflects the response when the participant recalls the face from their short-term memory storage.

B. ERP Amplitude and Latency

From our experiment, we obtained that covering the eye component significantly dropped the ERP components for both N170 and P300, where P3a indicated the attention process and P3b indicated the identification process perception. We found that the amplitude and latency of N170 peaks of the partial face cognition conditions when the eyes component was covered only was significantly decreased and delayed when compared to the full face cognition condition in the button press task. While the latencies of maximum N170 peaks got extended when the eyes and nose, and eyes and mouth were covered, the amplitude of N170 was significantly dropped only when the eyes and nose were covered. This result implies that the eyes are the most important. Apart from the eyes which are the structural information, having a nose would provide the configural information of the distance between eyes and nose and make the performance of face cognition better which is in line with the mentioned study [32]. Therefore, when the eyes and nose were covered, we could observe both declinations in amplitude and latency. However, this result was obtained only in the button press task, which means that the N170 component is the face characteristic ERP component and is sensitive to facial features only either when performing behavior response or when there was enough time for the participant to perceive the face image.

When we investigated the partial face cognition studies that found the delay and decrease in amplitude of N170, we found the common factor that could be the cause of such response changes in N170. The two factors were stimulus presentation time and the condition of behavior response. Eimer [16] and Eimer [33] performed the experiment with a stimulus presented at only 100 msec but the participant was asked to press the button. The studies showed that the N170 peak was delayed when using a face with no eyes images compared with the intact face [16] and the N170 peak was lower and delayed in the experiment using the face with no internal features and a face lacking external features when compared to the intact face [33]. Other studies showed that presenting the stimulus images at 250 msec and 500 msec showed the delay in N170 latencies regardless of having behavior response or not [15], [17], [19]. Furthermore, when the stimulus was presented randomly in the range from 500 msec to 1000 msec and the participant was asked to provide the behavior response, N170 peak latency was extended and the amplitude also became lower for the face with Gaussian noise when compared to the full-face [20], [21].

From these studies, stimulus presentation time and the condition of behavior response might be the factors that affect N170. Our assumption is that a missing eye component face image only delays the N170 latency if either the stimulus was presented more than 250 msec or the stimulus was presented at a duration lower than 250 msec but the participants were asked to perform a physical response. Furthermore, we assumed that if the stimulus was presented longer than 500 msec in the physical response task, both the amplitude and latency of N170 will be smaller and longer. This assumption could explain why the significant diminish observed in N170 peak, and latency delay only in the button press task in which we presented the stimulus for 1000 ms and asked the participant to provide button press feedback. Certainly, other factors may cause the N170 response to be prolonged and diminished such as the type of stimulus, the duration of the inter-stimulus interval, etc. which are left to be clarified.

C. Limitation and Future Work

In our experimental design, the presentation time changed drastically between the physical response tasks. This can cause a better perception of the button press task due to a longer time to expose to the stimulus and process the presented information of the image. Furthermore, we could not find the difference in the N170 peak which could be a hint for the partial face cognition in the no button press task due to the short duration of stimulus presentation. To overcome such problems, we should reconsider the better presentation time that diminishes any inconsistency in our experiment in order to investigate the effect of workload in button press task precisely. It was suggested that 500 msec could be the most suitable for the RSVP face cognition task [6].

In the data pre-processing methods, we have only a small number of trials for each type of face image: therefore we only corrected the eye movement using ICA without rejecting any epochs. Applying ICA has a risk of losing some ERP components. To prevent that, we only removed the eye movement artifacts referencing to the eye movement data we recorded along with the EEG. However, there might be some other artifacts caused by the movements and sweat of the participants left in the processed data. Ideally, we should increase the number of trials to at least 100 trials [34] per face image to obtain clean ERPs. However, this study contained two tasks. Each task contained seven blocks; consisting of 6 images presented 20 times per image. As a result, A total of 120 images were presented in each block, 840 images per task, and 1,680 images in the whole experiment. The time taken to present all these images including the fixation of 500 msec in between each stimulus was around 21 minutes in the button press task and 10 minutes in the no button press task. In addition to the experimental duration, the EEG headset and training tasks needed to be arranged prior to the experiment. Therefore, we could only perform 20 trials for each image as Cohen and Polich [35] suggested that the average of 20 trials can yield a stable ERP with high reliability while Rietdijk, Franken, and Thurik [36] suggested that it could be as low as 14 trials, and Jensen and MacDonald [37] suggested 13 trials is the minimum trials needed to achieve high power (> 0.8) for 16 participants in a within-participants experiment. Others argued on the appropriate trial for averaging ERP and cannot come up with a conclusion. Certainly, a greater number of trials could give more stability and reliability to the ERP obtained. However, the number of trials itself cannot be considered alone. There are many factors such as sample size and fatigue caused during the experiment can also affect the quality of EEG.

In the future, we plan to perform another experiment on no button press task with a larger trial number. We will also introduce the eye-tracker to confirm our EEG results with the eye gaze.

ACKNOWLEDGMENT

This work was supported in part by JST Moonshot R&D Grant Number JPMJMS2237.

REFERENCES

[1] J. Tanaka and M. Farah, "Parts and wholes in face recognition," *The Quarterly journal of experimental psychology. A, Human experimental psychology*, vol. 46, pp. 225–45, 1993. DOI: 10.1080/14640749308401045.

[2] K. M. Curby and I. Gauthier, "A visual short-term memory advantage for faces," *Psychonomic Bulletin & Review*, vol. 14, no. 4, pp. 620–628, 2007, ISSN: 1531-5320(Electronic),1069-9384(Print). DOI: 10.3758/BF03196811.

[3] K. Curby, K. Glazek, and I. Gauthier, "A visual short-term memory advantage for objects of expertise," *Journal of experimental psychology. Human perception and performance*, vol. 35, pp. 94–107, 2009. DOI: 10.1037/0096-1523.35.1.94.

[4] T. Nguyen and K. Pezdek, "Memory for disguised same- and cross-race faces: The eyes have it," *Visual Cognition*, pp. 1–8, 2017. DOI: 10.1080/13506285.2017.1329762.

[5] D. J. Carragher and P. J. B. Hancock, "Surgical face masks impair human face matching performance for familiar and unfamiliar faces," *Cognitive Research: Principles and Implications*, vol. 5, no. 1, p. 59, 2020, ISSN: 2365-7464. DOI: 10.1186/s41235-020-00258-x.

[6] S. Lees, N. Dayan, H. Cecotti, *et al.*, "A review of rapid serial visual presentation-based brain-computer interfaces," *J. Neural Eng*, vol. 15, no. 2, p. 021 001, 2018, ISSN: 1741-2552. DOI: 10.1088/1741-2552/aa9817.

[7] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors (Basel, Switzerland)*, vol. 12, no. 2, pp. 1211–1279, 2012, ISSN: 1424-8220. DOI: 10.3390/s120201211.

[8] R. A. Ramadan and A. V. Vasilakos, "Brain computer interface: Control signals review," *Neurocomputing*, vol. 223, pp. 26–44, 2017, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2016.10.024.

[9] S. Sur and V. K. Sinha, "Event-related potential: An overview," *Ind Psychiatry J*, vol. 18, no. 1, pp. 70–3, 2009, ISSN: 0972-6748 (Print) 0972-6748. DOI: 10.4103/0972-6748.57865.

[10] G. F. Woodman, "A brief introduction to the use of event-related potentials in studies of perception and attention," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2031–2046, 2010, ISSN: 1943-393X. DOI: 10.3758/BF03196680.

[11] L. A. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr Clin Neurophysiol*, vol. 70, no. 6, pp. 510–23, 1988, ISSN: 0013-4694 (Print) 0013-4694. DOI: 10.1016/0013-4694(88)90149-6.

[12] C. Joyce and B. Rossion, "The face-sensitive N170 and VPP components manifest the same brain processes: The effect of reference electrode site," *Clin Neurophysiol*, vol. 116, no. 11, pp. 2613–31, 2005, ISSN: 1388-2457 (Print) 1388-2457. DOI: 10.1016/j.clinph.2005.07.005.

[13] M. Eimer, "The face-sensitive N170 component of the event-related brain potential," in *The Oxford Handbook of Face Perception*. 2012, pp. 329–344. DOI: 10.1093/oxfordhb/9780199559053.013.0017.

[14] B. Cai, S. Xiao, L. Jiang, Y. Wang, and X. Zheng, "A rapid face recognition BCI system using single-trial ERP," in *2013 6th International IEEE/EMBS Confer-*

ence on Neural Engineering (NER), pp. 89–92, ISBN: 1948-3554. DOI: 10.1109/NER.2013.6695878.

[15] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, "Electrophysiological studies of face perception in humans," *Journal of cognitive neuroscience*, vol. 8, no. 6, pp. 551–565, 1996, ISSN: 1530-8898 0898-929X. DOI: 10.1162/jocn.1996.8.6.551.

[16] M. Eimer, "Does the face-specific N170 component reflect the activity of a specialized eye processor?" *Neuroreport*, vol. 9, no. 13, pp. 2945–8, 1998, ISSN: 0959-4965 (Print) 0959-4965. DOI: 10.1097/00001756-199809140-00005.

[17] D. Nemrodov, T. Anderson, F. F. Preston, and R. J. Itier, "Early sensitivity for eyes within faces: A new neuronal account of holistic and featural processing," *NeuroImage*, vol. 97, pp. 81–94, 2014, ISSN: 1053-8119 (Print) 1053-8119. DOI: 10.1016/j.neuroimage.2014.04.042.

[18] G. Prete, A. D'Anselmo, and L. Tommasi, "A neural signature of exposure to masked faces after 18 months of covid-19," *Neuropsychologia*, vol. 174, p. 108 334, 2022.

[19] A. Żochowska, P. Jakuszyk, M. M. Nowicka, and A. Nowicka, "Are covered faces eye-catching for us? the impact of masks on attentional processing of self and other faces during the covid-19 pandemic," *Cortex*, vol. 149, pp. 173–187, 2022, ISSN: 0010-9452. DOI: https://doi.org/10.1016/j.cortex.2022.01.015.

[20] G. A. Rousselet, R. A. Ince, N. J. van Rijsbergen, and P. G. Schyns, "Eye coding mechanisms in early human face event-related potentials," *Journal of Vision*, vol. 14, no. 13, pp. 7–7, 2014.

[21] R. A. Ince, K. Jaworska, J. Gross, *et al.*, "The deceptively simple N170 reflects network information processing mechanisms involving visual feature coding and transfer across hemispheres," *Cerebral Cortex*, vol. 26, no. 11, pp. 4123–4135, 2016.

[22] M. Boksem, T. Meijman, and M. Lorist, "Effects of mental fatigue on attention: An ERP study," *Brain research. Cognitive brain research*, vol. 25, pp. 107–16, 2005. DOI: 10.1016/j.cogbrainres.2005.04.011.

[23] I. Chanpornpakdi and T. Tanaka, "The role of the eyes: Investigating face cognition mechanisms using machine learning and partial face stimuli," *bioRxiv*, pp. 2023–05, 2023.

[24] A. Gramfort, M. Luessi, E. Larson, *et al.*, "MEG and EEG data analysis with MNE-python," *Frontiers in Neuroscience*, vol. 7, 2013, ISSN: 1662-453X. DOI: 10.3389/fnins.2013.00267.

[25] D. A. Jeffreys and E. S. Tukmachi, "The vertex-positive scalp potential evoked by faces and by objects," *Exp Brain Res*, vol. 91, no. 2, pp. 340–50, 1992, ISSN: 0014-4819 (Print) 0014-4819. DOI: 10.1007/bf00231668.

[26] J. Tanaka and L. Pierce, "The neural plasticity of other-race face recognition," *Cognitive, affective & behavioral neuroscience*, vol. 9, pp. 122–31, 2009. DOI: 10.3758/CABN.9.1.122.

[27] S. J. Wuttke and S. R. Schweinberger, "The p200 predominantly reflects distance-to-norm in face space whereas the N250 reflects activation of identity-specific representations of known faces," *Biol Psychol*, vol. 140, pp. 86–95, 2019, ISSN: 0301-0511. DOI: 10.1016/j.biopsycho.2018.11.011.

[28] E. Mercure, F. Dick, and M. H. Johnson, "Featural and configural face processing differentially modulate ERP components," *Brain Research*, vol. 1239, pp. 162–170, 2008, ISSN: 0006-8993. DOI: 10.1016/j.brainres.2008.07.098.

[29] J. Polich and J. R. Criado, "Neuropsychology and neuropharmacology of P3a and P3b," *International Journal of Psychophysiology*, vol. 60, no. 2, pp. 172–185, 2006.

[30] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007, ISSN: 1388-2457 1872-8952. DOI: 10.1016/j.clinph.2007.04.019.

[31] S. L. Bressler and M. Ding, "Event-related potentials," *Wiley encyclopedia of biomedical engineering*, 2006.

[32] T. Roberts and V. Bruce, "Feature saliency in judging the sex and familiarity of faces," *Perception*, vol. 17, no. 4, pp. 475–81, 1988, ISSN: 0301-0066 (Print) 0301-0066. DOI: 10.1068/p170475.

[33] M. Eimer, "The face-specific N170 component reflects late stages in the structural encoding of faces," *Neuroreport*, vol. 11, no. 10, pp. 2319–2324, 2000, ISSN: 0959-4965.

[34] W. Giroldini, L. Pederzoli, M. Bilucaglia, S. Melloni, and P. Tressoldi, "A new method to detect event-related potentials based on pearson's correlation," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2016, pp. 1–13, 2016.

[35] J. Cohen and J. Polich, "On the number of trials needed for P300," *International Journal of Psychophysiology*, vol. 25, no. 3, pp. 249–255, 1997.

[36] W. J. Rietdijk, I. H. Franken, and A. R. Thurik, "Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe," *PloS one*, vol. 9, no. 7, e102672, 2014.

[37] K. M. Jensen and J. A. MacDonald, "Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components," *Psychophysiology*, e14245, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Residual, Mixer, and Attention: The Three-way Combination for Streaming Wake Word Detection Framework

Sattaya Singkul* Theerat Sakdejayont* and Tawunrat Chalothorn*
* Innovation Research and Development, Kasikorn Labs, Nonthaburi, Thailand
E-mail: {sattaya.s,theerat.s,tawunrat.c}@kbtg.tech

***Abstract*—Speech interactions and digital assistants rely on effective wake word detection models to identify predefined words. In this study, we address the need for an efficient wake word detection model and propose a "three-way residual separable convolution network" (3W-ResSC) inspired by human multi-perspective learning. Our 3W-ResSC model combines three information: residual learning (ResNet), point-wise pattern, and depth-wise patterns from a CNN mixer (Mi). In addition, to conclude and emphasize information, our independent multi-view attention (iMVA) is proposed based attention mechanism (At). Based on ResNet, Mi, and At, we propose three different combinations for our model; ResNet+Mi, ResNet+At, and ResNet+Mi+At. By leveraging the strengths of these three combinations, our combinations are investigated to find the outstanding performance and pattern. To evaluate our model, we conduct experiments on multiple datasets, including Gowajee, HeyFireFox, HeySnips, and GSCv2. The results demonstrate that our 3W-ResSC outperforms baseline models in terms of equal error rate (EER), Matthews correlation coefficient (MCC), and false rejection rate (FRR) in multi-class classification. Additionally, we introduce a wake word detection framework specifically designed for streaming processing. This framework leverages an independent speech window, resembling a buffer-like streaming production process. Overall, our proposed 3W-ResSC model and wake word detection framework offer significant advancements in wake word detection, showcasing improved performance and efficiency, which discuss in various datasets and training sizes. These findings contribute to the development of more effective wake word detection and speech processing when using ResNet, Mi, and At in deep learning.**

I. Introduction

Speech interactions and digital assistants built into smartphones and voice command devices require a wake word detection system to anticipate the detection of predefined words [1]. For this, a lightweight model with small memory footprint, small computational cost, and low latency is needed [2] to quickly activate the device as soon as the wake word appears in streaming audio.

In recent work, a deeper model requires high computational resources [3]–[5] for training and inference based on a number of parameters and model complexity, such as Wav2KWS [6] and Keyword Transformer [7], which are not appropriate to deploy on mobile devices with limited processing unit and memory. For smaller model, the recurrent neural network (RNN), is sequential in nature and perform well with sequential frame information. However, it prevents the parallel chunk-streaming process on GPUs from being clogged by previous frames. Therefore, most current systems apply a convolutional neural network (CNN) for a frame-by-frame processing, which does not required previous frames. A CNN kernel is repeatedly applied by sliding through time or frequency and covers a small and fixed frame length. Each kernel can capture local patterns, which can expand the receptive field by stacking CNN layers. The expanded receptive field of the higher layers can "see" longer frame lengths compared to the lower layers, which capture more global patterns. These are why CNN is better utilized resource than RNN and deeper models on devices. However, when learning with highly concentrated deep details, the CNN model falls short. Highly CNN layers lose some information and may experience vanishing gradient problems [4], [8]. By adding function and skipping connections, ResNet [9] solves this problem. Res8 [8] model presents the dilated CNN pattern with ResNet to achieved high accuracy. Though, the Res8 performance depends on a sufficient amount of speech data, which may not be enough when the dataset size is small or complex data is present.

As previously mentioned, we proposed a method to improve the efficiency of Res8 [8] for wake word model. Our "three-way residual separable convolution network (3W-ResSC)" is proposed, inspired by the concept of human multi-perspective learning [10]; that is, "the human brain flexibly adapts to support the information-processing needs of different perspectives and can mix to extend performance," as the same way that Jääskeläinen [10] described. Responding to our inspired concept, we implemented a learning method for deep learning with three kinds of information like multi-perspective learning: The first information is the original input with residual learning (ResNet). For the second and third, point- and depth-wise patterns are mixed (Mi) based on 2D separable dilated CNN inspired by [11]. This process involves dividing a single convolution with dilation factor into two or more convolutions to produce the same output, enabling a more informative view for learning with a minimal computational requirements. Also, to conclude the human multi-perspective learning, our independent multi-view attention (iMVA) is proposed based on attention mechanism (At) that inspired by traditional MVA [12]. Our iMVA captures and learns three side-view components including channel, global, and local attention, specifically to emphasize and conclude the learning information. There-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

fore, our 3W-ResSC have three-way combinations to adding information like a human multi-perspective learning such a ResNet+Mi, ResNet+At, and ResNet+Mi+At. Furthermore, our wake word model is proposed for streaming processing with our wake word detection framework, which scope information for learning in an independent speech window as a buffer-like streaming production process. Besides, to prove the overall performance of our wake word model, we experiment with four datasets: Gowajee [13] for wake word data with one sequence and a low training size, HeyFireFox [14] for data with three sequences and a medium training size, HeySnips [15] for data with two sequences with a large training size, and Google Speech Command V2 (GSCv2) [16] for data with many keywords and a large training size. Furthermore, to provide further performance on varying training sizes, we specifically sampled HeySnips into HeySnipsV.Small and HeySnipsV.Medium to represent a small and medium training sizes.

Our main contributions of this paper are as follows: (i) 3W-ResSC wake word model was proposed. 3W-ResSC block and 3W-ResSC learning block (3W-ResSC-LB) were redesigned its internal network from Res8 which were imitated from human multi-perspective learning. (ii) The evaluation of overall and various training size performance objectives was proposed. (iii) Streaming inference for wake word detection framework was proposed.

II. Our Wake Word Detection Framework Based on Streaming Inference Support

Our wake word detection framework, a frame-level based system with simulated streaming inference, is shown in Fig. 1. The framework consists of three parts: (i) speech alignment for training, (ii) feature extraction, and (iii) wake word model.

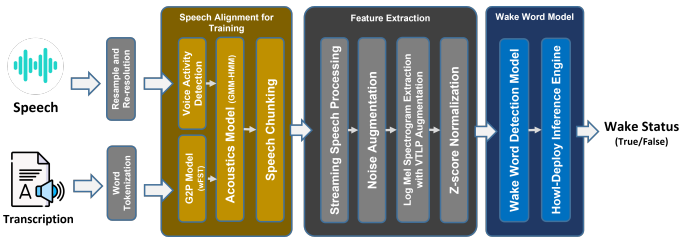


Fig. 1: A proposed wake word detection framework based on streaming inference support. Note that the transcription input is not required in the inference process.

A. Speech Alignment for Training

In general, wake word detection requires speech alignment for the training process to scope information before being chunked. This can provide word- or phoneme-based alignment. Phoneme-based alignment [17] involves intricate phonetic segmentation and matching, introducing computational complexity. In contrast, word-based alignment [17] offers computational efficiency by working with complete words. So, we focus on word-based alignment and aligned information is chunked to frame-based. Speech alignment requires two data inputs: speech and transcription. The speech data is resampled to 16 kHz for normal quality, and transcription is tokenized to word level for aligned speech preparation. We build our wake word detection framework on top of the speech alignment name's Montreal forced aligner (MFA)[1] [18]. The MFA, a traditional method that is processed in the same way as Kaldi frameworks, is chosen for its popularity with good performance and free license. It is used for tasks such as voice activity detection from energy-based to filter silence windows, grapheme-to-phoneme (G2P) model from weighted finite-state transducer (wFST) to extract phonetic of word, and acoustics model from Gaussian mixer model with hidden Markov model (GMM-HMM) to extract probability of sequential speech. The MFA version 2.0 is chosen for this experiment. Finally, the MFA is adapted to our framework, the speech is aligned and chunked for the feature extraction process.

B. Feature Extraction

As is commonly done in streaming speech processing, we chunked the speech before extracting features by frame-based length. The window size of 500 frames is set. The speech window uses noise augmentation by mixing the Microsoft scalable noisy speech dataset (MS-SNSD) [19] with original speech window. Then, the feature extraction from speech window is performed by using log Mel spectrogram (LMS), the parameters is 40 Mel filterbanks, 512 fast Fourier transform (FFT) size with 257 Fourier bins, and 200 hop length. All of these made the feature robust and streaming support. Besides, to learn more perspectives, the vocal tract length perturbation (VTLP) [20], [21] is applied to simulated new vocal track information in speech and increase the informative number of speakers. Finally, the feature is normalized by z-score to improved generalization, which is called Z-LMS. In summary, the previously mentioned feature dimension is define as "(batch, channel, time, number of Mel bin)".

C. Wake Word Model

Our 3W-ResSC, based on a non-sequential lightweight model, is performed by applying ResNet, mixer (Mi), and attention (At) to separable dilated CNN for learning. After that, our model is evaluated by the Howl-deploy inference engine[2] [14] to simulate streaming inference process use cases. This engine simulates the streaming inference process based on the PyAudio stream library. For the inference process, we use original speech data as input for PyAudio streaming pipeline, while the aligned speech data is not used, as if it were the actual streaming production process. The speech input is repeated play based on window length that is close to speech input from device. Then, the window is forwarded to feature extraction and wake word model process, respectively, to detect the wake word. In this paper, we set a 500 frames for window length and 16 window buffer size for the streaming process.

[1]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner.git
[2]https://github.com/castorini/howl.git

III. Our 3W-ResSC Model

We propose a model to improve Res8 performance by capturing more information based on human multi-perspective learning like three information. The first information captures overall information from original speech, like ResNet. Second and third information capture depth- and point- wise patterns to mix information using the CNN mixer concept (Mi). The original CNN mixer was initially proposed for image processing and harnesses the advantages of depth-wise features from the RGB channel, as well as point-wise features for each image layer. In our work, unlike the original CNN mixer, we do not utilized the RGB color channel. Instead, we extract the depth-wise features from the Mel bins [8], as explained in II-B subsection. Then, to conclude on a specific component from three information of ResNet and mixer, our independent multi-view attention (iMVA) is proposed based attention mechanism (At), that inspired by traditional MVA [12]. The traditional MVA is designed for the noise erasure task that is not for wake word detection, which is different from our task challenges. In our work, we adapt the concept of MVA to wake word detection task, specifically by scoping input into window levels that allow us to avoid the future context requirement. Thus, the attention mechanism of our iMVA is specifically focused on independent learning at the window level. Building upon these concepts, our 3W-ResSC is proposed with three combinations including ResNet+Mi, ResNet+At, ResNet+Mi+At.

First and foremost, we introduce the ResNet+Mi based 3W-ResSC block in Fig. 2 (purple rectangle). Dilated CNN [11], [22] with ResNet [9] is the main process in our model which use the convolution operation [22] to extract hidden feature and learning. Assume the input feature to a our model is $x_t = \{x_1, ..., x_T\} \in R^T$. The key operation of our model is the mapping feature of dilated convolution and is expressed as

$$F(x_{m,n}) = (x_{m,n} *_l k_{m,n}) = \sum_{j=-\inf}^{\inf} \sum_{i=-\inf}^{\inf} x_{i,j} \cdot k_{m-l_i,n-l_j} \tag{1}$$

where m, n indicates the 2D-position of $M \times N$ matrix, $x_{m,n}$ is the 2D Z-LMS feature input at m and n, $k_{m,n}$ is the kernel function at m and n, and $*_l$ is the dilated convolution [11] where l is a positive integer value called the dilation factor that apply the $k_{m,n}$ over the $x_{m,n}$ input. Further, to reduce loss of output information and overfitting problem, the ResNet concept is introduced to easier learn the residual by adding previous information to output information as $F(x^{in}_{m,n}) + x^{out}_{m,n}$, where $x^{in}_{m,n}$ denotes the feature input that refers to the $x_{m,n}$ above, $F(x^{in}_{m,n})$ is the feature output from convolution network, and $x^{out}_{m,n}$ is the input from ResNet's skip-connection (depicted by the green line), as refer to a sequence in Figure 2. This ResNet is the first information in our model.

To extend the mixed information (Mi) for the second and third information, in mathematically, commutative is fundamental property of convolution $(x_n * k_n = k_n * x_n)$, then

$$F(x^{in}_{m,n}) = (k_{m,n} *_l x^{in}_{m,n}) = \sum_{j=-\inf}^{\inf} \sum_{i=-\inf}^{\inf} k_{l_i,l_j} \cdot x^{in}_{m-i,n-j} \tag{2}$$

and, $k_{m,n}$ is separated to $M \times 1$ and $1 \times N$ as expressed in (3).

$$k_{m,n} = k^{depth}_m \cdot k^{point}_n \tag{3}$$

where k^{depth}_m indicates the kernel at m in a depth-wise pattern and k^{point}_n indicates the kernel at n in a point-wise pattern. Therefore, the 2D dilated convolution of (2) is modified to perform twice the dilated convolution by (3) in depth-wise and point-wise kernels same as separable convolution concept [23]:

$$(k_{m,n} *_l x^{in}_{m,n}) = \sum_{j=-\inf}^{\inf} k^{point}_{l_j} \sum_{i=-\inf}^{\inf} k^{depth}_{l_i} \cdot x^{in}_{m-i,n-j} \tag{4}$$

From (4), a convolution consists of two computed kernels. A depth-wise kernel (a spatial convolution performed independently over each channel of an input) is followed by a point-wise kernel (1x1 convolution), projecting the channel's output by the depth-wise convolution onto a new channel space.

Thus, to capture a point-wise pattern, a point-wise kernel is used instead of the normal kernel of (1) as expressed in (5).

$$f(x^{in}_{m,n}) = (x^{in}_{m,n} *_l k^{point}_{m,n}) = \sum_{j=-\inf}^{\inf} \sum_{i=-\inf}^{\inf} x^{in}_{i,j} \cdot k^{point}_{m-l_a,n-l_j} \tag{5}$$

In the same way, we used depth-wise kernel instead of the normal kernel of (1) as expressed in (6).

$$F(x^{in}_{m,n}) = (x^{in}_{m,n} *_l k^{depth}_{m,n}) = \sum_{j=-\inf}^{\inf} \sum_{i=-\inf}^{\inf} x^{in}_{i,j} \cdot k^{depth}_{m-l_a,n-l_j} \tag{6}$$

The output information is then embedded by adding functions with three information to improve Res8 performance, can be defined as

$$s_{m,n} = F(x^{in}_{m,n}) + f(x^{in}_{m,n}) + x^{out}_{m,n}. \tag{7}$$

$F(x^{in}_{m,n})$ is used to highlight the important prosody information of the Z-LMS feature (depth-wise CNN), then $x^{out}_{m,n}$ from ResNet skip-connection is used for reconstruction to reduce the lossy information, and $f(x^{in}_{m,n})$ deeply extracts the convolution output into point-wise information from the same $F(x^{in}_{m,n})$ input. All of which is a ResNet+Mi combination, as shown in the 3W-ResSC block (purple rectangle) of Fig. 2. In addition, we employ the Swish activation function [24] instead of ReLU to retain more informative negative values, coupled with the GELU activation function [25] for nonlinear transformations within the point-wise pattern. The depth-wise CNN involves significant computational complexity. Swish activation aids in the smoother flow of gradients during backpropagation, potentially resulting in faster convergence and enhanced training

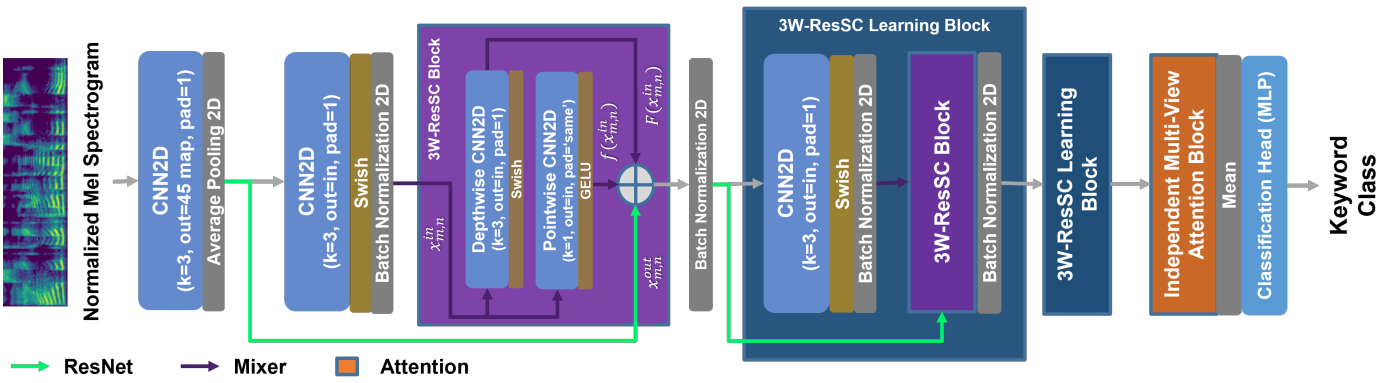


Fig. 2: Our proposed 3W-ResSC wake word model.

stability. Meanwhile, the use of GELU in the point-wise CNN has gained popularity due to its effectiveness in handling a broad range of activation values.

In addition, to emphasize and summarize the three information, attention (At) is applied in the final layer before the classification head to increase capturing and concluding capabilities. MVA based on attention mechanism is used to specifically capture the three side-view components [12] such as channel, global, and local attentions. Channel attention emphasizes representations from compressed time with channels. Global and local attention implemented through dual paths, efficiently reflect the long sequential features. However, in our work, we utilize iMVA instead of MVA due to the limitation of wake word detection in terms of streaming window forwarding, which is a crucial for proper real-world application in wake word detection area. In the real-world application, the speech window consistently forwarded to the wake word model at all times, leading to uncertainty about when the speech will end. As a result, long sequential features cannot be effectively utilized in this task. Therefore, our iMVA is proposed specifically for wake word detection tasks ensuring proper functionality, our iMVA employ the feature input in terms of window level instead of a long sequential feature. It supports streaming by learning the independent speech windows. Regardless, for efficient global and local attentions, overlapped chunking is adopted at the window level. Overlapped chunking involves chunking the window to a sequential sub-window, that can provide information relate to the long sequential feature of the original MVA, but in our iMVA is the sequential sub-window feature instead. The details of iMVA implementation setting refer to the original MVA [12]. As a result, our iMVA does not require future context for learning and prediction, as shown in Figure 3.

In our work, our 3W-ResSC has three-way combinations including ResNet+Mi, ResNet+At, and ResNet+Mi+At. ResNet+Mi incorperates three kinds of information like a human multi-perspective learning. Also, the added At is used for the informative conclusion and emphasis. ResNet+At and ResNet+Mi+At are proposed to investigate the impact of informative conclusion and emphasis effects on the perfor-

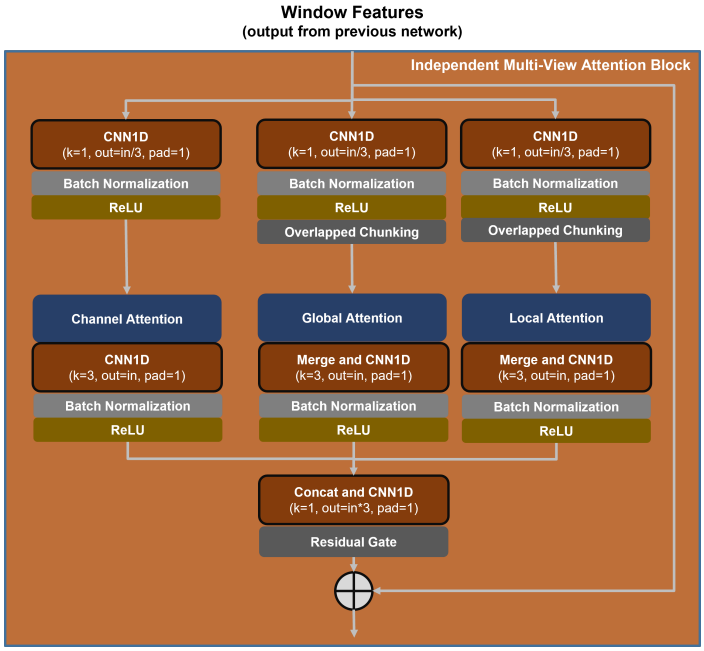


Fig. 3: A proposed independent multi-view attention (iMVA) block. Note that the input is window features from the previous output 3W-ResSC learning block.

mance of our model. Above mentioned are three combinations of "3W-ResSC" wake word model, the full combination (ResNet+Mi+At) as shown in Fig. 2. Besides, for the model parameter and computation (multiply-accumulate operations: MAC), Table I is reported. ResNet+Mi is the best with 56K parameters and 8M MACs. ResNet+At has 161K parameters and 20M MACs. ResNet+Mi+At has 161K parameters and 20M MACs. Although the addition of At in combination necessitates a larger model size and increase computation, which can be a double-edged sword. We refer the definition provided in [26] to establish the criteria for a streaming and effective wake word detection task. According to the baseline for streaming wake word detection of [26], the model should have fewer than 250K parameters and less than 101M MACs. Therefore, all of our 3W-ResSC meets the streaming wake

word detection task requirement, especially in ResNet+Mi. Our model is configured with the following hyper-parameters: 300 max epochs with saving, 0.001 learning rate, 0.98 learning rate decay, and Adam optimizer.

TABLE I: A number of parameters and model complexity of our wake word detection framework and baselines.

Model	Parameters	MACs
Res8 (ResNet) [8], [14]	109.84K	14.91M
MobileNetV2 [14], [27]	2,226.47K	9.21M
ConvMixer [28], [29]	119.20K	22.73M
Streaming Transformer [2]	58.59K	37.17M
3W-ResSC (ResNet+Mi)	56.43K	8.02M
3W-ResSC (ResNet+At)	161.80K	20.96M
3W-ResSC (ResNet+Mi+At)	161.84K	20.97M

IV. Experiments

The experiments were conducted on a desktop computer running Windows 10 OS, equipped with an Intel Core i7-11370H processor, 32 GB of RAM, and an NVIDIA GeForce RTX 3070 GPU with 8 GB of memory. Inference was simulated using our streaming framework, limited to multi-thread CPU without GPU utilization. For dataset experiments, we have the overall performance and training size performance objectives. The overall performance objective is used for our model and all baselines while the training size performance objective is used to investigate our combination model. In evaluation metrics, we use the EER and MCC for balanced performance evaluation. Also, FRR is used for multi-class classification. For model compression, all models in our study are dynamic quantized with Pytorch precision 16-bit.

A. Datasets

Four datasets were experimented to prove the model's performance for the waked keyword and dataset size challenges [30]. HeyFireFox and HeySnips were used to prove the more than one sequence wake keyword challenge, the prediction result will be true if all the sequential prediction result is an exact pattern. GSCv2 was used to prove the multiple wake keyword challenge similar to a multi-class classification. Additionally, Gowajee, a Thai language dataset, focuses on the one wake keyword with performance when using a small data size with a new accent, due to the limited resources on speech processing tasks.

HeyFireFox contains three sub-wake keywords: "Hey", "Fire", and "Fox". The dataset is split into two sets; the positive set contains only the denoted sequence in one file, and the negative sets are the rest. HeySnips and Gowajee are split like HeyFirefox, with the sequences "Hey Snips" and "Gowajee", respectively. For GSCv2, each file contains only one keyword. We select positive set as ten classes like a multi-classification including "yes", "no", "up", "down", "left", "right", "on", "off", "stop", and "go". The other classes are changed to "unknown", like a negative set. Besides, for the dataset size challenges, we set 3 criteria to define the training dataset size including small, medium, and large. In the positive set, the small size is less than 350 train utterances, the medium is between 350 to 3000, and the large is more than 3000. For the negative set, the small size is less than 10000, medium is between 10000 to 20000, and large is more than 20000. However, for the training process, the MFA is used for speech alignment. HeyFireFox, HeySnips, and GSCv2 datasets are in English. The nature of English transcription contains white space for word boundaries which easier to tokenize. In contrast, Thai transcription does not have. Therefore, in this experiment, the Attacut model [31] is used for tokenized Thai transcription, a state-of-the-art in Thai tokenization model that performs well with high speed. Furthermore, four datasets were resampled to a 16 kHz sampling rate and 16-bit PCM resolution. Subsequently, they were partitioned into train, dev, and test sets, following the methodology outlined in the original dataset paper [13]–[16].

In addition, each dataset contains varying uncontrolled factors such as recording system, speaker culture, and wake keyword challenge, that may not directly compare performance with each other datasets. Based on this issue, our models are experimented with two objectives. First is the overall performance based on various datasets which shows the model is robust in any dataset environment. Second is the training size performance based on different training sizes on the same dataset which shows each of our combination models is proper in differ model training limitation. For the second objective, we create HeySnipsV.Small and HeySnipsV.Medium from HeySnips dataset to experiment on this objective. HeySnipsV.Small experiments for small training size limitation that refers the sampling training size to match that of the Gowajee dataset in train and dev sets. HeySnipsV.Medium experiments for medium training size limitation that refers the sampling training size to match that of the HeyFireFox in train and dev sets. Also, the original HeySnips experiments for non training size limitation which has large training size. We conclude the statistical report based on positive and negative sets as shown in Table II.

TABLE II: Statistical dataset based on positive and negative sets.

Dataset	Positive set				Negative set			
	Train	Dev	Test	Size	Train	Dev	Test	Size
Gowajee [13]	336	104	17	small	14,518	1,640	979	medium
HeyFireFox [14]	502	76	54	medium	26,132	2,531	2,504	large
HeySnipsV.Small	336	104	1,121	small	14,518	1,640	6,730	medium
HeySnipsV.Medium	502	76	1,121	medium	26,132	2,531	6,730	large
HeySnips [15]	7,636	1,180	1,121	large	45,474	6,854	6,730	large
GSCv2 [16]	30,769	3,703	4,074	large	54,074	6,278	6,931	large

B. Evaluation Metrics

For evaluation performance of the overall performance and training size performance objectives, the experiments aim to evaluate two metric types; balanced and multi-class metric types. To measure balanced performances, the Matthews correlation coefficient (MCC) and equal error rate (EER) are used.

MCC is a reliable statistical rate that produces a high when the prediction obtains good results in all four categories of confusion matrix [32], as described in (8).

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

where TP represents the true positive value, FP represents the false positive value, TN represents the true negative value, and FN represents the false negative value.

For false errors, the EER reports the same rate as false rejection (FRR) and false acceptance (FAR), well-known in bio-matrices verification tasks. EER uses the inference threshold between 0.00 and 1.00 with a step of 0.05 to fine-tune the best threshold. However, for multi-class performances, the challenge is to avoid many rejection cases [33] in wake word detection. Therefore, only FRR is focused on to report the performance.

C. Results and Discussions

First, we evaluate the model's performance for the overall performance objective (see section IV-A). Table III shows the performance based on various datasets. Our 3W-ResSC (ResNet+Mi) shows the best performance of 0.72% EER and 73.52% MCC based on Gowajee. For HeyFireFox, the best EER of 4.60% on our 3W-ResSC (ResNet+Mi+At) and MCC of 73.67% on our 3W-ResSC (ResNet+At). For HeySnips, our 3W-ResSC (ResNet+At) achieves the best performance at 2.98% EER and 95.62% MCC. In addition, for multi-class classification based on GSCv2 in Table IV, ResNet+Mi achieved the best dev FNR of 1.99% and test FNR of 1.71%, demonstrating its suitability for multi-class classification.

However, in each combination, our 3W-ResSC results still differ, that is proper in different cases. Referring to Table II with previously mentioned results, ResNet+Mi is proper when the positive training size is small and the negative training size is medium, while ResNet+At performs better when the training size is large. ResNet+Mi+At is suitable for medium to large training sizes. Yet, each dataset contains the uncontrolled environment such as recording system, speaker culture, and wake keyword challenge. Based on this, HeySnipsV.Small and HeySnipsV.Medium are experimented for the same dataset in different training sizes, to address the second objective. Table V shows the training size performance based on HeySnips dataset. ResNet+Mi achieved the best performance of 5.04% EER and 91.65% MCC in small training sizes (HeySnipsV.Small). ResNet+Mi+At achieved the best performance of 4.21% EER and 93.72% MCC in medium training sizes (HeySnipsV.Medium). ResNet+At acheived the best performance of 2.98% EER and 95.62% MCC in large training sizes (HeySnips). As these training size result patterns illustrated in Figure 4 for EER and Figure 5 for MCC, ResNet+Mi is suitable for small, ResNet+Mi+At is suitable for medium, and ResNet+At is suitable for large training size. Besides, if the training size is increased, the combination of our 3W-ResSC can extend performance. ResNet+At is the best for extend performance when using large training sizes. Additionally, for multi-class classification based on GSCv2 in Table IV, ResNet+Mi achieved the best dev FNR of 1.99% and test FNR of 1.71%, demonstrating its suitability for multi-class classification.

In summary, our study involved experiments with three combinations of human multi-perspective learning concepts, namely ResNet+Mi, ResNet+At, and ResNet+Mi+At, where Mi stands for Mixer and At stands for attention. The results of our study showed that each combination performed best under different circumstances: (1) ResNet+Mi performed best with small training data and showed better performance in terms of multi-class classification on the GSCv2 dataset. (2) ResNet+At performed best with large training data and words with higher wake sequences. (3) ResNet+Mi+At performed best with medium training data and words with higher wake sequences. Furthermore, our study found that adding the At combination to each ResNet combination resulted in a noise-robust model, which is a evident when comparing the results of Res8, ResNet+Mi (without At), ResNet+At, and ResNet+Mi+At (with At).

TABLE III: Wake word detection performance on the test set.

Model	Gowajee		HeyFireFox		HeySnips	
	EER	MCC	EER	MCC	EER	MCC
Res8 [8], [14]	2.50	49.47	9.78	70.88	3.58	92.51
MobileNetV2 [14], [27]	40.81	4.84	7.38	65.81	5.02	92.38
ConvMixer [28], [29]	1.84	55.58	10.20	61.78	6.27	89.00
StreamingTransformer [2]	26.64	15.50	20.83	57.82	8.89	85.52
3W-ResSC (ResNet+Mi)	**0.72**	**73.52**	13.40	68.42	3.90	92.46
3W-ResSC (ResNet+At)	4.22	59.70	9.68	**73.67**	**2.98**	**95.62**
3W-ResSC (ResNet+Mi+At)	0.92	69.05	**4.60**	69.02	4.05	94.06

TABLE IV: Multi-class wake word classification performance based on GSCv2 dataset.

Model	Dev FRR	Test FRR
Res8 [8], [14]	2.48	2.12
MobileNetV2 [14], [27]	14.38	14.88
ConvMixer [28], [29]	3.71	4.38
StreamingTransformer [2]	7.54	7.69
3W-ResSC (ResNet+Mi)	**1.99**	**1.71**
3W-ResSC (ResNet+At)	2.80	2.64
3W-ResSC (ResNet+Mi+At)	2.19	2.14

TABLE V: Wake word detection performance with varied training sizes based on the HeySnips test set (original size).

Model	HeySnipsV.Small		HeySnipsV.Medium		HeySnips	
	EER	MCC	EER	MCC	EER	MCC
3W-ResSC (ResNet+Mi)	**5.04**	**91.65**	4.48	92.05	3.90	92.46
3W-ResSC (ResNet+At)	7.26	86.37	7.18	88.82	**2.98**	**95.62**
3W-ResSC (ResNet+Mi+At)	5.78	89.26	**4.21**	**93.72**	4.05	94.06

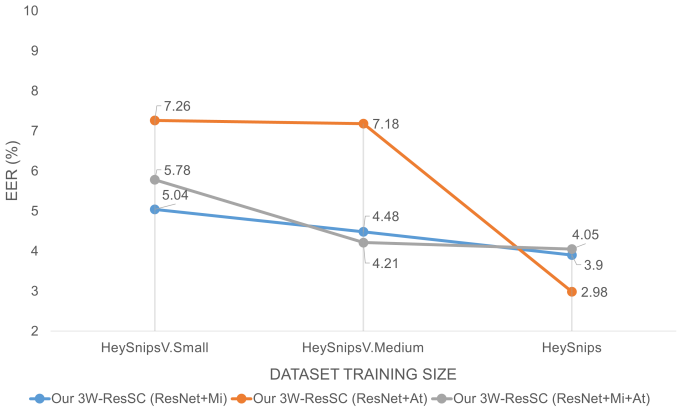


Fig. 4: The trend of EER evaluation results in our 3W-ResSC model with varied training sizes based on HeySnips dataset.

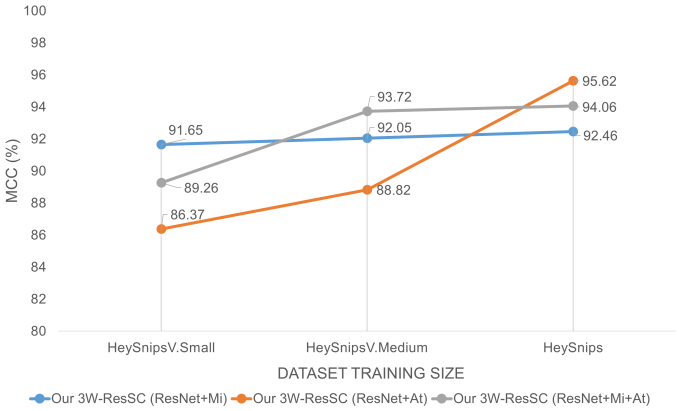


Fig. 5: The trend of MCC evaluation results in our 3W-ResSC model with varied training sizes based on HeySnips dataset.

V. Conclusions

In this paper, our 3W-ResSC wake word model is proposed to improve the Res8 model, which is imitated from the human multi-perspective learning concept like a three information. First information is ResNet. Second and third information are depth- and point- wise from CNN mixer (Mi) concept. In addition, for informative conclusion and emphasis, our iMVA based on attention mechanism (At) is proposed. Thus, from previously mentioned ResNet, Mi, and At, our 3W-ResSC has three-way combinations including ResNet+Mi, ResNet+At, ResNet+Mi+At. Our experiments on Gowajee, HeyFireFox, HeySnips, and GSCv2 showed that our 3W-ResSC significantly outperforms the baselines, which is related to our combinations. ResNet+Mi performs best with small training size and show better performance in multi-class classification task. ResNet+At perform best with large training sizes and ResNet+Mi+At perform best with medium training sizes. All of these make our proposed combination useful for training with limitations. Additionally, to support inference real-world application, our wake word detection framework is proposed to guide the training and inferencing in our model. We believe that this study of our 3W-ResSC with combinations is an effective tool for wake word detection, and may perhaps be used in future audio-focused machine learning work.

Acknowledgment

This work was supported by Kasikorn BusinessTechnology Group (KBTG).

References

[1] Y. Wang, X. Fan, I.-F. Chen, Y. Liu, T. Chen, and B. Hoffmeister, "End-to-end anchored speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 7090–7094.

[2] Y. Wang, H. Lv, D. Povey, L. Xie, and S. Khudanpur, "Wake word detection with streaming transformers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5864–5868.

[3] J. Amoh and K. M. Odame, "An optimized recurrent unit for ultra-low-power keyword spotting," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–17, 2019.

[4] S. Singkul, T. Chatchaisathaporn, B. Suntisrivaraporn, and K. Woraratpanya, "Deep residual local feature learning for speech emotion recognition," in *International Conference on Neural Information Processing*, Springer, 2020, pp. 241–252.

[5] S. Singkul and K. Woraratpanya, "Vector learning representation for generalized speech emotion recognition," *Heliyon*, vol. 8, no. 3, e09196, 2022.

[6] D. Seo, H.-S. Oh, and Y. Jung, "Wav2kws: Transfer learning from speech representations for keyword spotting," *IEEE Access*, vol. 9, pp. 80 682–80 691, 2021.

[7] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," in *Interspeech 2021*, ISCA, 2021, pp. 4249–4253.

[8] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5484–5488.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] I. P. Jääskeläinen and V. Kosonogov, "Perspective taking in the human brain: Complementary evidence from neuroimaging studies with media-based naturalistic stimuli and artificial controlled paradigms," *Frontiers in Human Neuroscience*, vol. 17, 2023.

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[12] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7842–7846.

[13] E. Chuangsuwanich, A. Suchato, K. Karunratanakul, *et al.*, "Gowajee Corpus," Chulalongkorn University, Faculty of Engineering, Computer Engineering Department, Tech. Rep., Dec. 2020, Version 0.9.2. [Online]. Available: https://github.com/ekapolc/gowajee_corpus.

[14] R. Tang, J. Lee, A. Razi, *et al.*, "Howl: A deployed, open-source wake word detection system," in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 2020, pp. 61–65.

[15] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6341–6345.

[16] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[17] S. F. Solórzano and R. Coto-Solano, "Comparison of two forced alignments systems for aligning bribri speech," *CLEI Electronic Journal*, vol. 20, no. 1, pp. 2–1, 2017.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[19] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. Interspeech 2019*, pp. 1816–1820, 2019.

[20] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, International Speech Communication Association (ISCA), 2014, pp. 16–23.

[21] K. Wongpatikaseree, S. Singkul, N. Hnoohom, and S. Yuenyong, "Real-time end-to-end speech emotion recognition with cross-domain adaptation," *Big Data and Cognitive Computing*, vol. 6, no. 3, p. 79, 2022.

[22] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495, 2017.

[23] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[24] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[25] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[26] M. O. Khursheed, C. Jose, R. Kumar, G. Fu, B. Kulis, and S. K. Cheekatmalla, "Tiny-crnn: Streaming wake-word detection in a low footprint setting," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 541–547.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[28] A. Trockman and J. Z. Kolter, "Patches are all you need?" *arXiv preprint arXiv:2201.09792*, 2022.

[29] D. Ng, Y. Chen, B. Tian, Q. Fu, and E. S. Chng, "Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3603–3607.

[30] A. Ghosh, M. Fuhs, D. Bagchi, B. Farahani, and M. Woszczyna, "Low-resource low-footprint wake-word detection using knowledge distillation," *arXiv preprint arXiv:2207.03331*, 2022.

[31] P. Chormai, P. Prasertsom, and A. Rutherford, "Attacut: A fast and accurate neural thai word segmenter," *arXiv preprint arXiv:1911.07056*, 2019.

[32] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[33] V. Z. Këpuska and T. Klein, "A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 71, no. 12, e2772–e2789, 2009.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Audio-to-Score Singing Transcription Based on Joint Estimation of Pitches, Onsets, and Metrical Positions With Tatum-Level CTC Loss

Tengyu Deng, Eita Nakamura, and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Japan
E-mail: deng@sap.ist.i.kyoto-u.ac.jp, eita.nakamura@i.kyoto-u.ac.jp, yoshii@i.kyoto-u.ac.jp

***Abstract*—This paper describes an end-to-end singing transcription method that directly estimates a musical score (a sequence of sung notes with metrical positions) from a music audio signal. The monotonicity of audio-to-score mapping naturally calls for the use of connectionist temporal classification (CTC). Inspired by the success of character-level automatic speech recognition, previous studies on CTC-based music transcription represent a musical score as a sequence of various kinds of symbols (*e.g.*, note pitches and values and barlines) defined in some music notation. Such a naive notation-respecting representation, however, does not fit the non-overlapping monotonic audio-to-symbol alignment and the positions of barlines and the durations of beats tend to be incoherent in the estimated score. To solve this problem, we propose a tatum-level singing transcription method that jointly estimates the pitch (including rest), onset flag, and metrical position at each tatum. Our approach enables the tatums to be monotonically aligned with regularly-spanned intervals of the music signal and the estimated notes are located on the estimated metrical positions that are encouraged to be periodic. Experimental results clearly showed that our proposed model reached comparable accuracy on score-level singing transcription with only unaligned training data, and the proposed tatum-level representation significantly improved the stability of the metrical structures in the estimated scores.**

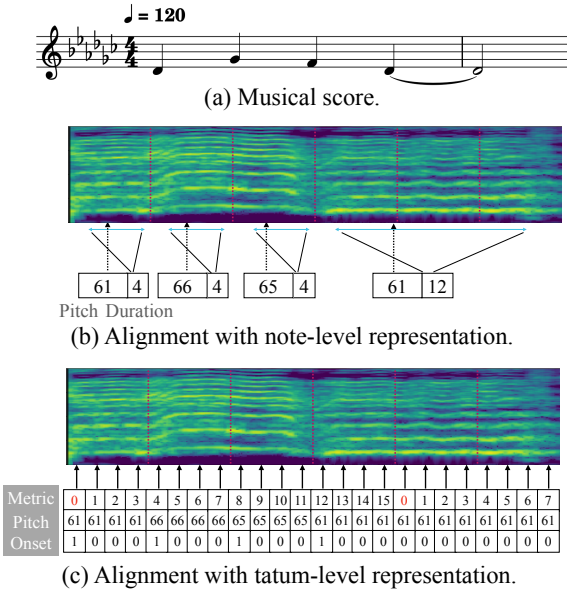


Fig. 1. Comparison of (b) conventional and (c) proposed methods of CTC-based audio-to-score transcription.

I. Introduction

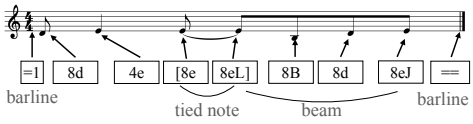
Automatic singing transcription (AST) is one of the most fundamental music analysis tasks in the field of music information retrieval (MIR). Its ultimate goal is to estimate a *musical score* of the vocal part (described in the MusicXML format) from a music audio signal. The estimated scores can be used for melody-based music search (*e.g.*, query-by-humming [1], [2]), score-guided singing voice separation [3], [4], and interpretable emotion recognition [5].

Most studies on AST have attempted to estimate a *piano roll* of the vocal part (described in the MIDI format) [6]–[10], where each sung note is represented by a semitone-level pitch and onset and offset times in seconds. The basic strategy is to estimate a continuous contour of fundamental frequencies (F0s) in Hz (*a.k.a.*, melody extraction) with a deep neural network (DNN) and then segment it to a series of sung and rest notes (discontinuous step function) with a note tracking method [7], [8]. Note that the piano roll differs from the musical score in the time units of the note onsets and offsets. The onsets and offsets are represented in frames or seconds in the piano roll representation, whereas they are represented in quantized units such as beats or fractions of beats in the score representation. To obtain a musical score from a piano roll, rhythm transcription methods [11] are required for quantizing the onset and offset times of notes onto a grid represented in musical metrics such as beats or tatums.

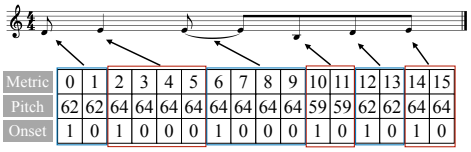
Only a few studies have tackled audio-to-score transcription [12], [13], which typically involve audio-to-MIDI transcription and MIDI-to-audio transcription. Such a cascading approach, however, suffers from several problems. Since F0 annotations aligned with music audio signals are limited both in size and variation, the accuracy of supervised F0 estimation is hard to improve. The complicated dynamics of the F0 contour (*e.g.*, vibrato and glissando) make it hard to perform frequency and temporal quantization without referring to the audio data [12]. This calls for the end-to-end (*i.e.*, audio-to-score) approach to AST that can make effective use of non-aligned pairs of musical scores and audio signals.

Considering the monotonicity of audio-to-score alignment, connectionist temporal classification (CTC) [14] has been used for audio-to-score automatic music transcription (AMT) [15],

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



(a) The Kern representation.



(b) The proposed tatum-level representation.

Fig. 2. Comparison of score representations.

[16] (Fig. 1). In the analogy with CTC-based automatic speech recognition (ASR) that converts a speech signal to a sequence of characters [17], one may try to convert a music signal to a sequence of various kinds of symbols (*e.g.*, note pitches, note values, and barlines) defined in some music notation (*e.g.*, Lilypond [18] and Kern [19], Fig. 2 (a)). It is, however, unclear how such instantaneous and temporal symbols are monotonically aligned with the audio signal in a non-overlapping manner (Fig. 1 (b)). Note that the basic CTC is intended for instantaneous symbols (*e.g.*, phonemes and note pitches) that can be estimated in local regions. To estimate temporal symbols (*e.g.*, note values and barlines), acoustic features should be aggregated from a nonlocal region. The mixture of these modes harms the stability of CTC-based training, leading to incoherent metrical structure in the estimated score. Indeed, previous CTC-based methods were tested only on synthetic data and evaluations on real audio data have not been reported in the literature.

To solve this problem, we propose a CTC-based AST method that converts a music audio signal to a tatum-level score representation with a metrical structure. The tatum is a minimum time unit on a score, which is assumed to be the sixteenth note length in this study. Representing a musical score as a sequence of tatums, each having a pitch (MIDI note number or 128 (rest)), an onset flag ($\in [0, 1]$), and a metrical position ($\in [0, 15]$) (Fig. 2 (b)), we can achieve stable monotonic audio-to-tatum alignment (Fig. 1 (c)). Specifically, a convolutional recurrent neural network (CRNN) that predicts the probabilities of pitches, onset flag, and metrical positions from a music signal at the frame level is trained such that the CTC loss is minimized. At run-time, the most likely tatum-level sequence is estimated with a Viterbi algorithm that explicitly considers the durations of sung notes and the metrical structure.

II. Related Work

Conventional AST approaches combine frame-level F0 estimation and note tracking, where the piano-roll representation is a transcription target [20], [21]. Recently, deep learning techniques have been used for F0 extraction tasks, remarkably improving the estimation accuracy [6]–[8]. A hidden semi-Markov model (HSMM) [12] was employed to quantize F0 contours into sequences of musical notes. Attempts to directly transcribe the musical signals into note events have also be made [9], using deep neural networks to overcome the error accumulation problem in cascading methods. To improve the accuracy of piano-roll transcription, the CTC loss was used as an auxiliary loss function in addition to the standard frame-wise cross-entropy loss [10].

In the context of AST that aims to estimate musical scores, rhythm transcription methods have been explored to quantize piano-roll representations into musical scores [11], [22], [23]. A hybrid DNN-HSMM model [13], for example, was proposed to avoid depending on the preprocessing step of F0 estimation, in the same way as the DNN-HMM approach to ASR. In these studies, representations of rhythmic information for music transcription have been studied. In general, it is possible to represent rhythmic information either at the tatum level using metrical positions or at the note level using note values. Studies using HMMs have shown that the tatum-level representation generally performs better than the note-level representation because certain logical structures of musical rhythms cannot be represented in the latter [23].

Several recent studies have attempted end-to-end AMT to address the error accumulation issue in cascading methods. Inspired by the success of end-to-end ASR with either the attention mechanisms or CTC, studies on end-to-end AMT have employed both attention mechanism [24] and CTC [15], [16] to directly transcribe audio signals into musical scores. To facilitate sequence-to-sequence learning in these methods, it is necessary to represent a musical score as a sequence of symbols. Previous CTC-based end-to-end AMT methods [15], [16] used a *note-level* score representation, where quantized durations of notes were directly estimated as output symbols. However, these methods have been shown to work only on synthetic audio signals made from MIDI-like data. Besides, an encoder-decoder model with an attention mechanism and a *tatum-level* score representation was also employed [24]. This study proposed a regularized training method that encourages the attention matrix to have monotonic and regular structures, aligning the tatums with the frames in a beat-synchronous manner. However, this approach often failed to find correct alignments because the attention matrix does not necessarily be monotonic in early epochs of non-regularized training.

III. Proposed Method

This section describes the proposed CTC-based end-to-end AST method that considers metrical structure.

A. Problem Specification

Given a music audio signal as input, the goal is to estimate a musical score of sung notes. Let $\mathbf{X} \triangleq \{\mathbf{x}_t\}_{t=1}^{T}$ be a frame-level sequence of acoustic features obtained by stacking the mel-spectrogram of the music signal and that of the singing voice separated with a source separation method called Open-Unmix [25], where T is the number of frames. The input to our system is $\mathbf{X}$.

In this paper, the music signal is assumed to have a time signature of 4/4, and the duration of the sixteenth note is regarded

as the minimum time unit called a tatum, *i.e.*, each measure includes 16 metrical positions (tatums). Let $\mathbf{Y} \triangleq \{y_n\}_{n=1}^N$ be a tatum-level score representation, where N is the number of tatums and $y_n \triangleq (b_n, p_n, o_n)$ is a triplet of a metrical position $b_n \in [0, 15]$ (relative position in the measure), a semitone-level pitch $p_n \in [0, 128]$ represented by a MIDI note number ($p_n = 128$ indicates the rest), and an onset flag $o_n \in \{0, 1\}$. To reconstruct a musical score from $\mathbf{Y}$, we first determine the positions of barlines from $\mathbf{B} \triangleq \{b_n\}_{n=1}^N$ under the assumption of the time signature and then the note pitches and values are determined from $\mathbf{P} \triangleq \{p_n\}_{n=1}^N$ and $\mathbf{O} \triangleq \{o_n\}_{n=1}^N$.

B. End-to-End Training

In the training phase, given paired data of the input $\mathbf{X}$ and the ground-truth output $\mathbf{Y} \triangleq (\mathbf{B}, \mathbf{P}, \mathbf{O})$, we train a DNN with the CTC loss in a supervised manner. The input length is assumed to be longer than the output length ($T > N$), which always holds in AST.

1) Single-Label CTC: We first explain the standard CTC. Suppose we aim to estimate only $\mathbf{B}$ from $\mathbf{X}$ ($\mathbf{P}$ and $\mathbf{O}$ are not considered here). Let K be the number of kinds of original output symbols (metrical positions) (*i.e.*, $K = 16$). The special blank symbol "*" working as a wild card is introduced.

A DNN is trained such that the posterior probability of the *tatum-level* sequence $\mathbf{B}$ given as the ground-truth data, denoted by $p(\mathbf{B}|\mathbf{X})$, is maximized. Let $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t=1}^T$ be a *frame-level* sequence such that $\mathbf{B} = \mathcal{M}(\boldsymbol{\pi})$, where $\pi_t \in [0, K-1] \cup \{*\}$ represents the output symbol at frame t, and $\mathcal{M}(\boldsymbol{\pi})$ is a reduction operator that annexes repeated symbols and removes all blank symbols from $\boldsymbol{\pi}$. If $\mathbf{B} = \{1, 2, 3\}$, for example, $\{1, 1, 1, *, *, 2, *, *, 3, *\}$ is a possible example of $\boldsymbol{\pi}$. The DNN is used for inferring metrical positions in frames as follows:

$$(\boldsymbol{\phi}^{\mathrm{B}}, \boldsymbol{\phi}^{*}) = \mathrm{DNN}(\mathbf{X}), \tag{1}$$

where $\boldsymbol{\phi}^{\mathrm{B}} \triangleq \{\phi_{tk}^{\mathrm{B}}\}_{t=1,k=1}^{T,K-1}$, $\boldsymbol{\phi}^{*} \triangleq \{\phi_t^{*}\}_{t=1}^{T}$, ϕ_{tk}^{B} is the posterior probability of metrical position k when $\pi_t \neq *$, and ϕ_t^{*} is that of $\pi_t = *$ at frame t.

$p(\mathbf{B}|\mathbf{X})$ is obtained by accumulating the posterior probabilities over all possible $\boldsymbol{\pi}$'s (frame-to-tatum alignments) that can be reduced to $\mathbf{B}$ as follows (Fig. 3 (a)):

$$p(\mathbf{B}|\mathbf{X}) = \sum_{\boldsymbol{\pi} \in \mathcal{M}^{-1}(\mathbf{B})} p(\boldsymbol{\pi}|\mathbf{X}), \tag{2}$$

where $\mathcal{M}^{-1}(\mathbf{B})$ is an expansion operator that maps $\mathbf{B}$ to a frame-level sequence and $p(\boldsymbol{\pi}|\mathbf{X})$ is the posterior probability of a sequence $\boldsymbol{\pi}$. Assuming the conditional independence over frames, $p(\boldsymbol{\pi}|\mathbf{X})$ is given by

$$p(\boldsymbol{\pi}|\mathbf{X}) = \prod_{t=1}^{T} p(\pi_t|\mathbf{X}), \tag{3}$$

where $p(\pi_t|\mathbf{X})$ is the local posterior probability given by

$$p(\pi_t|\mathbf{X}) = \begin{cases} (1 - \phi_t^{*})\phi_{t\pi_t}^{\mathrm{B}} & \text{if} \quad \pi_t \neq *, \\ \phi_t^{*} & \text{otherwise.} \end{cases} \tag{4}$$

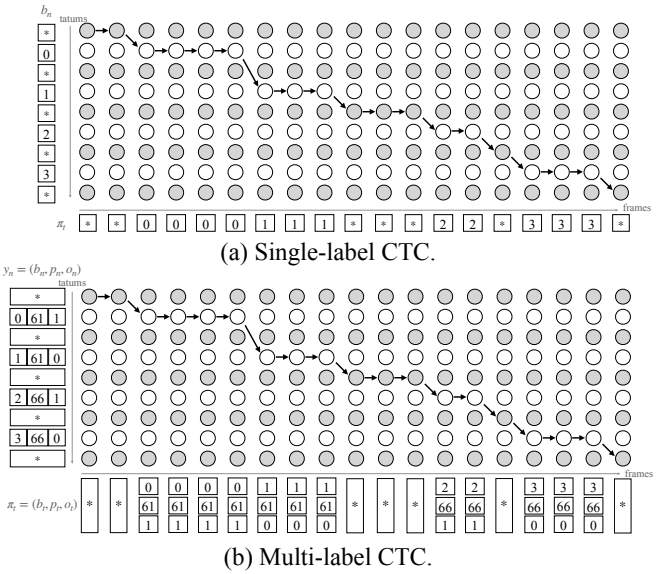


Fig. 3. Possible alignment paths in CTC.

2) Multi-Label CTC: In our study on joint pitch, onset, and metrical estimation, we use the multi-label CTC (MCTC) loss [26]. Let I be the size of the pitch vocabulary ($I = 129$).

A DNN is trained such that the posterior probability $p(\mathbf{Y}|\mathbf{X})$ of the *tatum-level* sequence $\mathbf{Y}$ given as the ground-truth data is maximized. Let $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t=1}^T$ be a *frame-level* sequence (alignment path) such that $\mathbf{Y} = \mathcal{M}(\boldsymbol{\pi})$, where $\pi_t \triangleq (b_t, p_t, o_t) \in \{(k, i, \{0, 1\})\}_{k=1,i=1}^{K-1,I-1} \cup \{*\}$ denotes the output symbol (triplet or $*$) at frame t, and $\mathcal{M}(\boldsymbol{\pi})$ is a reduction operator. The DNN is used for jointly inferring metrical positions, note pitches, and onset flags at the frame level as follows:

$$(\boldsymbol{\phi}^{\mathrm{B}}, \boldsymbol{\phi}^{\mathrm{P}}, \boldsymbol{\phi}^{\mathrm{O}}, \boldsymbol{\phi}^{*}) = \mathrm{DNN}(\mathbf{X}), \tag{5}$$

where $\boldsymbol{\phi}^{\mathrm{P}} \triangleq \{\phi_{ti}^{\mathrm{P}}\}_{t=1,i=1}^{T,I-1}$, $\boldsymbol{\phi}^{\mathrm{O}} \triangleq \{\phi_t^{\mathrm{O}}\}_{t=1}^{T}$, ϕ_{ti}^{P} is the posterior probability of pitch i when $\pi_t \neq *$, and ϕ_t^{O} is that of the onset presence at frame t.

In the same ways as Eq. (2), $p(\mathbf{Y}|\mathbf{X})$ can be computed with dynamic programming (DP) as follows:

$$p(\mathbf{Y}|\mathbf{X}) = \sum_{\boldsymbol{\pi} \in \mathcal{M}^{-1}(\mathbf{Y})} p(\boldsymbol{\pi}|\mathbf{X}), \tag{6}$$

where $\mathcal{M}^{-1}(\mathbf{Y})$ is an expansion operator. Assuming the conditional independence over frames and labels, the posterior probability $p(\boldsymbol{\pi}|\mathbf{X})$ can be computed in the same ways as Eq. (3), where $p(\pi_t|\mathbf{X})$ is given by

$$p(\pi_t|\mathbf{X}) = \begin{cases} (1 - \phi_t^{*})\phi_{tb_t}^{\mathrm{B}}\phi_{tp_t}^{\mathrm{P}}(\phi_t^{\mathrm{O}})^{o_t}(1 - \phi_t^{\mathrm{O}})^{1-o_t} & \text{if} \quad \pi_t \neq *, \\ \phi_t^{*} & \text{otherwise.} \end{cases} \tag{7}$$

C. Network Architecture

The DNN used in this study is based on an encoder-decoder architecture (Fig. 4). Since wide-band harmonic structures and long-term temporal structures are considered to be useful clues for pitch estimation and metrical analysis, respectively, two

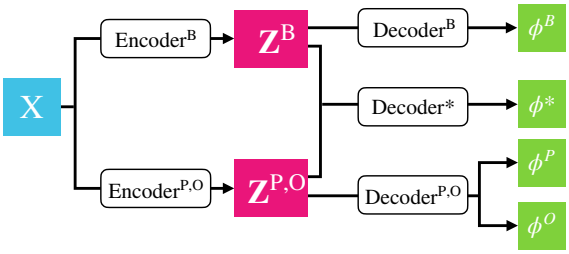


Fig. 4. Network architecture.

separate encoders are used to extract latent metrical features $\mathbf{Z}^{\mathrm{B}}$ and latent pitch features $\mathbf{Z}^{\mathrm{P,O}}$ as follows:

$$\mathbf{Z}^{\mathrm{B}} = \mathrm{Encoder}^{\mathrm{B}}(\mathbf{X}), \tag{8}$$

$$\mathbf{Z}^{\mathrm{P,O}} = \mathrm{Encoder}^{\mathrm{P,O}}(\mathbf{X}). \tag{9}$$

To estimate the pitch and onset presence at each frame, it is sufficient to focus on the local region of the singing voice around the frame. To estimate the metrical position, in contrast, it is necessary to collect long-term periodic information over multiple measures from the background music.

We then compute the posterior probabilities of the metrical positions, pitches, onset presence, and blank symbol, denoted by $\boldsymbol{\phi}^{\mathrm{B}}$, $\boldsymbol{\phi}^{\mathrm{P}}$, $\boldsymbol{\phi}^{\mathrm{O}}$, and $\boldsymbol{\phi}^{*}$, as follows:

$$\boldsymbol{\phi}^{\mathrm{B}} = \mathrm{Decoder}^{\mathrm{B}}(\mathbf{Z}^{\mathrm{B}}), \tag{10}$$

$$(\boldsymbol{\phi}^{\mathrm{P}}, \boldsymbol{\phi}^{\mathrm{O}}) = \mathrm{Decoder}^{\mathrm{P,O}}(\mathbf{Z}^{\mathrm{P,O}}), \tag{11}$$

$$\boldsymbol{\phi}^{*} = \mathrm{Decoder}^{*}(\mathbf{Z}^{\mathrm{B}}, \mathbf{Z}^{\mathrm{P,O}}). \tag{12}$$

Since the blank symbol tends to be selected at frames with small confidence about the pitch, onset, or metrical estimation, all the latent features are considered.

D. Decoding

After the CTC-based training, one can estimate the frame- and tatum-level sequences $\hat{\boldsymbol{\pi}}$ and $\hat{\mathbf{Y}}$ as follows:

$$\hat{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{\pi}} p(\boldsymbol{\pi}|\mathbf{X}), \tag{13}$$

$$\hat{\mathbf{Y}} = \mathcal{M}(\hat{\boldsymbol{\pi}}). \tag{14}$$

According to this method, however, the estimated metrical positions may not have a monotonically increasing order.

To impose the constraint of monotonically increasing metrical positions, we propose a refined decoding method based on an HSMM. Generally in music performances including singing, the durations of beats or tatums keep stable because of the regularity of metrical structure in music. To utilize this temporal stability, we introduce the durations of tatums $\mathbf{D} \triangleq \{\mathbf{d}_n\}_{n=1}^{N}$ measured in frames. Consider the frame-level input features and tatum-level output features described in Section. III-A, namely, input spectrograms $\mathbf{X} \triangleq \{\mathbf{x}_t\}_{t=1}^{T}$, and output metrical positions $\mathbf{B} \triangleq \{b_n\}_{n=1}^{N}$, pitches $\mathbf{P} \triangleq \{p_n\}_{n=1}^{N}$ and onset flags $\mathbf{O} \triangleq \{o_n\}_{n=1}^{N}$. We formulate the generative process of $\mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}$ and $\mathbf{X}$ as an HSMM, which only allows the metrical positions to make periodic progress, that is,

$$b_n = \begin{cases} b_{n-1} + 1 & b_{n-1} < 15, \\ 0 & b_{n-1} = 15, \end{cases} \tag{15}$$

and encourages the durations of metrical positions to be stable by assuming

$$p(d_n|d_{n-1}) \propto \exp\left(-\lambda \left| \frac{d_n}{d_{n-1}} - 1 \right| \right). \tag{16}$$

With the HSMM, one can efficiently decode the sequence of metrical positions, pitch values and onset flags with the Viterbi algorithm (see details in Appendix A).

IV. Evaluation

This section reports experiments conducted for evaluating the effectiveness of the proposed method.

A. Experimental Conditions

1) Dataset: We made a dataset consisting of 343 Japanese popular (JPOP) songs with the corresponding musical scores in the MusicXML format transcribed by experts and real audio recordings extracted from the original CDs. It was randomly split into a training set of 308 songs and a test set of 35 songs. Although this JPOP dataset is useful for evaluating the practical performance, it cannot be released due to copyright issues. We thus also used the publicly available RWC Popular Music dataset [27] for evaluation. In detail, we randomly selected 12 songs with the signature of 4/4 to form a test dataset (Nos. 7, 8, 13, 18, 20, 47, 63, 79, 80, 84, 90, and 100). The vocal scores of these songs were transcribed by experts in the MusicXML format as well.

Note that most common datasets (*e.g.*, MIR-ST500 [28]) consist of singing voices with annotations in the MIDI format, where the onset and offset times of each note are given in seconds. These datasets have thus been used for audio-to-MIDI AST, but are hard-to-use for audio-to-score AST (ours) due to the lack of ground-truth musical scores.

To reduce the computational cost, all the audio recordings were resampled at 22050 Hz and cut into segments of 8 seconds. To find the corresponding segments of scores, we synthesized waveforms from the scores and aligned them with the original signals with dynamic time warping (DTW). These alignment results were then used to segment the ground truth scores. The spectrogram of a music signal was obtained with short-time Fourier transform (STFT) with a window of 2048 samples and a hop length of 256 samples (11.61 ms). The mel spectrogram was then obtained with 128 mel filter banks.

2) Network Configuration: The CRNN used for evaluation was configured as follows. $\mathrm{Encoder}^{\mathrm{B}}$ and $\mathrm{Encoder}^{\mathrm{P,O}}$ were implemented with CNNs of the same structure. Each CNN had five layers and the numbers of channels for these layers were 64, 32, 32, 32, and 32, respectively. The kernel sizes for the hidden layers were 5, 5, 3, 3, and 3, respectively. Following the CNN, a fully-connected layer was used for reducing the hidden dimension to 256 at each frame. $\mathrm{Decoder}^{\mathrm{B}}$, $\mathrm{Decoder}^{\mathrm{P,O}}$ and $\mathrm{Decoder}^{*}$ were implemented with RNNs. Each RNN had two layers of BLSTMs, and the number of channels in the hidden layers was all 256. The CRNN was trained by an Adam optimizer with a learning rate of 1×10^{-4} for 20 epochs with the training set.

TABLE I
COMPARISON OF EDIT-DISTANCE-BASED ERROR RATES (%) OF DIFFERENT METHODS.

Test Dataset	Method	Pitch Error	Miss Rate	Extra Rate	Onset Error	Offset Error	Mean Error
JPOP-test	VOCANO [9] + quantization [22]	9.0	47.8	14.2	47.4	31.2	29.9
	Note-level CTC (baseline)	**6.9**	24.6	**7.5**	37.3	27.3	**20.7**
	Tatum-level CTC (proposed)	8.8	**23.4**	14.1	**35.1**	**26.4**	21.6
RWC-test	VOCANO [9] + quantization [22]	12.7	16.5	15.6	40.6	29.5	23.0
	Note-level CTC (baseline)	**9.1**	**12.1**	**11.7**	35.5	**24.0**	**18.5**
	Tatum-level CTC (proposed)	9.4	13.3	18.9	**32.3**	26.9	20.2

TABLE II
COMPARISON OF BEAT AND DOWNBEAT F SCORES (%) FOR DIFFERENT METHODS.

Test Dataset	Method	Beat			Downbeat		
		Precision	Recall	F score	Precision	Recall	F score
JPOP-test	VOCANO [9] + quantization [22]	53.3	42.8	45.7	15.3	15.7	14.8
	Note-level CTC (baseline)	50.7	28.0	34.6	13.0	6.9	8.6
	Tatum-level CTC (proposed)	**80.5**	**73.6**	**75.2**	**67.9**	**68.8**	**67.3**
RWC-test	VOCANO [9] + quantization [22]	39.5	39.3	39.0	24.7	25.2	24.6
	Note-level CTC (baseline)	41.0	26.7	31.5	15.4	8.1	10.37
	Tatum-level CTC (proposed)	**77.0**	**82.0**	**79.3**	**66.5**	**70.2**	**68.0**

The HSMM used in the decoding phase was configured as follows. The hyperparameter λ in the transition model was set to 100, the minimum and maximum pitch values were $p_{min} = 43$, $p_{max} = 79$, and the minimum and maximum tatum durations were $d_{min} = 6$, $d_{max} = 30$, corresponding to tempos from 215 bpm to 43 bpm.

B. Compared Methods

To evaluate the efficiency of the proposed method, we compared it with VOCANO [9], a singing transcription system which combined pitch extraction and note segmentation. The original training data used for VOCANO have no overlap with the test data in our experiments. Since VOCANO was designed for estimating a pitch contour from singing voice, we processed the frame-based result obtained by VOCANO with a rhythm quantization method based on an HMM [22].

For comparison, we configured a baseline model based on a note-level score representation that estimates the note value of each note instead of the metrical position. This model was similar to ones used in the literature [15], [16] and was almost the same as the proposed method, except that each output symbol was given by $y_n \triangleq (d_n, p_n, o_n)$, where p_n and o_n are the same as ones described in Section III-A and $d_n \in [1, 16]$ represents the note duration. Note that only the duration within a measure was considered, and notes longer than a measure were split into different symbols. For example, a note with the pitch value of 64 and a duration of 18 tatums was represented as $(16, 64, 1), (2, 64, 0)$. The note-level baseline model was trained with the same procedure and the same dataset as the proposed CTC-based method and decoded with the greedy algorithm as Eqs. 13,14.

C. Evaluation Measures

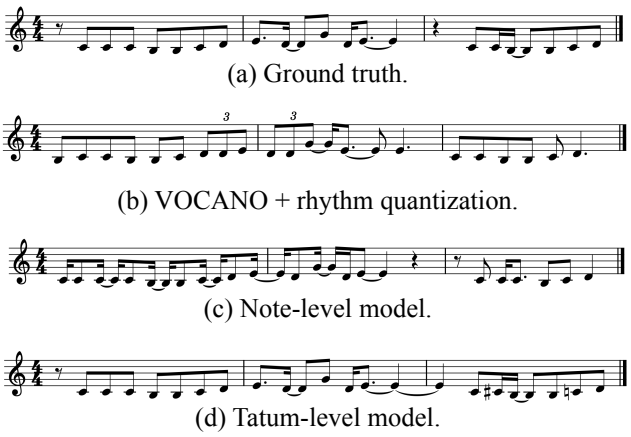
We evaluated the estimated results with edit-distance-based metrics [29] in a way similar to the word error rate used for ASR. These metrics were computed by performing alignment between the estimated notes and ground-truth notes with DP and then identifying five exclusive types of errors: pitch errors, missing notes, extra notes, onset time errors, and offset time errors. The onset time and offset time errors are defined in terms of score times in units of tatums, but they do not take into account metrical positions. To evaluate the metrical structure of estimated scores, we additionally calculated the precision, recall, and F-score by inspecting the beat and downbeat positions of the correctly aligned notes.

D. Experimental Results

Tables I and II show the evaluation results. The multi-step combination of VOCANO and rhythm quantization had a low overall accuracy in terms of both the edit-distance-based metrics and the beat and downbeat F scores. This shows the difficulty of quantizing the estimated frame-level F0 contours. The CTC-based method with the note-level representation and its variant with the tatum-level representation had similar mean error rates in Table I, and the latter had higher beat and downbeat F scores (Table II). This shows the effectiveness of the end-to-end approach to AST based on CTC with the tatum-level representation.

Fig. 5 shows examples of musical scores estimated for the same piece of music from the JPOP-test dataset. As shown in Fig. 5(b), the score estimated by VOCANO contained extra notes and missing notes as well as a number of rhythmic errors. The score estimated by the baseline model with the note-level representation had mostly correct pitches and note values, but suffered from severe misalignment of metrical positions. Since the note-level representation does not consider beat and downbeat positions, rhythmic errors were accumulated when estimated notes were concatenated together. Compared with the baseline model, the proposed tatum-level model significantly reduced errors in metrical positions while keeping errors in pitches and note values at the same level.

Fig. 6 shows the histograms of the durations of the estimated notes obtained by the note-level model and the tatum-level model on the JPOP-test dataset. We found that the CTC-based



(a) Ground truth.

(b) VOCANO + rhythm quantization.

(c) Note-level model.

(d) Tatum-level model.

Fig. 5. Examples of estimated scores.

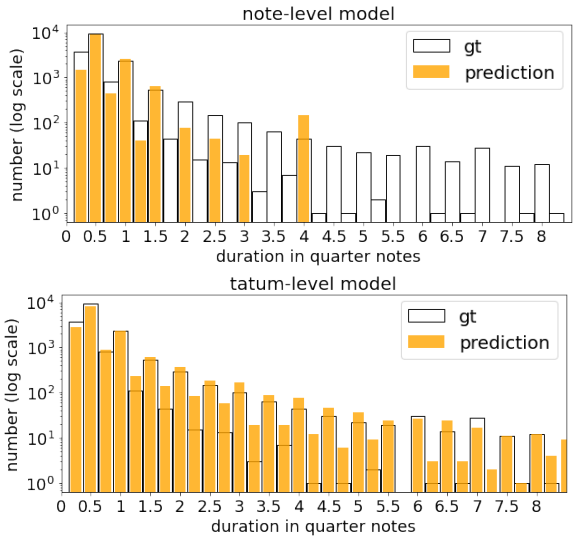


Fig. 6. Histograms of the note values estimated by the note-level (above) and tatum-level (below) CTCs, where "gt" stands for the ground truth data.

model with the note-level representation failed to correctly predict note values longer than a half note. Although the method was designed for dealing with note values longer than a whole note, the model failed to learn this property. With the proposed tatum-level representation, the histogram of the estimated note values was similar to that of the ground truth data. The proposed representation increased the rate of correctly estimating note values and contributed to correctly predicting metrical positions.

V. Conclusion

In this paper, we proposed an end-to-end AST method based on the CTC-based training with a tatum-level score representation consisting of positions of the tatums within measures, note pitches, and onset flags. The experimental results showed that the proposed method achieved better accuracy on the estimation of metrical positions of notes, and a mean edit-distance-based error rate similar to the CTC-based baseline method with a note-level score representation. We also found that the cascade of F0 extraction and rhythm quantization yielded a significantly worse result.

There are several issues left for future work. First, we assumed that the minimum time unit of a score is the sixteenth note, and we thus need to extend the model for fine-grained time units. Second, the proposed method can only transcribe songs with the time signature of 4/4, and other common meters should be properly dealt with. Additionally, popular music can have multiple meters within a song, and such variable meters should also be considered. Finally, in the decoding with the HSMM, a note-level musical language model is expected to be effective to improve the result, as in the case for ASR.

Acknowledgment

This study was partially supported by JSPS KAKENHI Nos. 21K12187, 21K02846, 22H03661, 20H00602, and 21H03572, JST PRESTO No. JPMJPR20CB, and JST FOREST No. JPMJPR226X. We thank Ryo Nishikimi for the constructive discussion with him on this study.

Appendix

A. HSMM-Based Decoding

1) Formulation of the HSMM: To impose the constraint of monotonically increasing metrical positions, we propose a refined decoding method based on an HSMM. Generally in music performances including singing, the durations of beats or tatums keep stable because of the regularity of metrical structure in music. To utilize this temporal stability, we introduce the durations of tatums $\mathbf{D} \triangleq \{\mathbf{d}_n\}_{n=1}^N$ measured in frames in the proposed HSMM. Consider the frame-level input features and tatum-level output features described in Section. III-A, namely, input spectrograms $\mathbf{X} \triangleq \{\mathbf{x}_t\}_{t=1}^T$, and output metrical positions $\mathbf{B} \triangleq \{b_n\}_{n=1}^N$, pitches $\mathbf{P} \triangleq \{p_n\}_{n=1}^N$ and onset flags $\mathbf{O} \triangleq \{o_n\}_{n=1}^N$. We formulate the generative process of $\mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}$ and $\mathbf{X}$ as

$$p(\mathbf{X}, \mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}) = p(\mathbf{X}|\mathbf{B}, \mathbf{P}, \mathbf{O}) p(\mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}). \quad (17)$$

In Eq. 17, $p(\mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D})$ represents the language model of the hidden variables. We assume the distributions of pitches $\mathbf{P}$ and onsets $\mathbf{O}$ to be uniform, and formulate the transition probabilities as

$$p(\mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}) = p(b_1, p_1, o_1, d_1) \times \prod_{n=2}^{N} p(b_n, p_n, o_n, d_n | b_{n-1}, p_{n-1}, o_{n-1}, d_{n-1}). \quad (18)$$

We assume the initial distribution of $\mathbf{B}$ and $\mathbf{D}$ is uniform, namely

$$p(b_1, p_1, o_1, d_1) = \frac{1}{K_b} \frac{1}{K_o} \frac{1}{p_{max} - p_{min} + 2} \frac{1}{d_{max} - d_{min} + 1}, \quad (19)$$

where $K_b = 16, K_o = 2$ are the number of distinct metrical positions and onset flags, p_{min} and p_{max} are assumed minimum and maximum pitch values other than the rest, and d_{min}

TABLE III
AN EXAMPLE OF THE FRAME-LEVEL VARIABLES.

Metrical Position $\bar{b}_t$	15	15	15	15	0	0	0	0	0	1	1	1	1	...
Tatum Duration $\bar{d}_t$	4	4	4	4	5	5	5	5	5	4	4	4	4	...
Tatum Counter $\bar{c}_t$	3	2	1	0	4	3	2	1	0	3	2	1	0	...
Pitch $\bar{p}_t$	62	62	62	62	62	62	62	62	62	59	59	59	59	...
Onset $\bar{o}_t$	1	1	1	1	0	0	0	0	0	1	1	1	1	...

and d_{max} are assumed minimum and maximum lengths of a tatum. The transition probabilities of b_n, d_n from b_{n-1}, d_{n-1} are determined by

$$\begin{aligned} &p(b_n, p_n, o_n, d_n | b_{n-1}, p_{n-1}, o_{n-1}, d_{n-1}) \\ &= \frac{1}{K_o} \frac{1}{p_{max} - p_{min} + 2} \delta_{16}(b_n, b_{n-1} + 1) p(d_n | d_{n-1}), \end{aligned} \quad (20)$$

where δ_{16} is the Kronecker's function modulo 16, that is

$$\delta_{16}(x, y) = \begin{cases} 1 & \text{if } x \equiv y \pmod{16}, \\ 0 & \text{if } x \not\equiv y \pmod{16}. \end{cases} \quad (21)$$

As suggested in [30], the transition probabilities of d_n from d_{n-1} are given by

$$p(d_n | d_{n-1}) \propto \exp\left(-\lambda \left| \frac{d_n}{d_{n-1}} - 1 \right| \right). \quad (22)$$

2) Frame-Level Formulation of the HSMM: The probabilistic model proposed in Section. A1 contains temporally asynchronous variables, $\mathbf{X}$ on the frame level and $\mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}$ on the tatum level. To formulate the model in a synchronous way, we introduce the frame-synchronized versions of metrical positions $\bar{\mathbf{B}} \triangleq \{\bar{b}_t\}_{t=1}^T$, pitches $\bar{\mathbf{P}} \triangleq \{\bar{p}_t\}_{t=1}^T$, onset flags $\bar{\mathbf{O}} \triangleq \{\bar{o}_t\}_{t=1}^T$, and tatum durations $\bar{\mathbf{D}} \triangleq \{\bar{d}_t\}_{t=1}^T$. These variables can be constructed from the corresponding frame-level variables with $\bar{b}_t = b_n$, $\bar{p}_t = p_n$, $\bar{o}_t = o_n$ and $\bar{d}_t = d_n$ when $u(n-1) + 1 \le t \le u(n)$, where

$$u(n) \triangleq \begin{cases} 0 & n = 0, \\ \sum_{j=1}^n d_j & n > 0, \end{cases} \quad (23)$$

is the index of the last frame for each tatum. In addition, a tatum counter $\bar{\mathbf{C}} \triangleq \{\bar{c}_t\}_{t=1}^T$ is used to represent the position of the current frame within the tatum, counting down from $d_n - 1$ to 0. Table III shows an example of the frame-level variables.

With the tatum-level variables the HSMM can be rewritten as

$$\begin{aligned} p(\mathbf{X}, \mathbf{B}, \mathbf{P}, \mathbf{O}, \mathbf{D}) &= p(\mathbf{X}, \bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}, \bar{\mathbf{D}}, \bar{\mathbf{C}}) \\ &= p(\mathbf{X} | \bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}) p(\bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}, \bar{\mathbf{D}}, \bar{\mathbf{C}}). \end{aligned} \quad (24)$$

In Eq. 17, $p(\bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}, \bar{\mathbf{D}}, \bar{\mathbf{C}})$ represents the language model of the frame-level hidden variables. It is similar to Eq. 18, except that the transition probabilities become

$$\begin{aligned} &p(\bar{b}_t, \bar{p}_t, \bar{o}_t, \bar{d}_t, \bar{c}_t | \bar{b}_{t-1}, \bar{p}_{t-1}, \bar{o}_{t-1}, \bar{d}_{t-1}, \bar{c}_{t-1}) \\ &= \begin{cases} \delta(\bar{b}_t, \bar{b}_{t-1}) \delta(\bar{p}_t, \bar{p}_{t-1}) \delta(\bar{o}_t, \bar{o}_{t-1}) \delta(\bar{d}_t, \bar{d}_{t-1}) \delta(\bar{c}_t, \bar{c}_{t-1} - 1), \\ \bar{c}_{t-1} > 0, \\ \frac{1}{K_o} \frac{1}{p_{max} - p_{min} + 2} \delta_{16}(\bar{b}_t, \bar{b}_{t-1} + 1) p(\bar{d}_t | \bar{d}_{t-1}) \delta(\bar{c}_t, \bar{d}_t - 1), \\ \bar{c}_{t-1} = 0. \end{cases} \end{aligned} \quad (25)$$

In Eq. 24, $p(\mathbf{X} | \bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}})$ represents the emission probabilities of the observed spectrograms given the metrical positions, pitch values and onset flags. We assume the conditional independence of $\mathbf{X}$ given $\bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}$, and we have

$$p(\mathbf{X} | \bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}) = \prod_{t=1}^T p(\mathbf{x}_t | \bar{b}_t, \bar{p}_t, \bar{o}_t), \quad (26)$$

According to Bayes' theorem, the emission probability of a frame can be written as

$$\begin{aligned} p(\mathbf{x}_t | \bar{b}_t, \bar{p}_t, \bar{o}_t) &= \frac{p(\bar{b}_t, \bar{p}_t, \bar{o}_t | \mathbf{x}_t) p(\mathbf{x}_t)}{p(\bar{b}_t, \bar{p}_t, \bar{o}_t)} \\ &\propto \frac{p(\bar{b}_t, \bar{p}_t, \bar{o}_t | \mathbf{x}_t)}{p(\bar{b}_t, \bar{p}_t, \bar{o}_t)}. \end{aligned} \quad (27)$$

We assume that $p(\bar{b}_t, \bar{p}_t, \bar{o}_t)$ is a uniform distribution and that $p(\bar{b}_t, \bar{p}_t, \bar{o}_t | \mathbf{x}_t)$ can be estimated with a DNN trained in an end-to-end manner as described in Section III-B. Under the assumption of conditional independence of $\bar{b}_t, \bar{p}_t, \bar{o}_t$ given $\mathbf{x}_t$, we put that

$$\begin{aligned} p(\bar{b}_t, \bar{p}_t, \bar{o}_t | \mathbf{x}_t) &= p(\bar{b}_t, | \mathbf{x}_t) p(\bar{p}_t, | \mathbf{x}_t) p(\bar{o}_t, | \mathbf{x}_t), \\ &= \phi^{\mathrm{B}}_{t\bar{b}_t} \phi^{\mathrm{P}}_{t\bar{p}_t} (\phi^{\mathrm{O}}_t)^{\bar{o}_t} (1 - \phi^{\mathrm{O}}_t)^{1-\bar{o}_t}, \quad (28) \\ p(b_n | \mathbf{x}_t) &= \phi^{\mathrm{B}}_{t\bar{b}_t}, \quad (29) \\ p(p_n | \mathbf{x}_t) &= \phi^{\mathrm{P}}_{t\bar{p}_t}, \quad (30) \\ p(o_n | \mathbf{x}_t) &= (\phi^{\mathrm{O}}_t)^{\bar{o}_t} (1 - \phi^{\mathrm{O}}_t)^{1-\bar{o}_t}. \quad (31) \end{aligned}$$

It is notable that only the outputs $\phi^{\mathrm{B}}, \phi^{\mathrm{P}}$ and ϕ^{O} are used here since they can be interpreted as the posterior probabilities when the neural network judges a frame to be non-blank.

3) Viterbi Decoding: With the probabilistic model described above, we can estimate the optimal metrical positions $\mathbf{B}^*$, pitch values $\mathbf{P}^*$ and onset flags $\mathbf{O}^*$ by first estimating

$$\bar{\mathbf{B}}^*, \bar{\mathbf{P}}^*, \bar{\mathbf{O}}^*, \bar{\mathbf{D}}^*, \bar{\mathbf{C}}^* = \operatorname{argmax} p(\mathbf{X}, \bar{\mathbf{B}}, \bar{\mathbf{P}}, \bar{\mathbf{O}}, \bar{\mathbf{D}}, \bar{\mathbf{C}}), \quad (32)$$

and then obtain $\mathbf{B}^*, \mathbf{P}^*, \mathbf{O}^*$ from $\bar{\mathbf{B}}^*, \bar{\mathbf{P}}^*, \bar{\mathbf{O}}^*$. Eq. 32 can be solved efficiently with the Viterbi algorithm.

REFERENCES

[1] Akinori Ito, Yu Kosugi, Shozo Makino, and Masashi Ito. A query-by-humming music information retrieval from audio signals based on multiple F0 candidates. In *ICALIP*, pp. 1–5, 2010.

[2] Martín Rocamora, Pablo Cancela, and Alvaro Pardo. Query by humming: Automatically building the database from music recordings. *Pattern Recognition Letters*, Vol. 36, pp. 272–280, 2014.

[3] Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, No. 6, pp. 1205–1215, 2011.

[4] Sebastian Ewert, Bryan Pardo, Meinard Muller, and Mark D. Plumbley. Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine*, Vol. 31, No. 3, pp. 116–124, 2014.

[5] Shreyan Chowdhury, Andreu Vall Portabella, Verena Haunschmid, and Gerhard Widmer. Towards explainable music emotion recognition: The route via mid-level features. In *ISMIR*, pp. 237–243, 2019.

[6] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *ICASSP*, pp. 161–165, 2018.

[7] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang. A streamlined encoder/decoder architecture for melody extraction. In *ICASSP*, pp. 156–160, 2019.

[8] Shuai Yu, Xiaoheng Sun, Yi Yu, and Wei Li. Frequency-temporal attention network for singing melody extraction. In *ICASSP*, pp. 251–255, 2021.

[9] Jui-Yang Hsu and Li Su. VOCANO: A note transcription framework for singing voice in polyphonic music. In *ISMIR*, 2021.

[10] Jun-You Wang and Jyh-Shing Roger Jang. Training a singing transcription model using connectionist temporal classification loss and cross-entropy loss. *TASLP*, 2023.

[11] Eita Nakamura and Kazuyoshi Yoshii. Musical rhythm transcription based on Bayesian piece-specific score models capturing repetitions. *Information Sciences*, Vol. 572, pp. 482–500, 2021.

[12] Ryo Nishikimi, Eita Nakamura, Masataka Goto, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Scale- and rhythm-aware musical note estimation for vocal F0 trajectories based on a semi-tatum-synchronous hierarchical hidden semi-Markov model. In *ISMIR*, pp. 376–382, 2017.

[13] Ryo Nishikimi, Eita Nakamura, Masataka Goto, and Kazuyoshi Yoshii. Audio-to-score singing transcription based on a CRNN-HSMM hybrid model. *APSIPA*, Vol. 10, p. e7, 2021.

[14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pp. 369–376, 2006.

[15] Miguel A Román, Antonio Pertusa, and Jorge Calvo-Zaragoza. An end-to-end framework for audio-to-score music transcription on monophonic excerpts. In *ISMIR*, pp. 34–41, 2018.

[16] Victor Arroyo, Jose J. Valero-Mas, Jorge Calvo-Zaragoza, and Antonio Pertusa. Neural audio-to-score music transcription for unconstrained polyphony using compact output representations. In *ICASSP*, pp. 4603–4607, 2022.

[17] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, pp. 1764–1772, 2014.

[18] Han-Wen Nienhuys and Jan Nieuwenhuizen. Lilypond, a system for automated music engraving. In *CIM*, pp. 167–171, 2003.

[19] Craig Stuart Sapp. Online database of scores in the humdrum file format. In *ISMIR*, pp. 664–665, 2005.

[20] Matthias Mauch and Simon Dixon. Pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *ICASSP*, pp. 659–663, 2014.

[21] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon. Computer-aided melody note transcription using the tony software: Accuracy and efficiency. In *TENOR*, 2015.

[22] Kentaro Shibata, Eita Nakamura, and Kazuyoshi Yoshii. Non-local musical statistics as guides for audio-to-score piano transcription. *Information Sciences*, Vol. 566, pp. 262–280, 2021.

[23] Eita Nakamura, Kazuyoshi Yoshii, and Shigeki Sagayama. Rhythm transcription of polyphonic piano music based on merged-output hmm for multiple voices. *TASLP*, Vol. 25(4), pp. 794–806, 2017.

[24] Ryo Nishikimi, Eita Nakamura, Masataka Goto, and Kazuyoshi Yoshii. End-to-end melody note transcription based on a beat-synchronous attention mechanism. In *WASPAA*, pp. 26–30, 2019.

[25] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-Unmix — a reference implementation for music source separation. *Journal of Open Source Software*, 2019.

[26] Curtis Wigington, Brian Price, and Scott Cohen. Multi-label connectionist temporal classification. In *ICDAR*, pp. 979–986, 2019.

[27] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *ISMIR*, 2002.

[28] Jun-You Wang and Jyh-Shing Roger Jang. On the preparation and validation of a large-scale dataset of singing transcription. In *ICASSP*, pp. 276–280, 2021.

[29] Eita Nakamura, Emmanouil Benetos, Kazuyoshi Yoshii, and Simon Dixon. Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization. In *ICASSP*, pp. 101–105, 2018.

[30] Florian Krebs, Sebastian Böck, and Gerhard Widmer. An efficient state-space model for joint tempo and meter tracking. In *ISMIR*, pp. 72–78, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Mask2Hand: Learning to Predict the 3D Hand Pose and Shape from Shadow

Li-Jen Chang*, Yu-Cheng Liao*, Chia-Hui Lin*, Shih-Fang Yang-Mao † and Hwann-Tzong Chen*
* National Tsing Hua University, Taiwan
E-mail: htchen@cs.nthu.edu.tw
† Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan.
E-mail: YangMao@itri.org.tw

***Abstract*—We present a self-trainable method, Mask2Hand, which learns to solve the challenging task of predicting 3D hand pose and shape from a 2D binary mask of hand silhouette/shadow without additional manually-annotated data. Given the intrinsic camera parameters and the parametric hand model in the camera space, we adopt the differentiable rendering technique to project 3D estimations onto the 2D binary silhouette space. By applying a tailored combination of losses between the rendered silhouette and the input binary mask, we are able to integrate the self-guidance mechanism into our end-to-end optimization process for constraining global mesh registration and hand pose estimation. The experiments show that our method, which takes a single binary mask as the input, can achieve comparable prediction accuracy on both unaligned and aligned settings as state-of-the-art methods that require RGB or depth inputs.**

I. Introduction

More than hundreds of years ago the historic art form of hand shadow puppetry had already been documented in many ancient countries. Since then, making hand shadow puppets has been a low-cost household entertainment activity that parents can play with their children. As far as human perception is concerned, it is interesting that humans can recognize the animals and characters portrayed by the hands and fingers from merely their projected shadows. It is also remarkable that people are able to develop and create different ways of making hand shadow puppets, which may involve the generative capability of imagining some target shape and adjusting the hands and fingers to match the imaginary shape with the shadow. Inspired by the discriminative and generative nature in the visual process of doing hand shadow puppets, we aim to explore the possibility of learning to estimate the 3D hand pose and shape from an input hand silhouette, through a similar trial-and-error manner of learning to do hand shadow puppets.

Our main goal in this endeavor is to train a renderer-equipped deep network that takes a binary mask of a hand's silhouette as the input and predicts the corresponding 3D hand shape that produces a silhouette very similar to the input. (Fig. 1 shows several example results obtained using our model.) To build such a model, we employ MANO [1] as the 3D hand-shape renderer in our deep network. We propose a new network, called Mask2Hand, which consists of an encoder, a MANO layer, a differentiable render, and the refinement module. Fig. 2 shows an overview of the proposed Mask2Hand network. Our network takes a binary mask as the input and uses the encoder to generate the required parameters for the MANO layer to render the hand mesh. The encoder learns to predict the pose-related principal components of MANO and the 3D global transformation of the hand from only a single input binary mask. Based on the principal components and global transformation, the MANO layer reconstructs the 3D hand mesh and the hand joint positions. We then use a differentiable renderer to generate the 2D projected silhouette from the hand mesh. Finally, the refinement module compares the projected silhouette with the input binary mask to refine the entire network. Three groups of loss functions are tailored for training the aforementioned modules, including the pose loss, silhouette loss, and mesh loss. The loss functions and the operations involved in the modules are all differentiable—we are able to train the entire network from end to end. Our code can be found at GitHub by searching Mask2Hand.

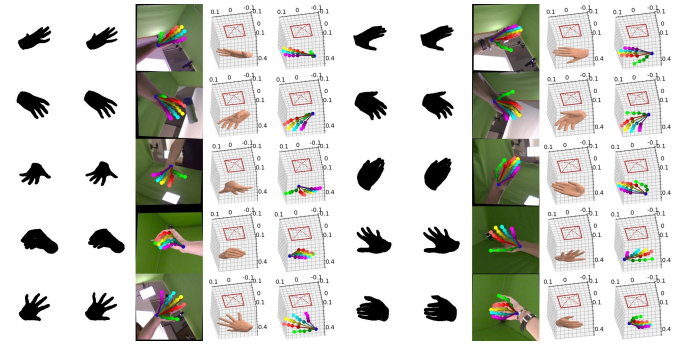


Fig. 1. The proposed method, Mask2Hand, learns to solve the challenging task of predicting the 3D hand pose and shape from a 2D binary mask of hand silhouette. Each row in the left or right half of the figure illustrates an example result of performing our method on real data. The first column shows the input of 2D binary mask. The second column shows the rendered mask using the predicted hand shape. The third column shows the predicted 2D skeleton depicted in the image space. Note that the color images are simply for visualization; our method does not use any RGB data. The last two columns show the predicted mesh and 3D skeleton.

As this work aims at addressing a rather new and challenging task of predicting the 3D hand pose from a single 2D binary mask, it is quite encouraging that the proposed Mask2Hand network can achieve comparable prediction accuracy on 3D hand pose and shape estimation as state-of-the-art methods, although we only need a single binary input

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

image while theirs require an RGB image or a depth map. Note that our model can be trained under supervision with manually annotated joint positions or in a fully self-supervised manner without human annotations. Furthermore, since our method learns to predict the global transformation of the hand, its evaluation performance under the unaligned setting of hand pose estimation does not degrade as much as other methods do. Such a property is particularly favorable as, in real applications, it is not allowed to align the prediction with the unknown ground truth.

In addition to the motivation of hand shadow puppets, there are some other advantages and real-world applications of working on binary silhouettes. First, in contrast to RGB input, binary images can be obtained with relatively inexpensive sensors, such as low-resolution range sensors and low-resolution infrared sensors. They are especially useful when the lighting condition is poor and not much color information is available. Second, a line of research [2]–[4] in the hardware domain explores the usage of ambient light as a ubiquitous sensing medium. Light sensors, such as photodiodes or solar cells, measure and track the shadow blockage that a user's hands cause with low cost and ultra-low power consumption, which have the potential to be widely used in IoT devices for shadow-based hand pose estimation systems. Finally, binary input is more robust to adversarial attacks of imperceptible pixel perturbation. From the literature on adversarial machine learning [5], we know that invisible image manipulation can seriously spoil the functionality of DNNs. However, this common type of attack is not directly applicable to the binary setting since small perturbation in pixel values is impossible with binary images. Also, pixel inversion in binary silhouettes can be easily identified and addressed.

II. Related Work

One of the research endeavors inspiring us is *Shadow Theatre* [6], which aims to reproduce computationally the shadow puppet theatre—an ancient shadow-oriented form of performance art—just like how we are motivated by hand shadow puppetry. Won and Lee [6] develop an algorithm to incorporate several features and heuristic rules that are related to shadow generation using human bodies. The proposed method optimizes the poses of human bodies to match the given 2D target shapes. The underlying nonlinear high-dimensional optimization is solved using the proposed heuristic strategies so that a plausible solution can be found with reasonable computing resources. Nevertheless, the computation still requires tens to hundreds of minutes to complete on GPU.

A. 3D Hand Pose and Shape Estimation

Various methods have been proposed to address the problem of 3D hand pose and shape estimation. Generally, the techniques that are used to solve 3D human pose estimation can also be applied to 3D hand pose estimation. Based on the input format, we consider the following four categories of methods on 3D hand pose and shape estimation: RGB-based, depth-based, event-camera-based, and silhouette-based methods. Among them, silhouette-based methods are most closely related to this work, where we seek to generate the 3D hand pose and mesh from merely a single binary-valued image of hand silhouette.

Despite the discrepancy in their input type, many of the methods take 2D joint detection as the basis. Such an approach is promising, and many existing deep-learning-based techniques can be employed. The following is a brief review of recent methods in the four categories.

a) RGB-based Methods.: Lin *et al*. propose a method called Mesh Transformer (METRO) [7] that explores non-local relationships between vertices and joints to reconstruct a 3D human pose from a single RGB image without using parametric mesh models. This method can also be extended to hand mesh reconstruction. Chen *et al*. [8] propose Camera-space Mesh Recovery (CMR) that generates 3D meshes from a single RGB image by extracting 2D cues including joint landmarks and silhouette. Chen *et al*. [9] present a self-supervised 3D hand reconstruction network called S^2HAND that estimates 3D poses from 2D detected keypoints. Liu *et al*. [10] propose a semi-supervised 3D hand-object pose estimation method leveraging spatial-temporal consistency in videos to obtain pseudo-labels for self-training. Li *et al*. [11] propose a hand pose estimation method by using multi-task learning to categorize joints into groups so that different features can be learned to recover the 3D joint locations in a group-wise manner.

b) Depth-based Methods.: Xiong *et al*. [12] propose the Anchor-to-Joint Regression Network (A2J) that exploits informative anchor points to enhance the generalization ability of 3D pose estimation. Fang *et al*. [13] present a dense-prediction-based 3D hand pose estimation method called JGR-P2O that uses a pixel-to-offset prediction network and a joint graph reasoning module to enable end-to-end training and improve computational efficiency.

c) Silhouette-based Methods.: Silhouette-based methods have not yet been widely applied to hand/body pose estimation. The task to be solved is more challenging than those in other settings of RGB-based and depth-based methods. The difficulty comes from the lack of information in the input data, where the shading, color, and depth of the 3D objects are absent in the binary-valued images. Lee *et al*. present the Silhouette-Net [14] method that generates 3D hand poses based on binary-valued silhouettes and can achieve a performance similar to or better than depth-based methods if multiview silhouettes are provided. However, their method requires the guidance of depth maps during the training stage. In contrast, our method does not require additional depth information during training, and it only needs a single-view 2D binary mask of the hand as the input. Despite the challenging setting, estimating the 3D hand pose and mesh from limited binary information in the input is our major goal in this work.

B. 3D Hand Models

Several hand mesh models have been proposed, such as MANO [1], Sphere-Meshes [15], and Convex Parts [16]. With

the help of these models, hand shape generation and pose estimation can be processed via parametric methods. Typically, a hand model like MANO is designed to parameterize a triangle mesh into pose, shape, and rotation parameters. MANO has been widely used by hand pose estimation methods [17]–[26] to train their deep networks; the deep networks learn to predict the hand shape in the target image by regressing to the principal components of the MANO model. As shown in Fig. 2, our method also uses the MANO model to construct the target hand shape. By learning to optimize the parameters of MANO meshes, our method achieves the goal of generating the 3D hand mesh from a binary-valued silhouette image.

III. Our Approach

To predict the hand pose and shape from a binary mask, we propose a two-phase end-to-end trainable network, *Mask2Hand*, that aims to reconstruct 3D hand meshes in the camera space. Fig. 2 illustrates the overall pipeline of our method. In what follows, we detail each module of the pipeline and describe how we formulate the loss functions.

A. Model Architecture

The core modules of our model include an encoder for estimating the MANO parameters, a MANO layer that reconstructs the 3D hand mesh, a differentiable renderer for optimizing the hand pose, and a RefineNet improving the accuracy.

a) Encoder.: Given a hand silhouette image presented as a binary mask, the encoder first uses a ResNet18 [27] backbone to extract features from the input. Then, these features are fed into four different estimators to encode 3D information into the PCA space of MANO [1] and predict the extrinsic camera parameters. More specifically, the output of the encoder consists of the first six or forty-five hand-pose-related principal components $\theta \in \mathbb{R}^6$ or $\mathbb{R}^{45}$ of MANO, shape parameters $\beta \in \mathbb{R}^{10}$, 3D rotation $R \in \mathbb{R}^3$ in axis-angle representation, and 3D translation $T \in \mathbb{R}^3$.

b) MANO Layer.: The 3D hand shape is represented by a triangle mesh that contains a vertex set $V \in \mathbb{R}^{778\times 3}$ and fixed faces F indicating the connection between them. In our network, the hand mesh and joints $J \in \mathbb{R}^{21\times 3}$ are reconstructed using MANO [1]. Being a low-dimensional parametric hand model, MANO can generate the required hand mesh from parameters including the hand-pose-related principal components θ, shape $\beta \in \mathbb{R}^{10}$, rotation R, and translation T.

c) Differentiable Renderer.: Given intrinsic camera parameters and hand meshes in the camera space, we use the differentiable rendering technique [28] to project 3D estimations into the 2D image space. By applying various losses between the rendered mask of the hand silhouette and the input binary image, we can integrate self-supervision into our end-to-end optimization process for constraining global mesh registration. Here we use the PyTorch3D [29] implementation of the differentiable renderer. The loss functions will be detailed later in the subsequent sections.

d) RefineNet.: In the previous stage, we only consider several pose-related principal components of MANO for parametrizing hands in a low-dimensional space. The lower degree of freedom may have an expressiveness problem in that the finer details of hand gestures cannot be ideally reconstructed. Therefore, we incorporate a small convolutional neural network, called the RefineNet module [1], into our model to deal with this issue. The network takes the concatenation of the input image and the projected silhouette as input, and it produces residual coordinate values that serve as the point-wise offsets to the previously constructed mesh. The final output of our model is the summation of the originally reconstructed hand mesh and the offset predicted by the RefineNet. The refined joints are then estimated by MANO's joint regressor given the refined mesh as its input.

B. Loss Functions

The complete objective function that we design for training our Mask2Hand model comprises three main loss terms:

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{silhouette}} + \mathcal{L}_{\text{mesh}} \,, \tag{1}$$

where we have the pose loss $\mathcal{L}_{\text{pose}}$ on the predicted joints, the silhouette loss $\mathcal{L}_{\text{silhouette}}$ on the 2D projected silhouette of the hand, and the mesh loss $\mathcal{L}_{\text{mesh}}$ on the reconstructed 3D mesh.

1) Pose Loss.: The pose loss in Eq. (1) further consists of two loss functions that use the unaligned and aligned joint positions to evaluate the preliminary and refined joint predictions by the squared L2-norm error:

$$\mathcal{L}_{\text{pose}} = \lambda_{\text{J}}\mathcal{L}_{\text{J}} + \lambda_{\text{alignJ}}\mathcal{L}_{\text{alignJ}} \,, \tag{2}$$

where we have $\mathcal{L}_{\text{J}}$ for the unaligned joint evaluation and $\mathcal{L}_{\text{alignJ}}$ for the aligned joint evaluation, with λ_{J} and λ_{alignJ} as the weight factors. The 3D positions of the K joints are represented as a K by 3 matrix, *i.e.*, $J \in \mathbb{R}^{K\times 3}$. The loss function $\mathcal{L}_{\text{J}}$ of the unaligned joint evaluation is computed as

$$\mathcal{L}_{\text{J}} = \frac{1}{K}\sum_{i=1}^{K}\left\|J_i - \widehat{J}_i\right\|_2^2 + \frac{1}{K}\sum_{i=1}^{K}\left\|J_i - \widehat{J}_i^{\text{ref}}\right\|_2^2 , \tag{3}$$

J_i is the ground-truth 3D position of joint i, $\widehat{J}_i$ is the preliminarily reconstructed position of joint i, and $\widehat{J}_i^{\text{ref}}$ is the final prediction after refinement.

For the aligned joint loss, the predicted joints are aligned with the ground truth using orthogonal Procrustes Analysis (PA) [31], and then the L2 loss is applied to the aligned and the ground-truth joints:

$$\begin{gathered}\mathcal{L}_{\text{alignJ}} = \frac{1}{K}\sum_{i=1}^{K}\left\|J_i - \widetilde{J}_i\right\|_2^2 + \frac{1}{K}\sum_{i=1}^{K}\left\|J_i - \widetilde{J}_i^{\text{ref}}\right\|_2^2 , \\ \text{where } \widetilde{J} = \text{PA}(J, \widehat{J}) \text{ and } \widetilde{J}^{\text{ref}} = \text{PA}(J, \widehat{J}^{\text{ref}}) \,.\end{gathered} \tag{4}$$

The procedure of alignment $\text{PA}(J, \widehat{J})$ can be described as follows.

[1]Our RefineNet module is different from the dense prediction network proposed in [30] despite the same name.

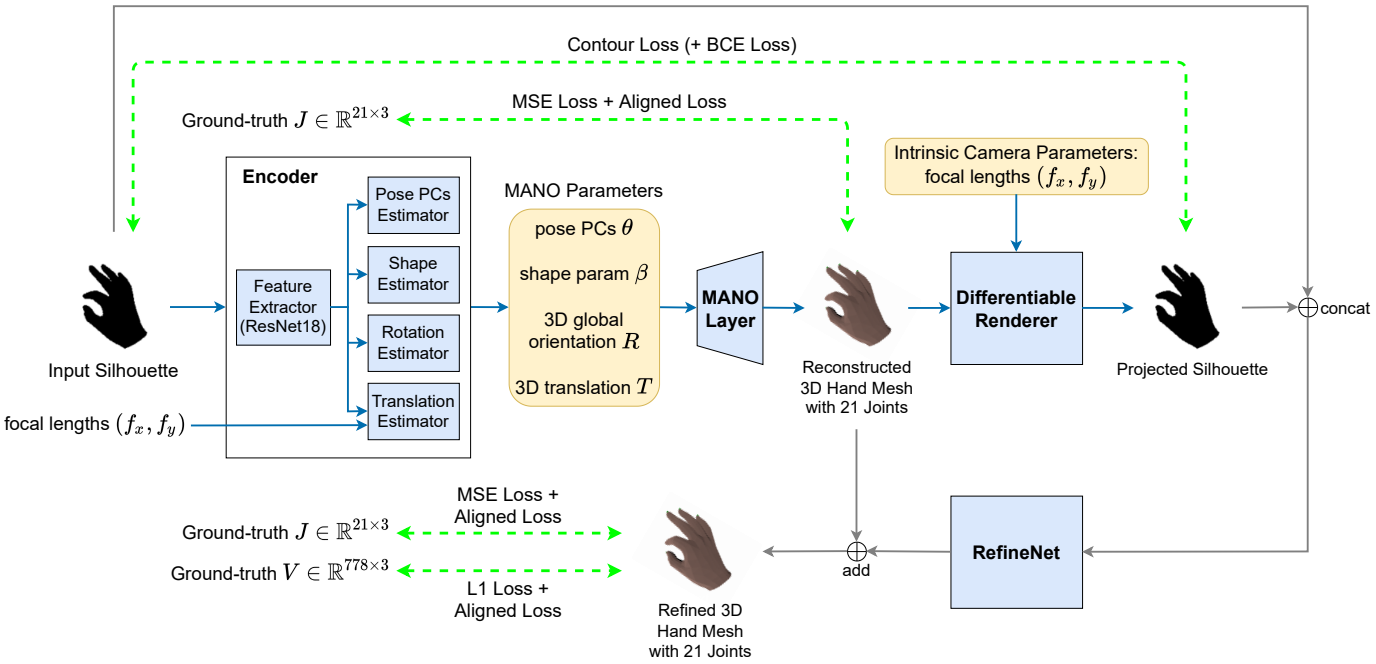


Fig. 2. The Mask2Hand network comprises the Encoder, MANO Layer, Differentiable Renderer, and RefineNet. The Encoder takes a binary mask as the input and generates the required parameters for the MANO Layer; the parameters include the shape, the six or forty-five pose-related principal components of MANO, the 3D global orientation, and translation of the hand. Based on these parameters, the MANO Layer reconstructs the 3D hand mesh (represented by 778 vertices) with the 21 positions of the hand joints. The Differentiable Renderer generates the 2D projected silhouette of the hand mesh, and the RefineNet can compare the projection with the input binary mask to refine the entire model. We design different loss functions for different modules in our model. The green dashed lines indicate what losses are applied to the associated modules. Please refer to the text for the detailed formulations of the loss functions.

1) Translate the joints so that their mean lies at the origin:

$$J \leftarrow J - \frac{1}{K}\sum_{i=1}^{K} J_i\,, \quad \widehat{J} \leftarrow \widehat{J} - \frac{1}{K}\sum_{i=1}^{K} \widehat{J}_i\,.$$

2) Divide the joints set by its Frobenius norm to remove the uniform scaling component:

$$J \leftarrow J/\|J\|_F\,, \quad \widehat{J} \leftarrow \widehat{J}/\|\widehat{J}\|_F\,.$$

3) Solve the orthogonal Procrustes problem: given matrices J and $\widehat{J}$, find an orthogonal matrix $Q \in \mathbb{R}^{3\times3}$ that maps J most closely to $\widehat{J}$, *i.e.*, $\min_Q \|JQ - \widehat{J}\|_F$ subject to $Q^\top Q = I_{3\times3}$. The solution consists of $Q = UV^\top$ and $s = \sum_i \sigma_i$, where $U\Sigma V^\top$ is the singular value decomposition of $J^\top \widehat{J}$ and σ_i's are the singular values.
4) $\widetilde{J} = \left(\widehat{J}Q^\top \times s\right) \times \|J\|_F + \left(\sum_{i=1}^{K} J_i\right)/K$, where the Procrustes output is scaled by the norm and shifted by the mean that we derive from J in the first two steps.

2) Silhouette Loss.: For the module of differentiable rendering, the ideal result is that the 2D projected hand silhouette matches the input binary mask exactly. Hence, our silhouette loss is based on the binary cross-entropy (BCE) and the contour loss:

$$\mathcal{L}_{\text{silhouette}} = \lambda_{\text{bce}}\text{BCE}(S, B) + \lambda_{\text{contour}}\text{ContourLoss}(S, B)\,, \tag{5}$$

where S is the rendered silhouette and B is the input binary mask. Our contour loss is a differentiable estimation of Chamfer distance between two contours, calculated as a pixel-wise multiplication between the contour of rendered silhouette and the pre-computed distance field of the ground-truth silhouette's contour. The detailed steps for computing the Chamfer distance are summarized as follows:

1) Compute the distance transform [32] for the contour ∂B of the ground-truth binary mask B in the dataset to get the distance field $\Psi_{\partial B}$.
2) Apply differentiable binarization to the rendered silhouette image S. In our case, we convert a pixel value $x \in [0, 1]$ in S to an approximately binary value as $\frac{1}{1+e^{-100\,(x-0.5)}} \simeq \{0, 1\}$.
3) Zero out the pixel values that are greater than the threshold 0.5 to deal with the noise near the wrist caused by PyTorch3D differentiable rendering. The result of Steps 2 & 3 is a clean and nearly binarized silhouette $\overline{S}$.
4) Apply Laplacian operator to $\overline{S}$ to get $\Delta\overline{S}$. The contour $\partial\overline{S}$ is then obtained by $\partial\overline{S} = \tanh(\max\{\Delta\overline{S}, 0\})$, which maps the values on the contour to near 1 and all the rest to 0.
5) Do element-wise product between the resulting contour $\partial\overline{S}$ and the distance field $\Psi_{\partial B}$ to retrieve the corresponding distance values and then take the sum to get $\text{ContourLoss}(S, B)$.

All of the aforementioned computations are differentiable. Therefore, we can readily include the silhouette loss in our method, and end-to-end train the network.

3) Mesh Loss.: We employ the mesh loss to measure the error in predicting the 3D vertices of the hand mesh. Like the pose loss, the mesh loss is also evaluated on both the unaligned and aligned predictions:

$$\mathcal{L}_{\text{mesh}} = \lambda_{\text{V}}\mathcal{L}_{\text{V}} + \lambda_{\text{alignV}}\mathcal{L}_{\text{alignV}} \,. \tag{6}$$

The unaligned vertex loss $\mathcal{L}_{\text{V}}$ consists of the L1 losses for the predictions from the preliminary and refined meshes:

$$\mathcal{L}_{\text{V}} = \frac{1}{M}\sum_{i=1}^{M}\left\|V_i - \widehat{V}_i\right\|_1 + \frac{1}{M}\sum_{i=1}^{M}\left\|V_i - \widehat{V}_i^{\text{ref}}\right\|_1 , \tag{7}$$

where $V \in \mathbb{R}^{M\times 3}$ is the ground-truth 3D vertices of the hand mesh, $\widehat{V}$ is the preliminarily reconstructed mesh, and $\widehat{V}^{\text{ref}}$ is the final prediction of vertices after refinement.
For the aligned vertex loss $\mathcal{L}_{\text{alignV}}$, the predicted vertices are aligned with the ground truth using orthogonal Procrustes analysis as described in the aligned pose loss Eq. (4), and then the L1 loss is applied to the aligned and the ground-truth mesh as

$$\mathcal{L}_{\text{alignV}} = \frac{1}{M}\sum_{i=1}^{M}\left\|V_i - \widetilde{V}_i\right\|_1 + \frac{1}{M}\sum_{i=1}^{M}\left\|V_i - \widetilde{V}_i^{\text{ref}}\right\|_1 , \tag{8}$$
$$\text{where } \widetilde{V} = \text{PA}(V, \widehat{V}) \text{ and } \widetilde{V}^{\text{ref}} = \text{PA}(V, \widehat{V}^{\text{ref}}) \,.$$

IV. Experiments

Since the problem formulation of the task that we aim to address is different from those commonly adopted in prior work, we have not found state-of-the-art methods that take exactly the same input setting and could therefore be considered for direct comparison with our approach. Note that the work of Lee *et al.* [14] is only posted on arXiv as a preprint without available code. Therefore, we do not include their method in the following comparisons. Moreover, their method requires depth maps during training, which is more restrictive than ours. In our experiments, we first compare our method with the state-of-the-art depth-based model A2J [12] on a synthetic dataset and show that transferring our task for an existing model to solve is probably not as straightforward as it may seem. Further, we use the real data in FreiHAND to compare our method with the three state-of-the-art RGB-based methods: I2L-MeshNet [22], MeshTransformer [7], and CMR (ResNet18) [8]. The results show that our method performs particularly well under the unaligned evaluation setting of 3D hand pose estimation, which is more practical for real applications.

A. Datasets

We conduct experiments on the following two datasets.

a) Synthetic Dataset.: This dataset contains 20,000 training and 2,000 test binary images synthesized using MANO [1], [33]. For each sample in the dataset, its hand-pose principal components $\theta \in \mathbb{R}^6$ are randomly sampled from the uniform distribution between $[-2.0, 2.0)$, and each rotation angle of its 3D global orientation R is randomly sampled from the uniform distribution $[-\pi, \pi)$. We use MANO to generate the corresponding hand shapes from the sampled parameters. The ground-truth 3D coordinates of the 21 hand joints can thus be automatically derived from the MANO rendered hand shapes.

b) FreiHAND [34].: It is a common real-world RGB dataset with 32,560 training samples and 3,960 test samples. We binarize the segmentation masks available in the training set and use them as the input to our model. However, since its original evaluation set does not provide segmentation masks, we resort to splitting its original training set into a random partition of 26,000 training, 3,280 validation, and 3,280 test images to achieve a reasonably fair evaluation of our method.

For quantitative evaluation, we report the standard metrics used by the FreiHAND dataset on the predictions of 3D joints and meshes:

- **MPJPE**: the *mean per joint position error*, which measures the Euclidean distance (in cm) between the ground-truth joints and the predicted joints.
- **MPVPE**: the *mean per-vertex position error*, which measures the Euclidean distance (in cm) between the ground-truth vertices and the predicted vertices of a mesh.
- **AUC of PCK**: the *area under the curve* of the *percentage of correct keypoints*, which is plotted using 100 equally-spaced thresholds between 0 cm to 5 cm.
- **AUC of PCV**: the *area under the curve* of the percentage of correct vertices.
- **PA-MPJPE** or **PA-MPVPE**: It first applies Procrustes alignment between the ground truth and the prediction, and then calculates MPJPE or MPVPE. This metric aims to measure reconstruction error that neglects the effect of global rotation, translation, and scaling.
- **mIoU** and **Dice coefficient**: In addition to the preceding metrics used by the FreiHAND dataset for the evaluation of joint and mesh predictions, we also use the *mean intersection over union* and the *Dice coefficient* to evaluate the similarity between the input binary mask and the mask projected from the predicted hand mesh.

TABLE I
Comparison with the depth-based model A2J [12] on the synthetic dataset. (PA means Procrustes Aligned)

Method	Input	MPJPE ↓	AUC$_J$ ↑	PA-MPJPE	PA-AUC$_J$
A2J [12]	Mean depth + 5-view aug.	7.73 cm	0.35	1.93 cm	0.62
Ours	Binary (0, 255) + 5-view aug.	**0.87 cm**	**0.83**	**0.53 cm**	**0.89**

B. Comparison with State-of-the-Art Depth-based Methods

To demonstrate that state-of-the-art depth-based models cannot be easily transferred to the task with binary input, we

TABLE II
FREIHAND (**Not Aligned**). NOTE THAT THE MIOU AND THE DICE COEFFICIENT ARE ONLY RELEVANT TO THE BINARY INPUT CASE.

Method	Input	MPJPE (cm)↓	AUC of PCK↑	MPVPE (cm)↓	AUC of PCV↑	mIoU↑	Dice↑
Mesh-Transformer [7]	RGB	68.59	0.00	68.58	0.00	–	–
	Binary	68.59	0.00	68.59	0.00	0.20	0.12
CMR (ResNet18) [8]	RGB	3.49	0.42	3.49	0.42	–	–
	Binary	4.31	0.35	4.31	0.35	**0.90**	**0.90**
Ours (6 PCs)	Binary	**3.56**	0.40	**3.57**	0.40	0.88	0.87
Ours (45 PCs)		**3.56**	**0.41**	**3.57**	**0.41**	0.88	0.88

TABLE III
FREIHAND (**Procrustes Aligned**). THE ADDITIONAL ENTRIES OF PA-MPJPE AND PA-MPVPE OF I2L-MESHNET WITH RGB INPUT ARE COPIED FROM THE ORIGINAL PAPER [22] AND POSTED HERE FOR THE EASE OF REFERENCE, WHICH CAN BE CONSIDERED AS A STRONG BASELINE.

Method	Input	PA-MPJPE (cm)↓	AUC of PCK↑	PA-MPVPE (cm)↓	AUC of PCV↑
I2L-MeshNet [22]	RGB	0.74	–	0.76	–
Mesh-Transformer [7]	RGB	0.54	0.89	0.58	0.88
	Binary	0.60	0.88	0.64	0.87
CMR (ResNet18) [8]	RGB	0.67	0.87	0.68	0.87
	Binary	0.77	0.85	0.77	0.85
Ours (6 PCs)	Binary	0.73	0.85	0.76	0.85
Ours (45 PCs)		0.68	0.86	0.69	0.86

compare our model with A2J [12] on the synthetic dataset. The input image of our model is binarized to $\{0, 255\}$ while that of A2J is binarized to 0 (as the background) and the mean depth of joints to treat the depth-based model fairly. This way, in spite of the fact that the input image is binary-valued, we still provide sufficient statistics of depth information to A2J. As shown in Table I, for the unaligned setting, our model achieves MPJPE of 0.87 cm, and A2J reaches MPJPE of 7.73 cm. The results indicate that A2J fails to learn to localize 3D hand joints accurately when the input is binary, even if the reference mean depth is given. In contrast, our method has the promising ability to recover the positions of the joints in the camera space. In terms of MPJPE after Procrustes alignment, our approach also outperforms A2J by a large margin, as presented in Table I. Some other depth-based models are not quantitatively compared with ours for the following reasons. V2V-PoseNet [35] takes a 3D voxelized depth map as its input, which is expected to fail for silhouette-based input since the voxelization process cannot work for binary images. Hand PointNet [36] converts a depth map into a 3D point cloud and then uses it as the model's input. Likewise, the conversion procedure cannot function properly for silhouettes, making it inadequate for the task of binary input.

C. Comparison with State-of-the-Art RGB-based Methods

For the comparison with state-of-the-art RGB-based methods, we select the leading models from the FreiHAND competition leaderboard at CodaLab [2], including CMR [8] and MeshTransformer [7]. To make a fair comparison with our method, we use the officially released code to re-train their models on the FreiHAND dataset randomly partitioned by us, as mentioned in Section IV-A, under the settings of both binary and RGB input. The former is for performance observation, while the latter is primarily for convincing proof that these models are not trained imperfectly. As presented in Table II, CMR achieves MPJPE and MPVPE of 4.31 cm with silhouette-based input, while MeshTransformer fails to localize 3D hand joints and mesh in the camera space. Our model reaches MPJPE of 3.56 cm and MPVPE of 3.57 cm, which outperforms the previous two methods by a large margin and is even comparable to the result of CMR with RGB input. The superior performance of our model in the unaligned case indicates that it has a better ability to reconstruct 3D hand mesh in the camera space, which is more practical in real applications. In terms of mIoU and Dice coefficient, our model achieves similar results as CMR.

[2]https://competitions.codalab.org/competitions/21238

TABLE IV
ABLATION STUDY. (MEANING OF ⋇ IN THE FIRST COLUMN: THE ALIGNED LOSS IS ONLY APPLIED TO THE REFINED JOINTS AND VERTICES. (A FOR ALIGNED LOSS, B FOR BCE LOSS, C FOR CONTOUR LOSS, R FOR REFINE NET, M FOR MESH LOSS.)

A	B	C	R	M	MPJPE (cm)↓	MPVPE (cm)↓	PA-MPJPE (cm)↓	PA-MPVPE (cm)↓	mIoU↑
✓		✓	✓	✓	3.55	3.56	0.78	0.79	0.87
✓		✓	✓		3.67	3.84	0.86	1.18	0.76
✓		✓			3.73	3.73	0.94	0.94	0.83
✓			✓	✓	3.60	3.61	0.80	0.81	0.87
		✓	✓	✓	3.83	3.91	1.36	1.56	0.86
⋇			✓	✓	3.66	3.67	0.79	0.80	0.86
✓	✓	✓	✓	✓	3.56	3.57	0.78	0.79	0.87
✓	✓	✓			3.72	3.73	0.90	0.93	0.83
	✓	✓	✓	✓	3.66	3.67	1.04	1.06	0.88

When the predicted joints and vertices are aligned with the ground truth by Procrustes analysis, our approach reaches PA-MPJPE of 0.68 cm and PA-MPVPE of 0.69 cm. These experimental results are significantly better than CMR's performance under the setting of binary input and are comparable to CMR's results with RGB input. Furthermore, unlike MeshTransformer, which totally fails in the unaligned case, our method performs stably well in both cases. Such a balanced behavior is preferable in our task since localization and detail recovery are vitally important for projecting a silhouette similar to the input one.

D. Implementation Details

Our encoder's backbone is based on ResNet18 [27]. PyTorch3D's implementation [29] is employed for our differentiable renderer. We use the Adam optimizer to train our network with a batch size of 32. The initial learning rate is set to 10^{-4}, and we use the ReduceOnPlateau scheduler to adjust the learning rate. The whole network is trained for 150 epochs, and the checkpoint with the lowest validation loss is selected for our final usage. For the FreiHAND dataset, we apply data augmentation of random rotation between $[-\pi, \pi)$ and scaling between $[\,0.9, 1.1)$ in the 2D image space. For the synthetic dataset, we randomly sample only one of the five views for each training data at every epoch to perform data

augmentation. We set $\lambda_{\mathrm{J}} = 2 \times 10^{-3}$, $\lambda_{\mathrm{alignJ}} = 2 \times 10^{-2}$, $\lambda_{\mathrm{contour}} = 10^{-4}$, $\lambda_{\mathrm{V}} = 0.1$, and $\lambda_{\mathrm{alignV}} = 1$ to balance each loss term. λ_{bce} is set to 0 for our final model and 0.5 for the ablation study.

E. Ablation Study

We evaluate different configurations of our method to analyze the effectiveness of each component and examine the effects of the loss functions.

1) Effects of Shape Parameters.: In our ablation study, we use six pose PCs and set β to the mean shape, *i.e.*, $\beta = \mathbf{0}$. By comparing the first row of Table IV and the penultimate row of Tables II & III, we can observe that regressing shape parameters in the first stage of our model helps reduce the burden of refinement and achieve better performance in the aligned case.

2) Effectiveness of Each Component.: As shown in Table IV, we explore the evaluation results on the FreiHAND dataset with the exclusion of different components. Our base model consists of Encoder, MANO Layer, and Differentiable Renderer. After incorporating the RefineNet module into our network with pose loss only, we observe that both MPJPE and PA-MPJPE decrease while MPVPE and PA-MPVPE undesirably increase. This result is in line with our expectation since the pose loss cannot impose strong enough constraints on the deformation of the preliminarily constructed hand mesh. Based on the observation, we further apply the mesh loss to the output of the RefineNet, which helps to strengthen the functionality of refinement and improve all evaluation metrics significantly.

To reveal the effects of other losses, we exclude them one at a time from our entire network. First, we remove the contour loss between the input binary image and the projected silhouette. The evaluation result shows that the model's performance on both hand pose and shape estimation degrades, especially the unaligned ones, demonstrating the effectiveness of the contour loss for achieving more accurate global mesh registration. Second, we eliminate the aligned losses of both joints and vertices. Our experimental result shows that MPJPE, MPVPE, PA-MPJPE, and PA-MPVPE increase by 0.28, 0.35, 0.58, and 0.77, respectively. Such a significant deterioration in performance indicates that the aligned losses not only play crucial roles in the reconstruction of fine-grained gestures but also help to improve mesh recovery in the camera space. Finally, we remove all losses inserted before RefineNet. The result shown in the sixth row of Table IV suggests that adding the pose loss and contour loss before the refinement stage can help our model improve the ability to localize the 3D hand mesh in the camera space.

3) Effects of Adding the BCE Loss.: We also do some experiments to see whether adding the binary cross-entropy loss between the input binary mask and the projected silhouette has any positive effect. As listed in Table IV, directly incorporating the BCE loss into our entire architecture contributes nothing to the final performance. To understand the root cause of this counterintuitive phenomenon, we take away the RefineNet module and the aligned loss one at a time with the presence of BCE loss. The experimental results show that the BCE loss can enhance our model's performance by a large margin when the aligned losses are absent. This implies that the BCE loss does not have much effect when the model learns the shape estimation task to a certain degree guided by the stronger training objectives, *i.e.*, the aligned losses of joints and vertices.

V. Conclusion

This paper introduces the Mask2Hand network that learns to predict the 3D hand pose and shape from a 2D binary mask without relying on any RGB or depth information. Despite the more challenging input setting of our task in comparison with prior work on hand pose and shape estimation, we show that the proposed method can achieve comparable performance as state-of-the-art RGB-based and depth-based methods. With the parametric hand model and the differentiable rendering technique, we integrate the self-supervised mechanism into our end-to-end training process without the need for human annotations. We propose several loss functions to model different aspects of 3D hand pose and shape estimation. Since our method explicitly learns to predict the global transformation of the hand, its better performance in the unaligned setting is thus promising for real applications.

Acknowledgments

This work was supported in part by Industrial Technology Research Institute, Taiwan. We are grateful to National Center for High-performance Computing for providing computational resources and facilities.

References

[1] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017.

[2] T. Li, X. Xiong, Y. Xie, G. Hito, X. Yang, and X. Zhou, "Reconstructing hand poses using visible light," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, 71:1–71:20, 2017.

[3] D. Ma, G. Lan, M. Hassan, *et al.*, "Solargest: Ubiquitous and battery-free gesture recognition using solar cells," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, 12:1–12:15.

[4] A. Varshney, A. Soleiman, L. Mottola, and T. Voigt, "Battery-free visible light sensing," in *Proceedings of the 4th ACM Workshop on Visible Light Communication Systems*, 2017, pp. 3–8.

[5] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.

[6] J. Won and J. Lee, "Shadow theatre: Discovering human motion from a sequence of silhouettes," *ACM Trans. Graph.*, vol. 35, no. 4, 147:1–147:12, 2016.

[7] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *CVPR*, 2021, pp. 1954–1963.

[8] X. Chen, Y. Liu, C. Ma, *et al.*, "Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration," in *CVPR*, 2021, pp. 13 274–13 283.

[9] Y. Chen, Z. Tu, D. Kang, *et al.*, "Model-based 3d hand reconstruction via self-supervised learning," in *CVPR*, 2021, pp. 10 451–10 460.

[10] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang, "Semi-supervised 3d hand-object poses estimation with interactions in time," in *CVPR*, 2021, pp. 14 687–14 697.

[11] M. Li, Y. Gao, and N. Sang, "Exploiting learnable joint groups for hand pose estimation," in *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021, pp. 1921–1929.

[12] F. Xiong, B. Zhang, Y. Xiao, *et al.*, "A2J: anchor-to-joint regression network for 3d articulated pose estimation from a single depth image," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 793–802.

[13] L. Fang, X. Liu, L. Liu, H. Xu, and W. Kang, "JGR-P2O: joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image," in *European Conference on Computer Vision (ECCV)*, 2020.

[14] K. Lee, S. Liu, H. Chen, and K. Ito, "Silhouettenet: 3d hand pose estimation from silhouettes," *CoRR*, vol. abs/1912.12436, 2019.

[15] A. Tkach, M. Pauly, and A. Tagliasacchi, "Sphere-meshes for real-time hand modeling and tracking," *ACM Trans. Graph.*, vol. 35, no. 6, 222:1–222:11, 2016.

[16] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 172–193, 2016.

[17] A. Boukhayma, R. A. de Bem, and P. H. S. Torr, "3D hand shape and pose from images in the wild," in *CVPR*, 2019, pp. 10 843–10 852.

[18] S. Baek, K. I. Kim, and T. Kim, "Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering," in *CVPR*, 2019, pp. 1067–1076.

[19] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *European Conference on Computer Vision (ECCV)*, 2020.

[20] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez, "GanHand: Predicting human grasp affordances in multi-object scenes," in *CVPR*, 2020, pp. 5030–5040.

[21] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *CVPR*, 2020, pp. 568–577.

[22] G. Moon and K. M. Lee, "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single RGB image," in *European Conference on Computer Vision (ECCV)*, 2020.

[23] M. Seeber, R. Poranne, M. Polleyfeys, and M. R. Oswald, "Realistichands: A hybrid model for 3d hand reconstruction," in *International Conference on 3D Vision (3DV)*, 2021, pp. 22–31.

[24] J. Wang, F. Mueller, F. Bernard, *et al.*, "Rgb2hands: Real-time tracking of 3d hand interactions from monocular RGB video," *ACM Trans. Graph.*, vol. 39, no. 6, 218:1–218:16, 2020.

[25] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular RGB image," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2354–2364.

[26] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu, "Monocular real-time hand shape and motion capture using multi-modal data," in *CVPR*, 2020, pp. 5345–5354.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[28] H. Kato, D. Beker, M. Morariu, *et al.*, "Differentiable rendering: A survey," *CoRR*, vol. abs/2006.12057, 2020.

[29] N. Ravi, J. Reizenstein, D. Novotný, *et al.*, "Accelerating 3d deep learning with pytorch3d," *CoRR*, vol. abs/2007.08501, 2020.

[30] G. Lin, F. Liu, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1228–1242, 2020.

[31] P. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[32] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," *Theory Comput.*, vol. 8, no. 1, pp. 415–428, 2012.

[33] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision (ECCV)*, 2020.

[34] C. Zimmermann, D. Ceylan, J. Yang, B. C. Russell, M. J. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single RGB images," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 813–822.

[35] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *CVPR*, 2018, pp. 5079–5088.

[36] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *CVPR*, 2018, pp. 8417–8426.

A Reversible Image Processing Method for Color Tone Control Using Data Hiding

Daichi NAKAYA* and Shoko IMAIZUMI†
* Graduate School of Science and Engineering, Chiba University, Chiba, Japan
E-mail: daidai2363@chiba-u.jp
† Graduate School of Engineering, Chiba University, Chiba, Japan
E-mail: imaizumi@chiba-u.jp

***Abstract*—Reversible image processing methods have been actively researched by using data hiding techniques. In these methods, additional information for reversibility should be embedded into the processed image. Their main targets for image processing are luminance and saturation. In this paper, we focus on tone control that specifically enhances blue or red tone with reversibility. Through our simulations, we prove that our method can effectively enhance color tones and the data embedding process does not severely affect the quality of output images.**

I. Introduction

With the development of the Internet and social networking services, we widely use image processing applications. In general image processing, however, it is difficult to restore the original image after processing. The processing history or original image itself should be stored separately, when we desire to reconstruct the original image. Recently, reversible image processing methods based on reversible data hiding (RDH) have been paid attention. In particular, reversibility is strongly required in medical, satellite and military images. Particularly, many contrast enhancement methods with reversibility have been proposed [1]–[4]. Wu et al. proposed a histogram shifting (HS) based RDH method for grayscale images to achieve reversible contrast enhancement [1]. In this method, additional information is embedded into the image histogram to ensure the reversibility. Further, Wu et al. introduced pairs of adjacent pixels and applied an HS-based RDH method to the two-dimensional histogram. The method allows for reversible contrast enhancement while preserving the overall average luminance [4]. However, these methods are the techniques for grayscale images. When we apply them to color images with three color components, the hue may be distorted. Therefore, Wu et al. proposed another method that enhances the contrast of color images by referring to the HSV color space [5]. The method could enhance the contrast without the severe distortion of the hue and saturation. Nevertheless, the method does not completely restore the original image due to the rounding errors in most cases.

Here, a processed image is defined as an image that has no information for the reversibility. In contrast, an output image represents an image containing the additional information for the reversibility in this paper. Sugimoto et al. ensured the reversibility by embedding the additional information into the processed images to correct the rounding errors [6], [7]. While most previous methods have only discussed on reversible contrast enhancement, the method [7] enables not only contrast enhancement but also multiple functions with the reversibility. Moreover, these functions are independently executed from each other. The method uses the prediction-error expansion with HS (PEE-HS) method [8] in order to embed the additional information. Although the method allows multiple processing for saturation and luminance, the hue control is not considered.

In this paper, we propose a reversible tone control method that effectively enhances blue or red tone. The proposed method further mitigates visual distortion; the additional information is embedded into the B component of RGB images in enhancing red tone, and vice versa. Simulation results show the effectiveness of our color tone enhancement and also evaluate the effects of the data embedding process on the output image quality.

II. Preparation

A. Previous method [7]

Sugimoto et al. proposed a reversible image processing method that enables multiple functions for the luminance and saturation [7]. This method has five functions for the luminance component: contrast enhancement, sharpening/smoothing, and increasing/decreasing. Additionally, we can also improve the saturation. These functions can be applied independently from each other, and multiple functions can be combined together.

The HSV color space is used in the previous method. This space consists of the hue (H), saturation (S), and luminance (V). The color space conversion from RGB to HSV is given by

$$\begin{cases} H = \begin{cases} 60 \times \frac{G-B}{Max-Min} & \text{if } \textit{Max = R} \\ 60 \times (\frac{B-R}{Max-Min} + 2) & \text{if } \textit{Max = G} \\ 60 \times (\frac{R-G}{Max-Min} + 4) & \text{if } \textit{Max = B} \\ \text{undefined} & \text{if } \textit{Max = 0}, \end{cases} \\ S = Max - Min, \\ V = Max, \end{cases} \tag{1}$$

where Max and Min represent the maximum and minimum values in the RGB components of each pixel, respectively.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

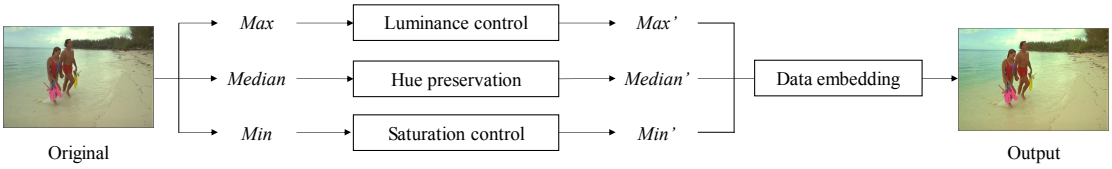


Fig. 1 Block diagram of previous method.

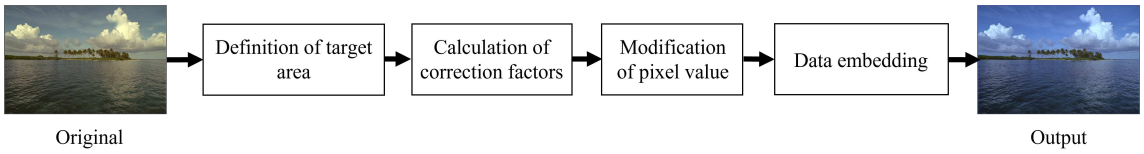


Fig. 2 Block diagram of tone control process.

Note that the median value is called $Median$. Figure 1 shows the block diagram of the previous method. The previous method indirectly controls HSV components by controlling Max, $Median$, and Min that are originally the R, G, or B value.

1) Luminance control: The luminance component is defined as an 8-bit grayscale image composed of the Max values of each pixel. The contrast is enhanced by applying the histogram equalization based RDH method [1] to the luminance component. The sharpening and smoothing functions are conducted by applying the specific spatial filter to the luminance component. Further, the luminance can be increased or decreased by raising or reducing Max. As can be seen in Eq. (1), however, Max affects not only the luminance component but also the saturation component. Therefore, the saturation is regularly distorted by changing Max simultaneously. In the previous method, Min should be adjusted to preserve the saturation before and after the process.

2) Saturation control: Based on Eq. (1), the saturation can be improved by increasing Max or decreasing Min. In the previous method, the saturation improvement is executed by decreasing Min in order to prevent from interfering with the luminance control. The intensity of improvement can be controlled by the decreasing amount of Min.

3) Hue preservation: Hue can be preserved by maintaining C_H. C_H is defined by

$$C_H = \frac{Median - Min}{Max - Min} \simeq \frac{Median' - Min'}{Max' - Min'}, \qquad (2)$$

where Max' and Min' refer to Max and Min after the processes in II-A1 and II-A2, respectively. $Median$ should be adjusted so as to preserve C_H without significant change.

4) Data embedding: Each of the above processes requires additional information to ensure the reversibility. For example, in section II-A2, Min is excluded from the process and stored as a part of additional information when Min should have a negative value after the process. The additional information is embedded by using the PEE-HS method [8].

The previous method achieved a full reversible image processing for the luminance and saturation without affecting the hue component. Thus, the hue control was not considered in the method. In the next section, we propose a reversible image processing method that controls blue or red tone.

B. $\mathrm{YD_bD_r}$ color space

The $\mathrm{YD_bD_r}$ color space consists of the luminance (Y), blue chrominance ($\mathrm{D_b}$), and red chrominance ($\mathrm{D_r}$). This space is used in JPEG2000 [11], which is one of the international standards on image coding. The color space conversion from RGB to $\mathrm{YD_bD_r}$ is given by

$$\begin{cases} Y = \lfloor \frac{R+2G+B}{4} \rfloor, \\ D_b = B - G, \\ D_r = R - G. \end{cases} \qquad (3)$$

The conversion between RGB and $\mathrm{YD_bD_r}$ can be conducted reversibly. Since $\mathrm{D_b}$ and $\mathrm{D_r}$ are obtained as integer values without causing rounding errors, we can flexibly control blue and red tones with the smaller amount of additional information.

III. Proposed Method

In this section, we propose a reversible image processing method that arbitrarily controls blue or red tone of color images. We describe the procedures of the tone control and restoration processes, respectively.

A. Tone control process

Figure 2 shows the block diagram of the tone control process in the proposed method. The RGB color space is first converted to the $\mathrm{YD_bD_r}$ color space, and then a target area is defined within the D_b-D_r plane. Here, the target area refers to an area, where the average color becomes an achromatic color after the tone control process. It should be noted that the $\mathrm{YD_bD_r}$ color space is used as a reference space, while the actual processing is performed in the RGB color space. The correction factors bring the average color of the target area

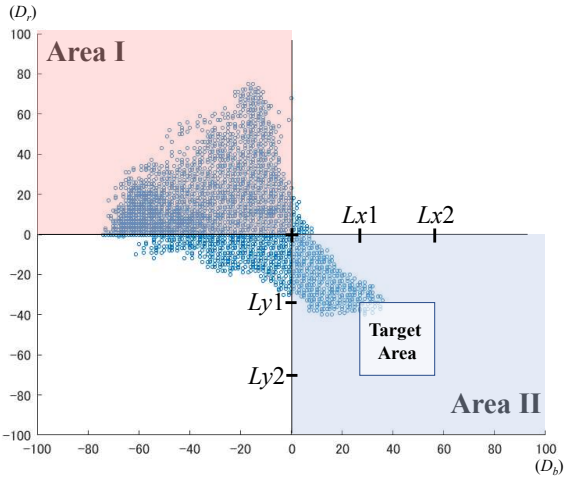


Fig. 3 Scatter plot of pixels in D_b-D_r plane.

to an achromatic color; they are calculated and multiplied by all pixels in the image. The RGB components of each pixel are represented as decimals, so they are rounded to integer values. Finally, the additional information required for the reversibility is embedded into the processed image using the PEE-HS method [8]. Detailed descriptions of each process are given below.

1) Definition of target area: The conversion from the RGB color space to the $\mathrm{YD_bD_r}$ color space is performed using Eq. (3). Figure 3 shows the scatter plot of pixels in the D_b-D_r plane. From this figure, the target area is determined from either area I or II. In defining the target area in area I, we enhance the blue tone. In contrast, we can enhance the red tone by defining the target area in area II. Additionally, the proposed method can control the enhancement intensity by adjusting the size and position of the target area using parameters *Lx*1, *Lx*2, *Ly*1, and *Ly*2. *Lx*1 and *Lx*2 ($|Lx1|<|Lx2|$) are located along the x-axis, while *Ly*1 and *Ly*2 ($|Ly1|<|Ly2|$) are located along the y-axis. Each value represents the distance from the origin. Therefore, these four parameters form a rectangular target area.

2) Calculation of correction factors: The averages of R, G, B, and Y of the pixels in the target area are represented as R_{ave}, G_{ave}, B_{ave}, and Y_{ave}, respectively. Correction factors are then calculated to ensure that R_{ave}, G_{ave}, and B_{ave} are analogous to each other. This leads to make the average color achromatic. R_{ave}, G_{ave}, and B_{ave} are rounded to integer values and stored as additional information. The correction factors for RGB components are obtained by

$$\begin{cases} K_r = (\frac{Y_{ave}}{R_{ave}}), \\ K_g = (\frac{Y_{ave}}{G_{ave}}), \\ K_b = (\frac{Y_{ave}}{B_{ave}}). \end{cases} \tag{4}$$

Here, K_r, K_g, and K_b are factors for the R, G, and B components, respectively. Since these factors are obtained as decimals, each pixel value should be rounded after multiplying by these factors. These rounding errors prevent the reversible restoration. For the reversibility, the locations of pixels affected by rounding should be stored as additional information called a location map. The location map is preliminarily compressed using JBIG2 [12]. However, K_g constantly exist in the range of 0.98 to 0.99 regardless of the enhancement intensity. This indicates that the effect by K_g on the multiplication process is significantly small. In order to reduce the amount of additional information, we define $K_g = 1.0$ in any case.

3) Modification of pixel value: The proposed method may cause overflows in the target component to be enhanced when the pixel value is multiplied by the correction factor. Thus, it is necessary to preprocess on the component to be enhanced. We first explore the maximum value P_{max} in the target component, which remains less than 255 after multiplication by the correction factors. All pixels, which exceeds P_{max}, are then replaced with P_{max}. The replaced pixels and their variations are stored in additional information. After preprocessing, the R and B components are multiplied by the correction factors so that the processed image with a red or blue tone enhancement is obtained.

4) Data embedding: For the perfect reversibility, additional information is embedded into the processed image using the PEE-HS method [8]. Based on Eq. (3), the G component more strongly affects the Y component than the other components. Namely, variations in the G component lead to severer visual distortion. In the data embedding process, the G component is thus excluded from the target. The additional information is embedded into the B component in enhancing red tone, and vice versa. The component data used to embed the additional information is stored in the least significant bit (LSB) of the top-left pixel of the G component. The LSB is replaced to 0 in embedding into the B component, and replaced to 1 in embedding into the R component.

The additional information consists of the followings: R_{ave}, G_{ave}, B_{ave}, the data to revert the preprocessing, and the original LSB of the top-left pixel of the G component. The proposed method can restore the original image using the additional information.

B. Restoration process

Figure 4 shows the block diagram of the restoration process in the proposed method. We describe each step in detail.

1) Data extraction: First, the component used to embed the additional information is identified by referring to the LSB of the top-left pixel of the G component. We extract the additional information, and the processed image excluding the additional information is reconstructed. The LSB of the top-left pixel of the G component is restored with the extracted data.

2) Calculation of correction factor: By assigning R_{ave}, G_{ave}, and B_{ave} to Eq. (3), the average luminance Y_{ave} of the target area is obtained. Correction factors K_r and K_b are then calculated by Eq. (4).

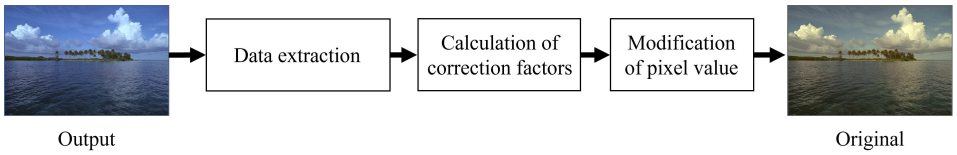


Fig. 4 Block diagram of restoration process.

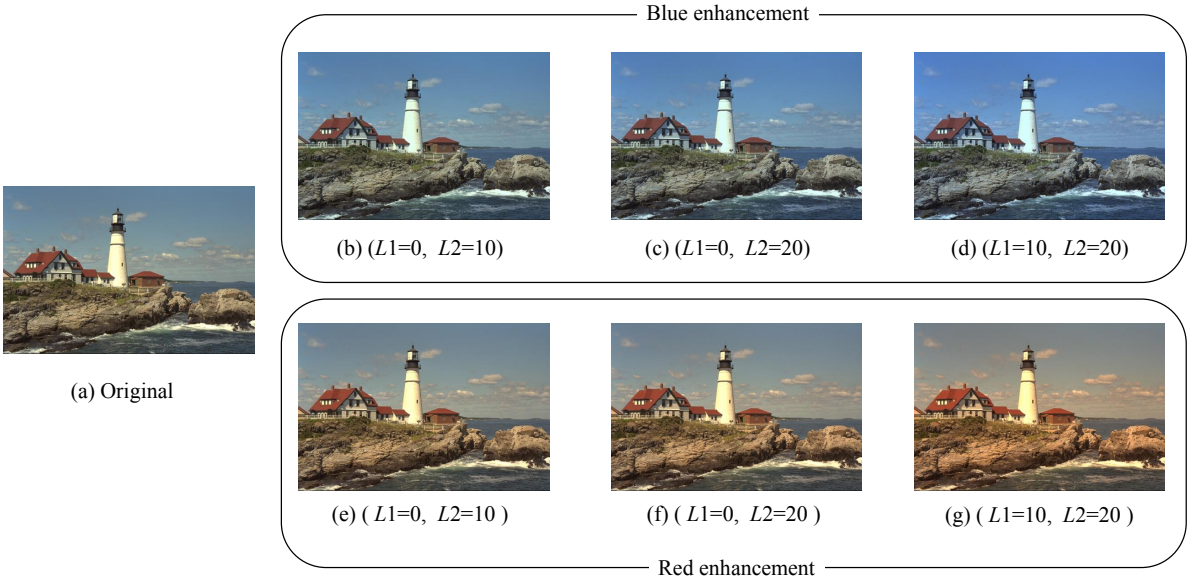


Fig. 5 Output images (kodim21).

3) Modification of pixel value: The R and B components of the processed image are divided by their respective correction factors so that the pixel values before correction are restored. By referring to the location map, the pixels affected by rounding are reverted to their originals. The pixels with P_{max} of the enhanced component are also reverted to the originals, and the original image is reconstructed.

IV. Simulations

In this section, we evaluate the image quality of output images through comparing scatter plots of pixels in the D_b-D_r plane between before and after the tone enhancement process.

A. Simulation conditions

We used five images with 512 × 768 pixels from the Kodak Lossless True Color Image Suite [13]. In this simulation, we defined each parameter as $|Lx1| = |Ly1|$ and $|Lx2| = |Ly2|$ to make the target area square, and denoted them as $L1$ and $L2$, respectively. The intensity of enhancement can be controlled by adjusting the size and position of the target area. We evaluated the output images obtained with parameters ($L1$, $L2$) = (0, 10), (0, 20), and (10, 20), respectively.

B. Evaluation of color tones for output images

Figure 5 shows the output images of kodim21 generated by the proposed method. The upper and lower images show the results of enhancing the blue or red tone, respectively. As the dimension of target area or the distance between the target area and origin is larger, the intensity of the enhancement strengthens. As can be seen from Figure 5, it is clear that ($L1$, $L2$) = (10, 20) is most effective for tone enhancement.

Figure 6 shows the scatter plots of pixels in the D_b-D_r plane for each image corresponding to Figure 5. The blue tone enhancement leads to the increment of D_b values and the decrement of D_r values, so the global distribution is shifted to the lower right direction. In contrast, in the red tone enhancement, the global distribution is shifted to the upper left direction. By increasing $L1$ or/and $L2$, the averages of D_b and D_r in the target area become farther from the origin. This provides the stronger enhancement for the target color tone.

Figure 7 shows the output images with ($L1$, $L2$) = (10, 20). In these output images, the enhancement of blue or red tone was effectively achieved. In each image, additional information was completely embedded, and thus the reversibility is guaranteed.

Table I shows the average correction factors calculated

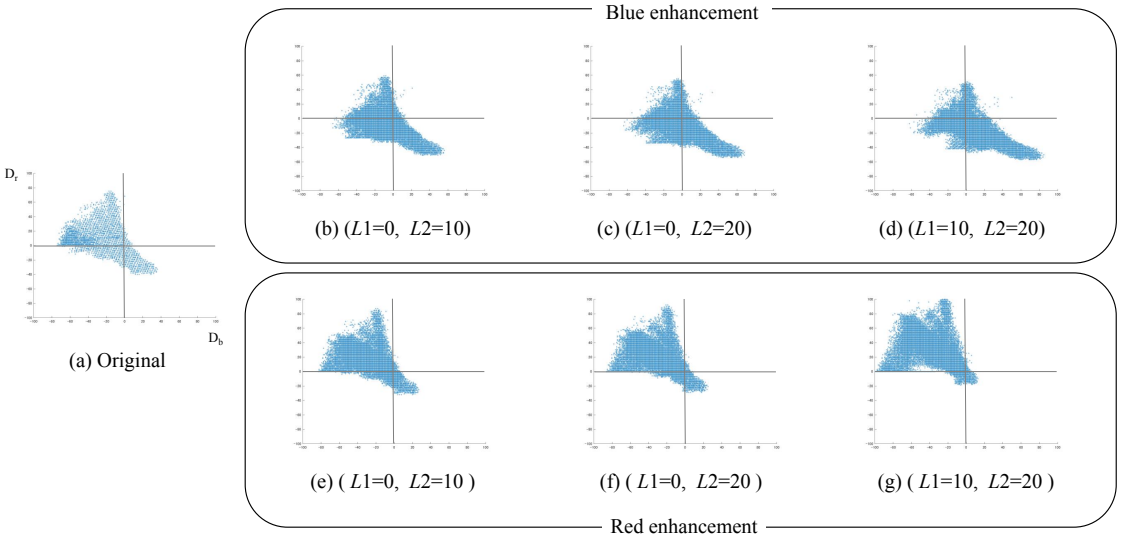


Fig. 6 Scatter plots of pixels in output images in D_b-D_r plane (kodim21).



Fig. 7 Output images in ($L1$, $L2$) = (10, 20).

for all the test images. It is obvious that the correction factor becomes larger/smaller by the increment of $L1$ than the increment of $L2$. This is because few numbers of pixels are distributed in the area far from the origin.

C. Effect of data embedding process on output image quality

We need to embed additional information into the processed image to ensure the reversibility. Generally, a data embedding process causes distortion in the marked image, i.e., output image. In this section, we examine the effect of the data embedding process on the output image quality. Table II shows the average values of PSNR and SSIM between the images before and after data embedding, i.e., the processed and output images. The quality of output images decreases with the increasing intensity of enhancement. As the correction factor becomes larger, the amount of additional information should increase. Nevertheless, the PSNR and SSIM values are relatively high, and it is difficult to visually recognize the artifacts in the output images.

V. Conclusions

We proposed a reversible tone control method that enhances blue or red tone in this paper. Since most previous methods aimed to only control the contrast enhancement, so the proposed method is a challenging approach. Through our

TABLE I Correction factors using tone control process.

	Blue enhancement		Red enhancement	
	K_b	K_r	K_b	K_r
($L1$, $L2$) = (0, 10)	1.1022	0.9289	0.9596	1.0632
($L1$, $L2$) = (0, 20)	1.1509	0.9075	0.9411	1.1126
($L1$, $L2$) = (10, 20)	1.3326	0.8277	0.8945	1.1715

TABLE II Quality assessment of output images.

	PSNR[dB]		SSIM	
	Blue enhancement	Red enhancement	Blue enhancement	Red enhancement
($L1$, $L2$) = (0, 10)	51.25	50.43	0.9968	0.9773
($L1$, $L2$) = (0, 20)	49.24	49.23	0.9930	0.9672
($L1$, $L2$) = (10, 20)	47.01	45.73	0.9974	0.9405

simulations, we confirmed that blue and red tones can be flexibly and reversibly controlled by using our method. In our future work, we will specifically investigate the effectiveness of the proposed method by using different image datasets and evaluation indices.

REFERENCES

[1] H.-T. Wu, J.-L. Dugelay, and Y.-Q. Shi, "Reversible image data hiding with contrast enhancement," IEEE Sig. Process. Lett., vol.22, no.1, pp.81-85, 2015.

[2] G. Gao and Y.-Q. Shi, "Reversible data hiding using controlled contrast enhancement and integer wavelet transform," IEEE Sig. Process. Lett., vol.22, no.11, pp.2078-2082, 2015.

[3] H. Chen, J. Ni, W. Hong, and T.-S. Chen, "Reversible data hiding with contrast enhancement using adaptive histogram shifting and pixel value ordering," Sig. Process. Image Commun., vol.46, pp.1-16, 2016.

[4] H. -T. Wu, X. Cao, R. Jia and Y. -M. Cheung, "Reversible data hiding with brightness preserving contrast enhancement by two-dimensional histogram modification," IEEE Trans. Circuits Syst. Video Technol., vol.32, no.11, pp.7605-7617, 2022.

[5] H.-T. Wu, Y. Wu, Z. Guan, and Y.-M. Cheung, "Lossless contrast enhancement of color images with reversible data hiding," Entropy, vol.21, no.910, 2019.

[6] Y. Sugimoto and S. Imaizumi, "An extension of reversible image enhancement processing for saturation and brightness contrast," J. Imaging, vol.8, no.27, 2022.

[7] Y. Sugimoto and S. Imaizumi,"Reversible image processing for color images with flexible control," Appl. Sci., vol.13, no.2297, 2023.

[8] D. M. Thodi and J. J. Rodriguez, "Expansion embedding techniques for reversible watermarking," IEEE Trans. Image Process., vol.16, no.3, pp.721-730, 2007.

[9] R. Motomura and S. imaizumi,"A reversible data-hiding method with prediction-error expansion in compressible encrypted images," Appl. Sci., vol. 12, no.19, 9418, 2022.

[10] F. Pasteau, C. Strauss, M. Babel, O. Deforges and L. Bedat, "Improved colour decorrelation for lossless colour image compression using the LAR codec,"17th Eur. Sig. Process. Conf., pp.2122-2126, 2009.

[11] "Information technology - JPEG 2000 image coding system - Part 1: Core coding system," International Standard ISO/IEC IS-15444-1, Dec. 2019.

[12] P. G. Howard, F. Kossentini, B. Martins, S. Forchhammer and W. J. Rucklidge, "The emerging JBIG2 standard," IEEE Trans. Circuits Syst. Video Technol., vol.8, no.7, pp.838-848, 1998.

[13] [Online] Available:URL:http://www.r0k.us/graphics/kodak/

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Image-Text Out-Of-Context Detection Using Synthetic Multimodal Misinformation

Fatma Shalabi*[1,2], Huy H. Nguyen*[2], Hichem Felouat[1,2], Ching-Chun Chang[2], and Isao Echizen[1,2,3]
*These authors contributed equally
[1]The Graduate University for Advanced Studies, SOKENDAI, Japan
[2]National Institute of Informatics, Japan [3]The University of Tokyo, Japan
E-mail: {fatmafaek, nhhuy, hichemfel, cchang, iechizen}@nii.ac.jp

***Abstract*—Misinformation has become a major challenge in the era of increasing digital information, requiring the development of effective detection methods. We have investigated a novel approach to Out-Of-Context detection (OOCD) that uses synthetic data generation. We created a dataset specifically designed for OOCD and developed an efficient detector for accurate classification. Our experimental findings validate the use of synthetic data generation and demonstrate its efficacy in addressing the data limitations associated with OOCD. The dataset and detector should serve as valuable resources for future research and the development of robust misinformation detection systems.**

I. Introduction

The widespread dissemination of misinformation online, whether in news articles or brief tweets, represents a serious threat to information security and undermines public confidence in online information. The spread of rumors and misinformation can negatively affect people in many ways, such as by being influenced by misleading information, by being victimized by hate speech, by being aggrieved by racism, and by suffering psychological distress [1], [2]. Those who generate online misinformation often utilize multimodal misinformation, such as images and texts, to maximize the effect [3].

Since misinformation is usually accompanied by images to capture the readers' attention and add more credibility, a term has emerged to describe the occurrence of an image or text being taken out of its original context or when an image or text is manipulated so as to create misunderstandings and misrepresentations: "multimodal misinformation". One common form of multimodal misinformation is Out-of-context (OOC) which refers to separating authentic images and texts from their original context to change (or eliminate) the intended meaning, which can deceive the audience by making the information appear more credible. Despite its prevalence, the use of out-of-context images and texts remains one of the easiest and most effective ways to create realistically appearing misinformation [4], [5].

Unlike the deepfake generation, the generation of OOC does not require extensive technical knowledge or experience. Moreover, this type of manipulation is harder to detect since neither the image nor the text is distorted. Furthermore, misleading content can be produced in numerous ways [6], [7]. Existing deepfake detection methods are not helpful since the out-of-context relationship between the image and text is at the semantic level rather than at the signal level [8]. As a result, verifying such content relies heavily on manual fact-checking, which can be costly, time-consuming, and error-prone [9]. Therefore, there is a pressing need to develop machine learning-based algorithms as an alternative solution for automatically detecting OOC.

Several recent studies on detecting inconsistencies between images and texts have demonstrated promising performance on benchmark datasets. Existing methods have taken various approaches, such as relying on entities, context, or social media responses/reactions to posts, to identify such misinformation [10]. Notwithstanding the advancements made in this field, OOCD still has limitations, including a lack of relevance to real-world situations due to random mismatches in the training and testing datasets, the need for interpretable explanations to verify accuracy and understand the reasoning, the need for scalability to handle the increasing volume of misinformation in real-time, the dependence on specific languages (e.g., English) limiting applicability to other languages, the scarcity of labeled multimodal datasets, and the challenge of detecting and adapting to adversarial attacks based on misinformation. These limitations highlight the need for research and development of more robust and comprehensive methods for OOCD [11], [12].

In the work reported here, we addressed the challenge of OOCD. **First**, to address the data limitations, we devised a novel training and evaluation method that uses synthetic data generation to enhance the accuracy of OOCD models. Generating synthetic data, which does not require human labeling effort, expands the diversity and complexity of the training data, enabling the detector to better identify and classify instances of information that deviate from their surrounding context. Furthermore, we can address the problem of explainability by generating a synthetic multimodal dataset that aids in understanding the reasoning behind the detection. **Second**, we created a novel dataset for OOCD tasks. This dataset should be a valuable resource for future research and benchmarking. **Finally**, we developed a detector that leverages state-of-the-art machine learning algorithms and techniques to accurately identify OOC multimodal information. This work offers several benefits to the image-text OOCD field by improving the reliability and credibility of image-text data analysis.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Related Work

A. Available Datasets

Gathering large-scale labeled multimodal misinformation datasets is difficult due to their scarcity and the substantial manual effort required. Some examples of datasets that have been created for detecting multimodal misinformation include the MultimodAI Information Manipulation (MAIM) dataset [13] and the Multimodal Entity Image Repurposing (MEIR) dataset [14]. The MAIM dataset includes over 239K image-caption pairs with randomly mismatched falsified media. The MEIR dataset introduces swaps over named entities for people, organizations, and locations. It is based on the assumption that, for each image caption (package), there is an unmanipulated related package (geographically near and semantically similar) in the reference set. This enables the integrity of a query package to be verified by retrieving a related package and comparing the two.

Another such dataset is the TamperedNews dataset [15] in which named entities specific to people, locations, and events are randomly swapped with other named entities within the article body. These text manipulations can lead to strong linguistic biases. Zellers *et al.* [16] replaced image captions with synthetic ones that matched the images, which remained relevant to the article's content. The usefulness of image analysis on this task is rather limited, whereas analyzing the captions and the article body results in the best detection performance. The recently developed COSMOS dataset [17] contains 200K images and 450K captions, where each image is associated with two captions from two sources. The randomly chosen text is used to generate mismatched image-caption pairs. Most of the data in COSMOS is not labeled, although a small subset has been manually annotated as in- or out-of-context.

Another recently developed dataset is the Visual News dataset [18], which contains more than one million publicly available news images paired with both captions and news article text collected on a diverse set of topics from news sources in English (The Guardian, the BBC, USA TODAY, and the Washington Post). It was leveraged to create the Visual News Captioner, which generates captions by attending to individual word tokens, named entities in the input news article text, and localized visual features. It was used to create the NewsCLIPpings [12] dataset, a large-scale dataset designed for automatic retrieval of image-text pairs [18]. It contains 988K image-caption pairs for news media mismatch detection. Several strategies can be used to automatically retrieve suitable images for the given captions, including caption-image similarity, caption-caption similarity, person match, and scene match, to identify cases in which there are inconsistent entities or inconsistent semantic context. Finally, the Twitter-COMMs [19] dataset consists of multimodal tweets covering such topics as COVID-19, climate change, and military vehicles. It was created using the Twitter API v2 [1]. Hard negatives are created by retrieving the image of the sample tweet with the most textual similarity with a given caption (using the "Semantics / CLIP Text-Text split" method of Luo *et al.* [12]), and random negatives are created by selecting an image for a given caption at random to resolve the class imbalance of data collected from Twitter.

Several studies have focused on human-made fake news detection, resulting in the development of such datasets as FakeNewsNet [20] and Fakeddit [21]. Another study focused on creating a videos and captions dataset, i.e., a video-based dataset [2], that contains 4,000 real-world Facebook news posts with captions randomly taken from other posts to generate mismatched samples. It was developed for detecting semantic inconsistency between videos and captions in social media posts. The datasets relevant to the work reported here are summarized in Table I. The key differences, such as the image-text mismatch procedure used in each dataset, are highlighted.

B. Image Captioning

Generating a high-quality description of an image using natural language is the goal of image captioning. This is a challenging task for both humans and machines [22], [23]. However, recent advancements in large-scale vision-language models (VLMs) and large-language models (LLMs) have made it possible for models to generate captions that are increasingly difficult to distinguish from those generated by humans. Some models, such as OFA [24], Flamingo [25], and BLIP-2 [26], have demonstrated impressive performance in image captioning tasks. In certain evaluations, several approaches have been rated as good as or better than human users.

Researchers have explored various approaches, such as "Plug-and-play VQA" (visual question answering) [27], and "PromptCap" (Prompt-guided image Captioning) [28], to generating natural language prompts in conjunction with a generative pre-trained transformer (GPT) to achieve state-of-the-art (SOTA) performance in visual question-answering. Mokady *et al.* [29] utilized continuous embedding as a prompt for the GPT-2 model [30] to achieve strong single-viewpoint image captioning performance. Zeng *et al.* [31] used a CLIP-based model [32] to extract key tags from an image and then used the GPT-3 model [33] with a specialized prompt to generate an appropriate image caption.

C. Text-to-Image Generation

Deep generative models have achieved tremendous success in text-to-image (T2I) generation tasks and have recently attracted intensive attention. Existing T2I generation models can be categorized as GAN-based, VAE-based, and diffusion-based. Although GAN-based [34] and VAE-based models [35] synthesize images with promising quality and diversity, they still cannot match user descriptions well.

Diffusion-based models [36] have demonstrated unprecedentedly high-quality and controllable image generation and have been broadly applied to T2I generation tasks. They were first introduced into T2I generation with classifier-free guidance. The introduction of the DALL-E deep learning

[1] https://developer.twitter.com/en/docs/twitter-api/gettingstarted/about-twitter-api

TABLE I
SUMMARY OF RELEVANT DATASETS. SIZE IS THE TOTAL NUMBER OF UNIQUE SAMPLES ACROSS ALL SPLITS.

Dataset	Year	Data	Source	Mismatch	Size
MAIM [13]	2017	Caption, Image	Flickr	Random	239k
MEIR [14]	2018	Caption, Image, GPS	Flickr	Text entity manipulation	57k
TamperedNews [15]	2020	Article, Image	BreakingNews	Text entity manipulation	776k
VisualNews [18]	2020	Article, Caption, Image, Metadata	4 news agencies: Guardian, BBC, USA Today, Washington Post	/	1.25M
COSMOS [17]	2021	Caption, Image	News Outlets	Two sources (3k labeled)	453k
NewsCLIPpings [12]	2021	Caption, Image	VisualNews	Automatic retrieval	988k
Twitter-COMMs [19]	2022	Caption, Image	Tweets: COVID-19, Climate Change, Military Vehicles	Random & Automatic retrieval	884k

model [37] made it possible to perform zero-shot T2I generation with arbitrary text input. It quickly inspired several follow-up models, including Glide [38], a text-guided diffusion model supporting both image generation and editing. It implements the text generation feature in the transformer blocks used in the denoising process. The CogView model [39] is a transformer with a vector quantized-variational autoencoder [40] tokenizer similar to that in DALL-E [37], and it supports Chinese text input. The Imagen model [41] has a T2I structure that does not use latent images and instead directly diffuses pixels using a pyramid structure. The Stable Diffusion model [42] is a large-scale implementation of the latent diffusion model (LDM) [43] that achieves T2I generation. These models use pre-trained large-scale text encoders to provide more controllable guidance signals. They either introduce super-resolution models or encode images into the low-dimensional latent space to achieve a trade-off between efficiency and high resolution.

D. Multimodal Misinformation Detection

Studies on MMD have focused on OOC image-text pairs or cross-modal named entity inconsistencies. Aneja *et al.* [17] proposed using a self-supervised approach to establish whether two captions accompanying an image are consistent. However, the unlabeled dataset they collected does not support veracity detection training and evaluation. Furthermore, the assumptions made about the availability of such a dataset may affect the model's feasibility in a generic test setting, and its disproportionate reliance on text mode may lead to an increased decision-level bias.

Luo *et al.* [12] leveraged the expressiveness of a large pre-trained contrastive model, CLIP [32], to classify mismatches on the basis of retrieval; in addition, they evaluated two VLMs, CLIP and VisualBERT [44], on the proposed dataset and achieved classification accuracies of 60.23% and 54.81%, respectively. They showed that both machine and human detection are limited, indicating that the task is challenging. Abdelnabi *et al.* [10] collected evidence for both visual and textual components for use in performing cycle consistency checks. However, relying on such external information for validity checks makes the entire approach very expensive, complex, memory intensive, and challenging to deploy in a generic test setting. Furthermore, in this use case setting, the primary objective was to verify the content of a news item quickly. Therefore, sufficient reliable information related to the news topic may not be available yet on the web.

Huang *et al.* [11] used two VLMs, CLIP and VinVL [45], for the multimedia inconsistency detection task. They encoded each image and its corresponding caption separately and then compared their embeddings; dissimilar embeddings suggested an out-of-context text-image pair. They evaluated their method on several large-scale datasets. However, their method does not address the potential biases in the VLMs used, which may have affected the accuracy of the OOCD. In addition, their method relies on the availability of textual information associated with the image, which may not always be available. The self-supervised distilled learner (SSDL) method devised by Mu *et al.* [46] uses two strategies to estimate the cross-modal consistency status in terms of a binary class label (Pristine or Falsified). However, the effectiveness of the SSDL approach relies heavily on the quality and accuracy of the annotated data used during the training phase. Errors and biases in the annotations can severely degrade the model's ability to learn and generalize effectively. Moreover, the SSDL approach fails to address potential biases or provide strategies for mitigating them. Furthermore, the training and fine-tuning of the self-supervised models are computationally intensive and take a considerable amount of time.

The method devised by Zhang *et al.* [47] detects OOC multimodal misinformation by using an interpretable neural-symbolic model. However, it does not address other types of misinformation and lacks empirical evaluation and comparison to other approaches. The model shows theoretical promise, but without thorough experimental validation, its effectiveness and superiority over other methods are uncertain. It is also subject to challenges like scalability, efficiency, and generalization across domains and languages, raising concerns about its practical viability. Table II summarizes existing methods for detecting multimodal misinformation.

III. PROPOSED APPROACH

Our proposed approach has distinctive features. Unlike existing approaches, it achieves remarkable results without relying on external sources or additional data to enhance and augment the learning process of the classifier. Furthermore, it eliminates the need to conduct Internet-based searches to validate the accuracy of the classifier's decision or to establish the connection between an image and the corresponding caption. These features will allow us to advance our approach further as the underlying models evolve.

A. Problem Statement

Detecting OOC is a challenging problem that not only stems from multimodal semantic mismatches but also from the lack

TABLE II
SUMMARY OF EXISTING MULTIMODAL MISINFORMATION DETECTION METHODS.

Work	Year	Models	Dataset	Method Description
Aneja *et al.* [17]	2021	spaCy NER, Mask-RCNN, Object Encoder, Universal Sentence Encoder, SBERT	COSMOS	Uses a self-supervised approach to establish whether two captions accompanying an image are consistent.
Luo *et al.* [12]	2021	CLIP, VisualBERT	VisualNews	Classifies mismatches on the basis of retrieval (CLIP and VisualBERT were evaluated on proposed dataset).
Abdelnabi *et al.* [10]	2022	ResNet152, sentence transformer, BERT + LSTM ,spaCy NER, CLIP	NewsCLIPpings	Collects evidence for both visual and textual components for use in cycle consistency checks.
Huang *et al.* [11]	2022	CLIP, VinVL	NewsCLIPpings & private datasets (DARPA)	Encodes an image and its corresponding caption separately and then compares their embeddings for multimedia inconsistency detection task.
Mu *et al.* [46]	2023	CLIP, SimCLRV2 , ResNet50	NewsCLIPpings	Uses two networks, Student Classifier (semi-supervised learning) and Teacher Network (self-supervised learning), for SSDL.
Zhang *et al.* [47]	2023	MLPs and SAFE, VisualBERT, CLIP, VINVL, FaceNet+BERT	NewsCLIPpings	Uses a hybrid model that combines a neural network and symbolic reasoning and captures context and semantics of content by using textual and visual information.

of mature problem definitions and datasets for evaluation. Deep neural networks require large amounts of data, and manual annotation is labor-intensive. Furthermore, the limited evaluation of real-world misinformation and the absence of comparisons with synthetic multimodal misinformation make it difficult to assess progress in the field. One of the main reasons for the spread of misinformation is the increasing use of generative AI to create multimodal misinformation. However, few methods exist to detect misinformation because annotated datasets are scarce. Our proposed approach is aimed at detecting OOC by using synthetic misinformation generated from the original dataset. This maximizes the utilization of limited annotated data by generating additional samples. We estimate the level of consistency between the multimodal data using the generated synthetic data, enabling us to assess the overall consistency within the original data. This estimation is performed as a binary classification (match or mismatch).

B. Data Preparation

Our approach involves the creation of synthetic data versions using the original image, caption, and label from the original dataset[12]. This process consists of two key steps. Firstly, Starting with the original image I, we utilize an image captioning model to generate a corresponding caption C'. Subsequently, employing a text-to-image T2I model, we generate an image I' based on the basis of the original caption C. As a result, our augmented dataset comprises the original images accompanied by their corresponding captions and the generated images and generated captions. The dataset preparation process is outlined in Figure 1.

1) Caption Generation: The process of generating a caption for the original image is performed using a pre-trained image captioning model to generate multiple captions on the basis of a given image. The captions produced are condensed into a single (summary) caption to create the output. Formally, given an image $I \in \mathbb{N}^{H \times W \times C}$, we aim to make a sequence of m tokens $x_1 ... x_m$ describing the image in some vocabulary V and produce a probability distribution $P(x_m \in V | I, x_1 ... x_{m-1})$, which is the probability of the next token in the sentence, given both the image and all the previous tokens.

We used the BLIP-2 [26] model and fine-tuned it specifically for the image captioning task. BLIP-2 is a vision-language pre-training framework designed to utilize large amounts of noisy

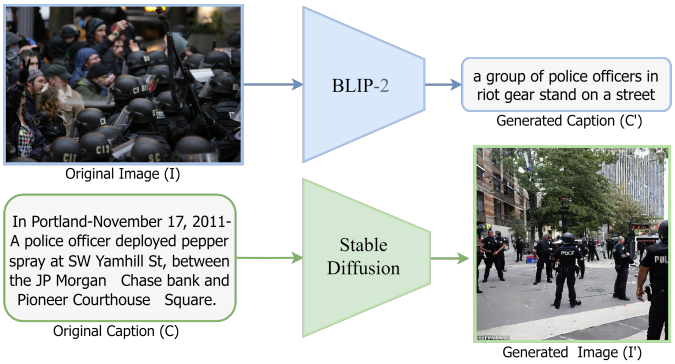


Fig. 1. Representation of steps involved in dataset preparation.

web data for effective pre-training. The model uses synthetic image-caption pairs from a seed captioning model and a filter to eliminate low-quality synthetic captions, resulting in a large dataset. BLIP-2 has shown remarkable transfer performance in many vision-language tasks, and it excels in zero-shot and fine-tuning scenarios for image captioning because the BLIP-2 model is trained on massive image and language datasets (the MS-COCO datasets [22]).

2) Image Generation: Our work uses an LDM [43] loaded with Stable Diffusion v1.4 weights to generate synthetic images. The Stable Diffusion model was trained on a subset of the LAION5B dataset [48] and generated (512×512) pixel images, which are then resized to the native image size of each dataset. To provide conditioning from text prompts, a frozen CLIP ViT-L/14 text encoder is used in conjunction with Stable Diffusion [41], [43]. Although LDMs can generate images from several different conditionings (image, text, semantic map), we only generated images using text prompts (captions). This enables zero-shot learning; it requires only a text prompt describing the contents of the image without any real images.

To generate images through text prompts, we require a prompt (caption) that describes the desired image's contents to guide the diffusion process. The prompt is first projected into an intermediate representation via CLIP's text encoder and then mapped to the intermediate layers of the LDM's denoising UNet using cross-attention [43]. This guidance helps the LDM diffuse a latent representation of an image starting

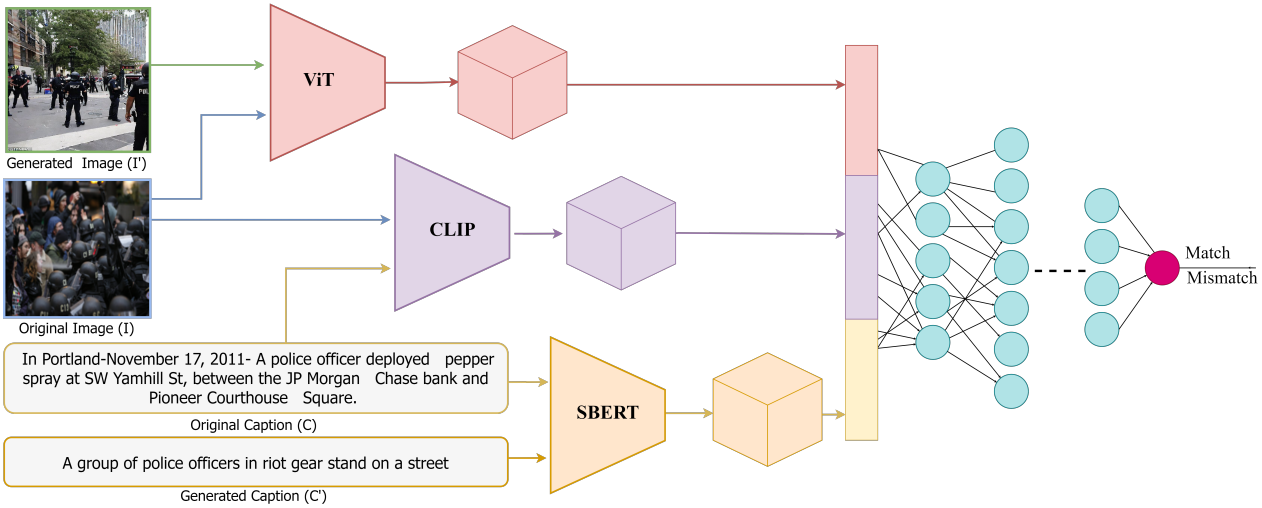


Fig. 2. Overview of our proposed approach.

from Gaussian noise. The resulting latent representation is then decoded back into the pixel domain to produce the final image.

Three primary hyperparameters control the generation process. One is the number of steps the denoising diffusion implicit model (DDIM) takes in the denoising process. More steps generally result in more realistic and coherent images, while fewer steps result in more disjointed and surreal images [36]. We generated synthetic images using 500 DDIM steps. The second hyperparameter is the unconditional guidance scale (UGC), which controls the scale between the precision of the generated image matching, the provided prompt, and generation diversity. This is accomplished by scaling between the jointly trained conditional and unconditional diffusion models [49]. A lower UGC value means less guidance and more diversity, while a higher value means more guidance and less diversity. The third hyperparameter is the Seed, which is the starting point for diffusion, generating the initial Gaussian.

C. Out-Of-Context Detection Model

Our proposed multimodal model, which effectively detects OOC by leveraging synthetic misinformation, is illustrated in Figure 2. Deep features are extracted from our dataset (I, C, I', C') and used by the classifier to more effectively train the model and to enhance its ability to discern patterns of consistency between pristine and falsified examples. These features are extracted from the original image I and caption C by using a pre-trained multimodal CLIP model [32]. Additionally, features are extracted from the original caption C and the generated caption C' by using Sentence-BERT [50]. A vision transformer (ViT) [51] is used to extract features from the original image I and the generated image I'. All of these features are concatenated and input into the classifier.

The output of the classification process yields one of two scenarios depending on the contextual relationship between the original image I and its corresponding caption C. In the first scenario, if the original image and caption (I, C) are within the same context, the generated image I' derived from the original caption will inherently share the context of the original caption C and align with the original image I. Similarly, the generated caption C' will be in the same context as the original image I, which is consistent with the original caption C. As a result, there is a high similarity between the original image and the generated image (I, I'), and between the original caption and the generated caption (C, C').

If the original image and caption (I, C) are out of context, the original image and the generated image $I, I')$ and also the original caption and the generated caption (C, C') will have low similarity scores, as follows:

$$((I, C, y) = \begin{cases} If \quad y = Match \\ High_{Similarity}((I, C), (C, C^{'}), (I, I^{'})) \\ \\ Otherwise \quad y = Mismatch \\ Low_{Similarity}((I, C), (C, C^{'}), (I, I^{'})) \end{cases}$$

where y denotes the label of the pair of the original image and caption (match or mismatch). $High_{Similarity}$ is defined as the highest similarity score among all pairs, indicating a high correlation between the image and caption. Conversely, $Low_{Similarity}$ is defined as the minimum similarity score among all pairs, indicating a low correlation between the image and caption.

IV. Experimental and Results

In this section, we describe our experimental setting and then present and analyze the results.

A. Dataset

We used the NewsCLIPpings [12] dataset, which consists of pristine and falsified OOC images. It was constructed on the basis of the VisualNews [18] corpus, which includes news articles from four major news outlets: the BBC, The Guardian, the Washington Post, and USA Today. The NewsCLIPpings dataset was constructed using various methods for automatically retrieving convincing images for a given caption, capturing cases with inconsistent entities or semantic context (such as text-text similarity or image-image similarity), etc. We use the Merged/Balanced dataset, which contains data representative of all retrieval methods and contains 85K balanced pristine and falsified examples, 71,072 training examples, 7,024 validation examples, and 7,264 test examples, divided into a 10:1:1 ratio. To initiate our generation-assisted detection approach, we used the original image-caption pairs as prompts to generate synthetic image-caption pairs, as illustrated in Figure 1.

B. Experimental Settings

We conducted two evaluation experiments to assess the merits of our proposed approach comprehensively. To establish the effectiveness and efficacy of our approach, we compared the results with those of previous studies. This comparative analysis serves as a robust validation of our proposed approach.

1) Experiment using Cosine Similarity: In the first experiment, we used CLIP (ViT-B/32) [32], a powerful model, to extract embedded features from the original image-caption pairs. This process yielded a 512-dimensional vector, represented as $(I_{embdCLIP}, C_{embdCLIP})$, for each pair. To quantify the similarity between the original image and caption (I, C), we used the cosine similarity metric to calculate $CLIP_{similarity}$. To evaluate the similarity between the original caption and the generated caption $(C, C^{'})$, we used Sentence-BERT [50] (all-mpnet-base-v2) for embedding both captions and obtained a 768-dimensional vector $(C_{embdSBERT}, C^{'}_{embdSBERT})$ for each pair. We then measured the similarity between their embeddings by using the cosine similarity metric, resulting in $SBERT_{similarity}$.

Furthermore, we used a vision transformer (ViT_L_16) [51] to extract the embeddings from both the original image and the generated image $(I, I^{'})$; this embedding yielded a 1,024 item vector, represented as $(I_{embdViT}, I^{'}_{embdViT})$, for each image. Using cosine similarity, we compared the embeddings of $(I, I^{'})$ to obtain $ViT_{similarity}$, thereby quantifying their similarity. Next, the three similarity values were fed into machine learning classifiers to distinguish patterns between the (I, C) pairs with coherent context and those with incoherent context. Table III presents the experimental results in comparison with those for four other classifiers.

2) Experiment using Feature Extraction: In the second experiment, we used feature extraction techniques to analyze the effectiveness of different feature map combinations. The feature maps were obtained from three models (see Figure 2), namely CLIP, Sentence-BERT, and Vision Transformer, as represented by $(I_{embdCLIP}, C_{embdCLIP})$, $(C_{embdSBERT}, C^{'}_{embdSBERT})$, and $(I_{embdViT}, I^{'}_{embdViT})$, respectively.

To analyze the feature maps, we followed a two-step approach. We first flattened each pair of feature maps and separately fed them into machine learning classifiers and a transformer designed for tabular data analysis. Next, we consolidated all the feature maps and fed them into both the machine learning classifiers and the transformer. To optimize the classification process, we utilized dimensionality reduction techniques on the combined set of concatenated features. This enabled us to reduce the complexity of the data while retaining important information. To thoroughly investigate the potential of different feature map combinations, we systematically evaluated multiple combinations. The results of these evaluations are summarized in Table IV.

TABLE III
DETECTION PERFORMANCE (IN %) BASED ON SIMILARITY SCORES

Model	Similarity					
	CLIP+SBERT		CLIP + ViT		CLIP+SBERT+ViT	
	ACC	AUC	ACC	AUC	ACC	AUC
SVM	66	72	65	72	67	73
Random Forest	62	71	62	67	67	72
XGBoost	66	72	64	72	67	73
MLP	67	73	66	73	**68**	**74**
Transformer	66	72	65	72	67	73

TABLE IV
DETECTION ACCURACY (IN %) BASED ON FEATURE MAPS

Model	Features		
	CLIP	CLIP+SBERT	CLIP+SBERT+ViT
SVM	50	50	54
Random Forest	49	50	52
XGBoost	52	53	55
MLP	53	55	56
Transformer	**56**	**60**	**62**

C. Discussion

OOCD presents substantial complexities due to the unaltered nature of the image and its accompanying caption. These two elements exist in separate contexts, often featuring the same names of individuals and places but with different temporal attributes, such as dates. This disparity between the image and its caption poses a considerable challenge in detecting the asymmetry, necessitating further efforts to develop robust models in the future. Despite these complexities, the results obtained with our proposed approach are highly encouraging, motivating the continued advancement of this approach.

Our proposed approach resulted in the highest accuracy among the approaches compared, surpassing other research in this field, as shown in Table V. This table compares our best experimental accuracy with those of previous studies using the same dataset without leveraging additional information to enhance the classification process. Further development of this approach is expected to yield even better results as future research introduces more powerful models capable of

accurately measuring the similarity between images or captions and providing improved image captioning.

The efficiency of our approach makes it potentially suitable for real-time operation, delivering prompt responses. This efficiency arises from the parallel implementation of the models used. Furthermore, alongside the OOCD model, we built a dataset of generated images from real captions that researchers in this and related fields can utilize as valuable resources in their work.

TABLE V
COMPARISON RESULTS OF OUR PROPOSED APPROACH WITH OTHER APPROACHES USING THE SAME DATASET.

Paper	**Year**	**Accuracy** (%)
Luo *et al.* [12]	2021	65.9
Huang *et al.* [11]	2022	65.2
Abdelnabi *et al.* [10]	2022	66.1
Zhang *et al.* [47]	2023	62.8
Ours	2023	**68.0**

V. CONCLUSION

Our proposed approach to Out-Of-Context detection uses synthetic data generation. The dataset and detector we developed enable the accurate identification of misinformation across various modalities. The results of extensive experimentation and evaluation validate the effectiveness of the proposed approach, with which a promising classification accuracy rate of 68% was achieved for misinformation detection. The dataset and detector should prove valuable for future research and facilitate the development of robust misinformation detection systems.

VI. ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grants JP18H04120, JP20K23355, JP21H04907, and JP21K18023, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, Japan.

REFERENCES

[1] S. Ali, "Combatting against covid-19 & misinformation: A systematic review," *Human Arenas*, vol. 5, no. 2, pp. 337–352, 2022.

[2] S. McCrae, K. Wang, and A. Zakhor, "Multi-modal semantic inconsistency detection in social media news posts," in *MMM*, Springer, 2022, pp. 331–343.

[3] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes.," in *CVPR-W*, vol. 1, 2019, p. 38.

[4] S. Aneja and M. Nießner, "Generalized zero and few-shot transfer for facial forgery detection," *arXiv preprint arXiv:2006.11863*, 2020.

[5] K. Wang and A. Zakhor, "Representation learning in video and text-a social media misinformation perspective," Ph.D. dissertation, MA thesis. EECS Department, University of California, Berkeley, 2022.

[6] A. Jaiswal, Y. Wu, W. AbdAlmageed, I. Masi, and P. Natarajan, "Aird: Adversarial learning framework for image repurposing detection," in *CVPR*, 2019, pp. 11 330–11 339.

[7] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[8] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *MIPR*, IEEE, 2018, pp. 384–389.

[9] R. Tan, B. Plummer, and K. Saenko, "Detecting cross-modal inconsistency to defend against neural fake news," in *EMNLP*, 2020, pp. 2081–2106.

[10] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in *CVPR*, 2022, pp. 14 940–14 949.

[11] M. Huang, S. Jia, M.-C. Chang, and S. Lyu, "Text-image de-contextualization detection using vision-language models," in *ICASSP*, IEEE, 2022, pp. 8967–8971.

[12] G. Luo, T. Darrell, and A. Rohrbach, "Newsclippings: Automatic generation of out-of-context multimodal media," in *EMNLP*, 2021, pp. 6801–6817.

[13] A. Jaiswal, E. Sabir, W. AbdAlmageed, and P. Natarajan, "Multimedia semantic integrity assessment using joint embedding of images and text," in *ACM Multimedia*, 2017, pp. 1465–1471.

[14] E. Sabir, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Deep multimodal image-repurposing detection," in *ACM Multimedia*, 2018, pp. 1337–1345.

[15] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, and R. Ewerth, "Multimodal analytics for real-world news using measures of cross-modal entity consistency," in *ICML*, 2020, pp. 16–25.

[16] R. Zellers, A. Holtzman, H. Rashkin, *et al.*, "Defending against neural fake news," *NeurIPS*, vol. 32, 2019.

[17] S. Aneja, C. Bregler, and M. Nießner, "COSMOS: Catching out-of-context misinformation with self-supervised learning," in *AAAI*, 2023.

[18] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual news: Benchmark and challenges in news image captioning," in *EMNLP*, 2021, pp. 6761–6771.

[19] G. Biamby, G. Luo, T. Darrell, and A. Rohrbach, "Twitter-comms: Detecting climate, covid, and military multimodal misinformation," in *NAACL*, 2022, pp. 1530–1549.

[20] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[21] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proceedings of the Twelfth Language*

Resources and Evaluation Conference, 2020, pp. 6149–6157.

[22] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, Springer, 2014, pp. 740–755.

[23] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018, pp. 2556–2565.

[24] P. Wang, A. Yang, R. Men, *et al.*, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *ICML*, PMLR, 2022, pp. 23 318–23 340.

[25] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: A visual language model for few-shot learning," *NeurIPS*, vol. 35, pp. 23 716–23 736, 2022.

[26] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[27] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training," *arXiv preprint arXiv:2210.08773*, 2022.

[28] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided task-aware image captioning," *arXiv preprint arXiv:2211.09699*, 2022.

[29] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[31] A. Zeng, A. Wong, S. Welker, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[32] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, PMLR, 2021, pp. 8748–8763.

[33] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[34] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[35] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[37] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation," in *ICML*, PMLR, 2021, pp. 8821–8831.

[38] A. Q. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*, PMLR, 2022, pp. 16 784–16 804.

[39] M. Ding, Z. Yang, W. Hong, *et al.*, "Cogview: Mastering text-to-image generation via transformers," *NeurIPS*, vol. 34, pp. 19 822–19 835, 2021.

[40] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *NeurIPS*, vol. 30, 2017.

[41] C. Saharia, W. Chan, S. Saxena, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, vol. 35, pp. 36 479–36 494, 2022.

[42] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[44] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[45] P. Zhang, X. Li, X. Hu, *et al.*, "Vinvl: Revisiting visual representations in vision-language models," in *CVPR*, 2021, pp. 5579–5588.

[46] M. Mu, S. Das Bhattacharjee, and J. Yuan, "Self-supervised distilled learning for multi-modal misinformation identification," in *WACV*, 2023, pp. 2819–2828.

[47] Y. Zhang, L. Trinh, D. Cao, Z. Cui, and Y. Liu, "Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model," *arXiv preprint arXiv:2304.07633*, 2023.

[48] M. Cherti, R. Beaumont, R. Wightman, *et al.*, "Reproducible scaling laws for contrastive language-image learning," in *CVPR*, 2023, pp. 2818–2829.

[49] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[50] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," in *EMNLP*, 2019.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Gait Recognition Scheme Focusing on Operating Characteristics at Feature Points Detected by OpenPose

Chinatsu Tanaka*[1] and Minoru Kuribayashi† and Nobuo Funabiki*
* Graduate School of Natural Science and Technology, Okayama University, Japan
E-mail: [1] pvhm32s3@s.okayama-u.ac.jp
† Center for Data-driven Science and Artificial Intelligence, Tohoku University, Japan

***Abstract*—The gait recognition system uses the characteristics of a person's gait to authenticate the individual. Generally, it applies a walking model to extract features from a captured video. In the conventional works, two types of methods have been studied: model-based methods, which extract features by fitting a person model on an image, and appearance-based methods, which extract features based on time-series silhouette images. In this study, we introduce a prediction operation to extract features in the model-based method. The proposed method consists of two machine learning models. One is a *keypoint prediction model* that uses OpenPose to estimate the skeleton and predict skeletal movements of a typical person. The other is an *identification model* that identifies individuals based on the difference between the keypoints and predictions of the target person. Our computer simulations show that the proposed method performs better in terms of identification accuracy than the method that directly uses skeletal coordinate data.**

I. Introduction

Since the stride length, arm swing, and posture during walking vary from person to person, gait is considered one of the biometric (biometa risk) methods for identifying individuals. In addition, in 2017, gait was taken up as a personal identification code in the enforcement order for laws and regulations concerning the protection of individuals under the revised Personal Information Protection Law in Japan, and is recognized as one of the biometric authentication methods. The method of authenticating individuals based on the individuality of their gait is called "gait authentication" [1]. Unlike biometric conditions such as fingerprints and iris, gait can be authenticated even when the subject is far away from the camera, making it the only modality that can authenticate individuals from a distance. For example, it is expected to be applied to forensic science using the images from security cameras, because it can be used even when face recognition is not possible due to low resolution.

The gait recognition is based on extracting feature expressions from video frames of the subject and matching them with stored data or using machine learning models to identify the subject. The approaches can be classified into two methods: model-based and appearance-based [2]. In model-based methods, gait parameters such as stride length and time series of joint nuclei are extracted as features by fitting a human body model to the target person in the input image. Appearance-based methods extract features from input video or silhouette image sequences by analyzing the components of variation over time and space. When using deep learning techniques to discriminate based on these features, high discrimination accuracy can be achieved if the features are computed in a feature space or metric space that is more suitable for discrimination.

Many model-based methods estimate skeleton by analyzing an video frames to extract features of them. One of popular skeletal estimation tool is OpenPose [3], [4]. Some studies perform repair and interpolation of skeletal estimation, while others discriminate gait features by latent intra-individual variation factors in the observed direction and gait recognition. In these studies, the feature points detected by the skeletal estimation are represented by XY coordinates and are directly used to the identification of individuals. However, due to characteristics such as camera angle and distance from the object, the amount of change in coordinates varies depending on the position of the person in the image, affecting the accuracy of identification.

In this study, we propose an identification method using the difference between the skeletal and predicted coordinates of the target person. The environment in which the proposed method is used is assumed to be an entry system in which a fixed camera is used to capture the entrance. Therefore, observers are limited to walking in one direction. Thus, we assume that a surveillance camera is capturing a pedestrian in front of it, the proposed gait recognition scheme attempts to identify the individual only from the video. In each frame of the captured video, the distance that the person's skeletal coordinates have moved horizontally and vertically is calculated. For feature points, preprocessing is performed such that the change in their coordinates is normalized. The proposed method composed of two machine-learning models: Keypoint Prediction Model (KPM) and Identification Model (IM). KPM is constructed by the long short term memory (LSTM) [5] trained to output the walking pattern of typical human. At first, we detect 25 skeleton keypoints by using OpenPose, and normalize them. Then, the future coordinates at each key point are then predicted, and the horizontal and vertical

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

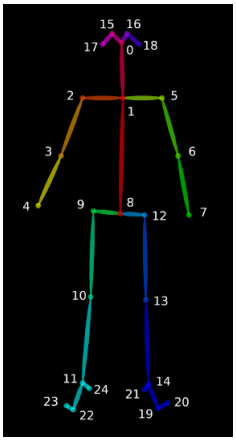


Fig. 1: 25 skeletal keypoints detected by OpenPose.

prediction errors are calculated as the feature values to be input to IM. Finally, individuals are identified from the IM output. Compared to methods that directly use the coordinates of skeletal keypoints, our computer simulations confirm that prediction errors are effective in obtaining feature components suitable for individual identification.

II. Related Studies

A. OpenPose

OpenPose [3], [4] is a real-time skeleton estimation algorithm developed at the Center for Technology Transfer and Enterprise Creation (CTTEC) at Carnegie Mellon University. It is capable of detecting a total of 130 key points in an image, including the human body, hands, and face. The execution time is constant regardless of the number of people in the image. It is a flexible, easy-to-configure, multi-threaded module that can save and load images in a variety of formats, including JSON, XML, PNG, and JPG. It uses Caffe, but the code can be ported to other frameworks such as Tensorflow and PyTorch. The keypoints of human the skeleton are the points estimated as joint points in the 25 regions shown in Fig. 1. In this study, we use these 25 keypoints for the Gait recognition.

B. LSTM

LSTM [5] is a machine learning technique that improves the layer that considers time series of RNN (Recurrent Neural Network) and eliminates the gradient loss problem. It solves the gradient loss problem by using gates to selectively retain relevant information and forget irrelevant information. Due to its low sensitivity to time differences, LSTM is better suited for analyzing time-series data than simple RNNs.

A typical LSTM block has memory cells, input gates, output gates, and forget gates, in addition to the hidden states in a conventional RNN. The memory cells store values over an arbitrary time interval. Weights and biases for the input gates control the amount of new values flowing into the cells. The weights and biases for the forget gates control how much value is retained in the cell, and the output gates control how much of the value in the cell is used to compute the activation state of the output of the LSTM block.

LSTM is widely used in natural language processing and has many other applications in speech and acoustic signal processing. We believe that LSTM can be applied to video images to effectively extract hidden features because the information in the frames changes over time.

The central idea of the LSTM architecture is a memory cell that can maintain its state over time and a nonlinear gate unit that controls the flow of information into and out of the cell. Several variants of the LSTM architecture have been investigated since the initial formulation of the LSTM architecture [5]. Due to the simplicity, we use the standard LSTM architecture implemented on the Keras environment in this study.

C. Previous Work

Many of the previous model-based studies on gait recognition have used skeletal information as it is. For example, there have also been studies on the influence of external factors such as camera position on identification accuracy [6], and identification using deep generative models that consider person labels such as age and gender as latent variables in the skeletal information [7]. These studies have improved identification accuracy by adding camera positions and latent variables as additional labels to skeletal information. Fig. 2 shows the flow of identification common to the previous studies. The video is divided into images at the frame rate, and skeletal estimation is performed. Since the video is divided into images, it is necessary to add the person ID, the video ID, and the time on the video as labels to the skeletal information for each image. In addition, the camera position and latent variables of the target person are added as labels. These datasets are identified using a discriminator.

III. Proposed Identification Method

A. Overview

The flow of the identification method proposed in this study is shown in Fig. 4 and Fig. 5. First, each frame of a video file is extracted with a specified frame rate using the FFMPEG tool[1]. Next, the human skeleton represented by 25 keypoints is detected by the OpenPose from each frame. Each keypoint is represented as XY coordinates $(x_i, y_i), (1 \leq i \leq 25)$ on the image. For each person, the coordinate of 25 keypoints at a certain time period is stored as a training data for keypoint prediction model (KPM). The model can predict the changes in 25 keyoints when a typical person walks. The difference between the predictions from LSTM and the keypoint coordinate values obtained directly from the training video by OpenPose (hereafter, the difference value) is calculated. Finally, the discriminator discriminates the video based on the difference values obtained from LSTM and the subject's label.

[1] https://ffmpeg.org/

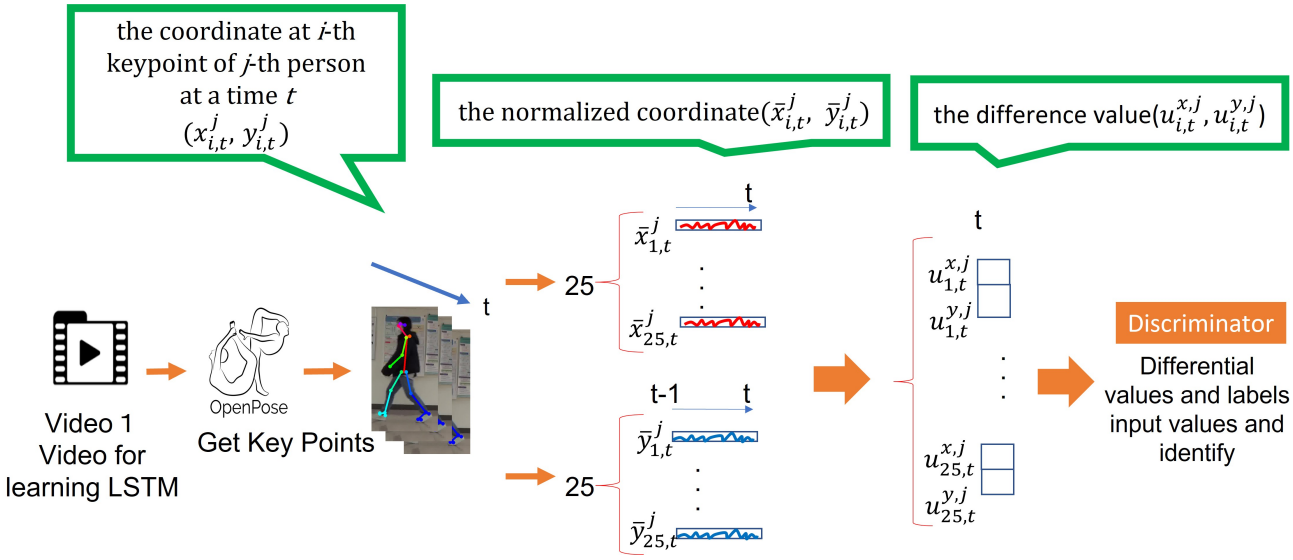


Fig. 2: Identification process using skeletal information.

B. Data Preprocessing

Skeletal coordinates obtained by OpenPose may be biased due to the difference of environmental perturbations. For example, If a person is walking left or right in front of the surveillance camera, the position of the person in the captured video frame will change, and the amount of difference in position will depend on the camera angle and distance from the object. Therefore, the range of coordinate changes will be biased. A normalization is performed on the skeletal coordinates of each person to cancel such a bias.

Suppose that a person is walking down a street on either side and the movements are captured by a surveillance camera at a certain continuous time. Let N be the number of individuals. The coordinate at i-th keypoint of j-th person at a time t is represented by $(x^j_{i,t}, y^j_{i,t})$. Then, the normalized coordinate $(\overline{x}^j_{i,t}, \overline{y}^j_{i,t})$ is calculated as follows:

$$\begin{cases} \overline{x}^j_{i,t} = \dfrac{x^j_{i,t} - x^j_{i,max}}{x^j_{i,max} - x^j_{i,min}} \\ \overline{y}^j_{i,t} = \dfrac{y^j_{i,t} - y^j_{i,max}}{y^j_{i,max} - y^j_{i,min}} \end{cases} \tag{1}$$

where

$$\begin{cases} x^j_{i,max} = \max\limits_t x^j_{i,t} \\ x^j_{i,min} = \min\limits_t x^j_{i,t} \\ y^j_{i,max} = \max\limits_t y^j_{i,t} \\ y^j_{i,min} = \min\limits_t y^j_{i,t} \end{cases} \tag{2}$$

C. Keypoint Prediction Model (KPM)

For predicting the changes of keypoint in captured video frames, we use two LSTM models for horizontal and vertical directions. In this study, we predict the XY coordinates at time t from the previous t_p consecutive coordinates. If these models are trained on a specific person's data set, high accuracy can be achieved in predicting XY coordinates. However, the need to train two LSTM models for each individual is time consuming and results in storage waste. The goal of KPM is to calculate horizontal and vertical prediction errors that are useful for individual identification. Therefore, we train these two LSTM models to output the general variation of pedestrians walking in front of a surveillance camera.

Two LSTM models shown in Fig. 3 are used to predict the time-series skeletal coordinate data. There are some hyper-parameters in LSTM in Keras environment. As 25 keypoints are detected by OpenPose, these LSTMs have a dimension of 25 for the input values. For its simplicity, we fix the number of nodes in hidden layer is 64. For these 25 coordinates at a time t, we use the coordinates at the last consecutive t_p time for prediction. The differences between the obtained 25 coordinates and predicted 25 coordinates are stored in a CSV file along with the person labels as prediction errors. This dataset is used as input values for the identification model.

The predicted coordinate $(\hat{x}^j_{i,t}, \hat{y}^j_{i,t})$ at i-th keypoint of j-th person at a time t is calculated by the following equation.

$$\begin{cases} \hat{x}^j_{i,t} = \text{LSTM}(\tilde{\boldsymbol{x}}^{\boldsymbol{j}}_{\boldsymbol{i,t-1}}) \\ \hat{y}^j_{i,t} = \text{LSTM}(\tilde{\boldsymbol{y}}^{\boldsymbol{j}}_{\boldsymbol{i,t-1}}) \end{cases} \tag{3}$$

where

$$\begin{cases} \tilde{\boldsymbol{x}}^{\boldsymbol{j}}_{\boldsymbol{i,t-1}} = \{\overline{x}^j_{i,t-t'} | 1 \le t' \le t_p\} \\ \tilde{\boldsymbol{y}}^{\boldsymbol{j}}_{\boldsymbol{i,t-1}} = \{\overline{y}^j_{i,t-t'} | 1 \le t' \le t_p\} \end{cases} \tag{4}$$

Then, the prediction errors $\delta^{x,j}_{i,t}$ and $\delta^{y,j}_{i,t}$ for horizontal and vertical directions are respectively obtained as follows:

$$\begin{cases} \delta^{x,j}_{i,t} = \overline{x}^j_{i,t} - \hat{x}^j_{i,t} \\ \delta^{y,j}_{i,t} = \overline{y}^j_{i,t} - \hat{y}^j_{i,t} \end{cases} \tag{5}$$

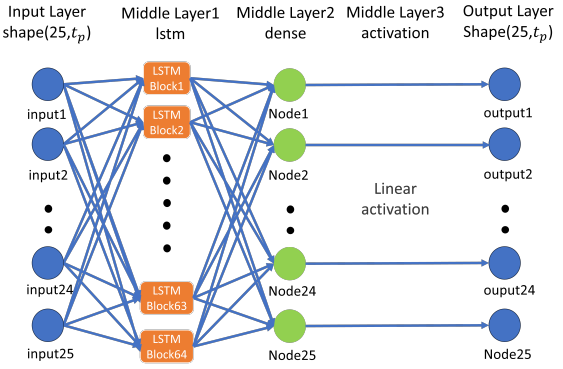


Fig. 3: Parameters in the LSTM model.

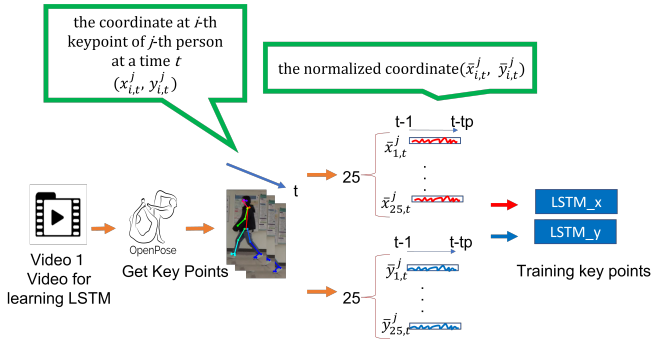


Fig. 4: Operation flow in the keypoint prediction model.

For training these LSTM models, the choice of optimization algorithm and loss function are important to improve the performance. In this study, Adam [8] is used as the optimization algorithm. Adam is a very popular optimization algorithm that combines conventional momentum and RMSprop [9]. Momentum is an algorithm that suppresses oscillations by moving averages, while RMSprop is an algorithm that suppresses oscillations by adjusting the learning rate.

We use the mean squared error (MSE) as the loss function, which is the sum of the squares of the differences between the correct and predicted values. Let $(\overline{x}_i, \overline{y}_i)$ and $(\hat{x}_i, \hat{y}_i)$ be the coordinates of the i-th keypoint for correct and predicted coordinates, respectively. Then, the MSE is calculated by the following equation.

$$MSE = \frac{1}{25}\sum_{i=1}^{25}\sqrt{(\overline{x}_i - \hat{x}_i)^2 + (\overline{y}_i - \hat{y}_i)^2} \tag{6}$$

This function overestimates errors because the larger the error, the easier it is to evaluate the error. On the other hand, it also overreacts to outliers in the dataset, which may result in learning that adapts to outliers (over-fitting).

D. Identification Model (IM)

IM identifies individuals based on the horizontal and vertical prediction errors obtained by KPM. The predictions (δ_x, δ_y) obtained by LSTM are converted to a 1D tensor and input to the discriminator. Without loss of generality, we assume that the video frame is consecutively captured at each time, and the horizontal and vertical directions of prediction error $\delta_i^{x,j}$ and $\delta_i^{y,j}$ are extracted during a time range from 1 to T.

$$\begin{cases} \Delta_i^{x,j} = (\delta_{i,1}^{x,j}, \ldots, \delta_{i,T}^{x,j}) \\ \Delta_i^{y,j} = (\delta_{i,1}^{y,j}, \ldots, \delta_{i,T}^{y,j}) \end{cases} \tag{7}$$

$$\Delta^j = \Delta_1^{x,j} || \Delta_1^{y,j} || \ldots || \Delta_{25}^{x,j} || \Delta_{25}^{y,j} \tag{8}$$

where "||" stands for a concatenation.

The discriminator identifies person labels (individual numbers) based on the difference values of skeletal coordinate data in XY coordinates. For this purpose, a 1D CNN as shown in Fig. 6 is used. In this study, since the identification target is not an image but XY coordinates, and the data is time series, we use 1D convolution and total. Since the identification target is not an image but XY coordinates and the data are time series, we use a combination of 1D convolution and all-connected layers [10]. Adam is used for optimization and categorical cross-entropy is used for the loss function in KPM.

IV. Experiments

In this experiment, we test whether the proposed method improves the accuracy of identification using difference values compared to identification using skeletal information as is. The same model discriminator was used for both methods, and the parameters were fixed. In addition, simulations were performed 10 times with the data randomly divided into 10 pieces. Experimental results are presented in terms of Accuracy and Loss for the Test of each simulation.

A. CASIA Gait

CASIA Gait [11] is a database created by the Chinese Academy of Sciences Institute of Automation (CASIA) for gait certification. There are three datasets in the database: Dataset A, Dataset B (multi-view dataset), and Dataset C (infrared dataset). The subjects in the video walk a predetermined straight course and are captured from various angles. The camera angle is recorded as 0° when horizontal and 90° when vertical to the direction of travel. In Database A, the camera angles are recorded at 0°, 45°, and 90°. In Database B, the camera angles are recorded from 0° horizontally to 180° in increments of 18°. Since the data used in this study is a walk from the right side of the screen to the left side, an angle of 90 degrees was used in the database. The camera used was a USB camera (Famtech 318SC) with a frame rate of 25 fps and an image size of 320×240 pixels. The data is stored in motion JPEG format. Details of the databases are shown in Table I. In this study, Dataset A and Dataset B are used.

B. Experiental Conditions

The CASIA Gait database was used in this simulation as shown in Table I. Two types of videos were used, DatasetA and DatasetB, each with 10 men and 10 women. In this study, LSTM learns up to the past 30 ($t_p = 30$). In IM, data are acquired with a sample time of 2 seconds ($T = 60$). Each video was divided into 2-second segments and treated as a single

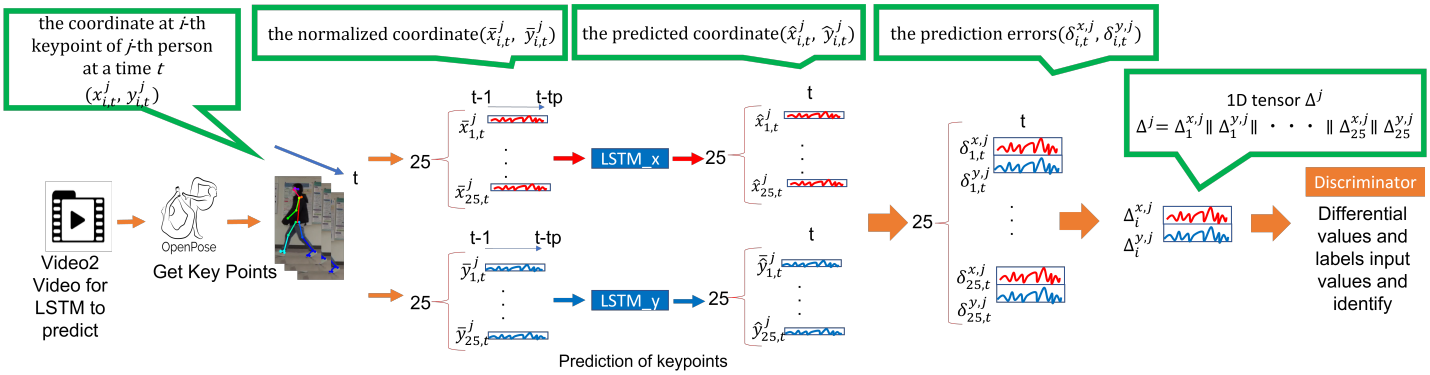


Fig. 5: Operation flow in the identification model.

TABLE I: CASIA Gait.

	Subjects	Sequences	Scenes	Condition change
DataSetA	20	240	Outdoors	Observation direction (3: front, oblique, side)
DataSetB	124	13640	Indoor	Observation direction (11:18° increments), clothing, objects carried
DataSetC	153	612	Night/Infrared	Speed, Objects carried

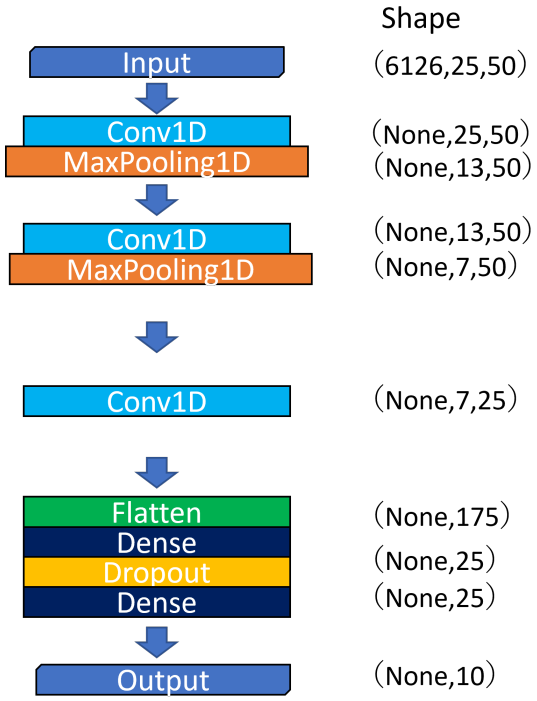


Fig. 6: Parameters in the discriminator.

data set. The total number of videos is 40 and the number of data is 2400. Since data augmentation was performed, the total number of data is 4800.

Two types of datasets are used in this study. One is the skeletal coordinate data and person labels of the videos used as the LSTM dataset, and the other is the prediction errors and person labels used as the discriminator dataset. The datasets are divided into train for training, validation for validation, and test for testing, respectively, to verify the accuracy and loss of the model. In this study, we use 10% of the dataset as validation for the LSTM dataset and 20% of the dataset as validation for the discriminator dataset. In addition, test is performed when all training has been completed.

The CASIA Gait video used as the database is a 412KB AVI file. The frame rate is 30 fps, and it is divided into images with a similar frame rate for skeletal estimation in OpenPose. The performance of the computer used is based on the recommended specifications of OpenPose. The performance of the computer used in this study is based on Windows 10 OS, Core(TM) i7-1165G7 CPU, and 16 GB memory.

C. Methods using Skeletal Information

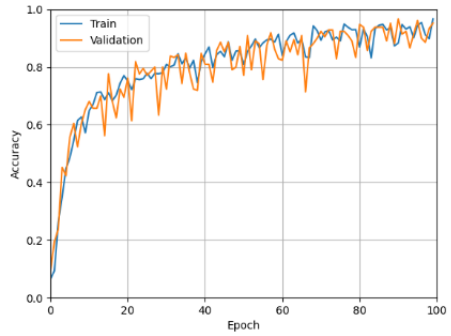
The flow of the identification method without LSTM (This method is henceforth referred to as Baseline.) is shown in Fig. 2. Compared to the proposed method, it does not calculate the prediction error in LSTM. This method uses only skeletal information obtained from the video using OpenPose. When the normalized coordinates are $(\overline{x}^j_{i,t}, \overline{y}^j_{i,t})$, the amount of movement $(u^{x,j}_{i,t}, u^{y,j}_{i,t})$ from the time $t-1$ to t is calculated as follows:

$$\begin{cases} u^{x,j}_{i,t} = \overline{x}^j_{i,t} - \overline{x}^j_{i,t-1} \\ u^{y,j}_{i,t} = \overline{y}^j_{i,t} - \overline{y}^j_{i,t-1} \end{cases} \tag{9}$$

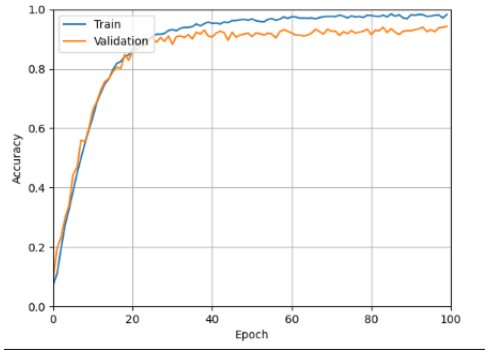
In addition, the same discriminator is used for both the conventional and the proposed methods in order to correctly detect the effect of LSTM on discrimination accuracy. In addition, the Baseline method uses the same video set that was used to train the IM for the prediction method. On the other hand, the Baseline method does not use the same dataset used in training KPM.

D. Identification Accuracy and Loss

The training accuracies of the proposed method and the method using skeletal information are shown in Fig. 7. From



(a) Baseline



(b) Proposed

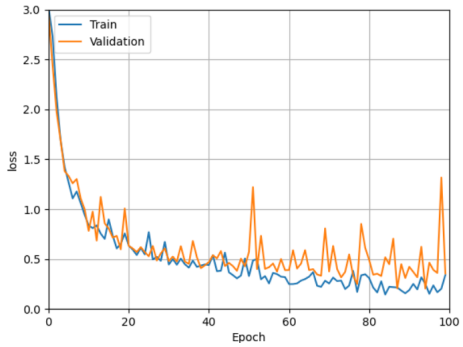
Fig. 7: Transition of discrimination accuracy in the training process.

TABLE II: Test accuracy in IM.

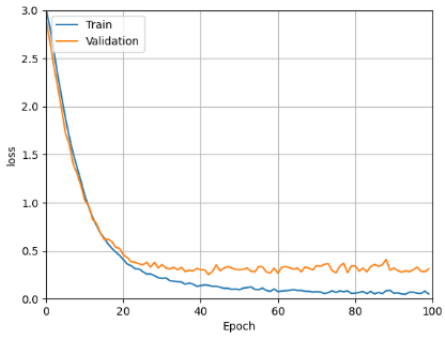
Proposed	0.9785
Baseline	0.9546

the trains in the figure, we can see that the accuracy of both methods is high during training. The accuracy of the skeleton-based method is higher than that of the proposed method at each epoch. This is due to differences in the input values used by each method. The proposed method uses prediction errors, therefore, fluctuations in accuracy and loss values are small. On the other hand, the Baseline method is varies widely even if normalization is performed. The same cause as the training accuracy can be considered from the loss during the training of Fig. 8.

The average of 10 trials in test accuracy is shown in Table II. The test accuracy of the proposed method is 2.5% higher than that of the method using the skeleton information. The mean values of the test losses are shown in Table III. Due to the fluctuations on the loss values in the Baseline method, the average of test loss becomes much larger than the proposed method in this experiment. Because of the stability of the performance, we can say that the prediction errors introduced in the proposed method is useful for the gait recognition system.



(a) Baseline



(b) Proposed

Fig. 8: Transition of loss values in the training process.

TABLE III: Test loss in IM.

Proposed	0.0819
Baseline	0.9710

E. Time Consumption

The proposed method and the Baseline method are shown in Fig. 9. From each method, the processing time is divided into the video segmentation time, OpenPose processing time, LSTM prediction time, and discriminator time. For both methods, OpenPose accounted for about 90% of the total processing time. On the other hand, LSTM's forecasting time was about 3% of the total. This result indicates that the processing time of the forecasting methods is about the same as that of the Baseline method. The processing time of the proposed method increased by about 4% compared to the processing time of the baseline method.

V. Conclusions

In this study, we proposed a method to identify skeletal coordinate data of video images by calculating prediction errors using LSTM. Specifically, we proposed a method for discriminating skeletal coordinate data from videos using OpenPose and LSTM, and compared the accuracy of the proposed method with a method using skeletal coordinate data from videos without LSTM. In the future, we aim to further improve the accuracy by adjusting the LSTM model to obtain the prediction error and the discriminator model. In this study,

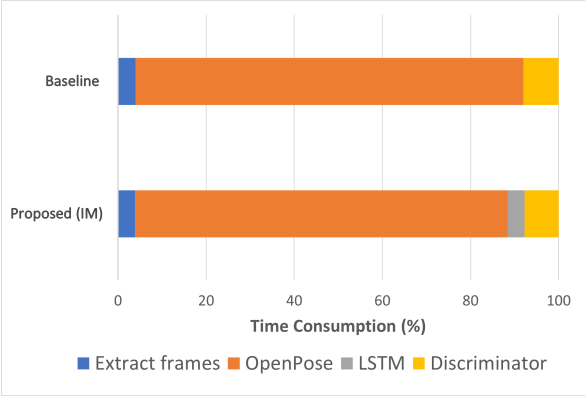


Fig. 9: Percentage of time consumption in each operations.

only time-series skeletal coordinate data and person labels were used to create the dataset. However, actual time-series skeletal coordinate data is strongly affected by intra-individual variables such as observation direction, clothing, and presence or absence of luggage. In the future, we aim to improve the discrimination performance by constructing a discrimination model that takes such factors into account.

Acknowledgment

This study was supported by the JSPS KAKENHI (22K19777), JST SICORP (JPMJSC20C3), JST CREST (JPMJCR20D3), Japan.

References

[1] Y. Makihara, D. Muramatsu, and Y. Yagi, "Walking admissibility and its application to forensic science," *IEICE Fundamental Review*, vol. 14, no. 4, pp. 318–328, 2021.

[2] K. Nakamura, Y. Utsumi, M. Iwamura, *et al.*, "Evaluation of the accuracy of person area extraction for higher accuracy of walker identification," *Research Report of Information Processing Society of Japan*, vol. 2017-CVIM-207, no. 18, pp. 1–7, 2017.

[3] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 7291–7299.

[4] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 01, pp. 172–186, 2021, ISSN: 1939-3539.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] H. Y. Toshimi Gunji, "Verification of the accuracy of gait verification by movement up stairs," *IPSJ Interaction 2021*, no. 3Q09, pp. 809–811, 2021.

[7] N. Asami, A. Moro, Y.-H. Ji, and K. Umeda, "Personal identification using deep generative models from time-series skeletal data acquired with OpenPose," *The 26th Symposium on Sensing via Image Information*, no. IS3-34, 2020.

[8] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *Published as a conference paper at ICLR 2015*, vol. 9, no. 01, pp. 1–15, 2017.

[9] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, no. 4, pp. 26–31, 2012.

[10] K. Shota, T. Yasushi, and M. Yoshiki, "A study on person identification based on walking motion," *Research Report of Information Processing Society of Japan*, vol. 2018-HCI-177, no. 31, pp. 1–4, 2018.

[11] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," *Proc. the 18th Int. Conf, On Pattern Recognition*, no. 4, pp. 441–444, 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Study on Eliminating Biased Node in Federated Learning

Reon Akai*[1] and Minoru Kuribayashi† and Nobuo Funabiki*
* Graduate School of Natural Science and Technology, Okayama University, Japan
E-mail: [1] py365q4n@s.okayama-u.ac.jp
† Center for Data-driven Science and Artificial Intelligence, Tohoku University, Japan

***Abstract*—Federated Learning (FL) is a method to perform a deep learning technique with distributed nodes without aggregating their individually collected data. Each node can keep its own dataset secret, and a central server aggregates the updated parameters in a deep learning model from each node. However, it is reported that FL is vulnerable to poisoning attacks on training data. In this study, we propose a method to exclude biased nodes by analyzing the statistical characteristics of the weight parameters sent by each node to the central server. In the computer simulation, we evaluate the detection accuracy of biased node, and conduct discussions.**

I. Introduction

With the improvement of data mining technology, various studies have been conducted on deep learning to extract useful information from data collected and accumulated everywhere. In fact, there is a growing trend of utilizing deep learning with the data possessed by companies and hospitals. However, there are limitations to the data that each organization can collect. Furthermore, there are challenges in using such data for training across organizations due to privacy concerns, as the data held by organizations often contains personal information.

To address this problem, the idea of Federated Learning (FL) was proposed by McMahan et al. in 2017 [1]. FL allows machine learning to be performed in a decentralized manner without aggregating the data. FL has been studied using various techniques [2], and one of the representative methods is Federated Averaging (FedAvg) [1]. In FedAvg, each node holds its own dataset for privacy protection, and communication between distributed nodes is facilitated through a central server. This central server distributes the parameters of the Deep Neural Network (DNN) model to each node and aggregates the updated parameters from each node.

However, FedAvg faces the problem of performance degradation when there are malicious nodes among the distributed nodes. If the parameters received from such malicious nodes are aggregated to update the DNN model, the overall performance of the system can be compromised. In fact, research by Bagdasaryan et al. [3] demonstrates the vulnerability of FL to poisoning attacks against the training data, and even more so to poisoning attacks against the learning model. As the number of nodes participating in FL increases, the threat of data poisoning, whether intentional or unintentional, is heightened. Some nodes may intentionally set a backdoor in their trained DNN model and upload the parameters to the central server. Additionally, there is a possibility that the collected data from each node may lack reliability due to fluctuations caused by device and environmental differences. Concerning poisoning by malicious nodes, there must be a significant bias in the weight parameters, as the model may be trained to classify samples with labels different from their true labels. On the other hand, addressing the impact of fluctuations caused by non-malicious nodes remains a challenge. Additionally, when the number of malicious or non-malicious nodes changes per round, the number of nodes to be removed may vary. It is necessary to calculate and consider the number of nodes to be removed for each round by taking into account this variation.

In this study, we propose and evaluate a method for excluding biased nodes in FL so that the DNN model is protected against malicious party and it is correctly updated in the distributed training process, where biased nodes includes malicious and non-malicious ones. It is assumed that the number of biased nodes varies per round. The proposed method attempts to find biased nodes whose proportion of weight parameters following a Gaussian distribution falls below a certain threshold. We focus on the weight parameters that each node sends to the central server and explore a method to prevent performance degradation by excluding nodes that exhibit outliers compared to legitimate nodes. To numerically validate the effectiveness of this proposed method, we perform statistical analysis using hypothesis testing on the received parameters from each node. We demonstrate the relationship between the analyzed parameter characteristics and the accuracy of the global model, thereby confirming the validity of the approach.

From our computer simulations, it is possible to determine with a proposed measurement for the confidence level whether biased nodes are participating. Furthermore, it has been confirmed that excluding such nodes with a significant number of weight parameters that do not follow a normal distribution among the weight parameters transmitted from each node during the training process of FedAvg is beneficial.

II. Federated Learning

FL [4] is a method of performing machine learning in a distributed manner without aggregating data.

A. Overview of FL

Consider a general FL system consisting of one server and N nodes as shown in Fig. 1. Let D_i denote the local database held by node C_i, and M_i denote the local model, where $i \in \{1, 2, \ldots, N\}$.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

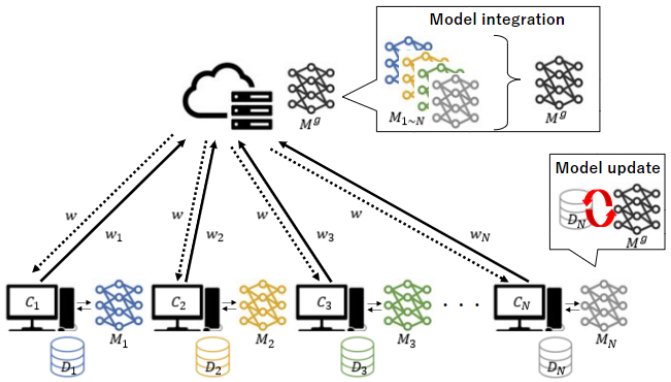


Fig. 1. Federated Learning

The purpose of the server is to train a model on the data present on N nodes. The active clients participating in local training need to find the weight vector $\boldsymbol{w}$ of the deep learning model that minimizes the loss function. Generally, the server aggregates the weight parameters received from N nodes as shown in the following equation.

$$\boldsymbol{w} = \sum_{i=1}^{N} p_i \boldsymbol{w_i}, \tag{1}$$

where $\boldsymbol{w_i}$ is the vector of weight parameters trained at node i, $\boldsymbol{w}$ is the parameter vector aggregated by the server, and p_i is calculated $p_i = \frac{|D_i|}{|D|} \geq 0 \quad (\sum_{i=1}^{N} p_i = 1)$, $|D| = \sum_{i=1}^{N} |D_i|$. Here, $|D_i|$ is the total amount of data samples held by node i. Furthermore, the optimization problem of these weight parameters can be formulated to find $\boldsymbol{w}^\star$ as shown in Eq. (2).

$$\boldsymbol{w}^\star = \arg\min_{w} \sum_{i=1}^{N} p_i F_i(\boldsymbol{w}, D_i), \tag{2}$$

where $F_i(\cdot)$ is the local loss function of node i, which is given by the local empirical risk. The learning process of such an FL system usually involves the following five steps, repeated to update the model. This sequence of steps is referred to as a round.

Step 1: Node selection: The server selects the nodes to participate in the round.
Step 2: Distribution of global model to selected nodes. (In the first round, the parameters of the global model are randomly determined.)
Step 3: Local model update: Nodes train and update the distributed model using their own locally held data. The updated model is referred to as the local model.
Step 4: Transmission of local model parameters to the server.
Step 5: Global model update: The server aggregates the transmitted parameters and updates the global model.

In FL, N nodes with the same data structure collaboratively train deep learning models using the server. After a sufficient number of local training and updates between the server and the nodes, the solution to the optimization problem in Eq. (2) can converge to the global optimal solution of the learning model. In this study, we used the Flower [5] FL library to conduct FL.

B. Federated Averaging

FedAvg [1] is one of the representative methods in the research on FL. In FedAvg, each node holds its own dataset for privacy protection, and communication among distributed nodes is facilitated through a central server. This server distributes the parameters of the learning model to each node and aggregates the updated parameters from each node. The aggregated parameters are then integrated into a single model for learning by taking the weighted average.

In FedAvg, the aggregated parameters are computed by taking the weighted average as shown in Eq. (3) in each round.

$$\boldsymbol{w^t} = \sum_{i=1}^{N} \frac{n_i}{n} \boldsymbol{w_i^t}, \tag{3}$$

where t represents the round number, $\boldsymbol{w^t}$ is the parameter of the global model integrated by the server in the t-th round, n_i is the number of data samples for node i, n is the total number of data samples, and $\boldsymbol{w_i^t}$ is the parameter of the local model trained by node i and sent to the server in the t-th round. The following shows Algorithm1 for FedAvg.

Algorithm 1 Calculate Federated Averaging. C is the fraction of nodes that perform computation on each round. The N nodes are indexed by i; B is the number of local epochs, and η is the learning rate. $P_i = n_i$. $\nabla l(\boldsymbol{w}; b)$ is the gradient for loss function.

Server executes:
  initialize $\boldsymbol{w^0}$
  for each round $t = 1, 2, \cdots$ **do**
    $m \leftarrow \max(C \cdot N, 1)$
    $S_t \leftarrow$ (random set of m nodes)
    for each node $i \in S_t$ in parallel **do**
      $\boldsymbol{w_i^t} \leftarrow \text{NodeUpdate}(i, \boldsymbol{w^{t-1}})$
    $\boldsymbol{w^t} \leftarrow \sum_{i=1}^{N} \frac{n_i}{n} \boldsymbol{w_i^t}$
NodeUpdates(i,w): //Run on node i
  $B \leftarrow$ (**split** P_i **into batches of size** B)
  for each local epoch k from 1 to E **do**
    for batch $b \in B$ **do**
      $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla l(\boldsymbol{w}; b)$
  return w **to server**

C. Existing Model Poisoning Attacks in Federated Learning

Various attacks [3],[6]–[9] have been proposed to poison the global model in FL, and all these attacks rely on compromised legitimate nodes. Based on the attacker's objective, these attacks can be categorized into two types: non-targeted model poisoning attacks [7],[9] and targeted model poisoning attacks [3],[6],[8]. Non-targeted model poisoning attacks aim to decrease the test accuracy of the global model. On the other hand, targeted model poisoning attacks, which is known as Byzantine attack [10] aim

to make the global model output the attacker's chosen target label for a selected target input. The objective of the attacker is to maximize the accuracy of the trained model on the specific data, while ensuring that the model performance on the other data does not degrade significantly. The attacker makes a backdoor in the model during the training phase by injecting the data which label is flipped to one of the labels. This research focuses on untargeted poisoning attacks.

D. Defense Strategies Against Model Poisoning Attacks

Several defense strategies[7],[10]–[16] have been proposed to mitigate model poisoning attacks in FL, and they can primarily be divided into two categories. The first type of defense strategies [10]–[14], [16] focuses on designing Byzantine-robust aggregation rules. Their idea is to reduce the influence of statistical outliers in the updates of local models. For example, Trimmed-mean [10] removes the maximum and minimum values from the local model updates before calculating the average. In addition, SparseFed [17] employs a technique where the central server calculates the aggregated gradients and only updates the parameters corresponding to the top k absolute values in the model. The second type of defense strategies [18],[19] aim to provide provable guarantees against poisoning attacks.

III. Proposed Method

In this study, we investigate methods to prevent the degradation of overall system performance in FedAvg caused by the bias in the weight parameters. First, we consider the biased nodes assumed in this study and examine the impact of their participation on the system performance. Next, we focus on the weight parameters sent to the central server and investigate the differences between biased nodes and legitimate nodes. We propose a method based on a statistical analysis to protect the model by excluding outliers in the weight parameters transmitted from each node. Specifically, we propose a method for excluding nodes in FedAvg where the number of biased nodes varies from round to round, and the number of nodes to be removed changes accordingly, by excluding nodes that have a proportion of parameters following a Gaussian distribution below a certain threshold.

To numerically validate the effectiveness of this proposed method, we will conduct statistical analysis using hypothesis testing on the parameters received from each node. We will demonstrate the relationship between the analyzed parameter characteristics and the accuracy of the global model, thus confirming its validity.

A. Biased Node

As the number of nodes grows, it becomes more difficult to manage the soundness of the FL framework. A malicious party could inject poisoning data into the training dataset at some nodes and secretly embed a backdoor into the model centralized on a central server. Even if all nodes are trusted, the local data obtained at each node could be biased by environmental differences or unintended bugs in the system.

We suppose to train a CNN-based image classifier which predicts a label of given image. It is assumed that the biased nodes include both malicious and non-malicious nodes. The difference of these nodes are described as follows.

- Malicious node
  Images which labels have been changed from ground-truth are intentionally added to the training dataset in the node.
- Non-malicious node
  Some images collected at the node may have labels different from the ground-truth due to some occasional reasons such as environmental noise or system bugs.

It is reasonable to assume that the number of malicious nodes is rather smaller than the non-malicious nodes. In addition, the number of mislabeled images in the dataset at the malicious nodes are assumed to small so that the amount of changes in the DNN model parameters is managed to be as small as possible.

Here, we consider the following two rates in the biased nodes.

- Poisoning Rate: R^{poison}
  It is the rate of images in the dataset that are mislabeled.
- Bias Rate: R^{bias}
  It is the rate of biased nodes among all nodes in FL framework.

In both cases, we attempt to detect the biased nodes by analyzing the irregularity in the weight parameters transmitted to a central server.

B. Consideration on Excluding Outlier Weight Parameters

It is conceivable that weight parameters transmitted from biased nodes within the distributed network deviate significantly from the weight parameters of legitimate nodes, potentially exhibiting outliers. In this study, we examine whether excluding nodes that transmit outlier weight parameters is a valid approach in FedAvg.

We focus on the weight parameters transmitted to the central server, and the differences between biased and legitimate nodes. At each epoch, the central server checks the validity of weight parameters uploaded from each node by calculating the difference between the weight parameters at the server in the previous epoch and the weight parameters updated at each node. In general, the differences are assumed to follows a Gaussian distribution with zero mean. If the dataset for training at a certain node is biased, the Gaussian assumption is no longer valid.

In this study, we use a CNN model as an image classifier. As shown in Fig. 2. We observe all weight parameters of the fully connected layer, which is composed of the nearest to the output of the CNN. The weight observations are performed at the same position for each node to ensure consistency.

The validity of the weight parameters transmitted from each node is determined by assessing whether they follow a normal distribution. The accuracy of the globally integrated model at the central server is used to perform a comparative analysis between models that include biased nodes and models composed only of legitimate nodes.

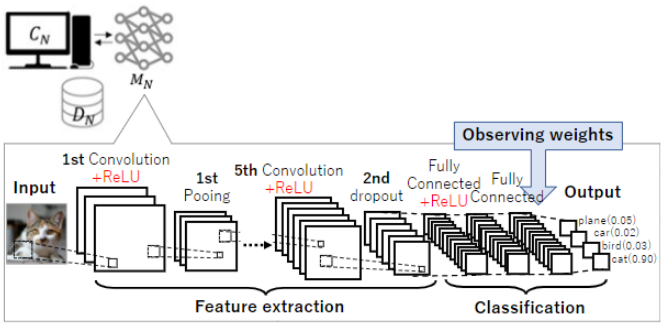


Fig. 2. Observation of weight parameters at the same location

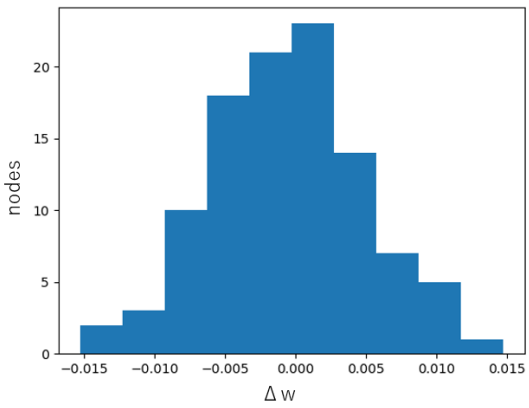


Fig. 3. Histogram of weight parameters of legitimate nodes.

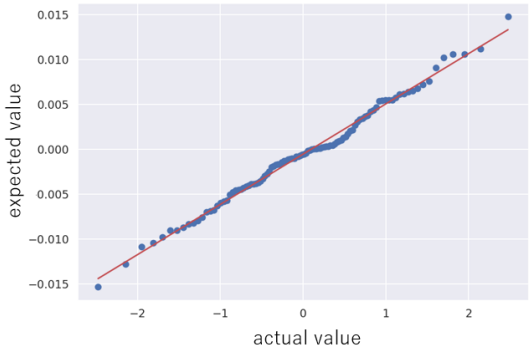


Fig. 4. Q-Q plot of weight parameters in a normal FedAvg.

C. Analyzing Weight Distributions

We investigate the statistical distribution of weight parameters received from each node. When evaluating the statistical distribution of weight values, we employ a histogram, quantile-quantile (Q-Q) plot, and the Shapiro-Wilk test. The Q-Q plot is a probability plot that takes the expected values of observations based on a normal distribution on the y-axis and the observations themselves on the x-axis. By determining the percentiles (cumulative probabilities) from the ranked order of observations and predicting the expected values using the inverse function of the probability density function of the normal distribution, we can assess if the observations follow a normal distribution. If the plot aligns along a straight line, it indicates that the observations follow a normal distribution.

The Shapiro-Wilk test [20] is a statistical method used to examine whether data follows a normal distribution by calculating the test statistic W and p-value and performing a hypothesis test to determine if the data follows a normal distribution. There are two types of tests for normality: the Shapiro-Wilk test and the Kolmogorov-Smirnov test. In this study, we utilize the Shapiro-Wilk test due to the small sample size (100 data points) obtained. The definitions of the test statistic W and p-value are provided in Eq. (4) and Eq. (5).

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{4}$$

where n represents the sample size, a_i denotes the ranks of the samples, x_i is the sample data, $x_{(i)}$ represents the ranked sample data, and $\bar{x}$ is the sample mean.

$$p = 1 - \left(\frac{\sum_{i=1}^{k} b_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right), \tag{5}$$

where k represents the sample size, b_i represents the standard normal percentiles of the ranks, x_i is the sample data, and $\bar{x}$ is the sample mean.

In this study, the null hypothesis (H0) and the alternative hypothesis (H1) are set as follows:

Null hypothesis (H0): The data follows a normal distribution.

Alternative hypothesis (H1): The data does not follow a normal distribution.

The confidence level is set to 95[%], and when $p < 0.05$, it is considered statistically significant. Thus, if the test result shows $p < 0.05$, it is concluded that the data does not follow a normal distribution, while $p \geqq 0.05$ indicates that it cannot be concluded that the data does not follow a normal distribution (i.e., it follows a normal distribution).

D. Preliminary Experiment

To examine the characteristics of the weight parameters transmitted from biased nodes, we conducted partial observations of the same positions of the parameters transmitted from legitimate nodes and the biased nodes subjected to an extreme poisoning attack.

In the case of legitimate nodes, the histogram and Q-Q plot results of the weight parameters are shown in Fig. 3 and Fig. 4. It is observed from the figures that the weight parameters follow normal distribution. The Shapiro-Wilk test calculated based on Eq. (5) yields $p = 0.8845 \geqq 0.05$, indicating that the weight parameters follow a normal distribution at a significance level of 5[%].

Next, we measure the weight parameters of biased nodes subjected to the case of $(R^{\text{bias}}, R^{\text{poison}}) = (1.0, 0.5)$, which

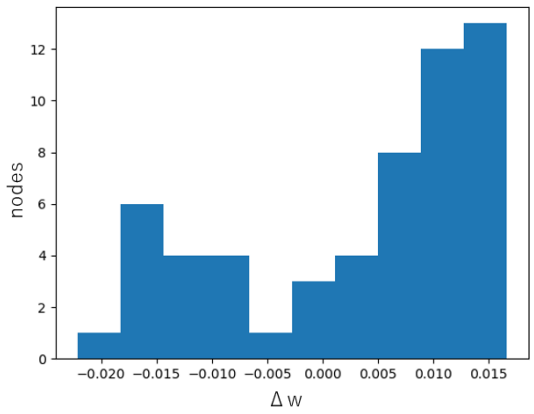


Fig. 5. Histogram of weight parameters in the case of $(R^{\text{bias}}, R^{\text{poison}}) = (1.0, 0.5)$.

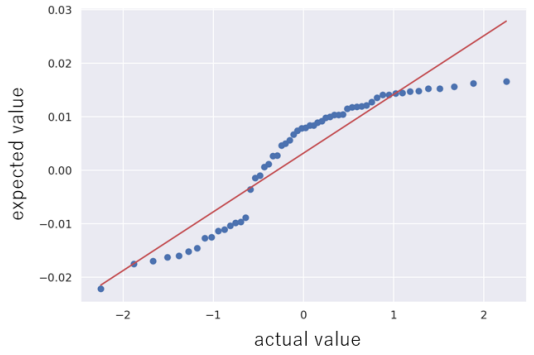


Fig. 6. Q-Q plot of weight parameters in the case of $(R^{\text{bias}}, R^{\text{poison}}) = (1.0, 0.5)$.

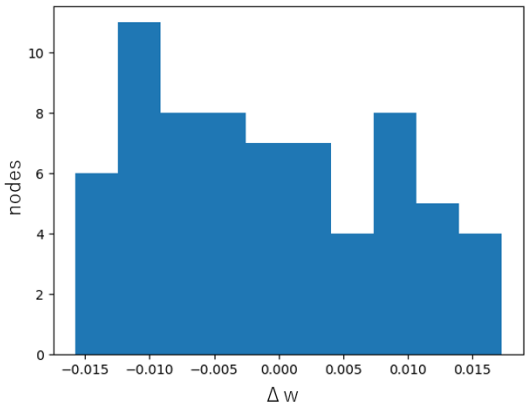


Fig. 7. Histogram of weight parameters in the case of $(R^{\text{bias}}, R^{\text{poison}}) = (0.5, 1.0)$.

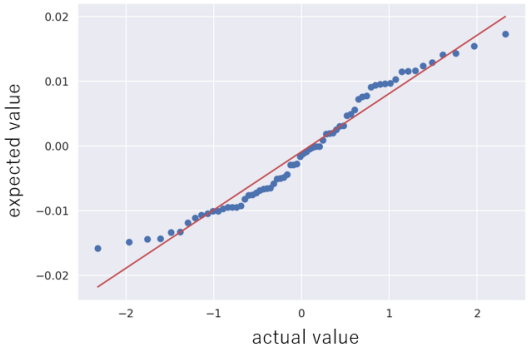


Fig. 8. Q-Q plot of weight parameters in the case of $(R^{\text{bias}}, R^{\text{poison}}) = (0.5, 1.0)$.

results are shown in Fig. 5 and Fig. 6. The Shapiro-Wilk test calculated using Eq. (5) gives $p = 0.0172 < 0.05$, indicating that the weight parameters do not follow a normal distribution at a significance level of 5[%]. In case of $(R^{\text{bias}}, R^{\text{poison}}) = (0.5, 1.0)$, the results are shown in Fig. 7 and Fig. 8. Since the Shapiro-Wilk test is $p = 0.00003 < 0.05$, the weight parameters also do not follow a normal distribution.

From these results, it can be speculated that as the poisoning rate increases for both nodes and data, the proportion of parameters in the fully connected layer of the CNN that do not follow a normal distribution increases. In other words, for each position of the parameters sent by all nodes, the number of parameters with Shapiro-Wilk test statistic $p < 0.05$ is increasing. Therefore, we examined the change in the proportion of normally distributed parameters for all parameters in each node as the poisoning rate varies.

E. Integration

Let $\boldsymbol{w}^t$ represent the weight parameters of the initial model (global model) distributed to the N nodes in the t-th round, and let $\boldsymbol{w}_i^t$ represent the weight parameters transmitted from the i-th node. The difference in weight parameters between the weight parameters transmitted from each node and the weight parameters of the initially distributed model, denoted as $\Delta \boldsymbol{w}_i^t$, can be calculated using Eq. (6).

$$\Delta \boldsymbol{w}_i^t = |\boldsymbol{w}_i^t - \boldsymbol{w}^t| \tag{6}$$

In this study, an analysis was conducted on all L weight parameters comprising the fully connected layers. The update amount, $\Delta \boldsymbol{w}_i^t$, of the L weight parameters $\boldsymbol{w}^t$ was calculated every N nodes. Subsequently, the normality of the update amounts, $\Delta \boldsymbol{w}_i^t$, for the same locations sent by each of the α nodes was determined using the Shapiro-Wilk test statistic, $\mathcal{P}' = \{p_i \mid p_i < 0.05, (1 \leq i \leq L)\}$. Then, the ratio of weight parameters with normality among all weight parameters in the fully connected layers, denoted as p_{rate}, can be calculated using the following Eq. (7):

$$p_{\text{rate}} = \frac{L - |\mathcal{P}'|}{L} \tag{7}$$

IV. Computer Simulation

We conducted experiments to measure the irregularity observed at the weight parameters transmitted from biased nodes.

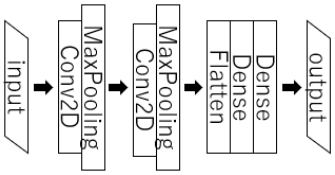


Fig. 9. Structure of the CNN model used in the experiment.

TABLE I
HYPER-PARAMETERS IN FEDAVG.

Number of nodes	100
Local mini-batch size	10
Local epoch	1
Round number	500
Learning rate	0.001

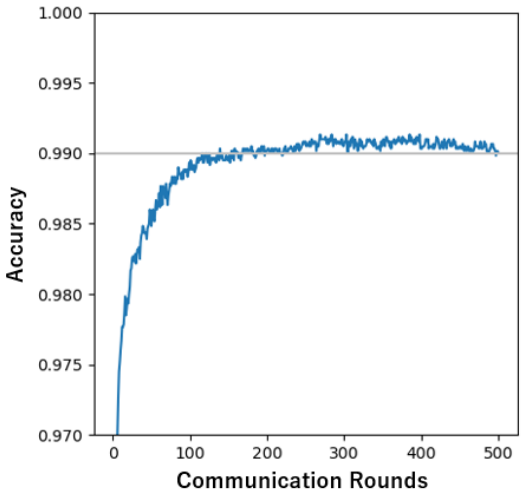


Fig. 10. Accuracy of 500 rounds with 100 nodes in FedAvg.

A. Experimental Conditions

In this study, a CNN image classifier with the structure shown in Fig. 9 was used for the classification of 10 classes using the MNIST dataset. The MNIST dataset consists of a total of 70,000 images and labels, with 60,000 images used for training and 10,000 images used for testing. Among the training data, 90[%] (54,000 images) were used for training, 10[%] (6,000 images) for validation, and all 10,000 images from the test data were used for testing. Additionally, these datasets were divided equally among the nodes of FL, meaning that each node used $54,000/N$ images for training, $6,000/N$ images for validation, and $10,000/N$ images for testing, where N represents the number of nodes. Therefore, each node had its own unique dataset, replicating the scenario where each node possesses distinct data in the FL process.

The FedAvg algorithm used in this study trained the model with $\alpha = 100$ nodes. The number of weight parameters observed in this experiment is $L = 5120$, where the last FC layers has 512 nodes and the output is 10 classes. Fig. 10 shows that the accuracy converged to 99.25[%] after 500 rounds of training with a local minibatch size of 10 and a local epoch of 1. The hyper-parameters used in the FedAvg implementation in this study are shown in Table I.

There are various types of malicious nodes that can be considered in FedAvg. In this study, we use FedAvg to train a deep learning model for image classification on the MNIST dataset. Therefore, the malicious nodes considered in this study are nodes that have maliciously modified the labels of the MNIST dataset's 10 classes. Such poisoned data are crafted by modifying the label number ℓ of the MNIST dataset to a new label number ℓ' as follows:

$$\ell' = (\ell + 1) \mod 10 \tag{8}$$

This modification changes the original label ℓ to a different label ℓ' by incrementing it by 1 and taking the modulo 10. In case of non-malicious node, we also use the same modification as the malicious node in the experiments. Thus, biased nodes with the bias rate R^{bias} are trained by the poisoned data crafted by Eq.(8) with the poisoning rate R^{poison}.

The change in image classification accuracy of the integrated model at the central server due to the participation of biased nodes, as compared to when only legitimate nodes participate, was examined in this study.

B. Experimental Results

Based on the proposed method described in Section III-C, we evaluate the changes in classification accuracy and the proportion of normally distributed parameters in the weight parameters using computer simulations by varying the poisoning rates of each node and dataset in FedAvg. For each node and its dataset, we observed the changes in the proportion of normally distributed parameters by varying the poisoning rates as described in Scenarios 1 and 2 below:

- Scenario 1: malicious
  The number of malicious nodes is small, and the biased rate is fixed to $R^{\text{bias}} = 0.01$. The poisoning rate is varying from the following range: $R^{\text{poison}} \in [0.02, 0.5]$.
- Scenario 2: non-malicious
  The number of non-malicious nodes is relatively large, and the poisoning rate and bias rate are varying from the following range: $R^{\text{poison}} \in [0.2, 1.0]$ and $R^{\text{bias}} \in [0.02, 0.5]$

The changes in classification accuracy and the proportion of normally distributed parameters for Scenarios 1 and 2 are shown in Fig. 11 and Fig. 12, respectively. Both values are calculated from an average of five simulations. As the poisoning rates increase, both the classification accuracy and the proportion of normally distributed parameters decrease.

When $R^{\text{bias}} = 0.01$, which means even when one node contains poisoned data, an increase in the data poisoning rate leads to a tendency of decreasing both the classification accuracy and the proportion of normal distribution parameters. Furthermore, it was observed that an increase in R^{bias}, rather than R^{poison}, leads to a decrease in the proportion of normal distribution parameters.

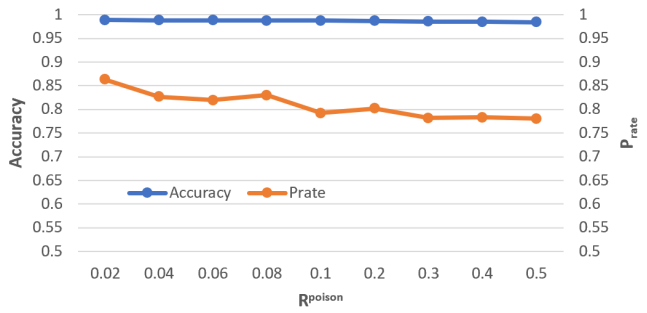


Fig. 11. Comparison of accuracy and p_{rate} in Scenario 1.

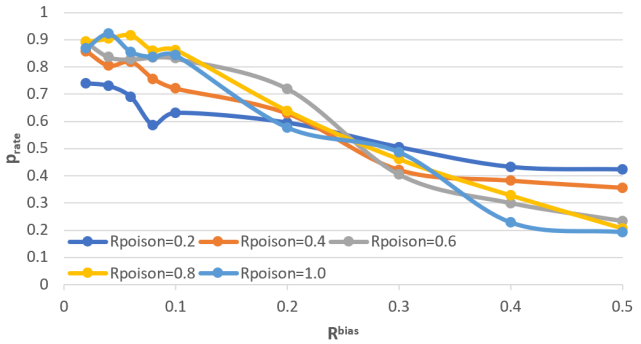


Fig. 12. Comparison of p_{rate} in Scenario 2.

C. Discussion and Challenges

In this study, we conducted an investigation and proposed a method to prevent the overall performance degradation of the system by excluding biased nodes in FedAvg.

Based on the results in Section IV-B, it was observed that the classification accuracy of the aggregated model decreases as the poisoning rates increase. Additionally, regarding the normality of weight values, it was shown that increasing poisoning rates lead to biases away from the normal distribution, resulting in a higher proportion of weight values with non-normal distributions.

The evaluation of the proportion of normally distributed parameters with respect to the poisoning rates indicated a tendency for more weight values to become non-normal as the poisoning rates increase.

From these results, it can be inferred that as the poisoning rates increase, the performance of the aggregated model decreases, and there is an increase in the proportion of weight values with non-normal distributions. Therefore, when the proportion of normally distributed parameters is low for all parameters in FedAvg, it indicates the presence of biased nodes.

To minimize the impact of poisoning, it is crucial to manage the poisoning rates of each node and dataset and limit the increase in non-normal weight values. One potential approach could be excluding nodes that contributed to lowering the proportion, i.e., nodes with a higher number of outlier parameters during p-value calculation.

Furthermore, considering the non-normal distribution of weight values, it is important to explore weight aggregation methods and retraining strategies that account for the impact of poisoning.

V. Conclusions

In this study, we investigated a method to protect the performance of the aggregated model in the FedAvg algorithm by excluding nodes that transmit outlier weight values, aiming to mitigate the impact of biased nodes.

Through the experiments, we demonstrated that the weight values transmitted from legitimate nodes in FedAvg tend to follow a normal distribution. This understanding of the statistical properties of weight distributions is valuable and can contribute to the development of techniques for detecting and handling outlier weight values.

Furthermore, we found that by examining the proportion of parameters that follow a normal distribution, we can consider the possibility of the participation of biased nodes. By excluding nodes that contribute to the decrease in the proportion of normally distributed parameters, which include more parameters that contribute to the decrease, we can effectively exclude biased nodes.

For future prospects, it is important to further research and improve methods for detecting and excluding biased nodes in federated learning. Exploring approaches that consider a larger number of weight locations can accurately identify outlier weight values, leading to more effective protection of model performance. Additionally, it is necessary to validate the effectiveness of these methods in FL approaches other than FedAvg. Moreover, we should consider extending experiments to different datasets, like RGB images (e.g., CIFAR-10). This will allow us to assess the generalizability of our proposed bias detection and exclusion methods across various data types. Additionally, it is crucial to evaluate the proposed method with different network architectures, such as ViT, to ensure its applicability and robustness in various neural network structures.

Acknowledgment

This study was supported by the JSPS KAKENHI(22K19777), JST SICORP (JPMJSC20C3), JST CREST (JPMJCR20D3), Japan.

References

[1] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. International Conference on Artificial intelligence and statistics (AISTATS’17)*, 2017, pp. 1273–1282.

[2] Q. Li, Z. Wen, Z. Wu, *et al.*, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowledge and Data Engineering*, pp. 1–1, 2021.

[3] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS'20)*, 2020, pp. 2938–2948.

[4] K. Wei, J. Li, M. Ding, *et al.*, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[5] D. J. Beutel, T. Topal, A. Mathur, *et al.*, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[6] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*, PMLR, 2019, pp. 634–643.

[7] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *Proceedings of the 29th USENIX Conference on Security Symposium*, 2020, pp. 1623–1640.

[8] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *International conference on learning representations*, 2020.

[9] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation," in *Uncertainty in Artificial Intelligence*, PMLR, 2020, pp. 261–270.

[10] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, PMLR, 2018, pp. 5650–5659.

[11] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[12] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–25, 2017.

[13] R. Guerraoui, S. Rouault, *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning*, PMLR, 2018, pp. 3521–3530.

[14] S. Rajput, H. Wang, Z. Charles, and D. Papailiopoulos, "Detox: A redundancy-based framework for faster and more robust gradient aggregation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] S. Shen, S. Tople, and P. Saxena, "Auror: Defending against poisoning attacks in collaborative deep learning systems," in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 2016, pp. 508–519.

[16] C. Xie, S. Koyejo, and I. Gupta, "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance," in *International Conference on Machine Learning*, PMLR, 2019, pp. 6893–6901.

[17] A. Panda, S. Mahloujifar, A. N. Bhagoji, S. Chakraborty, and P. Mittal, "Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 7587–7624.

[18] X. Cao, J. Jia, and N. Z. Gong, "Provably secure federated learning against malicious clients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 6885–6893.

[19] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "Crfl: Certifiably robust federated learning against backdoor attacks," in *International Conference on Machine Learning*, PMLR, 2021, pp. 11 372–11 382.

[20] P. Royston, "Approximating the shapiro-wilk w-test for non-normality," *Statistics and computing*, vol. 2, pp. 117–119, 1992.

Can StArtGAN withstand Image Processing Attacks?

Koi Yee Ng*, Simying Ong*, and Yuen Peng Loh†
* Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia.
E-mail: {wva190003@siswa., simying.ong@}um.edu.my
† Faculty of Computing and Informatics, Multimedia University, Malaysia.
E-mail: yploh@mmu.edu.my

***Abstract*—StArtGAN is a constructive-based coverless data hiding framework, that generates various classes of art images to represent secret bits. These generated art images will be collated to form a collage image that carries the secret data. In certain data hiding applications such as watermarking applications, robustness is an important factor to be considered in such a data hiding framework. Therefore, in this paper, we will investigate the robustness level of StArtGAN. Six common image processing operations with multiple variants are applied as the robustness attacks to the collage image output from StArtGAN to examine its classification accuracy (or secret data extraction accuracy) and its perceptual qualities after these attacks. The result proves that this framework is robust to Salt & Pepper Noise, Gaussian Noise, Sharpening, Gaussian Filter, and Compression Attacks.**

I. Introduction

Coverless image data hiding abolishes the utilization of cover images during the secret data embedding process. There are two types of coverless data hiding in image [1], namely the non-constructive- and the constructive-based methods.

The non-constructive-based coverless image data hiding methods usually involve selecting suitable images from the image dataset to represent secret data using special formulas [2], [3]. These methods can perform data hiding without embedding; however, a large image dataset is required to identify a set of suitable images. Besides, they also suffer from low embedding capacity and often have tedious output image transmission protocol since it involves more than one image.

On the other hand, constructive-based coverless data hiding approaches generate stego images from secret. The first approach was introduced by Otori and Kuriyama in 2007 [4] and 2009 [5]. Otori and Kuriyama propose to create a textual output image by generating local binary patterns in the blank image based on the secret data. In the early years of the constructive-based coverless data hiding domain, most of the works are focusing on generating pattern or texture images [6], [7], [8]. The application is rather limited due to the restriction of the constructed output.

Later, inspired by the generative adversarial network (GAN) [9], [10], various constructive-based data hiding models using GAN are proposed [11], [12], [13], [14]. However, [12] and [13] methods are not coverless as they used the GAN model to generate a cover image for data hiding to reduce the detection rate and enhance the robustness level. Hu et. al method [11] uses GAN to embed secret data directly while the image is generated; however, this method suffers from a high bit error rate. In 2021, Ng et. al proposed StArtGAN [14] that generates classes of tiny art images to represent secret bits. These art images will be selected and arranged according to the secret data, and combined as a collage image. This method is not only able to achieve scalable embedding capacity, but the bit error rate is also lower as compared to the other similar methods.

However, can StArtGAN withstand image processing attacks? High robustness to image processing attacks is crucial in some applications, such as watermarking applications, or applications that require noisy transmissions or storage in the cloud, to ensure that the embedded secret data remains recognizable or readable after some interruption. Hence, in this paper, we investigate the robustness level of StArtGAN against some common image processing attacks, which include various noises, filters, and image compression. Section II will present a brief review of the StArtGAN constructive-based coverless data hiding framework. Then, Section III will show the experiment results and analysis for robustness tests using various image processing operations. Section IV will discuss the overall robustness level of the StArtGAN and other similar methods. Finally, Section V concludes the findings of this paper.

II. Revisiting StArtGAN [14]

A. StArtGAN Framework

StArtGAN framework prioritizes the generation of ambiguous art images with high classification accuracy. It consists of a generator and a discriminator. The trained generator will be used as the encoder to generate the art images based on the secret classes, whereas the trained discriminator will act as the decoder to decode the classes of the generated image. For more information on the StArtGAN framework, we refer the readers to [14].

B. Embedding Steps

In this work, the grayscale secret image $G(i, j)$ consists of $W \times H$ pixels will be first converted into 8 bits binary secret, $\Gamma(i, j)$. The $\Gamma(i, j)$ is reshaped into 1 dimension array, ϕ, and divided into multiple chunks $\rho = \phi/\gamma$, where each chunk $\gamma = \log_2 \eta$ bits each and η is the total number of image classes. The ρ is relabelled into φ by performing an XOR operation using a key κ and mapped to the corresponding art image class c to be fed into trained StArtGAN for generation of art images G_A. The G_A are then combined into collage images G_C for transmission.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



Aerial Baboon Boat Cameraman Clock Lena Pepper Pirate Plan Woman

Fig. 1: The 10 secret images.

TABLE I: Sample binary secret relabelling using $\kappa = 101$ when given $\eta = 8$.

Secret Relabelling								
γ	000	001	010	011	100	101	110	111
φ	101	100	111	110	001	000	011	010
c	5	7	6	2	4	3	0	1

As an illustration, when $\eta = 8$ classes of art images are generated, each chunk $\gamma = \log_2 8$ represents 3 bits. By using $\kappa = 101$ to perform the XOR operation, every different 3 bits will be relabelled into φ and mapped to the corresponding classes. The sample of relabelling and mapping is illustrated in Table I.

C. Decoding Steps

At the receiver end, the G'_A will be extracted from the received G_C and fed into the trained discriminator of StArtGAN to output the corresponding image image class c'. The c' will be reversely mapped to the φ' and relabelled into γ' using κ. The γ will be joined into a string of ϕ'. Every 8 bits of the ϕ' will be converted into the corresponding pixels to regenerate the original grayscale image $G(i, j)$' with $W \times H$ pixels.

III. Experiment Results

In this experiment, 10 images are obtained from the USC-SIPI Image Dataset [15] and used as the secret image. To prepare these secret images, they are first transformed into 8-bit grayscale images of size 64×64 pixels. These transformed grayscale images are shown in Fig. 1. Then, for each image, the image pixels are collected and linearized in a top-down, left-right manner. Finally, the linearized pixels are converted into binary form as secret bits. The rest of the secret image embedding process will be similar to those steps mentioned in Section II. The StArtGAN is trained using the WikiArt Dataset [16] to generate the secret class into multiple 32×32 art images and combine them into collages.

A. The Image Processing Operations

In the experiments, the collage images generated from the StArtGAN model will be put on the robustness test. In particular, six common image processing operations are applied and act as an attack in the robustness experiments: (1) Salt & Pepper Noise, (2) Gaussian Noise, (3) Image Sharpening, (4) Median Filter, (5) Gaussian Filter, and (6) Image Compression.

Salt & Pepper Noise is an impulsive noise that presents itself as sparsely white and black pixels, that often mimic the unanticipated disturbance on the image, while Gaussian Noise is an additive noise that has a uniform (normal) distribution that models many random noise distributions in natural settings. On another spectrum, image sharpening is a special filter that emphasizes the image edges, while Median Filter and Gaussian Filter are the linear and non-linear smoothing techniques used to reduce noises in the image while preserving the edges. Finally, the last image processing operation used is image compression because it removes the insignificant parts of the image to reduce the file size, which applies to most of the cloud systems nowadays.

TABLE II: The Image Processing Attacks and their variants.

Attacks	Variance	Levels
Salt & Pepper Noise	Noise Density	0.01, 0.03, 0.05
Gaussian Noise	Noise Variance	0.01, 0.03, 0.05
Image Sharpening	Effect Strength	0.4, 0.6, 0.8
Median Filter	Neighbourhood Size	3×3, 5×5, 7×7
Gaussian Filter	Distribution S.D.	0.1, 0.3, 0.5
Image Compression	Quality Factor	50, 70, 90

The attacks using these image processing operations are implemented using MATLAB2023a. These operations are applied on the collage image (viz., the image with hidden secret image) using the functions in the Image Processing Toolbox, which includes the *imnoise* [17] function, *medfilt2* [18] function, *imgausfilt* [19] function, and *imsharpen* [20] function.

For each operation, two to three different variance levels are added to the images. For the details of the variants, please refer to Table II.

B. Secret Recovery

The decoding of the collage image will be performed using the trained decoder of the StArtGAN to recover the secret classes. The result of classifying the attacked collage images is shown in Table III.

In terms of Salt & Pepper Noise, it can almost perfectly recover the secret, which is up to 99% when the added noise intensity is as low as 0.01. This shows that the collage images with added Salt & Pepper noises with low intensity do not have much interference with the classification of the collage images for secret recovery since the Salt & Pepper noises are distributed in a sparse and tiny manner. However, when the noise intensity increases, it will consequently reduce the secret recovery. However, the secret recovery still can achieve an average above 92% even when the noise increases to 0.05.

Regarding the Gaussian noise, the secret recovery is lower compared to the added Salt & Pepper Noise. It achieves around 88% secret recovery for the intensity of the 0.01 added noise

TABLE III: Classification accuracy of attacked images [normalized to $[0, 1]$].

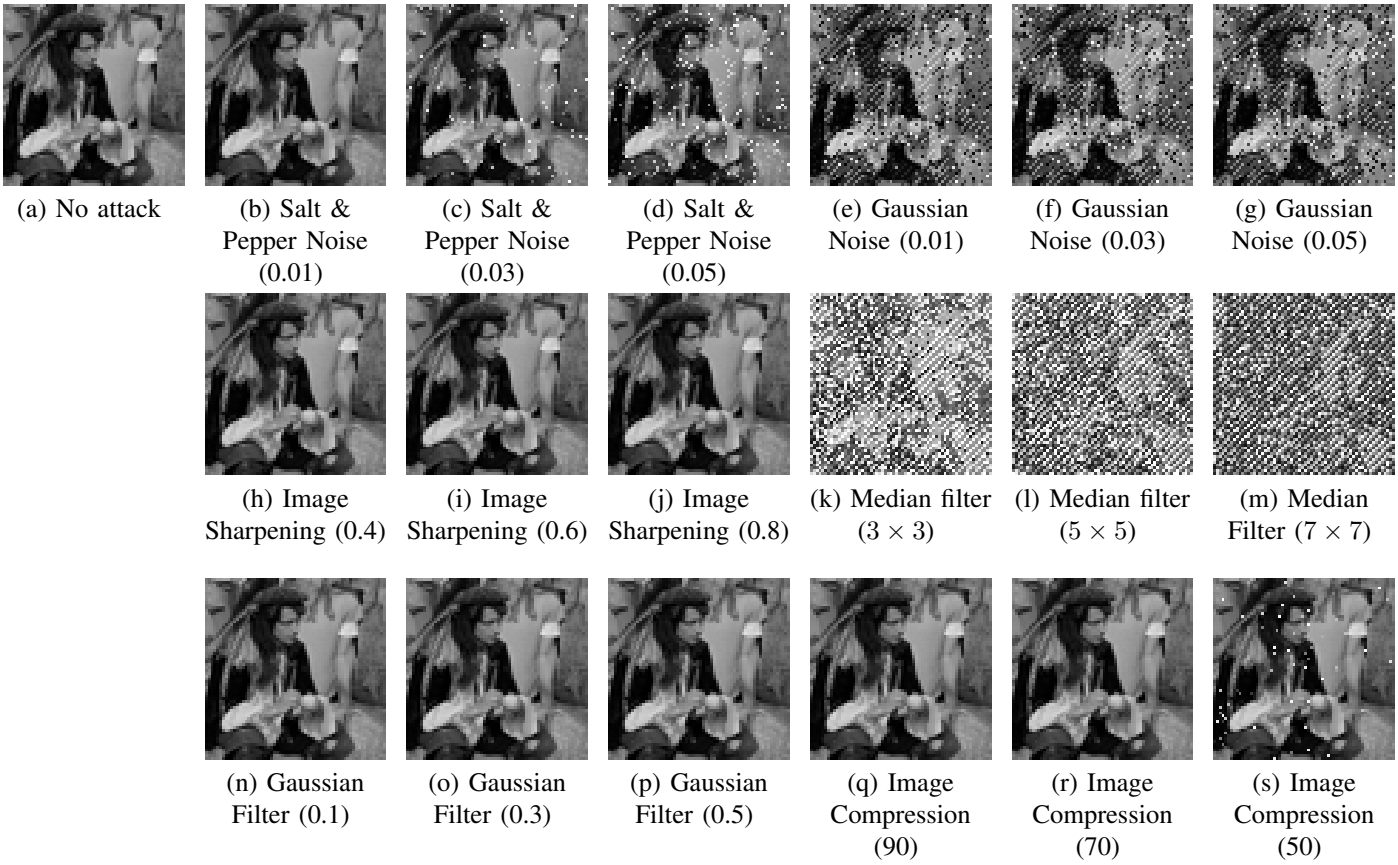
Image Processing Attacks	Variant	Secret Images									
		Aerial	Baboon	Boat	Cameraman	Clock	Lena	Pepper	Pirate	Plan	Woman
No Attack	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Salt & Pepper Noise	0.01	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	0.03	0.980	0.980	0.974	0.978	0.982	0.978	0.980	0.978	0.982	0.979
	0.05	0.942	0.937	0.922	0.935	0.942	0.930	0.929	0.932	0.942	0.932
Gaussian Noise	0.01	0.873	0.865	0.839	0.862	0.872	0.861	0.861	0.849	0.877	0.858
	0.03	0.871	0.863	0.837	0.859	0.869	0.859	0.858	0.847	0.872	0.857
	0.05	0.868	0.859	0.835	0.859	0.867	0.854	0.857	0.845	0.868	0.854
Sharpening	0.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Median Filter	3×3	0.655	0.639	0.634	0.626	0.665	0.629	0.639	0.582	0.658	0.669
	5×5	0.278	0.286	0.299	0.278	0.294	0.283	0.279	0.279	0.277	0.286
	7×7	0.159	0.169	0.188	0.168	0.160	0.171	0.169	0.180	0.157	0.173
Gaussian Filter	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Compression	$QF = 50$	0.997	0.998	0.997	0.995	0.998	0.996	0.996	0.993	0.997	0.997
	$QF = 70$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$QF = 90$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



(a) Original Collage Image
(b) Salt & Pepper Noise (0.01) [24.81, 0.93]
(c) Salt & Pepper Noise (0.03) [20.19, 0.81]
(d) Salt & Pepper Noise (0.05) [17.94, 0.72]
(e) Gaussian Noise (0.01) [20.41, 0.75]
(f) Gaussian Noise (0.03) [20.06, 0.74]
(g) Gaussian Noise (0.05) [19.47, 0.73]
(h) Image Sharpening (0.4) [28.68, 0.98]
(i) Image Sharpening (0.6) [25.60, 0.96]
(j) Image Sharpening (0.8) [23.51, 0.93]
(k) Median filter (3×3) [18.32, 0.65]
(l) Median filter (5×5) [16.30, 0.48]
(m) Median Filter (7×7) [14.89, 0.37]
(n) Gaussian Filter (0.1) [Inf, 1.00]
(o) Gaussian Filter (0.3) [54.63, 0.99]
(p) Gaussian Filter (0.5) [27.11, 0.96]
(q) Image Compression (90) [36.68, 0.99]
(r) Image Compression (70) [28.31, 0.95]
(s) Image Compression (50) [25.36, 0.90]

Fig. 2: Generated collage image samples of the original "Pirate" image after image processing is performed. The image quality of the images is indicated in the subfigure caption in the format of [PSNR[dB], SSIM].

since the added noise in a normal distribution will change the most pixels of the original collage image. When the intensity of the added noise increases, the effect of the wrong classification also slightly increases. Generally, the classification accuracy



(a) No attack (b) Salt & Pepper Noise (0.01) (c) Salt & Pepper Noise (0.03) (d) Salt & Pepper Noise (0.05) (e) Gaussian Noise (0.01) (f) Gaussian Noise (0.03) (g) Gaussian Noise (0.05)

(h) Image Sharpening (0.4) (i) Image Sharpening (0.6) (j) Image Sharpening (0.8) (k) Median filter (3×3) (l) Median filter (5×5) (m) Median Filter (7×7)

(n) Gaussian Filter (0.1) (o) Gaussian Filter (0.3) (p) Gaussian Filter (0.5) (q) Image Compression (90) (r) Image Compression (70) (s) Image Compression (50)

Fig. 3: Extracted secret image after various image processing attacks.

still can be maintained above 80%.

When sharpening is performed on the images, the collage images can fully recover the secret class regardless of the applied strength. This shows that sharpening the image helps emphasize the edges and image features, and would not degrade the image quality, thus ensuring the collage images are always classified correctly.

On the other hand, the median filter applied to the collage images causes a smoothening effect on them, especially on the edges, making the images unrecognizable by the decoder. Thus, the secret recovery is the lowest among all these attacks, which is below 70% secret recovery, even with the smallest window size of 3×3 is applied. It shows a drastic drop in secret recovery accuracy when the window size increases to 5×5 and 7×7, which is below 30% secret recovery.

The Gaussian filter preserves edges better compared to median filtering. Thus, the Gaussian filter applied to the images does not have a huge effect on the collage image where it shows perfect recovery.

Lastly, a compression attack has also been performed on the collage image. From the results, the attack does not show much impact when the quality factor (QF) is 70 and above. When it comes to $QF = 50$, it shows a slight error in secret recovery, which the error is below 1%.

All in all, the generated collage images using the StArtGAN framework are able to resist Salt & Pepper Noise, Gaussian Noise, Sharpening, Gaussian Filter, and Compression Attacks. However, it has the least resistance to median filtering since the smoothen effects that remove the edges and important features from images will cause misclassification of the decoder.

C. Image Quality

1) Image Quality of the Collage Image: One of the generated 8×8 original collage images and collage images after the application of image processing of size are shown in Fig. 2. From the figure, we can see that the Salt & Pepper Noise does not have too much effect on the collage images, and the noise is distributed in a sparse manner. Higher intensity of noise will have a more obvious effect on the collage image. Even though the noises are visible from the collage image, we can still see the underlying original images. However, when the noise intensities increase, the PSNR and SSIM scores reduce, showing more intense effects on the image.

Besides, the application of Gaussian Noise in a normal distribution causes manipulation in most of the image pixels. When it comes to Gaussian Noise, a more obvious effect is observed from the images. When the intensity increases, the visual effect slightly becomes more intense. Similarly, the PSNR and SSIM scores are likely the same with a slight dropping, indicating the increase in noise intensity does not

TABLE IV: PSNR values [dB] of the extracted secret image after attacks.

Image Processing Attacks	Variants	Secret Images									
		Aerial	Baboon	Boat	Cameraman	Clock	Lena	Pepper	Pirate	Plan	Woman
Salt & Pepper Noise	0.01	54.084	43.242	48.088	54.109	65.224	71.244	42.038	66.193	71.244	44.589
	0.03	27.480	26.082	26.549	27.415	28.823	24.848	26.153	24.511	24.848	27.682
	0.05	20.777	20.251	21.074	22.168	24.556	19.581	19.875	20.147	19.581	21.560
Gaussian Noise	0.01	17.363	17.539	16.749	17.539	17.304	16.874	17.034	17.021	16.874	17.061
	0.03	17.346	17.269	16.609	17.618	17.429	16.695	16.850	17.007	16.695	17.213
	0.05	17.206	17.317	16.739	17.655	17.293	16.769	16.993	16.854	16.769	17.034
Sharpening	0.4	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	0.6	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	0.8	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
Median Filter	3×3	12.034	11.144	10.762	10.093	13.783	10.653	10.653	8.747	10.653	12.097
	5×5	10.432	10.551	10.301	9.265	8.903	9.849	9.583	8.882	9.849	10.365
	7×7	10.326	11.048	10.422	9.312	7.132	10.172	9.714	9.251	10.172	10.432
Gaussian Filter	0.1	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	0.3	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	0.5	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
Compression	$QF = 50$	31.986	32.689	28.156	23.975	33.519	30.152	27.323	23.069	31.923	31.012
	$QF = 70$	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	$QF = 90$	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞

TABLE V: SSIM values of the extracted secret image after attacks.

Image Processing Attacks	Variants	Secret Images									
		Aerial	Baboon	Boat	Cameraman	Clock	Lena	Pepper	Pirate	Plan	Woman
Salt & Pepper Noise	0.01	0.999	0.995	0.996	0.999	0.999	0.999	0.998	0.999	0.999	0.999
	0.03	0.946	0.837	0.833	0.840	0.919	0.856	0.904	0.878	0.856	0.908
	0.05	0.789	0.615	0.636	0.659	0.837	0.648	0.718	0.726	0.648	0.708
Gaussian Noise	0.01	0.623	0.393	0.363	0.384	0.352	0.450	0.525	0.519	0.450	0.415
	0.03	0.618	0.368	0.360	0.389	0.361	0.431	0.512	0.523	0.431	0.418
	0.05	0.615	0.370	0.366	0.391	0.354	0.441	0.523	0.519	0.442	0.409
Sharpening	0.4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Median Filter	3×3	0.262	0.100	0.109	0.129	0.187	0.167	0.204	0.145	0.167	0.150
	5×5	0.067	0.034	0.032	0.046	0.047	0.045	0.056	0.083	0.044	0.026
	7×7	0.068	0.054	0.026	0.019	0.012	0.029	0.028	0.051	0.029	0.028
Gaussian Filter	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Compression	$QF = 50$	0.979	0.963	0.929	0.865	0.973	0.957	0.937	0.877	0.972	0.957
	$QF = 70$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$QF = 90$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

have a huge effect on the image.

Sharpening of strength 0.4, 0.6, and 0.8 are applied to the collage image. This effect enhances the edges and fine details to ensure the collage images have sharper image quality. The lower strength in sharpening can preserve the original image better, thus maintaining higher SSIM and PSNR scores.

In terms of the median filter, when the window sizes increase, the smoothening effect on the collage image becomes more obvious, and the edges of the images are less visible. It makes the collage image very different from the original collage image. By using a smaller neighborhood size for median filtering, the images will have less distortion and thus have higher PSNR and SSIM scores.

Gaussian Filter can completely maintain the quality of the original image when the Gaussian filter with a 0.1 S.D. is applied. Even when S.D. is increased to 0.3, it can still preserve the image quality with high PSNR and SSIM. However, when it comes to an S.D. of 0.5, the image quality significantly drops.

Lastly, in terms of Compression, the images can still preserve the image quality with high PSNR and SSIM scores. Hence, the effect of compression does not have much effect on the image quality.

In short, among all these methods, the Salt & Pepper Noise, Gaussian Noise, Image Sharpening, Gaussian Filter and Compression applied on images do not cause huge degradation in the image quality. On the other hand, when the median filter is applied, the loss in edges and lines of images will cause image distortion, causing low PSNR and SSIM scores.

2) Image Quality of the Recovered Image: The collage image will be decoded into their corresponding secret class and finally recover the secret image. Figure 3 shows the secret images before the attack and outputs of the secret image regenerations after the image processing is applied. From the recovered image, the effect of the image processing can also be

observed. In terms of Salt & Pepper, when the noise intensity increases, the effect seems to be more obvious and the images become harder to visible. Besides, for the Gaussian Noise with different intensities, the contents of the images are likely almost invisible. Besides, for the median filter, the content of the recovered images is totally unrecognizable. The image with Sharpening, Gaussian Filter, and Image Compression with a quality factor of 90 and 70, the recovered secret image is highly similar to the original image. Lastly, in image compression with a quality factor of 50, distortion is visible in the images as compared to the other quality factor.

The PSNR and SSIM scores between the secret image with the recovered image are evaluated and shown in Table IV and V. The results show that the secret image can be fully recovered after sharpening, a Gaussian filter, or compression with a quality factor of above 70 is applied. On the other hand, the images still have good performance in preserving the secret image after the Salt & Pepper attack and Compression attack with a quality factor of 70 is performed. Gaussian Noise has a lower recovery of the original image. Lastly, a median filter attack can preserve the lowest image quality among all these methods. From the result, low PSNR and SSIM scores are shown for the median filter with various neighborhood sizes.

D. Embedding Capacity and other Data Hiding Properties

In this experiment, the size of the secret image is 64×64 pixels, and they are converted into 8-bit grayscale images. Therefore, the total capacity that can be hidden in the collage image is 32,769 bits. However, StArtGAN is a generative data hiding model that can output collage images of various sizes in accommodating various capacities. For more information on the flexible embedding capacity and performance of other data hiding properties in StArtGAN when compared with relevant works, we refer the readers to [14].

IV. Discussion

From the result, the images that are applied with median filtering attacks are hard to identify. The smoothening effect performed on the images has reduced the visibility of the edges and thus, the decoder is unable to identify the image content. This consequently leads to the misclassification of the collage images, and finally the inability to recover the secret image. Sharpening will emphasize the image content, thus making the secret recovery can be performed easier. The compression performed on the images does not have much impact on degrading the image quality unless the quality factor achieves up to 50. For the Gaussian filter, the image quality can still be preserved even when the standard deviation achieves up to 0.5.

All in all, the experiment results indicate that StArtGAN can resist image processing attacks, especially in noise, sharpening, and image compression operations. It is weak against the median filter because it blurs the art images, and weakens the distinctive feature of each class. Low classification accuracy then severely degrades the image quality of the extracted secret image.

For non-constructive-based data hiding methods, most of them couldn't be able to resist image processing attacks due to the special formula that they used in selecting the images to represent secret data. When the pixel values change by the image processing operations, it affects the formula calculation and hence might cause wrong secret data extraction.

For constructive-based data hiding methods, patterns and texture image generation should be able to achieve similar performance as StArtGAN because the secret data is usually embedded on patch level ($w \times w$ pixels per unit) but not on pixel level (one pixel per unit). Therefore, secret data can still be accurately extracted if the majority of the pixels or structures stay intact.

V. Conclusion

In conclusion, the results show the StArtGAN [14] method can resist various image processing operations such as the Salt & Pepper noise, Gaussian noise, Sharpening, Gaussian Filter, and Compression attacks to classify the collage image using the trained decoder and to recover the original images. On the other hand, a median filtering attack smoothens the edges of images, consequently leading to the misclassification of collage images, causing low recovery of the original image.

Acknowledgment

This publication presents the outcome of the undergoing project, entitled Steganography by Image Synthesis: A Novel Model for Coverless Data Hiding, supported by Fundamental Research Grant Scheme (FRGS) MoHE Grant of ID FP053-2019A (Reference Code: FRGS/1/2019/ICT02/UM/02/1) awarded by the Malaysian Government.

References

[1] K. Y. Ng, S. Ong, and K. Wong, "Delving into the methods of coverless image steganography," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1763–1772, 2019.

[2] Z. Zhou, H. Sun, R. Harit, X. Chen, and X. Sun, "Coverless image steganography without embedding," in *Cloud Computing and Security* (Z. Huang, X. Sun, J. Luo, and J. Wang, eds.), (Cham), pp. 123–132, Springer International Publishing, 2015.

[3] J. Wu, Y. Liu, Z. Dai, Z. Kang, S. Rahbar, and Y. Jia, "A coverless information hiding algorithm based on grayscale gradient co-occurrence matrix," *IETE Technical Review*, vol. 35, no. sup1, pp. 23–33, 2018.

[4] H. Otori and S. Kuriyama, "Data-embeddable texture synthesis," in *Smart Graphics* (A. Butz, B. Fisher, A. Krüger, P. Olivier, and S. Owada, eds.), (Berlin, Heidelberg), pp. 146–157, Springer Berlin Heidelberg, 2007.

[5] H. Otori and S. Kuriyama, "Texture synthesis for mobile data communications," *IEEE Computer Graphics and Applications*, vol. 29, no. 6, pp. 74–81, 2009.

[6] K.-C. Wu and C.-M. Wang, "Steganography using reversible texture synthesis," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 130–139, 2015.

[7] Z. Qian, H. Zhou, W. Zhang, and X. Zhang, "Robust steganography using texture synthesis," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, pp. 25–33, Springer, 2017.

[8] W. K. Lee, S. Ong, K. Wong, and K. Tanaka, "A novel coverless information hiding technique using pattern image synthesis," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1122–1127, IEEE, 2018.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, p. 139–144, oct 2020.

[11] D. Hu, L. Wang, W. Jiang, S. Zheng, and B. Li, "A novel image steganography method via deep convolutional generative adversarial networks," *IEEE Access*, vol. 6, pp. 38303–38314, 2018.

[12] H. Naito and Q. Zhao, "A new steganography method based on generative adversarial networks," in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pp. 1–6, 2019.

[13] W. F. . C. X. Fu, Z., "The secure steganography for hiding images via gan," *EURASIP Journal on Image and Video Processing*, pp. 1687–5281, 2020.

[14] K. Y. Ng, S. Ong, Y. P. Loh, and C. S. Chan, "Relabel, scramble, synthesize:a novel coverless steganography approach via collage image," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1877–1882, 2021.

[15] "The usc-sipi image database," *http://sipi. usc. edu/database/*, 2006.

[16] W. R. Tan, C. S. Chan, H. Aguirre, and K. Tanaka, "Improved artgan for conditional synthesis of natural image and artwork," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 394–409, 2019.

[17] MATLABR2023a, "Image Processing Toolbox - imnoise function." https://www.mathworks.com/help/images/ref/imnoise.html, 2023. [Online; accessed 21-Jun-2023].

[18] MATLABR2023a, "Image Processing Toolbox - 2-D Median Filtering." https://www.mathworks.com/help/images/ref/medfilt2.html, 2023. [Online; accessed 21-Jun-2023].

[19] MATLABR2023a, "Image Processing Toolbox - 2-D Gaussian Filtering." https://www.mathworks.com/help/images/ref/imgaussfilt.html, 2023. [Online; accessed 21-Jun-2023].

[20] MATLABR2023a, "Image Processing Toolbox - Sharpen Image using Unsharp Masking." https://www.mathworks.com/help/images/ref/imsharpen.html, 2023. [Online; accessed 21-Jun-2023].

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition

Chang Wang*, Jun Du✉*, Hang Chen*, Ruoyu Wang*, Chao-Han Huck Yang†,
Jiangjiang Zhao‡, Yuling Ren‡, Qinglong Li‡ and Chin-Hui Lee†
* University of Science and Technology of China, Hefei, Anhui, PR China
† Georgia Institute of Technology, Atlanta, GA, USA
‡ China Mobile Online Services Company Limited, China

***Abstract*—In this paper, we investigate the effectiveness of using quantum machine learning for privacy protection in audio-visual speech processing. Quantum machine learning has made significant theoretical advancements and has been shown to possess natural advantages in privacy protection over conventional techniques. Here, we first apply quantum circuits to a word-level audio-visual speech recognition task. We then propose a novel metric, an inter-class intra-class similarity ratio, for measuring the privacy-protecting capabilities of quantum circuits. Finally, we conduct an in-depth analysis of the differences in privacy protection between quantum privacy methods and traditional methods, evaluating their working principles, strengths, and limitations. Experiments results on the LRW data set show that the quantum privacy-preserving approach performs well in word-level speech recognition tasks, demonstrating excellent privacy-preserving capabilities through selective retention of features.**

I. Introduction

With the rapid development of deep learning, face recognition, automatic speech recognition and audio-visual technology have been gradually applied in daily life scenarios in recent years. The sensitivity of face image data and speech data has led to growing concerns about data privacy. In this context, it is crucial to develop effective methods for protecting audio and video privatized data to comply with new privacy protection regulations, e.g. the EU General Data Protection Regulation (GDPR) [1].

There are many traditional algorithms in the field of privacy protection, such as homomorphic encryption [2] and differential privacy [3], [4]. Homomorphic encryption is a cryptographic technique based on the computational complexity theory of mathematical puzzles. This algorithm offers the property that the result of performing an operation on plaintext and then encrypting it is equivalent to the result of performing the corresponding operation on the ciphertext after encryption. Due to this desirable feature, it can maintain high recognition accuracy while protecting data privacy. However, its application scenarios are greatly limited due to the large amount of additional computation required for the encryption process. Differential privacy methods add specific noise to the input data to prevent attackers from inferring sensitive information based on the output. However, in practical applications, it is often caught in the dilemma of trade-off between privacy-preserving capability and downstream task accuracy. To address the above challenges, [5] proposed a differential privacy method based on frequency domain transform and learnable privacy budget, which demonstrated outstanding results in face recognition tasks.

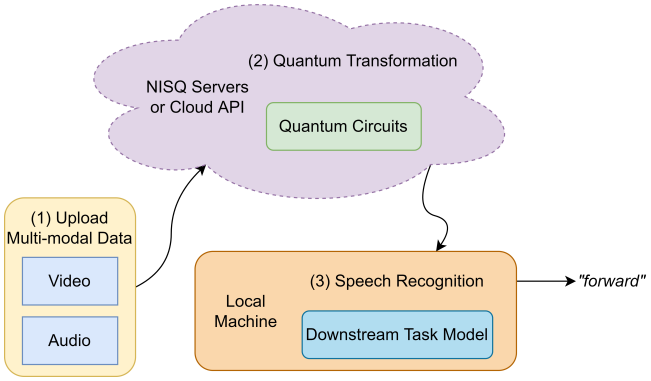


Fig. 1: Quantum privacy-preserving architecture for speech recognition task in a vertical federated learning process. Quantum circuits on the Noisy Intermediate-Scale Quantum (NISQ) server transform the input multi-modal data into quantum private data and transfer it to the speech recognition model on local machine.

In recent years, federated learning has also emerged as a common privacy-preserving approach. For example, vertical federated learning (VFL) [6] is a potential data protection strategy that decentralizes the end-to-end deep learning framework and separates feature extraction from the downstream task system. And with the recent advances in commercial quantum technologies, quantum machine learning (QML) [7] becomes an ideal building block for VFL due to its parameter encryption and data isolation advantages. The distinction between quantum machine learning and traditional machine learning lies in the utilization of quantum circuits. These circuits consist of a series of quantum gates that operate on quantum bits (qubits), which are the fundamental units of

✉ : Corresponding author

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

quantum information. The VFL architecture using quantum circuits as a building block are shown in Fig. 1. Based on the above, Yang et al. [8] proposed a decentralized acoustic modeling scheme and designed a quantum convolutional neural network, resulting in impressive outcomes by combining the learning paradigm of variable quantum circuit [7] and deep neural network.

The contributions of this paper are as follows:

• Demonstrates the effectiveness of quantum circuits-based methods for privacy preservation in a word-level audio-visual speech recognition task.

• Proposes the inter-class intra-class similarity ratio as a privacy protection effect metric, and use it to thoroughly analyze the differences between quantum privacy methods and traditional methods.

II. RELATED WORK

A. Quantum Machine Learning for Signal Processing

The current phase of quantum computing is often referred to as the noisy intermediate-scale quantum era. During this era, quantum processors consist of intermediate-scale qubits, which are not yet advanced enough to achieve fault-tolerance or reach the threshold for quantum supremacy. These processors are sensitive to external factors (noisy) and susceptible to quantum decoherence and not capable of continuous quantum error correction, which limits the potential applications of NISQ technology. However, Mitarai *et al.* [7] build a pioneering framework for building machine learning models on NISQ devices by the use of variable quantum circuit (VQC) [9]. VQC can be progressively iteratively optimized, allowing noise effects in NISQ devices to potentially be absorbed into these learned circuit parameters. Several successful machine learning applications based on VQC have been reported in the recent literature, such as transfer learning [10] and deep reinforcement learning [11]. It is worth noting that recent studies [12], [13] have shown that quantum machine learning (QML) has advantages over classical machine learning in terms of lower memory storage, secure encryption of model parameters, and good feature representation capabilities.

While still a nascent technology, quantum machine learning has already been applied in various fields. For instance, [14] successfully implemented image recognition using quantum deep convolutional networks, [15] proposed a speech recognition system using quantum backpropagation (QBP) simulated by fuzzy logic computation, and [16] successfully applied quantum transfer learning techniques to the synthetic speech detection task.

B. Privacy Preservation

In the realm of privacy protection, research can be broadly classified into three categories, depending on how the input data is processed: data anonymization methods, data encryption methods, and data perturbation methods. Data anonymization methods, such as k-anonymity [17] and t-closeness [18], work by desensitizing sensitive information to protect privacy. While these methods allow data to be analyzed and utilized to some extent, they may also undermine the integrity and accuracy of the data, reducing its validity. Homomorphic encryption and secure multi-party computation [19] are classical data encryption methods that effectively protect data security while achieving excellent results on downstream tasks. However, the high computational costs associated with encryption and decryption make these methods difficult to apply in practice. In contrast, data perturbation methods avoid the high costs of encryption by perturbing the original input. Differential privacy is a widely used data perturbation method. [20] employed differential privacy to perturb eigenfaces and distributed the privacy budget equally to each eigenface, which resulted in a significant loss of accuracy. Besides, [21] uses the Mixup [22] method to perturb the data, and [23] proposed a Gaussian noise perturbation method to suppress the unimportant pixels before sending them to the cloud. Recent studies [24] have started to investigate how to incorporate differential privacy into speech processing, but these efforts have not extended into the multi-modal learning setting. In summary, privacy concerns regarding audio-visual patterns deserve further investigation.

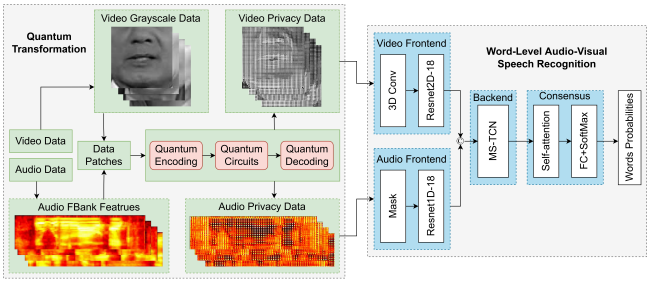


Fig. 2: Pipeline of quantum privacy-preserving method for word-level audio-visual speech recognition. © denotes the concatenate operation.

C. Word-Level Audio-Visual Speech Recognition

Traditional Audio-Visual Speech Recognition (AVSR) methods follow a two-step process that involves feature extraction and recognition [25]. However, recent advancements have significantly improved the performance of Visual Speech Recognition (VSR) and Automatic Speech Recognition (ASR) by combining feature extraction and recognition inside deep neural networks. In VSR, [26] proposed an end-to-end network called Visual to Phoneme. [27] proposed Spatio-Temporal Fusion Module to maintain the local spatial information and reduce the feature dimensions. [28] designs a novel deep learning architecture using hierarchical pyramidal convolution and self-attention, thereby enhancing the model's ability to discover fine-grained lip movements. In ASR, [29] demonstrated the superiority of deep representations of the network over hand-crafted features such as filterbank features.

III. METHOD

A. Audio-Visual Quantum Privacy Protection

Figure 2 illustrates the pipeline of the audio-visual quantum privacy protection method. It comprises two modules: quantum

transformation and word-level audio-visual speech recognition. In the quantum transformation module, we pre-process the audio-video data first. For audio data, we extract its FBank features; for video data, we read its grayscale image and split it into different frames for subsequent processing. Next, we chunk the processed audio and video data into patches and feed them into a quantum gate for quantum encoding into the initial quantum states. We then perform further quantum transformations using quantum circuits. Finally, the quantum state is decoded by a measurement function to output quantum privatized data.

For the word-level audio-visual speech recognition task, referring to the system in [28], [30], our system utilizes a 3D convolutional layer and an 18-layer residual network (ResNet-18) [31] as the frontend for the video modality, a mask module and a ResNet-18 as the frontend for the audio modality. The features extracted by the two frontends are concatenated and fed into a multi-scale temporal convolutional network (MS-TCN) backend, which then goes through the self-attention based consensus module to output the words probabilities.

B. Quantum Gates

In this section, we briefly introduce several basic quantum gates: R_x, R_y, R_z, and $CNOT$. Unlike a classical bit, which has two states (0 and 1), a quantum bit can be represented as $|\phi\rangle = \alpha|0\rangle + \beta|0\rangle, |\alpha|^2 + |\beta|^2 = 1$. We can also express it in the form $|\phi\rangle = \cos\theta|0\rangle + e^{-i\phi}\sin\theta|1\rangle, \theta, \phi \in R$. The density matrix of this quantum bit is:

$$\rho = |\varphi\rangle\langle\varphi| = \begin{bmatrix} \cos^2\frac{\theta}{2} & e^{-i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} & \sin^2\frac{\theta}{2} \end{bmatrix}. \quad (1)$$

$\langle\phi|$ represents the conjugate transpose of $|\phi\rangle$.

Alternatively, we can construct the density matrix space using the four Pauli matrices as its basis, which is:

$$I = \sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_x = \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$
$$\sigma_y = \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_z = \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Any matrix can be decomposed into $\rho = \sum_{i=0}^{4} a_i\sigma_i$. Considering $Tr\rho = \sum_{i=0}^{4} a_i Tr\sigma_i$, we can determine that $a_0 = 1/2$. Therefore, the density matrix can be expressed as:

$$\rho = \frac{1}{2}(I + \vec{r}\cdot\sigma) = \rho(\vec{r}),$$
$$\vec{r} = 2\,(a_1, a_2, a_3) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta).$$

Substituted the expression of $\vec{r}$ into the density matrix yields:

$$\rho = \frac{1}{2}\left(I + \begin{bmatrix} \cos\theta & e^{-i\phi}\sin\theta \\ e^{-i\phi}\sin\theta & -\cos\theta \end{bmatrix}\right)$$
$$= \begin{bmatrix} \cos^2\frac{\theta}{2} & e^{-i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} & \sin^2\frac{\theta}{2} \end{bmatrix}.$$

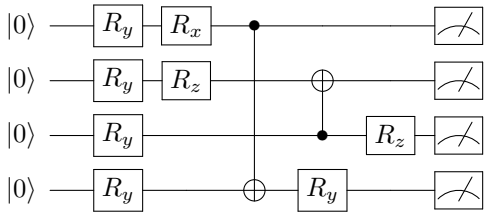


Fig. 3: An example of random quantum circuit.

This is equal to Equation 1, thus we establish the mapping relation from the quantum state to the 3D sphere, commonly known as the Bloch sphere [32].

Next, we define the quantum gates R_x, R_y, R_z to mean the representation of the quantum bit in the Bloch sphere rotating around the x, y, z axes respectively. It is then straightforward to determine their corresponding matrices as:

$$R_X(\theta) = \begin{bmatrix} \cos\frac{\theta}{2} & -i\sin\frac{\theta}{2} \\ -i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix}, R_Y(\theta) = \begin{bmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix},$$
$$R_X(\theta) = \begin{bmatrix} e^{-i\frac{\theta}{2}} & 0 \\ 0 & e^{-i\frac{\theta}{2}} \end{bmatrix}.$$

Similarly, with reference to the Exclusive-OR gate in classical logic gates, we can express the $CNOT$ gate as a matrix:

$$CNOT = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

C. Quantum Circuits

For better privacy protection effect, the structure and parameters of quantum circuits are randomly generated. Fig. 3 gives a diagram of a random quantum circuit, where only the most basic quantum gates R_x, R_y, R_z, and $CNOT$ are applied. The classical vectors are initially encoded into a quantum state $|\phi\rangle = |0000\rangle$, and the encoded quantum states go through the quantum circuit for the following steps:

Step 1: $\Phi_1 = R_y|0\rangle R_y|0\rangle R_y|0\rangle R_y|0\rangle$.
Step 2: $\Phi_2 = R_x R_y|0\rangle R_z R_y|0\rangle R_y|0\rangle R_y|0\rangle$.
Step 3: $\Phi_3 = R_x R_y|0\rangle R_z R_y|0\rangle R_y|0\rangle CNOT\,(R_y|0\rangle)$.
Step 4: $\Phi_4 = R_x R_y|0\rangle CNOT\,(R_z R_y|0\rangle)\,R_y|0\rangle R_y\,\mathrm{CNOT}\,(R_y|0\rangle)$.
Step 5: $\Phi_5 = R_x R_y|0\rangle CNOT\,(R_z R_y|0\rangle)\,R_z R_y|0\rangle R_y CNOT\,(R_y|0\rangle)$.

Furthermore, due to the potential for many unexpected noisy signals from $CNOT$ gates, particularly with our current non-error-corrected quantum devices and the connectivity of physical qubits, we limit the number of qubits in our random quantum circuit to ensure it falls within the noise tolerance capabilities of VQC.

IV. Privacy Protection Capability Measurement

A. Inter-class Intra-class Similarity Ratio

To quantitatively analyze the privacy protection capability of different methods, we introduce a new metric called the inter-

class intra-class similarity ratio (IISR).

To process video data, we group frames from the same video together as one class, while frames from different videos are treated as separate classes. Audio data is handled in a similar fashion. Suppose there are n video segments in total, and each video segment has m frames, the j-th frame image of the i-th video segment is $\mathrm{fig}_{i,j}$.

First, we utilize SSIM as the measure of the similarity between data. From there, we can calculate the intra-class similarity by employing the following formula:

$$\mathrm{ssim}_{\mathrm{intra}} = \frac{1}{n * m * (m-1)} \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right).$$

Similarly, the inter-class similarity function is:

$$\mathrm{ssim}_{\mathrm{inter}} = \frac{1}{n * (n-1) * m^2} \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right).$$

To calculate the IISR of SSIM, we can express it as the ratio of inter-class similarity to intra-class similarity:

$$\mathrm{IISR}_{\mathrm{ssim}} = \frac{(m-1) \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right)}{m * (n-1) \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right)}.$$

Similarly, the formulas for the IISR of RMSE and PSNR are:

$$\mathrm{IISR}_{\mathrm{rmse}} = \frac{(m-1) \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} 1/\mathrm{rmse}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right)}{m * (n-1) \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} 1/\mathrm{rmse}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right)}.$$

$$\mathrm{IISR}_{\mathrm{psnr}} = \frac{(m-1) \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} \mathrm{psnr}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right)}{m * (n-1) \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \mathrm{psnr}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right)}.$$

Inter-class and intra-class similarities respectively denote the proportion of shared information contained within inter-class and intra-class data. Greater similarity indicates more compact data. The IISR represents the proportion of common information contained within inter-class and intra-class data. A smaller ratio implies that the intra-class data is more compact than the inter-class data, making it easier to differentiate whether the two data belong to the same class, which in turn means more information is preserved.

B. Privacy Attack

To assess the effectiveness of privacy protection methods, privacy attacks [24] are often employed. It can be broadly classified into white-box attacks and black-box attacks.

1) White-box attack: White-box attacker has access to all of our operations. With this knowledge, in order to attack the quantum privacy protection method, they can take all parameters in the quantum circuits and inverse them to get inverse quantum circuits, and then feed the quantum privatized data into the inverse quantum circuits to get the output.

2) Black-box attack: A black-box attack assumes that the attacker does not possess any knowledge regarding the model's internal structure and parameters. Nevertheless, attackers can collect large amounts of audio and video data and input the data into the model to obtain processed data. Subsequently, they can train a decoder to restore the processed results to the original input. Finally, attackers can leverage the trained decoder to recover users' private data. For our subsequent experiments, we will use UNet [33] as our decoder to reconstruct the original data from the processed data.

V. Experiments

A. Experimental Setup

To evaluate the impact of two privacy-preserving methods on downstream tasks, we selected a moderately challenging yet feasible task: word-level speech recognition. For the dataset, we use the Lip Reading in the Wild (LRW) [34], which consists of about 500,000 audio-visual speech segments extracted from BBC TV broadcasts. It encompasses 500 target words and numerous speakers, and provides a word-level label for each audio-visual speech segment.

To account for practical application factors, we introduce noise to the original speech. A total of 115 noise types including 100 noise types from [35] and 15 homemade noise types corrupted the corpus in the training set at 5 SNR levels (i.e., 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB). The validation and test sets were corrupted by three unseen noise types at the aforementioned SNR levels, i.e., Speech Babble, Buccaneer1, and Destroyer Engine. All unseen noises were collected from the NOISEX92 corpus [36]. Following the addition of noise, we extracted its 40-dim Fbank features from the noisy speech as the input. Regarding the video modality, we utilize grayscale images directly as input.

We employed the method proposed by [5] as the baseline for privacy protection. In the frequency domain transform part, we use some functions in TorchJPEG [37] to transform the input data into BDCT coefficients and then remove its direct part. And for the initial value of the learnable budget allocation parameter, we set it to 0 so that the privacy budget of each location is equal in the initial stage.

For the quantum privacy-preserving method, we use Pennylane to build a 4-bits random quantum circuits and divide the input data into 2×2 patches into the quantum circuits to generate quantum privatized data.

B. Performance of Word-level Audio-visual Speech Recognition

We conducted a comparative analysis of the classification accuracy for speech recognition task using original data, baseline

TABLE I: Speech recognition accuracy(Acc) in different data types and modalities.

Acc	Original	Baseline	**Quantum**
ASR	69.00	65.92	**68.12**
VSR	86.43	84.23	**85.42**
AVSR	89.23	87.13	**88.26**

privatized data, and quantum privatized data, respectively. The results are shown in Table I.

Based on the table, it is apparent that the classification accuracy of the original data is the highest, irrespective of the modality. This implies that both privacy-preserving methods lead to some information loss. Nonetheless, we observed that the training results of the quantum privatized data are superior to the baseline approach, suggesting that the quantum privacy-preserving method loses less information for the downstream task required.

C. privatized data Analysis

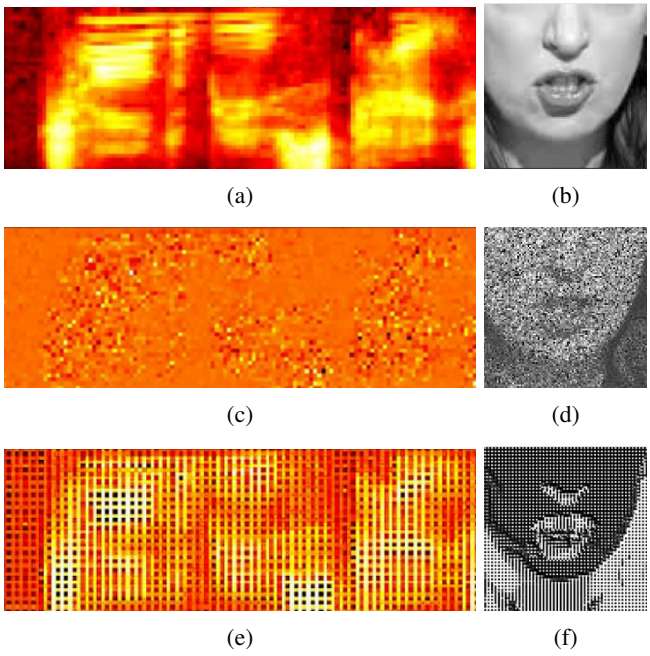


Fig. 4: Visualization of privatized data. (a) Spectrogram of original audio data; (b) Grayscale plot of original video data; (c) Spectrogram of audio baseline privatized data; (d) Grayscale plot of video baseline privatized data; (e) Spectrogram of audio quantum privatized data; (f) Grayscale plot of video quantum privatized data.

As the first step of our analysis, we performed visualization operations on the dataset. Specifically, We randomly select an audio-video segments from the original dataset, generated the grayscale map of the first frame of the video and the audio spectrogram, and repeated the same visualization procedure for the baseline privatized data and quantum privatized data

TABLE II: privatized data analysis.

Modality	Data_type	IISR_{rmse}	IISR_{psnr}	IISR_{ssim}
Video	Original	0.314	0.611	0.491
	Baseline	0.532	0.741	0.585
	Quantum	**0.539**	**0.772**	**0.680**
Audio	Original	0.611	0.824	0.407
	Baseline	0.855	0.925	0.675
	Quantum	**0.926**	**0.973**	**0.776**
Video_lip	Original	0.457	0.738	0.533
	Baseline	**0.693**	**0.856**	**0.663**
	Quantum	0.641	0.818	0.635

TABLE III: Data analysis of white-box attack in video modality.

Data_type	IISR_{rmse}	IISR_{psnr}	IISR_{ssim}
Original	0.314	0.611	0.491
Quantum	0.539	0.772	0.680
White-box Attack	0.536	0.761	0.662

corresponding to this segments. The resulting visualizations are shown in Fig. 4.

As per our analysis of the visualizations, we observed that the quantum privatized data the quantum privatized data better preserves the lip area compared to the baseline data while worse in other area. Similarly, we noted that quantum privatized data captures high-frequency information better than the baseline privatized data from the audio spectrograms.

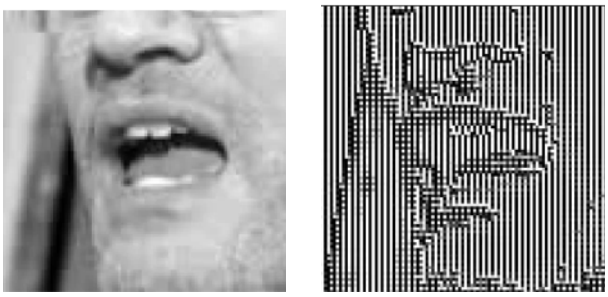
For further analysis, we employ the IISR metric. We randomly select 10 words and 30 audio-video segments for each word, and calculate IISR for the audio and video data separately. Moreover, given the significant reliance of the VSR task on the lip area in the video, we extracte the lip area of the video segments and compute its IISR. The results are shown in Table II.

From the first two rows of the table, we can see that for both the video and audio data, the IISR values for the original, baseline, and quantum privatized data follow the trend of $Original_{\text{IISR}} < Baseline_{\text{IISR}} < Quantum_{\text{IISR}}$, regardless of the IISR_{rmse}, IISR_{psnr} or IISR_{ssim}. This suggests that the quantum privatized data retains relatively less information compared to the baseline method, thereby providing stronger privacy protection. By further analyzing the IISR of lip area data, we can find that there is $Original_{\text{IISR}} < Quantum_{\text{IISR}} < Baseline_{\text{IISR}}$, which is different from previous results and indicating that quantum privatized data retain more information in the lip region, thus performing better in downstream tasks.

D. White-box Attacking Experiments

The white-box attack results is visualized in Fig. 5. Additionally, we calculated the IISR for the white-box attack results, with the results shown in Table III.

As we can see from the figure or the table, the impact of white-box attacks is very limited. The quantum privacy protection method exhibited strong resistance to white-box

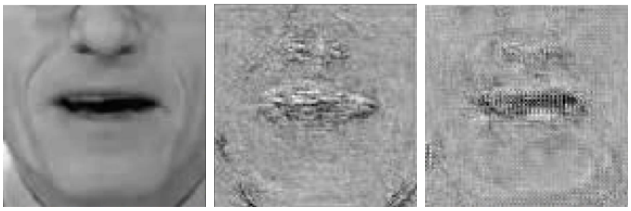


(a) (b)

Fig. 5: Visualization of white-box attack. (a) Grayscale plot of original video data; (b) Grayscale plot of white-box attack results of video quantum privatized data.

attacks, which is indicative of the irreversibility of quantum circuits.

E. Black-box Attacking Experiments



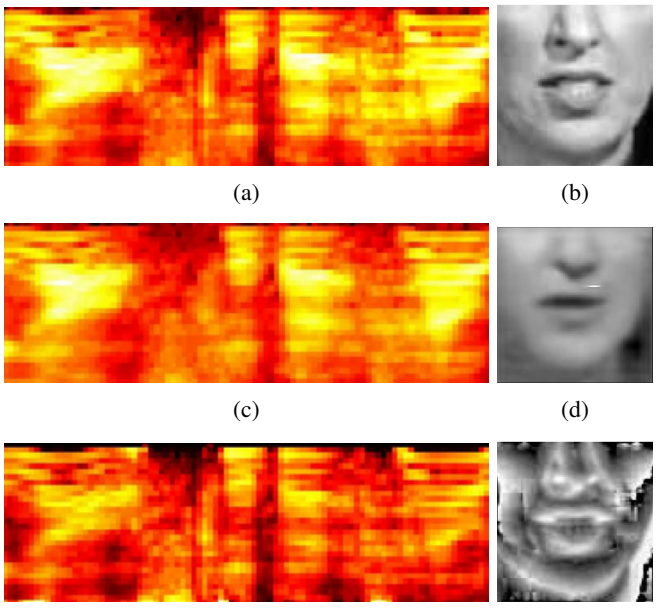
(a) (b) (c)

Fig. 6: Visualization of untrained black-box attack. (a) Grayscale plot of original video data; (b) Grayscale plot of untrained black-box attack results of video baseline privatized data; (c) Grayscale plot of untrained black-box attack results of video quantum privatized data.

In this section, we discuss the privacy-preserving reliability of the two approaches under black-box attacks.

Since the quantum privacy method does not require any trainable parameters and the number of parameters required for the baseline method is small, we examined the reconstruction results of the untrained attack network after initialization. The results are visualized in Fig. 6. From the figure, it is evident that compared to the baseline, the quantum privatized data reconstruction results are better restored near the lips, but worse restored in other regions.

The visualization results of the audio and video data reconstructed by the trained black-box attack network are also shown in Fig. 7. The video visualizations reveal that while both the quantum privacy protection method and the baseline method are effective at resisting black-box attacks, the restoration results for the video quantum privatized data are superior to those of the video baseline privatized data in the lip region. Similarly, the audio quantum privatized data exhibits better restoration capability for high-frequency information compared to the audio baseline privatized data.



(a) (b) (c) (d) (e) (f)

Fig. 7: Visualization of black-box attack. (a) Spectrogram of original audio data; (b) Grayscale plot of original video data; (c) Spectrogram of black-box attack results of audio baseline privatized data; (d) Grayscale plot of black-box attack results of video baseline privatized data; (e) Spectrogram of black-box attack results of audio quantum privatized data; (f) Grayscale plot of black-box attack results of video quantum privatized data.

TABLE IV: Data analysis of black-box attack.

Modality	Data_type	IISR_{rmse}	IISR_{psnr}	IISR_{ssim}
Video	Original	0.314	0.611	0.491
	Baseline	0.489	0.838	0.773
	Quantum	**0.632**	**0.867**	**0.892**
Audio	Original	0.611	0.824	0.407
	Baseline	0.798	0.956	0.882
	Quantum	**0.914**	**0.984**	**0.947**
Video_lip	Original	0.457	0.738	0.533
	Baseline	**0.589**	**0.864**	**0.894**
	Quantum	0.562	0.839	0.848

We compute the IISR values of the reconstructed audio and video data, with the results depicted in Table IV. From the table, we can see that for both audio and video modalities, the reconstruction results of quantum privatized data are inferior to those of the baseline privatized data, indicating that the quantum protection approach is more resistant to black-box attack than the baseline. Furthermore, we analyze the lip region data separately and find that the reconstruction results of quantum privatized data are better than those of the baseline privatized data. This observation provides further evidence to support the superior performance of the quantum privatized data for word-level speech recognition tasks.

VI. Conclusions

This paper successfully demonstrates the effectiveness of quantum circuits-based privacy-preserving methods in audio and video modalities. Through visualizations and the IISR metric, we are able to analyze the privacy-preserving ability of the two methods in terms of privatized data, resistance to white-box privacy attacks, and black-box privacy attacks. Furthermore, by analyzing the data in the lip region, we can find that the quantum privacy-preserving methods can maintain the features required for downstream tasks while discarding other information, thereby enabling excellent privacy-preserving capabilities while performing well in downstream tasks, which is inseparable from the advantage of quantum circuits in feature representation capability.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427.

References

[1] P. Voigt and A. V. dem Bussche, "The eu general data pro tection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[2] C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in *International Conference on the Theory and Application of Cryptographic Techniques*, 2011.

[3] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, 2008.

[4] C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Pate-aae: Incorporating adversarial autoencoder into private aggregation of teacher ensembles for spoken command classification," *Proc. of Interspeech*, 2021.

[5] J.-B. Ji, H. Wang, Y. Huang, *et al.*, "Privacy-preserving face recognition with learnable privacy budgets in frequency domain," in *European Conference on Computer Vision*, 2022.

[6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *arXiv: Artificial Intelligence*, 2019.

[7] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Physical Review A*, 2018.

[8] C.-H. H. Yang, J. Qi, S. Y.-C. Chen, *et al.*, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6523–6527, 2020.

[9] M. Benedetti, E. Lloyd, S. H. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Science and Technology*, vol. 4, 2019.

[10] J. Qi and J. Tejedor, "Classical-to-quantum transfer learning for spoken command recognition based on quantum neural networks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8627–8631, 2021.

[11] S. Y.-C. Chen, C.-H. H. Yang, J. Qi, P.-Y. Chen, X. Ma, and H.-S. Goan, "Variational quantum circuits for deep reinforcement learning," *IEEE Access*, vol. 8, pp. 141 007–141 024, 2019.

[12] V. Havlíek, A. D. Córcoles, K. Temme, *et al.*, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, pp. 209–212, 2018.

[13] C.-H. H. Yang, B. Li, Y. Zhang, *et al.*, "A quantum kernel learning approach to acoustic modeling for spoken command recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] Y. Li, R.-g. Zhou, R. G. Xu, J. Luo, and W. Hu, "A quantum deep convolutional neural network for image recognition," *Quantum Science & Technology*, vol. 5, 2020.

[15] F. Li, S. Zhao, and B.-y. Zheng, "Quantum neural network in speech recognition," *6th International Conference on Signal Processing, 2002.*, vol. 2, 1267–1270 vol.2, 2002.

[16] R. Wang, J. Du, and T. Gao, "Quantum transfer learning using the large-scale unsupervised pre-trained model wavlm-large for synthetic speech detection," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[17] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 10, pp. 557–570, 2002.

[18] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, 2007.

[19] F. Boemer, R. Cammarota, D. Demmler, T. Schneider, and H. Yalame, "Mp2ml: A mixed-protocol machine learning framework for private inference," *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020.

[20] P. C. M. Arachchige, P. Bertók, I. Khalil, D. Liu, and S. A. Çamtepe, "Privacy preserving face recognition utilizing differential privacy," *Comput. Secur.*, vol. 97, p. 101 951, 2020.

[21] G. McGraw and A. K. Ghosh, "Developing expertise in software security an outsider s perspective," 1996.

[22] H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *ArXiv*, vol. abs/1710.09412, 2017.

[23] F. Mireshghallah, M. Taram, A. Jalali, A. T. Elthakeb, D. M. Tullsen, and H. Esmaeilzadeh, "Not all features are equal: Discovering essential features for preserving

prediction privacy," *Proceedings of the Web Conference 2021*, 2021.

[24] C.-H. H. Yang, I.-F. Chen, A. Stolcke, S. M. Siniscalchi, and C.-H. Lee, "An experimental study on private aggregation of teacher ensemble learning for end-to-end speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 1074–1080.

[25] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multim.*, vol. 2, pp. 141–151, 2000.

[26] B. Shillingford, Y. Assael, M. W. Hoffman, *et al.*, "Large-scale visual speech recognition," *ArXiv*, vol. abs/1807.05162, 2018.

[27] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 713–722, 2019.

[28] H. Chen, J. Du, Y. Hu, L.-R. Dai, B.-C. Yin, and C.-H. Lee, "Automatic Lip-Reading with Hierarchical Pyramidal Convolution and Self-Attention for Image Sequences with No Word Boundaries," in *Proc. Interspeech 2021*, 2021, pp. 3001–3005.

[29] T. Parcollet, M. Morchid, and G. Linarès, "E2e-sincnet: Toward fully end-to-end speech recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7714–7718, 2020.

[30] B. Martínez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6319–6323, 2020.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[32] F. Bloch, N. INTRODUCTIO, R. Zacharias, *et al.*, "Nuclear induction," 2011.

[33] J. Qian, R. Li, X. Yang, *et al.*, "Hasa: Hybrid architecture search with aggregation strategy for echinococcosis classification and ovary segmentation in ultrasound images," *ArXiv*, vol. abs/2204.06697, 2022.

[34] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[35] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 2067–2079, 2010.

[36] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[37] M. Ehrlich, S.-N. Lim, L. S. Davis, and A. Shrivastava, "Quantization guided jpeg artifact correction," in *European Conference on Computer Vision*, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Interpretable Image Recognition in Hyperbolic Space

Irina Lebedeva* Mohamed Jaward Bah† and Taihao Li‡
* Research Institute of Artificial Intelligence, Zhejiang Lab, China
E-mail: irina@zhejianglab.com
† Research Center for Big Data Intelligence, Zhejiang Lab, China
E-mail: easybah@zhejianglab.com
‡ Research Institute of Artificial Intelligence, Zhejiang Lab, China
E-mail: lith@zhejianglab.com

***Abstract*—The lack of human-level interpretability in the decision-making process of computer vision is one of the obstacles to further its application and performance improvement. This issue is investigated by eXplainable Artificial Intelligence (XAI), and a range of interpretable XAI networks has been proposed to mitigate the black box issue. Most of them attempt to implement case-based reasoning in image recognition that is similar to humans. This cognitive process includes dissecting the image and identifying prototypical parts that distinct one class from another. However, only a few works take into consideration the hierarchical principle of this process and visual data. In contrast to previously introduced XAI approaches that utilize Euclidean space to design a network, hyperbolic space with its ability to embed hierarchical structures is exploited in this work. The proposed prototype-based hyperbolic network, namely ProtoHyp, first learns prototypes that play key roles in identifying classes. Then, using the similarity between them and an image, the human-understandable classification of an image is conducted in hyperbolic space. Experimental results on two image recognition datasets, Birds CUB-200-2011 and Stanford Dogs, have demonstrated the improvement of state-of-the-art interpretable image recognition in terms of both accuracy and interpretability.**

I. Introduction

Despite recent advances in computer vision, there is a big gap between the way machines and humans perceive an image and recognize objects [1], [2]. The black box issue of deep learning limits not only the improvement of computer vision but also its application [3], [4]. EXplainable Artificial Intelligence (XAI) addresses the black box issue and aims to bring the decision-making process in AI closer to human reasoning. The current XAI includes post-hoc and interpretable approaches. Although the former ones that produce posterior explanations of deep learning models have attracted a lot of interest, their results are often misleading and not reliable [1]. Interpretable XAI methods are superior due to their ability to automatically provide human-interpretable explanations alongside their predictions [3].

Most interpretable XAI approaches are represented by prototype-based networks [5], [6]. The underlying idea is to reproduce the human reasoning process in image classification. This process includes dissecting an image and identifying prototypical parts that distinguish one class from another [7], [8]. The principle of this cognitive process is illustrated in Fig. 1. After learning prototypical parts, the similarity between an input image and these prototypes is utilized to conduct human-explainable classification. The ProtoPNet network [5], the first attempt to implement this principle, has achieved reasonable accuracy and interpretability. However, it has a range of issues that have been addressed in the subsequent works. The deformable ProtoPNet exploits prototypes that adaptively change their relative positions instead of spatially rigid ones [9]. ProtoPShape [10] and ProtoPool [7] enhance the network by decreasing the number of prototypes and simplifying the training process.

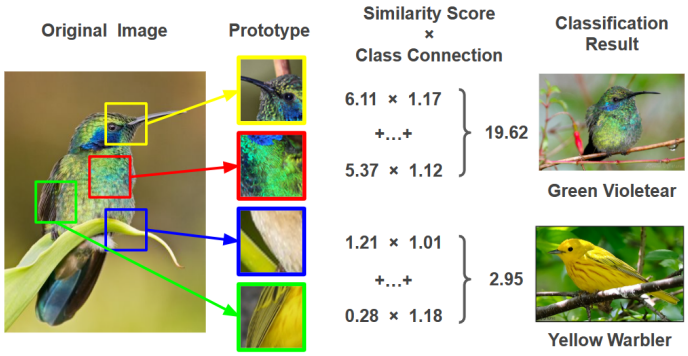


Fig. 1. The result of the proposed ProtoHyp network on Birds CUB-200-2011 - an example of case-based reasoning in image recognition. The interpretable classification is achieved by representing each prediction score as a linear combination of similarities between an image and prototypes learned in the training phase.

Although the case-based reasoning described above has the hierarchical nature [6], only a few works explore and exploit it to improve interpretable XAI. The hierarchy of class labels is utilized to interpretably recognize unseen objects at their parent node of the taxonomy [11]. Hierarchical trees are exploited for multi-grained interpretable recognition [12]. The hierarchical nature is inherent not only to reasoning in image recognition but is also a feature of visual data [13]. The consideration of hierarchical relations between images and their parts have already demonstrated the improvement of the state-of-the-art

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

in metric [14], [15] and few-shot recognition [13], [16].

Hyperbolic space (HS) with its negative curvature and exponential growth has a better ability to embed hierarchical structures than Euclidean space (ES). However, its usage has not yet been investigated for interpretable image recognition. Only polynomially expanding Euclidean space (ES) with zero curvature has been exploited to design interpretable image recognition networks.

In this work, this gap has been narrowed, and a novel interpretable image recognition network, namely ProtoHyp, is introduced. Deformable prototypes [9] are first learned. Then, the similarity score between an input image and the prototypes is obtained. The third step is exponential mapping from Euclidean space to hyperbolic space. Poincarè ball is selected as the hyperbolic space model due to the differentiability of its distance function and its capacity for gradient-based optimization [15]. Finally, the formalism of Möbius gyrovector spaces is utilized to carry out the interpretable classification process in Poincarè ball. Since, to the best of our knowledge, this work is the first attempt to utilize hyperbolic space for the purpose of XAI, only the last (interpretable classification) layer is implemented in HS for a better comparison to its performance in Euclidean space.

To evaluate the performance of the proposed ProtoHyp network, the following experiments are conducted on two image recognition datasets, Birds CUB-200-2011 [17] and Stanford Dogs [18]. Firstly, the results of the model with various backbone architectures, such as VGG-16 [19], ResNet34 [20], ResNet152 [20], and Dense161 [21], on cropped and full images are presented. Secondly, different combinations of hyperbolic parameters in Poincarè ball, such as its curvature and dimentionality, are evaluated as an ablation study. Thirdly, the quantitative and qualitative comparisons to the state-of-the-art XAI in terms of prediction accuracy and interpretability are demonstrated.

Our main contributions are summarized as follows:

- To the best of our knowledge, we have made the first attempt to explore hyperbolic space due to its ability to embed hierarchical structure for the purpose of interpretable image recognition.
- A novel interpretable XAI network, namely ProtoHyp, that imitates human reasoning process in image recognition is introduced. It makes an explainable decision in hyperbolic space using the similarity to prototypes that play key roles in the class definition.
- Extensive experiments on Birds CUB-200-2021 and Stanford Dogs datasets demonstrate that the proposed ProtoHyp overcomes the state-of-the-art XAI in terms of accuracy and interpretability.
- Various hyperbolic parameters of Poincarè ball, such as its curvature and dimensionality, have been evaluated. Their best combination, as well as backbone architectures, has been empirically selected for the purpose of interpretable recognition.

II. Related Work

In this section, we first review the relevant literature on interpretable XAI. Then, a brief survey of vision models implemented in hyperbolic space is presented.

A. Interpretable Vision Models

ProtoPNet [5] is the first interpretable network that implements case-based reasoning in image recognition similar to humans. It achieves interpretable image classification by comparing crucial image parts with a set of prototypes assigned to a class. However, the network has a range of issues that have been addressed in the following works. First of all, ProtoPNet uses spatially rigid prototypes that have negative influence on its performance even on images preprocessed by cropping. This shortcoming is addressed in [9], and a deformable prototypical network based on prototypes that adaptively change their relative spatial positions is proposed.

Another issue of the ProtoPNet network is the separate usage of prototypes for each class. ProtoPShape [10] and ProtoPool [7] have been designed in order to overcome this problem. The former network shares the prototypes between classes by pruning prototypical parts using data-dependent similarity. ProtoPool decreases the number of prototypes by reusing them in classes and allows fully automatic differentiable assignment of prototypes to classes to simplify the training process [7]. The issue of complex data structure and confusing the effect of individual prototypes on the output category is addressed in [8]. In this study, the interpretable network TesNet constructs prototypes on the Grassmann manifold with several regularization loss functions as constraints.

The hierarchical principle of human reasoning in image recognition is also explored in a range of studies. The predefined hierarchy of class labels is utilized to classify objects at each level and interpretably recognize images from unseen classes at their parent level in the taxonomy [11]. A multi-grained interpretable network that organizes classes into different granularities within a class hierarchy tree and provides decision pathways with different granularity explanations is proposed in [12]. An interpretable deep learning method that combines prototype-based XAI with decision trees is introduced in [6]. It can explain a single prediction by outlining a hierarchical path through the decision tree.

In contrast to previously proposed interpretable vision models that are fully designed in Euclidean space, the proposed ProtoHyp is constructed in hyperbolic space (Poincarè ball). Since HS has a better ability to embed hierarchical structure [13], its usage potentially improves the interpretable image classification in both accuracy and interpretability points of view.

B. Hyperbolic Vision Models

It is common in deep learning to represent data in Euclidean space due to its vectorial structure, continuity and closed-form formulas of distance and inner-product [22]. However, many scientific fields study data with non-Euclidean structures. Hyperbolic space with its tree-like properties that help embed

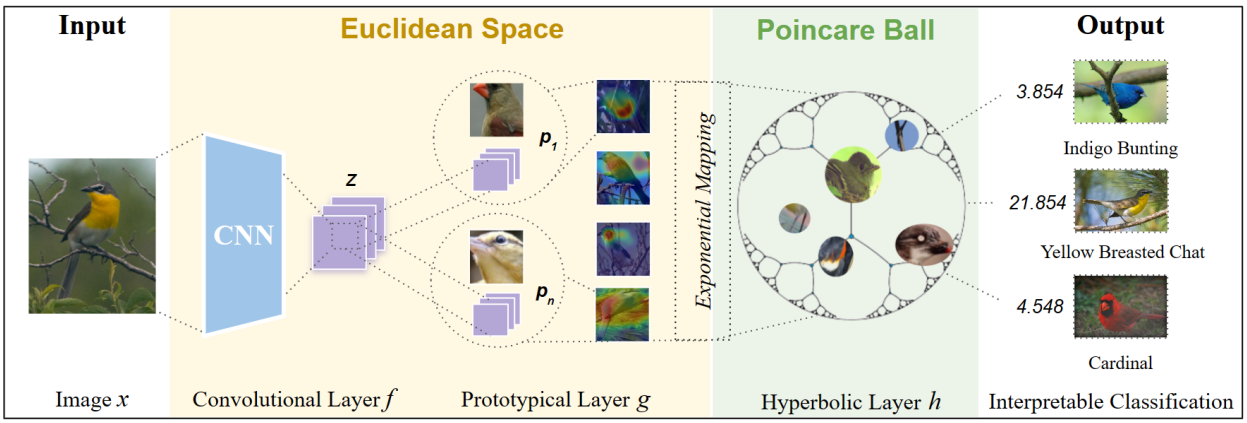


Fig. 2. The flowchart of the proposed ProtoHyp network. Its convolutional layer is responsible for feature extraction, prototypical one obtains image regions playing key roles in recognition, while the hyperbolic layer conducts interpretable classification.

taxonomies is one of the alternatives to ES [15]. Although the advantages of HS have successfully been applied in the NLP area, less investigation has been done to improve computer vision models by using hyperbolic space [13].

The first attempt to apply the machinery of hyperbolic networks to computer vision was made in [13]. The work demonstrates that HS is beneficial for visual data and exploits it for few-shot learning and person re-identification. Moreover, an approach for evaluation of the dataset hyperbolicity based on the concept of Gromov δ - hyperbolicity is introduced. Few-shot learning in HS is also explored in [16] where a hyperbolic knowledge transfer framework with hierarchical classification and relation reconstruction loss is proposed to learn the tree-like structure of class hierarchy.

A range of works have addressed metric learning in HS [15], [23]. The joint usage of vision transformers, hyperbolic embeddings, and pairwise cross-entropy loss is utilized in [14]. Contrastive learning is explored and a novel framework for self-supervised image representation learning is proposed in [15]. The noisy label issue in metric learning is addressed in [23], and an adaptive hierarchical similarity approach is introduced.

Compared to existing hyperbolic networks, in this work, the advantages of HS are explored for another task - interpretable image recognition. Hyperbolic space is utilized to design the ProtoHyp network that conducts human-explainable classification by using the similarity of prototypes that play key roles in the class definition. Since these prototypes have hierarchical principle, the usage of HS can improve the performance of interpretable recognition.

III. The ProtoHyp Network

In this section, the main points of the hyperbolic geometry that are necessary for the proposed approach are briefly introduced. The ProtoHyp architecture, including its prototype learning, mapping to Poincarè ball and hyperbolic layer, is described.

A. Hyperbolic Geometry

In contrast to Euclidean space, HS is not a vector space in the traditional sense. Thus, standard operations, such as multiplication and addition, cannot be utilized. The non-associative algebraic formalism of Möbius gyrovector space allows to generalize many operations to hyperbolic space [22].

Among the various isometric models used to describe hyperbolic spaces, Poincarè ball has to be selected due to the differentiability of its distance function and ability for gradient-based optimization [15]. The model is defined as $D_c^m = \{a \in R^m : c\|a\|^2 < 1, c \geq 0\}$, where c is the curvature of Poincarè ball, m - its dimensionality. The corresponding conformal factor that denotes the Euclidean metric tensor takes the form $\lambda_a^c = \frac{2}{1-c\|a\|^2}$.

For two vectors $a, b \in D_c^m$, the Möbius addition is defined as follows:

$$a \oplus_c b = \frac{(1 + 2c\langle a, b\rangle + c\|b\|^2)a + (1 - c\|a\|^2)b}{(1 + 2c\langle a, b\rangle + c^2\|a\|^2\|b\|^2)}. \quad (1)$$

The hyperbolic distance is calculated by:

$$d_c(a, b) = \frac{2}{\sqrt{c}}\text{arctanh}(\sqrt{c}\| - a \oplus_c b\|). \quad (2)$$

If $c = 0$, the HS model loses its curvature, and all the operations presented below take their usual Euclidean form $\lim_{c\to 0} d_c(a, b) = 2\|a - b\|$. The curvature parameter c allows the balance between Euclidean and hyperbolic geometries. In Section IV-C, the performance of ProtoHyp with various curvature parameters from 0.1 to 0.9 is demonstrated, and its influence on the results is discussed.

B. ProtoHyp

1) The overall architecture: The structure of the proposed network is illustrated in Fig. 2. The architecture of the ProtoHyp network is generally inspired by ProtoPNet, the fundamental work that made the first attempt to implement

case-based reasoning in image recognition [5]. Similar to ProtoPNet, our network includes the convolutional neural network f, prototypical layer g. However, in contrast, the last layer of the proposed network is a layer in hyperbolic space (III-B4) and the prototypical layer (III-B2) has a different structure as well. The mapping to Poincarè ball is also added to the architecture. Regarding the convolutional layer, various models such as VGG-16 [19], ResNet34 [20], ResNet152 [20], and Dense161 [21] are employed as its backbone networks. The performances of ProtoHyp with different models are discussed in Section IV-C.

2) Prototypical Layer: Let $[X, Y] = (x_i, y_i)_{i=1}^n$ be the training set, $f(x)$ the image features extracted by the convolutional layer. The next step is to obtain the deformable prototypes P in g. The approach to learning deformable prototypes presented in [9] is adapted to ProtoHyp. After the normalization and interpolation of the image features $f(x)$ as described in the work, we get the feature tensor z and compare it with a l-deformable prototype $p^{(k,l)}$. $z_{s,t}$ is used as the d-dimensional vector at the spatial location (s, t) of the feature tensor z. In contrast to ProtoPNet, where L^2-based similarity is utilized to compare a prototype and an image, ProtoHyp utilizes cosine similarity, and the final function of the prototypical layer is:

$$g(z_{s,t}^{(k,l)}) = p^{(k,l)} \cdot z_{s,t} = \cos(\theta(p^{(k,l)}, z_{s,t})) \tag{3}$$

The convolutional f and prototypical g layers of ProtoHyp are trained while keeping its hyperbolic layer fixed. For this purpose, stochastic gradient descent is performed, and features $z_{s,t}$ of class k are clustered around prototypes $p^{(k,l)}$ of the class. Cluster, separation, and orthogonality loss functions [9] are utilized for the training of these layers.

3) Mapping to Poincarè Ball: In order to perform operations in hyperbolic space, the mapping of Euclidean vectors to HS is defined. It is called the exponential map, while mapping from the Poincarè model to Euclidean space is termed the logarithmic map [13].

$$\exp_a^c(v) = a \oplus_c \Big(\tanh\Big(\sqrt{c}\frac{\lambda_a^c \|v\|}{2}\Big) \frac{v}{\sqrt{c}\|v\|} \Big), \tag{4}$$

where the notations are as same as in Section III-A. In this work, the practice of previous hyperbolic vision models is followed, and a is set to 0. Then, $\exp_a^c$ makes the calculation less cumbersome with little impact on the results [14], [15].

4) Hyperbolic Layer: The last layer of ProtoHyp produces interpretable prediction results in Poincarè ball by applying hyperbolic geometry described in Section III-A. Similar to existing interpretable image recognition networks that are designed in Euclidean space, a fully connected layer is considered and implemented in HS. For this purpose, the Möbius matrix is defined by the vector product:

$$M^{\oplus_c}(a) = \frac{1}{\sqrt{c}} \tanh\Big(\frac{\|Ma\|}{\|a\|}(\sqrt{c}\|a\|)\Big) \frac{Ma}{\|Ma\|}, \tag{5}$$

if $Ma = 0$, $M^{\oplus_c}(\mathrm{a}) = 0$. Then, the operation underlying the hyperbolic fully connected layer is given by $M^{\oplus_c}(\mathrm{a}) \oplus_c \mathrm{b}$ for a bias vector $\mathrm{b} \in D_c^n$.

The loss function utilized for the training of the hyperbolic layer is defined as:

$$loss_{hyp} = CE + \lambda \sum_{k,l} \sum_{k \neq k'} |w_h^{((k,l),k')}|, \tag{6}$$

where CE is the standard cross entropy loss, λ is a learning rate, w_h is the weights of the hyperbolic layer. The right part of the loss discourages the negative reasoning processes and is explained in [5].

The examples of the network output are illustrated in Fig. 1. The output is logits that represent each class, and the one with the highest score is selected as the result. Along with the predicted class, the approach demonstrates all the prototypes that influence the final decision and their "weights" in it.

IV. Experimental Results

In this section, the datasets utilized for experiments are first described. Then, the performance of the proposed interpretable neural network with various backbone architectures and preprocessing approaches is evaluated in Section IV-C. Thirdly, the ablation study with different combinations of hyperbolic parameters is demonstrated in IV-D. The quantitative IV-E and qualitative IV-F comparisons to the state-of-the-art of interpretable XAI architectures are also presented.

A. Datasets

1) Birds CUB-200-2011 [17]: The dataset includes 200 species of birds, presented by 11788 images in total. Annotations include bounding boxes and segmentation labels. The train/test separation is predefined. The dataset can be found by following this link[1].

2) Stanford Dogs [18]: It contains 120 breeds of dogs. There are 20580 images in all, with 12000 being the training set and 8580 constitute a testing set. The coordinates of bounding boxes are also provided [2].

B. Implementation Details

The number of prototypes is selected at 2000, following the previous prototype-based approaches [5]. The backbone networks utilized for the experiments, including the state-of-the-art comparison, are pretrained on ImageNet [24]. With regard to the training process, the number of epochs is set to 25, and the learning rate of 0.0001 is experimentally selected. All experiments in this work are run on Amazon Elastic Compute Cloud [3]. A GPU-based deep learning virtual machine with Ubuntu 20.04 and CUDA 11.7 are selected to conduct the experiments. The ProtoHyp network is implemented in Pytorch 1.13, Python 3.9.

TABLE I
PERFORMANCE OF THE PROPOSED PROTOHYP ON THE BIRDS AND DOGS DATASETS. THE ACCURACY ARE OBTAINED USING VARIOUS BACKBONE ARCHITECTURES, WITH (BB IMAGES) AND WITHOUT IMAGE CROPPING (FULL IMAGES)

Dataset	Full Images				BB images			
	VGG-16	*ResNet34*	*ResNet152*	*Dense161*	*VGG-16*	*ResNet34*	*ResNet152*	*Dense161*
Birds CUB-200	80.034	81.492	82.581	**83.125**	80.963	82.514	82.173	82.428
Stanford Dogs	78.591	81.527	**86.901**	84.305	78.848	82.035	86.181	84.140

TABLE II
ACCURACY OF PROTOHYP WITH VARIOUS COMBINATIONS OF HYPERBOLIC PARAMETERS OF POINCARÈ BALL.

Curvature	dim = 2	dim = 20	dim = 200	dim = 2000
0.1	79.541	81.873	82.729	81.904
0.2	79.583	82.400	83.270	82.353
0.3	80.983	82.983	83.612	82.735
0.4	80.970	83.239	**84.142**	83.031
0.5	81.003	83.080	83.829	82.917
0.6	81.192	82.793	83.650	83.049
0.7	80.895	82.482	83.399	82.873
0.8	80.704	82.531	83.274	82.760
0.9	80.930	82.472	83.221	82.612

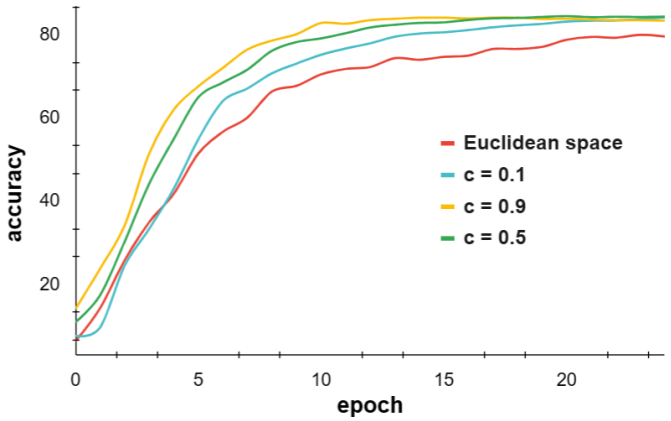


Fig. 3. The graph of the accuracy vs. the number of epochs. The red curve is associated with the accuracy during the training data in Euclidean space, while the blue, green, and yellow curve show the accuracy in Poincarè ball with the curvature value of $0.1, 0.5, 0.9$, respectively.

C. ProtoHyp Performance Evaluation

Table I demonstrates the accuracy of the proposed ProtoHyp with various backbone networks, including VGG-16 [19], ResNet34 [20], ResNet152 [20], and Dense161 [21]. The experiments have been conducted on full images, as well as cropped ones by using bounding boxes. The results obtained on Birds CUB-200-2011 and Stanford Dogs described in Section IV-A are presented. To begin, the parameters of hyperbolic space recommended by the authors in [13] and their implementation are utilized for the experiments. The dimensionality of Poincarè Ball is set to 500, while its curvature is selected to be 0.5.

Since image cropping is essential for some of the previously proposed interpretable neural networks [5], [7], skipping this step results in significantly lower classification accuracy, the effectiveness of ProtoHyp in both scenarios - with and without image preprocessing is presented. As we can see, the model demonstrates almost the same performance in all cases due to the usage of spatially flexible prototypes. It is quite predictable that the best accuracy among all the backbone architectures presented in Table I has been achieved with the network with the highest number of parameters - ResNet152 and Dense161. This difference is especially obvious on full images.

D. Ablation Study

Various parameters of the Poincarè ball model, such as its curvature, its dimensionality has been evaluated in order to find their optimal practical combination and compare the results with the theory given in previous works [13]. Table II demonstrates the performance of the model in terms of accuracy with different parameters on Birds CUB-200-2011. Since both datasets utilized in this work has similar consistency, fine-grained and in-the-wild, ablation study is conducted on the birds dataset, and the same hyperbolic parameters are applied to Stanford Dogs. The backbone architecture used in these experiments is Dense161. The highest accuracy achieved is with a curvature of 0.4. This value is close to the hyperbolicity value of the Birds CUB-200-2011 dataset [13]. The dimensionalities of 2, 20, 200, and 2000 is selected due to the lowest possible dimensionality of the Poincarè model that equals 2 (Poincarè Disk). Then, we multiply it by 10. The best performance is achieved with $dim = 200$, while it starts decreasing with a higher dimensionality.

The experiments have also revealed the dependency of the network convergence on the curvature of Poincarè ball. Fig. 3 demonstrates the training process of ProtoHyp with different curvature values. Moreover, it also shows the graph of the accuracy vs. the number of epochs during the training of the network in Euclidean space (red line). As it is mentioned in [13], the hyperbolic space with $c = 0$ is the analogy of ES, and $c = 0.1$ (blue line) demonstrates the most similar training process. The faster convergence with a slight loss in accuracy is produced by $c = 0.9$ (yellow line). It can be concluded that ProtoHyp and models in HS, as a whole, are superior to neural networks designed in ES in terms of convergence.

[1]https://www.vision.caltech.edu/datasets/
[2]http://vision.stanford.edu/aditya86/ImageNetDogs/
[3]https://aws.amazon.com/ec2/

TABLE III
COMPARISON TO THE STATE-OF-THE-ART OF INTERPRETABLE XAI MODELS IN TERMS OF CLASSIFICATION ACCURACY. FRAMEWORKS USING BOUNDING BOXES ARE MARKED (BB)

Model	Birds CUB-200				Stanford Dogs			
	VGG-16	*ResNet34*	*ResNet152*	*Dense161*	*VGG-16*	*ResNet34*	*ResNet152*	*Dense161*
ProtoPNet (bb) [5]	77.482	78.763	79.392	79.872	76.867	78.402	79.728	80.193
ProtoPool (bb) [7]	79.148	80.309	81.458	82.027	78.095	78.156	80.802	81.368
ProtoTree [6]	75.904	77.182	78.036	79.515	74.038	76.883	79.894	80.785
TesNet [8]	77.165	77.957	80.382	81.591	74.547	80.129	83.177	84.115
Def.ProtoPNet [9]	80.282	81.126	83.194	83.848	77.905	81.272	86.657	83.742
ProtoHyp	81.315	82.490	83.802	**84.026**	78.684	82.544	**87.348**	85.205

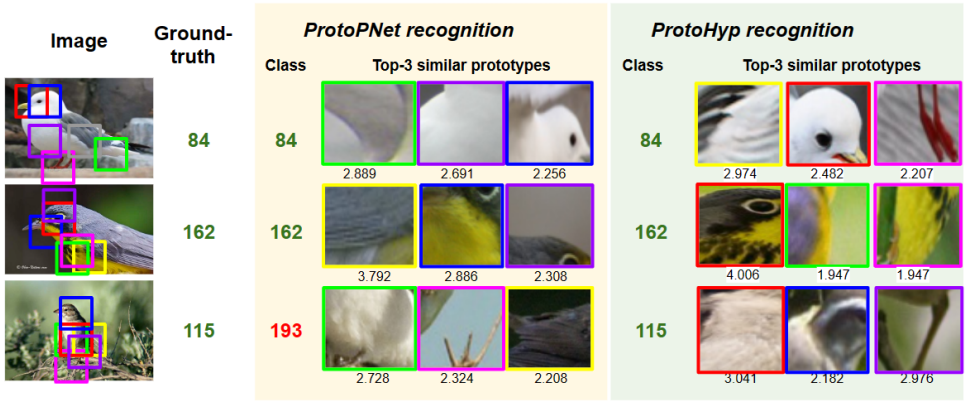


Fig. 4. Qualitative comparison of ProtoPNet and ProtoHyp on Birds CUB-200-2011. The left two columns are an image and its ground-truth. The middle and the right ones are prediction results and top 3 most similar prototypes obtained by ProtoPNet and ProtoHyp, respectively. The image parts corresponding to prototypes are marked with the same color.

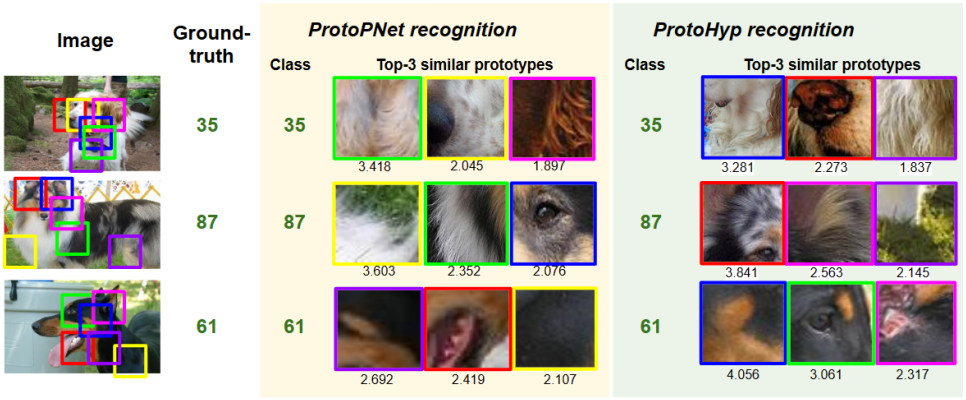


Fig. 5. Qualitative comparison of ProtoPNet and ProtoHyp on Stanford Dogs. The left two columns are an image and its ground-truth. The middle and the right ones are prediction results and top 3 most similar prototypes obtained by ProtoPNet and ProtoHyp, respectively. The image parts corresponding to prototypes are marked with the same color.

E. Quantitative Comparison to the State-of-the-Art

In order to demonstrate the advantages of the proposed ProtoHyp network and the usage of hyperbolic space for interpretable XAI as a whole, quantitative evaluations of existing interpretable networks in terms of accuracy are first presented in Table III. The reimplementation of the state-of-the-art approaches is conducted using the official codes created by the respective authors with their recommended hyperparameters. The fundamental case-based reasoning architecture ProtoPNet [5], its enhanced versions ProtoPool [7], TesNet [8], ProtoTree [6], Deformable ProtoPnet [9], are used for the comparison. More details about the works can be found in Section II-A. Their performances with various backbone networks achieved on the datasets are evaluated. The results of the proposed ProtoHyp with the hyperbolic parameters that demonstrated the highest accuracy in Section IV-D are selected for comparison to the state-of-the-art ($c = 0.4, dim = 200$).

Regarding the performances of ProtoPNet and ProtoPool presented in Table III, these results have been obtained by

using bounding boxes. In other words, cropped and augmented images of the datasets are used for training and testing (in this case, without augmentation). In contrast, other models presented in the table are able to produce reasonable accuracy on full images of the datasets without any preprocessing. However, only the deformable ProtoPNet that exploits flexible prototypes instead of spatially rigid overcomes the "in-the-wild" issue and demonstrates the highest classification accuracy among all the previously proposed models.

It must be mentioned that the main goal of ProtoPool network is to enhance ProtoPNet by reducing the number of utilized prototypes. While other frameworks presented in Table III are based on 2000 prototypes, ProtoPool produces similar accuracy on only 200 of them. The results of ProtoTree [6] have been achieved on a single interpretable tree, not on their ensembles.

The proposed ProtoHyp network has achieved the highest accuracy among all the presented case-based reasoning interpretable XAI models. It does not require any image preprocessing due to the usage of spatially flexible prototypes. In contrast to current interpretable frameworks, ProtoHyp is designed and trained by using hyperbolic space that allows it to outperform the state-of-the-art in interpretable image recognition.

F. Qualitative Comparison of Interpretability

Although there is still no benchmark approach and metrics to evaluate the interpretability of explainable AI models [4], attempts to compare the results of the proposed ProtoHyp network and fundamental ProtoPNet [5] have been made in this work. Fig. 4 and Fig. 5 illustrate the quantitative analysis of their performances on Birds CUB-200-2011 and Stanford Dogs, respectively. Three images are randomly selected from a testing set. Then, ProtoPNet and ProtoHyp are utilized for their classification. If the predicted class is not the same as its ground-truth, it turns red. Most importantly, the figures demonstrate three prototypes that the networks found the most similar to the images. Their corresponding similarity scores are also presented. Since the main goal of explainable AI is to produce a human-understandable decision, it can be concluded that the prototypes caught by ProtoHyp are more reasonable than ProtoPNet, and the proposed network has a higher level of interpretability.

V. Conclusion and Future Work

In this work, hyperbolic space with its ability to embed hierarchical structures has been utilized to design and train an interpretable image recognition network. Prototypes with the highest influence in the final decision are first learned. After mapping into Poincarè ball, an intepretable classification is conducted. Extensive experiments have been carried out on two image recognition datasets, Birds CUB-200-2011 and Stanford Dogs. Various backbone networks and hyperbolic parameters have been evaluated in order to find the optimal configuration of the network and HS for the XAI, as a whole. Comparison to the interpretable state-of-the-art has shown its improvement in terms of accuracy and interpretability.

The research in the interpretable image recognition area suffers from the lack of approaches to compare the interpretability level of benchmarks. Our future research is going to address this issue and a dataset augmented with human perception of image interpretability is going to be collected. The usage of HS not only in the interpretable classification layer but also in lower layers is another future research direction. Taking into account the effectiveness of few-shot and representation learning in hyperbolic space, attempts to achieve transparency and interpretability of this research by utilizing hyperbolic space will be explored.

Acknowledgement

This research is supported in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant No. 2021ZD0114303; in part by the Key Research Project of Zhejiang Lab under Grant No. 2020KB0AC01; in part by the Open Research Project of Shanghai Key Laboratory of Brain-Machine Intelligence for Information Behavior under Grant No. 2022KFKT002.

References

[1] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[2] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[3] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 573–584, 2020.

[4] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, "" help me help the ai": Understanding how explainability can support human-ai interaction," *arXiv preprint arXiv:2210.03735*, 2022.

[5] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[6] M. Nauta, R. Van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 933–14 943.

[7] D. Rymarczyk, Ł. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński, "Interpretable image classification with differentiable prototypes assignment," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, Springer, 2022, pp. 351–368.

[8] J. Wang, H. Liu, X. Wang, and L. Jing, "Interpretable image recognition by constructing transparent embedding space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 895–904.

[9] J. Donnelly, A. J. Barnett, and C. Chen, "Deformable protopnet: An interpretable image classifier using deformable prototypes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 265–10 275.

[10] D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński, "Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1420–1430.

[11] P. Hase, C. Chen, O. Li, and C. Rudin, "Interpretable image recognition with hierarchical prototypes," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 32–40.

[12] P. Yang, Z. Wen, and A. Mian, "Multi-grained interpre table network for image recognition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 3815–3821.

[13] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, "Hyperbolic image embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6418–6428.

[14] A. Ermolov, L. Mirvakhabova, V. Khrulkov, N. Sebe, and I. Oseledets, "Hyperbolic vision transformers: Combining improvements in metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7409–7419.

[15] Y. Yue, F. Lin, K. D. Yamada, and Z. Zhang, "Hyperbolic contrastive learning," *arXiv preprint arXiv:2302.01409*, 2023.

[16] B. Zhang, H. Jiang, S. Feng, X. Li, Y. Ye, and R. Ye, "Hyperbolic knowledge transfer with class hierarchy for few-shot learning," *IJCAI, Vienna, Austria*, vol. 7, pp. 3723–3729, 2022.

[17] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," in *Computation and Neural Systems Technical Report*, 2011.

[18] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, Citeseer, vol. 2, 2011.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[22] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[23] J. Yan, L. Luo, C. Deng, and H. Huang, "Adaptive hierarchical similarity metric learning with noisy labels," *IEEE Transactions on Image Processing*, vol. 32, pp. 1245–1256, 2023.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Low-light is More Than Darkness: An Empirical Study on Illumination Types and Enhancement Methods

Hui Sze Liew*[a], Yuen Peng Loh†, and Simying Ong*[b]
* Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia
E-mail: {[a]s2019982@siswa., [b]simying.ong@}um.edu.my
† Faculty of Computing and Informatics, Multimedia University, Malaysia
E-mail: yploh@mmu.edu.my

***Abstract*—Low-light images challenge both human perception and computer vision algorithms. Despite notable progress in this field, there are still various gaps that are yet to be investigated, such as the significance of low-light illumination characteristics towards image enhancement and object classification. Therefore, this paper details various analyses to study this phenomenon and provide insights for future developments of algorithms and solutions. Specifically, comparative analysis was done to investigate human and machine perception towards "low-light types", followed by empirical studies on the effect of illumination types towards state-of-the-art image enhancement quality and also their pre-processing capability for downstream task, namely object classification. It is found that illumination types significantly influences the performance of enhancement algorithms that tend to cater for a "general" type of low-light illumination. This lack of illumination type awareness therefore leads models to perform well in certain conditions, but severely underperforms in others. Thus, it is imperative for upcoming works to incorporate such illumination information for potential breakthroughs in this area.**

I. Introduction

Computer vision is an automated system modeled after human vision for a machine to visually perceive and interpret the environment via digital images or videos. However, visual data taken under substandard lighting conditions, namely low-light, suffer from compromised aesthetic quality and affects the perceptibility of information. Consequently, high-level tasks such as object recognition and tracking become more challenging, even for human vision. Nonetheless, there are extensive application possibilities for intelligent visual systems that is robust to low-light conditions, for instance, smart surveillance, self-driving vehicles, and more.

In light of such potential, interest in this research domain has been growing exponentially in the recent years, particularly in low-light image enhancement (LLIE), yet there is still a noticeable gap in the existing works in terms of understanding the low-light phenomenon in image data. Typically, low-light images can be caused by various factors, such as capturing them in environments with uneven lighting, light source intensity limitations, or even limited exposure time of the capturing process. This can result in notably different visual characteristics despite similar content (e.g. objects) in the images. For instance, images captured in a very dark indoor environment may have significant noise pixels due to an increase in camera ISO to capture more light particles, whereas an image captured on the street at nighttime may have less noise pixels but exhibits a warm color tone due to the yellowish or orange glow of the street lights. Such differing visual characteristics can be termed as illumination variations.

Despite such clear visual differences in low-light images, most LLIE works aim to improve the perceptual quality or decipherability of the content within, regardless of the variations in illuminations. As such, the enhancement outcomes of LLIE may not be effective for all illumination types, not only in terms of visual quality but also as input for downstream applications like object classification. This is because ineffective enhancement may not improve object visibility and also potentially distort object features, thus causing models to fail at the classification task. Even so, there is very limited research works that look into this problem in the LLIE domain.

Therefore, this work investigates the impact of this phenomena on LLIE to derive insights and directions for future studies. Specifically, we devised quantitative and qualitative experiments in order to empirically study the impact of varying low-light image illuminations on notable LLIE methods, as well as their impact on the effectiveness of LLIE in supporting downstream tasks. The contribution of this work is three-fold:

1) A comparative analysis to explore and compare human acumen with the acuity of unsupervised machine learning at distinguishing low-light illumination types.
2) An empirical analysis to provide a comprehensive view on the impact of low-light illumination types towards the resultant image quality of various state-of-the-art LLIE methods from quantitative and qualitative perspectives.
3) An empirical study on the impact of illumination types towards the effectiveness of LLIE methods as preprocessing for downstream application task, namely object image classification.

These analyses validate the hypothesis that low-light images has varying illumination types and these variations has notable impact on the LLIE outcomes in more ways than image quality. Thus, we discuss several open issues in this domain of research

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

and invite all to explore and devise more holistic solutions for future breakthroughs in this area.

II. Review of LLIE Methods

Low-light vision related research has been focused on image enhancement for years, and the most recently notable approaches can be distinguished into two main categories, Retinex-based methods [1] and Deep Learning-based methods. Firstly, Single Scale Retinex (SSR)[2] synthesizes contrast enhancement and color constancy by performing a non-linear spatial transform to mimic the human visual system. A combination of several SSR outputs [3] produces a single output image that has both good dynamic range compression, color constancy and good tonal rendition by using Gaussian distribution to determine the weighting factors. Alongside these methods, the illumination map estimation approach proposed by [4], LIME is one of the most effective at producing visually pleasing images by estimating a single illumination map, thus reducing computational cost. However, illumination map estimations by these methods do not take into consideration the relationship between colors and low-light thus significantly affects the enhancement outcomes when faced with significantly low-light images with limited color content.

Deep learning methods on the other hand can be sub-divided into supervised models [5]–[9], unsupervised models[10]–[12], and zero-shot models [13], [14]. Zero-shot models have shown significant potential as it strikes a balance between supervised and unsupervised methods with a loose requirement on reference images for training, yet achieve competitive enhancement outcomes. The ExCNet [14] is a method that demonstrates this concept by training image-specific curve estimation networks at testing time using a small amount of training images, whereas the ZeroDCE [13] focused on formulating a set of non-reference loss functions in order to train a network model to estimate pixel-wise high-order estimation curves. Nonetheless, the network architectures as well as loss functions that are still formulated to generalize the estimation curves across "all" low-light conditions without specific consideration of illumination differences, thus only reaching a standard outcome across the board.

III. Dataset and Experiment Design

The main goal of this study can be summed up as an investigation to verify the existence and impact of varying low-light illumination types on enhancement outcomes and downstream task performance, an aspect that has been mostly overlooked in LLIE research. To this end, this study can be distinguished into two major elements, (1) comparative analysis on perceptions of low-light illuminations between human vision and unsupervised machine learning, and (2) empirical analyses on LLIE performance differences in image quality and post-enhancement object image classification when subjected to different categories of low-light illuminations.

A. Dataset

The emphasis of this work is the analyses relating to specifically low-light illumination types, an element that is not present in most notable publicly available low-light image datasets catered for enhancement research, such as LoL [15], SID [16], and SICE [17]. Therefore, the ExDark [18] is selected as the dataset for this study because to the best of our knowledge, this is the only dataset which provides a collection of low-light images with illumination type annotation as well as image and bounding box level object class labels. Specifically, the ExDark contains up to 7363 challenging real-world low-light images with human categorized illumination classes, namely *low, ambient, object, single, weak, strong, screen, window, shadow* and *twilight* as shown in Fig. 1. This categorization would facilitate the comparative analysis as well as condition the empirical analyses to understand their impact towards performances. Alongside, the dataset image level annotation of twelve object classes, which are *bicycle, boat, bottle, bus, car, cat, chair, cup, dog, motorbike, people* and *table* will be used for the downstream task evaluation.

B. Human Perception vs. Unsupervised Machine Learning

The categorization of low-light images into the 10 illumination types has been briefly discussed in [18], whereby the categorization was done wholly on observable characteristics without statistical or mathematical basis. Therefore, this work further investigates this and contrast it with categorizations that is generated via machine learning in an unsupervised manner. In particular, the color histograms of these low-light images were studied in order to formulate statistical illumination features. These features are then used to train a clustering algorithm in order to form potential groupings, so as to find out if such illumination differences can be identified purely from the statistics. Subsequently, both perspectives of illumination categorizations are used in the empirical analyses to ascertain and compare the impact of illumination on image enhancement quality and object classification accuracy.

1) Feature Extraction: In order to formulate relevant features, the color histograms (i.e. intensity histograms of RGB channel) of the low-light images were explored and two key characteristics were found. Firstly, there is a notable amount of images with histograms having a large number "empty" bins in the higher intensity levels and high frequency in the lower intensity bins. This can be associated with the lack of color in very low-light images as the images contain mostly "black" pixels. Such pattern of histograms are observable in a few illumination types shown in Fig. 1, such as *low*, *ambient*, and *object*. This characteristic is therefore quantified using the sum of bins with frequencies between each logarithm of 10, from 10^0 till 10^4 including the count of "empty" bins of each color channel as a representation of histogram bins *sparsity* with 18-dimensions (6 for each channel). Secondly, it is also found that there are a few modalities in the histograms, for instance unimodal for types like *strong* and *object*, bimodal in *weak*, *window* and *shadow*, and multimodal for *single*, *screen*,

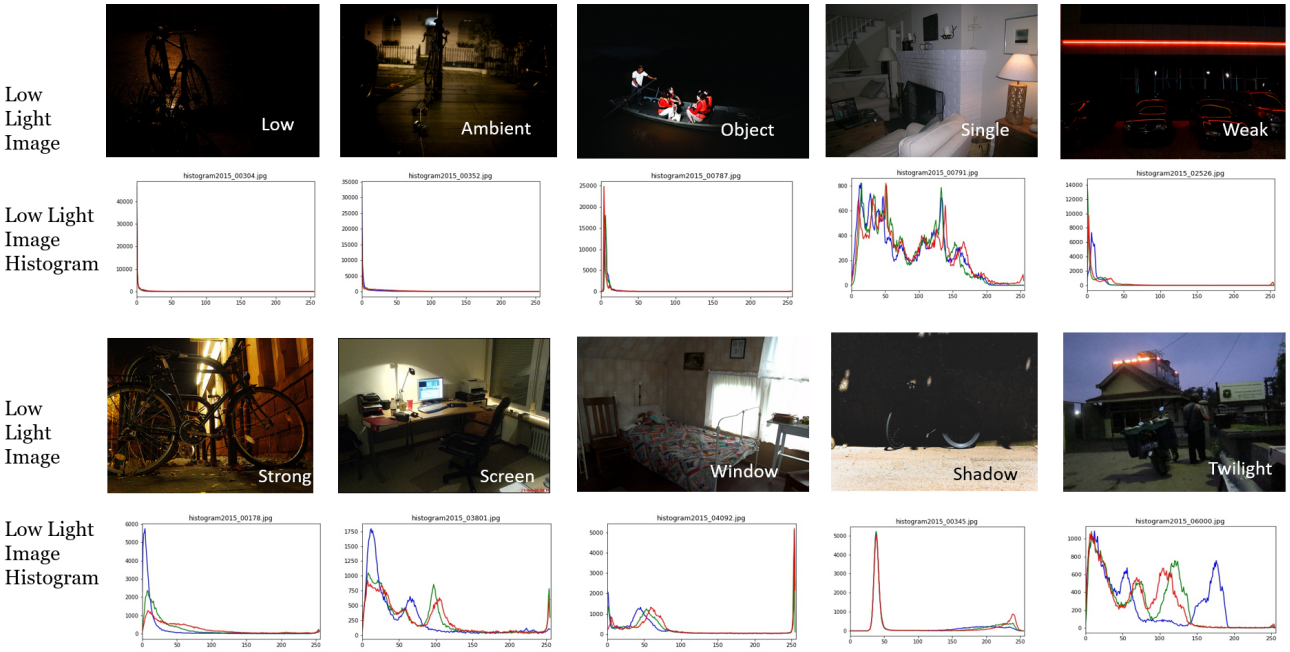


Fig. 1. ExDark Low-light images and its corresponding histogram. There are ten illumination types by human annotations namely low, ambient, object, single, weak, strong, screen, window, shadow and twilight. Each type of low-light image exhibit a different histogram behavior.

and *twilight*. Hence, another 3-dimensions are used to indicate the histograms' *modality*. Finally, for the overall feature, the intensity histograms for all three RGB channels are included as well to form a 789-dimension feature vector to represent the low-light images.

2) Clustering: The k-means algorithm is implemented to perform clustering with Euclidean Distance as the measure of dissimilarity in order to investigate the unsupervised clustering of low-light illumination types based on the formulated feature vector. The Elbow Approach is used to determine the best number of clusters that can be formed, based on Within-Cluster Sum of Square (WCSS) values as shown in Fig. 3. According to this assessment, 3 is found to be the optimal number of clusters based on the elbow of the curve which is a notable departure from the human categorized 10 illuminations. Thus, the comparative analysis in Section IV-A discusses the commonality and differences between these categorizations of illumination types.

C. Image Enhancement against Illumination Types

A main hypothesis in this research work is that current LLIE methods may lack of robustness towards illumination variations as there is a common general assumption that low-light itself is a "single" type of illumination. Thus, in order to explore and highlight potential key challenges in such models, state-of-the-art enhancement algorithms are implemented to enhance the ExDark images and evaluated both qualitatively and quantitatively, with respect to the 10 illumination types, as well as the 3 clusters found in Section III-B2. Specifically, the notable LIME [4] and Zero-DCE [13] are the selected methods for this evaluation, and the no-reference Image Quality Assessment (IQA) metric, BRISQUE [19] is adopted as the quantitative measure of enhanced image quality.

D. Object Classification against Illumination Types

Recent works in LLIE research also investigates their effectiveness at supporting downstream tasks like object classification and detection, but similarly, the impact of illumination types on such task has not been explicitly discussed [10], [20], [21]. Therefore, we approach this investigation by using object classification as the downstream task, and the classification accuracy as the task-driven metric to evaluate the influence of illumination as well as the effectiveness of the LLIE methods at supporting classification models in such conditions.

Particularly, the MobileNetv2 [22] classifier is selected as the model for object classification, following the 12 object classes provided in the ExDark. The experiments involve training 3 separate models of the classifier with different sets of images, namely, the original ExDark images, ExDark enhanced by LIME, and ExDark enhanced by Zero-DCE. Similar to the enhancement analysis, the classification performance of each of these models were scrunized with consideration of the 10 illuminations types provided by the ExDark, as well as the 3 cluster groups found in Section III-B2.

IV. Benchmarking and Analysis

This section provides the results of the comparative and empirical analyses, whereby the illumination types produced by clustering are contrasted with the 10 illumination types from the ExDark. This is followed by the empirical analysis on image quality enhancement using LIME and Zero-DCE and object classification performance of MobileNetV2 trained using different sets of training images. All of which will be discussed in view of the different illumination types.

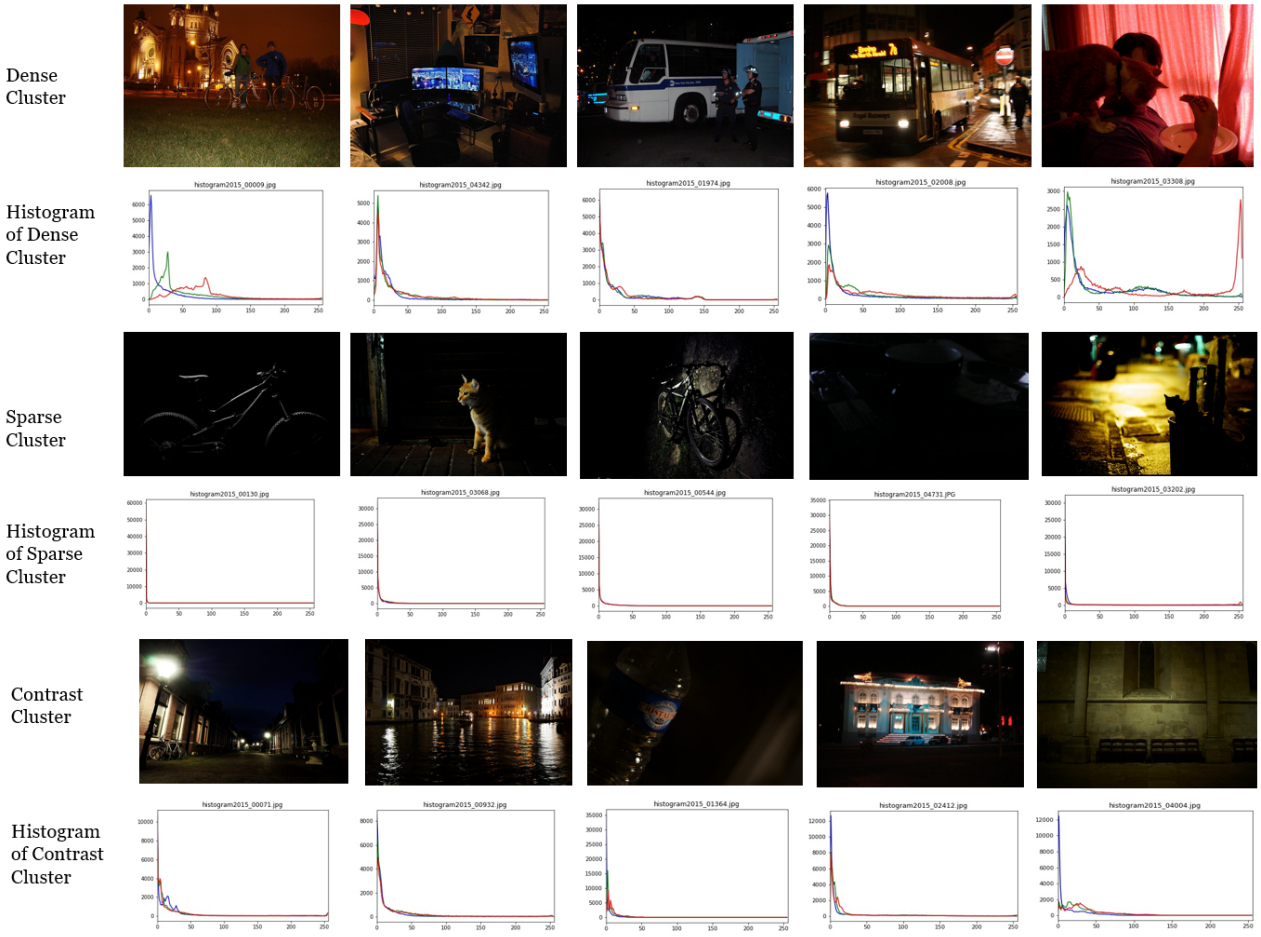


Fig. 2. Histogram of different clusters using ExDark Dataset.

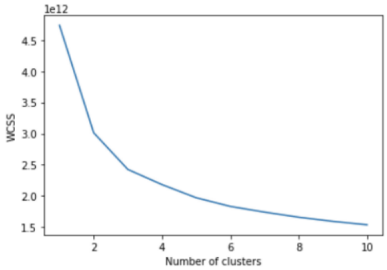


Fig. 3. ExDark Dataset is found to have 3 clusters based on the machine clustering illumination type using WCSS.

A. Comparative Analysis of Illumination Types

The 10 illuminations types specified in the ExDark dataset were manually categorized. Fig. 1 are selected samples of images from each illumination type with the respective histograms. It can be observed that there are certain patterns that distinguishes the types from one another. For instance, *low* and *ambient* types exhibit the highest bin sparsity and is positively skewed towards lower intensity values. The image itself exhibits with very few color shades, and the histograms of the RGB channels are very similar. On the other hand, *object, weak* and *strong* types also have high sparsity but with slightly wider distribution, and visually, the images have more color shades but still very similar histograms between the RGB channels. The *single* illumination type however has a histogram with notably wider distribution pattern, and visually more apparent colors though similar histogram patterns between the channels. This is similarly seen in the *twilight* type but the main distinguishing aspect is the significant differences between the histograms of the RGB channels. Lastly, the *shadow, screen* and *window* types has notable bimodal/multimodal histograms with distinct peaks on the two extremes of low and high intensity values.A lot more color shades can also be seen in such images as well for the parts with sufficient lighting.

Contrarily, the cluster groups formed by the k-means algorithm are more generalized in comparison. From observations, the clusters can be distinguished into *dense, sparse*, and *contrast* clusters as seen in the samples in Fig. 2. The largest cluster is the the *dense* cluster containing 4911 images and their histograms have relatively mixed patterns with wide distributions and the maximum frequency recorded in the histograms are lower in comparison to the images in the other two clusters. Moreover, images with marked differences in each of the RGB channel histograms are also clustered into this group. Visually, the images are relatively "bright" where the contents are still relatively easy to discern and more color shades can be observed. Comparing to the ExDark illumination types, it seems that the *single, strong, screen, window, shadow,*

and *twilight* types have been grouped into this cluster.

Separately, the *sparse* cluster is the smallest among the 3 clusters with only 620 images. Visually, the images in this cluster are darker and the objects can be "hidden" in the darkness, as shown in Fig. 2. The color histogram of these images has high bin sparsity, yet the maximum frequency of bins are the highest among the clusters. The histograms of the individual RGB channels are also very similar, signifying a lack of color shades as seen in the image samples. The characteristic of this cluster is very similar to the *low* and *weak* types in ExDark.

Lastly, the *contrast* cluster contains 1830 images with very similar histogram distributions to the *sparse* cluster but the maximum histogram frequency is lower and a slightly wider distribution. The images has more color shades as well, though there are still noticeably large dark regions. The *ambient* and *object* types appears to fall into this category. Overall, the clustering by k-means has some commonality with the types given by ExDark, but the groupings are less granular whereby the main segregation is mainly based on color shades and general content/object visibility as compared to more contextual categorization done by humans, e.g. twilight, screen, and shadow. This leads to the question, if there are notable differences in enhancement performance following these any of types of illumination.

B. Empirical Analyses of LLIE

1) Quantitative Assessment:: Table I shows the summary of the BRISQUE scores of the enhanced low-light images of the three clusters. Enhanced images from the *dense* cluster are found to have the best results irregardless of LIME or ZERO DCE as the enhancement approach as they have the lowest score across all statistics, implying best image quality as well as consistent enhancement results. This performance is followed by the *contrast* cluster and lastly the *sparse* cluster. We note that the quality of images from the *sparse* cluster are the worst with the average BRISQUE score that almost doubles that of the *dense* and *contrast* clusters. Moreover, the highest standard deviation also indicates that the scores of enhanced *sparse* cluster images have significant variations, implying an unstable enhancement outcome.

Conversely, Table II shows the summary of the BRISQUE comparison for the enhanced low-light images with respect to the 10 illumination types from ExDark. From these ten human annotated illumination types, the BRISQUE scores of the *low* type have the worst average score for both LIME and ZERO DCE enhancement. This is consistent with the results of the cluster groups as the *low* and *ambient* types are parts of the *sparse* cluster. Similarly, the enhanced results of the *weak* type are the third worst. The illumination type that records the best enhanced outcome is the *twilight* type which records the lowest BRISQUE scores for all statistical measures and approximately 2 times lower than the scores of enhanced images from the *low* and *ambient* types. This result is followed by the *strong* and *screen* types, both of which are also part of the *dense* cluster with the *twilight* type, thus concurring with the clusterring-based results in Table I.

However, an interesting finding is that the second worst enhancement comes from the *shadow* type images which falls into the *dense* cluster and mostly having bimodal histogram distribution, thus contradicting the results of Table I. This implies that there are more nuance in low-light illuminations that is not captured in histogram distributions, and the performance of image enhancement methods varies greatly depending on the illumination type.

As a whole, different types of low-light images have varying impact on the enhancement results as shown in Tables I and II. Both of the state-of-the-art enhancement models work better on images with more visible color shade and notably different histogram distributions in the RGB channels, e.g. images in the *dense* cluster or from the *twilight* and *strong* types. Images that have significant dark regions with less color shades and low difference in the RGB channel histograms record the worst enhancement results, with BRISQUE scores almost 2 times of the best enhanced group of images.

TABLE I
BRISQUE SCORE COMPARISON OF THREE CLUSTERS WITH KMEANS ALGORITHM USING EXDARK DATASET FOR DENSE CLUSTER, SPARSE CLUSTER AND CONTRAST CLUSTER.

Cluster (Number of images)	dense cluster (4911)		sparse cluster (620)		contrast cluster (1830)	
Algorithm	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE
Min	-28	-21	8	4	-8	-5
Max	102	99	183	179	136	142
Average	**31**	**30**	69	65	**39**	**38**
Std. Dev.	14	13	27	22	17	16

2) Qualitative Assessment: The qualitative evaluation using BRISQUE score has shown that the LIME and Zero-DCE shows varying enhancement performance when given images with differing types. This section takes a further look into the enhancement results to visually assess the image quality.

Fig. 4 shows visual samples of low-light images before and after enhancement using Zero-DCE and LIME. Notably, enhanced images from several illumination types have issues with enhanced artifacts, noise pixels, and over-enhancement. For instance, the samples of *low* and *object* types show significant noises after enhancement, whereas the *window* type is over-enhanced, and the *screen* example has lingering artifacts.

While different illumination types seem to exhibit quality issues at random, the general observations are that, enhancing significantly low-light images such as *low* commonly "reveals" noises and artifacts. This is also true for illumination types where the images are only partially dark but the intensities of those dark regions are severely low, such as *screen, window*, and even *shadow*. Thus the observation in Table II where *shadow* types recorded second worst average quality is due to the darker regions that affected the enhancement results.

On the contrary, enhancement results of images that has more color shades originally, as opposed to mostly dark regions, tend to yield better visual results. This is typically

TABLE II
BRISQUE SCORE COMPARISON OF THE ORIGINAL TEN ILLUMINATION TYPES USING EXDARK DATASET FOR LOW, AMBIENT, OBJECT, SINGLE AND WEAK ILLUMINATION TYPES.

	Low		Ambient		Object		Single		Weak	
Total Images	241		2098		498		762		704	
	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE
Min	14.6	14.5	-17.0	-14.5	-3.6	-1.3	-3.7	-3.7	0.3	3.05
Max	183.3	139.1	179.2	179.4	124.8	119.3	136.8	133.6	155.9	125.9
Average	54.8	50.8	37.8	36.9	35.7	33.6	37.8	36.9	39.3	38.5
Standard Deviation	23.8	21.7	19.9	18.0	20.4	18.7	20.0	18.2	20.5	17.9
	Strong		Screen		Window		Shadow		Twilight	
Total Images	2056		165		363		56		416	
	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE	LIME	ZERO DCE
Min	-27.6	-21.0	1.82	2.6	-3.8	3.0	7.2	5.0	-12.5	-10.3
Max	142.5	121.2	95.3	103.8	178.7	159.0	121.4	117.1	81.3	72.9
Average	32.4	31.3	33.3	33.3	38.0	37.7	40.8	38.4	**23.4**	**24.2**
Standard Deviation	15.8	14.0	15.2	103.8	21.8	19.8	25.9	24.2	13.2	12.9

observable in the *strong* type of images but other types may have similar results provided there is sufficient color.

In terms of the methods, the results of Zero-DCE tend to have less color and are whitish compared to LIME. Contrasting, LIME enhanced images tend to have more color but at the same time more noise as well. Nonetheless, both methods perform well on image types with more color originally like *twilight* and *strong* types but the quality deteriorates for darker types like *low*.

3) Object Image Classification: This section details the analysis using image classification accuracy as a task-driven metric to measure the effectiveness of image enhancement at supporting object classification performance in varying illumination conditions.

Fig. 5 illustrates the classification performance of the MobileNetV2 at classifying objects when trained using the original ExDark images, as well as images enhanced by LIME and Zero-DCE and tested on the respective test sets. The bars show the object classification accuracy sorted by the cluster groups. From this visualization, it can be seen that object classification accuracy is consistently highest for images from the dense cluster followed by the contrast cluster and lastly the sparse cluster, regardless if the images are enhanced or not. This observation is in line with the quality evaluations, however, it should be noted that the performance differences are not significant between the clusters. Thus, implies that the illumination may not have a significant impact.

Following that, Fig. 6 shows the same object classification results but with respect to the 10 illumination types of the ExDark. It can be clearly seen that the object classification accuracy varies significantly based on the image illumination type. Images from the *twilight* type is found to have the highest accuracy with an average of 89%, whereas *screen* and *single* are shown to have the lowest accuracy with averages of 50% to 60% respectively. This result clearly shows the significant impact that illumination types have on downstream task, as well as the relatively weak effect the enhancement of LIME and Zero-DCE has on the performance.

From these results, it can be stated that the current state-of-the-art image enhancement models are not optimized to handle different types of low-light images, especially those with high sparsity in the histogram pattern. This include images with large number of pixels that are dark and lack colors.

V. OPEN ISSUES

Based on the analyses, we present the following open issues in low-light research that we believe warrants attention:

1) **Over generalization of low-light**. Although significant progress has been made in various areas of low-light vision research in recent years, low-light is still addressed in a singular categorization which under-represents the various characteristics that are imperative to various tasks, including enhancement and recognition. This can be seen from the various notable datasets which provides "uniform low-light" images by varying exposures, as well as approaches utilizing synthetic data generated using power functions, both of which do not represent the diverse characteristics of low-light environments.
2) **Lack of low-light characteristics awareness in enhancement frameworks**. Due to the common over-generalization of low-light, image enhancement methods are often not adaptable towards specific illumination types. In our analyses, despite localized pixel-level enhancement estimations, current enhancement methods are unable to handle the "cross-illumination differences" in the images, thus resulting in sub-par enhancement results, like noise amplification as well as over- and under-enhancement for particular types of low-light images.
3) **Bottleneck for downstream tasks**. Downstream tasks like object classification are still under-performing in low-quality conditions like low-light, mainly due to the lack of data to train a sufficiently reliable model. The support of enhancement as pre-processing is still inferior in comparison to the outstanding performance in regular light conditions, especially considering the more "severe" low-light types. Hence, there is a need for more work to address the underperformance of enhancement methods in these cases to overcome this bottleneck.



Fig. 4. Visual results of original ExDark Dataset, ZERO DCE and LIME on Exark Dataset's ten types of low-light images namely low, ambient, object, single, weak, strong, screen, window, shadow and twilight

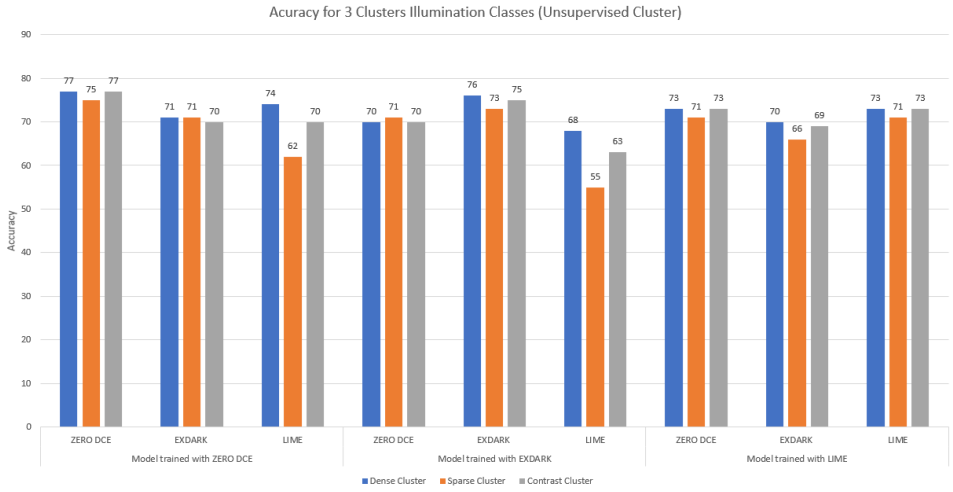


Fig. 5. ExDark Low-light images object classification with KMeans clusters (Machine)

VI. Conclusion

Low-light vision is a challenging topic with notable gaps still yet to be addressed. In this paper, we conduct several analyses to investigate and validate the presence of diverse illumination types in low-light environment. It is found that low-light can be diversified into various types and they have notable impact towards image enhancement performance and subsequently affect downstream tasks. Thus, we discuss several open issues that is worth further attention that can potentially push forward the research in this area.

References

[1] E. H. Land, "The retinex theory of color vision.," *Scientific American*, vol. 237 6, pp. 108–28, 1977.

[2] G. D. Hines, Z.-u. Rahman, D. J. Jobson, and G. A. Woodell, "Single-scale retinex using digital signal processors," 2005.

[3] D. Jobson, Z. Rahman, and G. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

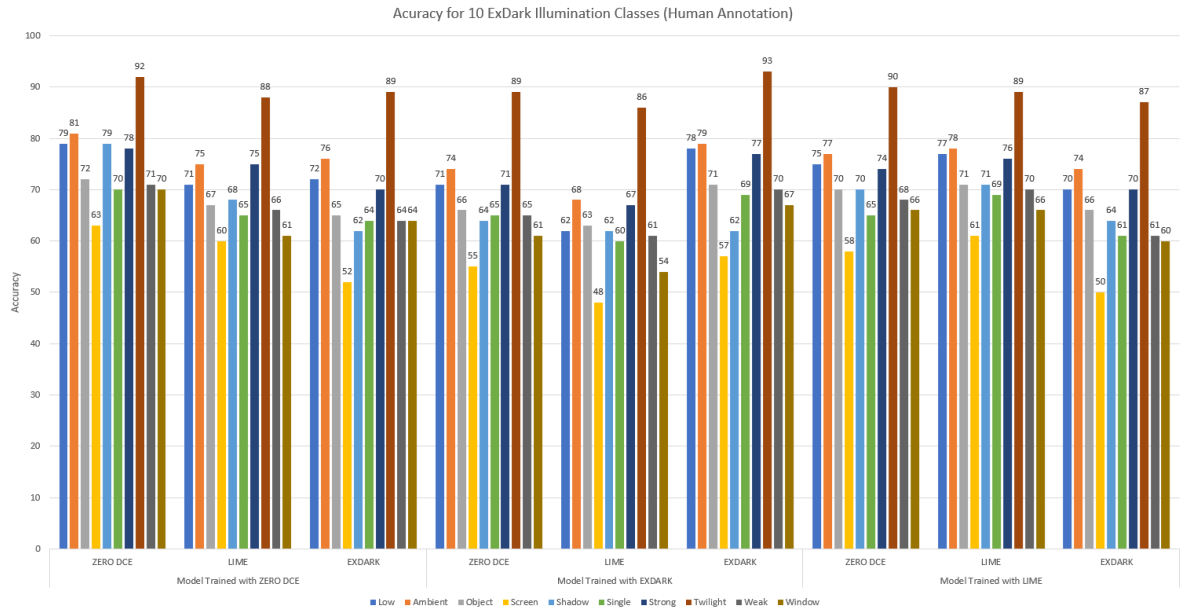


Fig. 6. ExDark Low-light images object classification with ExDark Ten Illumination types (Human Annotation).

[4] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Trans. on Image Processing*, vol. 26, pp. 982–993, 2017.

[5] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *ArXiv*, vol. abs/1511.03995, 2017.

[6] F. Lv, F. Lu, J. Wu, and C. S. Lim, "Mbllen: Low-light image/video enhancement using cnns," in *BMVC Conference*, 2018.

[7] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *AAAI Conference on Artificial Intelligence*, 2020.

[8] F. Zhang, Y. Li, S. You, and Y. Fu, "Learning temporal consistency for low light video enhancement from single images," in *IEEE/CVF CVPR*, 2021, pp. 4965–4974.

[9] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4227–4240, 2021.

[10] Y. Jiang, X. Gong, D. Liu, *et al.*, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Trans. on Image Processing*, vol. 30, pp. 2340–2349, 2019.

[11] J. Hu, X. Guo, J. Chen, G. Liang, F. Deng, and T. L. Lam, "A two-stage unsupervised approach for low light image enhancement," *IEEE Robotics and Automation Letters*, vol. 6, pp. 8363–8370, 2021.

[12] Y.-S. Chen, Y. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," *IEEE/CVF CVPR*, pp. 6306–6314, 2018.

[13] C. Guo, C. Li, J. Guo, *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," *IEEE/CVF CVPR*, pp. 1777–1786, 2020.

[14] L. Zhang, L. Zhang, X. Liu, Y. Shen, S. Zhang, and S. Zhao, "Zero-shot restoration of back-lit images using deep internal learning," *ACM Multimedia*, 2019.

[15] W. Chen, W. Wenjing, Y. Wenhan, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC Conference*, 2018.

[16] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *IEEE/CVF CVPR*, 2018, pp. 3291–3300.

[17] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[18] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[19] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[20] Y. P. Loh, X. Liang, and C. S. Chan, "Low-light image enhancement using gaussian process for features retrieval," *Signal Process. Image Commun.*, vol. 74, pp. 175–190, 2019.

[21] W. Wang, W. Yang, and J. Liu, "Hla-face: Joint high-low adaptation for low light face detection," in *IEEE/CVF CVPR*, 2021, pp. 16 195–16 204.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE/CVF CVPR*, 2018, pp. 4510–4520.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

MoMo Strategy: Learn More from More Mistakes

Sophia Chulif*†, Sue Han Lee*, Yang Loong Chang†,
Mark Tee Kit Tsun*, Kok Chin Chai† and Yi Lung Then‡
* Swinburne University of Technology Sarawak Campus, Malaysia
E-mail: schulif@swinburne.edu.my, shlee@swinburne.edu.my, mtktsun@swinburne.edu.my
† NEUON AI SDN. BHD., Malaysia E-mail: yangloong@neuon.ai, kc@neuon.ai
‡ Universiti Malaysia Sarawak, Malaysia E-mail: ylthen@unimas.my

***Abstract*—Training accurate convolutional neural networks (CNNs) is essential for achieving high-performance machine learning models. However, limited training data pose a challenge, reducing model accuracy. This research investigates the selection and utilization of misclassified training samples to enhance the accuracy of CNNs where the dataset is long-tail distributed. Unlike classical resampling methods involving oversampling of tail classes and undersampling of head classes, we propose an approach that allocates more misclassified training samples into the training process to learn more (namely, MoMo strategy), with ratios of 50:50 and 70:30 for the wrongly predicted and correctly predicted samples, respectively. Additionally, we propose incorporating a balanced sample selection method, whereby the maximum training sample per class in an epoch is assigned to address the long-tail dataset problem. Our experimental results on a subset of the current largest plant dataset, PlantCLEF 2023, demonstrate an increase of 1%-2% in overall validation accuracy and a 2%-5% increase in tail class identification. By selectively focusing on more misclassified samples in training, at the same time, integrating a balanced sample selection achieves a significant boost in accuracy compared to traditional training methods. These findings emphasize the significance of adding more misclassified samples into training, encouraging researchers to rethink the sampling strategies before implementing more complex and robust network architectures and modules.**

I. Introduction

Limited training samples are a common challenge in real-world machine-learning applications such as in face [1], traffic scene [2], medical image [3], and species recognition [4]. In reality, training data is typically scarce and follow a long-tail distribution. This distribution has fewer classes with many training samples and many classes with fewer training samples. When lacking training data, models may struggle to generalize and accurately classify new or unseen instances. Similarly, in datasets with a long tail distribution, the model may output higher prediction scores for the classes with many training samples (head classes) but lower prediction scores for the classes with fewer training samples (tail classes) [5]. This results in models being biased towards the head classes, often leading to poor performance in the tail classes, or rather the minority and rare.

In the context of plant identification, deep learning has shown significant progress in recent years and even performed comparably with human experts [6]. Nonetheless, this classification is supplied with ample training samples representing only tens of thousands of species, whereas the global plant species consists of approximately four hundred thousand [7].

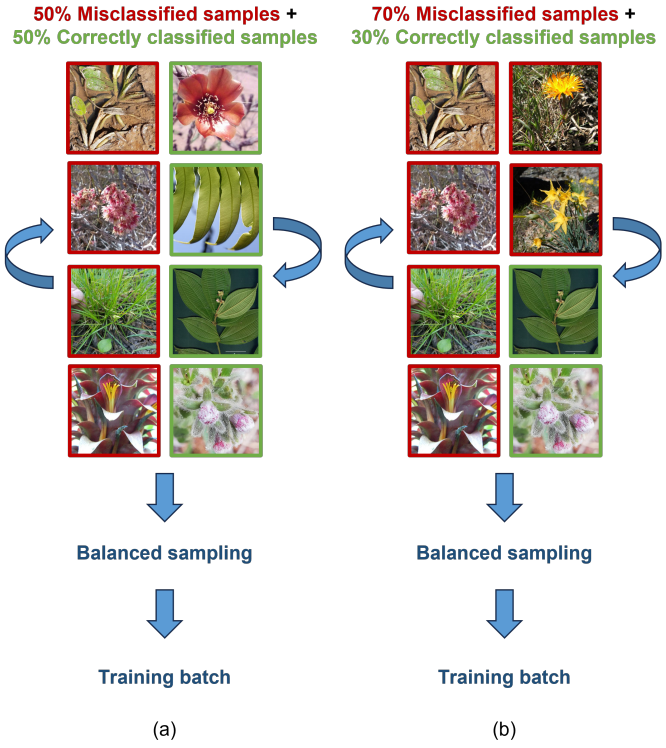


Fig. 1. An overview of the two different MoMo strategies experimented on. (a) illustrates the allocation of 50% misclassified training samples and 50% correctly classified training samples. (b) represents the allocation of 70% misclassified training samples and 30% correctly classified samples. Each batching will be further selected according to the limit set to ensure the samples are balanced in the epoch and finally used in training.

Most plant species universally are less observed and significantly lower in samples as they are located in remote areas that are difficult to access. Consequently, rare or new plant species are commonly under-represented in training data, making them difficult to be identified. It was shown in [7] where the training dataset size used is significantly lower per species and suffered from a higher degree of noise, accurately representing the challenges in global species identification.

Various methods have been proposed to address this long-tailed class imbalance issue, which includes class re-balancing [8], [9], information augmentation [10], and module improvement [11]. However, in the field of plant classification, more

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

work has yet to be explored to tackle this problem. Many researchers focused on feature extraction approaches [12], [13], altering model loss functions [14]–[16], data augmentation [17], [18], and using varying network architectures [19]–[21] but not directly taking on the long-tailed dataset itself at the sample level despite it being the most straightforward approach among them.

In this paper, we conduct extensive experiments on a subset of the current largest plant dataset, PlantCLEF 2023, where we randomly selected ten thousand species to investigate the significance of devising the batching samples in training long-tailed deep plant models. Classical resampling methods involve the random over-sampling (ROS) and random under-sampling (RUS) techniques where ROS repeats the samples of the tail classes in training, and RUS removes the samples of the head classes in training [22]. Although these methods ultimately result in good predictions, ROS might overfit the tail classes, while RUS might underfit the whole data distribution, reducing the performance of the head classes [23].

Conversely, instead of resampling or re-balancing the head and tail classes, we propose to use more misclassified training samples to allow the model to learn more from its mistakes, hereinafter referred to as MoMo strategy. Fig. 1 illustrates the overview of the strategy. We hypothesize that giving more attention to misclassified instances increases the representation of the difficult samples, allowing the model to allocate more learning resources to those instances. It can be particularly beneficial for the rare or underrepresented classes, as the model gets more opportunities to refine its decision boundaries and better capture their intricacies. Besides this, we hypothesize that it can help the model to focus on the regions of the feature space where it struggles the most, promoting better generalization and reducing the risk of overfitting the majority classes. In addition, we propose to set a limit to the maximum class samples for training to study the effects of MoMo strategy in a balanced and unbalanced scenario.

Our experiments show that our proposed MoMo strategy with a balanced class distribution achieves a significant boost in accuracy compared to traditional batching methods, particularly in tail class recognition. This finding emphasizes the importance of adding misclassified samples to enhance the training of deep learning models. It provides new experimental evidence and discussion for the need to leverage misclassified samples in addition to implementing more intricate approaches, such as network alteration and module improvement, when addressing the curse of long-tail datasets.

II. Methodology

A. Network Setup

The networks implemented in this work are classical CNNs optimized by a softmax loss function. They are constructed based on the Inception-ResNet-v2 [24] model and pre-trained from the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012) dataset [25] available in the TensorFlow-Slim image classification library [26]. We used a batch size of 128, an initial learning rate of 0.00001 for the front layers, 0.0001 for the classification layers, an Adam optimizer, a weight decay of 0.00004, and an image input size of 299 $\times$ 299 $\times$ 3. Inception-ResNet-v2 is adopted as our backbone (base) network because it achieved the lowest Top-1 Error and Top-5 Error on the non-blacklisted subset of the validation set of ILSVRC 2012, when proposed in [24]. In addition, the Inception-ResNet-v2 model is one of the highest performing and lighter pre-trained models in the TensorFlow-Slim image classification library without significantly compromising the accuracy. Lastly, it is among the top-performing models in the plant identification challenges, PlantCLEF 2017 [27], PlantCLEF 2019 [7], PlantCLEF 2020 [28], PlantCLEF 2021 [29], and PlantCLEF 2022 [30].

B. Training Strategies

Our training strategies emphasize the incorporation of misclassified training samples during model training. To implement our approach, we devise a sample selection strategy for each training epoch. We assign a ratio for misclassified and correctly classified samples, ensuring they receive suitable focus and attention in training. This prioritization is achieved by hard-coding the sampling strategy to include more misclassified instances than correctly classified samples. The ratios we experimented on are 50% and 70% for misclassified samples. The prioritized misclassified samples are then combined with correctly classified samples (50% and 30%) during each epoch to form each training batch. In addition, to maintain a balanced sample selection without over-picking the head classes, the training samples per class were limited to a maximum of 16 in each epoch. It aims to prevent classes with more samples from being selected more frequently to minimize class bias. In short, two kinds of experimental training strategies are carried out, one on the ratio of misclassified samples incorporated in the training and the other with or without a balanced sample selection between classes. Moreover, we employ data augmentation techniques to enhance the learning process further. Data augmentation introduces variations and perturbations to the samples, enabling the model to understand better and generalize from these challenging instances.

C. Inference Method

The model predictions are obtained using the softmax function where it transforms the output of the previous layers to the final layer into a probability distribution over the total classes. It can be represented in (1)

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \tag{1}$$

where p_i denotes the probability of the input belonging to the *i*-th class, e^{z_i} represents the exponential function applied to the *i*-th element of the input vector, z_i. The exponential function raises Euler's number (e) to the power of z_i, resulting in a positive value. Meanwhile, $\sum_{j=1}^{n} e^{z_j}$ denotes the sum of the exponential values across all elements of the input vector. It calculates the sum of the exponential values for each z_j. After

TABLE I
DETAILS OF THE TRAINING AND VALIDATION DATASETS.

Dataset	Number of species	Number of images
Training (Unbalanced)	10,000	298,669
Training (Balanced)	10,000	128,040
Validation (Overall)	9,402	9,402
Validation (Head)	1,238	1,238
Validation (Middle)	2,163	2,163
Validation (Tail)	6,001	6,001

TABLE II
THE LONG-TAIL TRAINING DATASET DISTRIBUTION.

Category	Number of training samples	Number of species
Tail	5 - 10	3,091
	11 - 20	2,201
	21 - 30	1,302
Middle	31 - 40	751
	41 - 50	603
	51 - 60	449
	61 - 70	364
Head	71 - 80	292
	81 - 495	495
	91 - 112	452

obtaining the probability of the predicted class, we used the Top-1 and Top-5 accuracy to measure the performance of the model. Prior to obtaining the prediction probability, we crop the target images into multiple crops, specifically ten crops (center, top-left, top-right, bottom-left, bottom-right, and each flipped horizontally) instead of a single crop during inference.

D. Dataset

The experiments are performed on a subset of the largest plant benchmark dataset, PlantCLEF 2023. The original dataset contains 4 million images of eighty thousand species from trusted academic sources and the web. However, we randomly selected only ten thousand species from their trusted dataset to reduce the computing resources needed. We also ensured that the dataset followed a long-tail distribution by confirming that the classes have a minimum of 5 training image samples and a maximum of 112 training image samples per class. Consequently, the training dataset totaled 298,669 images, while the validation totaled 9,408. On the other hand, when the training samples are limited to 16 per class in each epoch, the training dataset totaled 128,040. In addition, we grouped the validation dataset into head, middle, and tail categories to examine the performance of each category. The head category consists of classes with 71 to over 100 training samples; the middle category consists of classes with 31 to 70 training samples; and the tail consists of classes with 5 to 30 images. The training and validation datasets are detailed in Table I, while the long-tail training dataset and its head, middle, and tail classes are presented in Table II.

III. RESULTS AND ANALYSIS

A. Experimented Models

We conducted experiments on the MoMo strategy with two different misclassified sample ratios: 50% and 70%. Furthermore, we examined the difference between a balanced and unbalanced sample selection between classes in these settings. A balanced sample selection would have a maximum of 16 training samples per class in each epoch. In contrast, the unbalanced sample selection would use all the available training samples in each epoch. We describe our models as follows:

Model 1: The baseline model of our experiments. It is set up with the settings described in Section II. In contrast with the rest, the baseline is trained without the MoMo strategy and with an unbalanced strategy. The training samples are sampled randomly as conventional CNN training, where the head classes are naturally chosen more than the tail classes.

Model 2: This model is set up with the settings described in Section II, trained with MoMo (50% misclassified samples and 50% correctly classified samples), and a maximum of 16 training samples per class in each epoch.

Model 3: This model is set up with the settings described in Section II, trained with MoMo (70% misclassified samples and 30% correctly classified samples), and a maximum of 16 training samples per class in each epoch.

Model 4: This model is set up with the settings described in Section II and trained with MoMo (50% misclassified samples and 50% correctly classified samples) using all the available training samples per class in each epoch. Likewise, in the baseline model, the head classes are sampled more than the tail classes.

Model 5: This model is set up with the settings described in Section II and trained with MoMo (70% misclassified samples and 30% correctly classified samples) using all the available training samples per class in each epoch. Likewise, in the baseline model, the head classes are sampled more than the tail classes.

B. Results of the overall validation set

As shown in Table III, the MoMo strategy generally improves the overall plant classification to 1% - 2% in Top-1 accuracy when comparing the baseline model to models 2, 3, and 4. Fig. 2 shows the baseline model's mistakes corrected with the MoMo strategy. The misclassified validation samples show similar features to its training samples and should be easily corrected when fed more to the model. On the other hand, model 5 with 70% MoMo and an unbalanced approach showed no improvements in both Top-1 and Top-5 accuracies. We suggest that the extended 70% misclassified samples may not contain enough significant features for the overall identification and could be erroneous, as shown in Fig. 3 where the extended misclassified samples mainly constitute non-plant and plant images that are visibly unclear. In addition, MoMo with balanced strategies (models 2 and 3) also improved the Top-5 accuracy compared to the baseline model despite being trained with less training data. It demonstrates the ability of MoMo to generalize class features sufficiently with lower-class samples. Nonetheless, MoMo with an unbalanced strategy (models 4 and 5) does not improve the Top-5 accuracy for the overall validation set but is not too far off from the baseline model.

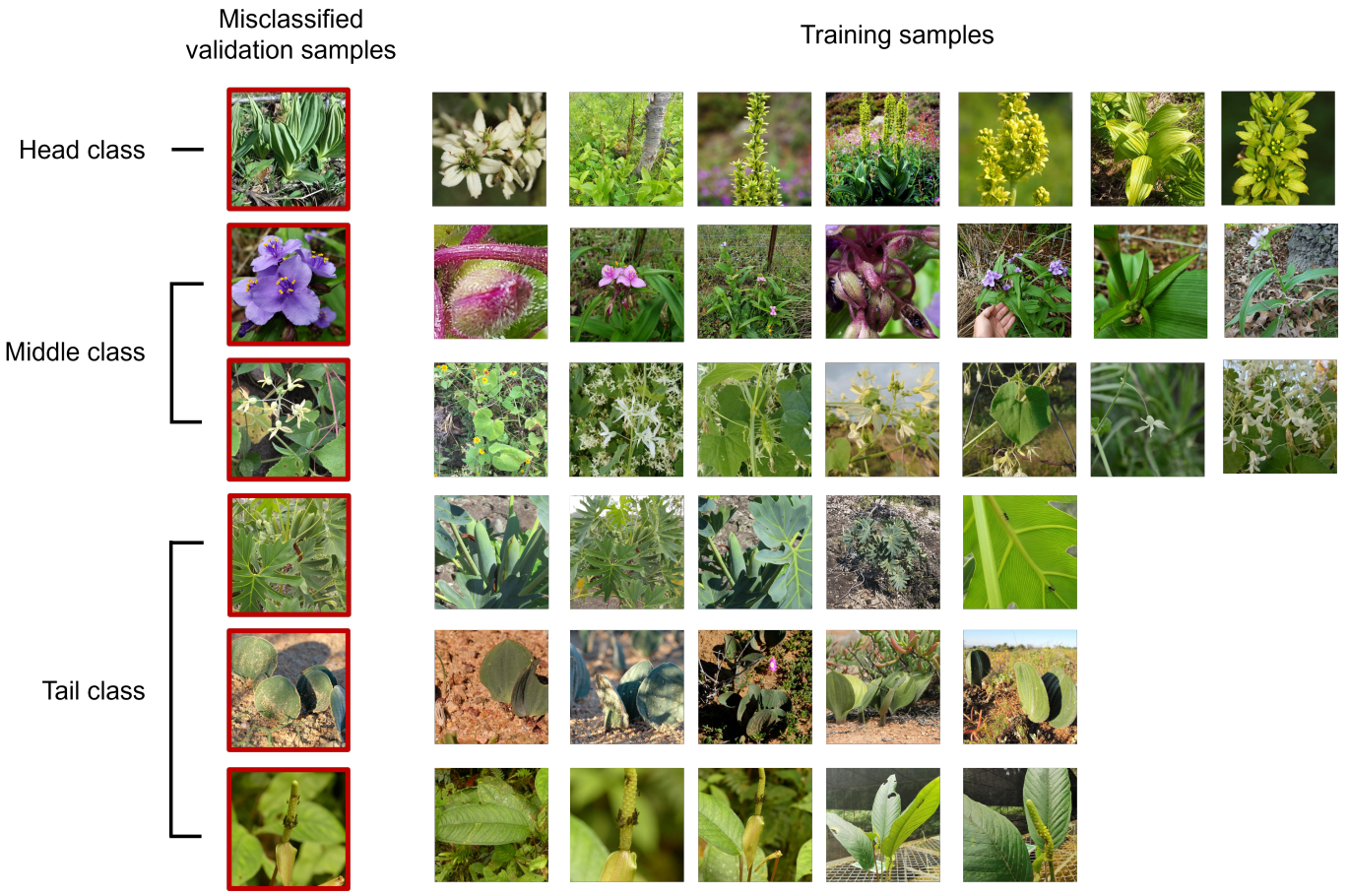


Fig. 2. Several head, middle, and tail class validation samples misclassified by the baseline model but correctly classified by the models with MoMo strategies. The right column of images without red borders represents their respective training samples and features in the dataset.

TABLE III
OVERALL RESULTS OF THE MODELS IN THE VALIDATION SET.

Model	Strategy	Top-1		Top-5	
1	Baseline, + unbalanced	0.6082	5718 / 9402	0.7964	7488 / 9402
2	**50% MoMo, + balanced**	**0.6209**	**5838 / 9402**	**0.7973**	**7496 / 9402**
3	70% MoMo, + balanced	0.6139	5772 / 9402	0.7966	7490 / 9402
4	50% MoMo, + unbalanced	0.6113	5747 / 9402	0.7924	7450 / 9402
5	70% MoMo, + unbalanced	0.6055	5693 / 9402	0.7899	7427 / 9402

TABLE IV
RESULTS OF THE MODELS BASED ON THE HEAD CLASSES IN THE VALIDATION SET.

Model	Strategy	Top-1		Top-5	
1	Baseline, + unbalanced	0.6971	863 / 1238	0.8611	1066 / 1238
2	50% MoMo, + balanced	0.6341	785 / 1238	0.8433	1044 / 1238
3	70% MoMo, + balanced	0.6446	798 / 1238	0.853	1056 / 1238
4	50% MoMo, + unbalanced	0.71	879 / 1238	0.87	1077 / 1238
5	**70% MoMo, + unbalanced**	**0.7132**	**883 / 1238**	**0.8708**	**1078 / 1238**

C. Results of the head classes

In contrast to the above, Table IV shows that MoMo with unbalanced class selection (models 4 and 5) outperforms MoMo with balanced class selection (models 2 and 3) as well as the baseline in both the Top-1 and Top-5 accuracies. It is expected as the unbalanced class selection method would naturally pick the head classes more during training, resulting in a higher head class performance. In other words, the unbalanced class selection MoMo model has seen more of the training data, therefore, it performs better than the balanced class selection MoMo model in this case. By comparing all the unbalanced models (models 1, 4 and 5), we can also confirm the advantage of implementing MoMo strategy in the head classes as it improves the accuracies of models 4 and 5 compared to the baseline.

D. Results of the middle classes

Similarly to the head class performances, MoMo with unbalanced class selection (models 4 and 5) generally performs best in both the Top-1 and Top-5 accuracies in the middle classes,

TABLE V
RESULTS OF THE MODELS BASED ON THE MIDDLE CLASSES IN THE VALIDATION SET.

Model	Strategy	Top-1		Top-5	
1	Baseline, + unbalanced	0.7	1514 / 2163	0.8451	1828 / 2163
2	50% MoMo, + balanced	0.705	1525 / 2163	0.8479	1834 / 2163
3	70% MoMo, + balanced	0.6828	1477 / 2163	0.8497	1838 / 2163
4	**50% MoMo, + unbalanced**	**0.7258**	**1570 / 2163**	0.8687	1879 / 2163
5	70% MoMo, + unbalanced	0.7157	1548 / 2163	**0.8692**	**1880 / 2163**

TABLE VI
RESULTS OF THE MODELS BASED ON THE TAIL CLASSES IN THE VALIDATION SET.

Model	Strategy	Top-1		Top-5	
1	Baseline, + unbalanced	0.5377	3227 / 6001	0.7404	4443 / 6001
2	**50% MoMo, + balanced**	**0.5879**	**3528 / 6001**	**0.7695**	**4618 / 6001**
3	70% MoMo, + balanced	0.5827	3497 / 6001	0.7659	4596 / 6001
4	50% MoMo, + unbalanced	0.5496	3298 / 6001	0.7489	4494 / 6001
5	70% MoMo, + unbalanced	0.5436	3262 / 6001	0.7447	4469 / 6001

as shown in Table V. It is also foreseen as the difference in the number of class samples between the middle and head classes is close to one another. Consequently, the middle class would be picked sufficiently, like in the head class. Models 2 and 3, with a balanced sample selection strategy, would perform slightly worse than the unbalanced strategy (models 4 and 5) because Models 2 and 3 are trained with reduced training samples.

E. Results of the tail classes

From Table VI, we see that the MoMo strategy significantly improves the tail class identification in both the Top-1 and Top-5 accuracies for the balanced and unbalanced sample selection strategies. Notably, with the balanced sample selection method (models 2 and 3), the accuracies increased 5% and 2% in Top-1 and Top-5 accuracies compared to the baseline model. On the other hand, without a balanced sample selection method (models 4 and 5), the MoMo strategy can only increase the performance up to 1%. Therefore, we can conclude that MoMo with a balanced sample selection strategy is essential and most effective for the tail class identification.

IV. CONCLUSION

This research aims to investigate the effects of incorporating more misclassified training samples into the training of a deep-learning plant model. Our findings indicate that this approach, namely MoMo, effectively improved the classification accuracy and generalization capabilities in the overall species identification. In addition, implementing a balanced sample selection batching strategy minimizes the impact of class imbalance during the training process, resulting in a more robust and accurate model that is less biased towards the head classes and enhances the tail classes. Combining both approaches led to further improvements in the model's performance, particularly in the tail classes. The MoMo strategy addresses the difficult feature representation during training, while the balanced batching method ensures a balanced local distribution of instances within each mini-batch or epoch. Notably, the unbalanced strategy works well with the head and middle classes, while the balanced strategy works well with the overall and tail classes. It is expected since the unbalanced class selection method can choose the head and middle classes more during training, resulting in a higher head and middle class performance. Nevertheless, this combined approach of MoMo and balanced strategy effectively maximizes the potential of the deep learning model and enhances its ability to learn and discriminate between different classes straightforwardly.

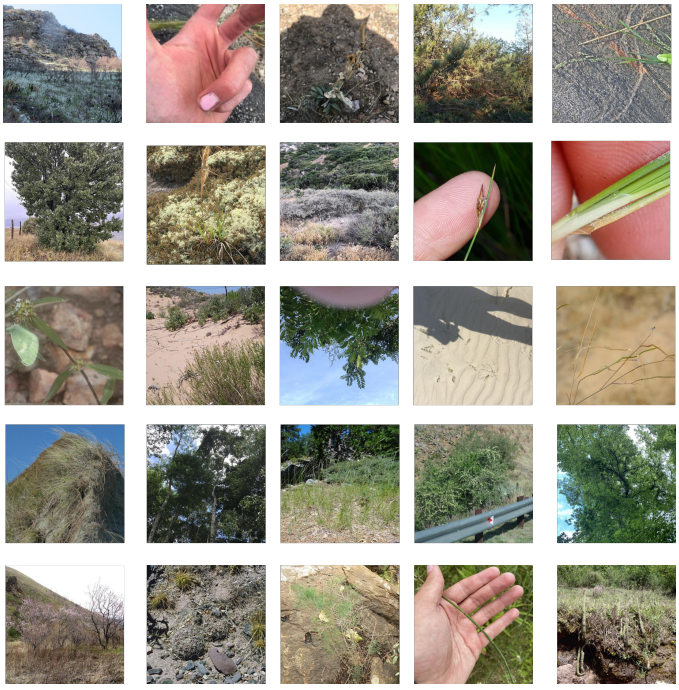


Fig. 3. The common misclassified samples in training and their similar attributes depicting non-plant images and plant samples that are difficult to view.

The findings of this study can be applied to various real-world applications where imbalanced datasets are common, particularly for long-tail distributed datasets. Future research can explore additional techniques for incorporating misclassified samples and advanced batching strategies into existing network architectures. Overall, this research sheds light on the advantages of simply utilizing the existing training samples to further improve the performance of the deep learning models. It encourages the need to leverage misclassified samples before implementing more complex and robust models.

ACKNOWLEDGMENT

This research was supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant No. (Ref:FRGS/1/2021/ICT02/SWIN/03/2), from the Ministry of Higher Education Malaysia and NEUON AI SDN. BHD., Malaysia.

REFERENCES

[1] Y. Ma, M. Kan, S. Shan, and X. Chen, "Learning deep face representation with long-tail data: An aggregate-and-disperse approach," *Pattern Recognition Letters*, vol. 133, pp. 48–54, 2020.

[2] J. Li, H. Zhang, and Y. Li, "Parallel detection architecture for long-tail problem in intelligent transportation systems," in *2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI)*, IEEE, 2021, pp. 238–241.

[3] A. G. Roy, J. Ren, S. Azizi, *et al.*, "Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions," *Medical Image Analysis*, vol. 75, p. 102 274, 2022.

[4] J. Kim, S. Woo, B. Park, and C. Kim, "Temporal flow mask attention for open-set long-tailed recognition of wild animals in camera-trap images," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 2152–2156.

[5] Y. Li, T. Wang, B. Kang, *et al.*, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 991–11 000.

[6] P. Bonnet, H. Goëau, S. T. Hang, *et al.*, "Plant identification: Experts vs. machines in the era of deep learning," in *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, and P. Bonnet, Eds. Cham: Springer International Publishing, 2018, pp. 131–149, ISBN: 978-3-319-76445-0. DOI: 10.1007/978-3-319-76445-0_8. [Online]. Available: https://doi.org/10.1007/978-3-319-76445-0_8.

[7] H. Goëau, P. Bonnet, and A. Joly, "Overview of lifeclef plant identification task 2019: Diving into data deficient tropical countries," in *CLEF 2019 Working Notes-Conference and Labs of the Evaluation Forum*, vol. 2380, 2019, p. 247.

[8] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2019.07.070. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519306838.

[9] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Information Sciences*, vol. 477, pp. 47–54, 2019, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2018.10.029. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025518308478.

[10] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2970–2979.

[11] Y. Zhang, B. Hooi, L. Hong, and J. Feng, "Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 077–34 090, 2022.

[12] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, and D. Mahajan, "Clusterfit: Improving generalization of visual representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6509–6518.

[13] Y. Du, J. Shen, X. Zhen, and C. G. Snoek, "Superdisco: Super-class discovery improves visual recognition for the long-tail," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 944–19 954.

[14] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[15] Z. Chen, J. Duan, L. Kang, and G. Qiu, "Class-imbalanced deep learning via a class-balanced ensemble," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 10, pp. 5626–5640, 2021.

[16] C. Garcin, M. Servajean, A. Joly, and J. Salmon, "Stochastic smoothing of the top-k calibrated hinge loss for deep imbalanced classification," in *International Conference on Machine Learning*, PMLR, 2022, pp. 7208–7222.

[17] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 694–710, ISBN: 978-3-030-58526-6.

[18] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu, "Rsg: A simple but effective module for learning imbalanced datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3784–3793.

[19] Y. Chen, J. Song, and M. Song, "Hierarchical gate network for fine-grained visual recognition," *Neurocomputing*, vol. 470, pp. 170–181, 2022.

[20] X. Du, T.-Y. Lin, P. Jin, *et al.*, "Spinenet: Learning scale-permuted backbone for recognition and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 592–11 601.

[21] M. Xu, S. Yoon, Y. Jeong, J. Lee, and D. S. Park, "Transfer learning with self-supervised vision transformer for large-scale plant identification," in *International conference of the cross-language evaluation forum for European languages (Springer;)*, 2022, pp. 2253–2261.

[22] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[23] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9719–9728.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017. DOI: 10 . 1609 / aaai . v31i1 . 11231. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11231.

[25] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. DOI: https://doi.org/10.1007/s11263-015-0816-y.

[26] Sergio Guadarrama, Nathan Silberman, *TensorFlow-Slim: A lightweight library for defining, training and evaluating complex models in tensorflow*, https://github.com/google-research/tf-slim, 2016. [Online]. Available: https://github.com/google-research/tf-slim.

[27] H. Goeau, P. Bonnet, and A. Joly, "Plant identification based on noisy web data: The amazing performance of deep learning (lifeclef 2017)," in *CLEF: Conference and Labs of the Evaluation Forum*, 2017.

[28] H. Goëau, P. Bonnet, and A. Joly, "Overview of lifeclef plant identification task 2020," in *CLEF 2020-Conference and labs of the Evaluation Forum*, vol. 2696, 2020.

[29] H. Goëau, P. Bonnet, and A. Joly, "Overview of plantclef 2021: Cross-domain plant identification," in *CLEF 2021-Conference and Labs of the Evaluation Forum*, vol. 2936, 2021, pp. 1422–1436.

[30] H. Goëau, P. Bonnet, and A. Joly, "Overview of plantclef 2022: Image-based plant identification at global scale," in *CLEF 2022-Conference and Labs of the Evaluation Forum*, vol. 3180, 2022, pp. 1916–1928.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Unveiling Robust Feature Spaces: Image vs. Embedding-Oriented Approaches for Plant Disease Identification

Hamza Ahmed Ishrat †*, Abel Yu Hao Chai†*, Sue Han Lee† and Patrick Hang Hui Then†
† Swinburne University of Technology Sarawak Campus, Malaysia
E-mail: 102765181@students.swinburne.edu.my, aychai@swinburne.edu.my, shlee@swinburne.edu.my, pthen@swinburne.edu.my

***Abstract*—Deep learning models, such as Convolutional Neural Networks (CNNs), possess the ability to learn robust features that enable effective plant disease classification. However, these models need large and diverse datasets that reflect the variations present in real-world scenarios in order to achieve high performance not only on classes that the model has already seen, but also on new classes that have not been part of the training. To address this challenge, the utilization of Conditional Generative Adversarial Network (CGAN) models offers significant advantages. CGANs can generate diverse synthetic data, thereby expanding our training dataset and enhancing the generalization capability of deep learning models. While the conventional approach typically focuses on training CGANs to generate synthetic plant disease images, our investigation goes a step further by exploring the effectiveness of training CGANs using embedded images as reference points. This research is driven by the challenge of training CGANs on high-dimensional color images, as opposed to their simpler embedding data with a low-dimension. We have discovered that CGAN trained easier in embedding space and performed better in plant disease classification for both seen and unseen compositions. Our visual analysis also showed that the small CGAN model is able to generate better synthetic embedding than the synthetic images, which leads to significant performance in the overall classification results.**

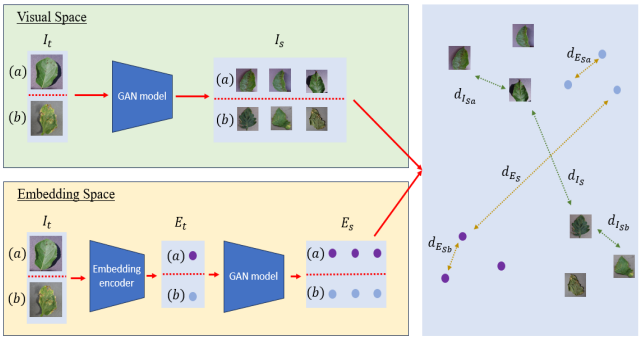


Fig. 1. Overview of our study. We train CGAN model in both visual and embedding space to evaluate the compatibility of CGAN model with deep learning model for downstream plant disease classification task. CGAN model is used to generate synthetic images, I_s and synthetic embeddings, E_s from original image, I_t and original embedding, E_t. This can increase the size and diversity of our training dataset. Furthermore, we also perform in-depth analyses on both synthetic data. The analysis are perform in various aspect which are inter-class separation for images, D_{Isa} / D_{Isb}, intra-class separation for images, D_{Is}, inter-class separation for embeddings, D_{Esa} / D_{Esb}, intra-class separation for embeddings, D_{Es},

I. Introduction

Plant diseases are characterized by deviations from a plant's normal state due to the intrusion of diverse pathogens. These diseases disrupt the plant's growth patterns, leading to diminished crop yields and substantial economic losses. However, these losses can be mitigated through early prevention and detection measures. Unfortunately, the vast knowledge gap between plant pathologists and the general public hinders individuals from independently classifying plant diseases. As a result, people depend on the expertise of plant pathologists or specialists to conduct accurate disease identification processes.

Deep learning models have become increasingly powerful tools for plant disease classification in recent years, as evidenced by various studies [1]–[7]. These models demonstrate remarkable success in extracting dominant features that differentiate between different plant species and disease classes. Despite their versatility, deep learning models heavily rely on data for their performance. To effectively classify objects, these models require a large and diverse training dataset that enables them to learn generalized features. However, real-world data often follows a long-tailed and open-ended distribution, where a small number of classes have a significant number of samples while others are associated with only a few samples. This long-tailed distribution is particularly evident in plant disease datasets due to the scarcity of labeled images and the high cost of data acquisition. Consequently, trained models may exhibit biases or overfitting tendencies towards specific classes, limiting their ability to learn generalized features for plant disease classifications. For a practical recognition system, it is essential not only to be able to classify majority classes effectively but also to account for the minority classes.

In theory, increasing the size of the training dataset could address this issue; however, the process of collecting such data is excessively costly and time-consuming. To this end, two main approaches can be employed to enhance model performance: (1) enhancing feature extraction techniques [8],

*These authors contributed equally to this work

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

[9] and (2) generating synthetic images to enrich the existing training dataset [10]–[14]. For the first main approach, feature extraction techniques continually undergo rapid advancements, giving rise to numerous state-of-the-art (SOTA) models, such as MobileNet, ResNet-152, and various vision transformer models. The fundamental principle underlying this approach revolves around enhancing feature extraction techniques to extract a more generalized features that benefit downstream classification tasks. However, this method encounters bottlenecks when the available samples within the training dataset are limited or unavailable.

Numerous scientific studies have successfully employed Generative Adversarial Network (CGAN) models to produce synthetic plant images, thereby enhancing training datasets for plant disease classification. However, formidable challenges persist, including mode collapse, limited variation, and training instability. Overcoming these hurdles necessitates meticulous architectural design, effective regularization techniques, and sophisticated training strategies. Particularly when dealing with plant disease data, characterized by diverse textures and appearances, imbuing CGAN models with the ability to capture intricate details for synthetic image generation becomes even more arduous. In practice, visual deep learning models are commonly employed, relying on original disease data as a reference for synthetic data. Consequently, CGAN models are compelled to generate visually higher-resolution synthetic images, leading to enlarged model sizes and introducing additional complexities during training. Moreover, generating a diverse array of synthetic images in the visual space poses an additional formidable task, as it requires the CGAN model to leverage both localization information and learned features from the images.

To overcome these obstacles, it is worthwhile to explore embedding-driven deep learning models, where the generative model utilizes embedded images as references instead of the original images for data generation. The utilization of precomputed image embeddings offers several benefits. It enhances training efficiency by leveraging the encoded knowledge within the embeddings, eliminating the need to start from scratch with the original images. This approach potentially simplifies the training of CGAN models, which focus on the embedding space, and improves compatibility with feature extractor networks. Additionally, pretrained embeddings capture relevant information from the data and encode semantic and contextual details, facilitating the extraction of meaningful features. However, one potential challenge is the issue of collapsing, wherein the output of an embedding collapses to a constant vector, deviating from the original distribution. Therefore, it is crucial to compare the performance of CGAN models operating in the visual space versus the embedding space for plant disease identification task. The contribution of this paper are:

1) To the best of our knowledge, our study represents the first evaluation of the effectiveness of CGAN (Conditional Generative Adversarial Network) models in generating representative data to improve plant disease classifiers for both seen and unseen classes. Based on our results, the classifier using the generated embedding data outperforms the classifier based on the generated image.
2) We have experimentally shown that the CGAN model and the deep learning-based classifier operating in the embedding space are able to learn more diverse and generalised features, showing a significant improvement of 30% for plant identification and a further 48% improvement for disease identification compared to models trained based on original images.
3) We present a visualisation analysis to interpret the reason for the better performance of the CGAN model operating in the embedding space. The results show that models operating in the embedding space exhibit better interclass and intraclass separation than those operating only in the image space. These visualisations provide valuable insights into the benefits of the embedding space approach.

The rest of the paper is organized as follows. Section II discusses the literature related to plant disease identification with various models. In section III, we present our methodology as well as the various models we propose to carry out our experiments where, in section IV, we present our findings and provide discussion. Section V concludes our research findings and presents future work.

II. Related work

A. Plant disease classification with deep learning network

Deep learning models have shown remarkable success in plant disease identification. These models leverage the power of deep neural networks to automatically learn and extract meaningful patterns from plant images, enabling robust disease classification. For instance, convolutional neural networks (CNNs) have been employed to perform classification tasks across various crops, such as maize [1], cassava[2], tomato [3], apple[4], wheat[5], citrus [6] or potato [7]. These studies primarily focus on individual crops, wherein the models solely need to learn disease-specific features to accomplish the classification tasks. However, in 2016, the scope of the research expanded to encompass multiple crops and diseases with the introduction of the PlantVillage (PV) dataset [15]. This dataset serves as a largest publicly available dataset, featuring multiple crops and diseases. It comprises 38 classes representing 21 diseases across 14 different crops. The PV dataset presents new challenges to the models as they are now required to simultaneously consider both plant and disease features for classification tasks. Previous studies in [16], [17] use this dataset to train CNN models for plant disease classifications. Although the PV dataset is the largest publicly available dataset, it does not cover the complete range of plant disease compositions. As a result, our study focuses on utilizing CGAN models to generate synthetic data. By doing so, we aim to augment the diversity of plant disease compositions without the requirement for manual acquisition. This approach enables us to expand the dataset

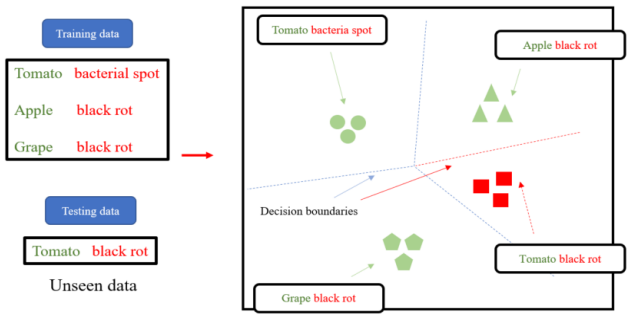


Fig. 2. The concept of unseen plant disease composition. *Tomato bacterial spot*, *Apple black rot* and *Grape black rot* are the seen composition from training data. *Tomato black rot* is the unseen composition from testing data. The individual *Tomato* species and *black rot* disease are exist in the training data.

and enhance the representation of plant diseases for improved training and analysis.

B. Plant disease classification with generative adversarial network

GANs were introduced in 2014 by [18] to generate synthetic images that exhibit a similar distribution to the original images. The synthetic images prove valuable in enriching our training dataset, particularly for classes that lack labeled data or are prohibitively expensive to obtain. This approach also helps prevent bias in our deep learning models towards dominant classes present in the training dataset. For instance, the PlantVillage dataset from [15] consists of 38 classes, with tomato species (10 classes) and healthy plants (11 classes) being the dominant categories. This skewed distribution may lead to a bias in our trained deep learning models, favoring tomato species and healthy plants. To overcome this issue, previous studies [10]–[14] have incorporated various types of GANs with deep learning models to enhance the performance of plant disease classification. These GAN models have been utilized to increase the diversity within each class of the training dataset and augment samples for classes with limited representation. While the conventional approach focuses on training GANs to generate synthetic plant disease images, our investigation takes a step further to assess the effectiveness of training CGANs using embedded images as reference points. This research endeavor is driven by the recognition of the complexities involved in GAN training for high-dimensional colored images, particularly in capturing intricate and diverse disease symptoms.

C. Zero shot classification for plant disease classification

In section II-A, we highlighted the limitation of the PV dataset in encompassing all possible plant disease compositions. Consequently, several studies have aimed to enhance the generalization capability of their models to effectively handle unseen plant disease compositions [17], [19]. The concept of unseen plant disease composition refers to situations where the individual plant species or disease types are present within the training dataset, but their combined composition as a specific plant disease is not represented. Fig. 2 illustrated the concept of unseen plant disease composition. A previous study by [19] introduced the CTML model, which improves upon conventional models by incorporating a conditional link that promotes knowledge sharing between plant features and disease features. However, this approach can only be effective if both plant and disease features are present and generalize well across unseen compositions. To address this challenge, our study takes a two-fold approach. Firstly, we generate synthetic data in both visual and embedding spaces to ensure that our training dataset is diverse enough for the deep learning model to learn generalized features. Secondly, we project both plant and disease features into individual plant and disease feature spaces for both seen and unseen plant disease classifications. By separating the features into distinct spaces, we enable the model to effectively analyze and classify plant diseases based on their unique characteristics, even when certain plant species or disease types are not present in our training dataset.

III. Method

A. Problem formulation

The goal of our plant disease classification is to recognise composition of both plant species and disease types. Our training dataset consist of different combined plant disease compositions, $c = (c_c, c_d) \in C$ where c_c and c_d corresponding to the plant species and disease types such as *Tomato healthy* and *Potato early blight*. Our testing dataset consist of seen compositions $c_s = (c_c, c_d) \in S$ and unseen compositions $c_u = (c_c, c_d) \in U$. Unseen compositions are the compositions that not present in training dataset such as *Potato healthy* but the individual *Potato* species, c_c and *healthy* disease, c_d are present in training dataset, C.

B. Overall framework

Our study aims to explore the compatibility of GAN model with deep learning model in both visual and embedding space for plant disease identification. Therefore, our architecture can be separated into visual-based and embedding-based as shown in Fig. 3. The architectures consist of three main modules: (1) feature extractor module, (2) Conditonal Generative Adversarial Network (CGAN) models for image and embedding space and (3) plant and disease classifiers for prediction. The feature extractor module extract features from images and convert the features from visual space to embedding space. CGAN model will generate synthetic images or embedding to enrich the training dataset. Plant and disease classifiers will performs different classification tasks based on visual images or embedding features.

1) Feature extractor module: The feature extractor module converts visual images into embedding spaces. The backbone of our feature extractors are built from MobileNet-v2 [20]. The classification head of MobileNet-v2 is removed to produce embedding output for our framework. Specifically, it consists of plant feature extractor, $g(\phi)$ and disease feature extractor, $f(\phi)$ as shown in Fig. 4 . Plant feature extractor, $g(\phi)$ takes

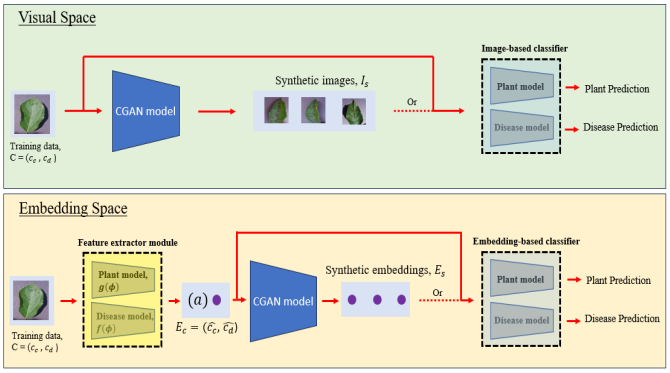


Fig. 3. Overall framework of our study. Our architecture can be separated into visual space and embedding space. For visual space, CGAN model will be used to generate synthetic images, I_s from training data, C. The image-based classifiers will either take real images or real images and synthetic images together to perform classification tasks. For embedding space, feature extractor module is used to convert the visual images from training data, C to embedding, E_c. The CGAN model will generate synthetic embeddings, E_s from real embeddings, E_c. The embedding-based classifiers will either take real embeddings or real and synthetic embeddings together to perform classification tasks.

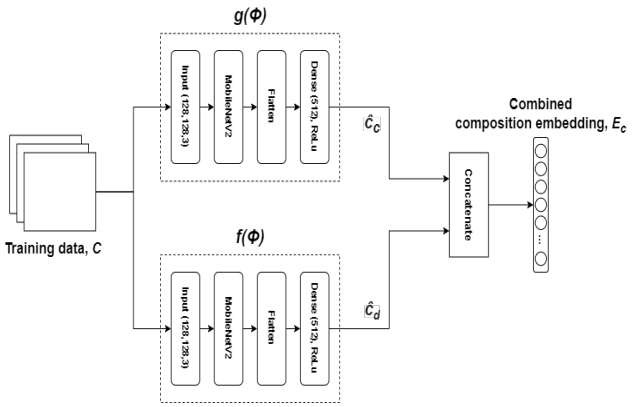


Fig. 4. To create the embedding dataset, we leverage a plant and disease model that use MobileNet-v2 as backbone without a classifier head. The plant and disease model will extract plant features and disease features respectively and will be concatenated to be used by the embedding CGAN model or classifiers.

input from training dataset, C and output plant features, $\hat{c}_c = g(c)$. Disease feature extractor, $f(\phi)$ takes input from training dataset, C and output plant features, $\hat{c}_d = f(c)$. Both feature extractors used pre-trained weight from ImageNet and finetuned with our training dataset by using cross-entropy loss. The combined composition embedding, $E_c = (\hat{c}_c, \hat{c}_d)$ will be generated from concatenation between plant and disease features and will be used by CGAN model or classifier in embedding spaces.

2) Conditional Generative Adversarial Network (CGAN) models: We use CGAN model to generate synthetic data for all classes in our training dataset. CGAN consist of two modules which are generator and discriminator modules. The generator takes class label and Gaussian noise as input and generate synthetic data. The discriminator distinguishes the synthetic data and eventually bring the distribution of

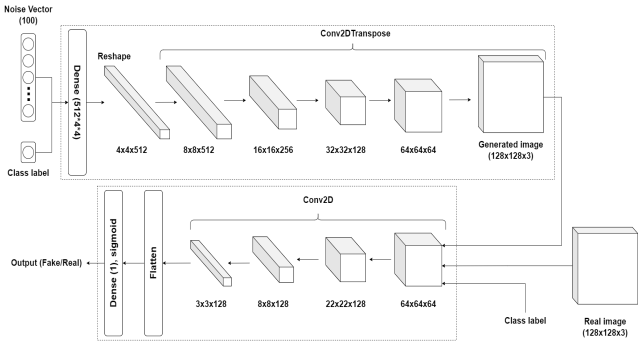


Fig. 5. Our CGAN architecture for generating RGB images of 128 X 128 images

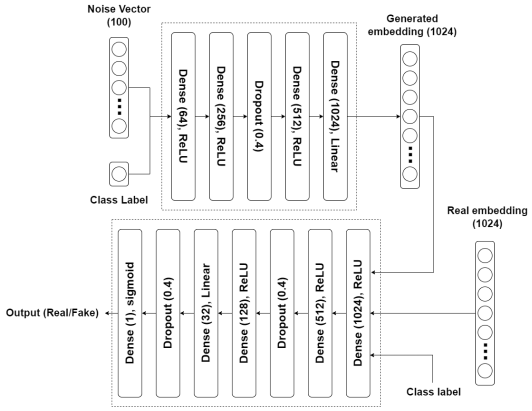


Fig. 6. Our CGAN architecture for generating embeddings of size 1024

synthetic data close to the real data.

Visual space CGAN. The CGAN architecture is shown in Fig. 5. The generator consists of 5 convolutional layers to up-sample the input and product synthetic images, I_s with size of 128x128x3. The discriminator consists of 4 convolutional layers to down-sample and distinguishes the synthetic data. The generator takes a class label and Gaussian noise as input. The discriminator takes real images from training dataset and synthetic images from generator as input.

Embedding space CGAN. The CGAN architecture is shown in Fig. 6. The generator consists of 5 fully connected layers to up-sample the input and product synthetic embedding, E_s with size of 1024. The discriminator consists of 7 fully connected layers to down-sample and distinguish the synthetic data. The generator takes a class label and Gaussian noise as input. The discriminator takes real embedding from feature extractor module and synthetic embedding from generator as input.

3) Plant and disease classifiers: A plant and disease classifier will be used to perform prediction in both the image and embedding space. MobileNet-v2 and 4 fully connected layers will be used for both image-based classifiers or embedding-based classifiers respectively. All classifiers will take in either

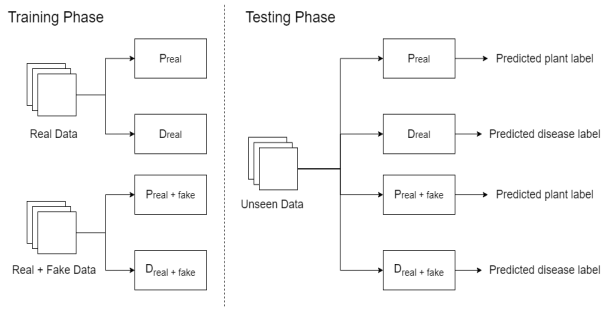


Fig. 7. Our evaluation model architecture, consisting of identical models built upon the MobileNet-v2 weights. P_{real} and D_{real} are trained on real images or embeddings, like wise $P_{real+fake}$ and $D_{real+fake}$ are trained on a combined dataset of real and generated images or embeddings. The models then predict the unseen data to identify the plant and disease

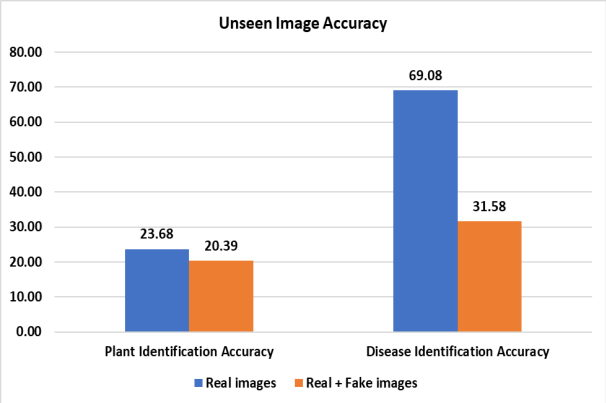


Fig. 8. Accuracy of identifying unseen plant label combination for real and real + fake image dataset

real or real and synthetic data together as training dataset to perform plant disease classification tasks.

IV. Experiments and Results Analysis

To explore the differences in the performance of image and embedding data, we used the open-source Plant Village dataset [15] that contains healthy and unhealthy images of leaves of various species and diseases. For this study, we selected 6 classes as seen training data, C and 1 class as unseen testing data U. Specifically, the 6 classes in our training data, C are *Peach bacterial spot*, *Peach healthy*, *Potato early blight*, *Tomato bacterial spot*, *Tomato early blight* and *Tomato healthy*. The unseen testing data is *Potato healthy*. We evaluate the quality of the generated images and embeddings via embedding space as well as utilising the plant and disease classifier described in Fig. 7.

A. Expt. 1: Analysis of the Classification Performance of the Image-Based Model

The first experiment involves exploring the performance of the deep learning models on unseen data with the aid of generated images. 2 sets of images would be prepared, 200 images of each class for the real dataset, and 200 images and 100 generated images via the CGAN per class for the real

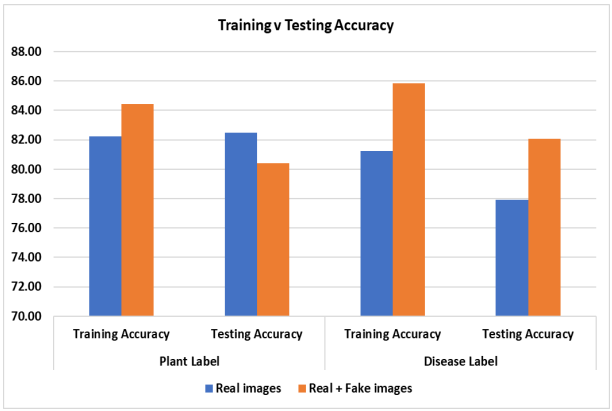


Fig. 9. Performance difference for training and testing accuracy of plant disease classifier on image data

and fake dataset. 2 pairs of models would be trained on each dataset, and then performance would be evaluated based on the accuracy of correctly identifying the plant and disease label of an unseen combination.

Fig. 8 shows that accuracy for identifying plant and disease for an unseen combination reduces with the introduction of generated images, however, the accuracy for the correct label was low to begin with. This may be due to the fact that, although the dataset included 3 different plant labels, analysis into the dataset shows that the distribution for plants is imbalanced. For instance, there are 3 classes with 'tomato' as the plant label, 2 for 'peach' and 1 for 'potato'. The ratio between tomato and potato sample sizes is 3 to 1, coupled with the fact that the leaves look visually similar, we believe the low accuracy of the unseen data may be attributed to this imbalance. Further evidence to this is revealed when the model wrongly predicts tomato with the highest confidence when testing with the unseen data.

We also attribute the reduction in accuracy to the quality of the images produced by the CGAN, which was trained on 2000 epochs. Various challenges arise when training GANs, especially when wanting to have it pick up on detailed features such as venation, particularly due to the limited resolution of the output image. Due to computational constraints, our initial dataset were images of size 256 by 256, but had to be resized to 128 by 128 and subsequently can produce only at 128 by 128. At a reduced size we suspect that finer details are hidden and removed and so the CGAN generates images that are not as detailed.

However, there was an increase in accuracy during training and testing evidenced in Fig. 9, which had increased performance when training on the real and fake dataset. This shows that the models are better fitted to seen classes, but are not sufficiently generalised to unseen data.

B. Expt. 2: Analysis of the Classification Performance of the Embedding-Based Model

Similar to experiment one, this experiment focuses on the performance via real and generated embeddings. The real

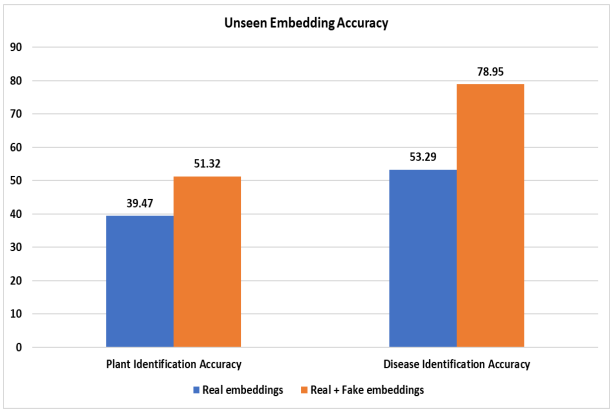


Fig. 10. Accuracy of identifying unseen plant label combination for real and real + fake embedding dataset

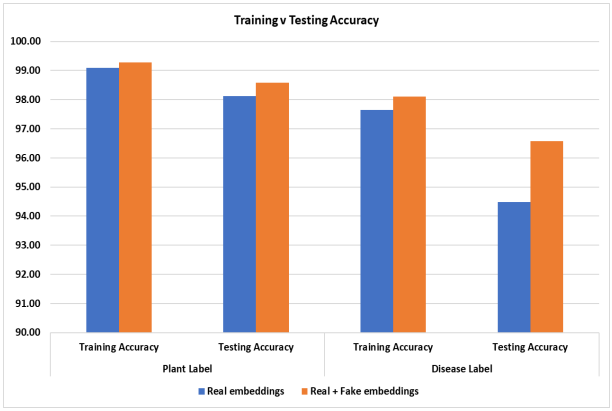


Fig. 11. Performance difference for training and testing accuracy of plant disease classifier on embedding data

dataset will consist of 200 embeddings per class and 200 real and 100 generated embeddings via CGAN for the real and fake dataset. The performance would also be tested on the same unseen combination, this time using unseen embeddings.

For our embedding CGAN, our results show a significant increase in performance when classifying an unseen combination of plant and disease, rising from 39.5% to 53.3% for plant accuracy and 51.3% to 79% for disease accuracy when trained on real embeddings and a combination of real and fake embeddings respectively, as shown in Fig. 10.

It also greatly improved training and testing accuracy on seen data, increasing for all values even when the baseline values were already more than 94% as shown in Fig. 11. We believe this is due to the way embeddings were extracted in the first place, capturing relevant information from each disease and plant species concept, encoding semantic and contextual details that could help CGAN better extract meaningful features.

We also found that training the embedding CGAN was significantly easier as compared to the image CGAN due to its smaller and simpler architecture as well as its lack of needing to extract features since it was already being fed extracted features. Thus, it made it easier for the generator to quickly

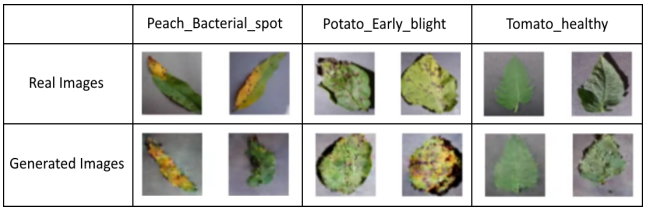


Fig. 12. Visual analysis of real images and generated images

catch on in 450 epochs in a short amount of time to be able to produce not only close to real embeddings, but also in variations as explored in the next section.

C. Expt. 3: Ablation Study on Embedding-Based CGAN Model Training with Increased Data, Including Synthetic Samples of Unseen Classes

Experiment 3 follows up the previous experiment by slightly modifying the embedding CGAN. Instead of taking in a single integer as the class, we decided to turn a class label into a two one hot vectors representing each part of the label which would then be concatenated together. For example, 'Tomato healthy' would have its one hot vector for 'tomato' being [1, 0, 0] and 'healthy' being [0, 1, 0]. The final class vector would then be [1, 0, 0, 0, 1, 0], which would be used as the class label for the generator. This approach allows us to generate unseen data by manipulating the input vectors. For this experiment, we generated 200 unseen embeddings and incorporated them into our pre-existing training dataset. The updated training dataset now comprises 200 real embeddings for each class along with the 200 previously generated unseen embeddings. Subsequently, we compared our findings with the experiment outlined in section IV-B.

Our findings indicate an accuracy of 52.6% for plant classification and 78.2% for disease classification, showcasing notable improvement in plant identification but limited progress in disease recognition. This discrepancy could be attributed to the fact that the generated embeddings are more proficient at capturing distinctive crop features compared to disease symptoms, as the differentiation of disease classes is highly dependent on the level of detail captured, making it a more difficult task.

D. Expt. 4: Qualitative Analysis

Experiment 4 explores the qualitative analysis of the generated images and embeddings against real images and embeddings respectively. Using t-sne analysis [21] to show the class-wise features in the embedding space we will explore the distribution and the separation of the features based on classes, real and fake data. We will also perform visual analysis on generated images discussing the differences between real and fake images.

Fig. 12 shows the real and generated data for 3 of the classes, and immediately we notice a lack of fine details in the generated images. Although low-level features are captured, such as the discolouration due to bacterial spot or early

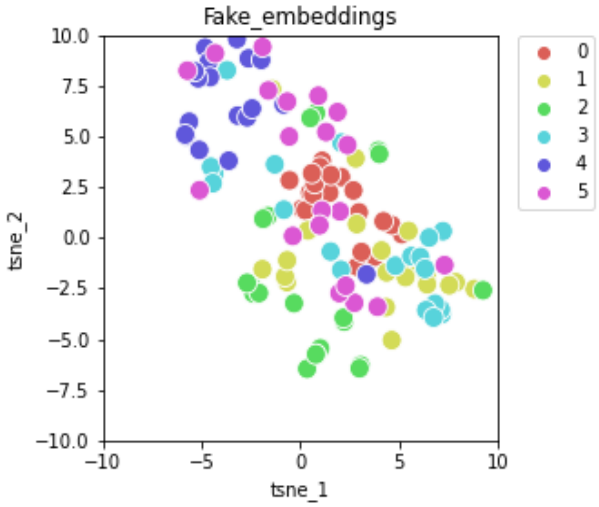


Fig. 13. T-SNE analysis in embedding space for generated images

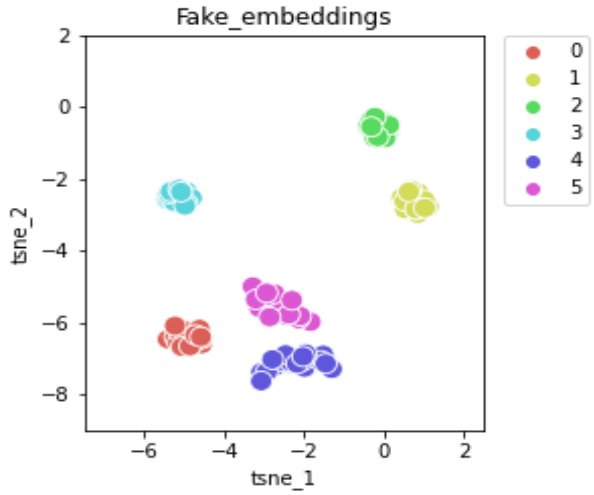


Fig. 14. T-SNE analysis in embedding space for generated embeddings

blight, higher level features such as the midrib or venation is not captured. it also has a tendency to confuse shadows as discoloration as evidenced in the second generated image for 'potato early blight'. The generated images are also blurry due to a lack of high-level features, and we suspect longer training times might alleviate this problem.

We built a feature extractor that utilised the MobileNetV2 model that was trained on real images and had its last layer removed to expose the features of shape 512. The generated images were then run through the feature extractor, and they were projected onto an embedding space in Fig. 13. The generated embeddings show almost no distinct clusters as compared to embedding clusters in Fig. 14. We again attribute this to not having enough high-level features for the classifier to accurately distinguish between the generated images. For fake embeddings, the projections look a lot more distinct for fake with 6 very obvious clusters being made with variations to show that the points, while similar enough to cluster, are not exact matches.

V. Conclusions

In this paper we investigated the capabilities of using CGANs to generate data, in terms of images and embeddings, to aid in the classification of unseen plant disease combinations. Our objective was to assess the effectiveness of CGAN-generated data in improving the performance of plant and disease classifiers, not only on seen data but also on previously unseen combinations, thereby enabling more accurate identification of unseen classes. We established that while image CGANs require more care and testing, embedding CGANs demonstrate a greater potential to aid in plant disease identification by being simpler, faster, and easier to train. By training plant and disease classifiers on both real and generated embeddings, we observed an increase in their overall performance for both seen and unseen combinations. This augmentation enabled the models to make more confident predictions on unseen combinations, resulting in enhanced plant and disease identification accuracy. In future work, we will incorporate semantic information, such as word embedding, to improve the learning of correlations between seen and unseen classes. By leveraging this additional contextual information, we anticipate further improvements in our model's performance.

Acknowledgment

This research was supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant No.(Ref: FRGS/1/2021/ICT02/SWIN/03/2), from the Ministry of Higher Education Malaysia

References

[1] T. Wiesner-Hanks, E. L. Stewart, N. Kaczmar, *et al.*, "Image set for deep learning: Field images of maize annotated with disease symptoms," *BMC research notes*, vol. 11, no. 1, pp. 1–3, 2018.

[2] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," *Frontiers in plant science*, vol. 8, p. 1852, 2017.

[3] A. F. Fuentes, S. Yoon, J. Lee, and D. S. Park, "High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank," *Frontiers in plant science*, vol. 9, p. 1162, 2018.

[4] B. Liu, Y. Zhang, D. He, and Y. Li, "Identification of apple leaf diseases based on deep convolutional neural networks," *Symmetry*, vol. 10, no. 1, p. 11, 2017.

[5] A. Picon, A. Alvarez-Gila, M. Seitz, A. Ortiz-Barredo, J. Echazarra, and A. Johannes, "Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild," *Computers and Electronics in Agriculture*, vol. 161, pp. 280–290, 2019.

[6] Z. Iqbal, M. A. Khan, M. Sharif, J. H. Shah, M. H. ur Rehman, and K. Javed, "An automated detection and classification of citrus plant diseases using image processing techniques: A review," *Computers and electronics in agriculture*, vol. 153, pp. 12–32, 2018.

[7] D. Oppenheim and G. Shani, "Potato disease classification using convolution neural networks," *Advances in Animal Biosciences*, vol. 8, no. 2, pp. 244–249, 2017.

[8] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Transformer neural network for weed and crop classification of high resolution uav images," *Remote Sensing*, vol. 14, no. 3, p. 592, 2022.

[9] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[10] H. M. Maniyar and S. V. Budihal, "Plant disease detection: An augmented approach using cnn and generative adversarial network (gan)," in *Information, Communication and Computing Technology: 5th International Conference, ICICCT 2020, New Delhi, India, May 9, 2020, Revised Selected Papers*, Springer, 2020, pp. 252–261.

[11] Y. Zhao, Z. Chen, X. Gao, *et al.*, "Plant disease detection using generated leaves based on doublegan," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 3, pp. 1817–1826, 2021.

[12] R. Rajalaxmi, R. Prabavathi, T. Suganya, S. Mohanapriya, K. Tamilselvi, C. Ebenezarkanmani, *et al.*, "Optimized deep convolutional generative adversarial network for tomato leaf disease identification," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, 2021, pp. 1301–1307.

[13] L. Bi and G. Hu, "Improving image-based plant disease classification with generative adversarial network under limited training set," *Frontiers in plant science*, vol. 11, p. 583 438, 2020.

[14] Q. Wu, Y. Chen, and J. Meng, "Dcgan-based data augmentation for tomato leaf disease identification," *IEEE Access*, vol. 8, pp. 98 716–98 728, 2020.

[15] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in plant science*, vol. 7, p. 1419, 2016.

[16] S. H. Lee, H. Goëau, P. Bonnet, and A. Joly, "New perspectives on plant disease characterization based on deep learning," *Computers and Electronics in Agriculture*, vol. 170, p. 105 220, 2020.

[17] S. H. Lee, H. Goëau, P. Bonnet, and A. Joly, "Attention-based recurrent neural network for plant disease classification," *Frontiers in Plant Science*, vol. 11, p. 601 250, 2020.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] S. H. Lee, H. Goeau, P. Bonnet, and A. Joly, "Conditional multi-task learning for plant disease identification," in *2020 25th international conference on pattern recognition (ICPR)*, IEEE, 2021, pp. 3320–3327.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Facial Expression Recognition as markers of Depression

Jia Xuan Gue* and Chun Yong Chong† and Mei Kuan Lim‡
* School of Information Technology, Monash University Malaysia, Selangor, Malaysia
E-mail: jgue0004@student.monash.edu Tel/Fax: +6012-3073800
† E-mail: chong.chunyong@monash.edu Tel/Fax: +6012-6760146
‡ E-mail: lim.meikuan@monash.edu Tel/Fax: +6012-6337702

***Abstract*—With research done to determine depression with different types of inputs, such as text, images, and mobile sensing, numerous distinct approaches have been devised and enhanced for identifying depression across diverse backgrounds. However, there is a significant lack of research in making use of only facial expressions to detect depression. Introducing such methods could be beneficial due to the large amount of research done to develop models capable of performing facial expression recognition. This project aims to utilize pre-existing facial expression models to determine depression based on the expression recognition capabilities, and to also determine the reliability and effectiveness of using only facial expression to detect depression. We have chosen to use EfficientNet, a state-of-the-art facial expression recognition model, to experiment with the different facial expressions that are shown based on images of clinically proven to be depressed patients. The findings from our work can further improve depression detection research, specifically focusing on making use of only information from shoulder up of patients, while also helping the community to improve mental health.**

I. Introduction

Depression is a mental illness that mainly affects mood, emotions, and behaviour. It usually results in sadness, despair, and the loss of interest in activities that once used to be enjoyable. It is currently one of the most common mental illnesses in the world, affecting an estimate of 3.8% of the population, and 5% of adults[1]. In addition, with the Covid-19 pandemic taking place for almost 2 years ever since December 2019, it has resulted in the worsening of depression, with an astounding 45% of infected patients having depression[2]. The increase in depression during the pandemic has resulted in doubling of individuals having suicidal thoughts, as compared to pre-pandemic[3]. There was also an upward trend of the number of suicide attempts during the Covid-19 period due to depression[3]. Despite how prevalent and devastating depression is today, many still are unaware of the symptoms of depression, and are also uninformed about how serious it is. This can result in depressed patients refrain from proactively seeking treatment. Furthermore, stigma surrounding depression can result in them to be afraid of judgement and discrimination and choose not to seek proper help and support. This issue is highly detrimental to society, and action needs to be taken to address it.

To date, various technologies have been proposed to create awareness and introduce intervention mechanisms to combat depression [4][5], from utilizing mobile applications for mobile health to electronic survey tools and screening assessments. With the rise of cloud-based collaborative video conferencing platforms such as Zoom or Microsoft teams due to the Covid-19 pandemic, patients with depression can seek treatment and care through such technologies. They can receive therapy from their therapists from what is usually known for their face-to-face practices. This is also known as teletherapy.

Some main teletherapy platforms include SimplePractice, TherapyNotes or BetterHelp, with main functions including HIPAA-compliant (The Health Insurance Portability and Accountability Act of 1996) video calls, appointment scheduling, secure messaging and recordings and providing different templates for presentation. With that, therapists can share content through different preferred presentation styles, while also having the ability to annotate for their patients to see. However, during a teletherapy session, therapists will have to juggle with many aspects of therapy, which includes providing a useful presentation, listening to what patients have to offer, as well as paying attention to patient's emotions, as well as non-verbal cues. Thus, there is a need for other forms of smart technology to be introduced to reduce the load of therapists in a tele-therapy setting.

A main communication type for video conferencing platforms would be the conversations that both parties will have through audio and chat. Work done by Deshpande, M. S.,& Rao, V. has utilized the emotion recognised from textual data to detect depression[6]. To the best of the author's understanding, work that utilises the emotion recognised from facial images or videos are underexplored to detect depression.

It is also known that video conferencing platforms have video as one of the main communication types, where both patients and therapists can see each other while communicating. It is common that most users use a webcam to stream their videos on conference calls. However, it is also common that most users will not be able to show their full body, usually only showing from the chest up, where the focus of the video is usually the face. While therapists are busy indulging in conversations with the patients, they may potentially overlook the expression of the subject, which may give insights to the engagement or feeling of the patient during the conversation. Thus, facial images can be an input used for therapists to determine how their patients are feeling and to potentially

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

make decisions that could help them or make them feel better.

Facial Expression Recognition using computer vision have also been researched and many state-of-the-art models have been created and collated due to the many challenges that have surfaced ever since 2011 [7], while also being a very prominent subject of discussion in the computer science field for almost 30 years [7]. Facial expression technology is nowhere new in Computer Vision and Artificial Intelligence but has not been utilized in depression detection. Thus, it would be beneficial to tap into such advanced systems and make use of them to aid in determining depression based on expressions. Therefore, it is essential to come up with an optimised solution for such technologies to aid therapists and medical professionals.

The contribution of this paper is two-fold:

1) The potential of using facial expressions as markers of depression is investigated.
2) Annotated images from videos from publicly available sources online.[1]

II. Related Work

The main bulk of related work will be split into 3 sections, works in teletherapy, facial expression recognition and depression detection.

A. Teletherapy

Teletherapy has surfaced ever since there was an option for therapists to have conversations with patients through the phone in the 1960s. It became increasingly popular, and even primary means of care for some individuals ever since the Covid-19 pandemic, with the main purpose to prevent the spread of the novel virus by discouraging face-to-face contact. There are a multitude of benefits and well as disadvantages of teletherapy. Having therapy online helps when patients have difficulties meeting therapists physically. There are many different factors that affect this, such as accessibility, where patients may be located a distance away from a suitable therapist. Other factors include anonymity, privacy as well as flexibility of time and space[8]. Most of the population now also has access to the internet and a smart device, while being proficient in using it, thus would also prefer attending therapy sessions through a phone or a computer.

However, not attending a therapy session physically has also resulted in certain problems, not just for the patients, but for therapists as well. Patients have reported that when comparing both types of therapy, some have felt that attending them online felt less engaging [9], some of them felt that their conversations are less private [9], or even issues with the internet or the hardware and software not being able to function as intended [9]. Therapists also felt that they are not able to pick up non-verbal cues while conducting sessions [9], while also struggling in certain moments to pick up markers of expressions, gestures, or postures [9][10]. Other therapists have also reported an increase in struggle to attune with their patients in an online setting, as compared to being in their office[9]. Thus, having external technologies can provide therapists with indicators of patient's mood, engagement and other non-verbal factors and fill in the gaps that teletherapy will not be able to achieve. With teletherapy on the rise, it is easier to obtain videos of teletherapy sessions as most video-telephony software provide a recording feature. While it is important to request for permission from all parties involved to ensure that data collection is ethical, this data collection is essential with the training and testing of models capable of depression detection.

B. Facial Expression Recognition

Facial expression analysis can be dated back many years, but the more prominent research done was by firstly Charles Darwin, where he explained and grouped facial expressions based on biological reasoning, which are making use of different muscle groups present on the face[11]. This is further researched by Ekman and Friesen, which they introduced the concept of the Facial Action Coding System[12]. The work done by these researchers has proved to be vital in the development of the first facial expression recognition models. These models make use of the facial action coding system to determine the different types of facial muscles used in the expression of facial expressions[12].

Facial features are another type of input that can be used to determine expressions in a person. This can include the eyes, lips, eyebrows, and cheeks[7]. There are also transient facial features, features that are not found in all people. This includes wrinkles and facial lines[7].

There are many types of facial expression recognition architectures present in the research field, including but not limited to: MobileNet[13], EfficientNet[14], ResNet[15]. These are computer vision models that can be used to train with annotated data, each with specialized methods to help with the development and training of deep learning models. These state-of-the-art models are also rather mature, thus the focus on this work is not to devise a new expression recognition model, but to utilize these pre-existing models.

There are also many datasets present due to the number of challenges present, with the more prominent ones being the EmotiW (Emotion Recognition in the Wild) challenge and AFEW (Acted Facial Expression in The Wild) challenge. AffectNet [16] is also another dataset used for emotion classification, where it consists of annotated images of people with their expressions. This manually annotated dataset is labelled with 7 different expressions, together with including valence and arousal of such expressions.

C. Depression Detection

Facial expressions can provide information about how one feels, and research has been done to prove that it can provide enough information with whether an individual is feeling depressed. Many studies have shown that the main parts of the face that can provide most of the information are usually the eyes and the mouth, where together with expressions, can help determine an emotion of a person[17][18]. Mehrabian, A.,

[1]Dataset will be made publicly available upon publication.

& Russell, J. A. did state that it is a lot more likely that one can determine emotions of someone else using facial expressions, as compared to using word of mouth[19]. This has resulted in more research done using facial features. Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. have done research using specifically eye movement analysis to detect depression in individuals[18], while Wang, Q., Yang, H., & Yu, Y. uses the overall facial expression to analyse for depression detection as well [17]. Wang, Q., Yang, H., & Yu, Y. have decided to classify this issue as a binary problem (e.g., depressed/ not depressed) [17], while Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. also classifies their models as binary [18].

Feature extraction for such research is more straightforward as it uses physical spots on the face to extract information. For instance, for eye movement analysis, features such as horizontal, vertical and eyelid movements can be extracted without much pre-processing needed [18]. However, normalisation is required as different people will have different eye sizes, and the eye sizes can also change depending on different factors like how far a person is located from the camera or the angle of the camera relative to the person. This similarly applies to any other frontal face features.

There are datasets present to help with depression detection. The more prominent ones would be the DAIC-WOZ dataset, which consists of annotated and transcribed audio recording, video recordings and questionnaire reports of clinically proven depressed patients. Another example dataset that is used for a challenge would be the SemEval-2018 dataset, which features a database of tweets from twitter and its multiple subtasks that determines the affectual state of the author based on their posted tweets. Most of these datasets do not make use of facial expressions to determine depressions, more frequently touching on speech and text as it has been proven to be highly accurate in detection. However, facial expressions are still indicators of depression, and thus should be considered while attempting to improve depression detection.

To date, there is research done to make use of facial expressions to determine depression [17][18]. Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. have identified indicators of depression mainly using facial features and movement of facial features, however, there is little to no use of facial emotions to aid with their research. Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. also investigated about how difference in emotions have led to no statistically significant differences in detecting depression, while Wang, Q., Yang, H., & Yu, Y. mentions about how cultural differences can result in different forms of depression. Since facial emotions are more universal and easier to understand for the average person, there is a need to utilize emotions to determine indicators of depression.

III. Methodology

A. Approach

To figure out which facial expressions a typical depressed person would evoke, as well as frequency of facial expressions

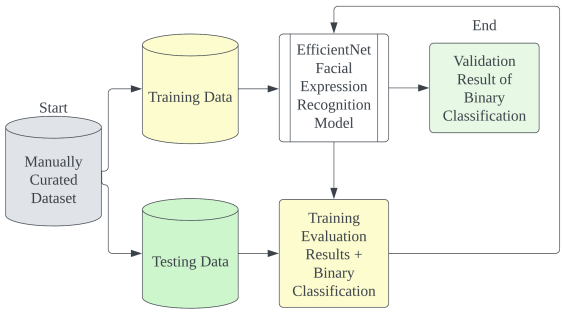


Fig. 1: Proposed Approach

used, it is necessary to collate existing data present online of images or videos of clinically proven depressed patients or make use of related datasets if present. It is mentioned in related work there is currently no such dataset present, thus the former option would be used. After collection of data, facial expression recognition models will be used to determine the expressions of the depressed patients. This is essential since expression recognition by other humans are highly inconsistent. After putting through a section images through the model, the results, which will be facial expressions, will be collected and presented in the results section. Results will be used to combine the 8 different classes of expressions into a binary classifier for depression. Then, determine how accurate the binary classifier is by testing on images of both depressed and non-depressed patients. The approach is also showcased in Figure 1.

B. Dataset

Images of depressed patients are collected from online sources, mainly making use of video clips of interviews or consultations of depressed patients, mainly from YouTube, with channels ranging from universities to individual channels that provide clips of interviews and consultations. For each clip of a specific person, to reduce randomness, 4 images will be taken from the video, focusing only on the face. These 4 images will be taken by splitting the video into 4 equal sections and selecting a specific timestamp. This first dataset will be for the training phase and will have around 100 images, representing 25 individuals. All the photos are extracted at points in the video where all their facial features are visible, where there is no obstruction or blocking of such to prevent any false detection or incorrect outcomes. The images are also cropped to ensure that the face is centered as well. An example of such an image is shown below in Figure 2, while Image II depicts annotated images by the expression recognition model. Figure 4 showcases the processing of an image in this research work.

For the testing phase, to test the reliability of the images taken, we will similarly be using interviews and consultations of depressed patients, but also include interviews with patients that are not depressed as well. This set of images will be around 100 images, split into half for depressed and not depressed.



Fig. 2: Pre-processed Image



Fig. 3: Annotated Image

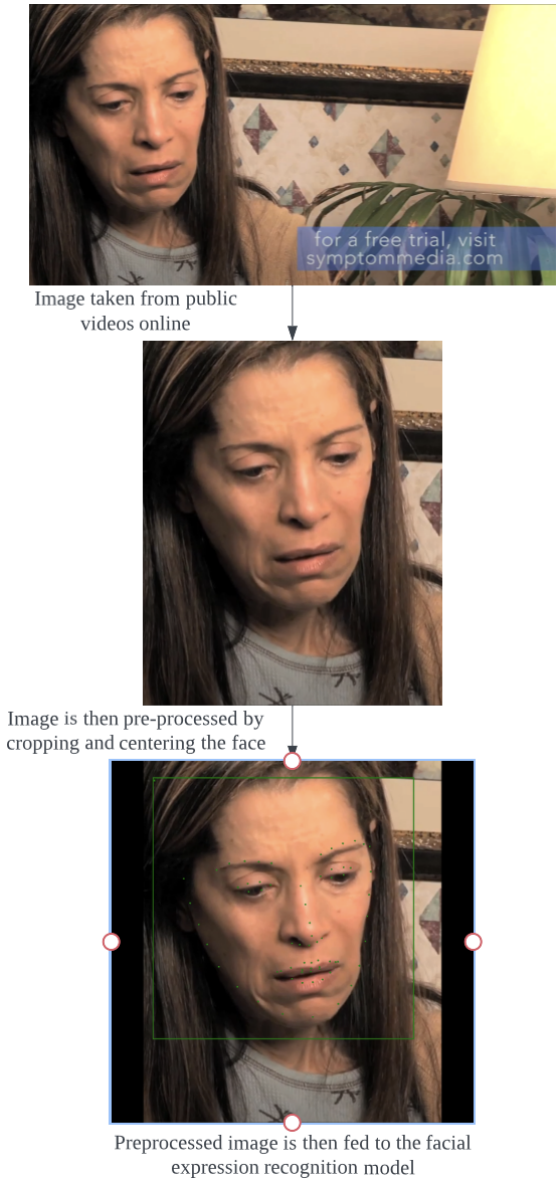


Fig. 4: Processing of Image

C. Model (EfficientNet Facial Expression Recognition)

A state-of-the-art facial expression recognition model will be used. Savchenko [20] has developed a facial expression and attributes recognition based on multi-task learning of lightweight neural networks. This model uses EfficientNet, which is known for their uniform scaling of dimensions using a user-specified compound coefficient[14]. This is because scaling up along single dimensions such as width, depth and resolution will saturate after reaching around 80%, as shown in equation (1)[14].

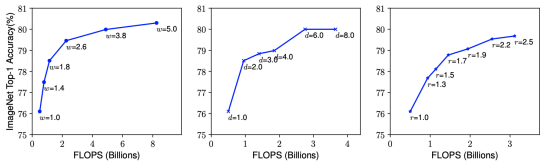


Fig. 5: Single Dimensional Scaling based on Width, Depth and Resolution respectively. Figure By: Tan, M. (2019, May 28). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv.org. https://arxiv.org/abs/1905.11946v5

Research work done by Tan has observed that balancing network, width, depth, and resolution is an integral piece of scaling for ConvNet to help improve accuracy[14]. This is done via a scaling method with a constant ratio. The constant ratio is determined by the parameters in equation (1) below, which includes alpha, beta and gamma. The ratio is ensured in a way such that alpha times beta squared times gamma squared is approximated to 2, where alpha, beta and gamma is constrained to be more than or equal to 1[14].

$$\begin{aligned} depth : d &= \alpha_\phi \\ width : w &= \beta_\phi \\ resolution : r &= \gamma_\phi \\ ratio : \quad \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2 \\ \alpha \geq 1, \beta \geq 1, \gamma &\geq 1 \end{aligned} \tag{1}$$

With added robustness for facial extraction and alignment, it has resulted in higher accuracy numbers of 61.32% for 8 classes of expressions with the using of AffectNet dataset [20] as compared to other methods of less than 60%. This pipeline also has better performance compared to its peers.

IV. Experimental Results

A. First Experiment

The first experiment is to feed the training set of images of depressed patients to the model and obtain the images and the images' facial expression recognition results. The results are shown in Table I. Expression row represents the expression classifier associated with an image, and the count is the number of occurrences where an expression is detected. The model is highly consistent, with retraining of same images will result in similar outcomes.

Based on the results collected from the first experiment, 80% of the depressed patients are likely to evoke expressions that are sadness, surprise, disgust or neutral. Sadness is understandably the one with the highest count, as depression is

Model: EfficientNet

Expression	Sadness	Surprise	Disgust	Neutral
Count	24	22	17	19
Expression	Happiness	Contempt	Anger	Fear
Count	7	8	2	1

Table I: Results depicting expression detected by model

usually linked to having a despairing expression. Even though neutral may be classified under depressed, it is not surprising as there are studies proving that depression is linked to the lack of emotion, which results in a lack of expressions[21]. A split of classifiers can be done to turn 8 separate classifiers into one binary classifier capable of determining depression. The split will be determined based on the count from Table I, where sadness, surprise, disgust and neutral will be classified under depressed, and happiness, contempt, anger, and fear will be classified under non- depressed. The split is represented in Table II below.

Classifier	Expressions Involved
Depressed	Sadness, Surprise, Disgust, Neutral
Not Depressed	Happiness, Contempt, Anger, Fear

Table II: Classifier Grouping

B. Second Experiment

The second experiment is to test our classifier concluded earlier, by testing it against our testing dataset. To test out the classification, a confusion matrix will be used, as shown below in Table III. TP is the number of true positives that denotes depressed patients that predicted as depressed, TN is the number of true negatives that denotes non-depressed patients that are predicted as not depressed, FN is the number of false negatives that denotes depressed patients that are predicted as not depressed, FP is the number of false positives that denotes non-depressed patients. With the confusion matrix, we then calculate the accuracy, recall and precision in equations (2),(3) and (4). The purpose of using accuracy is to determine out of the binary classifier, how many of the test data is predicted correctly. The purpose of using recall is to determine out of all depressed patients, how many of them are predicted correctly. The purpose of using precision is to determine from all the patients that have been predicted as depressed, how many are depressed.

		Prediction	
		Depressed	Not Depressed
Actual	Depressed	True Positive (TP)	False Negative (FN)
	Not Depressed	False Positive (FP)	True Negative (TN)

Table III: Confusion Matrix

$$Accuracy(\%) = \frac{TP + TN}{TP + FN + FP + TN} \times 100 \quad (2)$$

$$Recall(\%) = \frac{TP}{TP + FN} \times 100 \quad (3)$$

$$Precision(\%) = \frac{TP}{TP + FP} \times 100 \quad (4)$$

After running our testing dataset through the same model again, this is the overall result for the confusion matrix in Table IV, while table V showcases the accuracy, recall and precision of the model to determine depression. From the results, it can be determined that our method of classification has around 59% accuracy based on the testing dataset, with 70% recall and 57.38% precision. With this result, the conclusion can be made that this classifier is able to detect depression in depressed individual 70% of the time, while the classifier can accurately predict depression about 57% of the time. One possible justification for the low precision could be due to the test data obtained from patient interviews and consultations not relating to depression. Patients in patient interviews usually express themselves negatively, as patients only attend consultations when they are feeling unwell. Other depression detection research such as eye movement analysis for depression detection by Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. has about 60% recall accuracy[18], while Wang, Q., Yang, H., & Yu, Y. facial expression video analysis for depression detection has about 79% accuracy[17].

		Prediction	
		Depressed	Not Depressed
Actual	Depressed	35	15
	Not Depressed	26	24

Table IV: Confusion Matrix Results

Model	Results for detecting depression	
EfficientNet	Accuracy (%)	59
	Recall (%)	70
	Precision (%)	57.38

Table V: Results

C. Limitations

This method of determining expressions of depression in this research is very subjective since it is determined based on the results of the first experiment to craft a classifier for the second experiment. In addition, there are multiple studies online that concluded that depression varies from individual to individual. Some studies mentioned how depression is linked with reduced emotional exertion[21], while others discuss about how depression is linked to inconsistent emotion regulation[22]. Furthermore, other factors such as gender can also affect emotional language, where women tend to be more expressive with their emotions when depressed, while men are known to be less expressive[23]. This phenomenon is also largely affected by societal norms, family upbringing, or other potential mental issues[24]. With this in mind, there is a need in future work to utilise other expression features, with the

main aim targeting at micro-expressions, which current facial expression models are unable to achieve.

Also, the training and testing dataset might be too small since there are 8 different expressions associated with the expression recognition model, thus some of these expressions might have been overrepresented. However, there are a limited number of videos online that depict clinically proven depressed patients. To best improve this method, there is a need to create a new dataset from scratch. To improve the accuracy of this method, when time permits, the need for ablation studies on larger datasets is required to decide with classes more accurately constitutes depression.

D. Discussion and comparison of results with other related work

There are limited related work regarding solely on facial expression, with Wang, Yang and Yu[17] having research using it to detect depression in Chinese patients. Instead of using available data online, they have volunteered Chinese patients reacting to images, where they then use a camera to record facial expression. They followed up by using facial features points to track facial movement. While having a higher accuracy of around 79%, they have a more sophisticated methodology. They also do share a similar binary classification method instead of a scale to determine depression in users. This overall reduce the complicated process of depression detection, with less computation needed.

Looking at other widespread research in the depression detection field with other different inputs used for depression, multi-modality is another prospect that can further improve accuracy in such detection. An example would be research by Mohan et al. [25], where it focuses on using facial expression and sentiment analysis. This research uses more than facial expressions analysis, including sleep status as well as personality. Sentiment analysis uses user's provided texts to determine depressions through determining negative or positive sentiment. Together with sentiment analysis, the proposed method has seen around 73% accuracy.

V. Conclusions

In this paper, a state-of-the-art facial expression recognition model is used to determine expressions of depressed patients. Images of depressed patients are collected online through different video sources, and then fed through the model to determine patients' expressions. From the results, it is likely that depressed patients would usually show expressions of sadness, surprise, disgust and neutral. A binary classifier is then established and tested using the testing dataset, with the classifier having an accuracy of 60%. However, it is important to note that the datasets are small due to the lack of data present and could lead to bias. Another important thing to note is that based on the results, it is very difficult to only use facial expressions of depressed patients to determine depression, and other classifiers should be introduced and added to facial expressions to increase accuracy of model.

Acknowledgment

Thank you to Tomas Gajarsky for the open-source implementation of Savchenko's HSEmotion library through his facetorch library. This work was carried out within the framework of the research project FRGS/1/2020/ICT02/MUSM/03/5 under the Fundamental Research Grant Scheme provided by the Ministry of Higher Education, Malaysia.

References

[1] W. H. Organisation. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] J. Deng, F. Zhou, W. Hou, *et al.*, "The prevalence of depression, anxiety, and sleep disturbances in covid-19 patients: A meta-analysis," *Annals of the New York Academy of Sciences*, vol. 1486, no. 1, pp. 90–111, 2021.

[3] Y. Yan, J. Hou, Q. Li, and N. X. Yu, "Suicide before and during the covid-19 pandemic: A systematic review with meta-analysis," *International journal of environmental research and public health*, vol. 20, no. 4, p. 3346, 2023.

[4] V. Decker, M. Valenti, V. Montoya, A. Sikorskii, C. W. Given, and B. A. Given, "Maximizing new technologies to treat depression," *Issues in Mental Health Nursing*, vol. 40, no. 3, pp. 200–207, 2019.

[5] J. A. Callan, J. Wright, G. J. Siegle, R. H. Howland, and B. B. Kepler, "Use of computer and mobile technologies in the treatment of depression," *Archives of psychiatric nursing*, vol. 31, no. 3, pp. 311–318, 2017.

[6] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," in *2017 international conference on intelligent sustainable systems (iciss)*, IEEE, 2017, pp. 858–862.

[7] V. Bettadapura, "Face expression recognition and analysis: The state of the art," *arXiv preprint arXiv:1203.6722*, 2012.

[8] B. Buck, A. Chander, J. Tauscher, T. Nguyen, M. Monroe-DeVita, and D. Ben-Zeev, "Mhealth for young adults with early psychosis: User preferences and their relationship to attitudes about treatment-seeking," *Journal of technology in behavioral science*, vol. 6, pp. 667–676, 2021.

[9] N. Burgoyne and A. S. Cohn, "Lessons from the transition to relational teletherapy during covid-19," *Family process*, vol. 59, no. 3, pp. 974–988, 2020.

[10] S. Geller, "Cultivating therapeutic presence: Strengthening your clinical heart, mind, and practice," *Transformance*, vol. 10, pp. 1–25, 2020.

[11] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[12] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[13] A. G. Howard, M. Zhu, B. Chen, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[17] Q. Wang, H. Yang, and Y. Yu, "Facial expression video analysis for depression detection in chinese patients," *Journal of Visual Communication and Image Representation*, vol. 57, pp. 228–233, 2018.

[18] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *2013 IEEE International Conference on Image Processing*, IEEE, 2013, pp. 4220–4224.

[19] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* the MIT Press, 1974.

[20] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, IEEE, 2021, pp. 119–124.

[21] J. Rottenberg, "Mood and emotion in major depression," *Current Directions in Psychological Science*, vol. 14, no. 3, pp. 167–170, 2005.

[22] J. Joormann and I. H. Gotlib, "Emotion regulation in depression: Relation to cognitive inhibition," *Cognition and Emotion*, vol. 24, no. 2, pp. 281–298, 2010.

[23] U. Danielsson and E. E. Johansson, "Beyond weeping and crying: A gender analysis of expressions of depression," *Scandinavian Journal of Primary Health Care*, vol. 23, no. 3, pp. 171–177, 2005.

[24] R. Kosterman, J. D. Hawkins, W. A. Mason, T. I. Herrenkohl, L. J. Lengua, and E. McCauley, "Assessment of behavior problems in childhood and adolescence as predictors of early adult depression," *Journal of psychopathology and behavioral assessment*, vol. 32, no. 1, pp. 118–127, 2010.

[25] M. Mohan, A. Abhinav, A. Ashok, A. Akhil, and P. Achinth, "Depression detection using facial expression and sentiment analysis," in *2021 Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, 2021, pp. 1–6.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

How Transferable are Herbarium-Field Features in Few-Shot Plant Identification with Triplet Loss?

Sophia Chulif*†, Sue Han Lee*, Yang Loong Chang†,
Mark Tee Kit Tsun*, Kok Chin Chai† and Yi Lung Then‡
* Swinburne University of Technology Sarawak Campus, Malaysia
E-mail: schulif@swinburne.edu.my, shlee@swinburne.edu.my, mtktsun@swinburne.edu.my
† NEUON AI SDN. BHD., Malaysia E-mail: yangloong@neuon.ai, kc@neuon.ai
‡ Universiti Malaysia Sarawak, Malaysia E-mail: ylthen@unimas.my

***Abstract*—Due to the limited tropical plant field photos but increasing digitized herbarium specimens, cross-domain plant identification has been employed to investigate the use of herbarium specimens on automated plant identification. However, the domain shift between herbarium and field images makes identifying plant species across these two domains challenging. One of the recent papers shows the superiority of triplet loss in preserving similarities between the crossed domains. Nevertheless, the impact of data deficiency on the performance of this triplet loss model has yet to be studied in depth in this field. Specifically, the transferability of the model features trained from limited cross-domain data to target images in the field domain. Therefore, this paper investigates the robustness of cross-domain plant features learned using this triplet loss metric learning approach compared to the supervised classification approach under general and few-shot experimental settings. Detailed experiments show that the triplet loss metric approach outperformed the supervised classification approach in the few-shot setting and achieved comparable results in the general experimental setting. In addition, the feature dictionary generation schemes composed of various herbarium field feature combinations we proposed boost our models' performance significantly compared to a single feature type dictionary strategy.**

I. Introduction

Over the last few years, we have seen tremendous progress in deep learning algorithms over automated plant identification [1]–[4]. However, these methods require much training data to achieve favorable results [5]. It is often difficult to obtain a balanced training dataset representing each class in a real-world scenario. Such is the case for automated plant identification of tropical species, where plant photographs are scarce. Meanwhile, herbarium images which are dried plant photographs, have been increasingly made available. In fact, the herbarium is one of the primary references for plant experts to identify samples collected in the field [6]. Therefore, cross-domain plant identification appears as one of the recent applications [7], [8] of machine learning and computer vision on digitized herbarium specimens.

Nonetheless, it remains challenging as there is a change in distribution (domain shift) in the training and test data. The training data consists of dried herbarium specimens, while the test data consists of live photographed plants and vice versa, which are highly morphologically different. Furthermore, the various available views or features representing a species in both domains are only partially present. They can be missing, especially for rare species which are difficult to photograph. In the cross-domain plant identification challenges, PlantCLEF 2020 [9] and 2021 [10], which is the largest benchmark in the field, we encoded the inadequacy of convolutional neural networks (CNNs) trained in the supervised and unsupervised manner. We learned that the triplet loss metric learning approach [11], [12] performed the best under the difficult test set with the least photographed species. Moreover, the triplet loss metric learning approach has shown supremacy over several plant diseases [13]–[15] studies.

However, the feasibility of the triplet loss model adaptable to limited training data in the context of the few-shot cross-domain plant identification configuration, has yet to be thoroughly explored. In addition, the triplet network's discriminative herbarium-field feature embedding space has yet to be investigated to produce a more generalized plant feature representation.

Therefore, in this paper, we exploit the robustness of the features learned by the triplet loss metric learning approach and compare its performance to that of the CNN trained with a general supervised recognition method in a few-shot configuration mimicking a realistic scenario. Specifically, we aim to determine the extent to which the methods can generalize to seen and unseen species when training data is limited. We also explore various feature dictionary generation strategies to improve the herbarium-field feature representations explicitly catered for our triplet network.

In short, this paper has two main contributions. First, we experimentally show that the triplet loss metric approach outperforms the traditional supervised classification approach in both seen and unseen classes in the cross-domain few-shot configuration. Second, we empirically studied and implemented various compositions of feature dictionaries from separate domain network streams (herbarium and field) to further encapsulate the representation of each species.

II. Methodology

A. Network Setup

The One-Streamed Mixed (OSM network) comprises a single classical CNN that receives both herbarium and field

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

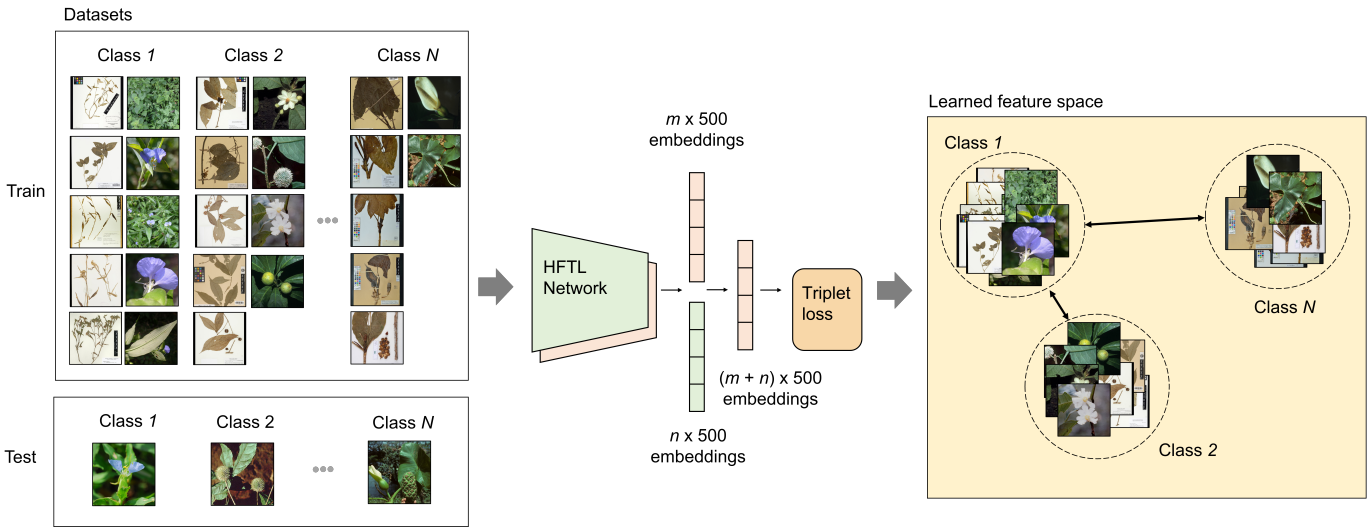


Fig. 1. Overview of the plant identification framework adopting the current state-of-the-art HFTL network to identify the imbalanced composition of herbarium and field images, producing a class discriminating feature space.

image inputs and is optimized by a softmax loss function. It serves as a benchmark.

The Herbarium-Field Triplet Loss (HFTL network) consists of two CNNs: herbarium CNN and field CNN, each receiving herbarium and field images as input, respectively, and each having an extractor layer. Fig. 1 shows the overview of the plant identification framework with the triplet network. The HFTL is optimized by a triplet loss function, and the network details can be found in [16]. Both OSM and HFTL networks use the same base CNN, which is Inception-ResNet-V2 [17]. However, they differ in their network compositions and training strategies.

B. Training Strategies

We initialized both OSM and HFTL networks on pre-trained ImageNet weights but trained HFTL networks in two stages, while OSM networks were in a single step. Since we designed our HFTL networks with two CNNs, the herbarium and field CNNs are each trained separately in the first stage. We then fine-tune the whole HFTL network with our training dataset in the second stage. In contrast, we directly train the OSM networks with our training datasets after initializing them with the pre-trained ImageNet weights. To reduce overfitting and improve generalization, we also apply data augmentation in all the training of the networks. We set up our experiments using the TensorFlow-Slim library with Tensorflow version 1.12.

C. Inference Method

Unlike conventional CNNs, where the classification layers are used to predict the class, deep metric learning models use its CNN (extractor layers) to generate discriminative features for classification. Similarly, we use the extractor layers of the HFTL and OSM networks to obtain the feature embeddings for identification. We do this by comparing the similarities between the extracted feature embeddings and a pre-saved

TABLE I
TRAINING PARAMETERS OF THE NETWORKS.

	OSM Network	HFTL Network
Batch size	128	128
Initial learning rate	0.0001	0.0001
Optimizer	Adam optimizer	Adam optimizer
Weight decay	0.00004	0.00004
Input image size	$299 \times 299 \times 3$	$299 \times 299 \times 3$
Loss function	Softmax cross entropy	Triplet loss

feature dictionary. We construct this pre-saved feature dictionary by assembling the average feature of each trained class from a set of training images. Before averaging each class feature, note that we also cropped each training image into ten variations: top-right, top-left, bottom-right, bottom-left, center and all five mentioned flipped horizontally to retain as much data. Consequently, we calculate the cosine similarity between the extracted feature and the feature dictionary to measure the feature similarities. It is then converted to probabilities using inverse distance weighting to rank the predictions via

$$P_i = \frac{(\frac{1}{d_i})}{\sum_k (\frac{1}{d_k^n})}, \tag{1}$$

where k and i represents the index of the class, d being the embedding distance, and n being the power in positive real number, which in our case is 5. We then use Mean Reciprocal Rank (MRR) score as an evaluation metric to measure the performance of our experiments. It is the average of the reciprocal ranks of the whole test set formulated as:

$$MRR = \frac{1}{\mid Q \mid} \sum_{i=1}^{|Q|} \frac{1}{rank_i}, \tag{2}$$

where $\mid Q \mid$ is the total plant occurrences, $rank_i$ is the predicted rank of the ground truth species for the ith plant

TABLE II
DETAILS OF THE DIFFERENT FEATURE DICTIONARY GENERATION METHODS.

Method	Feature type	Feature extractor	
		HFTL Network	OSM Network
1	Herbarium	Herbarium	OSM
2	Herbarium and field	Herbarium	OSM
3	Herbarium or field	Herbarium or field	OSM
4	Hebarium and field	Hebarium and field	OSM

TABLE III
DETAILS OF THE TRAINING DATASETS COMPRISING HERBARIUM AND FIELD IMAGES.

Dataset	Source	# of species	# of images
Initial	PlantCLEF 2021	435	193,210
435	PlantCLEF 2021, PlantCLEF 2022	435	203,959
608	PlantCLEF 2021, PlantCLEF 2022	608	243,617
435-5	PlantCLEF 2021	435	4,013
608-5	PlantCLEF 2021, PlantCLEF 2022	608	5,691

occurrence.

Feature Dictionary Generation Strategies. As opposed to traditional CNNs where a label is directly predicted, the triplet model relies on a discriminative embedding space for recognition. Therefore, a suitable feature dictionary is explored to boost the existing method. Our dictionary generation strategies include variations of the plant features to save (either the herbarium, field, or both) and the network used to obtain these features. Table II summarizes our feature dictionary generation scheme. *Feature type* indicates the feature to extract and save while *Feature extractor* denotes the network applied to generate these features. For HFTL network, the feature dictionary in method 1 is composed of herbarium features generated by its herbarium network. Method 2 dictionary is composed of herbarium and field features but generated by only its herbarium network. Method 3 dictionary is composed of either herbarium or field. Field features will be saved only if there are field samples available, otherwise herbarium features will be saved. Method 4 dictionary is composed of both herbarium and field. Unlike method 2, method 4 is generated by its herbarium and field networks respectively. These settings are similar for the OSM network. However, note that since OSM has only one stream, the feature extractor used is the same. Consequently, methods 2 and 4 are equivalent for the OSM network.

D. Datasets

We use the datasets from PlantCLEF 2021, PlantCLEF 2022, and our self-collected dataset from various reliable online queries to further analyze the impact of data scaling on model performance.

TABLE IV
DETAILS OF THE TEST DATASETS COMPRISING ONLY FIELD IMAGES.

Dataset	Source	# of species	# of images
Seen	PlantCLEF 2021	338	1,139
Unseen	Online queries	184	706

PlantCLEF 2021. This dataset is currently the largest benchmark dataset for cross-domain plant identification. It consists of 997 species represented by over 321,226 herbarium images and 5,824 field images. Note that only a subset of 435 species contains both herbarium and field training data, which we use for training. The dataset constructed from here is named "Initial" dataset in Table III.

PlantCLEF 2022. This dataset is currently the most extensive benchmark dataset for plant species identification. It consists of 80,000 species represented by four million field images. To build different dataset compositions with a varying number of training images compared to the initial cross-domain plant identification dataset, we exploit the PlantCLEF 2022 dataset. Specifically, we find that there are 185 shared species in PlantCLEF 2022, which corresponds to the 435 existing species in the PlantCLEF 2021 dataset. Here, we are able to build our "435" training set in Table III, accumulating up to 203,959 images from the "Initial" dataset. In addition, we also managed to find 173 additional species in this dataset that also exist in PlantCLEF 2021. Note that, these 173 species are species that contain herbarium images but no field images in PlantCLEF 2021. Therefore, we are able to generate our "608" training set in Table III, which increases the total number of images to 243,617.

Limited 5 training samples. To create a few-shot learning configuration, we construct two additional datasets with fixed limited five training samples. These are datasets "435-5" and "608-5" in Table III. As their name implies, "435-5" is the "435" dataset, but each species contains a maximum of five herbarium training samples and five field training samples. Similarly, "608-5" is the "608" dataset but with each species trained with a maximum of five herbarium samples and five field samples. We limit the number of samples to 5 for each class in order to test our hypothesis on the transferability of features across different domains using a triplet network, and to compare its performance with a supervised classification network.

Test sets. Our Seen test set is constructed from the PlantCLEF 2021 dataset. It is a set of 338 species from the 435 species that have both herbarium and field images. Meanwhile, our Unseen test set is generated from multiple online queries from sources such as GBIF and iNaturalist. Since we want this Unseen test set to contain species not in training, we collected 184 species neither from PlantCLEF 2021 nor 2022. Due to the difficulty of obtaining more field images, our test sets combined are only

TABLE V
RESULTS OF THE MODELS BASED ON THE FEW-SHOT SETTING IN THE SEEN TEST SET.

Model	Training dataset	Method 1	Method 2	Method 3	Method 4
OSM-435-5	435-5	0.3116	0.4791	**0.5733**	0.4791
OSM-608-5	608-5	0.1594	0.3472	**0.4525**	0.3472
HFTL-435-5	435-5	0.6962	0.6168	**0.7213**	0.7122
HFTL-608-5	608-5	0.5807	0.4375	0.4688	**0.6374**

TABLE VI
RESULTS OF THE MODELS BASED ON THE FEW-SHOT SETTING IN THE UNSEEN TEST SET.

Model	Training dataset	Method 1	Method 2	Method 3	Method 4
OSM-435-5	435-5	**0.0556**	0.046	0.0543	0.046
OSM-608-5	608-5	**0.0583**	0.0405	0.0354	0.0405
HFTL-435-5	435-5	0.0671	**0.0677**	0.0526	0.0558
HFTL-608-5	608-5	0.101	**0.1096**	0.1047	0.068

focused on over 500 species as shown in Table IV.

III. EXPERIMENTS AND RESULTS

Our primary investigation concerns the study of few-shot learning and classical learning in cross-domain plant identification. Typically, CNNs perform poorly with inadequate training data and favorably otherwise. Therefore, these settings are studied to investigate if the HFTL network serves as a conventional OSM network benchmark.

A. Few-shot Setting

To exploit the capability of the models adaptable to the few shot setting, we train and compare the performance of several models on training sets with a fixed limited amount of images as mentioned in Sec. II-D. Table V and VI show the classification results on the seen and unseen test set respectively. We found that the metric learning approach (HFTL-435-5, HFTL-608-5) performed better than the traditional supervised approach (OSM-435-5, OSM-608-5) in both the seen and unseen test sets. We see a performance gap of over 15% from the metric learning approach against the classical supervised approach in the seen test set and up to 5% performance gap in the unseen test set. It confirms the superiority of triplet loss over traditional supervised methods in preserving similarities between cross-domains and the feasibility of the model adapting to limited training data.

B. Classical Setting

On the other hand, we also train and compare the performance of the models on a larger scale training data ("Initial", "435" and "608"). Note that these datasets do not have fixed limits on the number of images and may be imbalanced. Tables VII and VIII show their classification results on the seen and unseen test sets, respectively. Naturally, with the increase of training data, the performance of both approaches shows improvements. However, unlike the few-shot setting, we found that the metric learning approach (HFTL, HFTL-435, HFTL-608) performed better in the seen test set, but the traditional supervised method (OSM, OSM-435, OSM-608) performed better in the unseen test set. In the seen test set, we see a performance gap of approximately 30% from the metric learning approach, while a performance gap of approximately 10% achieved by the classical supervised method in the unseen test set. It shows the advantage of the metric learning approach in generalizing seen species and the classical supervised approach in generalizing unknown species better when there is sufficient training data. In addition, we find the increase in training classes for both approaches in the seen test set; specifically models OSM-608 and HFTL-608, does not significantly improve the performance as expected. It may be due to the fact that the seen test set is based on the 435 species and the reason why models trained with 435 species (OSM-435, HFTL-435) generalize best to this seen test set.

TABLE VII
RESULTS OF THE MODELS BASED ON CLASSICAL SETTING IN THE SEEN TEST SET.

Model	Training dataset	Method 1	Method 2	Method 3	Method 4
OSM	Initial	0.4783	0.6277	**0.6596**	0.6277
OSM-435	435	0.5488	0.6906	0.6885	**0.6906**
OSM-608	608	0.5157	0.6379	**0.6546**	0.6379
HFTL	Initial	0.7361	0.6205	0.736	**0.7443**
HFTL-435	435	0.7407	0.7832	0.8318	**0.9909**
HFTL-608	608	0.7172	0.5649	0.5879	**0.7366**

TABLE VIII
RESULTS OF THE MODELS BASED ON THE CLASSICAL SETTING IN THE UNSEEN TEST SET.

Model	Training dataset	Method 1	Method 2	Method 3	Method 4
OSM	Initial	**0.1611**	0.1129	0.0608	0.1129
OSM-435	435	**0.1595**	0.0923	0.0584	0.0923
OSM-608	608	**0.2184**	0.1176	0.065	0.1176
HFTL	Initial	0.064	**0.0666**	0.0617	0.06
HFTL-435	435	0.0957	**0.115**	0.1055	0.0867
HFTL-608	608	0.107	**0.1209**	0.1062	0.0888

C. Feature Generation Scheme

Regarding the various dictionary generation methods experimented on, we find that methods 1 and 2 work best under the unseen test set, while methods 3 and 4 work best under the seen test set. Since method 3 and 4 introduces field features to the feature dictionary, they help match the learned field embeddings of the seen species against the dictionary. In contrast, since there is no field training images for the unseen species, method 3 and 4 does not introduce useful field images to the class representations. Consequently, methods 1 and 2 are better representations of the unseen test set. These findings can be used to construct suitable feature dictionary for supervised

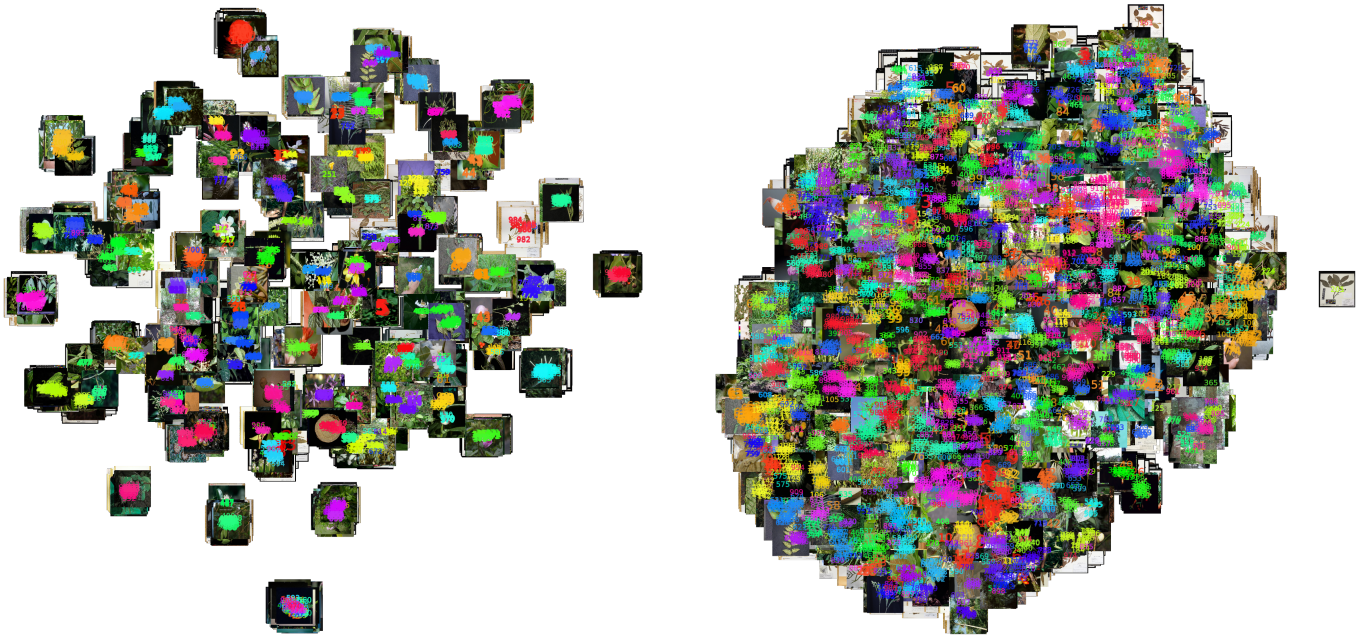


Fig. 2. T-SNE visualizations of 261 plant species and 4,402 images from the training dataset. (a) is generated from the few-shot HFTL-608-5 model. (b) is generated from the few-shot OSM-608-5 model.

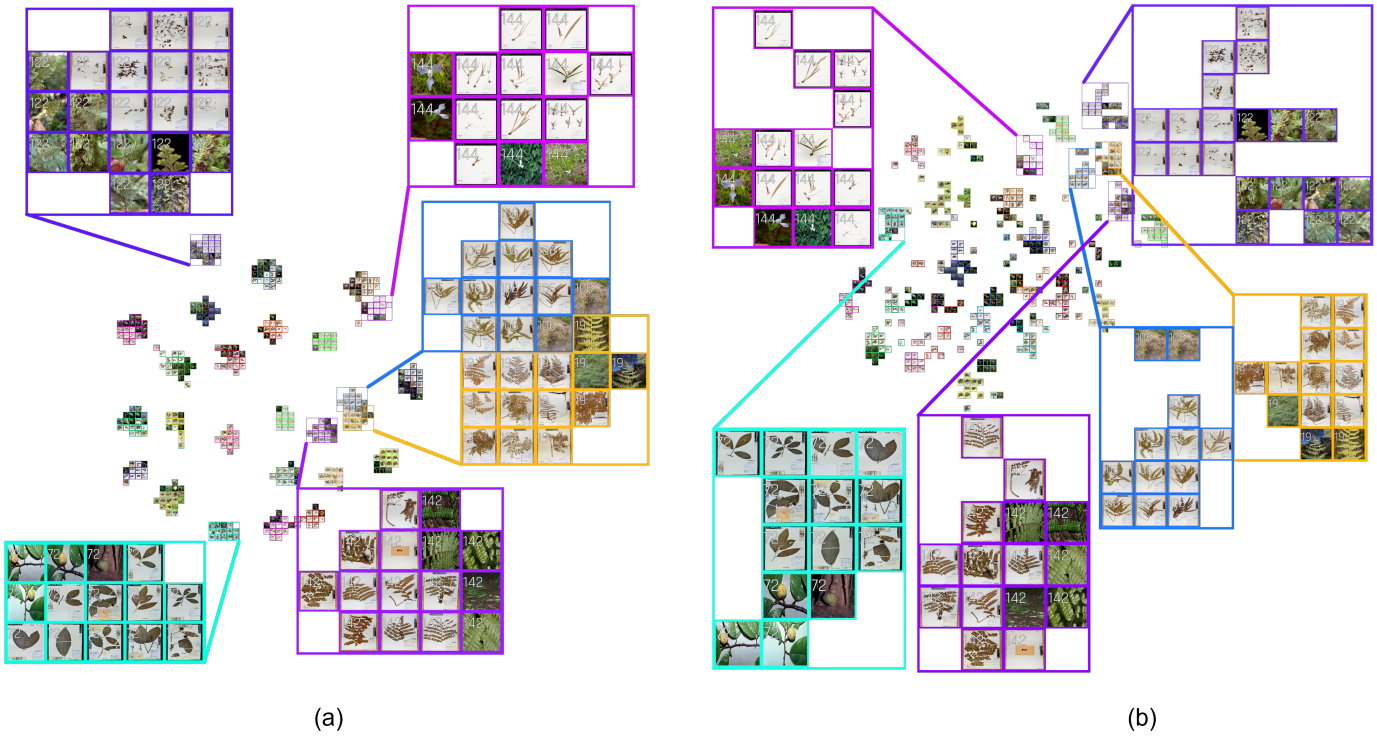


Fig. 3. A close-up T-SNE visualizations of 25 plant species and 392 images from the training dataset. (a) is generated from the few-shot HFTL-608-5 model. (b) is generated from the few-shot OSM-608-5 model.

and unsupervised cross-domain plant recognition whereby field features can be added to accommodate classification.

D. Visualizations

To visualize and compare the feature embeddings generated from the HFTL and OSM models, we use t-distributed Stochastic Neighbor Embedding (t-SNE) [18] to assess the performance of these models under few-show settings. t-SNE is a dimension reduction method that converts and visualizes high-dimensional data into a low two or three-dimensional space. Therefore, the feature embeddings generated can be represented and viewed in a two-dimensional image form. Fig. 2 presents the t-SNE visualizations showing 4,402 herbarium and field images (embeddings) of 261 species from the training dataset. We randomly selected two species from all the available family taxon to balance out the samples. In Fig. 2 , the images are accompanied by their respective species id. Each color represents a unique species id of the above mentioned 261 species. In addition, the images are grouped according to how similar their embeddings are to one another. The closer the images are stacked together, the more similar their embeddings are to one another. On the other hand, the further the images are from one another, the more different their embeddings are to one another.

Fig. 2 (a) is constructed from the HFTL-608-5 model, while Fig. 2 (b) is constructed from the OSM-608-5 model, in which both were trained under the few-shot settings. Similar to the results obtained in the experiments, HFTL-608-5 is better at distinguishing between species than OSM-608-5. As observed, the herbarium and field samples are better grouped according to their species in Fig. 2 (a) than in Fig. 2 (b). Based on the colored species ids, we can see that the HFTL-608-5 model is able to separate different species more clearly compared to the OSM-608-5.

In addition, we demonstrate the close-up t-SNE visualizations of a smaller training subset of 25 species and 392 images in Fig. 3. Unlike Fig. 2, the herbarium and field images are not computed to stack on top of one another to allow a clearer view. The close-up images' different border colors represent the different species. Meanwhile, the white numbered text in the close-up images represent the species id number. We randomly selected 25 species from 25 different family taxon to allow a clearer observation of the visualizations. Fig. 3 (a) is constructed from HFTL-608-5, while Fig. 3 (b) is constructed from OSM-608-5. Likewise, the HFTL-608-5 model shows better species recognition as the embeddings are well separated and clustered compared to OSM-608-5. Although the HFTL model can better recognize the plant species and give a distinct separation between different class embeddings, it is also noted that when the samples are very similar, the separation is weaker. It can be observed in the close-up samples with blue and yellow borders, which are visually similar. Consequently, we look forward to enhancing the maximization of the embedding distance between visually similar species for the triplet network by using their higher taxonomy classes to overcome this in future experiments.

IV. CONCLUSION

This work investigated and compared the feature generalization and transferability between the triplet network and the classical supervised learning network in this few-shot cross-domain plant identification. We found that the triplet network could learn more robust features despite a limited training sample and could further harness the similarity between different domains with our proposed feature generation scheme. This result is an important reference for researchers working in this field and a reliable reference for processing smaller plant data for practical applications. In future, we will extend the work to incorporate and exploit the potential power of the supervisory network to leverage this distance metric approach for better feature generalization to both seen and unseen data.

ACKNOWLEDGMENT

This research was supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant No. (Ref:FRGS/1/2021/ICT02/SWIN/03/2), from the Ministry of Higher Education Malaysia and NEUON AI SDN. BHD., Malaysia.

REFERENCES

[1] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, "Deep-plant: Plant identification with convolutional neural networks," in *2015 IEEE international conference on image processing (ICIP)*, IEEE, 2015, pp. 452–456.

[2] S. H. Lee, Y. L. Chang, C. S. Chan, and P. Remagnino, "Hgo-cnn: Hybrid generic-organ convolutional neural network for multi-organ plant classification," in *2017 IEEE international conference on image processing (ICIP)*, IEEE, 2017, pp. 4462–4466.

[3] J. Krause, K. Baek, and L. Lim, "A guided multi-scale categorization of plant species in natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[4] S. H. Lee, C. S. Chan, and P. Remagnino, "Multi-organ plant classification based on convolutional and recurrent neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4287–4301, 2018.

[5] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4918–4927.

[6] A. López and A. B. Sassone, "The uses of herbaria in botanical research. a review based on evidence from argentina," *Frontiers in Plant Science*, vol. 10, p. 1363, 2019.

[7] B. R. Hussein, O. A. Malik, W.-H. Ong, and J. W. F. Slik, "Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review," *Ecological Informatics*, p. 101 641, 2022.

[8] J. Carranza-Rojas, H. Goeau, P. Bonnet, E. Mata-Montero, and A. Joly, "Going deeper in the automated identification of herbarium specimens," *BMC evolutionary biology*, vol. 17, no. 1, pp. 1–14, 2017.

[9] A. Joly, H. Goëau, S. Kahl, *et al.*, "Overview of lifeclef 2020: A system-oriented evaluation of automated species identification and species distribution prediction," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, Springer, 2020, pp. 342–363.

[10] A. Joly, H. Goëau, S. Kahl, *et al.*, "Overview of lifeclef 2021: An evaluation of machine-learning based species identification and species distribution prediction," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, Springer, 2021, pp. 371–393.

[11] S. Chulif and Y. L. Chang, "Herbarium-field triplet network for cross-domain plant identification. neuon submission to lifeclef 2020 plant.," in *CLEF (Working Notes)*, 2020.

[12] S. Chulif and Y. L. Chang, "Improved herbarium-field triplet network for cross-domain plant identification: Neuon submission to lifeclef 2021 plant.," in *CLEF (Working Notes)*, 2021, pp. 1526–1539.

[13] D. Argüeso, A. Picon, U. Irusta, *et al.*, "Few-shot learning approach for plant disease classification using images taken in the field," *Computers and Electronics in Agriculture*, vol. 175, p. 105 542, 2020.

[14] T. V. Nguyen and I. Paik, "Feature extraction with triplet loss to classify disease on leaf data," in *2020 11th International Conference on Awareness Science and Technology (iCAST)*, IEEE, 2020, pp. 1–5.

[15] A. Afifi, A. Alhumam, and A. Abdelwahab, "Convolutional neural network for automatic identification of plant diseases with limited data," *Plants*, vol. 10, no. 1, p. 28, 2020.

[16] S. Chulif, S. H. Lee, Y. L. Chang, and K. C. Chai, "A machine learning approach for cross-domain plant identification using herbarium specimens," *Neural Computing and Applications*, pp. 1–23, 2022.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017. DOI: 10 . 1609 / aaai . v31i1 . 11231. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11231.

[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Resolution-Adaptive Lossless Image Compression Using Frequency Decomposition Network

Hochang Rhee* and Nam Ik Cho†
* Department of ECE, INMC, Seoul National University, Seoul, Korea
E-mail: hochangispl@gmail.com Tel/Fax: +82-2-880-8480
† Department of ECE, INMC, Seoul National University, Seoul, Korea
E-mail: nicho@snu.ac.kr Tel/Fax: +82-2-880-1810

***Abstract*—This paper introduces a new lossless image compression method that aims to effectively utilize image correlations from two distinctive perspectives. Firstly, a hierarchical approach is employed to decompose the input image into subimages, and a new encoding order is designed for efficient subimage encoding. We enable the method to fully exploit the horizontal, vertical, and diagonal components of previously encoded subimages. Secondly, each subimage is further decomposed into low and high-frequency components, where the low-frequency components are first compressed. Then the low-frequency components are used as additional input for compressing the remaining high-frequency ones. In this scenario, the low-frequency ones serve as strong prior for encoding the higher frequencies. In addition to improving the compression performance, we make our method adapted to image sizes, specifically the performance differences depending on the input resolution, by introducing resolution-specific kernels. Experiments show that the proposed method achieves state-of-the-art performance for high and low-resolution benchmarks.**

I. Introduction

There have been some progresses in lossless image compression performances by using deep neural networks (DNNs) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. Based on previous conventional non-learning methods (BPG, PNG, JPEG-LS, JPEG-XL) [13], [14], [15], [16] using the closed-loop pixel prediction, learning-based lossless compression methods [1], [2], [3], [4], [5], [6] replace the predictor with powerful DNNs to make a more precise prediction. These methods, which encode pixels one by one, are called autoregressive models. Early autoregressive methods such as PixelCNN [1], PixelCNN++ [2], and PixelRNN [4] predict the probability of an encoding pixel conditioned on all previous ones. However, these methods have two limitations. First, the inference time is impractical since they require neural network computations for the number of whole pixels. Next, these methods proceed in the raster scan order while there can be other orders that can exploit more correlated pixels.

To reduce the inference time, recent learning-based methods [3], [5], [6], [37] proceed with the encoding in the unit of subimages rather than individual pixels. For instance, MS-PixelCNN [3] divides the input into subimages and sequentially compresses them one by one. L3C [5] estimates the probability distribution starting from a uniform distribution and progressively upsamples to the final image. RC [6] divides the image into a lossy compressed image and its residuals, where the residuals are losslessly encoded. Although these methods achieve reasonable inference time, they demonstrate lower compression efficiency compared to start-of-the-art conventional methods, such as FLIF and JPEG-XL.

Furthermore, it is important to note that learning-based methods exhibit limited robustness to variations in image resolution. While these methods may demonstrate comparable performance when applied to high-resolution images (2K), their compression efficiency significantly deteriorates when dealing with low-resolution counterparts, such as 512×512. This decline in compression efficiency arises from the training bias of learning-based methods towards high-resolution images, which restricts their adaptability.

To address the robustness of encoding performance against image resolution variations, this paper presents a resolution-adaptive lossless compression scheme, based on our previous compression method using frequency decomposition network (LC-FDNet) [37]. The LC-FDNet provides state-of-the-art lossless compression performances by developing a hierarchical compression scheme that maximizes the correlation between previous and current subimages. The approach begins by decomposing the input into subimages based on spatial location and color channel. A hierarchical scheme is then employed to determine the compression order of these subimages, effectively utilizing the horizontal, vertical, and diagonal components for predicting the current encoding pixel. For each subimage, we further decompose it into low and high-frequency components to leverage a coarse-to-fine scheme. This is motivated by the observation that separate predictors are more effective for low and high-frequency regions, as high-frequency regions typically exhibit higher prediction errors. By compressing the low-frequency regions first and utilizing them as additional input for predicting the remaining high-frequency regions, we exploit the strong prior provided by the low-frequency regions. Moreover, the frequency decomposition process is designed to be image-specific, allowing the network to determine the optimal ratio of low and high-frequency regions for each image. On the basis of the LC-FDNet, this paper introduces a novel Resolution Refinement Block (RRB) to address the domain gap resulting from resolution variations across datasets. This block leverages image resolution information to dynamically generate resolution-specific convolution kernels. By applying these resolution-specific kernels to the intermediate features of the network, we enhance the network's

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

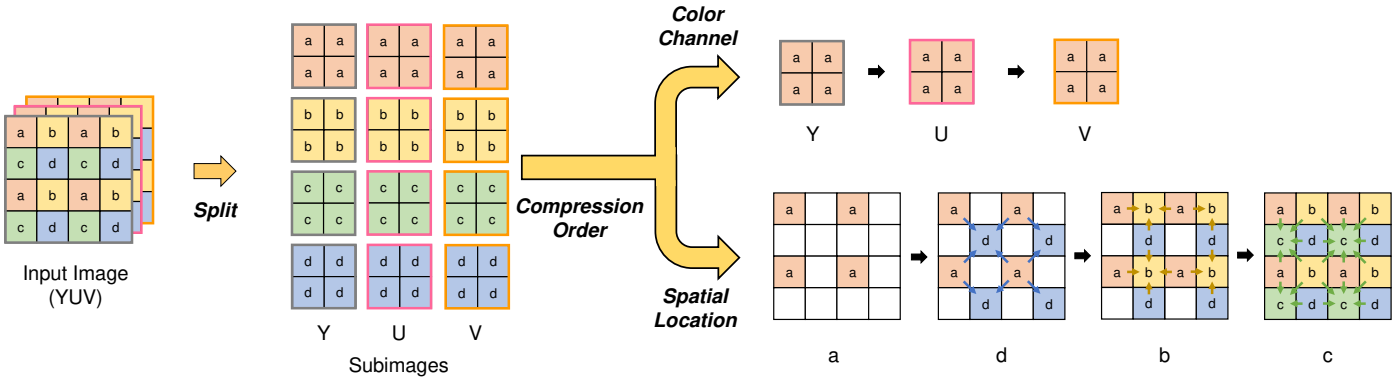


Fig. 1: Illustration of the overall compression scheme. We first split the input into 12 subimages (four for each channel) depending on spatial location and color channel, where the border color of a subimage represents the color channel. Then, we compress the subimages one by one, where the order of the compression is determined based on color channel and spatial location.

ability to handle image resolution dependencies and improve overall performance.

In summary, our method proceeds with the encoding in a coarse-to-fine manner in two aspects. First, we decompose an input image in a hierarchical fashion to maximally exploit the correlation between the decomposed subimages. Secondly, we decompose the subimages into low and high-frequency regions and use low-frequency components to compress the high-frequency ones better. Furthermore, we use image resolution as additional input to refine the intermediate features, which drives our network to be robust to image resolution. Experiments show that the proposed method achieves state-of-the-art performance for benchmark datasets with reasonable inference time.

II. Method

A. Overview

The framework of our method is illustrated in Fig. 1. We first convert an input RGB image $I \in \mathcal{R}^{H\times W\times 3}$ into a YUV through a reversible color transform [26], where H and W are the height and width of the image, respectively. Then, we split each of the YUV into 4 subimages, resulting in total of 12 subimages $x_{c,s} \in \mathcal{R}^{\frac{H}{2}\times\frac{W}{2}\times 1}$ based on the color channel and spatial location, where c is the channel index ($c \in \{Y, U, V\}$) and s is the spatial location index ($s \in \{a, b, c, d\}$). The subimages are compressed one by one, where the compression order of the subimages is designed to maximize the compression efficiency. The details of the spatial location index and compression order will be described in the next subsection. We intially compress 3 subimages $\{x_{Y,a}, x_{U,a}, x_{V,a}\}$ using a conventional compression algorithm. Then, the remaining subimages are sequentially compressed using learned networks, where the previously encoded subimages and the image resolution (HW) are given as input.

B. Compression Order

We decompose each color channel into 4 subimages indexed as a, b, c, d, which is shown in Fig. 1. The encoding of a subimage is conditioned on the previously encoded subimages. Specifically, for compressing the N-th subimage $y \in \mathcal{R}^{\frac{H}{2}\times\frac{W}{2}\times 1}$, our system takes as input the $N-1$ subimages that have already been encoded ($x_{in} \in \mathcal{R}^{\frac{H}{2}\times\frac{W}{2}\times(N-1)}$) along with the image resolution (HW). The set of previously encoded subimages (x_{in}) serves as strong prior for predicting and compressing the N-th subimage y, while the image resolution is utilized for refining the intermediate features. The correlation between the input subimages x_{in} and the N-th subimage y plays a crucial role in compression performance, as higher correlation leads to better estimation and enhanced compression efficiency. Therefore, arranging the subimages in an optimal order is essential to maximize correlation utilization.

We determine the order of the subimages by considering both color channel and spatial order. Regarding the color channel order, since Y contains more significant features than U and V, a natural choice is to compress them in the order of $Y \rightarrow U \rightarrow V$. Regarding the spatial order, MS-PixelCNN [3] uses a similar subimage splitting approach proceeding in the naive scan order of $a \rightarrow b \rightarrow c \rightarrow d$. We have observed that this order may not maximize the correlation between spatial locations. We observe that $a \rightarrow d$ maximizes the correlation compared to $a \rightarrow b$, where d fully exploits the information of a both horizontally and vertically, while b benefits from a but fails to utilize the vertical components fully. Based on this observation, we set the order as $a \rightarrow d \rightarrow b \rightarrow c$. In conclusion, we proceed with the compression of subimages in the order of $x_{Y,a} \rightarrow x_{U,a} \rightarrow x_{V,a} \rightarrow x_{Y,d} \rightarrow x_{U,d} \rightarrow \cdots \rightarrow x_{V,c}$, where we encode the initial subimages $x_{Y,a}, x_{U,a}, x_{V,a}$ using a conventional codec, such as JPEG-XL [16].

C. Architecture

This section presents the architecture of the proposed method illustrated in Fig. 2, which shows the process of compressing the N-th subimage y given the input x_{in} (concatenation of $N-1$ previously encoded subimages) and image resolution (HW). As shown in the figure, our method consists of the following steps: Backbone Network, Adaptive

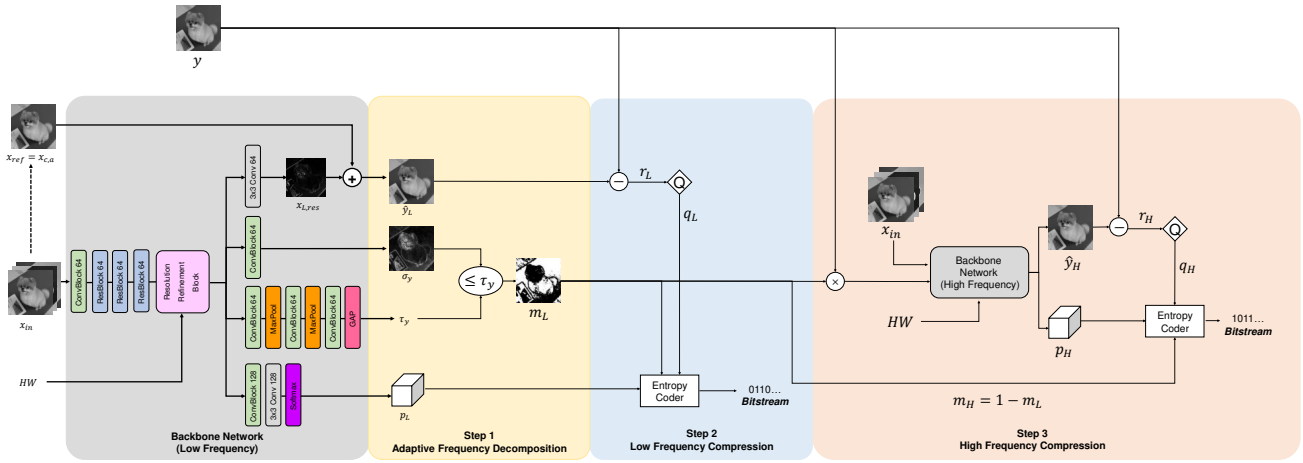


Fig. 2: The encoder architecture of the proposed method. In this figure, we consider the case of compressing $y = x_{Y,d}$ given $x_{in} = \{x_{Y,a}, x_{U,a}, x_{V,a}\}$. The input first goes through a backbone network (Fig. 3) that outputs several parameters and features for the next steps of encoding. Step 1, named adaptive frequency decomposition (AFD), receives the outputs from the backbone and determines whether each pixel belongs to either low or high-frequency regions. Then, Step 2 (low-frequency compression (LFC)) encodes the low-frequency region of the subimage y. The final Step 3 (high-frequency compression (HFC)) receives the encoded low-frequency region as additional input and compresses the remaining high-frequency region.

Frequency Decomposition (AFD), Low-Frequency Compression (LFC), and High-Frequency Compression (HFC). AFD decomposes the subimage by frequency, classifying each pixel into either low or high frequency. Then in LFC, only the low-frequency components are compressed. Finally, the remaining high-frequency components are compressed using the already encoded low-frequency region as additional input to the HFC. Throughout the paper, notations L and H denote low and high-frequency, respectively.

Initially, the already encoded subimages and the image resolution are fed to the backbone network and generates four intermediate outputs: 1) subimage prediction $\hat{y}_L$, 2) probability distribution p_L, 3) error variance map σ_y, and 4) error variance threshold τ_y. In the backbone, x_{in} first passes through one ConvBlock and three ResBlocks and generates an intermediate feature. The ConvBlock is composed of 3×3 convolution layer, Batch Normalization, and ReLU. The detailed architecture of the ResBlock is illustrated in Fig. 3. The intermediate feature is used to generate four outputs. However, we observe from experiments that the generated intermediate feature lacks the robustness to image resolutions.

Thus, we add a Resolution Refinement Block (RRB) (illustrated in Fig. 3) that explicitly uses the image resolution to refine the intermediate feature. We feed the information of image resolution as the number of pixels in the image *i.e.*, HW. In the RRB, we begin by applying positional embedding [43] to expand the dimensionality of the image resolution. Prior researches, such as in NeRF [44] have demonstrated that mapping the inputs to a higher dimensional space using high frequency functions can enhance the fitting of data with high-frequency variations. Afterwards, the embedding vector is passed through two MLPs to generate 3×3 convolution

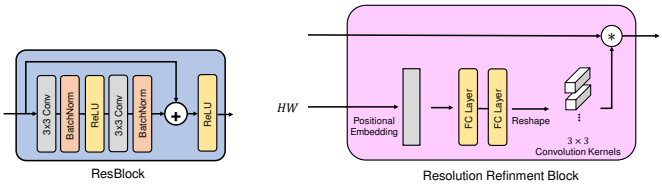


Fig. 3: Architecture of Backbone Network (Low Frequency)

kernels, which are then applied to the intermediate feature. If the channel dimension of the intermediate feature is C, then the output of the MLPs has a dimension of $3\times3\times C\times C$. The generated convolution kernels drive the intermediate feature to reflect the image resolution.

The refined feature is further processed through additional convolution layers to generate four outputs. Firstly, $\hat{y}_L \in \mathcal{R}^{\frac{H}{2}\times\frac{W}{2}\times1}$ represents the prediction of y, with better predictions resulting in more compact compression. Following the standard practice in lossless compression, we calculate the residual $r_L = \hat{y}_L - y$. As the residual is not in the form of integers, we quantize (round) it, denoting the quantized residual as q_L, which is then passed to the entropy coder.

Secondly, p_L is the estimated probability distribution of the quantized residual q_L. The probability distribution is directly estimated as the probability mass function (pmf), and thus the dimension of p_L is $\frac{H}{2}\times\frac{W}{2}\times K$, where K is 511 for the Y channel and 1021 for the U, V channels. To ensure that the probabilities sum up to 1, we apply the softmax operation to derive p_L.

For classifying each pixel as belonging to low or high-frequency regions, we introduce an error variance map $\sigma_y \in \mathcal{R}^{\frac{H}{2}\times\frac{W}{2}\times1}$, which is obtained through learning with an ap-

propriate loss function. The map intensity reflects the pixel's position in a high-frequency or low-frequency region, with higher intensity indicating a high-frequency pixel. To design the error variance map, we utilize the magnitude of the prediction error, and the loss function used for this purpose is defined as:

$$L_{ev} = \|\sigma_y - |y - \hat{y}_L|\|1. \tag{1}$$

Lastly, we learn a threshold $\tau_y \in \mathcal{R}$, which is applied to the error variance map to classify pixels as low or high-frequency regions. We have observed that the error variance heavily depends on the color channel, spatial location, and image characteristics. To adapt to different inputs, the network generates threshold values, with each subimage having its own threshold. In total, nine threshold values are generated.

Especially for the generation of the subimage prediction $\hat{y}_L$, we adopt a residual scheme. That is, instead of directly estimating the pixel value of the subimage y, the network estimates $x_{L,res}$, which is the difference between y and a reference image x_{ref} (set as $x_{c,a}$). For instance, in the case of compressing $y = x_{Y,d}$, the network sets $x_{Y,a}$ as offset and estimates the residual between $x_{Y,a}$ and $x_{Y,d}$. The final output $\hat{y}_L$ is derived as $x_{L,res} + x_{Y,a}$. In other words, the network estimates the residual with respect to x_{ref}. Since x_{ref} and y tend to have a high correlation, their difference usually has low variance, with the mean close to zero. This leads to stable training and performance enhancement.

After obtaining the above outputs, our method categorizes each pixel as either low or high-frequency in Step1 (AFD) of Fig. 2. With σ_y and τ_y in hand, this step generates a binary map as

$$m_L^i = \begin{cases} 1 & \text{if } \sigma_y^i \leq \tau_y \\ 0 & \text{else,} \end{cases} \tag{2}$$

indicating which components are considered by the LFC, where i denotes the pixel index.

Then, in the LFC (Step 2), we only compress the low-frequency components, specifically the pixels corresponding to $m_L^i = 1$. The quantized residual q_L is converted into bitstreams using the corresponding probability distribution p_L. We follow a low to high frequency compression order since the low-frequency region is more easily compressible. Given the input x_{in}, the compression efficiency in the low-frequency region is already optimized, and providing high-frequency components as additional input would only yield marginal performance enhancement. The significant gain in compression efficiency is achieved when the low-frequency components serve as an additional input for the high-frequency ones.

The HFC (Step 3) follows a similar pipeline to the LFC but introduces an additional input ($y \odot m_L$) alongside the inputs of the LFC (x_{in} and HW). The input $y \odot m_L$ represents the low-frequency components of the current subimage. In the HFC, these inputs are fed into the backbone network, which has the same architecture as the LFC but without the branches for deriving σ_y and τ_y. Consequently, we obtain two outputs: 1) $\hat{y}_H$, which represents the prediction of y in the high-frequency region, and 2) p_H, the probability distribution of the quantized residual q_H. The utilization of low-frequency components as a strong prior enables more accurate predictions in the high-frequency region.

Finally, $\hat{y}_H$ and p_H are fed into the entropy coder, allowing us to encode the high-frequency components, specifically the pixels where $m_L^i = 0$ or $m_H^i = 1 - m_L^i = 1$. The HFC can disregard the estimation of low-frequency components and solely focus on the high-frequency ones.

D. Loss Function

Our network is trained with the following losses. First is the reconstruction loss that derives the predicted subimage to match the ground truth. The reconstruction loss is defined as the L1 loss:

$$L_{rec} = m_L \cdot \|y - \hat{y}_L\|_1 + m_H \cdot \|y - \hat{y}_H\|_1. \tag{3}$$

We multiply the corresponding frequency mask to the reconstruction loss of LFC and HFC. By doing so, we ensure that only low-frequency components contribute to the reconstruction loss in LFC, while the high-frequency components contribute solely to the reconstruction loss in HFC.

Additionally, we incorporate the bitrate loss, which aims to minimize the cross-entropy between the true probability distribution of the quantized residual (p_{q_L}, p_{q_H}) and the estimated distributions (p_L, p_H) respectively. It is defined as

$$L_{br} = m_L \cdot \|-\log p_L(q_L)\|_1 + m_H \cdot \|-\log p_H(q_H)\|_1. \tag{4}$$

The bitrate loss encourages the model to accurately predict the probability distribution of the quantized residuals, thereby improving the overall compression performance. To restrict the contribution of each frequency component as in the reconstruction loss, we multiply the frequency masks to the corresponding reconstruction loss.

Altogether, we train our network with the loss:

$$L = L_{br} + \lambda_{rec} L_{rec} + \lambda_{ev} L_{ev} \tag{5}$$

where L_{ev} is the error variance loss mentioned in Eq. 1. λ_{ev} and λ_{rec} are the balancing hyperparmeters. In our experiments, we set both λ_{ev} and λ_{rec} as 1.

III. Experiments

A. Experiment Setup

Implementation Detail For the quantization of residuals r to q, we use the round function *i.e.*, $q = round(r)$. In this case, gradient-based optimization cannot be used since the derivative is zero except at integers. Thus, we approximate the round function as STE [38] *i.e.*, $q = r$ in the backward pass. Although other quantization approximation methods can also be applied, [39] has shown that various methods have a minor effect on the compression performance. In addition, deriving m_L with

TABLE I: Comparison of our method and other non-learning and learning-based codecs on high-resolution benchmark datasets. We measure the performances in bits per pixel (bpp). The difference in percentage to our method is highlighted in green.

Method	CLIC.m	CLIC.p	DIV2K
PNG [14]	11.79 +71.4%	11.79 +51.9%	12.69 +57.6%
JPEG-LS [15]	7.59 + 10.3%	8.46 +9.0%	8.97 +11.4%
JPEG2000 [34]	8.13 +18.2%	8.79 +13.3%	9.36 +16.3%
WebP [36]	8.19 +19.0%	8.70 +12.1%	9.33 +15.9%
BPG [13]	8.52 +23.8%	9.24 +19.1%	9.84 +22.2%
LCIC [12]	7.88 +14.5%	9.02 +16.2%	9.35 +16.1%
FLIF [27]	7.44 +8.1%	8.16 +5.2%	8.73 +8.4%
JPEG-XL [16]	7.20 +4.7%	8.19 +5.5%	8.49 +5.5%
L3C [5]	7.92 +15.1%	8.82 +13.7%	9.27 +15.2%
RC [6]	7.62 +10.8%	8.79 +13.3%	9.24 +14.8%
Near-Lossless [7]	7.53 +9.4%	7.98 +2.8%	8.43 +4.7%
LC-FDNet[37]	6.90 +0.3%	7.86 +1.3%	8.14 +1.1%
Ours	6.88	7.76	8.05

Eq. 2 introduces the same problem, where the derivative is zero except for the threshold. In this case, we approximate as $m_L = \text{sigmoid}(-(\sigma_y - \tau_y))$ in the backward pass. For the entropy coder, we employ "torchac," which is a fast arithmetic coding library for PyTorch developed by the authors of L3C [5].

Evaluation Dataset For the 2K high-resolution evaluation dataset, we use the following three benchmark datasets, CLIC.m, CLIC.p, and DIV2K. The CLIC mobile (CLIC.m) and CLIC professional (CLIC.p) are datasets released as part of the "Workshop and Challenge on Learned Image Compression" [28]. The CLIC.m consists of 61 evaluation images taken by mobile phones. The CLIC.p contains 41 evaluation images by DSLRs. The image resolutions in the CLIC datasets are mostly 2K, but some have low resolutions down to 512×384. The DIV2K [29] is a super-resolution dataset comprising 2K resolution high-quality images, where 100 images are provided for evaluation.

For the low-resolution evaluation dataset, the following three datasets are used: Kodak [41], McMaster [40], and Classic. Kodak contains 24 images with a resolution of 512×768. McMaster includes 18 images of size 500×500, which are cropped from high-resolution images. The Classic dataset consists of 12 popular images, which generally have a resolution of 512×512. For the Classic dataset, we use the following images: airplane, Barbara, coastguard, comic, flowers, goldhill, Lena, mandrill, monarch, pepper, ppt3, and zebra.

Training We train our network with the Flickr2k [32] dataset having 2,000 high-quality images. In addition, we use data augmentation by randomly downsampling the images by the ratios of 2, 3, and 4 for the robustness to resolution dependency. We randomly extract a patch of size 128×128 from the input image during training. Adam optimizer [33] is used for the training, with a batch size of 24 for 3,000 epochs. The learning rate is initially set as 1×10^{-3} and decays by a factor of 0.1 for every 1,000 epochs. The training takes 72 hours on a GeForce GTX 1080 Ti.

Evaluation We compare our method with learned and conventional non-learned compression algorithms. Specifically, the following conventional lossless image codecs are compared: PNG [14], JPEG-LS [15], JPEG2000 [34], WebP [36], BPG [13], LCIC [12], FLIF [27], and JPEG-XL [16]. As for the learned methods, L3C [5], RC [6], Near-Lossless [7], and LC-FDNet [37] are considered. L3C and RC are trained with Open Images dataset [42] consisting of 300,000 images, and Near-Lossless is trained with the DIV2K dataset. For the evaluation metric, *bits per pixel* (bpp) is used.

B. Compression Result

Results for High Resolution Images We present the comparison for 2K images in Table I, which shows that our method outperforms the conventional non-learning codecs and existing practical-time learned codecs. Compared to JPEG-XL, which yields better performance than all the traditional codecs and even better than some learning-based methods, our method shows 4.7%∼5.5% gains. In comparison to learning-based methods, our method shows superior performance against others. Note that our method demonstrates improved coding efficiency compared to LC-FDNet by 0.3%∼1.3%. In addition, it should be noted that in the case of CLIC.m, non-learning codecs such as FLIF and JPEG-XL also outperform existing learning-based methods. It can be interpreted that learning-based methods are difficult to be generalized to CLIC.m. Nevertheless, our method presents a 4.7% gain compared to JPEG-XL.

Results for Low Resolution Images We first show the compression results on the downsampled dataset and illustrate that our method is robust to image resolution. Then, we present the compression results when applied to the actual low-resolution dataset.

In Table II, we show the compression performance for the downsampled datasets. Specifically, we downsample the evaluation dataset of CLIC.m, CLIC.p, DIV2K by the ratios of 2, 3, and 4. We set JPEG-XL as the basis and address the percentage with respect to JPEG-XL. Comparing JPEG-XL and LC-FDNet, the performance gap generally decreases as the downsampling ratio increases, demonstrating that the LC-FDNet suffers from the domain gap in image resolutions.

TABLE II: Compression results for downsampled datasets. $\downarrow s$ denotes a downsampling ratio by s. LC-FDNet* indicates that the network is additionally trained with downsampled images. The difference in percentage is with respect to JPEG-XL.

(a) CLIC.m result

Method	CLIC.m	CLIC.m ↓2	CLIC.m ↓3	CLIC.m ↓4
JPEG-XL	7.20	9.55	10.32	10.93
LC-FDNet	6.90 +4.2%	9.22 +3.4%	9.94 +3.6%	10.53 +3.7%
LC-FDNet*	6.91 +4.0%	9.15 +4.1%	9.80 +5.0%	10.37 +5.1%
Ours	6.88 +4.4%	9.01 +5.7%	9.68 +6.2%	10.22 +6.5%

(b) CLIC.p result

Method	CLIC.p	CLIC.p ↓2	CLIC.p ↓3	CLIC.p ↓4
JPEG-XL	8.19	10.30	11.13	11.69
LC-FDNet	7.86 +4.0%	9.91 +3.8%	10.70 +3.9%	11.20 +4.2%
LC-FDNet*	7.82 +4.5%	9.74 +5.5%	10.59 +4.8%	11.02 +5.7%
Ours	7.76 +5.3%	9.64 +6.4%	10.36 +6.9%	10.82 +7.4%

(c) DIV2K result

Method	DIV2K	DIV2K ↓2	DIV2K ↓3	DIV2K ↓4
JPEG-XL	8.49	10.71	11.72	12.39
LC-FDNet	8.14 +4.2%	10.32 +3.6%	11.30 +3.6%	11.96 +3.5%
LC-FDNet*	8.11 +4.5%	10.22 +4.6%	11.13 +5.0%	11.72 +5.4%
Ours	8.05 +5.2%	10.03 +6.3%	10.93 +6.7%	11.52 +7.0%

TABLE III: Compression result on low-resolution dataset.

Method	Kodak	McMaster	Classic
PNG	13.05 +51.2%	13.97 +30.2%	14.35 +35.0%
JPEG-LS	9.48 +9.8%	12.68 +18.2%	13.59 +27.8%
JPEG2000	9.57 +10.9%	12.50 +16.5%	11.93 +12.2%
WebP	9.62 +11.5%	11.86 +10.5%	11.37 +7.0%
LCIC	8.86 +2.7%	12.30 +14.6%	11.92 +12.1%
FLIF	8.72 +1.0%	11.59 +8.0%	11.10 +4.4%
JPEG-XL	8.93 +3.5%	11.35 +5.8%	11.79 +10.9%
L3C	10.64 +23.3%	12.24 +14.1%	12.42 +16.8%
Near-Lossless	9.12 +5.7%	11.90 +10.9%	12.18 +14.6%
LC-FDNet	8.76 +1.5%	10.86 +1.2%	10.85 +2.1%
Ours	8.63	10.73	10.63

Since LC-FDNet is trained only with 2K resolution images, the performance decreases when the training and evaluation domain are not matched. For a fair comparison, we additionally train LC-FDNet with a downsampled version of FLICKR2K and denote it as LC-FDNet*, which shows somewhat larger performance gaps in the downsampled datasets compared to the original datasets. In the case of our method, we added Resolution Refinement Block (RRB), which utilizes image resolution to generate resolution-specific kernels. As a result, our method yields better compression efficiency compared to LC-FDNet*, especially for downsampled datasets where performance gaps up to 2.1% are achieved. In summary, adding downsampled images as an additional training dataset and RRB both have valid contributions, making the network robust to image resolution.

Table III exhibits the compression results on the actual low resolution dataset. It can be observed that our method achieves state-of-the-art performance for all three datasets. Note that learning-based methods such as L3C and Near-Lossless exhibit a large performance drop compared to conventional methods such as FLIF and JPEG-XL. It can be implied that learning-based methods typically suffer from generalization problems. On the other hand, our method outperforms FLIF and JPEG-XL, demonstrating steady performance over the image resolution changes.

TABLE IV: Ablation study of using difference components in RRB.

Method	DIV2K	DIV2K↓2	Kodak
Fully Connected Layer	8.10	10.21	8.78
One-hot Encoding	8.07	10.10	8.68
Positional Embedding	8.05	10.03	8.63

C. Ablation Study

In this section, we provide a motivation for our design choice to utilize positional embedding in the Resolution Refinement Block (RRB). We conduct an ablation study by substituting the positional embedding with fully connected layers or one-hot encoding to evaluate the impact of each components. For the one-hot encoding approach, we convert the input image resolution HW into a one-hot vector, where each class represents a specific downsampling ratio applied to a 2K resolution. The classes correspond to resolutions of 2K (2048×1080), 1K (1024×768), and so on. We use a total of four classes in this experiment.

Table. IV presents the results obtained using difference components for RRB across various datasets, ranging from high to low resolutions. The result demonstrates that positional embedding yields the best performance, validating our design choice. On the other hand, employing fully connect layers leads to the poorest performance, almost comparable to not using RRB at all. Although the MLP approach allows more flexibility, it fails to capture the relationship between image resolution and intermediate features. The one-hot encoding approach shows improved performance compared to the MLPs, yet it still falls short of positional embedding. In conclusion, the ablation study confirms that incorporating positional embedding to convey image resolution information is the optimal choice for our proposed method.

IV. Conclusion

We have presented a new resolution-adaptive lossless image compression method by adding a Resolution Refinement Block to our previous frequency decomposition network [35]. The overall compression scheme operates at the subimage level, aiming to maximize the correlation between consecutive subimages. Our approach carefully arranges the order of subimages to ensure that the current subimage fully utilizes the horizontal and vertical components of the previous subimages. To compress each subimage, we employ a coarse-to-fine strategy by decomposing it into low and high-frequency regions. The compression process begins with the compression of low-frequency components, which serve as strong priors

for compressing the high-frequency regions. The frequency decomposition method is adaptive to the specific characteristics of each image, enabling the derivation of an optimal ratio of low and high-frequency components. The Resolution Refinement Blocks generate resolution-specific kernels, enabling the network to be adaptive to image resolution variations. Experimental results demonstrate that our method achieves state-of-the-art performance across various datasets, independent of image resolutions.

ACKNOWLEDGMENT

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01062), and in part by the Technology Innovation Program (ATC+ program, 20014131, 25nm X-ray Inspection System for Semiconductor Backend Process) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).

REFERENCES

[1] Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A. & Kavukcuoglu, K. Conditional image generation with pixelcnn decoders. *ArXiv Preprint ArXiv:1606.05328*. (2016)

[2] Salimans, T., Karpathy, A., Chen, X. & Kingma, D. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *ArXiv Preprint ArXiv:1701.05517*. (2017)

[3] Reed, S., Oord, A., Kalchbrenner, N., Colmenarejo, S., Wang, Z., Chen, Y., Belov, D. & Freitas, N. Parallel multiscale autoregressive density estimation. *International Conference On Machine Learning*. pp. 2912-2921 (2017)

[4] Van Oord, A., Kalchbrenner, N. & Kavukcuoglu, K. Pixel recurrent neural networks. *International Conference On Machine Learning*. pp. 1747-1756 (2016)

[5] Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R. & Gool, L. Practical full resolution learned lossless image compression. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 10629-10638 (2019)

[6] Mentzer, F., Gool, L. & Tschannen, M. Learning better lossless compression using lossy compression. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 6638-6647 (2020)

[7] Bai, Y., Liu, X., Zuo, W., Wang, Y. & Ji, X. Learning Scalable lY=-Constrained Near-Lossless Image Compression via Joint Lossy Image and Residual Compression. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 11946-11955 (2021)

[8] Rhee, H., Jang, Y., Kim, S. & Cho, N. Lossless Image Compression by Joint Prediction of Pixel and Context Using Duplex Neural Networks. *IEEE Access*. (2021)

[9] Hoogeboom, E., Peters, J., Berg, R. & Welling, M. Integer discrete flows and lossless compression. *ArXiv Preprint ArXiv:1905.07376*. (2019)

[10] Schiopu, I. & Munteanu, A. Deep-learning-based lossless image coding. *IEEE Transactions On Circuits And Systems For Video Technology*. **30**, 1829-1842 (2019)

[11] Cao, S., Wu, C. & Krähenbühl, P. Lossless image compression through super-resolution. *ArXiv Preprint ArXiv:2004.02872*. (2020)

[12] Kim, S. & Cho, N. Hierarchical prediction and context adaptive coding for lossless color image compression. *IEEE Transactions On Image Processing*. **23**, 445-449 (2013)

[13] "BPG image format, Download from," https://bellard.org/bpg.

[14] Boutell, T. & Lane, T. Png (portable network graphics) specification version 1.0. *Network Working Group*. pp. 1-102 (1997)

[15] Weinberger, M., Seroussi, G. & Sapiro, G. The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Transactions On Image Processing*. **9**, 1309-1324 (2000)

[16] Alakuijala, J., Asseldonk, R., Boukortt, S., Bruse, M., Comșa, I., Firsching, M., Fischbacher, T., Kliuchnikov, E., Gomez, S., Obryk, R. & Others JPEG XL next-generation image compression architecture and coding tools. *Applications Of Digital Image Processing XLII*. **11137** pp. 111370K (2019)

[17] Schwartz, E., Giryes, R. & Bronstein, A. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions On Image Processing*. **28**, 912-923 (2018)

[18] Ignatov, A., Van Gool, L. & Timofte, R. Replacing mobile camera isp with a single deep learning model. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops*. pp. 536-537 (2020)

[19] Kim, S. & Cho, N. Lossless compression of color filter array images by hierarchical prediction and context modeling. *IEEE Transactions On Circuits And Systems For Video Technology*. **24**, 1040-1046 (2014)

[20] Zhang, N. & Wu, X. Lossless compression of color mosaic images. *IEEE Transactions On Image Processing*. **15**, 1379-1388 (2006)

[21] Chung, K. & Chan, Y. A lossless compression scheme for Bayer color filter array images. *IEEE Transactions On Image Processing*. **17**, 134-144 (2008)

[22] Bazhyna, A. & Egiazarian, K. Lossless and near lossless compression of real color filter array data. *IEEE Transactions On Consumer Electronics*. **54**, 1492-1500 (2008)

[23] Hernandez-Cabronero, M., Marcellin, M., Blanes, I. & Serra-Sagrista, J. Lossless compression of color filter array mosaic images with visualization via JPEG 2000. *IEEE Transactions On Multimedia*. **20**, 257-270 (2017)

[24] Lee, Y. & Hirakawa, K. Shift-and-decorrelate lifting: CAMRA for lossless intra frame CFA video compression. *IEEE Signal Processing Letters*. **27** pp. 461-465 (2020)

[25] Richter, T., Fößel, S., Descampe, A. & Rouvroy, G. Bayer CFA pattern compression with JPEG XS. *IEEE Transactions On Image Processing*. **30** pp. 6557-6569 (2021)

[26] Pei, S. & Ding, J. Improved reversible integer-to-integer color transforms. *2009 16th IEEE International Conference On Image Processing (ICIP)*. pp. 473-476 (2009)

[27] Sneyers, J. & Wuille, P. FLIF: Free lossless image format based on MANIAC compression. *2016 IEEE International Conference On Image Processing (ICIP)*. pp. 66-70 (2016)

[28] "Workshop and Challenge on Learned Image Compression," https://www.compression.cc/challenge/.

[29] Agustsson, E. & Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops*. pp. 126-135 (2017)

[30] Abdelhamed, A., Lin, S. & Brown, M. A high-quality denoising dataset for smartphone cameras. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1692-1700 (2018)

[31] Bychkovsky, V., Paris, S., Chan, E. & Durand, F. Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs. *The Twenty-Fourth IEEE Conference On Computer Vision And Pattern Recognition*. (2011)

[32] Lim, B., Son, S., Kim, H., Nah, S. & Mu Lee, K. Enhanced deep residual networks for single image super-resolution. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops*. pp. 136-144 (2017)

[33] Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*. (2014)

[34] Rabbani, M. JPEG2000: Image compression fundamentals, standards and practice. *Journal Of Electronic Imaging*. **11**, 286 (2002)

[35] Rhee, H., Jang, Y., Kim, S. & Cho, N. LC-FDNet: Learned Lossless Image Compression with Frequency Decomposition Network. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 6033-6042 (2022)

[36] Version, W. 1.0 Hardware Reference Guide, XP-002202892, NetWork Engines. *Inc., Jun*. **1** pp. 92 (2000)

[37] Rhee, H., Jang, Y., Kim, S. & Cho, N. LC-FDNet: Learned Lossless Image Compression With Frequency Decomposition Network. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 6033-6042 (2022,6)

[38] Bengio, Y., Léonard, N. & Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv Preprint ArXiv:1308.3432*. (2013)

[39] Cheng, Z., Sun, H., Takeuchi, M. & Katto, J. Learning image and video compression through spatial-temporal energy compaction. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 10071-10080 (2019)

[40] Zhang, L., Wu, X., Buades, A. & Li, X. Color demosaicking by local

directional interpolation and nonlocal adaptive thresholding. *Journal Of Electronic Imaging*. **20**, 023016 (2011)
[41] "Kodak Lossless True Color Image Suite, Download from," http://r0k.us/graphics/kodak/.
[42] Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A. & Others Open-images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset Available From Https://github. Com/openimages*. **2**, 18 (2017)
[43] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)
[44] Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R. & Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications Of The ACM*. **65**, 99-106 (2021)

Implementation and Analysis on Backpropagating Refinement Scheme for Interactive Image Segmentation

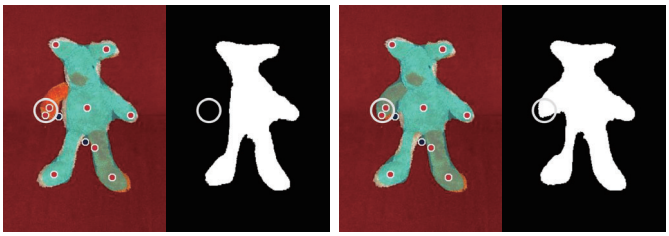
Chaewon Lee*, Won-Dong Jang†, and Chang-Su Kim*
* School of Electrical Engineering, Korea University, Seoul, Korea
†Amazon, Los Angeles, California
E-mail: chaewonlee@mcl.korea.ac.kr, dotol1216@gmail.com, changsukim@korea.ac.kr

***Abstract*—BRS is the first CNN-based interactive image segmentation algorithm to refine segmentation results based on a backpropagation scheme. In this paper, we give a detailed description of how BRS operates and demonstrate how to implement the algorithm. In BRS, user-provided clicks are first converted into interaction maps, which are then concatenated with the RGB image and provided as input to a segmentation network. In the test phase, performing the forward pass in the network generates an initial segmentation map. However, the user-annotated pixels may be mislabeled in the initial result. BRS refines this result by correcting the mislabeled pixels. We implement this BRS algorithm in PyTorch and publish the source codes. Moreover, we first show that BRS can reach the perfect IoU ratio of 1.0 in most cases and delineate objects more accurately than a variant of BRS, called f-BRS.**

I. Introduction

Interactive image segmentation is a task that enables users to segment objects of interest from the background. In this task, a target object is annotated interactively by a user in forms of either scribbles [1], [2], boxes [3]–[5], or clicks [6]–[8]. The user provides annotations until a desirable segmentation result is obtained. Interactive image segmentation allows users to select objects and delineate them easily, so it is useful in many applications such as image editing and medical image analysis. Recently, with the development of deep-learning-based algorithms for dense prediction tasks, demands for annotated data also have increased significantly. However, pixel-level annotations are costly due to their laborious and time-consuming traits. This labeling cost could be reduced by employing interactive image segmentation techniques.

Research on interactive image segmentation has been done since decades ago, and various approaches have been proposed to solve the problem. Recently, deep-learning-based methods [6]–[9] have begun to emerge for interactive segmentation. Xu *et al.* [6] first introduced deep learning into interactive segmentation. Adopting clicks as user annotations, they converted foreground and background clicks into interaction maps using the distance transform. They trained a CNN model, which takes the concatenation of the interaction maps and the RGB image as input. Since then, the idea of taking interaction maps as network input has become a conventional approach [7]–[9]. Although these interaction maps clearly represent the annotated labels in user-clicked locations, the output prediction may be inconsistent with the user-provided labels. In other words, the segmentation result of the CNN may have incorrect labels in user-clicked locations.



(a) Before BRS (b) After BRS

Fig. 1. (a) Before applying BRS, user-clicked pixels have incorrect labels in the segmentation result. Red and blue circles depict foreground and background clicks, respectively. (b) The mislabelled pixels within the gray ring are corrected after BRS.

Backpropagating refinement scheme (BRS) [7] is the first attempt to employ a backpropagating refinement strategy to address this issue. It enforces the network to yield correct labels in user-clicked locations, by modifying the input interaction maps, as illustrated in Fig. 1. Inspired by this research, subsequent methods [9], [10] have been developed; Sofiiuk *et al.* [9] and Lin *et al.* [10], respectively, proposed modifying intermediate feature maps of the segmentation networks, instead of input interaction maps.

In this paper, we focus on Jang *et al.*'s original BRS algorithm [7] and describe how to implement it. In [7], the codes were written in `Caffe`, which is now obsolete. Hence, we implement the algorithm in `PyTorch` and make the source codes publicly available at https://github.com/cwlee00/BRS-PyTorch. Also, we describe the different variations [7], [9], [10] of BRS comparatively using a unified set of notations. Finally, through experiments, we first show that the BRS algorithm can be used effectively for image segmentation, enabling maximum IoU ratios to even reach the perfect score of 1.0 in most cases.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

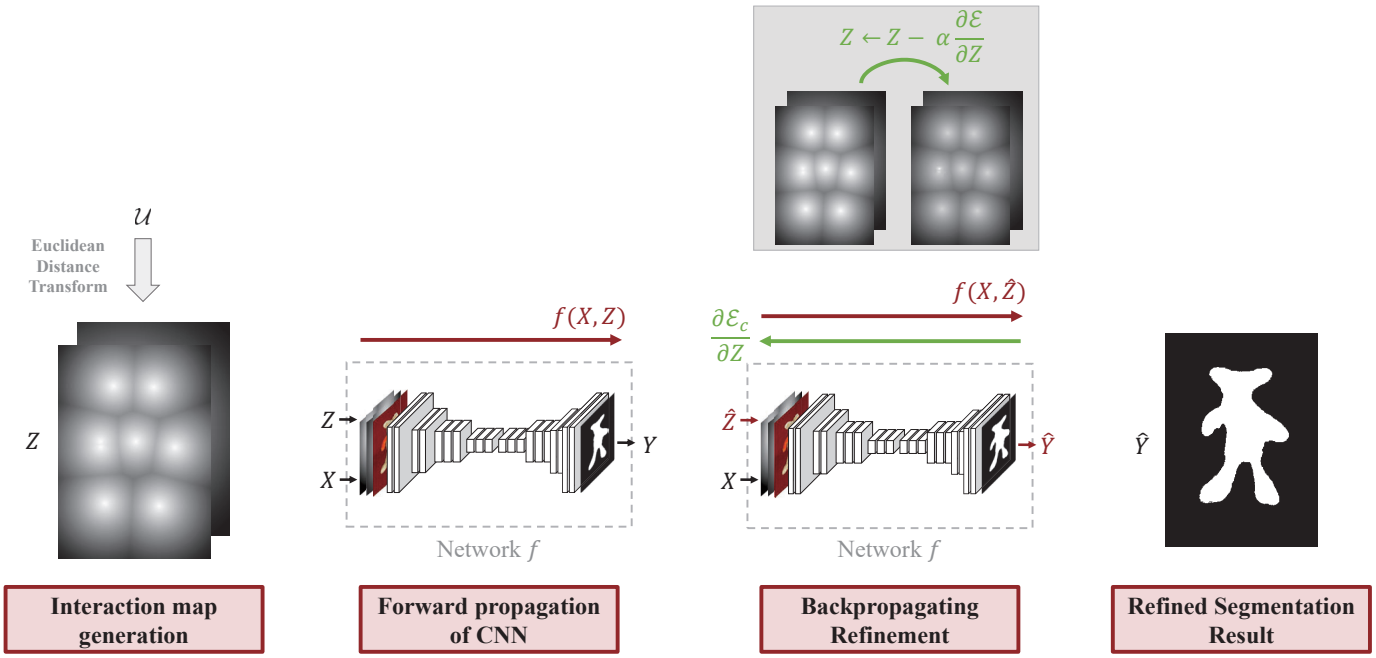


Fig. 2. Overview of the BRS algorithm.

II. Algorithms

A. BRS: Backpropgating Refinement Scheme

Fig. 2 shows an overview of the original BRS algorithm [7]. BRS allows a user to segment an object out by successively placing foreground and background clicks. Let $\mathcal{U}$ denote the set of user-provided clicks. Given $\mathcal{U}$, we first generate foreground and background interaction maps using the distance transform as in [6]. Each interaction map has the same height and width as the input image $X \in \mathbb{R}^{3\times H\times W}$. We stack the two maps into $Z \in \mathbb{R}^{2\times H\times W}$. We then feed the concatenation of X and Z into a CNN. Thresholding the output probability map of the CNN, we produce an initial segmentation map Y. Although the forward pass of the network usually provides a decent segmentation quality, it is still possible that the clicked pixels have incorrect labels as illustrated in Fig. 1. BRS aims to achieve more accurate segmentation by correcting these mislabeled pixels. More specifically, rather than fine-tuning the entire network to yield better segmentation results, BRS adjusts the interaction maps so that the output predictions are consistent with the input clicks.

Let us define notations for describing the BRS algorithm. The entire network can be represented by the composite function

$$f = f^R \circ f^{R-1} \circ \cdots \circ f^2 \circ f^1 \tag{1}$$

where f^r is the rth layer in the network, $r = 1, 2, \ldots, R$. Let y^r denote the response of layer r, and let θ^r be the set of parameters in layer r. Then, y^r can be written as

$$y^r = f^r(y^{r-1}; \theta^r). \tag{2}$$

Hence, y^0 and y^R are the input and output of the network. Since the input is the concatenation of X and Z, we have

$$y^0 = X \oplus Z. \tag{3}$$

The goal of BRS is to assign correct labels at user-annotated pixels by optimizing the input interaction maps. The optimization is done to minimize the energy function, given by

$$\mathcal{E}(Z) = \mathcal{E}_{\mathrm{C}}(Z) + \lambda \mathcal{E}_{\mathrm{I}}(Z) \tag{4}$$

where $\mathcal{E}_{\mathrm{C}}(Z)$ is the corrective energy, $\mathcal{E}_{\mathrm{I}}(Z)$ is the inertial energy, and λ is a weight. First, the corrective energy enforces the network output to yield correct labels in user-annotated locations, which is defined as

$$\mathcal{E}_{\mathrm{C}}(Z) = \sum_{\mathbf{u}\in\mathcal{U}} \left(l(\mathbf{u}) - y^R(\mathbf{u})\right)^2 \tag{5}$$

where $l(\mathbf{u})$ denotes a user-annotated binary label: 1 for foreground and 0 for background. Second, to prevent excessive perturbations of the interaction maps, we also minimize the inertial energy

$$\mathcal{E}_{\mathrm{I}}(Z) = \sum_{\mathbf{x}\in\mathcal{N}} \left(Z(\mathbf{x}) - Z_{\mathrm{init}}(\mathbf{x})\right)^2 \tag{6}$$

where Z_{init} is the initial interaction map before the modification, and $\mathcal{N}$ is the set of coordinates in the interaction maps.

As shown in Fig. 2, to perform BRS, we should compute the gradient of the energy function,

$$\frac{\partial \mathcal{E}}{\partial Z} = \frac{\partial \mathcal{E}_{\mathrm{C}}}{\partial Z} + \lambda \frac{\partial \mathcal{E}_{\mathrm{I}}}{\partial Z}. \tag{7}$$

The gradient $\frac{\partial \mathcal{E}_{\text{C}}}{\partial Z}$ of the corrective energy can be obtained via backpropagation. First, we compute the gradient of $\mathcal{E}_{\text{C}}$ with respect to the output segmentation map y^R,

$$\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^R} = -2 \times \sum_{\mathbf{u} \in \mathcal{U}} \big(l(\mathbf{u}) - y^R(\mathbf{u})\big). \tag{8}$$

Then, by differentiating y^r in (2) with respect to the input y^{r-1}, we compose the derivative matrix $\frac{\partial y^r}{\partial y^{r-1}}$ or the transposed Jacobian matrix $(J_y^r)^T$ at layer r by

$$\frac{\partial y^r}{\partial y^{r-1}} = (J_y^r)^T = \begin{bmatrix} \frac{\partial y_1^r}{\partial y_1^{r-1}} & \cdots & \frac{\partial y_{k_r}^r}{\partial y_1^{r-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1^r}{\partial y_{k_{r-1}}^{r-1}} & \cdots & \frac{\partial y_{k_r}^r}{\partial y_{k_{r-1}}^{r-1}} \end{bmatrix} \tag{9}$$

where k_r is the number of activations at layer r. Using the Jacobian matrix, we perform backpropagation by

$$\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^r} = (J_y^{r+1})^T \frac{\partial \mathcal{E}_{\text{C}}}{\partial y^{r+1}}. \tag{10}$$

We begin the backpropagation by computing the gradient $\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^R}$ in (8) at the last layer of the network. Then, we compute the gradient $\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^r}$ in (10) recursively from $r = R-1$ to 0. As a result, we obtain $\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^0}$. Finally, as y^0 is the concatenation of X and Z in (3), we can easily obtain $\frac{\partial \mathcal{E}_{\text{C}}}{\partial Z}$ from $\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^0}$.

Also, we compute the gradient of the inertial energy by

$$\frac{\partial \mathcal{E}_{\text{I}}}{\partial Z} = 2 \times \sum_{\mathbf{x} \in \mathcal{N}} (Z(\mathbf{x}) - Z_{\text{init}}(\mathbf{x})). \tag{11}$$

Then, via the gradient descent

$$Z \leftarrow Z - \alpha \frac{\partial \mathcal{E}}{\partial Z}, \tag{12}$$

we can optimize the interaction maps until mislabeled pixels are corrected. We use the L-BFGS algorithm [11] to conduct the gradient descent. Fig. 2 includes an example of how BRS modifies the initial interaction map and improves the segmentation results.

B. f-BRS: Feature Backpropgating Refinement Scheme

With the use of backpropagation, BRS shows great improvement in segmentation results. However, Sofiiuk *et al.* [9] points out its drawback in terms of computational speed. Since BRS should optimize input interaction maps, it needs to run both foreground and background passes through the entire network. To alleviate this problem, they propose optimizing intermediate features in the network instead of interaction maps. Their f-BRS algorithm results in a faster speed, for it computes gradients through only a small part of the network.

They introduce a set of auxiliary parameters P to modify an intermediate feature layer y^m of the network. Let G be the modifying function for y^m. Then, the modified y^m can be written as $G(y^m, P)$. Note that the network output y^R is a function of y^m. Thus, after the modification, y^R is given by

$$y^R = f^R \circ f^{R-1} \circ \cdots \circ f^{m+1}(G(y^m, P); \theta^{m+1}). \tag{13}$$

Then, instead of (6), the inertial energy is reformulated as

$$\mathcal{E}_{\text{I}}(P) = \sum_{p \in P} \|p - p_{\text{init}}\|^2. \tag{14}$$

For auxiliary parameters P, Sofiiuk *et al.* [9] suggest using scale and bias terms. Suppose $y^m \in \mathbb{R}^{c_m \times h_m \times w_m}$, where c_m, h_m, and w_m are the number of channels, height, and width of the intermediate feature map y^m. Also, let $s \in \mathbb{R}^{c_m}$ and $b \in \mathbb{R}^{c_m}$ denote the scale and bias vectors. Thus, $P = \{s, b\}$. In f-BRS, the intermediate feature map y^m is modified by

$$G_{\text{f-BRS}}(y^m, P) = s \dot{\times} y^m + b \dot{\times} o \tag{15}$$

where $\dot{\times}$ denotes channel-wise multiplication, and o is the constant map whose all element is 1. Initially, they set $s = \mathbf{1}$ and $b = \mathbf{0}$ to ensure that there is no modification, *i.e.* $G_{\text{f-BRS}}(y^m, P) = y^m$. By controlling the scale and bias vectors, they attempt to obtain segmentation results that better align with user clicks.

To optimize each parameter $p \in P$, the following gradient should be computed:

$$\frac{\partial \mathcal{E}}{\partial p} = \frac{\partial \mathcal{E}_{\text{C}}}{\partial p} + \lambda \frac{\partial \mathcal{E}_{\text{I}}}{\partial p} \tag{16}$$

$$= \frac{\partial y^m}{\partial p} \times \frac{\partial \mathcal{E}_{\text{C}}}{\partial y^m} + \lambda \frac{\partial \mathcal{E}_{\text{I}}}{\partial p}. \tag{17}$$

Note that $\frac{\partial \mathcal{E}_{\text{C}}}{\partial y^m}$ in (17) can be computed by employing the backward recursion in (10). Compared to the original BRS, we see that the backward operation needs to be performed only to the intermediate layer y^m, rather than to the input layer y^0, which results in faster computation.

C. g-BRS: Generalized Backpropgating Refinement Scheme

The auxiliary vectors s and b in (15) in f-BRS are spatially invariant, *i.e.* oblivious to the position in an image. Therefore, adjusting these parameters tends to cause only global changes in the output and may not enable local refinement. To address this issue, Lin *et al.* [10] propose auxiliary parameters for both global and local refinement. They introduce three re-parameterization methods.

First, they introduce the method of adding an additional bias map $b_m \in \mathbb{R}^{h_m \times w_m}$ before applying channel-wise scale and bias. They channel-wise scale b_m using $w_c \in \mathbb{R}^{c_m}$, which allows each channel of b_m to shift differently. In this approach, the intermediate feature map is modified to

$$G_{\text{g-BRS1}}(y^m, P) = s \dot{\times} (y^m + w_c \dot{\times} b_m) + b \dot{\times} o. \tag{18}$$

Thus, the set of auxiliary parameters is $P = \{s, b, w_c, b_m\}$.

Second, they also suggest modifying y^m as follows:

$$g_1(y^m) = s \dot{\times} y^m + b \dot{\times} o, \tag{19}$$

$$g_2(y^m) = y^m + w_c \dot{\times} b_m, \tag{20}$$

$$G_{\text{g-BRS2}}(y^m, P) = w \cdot g_1(y^m) + (1 - w) \cdot g_2(y^m), \tag{21}$$

where $P = \{s, b, w_c, b_m, w\}$. Since g_1 in (19) enables global modification and g_2 in (20) allows local modification, fusing these two with a learnable weight $w \in [0, 1]$ regulates the trade-off between global and localized changes.

Third, they introduce another re-parameterization method using a 1×1 convolutional layer,

$$G_{\text{g-BRS3}}(y^m, P) = \phi(y^m + 10w_c \dot{\times} b_m) + b \dot{\times} o \tag{22}$$

where $P = \{b, b_m, w_c, \phi\}$, and ϕ denotes the 1×1 convolution with $c_m \times c_m$ parameters. Note that in `PyTorch` the second term $b \dot{\times} o$ in (22) can be implemented as biases in the 1×1 convolution layer ϕ.

For all three re-parameterization methods, the auxiliary parameters in P are optimized via gradient descent using the gradient computation in (16).

III. Implementation Details

Let us describe in detail how to implement the original BRS algorithm, which is summarized in **Algorithm 1**. First, in line 1, we transform the given set $\mathcal{U}$ of user-provided clicks into interaction maps Z using the Euclidean distance transform (EDT). In other words, interaction maps are obtained by computing the distance of each pixel to the nearest foreground and background clicks. The maximum distance in Z is set to 255 for efficient data storage. Then, in line 2, we feed the input image X and the interaction maps Z into the network f to get a probability map of the target object. We obtain the segmentation result Y by thresholding this probability map. The threshold value is 0.5, so each pixel is determined as foreground if its probability is higher than 0.5, and background otherwise.

Since the goal of BRS is to correct mislabeled pixels, BRS is activated when any clicked pixel has a wrong label. Line 3 shows the condition for activating BRS. Y_{u_i,v_i} is the output label at the ith click location (u_i, v_i), and l_i is the ith click label. Thus, if $l_i = 1$ and $Y_{u_i,v_i} = 0$, or if $l_i = 0$ and $Y_{u_i,v_i} = 1$ for any i, we refine Z by performing BRS.

Lines 4∼8 in **Algorithm 1** show the process of refining Z using backpropagation. We first formulate the energy function to be minimized in lines 3∼6. Then, we optimize the interaction maps according to the gradient of the energy function. Therefore, we should compute $\frac{\partial \mathcal{E}}{\partial Z}$ through backpropagation as described in Section II-A. To compute this gradient, we use the built-in differentiation engine in `Pytorch`. Then, we apply the L-BFGS algorithm [11] to optimize the interaction maps Z using gradient descent. To this end, we use the `fmin_lbfgs_b` function in the `scipy` library. Once the refined interaction map $\hat{Z}$ is obtained, we execute the forward pass again. Finally, we threshold the probability map yielded by the network to get a refined segmentation result $\hat{Y}$.

Algorithm 1 Procedure of proposed algorithm

Input: Input image X, user clicks $\mathcal{U}$

1: $Z \leftarrow \text{EDT}(\mathcal{U})$
2: $Y \leftarrow \text{Threshold}(f(X, Z))$
3: **if** $l_i \neq Y_{u_i,v_i}$ for any i **then**
4: $\quad \mathcal{E}_C \leftarrow \sum_i (l_i - f(X, Z)_{u_i,v_i})^2$ ▷ (5)
5: $\quad \mathcal{E}_I \leftarrow \sum_x (Z(x) - Z_{\text{init}}(x))^2$ ▷ (6)
6: $\quad \mathcal{E} \leftarrow \mathcal{E}_C + \lambda \mathcal{E}_I$ ▷ (4)
7: $\quad$ Compute $\frac{\partial \mathcal{E}}{\partial Z}$ ▷ (7)∼(11)
8: $\quad \hat{Z} \leftarrow \text{L-BFGS}(Z, \mathcal{E}, \frac{\partial \mathcal{E}}{\partial Z})$
9: $\quad \hat{Y} \leftarrow \text{Threshold}(f(X, Z))$
10: **end if**

Output: Refined segmentation result $\hat{Y}$

IV. Experimental Results

A. Training and Test Details

Training: We adopt the network architecture in [7] but take VGG-16 [12] as the backbone. We also do not make use of the fine decoder. For training, we use the SBD dataset [13]. We crop 400 x 400 patches randomly around each object instance and apply horizontal flips for data augmentation. To simulate user interactions, annotations are automatically generated as follows. Pixels in a ground-truth mask are set as foreground candidates, while those at least 5 and at most 40 pixels away from the mask boundary are set as background candidates. We randomly sample foreground and background clicks from these candidates. The numbers of foreground and background clicks are determined randomly within [1, 10] and [0, 10], respectively. We employ the cross-entropy loss between a ground-truth mask and its predicted probability map. We use the Adam optimizer with a learning rate of 10^{-4} and train the network for 20 epochs.

Evaluation: For evaluation, we use the GrabCut [5] and Berkeley [14] datasets. In interactive segmentation, the segmentation result varies depending on how a user places clicks. Thus, for reliable assessment, we use an automatic clicking strategy as done in [6], [8]. This clicking strategy is designed based on the assumption that a typical user would provide clicks near the centers of big erroneous regions.

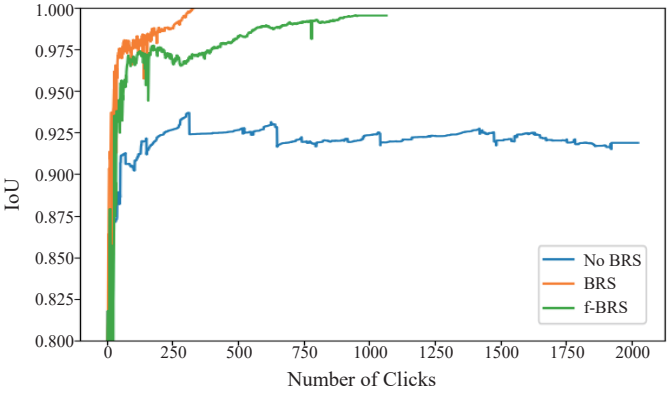


Fig. 3. IoU ratios according to the number of clicks for a sample image from the GrabCut dataset.

TABLE I
EVALUATION RESULTS ON THE GRABCUT AND BERKELEY DATASETS USING THE METRICS OF THE AVERAGE MAX IOU, NOF@99, NOF@100, AND THE AVERAGE RUNTIME TO PROCESS AN IMAGE.

Algorithm	GrabCut				Berkeley			
	Max IoU (↑)	NoF@99 (↓)	NoF@100 (↓)	Time (s)	Max IoU (↑)	NoF@99 (↓)	NoF@100 (↓)	Time (s)
No BRS	0.9604	44	50	**168.49**	0.9100	100	100	**139.90**
BRS	**0.9999**	**0**	**2**	366.42	**0.9998**	**0**	**63**	1340.84
f-BRS	0.9962	7	39	209.94	0.9807	68	99	388.07

B. Performance Assessment

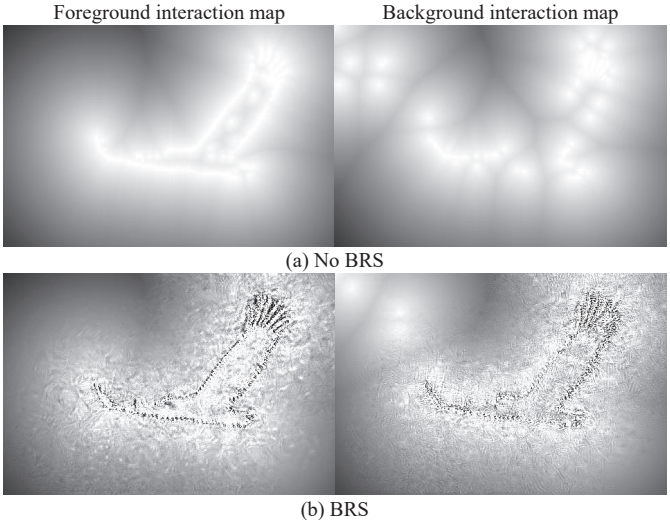
Experimental Conditions: Existing methods for interactive image segmentation usually limit the maximum number of clicks to 20. In this work, we remove this constraint and explore whether increasing the number of clicks could improve segmentation results. Thus, for all the experiments below, we set the maximum number of clicks to be the total number of pixels composing the target instance. As exemplified in Fig. 3, when only the forward pass is executed, increasing the number of clicks does not necessarily lead to more accurate results. However, with the use of BRS, additional clicks enable the maximum IoU ratio to reach the perfect 1.0. As the upper bound for the number of clicks is very large, we set stopping conditions for clicks: When the algorithm reaches an IoU ratio of 1.0, we stop at that click. Also, when the IoU ratios are all the same for recent 100 clicks, we assume convergence and stop.

Performance Comparison: We assess different variations of BRS comparatively in terms of the average of the maximum IoU ratios achieved. We also report NoF@IoU indicating the number of images for which the algorithm fails to reach a certain IoU. We do not include results of g-BRS [10], as there are no source codes available for the algorithm. Thus, we compare the original BRS algorithm [7], f-BRS [9], with the case when no backpropagation is applied (*i.e.* when only the forward pass is executed). Table I compares the results.

We see that in terms of accuracy, BRS outperforms f-BRS. Applying BRS results in higher maximum IoU scores and fewer failure cases to reach a certain IoU. Also note that when no backpropagating refinement scheme is applied, the model cannot yield desired segmentation results. Qualitative comparisons are given in Fig. 6. In Table I, we also measure the average runtime to process an image. As compared with BRS, f-BRS is about 2∼4 times faster, thus f-BRS shows greater performance in terms of speed. However, it is inferior to the original BRS in terms of accuracy.

C. Analysis and More Results

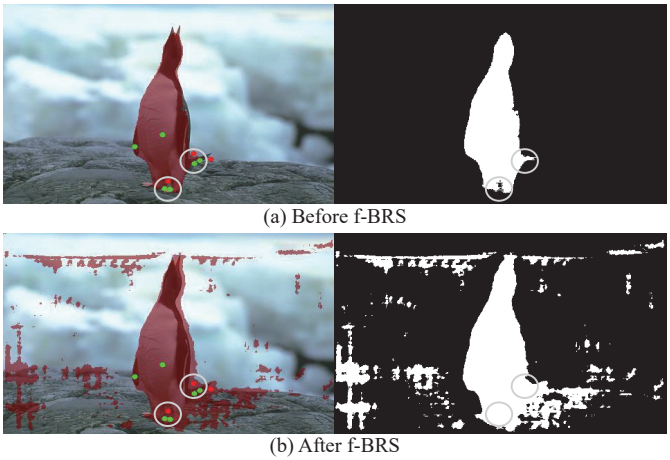
Fig. 4 visualizes examples of interaction maps. We see that the interaction maps using BRS contain much more information on elaborate details of the object. It is hard to find the details of the bird's wings in the interaction maps without BRS. This is one of the reasons why BRS yields a better segmentation mask than the algorithm that executes the forward pass only.



(a) No BRS

(b) BRS

Fig. 4. Visualization of interaction maps.


(a) Before f-BRS

(b) After f-BRS

Fig. 5. A failure case of f-BRS. Gray rings emphasize the regions where clicked pixels are corrected after f-BRS.

Also, we analyze why f-BRS is less accurate than BRS in some cases. We observed that f-BRS sometimes provides undesirable errors as in Fig. 5. In this example, although the labels of clicked pixels are corrected after f-BRS, many unwanted artifacts are generated in the background. This is because f-BRS optimizes intermediate feature maps of

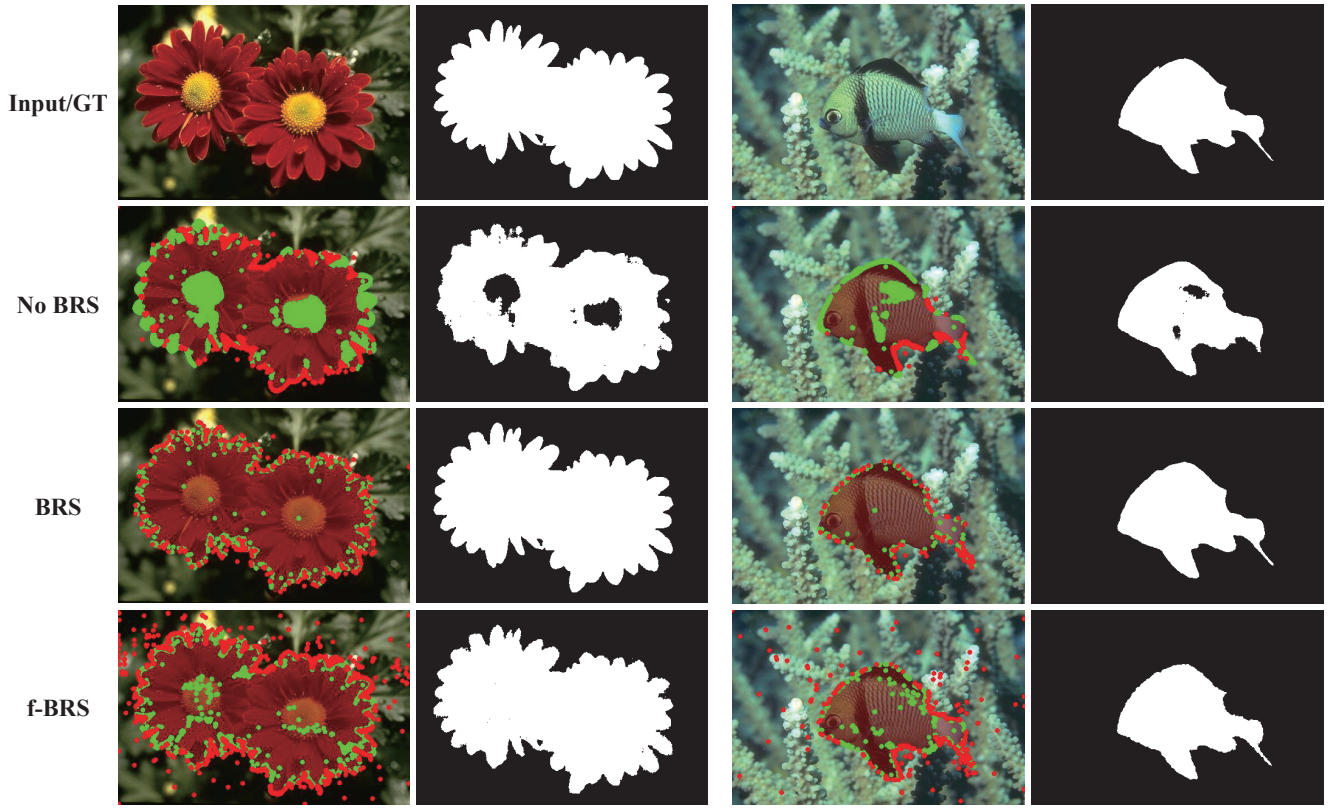


Fig. 6. Qualitative comparison of segmentation results. The left column shows an input image with clicks, while the right column shows the corresponding segmentation result. Foreground clicks are depicted by green circles, and background clicks by red circles.

the network via channel-wise operations. The channel-wise perturbations modify segmentation results globally. Consequently, along with the correction of clicked pixels, there are many changes in regions far from the clicks leading to errors. Therefore, f-BRS requires much more clicks and effort to correct such unwanted results. This also explains why f-BRS cannot reach the perfect IoU ratio of 1.0 in many cases in Table I.

V. Conclusions

In this paper, we provided a detailed description of the BRS algorithm and demonstrated how to implement the algorithm. In BRS, interaction maps are optimized to enforce output predictions to be consistent with user annotations. We implemented this BRS algorithm in PyTorch and made the source codes publicly available. Also, we described different variants of BRS comparatively using a unified set of notations, and showed that the original BRS algorithm can be used to segment out objects more accurately than f-BRS does.

VI. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2021R1A4A1031864 and No. NRF-2022R1A2B5B03002310)

References

[1] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, 2006.
[2] J. Bai and X. Wu, "Error-tolerant scribbles based interactive image segmentation," in *CVPR*, 2014, pp. 392–399.
[3] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu, "MILCut: a sweeping line multiple instance learning paradigm for interactive image segmentation," in *CVPR*, 2014, pp. 256–263.
[4] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *ICCV*, 2009, pp. 277–284.
[5] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut" — Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graphics*, vol. 23, no. 3, pp. 309–314, 2004.
[6] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, "Deep interactive object selection," in *CVPR*, 2016, pp. 373–381.
[7] W.-D. Jang and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme," in *CVPR*, 2019, pp. 5297–5306.
[8] Z. Li, Q. Chen, and V. Koltun, "Interactive image segmentation with latent diversity," in *CVPR*, 2018, pp. 577–585.
[9] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "f-BRS: Rethinking backpropagating refinement for interactive segmentation," in *CVPR*, 2020, pp. 8623–8632.
[10] F. Lin, B. Price, and T. Martinez, "Generalizing interactive backpropagating refinement for dense prediction networks," in *CVPR*, 2022, pp. 773–782.
[11] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[13] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011, pp. 991–998.

[14] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, vol. 2, 2001, pp. 416–423.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Implicit Neural Representation for Video Coding Through Progressive Feature Extraction

Jihoo Lee* and Je-Won Kang*
* Department of Electronic and Electrical Engineering, Ewha W. University, Seoul, Korea
Graduate Program in Smart Factory, Ewha W. University, Seoul, Korea
E-mail: jihoo99@ewhain.net; jewonk@ewha.ac.kr

***Abstract*—Implicit neural representation for video (INRV) utilizes a neural network to represent a video without relying on an explicit pixel-wise representation of RGB values, and, therefore an INRV-based video coding method aims to compress a neural network and transmit its parameters through a bit-stream. In this paper, we propose a novel rate-distortion (R-D) optimized INRV-based video coding method through a progressive feature extraction module (PFEM). The PFEM consists of a series of residual blocks (RBs) that can progressively control the quality of the reconstruction, by searching the optimal network architecture. By increasing the number of RBs, the quality of a reconstruction frame is improved. However, the quality comes at the expense of overheads in transmission. In the proposed method, the number of the RBs are decided to provide the best trade-off between the frame quality and bits. Our decision model is based upon the R-D theory. Experimental results demonstrated that the proposed method provides a superior coding gain to the conventional video coding standards.**

I. Introduction

As the volume of video data rapidly increases, extensive studies on video compression are conducted for efficient storage and transmission of visual data. Although conventional video codecs such as H.264/AVC [1] and H.265/HEVC [2] have been widely used in mobile devices and streaming services, the handcrafted features in the coding methods would limit compression efficiency. There have been many attempts to improve compression efficiency, by replacing conventional compression pipelines with neural networks [3]–[9]. However, these methods used large-scaled neural networks, incurring significant computational complexity and challenges in training.

Implicit neural representation for video (INRV) [10]–[13] has gained significant attention for video coding to address the problems of the conventional approaches. INRV utilizes a neural network to represent a video without relying on an explicit pixel-wise representation of RGB values. Therefore, the INRV-based video coding approach compresses a neural network and its parameters and transmits them through a bit-stream. In an encoder side, a network is configured to fit video data with spatial and temporal positional embedding [10]. In a decoder side, a network is reconstructed to produce an RGB value with the spatial and temporal indices of a frame as input.

Early INRV studies tried to embed all the (x, y, t) positions of a video sequence [14]–[16]. However, the model complexity has been substantially increased. In recent, NeRV [10] proposed an image-wise neural representation model, in which a frame was reconstructed using only the temporal index t instead of using all the temporal and spatial pixel indexes as input. The NeRV decoder was designed to reconstruct a video frame from the temporal embedding, and, thus it substantially reduced computational complexity. More recently, CNeRV [11] and HNeRV [12] used content-adaptive embedding to improve the quality of a reconstructed video, while keeping the same complexity. The decoder used a series of convolution layers to keep the high fidelity of video frames.

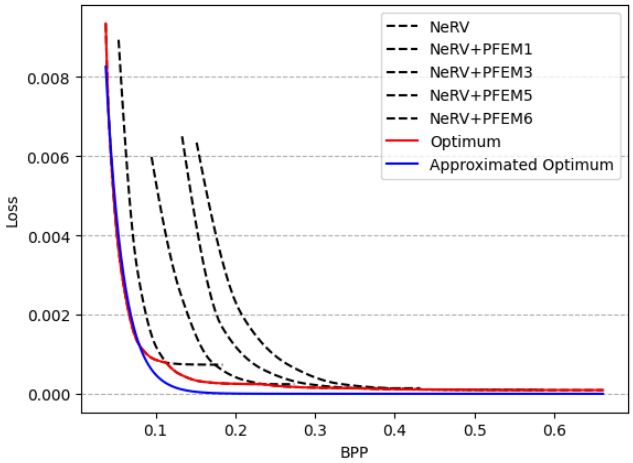


Fig. 1. Dotted graphs represent the R-D performance according to the number of RBs in the PFEM. The red graph represents the optimal R-D curve to achieve the best trade-off between frame quality and bits.

In this paper, we propose a rate-distortion (R-D) optimized INRV-based video coding method through a progressive feature extraction module (PFEM). The PFEM consists of a series of residual blocks (RBs) that can be adaptively configured to provide the best R-D performance. Fig. 1 displays the intuition of the proposed method. By increasing the number of RBs, the quality of a reconstruction frame is improved due to increasing representation capability of the INRV. However, the capability comes at the expense of overheads in transmission. In the proposed method, the number of the RBs are decided to provide the best trade-off between the frame quality and bits used to transmit a network architecture. Our decision model is based upon the R-D theory [17] and improves coding performance over the conventional video coding standards.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

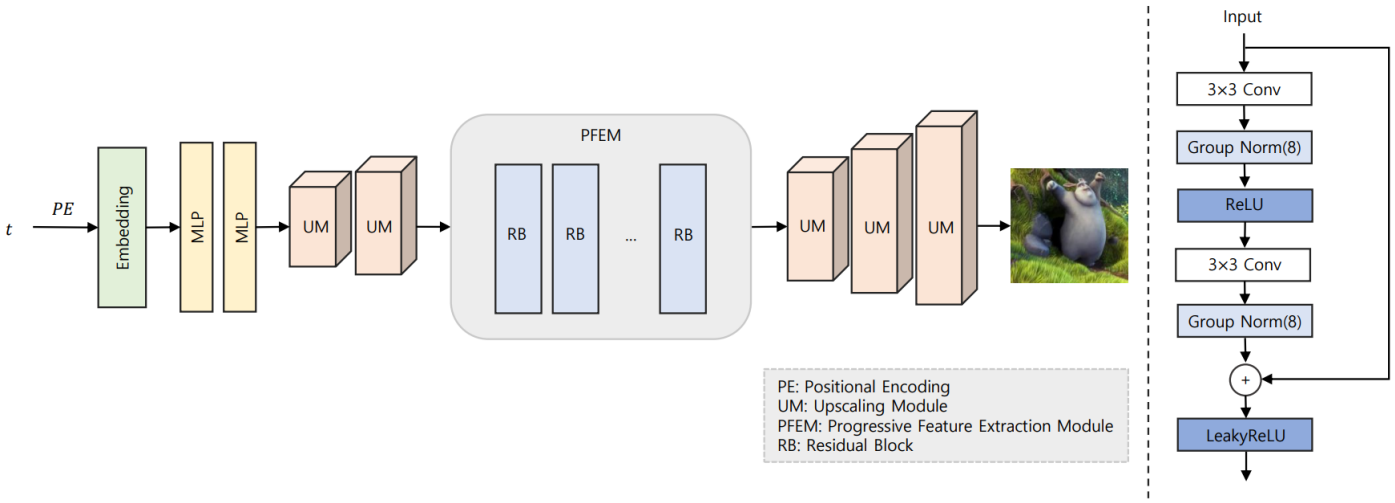


Fig. 2. INRV-based video coding model, including the PFEM (left). The PFEM consists of several RBs to control the trade-off between the size of the architecture and its reconstruction capability. The architectural details of the PFEM is presented (right).

II. Related Works

A. Implicit Neural Representation for Video Compression

The INRVs have been demonstrated as powerful yet compact parameterized representations for 2D and 3D video data, in which a neural function maps spatial and temporal coordinates to density or RGB values [18]–[24]. The INRVs are actively used for various image processing and computer vision tasks, including video compression [10]–[13], 3D novel view synthesis [18], [19], video generation [25], and video pre/post-processing techniques [20]–[24] such as video interpolation and super-resolution.

Most of the INRV methods used three steps in their pipelines, including an input embedding, video overfit, and model compression. The NeRV applied a positional encoding (PE) with high-frequency variations as an embedding function to memorize the spatial coordinates in high-dimensional embedding space and avoid poor representation [10]. The network architecture utilized several multi-layer perceptrons (MLPs) and convolution layers to fit a target video content. MLPs consumed a large number of network parameters to directly output pixel values. Thus, they proposed a NeRV block with convolutions and PixelShuffle [26] for upscaling to save the parameters. In model compression, they applied a model pruning, model quantization, and entropy coding.

Various NeRV-like methods have been proposed for video coding. D-NeRV [23] was designed to effectively encode large-scale and diverse videos. It reduced memorization complexity among subsequent video frames, by providing clip-specific visual content and global temporal dependency information. In addition, task-oriented flow was predicted and used instead of directly predicting raw pixel values, thereby reducing complexity by preventing memorization of the same pixel values in different frames. HNeRV [12] proposed a hybrid INRV model that provided visual prior for better reconstruction and compression performance by using tiny content-adaptive frame embedding as inputs of a learned neural decoder. It was used as visual priors and increased the efficiency of the model, while keeping the same amount of information about the target visual data. The methods used pruning, quantization, and entropy encoding for model compression[12], [23]. E-NeRV [13] disentangled spatial and temporal contexts in the INRV to improve coding performance.

B. Neural Network Compression (NNC)

Neural network compression (NNC) is the process of reducing the size or complexity of a neural network model, while avoiding significant loss in performance. There are several methods commonly used for NNC, including pruning [27], quantization [28], and low-rank factorization [29].

Pruning is used to build a compact network with a less number of parameters [27], [30]. The pruning process eliminates the connections or weights with small magnitudes, while giving a minimal impact on the efficiency of the network. Quantization [28], [31] is used to lower the precision of the weights. Low-rank factorization algorithms approximate the weight matrices of a neural network into a more lightweight form and allow for more effective computations.

III. Proposed Method

A. Overview

We propose an improved INRV-based video coding method with an R-D optimization of a target neural network, based on the NeRV [10]. Most of the INRV-based video coding methods are built on a reconstruction module to fit video data to network parameters and recover the original spatial dimension in the output. We use the conventional INRV pipeline as a baseline, as shown in Fig. 2. It uses a PE, MLPs, and upscaling modules (UM) for the reconstruction. The MLP uses two linear layers.

The UMs are used to recover the dimension of the extracted features to the original dimension, consisting of five upscaling modules. The network parameters are quantized and encoded through an entropy coding.

This paper develops a novel R-D optimization technique that progressively controls the quality of the reconstruction, by searching the optimal network architecture, in addition to the conventional compression pipeline. For this purpose, a PFEM is incorporated into the baseline. The PFEM consists of a stack of RBs. When the depth of a model increases with the number of RBs, the representation capability improves. However, this causes a problem that the size of the model becomes large. In the proposed method, the quality is controlled through the number of RBs in the PFEM, because the RBs directly control the quality of the output frame and require more bits in transmission, as shown in Fig. 3.

B. Progressive Feature Extraction Module (PFEM)

The PFEM uses the stack of RBs consisting of two 3×3 convolution layers, as shown in Fig. 2. The RBs have been widely used for the tasks of image reconstruction and restoration. By using the RB, the gradient loss problem is mitigated with skip connections, by allowing the gradient to bypass a specific layer. In addition, by reducing the duplication of model representation with such skip connections, although the size of a target model decreases with model compression, the representation capability can be maximally maintained.

The PFEM is placed between the upscaling process in Fig. 2. The upscaled features are used as the input of the PFEM so that the spatial information of the features can be learned effectively. Through this, the baseline can further utilize spatial information, although it used only a temporal index as input. By placing the PFEM in the middle of the upscaling process, the size of the model is effectively adjusted by passing the features to the reconstruction process.

Fig. 3 displays the R-D performance with respect to the number of RB. As the number of RBs increases, the number of parameters increases, improving both peak signal-to-noise ratio (PSNR) and multi-scale structural similarity index (MS-SSIM). In addition, it can be seen that the rate increases as the number of parameters increases.

C. Optimal PFEM Selection

The number of the RBs consisting of PFEM is chosen with R-D optimization. To this aim, we define a Lagrangian cost J, defined as,

$$J(n) = D(n) + \lambda R(n), \tag{1}$$

where n is the number of the RBs, D is the distortion, and R is the bit-rates. λ is the Lagrangian multiplier. In our experiments, λ is set up as follows:

$$\lambda = \alpha \cdot 2^{\beta \times R}, \tag{2}$$

where α is 8.5, and β is -96.8, which are obtained empirically.

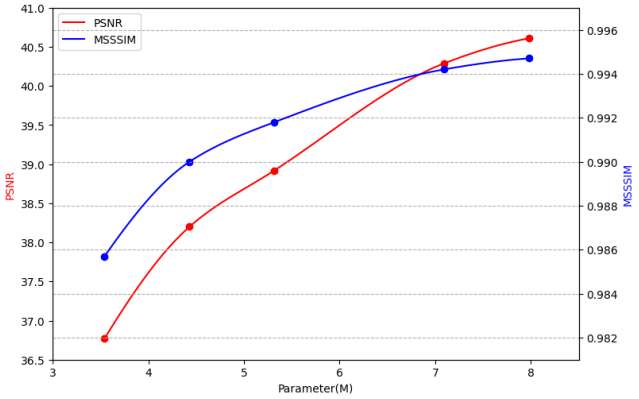


Fig. 3. The quality of the reconstructed frame in terms of PSNR and MS-SSIM and increment of bit-rates according to the number of RBs in the PFEM.

The optimal size is determined by minimizing J as follows:

$$n^* = \arg\min_{n \in N} J(n), \tag{3}$$

where n^* is the solution, and N is the set of the solutions. In the experiments, N is $\{12, 24, \cdots, 72\}$ with multiples of 12.

IV. EXPERIMENTAL RESULTS

A. Dataset and Experimental Setting

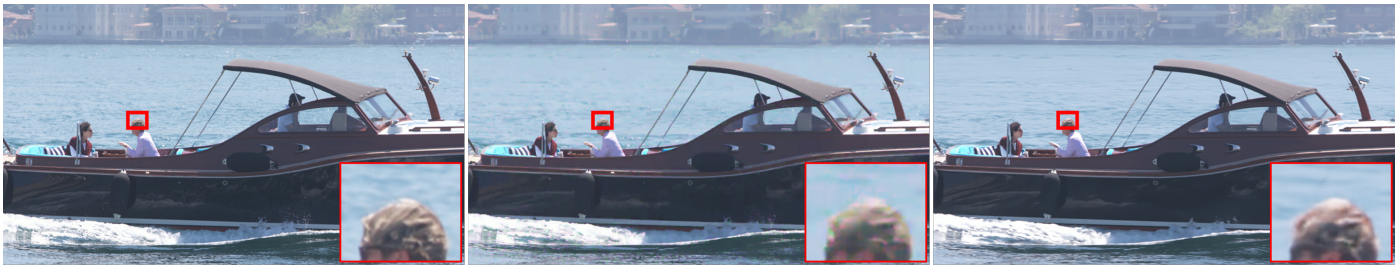
We conduct comparative experiments to evaluate the performance of the proposed method. We employ the conventional video codecs, H.264/AVC [1] and H.265/HEVC [2], and the baseline [10] for comparisons. We use "Big Buck Bunny", UVG [32], and DAVIS [33] dataset in the experiments.

Adam optimizer [34] is used to train the networks, and a learning rate is set to 0.0005 along with the cosine annealing learning rate scheduler. The number of training epochs was set to 1,000, and the batch size was set to 1.

B. Performance Evaluation

1) Quantitative evaluations: Table I displays a quantitative evaluation by calculating the Bjontegaard delta bit rate (BDBR) for each of NeRV and the proposed method, when H.264 is used as an anchor. The proposed method shows significantly improved coding performance as compared with NeRV for all the test sequences in the UVG dataset. Our method provides approximately 15.3% BD-rate reduction on average, as compared with H.264/AVC and 54.2% BD-rate reduction on average, as compared with NeRV as a baseline. Fig. 5 shows the R-D curves of H.265/HEVC, H.264/AVC, NeRV, and the proposed method (NeRV+PFEM). The proposed method provides a superior coding gain to the NeRV in all the ranges of the rates.

2) Qualitative evaluations: Fig. 4 shows the qualitative evaluation comparing H.264/AVC and the proposed method in the same bits per pixel (bpp) on "Yacht" test video sequence. The most left figure shows the ground truth. In the subjective comparisons, it was demonstrated that the proposed method maintains fine texture details after the compression.



(a) Ground Truth (b) H.264 (c) Ours

Fig. 4. Subject quality comparisons of the proposed method and H.264/AVC on "Yacht" test video sequence.

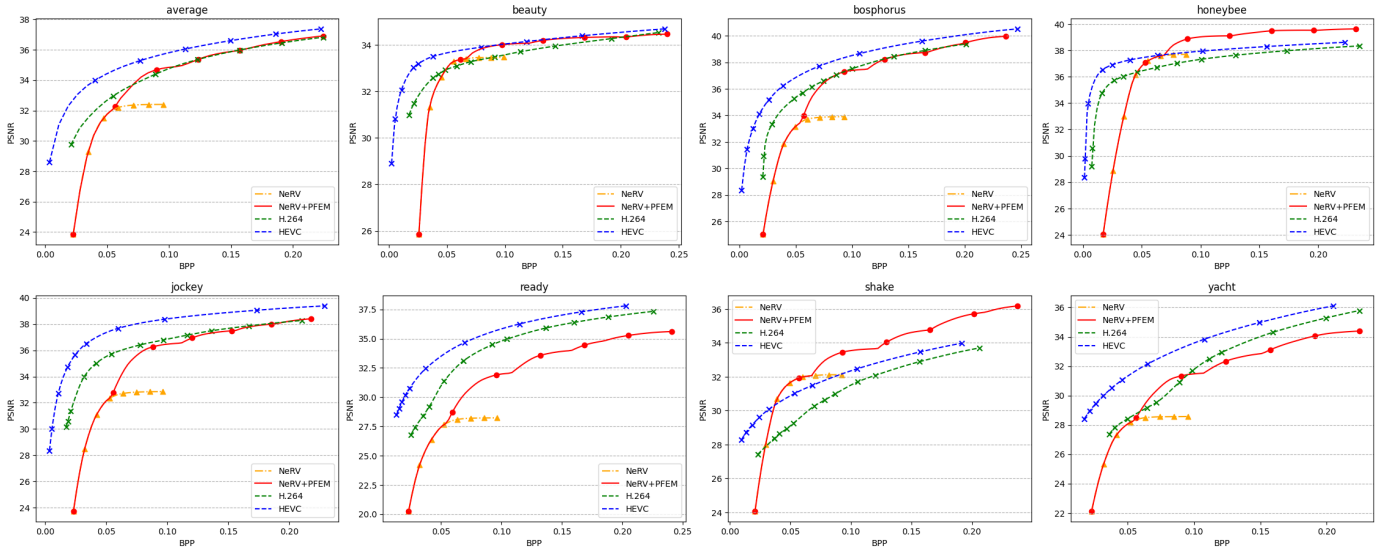


Fig. 5. R-D curves of H.264/AVC, H.265/HEVC, NeRV, and the proposed method (NeRV+PFEM) on UVG dataset.

TABLE I
COMPARISON OF R-D PERFORMANCE (BDBR) ON NERV AND NERV+PFEM ON UVG DATASET. H.264 IS AN ANCHOR.

Video	BDBR (%)	
	NeRV [10]	Ours
beauty	5.13	-29.09
bosphorus	80.66	2.00
honeybee	-13.92	-96.38
jockey	121.70	4.38
ready	80.88	52.82
shake	-26.09	-38.77
yacht	23.85	-2.14
avg.	38.89	-15.31

V. CONCLUSIONS

In this paper, we proposed an implicit neural representation model for video compression with PFEM to improve the reconstruction quality while optimizing the trade-off between model size and compression performance. Also, by defining the equation for the R-D optimization with λ, we demonstrated the way to select the PFEM size according to the trade-off between model size and compression performance. In the experimental results, quantitative and qualitative results for video reconstruction have been enhanced.

ACKNOWLEDGMENT

This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and partly supported by the NRF grant funded by MSIT (No.NRF2022R1A2C4002052).

REFERENCES

[1] T. Wiegand and G. J. Sullivan, "The h. 264/avc video coding standard [standards in a nutshell]," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 148–153, 2007.

[2] P. Zhao, H. Zhu, H. Li, and T. Shibata, "A directional-edge-based real-time object tracking system employing multiple candidate-location generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 3, pp. 503–517, 2012.

[3] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[4] O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3454–3463.

[5] H. Liu, M. Lu, Z. Ma, *et al.*, "Neural video coding using multiscale motion compensation and spatiotemporal context model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3182–3196, 2020.

[6] J.-K. Lee, N. Kim, S. Cho, and J.-W. Kang, "Convolution neural network based video coding technique using reference video synthesis," in *Proceedings, APSIPA Annual Summit and Conference*, 2018, pp. 12–15.

[7] J.-K. Lee, N. Kim, S. Cho, and J.-W. Kang, "Deep video prediction network-based inter-frame coding in hevc," *IEEE ACCESS*, vol. 8, pp. 95 906–95 917, 2020.

[8] N. K. Jung-Kyung Lee and J.-W. Kang, "Rate-distortion optimized temporal segmentation using reinforcement learning for video coding," in *Asia Pacific Signal and Information Processing (APSIPA)*, IEEE, 2021, pp. 1–4.

[9] N. K. J. K. Lee and J. W. Kang, "Reinforcement learning for rate-distortion optimized hierarchical prediction structure," *IEEE ACCESS*, vol. 11, pp. 20 240–20 253, 2023.

[10] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "Nerv: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 557–21 568, 2021.

[11] H. Chen, M. Gwilliam, B. He, S.-N. Lim, and A. Shrivastava, "Cnerv: Content-adaptive neural representation for visual data," *arXiv preprint arXiv:2211.10421*, 2022.

[12] H. Chen, M. Gwilliam, S.-N. Lim, and A. Shrivastava, "Hnerv: A hybrid neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 270–10 279.

[13] Z. Li, M. Wang, H. Pi, K. Xu, J. Mei, and Y. Liu, "E-nerv: Expedite neural video representation with disentangled spatial-temporal context," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, Springer, 2022, pp. 267–284.

[14] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5939–5948.

[15] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[16] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[17] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE signal processing magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[18] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[19] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "Pixelnerf: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4578–4587.

[20] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8628–8638.

[21] Z. Chen, Y. Chen, J. Liu, *et al.*, "Videoinr: Learning video implicit neural representation for continuous space-time super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2047–2057.

[22] D. Rho, J. Cho, J. H. Ko, and E. Park, "Neural residual flow fields for efficient video representations," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3447–3463.

[23] B. He, X. Yang, H. Wang, *et al.*, "Towards scalable neural representation for diverse videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6132–6142.

[24] Y. Lu, Z. Wang, M. Liu, H. Wang, and L. Wang, "Learning spatial-temporal implicit neural representations for event-guided video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1557–1567.

[25] Y. Du, Y. Zhang, H.-X. Yu, J. B. Tenenbaum, and J. Wu, "Neural radiance flow for 4d view synthesis and video processing," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2021, pp. 14 304–14 314.

[26] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[27] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.

[28] B. Jacob, S. Kligys, B. Chen, *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[29] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," *arXiv preprint arXiv:1405.3866*, 2014.

[30] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on cpus," 2011.

[31] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," *arXiv preprint arXiv:1806.08342*, 2018.

[32] A. Mercat, M. Viitanen, and J. Vanne, "Uvg dataset: 50/120fps 4k sequences for video codec analysis and development," in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, pp. 297–302.

[33] H. Wang, W. Gan, S. Hu, *et al.*, "Mcl-jcv: A jnd-based h. 264/avc video quality assessment dataset," in *2016 IEEE international conference on image processing (ICIP)*, IEEE, 2016, pp. 1509–1513.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Deep Unfolded Underwater Image Enhancement Based on Extreme Channels Prior

Thuy Thi Pham, Truong Thanh Nhat Mai, and Chul Lee
Department of Multimedia Engineering, Dongguk Univerisity, Seoul, Korea
E-mail: {pham.thuy, mtntruong}@mme.dongguk.edu, chullee@dongguk.edu

***Abstract*—We propose a deep unrolling approach for underwater image enhancement using extreme channels prior. First, we formulate underwater image enhancement as a joint optimization problem that incorporates an underwater-related extreme channels prior and implicit regularization functions. Then, we solve the optimization problem iteratively and develop an unfolded deep neural network, where each block of the network represents an iteration in which the optimization variables and regularizers are updated using closed-form solutions and learned proximal operators, respectively. Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art underwater image enhancement algorithms in both quantitative and qualitative comparisons.**

I. Introduction

Underwater image enhancement is an important preprocessing task for many applications in marine science and technology because the visual quality of captured underwater images degrades due to light absorption, refraction, and scattering. Therefore, underwater images exhibit low contrast, low brightness, blurry textures, and color distortion. To address this challenge, various underwater image enhancement algorithms have been developed, which can be broadly categorized into model- and learning-based approaches.

Earlier model-based approaches focused on directly enhancing the pixel intensities in underwater images without considering specific underwater imaging models [1]–[3]. However, these algorithms often lead to undesirable over- and under-enhancements, as they lack an understanding of the characteristics of underwater images. To overcome this limitation, researches have been made to formulate underwater image enhancement as an inverse problem by incorporating physics-based models. They employed image priors, *e.g.*, the dark channel [4], light attenuation [5], and hyper-Laplacian reflectance priors [6], to estimate the parameters for underwater imaging models. However, these approaches may be inaccurate due to manually designed priors.

Recently, significant advancements have been made in underwater image enhancement using deep learning techniques, *i.e.*, convolutional neural networks (CNNs) and generative adversarial networks (GANs). For example, Li [7] developed a baseline CNN for underwater image enhancement. Islam *et al.* [8] proposed a GAN-based algorithm to enhance underwater images in real-time using both paired and unpaired training. Liu *et al.* [9] proposed an object-guided twin-adversarial contrastive learning approach for enhancing underwater images. Although learning-based algorithms provide flexibility, adaptability, and efficient inference, their behaviors are difficult to analyze and interpret because of limited domain knowledge on to underwater images.

To overcome the limitations of both model- and learning-based approaches while leveraging their strengths, a model-based deep learning technique, which is known as algorithm unrolling, has recently been developed. Algorithm unrolling involves unfolding an iterative optimization algorithm into a series of layers within deep neural networks; each layer of the network corresponds to an iteration of the model-based iterative algorithm. Thus, algorithm unrolling can benefit from both approaches—interpretability from the model-based approach and data-driven adaptability from the deep learning-based approach. Because of this advantage, algorithm unrolling has recently been widely employed in various image processing and low-level computer vision tasks [10]–[12].

In this work, we propose an algorithm unrolling-based end-to-end underwater image enhancement algorithm using extreme channels prior. First, we formulate an inverse problem for underwater image enhancement as a joint optimization that incorporates physics-based constraints and underwater-related priors. In addition, we add implicit regularizers to capture prior knowledge of underwater images and compensate for modeling inaccuracies. To solve the optimization problem, we employ the augmented Lagrange multiplier (ALM) method, which divides the problem into subproblems that can be iteratively minimized. At each iteration, we update the optimization variables and regularizers using closed-form solutions and deep networks, respectively. Experimental results demonstrate that the proposed algorithm significantly outperforms state-of-the-art algorithms in both quantitative and qualitative comparisons.

II. Proposed Algorithm

A. Problem Formulation

The underwater image acquisition model is derived from the light transmission model, which is given by

$$\mathbf{I}(x) = \mathbf{J}(x)\mathbf{t}(x) + \mathbf{B}\left(1 - \mathbf{t}(x)\right), \tag{1}$$

where $\mathbf{I}$ and $\mathbf{J}$ denote the observed underwater and latent clean images, respectively; $\mathbf{t}$ and $\mathbf{B}$ are the transmission map and

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1F1A1074402).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

background light, respectively; x is the pixel location.

Given an observed underwater image $\mathbf{I}$, the underwater image enhancement task for restoring $\mathbf{J}$ is formulated as a joint optimization as

$$\min_{\mathbf{J},\mathbf{t},\mathbf{B}} \frac{1}{2}\|\mathbf{J}\mathbf{t}+\mathbf{B}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2, \tag{2}$$

where $\mathbf{1}$ denotes a matrix of all ones, and the multiplication is element-wise. However, because the problem in (2) attempts to recover $\mathbf{J}$, $\mathbf{B}$, and $\mathbf{t}$ from a single input underwater image $\mathbf{I}$, it is highly ill-posed. Thus, it is essential to develop effective underwater-related priors that can constrain the inverse problem.

To this end, we first impose constraints on transmission map $\mathbf{t}$ and background light $\mathbf{B}$ based on the maximum attenuation model [13]. Then, the optimization problem in (2) becomes

$$\begin{aligned}\min_{\mathbf{t},\mathbf{B},\mathbf{J}} \frac{\lambda_1}{2}\|\mathbf{J}\mathbf{t}+\mathbf{B}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2+\frac{\lambda_2}{2}\left\|\mathbf{t}-\tilde{\mathbf{t}}\right\| \\ +\frac{\lambda_3}{2}\|\mathbf{B}-\widetilde{\mathbf{B}}\|_2^2,\end{aligned} \tag{3}$$

where λ_1, λ_2, and λ_3 are the regularization parameters, and $\tilde{\mathbf{t}}$ and $\widetilde{\mathbf{B}}$ are the initial estimations of $\mathbf{t}$ and $\mathbf{B}$, respectively. We determine $\tilde{\mathbf{t}}$ and $\widetilde{\mathbf{B}}$ using the maximum attenuation model [13]. The optimal transmission map and background light are expected to be close to their initial estimations.

To improve the enhancement performance, we incorporate two extreme channels priors—the dark channel prior (DCP) [14] and bright channel prior (BCP) [15]—into (3) to better model the properties of natural clean images. In addition, we add two implicit regularization functions, $f(\cdot)$ and $g(\cdot)$, to capture the prior knowledge of clean image $\mathbf{J}$ and transmission map $\mathbf{t}$, respectively. Then, the optimization problem in (3) can be rewritten as

$$\begin{aligned}\min_{\mathbf{t},\mathbf{B},\mathbf{J}} \frac{\lambda_1}{2}\|\mathbf{J}\mathbf{t}+\mathbf{B}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2+\frac{\lambda_2}{2}\left\|\mathbf{t}-\tilde{\mathbf{t}}\right\| \\ +\frac{\lambda_3}{2}\|\mathbf{B}-\widetilde{\mathbf{B}}\|_2^2+\lambda_4\|\mathcal{D}(\mathbf{J})\|_1 \\ +\lambda_5\|\mathbf{1}-\mathcal{B}(\mathbf{J})\|_1+f(\mathbf{J})+g(\mathbf{t}),\end{aligned} \tag{4}$$

where $\mathcal{D}(\mathbf{J})$ and $\mathcal{B}(\mathbf{J})$ denote the dark and bright channels of image $\mathbf{J}$, respectively, and λ_4 and λ_5 are the regularization parameters.

B. Optimization Algorithm

We employ the ALM method to solve the problem in (4). To this end, we first introduce auxiliary variables $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{G}$, and $\mathbf{H}$ to split variables $\mathcal{D}(\mathbf{J})$, $\mathbf{1}-\mathcal{B}(\mathbf{J})$, $\mathbf{J}$, and $\mathbf{t}$, respectively. Then, the optimization problem in (4) can be reformulated as

$$\begin{aligned}\min_{\mathbf{t},\mathbf{B},\mathbf{J},\mathbf{P},\mathbf{Q},\mathbf{G},\mathbf{H}} \ & \frac{\lambda_1}{2}\|\mathbf{J}\mathbf{t}+\mathbf{B}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2+\frac{\lambda_2}{2}\left\|\mathbf{t}-\tilde{\mathbf{t}}\right\| \\ & +\frac{\lambda_3}{2}\|\mathbf{B}-\widetilde{\mathbf{B}}\|_2^2+\lambda_4\|\mathbf{P}\|_1+\lambda_5\|\mathbf{Q}\|_1 \\ & +f(\mathbf{G})+g(\mathbf{H}) \\ \text{subject to} \ & \mathbf{P}=\mathcal{D}(\mathbf{J}),\ \mathbf{Q}=\mathbf{1}-\mathcal{B}(\mathbf{J}), \\ & \mathbf{G}=\mathbf{J},\ \mathbf{H}=\mathbf{t}.\end{aligned} \tag{5}$$

Next, we define the augmented Lagrangian function for (5) as

$$\begin{aligned}&\mathcal{L}(\mathbf{B},\mathbf{t},\mathbf{J},\mathbf{P},\mathbf{Q},\mathbf{G},\mathbf{H},\mathbf{\Gamma},\mathbf{\Upsilon})=\frac{\lambda_1}{2}\|\mathbf{J}\mathbf{t}+\mathbf{B}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2 \\ &+\frac{\lambda_2}{2}\left\|\mathbf{t}-\tilde{\mathbf{t}}\right\|+\frac{\lambda_3}{2}\|\mathbf{B}-\widetilde{\mathbf{B}}\|_2^2+\lambda_4\|\mathbf{P}\|_1+\frac{\gamma_1}{2}\|\mathbf{P}-\mathcal{D}(\mathbf{J})\|_2^2 \\ &+\lambda_5\|\mathbf{Q}\|_1+\frac{\gamma_2}{2}\|\mathbf{Q}-(\mathbf{1}-\mathcal{B}(\mathbf{J}))\|_2^2+f(\mathbf{G})+g(\mathbf{H}) \\ &+\frac{\gamma_3}{2}\|\mathbf{G}-\mathbf{J}\|_2^2+\langle\mathbf{\Gamma},\mathbf{J}-\mathbf{G}\rangle+\frac{\gamma_4}{2}\|\mathbf{H}-\mathbf{t}\|_2^2+\langle\mathbf{\Upsilon},\mathbf{t}-\mathbf{H}\rangle,\end{aligned} \tag{6}$$

where $\gamma_1,\cdots,\gamma_4$ denote the penalty parameters, and $\mathbf{\Upsilon}$ and $\mathbf{\Gamma}$ are the Lagrange multiplier matrices.

The solutions to the optimization problem in (5) can be obtained by minimizing the function in (6). To this end, we employ the alternating direction method of multipliers [16], which splits the optimization into subproblems and solves them iteratively and individually. We describe how each optimal solution is obtained at each iteration k.

B- and t-subproblems: Given $\mathbf{J}_k$ and $\mathbf{t}_k$ at the k-th iteration, $\mathbf{B}_{k+1}$ and $\mathbf{t}_{k+1}$ are obtained by

$$\mathbf{B}_{k+1}=\arg\min_{\mathbf{B}} \frac{\lambda_1}{2}\|\mathbf{J}_k\mathbf{t}_k+\mathbf{B}(\mathbf{1}-\mathbf{t}_k)-\mathbf{I}\|_2^2+\frac{\lambda_3}{2}\|\mathbf{B}-\widetilde{\mathbf{B}}\|_2^2, \tag{7}$$

$$\begin{aligned}\mathbf{t}_{k+1}=\arg\min_{\mathbf{t}} \frac{\lambda_1}{2}\|\mathbf{J}_k\mathbf{t}+\mathbf{B}_{k+1}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2+\frac{\lambda_2}{2}\|\mathbf{t}-\tilde{\mathbf{t}}\|_2^2, \\ +\frac{\gamma_4}{2}\|\mathbf{t}-\mathbf{H}_k+\gamma_4^{-1}\mathbf{\Upsilon}_k\|_2^2,\end{aligned} \tag{8}$$

respectively. As both subproblems in (7) and (8) are quadratic problems, their closed-form solutions are readily available.

J-, P-, and Q-subproblems: Next, the solutions for the extreme channel priors $\mathbf{J},\mathbf{P}$, and $\mathbf{Q}$ are obtained by solving

$$\begin{aligned}\min_{\mathbf{J},\mathbf{P},\mathbf{Q}} \frac{\lambda_1}{2}\|\mathbf{J}\mathbf{t}+\mathbf{B}(\mathbf{1}-\mathbf{t})-\mathbf{I}\|_2^2+\lambda_4\|\mathbf{P}\|_1+\frac{\gamma_1}{2}\|\mathbf{P}-\mathcal{D}(\mathbf{J})\|_2^2 \\ +\lambda_5\|\mathbf{Q}\|_1+\frac{\gamma_2}{2}\|\mathbf{Q}-(\mathbf{1}-\mathcal{B}(\mathbf{J}))\|_2^2 \\ +\frac{\gamma_3}{2}\|\mathbf{J}-\mathbf{G}+\gamma_3^{-1}\mathbf{\Gamma}_{k-1}\|_2^2.\end{aligned} \tag{9}$$

We iteratively solve the optimization in (9) by employing a linear mapping of pixels to their dark channel [15].

G- and H-subproblems: We employ deep networks G-Net and H-Net to adaptively learn the physics-based priors associated with explicit regularization functions $f(\cdot)$ and $g(\cdot)$. Specifically, $\mathbf{G}$ and $\mathbf{H}$ are updated as

$$\begin{aligned}\mathbf{G}_{k+1}&=\arg\min_{\mathbf{G}} \frac{\gamma_3}{2}\|\mathbf{G}-(\mathbf{J}_{k+1}+\gamma_3{}^{-1}\mathbf{\Gamma}_k)\|_2^2+f(\mathbf{G}) \\ &=\text{G-Net}_k(\mathbf{J}_{k+1}+\gamma_3^{-1}\mathbf{\Gamma}_k),\end{aligned} \tag{10}$$

$$\begin{aligned}\mathbf{H}_{k+1}&=\arg\min_{\mathbf{H}} \frac{\gamma_4}{2}\|\mathbf{H}-(\mathbf{t}_k+\gamma_4{}^{-1}\mathbf{\Upsilon}_{k-1})\|_2^2+g(\mathbf{H}) \\ &=\text{H-Net}_k(\mathbf{t}_{k+1}+\gamma_4^{-1}\mathbf{\Upsilon}_k).\end{aligned} \tag{11}$$

Finally, the Lagrange multiplier matrices are updated as

$$\mathbf{\Gamma}_{k+1}=\mathbf{\Gamma}_k+\gamma_3(\mathbf{J}_{k+1}-\mathbf{G}_{k+1}), \tag{12}$$

$$\mathbf{\Upsilon}_{k+1}=\mathbf{\Upsilon}_k+\gamma_4(\mathbf{t}_{k+1}-\mathbf{H}_{k+1}). \tag{13}$$

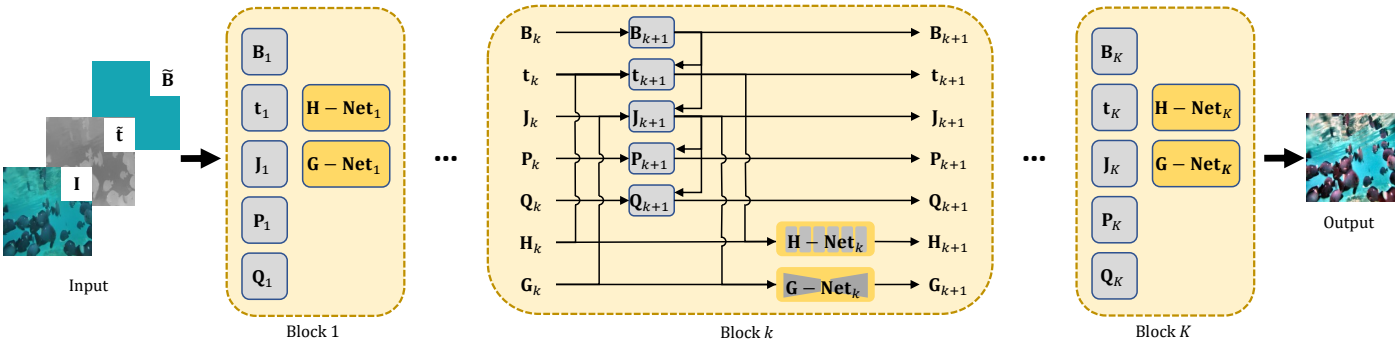


Fig. 1. Overview of the proposed algorithm, which is composed of K blocks.

TABLE I
QUANTITATIVE COMPARISON OF UNDERWATER IMAGE ENHANCEMENT PERFORMANCE. FOR EACH METRIC, THE **BOLDFACED** AND UNDERLINED NUMBERS DENOTE THE BEST AND THE SECOND-BEST RESULTS, RESPECTIVELY.

	UIEB						UCCS		
	PSNR	SSIM	LPIPS(↓)	UCIQE	NIQE(↓)	BRISQUE(↓)	UCIQE	NIQE(↓)	BRISQUE(↓)
GDCP [4]	14.07	0.7554	0.2032	0.6142	4.6391	18.08	0.5372	4.2953	37.74
HLRP [6]	12.55	0.2940	0.4534	**0.6225**	4.7266	22.37	0.5476	4.2630	19.87
MLLE [5]	20.25	0.8663	0.1433	0.5926	5.1330	13.21	0.4752	4.6025	35.02
FUnIE-GAN [8]	16.97	0.7412	0.2719	0.5600	4.7282	**12.96**	0.5034	4.3445	**18.64**
Water-Net [7]	15.76	0.6755	0.2058	0.5285	4.9934	18.80	0.5539	**4.0398**	20.15
Ucolor [17]	21.08	0.8767	0.1081	0.5710	4.7318	14.24	0.5010	4.7182	26.05
TACL [9]	23.25	0.8590	0.1563	0.6100	4.4941	13.18	**0.5768**	4.5761	26.99
Proposed	**24.27**	**0.9281**	**0.0791**	0.6183	**4.3915**	14.17	0.5610	4.1928	21.05

C. Unrolled Network

We adopt the algorithm unrolling strategy to design an unfolded deep network for the proposed iterative optimization algorithm. Fig. 1 shows the architecture of the proposed deep unfolded network. The network takes underwater image $\mathbf{I}$, initial transmission map $\tilde{\mathbf{t}}$, and initial background light $\widetilde{\mathbf{B}}$ as inputs and processes them through K blocks; each block corresponds to an iteration of the iterative optimization algorithm.

At each block k, optimization variables $\mathbf{B}$, $\mathbf{J}$, $\mathbf{t}$, $\mathbf{P}$, and $\mathbf{Q}$ are updated using closed-form solutions, and regularization variables $\mathbf{G}$ and $\mathbf{H}$ are updated using two specific CNNs: G-Net$_k$ and H-Net$_k$. We employ an informative proximal mapping module (IPMM) [18] and a residual dense network (RDN) [19] for G-Net$_k$ and H-Net$_k$. This process is repeated until the final block K, where the enhanced image $\widehat{\mathbf{J}}$ is obtained as the network output.

D. Training

We train the proposed algorithm using the l_1-norm between the ground-truth image $\mathbf{J}_{\text{gt}}$ and an enhanced image $\widehat{\mathbf{J}}$, which is defined as

$$L = \|\widehat{\mathbf{J}} - \mathbf{J}_{gt}\|_1. \tag{14}$$

We use the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and set the initial learning rate to 10^{-4}. The number of blocks, K, is fixed to 5 in all experiments. We set the regularization parameters, λ_1, λ_2, λ_3, λ_4, and λ_5, in (6) to 1.0, 0.7, 0.3, 0.5, and 0.5, respectively, whereas the penalty parameters, $\gamma_1, \ldots, \gamma_4$, are all learned from the training data.

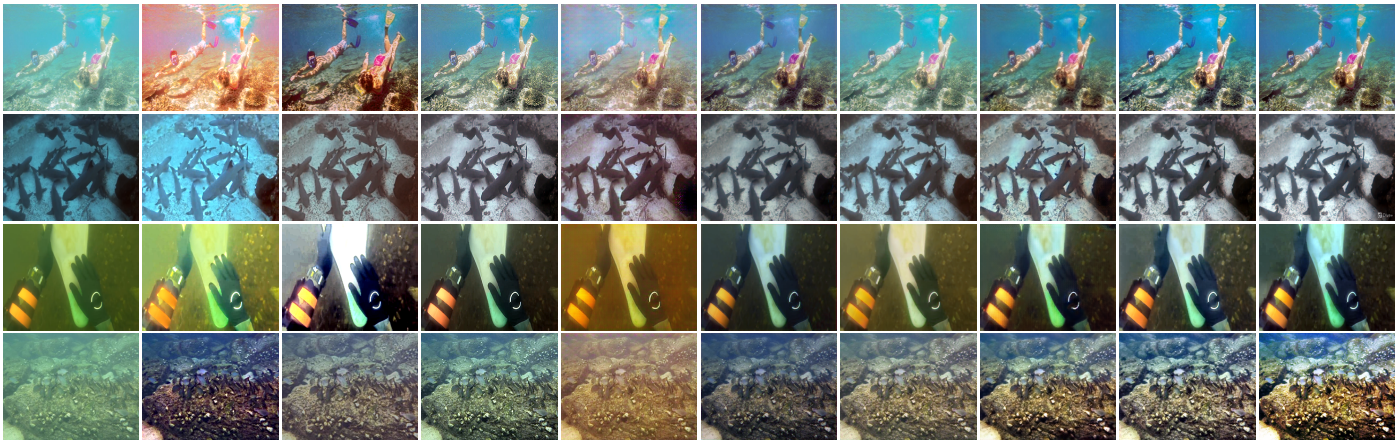
Training dataset: We randomly selected 800 image pairs from the UIEB dataset [7], and the remaining 90 image pairs were used as the test set.

III. EXPERIMENTAL RESULTS

A. Quantitative and Qualitative Comparison

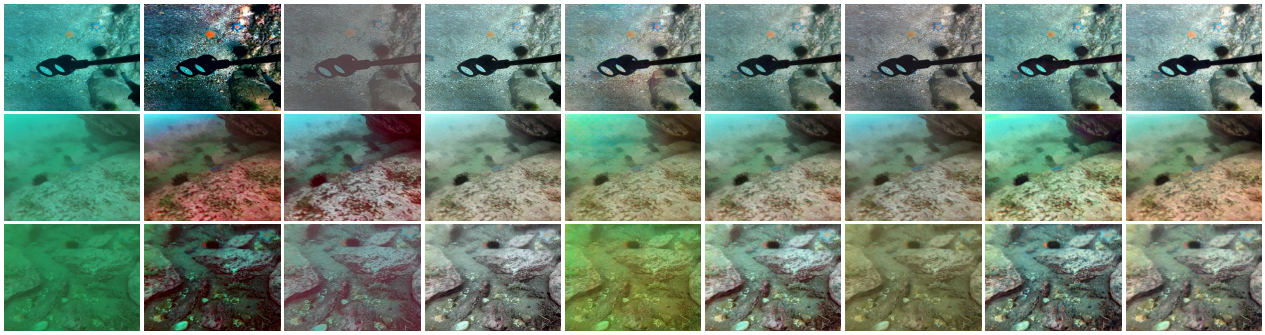
We compare the performance of the proposed algorithm with those the of state-of-the-art algorithms—GDCP [4], HLRP [6], MLLE [5], FUnIE-GAN [8], Water-Net [7], Ucolor [17], and TACL [9]. We use 90 image pairs from the UIEB dataset and 300 referenceless images from the UCCS dataset [20]. We use six quality metrics: PSNR, SSIM, LPIPS [21], UCIQE [22], NIQE [23], and BRISQUE [24]. PSNR, SSIM, and LPIPS are full-reference metrics, whereas UCIQE, NIQE, and BRISQUE are no-reference metrics. Higher PSNR, SSIM, and UCIQE scores indicate better performance, whereas lower NIQE, LPIPS, and BRISQUE scores imply better performance.

Table I quantitatively compares the underwater image enhancement performance. The proposed algorithm achieves the highest scores for PSNR, SSIM, LPIPS, and NIQE and provides comparable UCIQE and BRISQUE scores on the UIEB dataset. For example, the proposed algorithm yields 1.02 dB and 0.0514 higher PSNR and SSIM scores than the second-best TACL and Ucolor, respectively. In addition, the proposed algorithm achieves the second-best scores for the no-



(a) Input (b) GDCP [4] (c) HLRP [6] (d) MLLE [5] (e) F-GAN [8] (f) Water-Net [7] (g) Ucolor [17] (h) TACL [9] (i) Proposed (j) Ground-truth

Fig. 2. Comparison of the underwater image enhancement results on the UIEB dataset.



(a) Input (b) GDCP [4] (c) HLRP [6] (d) MLLE [5] (e) F-GAN [8] (f) Water-Net [7] (g) Ucolor [17] (h) TACL [9] (i) Proposed

Fig. 3. Comparison of the underwater image enhancement results on the UCCS dataset.

reference metrics, UCIQE and NIQE, on the UCCS dataset. This indicates that the proposed algorithm is effective in producing visually appealing and high-quality images consistent with human perception.

Figs. 2 and 3 compare the enhancement results of each algorithm on the UIEB and UCCS datasets, respectively. In Figs. 2(b), (c), and (d), GDCP, HLRP, and MLLE, respectively, fail to recover the colors of the objects accurately, *e.g.*, the first and third rows. Funie-GAN [8] in Fig. 2(e) generates severe color distortions, significantly altering the visual characteristics of the images. In Figs. 2(f) and (g), Water-Net and Ucolor, respectively, provide blurry and low-contrast results with unnatural color tones. In contrast, both TACL and the proposed algorithm in Figs. 2(h) and (i), respectively, provide restored images closest to the ground-truths in Fig. 2(j) without visible color changes. Furthermore, Fig. 3 confirms the effectiveness of the proposed algorithm in enhancing real-world underwater images, producing more natural-looking results than the conventional algorithms.

B. Effectiveness of Extreme Channels Prior

We analyze the effectiveness of DCP and BCP using the UCCS dataset by training the proposed network under different settings. Table II compares the enhancement performance. The proposed algorithm with both DCP and BCP yields higher NIQE and BRISQUE scores than the models with only either DCP or BCP. This implies that both DCP and BCP can improve the enhancement performance.

TABLE II
EFFECTS OF THE EXTREME CHANNELS PRIOR ON ENHANCEMENT PERFORMANCE.

BCP	DCP	UCIQE	NIQE(↓)	BRISQUE(↓)
	✓	**0.5625**	4.3243	21.37
✓		0.5580	4.2300	21.18
✓	✓	0.5610	**4.1928**	**21.05**

IV. CONCLUSIONS

We developed an unfolding approach for underwater image enhancement. We first formulated a joint optimization problem that incorporates physics-based constraints, underwater-related priors, and implicit regularizers. Then, we iteratively solved this problem using the ALM method. Next, we unfolded the iterative algorithm into a series of blocks in a deep network, in which the optimization variables and regularizers are updated using closed-form solutions and neural networks, respectively. Experimental results demonstrated that the proposed algorithm provides superior enhancement performance to state-of-the-art algorithms.

REFERENCES

[1] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 81–88.

[2] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 4572–4576.

[3] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, Jan. 2018.

[4] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856–2868, Jun. 2018.

[5] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Trans. Image Process.*, vol. 31, pp. 3997–4010, Jun. 2022.

[6] P. Zhuang, J. Wu, F. Porikli, and C. Li, "Underwater image enhancement with hyper-Laplacian reflectance priors," *IEEE Trans. Image Process.*, vol. 31, pp. 5442–5455, Aug. 2022.

[7] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, Nov. 2020.

[8] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3227–3234, Apr. 2020.

[9] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Trans. Image Process.*, vol. 31, pp. 4922–4936, Jul. 2022.

[10] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3217–3226.

[11] T. T. N. Mai, E. Y. Lam, and C. Lee, "Deep unrolled low-rank tensor completion for high dynamic range imaging," *IEEE Trans. Image Process.*, vol. 31, pp. 5774–5787, Sep. 2022.

[12] T. T. Pham, T. T. N. Mai, and C. Lee, "Model-driven deep unfolding approach to underwater image enhancement," in *Proc. Int. Workshop Advanced Image Technol.*, Jan. 2023, vol. 12592, p. 1259202.

[13] Z. Liang, Y. Wang, X. Ding, Z. Mi, and X. Fu, "Single underwater image enhancement by attenuation map guided color correction and detail preserved dehazing," *Neurocomputing*, vol. 425, pp. 160–172, Feb. 2021.

[14] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[15] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4003–4011.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[17] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Trans. Image Process.*, vol. 30, pp. 4985–5000, May 2021.

[18] C. Mou, Q. Wang, and J. Zhang, "Deep generalized unfolding networks for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17399–17410.

[19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[20] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2822–2837, Aug. 2021.

[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[22] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062–6071, Dec. 2015.

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[24] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Low-Light Image Enhancement via Distillation of NIR-to-RGB Conversion Knowledge

Young-Min Jeong*, Tae-Sung Park, Jeong-Hyeok Park and Jong-Ok Kim
School of Electrical Engineering, Korea University, Seoul, Korea
E-mail: {ymin2570, pts322, ghjon37, jokim}@korea.ac.kr

***Abstract*—In this paper, we propose a novel approach for low-light image enhancement via distillation of NIR-to-RGB conversion network's knowledge. NIR images contain abundant detail information that is independent of lighting conditions. We leverage the knowledge of the NIR-to-RGB conversion network, which effectively preserves detail information from NIR while generating accurate colors, as it shares similarities with low-light image enhancement. By using the rich detail information provided by the NIR-to-RGB conversion for low-light image enhancement, we are able to reduce color distortion and noise while preserving detail information. Our experimental results show the superior performance of our proposed network.**

I. Introduction

When a digital image is acquired, it can be lack of lighting due to low-light environments, performance limitations of photography equipment and its improper settings. Low-light images contain low contrast, massive noise, and significant degradation of detail information, leading to a significant decrease in visibility. Consequently, they diminish the performance of various computer vision techniques designed for normal light conditions, making it difficult to apply those techniques. Therefore, enhancing low-light images is crucial for applying various computer vision techniques such as object and motion recognition in nighttime environments, and 24-hour surveillance.

Near-Infrared (NIR) images are commonly captured in the wavelength range of 700nm to 1000nm as a single-channel representation. Unlike RGB images, they preserve detail and texture well in even low-light environments, making them suitable for various nighttime computer vision applications. However, due to the lack of color information in NIR images, they appear unnatural to human visibility and it is challenging to directly apply them to existing various computer vision tasks. In recent years, NIR colorization research that converts NIR images into RGB has been actively conducted [1]–[3]. It is a challenging task that is involved in generating color information not present in the NIR images.

In this paper, we propose a novel technique to enhance low-light images by leveraging knowledge distillation from an NIR-to-RGB conversion network. The knowledge from the NIR colorization network is distilled to the low-light image enhancement network via the student-teacher model. The proposed method is based on the idea that the NIR-to-RGB conversion network extracts structural information from an NIR images and additionally estimates colors to generate an RGB image. This restoration process is similar to the low-light image enhancement in term of detail and color estimation.

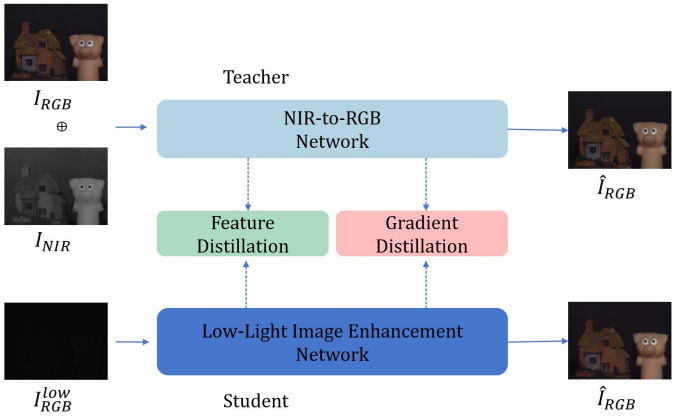


Fig. 1. Concept diagram of the proposed method

Fig. 1 illustrates the concept of the proposed technique which is to distill the knowledge from the NIR-to-RGB conversion network to the low-light image enhancement network in the student-teacher model. The key process in the NIR colorization task is to restore image details and colors, and it is similar to the restoration process in the low-light enhancement task. This observation motivates us to add the NIR input to the conventional student-teacher model for low-light enhancement. Additionally, we introduce a gradient distillation loss to effectively transfer texture information from the teacher network. A dataset with a pair of RGB and NIR images was built, and it was used for evaluating the proposed method. Experimental results show the superior performance of the proposed method.

II. Related Work

In recent years, significant advancements have been made in low-light image enhancement techniques. Many of the low-light image enhancement studies are being conducted based on the Retinex theory [4]. It states that an image can be decomposed into illumination and reflectance components. LIME [5] estimated illumination component only which is just fused with reflectance component. Retinex-Net [6] is a deep network-based method, which consists of image decomposition, illumination enhancement and denoising to the reflectance. KinD [7] and KinD++ [8] use sub-networks for each component

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

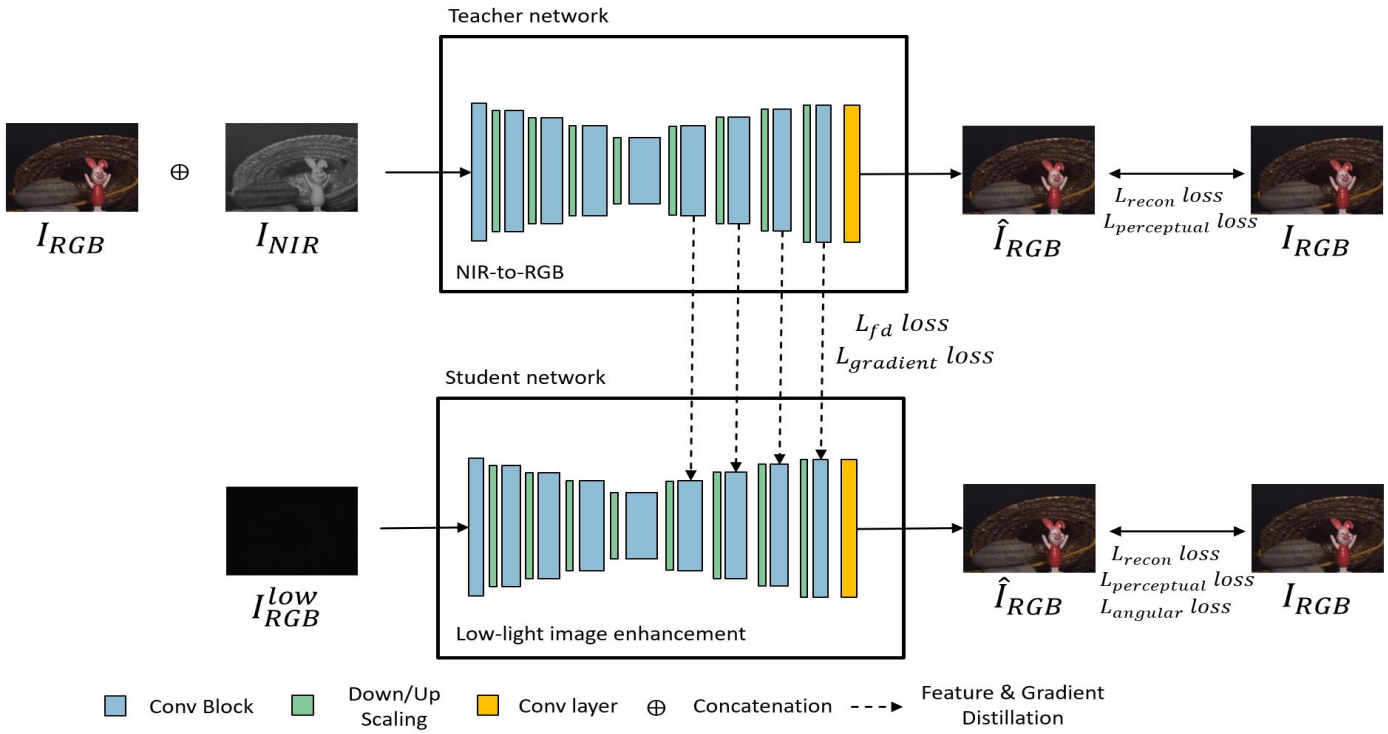


Fig. 2. The overall architecture of the proposed network

after decomposing an image to improve both the reflectance and illumination components. [9] attempts to further enhance low-light images by dual-teacher knowledge distillation. SNR-aware [10] utilizes the correlation between signal and noise to compute attention for patch-based enhancement of the low-light. [11] and [12] leverage the characteristics of near-infrared images that retain rich detail information even in extremely low-light conditions. These techniques have shown promising results and further research is being actively conducted.

III. The Proposed Method

A. Knowledge Distillation

Knowledge distillation [13] is a deep learning model that transfers knowledge from a teacher network to a student network. The student network can effectively learn the feature maps transferred by the teacher network, resulting in the improved performance. We further enhance the performance of a low-light image enhancement network through knowledge distillation of NIR-to-RGB conversion network. The proposed network utilizes two types of knowledge distillation techniques: feature distillation and gradient distillation. Through feature distillation, the overall detail information in the feature maps of the teacher network is learned, while the addition of gradient distillation enables the student network to learn more clear texture information, resulting in the enhanced textures in the generated images.

B. NIR colorization

Near-Infrared (NIR) images preserve abundant detail and texture information well in even low-light environments. Therefore, they have been popularly studied in various night-time computer vision tasks. NIR colorization is a task to aim at generating colorful RGB images, preserving rich details from NIR and restoring colors that appear natural to human visibility. This task is quite challenging as it is involved in generating color information that is not present in the original NIR images. This paper proposes a novel deep network that adopts the NIR-to-RGB conversion network as a teacher.

C. The Proposed Network

Fig. 2 illustrates the architecture of the proposed network which consists of a teacher network for NIR-to-RGB conversion, a student network for low-light image enhancement, and knowledge distillation between the teacher and the student. The teacher network takes both RGB and NIR images as inputs in order to accurately learn NIR colorization, unlike the conventional NIR colorization task which accepts RGB only. It learns the relationship between RGB and NIR images to generate color information resembling the real scene and to learn rich detail information present in NIR images. The teacher network is pre-trained and then, the student is learned while the learned teacher is kept fixed. The student network takes low-light image as input. It learns the relationship between different lighting images and generates normal-light images with rich colors and details close to ground truth images. Both

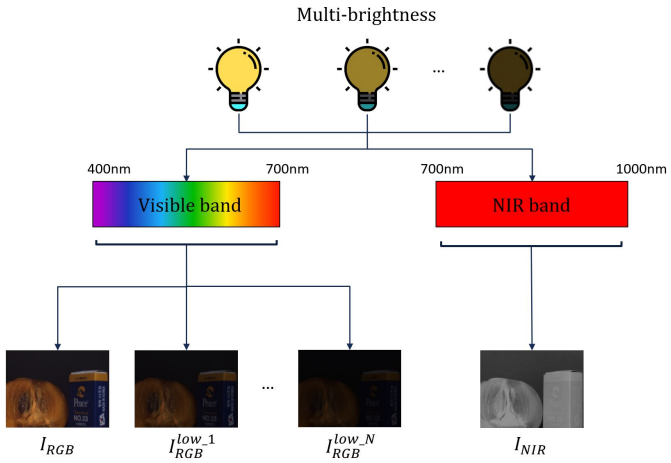


Fig. 3. Constructing dataset under different lighting conditions

the teacher and student networks follow the structure of U-Net [14].

The network for NIR-to-RGB conversion aims to preserve the rich structural and detail information of NIR images while generating realistic colors. The purpose of the network for enhancing low-light images is also to restore detail information by removing noise and generate realistic colors. Based on the idea that the objectives of both tasks are similar, we propose a technique to distill knowledge between the two networks. By distilling the rich details from NIR images in the teacher network, the student can remove noises and enhance detail regions in the low-light images. Similarly, by distilling the abundant color information from RGB images, it can suppress color distortion and improve color restoration capability. The improvement effects of the student network, such as noise reduction, detail enhancement, and color restoration for low-light images, are demonstrated in experimental results.

D. Loss function

The loss function of the proposed network consists of a reconstruction loss, a perceptual loss, an angular loss, a feature distillation loss, and a gradient distillation loss.

The reconstruction loss is to minimize the pixel value difference and is defined as follows:

$$L_{recon} = \|\hat{I}_{RGB} - I_{RGB}\|_1 \tag{1}$$

where $\hat{I}_{RGB}$ and I_{RGB} represent the RGB image generated by the network and the corresponding ground truth image, respectively.

The perceptual loss is to preserve high-level content information and generate more natural-looking images. It is defined as follows:

$$L_{perceptual} = \sum_{layer} |VGG(\hat{I}_{RGB}) - VGG(I_{RGB})|_{layer} \tag{2}$$

The above loss function calculates the difference in feature maps by layer using a pretrained VGG network.

The angular loss is aims to generate more natural colors by employing the Angular error [15], which quantifies the color restoration quality. It is defined as follows:

$$L_{angular} = arccos(\frac{\hat{I}_{RGB} \cdot I_{RGB}}{\|\hat{I}_{RGB}\|\|I_{RGB}\|}) \tag{3}$$

The feature distillation loss is used to train the student network to have similar feature maps to the teacher network. It is defined as follows:

$$L_{fd} = \sum_{layer} \|\alpha(F^S) - F^T\|^2_{layer} \tag{4}$$

where α represents the meta network and is used to reduce the difference in feature maps, effectively facilitating the learning of information from the teacher network.

The gradient distillation loss is used for the student network to learn texture information from the feature maps of the teacher network. It is defined as follows:

$$L_{gradient} = \sum_{layer} \|\nabla F^S - \nabla F^T\|^2_{layer} \tag{5}$$

where ∇ represents the first-order differential operator that includes ∇_x(in the horizontal direction) and ∇_y(in the vertical direction).

The loss function $L_{teacher}$ for the teacher network is defined as follows:

$$L_{teacher} = L_{recon} + L_{perceptual} \tag{6}$$

The loss function $L_{student}$ for the student network is defined as follows:

$$\begin{aligned} L_{student} =& L_{recon} + L_{perceptual} + \lambda_1 L_{angular} \\ &+ \lambda_2 L_{fd} + \lambda_3 L_{gradient} \end{aligned} \tag{7}$$

where λ_1, λ_2 and λ_3 represent hyperparameters.

IV. Experimental Result

A. Dataset

The dataset used for the evaluation of the proposed method consists of pairs of RGB and NIR images. It was created by controlling the intensity of the light to generate five levels of low-light conditions, ranging from approximately 0.1 Lux to 6 Lux. A total of 180 scenes were included in the training dataset, with a total of 900 images across different lighting conditions. The test dataset includes 20 scenes, with a total of 100 images across different lighting conditions. The RGB and NIR images were simultaneously captured for various scenes with multiple objects in a darkroom environment using a 4-channel camera, specifically the JAI's AD-130 GE. The camera captures the reflected light in a single channel format across all NIR wavelength bands.

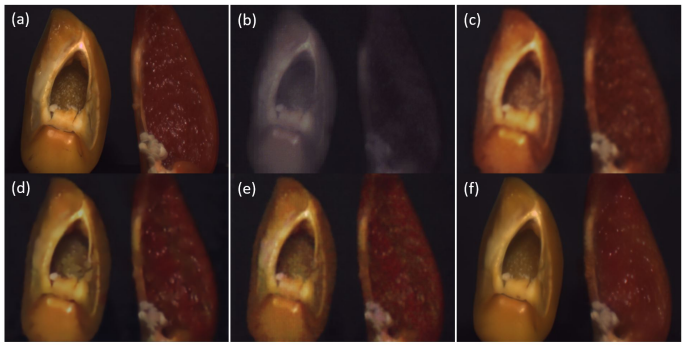


Fig. 4. Visual quality comparison. (a) Ground Truth (b) Retinex-Net [6] (c) KinD [7] (d) KinD++ [8] (e) SNR-Aware [10] (f) Proposed

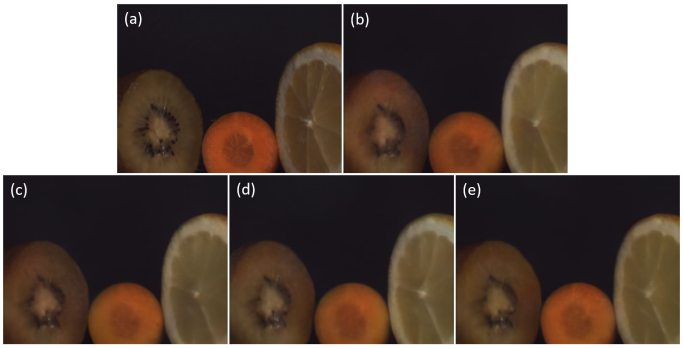


Fig. 5. Visual quality comparison of ablation study. (a) Ground Truth (b) Student only (c) w/o NIR (d) w/o gradient distillation (e) Proposed

TABLE I
QUANTITATIVE EVALUATION OF COMPARATIVE METHODS

Methods	PSNR(dB)	SSIM	AE [15]
Retinex-Net [6]	21.5224	0.8094	8.9735
KinD [7]	25.4068	0.8745	3.8156
KinD++ [8]	26.0031	0.8952	3.4938
SNR-Aware [10]	27.2424	0.8860	3.1859
Proposed	**28.2490**	**0.9086**	**2.6996**

TABLE II
QUANTITATIVE EVALUATION OF ABLATION STUDY

Methods	PSNR(dB)	SSIM	AE [15]
Student only	26.6659	0.9014	2.9089
w/o NIR	27.5823	0.9061	2.9932
w/o gradient distillation	27.6036	0.9064	2.8489
Proposed	**28.2490**	**0.9086**	**2.6996**

B. Performance Evaluation

For the experiments, the hyperparameters λ_1, λ_2 and λ_3 were set to 0.1, 1, and 1, respectively. To evaluate the results of the low-light image enhancement experiments, both quantitative and qualitative evaluations were conducted. Fig. 4 shows the results of the qualitative evaluation. In Fig. 4, Retinex-Net [6] shows very poor performance with inadequate color improvement in the extremely low-light image. KinD [7] appears highly blurred, indicating the loss of detail information. KinD++ [8] shows improved detail information but exhibits darker restored colors compared to the actual ones and suffers from noises. SNR-Aware [10] exhibits unnatural color restoration and noise occurrence. Our result has less noise, preserves details, and produces more realistic colors and achieves superior performance compared to the existing methods. In Table I, all metrics related to structural and color information indicate the superior performance of the proposed method.

C. Ablation study

To evaluate the performance of the proposed teacher network, we conducted a couple of ablation studies such as 'without the teacher', 'without NIR in the teacher', and 'without gradient distillation'. In Fig. 5, (b) represents the case without the teacher network. It is observed that the color distortion and detail loss occurred in the carrot and lemon regions. This demonstrates that the knowledge of a teacher network that converts NIR-to-RGB reduces color distortion and conveys rich detail information. (c) shows the case where the NIR image is removed from the teacher network. It generates colors different from the actual ones and the seed parts of the kiwi and carrot were blurred, resulting in insufficient details. It is demonstrated that the NIR input enhances the restoration ability of detail information and color. (d) represents the case where gradient distillation is removed. Although there was an improvement in detail using the NIR image, the texture visual quality in the lemon and carrot part is insufficient. This indicates that gradient distillation effectively transfers texture information. In Table II, the evaluation of only the student network reveals that the NIR-to-RGB teacher network significantly enhances low-light enhancement performance. The absence of NIR in the teacher's input leads to a decrease in all metrics, indicating that the teacher network effectively utilizes details and structural information from the NIR. Additionally, gradient distillation further conveys the texture information, contributing to the overall improvement of low-light enhancement performance.

V. CONCLUSIONS

Low-light images suffer from a large amount of noise, and the loss of structure detail and color information, leading to a significant decrease in visibility. Therefore, low-light image enhancement techniques are essential in computer vision tasks. In this paper, we proposed a deep learning network that improves low-light images through knowledge distillation from the NIR-to-RGB network. The proposed method is designed based on the idea that low-light image enhancement and NIR-to-RGB conversion share similar objectives such as restoration of image details and color information. By leveraging the rich details of the network that converts NIR to RGB, the proposed method reduces noise and color distortion in low-light images and enhances color restoration capability. Quantitative and

qualitative evaluations of the experimental results demonstrate the superior performance of the proposed method.

ACKNOWLEDGMENT

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Open Field Smart Agriculture Technology Short-term Advancement Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (322033-3).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C2003554).

REFERENCES

[1] Z. Dong, S.-i. Kamata, and T. P. Breckon, "Infrared image colorization using a s-shape network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 2242–2246.

[2] M.-J. Park, J.-H. Lee, S.-H. Lee, and J.-O. Kim, "Multi-band nir colorization using structure-aware network," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, pp. 1682–1686.

[3] T.-S. Park, T.-H. Kim, and J.-O. Kim, "Feature distillation network for multi-band nir colorization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1874–1878.

[4] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[5] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016.

[6] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[7] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1632–1640.

[8] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *International Journal of Computer Vision*, vol. 129, pp. 1013–1037, 2021.

[9] J.-H. Park, T.-H. Kim, and J.-O. Kim, "Dual-teacher distillation for low-light image enhancement," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1351–1355.

[10] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 714–17 724.

[11] J. Zheng, C. Jung, and S. Yu, "Low light image enhancement by multispectral fusion of rgb and nir images," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 2541–2545.

[12] S. Jin, B. Yu, M. Jing, Y. Zhou, J. Liang, and R. Ji, "Darkvisionnet: Low-light imaging via rgb-nir fusion with deep inconsistency prior," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 1104–1112, Jun. 2022.

[13] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1921–1930.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[15] G. D. Finlayson, R. Zakizadeh, and A. Gijsenij, "The reproduction angular error for evaluating the performance of illuminant estimation algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1482–1488, 2016.

3D Human Skeleton Estimation from Single RGB Image Based on Fusion of Predicted Depths from Multiple Virtual-Viewpoints

Wen-Nung Lie[1,2,3], Veasna Vann[1]
[1]Department of Electrical Engineering,
[2]Center for Innovative Research on Aging Society (CIRAS),
[3]Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI),
National Chung Cheng University (CCU), Taiwan
E-mail: ieewnl@ccu.edu.tw

***Abstract* — Estimating 3D human pose from a single RGB image is still challenging in computer vision. Driven by multi-view approach, we propose a model to estimate multiple virtual-view skeletons from a single real-view image which are then fused to derive the final 3D human skeleton. Our network is composed of two stages. The first-stage is a two-streams network, where a Real-Net stream predicts 2D image coordinates and relative depth for each joint from the real-view image and a Virtual-Net stream predicts the relative depths in virtual viewpoints for the same joints. Our second-stage network consists of depth-denoising module and fusion network, where the outputs of the first-stage network are concatenated, depth-denoised, and converted into the predicted 3D human skeleton. The experimental results show that our technique has achieved a performance of MPJPE=47.48 mm, which is comparable to other state-of-the-art methods based on a much longer image sequence (9~243 frames).**

I. INTRODUCTION

Extracting 3D human skeleton (or, the pose) from a single RGB image is challenging in computer vision, which draws many researchers' attention to focus on it. In recent years, 2D and 3D human skeletons have been used in numerous applications in order to achieve the goals of behavior analysis in video surveillance [2], action recognition [1], virtual/ augmented reality (VR/AR), human-computer interaction (HCI), human robot collaboration (HRC) [3], and elderly caring.

3D human skeleton sequence is one kind of input that are crucial to the achievement of action recognition. The recognition accuracy rely on how well the 3D human skeletons are estimated, in other words, the more precise the 3D human skeleton is, the more accuracy the action will be recognized.

For 3D human skeleton estimation, several types of input have been proposed and used, such as single monocular RGB image [4-5], multi-views RGB image [6, 7], RGB image sequence [8, 23] and RGBD images [9]. Among these kinds of inputs, RGB image sequence and multi-view RGB images gained more attention due to their more abundant information as input and the predicted accuracy is higher than monocular single view input. Nevertheless, single image input approach is still popular even their performance is lower compared to multi-view and monocular sequence approaches. This is because single image input has the advantages of a reduced or least amount of hardware or device required, i.e., being cost effective when applying to the real life. On the contrary, sequence image input will certainly increase the software/hardware design complexity, while multi-view image input approach can be only applied to specific scenarios (e.g., athlete training, film production), where multiple cameras from different viewpoints are available.

By surveying numerous papers, it was observed that the accuracy of 3D human skeleton estimation from multi-views [7, 10] is significantly higher than that from single-view [4, 5], due to the fact that multi-view input will provide more useful information that could be incorporated into a neural network model for generating a 3D human skeleton of great precision. To make a compromise, we would like to propose a new method which makes a compromise between the single-view and the multi-views approaches.

The contributions of our work are summarized as follows:

- Our method requires only a single-view RGB image as the input and predicts 2D skeletons and multiple depth information from different virtual viewpoints, which are then fused and used as the input of a fusion network to

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

regress the final 3D human skeleton.

- Our two-stream network predicting multiple depths in several virtual viewpoints is simple and effective.
- A depth-denoising algorithm was adopted in the training procedure, thus reducing the noise effect in following inferences.
- According to the experiments, it was shown that our algorithm provides a substantial improvement compared to the baseline method and even outperforms some methods based on input image sequence.

II. Related Work

A. Single-view methods

Sun et al. [11] proposed a method to predict 3D human pose from a single view input by predicting 3D heatmaps directly first and then apply integral operation (*soft-argmax*) on those heatmaps to obtain 3D joints location. They proposed this method to overcome the problem of non-differentiable post-processing and quantization error of heatmap. To overcome depth ambiguity, some researchers [12, 13] proposed a two-stage approach to predict 2D joints with or without relative depths (e.g., the densely-generated limb depth map [13]) first and then feed them into 2D-to-3D lifting network to estimate 3D skeleton. In [12], a simple method was proposed to lift the detected 2D joints to 3D space by introducing the algorithm of Part-Centric Heatmap Triplets (HEMlets), where three joint-heatmaps were used to represent the relative depth information of the end-joints for each skeletal body part. The outputs of HEMlets are then fused by up-sampling the features and sent to a regression network for estimating the 3D skeleton.

B. Multi-views methods

Since the information from a single monocular RGB image is limited in resolving inherent depth ambiguity, recent researches focus on utilizing information from multiple synchronized cameras to solve this problem. Iskakov et al. [7] proposed to learn differentiable algebraic triangulation and perform a volumetric aggregation of 2D feature maps produced by a 2D backbone. However, their proposed volumetric approach is computationally demanding. In order to avoid intensive computations caused by volumetric grids, Remelli et al. [14] proposed to fuse input images into a unified latent representation of pose via 3D geometry. On the other hand, Qiu et al. [6] presented an approach to predict 2D skeletons from multiple views and then used recursive Pictorial Structure Model (PSM) to recover 3D pose. Different from these methods that require multi-view images as input (hence surrounding in-ward cameras) to predict 3D skeleton, our method requires only single-view image as input (hence, a single camera).

III. Our Method

A. Two-Stage and Two-Streams Network Architecture

In this section, we provide the detailed design of our proposed method. An illustration of the geometry for multiple virtual-viewpoints is presented in Fig. 1, where a real camera and N virtual cameras from different viewpoints surrounding the human are indicated. All the $N+1$ cameras are arranged to viewpoints rotating around the z-axis of the coordinate system.

Hopefully, the goal of our design is to predict all 3D human skeletons in varying viewpoints from a single real RGB image as the input. Instead, we predict, as tradition, the 2D skeleton in the real viewpoint, together with the joint depths (relative to a root joint) along the y-axis of each camera coordinate system. Due to the lack of multi-view image information, joint depths in virtual viewpoints are embedded into the 2D joints of the real viewpoint. Hence, a total of real-view 2D skeleton and all-view joint depths can be predicted. This concept is illustrated in the architecture of Fig. 2, where a two-stage network is presented. In the first stage (Fig. 2(a)), it is composed of a two-streams network, the Real-Net stream and the Virtual-Net stream.

Real-Net is designed to predict the 2D skeletal joint and its relative depths in the real viewpoint v_0, denoted as $(\hat{u}_{v_0}, \hat{v}_{v_0}, \hat{d}_{v_0})$ (the joint index is ignored here), where the relative depth is measured with respect to a root joint (Pelvis). It outputs multiple 3D heatmaps $(J \times 64 \times 64 \times 64)$, J=18 is the number of skeletal joints in our model. Each 3D heatmap is used to predict the u, v, and d components, each of which is quantized into 64 levels for speed and memory consideration. 2D joints and their relative depth values can be obtained by applying *soft-argmax* operation on each 3D heatmaps.

The main goal of Virtual-Net is to estimate joints' relative depths $\{\hat{d}_{v_n}\}_{n=1}^{N}$ in the virtual viewpoints v_n. However, they must be embedded in the corresponding joint $(\hat{u}_{v_0}, \hat{v}_{v_0})$ of the real-view 2D skeleton to form $(\hat{u}_{v_0}, \hat{v}_{v_0}, \hat{d}_{v_n}), n = 1, \dots, N$. This network contains two outputs: (1) a set of 2D heatmaps (64×64×18×N) for predicting $(\hat{u}_{v_0}, \hat{v}_{v_0})$ and (2) a set of depth maps (64×18×N) for predicting $\{\hat{d}_{v_n}\}_{n=1}^{N}$. The 2D heatmaps are applied with *soft-argmax* operation to obtain 2D skeletal joints $(\hat{u}_{v_0}, \hat{v}_{v_0})$, as in the Real-Net, while joint depth values $\{\hat{d}_{v_n}\}_{n=1}^{N}$ predicted in the depth maps could be extracted under the guidance of the output $\hat{u}_{v_0}$ from the Real-Net (by using each joint number as row index and each predicted $\hat{u}_{v_0}$ from the Real-Net stream as the column index). In this paper, we

choose ResNeXt-50 as the backbone of both Real-Net and Virtual-Net.

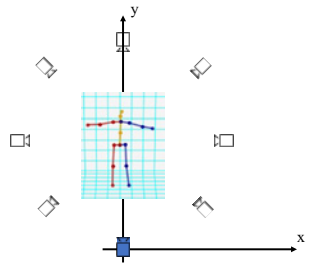


Fig. 1 Geometry of multiple virtual viewpoints (1 real camera (blue one) and N-1 virtual cameras) for human skeleton estimation.

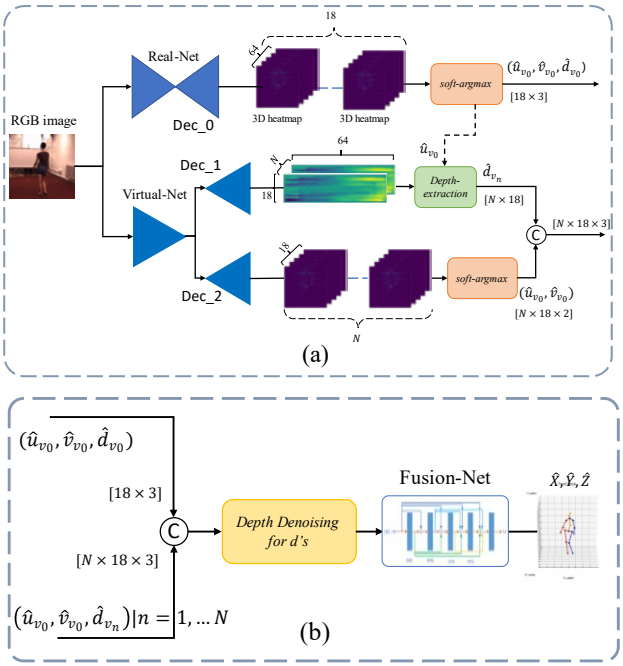


Fig. 2 Architecture design of our proposed approach. (a) 1st stage network. (b) 2nd stage network composed of depth-denoising and fusion-net modules. © is a concatenation operator.

In our design, the N+1 2D skeletons predicted from the Real-Net and Virtual-Net are expected to be the same and hence the two networks could be trained using the same ground truth from the real viewpoint. On the other hand, ground truth of depths in virtual viewpoints can be derived by geometrically rotating the real-view human skeleton for network training.

The second-stage network (Fig. 2(b)) consists of a depth-denoising module and a Fusion-Net which accepts N+1 concatenated 2D skeletons (with denoised depth values) as input and outputs the final 3D skeleton. The "depth denoising" is to disturb the predicted depths by adding Gaussian noises so that the network can have better robustness in test/inference. Our Fusion-Net is constructed by using a dense-connected structure of FC (fully connected) layers as in DenseNets [15] and using PReLU as the activation function.

B. Data Preprocessing and Depth Denoising

Normalization is a standard procedure as data pre-processing, without distorting or information loss, to transform data to have a common range of values. In ground truth preparation for the 1st stage network, the pelvis joint is chosen as the root for calculating the relative depth of each joint, which accompanies with the joint's 2D joint coordinates, are each quantized to a range of [0, 63] (i.e., the size of the 3D heatmap). For the 2nd stage network, ground truth of the 3D coordinates for each joint are normalized to be in the range of [-1, 1].

To derive ground truths of joint depths in virtual viewpoints, 3D skeletons in virtual-views should be generated. Let $S_r^j = (X_r^j, Y_r^j, Z_r^j, 1)$ be a given 3D joint from the real-view, and $S_{v_n}^j = (X_{v_n}^j, Y_{v_n}^j, Z_{v_n}^j, 1)$ is the 3D skeleton in virtual-view v_i. To obtain $S_{v_i}^j$, the equation below is used:

$$S_{v_n}^j = [R|T]S_r^j \tag{1}$$

where j represents the joint index, and R and T are the 3×3 rotation and 3×1 translation matrices (geometrical transformation from the real to the virtual camera v_n). From $S_{v_n}^j$, we can derive the joint depth $d_{v_n}^j$ by subtracting the Z coordinate of the root joint from $Z_{v_n}^j$.

Gaussian noise is commonly used in denoising [24]. Depth denoising module is designed by simply generating a Gaussian noise and add to the predicted depths from the first-stage network. It is represented as follow:

$$\tilde{d}_{v_n}^j = \hat{d}_{v_n}^j + N(\mu, \sigma^2) \tag{2}$$

where $N(\mu, \sigma^2)$ stands for the normally distributed noise generated at a mean μ and a variance σ^2, $\hat{d}_{v_n}^j$ is the predicted depth value for the j-th joint in virtual viewpoint v_n, and $\tilde{d}_{v_n}^j$ is the denoised depth value. Notice that denoising is only performed in training and will be ignored in testing phase.

C. Loss Functions

In the first-stage network, Real-Net is trained by using the Mean Absolute Error (MAE) for the predicted 2D joints and their associated relative depths by:

$$L_{real-net} = \sum_{j=1}^{J} |\hat{s}^j - s^j| \tag{3}$$

where $\hat{s}^j = (\hat{u}^j, \hat{v}^j, \hat{d}^j)$ stands for the predicted 2D coordinate and relative depth, and $s^j = (u^j, v^j, d^j)$ is the ground truth and j is the joint index.

The loss function for training the Virtual-Net is composed of 2D joint loss and relative depth value loss:

$$L_{virtual-net} = \alpha \cdot L_{2D} + \beta \cdot L_{depth} \quad (4)$$

where α and β are hyper-parameters and

$$L_{2D} = \sum_{n=1}^{N} \sum_{j=1}^{J} |\hat{p}_{v_n}^j - p_{v_0}^j| \quad (5)$$

$$L_{depth} = \frac{1}{N.J} \sum_{n=1}^{N} \sum_{j=1}^{J} |\hat{d}_{v_n}^j - d_{v_n}^j| \quad (6)$$

where j is the joint index, $\hat{p}_{v_n}^j = (\hat{u}_{v_n}^j, \hat{v}_{v_n}^j)$ is the predicted 2D coordinates for the j-th joint in virtual viewpoint v_n and $p_{v_0}^j = (u_{v_0}^j, v_{v_0}^j)$ is the ground truth, $\hat{d}_{v_n}^j$ is the predicted relative depth in virtual viewpoint v_n where n is the virtual view index and $d_{v_n}^j$ is its corresponding ground truth.

In the second-stage network, the loss function is composed of 3D joint position loss and bone vector loss, as below:

$$L_{fusion} = \lambda_{bone} \, . \, L_{bone} + \lambda_{joint} \, . L_{joint} \quad (7)$$

where

$$L_{joint} = \sum_{j=1}^{J} smoothL1(\hat{S}^j, S^j) \quad (8)$$

$$L_{bone} = \sum_{k=1}^{K} smoothL1(\hat{B}^k, B^k) \quad (9)$$

where $smoothL1$ is the smooth L1 loss function, $\hat{S}^j$ and S^j are predicted and ground-truth 3D coordinates for joint j, respectively; $\hat{B}^k$ and B^k are the predicted and ground-truth bone vector for k-th bone, respectively; λ_{bone} and λ_{joint} are scalers for bone vector loss and joint position loss, respectively. In this paper, bones consist of some selected virtual bones [23] and physical bones in human skeleton model.

IV. Experimental Results

A. Experimental Settings

In this paper, 3D skeletons in the virtual viewpoints are generated by arranging virtual cameras around the z-axis with equal angular intervals between them, as illustrated in Fig. 1.

Our proposed method is implemented using PyTorch platform. Our first-stage network is trained using initial learning rate of 0.001 and Adam optimizer, where Real-Net is trained for 50 epochs and Virtual-Net is trained for 100 epochs. For the second-stage network, it is trained for 250 epochs and optimized by using AdamW optimizer with an initial learning rate of 0.001. The other hyper-parameters are tuned as: $\alpha = 1$, $\beta = 1, \lambda_{bone}$=10, and $\lambda_{joint} = 1$.

Our first-stage network model is trained by using two popular datasets, Human3.6M [16] and MPII [17]. Adding MPII dataset in the first-stage training will make the models more robust and also applicable to the outdoor scenarios. In the second-stage training, we used only Human3.6M dataset. As in other references, we evaluate the accuracy of our proposed method by using two standard Protocols, Protocol #1 refers to Mean Per Joint Position Error (MPJPE) and Protocol #2 refers to P-MPJPE (MPJPE after performing rigid transformation via Procrustes Analysis). For Human3.6M dataset, 5 subjects (S1, S5, S6, S7, S8) are used for training and 2 subjects (S9, S11) are used for testing.

B. Ablation Study

To verify the impacts of multiple virtual viewpoints arrangement and depth denoising in the proposed method, we conduct ablation studies on different number N of virtual viewpoints and strength/ variance of Gaussian noise for depth denoising.

Table 1 shows the results of our method with different number of virtual viewpoints without depth denoising. It can be seen that when N is not zero, the MPJPE is decreased, meaning that the relative depths from virtual viewpoints help to lift the performance of the predicted 3D skeleton. It is also found that for N={1,2,3,7} the differences in MPJPE values are small, thus we select N={1,2,3} to further perform experiments on combinations of virtual cameras positioning. Finally, we found that when N=2 is used and the two virtual cameras are placed at an left-90 and right-90 degrees interval (i.e., the 2nd and 6-th clockwise positions starting from the real camera in Fig. 1), respectively, the best result can be achieved and denoted as (r, vv1, vv2) in Table 1.

To evaluate the performances of different depth denoising parameters, we choose the best virtual-view configuration N=2 (r, vv1, vv2) in Table 1 and apply the depth denoising to the real-view and virtual-view depths separately, as illustrated in Table 2, where σ_r and σ_v are the noise variances added to the joint depths predicted from the Real-Net and Virtual-Net, respectively. From the experiments, it is found that when $\sigma_r = 0.022$ and $\sigma_v = 0.01$, our proposed method achieves the best performance.

TABLE 1. Comparison of different number of virtual-views. This result is obtained without depth-denoising. (r, vv1, vv2) denotes real view, virtual views in a clockwise rotating angle around z-axis by 90- and 270-degree, respectively.

Number of virtual - views (N)	MPJPE (mm)	PA-MPJPE (mm)
0 (real view only)	50.6	38.1
1	49.2	37.5
2	49.2	37.6
3	49.2	37.7
4	49.3	37.8
5	49.3	37.6
6	49.3	37.6
7	49.2	37.7
8	50.0	38.3
9	49.3	37.7
10	49.3	37.7
2 (r, vv1, vv2)	**49.0**	**37.5**

TABLE 2. Comparison of different σ value for depth-denoising at the (r, vv1, vv2) configuration. σ_r: for real-view depth, σ_v: for virtual-view depths.

σ_r (noise on real)	σ_v (noise on virtual)	MPJPE (mm)	PA-MPJPE (mm)
0 (none)	0 (none)	49.00	37.49
0.022	0 (none)	47.52	37.22
0.022	0.001	47.60	37.17
0.022	0.005	47.50	**36.97**
0.022	0.010	**47.48**	37.50
0.022	0.015	47.78	37.30
0.022	0.020	47.81	37.12

C. Comparison with State-of-the-art Methods

Table 3 shows the comparison of proposed method with state-of-the-art methods on Human3.6M dataset. The T values indicate the number of consecutive frames used for 3D skeleton estimation. It can be seen that our proposed method (47.48mm in MPJPE and 37.5mm in PA-MPJPE) outperforms methods which also adopt a single image (T=1) as the input and achieves a comparable performance with respect to methods which adopt sequence images (T=9~243) as the input. For example, even there is a 2D pose refinement stage in [19], our method outperforms theirs by 1.7mm in MPJPE. Reference [20] proposed a PoseAug to adjust various geometry factors of a pose through differentiable operations, while our depth-denoising module designed by simply using Gaussian noise without any learned parameters can achieve a similar function and has a gain of 2.7mm in MPJPE.

In comparison to [18] and [23] which adopt a shorter length of sequence (*T*=9), our performance is inferior by only 1.7mm and 0.2mm in MPJPE, respectively. Actually, in practice viewpoint, adopting a long sequence of frames (e.g., up to *T*=96 [25] or even 243 [22]) as the input will face disadvantages of heavy network size, large memory requirement, and long delay in response time. On the other hand, our design adopting a single frame as the input not only results in a simplified system, but also achieves a quick response time and hence is a more cost-effective design. The qualitative results of our proposed method are illustrated in Fig. 3. Our network produces plausible 3D human skeleton which similar to the ground truth.

TABLE 3. Comparison of our result with other approaches under Protocol #1 (top) and Protocol #2 (bottom). (*T*: number of temporal frames used for estimation). **Bold**: best; <u>Underlined</u>: second best.

Protocol #1	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg
Gong et al. (CVPR'21) [20] (T=1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50.2
Zou et al. (ICCV'21) [21] (T=1)	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4
Xu et al. (CVPR'20) [19] (T=1)	40.6	47.1	45.7	46.6	50.7	63.1	45.0	47.7	56.3	63.9	49.4	46.5	51.9	38.1	42.3	49.2
Wang et al. (ECCV'20) [25] (T=96)	41.3	43.9	44.0	42.2	48.0	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
Zeng et al. (ICCV'21) [18] (T=1)	43.1	50.4	43.9	45.3	46.1	57.0	46.3	47.6	56.3	61.5	47.7	47.4	53.5	35.4	37.3	47.9
(T=9)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<u>45.8</u>
Chen et al. (TCSVT'22) [22] (T=243)	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	**44.1**
Lie et al. (MMSP'22) [23] (T=31)	40.8	46.0	41.3	57.1	47.0	52.8	39.9	42.3	55.2	72.4	44.7	53.3	47.7	33.3	34.9	47.3
Ours (T=1)	38.0	44.3	41.6	60.4	46.1	38.8	41.1	54.5	76.2	44.2	49.3	57.0	34.8	46.4	36.6	47.48
Protocol #2	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD	Walk	WalkT	Avg
Gong et al. (CVPR'21) [20] (T=1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.1
Zou et al. (ICCV'21) [21] (T=1)	35.7	38.6	36.3	40.5	39.2	44.5	37.0	35.4	46.4	51.2	40.5	35.6	41.7	30.7	33.9	39.1
Xu et al. (CVPR'20) [19] (T=1)	33.6	37.4	37.0	37.6	39.2	46.4	34.3	35.4	45.1	52.1	40.1	35.5	42.1	29.8	35.3	38.9
Wang et al. (ECCV'20) [25] (T=96)	32.9	35.2	35.6	34.4	36.4	42.7	31.2	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	<u>35.5</u>
Zeng et al. (ICCV'21) [18] (T=1) (T=9)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chen et al. (TCSVT'22) [22] (T=243)	32.6	35.1	32.8	35.4	36.3	40.4	32.4	32.3	42.7	49.0	36.8	32.4	36.0	24.9	26.5	**35.0**
Lie et al. (MMSP'22) [23] (T=31)	31.6	35.1	34.2	38.6	37.3	38.9	31.3	32.9	45.2	51.1	36.3	37.0	35.8	24.8	26.9	35.8
Ours (T=1)	31.5	35.3	35.3	41.2	37.9	31.6	32.8	46.2	54.1	37.5	39.1	39.5	27.2	37.2	30.1	37.5

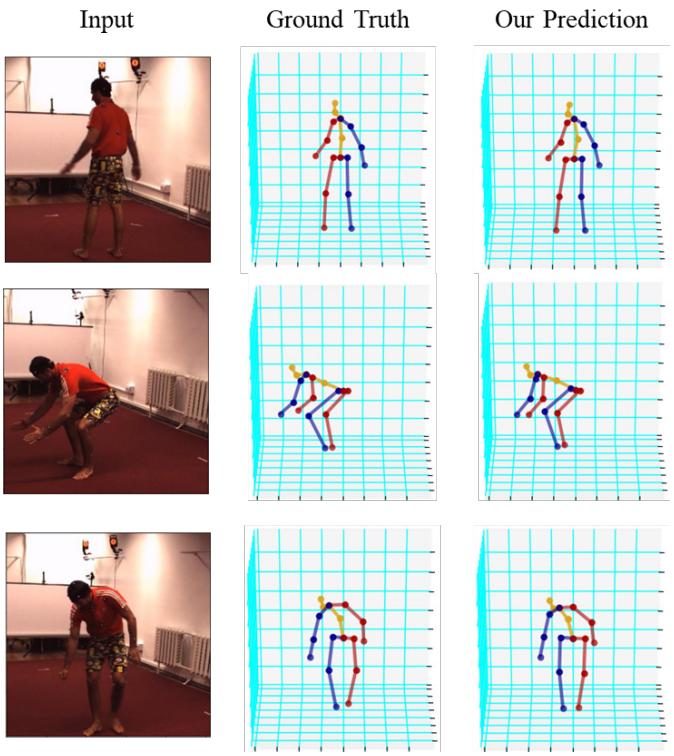


Fig. 3 Qualitative results of our method on Human3.6M dataset.

V. CONCLUSIONS

In this paper, we proposed a two-stage approach for 3D human skeleton estimation. In the first-stage, we proposed a two-streams network to learn joint depths in multiple viewpoints from a single RGB image, including one from real-view and *N* from virtual-views. In the second stage, we propose a simple yet effective method, named depth-denoising, to denoise the predicted depths from the first-stage network. Extensive experiments show that our techniques of fusion from multiple virtual viewpoints and depth denoising help to promote the overall performance. Our design has the advantages of low complexity and fast response time due to the use of a single image as the input. In the future work, the Fusion-Net can be re-designed to further promote the estimation accuracy.

REFERENCES

[1] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963-1978, 2019.

[2] K. Boekhoudt, A. Matei, M. Aghaei, and E. Talavera, "HR-Crime: Human-Related Anomaly Detection in Surveillance Videos," *International Conference on Computer Analysis of Images and Patterns*, pp. 164-174, 2021.

[3] K. Peppas, K. Tsiolis, I. Mariolis, A. Topalidou-Kyniazopoulou, and D. Tzovaras, "Multi-modal 3D Human Pose Estimation for Human-Robot Collaborative Applications," *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops*, pp. 355-364, 2021.

[4] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, "Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2325-2334, 2019.

[5] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic Graph Convolutional Networks for 3D Human Pose Regression," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3420-3430, 2019.

[6] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross View Fusion for 3D Human Pose Estimation," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4341-4350, 2019.

[7] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable Triangulation of Human Pose," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7717-7726, 2019.

[8] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3D Human Pose Estimation Using Spatio-Temporal Networks with Explicit Occlusion Training," *Proc. of the AAAI Conference on Artificial Intelligence,* pp. 10631-10638, 2020.

[9] J. Chun, S. Park, and M. Ji, "3D Human Pose Estimation from RGB-D Images Using Deep Learning Method," *Proc. of the 2018 International Conference on Sensors, Signal and Image Processing*, 2018.

[10] G. Hua, H. Liu, W. Li, Q. Zhang, R. Ding and X. Xu, "Weakly-supervised 3D Human Pose Estimation with Cross-view U-shaped Graph Convolutional Network," *IEEE Transactions on Multimedia*, 2022.

[11] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral Human Pose Regression," *Proc. of the European Conference on Computer Vision (ECCV),* pp. 529-545, 2018.

[12] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2344-2353, 2019.

[13] H. Wu, and B. Xiao. "3D Human Pose Estimation via Explicit Compositional Depth Maps." *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 12378-12385. 2020.

[14] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang, "Lightweight Multiview 3D Pose Estimation through Camera-

Disentangled Representation," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6040–6049, 2020.

[15] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261-2269, 2017.

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 7, pp. 1325-1339, 2014.

[17] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686-3693, 2014.

[18] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu and Q. Xu, "Learning Skeletal Graph Neural Networks for Hard 3D Pose Estimation," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11416-11425, 2021.

[19] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang, "Deep Kinematics Analysis for Monocular 3D Human Pose Estimation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 896-905, 2020.

[20] K. Gong, J. Zhang and J. Feng, "PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8571-8580, 2021.

[21] Z. Zou and W. Tang, "Modulated Graph Convolutional Network for 3D Human Pose Estimation," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11457-11467, 2021.

[22] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen and J. Luo, "Anatomy-Aware 3D Human Pose Estimation With Bone-Based Pose Decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 198-209, 2022.

[23] Wen-Nung Lie, Pei-Hsuan Yang, Veasna Vann, Lee Aing, Jui-Chiu Chiang, "3D Human Skeleton Estimation Based on RGB Image Sequence and Graph Convolution Network," *Proc. of IEEE 24th International Workshop on Multimedia Signal Processing*, MMSP-2022, Shanghai, China, Sept. 26-28, 2022.

[24] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proc. of the 25th International Conference on Machine Learning (ICML)*, pp. 1096-1103, 2008.

[25] J. Wang, S. Yan, Y. Xiong, and D. Lin, "Motion Guided 3D Pose Estimation from Videos, " *European Conference on Computer Vision*, pages 764–780. Springer, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

GNN-Based Small-Data Learning with Area-Control Mechanism for Hyperspectral Satellite Change Detection

Tzu-Hsuan Lin†, Chia-Hsiang Lin*†, and Si-Sheng Young†
* Miin Wu School of Computing, National Cheng Kung University, Taiwan
† Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Taiwan
E-mail: q38091518@gs.ncku.edu.tw; chiahsiang.steven.lin@gmail.com; q36111037@gs.ncku.edu.tw

***Abstract*—Hyperspectral change detection (HCD) has a profoundly impactful contribution to the remote sensing field, such as environmental monitoring, agriculture investigation, disaster assessment, and urban planning. Due to the large-scale data of hyperspectral images (HSIs), conventional regularizers needed in convex optimization (CO), which are typically mathematically sophisticated (e.g., total variation), could lead to a slow computational time. On the other hand, the lack of big data HSIs obstructs the application of deep learning (DE) methods to solve the challenging HCD problem. In this paper, we propose an HCD algorithm based on the CODE theory developed for hyperspectral inverse imaging problems and achieving small-data learning by combining CO and DE. In the proposed small-data learning HCD algorithm, a simple graph change detection network structure (SGCDNet) that learns deep features from limited labeled samples is sufficient. Then, a deep regularizer with a convex form is designed via the Q-quadratic norm function to extract useful information from the rough solution generated by SGCDNet and guide the convex data-fitting, designed with an area-control mechanism, to select a brilliant solution. The proposed convex formulation can be implemented by convex optimizers, such as the alternating direction method of multiplier (ADMM), to achieve high computational efficiency with convergence guaranteed. The real-data experiment shows the superiority of the proposed HCD algorithm.**

I. Introduction

Change detection (CD) is a technique to identify the changed areas among multi-temporal images, which makes significant contributions to environmental monitoring, agriculture investigation, disaster assessment, and urban planning [1]–[4]. Changed pixels of multi-temporal images typically come with material variations at the same position on different dates. Compared to conventional SAR and multispectral images, the high-resolution spectral information of hyperspectral images (HSIs) can provide endmember identifiability [5], [6], which promotes the development of HSIs in the CD field. In this paper, we aim to solve the highly challenging hyperspectral change detection (HCD) problem.

Traditional CD methods mainly focus on analyzing the pixel-level information of the multi-temporal images to detect the changed areas. For example. multivariate alteration detection (MAD), a transformation based on correlations analysis, analyzes the multi-temporal images at the pixel-level [7]. Change vector analysis (CVA) [8] determines the changed areas by investigating the change vector of the multi-temporal images with the intensity of the change vector representing the change probability. In [9], a CD method first adopts principal component analysis (PCA) to learn information from non-overlapping blocks of multi-temporal images, then detect the changed pixels using k-means [10]. In order to enhance the accuracy, HSIs have been adopted for the CD task in recent years. Thanks to the abundant spectral bands of the HSIs, advanced models that developed to fully exploit the spectral information achieve highly precise detection in HCD problems. For example, the HCD algorithm in [11] detects the change map by analyzing the difference of the abundance maps of the bi-temporal HSIs generated by the sparse unmixing model. PTCD exploits the local similarity using a non-overlapping patch strategy followed by the tucker decomposition eliminating the influence of the HSIs, and a detector is designed to generate the change map [12].

Recently, convex optimization (CO), widely used in remote sensing imaging problems, has been adopted for multi-temporal satellite imaging CD. An unsupervised convex relaxation method, called CD-ADMM [13], explicitly defines a convex criterion for the CD problem. Therefore, the convex problem can be solved using the convex optimization solver, alternating direction method of multipliers (ADMM) [14], [15], with convergence guaranteed. Since the inverse problem of CD is ill-posed with non-unique solutions, regularization terms are necessary to mitigate the ill-posedness. However, conventional regularizers are unable to fit various scenarios and could come with time-consuming computing procedures. For example, CD-ADMM adopts the total variation (TV) regularization to preserve the piecewise smoothness of the output change map. When the changed pixels are scattered distribution, the performance of CD-ADMM could be significantly affected. Moreover, the large-scale data of HSIs causes a slow computational time when solving the proximal operator corresponding to the regularization term.

In recent years, deep learning (DE) has drawn attention

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

due to its ability to solve various computer vision tasks by learning deep features embedded in images. To solve HCD problems, convolutional neural network (CNN) structures have been designed to dig for the information and the relationship of the multi-temporal images. In the general end-to-end 2-D convolutional neural network (GETNET), cross-channel gradient information is extracted using a simple mixed-affinity matrix, then the CNN is proposed to generate the change map for the hyperspectral datasets [16]. A siamese 2-D CNN structure is designed in CSANet, which explores the spatial-spectral–temporal features embedded in the multi-temporal HSIs using attention modules [17]. HyperNet proposes a spatial-spectral attention module to investigate the spatial correlation and discriminative spectral features of multi-temporal HSIs [18]. Though DE-based methods have achieved state-of-the-art performance under sufficient training samples, the acquisition and labeling of HSIs are expensive and labor-consuming, making the applications of DE-based algorithms impracticable.

In light of this, we propose an HCD algorithm by introducing a small data-learning theory blending the advantages of CO and DE, called CODE, which has been designed for solving hyperspectral inverse imaging problems and has successfully tackled the highly challenging hyperspectral inpainting problem [19]. As aforementioned, the regularization term in CO could lead to math-heavy optimization procedures, while DE typically relies on big data to learn sufficient information to fit the problem. The CODE theory combines CO and DE to create a learning framework suitable for small data scenarios. Specifically, a simple network structure trained by small data generates a rough solution. Then, the simple convex criterion of the inverse problem obtains the final output by leveraging the deep features embedded in the rough estimate. In the proposed method, a simple graph change detection network (SGCDNet) based on graph neural network (GNN) learns deep features with limited training samples and generates a roughly estimated detection map. Note that an unadorned network structure is sufficient for the CODE theory. Then the convex criterion designed for small-data learning HCD detects the change map with an area-control mechanism controlling the changed area and a simple convex deep regularizer designed via the $\boldsymbol{Q}$-quadratic norm function extracting useful information from the rough output. Such an HCD criterion is implemented by ADMM to achieve a computationally efficient algorithm. The experimental results show the superiority of the proposed HCD method.

Notations: $\mathbb{R}^{m\times n}$ is the set of all the $m \times n$ real-valued matrices. $\mathbb{R}^n$ denotes the n-dimensional Euclidean space. $\|\cdot\|_p$ and $\|\cdot\|_F$ denote the p-norm and Frobenius norm, respectively. $\mathbf{1}$ is the all-one vector with dimension specified in the subscript. vec$(\boldsymbol{M})$ is the vectorization of matrix $\boldsymbol{M}$. $[\boldsymbol{v}]_i$ is the ith entry of vector $\boldsymbol{v}$. $[\boldsymbol{M}]_{(i,j)}$ denotes the (i,j)th entry of matrix $\boldsymbol{M}$. Diag$(\boldsymbol{v})$ denotes the diagonal matrix whose ith diagonal entry is $[\boldsymbol{v}]_i$. diag$(\boldsymbol{M})$ denotes the column vector whose ith entry is $[\boldsymbol{M}]_{(i,i)}$. conv is the convex hull [15]. $\mathcal{I}_Y$ denotes $\{1,\ldots,Y\}$ for any given positive integer Y. $\odot$ denotes the Hadamard product.

II. Proposed method

A. Convex HCD Criterion with Area-Control Mechanism

Let $\boldsymbol{X}_1 \in \mathbb{R}^{M\times L_1 L_2}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{M\times L_1 L_2}$ be the HSIs covering the same area acquired at different dates, with M hyperspectral bands and $L = L_1 \times L_2$ pixels. Since the properties of an unchanged location across different dates should be similar, the distance between the unchanged pixels of the bi-temporal HSIs $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is assumed to be small based on the different image (DI) analysis [9]. Then, the criterion of the HCD problem generating the change map $\boldsymbol{Z} \in \mathbb{R}^{L_1\times L_2}$ can be formulated as

$$\min_{[Z]_{(i,j)}\in\{0,1\},\forall(i,j)} \Big\{ \big\|(\boldsymbol{X}_1-\boldsymbol{X}_2)\text{Diag}\left(\text{vec}(\mathbf{1}_{L_1}\mathbf{1}_{L_2}^T-\boldsymbol{Z})\right)\big\|_F^2 + \eta\mathbf{1}_{L_1}^T\boldsymbol{Z}\mathbf{1}_{L_2} + \lambda\text{REG}(\boldsymbol{Z})\Big\}. \quad (1)$$

The first and second terms are the data-fitting DF$(\cdot)$, minimizing the difference of the bi-temporal unchanged pixels and constraining the area of the changed pixels, respectively. The last term REG$(\cdot)$ is the regularizer describing the characteristic of the change map to mitigate the ill-posedness.

While the data-fitting domain in (1) introduces non-convexity, preventing from the adoption of highly robust convex solvers (e.g., ADMM) to implement the algorithm. A convex relaxation is needed to design a convex criterion with convergence guaranteed. Therefore, the constraint can be relaxed as $\boldsymbol{Z} \in \text{conv}\left\{\{0,1\}^{L_1\times L_2}\right\} = [0,1]^{L_1\times L_2}$, which is its convex hull, the continuous probability domain. Then the HCD criterion can be rewritten into a convex form as

$$\min_{\boldsymbol{z}\in[0,1]^L} \|(\boldsymbol{X}_1-\boldsymbol{X}_2)\text{Diag}\left(\mathbf{1}_L-\boldsymbol{z}\right)\|_F^2 + \eta\mathbf{1}_L^T\boldsymbol{z} + \lambda\text{REG}(\boldsymbol{Z}), \quad (2)$$

where $\boldsymbol{z} = \text{vec}(\boldsymbol{Z})$. Though the formulation of the HCD problem has been represented as a convex form, which can be solved by convex optimizers with convergence guaranteed, the large-scale optimization procedure with complex mathematics introduced by the conventional regularizers (e.g., TV) is still time-consuming.

B. Small-Data Learning Criterion via CODE Theory

In order to solve the HCD problem under a limited training sampling rate with high computational efficiency, CODE theory is introduced to rule out the sophisticated mathematics in the regularization design. CODE, recently developed for hyperspectral inverse imaging problems, judiciously avoids sophisticated mathematics by combining DE into CO criterion [19]. Specifically, a simple convex form regularizer is designed via $\boldsymbol{Q}$-quadratic norm utilizing deep information embedded in small training data extracted by a simple network structure. The general criterion of the CODE theory can be formulated as

$$\boldsymbol{z}^\star = \arg\min_{\boldsymbol{z}} \ \text{DF}(\boldsymbol{z}) + \frac{\lambda}{2}\|\boldsymbol{z}-\boldsymbol{z}_{\text{DE}}\|_{\boldsymbol{Q}}^2, \quad (3)$$

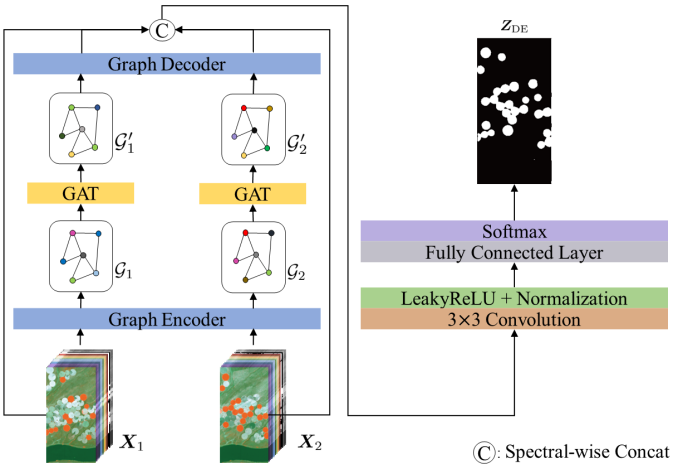


Fig. 1. Architecture of SGCDNet.

where the $\boldsymbol{Q}$-quadratic norm of the feature vector $\boldsymbol{v}$ is defined by $\|\boldsymbol{v}\|_{\boldsymbol{Q}}^2 := \boldsymbol{v}^T\boldsymbol{Q}\boldsymbol{v}$ [20]. Experiment evidence has suggested that the principal components analysis (PCA) is suitable for the design of the PSD matrix $\boldsymbol{Q}$ in $\|\cdot\|_{\boldsymbol{Q}}^2$ and has achieved state-of-the-art performance in the highly challenging hyperspectral inpainting problem [19]. For the HCD problem whose output is a gray-scale map, we design $\boldsymbol{Q}$ using the one-dimensional PCA, which is equivalent to the identity mapping $\boldsymbol{I}_L$. Then, the regularization term can be rewritten as

$$\text{REG}(\boldsymbol{z}) = \|\boldsymbol{z} - \boldsymbol{z}_{\text{DE}}\|_2^2. \tag{4}$$

Such a regularizer utilizes useful information in the training samples and guides the data-fitting DF$(\cdot)$ to select a suitable solution. To complete the small-data learning HCD algorithm based on the CODE theory, a simple graph change detection network (SGCDNet) is needed to be designed to acquire the rough solution $\boldsymbol{Z}_{\text{DE}}$, which will be detailed in Section II-C.

C. SGCDNet Design

As mentioned in Section II-B, a neural network is necessary for CODE theory to extract the deep information embedded in the labeled samples. To be more specific, we design a simple graph change detection network SGCDNet f_θ, where θ denotes the network parameters, to acquire the rough estimate $\boldsymbol{Z}_{\text{DE}} = f_\theta(\boldsymbol{X}_1, \boldsymbol{X}_2)$. By observation, the changed pixels tend to form small regions, which motivates us to study changed areas at the superpixel level. We believe that the relationship between the changed area and its neighbors will change when the region is changed across different dates. Therefore, the graph attention network (GAT) module [21], which automatically measures the importance of the neighboring superpixels by the attention mechanism, is adopted in SGCDNet.

The network structure of SGCDNet is shown in Fig. 1. In order to study at the superpixel level and reduce the data volume, the bi-temporal HSIs are firstly divided into superpixels by the simple linear iterative clustering (SLIC) method [22]. Linear discriminant analysis (LDA) [23] is applied to achieve dimension reduction before the superpixel segmentation. Then the graph $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}),\ \forall i \in \mathcal{I}_2$, representing the bi-temporal HSIs, can be constructed by graph encoder [24], where $\mathcal{V}$ and $\mathcal{E}$ denote the nodes and edges, respectively. After constructing the graph, a single-head GAT layer [21] is adopted to extract the relationship between first-order neighborhood superpixels of $\mathcal{G}_1$ and $\mathcal{G}_2$. Then the feature graphs $\mathcal{G}_1'$ and $\mathcal{G}_2'$ are converted back to pixel level by graph decoder [24] and concatenated along the spectral dimension. To extract features at the pixel level, a convolution layer containing a 3×3 kernel with the Leaky Rectified Linear Unit (LeakyReLU) activation function is employed. Finally, the change map $\boldsymbol{Z}_{\text{DE}}$ is detected by a fully connected layer followed by a softmax function. The loss function of SGCDNet is designed based on cross-entropy, which is defined as

$$loss(\boldsymbol{Z}_{\text{GT}}, \boldsymbol{Z}_{\text{DE}}) = -\frac{1}{N}\sum_{i=1}^{N}\left([\boldsymbol{z}_{\text{GT}}]_i \log\left([\boldsymbol{z}_{\text{DE}}]_i\right)\right), \tag{5}$$

where N denotes the number of training pixels, and $[\boldsymbol{z}_{\text{GT}}]_i$ and $[\boldsymbol{z}_{\text{DE}}]_i$ represent the ith training pixels from the ground truth $\boldsymbol{Z}_{\text{GT}}$ and the network output $\boldsymbol{Z}_{\text{DE}}$, respectively.

D. Small-Data Learning HCD Algorithm and Implementation

By combining the convex deep regularizer (4) into the convex HCD criterion (2), the small-data learning HCD criterion can be written as

$$\min_{\boldsymbol{z}\in[0,1]^L} \|(\boldsymbol{X}_1 - \boldsymbol{X}_2)\text{Diag}\left(\boldsymbol{1}_L - \boldsymbol{z}\right)\|_F^2 + \eta\boldsymbol{1}_L^T\boldsymbol{z} + \lambda\|\boldsymbol{z} - \boldsymbol{z}_{\text{DE}}\|_2^2. \tag{6}$$

Note that both the data-fitting term and the regularization term are convex, leading to a convex criterion that can be solved by convex optimizers, e.g. ADMM. Then, we first define $\boldsymbol{Y} := \boldsymbol{X}_1 - \boldsymbol{X}_2 \equiv [\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(M)}]^T$, $\Psi := [\text{Diag}(\boldsymbol{y}^{(1)}), ..., \text{Diag}(\boldsymbol{y}^{(M)})]^T$, and $\psi := \text{diag}(\boldsymbol{Y}^T\boldsymbol{Y})$. The first term of criterion (6) can be rewritten as $\|(\boldsymbol{X}_1 - \boldsymbol{X}_2)\text{Diag}\left(\boldsymbol{1}_L - \boldsymbol{z}\right)\|_2^2 = \|\boldsymbol{Y}\|_F^2 + \|\Psi\boldsymbol{z}\|_2^2 - 2\psi^T\boldsymbol{z}$. To this end, we reformulate criterion (6) into general ADMM form as

$$\begin{aligned}&\min_{\boldsymbol{z}} \frac{1}{2}\|\Psi\boldsymbol{z}_1\|_2^2 + \left(\frac{\eta}{2}\boldsymbol{1}_L - \psi\right)^T\boldsymbol{z}_2 + \frac{\lambda}{2}\|\boldsymbol{z}_3 - \boldsymbol{z}_{\text{DE}}\|_2^2 + I_{\mathcal{Z}}(\boldsymbol{z}_4),\\ &\text{s.t. } \boldsymbol{z} = \boldsymbol{z}_i,\ i = 1, \ldots, 4,\end{aligned} \tag{7}$$

where the indicator function $I_{\mathcal{Z}}(\boldsymbol{z}_4)$ has value 0 if $\mathcal{Z} \in [0,1]^L$ (otherwise, ∞). Before solving criterion (7) via ADMM, we first write down the augmented lagrangian

$$\begin{aligned}&\mathcal{L}(\boldsymbol{z}, \{\boldsymbol{z}_i\}_{i=1}^4, \{\boldsymbol{d}\}_{i=1}^4)\\ &= \frac{1}{2}\|\Psi\boldsymbol{z}_1\|_2^2 + \left(\frac{\eta}{2}\boldsymbol{1}_L - \psi\right)^T\boldsymbol{z}_2 + \frac{\lambda}{2}\|\boldsymbol{z}_3 - \boldsymbol{z}_{\text{DE}}\|_2^2 + I_{\mathcal{Z}}(\boldsymbol{z}_4)\\ &+ \sum_{i=1}^{4}\frac{\mu}{2}\|\boldsymbol{z} - \boldsymbol{z}_i - \boldsymbol{d}_i\|_2^2,\end{aligned}$$

where $\mu > 0$ is the penalty parameter and $\boldsymbol{d}_i \in \mathbb{R}^L, \forall i \in \mathcal{I}_4$, are the scaled dual variable. Based on the ADMM theory [15],

the global optimal solution $\boldsymbol{z}^\star$ can be obtained by iteratively updating the following subproblems until convergence:

$$\begin{aligned}\{\boldsymbol{z}_i^{q+1}\}_{i=1}^4 &= \arg\min_{\{\boldsymbol{z}_i\}} \mathcal{L}(\boldsymbol{z}^q, \{\boldsymbol{z}_i\}_{i=1}^4, \{\boldsymbol{d}^q\}_{i=1}^4),\\ \boldsymbol{z}^{q+1} &= \arg\min_{\boldsymbol{z}} \mathcal{L}(\boldsymbol{z}, \{\boldsymbol{z}_i^{q+1}\}_{i=1}^4, \{\boldsymbol{d}^q\}_{i=1}^4),\\ \boldsymbol{d}_i^{q+1} &= \boldsymbol{d}_i^q + \boldsymbol{z}_i^{q+1} - \boldsymbol{z}^{q+1},\ \forall i \in \mathcal{I}_4,\end{aligned}$$

where the subscript q is the iteration number. The closed-form solutions of the subproblems can be easily derived as

$$\begin{aligned}\boldsymbol{z}_1^{q+1} &= \mu[(\|\boldsymbol{y}_1\|_2^2+\mu)^{-1},\ldots,(\|\boldsymbol{y}_L\|_2^2+\mu)^{-1}]^T \odot (\boldsymbol{z}^q - \boldsymbol{d}_1^q),\\ \boldsymbol{z}_2^{q+1} &= \frac{1}{\mu}(\frac{\eta}{2}\mathbf{1}_L - \psi) + \boldsymbol{z}^q - \boldsymbol{d}_2^q,\\ \boldsymbol{z}_3^{q+1} &= \frac{1}{\mu}\boldsymbol{z}_{\mathrm{DE}} + \frac{1}{\lambda}(\boldsymbol{z}^q - \boldsymbol{d}_3^q),\\ [\boldsymbol{z}_4^{q+1}]_\ell &= \begin{cases} [\boldsymbol{z}^q - \boldsymbol{d}_4^q]_\ell, \text{ if } [\boldsymbol{z}^q - \boldsymbol{d}_4^q]_\ell \in [0,1],\\ 0, \text{ if } [\boldsymbol{z}^q - \boldsymbol{d}_4^q]_\ell < 0,\\ 1, \text{ otherwise,}\end{cases} \quad \forall \ell \in \mathcal{I}_L\ ,\\ \boldsymbol{z}^{q+1} &= \frac{1}{4}\sum_{i=1}^4 (\boldsymbol{z}_i^{q+1} + \boldsymbol{d}_i^q).\end{aligned}$$

We have completed the proposed small-data learning hyperspectral change detection (SDL-HCD) algorithm, whose superiority will be experimentally demonstrated next in Section III.

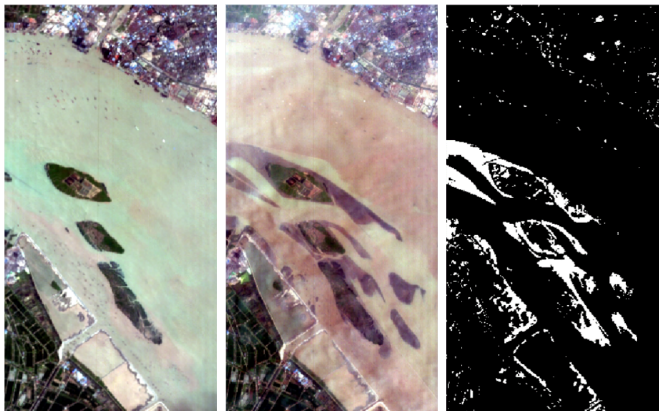
III. Experimental results

A. Experimental Setting

In this section, we compare the proposed SDL-HCD with benchmark HCD methods, including MAD [7], PCTD [12], CD-ADMM [13], CSANet [17], and HyperNet [18]. Due to the lack of labeled HSIs, the experiment conducted on real hyperspectral datasets under a limited labeled rate condition set as 0.1% demonstrates the performance of the HCD algorithms in a practical scenario. For MAD, PCTD, CD-ADMM, HyperNet, and the proposed SDL-HCD, the final output binary change map is generated by the k-means classifier. The penalty parameter μ is set as 0.001 in this paper; note that if $\mu > 0$, the convergence of the algorithm is theoretically guaranteed, which only affects the convergence speed without changing the performance. The parameters η and λ, which govern the strength of regularization, are subjected to experimentation across a spectrum of values $\{10^n \mid n = -2, -1, 0, 1, 2, 3\}$. For the real-data experiment in this paper, the parameter set $(\eta, \lambda) = (100, 100)$ is cherry-picked from this testing phase due to its stable performance across diverse datasets. The superpixel scale, which controls the number of superpixels, is set as 60 for SGCDNet. SGCDNet is trained using Adam optimizer [25] with an initial learning rate of 0.0001, a decay of 0.9 every 10 epochs, and the number of epochs set to 450.

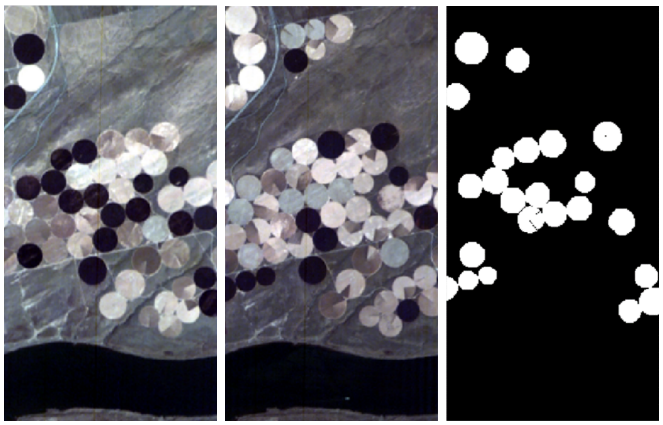
B. Datasets Description

In this paper, two widely used hyperspectral datasets, both containing two aligned bi-temporal HSIs and a ground-truth map marking the changed pixels and unchanged pixels with 1 and 0, respectively, are selected to investigate the efficacy of the HCD methods.



(a) (b) (c)

Fig. 2. Pseudo-color illustrations of the River dataset. (a) HSI acquired on May 3rd, 2013. (b) HSI acquired on December 31st, 2013. (c) Ground-truth change map.



(a) (b) (c)

Fig. 3. Pseudo-color illustrations of the Hermiston dataset. (a) HSI acquired on May 1st, 2004. (b) HSI acquired on May 8th, 2007. (c) Ground-truth change map.

- The River dataset with a dimension of $L = 463 \times 241$ pixels and 242 spectral bands, ranging from 350 to 2580 nm, was captured by the Hyperion sensor over Jiangsu province, China, on May 3rd, 2013, and December 31st, 2013, respectively [16]. After removing low-quality bands, 128 high SNR bands are used in this paper. The main change area is focus on substance in the river. The pseudo-color images of the bi-temporal HSIs and the ground-truth map are shown in Fig. 2.
- The Hermiston dataset with a dimension of $L = 390 \times 200$ pixels and 242 spectral bands, ranging from 350 to 2580 nm, was captured by the Hyperion sensor over Hermiston city in Umatilla County, on May 1st, 2004, and May 8th, 2007, respectively [12]. After removing low-quality bands, 154 high SNR bands are used in this paper. The changed areas of this dataset are primarily rural change, including crops, soil, and water. The pseudo-color images of the bi-temporal HSIs and the ground-truth map are shown in Fig. 3.

C. Quantitative Assessment

To fairly compare the performance of the HCD methods, three representative quantitative assessments, i.e., overall accuracy (OA), kappa coefficient κ, and F1-score (F1), are adopted. Before introducing the quantitative indices, we first define TP, FP, TN, and, FN standing for true positives, false positives, true negatives, and false negatives, respectively. OA is the total accuracy of all the classified pixels and can be represented as

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}.$$

Kappa coefficient κ is widely used to measure the classification results, which takes into account the inter-rater reliability, and provides a more robust evaluation than OA. We first define the hypothetical probability of chance agreement P_e as

$$\mathrm{P}_e = \frac{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FN} + \mathrm{TN})(\mathrm{FP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})^2}.$$

Then, κ can be written as

$$\kappa = \frac{\mathrm{OA} - \mathrm{P}_e}{1 - \mathrm{P}_e}.$$

F1 is a widely used metric combining precision and recall indicators. The precision indicator is the ability of the algorithm to correctly identify positive pixels out of all predicted positive pixels, which can be defined as

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

On the other hand, the recall indicator is the ability of the algorithm to correctly identify positive pixels out of all actual positive pixels, and can be defined as

$$\text{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

F1-score is the harmonic mean of Pr and Re, particularly useful when dealing with imbalanced datasets, and can be written as

$$\mathrm{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Computational time (measured in seconds) is used to measure the computational efficiency. In this paper, all the experiments are conducted on the computer facility equipped with NVIDIA RTX 3090 GPU, AMD Ryzen 9 5950X CPU with 3.40-GHz speed, and 24-GB random access memory. All the DE-based algorithms are implemented on Python 3.10.9 and PyTorch 2.0.0. On the other hand, mathematical-based algorithms are performed on MATLAB R2023a.

D. HCD Performance Under Limited Labeled Samples

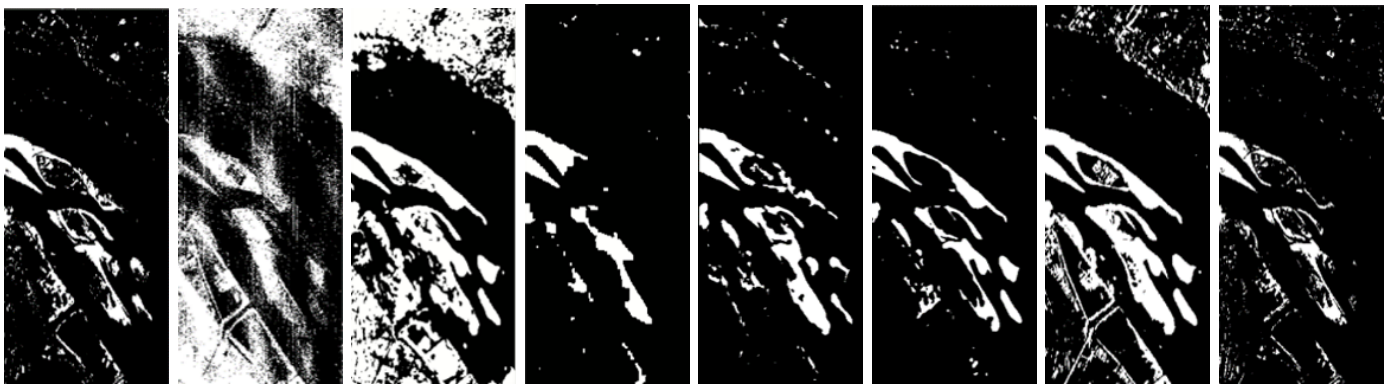
Although DE-based algorithms trained by sufficient data have demonstrated their effectiveness in HSI imaging tasks, acquiring hyperspectral big data is very costly in practice, promoting the development of small-data learning and unsupervised learning. In light of this, we investigate the capability of the HCD algorithms, including unsupervised, self-supervised, and semi-supervised ones, under a limited sampling ratio scenario.

TABLE I
Quantitative comparison on the River hyperspectral datasets.

Method	OA↑	κ ↑	F1↑	time (sec.)
MAD	0.636	0.126	0.251	0.75
PCTD	0.709	0.263	0.367	240.05
CD-ADMM	0.950	0.637	0.663	411.90
CSANet	0.938	0.594	0.626	96.27
HyperNet	0.933	0.587	0.623	40.05
SGCDNet	0.815	0.365	0.446	33.67
SDL-HCD	**0.966**	**0.789**	**0.808**	34.12

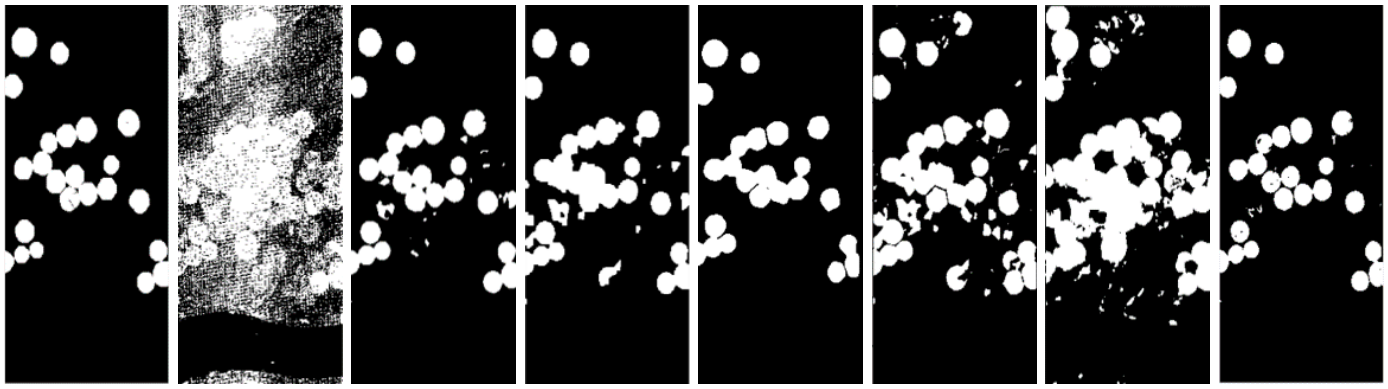
1) Experiment on the River Dataset: The output binary change maps of the HCD algorithms are shown in Fig. 4, and the quantitative comparison is shown in Table I. As shown in Fig. 4(b), MAD is able to detect the changed pixels in the center of the river; however, there are too many false positives, especially at the river bank. MAD only relies on correlation analysis and does not make use of advanced spectral and spatial characteristics. This false-positives phenomenon also happens in Fig. 4(c), the change map of PCTD, which uses tucker decomposition as a high-order principle component analysis (PCA) to eliminate spectral variations between unchanged pixels. In the detection process, the tensor-based algorithm PCTD comes with a slow computational time, as shown in Table I. The CO-based algorithm CD-ADMM is able to avoid false positives at the river bank and detect most of the changed pixels, while the details of the changed regions on the river bank are lost, which is illustrated in Fig. 4(d). The performance is also reflected in the quantitative indices shown in Table I, achieving the second on all three indicators. However, the TV regularizer in CD-ADMM not only reduces the capacity to detect scattered distributed changed pixels but causes a time-consuming computational procedure, especially in handling the large-scale hyperspectral optimization problem. Though under the situation of limited training samples, DE-based algorithms are able to generate finer details of changed areas, thanks to the ability to learn deep features from labeled data. CSANet and HyperNet have decent performance, but the details are not well-detected due to the lack of training samples, as shown in Table I and Figs. 4(e) and (f). Both CSANet and HyperNet exploit spectral and spatial attention of the bi-temporal HSIs; however, CSANet further utilizes temporal attention to focus on the temporal variations, leading to a more precise detection than HyperNet. The change map of the simple network structure SGCDNet is shown in Fig. 4(g). SFGCDNet can detect the approximate changed regions but loses the details of the change map, owing to its simple structure. Though the output of SGCDNet is a rough solution, SDL-HCD can still utilize the useful information embedded in the rough estimate and achieve state-of-the-art performance, as shown in Table I. Note that the additional computational time brings by the optimization procedure of SDL-HCD is within one second, thanks to the judicious design of the deep regularization with a simple convex form.

2) Experiment on the Hermiston Dataset: The output binary change maps of the HCD algorithms are shown in Fig. 5, and



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. Change detection results of the CD algorithms on the River dataset. (a) Ground Truth. (b) MAD. (c) PCTD. (d) CD-ADMM. (e) CSANet. (f) HyperNet. (g) SGCDNet. (h) SDL-HCD.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. Change detection results of the CD algorithms on the Hermiston dataset. (a) Ground Truth. (b) MAD. (c) PCTD. (d) CD-ADMM. (e) CSANet. (f) HyperNet. (g) SGCDNet. (h) SDL-HCD.

TABLE II
Quantitative comparison on the Hermiston hyperspectral datasets.

Method	OA↑	κ ↑	F1↑	time (sec.)
MAD	0.609	0.215	0.375	0.61
PCTD	0.976	0.897	0.911	100.89
CD-ADMM	0.960	0.838	0.861	143.82
CSANet	0.980	0.915	**0.927**	84.37
HyperNet	0.893	0.638	0.697	25.99
SGCDNet	0.927	0.736	0.776	23.45
SDL-HCD	**0.982**	**0.916**	**0.927**	23.76

the quantitative comparison is shown in Table II. MAD tends to generate massive false positives despite the computational time being fast, as shown in Fig. 4(b). CD-ADMM and PCTD both detect most of the changed pixels, while the false detection of CD-ADMM is a little more than PCTD in the center region of the image, as shown in Figs. 4(c) and (d). The piecewise smoothness promotion in CD-ADMM may lead to the unclearly separated edges of changed and unchanged regions, and cause a computational time about six times more than the proposed SDL-HCD, as demonstrated in Table II. As shown in Fig. 5(e), CSANet can detect the changed areas with the only defect that the boundary of the changed region is unclear. HyperNet generates false positives, as illustrated in Fig. 5(f), mainly due to the self-supervised training scheme. As shown in Fig. 5(g), there are massive false positives in the change map of SGCDNet, leading to lower κ and F1 values. SDL-HCD successfully extracts useful information from the rough estimate produced by SGCDNet and achieves state-of-the-art with a computational time about four times quicker than CSANet.

IV. Conclusions

In this paper, we propose a small-data learning algorithm with an area-control mechanism to solve the HCD problem, lacking big data in practical scenarios, by introducing the small-data learning CODE theory [19], which has been applied to highly challenging hyperspectral inverse problems, such as hyperspectral inpainting. A simple graph learning network SGCDNet, which can learn information from limited labeled samples using just a low inference time, is proposed based on the nature of CODE theory that a simple network structure is sufficient to guide the convex criterion to select a brilliant solution. Then, a simple convex deep regularizer, judiciously designed based on the $\boldsymbol{Q}$-quadratic norm function, trickly rules out the large-scale optimization procedure commonly coming with the traditional regularization term (e.g., TV) and successfully distills useful information from the roughly estimated solution generated by SGCDNet. Thanks to the

simple mathematical convex criterion, SDL-HCD can be implemented by the well-known convex solver ADMM to achieve a high computational algorithm. The proposed SDL-HCD outperforms benchmark methods in real-data experiments and shows its efficacy under a limited labeled sampling rate.

ACKNOWLEDGMENT

This study was supported partly by the EINSTEIN Program of National Science and Technology Council (NSTC), Taiwan, under Grant MOST 111-2636-E-006-028; and partly by the Emerging Young Scholar Program of NSTC, Taiwan, under Grant NSTC 112-2628-E-006-017. We thank the Center for Data Science at NCKU, and National Center for High-performance Computing (NCHC) for providing the computing resources.

REFERENCES

[1] J.-F. Mas, "Monitoring land-cover changes: A comparison of change detection techniques," *International Journal of Remote Sensing*, vol. 20, no. 1, pp. 139–152, 1999.

[2] O. Abd El-Kawy, J. Rød, H. Ismail, and A. Suliman, "Land use and land cover change detection in the western nile delta of egypt using remote sensing data," *Applied Geography*, vol. 31, no. 2, pp. 483–494, Apr. 2011.

[3] J. Sublime and E. Kalinicheva, "Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the tohoku tsunami," *Remote Sensing*, vol. 11, no. 9, p. 1123, May 2019.

[4] L. Yang, G. Xian, J. M. Klaver, and B. Deal, "Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data," *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 9, pp. 1003–1010, 2003.

[5] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, "Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 10, pp. 5530–5546, May 2015.

[6] C.-H. Lin, R. Wu, W.-K. Ma, C.-Y. Chi, and Y. Wang, "Maximum volume inscribed ellipsoid: A new simplex-structured matrix factorization framework via facet enumeration and convex optimization," *SIAM Journal on Imaging Sciences*, vol. 11, no. 2, pp. 1651–1679, Jun. 2018.

[7] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, Apr. 1998.

[8] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, Jan. 2007.

[9] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[10] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009.

[11] A. Ertürk, M.-D. Iordache, and A. Plaza, "Sparse unmixing with dictionary pruning for hyperspectral change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 321–330, Jan. 2017.

[12] Z. Hou, W. Li, and Q. Du, "A patch tensor-based change detection method for hyperspectral images," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Brussels, Belgium, Jul. 11–16, 2021, pp. 4328–4331.

[13] W.-C. Zheng, C.-H. Lin, K.-H. Tseng, *et al.*, "Unsupervised change detection in multitemporal multispectral satellite images: A convex relaxation approach," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, Jul. 28-Aug. 02, 2019, pp. 1546–1549.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, Jul. 2011.

[15] C.-Y. Chi, W.-C. Li, and C.-H. Lin, *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications*. CRC Press, Boca Raton, FL, 2017.

[16] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, Jan. 2019.

[17] R. Song, W. Ni, W. Cheng, and X. Wang, "CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, May 2022.

[18] M. Hu, C. Wu, and L. Zhang, "HyperNet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, Nov. 2022.

[19] C.-H. Lin, Y.-C. Lin, and P.-W. Tang, "ADMM-ADAM: A new inverse imaging framework blending the advantages of convex optimization and deep learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[21] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, Apr. 30-May 3, 2018.

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[23] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," in *Robust Data Mining*, New York, NY: Springer New York, 2013, pp. 27–33, ISBN: 978-1-4419-9878-1. DOI: 10.1007/978-1-4419-9878-1_4.

[24] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 1559–1572, Jan. 2022.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7-9, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Efficient Constraint-Aware Neural Architecture Search for Object Detection

Egor Poliakov* Wei-Jie Hung†and Ching-Chun Huang‡
* National Yang Ming Chiao Tung University, Taiwan
E-mail: poliakovegor.ee09@nycu.edu.tw Tel/Fax: +886-3-5712121#54736
† National Yang Ming Chiao Tung University, Taiwan
E-mail: imbigking12@gmail.com Tel/Fax: +886-3-5712121#54736
‡ National Yang Ming Chiao Tung University, Taiwan
E-mail: chingchun@nycu.edu.tw Tel/Fax: +886-3-5712121#54736

***Abstract*—We propose an efficient Neural Architecture Search (NAS) method, named Zero-DNAS, for object detection tasks capable of discovering a suitable architecture under given memory and FLOPs constraints. NAS aims to explore the search space automatically to discover the best-performing network architectures for a given task. However, NAS is resource-consuming and usually requires hundreds of hours of GPU computations to discover a neural network architecture with good performance. Especially for more extensive and complex computer vision tasks such as object detection, the computing time and memory when conducting NAS would increase dramatically. Practically, the conventional sampling-based NAS methods do not guarantee the best possible solutions, whereas differentiable methods require substantial memory resources, making it challenging to apply in macro search space settings. In comparison, we propose a differentiable NAS paradigm with zero-cost proxy metrics and aims to determine the architecture within the constraints of memory and FLOPS. The experiments on object detection datasets show that our proposed algorithm can discover more accurate and faster architectures in a heavy macro search space in less than 2 NVIDIA 2080TI GPU hours.**

I. Introduction

With the rapid progress in deep learning, numerous neural network architectures are required to satisfy particular demands in terms of performance and efficiency. Under the settings of a specific task, Neural Architecture Search (NAS) can relieve researchers from the complexities of architectural design by automatically exploring the search space to discover the best-performing network architecture. The solutions found must also be compact enough for practical scenarios, such as self-driving cars or other tracking[1][2] algorithms that require real-time inference on edge devices under hardware constraints. Hence, NAS algorithms must be hardware-aware and capable of searching for network architectures based on specific targets (i.e., hardware constraints), such as the number of floating-point operations per second (FLOPS) and the number of network parameters.

In addition, NAS is a computationally demanding technique that may require thousands of hours of GPU computations to discover a compact and objective neural network architecture. With recent advances in the NAS field, the time required for performing a search has been significantly reduced. One-shot sampling-based methods, such as [3]–[5], have substantially accelerated the NAS process by sharing weights across multiple candidates and avoiding training every architecture from scratch. However, such sampling-based algorithms still need to be faster and may lead to discovering suboptimal architectures because they treat NAS as a black-box optimization problem over a discrete sample domain, requiring numerous architecture evaluations to generate a solution.

Some studies propose differential NAS (DNAS)[6], [7] formulating the search problem in differentiabl way. Continuous relaxation of the search space allows gradient descent-based architecture discovery, reducing search time while maintaining performance. However, DNAS's fixed cell structure across all layers may not be optimal due to varied layer impacts[8]. Moreover, DNAS is usually unsuitable for a macro-architecture search space like FBNet[9] because parallel feature processing at every layer consumes considerable memory resources and substantially restricts its utility.

Recently, pruning at initialization techniques[10]merged into NAS, leading to new research field in NAS, Zero-Cost Proxies. Despite mathematical advances[11][12]. most focus on small spaces and simple tasks[10][13], and didn't prove the effectiveness on large search space and complex task like object detection.

This paper addresses the crucial and challenging task of developing an efficient and reliable NAS algorithm, specifically designed for demanding tasks such as object detection. Our proposed method builds upon the differentiable NAS paradigm, offering significant enhancements. Firstly, we incorporate zero-cost proxy performance metrics into a differentiable NAS algorithm, thereby eliminating the need for expensive network training. This design opens up the possibility for differentiable NAS to be applied in large macro-architecture search spaces, resulting in a greatly accelerated search process. Furthermore, we introduce a mechanism to meet user-defined computational and memory constraints, ensuring the discovered architecture is tailored to specific requirements. Lastly, we propose a novel layer-depth regulation strategy, forcing the algorithm to discover architectures that provide higher inference speeds.

The primary contributions of this work can be summarized

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

as follows:

- We introduce a novel zero-cost Differentiable Neural Architecture Search algorithm (Zero-DNAS), which is capable of efficiently discovering architectures for object detection tasks while keeping search time and memory consumption minimal. The resulting architectures exhibit superior performance and high inference speed.
- Zero-DNAS is specifically designed to satisfy the user-defined parameters and FLOPS constraints, ensuring the discovery of the most accurate model tailored to individual needs.
- We also propose a layer-depth constraint to ensure the final discovered architecture provides a rapid inference speed. Our findings indicate that maintaining a low layer count helps to discover faster models.

II. Related Work

This section discusses sampling-based search algorithms that treat the exploration procedure as a black-box optimization task over a discrete space. Moreover, we also detail differentiable solutions that address the search problem in a differentiable way using gradient descent. Also, we review the NAS algorithms specifically designed to handle object detection tasks.

A. Sampling-based search

In early Neural Architecture Search (NAS) studies like Zoph et al.[14], Reinforcement Learning (RL) methods were used with a recurrent neural network (RNN) controller for sampling and training candidate architectures. However, this approach proved computationally expensive and time-consuming due to training numerous candidate networks from scratch.

Efficient Neural Architecture Search (ENAS) [3] introduced a weight-sharing mechanism. The RL controller samples a candidate network, inheriting weights from a pre-existing larger network (supernet). After training the candidate network, weights and validation accuracy update the supernet's weights and RL controller.

However, ENAS[3] faces a weight coadaptation problem, leading to suboptimal results. To address this, SPOS[4] suggested a uniform sampling training procedure to replace the RL controller. Additionally, AttentiveNAS[15] improved the sampling procedure by considering the Pareto front of subnetwork performance. Notably, search algorithms in [4][15] use evolution algorithms, adding a stage to the overall architecture search process.

For enhanced efficiency in searching, Cream[16] proposed prioritized paths based on subnetwork accuracy during training. Real-time prioritization, considering performance and complexity, eliminates the need for complex RL or evolutionary searches.

B. Differentiable search

Contrary to traditional methods that approach architecture search as a non-differentiable problem involving a discrete search space, DNAS-based methods offer an innovative approach by framing the search process as a continuous, differentiable optimization problem. A pioneer work in this area, DARTS[6], utilized the concept of continuous relaxation to represent architectural configurations, thus enabling the application of gradient descent techniques for architecture search. However, DARTS[6] represents its architectures at the cell level and utilizes complex directed acyclic graphs for representation. Despite the flexibility, it has been shown to lead to efficiency issues[9]. Also, this cell-level structure tends to neglect the possibility of varied structures at different network depths.

In a different approach, Wu et al. [9] designed a layer-wise search space that incorporates a pre-determined macro-architecture. Here, each layer has the freedom to select distinct blocks, whose types are determined dynamically during the search process. This flexible architecture setup is advantageous for the discovery of more precise models. However, despite the impressive performance of DNAS methods in classification tasks, they still require training the network parameters, which is time-consuming, memory-inefficient, and challenging to apply in more complex tasks such as object detection.

C. Object detection NAS

Most NAS methods primarily focus on architecture design for image classification and directly apply the discovered configurations as backbones for object detection. However, this approach can lead to suboptimal outcomes due to a task gap. Recently, a subset of Neural Architecture Search (NAS) methods targeting object detection architectures has emerged, including [17][18][19][20][21]. These methods harness object detection loss to identify superior architectures, but they are computationally demanding, relying on powerful GPUs with substantial memory to complete the search.

OPANAS[17]proposes a complex feature pyramid network (FPN) search space involving a densely connected directed acyclic graph representing diverse FPN structures. However, the intricate FPN structure is unsuitable for real-time applications. NAS-FCOS[18] aims to discover architectures for FPNs and detection heads, inspired by the anchor-free object detector FCOS[22] using an RL paradigm.

In contrast, DetNAS[19] employs a sampling-based training and evolution algorithm to search for the detection backbone. Although both methods[18][19] achieve high accuracy, their focus is limited to specific parts of the detection module.

SM-NAS[20] strives to identify efficient module combinations and modular-level architectures for object detection. This method covers multiple stages in the object detection pipeline, including the backbone, feature fusion neck, RPN, and detection head. The authors introduce a two-stage structural-to-modular strategy to optimize module combinations and architecture.

III. Proposed Method

This section offers an in-depth explanation of our proposed search algorithm. We begin by defining the problem

TABLE I: The candidate configurations of each block within the proposed ScaledYOLOv4 [23] and YOLOv5 [24] search spaces. Each "backbone" CSP block (or C3 block for YOLOv5) has 12 distinct choices and each "neck" CSP2 block has 4 variations for choosing. Note that YOLOv4-CSP large is used to search for larger and accurate model while YOLOv4-CSP small is for smaller and compact model. Please refer to Fig. 2 for a detailed explanation of the channel split (γ) and depth (d).

Model part	Block type	YOLOv4-CSP large		YOLOv4-CSP small		YOLOv5	
		γ	d	γ	d	γ	d
Backbone	CSP/C3	0.25	2	0.25	0	0.25	0
Backbone		0.25	4	0.25	2	0.25	3
Backbone		0.25	6	0.25	4	0.25	6
Backbone		0.25	8	0.25	6	0.25	9
Backbone		0.50	2	0.50	0	0.50	0
Backbone		0.50	4	0.50	2	0.50	3
Backbone		0.50	6	0.50	4	0.50	6
Backbone		0.50	8	0.50	6	0.50	9
Backbone		0.75	2	0.75	0	0.75	0
Backbone		0.75	4	0.75	2	0.75	3
Backbone		0.75	6	0.75	4	0.75	6
Backbone		0.75	8	0.75	6	0.75	9
Neck	CSP2/C3	0.50	2	0.50	2	0.50	0
Neck		0.50	4	0.50	4	0.50	3
Neck		0.50	6	0.50	6	0.50	6
Neck		0.50	8	0.50	8	0.50	9

setting. Subsequently, we present the structure search space, encompassing the macro-architecture and the possible block configurations. After that, we detail the search process in our proposed Zero-DNAS algorithm. Lastly, we elaborate on the hardware constraints and our newly proposed depth-aware loss.

A. Problem definition

The overall objective of the classical NAS problem is formulated as follows:

$$\min_{\alpha \in A} \min_{w_\alpha} L(\alpha, w_\alpha) \tag{1}$$

where A is a search space and α is an architecture candidate from the search space. The goal is to determine an architecture α that yields the best performance after training its weights w_α. In most cases, the main bottleneck in computation time and memory capacity is the inner part of network weight training; hence, the focus is on optimizing this part of the problem. To validate these ideas, we designed a search space, which is introduced in the following subsection.

B. Search space

Besides the cell-level search spaces [26], [27], this work employs a layer-wise search space with a predetermined macro-architecture. This architecture is segmented into stages, with each stage aligning with a layer. We demonstrate our algorithm's versatility and draw on prior handcrafted models by independently crafting search spaces for ScaledYOLOv4[23] and YOLOv5[24].

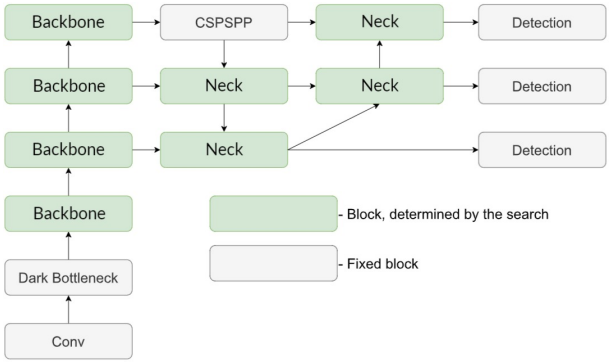


Fig. 1: The layer-wise macro-architecture search space for ScaledYOLOv4 [23], where NAS determines the design of the eight searchable (highlighted in green) blocks. Each block (also known as a layer or stage) is optimized to find a suitable network configuration from the available candidate sets. Refer to Fig. 2 for a detailed structure of the searchable block.

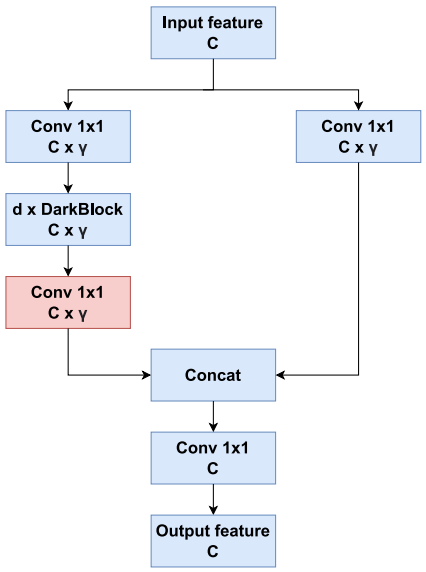


Fig. 2: The illustration of the CSP structure used in ScaledYOLOv4 search space. Here, C is the channel number, while γ is the reduction ratio of the channel for two parallel branches. Also, d determines the number of Darknet53 [25] modules used to extract more informative features. Note that the C3 block used in the YOLOv5 search space has a similar structure, except it does not include the module depicted in red.

Fig.1 introduces the ScaledYOLOv4[23] macro-architecture search space. YOLOv5[24] macro-architecture isn't depicted due to space limits, but both have 8 searchable stages (4 for backbone and 4 for neck blocks). Our algorithm seeks the optimal block structure in each green stage, while the gray stages remains fixed. The structure of the green block with adjustable parameters is shown in Fig. 2. As outlined in Table I, each backbone block offers 12 candidate configurations, and each neck block has 4 choices. The proposed object detector architecture search space comprises 8 layers, each selecting

a candidate block. This yields approximately $5 \cdot 10^6$ possible architectures (i.e., $12^4 \cdot 4^4$) for both ScaledYOLOv4[23] and YOLOv5[24] search spaces, making architecture optimization extremely demanding.

C. Search procedure

1) Preliminary of DNAS: We adopt a DNAS approach to solve Eq. (1). At each layer, a block b_l is independently sampled from a set of candidate blocks $b_{l,i}$ according to the probability $P_{\theta_l}(b_l = b_{l,i})$ (which we aims to optimize). The probability of sampling a complete architecture α can be computed as

$$P_\theta(\alpha) = \prod_l P_{\theta_l}(b_l = b_{l,i}). \tag{2}$$

To make the procedure differentiable, we use Gumbel softmax[28], introducing Gumbel random noise $g_{l,i}$ to $\theta_{l,i}$. θ represents an architecture importance matrix, $\theta_{l,i}$ signifies the importance of choosing the ith block at layer l, and τ controls Gumbel softmax softness. Throughout training, τ decreases to reach a one-hot value (selecting one configuration per block). $P_{\theta_l}(b_l = b_{l,i})$ is denoted $m_{l,i}$ in Eq. (3) for symbol simplification.

$$\begin{aligned} P_{\theta_l}(b_l = b_{l,i}) = m_{l,i} &= \text{GumbelSoftmax}(\theta_l) \\ &= \frac{\exp\left[(\theta_{l,i} + g_{l,i})/\tau\right]}{\sum_i \exp\left[(\theta_{l,i} + g_{l,i})/\tau\right]}. \end{aligned} \tag{3}$$

To facilitate forward and backward propagation of DNAS (Fig. 3), DNAS generates the feature map, x_{l+1}, for layer $l+1$ by summing weighted feature maps from different candidate blocks $b_{l,i}(x_l)$. This can be represented as:

$$x_{l+1} = \sum_i m_{l,i} \cdot b_{l,i}(x_l). \tag{4}$$

With these modifications, the DNAS problem is relaxed to learn the probability distribution P_θ of the block importance parameter θ and the network parameters w_α in the supernet (which combines all block configurations), aiming to minimize expected loss. This leads to an optimization problem formulated in the following:

$$\min_\theta \min_{w_\alpha} E_{\alpha \sim P_\theta} L(\alpha, w_\alpha). \tag{5}$$

However, the training procedure to determine network weights (w_α) is time-consuming and limits the size of candidate pools in DNAS methods. Also, in DNAS settings, all candidate blocks are trained simultaneously, making it require substantial memory resources and infeasible on a typical GPU device.

2) Zero-DNAS algorithm: To mitigate the issues of training time and memory consumption, we propose an innovative solution to the problem defined in (5). We score each block configuration within the supernet using a zero-cost metric, WOT (WithOut Training) [29]. Intuitively, a superior network architecture for downstream tasks should effectively differentiate between two different input images, even when using randomly assigned network weights. Under the assumption, WOT assesses the similarity between the activation outputs of a **randomly initialized** network subjected to varying inputs.

By integrating WOT into the DNAS process, we establish our Zero-DNAS algorithm which eliminates the costly inner optimization (i.e., finding the optimized w_α) and substantially reduce the search time and memory cost. Compared to the original NAS problem (equation (1)), after employing the WOT metric, our new optimization objective aims to learn the distribution of the architecture importance parameters(P_θ) and maximize the following expected score:

$$\max_\theta E_{\alpha \sim P_\theta} WOT(\alpha) \tag{6}$$

The WOT metric for each possible block wot_l, i is computed at the network initialization stage and requires only one forward propagation pass through the randomly initialized network. Thus, our method makes the search space for DNAS more flexible and allows the use of larger block configuration pools.

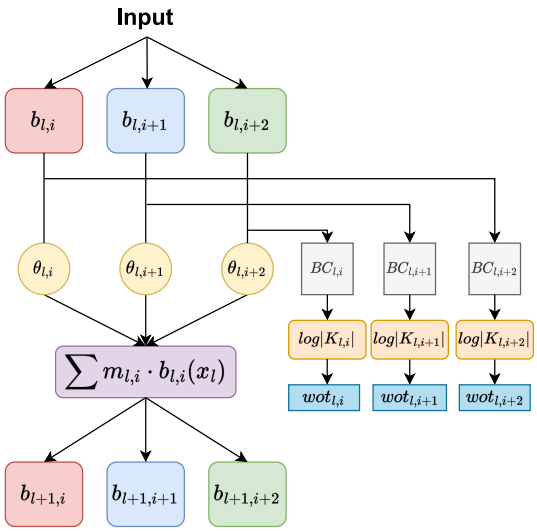


Fig. 3: The illustration of the block score $wot_{l,i}$ computation. First, activation patterns from a block $b_{l,i}$ are represented as binary codes ($BC_{l,i}$). Then, the correlation matrix ($K_{l,i}$) is formed based on the Huffman distances between the binary codes. Ultimately, the $wot_{l,i}$ is computed by taking the logarithm of the derivative of $K_{l,i}$.

At the beginning of the search procedure, we set all architecture importance parameters (i.e., θ) the same for every block. Next, we performed one forward pass through the supernet, computed the block WOT score $wot_{l,i}$ for every $b_{l,i}$ (Fig. 3) and stored them in a look-up table. Specifically, to compute $wot_{l,i}$, we binarize block activation outputs of k input images into k long binary vectors. Next, we build a $k \times k$ correlation matrix, $K_{l,i}$, for each block by calculating the Hamming distances among those binary vectors. The lower the correlation of the matrix, the better the performance of a candidate architecture is expected. Thus, $wot_{l,i}$ is computed by taking the logarithm of the derivative of $K_{l,i}$. Finally, the

WOT metric of the whole supernet is formulated as follows:

$$WOT(\alpha) = \sum_{l}\sum_{i} m_{l,i} \cdot wot_{l,i}$$
$$\text{where} \quad wot_{l,i} = log|K_{l,i}|. \tag{7}$$

Finally, the importance parameter $\theta_{l,i}$ is updated by backward propagation using the gradient of $WOT(\alpha)$ w.r.t $\theta_{l,i}$, which flows through the $m_{l,i}$ continuous variable in (3).

D. Hardware constraints

The optimization of Eq. (6) focuses solely on accuracy. However, the complexity of the searched architecture must also align with the target device. Thus, we introduced two terms to align the search algorithm with specified hardware constraints. We employed FLOPS count for computation cost and parameter count for memory footprint. The FLOPS loss is formulated by using mean squared error (MSE):

$$L_{flops}(\alpha) = (f_t - \log(f(\alpha)^{\beta}))^2 \tag{8}$$

where f_t is the specified target FLOPS, $f(\alpha)$ denotes the overall FLOPS of the model, and β is a hyper-parameter.

Similarly, the parameter loss term is computed as follows:

$$L_{params}(\alpha) = (p_t - \log(p(\alpha)^{\lambda}))^2 \tag{9}$$

where p_t corresponds to the target parameter number, $p(\alpha)$ represents the overall number of parameters, and λ is a hyper-parameter. Here, $f(\alpha)$ and $p(\alpha)$ are computed as

$$f(\alpha) = \sum_{l}\sum_{i} m_{l,i} \cdot f_{l,i} \tag{10}$$

$$p(\alpha) = \sum_{l}\sum_{i} m_{l,i} \cdot p_{l,i} \tag{11}$$

, where the precomputed $p_{l,i}$ and $f_{l,i}$ correspond to the parameters and FLOPS of a block $b_{l,i}$, respectively.

E. Depth-aware loss

Although the number of FLOPS correlates with model efficiency, architectures with fewer FLOPS don't always equate to faster inference speed. We find shallow and wide models are usually faster than their deeper and narrower counterparts. Thereby, we incorporate the following depth-aware loss into the optimization procedure to favor computationally efficient designs.

$$L_{depth}(\alpha) = \sum_{l}\sum_{i} d(\alpha) \cdot \omega \tag{12}$$

$$d(\alpha) = \sum_{l}\sum_{i} m_{l,i} \cdot d_{l,i} \tag{13}$$

where ω is a hyper-parameter to control this term's influence, $d_{l,i}$ denotes the depth of i^th candidates block in l^th layer.

In the end, the final loss function for optimization takes the following form:

$$L(\alpha) = -WOT(a) \cdot \psi + L_{flops}(\alpha) + L_{params}(\alpha) + L_{depth}(\alpha). \tag{14}$$

Also, the original objective function in (6) is replaced by the constraint-aware optimization problem, defined as

$$\min_{\theta} E_{\alpha \sim P_\theta} L(\alpha). \tag{15}$$

Algorithm 1 outlines the step-by-step search process. The searching algorithm takes the following items as the inputs: (a) the target/hardware constraints for FLOPS (f_t) and parameters (p_t), (b) the pre-defined searching space defining the number of searchable layers and the candidate network blocks for each layer, and (c) the pre-calculated metrics dictionary, $dict = (F_s, P_s, D_s, W_s) = (\{f_{l,i}\}, \{p_{l,i}\}, \{d_{l,i}\}, \{wot_{l,i}\})$, including the cost sets from the FLOPS constraints, the parameter constraints, the depth constraints, and the WOT values. It finally outputs the most fitting neural architecture as a result.

During the search process, the metrics for the whole supernet are computed (lines 4–8 of Algorithm 1). Next, the overall loss is calculated (lines 9–13 of Algorithm 1), and updating architecture importance parameters θ (line 14). Finally, the best architecture is acquired by sampling the blocks at every layer according to the θ distribution (line 18).

Algorithm 1 The proposed architecture search

Input: Target constraints: f_t, p_t; Pre-defined Searching Space: (l, i) indicates the ith candidate network block in the lth layer; Initialize equal importance θ for all candidate blocks; Pre-calculated Metrics dictionary: $dict = (F_s, P_s, D_s, W_s) = (\{f_{l,i}\}, \{p_{l,i}\}, \{d_{l,i}\}, \{wot_{l,i}\})$; Temperature τ;
Output: Suitable architecture (α^*)

1: **for** $epoch$ in $epochs$ **do**
2: **for** $iter$ in $maxIter$ **do**
3: $f, p, WOT, d \leftarrow 0$
4: $m \leftarrow computeM(\theta, \tau)$ (3)
5: $f \leftarrow computeFlops(F_s, m)$ (10)
6: $p \leftarrow computeParams(P_s, m)$ (11)
7: $WOT \leftarrow computeWot(W_s, m)$ (7)
8: $d \leftarrow computeDepth(D_s, m)$ (13)
9: $LossF \leftarrow computeLossFlops(f, f_t)$ (8)
10: $LossP \leftarrow computeLossParams(p, p_t)$ (9)
11: $LossWOT \leftarrow -WOT \cdot \psi$
12: $LossD \leftarrow computeLossDepth(d)$ (12)
13: $Loss \leftarrow computeLoss(LossF, LossP, LossWOT, LossD)$ (14)
14: $\theta \leftarrow updateTheta(\theta, Loss)$
15: **end for**
16: reduce τ
17: **end for**
18: $\alpha^* \leftarrow sampleAccordingTo(\theta)$ (6)

IV. Experimental Results

This section demonstrates the experimental results and compares the proposed method with previous NAS algorithms under different object detection architectures. The detection

performance is evaluated using multiple datasets, which are described in the first subsection. Next, we provide details about the evaluation metrics used in the experiments and the comparison. Following that, we present the implementation details for the search method and hyper-parameters used in the experiments. Finally, we showcase the comparison results and ablation studies, demonstrating that the proposed method can discover hardware-aware architectures using significantly less GPU time and memory resources.

A. Datasets & evaluation metrics

To prove these concepts, we conducted experiments on two benchmark datasets for object detection tasks: "Common Objects in Context (MSCOCO)" and "VOC". The MSCOCO dataset includes images of various scenes containing 80 classes of objects in their natural environments. Overall, it consists of 118,000 images in the training set and 20,000 images for testing. The VOC dataset with 20 classes, includes 5,000 images for the training and testing set respectively.

To evaluate the proposed algorithm and compare it with other approaches, we utilized the overall search time to assess the efficiency of NAS methods. Additionally, we employed mean average precision (AP) and average precision at $IoU = 50$ (AP50) to measure the detection performance. To estimate the efficiency of a network architecture, we calculated three metrics: (a) the number of FLOPS for evaluating computational cost, (b) the number of parameters for memory-cost evaluation, and (c) the number of frames per second (FPS) that the model can process, serving as a metric for inference speed.

B. Implementation details

For Zero-DNAS implementation, we used these hyper-parameters: $\beta = 0.01$, $\gamma = 0.03$, $\psi = 0.01$, $\omega = 0.005$, $\tau = 3$, and a τ decay of 0.065. During the search process, the Mish and Swish activations were substituted with ReLU to compute the score, as this has a minimal impact on network representation power [30]. The search ran for 40 epochs, 2,880 iterations per epoch, with τ decreasing every epoch.

Next, we constructed model architectures based on the search results and trained them from scratch to estimate their final performance. Following the search phase, we conducted model training for 300 epochs using the stochastic gradient descent optimizer with a momentum of 0.937 and a learning rate of 0.01. Moreover, we followed a cosine annealing schedule and set the weight decay to 0.0005. The same data augmentation techniques employed in the ScaledYOLOv4 and YOLOv5 algorithms were applied to train models derived from various search spaces. For estimating model inference speeds, we executed the models on the NVIDIA 2080Ti GPU.

C. Results and Comparison

1) Comparison with the manually designed configurations: In this subsection, We compare our method with manual designs from ScaledYOLOv4 and YOLOv5. Table II shows better accuracy and latency of our discovered architectures on the VOC dataset. Notably, the Zero-DNAS architecture, discovered from the 'YOLOv4-CSP small' search space (as defined in Table I), demonstrates a 0.6 higher average precision (AP) and approximately a 32% faster inference compared to ScaledYOLOv4 [23]. Furthermore, Zero-DNAS-L (discovered from the "YOLOv4-CSP large" search space) achieves a 0.5 higher AP and approximately 48.64% faster inference compared to ScaledYOLOv4-P5 [23]. In the YOLOv5 space, v5-Zero-DNAS-L and v5-Zero-DNAS-M match accuracy with YOLOv5-L/M, but with significantly faster speeds. Notably, fairness ensured with similar FLOPS and parameters.

2) Comparison with other NAS algorithms: Next, we conduct comparisons with other sampling-based and differentiable NAS methods. As shown in Table III, Zero-DNAS outperforms other NAS methods. It significantly reduces the search time, outperforming recent methods by 60x, while achieving a more efficient and accurate model configuration. Furthermore, unlike the DNAS algorithms with macro-architecture search spaces such as FBNet [9], which encounter memory exhaustion in complex object detection tasks, our method successfully completes the search procedure.

We further compared the models discovered using the proposed Zero-DNAS method with other recent NAS-designed object detection architectures on the MSCOCO dataset. The results are presented in Table IV. Our algorithm excels in both time and GPU memory usage, operating 50x faster and utilizing 10x less GPU memory, all while maintaining high performance and efficiency in the resulting architecture.

3) Ablation study of hardware constraints: In this subsection, we analyze the impact of hardware constraints on the search procedure and the discovered architecture. Table V presents the effects of different losses on the search process and showcases the performance of the discovered architecture. The results indicate that the algorithm closely follows the chosen target constraints. Nevertheless, amalgamating the FLOPS, parameters, and depth losses offers greater control over the search process. When the depth constraint is removed, the algorithm tends to prioritize deeper networks at the cost of a drop in FPS. Similarly, relaxing the parameter constraint results in a higher AP but leads to a larger model size.

V. Conclusions

This work introduces an efficient and differentiable Neural Architecture Search (NAS) algorithm for object detection, capable of identifying precise and efficient architectures under specific hardware constraints in less than 2 hours. Utilizing standard GPUs, such as the NVIDIA 2080Ti, our method demonstrates the feasibility of conducting NAS for complex object detection tasks, providing an advantage over traditional DNAS methods that demand high-end GPUs. By leveraging a training-free metric (WOT) to score network initialization within a differentiable architecture search framework, we can discover highly accurate object detection models subjected to specific FLOPS and parameter requirements. Through the incorporation of a depth-aware loss, the proposed approach ensures enhanced inference speeds. The resulting object de-

TABLE II: The performance of architectures discovered in the ScaledYOLOv4 and YOLOv5 search spaces was evaluated on the VOC dataset using an NVIDIA 2080Ti device. For fair comparisons, the FLOPS and parameters of our models configurations were set to match the manually designed counterparts: ScaledYOLOv4-P5, ScaledYOLOv4, YOLOv5-L, and YOLOv5-M.

Model	AP50	AP	FLOPS	Target FLOPS	Parameters	Target Parameters	FPS
ScaledYOLOv4-P5	71.6	47.5	165G	165G	70.2M	70.2M	37
Zero-DNAS-L (ours)	**71.8**	**48**	**156.1G**	165G	**69.5M**	70.2M	**55**
ScaledYOLOv4	71.5	46.9	119G	119G	52.5M	70.2M	47
Zero-DNAS (ours)	**71.7**	**47.5**	**117.7G**	119G	**51.5M**	52.5M	**62**
YOLOv5-L	65.6	42.9	114G	114G	46.6M	46.6M	82
v5-Zero-DNAS-L (ours)	**65.8**	**43**	**113.4G**	114G	**45.7M**	46.6M	**85**
YOLOv5-M	64.7	41	51G	51G	21M	21M	99
v5-Zero-DNAS-M (ours)	64.3	40.6	54G	51G	**16.8M**	21M	**120**

TABLE III: Comparison results on the VOC dataset. It shows that our search method is faster and results in more accurate and fast architectures than other NAS algorithms. The discovered architecture setting also performed better than the original ScaledYOLOv4. The search was conducted using the same ScaledYOLOv4 search space. Device: NVIDIA 2080Ti.

Model	Search time	AP50	AP	FLOPS	Parameters	FPS
ScaledYOLOv4[23]	-	71.5	46.9	119.3G	52.5M	48
Cream[16]	108 h	69	45.5	119.3G	44.9M	50
AttentiveNAS[15]	104 h	69.5	45.7	96.4G	40M	53
FBNet[9]	Out of memory	-	-	-	-	-
Zero-DNAS	**1.8 h**	**71.7**	**47.5**	**117.7G**	**51.5M**	**62**

TABLE IV: Comparison results on the MSCOCO dataset. Our Zero-DNAS method is faster and can discover more accurate models than other object detection NAS algorithms. Moreover, Zero-DNAS requires fewer and less powerful GPUs to search.

Model	Search time	AP	AP50	FLOPS	Parameters	GPU	Overall Memory
NAS-FCOS[18]	672 h	39.8	-	189.6G	38.4M	v100 (32GB)	256GB
OPANAS[17]	96 h	39.6	-	197G	35.5M	v100 (32GB)	128 GB
SM-NAS[20]	4512 h	40.1	58.2	-	-	v100 (32GB)	>256GB
Hit-detector[21]	> 100 h	41.4	62.4	272G	**27.12M**	v100 (32GB)	256GB
Zero-DNAS	**1.8 h**	**42.1**	**63**	**117.4G**	51.5M	**2080Ti (11GB)**	**22GB**

TABLE V: Ablation study of the influence of different constraints on the discovered architecture. Device: NVIDIA 2080Ti.

Method	FLOPS loss	Parameters loss	Depth loss	FLOPS	Target FLOPS	Parameters	Target Parameters	AP	AP50	FPS
ScaledYOLOv4	-	-	-	119.0G	-	52.5M	-	46.9	71.5	48
Zero-DNAS	✓	X	X	119.8G	119G	62.0M	-	47.2	71.4	35
	✓	✓	X	119.6G	119G	50.7M	52.5M	43.3	67	38
	✓	✓	✓	117.4G	119G	51.5M	52.5M	47.5	71.7	62

tection architectures surpass both manually-designed models and those discovered by other NAS-based methods.

VI. Acknowledgments

This work was funded in part by Qualcomm through a Taiwan University Research Collaboration Project and also supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC-110-2221-E-A49-066-MY3, Grant NSTC-109-2221-E-009-112-MY3, Grant NSTC-111-2218-E-A49-029-, Grant NSTC-110-2221-E-A49-065-MY3, Grant NSTC-112-2425-H-A49-001-, and Grant NSTC-111-2634-F-A49-010; and in part by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and the Ministry of Education (MOE), Taiwan.

References

[1] W. Zhang, K. Song, X. Rong, and Y. Li, "Coarse-to-fine uav target tracking with deep reinforcement learning," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1522–1530, 2019. DOI: 10.1109/TASE.2018.2877499.

[2] Y. Yin, D. Xu, X. Wang, and L. Zhang, "Adversarial feature sampling learning for efficient visual tracking," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 2, pp. 847–857, 2020. DOI: 10.1109/TASE.2019.2948402.

[3] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *International conference on machine learning*, PMLR, 2018, pp. 4095–4104.

[4] Z. Guo, X. Zhang, H. Mu, *et al.*, "Single path one-shot neural architecture search with uniform sampling," in *European Conference on Computer Vision*, Springer, 2020, pp. 544–560.

[5] Z. Ding, Y. Chen, N. Li, D. Zhao, Z. Sun, and C. P. Chen, "BNAS: Efficient neural architecture search us-

ing broad scalable architecture," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[6] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," *arXiv preprint arXiv:1806.09055*, 2018.

[7] Y. Zhou, X. Xie, and S.-Y. Kung, "Exploiting operation importance for differentiable neural architecture search," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[8] M. Tan, B. Chen, R. Pang, *et al.*, "MnasNet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828.

[9] B. Wu, X. Dai, P. Zhang, *et al.*, "FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 734–10 742.

[10] M. S. Abdelfattah, A. Mehrotra, L. Dudziak, and N. D. Lane, "Zero-cost proxies for lightweight nas," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=0cmMMy8J5q.

[11] W. Chen, X. Gong, and Z. Wang, *Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective*, 2021. arXiv: 2102.11535 [cs.CV].

[12] M. Lin, P. Wang, Z. Sun, *et al.*, *Zen-nas: A zero-shot nas for high-performance deep image recognition*, 2021. arXiv: 2102.01063 [cs.CV].

[13] A. Krishnakumar, C. White, A. Zela, R. Tu, M. Safari, and F. Hutter, "NAS-bench-suite-zero: Accelerating research on zero cost proxies," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=yWhuIjIjH8k.

[14] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[15] D. Wang, M. Li, C. Gong, and V. Chandra, "AttentiveNas: Improving neural architecture search via attentive sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6418–6427.

[16] H. Peng, H. Du, H. Yu, Q. Li, J. Liao, and J. Fu, "Cream of the Crop: Distilling prioritized paths for one-shot neural architecture search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 955–17 964, 2020.

[17] T. Liang, Y. Wang, Z. Tang, G. Hu, and H. Ling, "OPANAS: One-shot path aggregation network architecture search for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 195–10 203.

[18] N. Wang, Y. Gao, H. Chen, *et al.*, "NAS-FCOS: Fast neural architecture search for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 943–11 951.

[19] Y. Chen, T. Yang, X. Zhang, G. MENG, X. Xiao, and J. Sun, "Detnas: Backbone search for object detection," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[20] L. Yao, H. Xu, W. Zhang, X. Liang, and Z. Li, "SM-NAS: Structural-to-modular neural architecture search for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 12 661–12 668.

[21] J. Guo, K. Han, Y. Wang, *et al.*, "Hit-Detector: Hierarchical trinity architecture search for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 405–11 414.

[22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[23] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2021, pp. 13 029–13 038.

[24] G. Jocher, A. Stoken, J. Borovec, *et al.*, *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*, version v3.1, Oct. 2020. DOI: 10.5281/zenodo.4154370. [Online]. Available: https://doi.org/10.5281/zenodo.4154370.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[26] X. Chen, L. Xie, J. Wu, and Q. Tian, "Progressive differentiable architecture search: Bridging the depth gap between search and evaluation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1294–1303.

[27] X. Dong and Y. Yang, "One-shot neural architecture search via self-evaluated template network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3681–3690.

[28] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[29] J. Mellor, J. Turner, A. Storkey, and E. J. Crowley, "Neural architecture search without training," in *International Conference on Machine Learning*, PMLR, 2021, pp. 7588–7598.

[30] M. Lin, P. Wang, Z. Sun, *et al.*, "Zen-NAS: A zero-shot nas for high-performance image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 347–356.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Reliable Feature-Based Framework for Vehicle Tracking in Advanced Driver Assistance Systems

Ngoc-Quan Ha-Phan[1], Thanh-Nguyen Truong[2], Vu-Hoang Tran[3], Ching-Chun Huang[4]
[1 2 3] Ho Chi Minh City University of Technology and Education, Vietnam
[1] E-mail: 19145008@student.hcmute.edu.vn
[2] E-mail: 19145158@student.hcmute.edu.vn
[3] E-mail: hoangtv@hcmute.edu.vn
[4] National Yang Ming Chiao Tung University, Taiwan
E-mail: chingchun@nycu.edu.tw

***Abstract* — Vehicle tracking has always been a vital aspect of modern transportation systems. This phenomenon has gained even more interest with the introduction of Advanced Driver Assistance Systems (ADAS) and Autonomous Vehicles. Most state-of-the-art (SOTA) vehicle trackers, and their enhanced versions, commonly rely on mathematical motion models (e.g., Kalman Filter) as the core information. However, these models may produce unreliable outputs, especially when objects exhibit complex motion patterns. Hence, we propose a reliable feature-based tracking framework that fully exploits distinct vehicle appearance and conduct a comparative analysis with classic motion-based trackers. Additionally, we revisit previously proposed track handling strategies to incorporate a specially designed track management system for feature-based tracking. The proposed method achieves the highest score on all selected multi-object-tracking (MOT) evaluation metrics compared to the current SOTA methods on the KITTI dataset. Notably, our approach experienced significantly low False Positive (FP) errors, ensuring its performance in minimizing unreliable information.**

I. INTRODUCTION

Multi-object tracking (MOT) has gained significant popularity in the field of computer vision, finding applications in various domains such as robotic perception, complex medical imaging, and surveillance management. In the context of Autonomous Vehicles and ADAS, object tracking plays a critical role. By continuously monitoring the positions and trajectories of vehicles, pedestrians, and other objects, the tracking framework provides essential information for autonomous systems to make decisions and take appropriate actions. It is a vital component that contributes to the perceptual aspects of autonomous vehicle systems, including navigation, path planning, and even the advanced Vehicle-to-Infrastructure (V2I) tasks related to traffic management.

Vehicle Tracking has been dominated by the tracking-by-detection (TBD) paradigm, which leverages the innovative performance of modern object detectors to enhance tracking accuracy. Specifically, TBD can be seen as a data association approach that aims to match the detections across frames in a sequence. Currently, two popular TBD methods are being extensively researched and developed: motion-based tracking and feature-based tracking.

Motion-based tracking has recently gained favor in TBD field. This approach heavily relies on mathematical motion models, specifically the Kalman Filter, to calculate motion cues such as velocity, direction, and acceleration. These cues are then used to predict the object's location within a frame. Despite its inability to accurately estimate object location, this method is popularly employed due to its computational efficiency and robust performance in the presence of noise, such as changes in lighting conditions or object appearance variations.

One example of a motion-based tracking method is SORT [1], which utilizes Kalman Filter to predict object's state and adopts the Intersection over Union (IoU) and Hungarian algorithm to associate tracks across a video sequence. Although SORT performs well in real-time scenarios, it faces challenges in efficiently tracking occluded objects, leading to a relatively high number of identity switches. Addressing the occlusion issue, DeepSORT [2] builds upon SORT and takes object appearance into account. It integrates a pre-trained Convolution Neural Network (CNN) into the association process and introduces a matching cascade mechanism that computes appearance similarities before employing IoU and Hungarian matching. DeepSORT demonstrates relatively low identity switches, showcasing its strong performance in

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

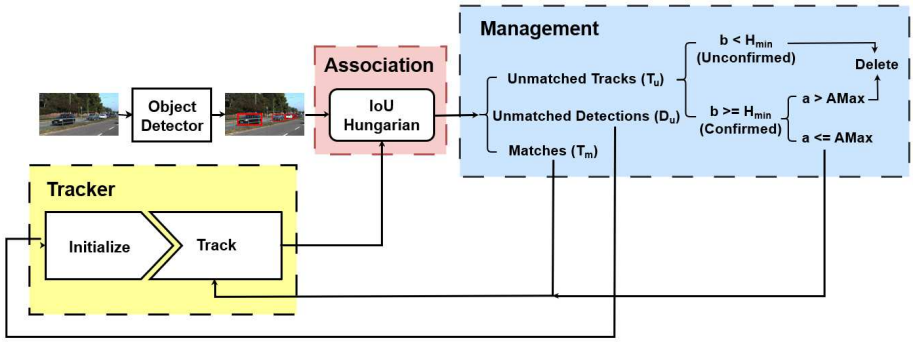


Fig. 1 System overview of proposed tracking framework

occlusion scenarios. Drawing upon the foundations laid by SORT and DeepSORT, several proposed approaches have been developed to overcome the remaining limitations, such as improving detection quality [3], handling non-linear motion [4], or addressing missing detection problems [5].

However, the heavy reliance on mathematical models introduces a level of unreliability to the output of trackers in ADAS. The assumptions made by these models often fail to accurately capture all the complex and diverse motions exhibited by on-road vehicles in real-world scenarios. For instance, situations of abnormal vehicle behaviors such as sudden lane change, hard braking, rapid acceleration as well as variation in camera motion contribute to the varied range of different vehicle motions that usually happen. Additionally, cameras in ADAS systems are not stationary but follow the vehicle, which creates additional challenges for motion-based tracking systems. Moreover, the appearance of on-road vehicles remains relatively consistent in terms of size, shape, color, and texture, and these characteristics do not undergo significant changes during tracking. Therefore, feature-based tracking can fully exploit these consistent appearance details and leverage them for accurate tracking.

In light of this, we propose a fully feature-based multi-object tracking framework that shifts the focus from mathematical models to leveraging the object appearance as the primary information for tracking. Our approach builds upon a previously proposed tracker [6] that analyzes rich vehicle features such as shapes, colors, or textures. By utilizing these insights, our framework can effectively handle complex real-world scenarios, including on-road objects' varied motion patterns or occlusions. To enhance the track management strategy, we also revisit SORT and incorporate an adaptive thresholding approach specifically designed for feature-based tracking. This approach aims to reduce the occurrence of common errors associated with feature-based vehicle tracking, such as out-of-frame handling, and feature loss. In summary, our contributions can be summarized as follows:

1. **A novel Feature-based tracking framework:** our proposed tracking framework fully exploits object appearance as tracking information, leading to reliable and accurate information.
2. **A Track Management System:** we revisit SORT [1] and design a track management system specialized for feature-based tracking.
3. **Adaptive Thresholding Approach:** we proposed an adaptive thresholding approach to handle the problems of feature-based trackers such as out-of-frame or feature loss scenarios.

II. Proposed Method

In this section, we present a novel vehicle tracking framework, as shown in Fig. 1, which incorporates an Object Detector, a Feature-based Tracker, a Data Association mechanism, and a Track Management strategy. Notably, we introduce adaptive thresholding approaches specifically designed for feature-based tracking methods. The following subsections provide a comprehensive description of these components, offering detailed insights into their methodologies.

A. Object Detector

Firstly, the accurate identification and precise localization of tracking targets in image sequences are crucial for achieving high-quality tracking performance. This object detection task (OD) plays a pivotal role in determining the overall accuracy and effectiveness of the tracking results. Furthermore, in ADAS applications, it is essential to meet strict processing time constraints. To meet these requirements, we evaluated several single-stage OD models from the YOLO family, known for their real-time capabilities, simplicity, and efficiency [7]. Among these models, YOLOv7 stood out due to its outstanding accuracy on the Microsoft COCO dataset [8], as shown in TABLE 1. Consequently, we have chosen YOLOv7 as our object detector to ensure accurate and efficient object detection,

thereby guaranteeing reliable tracking in ADAS systems. By leveraging the capabilities of YOLOv7, we aim to enhance the overall performance of our tracking framework and provide dependable object detection in real-time scenarios.

TABLE 1
COMPARISON OF BASELINE OBJECT DETECTORS [9]

Model	#Param. ↓	FLOPs↓	Size	AP^{val}↑	AP^{val}_{50}↑	AP^{val}_{75}↑
YOLOv4 [10]	64.4M	142.8G	640	49.7%	68.2%	54.3%
YOLOR-u5 (r6.1) [11]	46.5M	109.1G	640	50.2%	68.7%	54.6%
YOLOv4-CSP [12]	52.9M	120.4G	640	50.3%	68.6%	54.9%
YOLOR-CSP [11]	52.9M	120.4G	640	50.8%	69.5%	55.3%
YOLOv7 [9]	**36.9M**	**104.7G**	640	**51.2%**	**69.7%**	**55.5%**

B. *Feature-based Tracker*

The primary objective of the tracker is to estimate the position of the object based on information from the previous frame. In this paper, we utilize the idea of DiMP-50 Tracker [6] to build our feature-based tracker. The tracker's workflow consists of two primary steps: Initialize and Track. In the following parts of this subsection, we delve into the methodology of these components, providing a detailed overview of their functioning. Moreover, an additional aspect of the tracker, which involves bounding box refinement using IoUNet prediction, will be described.

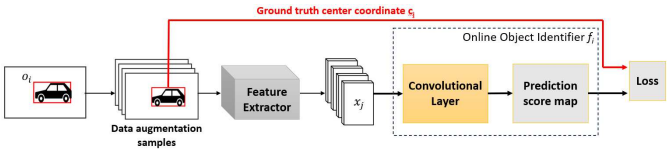


Fig. 2 Initialize component of tracker.

Initialize: The initialize step is responsible for establishing object identifiers when they first appear, as shown in Fig. 1, based on their characteristics. Fig. 2 illustrates the initialize process, the input to the component is an image containing information about a newly detected bounding box o_i, obtained from the object detector. Some data augmentation methods such as: translation, rotation, blur and dropout are then performed to create a set of augmented image samples that accounts for potential variations in object appearance. The Initialize process continues with feature extractor, here using ResNet-50, to create a set of deep feature maps x_j from the augmented image samples. The set is combined with associated bounding box centroid c_i to create a set of training samples, denoted as $S_{train} = \{(x_j, c_j\}_{j=1}^{n}$, the training samples are then used to train the online object identifier f_i, which consists of a single convolutional layer responsible for detecting the possibility of an input window containing an object.

The object identifier f_i is designed as a compact network that can be trained online for the purpose of being able to continuously and rapidly update changes in the object's appearance. It will be trained 10 epochs for the first frame, and then every 20 frames it will be trained 20 epochs on the updated training dataset S_{Train}. The object identifier is trained using the regularized loss is described in (1).

$$L = \frac{1}{|S_{Train}|} \sum_{(x,c)\in S_{Train}} \|r(f_\theta(x), c)\|^2 + \lambda\|\theta\|^2, \quad (1)$$

where, θ denotes the weight of the object identifier f and λ is the regularization factor. The function r captures the difference, or errors, between the predicted target confidence score $f_\theta(x)$ and the ground-truth bounding box centroid c at each spatial location. While r can be straightforwardly defined as the difference between the two inputs, to address the issue of data imbalance and enhance the model's emphasis on positive data, we opt to utilize a combination of least-squares

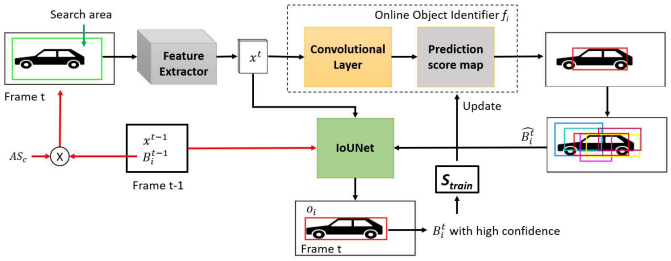


Fig. 3 Track component of tracker.

regression and hinge loss for r, as proposed in [6].

Track: After establishing initial information, the tracker proceeds to the next task of estimating object states in the subsequent frames, which is dealt with by the Track component. Fig. 3 describes the operational mechanism of the track component on the subsequent frames in a video sequence for an object o_i. To save processing time, the system will rely on the bounding box of the object in the previous frame B_i^{t-1} to determine the search area for the current frame t, and the feature x^t is only extracted in this search area. This area delimits the most probable position that the object exists in the current frame, enable the tracker concentrate only on the potential location of the object in the frame. Unlike in [6], where the search area is determined by multiplying B_i^{t-1} by a fixed scaling factor S_c, we will change S_c depending on the relative position of the object in previous frame as calculated in (2).

$$AS_c = \frac{min(d_x, d_y)}{D_{max}} S_c, \quad (2)$$

where, AS_c denotes adaptive search area scale, S_c represents the pre-defined default scaling factor, and d_x and d_y indicate the closest distances from the center of the bounding box in the previous frame to the edges of the input image, respectively. D_{max} signifies the maximum distance from the bounding box center to the image's edge. If the bounding box center in the

previous frame is closer to the vertical edge, D_{max} is calculated as half the image's width; otherwise, it is calculated as half the image's height. The adaptive search area scale dynamically adjusts the search area when the object approaches the image's edge, causing a partial disappearance and altering the object's appearance. By restricting the search area, the system focuses on the most relevant and reliable object features, enabling the tracker to adapt better to appearance changes and reduce false outputs.

Using the extracted features x^t, the online object identifier f_i is employed to generate a confidence score map across the entire search area. This map reflects the likelihood of the object's presence at different locations. The location with the highest score is chosen as the centroid of the object's bounding box in the current frame, representing the most probable location of the object. By combining this centroid with the width and height of the bounding box from the previous frame, a coarse estimation of the object o_i in current frame is obtained.

Although the tracking module performs well in determining the location of objects within the frame, its accuracy is limited as it relies on a coarse estimation derived solely from the size of the previous bounding box. To address this limitation, the IoUNet prediction approach, introduced in [13], is employed to refine and regress a more precise target box. As shown in Fig. 3, after obtaining a coarse estimation of the object's location in terms of the bounding box centroid together with the bounding box width and height from the previous frame, ten bounding box proposals, denoted as $\widehat{B_i^t}$, are generated by adding uniform noise, encompassing a range of potential object positions surrounding the coarse estimation. Using the extracted features x^{t-1} and bounding box B_i^{t-1} from previous frame, as well as current extracted features x^t, IoUNet predict IoU score for each proposal in $\widehat{B_i^t}$. The final estimation, B_i^t, is obtained by averaging of three proposals with the highest IoU score.

Furthermore, in order to maintain the object identifier's effectiveness, new training samples will be incorporated into the S_{train} dataset if they are predicted with a satisfactory level of confidence. To prevent an excessive increase in memory size, the oldest samples in S_{train} will be removed to make room for the new ones. This process ensures that the object identifier remains up to date and capable of accurately identifying objects.

C. Data Association Method

The core requirement of any tracking-by-detection method is the ability to assign detections to existing targets. In detail, produced output of the tracker, which includes bounding boxes of targets, is associated with measured bounding boxes given by an object detector. We tackle the assignment problem by using a similar approach to SORT [1], which employs Intersection over Union (IoU) and the Hungarian matching algorithm as shown in Fig. 1.

Given N detected boxes (D) from the object detector and M generated track boxes (T) from the tracker, the IoU distance between each detected box and all generated boxes resulting in cost matrix, $mIoU\ (D,T)$, as described in (3):

$$mIoU(D,T) = \begin{bmatrix} IoU(D_1,T_1) & \cdots & IoU(D_1,T_M) \\ \vdots & \ddots & \vdots \\ IoU(D_N,T_1) & \cdots & IoU(D_N,T_M) \end{bmatrix}, \quad (3)$$

where, IoU corresponds to the IoU score between chosen detected box D and generated box T is calculated as in (4).

$$IoU(D,T) = \frac{D \cap T}{D \cup T}. \quad (4)$$

Given the cost matrix $mIoU$, the Hungarian Algorithm [14] witnessed the final stage of the assignment process resulting in three possible outputs, namely matches (T_m), unmatched tracks (T_u), and unmatched detections (D_u) as shown in Fig. 1. Each T_m represents the pairings of D and T that are matched with the highest IoU score. T_u indicates tracks that have an overlap less than a predefined IoU_{min} threshold and, similarly, D_u refers to unmatched detections, which correspond to detections of new objects.

D. Track Management

Dealing with the set of T_m, T_u, and D_u, our framework employs a modified version of track handling approach as in [1], [2], as shown in Fig. 1. Additionally, we also propose an adaptive thresholding method designed for the proposed system, which can also be used for other feature-based trackers.

Track Initialization and Rejection: New tracks are initialized when there exists a set of unmatched detection D_u, each element in D_u will be initialized as new track T with the associated parameter $(c, \gamma, h, a, b, S_c)$ that contains bounding box center position c, aspect ratio γ, height h followed by the track management parameters such as track age a, track hit b, and predefined search area scale S_c.

After initialization, each T undergoes data association for detected boxes in the subsequent frame. The counter b is incremented following a successful association. In contrast, for every unsuccessful association of tracks, a and S_c are incremented to accommodate potential object re-entry and are reset to the initial value if any track subsequently rematches with a detected box. Any track with the related a that exceeds a maximum age threshold $AMax$ is considered to have

permanently left the frame and is therefore rejected. Unlike SORT and DeepSORT, which rely on a fixed age threshold, we propose to use an adaptive thresholding approach specifically designed for feature-based trackers, aiming to minimize associated errors. The $AMax$ is defined similarly to AS_c as described in (2). By employing the Adaptive Maximum Age mechanism, when a vehicle's trajectory starts to extend beyond the frame boundaries, $AMax$ gradually diminishes, facilitating the quick deletion of these tracks. This adaptive approach ensures that the tracker effectively handles objects that move out of the frame, enhancing overall tracking performance.

Track Confirm Mechanism: In tracking applications, scenarios when objects only appear in a few initial frames, should also be considered. Therefore, as described in [2], tracks at creation are considered to be tentative during their first H_{min} frames and is confirmed after consecutive successful associations. During H_{min}, tentative tracks will be deleted right after an unsuccessful association. By requiring consecutive matches, this method helps minimize related errors caused by false or short-lived detections.

III. Experimental Results

In this section, we presented an in-depth analysis of tracking evaluation and performance comparison to assess the performance of the proposed tracking framework compared to existing methods on the KITTI Object Tracking Benchmark [15]. For evaluation, the proposed method is tested on a chosen set of sequences specialized for vehicle tracking. For comparison, we executed the proposed tracking framework as well as reproduced results from DeepSORT [2] and SORT [1] using the same detection input, the results are evaluated on different selected metrics. Furthermore, performance analysis and ablation studies are accomplished for deeper insight.

A. Dataset

Our tracking framework is evaluated on KITTI Object Tracking Benchmark [15], a popular dataset utilized for many tasks in Autonomous Vehicle applications. The dataset incorporates data from various sensors including Color and Greyscale Cameras, a Laser Scanner, and GPS/IMU navigation systems. KITTI Image Dataset comprises a total of 21 Training and 29 Testing sequences, all captured at a common frame rate of 10 FPS using vehicle-mounted color cameras positioned at a height of 1.65 meters above the ground.

While the KITTI Dataset can be used for tracking applications on either vehicles or pedestrians, our framework specifically focuses on vehicle tracking. Consequently, a total of nine sequences, primarily composed of on-road vehicles, are selected for evaluation. These chosen sequences cover a diverse range of on-road scenarios spanning from highways to urban areas for a realistic evaluation of the framework's performance.

B. Evaluation Metrics

We use the metrics MOTA, HOTA, and IDF1 [16]–[18] to evaluate our tracking performance, with each metric providing different criteria for tracking performance assessment.

MOTA [16] measures the tracking performance at the detection level by calculating the matching between each ground-truth detection $(gtDet)$ and predicted detection $(prDet)$. From that approach, the number of matched pair of $gtDet$ and $prDet$ is regarded as True Positive (TP). The number of times a $prDet$ is unmatched results in the number False Positive (FP) while the number missed $gtDet$ corresponds to False Negative (FN). Moreover, association performance in MOTA is measured with the concept of Identity Switch $(IDSW)$, an $IDSW$ is counted when objects unexpectedly swap their identity during tracking. The final score of MOTA is calculated as in (5).

$$MOTA = 1 - \frac{|FP| + |FN| + |IDSW|}{|gtDet|}. \quad (5)$$

On the other hand, IDF1 [18] acquires tracking errors at the association level by measuring bijective matching between sets of ground-truth trajectories $(gtTraj)$ and prediction trajectories $(prTraj)$. Matching between each $gtTraj$ and $prTraj$ results in ID-True Positive $(IDTP)$. The counts of ID-False Negative $(IDFN)$ and ID-False Positive $(IDFP)$ represent the counts of remaining unmatched $gtDet$ and $prDet$, respectively. Final IDF1 score is calculated as in (6).

$$IDF1 = \frac{|IDTP|}{|IDTP| + 0.5|IDFN| + 0.5|IDFP|}. \quad (6)$$

HOTA [17] is a recently proposed metric that, unlike MOTA and IDF1, considers errors at both mentioned levels. This metric balances the effects of detection and association errors on tracking performance. At the detection level, HOTA acquires the instances of FN, FP, and TP in a similar way to MOTA to calculate Detection Accuracy $(DetA)$. Additionally, by using the method of matching the set of ground-truth Identity $(gtID)$ with prediction Identity $(prID)$ over a set of TPs, HOTA proposed an association-level evaluation that acquires True Positive Association (TPA), False Negative Association (FNA), and False Positive Association (FPA), these three concepts are used to measure Association Accuracy

($AssA$). As a result, the HOTA score over different localization thresholds (0.05 to 0.95 in 0.05 intervals) is calculated as in (7).

$$HOTA = \int_0^1 HOTA_\alpha \, d\alpha \approx \frac{1}{19} \sum_{\substack{\alpha=0.05 \\ \alpha+=0.05}} HOTA_\alpha \,, \quad (7)$$

where, $HOTA_\alpha$, calculated in (8), is the HOTA score over an α, which is the geometric mean of $DetA$ and $AssA$ over α.

$$HOTA_\alpha = \sqrt{DetA_\alpha \cdot AssA_\alpha}, \quad (8)$$

where, $DetA_\alpha$ and $AssA_\alpha$ are $DetA$ and $AssA$ calculated with localization threshold α as in (9) and (10).

$$DetA_\alpha = \frac{|TP|}{|TP| + |FN| + |FP|}, \quad (9)$$

$$AssA_\alpha = \frac{1}{|TP|} \sum_{c \in TP} \frac{|TPA|}{|TPA| + |FNA| + |FPA|} . \quad (10)$$

Note that the concepts of TP, FN, FP, TPA, FNA and FPA in (9) and (10) are calculated for a particular value of α. However, for clarity, the subscript is omitted.

C. Results

For comparison, we also reproduced two popular motion-based tracking methods that employ similar association strategies as our proposed method, namely DeepSORT [2] and SORT [1].

TABLE 2 and TABLE 3 illustrate the results of our evaluation on MOTA and IDF1 metrics, respectively. It is evident that our method notably improved tracking performance in the overall MOTA score. In detail, our method ranked first in FN and FP, surpassing the second-best method (SORT) with a slight reduction in FN and a complete out-performance in FP (from 775 to 465). The proposed method also secured the top position on other associated parameters, which are Mostly Track Ratio (MT), Mostly Lost Ratio (ML), and Tracking Precision (MOTP). Similarly, at the association level, the proposed method is dominant in IDF1 due to the relatively low IDFP value in comparison two other methods.

TABLE 2
EVALUATION RESULTS OF THREE METHODS ON MOTA METRIC.

	MOTA↑	MOTP↑	MT↑	ML↓	FN↓	FP↓	IDSW↓	Frag↓
SORT [1]	66.9	75.5	45.2	11.5	2928	775	153	202
DeepSORT [2]	63.3	74.7	44.9	12.5	2970	1169	**137**	**178**
Proposed	**70.1**	**80.5**	**48**	**10**	**2868**	**465**	152	211

TABLE 3
EVALUATION RESULTS OF THREE METHODS ON IDF1 METRIC.

	IDF1↑	IDTP↑	IDFN↓	IDFP↓
SORT [1]	70.52	7449	4190	2037
DeepSORT [2]	74.23	7971	3668	1867
Proposed	**76.68**	**8004**	**3635**	**1232**

To conclude, those mentioned enhancements prove the accuracy and reliability of our approach, the results on HOTA metrics (TABLE 4), additionally, contribute to this fact as the corresponding performance in both detection (*DetA)* and association (*AssA)* is leading in comparison with either SORT or DeepSORT.

TABLE 4
EVALUATION RESULTS OF THREE METHODS ON HOTA METRIC.

	HOTA↑	DetA↑	AssA↑
SORT [1]	54.65	53.85	56.62
DeepSORT[2]	56.46	52.21	62.62
Proposed	**60.77**	**58.42**	**63.84**

D. Ablation Studies

TABLE 5 summarizes the comparison of the proposed method on different improvement proposals, we compare our baseline method with two other ones with strategy improvements, namely Proposedv1 and Proposedv2.

Proposedv1 updated the Baseline method with Adaptive Maximum Age Threshold (AMax), resulting in a reduction in IDSW (-185) and Frag (Fragmentation) (-23). This approach also demonstrates the best performance in overall HOTA and MOTA scores. Here, the tendency indicates that the proposed method benefits considerably from AMax mechanism.

On the other hand, Proposedv2 integrated Adaptive Search Scale (ASc) and Track Confirm Mechanism (Cfm). ASc is implemented to minimize errors that may arise when objects tend to move out of frame, potentially resulting in feature loss and causing confusion for the tracker. The track confirm mechanism, on the other hand, is used to minimize errors for scenarios when tracks only appear for a few initial frames. By applying these two strategies, Proposedv2 significantly reduced *IDSW* and *Frag* errors, and achieved the highest association-level score (IDF1). Still, the improvement in IDF1(+0.04) compared to Proposedv1 is not considered significant, and there is a slight decrease in the overall HOTA and MOTA scores. This trend can be attributed to the fact that Proposedv2 focuses only on, as much as possible, reducing errors caused by objective uncertainties.

TABLE 5
ABLATION STUDY ON HOTA, MOTA, IDF1 FOR VARIOUS STRATEGIES: ADAPTIVE MAXIMUM AGE (**AMax**), ADAPTIVE SEARCH SCALE (**ASc**), TRACK CONFIRM (**Cfm**).

	AMax	*ASc*	*Cfm*	HOTA↑	IDF1↑	MOTA↑	IDSW↓	Frag↓
Baseline	-	-	-	58.21	71.68	69.71	396	294
Proposedv1	✓	-	-	**61.77**	76.64	**71.24**	211	271
Proposedv2	✓	✓	✓	60.77	**76.68**	70.06	**152**	**211**

E. Limitations

Although the proposed method has successfully witnessed improvements in accuracy and reliability, it still exhibits a certain limitation. The concern is that performance in IDSW

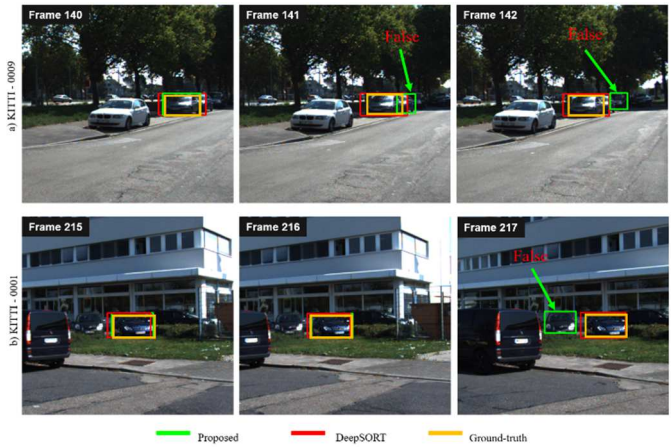


Fig. 4 Types of false output generated by Proposed method

and Frag is still poor compared to previous methods (especially DeepSORT). We consider this to be a distinctive issue in feature-based tracking, particularly evident in crowded scenarios, where uncertainties in object searching are relatively high. For example, Fig. 4 illustrates a situation in which vehicles are in close proximity and have identical appearances. The proposed method struggles to differentiate between the objects, leading to uncertain and incorrect outputs. In contrast, DeepSORT effectively handles the situation by leveraging the accurate predictions generated by the Kalman Filter. These findings highlight that future developments should focus on enhancing the proposed method's capability to predict object states in crowded environment.

IV. Conclusions

In this paper, we have introduced a novel feature-based tracking framework that leverages appearance information to enhance tracking performance. We have also presented a new track management system and an adaptive thresholding approach. These techniques have been specifically designed to address the requirements of vehicle tracking in ADAS applications. Through extensive evaluation on the well-established KITTI benchmark, our proposed method has demonstrated its effectiveness by surpassing existing approaches in terms of tracking accuracy and robustness. A series of ablation studies were conducted to further emphasize the advantages of the integrated track management system and adaptive thresholding approach. Overall, our findings underscore the importance of considering appearance information and implementing advanced track management strategies in feature-based tracking. This paper contributes to the existing body of knowledge in the field of vehicle tracking and provides valuable insights for future developments in ADAS applications.

References

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," *arXiv [cs.CV]*, 2016.

[2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and Realtime Tracking with a deep association metric," *arXiv [cs.CV]*, 2017.

[3] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu and X. Wang , "ByteTrack: Multi-object tracking by associating every detection box," *arXiv [cs.CV]*, 2021.

[4] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," *arXiv [cs.CV]*, 2022.

[5] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong and H. Meng, "StrongSORT: Make DeepSORT Great Again," *arXiv [cs.CV]*, 2022.

[6] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," *arXiv [cs.CV]*, 2019.

[7] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *arXiv [cs.CV]*, 2021.

[8] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common objects in context," *arXiv [cs.CV]*, 2014.

[9] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv [cs.CV]*, 2022.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv [cs.CV]*, 2020.

[11] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You Only Learn One Representation: Unified Network for Multiple Tasks," *CoRR*, vol. abs/2105.04206, 2021.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," *arXiv [cs.CV]*, 16-Nov-2020.

[13] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," *arXiv [cs.CV]*, 2018.

[14] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1--2, pp. 83--97, 1955.

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[16] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 2008.

[17] J. Luiten, A. Ošep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 548–578, 2021.

[18] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *arXiv [cs.CV]*, 2016.

Light-weight Zero-Reference-based Image Enhancement for Low-Light Images

Jie-Fan Chang[1], Kuan-Ting Lai[2], Cheng-Xuan Zhuang[2], Guo-Shiang Lin[2], and Ku-Yaw Chang [2]

[1] National Taiwan University, Taiwan

E-mail: jfanc@ntu.edu.tw, Tel: +886-2-33664888 ext.313

[2] National Chin-Yi University of Technology, Taiwan

E-mail: gslin@ncut.edu.tw, Tel: +886-4-23924505 ext. 8716

***Abstract* —** Images captured under low-light conditions often suffer from overall darkness and poor visual quality. In this study, we propose a light-weight zero-reference image enhancement method to improve the visual quality of low-light images. The proposed method is composed of some parts: image partition, transformation map estimation, and pixel value adjustment. To achieve transformation map estimation, a light-weight network, CSPDCE-Net, is developed. For performance evaluation, two image datasets, SICE and LOL, are selected here. For visual quality, the average PSNR and SSIM of the proposed method can achieve 17.01dB and 0.71, respectively. Compared with the existing methods, the proposed method can achieve better performance in terms of PSNR and SSIM. The experimental results show that the proposed method not only can process low-light images well but also is superior to some existing methods.

I. INTRODUCTION

With the rapid development of computer vision technology, digital image processing systems have been widely applied in various fields such as video surveillance, industrial production, and intelligent transportation. However, during the image acquisition process, the lighting conditions of images often vary, such as sunny, cloudy, rainy, and nighttime scenarios. However, images captured under poor lighting conditions, such as nighttime or cloudy conditions, may have poor visual quality. The image captured in relatively low-light environments is commonly referred to as a low-light image [1][2]. Figure 1 shows two low-light images. As shown in Fig. 1, it is quite difficult to recognize the image content by humans. This means that it would be more difficult for an intelligent system to detect and recognize human faces or license plates in low-light images. Therefore, improving the visual quality of low-light images plays a crucial role in many computer vision applications [2].



(a) (b)

Fig. 1 Examples of low-light images.

In fact, there are some traditional image enhancement techniques to deal with low-light images. For example, many histogram-based methods had been developed [14]. The basic concept of histogram-based methods is to modify pixel values to increase the contrast of input images. Recently, there are advancements in low-light image enhancement techniques developed based on deep convolutional neural networks (CNNs) [3]. However, most supervised-learning-based CNN-based image enhancement methods require a large amount of image pairs, i.e., low-light and normal-light images. In fact, capturing simultaneously both low-light and normal-light images in the same scene or synthesizing image pairs is an immense challenge. To overcome the constraint, several unsupervised deep learning methods have been proposed [4,5,6,7]. For example, the unsupervised-learning-based methods was developed based on a generative adversarial network and Retinex model [6]. In [6], a two-layer decomposition network decomposes a low-light image into an illumination map and a reflection map. Then, the Retinex model with correction coefficients is exploited to obtain the enhanced images. However, unpaired low/normal-light images are necessary for model training in [6]. The requirement makes the existing method infeasible.

To avoid the requirement of paired/unpaired low/normal-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

light images, an unsupervised-learning-based method was devised in [8]. The existing method called Zero-DCE [8] is designed to estimate the mapping function for modifying pixel values. The main feature of Zero-DCE is that paired or unpaired images are not needed for model training. On the other hand, some real applications such as video surveillance, intelligent factory, and autonomous driving, require intelligent image processing systems deployed in mobile or embedded devices. Therefore, this motivates us to develop a light-weight zero-reference network to deal with low-light images.

The rest of this paper is organized as follows. In Section II, we describe the proposed light-weight image enhancement method. Experimental results are shown in Section III. Section IV contains the conclusion.

II. PROPOSED METHOD

Image enhancement can be considered as the problem of finding an appropriate mapping function between input pixel values and the output results. In recent years, many low-light image enhancement methods have been developed based on deep neural networks such as U-Net [9] and its variants to produce visually improved results. However, these existing methods often have a large number of model parameters and computational requirements and then make them inappropriate to run on edge devices [10].

In fact, it is an important issue that a network should be light-weight to process low-light images due to limited storage and computational power of embedded/edge devices in existing applications. However, the existing methods may suffer from the issue of limited storage and computational power. Unlike U-Net, DCE-Net [11] does not adopt the down-sampling operation to reduce the size of feature maps for computing a suitable transformation function. Unfortunately, this leads to the need of more memory and computational complexity for processing these feature maps and then makes DCE-Net unsuitable for embedded/edge devices.

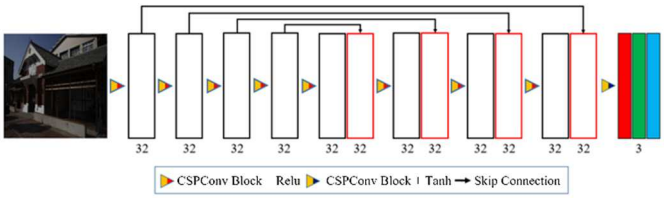
To effectively address the above issues, we propose a new light-weight image enhancement method to deal with low-light images. The proposed method is composed of three parts: image partition, transformation map estimation, and pixel value adjustment. In order to reduce the memory usage during runtime, the proposed method splits the input image into four regions. After image partition, the proposed network, CSPDCE-Net, is used to estimate the transformation map for these regions. Finally, pixel values in each region are adjusted based on the estimated transformation map to obtain the enhanced image. We elaborate each part in the following.

A. *Image partition*

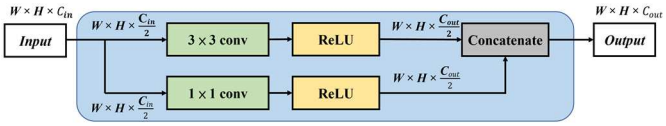
To reduce the memory requirement during model execution, this study divides the input low-light image into four equal parts, denoted as I_1^O, I_2^O, I_3^O, and I_4^O, before feeding them into the CSPDCE-Net for transformation map estimation. Each part is separately processed by the CSPDCE-Net, resulting in transformation parameter maps $\mathcal{A}_1$, $\mathcal{A}_2$ $\mathcal{A}_3$, and $\mathcal{A}_4$. Finally, these individual maps are merged to obtain the combined transformation parameter map $\mathcal{A}$ for pixel value adjustment. It is expected that this approach would reduce the memory usage during the subsequent transformation map estimation.

B. *Transformation map estimation*

To estimate the transformation map for each region, a light-weight network, CSPDCE-Net, is proposed. Similar to DCE-Net, the architecture of the proposed CSPDCE-Net model is shown in Fig. 2(a). As shown in Fig. 2(a), similar to DCE-Net, CSPDCE-Net consists of some convolutional layers. However, the main difference between DCE-Net and CSPDCE-Net is that each 3×3 convolutional layer in DCE-Net is replaced with a CSPConv Block, forming a new model architecture called CSPDCE-Net. In addition, the main concept of CSPNet is adopted to establish CSPDCE-Net in this study.



(a)



(b)

Fig. 2 CSPDCE-Net architecture: (a) CSPDCE-Net and (b) proposed CSPConv block.

As shown in Fig. 2(b), the input of a CSPConv block is divided into two parts. The two Conv1×1 layers are used to divide the input feature maps into two parts. In Fig. 2(b), C_{in} and C_{out} represent the number of the input and output channels in a CSPConv block, respectively. In one part, the 3×3 convolutional layer is performed by using depth-wise convolution. Since only a half of input feature maps are

processed by the 3×3 convolution, it is expected that the computational complexity of the proposed method can be reduced. Then the two parts are concatenated as the output in a CSPConv block. It is expected that the CSPDCE-Net can learn the transformation map after network training with a suitable loss function.

C. *Pixel value adjustment*

After generating the transformation map by using CSPDCE-Net, the next step is to adjust pixel values of the low-light images. This means that the visual quality of a low-light image can be improved based on the estimated transformation map. Therefore, the pixel value adjustment function $LE_n(\mathrm{x})$ is described as follows:

$$LE_n(x) = LE_{n-1}(x) + \mathcal{A}(x)LE_{n-1}(x)\big(1 - LE_{n-1}(x)\big) \qquad (1)$$

where x represents the input, $\mathcal{A}$ represents the transformation map estimated by CSPDCE-Net, and $LE_{n-1}(x)$ represents the result. The iteration is performed for a total of 8 times.

D. *Network training*

One key issue is to suitably estimate the transformation map for pixel value adjustment. To obtain a trained network for transformation map estimation, we adopted an unsupervised training framework using multiple exposures and low-light images. The proposed framework of network training is illustrated in Fig. 3. The following describes how the model is trained and the design of the loss functions.

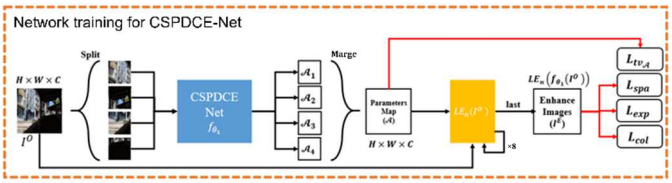


Fig. 3 Illustration of the network training for CSPDCE-Net.

Similar to Zero-DCE [8], the loss function used in the training process considers four characteristics:

1. Spatial Consistency
2. Exposure Control
3. Color Constancy
4. Illumination Smoothness

Spatial Consistency Loss (L_{spa})

The spatial consistency loss L_{spa} promotes spatial consistency in the enhanced images by preserving the differences between neighboring regions of the input image and its enhanced version. The loss L_{spa} is defined below:

$$L_{spa} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\alpha(i)}\big(\big|\big(I_i^E - I_j^E\big)\big| - \big|\big(I_i^O - I_j^O\big)\big|\big)^2, \qquad (2)$$

where K represents the number of local regions, $\alpha(i)$ denotes the four neighboring regions (up, down, left, right) centered around region i, and $|\cdot|$ represents the L_1 norm. I^E and I^O represent the average intensity values of the local regions in the enhanced and input images, respectively.

Exposure Control Loss (L_{exp})

To adjust the exposure level of the image, this study utilizes the exposure control loss (ECL) L_{exp} [8] to regulate the exposure level. The ECL is used to measure the distance between the average intensity values of local regions and the desired exposure level E. The loss L_{exp} is expressed below:

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M}\big|I_k^E - E\big|, \qquad (3)$$

where M means the number of non-overlapping local regions with a size of 16×16, the average intensity value of local regions in the enhanced image is denoted as I_k^E.

Color Constancy Loss (L_{col})

According to the gray world color constancy assumption [11], the average color of each sensor channel in the entire image is gray. In this study, color constancy loss [8] is utilized to correct potential color deviations in the enhanced image and establish relationships between the three color channels. The color constancy loss L_{col} can be expressed below:

$$L_{col} = \sum_{\forall(p\ ,\ q)\in\varepsilon}(I_p^E - I_q^E)^2, \varepsilon = \{(R,G),(R,B),(G,B)\} \qquad (4)$$

where I_p^E and I_q^E are the average intensity values of channels p and q in the enhanced image, and R, G, and B are the color channels of I^E.

Illumination Smoothness Loss ($L_{tv_{\mathcal{A}}}$)

To maintain smoothness among neighboring pixels, this study utilizes illumination smoothness loss to constrain the transformation map $\mathcal{A}$ and prevent excessive variations in illumination between adjacent pixels. The illumination smoothness loss $L_{tv_{\mathcal{A}}}$ is defined as follows:

$$L_{tv_{\mathcal{A}}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\beta}\big(|\nabla_x\mathcal{A}_n^c| + |\nabla_y\mathcal{A}_n^c|\big)^2, \beta = \{R,G,B\} \qquad (5)$$

where N denotes the number of iterations, ∇x and ∇y denote the horizontal and vertical gradient operations.

Total Loss

According to Eqs. (2)-(5), the total loss can be expressed

as follows:

$$L_{total} = \omega_{spa}L_{spa} + \omega_{exp}L_{exp} + \omega_{col}L_{col} + \omega_{tv_A}L_{tv_A} \quad (6)$$

where the weights, ω_{spa}, ω_{exp}, ω_{col}, and ω_{tv_A}, are used to balance these losses.

III. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, some image datasets, SICE and LOL, are selected here. The SICE Dataset is used for model training. The SICE Dataset consists of 589 high-resolution images and their corresponding multiple exposure versions, so there are a total of 4,413 images. The multiple exposure versions are illustrated in Fig. 4. In addition, to evaluate image visual quality, we consider the performance index, Peak Signal-to-Noise Ratio (PSNR) [12] and Structural Similarity Index (SSIM) [13].



Fig. 4 Examples of the SICE Multiple Exposure Image Dataset.

A. Performance of visual quality

Two low-light images are selected for visual quality evaluation. Fig. 5 shows two low-light images and the corresponding results. Fig. 5(a) are test images and the image content is not clear and difficult to be recognized. Fig. 5(b) shows the proposed method enhanced results. Compared to Fig. 5(a), the visual quality of the enhanced images is better than that of test images. Compared to the proposed method, in Fig. 5(c1), the results of Zero-DCE show insufficient brightness recovery. In Fig. 5(d1), the images generated using HE exhibit a yellowish color bias, and the sky are not brightened. In Fig. 5(c2), the results of Zero-DCE and Fig. 5(d2), which is the result of HE, there is some noise in the night sky. The result shows that the proposed method can raise visual quality of low-light images and surpass other methods.

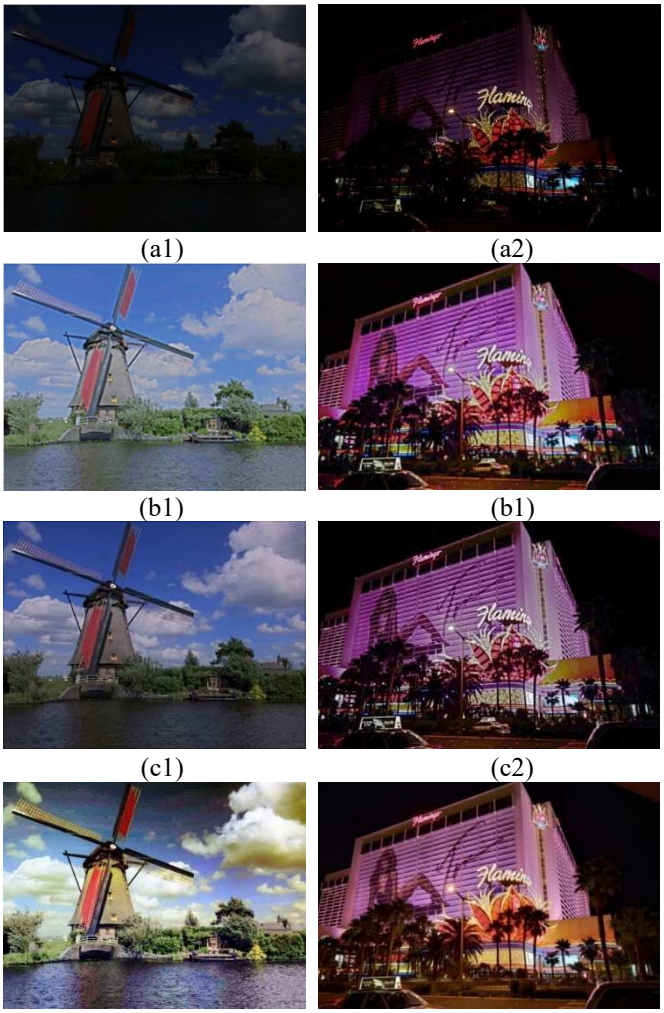


Fig. 5 Results of image enhancement: (a1)(a2) input images, (b1)(b2) proposed, (c1)(c2) Zero-DCE [8] and (d1)(d2) HE [14] .

B. Objective evaluation

Here we make a comparison with some existing methods [8] and [14]. As shown in Table I, the average PSNR and SSIM of the proposed method can achieve 17.01dB and 0.71, respectively. Compared with the existing methods [8] and [14], the proposed method can have better results in terms of PSNR and SSIM. In addition, the parameter size of the proposed is smaller than that of the existing method [8]. Therefore, the experimental results indicate that the proposed method is superior to the existing method [8].

Table I Quantitative comparison of existing method and the proposed CSPDCE-Net in terms of PSNR and SSIM for LOL Dataset

Method	SSIM	PSNR(dB)	Parameter(M)
Input	0.195	7.77	-
HE [14]	0.496	14.10	-
Zero-DCE [8]	0.532	16.88	0.07
CSPDCE-Net (Proposed)	0.71	17.01	0.04

IV. CONCLUSIONS

In this paper, we propose a light-weight zero-reference image enhancement method to improve the visual quality of low-light images. The proposed CSPDCE-Net is composed of several parts: image partition, transformation map estimation, and pixel value adjustment. To achieve transformation map estimation, a light-weight network, CSPDCE-Net is developed. To perform network training, an unsupervised learning approach is adopted.

To evaluate the performance of the proposed low-light image enhancement method, the LOL dataset was used for testing. Compared to the existing methods, the proposed CSPDCE-Net can achieve better performance in terms of PSNR and SSIM. The experimental results indicate that the CSPDCE-Net not only can process low-light images well but also is superior to the other methods.

V. ACKNOWLEDGMENT

This work was supported in part by National Science and Technology Council, Taiwan, under the Grants MOST 111-2221-E-167 -006 -MY2 and MOST 112-2410-H-002-011-MY2.

References

[1] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," *IEEE Access,* vol. 8, pp. 87884-87917, 2020.

[2] W. Ren *et al.*, "Low-light image enhancement via a deep hybrid network," *IEEE Transactions on Image Processing,* vol. 28, no. 9, pp. 4364-4375, 2019.

[3] Y. Wang, W. Xie, and H. Liu, "Low-light image enhancement based on deep learning: a survey," *Optical Engineering,* vol. 61, no. 4, pp. 040901-040901, 2022.

[4] Y. Jiang *et al.*, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing,* vol. 30, pp. 2340-2349, 2021.

[5] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," *arXiv preprint arXiv:2002.11300,* 2020.

[6] T. Ma *et al.*, "RetinexGAN: Unsupervised low-light enhancement with two-layer convolutional decomposition networks," *IEEE Access,* vol. 9, pp. 56539-56550, 2021.

[7] Q. Jiang, Y. Mao, R. Cong, W. Ren, C. Huang, and F. Shao, "Unsupervised decomposition and correction network for low-light image enhancement," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 10, pp. 19440-19455, 2022.

[8] C. Guo *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1780-1789.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015: Springer, pp. 234-241.

[10] J. Jo, S. Jeong, and P. Kang, "Benchmarking gpu-accelerated edge devices," in *2020 IEEE international Conference on Big Data and Smart Computing (BigComp)*, 2020: IEEE, pp. 117-120.

[11] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin institute,* vol. 310, no. 1, pp. 1-26, 1980.

[12] A. Horé and D. Ziou, "Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure?," *IET Image Processing,* vol. 7, no. 1, pp. 12-24, 2013.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, no. 4, pp. 600-612, 2004.

[14] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing,* vol. 39, no. 3, pp. 355-368, 1987.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Classwise Self-Paced Self-Training for Semi-Supervised Image Classification

Cheng-Yu Lu*, Heng-Cheng Hsu†, and Chen-Kuo Chiang‡
* Dept. of Computer Science and Information Engineering, National Chung Cheng University, Taiwan
E-mail: yun80765@gmail.com
† Dept. of Computer Science and Information Engineering, National Chung Cheng University, Taiwan
E-mail: g11410122@ccu.edu.tw
‡ Dept. of Computer Science and Information Engineering, National Chung Cheng University, Taiwan
E-mail: ckchiang@cs.ccu.edu.tw

***Abstract*—In recent years, self-training has become more and more popular in model learning because it only requires a small amount of labeled data to reach high classification accuracy. Since the conventional self-training does not perform in a class-specific manner, class imbalance problem happens frequently as self-training tends to select more pseudo labeled samples from easy classes. It becomes even worse when training dataset is originally class-imbalanced. In this paper, classwise self-paced self-training is proposed to solve this problem. A classwise pseudo-labeling method ensures that sufficient data are selected for each class in a self-pace manner. To avoid increasing number of noisy samples, two indicators, confidence score and uncertainty score, are proposed based on dropout prediction for noisy label filtering. Experimental results demonstrate that the proposed method achieves very competitive performance on CIFAR-10 and ISIC2018 datasets. Keyword: Deep Learning, Image Classification, Semi-Supervised Learning, Self-Training**

I. Introduction

A. Introduction

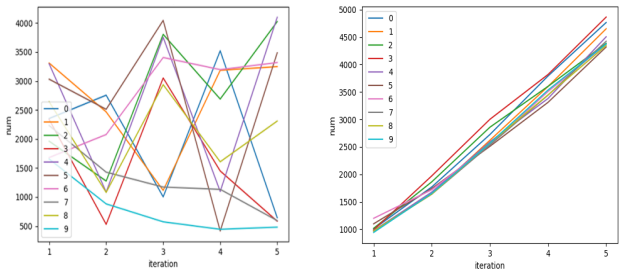
Semi-supervised learning has attracted attention in recent years. Since supervised learning approaches require a large number of labeling data, total time of data acquisition and manual-labeling has also increased significantly. Semi-supervised learning effectively exploits unlabeled data to improve the accuracy of training models. Recent semi-supervised learning methods can be classified into two categories: self-training framework [1] and consistency regularization [7] [10][11][12].

In self-training, the model is initially trained with just a few labeled data. Then, the trained model is used to predict the pseudo-labels of unlabeled data. The set of unlabeled data with pseudo-labels is filtered based on the predicted confidence score by a given threshold. Lastly, the labeled data and the selected pseudo-labeled data are employed as the new training data to retrain the model. This loop repeats until the classifier can no longer find any more unlabeled data higher than the threshold or reach the maximum iteration number. Consistency regularization aims to train a model that is invariant to various data augmentations, such as weak augmentations or strong augmentations. Famous methods includes FixMatch [11] and its variant Flexmatch [12]. In this paper, we focus on the first category, self-training based methods.

Most self-training methods rely on a fixed threshold to select unlabeled data based on the confidence score. However, the selection of threshold is not straightforward. It does not take the easy/hard classes into account, and thus leads to the problem of class imbalance. Training samples of easy classes may have larger number higher than the threshold compared to those in the hard classes during the iteration of self-training.

In Fig. 1(a), the number of pseudo-labeled samples selected for each class (10 classes in total) is unstable in each iteration. It makes the model learning more concentrating on classes with more samples to reduce the total loss and neglects the training for classes with few training samples. This motivates us to proposed a mechanism for self-training to maintain stable and balanced growth of pseudo-labeling sample selection for each class, as depicted in Fig. 1(b).

To solve the class imbalance problem, classwise self-paced self-training is proposed. Rather than using a fixed global threshold, the selection of unlabeled data is performed dynamically within each class and ensures the amount of samples in each class. Two indicators, confidence and uncertainty score, are employed to avoid increasing number of noisy samples. The system framework is depicted in Fig. 2.



(a) Imbalanced self-training. (b) Proposed method.

Fig. 1. Number of pseudo-labeling samples selected for each category during each iteration. 0-9 refers to the class index. The number of pseudo-labeling samples selected for each class changes drastically in conventional self-training framework whereas it maintains stable growth by our method.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

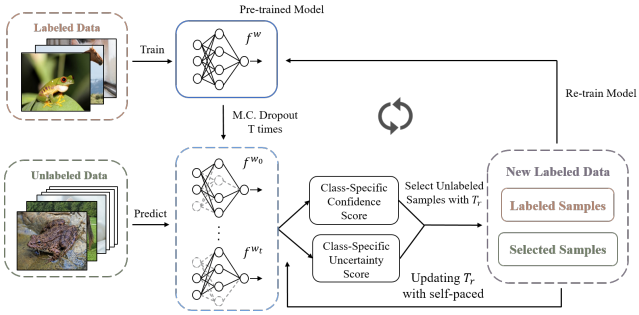


Fig. 2. System Framework

II. Related Work

A. Noisy Student

Xie et al. [6] [6] proposes a framework that combines knowledge distillation to improve self-training, which differs from self-training in that for each iteration of training, [6] uses a new and larger student model to The purpose of this is to allow the student model to have greater potential than the teacher model. In order to make the student model have better generalization ability, three kinds of noisy will be added in the training process of the student model: data augmentation, dropout, and stochastic depth. Raise the difficulty of student model learning, force the model to learn more generalized knowledge, beyond the teacher model.

B. Curriculum Labeling

Cascante-Bonilla, Paola, et al. [2] has improved the problem that the original selftraining method is not flexible enough when selecting unlabeled data. Inspired by curriculcum learning, let the model learn from simple samples first and then gradually progress to difficult samples. The model will have better judgment after learning simple samples, and the learning effect will be better when adding difficult samples than adding them at the beginning, which is a human-like learning method. This paper use extreme value theory as the core concept. They analyze the prediction distribution of unlabeled data and proposes Curriculum Labeling, which can effectively train the model in the self-training process and reduce the chance of false labeling. The implementation method is to sort the prediction results when selecting the unlabeled data stage, and then choose to select the data in a proportional way to join the next round of training.

C. Fixmatch

Sohn, Kihyuk, et al. [11] propose a semi-supervised learning method which based on both consistency regularization and pseudo-labeling. This method performs two types of Data augmentation on the input image simultaneously. 1) Weak augmentation :This is performed with operations such as flip and shift, and the result is used to generate a pseudo-label, which is used as a target for another type of augmentation. 2) Strong augmentation : Inspired by previous studies, this side uses more complex augmentation methods such as Cutout [13], CTAugment [14], and RandAugment [15], which usually result in a higher degree of image distortion. Unlabeled data enters the model judgment after weak-augmented, and gets prediction. If the prediction of this data is higher than the fix ed threshold 0.95, a pseudo-label will be given to the data. Then pass this data through strong-augmented and calculate cross entropy with pseudo-label.

D. Monte Carlo Dropout

For the same amount of data, we may train neural networks with different weights, but both can solve the problem. Gal, Yarin, and Zoubin Ghahramani. [5] study in detail the dropout layer in deep learning, interpreting it as the approximate bayesian inference in the deep gaussian process. This interpretation allows us to model the uncertainty in deep learning models through dropout and extract information from the discarded parts of the model. This paper studies in detail the nature of this uncertainty by combining bayesian models with deep learning models. Through a simple hands-on approach, model uncertainty can be estimated without additional complex computations, modifications to existing models, or sacrificing test accuracy.

III. Proposed Method

A. Confidence and Uncertainty Scores

In semi-supervised learning, only limited labeled samples are available. The classification model is usually pre-trained on ImageNet dataset. Then, label data of target dataset is employed to fine-tune the model. The model trained on labeled data is used to predict unlabeled data and assign pseudo-labels. Since the predictions may have a high bias during early iteration of model training, prediction with high confidence score can be incorrect. Two indicators: *confidence score* and *uncertainty score* are proposed to improve the accuracy of pseudo labels.

Define labeled data $D_L = \{(x_m, y_m) \mid m \in (1, \ldots, M)\}$. M is the number of labeled data. Unlabeled data $D_U = \{x_n \mid n \in (1, \ldots, N)\}$. N is the number of unlabeled data. Dropout is employed to filter noisy labels and reduce the bias for pseudo-label prediction. Dropout is applied during the inference stage while other parameters are frozen. As a result, multiple predictions can be made for each data sample. Denote q_θ^* the dropout layer. $\widehat{W}_t$ is the model weights by randomly discarding some weights through dropout layer. $t \in 1, \ldots, T$ is the index of forward passes. Define total T times forward pass through the network with dropout. The model weight $\widehat{W}_t$ can be defined as follows.

$$\widehat{W}_t = q_\theta^*(W) \tag{1}$$

W is the model weights before dropout layer.$\widehat{W}_t$ is different at each time, and so do the prediction result of times t as $y_t^* \in R^C$, and C is the total number of predicted class. y_t^* can be defined as follows.

$$y_* = \operatorname{softmax}\left(f^{W_t}(x)\right), \quad t \in 1, \ldots, T \tag{2}$$

By averaging the predictions y_t^* of T times, the prediction can be made with a lower bias. The maximum value of the average prediction vector y_t^* can be exploited as the *confidence score* P of this data sample.

$$P = \max\left(\frac{1}{T}\sum_{t=1}^{T} y_t^*\right) \tag{3}$$

The class index of the maximum value is then used as the pseudo-labels of data sample as follows.

$$\hat{y} = \underset{i}{\operatorname{argmax}}\left(\frac{1}{T}\sum_{t=1}^{T} y_t^*\right) \tag{4}$$

In addition, the variance of multiple dropout prediction is also meaningful. If the variance is high, it represents the prediction is diverse and highly uncertain. The pseudo-label can be potentially noisy. The variance for dropout predictions of T times can be calculated given the first prediction $\hat{y}$. The *uncertainty score* U can be defined as follows.

$$U = \operatorname{Var}\left(y^* \mid \hat{y}\right) \tag{5}$$

With these two indicators, noisy samples can be filtered out by keeping only samples with high confidence and low uncertainty scores.

B. Self-paced Classwise Sample Selection

In the framework of self-training, a critical issue is to select good and enough number of pseudo-labeled samples. However, maintaining quality and quantity are a trade-off problem. If the threshold is set too high, the number of selected samples can be relatively small and the training iterations will be long. In addition, it may tend to selection more samples from easy class and lead to class imbalance problem. On the contrary, if the threshold is set too low, many samples will be selected for training. The probability of incorrect pseudo-labeled sample will increase, and the quality of training data cannot be guaranteed. In addition, the fixed threshold relies empirically on experiments to find the better value and is not universal across datasets. In order to these problems, a self-paced classwise sample selection method is proposed via two indicators, confidence score and uncertainty score.

1) Self-Paced Samples Selection

Inspired by self-paced learning [16], our method calculates pseudo labels and examines samples in a fixed pace, which ensures that all samples can be seen at a fixed iteration to end the training. T_r represents the percentage of pseudo-labeled samples that should be added to training in current iteration. The initial value 100 represents that no pseudo-label sample is currently added. A hyperparameter T_{step} is defined to set the stepping percentile for thresholding after each iteration:

$$T_r -= T_{step} \tag{6}$$

In our experiments, T_{step} is set to 20. In each new iteration, (6) is executed until $T_r = 0$. That is, after all the pseudo-labeled data are examined to the training, the whole training process is finished.

2) Classwise Selection with Noisy Label Filtering

To ensure that sufficient data are sampled for each category, a class-specific sampling method is applied. However, the sampling thresholds in difficult classes are likely to be lower than that of the global selection method. Compared with a fixed yet higher threshold in the global selection method, classwise sampling method leads to higher probability of sampling more noisy labeled samples into the training. Therefore, a classwise sample selection method is proposed based on noisy label filtering is proposed.

Denote the confidence score P as $\{p_n : n \in 1, \ldots, N\}$, and uncertainty score U as $\{u_n : n \in (1, \ldots, N)\}$, where N is the number of unlabeled data. Split the confidence score P and uncertainty score U into subsets according to the pseudo-label $\hat{y}$ of each sample : $P = \{P_1 \cup P_2 \ldots \cup P_C\}$, $U = \{U_1 \cup U_2 \ldots \cup U_C\}$, where C is the number of classes in the dataset. As mentioned earlier, self-paced sample selection uses T_r to select a certain percentage of samples. The percentage is used to determine the class-specific threshold and filter samples. For each subset of confidence score P_c and uncertainty score U_c, threshold of cutting line is calculated respectively for P_c and U_c as follows.

$$T_p^c = \text{ Quantile } (P_c, T_r, descending = False) \tag{7}$$

$$T_u^c = \text{ Quantile } (U_c, T_r, descending = True) \tag{8}$$

The Quantile function first sorts the input in descending or ascending order based on parameter $descending$ is $True$ or $False$. The value of the T_r-th percentile is returned. For example, if the input set contains values of 1 to 10 and T_r is set to 80 (which means 80%), 8 is returned as threshold if $descending = False$. Only values larger than 8 will be chosen. On the other hand, Quantile function return 3 if $descending = True$ and 1 and 2 are selected. T_p^c and T_u^c represents the thresholds of the confidence and uncertainty score of class c. In each iteration, only samples above the confidence threshold and below the uncertainty threshold are selected into training in the next iteration.

C. Model Re-training

To avoid confirmation bias of model, re-initialing is adopted before new iteration rather than continuous fine-tuning the model. It reduces the impact of incorrect pseudo-labeled samples in the early training period. The process terminates when no more unlabeled data can be added into the model training. The complete algorithm is described in Algorithm 1.

IV. Experimental Results

Backbone model WideResNet-28-2[17] and ResNet-18[18] are used for CIFAR-10 and ISIC2018, respectively. Both of

Algorithm 1 Classwise Self-Paced Self-Training

Input: $D_L = \{(x_m, y_m) \mid m \in (1, \ldots, M)\}$
$D_U = \{x_n \mid n \in (1, \ldots, N)\}$
T_{step} : Stepping threshold percent
$T_{dropout}$: Times of M.C. dropout
f^W ←Train model with parameters W on task-specific labeled data D_L
$T_r = 100$
while $T_r > 0$ **do**
  T_r -= T_{step}
  for $t \leftarrow 1$ to $T_{dropout}$ **do**
    $\mathcal{W}_t \sim \text{Dropout}(\mathcal{W})$
    $y_t^* = \text{softmax}\left(f^{W_t}(x)\right)$
  end for
  $\mu = \text{mean}(y^*)$
  $P = \max(\mu)$
  $\hat{y} = \text{argmax}(\mu)$
  $U = \exp(Var(y^* \mid \hat{y}))$
  $X_t = D_L$
  for $c \leftarrow 1$ to C **do**
    $P^c = \mathbb{1}(\hat{y} = c) \cdot P$
    $U^c = \mathbb{1}(\hat{y} = c) \cdot U$
    $T_p^c = \text{Quantile}(P^c, T_r)$
    $T_u^c = \text{Quantile}(U^c, T_r)$
    $D_U^c = \mathbb{1}(\hat{y} = c) \cdot D_U$
    $D_S^c = \mathbb{1}(P^c > T_p^c)\mathbb{1}(U^c < T_u^c) \cdot D_U^c$
    $X_t = X_t \cup D_S^c$
  end for
  Re-train model on X_t and update parameters W
end while
Output : f^W =0

them are pretrained on ImageNet. Fast Autoaugment[19] and Mixup[20] are employed as data augmentation methods during the model training. The dropout is set to 10.

A. Dataset

CIFAR-10[21] contains 60,000 32×32 color images in 10 classes. Each class has 6,000 images. 4,000 images are used as labeled data for training, 41,000 images as unlabeled data. The number of validation and test set is set to 5,000, and 10,000, respectively.

SVHN[22] is a Google Street View House Number classification dataset. There are 10 categories, representing numbers 0 to 9. In the experiment, the training set is divided into 1,000 labeled data and 71,257 unlabeled data, the validation set was 1,000, and the test set is 26,032 images.

ISIC2018[23] is a skin lesion dataset, containing 10,208 600×450 color images in 7 classes. 1,000 images are used as labeled data for training, and 8,500 images as unlabeled data. The validation set contains 515 images, and 193 in the test set.

TABLE I
CLASSIFICATION ERROR ON CIFAR-10 DATASET.

Approach	Method	Error rate
Pseudo-Labeling	PL	17.78
	CL	5.62
	PL-CB	6.28
	UPS	6.42
	Ours	4.82

B. Experimental Results on CIFAR-10

1) Comparison with other method

Table I presents the classification error comparison with other methods on CIFAR-10 dataset. Recent pseudo-labeling methods have been chosen for comparison, including **PL** [1], **CL**, [2], **UPS** [3] and **PL-CB** [4]. The proposed method achieves the best results on CIFAR-10 dataset.

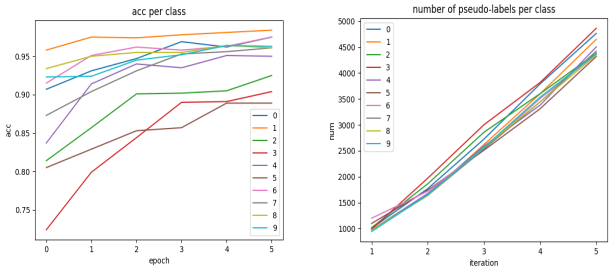
2) Visualization of Noisy Label Filtering

The visualization noisy label filtering is depicted in Fig. 4. The leftmost of each row is the predicted pseudo-label for that row of pictures. CNF on the right of the picture represents the confidence score, and the UNC is the uncertainty score. In this cases, the pseudo labels are incorrect but having high confidence score. Since their uncertainty score is higher than threshold, they are considered as noisy sample by our method.

Fig. 3(a) illustrates the test accuracy of each category at different iterations. Compared with Fig. **??** ,our method has a significant improvement in some categories. Fig. is a graph of the number of pseudo-labels for each category at different iterations. In the original method (Fig. 1.4), due to the use of strong augmentation such as Mixup, the number of pseudo-label data of each category is extremely unstable in different iterations, the changes are very drastic, and the difference between the number of data each class is very large. However, it can be seen from Fig. 5.4 that our method can effectively improve this problem. Even using the strong augmentation, it can maintain stable and balanced growth, and the difference between the number of the most and the least pseudo-labels in each class of the data witch at the last iteration does not exceed 1,000.

3) Ablation Study

Table II evaluates three components of our method. The baseline method PL uses a fixed threshold of 0.9 to select the pseudo-labeled data and achieves classification error of 17.78.



(a) Test Accuracy on CIFAR-10. (b) Number of the pseudo-labels.

Fig. 3. Visualization of the experiment result on CIFAR-10.

TABLE II
ABLATION STUDY ON CIFAR-10 DATASET.

Methods	Error rate
PL	17.78
PL + FA + Mixup	8.59
PL + FA + Mixup + NLF	8.18
PL + FA + Mixup + SP	5.13
PL + FA + Mixup + SP + NLF	4.78

TABLE III
CLASSIFICATION ERROR ON SHVN DATASET.

Approach	Method	Error rate
Pseudo-Labeling	PL	9.94
	CL	3.83
	Ours	3.77

After adding two data augmentation methods, Fast Autoaugment (FA) and Mixup, the error rate can be decreased to 8.59. Next, we evaluate the impact of noisy label filtering (NLF) and self-paced samples selection (SP), respectively. When including noisy label filtering and self-paced samples selection into self-training, the error rates can be further decreased to 8.18 and 5.13. Self-paced samples selection significantly reduces classification error than the other components. Finally, adding all the components together, it achieves the error rate of 4.78.

C. Experimental Results on SVHN

The test accuracy and the number of the pseudo-labels for each category are shown in Fig. 5(a) and Fig. 5(b). The biggest difference from the results in CIFAR-10 is that the SVHN dataset is not a category-average dataset.Owing to the proportional selection relationship, our method can adjust the amount of pseudo-label data for each category according to the distribution of the dataset. It allows for flexibility in the data set requirements, eliminating the need for category averages.

Table III shows the different pseudo-labeling-based methods experiment result on SVHN. Our method get an error rate of 3.77, that is superior to other proposed method.

D. Experimental Results on ISIC2018

Table IV presents the comparison with other methods on ISIC2018 dataset. Our method achieves classification error 26.43, which is very competitive to recent methods. Although

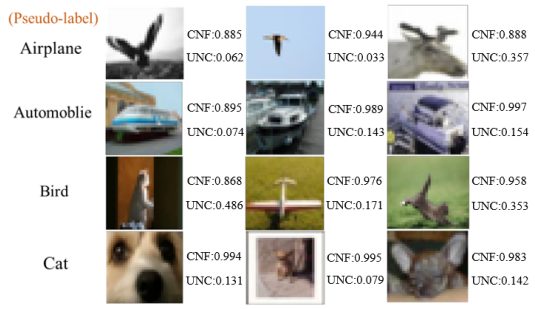
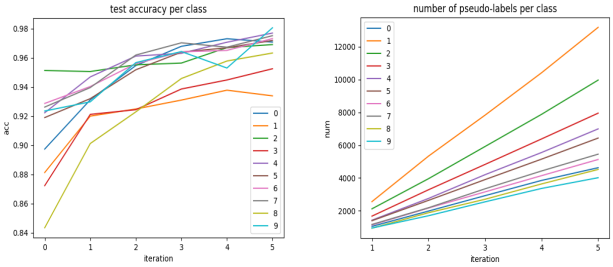


Fig. 4. Visualization of noisy pseudo-labeled sample filtering.



(a) Test Accuracy on SVHN. (b) Number of the pseudo-labels.

Fig. 5. Visualization of the experiment result on SVHN.

TABLE IV
CLASSIFICATION ERROR ON ISIC2018 DATASET.

Approach	Method	Error rate
Pseudo-Labeling	PL	36.27
	CL	32.13
	Ours	26.43

ISIC2018 is not a category-average dataset, the proposed method outperforms other methods.

V. CONCLUSION

A classwise self-paced method is proposed for self-training. Self-paced learning is employed to improve the selection of pseudo-labeled sample in self-training. This enable the model training from simple and reliable data first, and then gradually adding difficult data in to the training process. Dropout is exploited to evaluate data uncertainty. As a noisy label filtering, it helps pseudo-labeled sample selection be more accurate. In a class-specific selection fashion, these two methods can be combined to solve the class imbalance problem that is prone to occur frequently when self-training is applied. This also improves the accuracy of each category in the testing phase, and reduces the gap in accuracy between hard and easy categories. In the future, we hope that the concept of classwise self-paced samples selection can be combined with other semi-supervised learning methods, such as consistency regularization, to achieve higher classification accuracy.

ACKNOWLEDGEMENT

This work was partially supported by the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and Center for Innovative Research on Aging Society (CIRAS). This work was also supported in part by the National Science and Technology Council, Taiwan under Grant NSTC 111-2634-F-006-012 and NSTC 111-2634-F-194-003.

REFERENCES

[1] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3, p. 896.

[2] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 6912–6920.

[3] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh SRawat, and Mubarak Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv:2101.06329*, 2021.

[4] Eric Arazo, Diego Ortego, Paul Albert, Noel EO'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[5] Yarin Gal, Zoubin Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", in *international conference on machine learning*, pp. 1050–1059, PMLR, 2016.

[6] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, andQuoc V. Le, "Self-training with noisy student improves imagenet classification," in *2020 Proceedings of the IEEE/CVF conference oncomputer vision and pattern recognition(CVPR)*, 2020, pp. 10687–10698.

[7] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," *arXiv:1610.02242*, 2016.

[8] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol .30, 2017.

[9] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[10] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[11] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[12] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18408–18419, 2021.

[13] Terrance DeVries, Graham W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv::1708.04552*, 2017.

[14] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, Colin Raffel, "ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring" *International Conference on Learning Representations(ICLR)*, 2020

[15] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le, "Randaugment: Practical automated data augmentation with a reduced search space" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR) workshops*, pp. 702–703, 2020

[16] Hao Li and Maoguo Gong, "Self-paced convolutional neural networks," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra, Ed. 2017, pp. 2110–2116, ijcai.org.

[17] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv:1605.07146*, 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Los Alamitos, CA, USA, jun 2016, pp. 770–778, IEEE Computer Society.

[19] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim, "Fast autoaugment," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv:1710.09412*, 2017.

[21] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[22] Yuval Netzer Tao Wang Adam Coates Alessandro Bissacco Bo Wu Andrew Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[23] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv:1902.03368*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

CapFormer: A Space-Time Video Description Model using Joint-Attention Transformer

Mahamat Moussa, Chern Hong Lim and KokSheik Wong
School of Information Technology, Monash University Malaysia
E-mail: {mahamat.moussaabbasal, lim.chernhong, wong.koksheik}@monash.edu

***Abstract*—Transformers in video understanding are becoming popular due to the recent success of vision transformers. However, video transformers are still emerging and require various tricks to understand video tasks, such as action detection, classification, and description. The need for such tricks is due to videos' temporal and spatial space, which requires different techniques to understand the context better. The most critical component is the attention block, which gives different results if tackled differently. In video description techniques, the model designs are mainly complex due to the integration needed between visual and language contexts. In this work, we proposed a simple yet efficient video description model that relies entirely on an attention mechanism. The model uses a joint-attention mechanism to learn the spatial and temporal context of video frames with their description context. Surprisingly, this integration setting suggests a more straightforward way to achieve the same task achieved by complex networks, and thus, it is efficient in terms of training. To validate the design, we evaluated the proposed architecture on a large video description dataset (MSR-VTT) and compared its result with various works, and it showed promising results over other designs in terms of the ROUGE metric.**

I. Introduction

In recent years, many deep learning video understanding techniques have been developed to endow machines with the capability to understand the visual context at a high level [1]. Combining videos and natural language processing (NLP) has become a promising method in video description and video question answering (VQA) [2]. The success behind this combination is that video and NLP share similar high-level characteristics [3], as video and NLP are processed as a sequence of frames and sequence of words, respectively.

Previously, video description was achieved by combining multiple models such as convolutional neural networks (CNN), Recurrent Neural Networks (RNN), and Long-Short Term Memory (LSTM). The introduction of Vision Transformers (ViT) pushed progress positively by replacing CNN with Transformers. Furthermore, [3] introduced TimeSformer, a convolution-free network built on temporal and spatial self-attention transformer. The model extends the concept of ViT into video classification and achieves state-of-the-art performance [3]. ViT and TimeSformer have a similar architecture, demonstrating the effectiveness of simplified neural networks by utilizing attention mechanisms to achieve tasks that used to be done in complex ways. However, the works mentioned above consider only the video classification task, which is only halfway to describing a video. To achieve this feature, the model needs to integrate a language model to describe the content and events in the video. In this regard, the work of [3] is limited because the model is not designed to describe videos due to its architecture that consists of an encoder only, which is similar to [4].

Motivated by this, we introduce "CapFormer" a video description model using a joint-attention mechanism to learn the video features and their word embedding in a unified design. The architecture of CapFormer consists of two parts: 1) Encoder that uses divided self-attention over spatial (space) and temporal (time) mechanisms to learn spatial-temporal features from videos, and 2) Decoder that jointly learns the context of video concerning its word embedding in the description text. Such integration enables the model to generate a sentence that best describes the video content. CapFormer introduces a simplified concept to achieve a video description task entirely based on attention mechanism, unlike most video description networks that are complex in design or combine different networks [5], [6]. Our simplified design network shows its efficiency with various experiments by processing videos with higher frame rates and more extended duration with a good performance.

We evaluate our model on MSR-VTT [7] (a large video description dataset) and compare its performance against state-of-the-art video description methods. Results show that our model outperforms the conventional methods and transformer-based architecture that are complex in design.

II. Related Work

Many video understanding techniques have been developed based on transformer architecture. These techniques can be highlighted based on 1) Video attributes understanding techniques to perform detection, classification, and recognition and 2) Video attributes Integration with language to perform captioning, description, and question answering.

Video attributes understanding Attributes understanding in videos has been progressing since the introduction of vision transformer (ViT) by [4], which has contributed towards developing more video understanding models and frameworks that use transformers as main or part of their designs, achieving state-of-the-art performance in many visual understanding tasks due to its attention mechanism [8]. The work of video attribute understanding and learning covers various domains such as video understanding [9], video classification [3], action and activity recognition [10]–[12]. Although these works are

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

transformers-based, their designs are limited to attribute understanding only, facilitating and supporting visual understanding. Such design is mainly interested in the encoder part to perform classification or prediction.

Video attributes integration with language The video and language integration aims to understand video attributes and learn the contextual relationship between videos and their descriptions. Such concept enabled many applications in the video domain such as video description and captioning [13], [14], video question-answering [15]–[17], and video retrieval [18], [19] all based on transformers.

Regarding network design, integrating video and language requires different approaches than learning video attributes only. One common approach is to combine visual understanding networks such as CNN with transformers [20]. However, many recent approaches show that ViT can replace CNN to perform video learning tasks [3]. The focus then shifted to designing more efficient Video Transformer networks. These works include: Spare Transformers [21], Hierarchical Transforms [22], [23], Query-driven Transformers [24], [25], Local Transformers [26], Memory efficient Transformers [27], and more.

For instance, [22] proposed a multi-modal hierarchical temporal aware architecture that utilizes two encoders: one for video understanding and another for text. Their technique relies on extracting two views from the video, long view and short view, then learning the features of each and combining them to do cross-modal learning between text and video. Similarly, [23] adopted a design that uses two encoders to learn textual and visual data in a multi-modal design. Their design is based on pixel-level visual understanding, which neglects the local (within a frame) and global (among frames) context of the visual data, especially in long videos.

Another design that integrates language with video is the work by [24]. In their design, they use a transferable visual model from the natural language supervision approach by generating video captions using GPT-2, then query the result with ground truth description and perform video caption interaction between the ground truth description and the generated one. This work aims to contribute to text-video retrieval progress.

The common problem identified in these works is the complexity of design. Such complexity can be understood prior to transformers; however, with the presence of ViT, we aim to find a more straightforward and efficient solution to do the same task. We design our architecture to leverage the attention mechanism for the entire learning journey. We use the space-time technique to enable our model to learn the local and global context of the video and the relationship with their description words, making it a simple video description architecture that extends the attention over space, time, and word embedding.

III. CapFormer Development

Using a joint-attention mechanism, we put forward a video description model named CapFormer . Our model integrates an encoder for video understanding and a decoder for language generation to perform video description.

The model generates a sentence that describes the learned video features. In detail, the model takes a video clip X and its description T as input. The clip $X \in \mathbb{R}^{H\times W\times C\times N}$ has F frames of a dimension $H \times W$ pixels and C channels, where $C = 3$ denotes a color frame (RGB). The description T is a sentence consisting of i words, where each sentence is tokenized with a fixed-length vector L. The overall architecture of the model is illustrated in Fig. 1.

A. CapFormer Encoder

The encoder of CapFormer learns video representation features at a given time step (t). Each frame of X is divided into N non-overlap patches of size $P \times P$ as $N = HW/P^2$. These patches are flatten as vectors $\mathbf{x}_{(p,t)} \in \mathbb{R}^{3P^2}$ where $p = 1, \ldots, N$ denotes the spatial position and $t = 1, \ldots, F$ denotes the temporal position over frames. Each patch of $\mathbf{x}_{(p,t)}$ is mapped into an embedding vector $\mathbf{z}^{(0)}_{(p,t)} \in \mathbb{R}^D$ linearly using Learanble Matrix $(LM) \in \mathbb{R}^{D\times 3P^2}$ as given by (1).

$$\mathbf{z}^{(0)}_{(p,t)} = LM\mathbf{x}_{(p,t)} + \mathbf{e}^{pos}_{(p,t)} \tag{1}$$

The $\mathbf{e}^{pos}_{(p,t)} \in \mathbb{R}^D$ is the positional embedding for the spatial p and temporal t of each patch. The resulting linear embedding vector $\mathbf{z}^{(0)}_{(p,t)}$ represents the input to the encoder. At each patch p of each layer l of the encoder, a vector of query, key, and value is computed from $\mathbf{z}^{(\ell-1)}_{(p,t)}$ that is encoded by a previous block $\ell - 1$, as given by (2) where $\wp$ represents query (q), key (k), or value (v). The LN denotes layer normalization, a denotes a position through multiple attentions, and D_h indicates the latent dimension of each attention.

$$\wp^{(\ell,a)}_{(p,t)}\& = W^{(\ell,a)}_{\wp}\mathrm{LN}\left(\mathbf{z}^{(\ell-1)}_{(p,t)}\right) \in \mathbb{R}^{D_h} \tag{2}$$

Next, we compute the attention $\boldsymbol{\alpha}$ over the temporal and spatial position of X using (3) and apply softmax (SM). The $\measuredangle$ represents both temporal and spatial position but in a sequential process. First, the temporal attention is calculated (hence $\measuredangle$ = temporal), and then the spatial attention is calculated (hence $\measuredangle$ = spatial), and the En indicates the end of the current process.

$$\boldsymbol{\alpha}^{(\ell,a)\measuredangle}_{(p,t)} = \mathrm{SM}\left(\frac{\mathbf{q}^{(\ell,a)\top}_{(p,t)}}{\sqrt{D_h}} \cdot \left[\mathbf{k}^{(\ell,a)}_{(0,0)}\left\{\mathbf{k}^{(\ell,a)}_{(p,t')}\right\}_{\measuredangle'=1,\ldots,En}\right]\right) \tag{3}$$

After that, the final learned weights of X are computed by concatenating vectors of all attention heads $\mathcal{A}$ and perform a dot product on the weight matrix for the linear projection of the concatenated outputs W_O as in (4) then pass it through a multilayer perceptron (MLP) in (5) to obtain the resulting embedding $\mathbf{Z}$ of the learned video attributes, which will be passed to the bimodal block in the decoder as key, value.

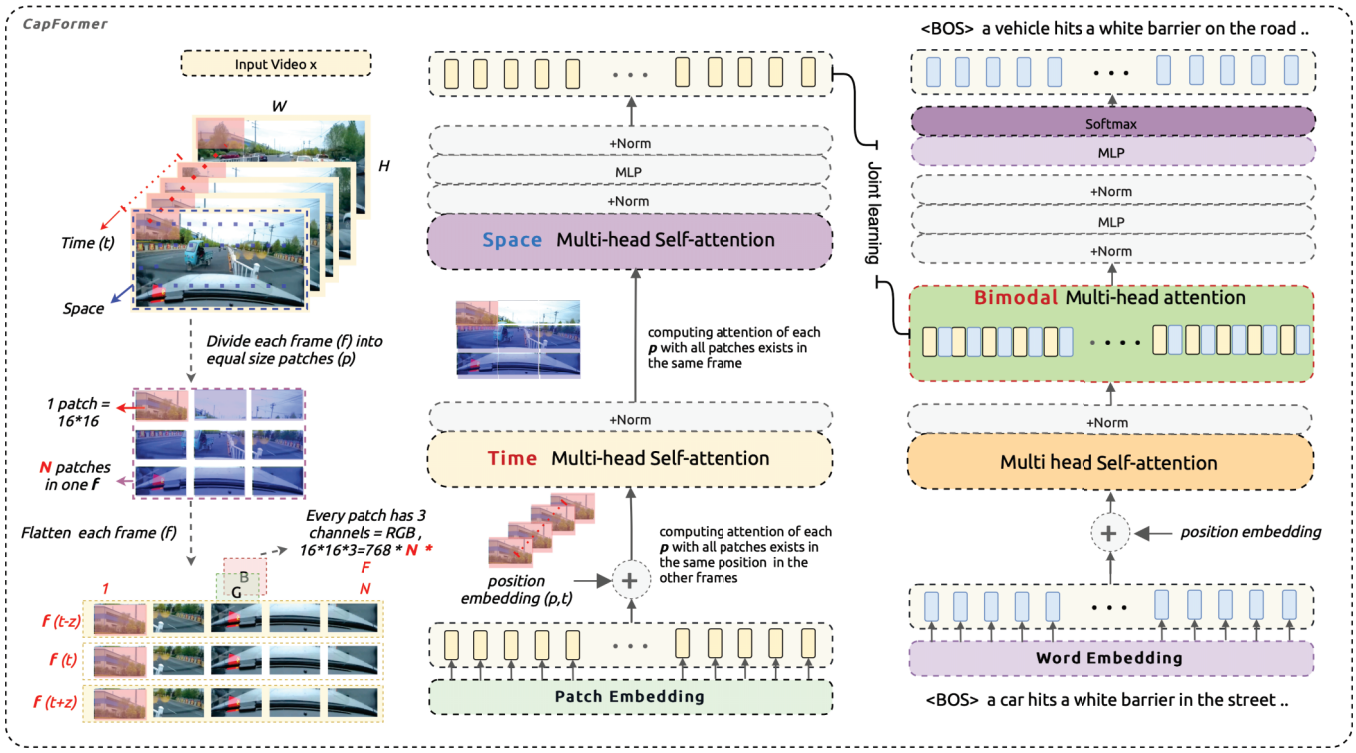


Fig. 1. The Architecture of CapFormer.

$$\mathbf{z'}^{(\ell)}_{(p,t)} = W_O \begin{bmatrix} \mathbf{s}^{(\ell,1)}_{(p,t)} \\ \vdots \\ \mathbf{s}^{(\ell,A)}_{(p,t)} \end{bmatrix} + \mathbf{z}^{(\ell-1)}_{(p,t)} \tag{4}$$

$$\mathbf{Z} = \mathrm{LN}\left(\mathrm{MLP}\left(\mathbf{z'}^{(\ell)}_{(p,t)}\right)\right) + \mathbf{z'}^{(\ell)}_{(p,t)} \tag{5}$$

B. CapFormer Decoder

The decoder of CapFormer takes the resulting embedding from (5) as key, value, and the description $\mathbf{T}$ that contains i words. The $\mathbf{T}$ is converted into a vector T_v consisting of i words, tokenized in a fixed-length vector using an embedding layer: $T_v = \{i_1, i_2, \ldots, i_C\}$ where i_C is the length of T_v as illustrated in figure 2.

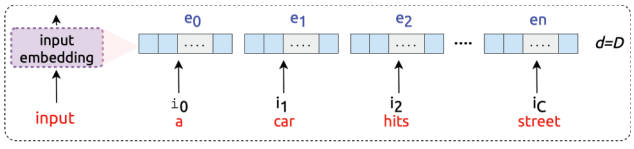


Fig. 2. Video description embedding illustration

The T_v is projected into a query (q), key (k), and value (v) to compute its self-attention At using (6).

$$At(q_i, k_j) = \left(\frac{q_i.k_j}{\sqrt{d_k}}\right) \tag{6}$$

Next is to obtain the computed attention weight AW using (7) which is then used to obtain the weighted sum. Finally, the attention heads are concatenated and linearly projected to obtain the final output query (q) = A_q as shown in (8), which is used as input for the bimodal block.

$$AW = \mathrm{SM}\left(At(q_i, k_1), At(q_i, k_2), \ldots, At(q_i, k_T)\right) \tag{7}$$

$$\mathbf{A_q} = W_O \begin{bmatrix} head_1 \\ \vdots \\ head_n \end{bmatrix} \tag{8}$$

Computing Joint-attention of X and T: The bimodal multi-head attention is the attention block where the interaction between the output of the encoder Z which is projected into key k_e and value v_e, and the output of the multi-head self-attention block of the decoder A_q occurs. Since the video description is generated token by token, the bimodal multi-head attention considers the temporal aspect of the decoder by concatenating the query vector at a time step $A_{q_{(t)}}$ in the decoder with the encoder's k_e and v_e as shown in (9),

$$Bimodal(A_{q_{(t)}}, k_e, v_e) = W_O \begin{bmatrix} head_1 \\ \vdots \\ head_h \end{bmatrix} \tag{9}$$

where the $head_n$ is defined in the Bimodal Scaled Dot Product Attention (BSA) function in (10), n denotes the

attention heads, while the full expression of BSA defined in (11).

$$head_n = BSA(A_{q_{(t)}}, k_e, v_e) \tag{10}$$

$$BSA(A_{q_{(t)}}, k_e, v_e) = SV\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{11}$$

Let us define the resulting attentive context vector of the bimodal from (9) as C_t for time step t; we use C_t as input to pass it through an MLP with the residual connections and then apply softmax function to generate the probability distribution P_t over the vocabulary vector at time step t until the stop token is generated. As a result, the complete sentence that describes the input video X is generated as shown in (12)

$$P_t = \mathrm{SM}\left(MLP(LN(C_t)) + C_t\right) \tag{12}$$

As a result, a full predicted sentence describing the video's learned attributes is generated.

IV. Experiments

We evaluated the proposed CapFormer model on a Large Video Description Dataset: MSR-VTT [7] to examine its performance compared to other models that use the same dataset for video description tasks. MSR-VTT consists of 10k video clips. We split the data as follows: 65% for training (viz., 6513 clips), 30% for testing (viz., 2990 clips), and 5% for validation (viz., 497 clips). Each clip has 20 sentences of description available.

Our model is configured with a stack of 12 encoder layers and six decoder layers. However, depending on the experiment, the number of layers could be changed empirically. We use a video of size $W \times H \times F = 224 \times 224 \times 8$, where each frame is divided into patches of 16×16 pixels. This results in 14×14 patches, each with a size of $16 \times 16 \times 3$ pixels.

A. Training and Hyperparameters

CapFormer is trained on an Advanced Computing Platform (HPC) powered by NVIDIA A100 GPU for 30 training epochs using the PyTorch framework.

The learning rate (lr) was set as $2e-5$ and optimized with the "Adam" optimizer, and cross-entropy was adopted as the loss function, with batch size 64. To optimize our training progress for video features, we utilize a pre-trained video understanding model by [3] as the starting point. Table I summarizes the hyperparameters of CapFormer .

B. Evaluation Metrics

We use Recall-Oriented Understudy for Gisting Evaluation (ROUGE) as the evaluation metrics. In particular, we use Rouge-L: Longest Common Subsequence (LCS) measure. Rouge-L has two types of evaluation: sentence-level LCS and summary-level LCS. We adopt sentence-level LCS using LCS-based F-measure, calculated between the original video description S of length m and the generated description D of length n. Specifically,

$$R_{lcs} = \frac{LCS(S,D)}{m}, \tag{13}$$

$$P_{lcs} = \frac{LCS(S,D)}{n}, \tag{14}$$

$$F_{lcs} = \frac{\left(1+\beta^2\right)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}, \tag{15}$$

where $LCS(S,D)$ is the length of a longest common subsequence of S and D, and $\beta = P_{lcs}/R_{lcs}$ when $F_{lcs}/R_{lcs} = F_{lcs}/P_{lcs}$. The R, P, and F refer to the Recall, Precision, and F-measure, respectively. Using these metrics, we report the outcome of Eq. (15) as the result of our model, referred to as ROUGE-L.

TABLE I
The Hyperparameters for CapFormer.

Key	Value
Number of frames	32
Spatial crop	224
Batch size	64
Optimizer	adam
Loss function	cross entropy
Epochs	30
Dropout	0.1

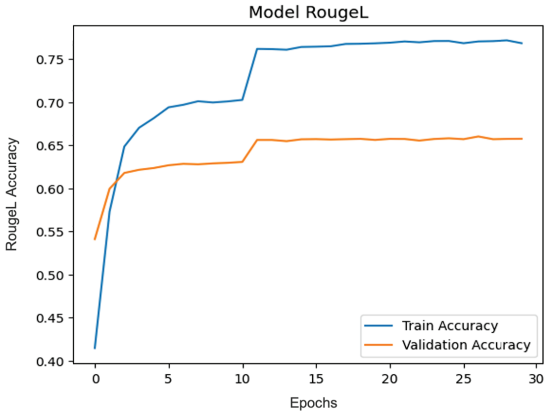


Fig. 3. The training and validation accuracy scores of the model over 30 epochs.

V. Results

Table II records the ROUGE-L score for CapFormer compared with other models tested on the MSR-VTT dataset. Results show that the CapFormer model outperforms other methods considered for comparison. Specifically, we achieved 66.4% for ROUGE-L, the highest among all considered methods at the time of the experiment.

A representative description output by CapFormer is shown in Fig. 4. Here, the predicted sentence and its reference are shown for comparison purposes. The samples show various outputs that can be observed in three groups: 1) relevant description (a,b,h), 2) relevant but not complete (e,f), and 3)

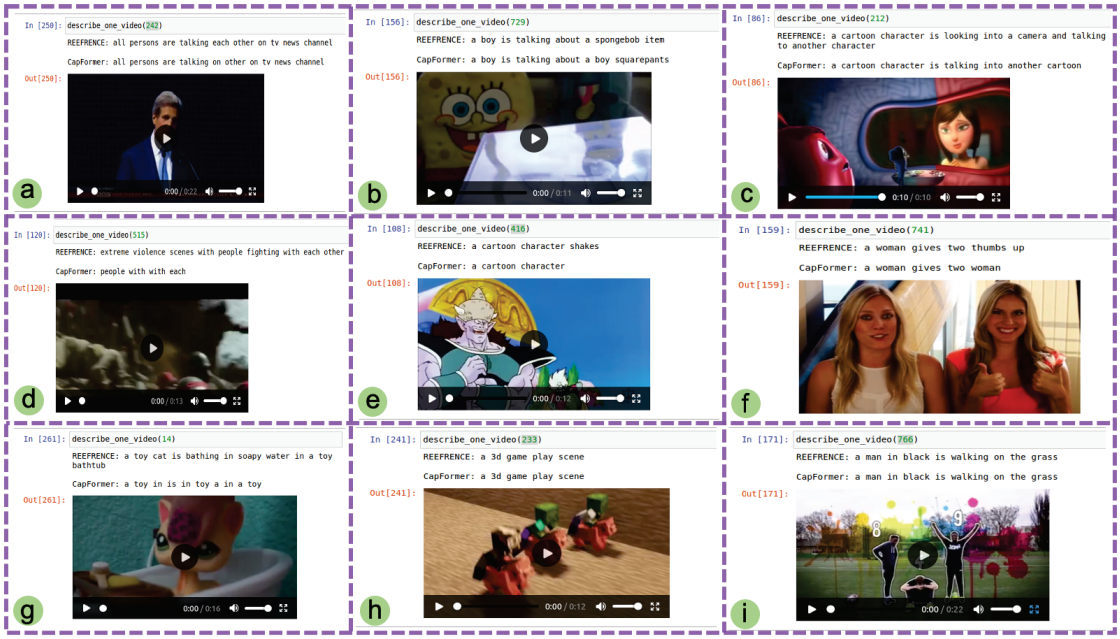


Fig. 4. Samples of CapFormer video description output on the MSR-VTT testing dataset.

TABLE II
PERFORMANCE OF VARIOUS VIDEO DESCRIPTION METHODS ON MSR-VTT DATASET EVALUATED USING ROUGE-L.

Method / Models	**Year**	**ROUGE-L %**
v2t-navigator [28]	2016	61.1
CST-GT-None [29]	2017	62.4
HRL [30]	2018	61.7
POS+VCT [31]	2019	62.0
TransEDRL [32]	2020	61.4
MV-GPT [23]	2022	64.0
HiTeA [22]	2022	65.0
CapFormer	**2023**	**66.4**

irrelevant descriptions (g,d), which overall attributed to the accuracy of the model.

The improvement over other methods is potentially attributed to using divided space-time attention and description attention by jointly learning the context of words and their associated visual features using attention only, implying that the CapFormer approach may produce the same or better results than other methods that complex in design and may combine multiple methods apart from attention. Furthermore, the design of CapFormer suggests that such a simple design could also solve similar problems, such as video question-answering and video retrieval, due to the power of the attention method and similarities in the process. From the preliminary experiments, we also discovered that the performance of CapFormer can be further improved with more hyperparameter tuning, computational power, and data. We believe this because we managed to improve the result from the initial experiment (61.2%) to (66.4%) with parameters tuning, implying that more experiments could potentially enhance the model even better. Table I reported the set of hyperparameters we used to achieve our best result (66.4%).

VI. CONCLUSION

The main goal of this paper is to propose a simplified video description model based on the joint-attention mechanism. Principally, we argued that transformers in the video domain are emerging, and the need for efficient design to perform video descriptions is paramount. There are various designs and suggestions in the literature; however, processing the spatial and temporal along with their description requires complex design. The proposed joint-attention mechanism is a powerful yet simple technique in this regard. Our CapFormer architecture consists of an encoder to learn the visual features of the video and a decoder to jointly learn how to decode a relevant sentence to describe these features. When evaluated and validated on a large video dataset (MSR-VTT), the design has shown promising results. We learned from the experiments that CapFormer could generate relevant video descriptions with a better result than the other models when tested on the MSR-VTT dataset evaluated on the ROUGE-L metric. The results also suggest that more experiments could lead to even better results, which we shall explore in our future work.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Higher Education (MOHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS) (Ref: FRGS/1/2020/ICT02/MUSM/03/1).

REFERENCES

[1] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 433–440.

[2] S. Aditya, Y. Yang, and C. Baral, "Integrating knowledge and reasoning in image understanding," *arXiv preprint arXiv:1906.09954*, 2019.

[3] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" *arXiv preprint arXiv:2102.05095*, 2021.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, Weissenborn *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] V. Iashin and E. Rahtu, "Multi-modal dense video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[6] X. Ai, H. Wang, H. Chen, and W. Xu, "Video captioning method based on multi-modal information fusion," in *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2021.

[7] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language." IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[9] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unifying convolution and self-attention for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[10] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[12] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 505–13 515.

[13] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

[14] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "Grit: Faster and better image captioning transformer using dual visual features," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*. Springer, 2022, pp. 167–184.

[15] J. Lei, T. L. Berg, and M. Bansal, "Revealing single frame bias for video-and-language learning," *arXiv preprint arXiv:2206.03428*, 2022.

[16] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvqa+: Spatio-temporal grounding for video question answering," *arXiv preprint arXiv:1904.11574*, 2019.

[17] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," *arXiv preprint arXiv:2005.00200*, 2020.

[18] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1728–1738.

[19] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[20] P. Dandwate, C. Shahane, V. Jagtap, and S. C. Karande, "Comparative study of transformer and lstm network with attention mechanism on image captioning," *arXiv preprint arXiv:2303.02648*, 2023.

[21] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "Swinbert: End-to-end transformers with sparse attention for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.

[22] Q. Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, and F. Huang, "Hitea: Hierarchical temporal-aware video-language pre-training," *arXiv preprint arXiv:2212.14546*, 2022.

[23] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 959–17 968.

[24] W. Wu, H. Luo, B. Fang, J. Wang, and W. Ouyang, "Cap4video: What can auxiliary captions do for text-video retrieval?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 704–10 713.

[25] H. Seong, J. Hyun, and E. Kim, "Video multitask transformer network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[26] C.-F. Chen, R. Panda, and Q. Fan, "Regionvit: Regional-to-local attention for vision transformers," *arXiv preprint arXiv:2106.02689*, 2021.

[27] Y. Zhao, L. Dong, Y. Shen, Z. Zhang, F. Wei, and W. Chen, "Memory-efficient differentiable transformer architecture search," *arXiv preprint arXiv:2105.14669*, 2021.

[28] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann, "Describing videos using multi-modal fusion," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1087–1091.

[29] S. Phan, G. E. Henter, Y. Miyao, and S. Satoh, "Consensus-based sequence training for video captioning," *arXiv preprint arXiv:1712.09532*, 2017.

[30] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4213–4222.

[31] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8918–8927.

[32] Y. Pan, Y. Li, J. Luo, J. Xu, T. Yao, and T. Mei, "Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training," *arXiv preprint arXiv:2007.02375*, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Local Contrast Enhancement with Multiscale Filtering

Kohei Hayashi*, Yoshihiro Maeda† and Norishige Fukushima*
* Nagoya Institute of Technology, Japan
Web: https://fukushima.web.nitech.ac.jp/en/
† Tokyo University of Science, Japan

***Abstract*—Multiscale processing is fundamental for image processing. Most multiscale processing splits signals into base and detail signals and then manipulates the detail signal. Local Laplacian filtering (LLF), multiscale processing, has an unusual form: manipulation and then separation. This structure brings out the high performance of the filter. However, LLFs are limited to the form of Laplacian pyramids, which requires downsampling to create image pyramids; it causes misalignments around edges. In this study, we break away from the pyramid and extend the interpretation to multiscale filters that locally enhance the contrast. We propose a local difference of Gaussian (DoG) filter (LDF) for the solution. DoG filtering can remove the downsampling in the filter. Moreover, we further improve the computational speed using the precomputing technique introduced in LLF. Furthermore, to accelerate it, we utilize $O(1)$ Gaussian filtering for large-scale filtering. Experimental results show that the proposed method suppresses overshoot and undershoot around sharp edges caused by the misalignment. Furthermore, the proposed method can keep computational speed even when the convolutional radius is large by $O(1)$ Gaussian filtering.**

I. Introduction

When capturing images, the contrast of the subject cannot be perfectly transferred due to diffraction of light, aberrations, surface reflections, and other physical characteristics of light and lenses. In particular, contrast is low in regions of high spatial frequency. The modulation transfer function expresses this relationship. Therefore, the image is multiscaled, and a nonlinear contrast transformation function enhances the contrast. In this study, the value referenced by the contrast transformation function is changed for each pixel to improve the local contrast. By improving the contrast at a specific spatial frequency, a locally good-looking image is generated, and by multiscaling the image and retaining only the coefficients of the relevant pixel among the coefficients after processing, more precise edge information can be obtained than by performing contrast transformation after multiscaling. In this way, an image with improved contrast at all positions in the image is obtained by processing each pixel so that the modulation transfer function is changed for each pixel.

Manipulation of images with multiscale analysis is a challenging task. For multiscale processing, an input image is decomposed into multiple layers by various methods: Gaussian and Laplacian pyramids [1], wavelet transform [2], and difference of Gaussians (DoG) form the scale-space analysis [3]. Multiscale processing manipulates coefficients of pyramid, wavelet, or scale-space representations. Then, the layered signals are collapsed to obtain an output image.

Unsharp masking is a straightforward multiscale process. The method has two layers: a base layer and a detail layer. The base layer is a Gaussian/box-filtered image, and the detail layer is the subtraction of the base layer from the input image. The detail layer is linearly multiplied to control the enhancement level. However, this approach produces large halos.

The Laplacian pyramid is utilized to extend two layers to multiple layers. Pyramid-based processing directly enhances the pyramid coefficient [4]–[6]. Instead of linear enhancement of coefficients, S-tone or Gaussian weighting is used to suppress large amplitude. In the Gaussian pyramid-based approach, gradients in pyramid images are optimized to scale factors to avoid halos [7].

The signal decomposition is extended to have an edge-preserving property. Seminal work uses bilateral filtering [8] to generate base signals in two-layer decomposition [9] instead of linear filtering. The cost-consuming part of bilateral filtering is accelerated by Fast Fourier transform (FFT) [9], and further accelerated by following accelerated methods [10], [11]. The two-layer decomposition based on bilateral filtering is extended to multiple layers using iterative bilateral filtering. Multiscale bilateral decomposition is accelerated by skipped filtering [12]. Skipping is further accelerated by hardware [13]. This approach has difficulty in determining parameters to suppress halo and other edge-reversal artifacts. The following work computes a multiscale edge-preserving decomposition with a least-square scheme instead of bilateral filtering [14]. The other extension uses iterative local linear regression filtering [15], similar to guide image filtering [16], [17].

The edge-avoiding wavelet [18], [19] is an edge-preserving extension of the wavelet transform. The method constructs a basis based on the edge content of the images. The wavelet becomes non-linear filtering. Also, iterative bilateral filtering-based decomposition [12] is one of À-Trous wavelet methods [20]. The method is extended to multilateral filtering for the computer graphics context [21].

The previous edge-preserving scale-space methods have difficulty handling each iterative and enhancement processing

This work was supported by JSPS KAKENHI (21H03465, 21K17768) and Environment Research and Technology Development Fund (JPMEERF20222M01) of the Environmental Restoration and Conservation Agency of Japan.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

parameter. Local Laplacian filtering (LLF) [22] solves the difficulty. LLF is an extension of the Laplacian pyramid [23] to have an edge-aware pyramid. This approach locally enhances the contrast quickly and generates Laplacian pyramids for each pixel for image detail control. LLF has better visual quality than other multiscale processing; however, LLF has a high processing cost. The following works accelerate the construction per pixel of the Laplacian pyramid: fast LLF [24] and Gaussian Fourier pyramid [25]. Also, LLF is accelerated by FPGA [26]. Edge-aware upsampling can also accelerate filtering, e.g., joint bilateral upsampling [27], guided image upsampling [28], bilateral guided upsampling [29], and local-LUT upsampling [30].

The processing flow is the most significant difference between the previous methods and LLF. Conventional methods, such as linear filter, edge-preserving smoothing, and wavelet methods, all perform the following steps: 1) separation, 2) enhancement, and 3) integration. However, the structure has difficulty controlling the smoothing level of scale-space and amplifying each level. Also, it has drawbacks of edge ambiguity in downsampling, and downsampling-based approaches do not preserve the translation invariance property.

In contrast, LLF is different: 1) local enhancement, 2) separation, and 3) integration, where the order of enhancement and separation is reversed. This processing flow more easily controls the parameters. In this paper, we break away from the concept of pyramids and consider a new edge-preserving scale-space analysis based on the concept of the different processing orders. Moreover, the new scale-space method solves the inevitable aliasing problem of pyramid processing, like the concept of stationary wavelet transforms [31].

The contributions of this paper are as follows:

- We replace the Laplacian pyramid with the multiscale filtering of DoG in LLF to increase the controllability and generality of scale-space named local DoG filter (LDF).
- We accelerate the DoG computation by recent $O(1)$ Gaussian filtering (GF) with a fast LLF formulation to accelerate LDF.

II. Preliminary

A. Gaussian Pyramid and Laplacian Pyramid

In this section, we review the Laplacian pyramid. Laplacian pyramid is used for various applications, such as compression [23], texture synthesis [32], and harmonization [33].

Here, we introduce the Gaussian pyramid. Let I be an input image, and let the Gaussian pyramid be defined as the set of $G_\ell[I]$, where $\ell \in \mathcal{L} = \{0, 1, 2, \ldots, \ell_{\max}\} \subset \mathbb{N}$. The lowest level of the Gaussian pyramid is $G_0[I] = I$, and the other level of pyramids $G_{\ell+1}[I]$ are equal to the downsampled blurred image of $G_\ell[I]$:

$$G_\ell[I] = \mathrm{downsampling}(\mathcal{G}_\sigma * G_{\ell-1}[I]), \tag{1}$$

where $\mathcal{G}_\sigma *$ means Gaussian convolution with standard deviation σ. The image of $G_{\ell+1}[I]$ has half the width and height of $G_\ell[I]$. Usually, the Gaussian convolution for pyramid building is implemented using the binomial distribution with five taps for integer operations, and the regarded standard deviation is about $\sigma = 1$.

The difference between the current-level pyramid and the upsampled next-level pyramid defines the Laplacian pyramid:

$$L_\ell[I] = G_\ell[I] - \mathrm{upsampling}(G_{\ell+1}[I]), \tag{2}$$

where $\mathrm{upsampling}(\cdot)$ is an operator that doubles the width and height of the image size by using a smoothing kernel. Usually, the smoothing kernel is the same kernel as $\mathcal{G}_\sigma$. Note that the highest level of the Laplacian pyramid is $L_{\ell_{\max}}[I] = G_{\ell_{\max}}[I]$.

In the Gaussian pyramid, the filtering parameter of each level, σ, is fixed. Iterative processing and downsampling change each level's actual parameter, i.e., equivalent to the full resolution parameter. The relationship between each scale and actual σ_ℓ at each level is essential for scale-space analysis.

Here, we show the relationships. Iterative Gaussian convolutions can be merged and represented by one large standard deviation value by the semi-group property of Gaussian:

$$\mathcal{G}_{\sigma 1} * \mathcal{G}_{\sigma 2} = \mathcal{G}_\sigma, \quad \sigma = \sqrt{\sigma_1^2 + \sigma_2^2}. \tag{3}$$

The Gaussian convolution is applied for each level. The σ for each level becomes the sum of each variance until the level. Note that the image is downsampled for each level; thus, the supporting distribution is doubled by the level. The actual standard deviation for each level σ_ℓ is defined as follows:

$$\sigma_\ell = \sigma \sqrt{\sum_{l=1}^{\ell} 2^l} \quad (\ell > 0), \quad \sigma_0 = 0, \tag{4}$$

where σ is the base value of a standard deviation for constructing Gaussian and Laplacian pyramids.

B. Manipulating Laplacian Pyramid

We show the example of manipulating the Laplacian pyramid for detail enhancement.

We boost the pyramid coefficients for multiscale detail enhancement except for the coarsest layer of $\sup \mathcal{L}$. We usually use an S-tone-like function to suppress overshoot for boosting, while the most straightforward function is $r(x, 0) = kx$, where $k > 1$ is a constant value. If $k = 1$, there is no boosting, and the resulting image is the input image. Note that the representation of the argument 0 in $r(x, 0)$ is unnecessary for this case; however, it is used to match the form of the latter remapping function for the local Laplacian filtering. The later section refers to the remapping function. Finally, remapped signals are collapsed to obtain the output image:

$$O = L_{\ell_{\max}}[I] + \sum_{\ell \in \mathcal{L} \setminus \{\ell_{\max}\}} r(L_\ell[I], 0). \tag{5}$$

C. Local Laplacian Filtering

We review the basic structure of local Laplacian filtering (LLF). In LLF, an output pixel value O_p at pixel $p = (x, y)$ is defined by constructing a pixel-dependent Laplacian pyramid of the output image, $L[O]_p$, called the local Laplacian pyramid.

Each level ℓ of pyramid $L_\ell[O]_p$ is constructed per pixel. Finally, the output pixel O_p is the summation of the value on p for each level, i.e., collapsing pyramid:

$$O_p = \sum_{\ell \in \mathcal{L}} L_\ell[O]_{p_\ell}, \tag{6}$$

where p_ℓ is the relative position of p for each level ℓ, and becomes sub-pixel. The size of the next level image in the pyramid is half, so the accessing position of pixel p also becomes half. For assessing sub-pixel, we interpolate intensity.

There are three steps to compute the coefficient itself $L_\ell[O]_{p_\ell}$. First, a remapping function $r(\cdot)$ is applied for the input image I and then creates the intermediate image $\tilde{I} = r(I, G_\ell[I]_p)$, where $r(\cdot)$ depends on the coefficient of the Gaussian pyramid at level ℓ at p. We will discuss the details of the remapping function later.

Next, we construct a Laplacian pyramid of the remapped image $L_\ell[\tilde{I}]$, and then we select a pyramid coefficient at p that is $L_\ell[\tilde{I}]_p$ as an output coefficient of $L_\ell[O]_p$. By repeating this process in all positions p and levels ℓ, we can get the Laplacian pyramid of the output pyramid $L[O]$.

Finally, we get the output image by collapsing the pyramid $L[O]$. The computational cost of this process is $O(\sigma|\mathcal{L}|N \log N)$ per pixel, where N is the number of image pixels. The separable Gaussian filtering requires $O(\sigma)$ for each level $|\mathcal{L}|$, and a pyramid requires contains $N \log N$ pixels.

Careful implementation can reduce the processing pyramid size. Let $R \in \mathbb{N}$ be the convolution radius: $R = \lceil n\sigma \rceil, n \in \mathbb{R}$, usually $n = 3.0$. The deep pyramid layer hierarchy requires outside pixels of the shallower layers. The required outside radius at level ℓ is defined as follows:

$$R_l = R \sum_{\ell=0}^{l} 2^\ell, \tag{7}$$

Therefore, we require $2 * R_l + 1$ pixels to build a pyramid. For example, $R_l \in \{1, 3, 7, 15, 31, 63, 127, \ldots\}$. R_l is exponentially increasing; thus, the trimmed pyramid size can rapidly reach the maximum size of N.

The approximated acceleration of LLF, called fast LLF [24], can reduce the order $O(\sigma|\mathcal{L}||\mathcal{I}|)$ per pixel, where $|\mathcal{I}| \in \mathbb{N}$ is the number of approximation samples. The method builds a limited number of pyramids and then blends the pyramid to approximate the result.

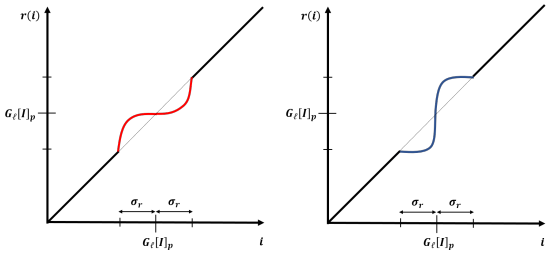
D. Remapping Function

The remapping function is defined as:

$$r(x, g) = x - (x - g)f(x - g), \tag{8}$$

where, x is input signal intensity, and g is the controlling center value. This function enhances/smooths signals. In LLF, $g = G_\ell[I]_p$, where $G_\ell[I]_p$ is a coefficient of a Gaussian pyramid, while f is specifically defined for each application. This paper used the Gaussian function with multiplying factor m as f.

$$f(x - g) = m \exp\left(\frac{(x - g)^2}{-2\sigma_r^2}\right) \tag{9}$$



(a) detail smoothing (b) detail enhancement

Fig. 1: . Remapping functions ($m = 1$).

The sign of m switches enhancing or smoothing signals, and the magnitude control strength of signal control. Figure 1 shows an example of remapping functions.

III. Proposed Method

We extend the concept of LLF to a multiscale filter that manipulates the contrast in advance, named Local DoG Filter (LDF). Furthermore, the LDF is accelerated using the fast LLF form. Moreover, we use $O(1)$ Gaussian filtering (GF) for further acceleration.

A. Definition

Under the concept of LDF, The difference between LDF and LLF is the scale-space structure: pyramid or DoG. In conventional LLF, the image at level $\ell + 1$ is a quarter-size image at ℓ due to downsampling. In contrast, the LDF's image size of each level ℓ is equal due to no downsampling.

First, we build the scale-space blurring representation instead of the Gaussian pyramid. The definition of scale-space $G_\ell[I]$ is as follows:

$$G_\ell[I] = \begin{cases} I & (\ell = 0) \\ \mathcal{G}_{\sigma_\ell} * I & (\text{otherwise}). \end{cases} \tag{10}$$

Here, we redefine $L_\ell[O]$ at a pixel p by using DoG. DoG filtering means the difference between two GFs that have different σ values; thus, $L_\ell[O]_p$ is defined as follows:

$$L_\ell[O]_p = \begin{cases} (\mathcal{G}_{\sigma_\ell} * r(I, G_\ell[I]_p) - \mathcal{G}_{\sigma_{\ell+1}} * r(I, G_\ell[I]_p))_p \\ \qquad\qquad (0 \le \ell \le \ell_{\max} - 1) \\ (\mathcal{G}_{\sigma_\ell} * I)_p \qquad (\ell = \ell_{\max}), \end{cases} \tag{11}$$

where σ_ℓ is the Gaussian scale parameter for each level ℓ. Steps between each level σ_ℓ depend on the applications and have more flexibility than the Laplacian pyramid. For example, we can set a small step when fine scaling is required. In this paper, however, we apply (4) for comparison with conventional LLF.

Finally, an output value O_p is obtained by collapsing the DoG stack:

$$O_p = \sum_{\ell \in \mathcal{L}} L_\ell[O]_p. \tag{12}$$

We do not need upsampling and downsampling, unlike the pyramid-based LLF.

The required number of pixels for a DoG kernel is σ^2, and each level does not have dependency. Moreover, we can compute a pixel using a convolution. In this case, separable filtering is ineffective since separable filtering is used to convolute whole pixels. Therefore, the order is $O(|\mathcal{L}|\sigma^2)$ per pixel. LLF requires $O(\sigma|\mathcal{L}|N\log N)$, larger than our representation.

B. Acceleration

Here, we rewrite the LDF representation for acceleration. We call the accelerated method *fast LDF*. The concept is that $r(I, G_\ell[I]_p)$ in (11) has a limited number of variations. Let a sample i be an integer value, $i \in \mathcal{I} \subset \mathbb{Z}$. First, remapping images $\mathcal{R}_i$ are prepared for each intensity i:

$$\mathcal{R}^i = r(I, i) \quad \forall i \in \mathcal{I}. \tag{13}$$

Second, the DoGs for each remapping image are built:

$$L_\ell[\mathcal{R}_i] = \mathcal{G}_{\sigma_\ell} * \mathcal{R}^i - \mathcal{G}_{\sigma_{\ell+1}} * \mathcal{R}^i \quad \forall i \in \mathcal{I}. \tag{14}$$

Eqs. (13) and (14) do not exist pixel accessing operator p since processing for each pixel is the same. Thus, the GF in (14) can be regarded as the usual GF for entire images. GF can be accelerated using the usual approaches, such as FFT, discussed in the following.

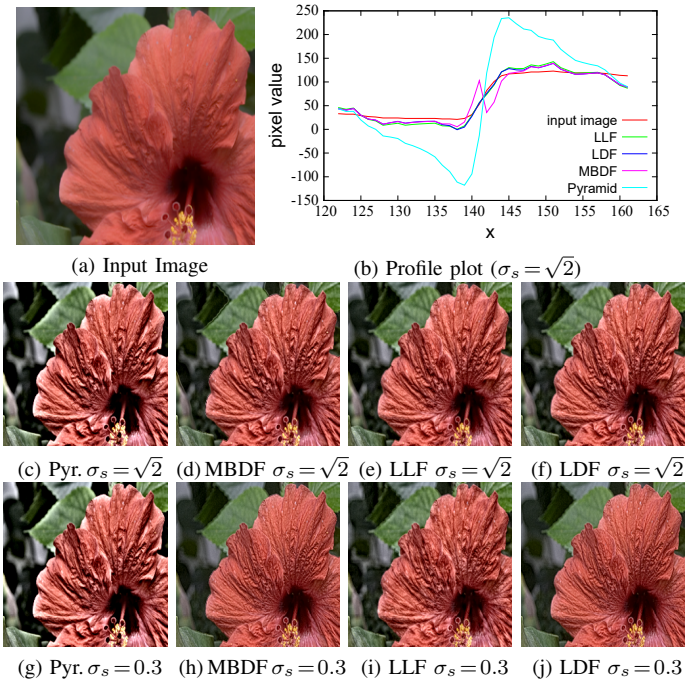
Next, DoGs of the remapped images are combined with the local Laplacian DoG by referring to the Gaussian stacks of (10):

$$L_\ell[O]_p = \sum_{i\in\mathcal{I}} w(G_\ell[I]_p, \mathcal{R}^i_p) L_\ell[\mathcal{R}^i]_p, \tag{15}$$

where w is the weight function for the linear interpolation. Equation (14) sums up each DoG of the remapped image weighted by w. Finally, we can obtain an output value O_p by collapsing the DoG based on (12).

Here, we discuss Gaussian convolution in (14). The order of fast LDF with the naïve implementation of GF is $O(|\mathcal{I}|\sigma^2)$. We perform $|\mathcal{I}|$ times GFs for each pixel, and the order of GF is $O(\sigma^2)$ per pixel. A separable representation can reduce the GF order to $O(\sigma)$. In this case, the LLF order is $O(N|\mathcal{I}|\sigma)$, which is the same as the conventional method of "fast" LLF. The GF is further accelerated by constant-time algorithms, which have two representations: infinite impulse response (IIR) and finite impulse response (FIR).

Typical examples of the IIR filter are the addition model [34] and the multiplication model [35]. The order of these methods is $O(K)$ per pixel, where K is the approximate order. However, the implementation of IIR is not faster than the implementation of recursive FIR [36]. In terms of FIR approximation methods, sliding DCT-based GF has been proposed [37]–[43]. The computational order of this method is also $O(K)$. Currently, sliding DCT-based GF has the best performance in terms of accuracy and processing speed [42]. Therefore, we use the sliding DCT-based GF for the implementation of LDF. The parameter of K is independent of the filtering parameter of σ, so the approximation has a constant-time property for σ. Using constant-time GF, the order of LDF is $O(|\mathcal{I}|K)$, *named fast $O(1)$ LDF* because the algorithm is constant for σ.

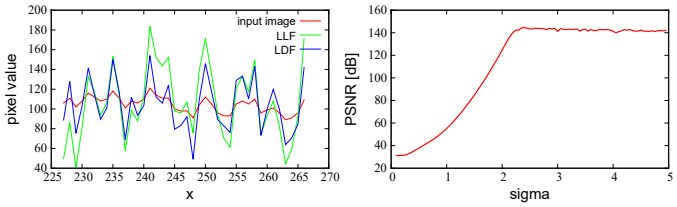


(a) Input Image (b) Profile plot ($\sigma_s = \sqrt{2}$)

(c) Pyr. $\sigma_s = \sqrt{2}$ (d) MBDF $\sigma_s = \sqrt{2}$ (e) LLF $\sigma_s = \sqrt{2}$ (f) LDF $\sigma_s = \sqrt{2}$

(g) Pyr. $\sigma_s = 0.3$ (h) MBDF $\sigma_s = 0.3$ (i) LLF $\sigma_s = 0.3$ (j) LDF $\sigma_s = 0.3$

Fig. 2: Results of each method ($\sigma_r = 20$, $\ell_{\max} = 2$).



(a) Profile plot ($\sigma_s = 0.3$) (b) PSNR

Fig. 3: a) Profile plot of Fig. 2i, Fig 2j, and the input image. b) PSNR between naïve Gaussian filtering and upsampled Gaussian filtering in the Gaussian pyramid.

IV. Experimental Results

We demonstrate the performance of the LDF, fast LDF, and fast $O(1)$ LDF on accuracy and speed. The code is written in C++ with OpenMP parallelization and AVX vectorization. The CPU tested is the Intel Core i9–9980XE (3.0 GHz, 18-cores/36-threads). We remove subnormal numbers in Gaussian distributions for acceleration based on [44].

First, we compare the output images of various methods: pyramid-enhancement of Eq. (5), multiscale bilateral decomposition filter (MBDF) [12], LLF [22] and the proposed LDF. Pyramid and MBDF are preenhancement methods. Figure 2 shows the output of these images and the profile plots of each signal. These images have two σ_s settings: $\sigma_s \in \{0.3, \sqrt{2}\}$. The pyramid has halos, while the other methods solve the problem. MBDF has edge reversals (e.g., the leaf contour in the image's upper left corner), while LLF and LDF solve the problem. LLF and LDF in $\sigma = \sqrt{2}$ have almost the same response, whereas, in the smaller case of $\sigma = 0.3$, the results are different.

Figure 3(a) shows the profile plots then $\sigma = 0.3$. The

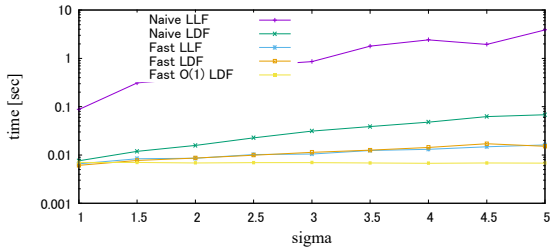


Fig. 4: Processing time ($\ell_{\max} = 2$, $K = 3$). σ_s is the standard deviation of $\ell = 1$, and the radius of the kernel is truncated by $3\sigma_\ell$ for each level in LDF. For conventional LLF, we used a fixed kernel radius, $3\sigma_s$, for each downsampled image. Test images are Kodak 24 images (768×512).

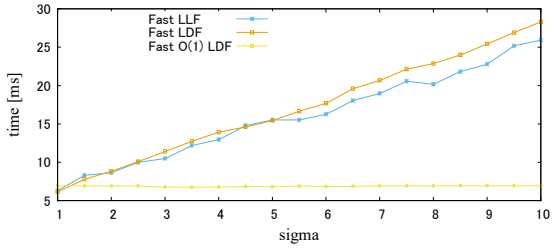


Fig. 5: Processing time of selected methods in Fig. 4. The Y scale is *not the log scale*, and the unit is *ms*.

responses differ, while the case $\sigma = \sqrt{2}$ is almost the same. The difference between high-level signals and low-level signals generates a Laplacian pyramid. The low-level signals are an approximation of Gaussian filtering using the semigroup property and downsampling. Figure 3(b) shows the approximation accuracy of Gaussian filtering. Due to downsampling, aliasing is produced in a small σ_s case; therefore, the accuracy of the approximation is low. The low approximation accuracy produces the difference between LLF and LDF.

Figure 4 shows the computational time (log scale in Y) of each method: naïve LLF [22], fast LLF [24], naïve LDF, fast LDF, fast O(1) LDF. Naïve LLF requires many processing areas, while naïve LDF needs less area; thus, the processing time of LDF is faster than the naïve LLF. For the other case, Fig. 5 shows faster plots with the removal of the logarithmic scale. The computational time of fast LDF and fast LLF is proportional to σ, while fast $O(1)$ LDF exhibits a constant time property. Except for the tiny sigma case (around $\sigma = 1$), the proposed method is faster than the other. Note that due to cache efficiency, when σ is small, the naïve GF implementation is faster than the $O(1)$ GF-based one.

V. Conclusion

This study proposed a local DoG filter (LDF) that creates image pyramids without downsampling. The pure definition of the proposed method generates DoG per pixel; however, it is more efficient than naïve LLF. Furthermore, we accelerate it by changing the form to share the convolution between pixels. Moreover, we use constant-time Gaussian filtering to accelerate large-scale filtering. In our future work, we extend the method to have color space [45] and to perceptually enhance images [46].

References

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, 1984.

[2] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[3] T. Lindeberg, *Scale-space theory in computer vision*. Springer Science & Business Media, 2013, vol. 256.

[4] P. Vuylsteke and E. P. Schoeters, "Multiscale image contrast amplification (musica)," in *Proceedings of SPIE*, vol. 2167, 1994, pp. 551–560.

[5] S. Dippel, M. Stahl, R. Wiemker, and T. Blaffert, "Multiscale contrast enhancement for radiographies: Laplacian pyramid versus fast wavelet transform," *IEEE Transactions on Medical Imaging*, vol. 21, no. 4, 2002.

[6] Y. Li, L. Sharan, and E. H. Adelson, "Compressing and companding high dynamic range images with subband architectures," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 836–844, 2005.

[7] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 249–256, 2002.

[8] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 1998, pp. 839–846.

[9] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 257–266, 2002.

[10] K. Sugimoto, N. Fukushima, and S. Kamata, "200 fps constant-time bilateral filter using svd and tiling strategy," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2019.

[11] Y. Sumiya, N. Fukushima, K. Sugimoto, and S. Kamata, "Extending compressive bilateral filtering for arbitrary range kernel," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2020.

[12] R. Fattal, M. Agrawala, and S. Rusinkiewicz, "Multiscale shape and detail enhancement from multi-light image collections.," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 51, 2007.

[13] N. Fukushima, T. Tsubokawa, and Y. Maeda, "Vector addressing for non-sequential sampling in fir image filtering," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

[14] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 1–10, 2008.

[15] B. Gu, W. Li, M. Zhu, and M. Wang, "Local edge-preserving multiscale decomposition for high dynamic range image tone mapping," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 70–79, 2012.

[16] K. He, J. Shun, and X. Tang, "Guided image filtering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[17] N. Fukushima, K. Sugimoto, and S. Kamata, "Guided image filtering with arbitrary window function," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[18] R. Fattal, "Edge-avoiding wavelets and their applications," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–10, 2009.

[19] Y. Sumiya, H. Kamei, K. Ishikawa, and N. Fukushima, "Vectorized computing for edge-avoiding wavelet," in *International Workshop on Advanced Image Technology (IWAIT)*, vol. 12177, 2022, pp. 23–28.

[20] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, Springer, 1990, pp. 286–297.

[21] H. Dammertz, D. Sewtz, J. Hanika, and H. P. A. Lensch, "Edge-Avoiding A-Trous Wavelet Transform for fast Global Illumination Filtering," in *Proc. ACM SIGGRAPH/EUROGRAPHICS Conference on High Performance Graphics*, The Eurographics Association, 2010.

[22] S. Paris, W. Hasinoff, and J. Kautz, "Local laplacian filters: Edge-aware image processing with a laplacian pyramid," *ACM Trans. on Graphics*, vol. 30, no. 4, 2011.

[23] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.

[24] M. Aubry, S. Paris, J. Kautz, and F. Durand, "Fast local laplacian filters: Theory and applications," *ACM Trans. on Graphics*, vol. 33, no. 5, 2014.

[25] Y. Sumiya, T. Otsuka, Y. Maeda, and N. Fukushima, "Gaussian fourier pyramid for local laplacian filter," *IEEE Signal Processing Letters*, vol. 29, pp. 11–15, 2022.

[26] S. Khandelwal, Z. Choudhury, S. Shrivastava, and S. Purini, "Accelerating local laplacian filters on fpgas," in *International Conference on Field-Programmable Logic and Applications (FPL)*, 2020, pp. 109–114.

[27] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.

[28] K. He and J. Sun, "Fast guided filter," *CoRR*, vol. abs/1505.00996, 2015.

[29] J. Chen, A. Adams, N. Wadhwa, and S. W. Hasinoff, "Bilateral guided upsampling," *ACM Transactions on Graphics*, vol. 35, no. 6, 2016.

[30] T. Tsubokawa, H. Tajima, Y. Maeda, and N. Fukushima, "Local look-up table upsampling for accelerating image processing," *Multimedia Tools and Applications*, 2023.

[31] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Wavelets and Statistics*, Springer, 1995, pp. 281–299.

[32] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. annual conference on Computer graphics and interactive techniques*, 1995, pp. 229–238.

[33] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister, "Multi-scale image harmonization," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 1–10, 2010.

[34] R. Deriche, "Fast algorithms for low-level vision," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 12, pp. 78–87, 1990.

[35] L. J. van Vliet, I. T. Young, and P. W. Verbeek, "Recursive gaussian derivative filters," in *Proc. International Conference on Pattern Recognition (ICPR)*, 1998.

[36] H. Takagi and N. Fukushima, "An efficient description with halide for iir gaussian filter," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2020.

[37] K. Sugimoto and S. Kamata, "Fast gaussian filter with second-order shift property of dct-5," in *Proc. IEEE International Conference on Image Processing*, 2013.

[38] K. Sugimoto and S. Kamata, "Efficient constant-time gaussian filtering with sliding dct/dst-5 and dual-domain error minimization," *ITE Transactions on Media Technology and Applications*, vol. 3, pp. 12–21, 2015.

[39] D. Charalampidis, "Recursive implementation of the gaussian filter using truncated cosine functions," *IEEE Transactions on Signal Processing*, vol. 64, 2016.

[40] K. Sugimoto, S. Kyochi, and S. Kamata, "Universal approach for dct-based constant-time gaussian filter with moment preservation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1498–1502, 2018.

[41] N. Fukushima, Y. Maeda, Y. Kawasaki, *et al.*, "Efficient computational scheduling of box and gaussian fir filtering for cpu microarchitecture," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2018.

[42] T. Otsuka, N. Fukushima, Y. Maeda, K. Sugimoto, and S. Kamata, "Optimization of sliding-dct based gaussian filtering for hardware accelerator," in *Proc. International Conference on Visual Communications and Image Processing (VCIP)*, 2020.

[43] K. Ishikawa, Y. Sumiya, and N. Fukushima, "Polynomial fitting for period prediction in sliding-dct-based filtering," in *International Workshop on Advanced Image Technology (IWAIT),*, International Society for Optics and Photonics, vol. 12177, SPIE, 2022, pp. 143–148.

[44] Y. Maeda, N. Fukushima, and H. Matsuo, "Effective implementation of edge-preserving filtering on cpu microarchitectures," *Applied Sciences*, vol. 8, no. 10, 2018.

[45] Y. Maeda, N. Fukushima, and T. Hamamoto, "Color transformation for compressive computing in image filtering," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021.

[46] S. Oishi and N. Fukushima, "Retinex-based relighting for night photography," *Applied Sciences*, vol. 13, no. 3, 2023.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Marine Snow Removal Benchmarking Dataset

Reina Kaneko*, Yuya Sato†, Takumi Ueda†, Hiroshi Higashi*, and Yuichi Tanaka*
* Osaka University, Japan
† Tokyo University of Agriculture and Technology, Japan
E-mail: r.kaneko@sip.comm.eng.osaka-u.ac.jp, {higashi, ytanaka}@comm.eng.osaka-u.ac.jp

***Abstract*—This paper introduces a new benchmarking dataset for marine snow removal of underwater images. Marine snow is one of the main degradation sources of underwater images that are caused by small particles, e.g., organic matter and sand, between the underwater scene and photosensors. We mathematically model two typical types of marine snow from the observations of real underwater images. The modeled artifacts are synthesized with underwater images to construct large-scale pairs of ground truth and degraded images to calculate objective qualities for marine snow removal and to train a deep neural network. We propose two marine snow removal tasks using the dataset and show the first benchmarking results of marine snow removal. The *Marine Snow Removal Benchmarking Dataset* is publicly available online.**

I. Introduction

Image restoration has been one of the main topics of computer vision for decades [10], [11], [28], [35]–[38]. In the current era of deep learning, the quality of image restoration for various tasks has been significantly improved under a sufficient number of pairs of ground truth and degraded images. Accordingly, the current focus on image restoration will be for images taken under extreme situations such as underwater, satellite, and medical images [1], [3], [6], [12], [14], [20], [23], [24], [26], [27], [31]–[34], [39], [41]. This problem is called *extreme image restoration* herein. Extreme images often have an untypical degradation beyond that of a Gaussian model, and the numbers of available images for training and modifying restoration algorithms are also limited. As a result, extreme image restoration is a challenging problem.

In this paper, we focus on underwater image enhancement. Its main challenge is mostly due to the fact that we generally have no ground truth images. This makes it difficult to evaluate the objective qualities of the restored images and measure the restoration performance. Moreover, the lack of pairs of ground truth and degraded image causes a difficulty in training a neural network (if we consider using a deep neural network for restoration) because the loss functions of the image restoration typically evaluate some objective image quality metrics such as mean squared error (MSE) and structured similarity index (SSIM) [40]. Hence, unsupervised restoration methods, including linear and nonlinear filtering, are still popular in underwater image enhancement [2], [6], [24].

This work is supported in part by Japan Science and Technology Agency (JST) Advanced International Collaborative Research Program (AdCORP) under Grant JPMJKB2307 and JSPS KAKENHI under Grant 23K17461.

Several studies for underwater image enhancement with deep neural networks have been proposed [14], [18], [20], [22], [23], [32], [33], [42]. Most focus on removing the color shift from underwater images. That is, the blueish underwater images are restored by enhancing the red and green channels. Conventional underwater image enhancement methods using deep neural networks tackle the problem on the lack of large-scale underwater image datasets in various ways. For example, the studies in [4], [14], [32], [33] synthetically generate degraded images from clean images taken on the ground by simulating the degradation processes of absorption and scattering. Moreover, [15], [20], [25] use generative adversarial networks. And a number of challenges to improving the restoration performance still exist, some promising results have been presented. However, such studies ignore another main source of degradation for underwater images: *Marine snow*.

In this paper, we consider a marine snow removal (MSR) problem for underwater images by constructing a large-scale dataset containing underwater image pairs. Marine snow is a typical underwater image degradation, examples of which are shown in Fig. 1. Marine snow artifacts are visible in digital images when we take an underwater image using a flashlight. Small particles in the scene, e.g., organic matter and sand, cause marine snow. The particles reflect the light from the flashlight, and the reflected light will be captured by the photosensors. Marine snow is often annoying in underwater photography because it affects the overall image quality. However, the particles are unevenly distributed in the scene, and the degradation process is difficult to model completely. Therefore, standard methods for marine snow removal are still limited to simple median-filter-based approaches [5], [13], [18], [21]. Moreover, as with standard underwater image enhancement, we do not have a pair of ground truth and degraded images to calculate the objective image qualities and tune the parameters of the restoration algorithms.

We propose marine snow synthesis methods for solving the above-mentioned problems of MSR. In other words, we *intentionally* append the marine snow artifacts to real underwater images without marine snow and generate pairs of ground truth images and degraded images (i.e., clean underwater images and their corresponding images with marine snow artifacts). The key for synthesizing artifacts is to mathematically model marine snow in digital images as accurately as possible. From actual marine snow images, we discover that many marine snow artifacts can be modeled through two typical pixel value

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

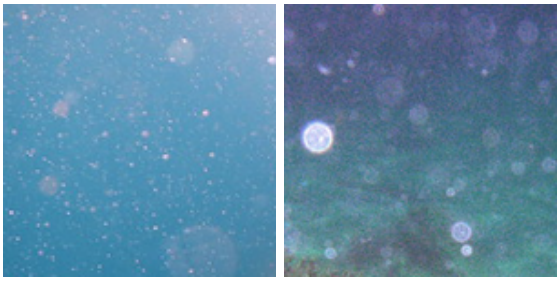
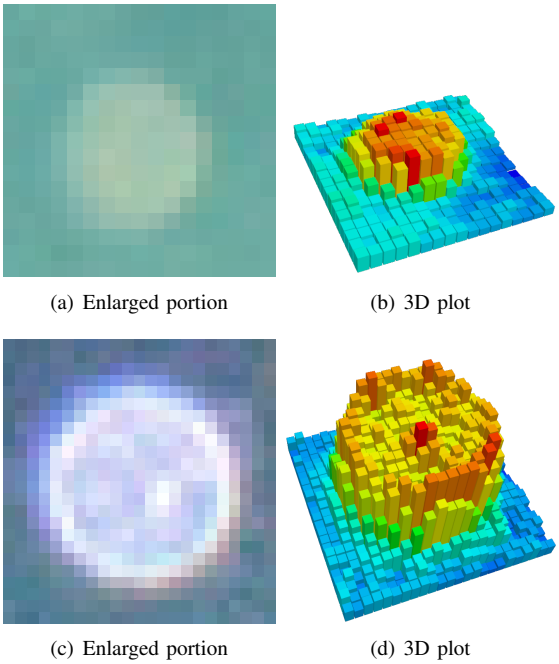


Fig. 1: Marine snow artifacts in real underwater images.



(a) Enlarged portion (b) 3D plot

(c) Enlarged portion (d) 3D plot

Fig. 2: Marine snow examples in real underwater images. Top: Highland type (type H). Bottom: Volcanic Crater type (type V). The 3D plots correspond to the grayscale images.

distributions. By utilizing this finding, we can synthetically generate many pairs of underwater images with/without marine snow artifacts.

We also propose two marine snow removal tasks and develop datasets corresponding to these tasks. This *Marine Snow Removal Benchmarking Dataset* (MSRB Dataset) is the first attempt in this field and is publicly available online[1]. We also present the first MSR benchmarking results using the MSRB Dataset: We compare the objective performance of marine snow removal among median filters and a deep neural network. Real MSR results as well as the limitations are also described.

The remainder of this paper is organized as follows: Section II reviews several related studies for underwater image enhancement. Two representative marine snow models that we discovered are presented in Section III along with their synthesizing method. The MSRB Dataset specifications are introduced in Section IV and include the description of the two tasks. The first MSR benchmarking results with objective image quality metrics are shown in Section V. Real MSR results and limitations are described in Section VI. Finally, we provide some concluding remarks in Section VII.

II. Related Studies

There have been few studies on modeling marine snow in digital images. Its seminal study is described in [7], [8]. A marine snow artifact is modeled using a Gaussian function in which the artifact is less transparent in its center than in its surrounding area. This method is better than the traditional salt and pepper noise; however, it does not simulate the actual shapes of marine snow artifacts. As a result, the center part of marine snow becomes excessively thick and the generated images are far from the real ones.

MSR is also an underrepresented problem in underwater image enhancement. Here, we briefly review some existing approaches.

A widely used method for MSR is a median filter (MF) [9], [19]. If marine snow artifacts can be assumed sufficiently small (typically 1–3 pixels in diameter), MF is expected to work because small marine snow artifacts can be considered to have a salt-and-pepper effect. However, it significantly blurs the entire image, particularly when we use a large filter kernel size.

Specific to MSR, a modified version of a MF is proposed in [5], [17]. This method applies the MF selectively if the target pixel has a higher intensity than the surrounding pixels. However, it is still difficult to remove large marine snow artifacts.

A few MSR methods for video sequences have also been proposed. They utilize the fact that marine snow artifacts continuously move in consecutive video frames. In [16], background modeling is used to remove marine snow artifacts from a static scene. Marine snow artifacts are tracked and a customized MF is applied to the detected artifacts in [13]. However, they are not applicable to a single underwater image. Importantly, all methods above are based on a model-based approach. This is mainly due to a lack of high-quality datasets, as mentioned in Section I.

A deep learning-based MSR method is proposed in [18]. Its training dataset consists of synthesized marine snow images, corresponding ground truth images, marine snow masks. This dataset also contains real underwater images with marine snow (without ground truth). Marine snow particles are manually made with Adobe Photoshop: That results in a lack of randomness and scalability.

In the following, we address the contributions of this study; marine snow models and the dataset specifications.

[1] https://github.com/ychtanaka/marine-snow

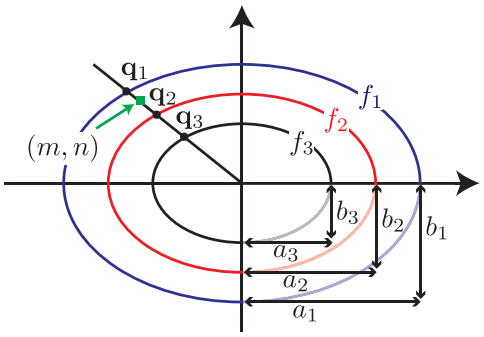


Fig. 3: Ellipses used for marine snow synthesis.

III. Marine Snow Models

There are various sources of marine snow and it is impractical to estimate the sources of all particles from a single underwater image. Instead of the estimation of the sources, we model the pixel value distributions of marine snow artifacts from observations of underwater images.

A. Real Marine Snow Examples

First, we start by showing marine snow examples in real underwater images. Figs. 2(a) and (c) show enlarged portions of representative marine snow artifacts cropped from real underwater images. Although they look similar, their pixel value distributions are slightly different.

Taking a closer look, the 3D plots of Figs. 2(a) and (c) are shown in Figs. 2(b) and (d), respectively. As clearly observed, they do *not* have a shape like a Gaussian function in contrast to the conventional assumption in [7], [8]. Rather than a Gaussian function, these 3D plots are similar to *elliptic conical frusta*, i.e., sliced elliptic cones. Furthermore, the top surfaces of the frusta have different characteristics between Figs. 2(b) and (d). In our preliminary observation, most marine snow artifacts can be classified into these two representative shapes. In the following, we present the mathematical models of marine snow artifacts that reflect the above-mentioned observations.

B. Marine Snow Model 1: Highland Type

The first marine snow model is the *Highland Type* (type H). The type H marine snow corresponds to Fig. 2(a) and can be modeled as an elliptic conical frustum with a rough surface.

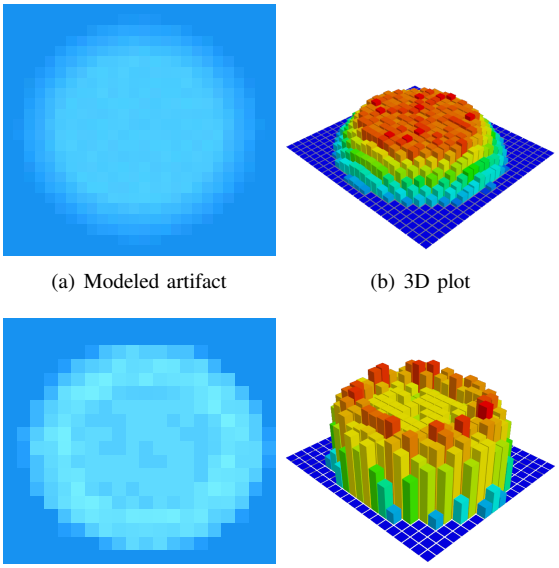
For modeling type H marine snow, we first suppose to have two ellipses f_1 and f_2, both centered at the coordinates (k, l). As shown in Fig. 3, we assume the focal points of f_1 and f_2 are located on the horizontal axis, i.e., the ellipses are wider, for simplicity. Let a_i and b_i be the semi-major and semi-minor axes of f_i where $i \in \{1, 2\}$, respectively. Suppose that f_1 is larger than f_2, i.e., $a_1 > a_2$ and $b_1 > b_2$ and the eccentricities of f_1 and f_2 are identical, i.e., $\frac{a_1}{b_1} = \frac{a_2}{b_2}$.

For notational simplicity, (k, l) is set to $(0, 0)$ hereafter. The (m, n)th pixel value of the type H marine snow $h(m, n)$ is formulated as follows:

$$h(m,n) = \begin{cases} c + \epsilon & \frac{m^2}{a_2^2} + \frac{n^2}{b_2^2} \le 1 \\ c\frac{d(\mathbf{p},\mathbf{q}_1)}{d(\mathbf{q}_1,\mathbf{q}_2)} + \epsilon & 1 < \frac{m^2}{a_2^2} + \frac{n^2}{b_2^2} \wedge \frac{m^2}{a_1^2} + \frac{n^2}{b_1^2} \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where c is a constant that determines the transparency of the marine snow, $d(\cdot, \cdot)$ calculates the Euclidean distance between two points, and ϵ is a small perturbation that mimics a rough surface. Furthermore, $\mathbf{p} := [m, n]^\top$ and $\mathbf{q}_i$ is the intersection of the ellipse f_i and the straight line crossing (m, n), as illustrated in Fig. 3.

An example of the synthesized type H marine snow artifacts is shown in Fig. 4(a) along with its 3D plot in Fig. 4(b). See the similarity in Figs. 2(a) and (b). For creating a dataset, the major axis of $h(m, n)$ is randomly rotated and the transparency c is also randomly chosen. These specifications are shown later in Section IV.



(a) Modeled artifact (b) 3D plot

(c) Modeled artifact (d) 3D plot

Fig. 4: Proposed marine snow models. Top: Type H marine snow. Bottom: Type V marine snow.

C. Marine Snow Model 2: Volcanic Crater Type

The second marine snow model is the volcanic crater type (type V). An example of type V marine snow is shown in Figs. 2(c) and (d). This can be modeled by a modified version of the type H with an overshoot top edge of the frustum.

For the type V marine snow, we consider another ellipse f_3, where a_3 and b_3 are its semi-major and semi-minor axes, respectively. It is located inside f_2, as illustrated in Fig. 3. As for the type H, focal points of all ellipses are assumed to be located along the horizontal axis and their eccentricities are identical, i.e., $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_3}{b_3}$.

Based on the type H function $h(m, n)$, the (m, n)th pixel value of the type V marine snow $v(m, n)$ is formulated as

TABLE I: Parameter settings for marine snow synthesis where $\mathcal{U}(v_{\min}, v_{\max})$ is a continuous uniform distribution between $v_{\min}$ and $v_{\max}$.

Parameter	Setting
c in (1)	$\mathcal{U}\{10, 80\}^*$
c_r in (2)	$\mathcal{U}\{1, 40\}^*$
Gaussian filter size in (3)	$(2r+1) \times (2r+1)$ where $r \sim \mathcal{U}(1, 6)$
a_2	$\begin{cases}\mathcal{U}(1,3) & \text{for small marine snow artifacts}\\ \mathcal{U}(3,16) & \text{for large marine snow artifacts}\end{cases}$
b_2	$\begin{cases}\mathcal{U}(0.3a_2, a_2) & \text{for small marine snow artifacts}\\ \mathcal{U}(0.8a_2, a_2) & \text{for large marine snow artifacts}\end{cases}$
a_1	$\mathcal{U}(a_2, 2a_2)^\dagger$
a_3	$\mathcal{U}(0.5a_2, a_2)^\ddagger$

† b_1 is obtained from the eccentricity of $f_2(m, n)$
‡ b_3 is obtained from the eccentricity of $f_2(m, n)$

$v(m,n) = h(m,n) + g(m,n)$, where $g(m,n)$ is the *rim* term defined as

$$g(m,n) = \begin{cases} c_r\left(1 - \frac{d(\mathbf{p}, \frac{1}{2}(\mathbf{q}_2+\mathbf{q}_3))}{\frac{1}{2}d(\mathbf{q}_2,\mathbf{q}_3)}\right) & 1 < \frac{m^2}{a_3^2} + \frac{n^2}{b_3^2} \wedge \frac{m^2}{a_2^2} + \frac{n^2}{b_2^2} \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where c_r is the constant for the maximum rim height and $\mathbf{q}_3$ indicates the coordinates of the intersection of the ellipse f_3 and the straight line crossing (m, n).

The synthesized type V marine snow artifact is shown in Figs. 4(c) and (d), which are similar to the real versions in Figs. 2(c) and (d).

D. Synthesizing Images with Marine Snow

By using the marine snow models introduced in the previous subsections, we synthesize the marine snow with underwater images.

Let $U \in \mathbb{R}^{H \times W \times 3}$ be an original underwater image. First, we randomly select a target pixel coordinate (k, l) to append the marine snow. Second, we randomly select the major and minor axes of the ellipses a_i and b_i. In addition, the direction of the major axis is randomly set. Third, for each channel, type H or V marine snow is independently generated, where the parameters c and c_r are independently set. We denote an image with one marine snow artifact as M, where pixel values in M are all-zero except the marine snow artifact. Finally, U and M are combined as follows to yield the synthesized image I:

$$I_{k,l,t} = \begin{cases} M_{k,l,t} + G_{k,l,t} & M_{k,l,t} \geq 0 \\ U_{k,l,t} & \text{otherwise} \end{cases} \tag{3}$$

where $G \in \mathbb{R}^{H \times W \times 3}$ is the Gaussian filtered version of U that reflects blur owing to marine snow.

We iterate the process until a predetermined number of marine snow artifacts are appended.

IV. Dataset Specifications

In this section, we present the specifications of synthesized marine snow artifacts in the MSRB Dataset. It is designed for two tasks:

Task 1: Removal of small-sized marine snow artifacts.
Task 2: Removal of various-sized marine snow artifacts.

Clearly, Task 2 is more difficult than Task 1.

First, detailed parameters and setups shared in both tasks are presented. Second, we introduce MSR Tasks with corresponding synthesized images.

A. General Setup

Each sub-dataset corresponding to an MSR Task contains $2,300$ training image pairs and 400 test image pairs, all having a pixel resolution of 384×384. An image pair contains one original underwater image and one image containing synthesized marine snow artifacts. All original images are collected from flickr[2] under a Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic (CC BY-NC-SA 2.0) License.

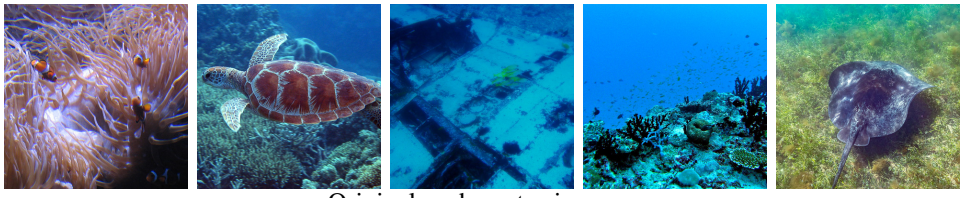
Each synthesized image contains N marine snow particles where N is chosen from $\mathcal{U}\{100, 600\}$, in which $\mathcal{U}\{v_{\min}, v_{\max}\}$ is a discrete uniform distribution between $v_{\min}$ and $v_{\max}$. In each synthesized image, type H and V marine snow particles are randomly generated with a probability of 0.7 for type H and 0.3 for type V, according to our preliminary observations. The other representative parameters introduced in the previous section are listed in Table I. Most of the parameters are determined according to our preliminary observations of real underwater images with marine snow artifacts.

B. MSR Task 1: Removing Small-Sized Marine Snow Artifacts

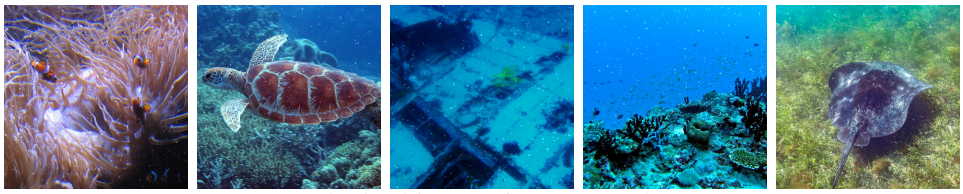
We introduce the first MSR Task: Removal of small-sized marine snow artifacts. In Task 1, the maximum width/height of the artifacts is restricted to 6 pixels, which corresponds to roughly 1.6% of the image width/height. Task 1 is designed for underwater scenes where particles are relatively far from the photosensors. Because the task is relatively simple, the conventional MF approach is expected to be successfully applied (at least partially).

Examples of test image pairs in MSR Task 1 are shown in Fig. 5. As clearly visualized, the synthesized images have various sized (but small) marine snow artifacts.

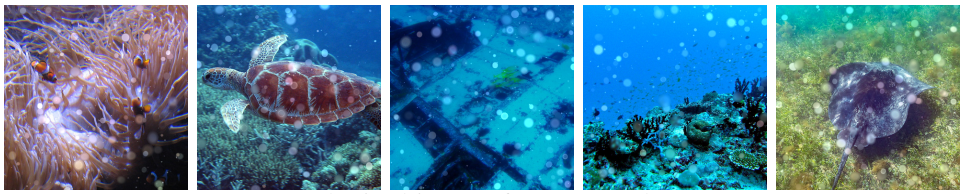
[2] https://www.flickr.com/



Original underwater images.

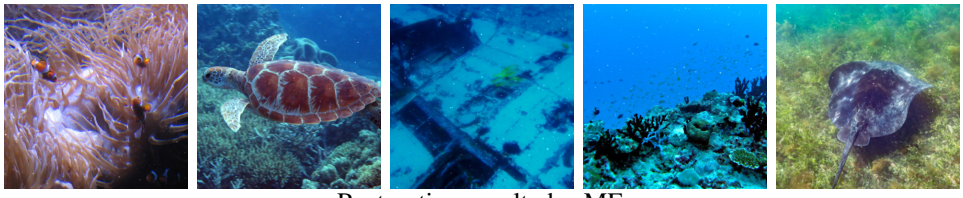


Synthesized images for MSR Task 1.

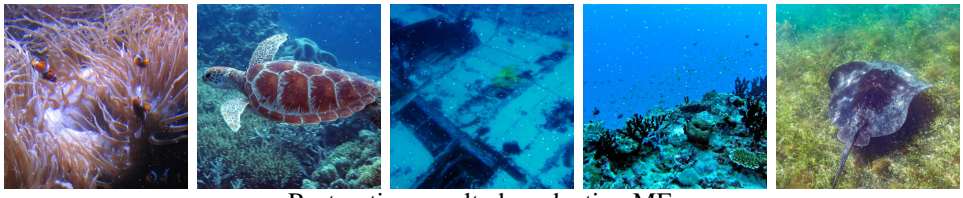


Synthesized images for MSR Task 2.

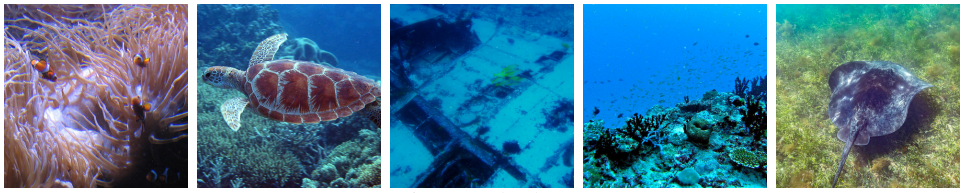
Fig. 5: Example images in MSRB dataset.



Restoration results by MF.



Restoration results by adaptive MF.



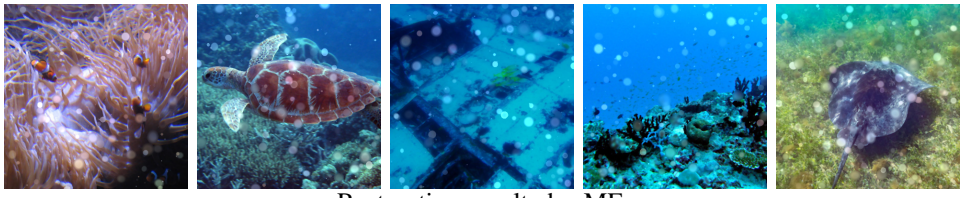
Restoration results by U-Net.

Fig. 6: MSR Task 1 results.

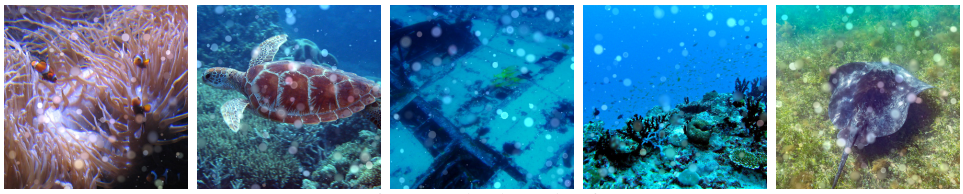
C. MSR Task 2: Removing Various-Sized Marine Snow Artifacts

The second MSR task is designed for underwater scenes containing particles in various distances. As a result, the synthesized images for Task 2 have small- and large-sized marine snow artifacts. For large-sized artifacts, we set the largest width/height of marine snow to 32 pixels, which corresponds to 8.3% compared to the image width. Furthermore, the probabilities of small- and large-sized artifacts are set to 0.7 and 0.3, respectively.

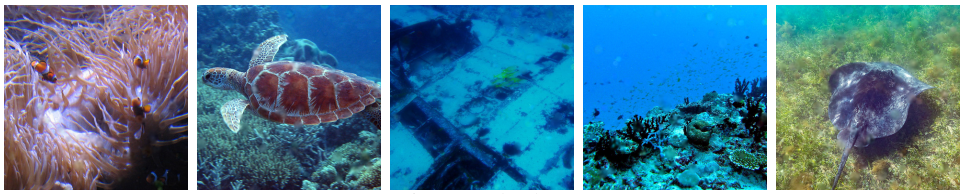
Training image examples for Task 2 are also shown in Fig. 5. The images have various-sized artifacts and the degraded areas



Restoration results by MF.



Restoration results by adaptive MF.



Restoration results by U-Net.

Fig. 7: MSR Task 2 results.

are larger than the images in Task 1.

In the following, we show the benchmarking results of MSR as well as the real MSR.

V. MSR Benchmarking Results

In this section, we present the first benchmarking performance for MSR Tasks 1 and 2. The methods used for the benchmarking are 1) MF [19], 2) adaptive MF [5], and 3) U-Net [29]. The kernel size of MFs is set to 3×3 or 5×5 pixels. Thanks to the MSRB Dataset, we can train a deep neural network for removing (synthesized) marine snow. For a benchmarking purpose, we train U-Net [29] using the dataset. The parameter settings are slightly different from the original.

The MSR results for Tasks 1 and 2 are shown in Figs. 6 and 7, respectively. As clearly observed, MF oversmoothes the images, which results in blur even for regions without marine snow artifacts. The adaptive MF suppresses blur; however, many artifacts are not removed mainly owing to the threshold-based algorithm. Furthermore, both MFs do not operate successfully for large marine snow artifacts in Task 2. By contrast, U-Net successfully removes marine snow for both tasks. Even for Task 2, it is able to suppress the large marine snow artifacts because of its multiresolution structure and the large number of trainable parameters.

For an objective comparison, we compute the average PSNRs and SSIMs over the test datasets. We summarize the results in Table II. The objective measure indicates that U-Net is superior to the MFs for both MSR tasks. Note that all MFs present smaller PSNRs/SSIMs than those of the synthesized images because they filtered areas without marine snow. By contrast, U-Net achieves a better objective quality than the synthesized images, which implies the effect of a deep neural network for MSR.

TABLE II: Average PSNRs and SSIMs over the test dataset.

	Task 1		Task 2	
Method	PSNR	SSIM	PSNR	SSIM
MF (3×3)	28.55	0.846	22.81	0.770
MF (5×5)	25.98	0.711	21.93	0.645
Adaptive MF (3×3)	29.88	0.910	23.35	0.842
Adaptive MF (5×5)	28.08	0.861	22.83	0.794
U-Net	**36.82**	**0.978**	**30.95**	**0.932**
Synthesized image	32.20	0.945	23.83	0.876

VI. Real Marine Snow Removal Examples

We construct the MSRB Dataset to mimic real marine snow artifacts. In this section, we show the real MSR results and present limitations of the dataset.

A. Real Marine Snow Removal

Fig. 8 shows MSR results for some real images. We use MF, adaptive MF, and U-Net as in the previous section. In addition, Deep WaveNet [30] is applied to validate the regular underwater enhancement model works for marine snow removal. U-Net is trained using MSRB Task 2 Dataset. The trained model of Deep WaveNet is taken from the original authors' GitHub repository.

As clearly observed, MFs cannot suppress marine snow artifacts because of their various sizes and different transparencies. Deep WaveNet also fails to remove marine snow artifacts. That means marine snow cannot be suppressed by underwater image enhancement techniques alone. U-Net presents a better

restoration image quality than MFs: It (partly) suppresses artifacts having both small and large sizes. This implies the effectiveness of deep-learning-based methods trained using the MSRB Dataset for real MSR.

Note that thick marine snow artifacts remain in the restored images. The reason for this could be two-fold: First, U-Net is not specifically designed for MSR although we slightly customized its structure. A neural network designed specifically for MSR would improve the performance. Second, our proposed model classifies marine snow artifacts into two representative types. Marine snow artifacts not fitted to these types are not well removed. This limitation is described in the next section.

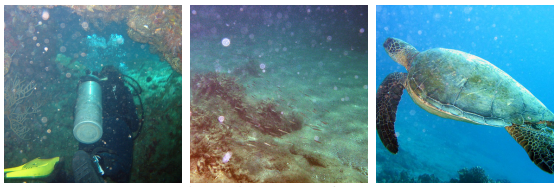
B. Limitations

Because our MSR benchmarking dataset is based on synthesizing marine snow artificially into real underwater images, some limitations exist. First, the dataset is designed for removing marine snow artifacts and does not aim at achieving other underwater image enhancements such as color correction. The design of a simultaneous restoration of MSR and color correction is an interesting area of future study.

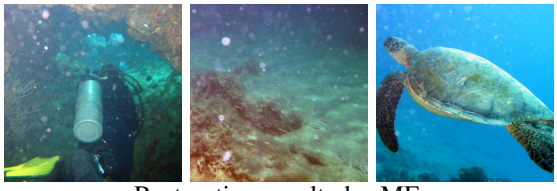
Although U-Net trained using the MSRB Dataset presents promising results, the dataset itself is imperfect. Specifically, real MSR performances are occasionally limited if MSR artifacts are not fitted to our marine snow models. Examples are shown in Fig. 9, which compares real underwater images and their restoration results using U-Net. Typically, U-Net fails to suppress artifacts when marine snow artifacts are larger and denser than our MSR benchmarking dataset. Moreover, if the lighting conditions differ from the dataset, the current version of U-Net will fail. Through a future study, this can be improved by carefully updating the MSRB Dataset to reflect various underwater scenes.

VII. Conclusions

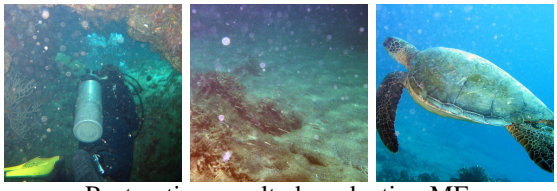
In this paper, the *Marine Snow Removal Benchmarking Dataset* was proposed. We mathematically modeled two representative marine snow artifacts and synthesized them in real underwater images. Two MSR tasks, the removal of small-sized and various-sized marine snow artifacts were also proposed. The first benchmarking results were also shown, which revealed the effectiveness of a deep neural network for MSR compared to median filter-based methods. MSR for real underwater images indicates that our MSRB Dataset contributes to real MSR problems. Our future studies will include improving the marine snow models as well as designing deep neural networks specifically designed for MSR.



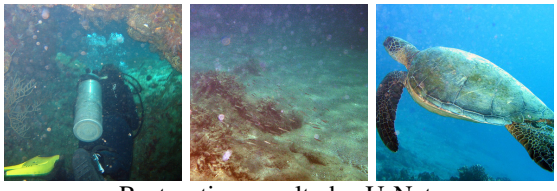
Real underwater images.



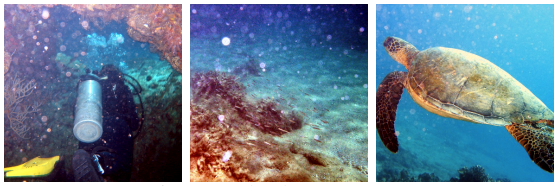
Restoration results by MF.



Restoration results by adaptive MF.



Restoration results by U-Net.



Restoration results by Deep WaveNet.

Fig. 8: Real MSR results.

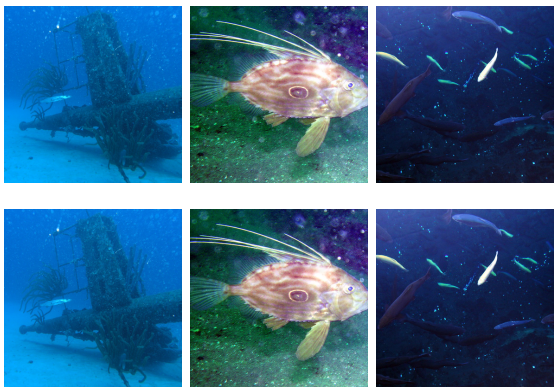


Fig. 9: Limitations. Top: Original images. Bottom: Restoration results by U-Net.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6723–6732, 2018. 1

[2] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2019. 1

[3] Saeed Anwar and Chongyi Li. Diving deeper into underwater image enhancement: A survey. *Signal Processing: Image Communication*, 89:115978, Nov. 2020. 1

[4] Saeed Anwar, Chongyi Li, and Fatih Porikli. Deep Underwater Image Enhancement. *arXiv:1807.03528*, July 2018. 1

[5] Soma Banerjee, Gautam Sanyal, Shatadal Ghosh, Ranjit Ray, and Nath Shome, Shome. Elimination of Marine Snow effect from underwater

image - An adaptive probabilistic approach. In *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science*, pages 1–4, Mar. 2014. 1, 2, 6

[6] Dana Berman, Tali Treibitz, and Shai Avidan. Diving into haze-lines: Color restoration of underwater images. In *Proc. British Machine Vision Conference (BMVC)*, volume 1, 2017. 1

[7] Matthieu Boffety and Frédéric Galland. Phenomenological marine snow model for optical underwater image simulation: applications to color restoration. In *2012 Oceans-Yeosu*, pages 1–6, 2012. 2, 3

[8] Matthieu Boffety, Frédéric Galland, and Anne-Gaëlle Allais. Color image simulation for underwater optics. *Applied optics*, 51(23):5633–5642, 2012. 2, 3

[9] David RK Brownrigg. The weighted median filter. *Communications of the ACM*, 27(8):807–818, 1984. 2

[10] Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016. 1

[11] Chang Chen, Zhiwei Xiong, Xinmei Tian, and Feng Wu. Deep boosting for image denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018. 1

[12] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10):1797–1801, 2014. 1

[13] Boguslaw Cyganek and Karol Gongola. Real-time marine snow noise removal from underwater video sequences. *Journal of Electronic Imaging*, 27(4):043002, July 2018. 1, 2

[14] Akshay Dudhane, Praful Hambarde, Prashant Patil, and Subrahmanyam Murala. Deep underwater image restoration and beyond. *IEEE Signal Processing Letters*, 27:675–679, 2020. 1

[15] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing Underwater Imagery Using Generative Adversarial Networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165, May 2018. 1

[16] Fahimeh Farhadifard, Martin Radolko, and Uwe Freiherr von Lukas. *Marine Snow Detection and Removal: Underwater Image Restoration Using Background Modeling*. Václav Skala - UNION Agency, 2017. 2

[17] Fahimeh Farhadifard, Martin Radolko, and Uwe von Lukas. Single image marine snow removal based on a supervised median filtering scheme. In *12th International Conference on Computer Vision Theory and Applications, VISIGRAPP 2017. Proceedings. Vol.4: VISAPP*, pages 280–287, 2017. 2

[18] Dongsheng Guo, Yiqing Huang, Tianshun Han, Haiyong Zheng, Zhaorui Gu, and Bing Zheng. Marine snow removal. In *OCEANS 2022 - Chennai*, pages 1–7, 2022. 1, 2

[19] Thomas Huang, GJTGY Yang, and Greory Tang. A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(1):13–18, 1979. 2, 6

[20] Qin Jiang, Yang Chen, Guoyu Wang, and Tingting Ji. A novel deep neural network for noise removal from underwater image. *Signal Processing: Image Communication*, 87:115921, 2020. 1

[21] Michał Koziarski and Bogusław Cyganek. Marine Snow Removal Using a Fully Convolutional 3D Neural Network Combined with an Adaptive Median Filter. In *Pattern Recognition and Information Forensics*, pages 16–25, 2019. 1

[22] Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, 30:4985–5000, 2021. 1

[23] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020. 1

[24] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020. 1

[25] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robotics and Automation Letters*, 3(1):387–394, Jan. 2018. 1

[26] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[27] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, 2016. 1

[28] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. 1

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. 6

[30] Prasen Sharma, Ira Bisht, and Arijit Sur. Wavelength-based attributed deep neural network for underwater image restoration. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1):1–23, 2023. 6

[31] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017. 1

[32] Takumi Ueda, Koki Yamada, and Yuichi Tanaka. Underwater image synthesis from rgb-d images and its application to deep underwater image restoration. In *Proc. 2019 IEEE International Conference on Image Processing (ICIP)*, pages 2115–2119. IEEE, 2019. 1

[33] Yang Wang, Jing Zhang, Yang Cao, and Zengfu Wang. A deep CNN method for underwater image enhancement. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1382–1386, Sept. 2017. 1

[34] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15(12):1090–1097, 2018. 1

[35] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25:341–349, 2012. 1

[36] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pages 1790–1798, 2014. 1

[37] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 1

[38] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, July 2017. 1

[39] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016. 1

[40] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004. 1

[41] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 1

[42] Peixian Zhuang, Jiamin Wu, Fatih Porikli, and Chongyi Li. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31:5442–5455, 2022. 1

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Cross-Frame Foreground Structural Similarity Modeling by Convolutional Sparse Representation

Kazuki Naganuma* and Shunsuke Ono*
* Tokyo Institute of Technology, Kanagawa, Japan
E-mail: naganuma.k.aa@m.titech.ac.jp; ono@c.titech.ac.jp

***Abstract*—In this paper, we propose a foreground-background separation (FBS) method using a novel foreground model based on convolutional sparse representation (CSR). Existing FBS methods are based on the assumption that the foreground component is sparse or continuous in time, resulting in poor separation performance for videos containing sparse noise/fast moving objects or for low frame rate ones. Therefore, we propose to use CSR to characterize the similar structure of the foreground component scattered in each frame. This CSR-based model can adaptively capture the "shape" of the foreground component and thus can accurately separate it even under the above difficulties. We formulate FBS as an optimization problem incorporating the CSR-based model and develope an algorithm to solve it based on alternating minimization. Experiments using an infrared video and an electron microscope video show that the proposed foreground model is superior to existing models.**

I. Introduction

Foreground-background separation (FBS) is a preprocessing technique for decomposing video data into dynamic foreground and static background components [1]. These components contain helpful information for subsequent processing, such as motion detection [2], moving object detection [3], and background subtraction [4]. However, since video data (especially those obtained from specialized measurements such as infrared videos and microscope videos) often contains various types of noise, it is very important to establish noise-robust FBS techniques.

To cope with the presence of noise and the lack of ground-truth data, many FBS methods take the approach of solving optimization problems that incorporate functions to characterize each component. There are two main categories of existing foreground models: the sparsity-based model [5], [6] and the total variation model [7]–[11]. The first model captures the sparsity of the foreground component but does not distinguish it from sparse noise. The second model characterizes smoothly moving objects based on their spatial correlation and temporal continuity. However, this model is unable to separate the foreground component from low frame-rate videos containing fast-moving objects.

Most types of foreground objects have similar spatial structure across frames (we call it *cross-frame similar structure*). As a typical example, Figure 1 shows that the shape of the airplane in the foreground component does not change over time. By capturing the cross-frame similar structure, the foreground component can be separated from noisy video data where the positions of foreground objects change significantly.

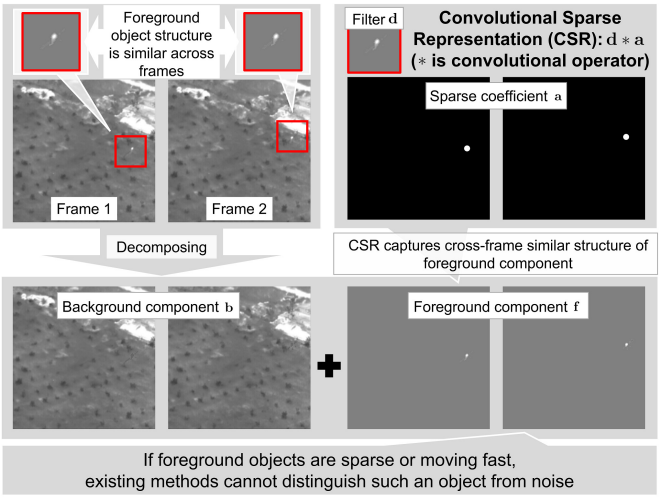


Fig. 1. Cross-frame similar structure in video data. The video data includes the complex ground scenes in the background component and an airplane in the foreground component.

In this paper, we propose a FBS method using a foreground model based on convolutional sparse representation (CSR) [12]–[15]. CSR can capture the specific spatial structure scattered throughout data [16]–[18]. The main contributions are as follows.

- We propose to use CSR to characterize the cross-frame similar structure of the foreground component scattered in each frame. This CSR-based model can adaptively capture the "shape" of the foreground component and thus can accurately separate it even under the aforementioned difficulties.
- We formulate FBS as an optimization problem incorporating CSR. The formulation is designed to separate the foreground and background components while removing sparse and Gaussian-like random noise.
- We develop an algorithm for solving optimization problems based on alternating optimization (ALM). ALM requires solving subproblems with certain variables fixed, and in our case, solvers for these subproblems are established based on the preconditioned primal-dual splitting algorithm (P-PDS) [19]–[21] and the fast iterative shrinkage-thresholding algorithm (FISTA) [22]. P-PDS can automatically adjust the appropriate stepsizes based on the problem structure, while FISTA converges at a

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

very fast rate, allowing each subproblem to be solved efficiently.

Experiments using an infrared (IR) video and an electron microscope (EM) video show that the proposed foreground model is superior to existing models.

II. Preliminaries

A. Notation

The ℓ_1, ℓ_2, and operator norms are defined as $\|\mathbf{x}\|_1 := \sum_i |\mathbf{x}_i|$, $\|\mathbf{x}\|_2 := \sqrt{\sum_i \mathbf{x}_i^2}$, and $\|\mathfrak{L}\|_{\mathrm{op}} := \sup_{\mathbf{x}\neq\mathbf{0}} \|\mathfrak{L}(\mathbf{x})\|_2/\|\mathbf{x}\|_2$, respectively. We denote the ℓ_p ball centered by $\mathbf{c}$ with a radius α as $B^{\mathbf{c}}_{p,\alpha} := \{\mathbf{y} \,|\, \|\mathbf{y}-\mathbf{c}\|_p \leq \alpha\}$. The adjoint operator of a linear operator $\mathfrak{L}$ is denoted as $\mathfrak{L}^*$.

B. Convolutional Sparse Representation (CSR) [14]

CSR [14] models data $\mathbf{x}$ as the sum of convolutions between dictionary filters $\mathbf{d} := \{\mathbf{d}_d \,|\, d = 1,\dots,D\}$ and sparse coefficient $\mathbf{a} := \{\mathbf{a}_d \,|\, d = 1,\dots,D\}$:

$$\begin{aligned}\min_{\mathbf{d},\mathbf{a}} \frac{1}{2}\|\mathbf{x} - \sum_{d=1}^{D} \mathbf{d}_d * \mathbf{a}_d\|_2^2 + \lambda \sum_{d=1}^{D} \|\mathbf{a}_d\|_1 \\ \text{s.t. } \mathbf{d}_d \in S_{2,1} := \{\mathbf{y} \,|\, \|\mathbf{y}\|_2 = 1\}\,(d = 1,\dots,D),\end{aligned} \quad (1)$$

where $*$ denotes the convolution operator. The first term guarantees the fidelity between the data $\mathbf{x}$ and the sum of convolutions. The second term promotes the sparsity of the coefficient $\mathbf{a}$. By combining these terms using the appropriate balancing parameter $\lambda > 0$, CSR can capture specific spatial structure scattered throughout the data $\mathbf{x}$. The unit ℓ_2 sphere constraints on filters $\mathbf{d}_d$ allows us to avoid ambiguity between filters and coefficients. The standard approach to solve Prob. (1) is ALM with respect to filters and coefficients.

III. Proposed Method

A. Problem Formulation

Consider that an observed video (of size $n_1 \times n_2 \times n_3$) $\mathbf{v} \in \mathbb{R}^{n_1 n_2 n_3}$ is modeled by

$$\mathbf{v} = \bar{\mathbf{f}} + \bar{\mathbf{b}} + \bar{\mathbf{s}} + \mathbf{n}, \quad (2)$$

where $\bar{\mathbf{f}}$, $\bar{\mathbf{b}}$, $\bar{\mathbf{s}}$, and $\mathbf{n}$ are the true foreground component, the true background component, sparsely distributed noise such as outliers, and random noise, respectively.

Based on the model in (2), we newly formulate FBS as the following optimization problem:

$$\begin{aligned}\min_{\mathbf{f},\mathbf{b},\mathbf{s},\mathbf{d},\mathbf{a}} \frac{1}{2}\|\mathbf{f} - \sum_{d=1}^{D} \mathbf{d}_d * \mathbf{a}_d\|_2^2 + \lambda \sum_{d=1}^{D} \|\mathbf{a}_d\|_1 + \sum_{i=1}^{K} R_i(\mathbf{L}_i\mathbf{b}) \\ \text{s.t. } P_M(\mathbf{f}+\mathbf{b}+\mathbf{s}) \in B^{\mathbf{v}_M}_{2,\varepsilon},\ \mathbf{s} \in B^{\mathbf{0}}_{1,\eta_{\mathbf{s}}},\ \mathbf{f} \in B^{\mathbf{0}}_{1,\eta_{\mathbf{f}}}, \\ \mathbf{d}_d \in S_{2,1}\,(d = 1,\dots,D).\end{aligned} \quad (3)$$

For handling dynamic background scenes arising from a moving camera, Prob. (3) contains the orthogonal projection P_M. This projection maps panoramic video data onto a subspace M that represents dynamic scenes [8]. The first and second terms and the fourth constraint build CSR that captures the cross-frame similar structure of foreground objects. The third term is a general form of background modeling including linear operators $\mathbf{L}_i$ and functions R_i.[1] The first constraint guarantees the data-fidelity to projected observed data $\mathbf{v}_M = P_M(\mathbf{v})$. The second and third constraints guarantee the sparsity of noise and the foreground component. Using these constraints instead of an data-fidelity and a sparse terms allow us to easily adjust hyperparameters since parameters ε, $\eta_{\mathbf{s}}$, and $\eta_{\mathbf{f}}$ can be determined based only on noise and foreground component intensities. Indeed, this kind of constrained formulation has played an important role in facilitating the parameter setting of signal recovery problems [23]–[29].

Algorithm 1 Solver for the main FBS problem (3)

Input: $\mathbf{f}^{(0)}, \mathbf{b}^{(0)}, \mathbf{s}^{(0)}, \mathbf{a}^{(0)}, \mathbf{d}^{(0)}$
Output: $\mathbf{f}^{(t)}, \mathbf{b}^{(t)}, \mathbf{a}^{(t)}$

1: Initialize $t = 0, \mathbf{f}^{(0)}, \mathbf{b}^{(0)}, \mathbf{s}^{(0)}, \mathbf{d}^{(0)}, \mathbf{a}^{(0)}$;
2: **while** A stopping criterion are not satisfied **do**
3: Updating $\mathbf{f}$, $\mathbf{b}$, $\mathbf{s}$, $\mathbf{a}$ with fixed $\mathbf{d}^{(t)}$ by solving Prob. (4) via Algorithm 2;
4: Updating $\mathbf{d}$ with fixed $\mathbf{f}^{(t+1)}$, $\mathbf{b}^{(t+1)}$, $\mathbf{s}^{(t+1)}$, $\mathbf{a}^{(t+1)}$ by solving Prob. (10) via Algorithm 3;
5: $t \leftarrow t+1$;

B. Optimization

To solve Prob. (1), we alternately update the variables $\mathbf{f}$, $\mathbf{s}$, $\mathbf{b}$, $\mathbf{a}$ and the variable $\mathbf{d}$ as shown in Algorithm 1. The details of algorithms solving each subproblem are as follows.

1) Updating $\mathbf{f}$, $\mathbf{s}$, $\mathbf{b}$, *and* $\mathbf{a}$*:* With $\mathbf{d}$ fixed, the subproblem with respect to $\mathbf{f}$, $\mathbf{s}$, $\mathbf{b}$, and $\mathbf{a}$ in Prob. (3) can be written as the following optimization problem:

$$\begin{aligned}\min_{\mathbf{f},\mathbf{b},\mathbf{s},\mathbf{a}} \frac{1}{2}\|\mathbf{f} - \sum_{d=1}^{D} \mathbf{d}_d * \mathbf{a}_d\|_2^2 + \lambda \sum_{d=1}^{D} \|\mathbf{a}_d\|_1 + \sum_{i=1}^{K} R_i(\mathbf{L}_i\mathbf{b}) \\ \text{s.t. } P_M(\mathbf{f}+\mathbf{b}+\mathbf{s}) \in B^{\mathbf{v}_M}_{2,\varepsilon}, \mathbf{s} \in B^{\mathbf{0}}_{1,\eta_{\mathbf{s}}}, \mathbf{f} \in B^{\mathbf{0}}_{1,\eta_{\mathbf{f}}}.\end{aligned} \quad (4)$$

Using indicator functions[2] $\iota_{B^{\mathbf{v}_M}_{2,\varepsilon}}$, $\iota_{B^{\mathbf{0}}_{1,\eta_{\mathbf{f}}}}$, and $\iota_{B^{\mathbf{0}}_{1,\eta_{\mathbf{s}}}}$ of norm balls $B^{\mathbf{v}_M}_{2,\varepsilon}$, $B^{\mathbf{0}}_{1,\eta_{\mathbf{s}}}$, and $B^{\mathbf{0}}_{1,\eta_{\mathbf{f}}}$, we translate Prob. (4) into the following equivalent optimization problem:

$$\begin{aligned}\min_{\mathbf{f},\mathbf{b},\mathbf{s},\mathbf{a}} \iota_{B^{\mathbf{0}}_{1,\eta_{\mathbf{f}}}}(\mathbf{f}) + \iota_{B^{\mathbf{0}}_{1,\eta_{\mathbf{s}}}}(\mathbf{s}) + \lambda \sum_{d=1}^{D} \|\mathbf{a}_d\|_1 + \sum_{i=1}^{K} R_i(\mathbf{L}_i\mathbf{b}) \\ + \frac{1}{2}\|\mathbf{f} - \sum_{d=1}^{D} \mathbf{d}_d * \mathbf{a}_d\|_2^2 + \iota_{B^{\mathbf{v}_M}_{2,\varepsilon}}(P_M(\mathbf{f}+\mathbf{b}+\mathbf{s})).\end{aligned} \quad (5)$$

Prob. (5) can be solved by P-PDS [19], [20].

[1] For example, if we respectively set $\mathbf{L}$ and R to the identity operator and the nuclear norm, the third term in (3) characterizes the foreground component as low-rank, as seen in [5].

[2] For a given nonempty closed set C, the indicator function of C is defined as $\iota_C(\mathcal{X}) := 0$, if $\mathcal{X} \in C$; ∞, otherwise.

Algorithm 2 P-PDS solver for the subproblem (4)

Input: $\mathbf{f}^{(0)}, \mathbf{b}^{(0)}, \mathbf{s}^{(0)}, \mathbf{a}^{(0)}, \mathbf{d}^{(t)}$
Output: $\mathbf{f}^{(l)}, \mathbf{b}^{(l)}, \mathbf{a}^{(l)}$

1: Initialize $l = 0, \mathbf{z}_{1,1}, \ldots, \mathbf{z}_{1,K}, \mathbf{z}_2, \mathbf{z}_3$;
2: Setting stepsizes $\gamma_{\mathbf{f}}, \gamma_{\mathbf{b}}, \gamma_{\mathbf{s}}, \gamma_{\mathbf{a}}, \gamma_{\mathbf{z}}$ as (6);
3: **while** A stopping criterion is not satisfied **do**
4: $\quad \mathbf{f}' \leftarrow \mathbf{f}^{(l)} - \gamma_{\mathbf{f}}(\mathbf{z}_2^{(l)} + P_M(\mathbf{z}_3^{(l)}))$;
5: $\quad \mathbf{f}^{(l+1)} \leftarrow \mathrm{prox}_{\gamma_{\mathbf{f}} \iota_{B^{\mathbf{0}}_{1,\eta_{\mathbf{f}}}}}(\mathbf{f}')$ by the algorithm in [30];
6: $\quad \mathbf{b}^{(l+1)} \leftarrow \mathbf{b}^{(l)} - \gamma_{\mathbf{b}}(\sum_{i=1}^{K} \mathbf{L}_K^{\top} \mathbf{z}_{1,K}^{(l)} + P_M(\mathbf{z}_3^{(l)}))$;
7: $\quad \mathbf{s}' \leftarrow \mathbf{s}^{(l)} - \gamma_{\mathbf{s}} P_M(\mathbf{z}_3)$;
8: $\quad \mathbf{s}^{(l+1)} \leftarrow \mathrm{prox}_{\gamma_{\mathbf{s}} \iota_{B^{\mathbf{0}}_{1,\eta_{\mathbf{s}}}}}(\mathbf{s}')$ by the algorithm in [30];
9: $\quad$ **for** $d = 1, \cdots, D$ **do**
10: $\quad\quad \mathbf{a}_d' \leftarrow \mathbf{a}_d^{(l)} - \gamma_{\mathbf{a}_d}(\mathfrak{C}^*_{\mathbf{d}_d^{(t)}}(-\mathbf{z}_2^{(l)}))$;
11: $\quad\quad \mathbf{a}_d^{(l+1)} \leftarrow \mathrm{prox}_{\gamma_{\mathbf{a}_d} \lambda \|\cdot\|_1}(\mathbf{a}_d')$;
12: $\quad$ **for** $i = 1, \ldots, K$ **do**
13: $\quad\quad \mathbf{z}_{1,i}' \leftarrow \mathbf{z}_{1,i}^{(l)} + \gamma_{\mathbf{z}} \mathbf{L}_i (2\mathbf{b}^{(l+1)} - \mathbf{b}^{(l)})$;
14: $\quad\quad \mathbf{z}_{1,i}^{(l+1)} \leftarrow \mathbf{z}_{1,i}' - \gamma_{\mathbf{z}} \mathrm{prox}_{\frac{1}{\gamma_{\mathbf{z}}} R_i}(\frac{1}{\gamma_{\mathbf{z}}} \mathbf{z}_{1,i}')$;
15: $\quad \mathbf{z}_2' \leftarrow (2\mathbf{f}^{(l+1)} - \mathbf{f}^{(l)}) - \sum_{d=1}^{D} \mathfrak{C}_{\mathbf{d}_d^{(t)}}(2\mathbf{a}_d^{(l+1)} - \mathbf{a}_d^{(l)})$;
16: $\quad \mathbf{z}_2^{(l+1)} \leftarrow (\mathbf{z}_2^{(l)} - \gamma_{\mathbf{z}} \mathbf{z}_2') - \gamma_{\mathbf{z}} \mathrm{prox}_{\frac{1}{2\gamma_{\mathbf{z}}} \|\cdot\|_2^2}(\frac{1}{\gamma_{\mathbf{z}}}(\mathbf{z}_2^{(l)} - \gamma_{\mathbf{z}} \mathbf{z}_2'))$;
17: $\quad \mathbf{z}_3' \leftarrow \mathbf{z}_3^{(l)} + \gamma_{\mathbf{z}} P_M(2(\mathbf{f}^{(l+1)} + \mathbf{b}^{(l+1)} + \mathbf{s}^{(l+1)}) - (\mathbf{f}^{(l)} + \mathbf{b}^{(l)} + \mathbf{s}^{(l)}))$;
18: $\quad \mathbf{z}_3^{(l+1)} \leftarrow \mathbf{z}_3' - \gamma_{\mathbf{z}} \mathrm{prox}_{\frac{1}{\gamma_{\mathbf{z}}} \iota_{B^{\mathbf{v}_M}_{2,\varepsilon}}}(\frac{1}{\gamma_{\mathbf{z}}} \mathbf{z}_3')$;
19: $\quad l \leftarrow l + 1$;

We show the detailed algorithm in Algorithm 2. Following a method [20], [21] that selects reasonable stepsize, the parameters $\gamma_{\mathbf{f}}, \gamma_{\mathbf{b}}, \gamma_{\mathbf{s}}, \gamma_{\mathbf{a}}, \gamma_{\mathbf{z}}$ are determined as follows:

$$\begin{aligned} &\gamma_{\mathbf{f}} = \tfrac{1}{2}, \gamma_{\mathbf{b}} = \tfrac{1}{1+\sum_{i=1}^{K} \|\mathbf{L}_i\|_{\mathrm{op}}^2}, \gamma_{\mathbf{s}} = 1, \\ &\gamma_{\mathbf{a}} = \tfrac{1}{\sum_{d=1}^{D} \|\mathfrak{C}_{\mathbf{d}_d}\|_{\mathrm{op}}^2}, \gamma_{\mathbf{z}} = \tfrac{1}{3+D}, \end{aligned} \tag{6}$$

where $\mathfrak{C}_{\mathbf{d}_d}(\mathbf{a}) = \mathbf{d}_d * \mathbf{a}$. The proximity operators[3] of $\|\cdot\|_1$, $\frac{1}{2}\|\cdot\|_2^2$, and $\iota_{B^{\mathbf{v}_M}_{2,\varepsilon}}$ (in lines 11, 16, and 18 of Algorithm 2) are respectively calculated as

$$[\mathrm{prox}_{\alpha\|\cdot\|_1}(\mathbf{x})]_i = \mathrm{sgn}(x_i) \max\{|x_i| - \alpha, 0\}, \tag{7}$$

$$\mathrm{prox}_{\frac{\alpha}{2}\|\cdot\|_2^2}(\mathbf{x}) = \frac{\mathbf{x}}{1+\alpha}, \tag{8}$$

$$\mathrm{prox}_{\alpha \iota_{B^{\mathbf{v}_M}_{2,\varepsilon}}}(\mathbf{x}) = \begin{cases} \mathbf{x}, & \text{if } \mathbf{x} \in B^{\mathbf{v}_M}_{2,\varepsilon}; \\ \mathbf{v}_M + \frac{\varepsilon(\mathbf{x}-\mathbf{v}_M)}{\|\mathbf{x}-\mathbf{v}_M\|_2}, & \text{otherwise.} \end{cases} \tag{9}$$

2) Updating $\mathbf{d}$*:* With $\mathbf{f}$, $\mathbf{s}$, $\mathbf{b}$, and $\mathbf{a}$ fixed, the subproblem with respect to $\mathbf{d}$ in Prob. (3) can be written as the following optimization problem:

$$\min_{\mathbf{d}} \frac{1}{2}\|\mathbf{f} - \sum_{d=1}^{D} \mathbf{d}_d * \mathbf{a}_d\|_2^2 \,\mathrm{s.t.}\, \mathbf{d}_d \in S_{2,1}\ (d = 1, \ldots, D). \tag{10}$$

[3]The proximity operator of a proper lower semicontinuous convex function f with a parameter $\alpha > 0$ is defined as $\mathrm{prox}_{\alpha f}(\mathbf{x}) := \mathrm{argmin}_{\mathbf{y}} f(\mathbf{y}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{y}\|_2^2$.

Algorithm 3 FISTA solver for the subproblem (10)

Input: $\mathbf{d}^{(0)}, \mathbf{f}^{(t)}, \mathbf{a}^{(t)}, \gamma$
Output: $\mathbf{d}^{(l)}$

1: Initialize $l = 0, \tau^{(0)} = 1, \mathbf{z}_1, \ldots, \mathbf{z}_D$;
2: **while** A stopping criterion is not satisfied **do**
3: $\quad \tau^{(l+1)} \leftarrow \frac{1+\sqrt{1+4\tau^{(l)}\tau^{(l)}}}{2}$;
4: $\quad$ **for** $d = 1, \ldots, D$ **do**
5: $\quad\quad \mathbf{d}_d' \leftarrow \mathbf{z}_d^{(l)} - \gamma \mathfrak{A}^*_{\mathbf{a}_d^{(t)}}(\sum_{d=1}^{D} \mathfrak{A}_{\mathbf{a}_d^{(t)}}(\mathbf{d}_d^{(l)}) - \mathbf{f}^{(t)})$;
6: $\quad\quad \mathbf{d}_d^{(l)} \leftarrow \mathrm{prox}_{\gamma \iota_{S_{2,1}}}(\mathbf{d}_d')$;
7: $\quad\quad \mathbf{z}_d^{(l+1)} \leftarrow \mathbf{d}_d^{(l+1)} + \left(\frac{\tau^{(l)}-1}{\tau^{(l+1)}}\right)(\mathbf{d}_d^{(l+1)} - \mathbf{d}_d^{(l)})$;
8: $\quad l \leftarrow l + 1$;

Using indicator function of $S_{2,1}$, we translate Prob. (10) into the following equivalent optimization problem:

$$\min_{\mathbf{d}} \frac{1}{2}\|\mathbf{f} - \sum_{d=1}^{D} \mathbf{d}_d * \mathfrak{A}_d\|_2^2 + \sum_{d=1}^{D} \iota_{S_{2,1}}(\mathbf{d}_d). \tag{11}$$

Prob. (11) is non-convex, but it can be approximately solved by FISTA [22].

We show the detailed algorithm in Algorithm 3. The stepsize parameter γ is set as $\gamma = \frac{1}{\sum_{d=1}^{D} \|\mathfrak{A}_{\mathbf{a}_d}\|_{\mathrm{op}}^2}$, where $\mathfrak{A}_{\mathbf{a}_d}(\mathbf{d}) = \mathbf{d} * \mathbf{a}_d$. The proximity operator of $\iota_{S_{2,1}}$ can be calculated as $\mathrm{prox}_{\gamma \iota_{S_{2,1}}}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ if $\mathbf{x} \neq \mathbf{0}$; $\frac{1}{n_d}\mathbf{1}$, otherwise [14], where n_d is the size of filters $\mathbf{d}_d$ and $\mathbf{1}$ is the all-one vector.

IV. Experiments and Discussion

A. Experimental Settings

To demonstrate the effectiveness of our CSR-based model, we compare it with existing foreground models: the sparsity-based model (SP) [5] and the total variation model (TV) [8]. To fairly compare the effectiveness of the foreground models, the same background model was used for all methods. Specifically, the static-scene constraint was used because it can exactly capture the static component [31]. We adjusted the parameters of the existing methods to achieve the best performances. All experiments were performed using MATLAB (R2022b) on a Windows 10 computer with an Intel core i9-10900K 3.7GHz processor, 32GB of RAM, and NVIDIA GeForce RTX 3090 GPU.

As the ground-truth IR and EM videos, we used Airplane video data (cropped into a size of $256 \times 256 \times 30$) from Science Data Bank [32] and T cell video data (cropped into a size of $100 \times 100 \times 30$) [33], respectively. The pixel values were normalized into $[0, 1]$. Contaminating these ground-truth videos by white Gaussian noise with 0.05 and 0.1 of the standard deviation σ and Salt & Pepper noise with 0 and 0.05 of the rate $p_{\mathbf{s}}$, we generated observed videos.

For Algorithm 1, we set the max iteration number to 300 and the stopping criterion to $\|\mathbf{b}^{(n+1)} - \mathbf{b}^{(n)}\|_2 / \|\mathbf{b}^{(n)}\|_2 \leq 10^{-5}$ and $\|\mathbf{f}^{(n+1)} - \mathbf{f}^{(n)}\|_2 / \|\mathbf{f}^{(n)}\|_2 \leq 10^{-5}$. Moreover, we set the maximum iterations of Algorithms 2 and 3 to 1000 and 10, respectively. The parameters $\eta_{\mathbf{s}}$ and ε in (3) were

TABLE I
PSNRS AND SSIMS OF THE FBS RESULTS.

Video	σ	$p_{\mathbf{s}}$	Measure	Results of $\mathbf{f}+\mathbf{b}$			Results of $\mathbf{f}$			Results of $\mathbf{b}$		
				SP	TV	Ours	SP	TV	Ours	SP	TV	Ours
Airplane	0.05	0	PSNR	28.56	30.29	**30.60**	31.11	45.62	**45.86**	26.14	30.14	**30.43**
			SSIM	0.8143	0.8885	**0.8936**	0.3169	0.9821	**0.9830**	0.6773	0.8864	**0.8912**
	0.1	0	PSNR	27.18	28.58	**29.32**	34.12	44.47	**44.67**	26.00	28.55	**29.31**
			SSIM	0.7991	0.8398	**0.8651**	0.6460	0.9792	**0.9803**	0.8443	0.8415	**0.8671**
	0.05	0.05	PSNR	24.70	**28.29**	28.02	26.89	42.13	**43.20**	20.74	**28.19**	27.98
			SSIM	0.7168	**0.8703**	0.8556	0.2872	0.9787	**0.9870**	0.6749	**0.8689**	0.8552
	0.1	0.05	PSNR	22.85	25.90	**25.97**	30.00	41.83	**43.30**	20.48	25.77	**25.94**
			SSIM	0.6878	0.8116	**0.8406**	0.4316	0.9769	**0.9902**	0.6707	0.8088	**0.8402**
T cell	0.05	0	PSNR	30.95	**33.80**	33.28	31.49	**34.39**	33.93	41.72	**51.86**	47.11
			SSIM	0.7280	0.7984	**0.8027**	0.6494	0.7410	**0.7496**	0.9998	**1.0000**	0.9999
	0.1	0	PSNR	28.46	31.16	**31.66**	28.83	31.61	**31.94**	40.24	**45.89**	43.46
			SSIM	0.5934	0.6920	**0.7254**	0.5059	0.6105	**0.6562**	0.9997	**0.9999**	**0.9999**
	0.05	0.05	PSNR	30.25	**33.64**	33.26	30.68	**34.26**	34.23	41.73	**55.37**	54.25
			SSIM	0.7376	0.8005	**0.8021**	0.6569	0.7454	**0.7960**	0.9998	**1.0000**	**1.0000**
	0.1	0.05	PSNR	29.51	**32.11**	31.74	30.12	32.69	**32.86**	40.51	**49.86**	47.44
			SSIM	0.7070	0.7543	**0.7589**	0.6679	0.6675	**0.7473**	0.9998	**1.0000**	0.9999

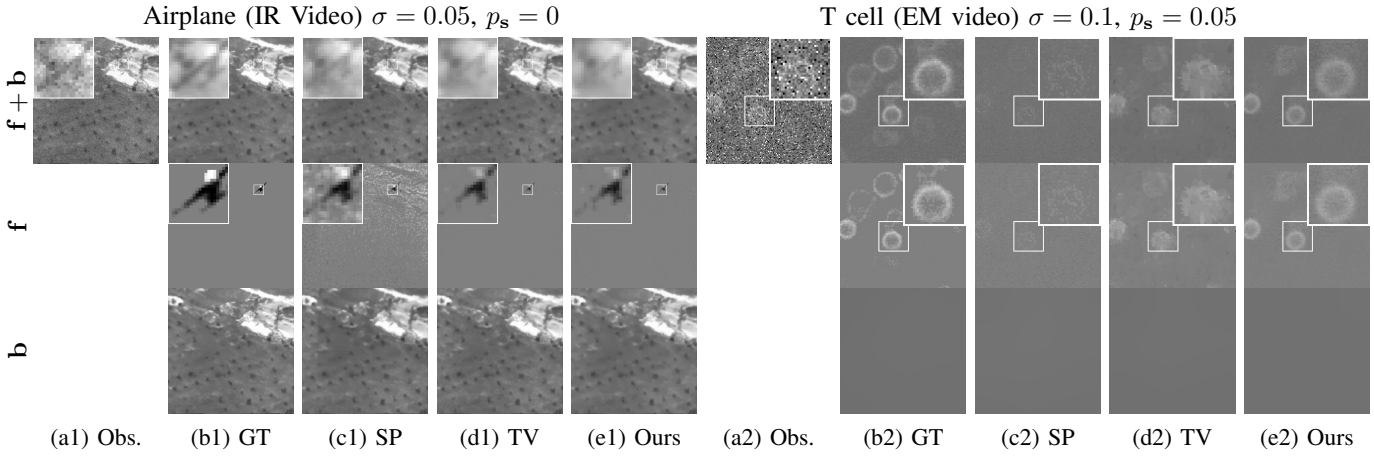
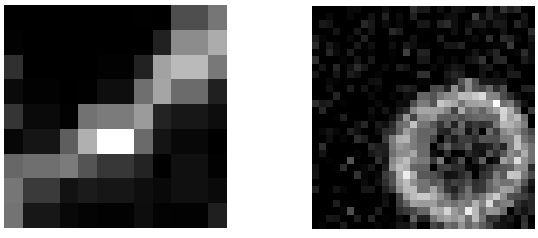


Fig. 2. FBS results. Left five columns are Airplane results ($\sigma = 0.05$ and $p_{\mathbf{s}} = 0$). Right five columns are T cell results ($\sigma = 0.1$ and $p_{\mathbf{s}} = 0.05$).



(a) Airplane (IR Video) (b) T cell (EM video)

Fig. 3. Some filters $\mathbf{d}$ obtained in the FBS process with our model. (a) shows airplane wing. (b) shows T cell.

set as $\eta_{\mathbf{s}} = 0.5 p_{\mathbf{s}} n_1 n_2 n_3$ and $\varepsilon = 0.9\sigma\sqrt{(1-p_{\mathbf{s}})n_1 n_2 n_3}$. The balancing parameter λ was manually adjusted to achieve the best performance. Specifically, we adopted $\lambda = 0.005$, $\lambda = 0.1$, $\lambda = 0.05$, and $\lambda = 0.1$ for Airplane ($p_{\mathbf{s}} = 0$), Airplane ($p_{\mathbf{s}} = 0.05$), T cell ($p_{\mathbf{s}} = 0$), and T cell ($p_{\mathbf{s}} = 0.05$), respectively. As the size of filters $\mathbf{d}_d$, 9×9 and 31×31 were used for Airplane and T cell, respectively. The parameter $\eta_{\mathbf{f}}$ determined from the size and intensity of foreground objects as $\eta_{\mathbf{f}} = 1080$ and $\eta_{\mathbf{f}} = 6615$ for Airplane and T cell, respectively. For the quality measures, we used the mean of the peak-signal-to-noise ratio (PSNR) [dB]: $\mathrm{PSNR}(\mathbf{q}, \bar{\mathbf{q}}) := 1/n_3 \sum_{k=1}^{n_3} 10\log_{10}(n_1 n_2 / \|\mathbf{q}_{n_3} - \bar{\mathbf{q}}_{n_3}\|_2^2)$, and the mean of structural similarity (SSIM) [34] for $\mathbf{f}+\mathbf{b}$, $\mathbf{f}$, and $\mathbf{b}$, where $\mathbf{q}_k$ is a kth frame of a video to be evaluated and $\bar{\mathbf{q}}_k$ is a reference frame to $\mathbf{q}_k$.

B. Experimental Results and Discussion

Table I shows PSNRs and SSIMs of FBS results, where the best values are highlighted in bold. Figure 2 depicts the FBS results of Airplane ($\sigma = 0.05$ and $p_{\mathbf{s}} = 0$) and T cell ($\sigma = 0.1$ and $p_{\mathbf{s}} = 0.05$). SP did not separate noise and foreground objects (see the middle row of Figs. 2 (c1) and (c2)), resulting in low PSNRs and SSIMs. TV resulted in higher PSNRs in some cases, but our method was better in

SSIMs for almost all cases. In particular, TV did not restore the spatial structure of airplane wing tips and cells (see Figs. 2 (d1) and (d2)). This is because the airplane and cells are not temporally continuous due to fast movement. In contrast, with our model, the foreground component was accurately separated (see Figs. 2 (e1) and (e2)). This is because our model characterizes the foreground component as the sum of convolutions between filters (Fig. 3) and sparse coefficients, and captures the cross-frame similarity structure.

V. Conclusion

We have proposed a novel FBS method incorporating CSR, which captures the cross-frame similar structure of the foreground component. By obtaining filters representing the shape of foreground objects, our method separates the foreground and background components from noisy video. Moreover, we have formulated the FBS as an optimization problem, and have developed an ALM algorithm for solving the problem. The experimental results have illustrated the superiority of our model to existing foreground models.

Acknowledgment

This work was supported Grant-in-Aid for JSPS Fellows under Grant 23KJ0912, in part by JST PRESTO under Grant JPMJPR21C4 and JST AdCORP under Grant JPMJKB2307, and in part by JSPS KAKENHI under Grant 22H03610, 22H00512, and 23H01415.

References

[1] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, and R. Otazo, "On the applications of robust pca in image and video processing," *Proc. IEEE*, vol. 106, no. 8, pp. 1427–1457, Aug. 2018.

[2] S.-C. Huang, "An advanced motion detection algorithm with video quality analysis for video surveillance systems," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 21, no. 1, pp. 1–14, Jan. 2011.

[3] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[4] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vis. Image Understanding*, vol. 122, pp. 4–21, May 2014.

[5] E. J. Candes, X. Li, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, Jun. 2011.

[6] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis with a new tensor nuclear norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 925–938, Apr. 2020.

[7] X. Cao, L. Yang, and X. Guo, "Total variation regularized RPCA for irregularly moving object detection under dynamic background," *IEEE Trans. Cybern.*, vol. 46, no. 4, pp. 1014–1027, Apr. 2016.

[8] B. E. Moore, C. Gao, and R. R. Nadakuditi, "Panoramic robust PCA for foreground-background separation on noisy, free-motion camera video," *IEEE Trans. Comput. Imag.*, vol. 5, no. 2, pp. 195–211, Jun. 2019.

[9] S. Javed, A. Mahmood, S. Al-Maadeed, T. Bouwmans, and S. K. Jung, "Moving object detection in complex scene using spatiotemporal structured-sparse RPCA," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 1007–1022, Feb. 2019.

[10] B. Shen, R. R. Kamath, H. Choo, and Z. Kong, "Robust tensor decomposition based background/foreground separation in noisy videos and its applications in additive manufacturing," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 1, pp. 583–596, 2023.

[11] B. Shen, W. Xie, and Z. J. Kong, "Smooth robust tensor completion for background/foreground separation with missing pixels: Novel algorithm with convergence guarantee," *J. Mach. Learn. Res.*, vol. 23, no. 217, pp. 1–40, Jan. 2022.

[12] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 391–398.

[13] B. Wohlberg, "Efficient convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 7173–7177.

[14] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[15] C. Garcia-Cardona and B. Wohlberg, "Convolutional dictionary learning: A comparative review and new algorithms," *IEEE Trans. Comput. Imag.*, vol. 4, no. 3, pp. 366–381, Sep. 2018.

[16] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014.

[17] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[18] B. Tu, J. Wang, G. Zhang, X. Zhang, and W. He, "Texture pattern separation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3602–3614, Sep. 2019.

[19] T. Pock and A. Chambolle, "Diagonal preconditioning for first order primal-dual algorithms in convex optimization," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1762–1769.

[20] K. Naganuma and S. Ono, "Operator-norm-based variable-wise diagonal preconditioning for automatic stepsize selection of a primal-dual splitting algorithm," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2022, pp. 2041–2045.

[21] K. Naganuma and S. Ono, "Variable-wise diagonal preconditioning for primal-dual splitting: Design and applications," *arXiv:2301.08468*, 2023. [Online]. Available: https://arxiv.org/abs/2301.08468

[22] A. Beck and T. M., "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[23] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, Mar. 2011.

[24] G. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, "Epigraphical projection and proximal tools for solving constrained convex optimization problems," *Signal, Image Video Process.*, vol. 9, no. 8, pp. 1737–1749, 2015.

[25] S. Ono and I. Yamada, "Signal recovery with certain involved convex data-fidelity constraints," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6149–6163, Nov. 2015.

[26] S. Ono, "Primal-dual plug-and-play image restoration," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1108–1112, Aug. 2017.

[27] S. Ono, "Efficient constrained signal reconstruction by randomized epigraphical projection," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2019, pp. 4993–4997.

[28] K. Naganuma and S. Ono, "A general destriping framework for remote sensing images using flatness constraint," *IEEE Trans. Geosci. and Remote Sens.*, vol. 60, pp. 1–16, Feb. 2022, Art no. 5525016.

[29] S. Takemoto, K. Naganuma, and S. Ono, "Graph spatio-spectral total variation model for hyperspectral image denoising," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, Jul. 2022, Art no. 6012405.

[30] L. Condat, "Fast projection onto the simplex and the l1 ball," *Math. Program.*, vol. 158, no. 1, pp. 575–585, Jul. 2016.

[31] K. Naganuma and S. Ono, "Static-scene constrained optimization for matrix/tensor-decomposition-free foreground-background separation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.

[32] B. Hui, Z. Song, H. Fan, and et al., "A dataset for infrared image dim-small aircraft target detection and tracking under ground / air background," in *Sci. Data Bank*, 2020.

[33] R. Murphy, "T cell signaling proteins 3D confocal microscope movies," in *Carnegie Mellon University*, Nov. 2019, available: https://kilthub.cmu.edu/articles/dataset/T_cell_signaling_proteins_3D_confocal_microscope_movies/9963566 (accessed on Oct. 10, 2022.).

[34] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

JPEG Artifact Removal for Hyperspectral Images Based on Spatial-Spectral Regularization

Ryunosuke Eguchi*‡, Iori Kobayashi*, Shunsuke Ono†, and Ryo Matsuoka*‡
*The University of Kitakyushu, Fukuoka, Japan
†Tokyo Institute of Technology, Tokyo, Japan
E-mail: ‡r-matsuoka@kitakyu-u.ac.jp

***Abstract*—A hyperspectral image (HSI) contains vast data, storing hundreds of spectral information with higher resolution than an RGB image. Therefore, considering practical aspects such as data transmission and storage, it is essential to explore the application of standard JPEG compression to an HSI. Conventional JPEG artifact removal methods, such as block and mosquito noise, recover a JPEG image by improving spatial smoothness under the quantization constraint of the DCT coefficients. However, these methods do not consider the application of JPEG compression to HSIs and the low rank of the spectral characteristics of HSIs. This paper proposes a novel optimization method for removing JPEG artifacts from an HSI. Specifically, we formulate a convex optimization problem by introducing a regularization term based on the nuclear norm to the conventional optimization problems. This regularization promotes low rankness in the spectral domain. To solve the proposed problem, we utilize the alternating direction method of multipliers (ADMM). Experiments demonstrate the effectiveness of the proposed method compared to several existing methods applied to both remote and non-remote sensing HSIs.**

***Index Terms*—Hyperspectral image, JPEG compression, total variation, low rank.**

I. Introduction

A hyperspectral image (HSI) is three-dimensional data from tens to hundreds of two-dimensional band images, each taken from visible to near-infrared light using a narrow-band spectral filter, and has much higher wavelength resolution than that of an RGB image [1]. Since different materials generally have different reflectance spectra. The analysis and classification of HSI pixels have applications in earth observation, environmental protection, and disaster monitoring. However, because an HSI often contains a large amount of data, it is expensive to store and transmit, and its applications are limited.

JPEG2000 has been widely adopted to achieve high compression efficiency and quality preservation of HSIs [2]–[7]. JPEG2000 uses pre-processing methods such as spectral direction and spatial direction filtering, which are suited to the characteristics of HSIs, to extract and encode information effectively during compression. However, JPEG2000 may have limited hardware and software compatibility. On the other hand, JPEG, the standard image compression technique, is widely used for its convenience and ability to compress images while maintaining visual quality through simple algorithms efficiently. Applying JPEG to HSIs needs to be considered to extend the range of applications in remote sensing and industrial fields such as visual inspection. However, at higher compression ratios on JPEG, artifacts such as block and mosquito noise may occur, leading to a degradation in image quality.

JPEG artifact removal methods based on optimization have been proposed in [8]–[12]. Bredies and Holler introduced a JPEG decompression model based on a convex optimization problem [8]. The optimization problem consists of the total variation (TV) regularization, which promotes image smoothness, and the block-wise DCT quantization constraint. They solved the problem using a primal-dual algorithm. However, the resulting images often have an undesirable staircasing effect, a well-known TV regularization problem. Ono and Yamada have used the total generalized variation (TGV) regularization instead of the TV regularization [9] to reduce these artifacts and considered the JPEG decompression model with super-resolution interpolation. TGV considers higher-order derivatives and can reduce the staircasing effect while preserving image gradation. However, when these methods are applied directly to each band of the HSI, they cannot sufficiently recover the JPEG artifact-free HSI. This limitation arises because they focus only on promoting smoothness in the spatial domain and do not consider the low rankness of the HSI in the spectral domain.

There have been several attempts to handle the JPEG artifact removal problem by deep convolutional neural networks (CNN) [13]–[15]. Learning-based approaches are fragile to degradation not found in a training data set. When recovering HSIs using CNN-based methods, it is difficult and impractical to construct training data sets and networks. For example, since exposure time, capturing spectral bands, and object details vary greatly depending on the scenes, it is necessary to construct a plurality of networks suitable for each scene and appropriately replace them according to an input scene.

In this paper, we propose a JPEG artifact removal method for HSIs based on spatial-spectral regularization. By incorporating a regularization term with the nuclear norm, which promotes the low rankness of HSIs in the spectral domain, we extend the conventional optimization problem consisting of the TV-based spatial regularization and block-wise DCT quantization constraint to HSIs. The proposed problem is solved using the alternating direction method of multipliers (ADMM). In the experiments, we show the effectiveness of the proposed method compared with several existing methods.

The paper is organized as follows. In Section II, we present

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

several mathematical preliminaries to introduce the proposed method. Section III introduces a proposed JPEG artifact removal method for HSIs. In Section IV, several examples are shown to verify the validity of the proposed method. Finally, we conclude this paper in Section V.

II. Preliminaries

A. Total Variation

TV is defined as the sum of gradient intensities of an image, which represents the characteristic of the local smoothness of natural images. The TV regularization can promote image smoothness while preserving edges and is often used for image recovery problems such as noise/blur removal and missing pixel recovery.

By letting $\mathbf{D}_v$ and $\mathbf{D}_h \in \mathbb{R}^{N\times N}$ be the vertical and horizontal first-order differential operators, respectively, with Neumann boundaries, the differential operator is expressed by $\mathbf{D} := [\mathbf{D}_v^\top \ \mathbf{D}_h^\top]^\top \ \left(\in \mathbb{R}^{2N\times N}\right)$ for a vectorized gray image with N pixels, and then TV [16], [17] is defined as

$$\begin{aligned}\|\mathbf{x}\|_{\mathrm{TV}} := \|\mathbf{D}\mathbf{x}\|_{1,2}^{(2N)} &= \sum_{i=1}^{N}\sqrt{\sum_{k=0}^{1}(\mathbf{D}\mathbf{x})_{i+kN}^2},\\ &= \sum_{i=1}^{N}\sqrt{(\mathbf{D}_v\mathbf{x})_i^2+(\mathbf{D}_h\mathbf{x})_i^2},\end{aligned} \tag{1}$$

where $\|\mathbf{x}\|_{1,2}^{(KN)} := \sum_{i=1}^{N}\sqrt{\sum_{i=1}^{K-1}x_{i+kN}^2}$ is the $\ell_{1,2}$-norm, and $(\mathbf{D}_v\mathbf{x})_i$, $(\mathbf{D}_h\mathbf{x})_i$ are the i-th element of $\mathbf{D}_v\mathbf{x}$, $\mathbf{D}_h\mathbf{x}$, respectively.

B. Total Generalized Variation

TGV is an extension of TV that considers higher-order derivatives. It is a regularization technique widely used in image processing and computer vision.

TGV [18] in the case of second-order is given by

$$\|\mathbf{x}\|_{\mathrm{TGV}}^{(\alpha)} := \min_{\mathbf{D}\mathbf{x}=\mathbf{p}+\mathbf{q}}\left\{\alpha\|\mathbf{p}\|_{1,2}^{(2N)}+(1-\alpha)\|\mathbf{G}\mathbf{q}\|_{1,2}^{(3N)}\right\}, \tag{2}$$

where $\alpha > 0$ is the weight that controls the balance between the first-order and second-order derivative the total variation costs, and

$$\mathbf{G} := \begin{bmatrix}-\mathbf{D}_v^\top & \mathbf{O}\\ -\mathbf{D}_h^\top & -\mathbf{D}_v^\top\\ \mathbf{O} & -\mathbf{D}_h^\top\end{bmatrix} \in \mathbb{R}^{3N\times 2N}.$$

Minimizing TGV regularization with image fidelity can promote local flatness while preserving gradation, resulting in staircasing effect-free smoothing.

C. Alternating direction method of multipliers

ADMM [19] is an algorithm for solving convex optimization problems expressed as follows

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x})+g(\mathbf{z}) \quad \text{s.t. } \mathbf{z}=\mathbf{L}\mathbf{x} \tag{3}$$

where f is a differentiable convex function, g is a non-differentiable convex function, and $\mathbf{L}$ is a linear operator. To find the solution of the unknown $\mathbf{x}$ and the auxiliary variables $\mathbf{z}$, the ADMM algorithm is given for any $\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)}$ and $\rho > 0$ by

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}}\left\{f\left(\mathbf{x}^{(k)}\right)+\tfrac{\rho}{2}\left\|\mathbf{z}^{(k)}-\mathbf{L}\mathbf{x}^{(k)}-\mathbf{d}^{(k)}\right\|_2^2\right\}\\ \mathbf{z}^{(k+1)} &= \arg\min_{\mathbf{z}}\left\{g\left(\mathbf{z}^{(k)}\right)+\tfrac{\rho}{2}\left\|\mathbf{z}^{(k)}-\mathbf{L}\mathbf{x}^{(k+1)}-\mathbf{d}^{(k)}\right\|_2^2\right\}\\ \mathbf{d}^{(k+1)} &= \mathbf{d}^{(k)}+\mathbf{L}\mathbf{x}^{(k+1)}-\mathbf{z}^{(k+1)}\end{aligned} \tag{4}$$

where ρ is a parameter to adjust the weight of the penalty term, k is the number of iterations of the algorithm, and $\mathbf{d}$ is a dual variable. The algorithm is guaranteed to converge to a globally optimal solution after a sufficient number of iterations, regardless of the initial values of the objective function.

III. Proposed Method

A. Problem formulation

Let $\mathbf{u}_m \in \mathbb{R}^N (m = 1,\ldots,B)$ be a vectorized m-th band of an unknown HSI before JPEG compression, where B is the number of the spectral bands. Also let $\mathbf{\Phi}$ and $\mathbf{\Psi} \in \mathbb{R}^{N\times N}$ be the forward and inverse blockwise discrete cosine transform (DCT), respectively. We assume that JPEG compression is applied to each band independently for $m = 1,\ldots,B$ as

$$\mathbf{v}_m = \mathbf{\Psi}\mathcal{Q}(\mathbf{\Phi}\mathbf{u}_m), \tag{5}$$

where $\mathbf{v}_m \in \mathbb{R}^N$ is the m-th band of a given HSI, and $\mathcal{Q}(\cdot)$ is a procedure that performs quantization and dequantization[1].

B. Arranged spectral nuclear norm

To promote the low rankness of an HSI in the spectral domain, we define the arranged spectral nuclear norm (ASNN) for the vectorized HSI $\mathbf{u} := [\mathbf{u}_1^\top,\ldots,\mathbf{u}_B^\top]^\top \in \mathbb{R}^{NB}$ by

$$\|\mathbf{u}\|_{\mathrm{ASNN}} := \|T(\mathbf{u})\|_* = \|[\mathbf{u}_1,\ldots,\mathbf{u}_B]\|_*, \tag{6}$$

where $T(\cdot)$ is the operator for converting a vector into a matrix form, and $\|\cdot\|_*$ is the nuclear norm defined by the sum of all the singular values of $(\cdot)$. ASNN computes the nuclear norm of the horizontally concatenated matrix of vectors for each band of the HSI. Minimizing ASNN regularization can promote the low rankness of the HSI.

C. Optimization problem

Our goal is to recover a JPEG artifact-free HSI from a given JPEG compressed HSI. Conventional optimization problems do not consider applying to multi-/hyper-spectral images. Therefore, we focus on the low-rank spectral features of HSIs and add ASNN regularization into the existing optimization problem to remove JPEG artifacts in the HSI.

The proposed optimization problem is defined by

$$\begin{aligned}&\min_{\mathbf{u}}\|\mathbf{u}\|_{\mathrm{ASNN}}+\lambda\sum_{m=1}^{B}\mathcal{S}(\mathbf{u}_m)\\ &\text{s.t. } \mathbf{u}_m \in C_{[0,r_{\max}]}, \mathbf{u}_m \in C_{\mathbf{v}_m}^{\mathbf{Q}} \ \ (m=1,2,\ldots,B),\end{aligned} \tag{7}$$

[1]We consider a similar JPEG acquisition model formulated in [9], but we do not consider a downsampling operator and a linear blur operator that must be considered in super-resolution interpolation.

TABLE I: Details of the input HSIs.

	Size	Band	Bit
Pavia	256 × 256	102	13
PaviaU	256 × 256	103	13
Salinas	512 × 216	224	14
Indian pines	144 × 144	220	14
Books	536 × 640	65	16
Fruits	536 × 640	65	16
Red objects	536 × 640	65	16
Snacks	536 × 640	65	16

where $\mathcal{S}(\mathbf{u}_m)$ is a regularization function for the spatial smoothness of each band, and $\lambda > 0$ is a balancing weight that controls the smoothness and low rankness of the HSI. Our experiments considered TV and TGV for $\mathcal{S}$. The convex sets $C_{[0,r_{\max}]}$ and $C^{\mathbf{Q}}_{\mathbf{v}_m}$ are defined as

$$C_{[0,r_{\max}]} := \left\{\mathbf{x} \in \mathbb{R}^N \mid 0 \le x_n \le r_{\max}\ (n = 1, \ldots, N)\right\}, \tag{8}$$

$$C^{\mathbf{Q}}_{\mathbf{v}_m} := \left\{\mathbf{x} \in \mathbb{R}^M \mid \mathcal{R}\left(\boldsymbol{\Theta}^{(p)}_{\mathbf{x}}./\mathbf{Q}\right) = \mathcal{R}\left(\boldsymbol{\Theta}^{(p)}_{\mathbf{v}_m}./\mathbf{Q}\right) \left(p = 1, \ldots, M/8^2\right)\right\}. \tag{9}$$

The first constraint limits the pixel range of $\mathbf{u}_m$ in the $[0, r_{\max}]$, where $r_{\max}$ is set to $2^q - 1$ in the case of a q-bit HSI. The second constraint is the quantization constraint on block-wise DCT coefficients used in [8], [9][2]. This constraint makes the quantized DCT coefficients of the unknown image equal to those of the input JPEG image.

By solving (7), we obtain the JPEG artifact-free HSI. The proposed convex Optimization problem can be solved using the ADMM algorithm discussed in Section II-C.

IV. Experiments

To demonstrate the effectiveness of the proposed method, we applied it to remote and non-remote sensing HSIs. Table I shows the size, number of bands, and bits of HSIs. The first to fourth rows correspond to remote sensing examples, while the fifth to eighth rows correspond to non-remote sensing examples[3]. Quantitative evaluation was performed using PSNR and SSIM calculated from the original and recovered HSIs.

The results of our method were compared with those of conventional optimization methods, i.e., the TV- and TGV-based JPEG decompression methods[4]. We set TV and TGV to $\mathcal{S}(\mathbf{x})$ in the proposed optimization problem (7). We call the two approaches "Ours-TV" and "Ours-TGV". For the parameters of our methods, we found the visually best results by adjusting λ. Furthermore, the results of our method were presented in the case of $\lambda = 0$. We call this case "ASNN". That is, this method considers only the spectral low-rankness of the HSIs.

[2] $\Theta^{(p)}_{\mathbf{x}}, \Theta^{(p)}_{\mathbf{v}_m}$ denote the DCT coefficients of the p-th 8×8 block of unknown image $\mathbf{x}$ and input image $\mathbf{v}_m$, respectively. $\mathbf{Q}$ denotes the 8×8 quantization matrix used for quantization of DCT coefficients. $\mathcal{R}(\cdot)$ denotes the rounding process to integer values [9].

[3] We used the remote sensing images provided by https://www.ehu.eus/ccwintco. The non-remote sensing images were taken in our laboratory with HS cameras.

[4] These conventional methods are equivalent to removing ASNN and setting TV and TGV to $\mathcal{S}$, respectively, as $\mathcal{S}(\mathbf{x}) = \|\mathbf{x}\|_{\mathrm{TV}}$ or $\|\mathbf{x}\|_{\mathrm{TGV}}$.

TABLE II: PSNR [dB] comparison results.

	JPEG	TV	TGV	ASNN	Ours-TV	Ours-TGV
Pavia	31.86	30.58	30.55	32.44	**33.13**	32.40
PaviaU	31.91	31.15	31.08	33.06	**33.17**	33.05
Salinas	39.96	39.60	39.50	42.34	**42.58**	42.43
Indian pines	36.48	35.82	35.80	37.71	**37.86**	37.80
Books	31.95	32.17	32.09	33.23	**33.47**	33.14
Fruits	35.81	35.47	35.39	36.73	**36.93**	36.81
Red objects	36.01	35.53	35.51	37.08	**37.26**	37.15
Snacks	32.98	32.52	32.50	34.25	**34.36**	34.15

TABLE III: SSIM comparison results.

	JPEG	TV	TGV	ASNN	Ours-TV	Ours-TGV
Pavia	0.8637	0.8345	0.8339	0.8978	**0.8983**	0.8962
PaviaU	0.8667	0.8452	0.8400	0.8986	**0.9006**	0.8988
Salinas	0.9578	0.9606	0.9603	0.9814	**0.9823**	0.9819
Indian pines	0.9426	0.9396	0.9373	0.9599	**0.9619**	0.9613
Books	0.9560	0.9650	0.9628	0.9675	**0.9708**	0.9699
Fruits	0.9663	0.9697	0.967	0.9756	**0.9774**	0.9769
Red objects	0.9660	0.9693	0.9673	0.9750	**0.9770**	0.9769
Snacks	0.9737	0.9768	0.9759	0.9810	**0.9824**	0.9821

Tables II and III show the average of PSNR and SSIM of all bands of the recovered HSIs obtained by the proposed and the conventional methods, respectively. One can observe from both tables that Ours-TV has the highest values compared with the conventional methods.

Figures 1 and 2 show the original and input HSIs, and the results of JPEG artifact removal. One can see from Figures 1 and 2 that the proposed methods outperform the conventional methods. The results of the TV- and the TGV-based methods can be removed the block noise. However, details of the HSIs are over smoothed; complex textures are lost, especially in Figures 1(d) and (e). Although ASNN can reduce JPEG artifacts without spatial regularization from Figures 1 and 2(f), adding TV and TGV as regularization in addition to ASNN can reduce block noise more effectively (see Figures 1 and 2(g) and (h)).

V. Conclusion

We proposed the JPEG artifact removal method for HSIs. The proposed optimization problem was defined by spatial-spectral regularization with pixel range constraint and block-wise DCT quantization constraints. The proposed problem can be solved by the ADMM. Experiments confirmed that the proposed methods with TV and TGV outperformed existing TV- and TGV-based JPEG artifact removal methods for several remote and non-remote sensing HSIs.

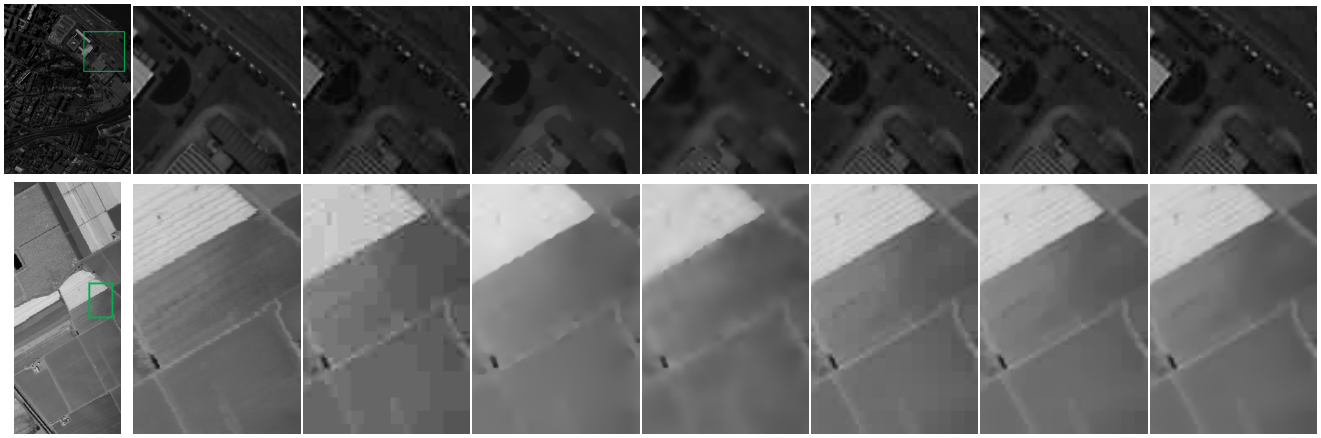
In future work, we will extend the proposed method to recover the artifacts of JPEG2000 compression, commonly used for HSI compression.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 21K17767.

References

[1] R. Pizzolante and B. Carpentieri, "Visualization, band ordering and compression of hyperspectral images," *Algorithms*, vol. 5, no. 1, pp. 76–97, 2012.



(a) Original (b) Enlarged (c) Input (d) TV (e) TGV (f) ASNN (g) Ours-TV (h) Ours-TGV

Fig. 1: Results for remote sensing: (upper) Pavia and (lower) Salinas.



(a) Original (b) Enlarged (c) Input (d) TV (e) TGV (f) ASNN (g) Ours-TV (h) Ours-TGV

Fig. 2: Results for non-remote sensing: (upper) Books and (lower) Red objects.

[2] H. S. Lee, N. H. Younan, and R. L. King, "Hyperspectral image cube compression combining jpeg-2000 and spectral decorrelation," in *IEEE international geoscience and remote sensing symposium*, vol. 6. IEEE, 2002, pp. 3317–3319.

[3] X. Tang, W. A. Pearlman, and J. W. Modestino, "Hyperspectral image compression using three-dimensional wavelet coding," in *Image and Video Communications and Processing 2003*, vol. 5022. SPIE, 2003, pp. 1037–1047.

[4] Q. Du and J. E. Fowler, "Hyperspectral image compression using jpeg2000 and principal component analysis," *IEEE Geoscience and Remote sensing letters*, vol. 4, no. 2, pp. 201–205, 2007.

[5] Q. Du and J. E. Fowler, "Low-complexity principal component analysis for hyperspectral image compression," *The International Journal of High Performance Computing Applications*, vol. 22, no. 4, pp. 438–448, 2008.

[6] D. Báscones, C. González, and D. Mozos, "Hyperspectral image compression using vector quantization, pca and jpeg2000," *Remote sensing*, vol. 10, no. 6, p. 907, 2018.

[7] R. Nagendran and A. Vasuki, "Hyperspectral image compression using hybrid transform with different wavelet-based transform coding," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 18, no. 01, p. 1941008, 2020.

[8] K. Bredies and M. Holler, "A total variation–based jpeg decompression model," *SIAM Journal on Imaging Sciences*, vol. 5, no. 1, pp. 366–393, 2012.

[9] S. Ono and I. Yamada, "Optimized jpeg image decompression with super-resolution interpolation using multi-order total variation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2013, pp. 474–478.

[10] I. Kobayashi and R. Matsuoka, "A study on jpeg artifact removal by four-directional difference-based $L_{0,1}$ norm regularization," in *2022 IEEE 11th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2022, pp. 724–725.

[11] X. Liu, G. Cheung, X. Wu, and D. Zhao, "Random walk graph laplacian-based smoothness prior for soft decoding of jpeg images," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 509–524, 2016.

[12] F. Alter, S. Durand, and J. Froment, "Adapted total variation for artifact free decompression of jpeg images," *Journal of mathematical Imaging and Vision*, vol. 23, no. 2, pp. 199–211, 2005.

[13] L. Cavigelli, P. Hager, and L. Benini, "Cas-cnn: A deep convolutional neural network for image compression artifact suppression," in *Proc. Int. Joint Conf. Neural Net. (IJCNN)*. IEEE, 2017, pp. 752–759.

[14] X. Zhang, W. Yang, Y. Hu, and J. Liu, "Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2018, pp. 390–394.

[15] J. Li, Y. Wang, H. Xie, and K.-K. Ma, "Learning a single model with a wide range of quality factors for jpeg image artifacts removal," *IEEE Trans. Image Process.*, vol. 29, pp. 8842–8854, 2020.

[16] X. Bresson and T. F. Chan, "Fast dual minimization of the vectorial total variation norm and applications to color image processing," *Inverse Probl. Imag.*, vol. 2, no. 4, pp. 455–484, 2008.

[17] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Probl.*, vol. 24, no. 6, p. 065014, 2008.

[18] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 492–526, 2010.

[19] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & mathematics with applications*, vol. 2, no. 1, pp. 17–40, 1976.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Data Driven Multiband Image Fusion That Preserves Wavelength-Specific Image Features

Hsuan Lin* and Keigo Hirakawa†
* University of Dayton, United States
E-mail: hsuanlin34@gmail.com
† University of Dayton, United States
E-mail: khirakawa1@udayton.edu

***Abstract*—In this paper, we propose a data-driven multiband image fusion approach called FusionUNet aimed at preserving image features that are specific to the wavelengths of the source images. Specifically, we extract low-level wavelength-specific image features from the source images and use them to guide the training process of the U-Net structure designed to combine source images of various wavelengths. We conducted extensive experiments on public datasets, and the results demonstrate that our algorithm outperforms the state-of-the-art approaches in both qualitative and quantitative aspects.**

***Index Terms*—Image fusion, convolution neural network, infrared image, visible image.**

I. Introduction

Image fusion refers to the task of combining images captured by multiple sensors, which typically measure different wavelength ranges. Compared to an image obtained by a single sensor, a sensor-fused image can reveal more salient details about objects represented in the scene using different wavelengths. For example, longwave infrared (LWIR) sensors capture the thermal radiation emitted from objects directly, which is useful for detecting humans and animals. However, LWIR sensors, such as microbolometers, often have low resolution, and LWIR images lack textural details.

In contrast, semiconductor-based detectors like CCD and CMOS imagers sense reflected visible wavelength (VIS) light from objects. These imagers are a mature technology that allows for densely packed pixels, resulting in high-resolution images with well-controlled pixel uniformity. However, visible wavelength imagery is susceptible to environmental factors such as sunlight, illumination intensity, and smoke/haze, which can negatively impact target detection and tracking by both human observers and computer vision algorithms.

An effective image fusion technique combines the complementary advantages of the source images collected at different wavelengths. For instance, LWIR-VIS image fusion should be robust to environmental factors, enhance the visibility of humans and animals, and preserve high-resolution details. In this paper, we propose a novel framework for a data-driven image fusion neural network called FusionUNet aimed at maintaining the visual cues of wavelength-specific features of VIS and LWIR images. Leveraging a neural network pretrained for general-purpose object recognition classifier (such as VGG16 [1]) as a proxy to a human observer, we posit that a *fusion of the low-level feature maps* extracted from source LWIR and VIS images represent the salient spatial or geometric structures that we want the fused LWIR-VIS image to represent. Under this conjecture, low-level feature maps extracted from a successfully fused LWIR-VIS image will match the low-level wavelength-specific features extracted from source LWIR and VIS images.

II. Background and Related Work

A. Image Fusion Metrics

Denote by $\{I_1, \ldots, I_K\}$ the K source images (possibly from different sensors and wavelength ranges), and $I_F = \mathrm{ImageFusion}\{I_1, \ldots, I_K\}$ represent the output from the image fusion tasks. The loss function is the key part of deep learning-based and non-learning-based image fusion algorithms alike, determining how to retain the source image information in the output fused image.

Two categories of loss/metrics have been used today. First, *statistical attributes* are used to measure the amount of information/context present in the fused images. For instance, entropy is employed to quantify the level of randomness [2], while standard deviation measures the degree of spread in the fused image [3], [4].

The second type of image fusion metric is designed to compare the fused image with each of the source images. For instance, metrics such as mean squared error (MSE), structural similarity index measure (SSIM) [5], and mutual information (MI) [3] are computed between the source image I_k and the fused image I_F to assess the similarity of pixel intensity patterns. Consider the sum of loss metrics (such as mean squared error, mutual information, structural similarity index) comparing the fused image to each of the source images:

$$L = Loss(I_F, I_1) + Loss(I_F, I_2) + \cdots + Loss(I_F, I_K). \tag{1}$$

The sum of losses is sensitive to the degradation of source image details present in the fused image. Considering the fact that sensitivity to target objects varies with the measured wavelengths, it is important to note that the behavior becomes unpredictable when the source images are dissimilar to each other. In other words, there is no fused image I_F that can simultaneously optimize the losses

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

$Loss(I_F, I_1), \ldots, Loss(I_F, I_K)$ in order to optimize the overall loss score L.

B. Image Fusion Approaches

Traditional methods for fusing multiple source images include multiscale transforms [6], [7], sparse dictionaries [8], saliency map-based techniques [9], and others [10].

Multiscale approaches involve decomposing source images into base and detail components, fusing each component separately using rule-based heuristics, and then combining these representations to produce a fused image [11]. Sparse representation methods employ a sliding window strategy to decompose an input image into overlapping patches, using an over-complete dictionary to achieve sparse representation coefficients. These coefficients are fused to reconstruct the final image. Some representations leverage overcompleteness and sparsity to enhance robustness against misregistration and noise [12]. In saliency map-based methods, rule-based heuristics for combining representations are manually crafted, often relying on the principles of saliency maps to highlight important objects for the human visual system or on the sparsity of real-world scene representations [9]. One challenge in designing rule-based fusion methods is addressing the imbalance between source images. For instance, the lack of detail in the visible source image can diminish the distinctive features of warm-body objects present in the LWIR source image [13].

Drawing on the recent advancements in machine learning, image fusion techniques in recent years have benefited from data-driven deep learning-based approaches [14], [15]. GAN-based approaches utilize an adversarial game between the fused image and the source images to retain the characteristic features of different sensors/wavelength modalities [16]–[18]. Autoencoder-based and regression-based methods, on the other hand, learn a mapping to extract salient features and then reconstruct the fused image by minimizing loss functions [19]–[22].

The lack of a "ground truth" fused image makes it challenging to learn the weights and biases of neural networks in a supervised manner. As an alternative, the method proposed in [20] adopts a semi/unsupervised approach by utilizing a pre-trained VGG19 model to extract deep features, which are then used to determine the fusion weights. Similarly, in the field of remote sensing, the method described in [23] employs a U-Net to extract and combine the spectral features from hyperspectral images and spatial representation from multispectral images to yield a sharp/high-resolution hyperspectral images.

III. Proposed Method

A. New Loss Function

The loss function plays a critical role in deep learning-based image fusion algorithms as it determines the information preserved in the fused image. Because there exists no notion of ground truth in image fusion, we characterize some desirable attributes of the fused images below.

We would like objects-of-interests detected from various wavelength source images to also be present in the combined fusion image $I_F = \mathrm{ImageFusion}(I_1, \ldots, I_K)$. However, we do not want to rely on the state-of-the-art data-driven object detection algorithms, such as the *supervised* machine learning methods that rely on a very large dataset with substantial "labeling costs." These labels are object-specific, and they do not generalize across applications.

In this paper, we develop a novel loss function aimed at preserving wavelength-specific features. Let $F_k = \mathrm{Feature}(I_k)$ denote shallow level VGG16 features extracted from kth source image I_k. We introduce a notion of "feature fusion" as follows:

$$\begin{aligned}\mathrm{FeatureFusion}\{I_1, \ldots, I_K\} &= \mathrm{softmax}\{F_1, \ldots, F_k\} \\ &= \frac{\sum_{k=1}^{K} F_k e^{\frac{F_k}{\alpha}}}{\sum_{k=1}^{K} e^{\frac{F_k}{\alpha}}},\end{aligned} \tag{2}$$

where the *softness parameter* α controls the degree to which the small positive-valued feature is preserved. The idea behind $\mathrm{FeatureFusion}\{\cdot\}$ is that if F_k represents features extracted immediately after ReLU layer, then the feature magnitude represents the significance of those feature map coefficients (since small values would have had a high risk of being thresholded by ReLU operator). Since typical feature maps are sparse, the softmax is approximately the union of the wavelength-specific features of the source images. In the case that the supports of feature maps overlap, (2) prioritizes the larger feature values likely indicating visually significant intensity patterns present in each of the images $\{I_1, \ldots, I_K\}$.

An example of the feature fusion is shown in Fig. 1. As evidenced in Fig. 1(e), the merged feature $\mathrm{FeatureFusion}(F_1, F_2)$ successfully represents salient details represented in both LWIR (I_1) and VIS (I_2) source images. For implementation, we used VGG16 in its $M = 5$th layer, which has the size $224 \times 224 \times 64$ and follows the convolutional layer, ReLU layer, convolutional layer, and ReLU layer.

Thus $\mathrm{FeatureFusion}\{I_1, \ldots, I_K\}$ represents the wavelength-specific attributes of the image we would like the fused image $I_F = \mathrm{ImageFusion}\{I_1, \ldots, I_K\}$ to inherit. For this reason, we define the "feature loss" function of fused image I_F as follows:

$$Loss(I_F)\{I_1, \ldots, I_K\} = \|F_F - \mathrm{FeatureFusion}\{I_1, \ldots, I_K\}\|^2, \tag{3}$$

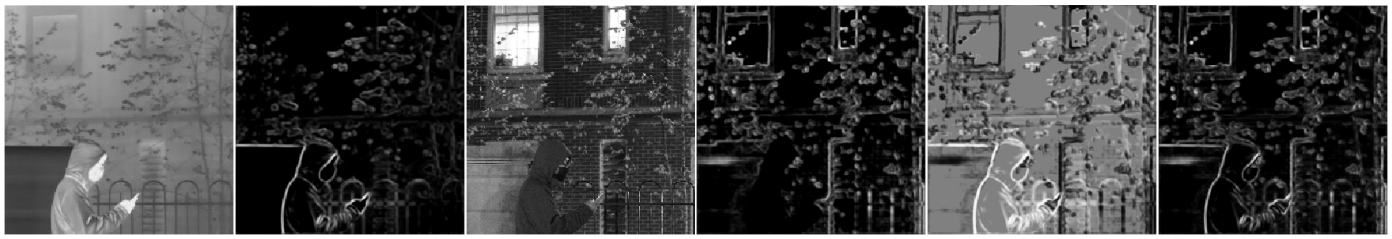
where F_F are the feature map vectors extracted from fused image I_F. For instance, if image fusion is a parametric function:

$$I_F = \mathrm{ImageFusion}_\theta(I_1, \ldots, I_K), \tag{4}$$

then we perform the following optimization task

$$\begin{aligned}\min_\theta \| &\,\mathrm{Feature}(\mathrm{ImageFusion}_\theta(I_1, \ldots, I_K)) \\ &- \mathrm{FeatureFusion}\{I_1, \ldots, I_K\}\|^2.\end{aligned} \tag{5}$$

When this loss function is minimized successfully, we assume that the fused image (4) preserved the wavelength-specific



(a) LWIR Image (b) LWIR Feature (c) VIS Image (d) VIS Feature (e) Softmax Weight (f) Fused Feature

Fig. 1: Infrared- and visible-specific feature maps merged by softmax function. (b)×(e) + (d)×(1-(e)) = (f).

attributes of the source images $\{I_1, \ldots, I_K\}$. It is important to emphasize that the feature extraction layers used in $\mathrm{Feature}(\cdot)$ and fusion feature map $\mathrm{FeatureFusion}(\cdot)$ is fixed feature extraction layers, and only parameter θ of the image fusion technique $\mathrm{ImageFusion}_\theta(\cdot)$ is optimized.

B. Network Architecture and Training Details

For the implementation of $\mathrm{ImageFusion}_\theta(\cdot)$ in (5), we leverage a regression architecture called U-Net, which is a variation on an auto-encoder network where the skip connections retain fine-level localized details while the deeper layers with contracting paths process coarse-level details [24]. U-Net can be trained with limited data and achieve excellent performance, and it has been proven useful in image segmentation[25], denoising [26] as well as in image fusion [14].

The exact configuration of the proposed FusionUNet is shown in Fig. 3, designed to combine two source images in LWIR and VIS. The input is LWIR and VIS source images that are first standardized and then concatenated. The network consists of an encoder (left side) responsible for feature extraction and a decoder (right side) that performs reconstruction. The encoder uses the classic architecture of a convolutional network—repeated two 3×3 convolutions (padding option = same, padding value =0) followed by ReLU. Each downsampling layer is a convolutional layer (Stride = 2×2) rather than the max pooling. We performed five levels of downsampling steps to ensure that the network extracts deep features, and the number of feature channels at each average pool operation was doubled. In the decoder, use the same repeated two 3×3 convolutions followed by ReLU, and the transposed convolutions for upsampling step and reduce half feature channels. The skip connection feature channels were concatenated from the corresponding U-Net encoder level. The final layer is 1×1 convolution, which outputs the final fused image whose size is the same as the input image. The output of FusionUNet is the fused image I_F.

The training process consists of two stages. In the first stage, we pre-train the U-Net to generate the average of the LWIR and VIS source images. This pre-training stage follows a standard U-Net-based regression approach (learning rate=10-5, epoch=30). In the second training stage, we refine the FusionUNet by treating the fused image I_F as the input of pre-trained VGG16 (i.e. fixed) as shown by Fig. 2. We extract the fifth layer of the pre-trained VGG16 network to compute feature loss in (5), which is backpropagated through the U-Net to update the network coefficients (learning rate=10-6 and epoch=15). We empirically found that the pre-training makes the overall process more stable and converges faster. The second training stage improves the image fusion quality by highlighting salient thermal targets and enhancing fine textures.

IV. Experimental Results

A. Experiment Configuration

We evaluated the efficacy of the proposed FusionUNet using the standard metrics described in Section II-A, including entropy, standard deviation, structural similarity index, and mutual information [27]. In addition, we report the feature fusion loss metric in (3) at the fifth layer of VGG16, as proposed in Section III-A. Even though it was designed for training FusionUNet, it is an excellent metric for assessing the low-level edges, shapes, and textures at test time. We standardize the input to VGG16 by subtracting out the ensemble pixel mean of the image.

For benchmarking purposes, the proposed FusionUNet was compared to seven state-of-the-art image fusion methods: IFE-VIP [10], MST_SR [28], TIF [29], VSMWLS [30], LatLRR [20], MDLatLRR [13], and FusionGAN [18]. All the experiments are conducted on a Windows computer with Intel(R) Core(TM) i7-10700 CPU at 2.9 GHz, 64 GB memory, and NVIDIA GeForce GTX 1660Ti.

B. Quantitative Evaluation

The performance metrics of the proposed FusionUNet are reported in Table I. Although no one state-of-the-art image fusion algorithm dominates over the rest, the proposed FusionUNet is the only method to consistently rank within the top three in all metrics (feature loss, entropy, standard deviation, SSIM, and mutual information). We attribute this gain to the fact that the feature loss promotes high fidelity by ensuring consistency between the fused image and the source images, which in turn encourages better SSIM and mutual information scores. Furthermore, the softmax operation in feature fusion leads to a higher degree of information and context in the fused images, as it discards smaller feature map coefficients, resulting in improved entropy and standard deviation scores. Unsurprisingly, FusionUNet yielded the best score for VGG16 feature loss, as it is optimized for this metric.

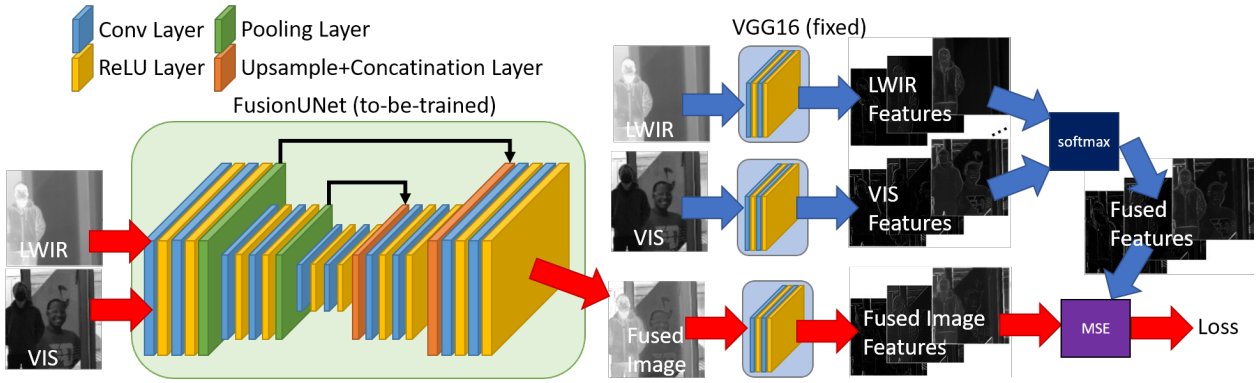


Fig. 2: The architecture of training of proposed method

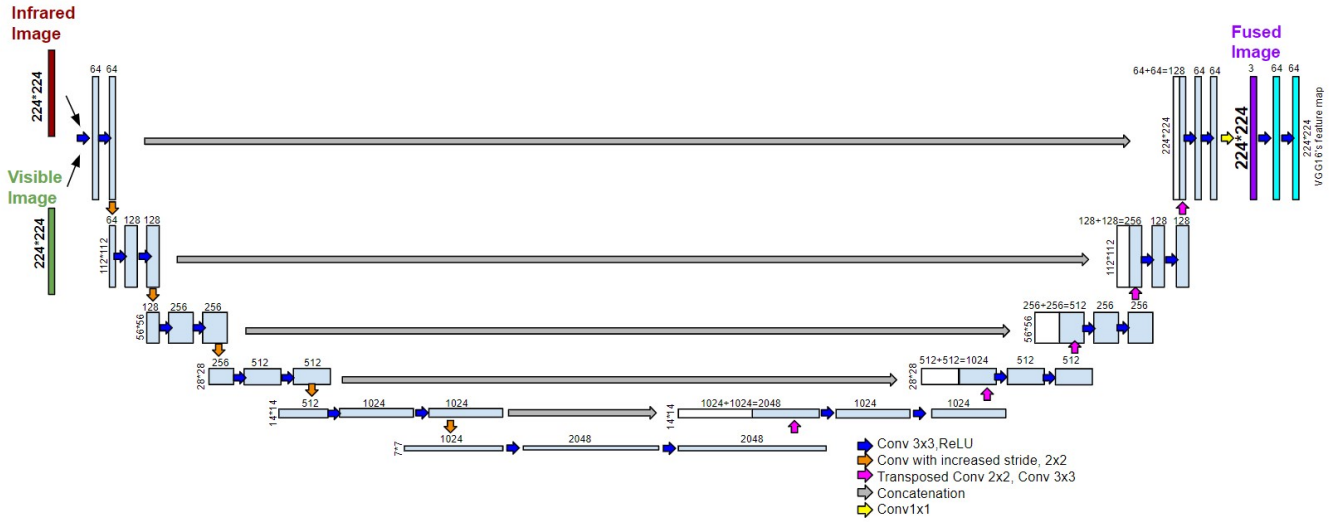


Fig. 3: The architecture of FusionUNet

C. Qualitative Evaluation

Fig. 4 provides visual comparisons of FusionUNet with various state-of-the-art LWIR-VIS fusion methods. For this qualitative evaluation, we selected four scenes from TNO dataset. In the first column of Fig. 4, it can be observed that MST_SR and MDLatLRR methods weaken the thermal salient target. Additionally, FusionGAN's result makes it difficult to identify the VIS environmental details. However, VSMWLS and FusionUNet effectively preserve both LWIR and VIS important objects in this scene. Moving to the second column, it can be seen that IFEVIP and LatLRR methods change the brightness during the fusion process. In contrast, FusionUNet maintains the source information properly and produces a fused image with sharp thermal targets and richer background information. In the third column, MST_SR, VSMWLS, and the proposed FusionUNet interestingly superimpose facial details captured in the VIS source image onto the thermally enhanced LWIR face. Lastly, in the final column examples, FusionUNet outperforms all other state-of-the-art methods in preserving environmental details such as branches and buildings, while still clearly displaying the thermal targets. Additionally, FusionUNet exhibits the least amount of ghosting/halo effect near high contrast LWIR-VIS edges overall.

In terms of appearance, the proposed FusionUNet was most similar to IFEVIP [10]. IFEVIP extracts infrared features through a process involving background subtraction via quadtree decomposition and interpolation, which is then combined with visible information. However, IFEVIP is tailored to low-light conditions, leading to extracting excessive infrared background, degrading the contour information in the fused image. This is evidenced in the second column of Fig. 4, where details of a street lamp are compromised, for example. By contrast, the proposed FusionUNet retains high fidelity on critical spatial/shape details present in both LWIR and VIS images.

V. Conclusion

In this paper, we propose a powerful image fusion algorithm designed to preserve the wavelength-specific features of the source images. Specifically, we apply softmax to approximate the combination of shallow-layer pre-trained VGG16 feature



Fig. 4: Qualitative results of the image fusion on TNO dataset

TABLE I
AVERAGE QUANTITATIVE EVALUATION ON 13 IMAGE PAIRS FROM TNO DATASET

Method	Feature Loss(10^4)	Entropy	Standard Deviation	Structural Similarity Index	Mutual Information
IFEVIP [10]	3.13	6.62	38.78	0.4966	2.20
MST_SR [28]	3.90	6.63	30.20	0.4840	1.07
TIF [29]	4.05	6.73	29.52	0.4804	1.06
VSMWLS [30]	3.45	6.85	34.32	0.5100	1.27
LatLRR [20]	4.30	6.71	42.19	0.4215	1.14
MDLatLRR [13]	4.53	6.29	24.22	0.5120	1.25
FusionGAN [18]	4.99	6.43	26.85	0.3262	1.46
FusionUNet (Ours)	2.92	6.91	41.39	0.5174	1.59

maps extracted from the source images. We then learn the mapping from the source images to a fused image that preserves the fused features using U-Net. Our experimental results demonstrate that our algorithm successfully fuses visible wavelength and longwave infrared images, resulting in a fused image that is sensitive to humans/animals, robust to environmental factors, and preserves high-resolution image features. In comparison to other state-of-the-art algorithms, our fusion results yield higher contrast images with finer spatial details.

The proposed FusionUNet is able to preserve wavelength-specific image features of the source images by extracting and combining feature maps extracted from general-purpose object recognition classifiers. Because of this, FusionUNet has the potential to improve the object detection performance in computer vision applications by ensuring that the targets visible in LWIR and VIS images are detected in the fused image also. The reference code is available for download at `issl.udayton.edu`.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, p. 023 522, 2008.

[3] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153–178, 2019.

[4] Y.-J. Rao, "In-fibre bragg grating sensors," *Measurement science and technology*, vol. 8, no. 4, p. 355, 1997.

[5] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[6] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Information Sciences*, vol. 508, pp. 64–78, 2020.

[7] H. Yan and Z. Li, "Infrared and visual image fusion based on multi-scale feature decomposition," *Optik*, vol. 203, p. 163 900, 2020.

[8] S. Zhang, F. Huang, B. Liu, *et al.*, "A multi-modal image fusion framework based on guided filter and sparse representation," *Optics and Lasers in Engineering*, vol. 137, p. 106 354, 2021.

[9] S. Minghui, L. Lu, P. Yuanxi, J. Tian, and L. Jun, "Infrared & visible images fusion based on redundant directional lifting-based wavelet and saliency detection," *Infrared Physics & Technology*, vol. 101, pp. 45–55, 2019.

[10] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infrared Physics & Technology*, vol. 83, pp. 227–237, 2017.

[11] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 203–209, 2015.

[12] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Information Fusion*, vol. 40, pp. 57–75, 2018.

[13] H. Li, X.-J. Wu, and J. Kittler, "Mdlatlrr: A novel decomposition method for infrared and visible image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4733–4746, 2020.

[14] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "Didfuse: Deep image decomposition for infrared and visible image fusion," *arXiv preprint arXiv:2003.09210*, 2020.

[15] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Information Fusion*, vol. 42, pp. 158–173, 2018.

[16] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

[17] J. Hou, D. Zhang, W. Wu, J. Ma, and H. Zhou, "A generative adversarial network for infrared and visible image fusion based on semantic segmentation," *Entropy*, vol. 23, no. 3, p. 376, 2021.

[18] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "Fusiongan: A generative adversarial network for infrared and visible image fusion," *Information fusion*, vol. 48, pp. 11–26, 2019.

[19] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[20] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *2018 24th international conference on pattern recognition (ICPR)*, IEEE, 2018, pp. 2705–2710.

[21] F. C. Ataman and G. B. Akar, "Visible and infrared image fusion using encoder-decoder network," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 1779–1783.

[22] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "Stdfusionnet: An infrared and visible image fusion network based on salient target detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[23] X. Wang, X. Wang, K. Zhao, X. Zhao, and C. Song, "Fsl-unet: Full-scale linked unet with spatial–spectral joint perceptual attention for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[25] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82 031–82 057, 2021.

[26] R. W. Baldwin, R. Liu, M. Almatrafi, V. Asari, and K. Hirakawa, "Time-ordered recent event (tore) volumes for event cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2519–2532, 2022.

[27] X. Zhang, P. Ye, and G. Xiao, "Vifb: A visible and infrared image fusion benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 104–105.

[28] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information fusion*, vol. 24, pp. 147–164, 2015.

[29] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[30] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

Shot-Noise-Aware Image Signal Restoration for Photoelectronic Charge-Based Sensors

Seishi Takamura
Hosei University, Japan
E-mail: s.takamura@hosei.ac.jp Tel: +81-42-387-4534

***Abstract* — It is widely presumed that digital image sensors can't obtain the brightness of too-bright image information that yields a saturated digital value, e.g., 1023 for 10-bit sensors and 4095 for 12-bit sensors, without the help of brightness reduction such as shortening the shutter duration, applying the optical filter. However, suppose it is allowed to capture the still object multiple times. In that case, we have previously shown that the inevitable existence of optical shot noise enables us to estimate a brightness value greater than seemingly-saturated brightness. That study was purely theoretical, i.e., it was conjectured by computer simulation, not with actual camera-captured data. In this paper, we formulate a precise photoelectronic charge-based sensing model, and based on that, we try to verify the above theoretical conjecture with actual camera-captured data. The experimental results successfully prove that restoring a brightness value greater than the digital count limit is possible.**

I. Introduction

The performance of image sensors has been increasing in recent years due to technological innovations. In CMOS image sensors, the reduction of dark current noise and readout noise, two of the three major noise factors in image sensors, has been affirmatively resolved [1], leaving only optical shot noise as the remaining noise factor. In general, noise harms image/video coding, image processing, video analysis, scene recognition, etc. However, the existence of optical shot noise is not always a nuisance. In [2], we have demonstrated that the presence of noise can reduce the quantization (digitalization) noise more, after averaging multiple observations, than the noiseless situation. We also have numerically shown that precise modeling of optical shot noise allows us to raise the upper limit of digital value (e.g., 255 for 8-bit sensors, 1023 for 10-bit sensors). The mathematical model used to formulate image sensing was effective but relatively simple; e.g., it was assumed that a certain integer number (no fractional number) of photoelectrons increase the digital count by one with no offset). And the study was purely theoretical. Therefore, simulation-based results were demonstrated. In [3], the variance and mean values of digital signals are measured with a real camera, and their relation is formulated considering dark current noise, readout noise, and optical shot noise. However, no connection with the number of photoelectrons nor A/D converter (ADC) properties was considered. In [4], a naïve model between the variance and average values of digital signals for a real camera is reported. However, the empirical result does not fit well in very dark and bright pixels.

This paper will formulate a more precise photoelectronic charge-based image sensing model than the former work [2-4]. Then, we experiment using an actual camera to capture the natural scene and verify the above theoretical conjecture.

II. Photoelectronic Charge-Based Sensing Model and Optical Shot Noise

Fig. 1 shows the typical image sensor model based on photon-to-electron conversion and charging. The photons pour down to the sensor of one pixel. The sensor (photodiode) absorbs the photons and emits electrons (photoelectrons). The ratio of incoming photons and outgoing electrons is called quantum efficiency. The photoelectrons (let the number be X_e) are stored in a well and read out by the ADC to yield an observed digital count (X_d) every frame period. Suppose we

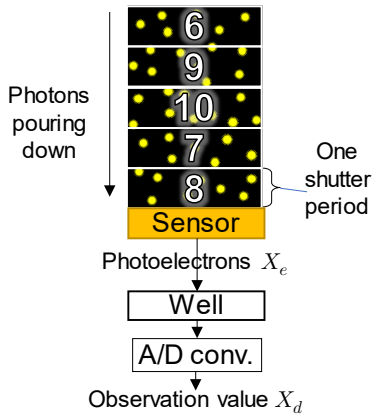


Fig. 1 Photoelectronic charge-based image sensing.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

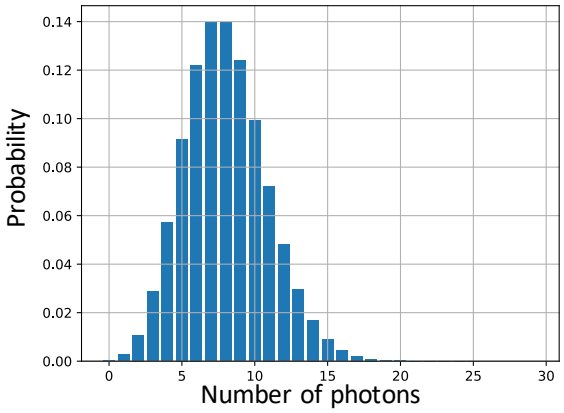


Fig. 2 Poisson distribution with mean 8.

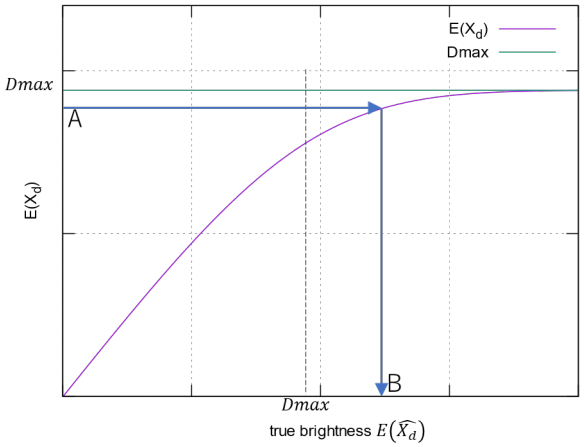


Fig. 3 True brightness vs. $E(X_d)$ and estimating process of true brightness (B) from averaged digital count (A).

have an utterly stationary light source whose average number of photons per frame is eight (8), measuring the photons flying from the source. Since photons come independently of each other, the number of photons per frame varies 8, 7, 10, 9, or 6, depending on the shooting time. This variation, called "optical shot noise, " is an unavoidable and dominant noise in digital cameras and strictly follows the Poisson distribution. Fig. 2 shows the probability mass function of the Poisson distribution with a mean of 8. Sometimes the number is zero, and sometimes the number is 20, 30, or even more. Because the number of photons obeys the Poisson distribution, the number of photoelectrons also obeys the Poisson distribution. One of the essential properties of Poisson distribution is that its mean and variance are the same. Therefore, the brighter the object is, the more noise it contains. Equation (1) shows the formulation of Poisson distribution, where X is a stochastic variable following Poisson distribution of the mean (and the variance) λ.

$$Po(\lambda, X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{1}$$

III. Precise Sensing Model

A. ADC Model and Parameter Formulation

In our preliminary experiment, it was observed that the digital count from the camera and the variance of the digital count have a linear relationship with bias. This section will formulate a precise digital count and photoelectron count model for the ADC inside the image sensor. In the following, we will assume ADC gain to be fixed to 0 dB (i.e., no gain), no tone curve, no gamma correction, and no automatic gain control.

Let an integer X_e denote the stochastic variable for the number of photoelectrons in a well to be converted to the digital count by the ADC, and let an integer X_d denote the stochastic variable associated to the digital count. The ADC property can be formulated as

$$X_d = Digital\left(\frac{X_e - b}{a}\right), \tag{2}$$

where

$$Digital(x) = \min(\lfloor x + 0.5 \rfloor, Dmax)$$

is the rounding and clipping function, a represents the average number of photoelectrons that increases the digital count by one, b represents the offset of the ADC, $Dmax$ is the maximum value of digital output. From (2),

$$E(X_d) = \frac{E(X_e) - b}{a} \tag{3}$$

$$V(X_d) = \frac{V(X_e)}{a^2} + \frac{1}{12} \tag{4}$$

are derived, where $E(\cdot)$ and $V(\cdot)$ represent the mean and variance of the given stochastic variable, respectively. Here we ignored clipping effect of $Digital(\cdot)$, but took its rounding effect into account (i.e., adding a variance of 1/12). We also know that

$$E(X_e) = V(X_e) \tag{5}$$

because X_e follows the Poisson distribution. So, we have

$$a^2\left(V(X_d) - \frac{1}{12}\right) = aE(X_d) + b \tag{6}$$

and thus, it is derived that

$$V(X_d) = \frac{E(X_d)}{a} + \frac{b}{a^2} + \frac{1}{12}. \quad (7)$$

We will confirm those relations in Section IV.

B. Estimation of True Brightness from Digital Observations

Let us consider capturing a bright object whose digital count is as high as $Dmax$. Suppose we have a real number $\lambda > 0$ and $X_e \sim Po(\lambda)$. According to (2), X_d is capped by $Dmax$.

Table I Properties of IMX547 Image Sensor [5]

Resolution	2448x2048
Pixel Size (μm)	2.74
ADC Bit Depth	12
Quantum Efficiency Mono (% at 525 nm)	67.99
Temporal Dark Noise (Read Noise) (e-)	2.49
Temporal Dark Noise (Read Noise) (DN)	16.18
Signal to Noise Ratio Maximum (dB)	39.89
Signal to Noise Ratio Maximum (Bits)	6.63
Absolute Sensitivity Threshold (γ)	4.4
Absolute Sensitivity Threshold (e-)	2.99
Saturation Capacity (Well Depth) (e-)	9748
Saturation Capacity (Well Depth) (γ)	14337
Dynamic Range (dB)	70.26
Dynamic Range (Bits)	11.67
Gain (e-/ADU)	0.15



Fig. 4 Camera and object setup. The red box roughly corresponds to the captured image (Fig. 5).

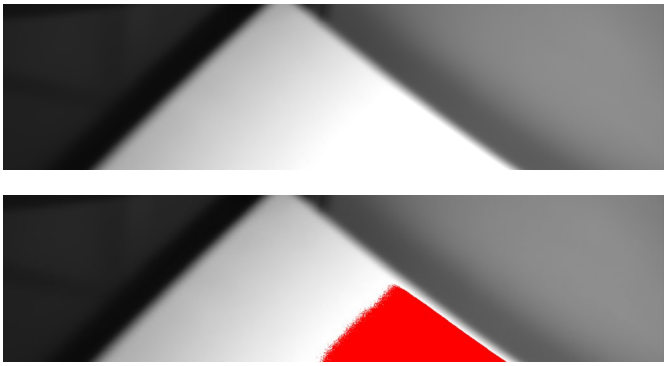


Fig. 5 Top: Captured image (averaged across frames, i.e. $EXd_{x,y}$). Bottom: same as top, constantly saturated pixels (i.e., $\forall f : P^f_{x,y} = Dmax$) are marked with red.

However, we can imagine $\widehat{X_d}$ beyond $Dmax$ based on the number of electrons fed into the well, i.e.,

$$\widehat{X_d} = \frac{X_e - b}{a}. \quad (8)$$

And its mean is

$$E\left(\widehat{X_d}\right) = \frac{\lambda - b}{a}, \quad (9)$$

which can be considered as "true brightness".

By ranging λ, $E(X_d)$ and $E(\widehat{X_d})$ can be numerically calculated based on (1) and (9). When a stationary light source is observed enough number of times, we can estimate $E(X_d)$ by averaging observed digital count. Fig. 3 shows the example of the relation between $E\left(\widehat{X_d}\right)$ and $E(X_d)$. Once the relation of those two is obtained, we can estimate the true brightness (B in the Figure) from averaged digital count (A in the Figure). Sometimes B can exceed $Dmax$. This is our proposal how to estimate true brightness from digital observations.

IV. Experimental Results

A. Camera Setup

We have conducted the experiment using real image and real sensor to validate the discussions in previous section. The camera is monochrome LUCID TRIO51S-MC whose image sensor is SONY IMX547. Some of its specifications are shown in Table I.

The experimental setup is shown in Fig. 4. We captured a static scene (Fig. 5) containing a saturated pixels (marked with red) and is used for the experiment hereafter. The focus is intentionally blurred to prevent pixel value fluctuation due to slight camera movements caused by subtle ambient vibration.

In total, 10000 frames, a frame size of 2048x512, and 12-bit pixel data were captured. According to the experiment, $Dmax$ of this camera turned out to be 4094, not 4095 (i.e., $2^{12} - 1$).

Fig. 6 shows a horizontal cross-section of the bottom line of Fig. 5, which has saturation. Fig. 7 shows the cross-section of the gray-scaled version of Fig.4, which has no saturation. Those two graphs are taken from roughly the same area to see the signal trend of the area of saturated pixels.

B. Measurement of ADC Properties

Let $P^f_{x,y}$ be the pixel value at position (x, y) of frame order f. We obtain temporal mean $EXd_{x,y}$ and temporal variance $VXd_{x,y}$ for each position (x, y). Then we obtain the level-wise average variance VXd_i for each digital level i $(i = 0,1,\ldots, Dmax)$. The formula are as follows:

$$EXd_{x,y} = Digital(mean_f\ P^f_{x,y})$$

$$VXd_{x,y} = variance_f\ P^f_{x,y}$$
$$VXd_i = mean_{x,y}(VXd_{x,y} | EXd_{x,y} = i)$$

$E(X_d)$ and $V(X_d)$ can be approximated by i and VXd_i respectively. Fig. 8 shows the linear relation of those based on actual data (i.e., Fig. 5), which coincides nicely with (7). The variance rapidly drops in very bright range (digital count > 3960). This is because most of the observation adheres to $Dmax$.

Dark range: For dark range (digital count < 250), the fitting result based on (7) was $a = 2.36398 \pm 0.003345$ and $b = 2.79738 \pm 0.6682$. Remembering that a represents the average number of photoelectrons that increases the digital count by one, $Dmax \times a = 9678.13 \pm 13.69$ is comparable with the saturation capacity (well depth), which is 9748 (cf. Table I). It is consistent with the fact that the ADC usually has less saturation capacity than well depth.

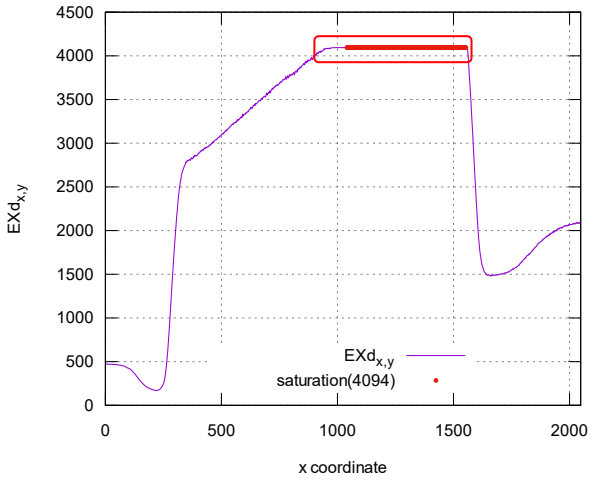


Fig. 6 Horizontal cross-section of the bottom line of Fig. 5. Red points represent saturation (i.e., all values across frames are $Dmax$). Signals in the red box area will be restored later (see. Fig. 9).

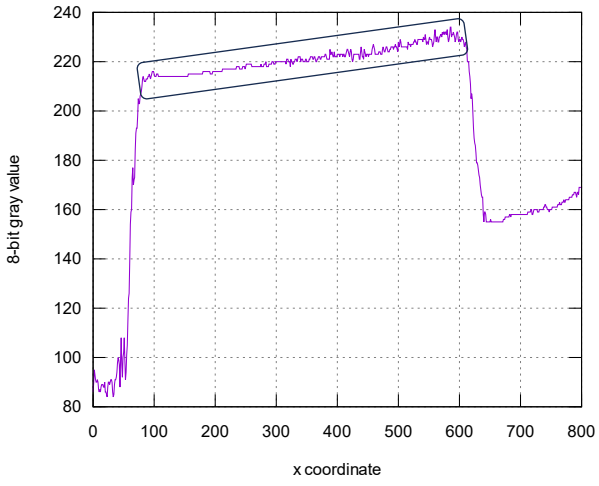


Fig. 7 Cross-section of the gray-scaled version of Fig.4, which does not precisely match the cross-section in Fig. 6; the resolution is also different. However, please note that the plateau (marked with a black box) is monotonically increasing.

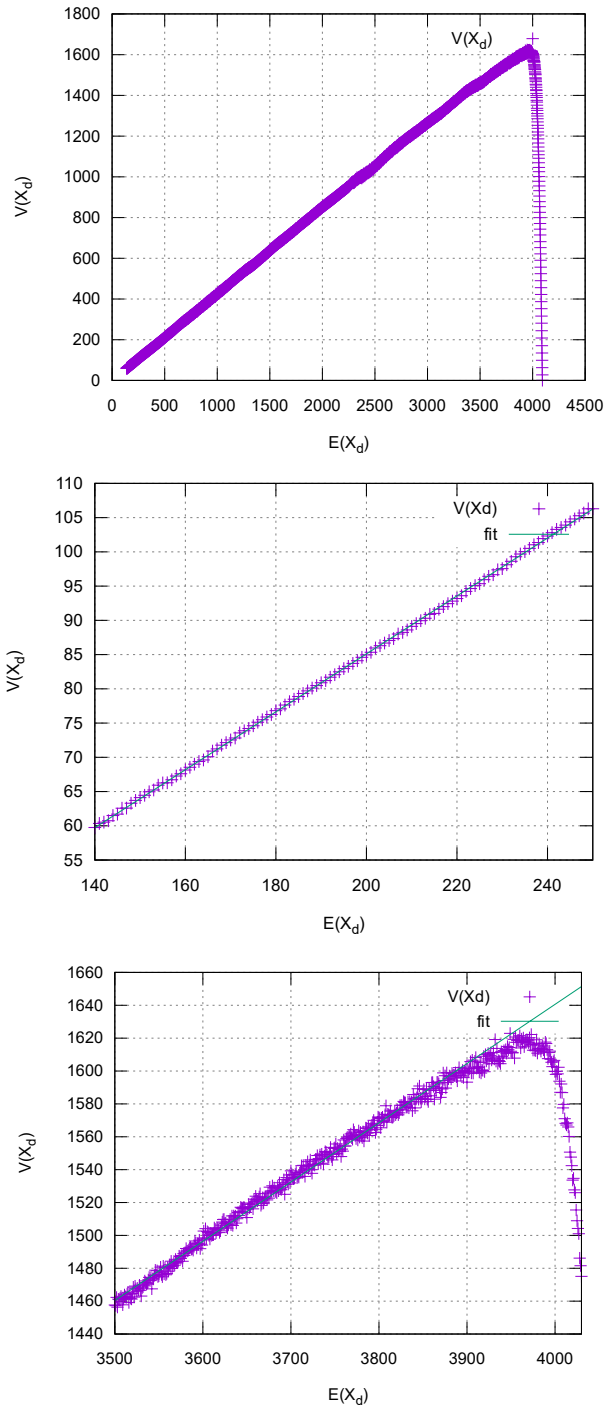


Fig. 8 Observed digital value vs. its variance. Top: whole range. Middle: dark range. Bottom: bright range.

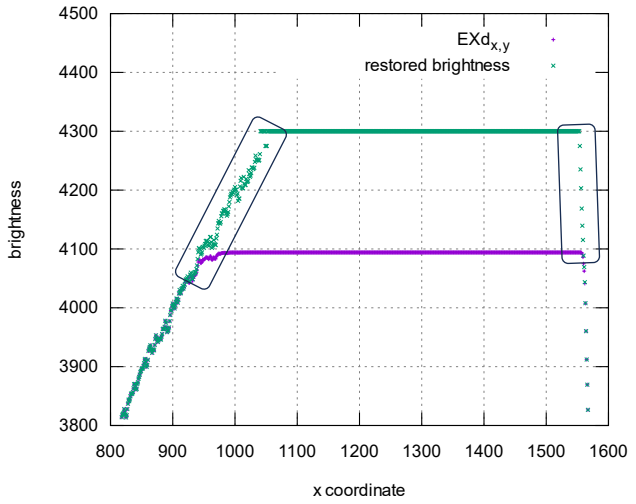


Fig. 9 Result of restoration of bright pixels (corresponding to the red box area of Fig. 6). Pixel values in black boxes are successfully restored.

Bright range: In our experiment, the bright pixels (digital count > 3500) do not precisely follow the dark pixels' ADC parameters. We suspect that the image sensor we used automatically applies some gain/offset control at the bright range. So, we fit another parameter and obtained $a = 2.77502 \pm 0.01015$ and $b = 1533.58 \pm 48.75$. The reason for the plot be more noisy than dark range's plot (Fig. 8 middle) is due to the nature of the Poisson distribution (larger mean leads larger noise variance).

C. *Restoration of True Brightness of Bright Pixels*

Based on the ADC parameters (a, b) obtained from bright range, we created a relation curve of true brightness vs. $E(X_d)$ (cf. Fig. 3) and conducted the estimation of true brightness from the data in Fig. 6 using the technique described in III.B. The maximum restoration value is chosen to be 4300. The result is shown in Fig. 9. It is clearly observed that quite a few bright pixels are successfully restored to have values greater than $Dmax$. Please note that the increasing trend seen at $800 < x < 950$ is also seen at $950 < x < 1050$ (in left black box). The same applies to the decreasing trend at $1550 < x < 1580$ (in right black box). The pixels at $1050 < x < 1550$ are *saturated*, i.e., all observed values across 10000 frames are constantly equal to $Dmax$, so there is no clue to estimate true brightness. However, if we increase the number of observations to more than 10000, we have more chance to recover more pixels.

D. *Discussion*

In our specific experiment, true brightness restoration was conducted at a flat region. However, true values of bright pixels at arbitrarily-textured region can be restored in the same manner. It should be noted that the true values of the pixels of middle brightness can practically be obtained with simple averaging along frames. Also, our proposed restoration method can coexist with high dynamic range (HDR) image acquisition because our method can expand the upper limit of each exposure of HDR image acquisition. This would be one of the applications of our technique.

V. Conclusion

In this paper, we scrutinized the temporal brightness fluctuations that light always has, and formulated precise ADC parameters which can be estimated from multiple digital camera observations. Based on those parameters, we also revealed the possibility of sensing beyond the upper limit of digital value range. The restriction of our method is, that the scene should be completely stationary – i.e., if the object moves or the illumination changes (such as fluorescent lamps) during image acquisition, our method fails.

Our future work focuses on the investigation of the restoration of dark pixel's value and to examine an analytical formulation of the relation between $E\left(\widehat{X_d}\right)$ and $E(X_d)$, which is currently empirically provided.

Acknowledgment

This work was supported by the JSPS KAKENHI Grant Number JP 23K11155.

References

[1] F. Nagumo, "Performance evolution of CMOS image sensors", O plus E, vol. 43, no. 2, pp. 168-174 (Mar.-Apr. 2021).

[2] S. Takamura, "Noise reduction based on quantization-aware multiple image averaging", IWISS2022, Dec. 2022

[3] K Irie, A. E. McKinnon, K. Unsworth, I. M. Woodhead, "A technique for evaluation of CCD video-camera noise", IEEE Trans. CSVT, vol. 18, no. 2, pp. 280-284 (Feb. 2008).

[4] R. Shimada, "Noise evaluation of DLP/EQUULEUS camera", Master's Thesis, University of Electro-Communications (Jan. 2018).

[5] FLIR, EMVA 1288 Imaging Performance of Blackfly S BFS-PGE-50S4M, http://softwareservices.flir.com/BFS-PGE-50S4/latest/EMVA/EMVA.html

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Generative Adversarial Network-Based Frame Interpolation with Multi-Perspective Discrimination

Quang Nhat Tran and Shih-Hsuan Yang*
Family Technology Company Limited, Vietnam
E-mail: tnquang1416@gmail.com Tel: +87-2436616333
* National Taipei University of Technology, Taiwan
E-mail: shyang@mail.ntut.edu.tw Tel: +886-2-27712171 ext. 4211

***Abstract*— Video frame interpolation plays an important role in video applications. This paper presents a new deep learning-based frame interpolation model under the generative adversarial networks (GANs) framework. The devised generator cascades three multi-scale sub-modules to capture coarse-to-fine visual characteristics. Furthermore, the generator is trained with an adequate loss function consisting of two adversarial and four reconstruction losses. Additional to the conventional discriminator that derives the spatial discrepancy, a new discriminator is employed to strengthen the temporal consistency among adjacent frames. The proposed method not only inherits the merit of GANs in processing speed but also produces high-quality video frames. Experimental results show that the proposed method achieves the most favorable image quality in PSNR and SSIM with a considerably short runtime compared to other state-of-the-art frame interpolation approaches.**

I. INTRODUCTION

Video frame interpolation is a technique that synthesizes intermediate frames from existing frames. Video frame interpolation finds many applications, including video creation, frame rate-up conversion, and video coding. Since the boom of deep learning, deep learning-based frame interpolation has caught much attention. Most frame interpolation networks are pixel-wise, and the most well-known approaches include kernel-based or optical flow-based ones. The kernel-based methods [1-3] derive texture effectively by aggregating neighbor pixels to estimate spatial kernels. Due to kernel size limitations, these approaches usually fail on large motions. The optical flow-based methods [4-7] derive texture effectively by aggregating neighbor pixels to estimate spatial kernels. Due to kernel size limitations, these approaches usually fail on large motions.

The Generative adversarial network (GAN) [8] is a deep learning data generation framework that iteratively trains adversarial networks, which are generators and discriminators. The generator synthesizes new data, deluding the discriminator, while the discriminator aims to recognize the corrected data. Based on that tactic, GAN has been applied in many images and video applications, including inpainting, random image generation, text-to-image translation, and video frame interpolation. Further improvement that conditions the data generation [9] with auxiliary data was proposed. Instead of only using latent vectors [10], labels, maps, and optical flow can be used as input. In video frame interpolation, GAN research also considers using attention maps, occlusion maps, and linear data as auxiliary inputs.

Considering quality and complexity, we introduce a GAN-based framework for frame interpolation comprising a generator and two discriminators (Fig. 1) without additional input requirements. The generator applies a three-scale architecture to capture multi-perspective features for deriving a frame from two adjacent ones. The discriminators consider

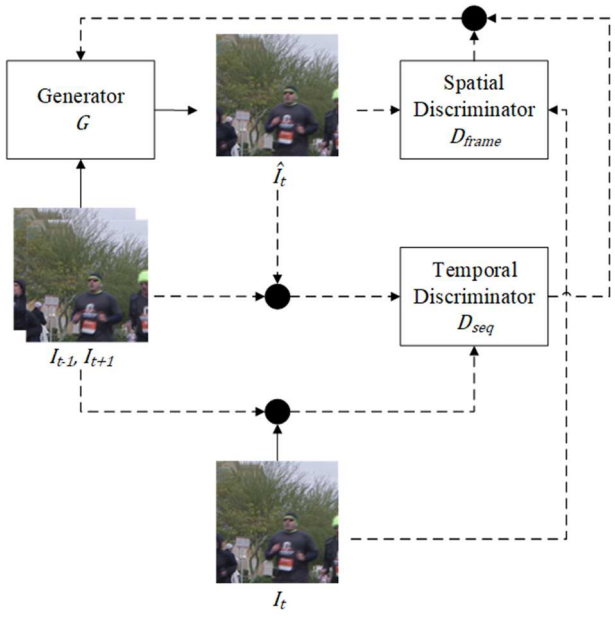


Fig. 1 The overall architecture of the proposed framework, in which two discriminators, D_{frame} and D_{seq}, facilitate a multi-scale generator G spatially and temporally. The solid and dashed arrows depict the flows of generation and discrimination, respectively.

This work was financially supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. , under Grants MOST 110-2221-E-027-083-MY2.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

spatial and temporal perspectives corresponding to frame-based and sequence-based models.

The main contributions of this study are summarized below:

- We design and implement a GAN-based frame interpolation framework, considering spatial correctness and temporal consistency.
- We introduce a training loss function for the generator consisting of adversarial and reconstruction loss ones to measure the reality and correctness of the output.
- We indicate each incorporated element's contribution with ablation studies and show the performance with a comparison with state-of-the-art solutions.

II. Related Works

Most recent deep learning frame interpolation solutions fall into the kernel- and optical flow-based categories. Niklaus *et al*. [1] introduced a kernel-based neural network for synthesizing intermediate frames via adaptive separable kernel estimation. Niklaus and Liu [13] manipulated optical flows and interpolation kernels with their proposed SoftSplat with forward warping operations. Besides, Liu *et al*. [4] used optical flows to copy pixels from adjacent frames for interpolation. Jiang *et al*. [5] introduced an end-to-end multi-frame interpolation network utilizing convolutional neural networks (CNNs) for jointly estimating bi-directional optical flow and occlusion mask. They also addressed the motion artifacts with visibility maps predicted by a U-Net [11]-based network.

Recently, several methods for frame interpolation have incorporated supplementary information, such as optical flows, depth maps, and occlusion maps, for high frame generation performance. Park *et al.* [7] proposed a frame interpolation network jointly estimating bilateral motion with bilateral cost volume to facilitate the optical flow estimation and dynamic filter generation network. Bao *et al.* [12] proposed an adaptive warping layer to enhance frame interpolation by integrating optical flows and interpolation kernels for motion estimation and motion compensation-driven design. From a similar point of view, Bao *et al.* [6] further used depth maps to reduce occluded objects' affection with their depth-aware projection layer. Kong *et al.* [14] introduced a pyramid frame interpolation network that refines the bilateral intermediate flow fields and features iteratively. The authors further trained IFRNet with a task-oriented optical flow distillation loss for higher performance and a geometry consistency regularization loss for enhancing structure layout. Huang *et al.* [15] introduced a frame interpolation network that utilizes intermediate optical flows with a coarse-to-fine strategy and leverages knowledge distillation from the teacher flow network.

GANs with an adversarial training strategy enable a different direction for high-quality and robust frame interpolation. GAN-based frame interpolation frameworks condition the generation to derive intermedia data by taking adjacent frames and other auxiliary data as input. Amersfoort *et al.* [16] considered a pyramid architecture to design a GAN-based frame interpolation framework. Wen *et al.* [17] proposed a multi-frame interpolation framework by concatenating two pairs of generator and discriminator aided with linear input data. Xiao and Bi [18] introduced a GANs frame interpolation framework utilizing multi-scale architecture and residual connections. Their design utilizes four pairs of generator-discriminator and incorporates attention map-based frame synthesis to enable the model to focus more on relevant regions of features based on its own estimated attention map. Tran and Yang [19] proposed a framework incorporating a two-scale generator and a discriminator to synthesize an intermediate frame from the adjacent ones.

Inspired by the previous works, the proposed method utilizes a generator for multi-perspective features and two discriminators for distinguishing spatially and temporally.

III. Proposed Methods

In this research, the proposed method aims to synthesize an intermediate frame from two given ones, as below equation:

$$\hat{I}_t = G(I_{t-1}, I_{t+1}) \tag{1}$$

Here, the target data belong to the $C \times H \times W$ vector space, in which C, H, and W represent the picture's channel, height, and width.

In the rest of this section, we describe the proposed GAN-based frame interpolation framework (Fig. 1) and explain the design for balancing generation and discrimination. We also provide a detailed description of the training loss function comprising adversarial and reconstruction loss terms.

A. Frame Interpolation Framework

Following conditional GAN-based design, the proposed method utilizes a multi-scale generator for frame interpolation and two discriminators for adversarial training. Here, the generator takes the feedback from both the discriminators, in which one considers the spatial correctness and another also pays attention to temporal consistency.

Multi-scale Generator. To explicit the multi-scale features from input and synthesize output from coarse to fine, the generator (Fig. 2) incorporates two different sub-modules deriving frames at different resolutions, as below equations:

$$\hat{I}_t^2 = G_2(I_{t-1}^2, I_{t+1}^2), \tag{2}$$

$$\hat{I}_t^1 = G_1\left(I_{t-1}^1, S(\hat{I}_t^2), I_{t+1}^1\right), \tag{3}$$

$$\hat{I}_t^0 = G_0\left(I_{t-1}^0, S(\hat{I}_t^1), I_{t+1}^0\right), \tag{4}$$

where I_t^i and $\hat{I}_t^i$ (i = 0, 1, 2) represent the existing frame and generated frame at time step t at the i^{th} level, respectively, and

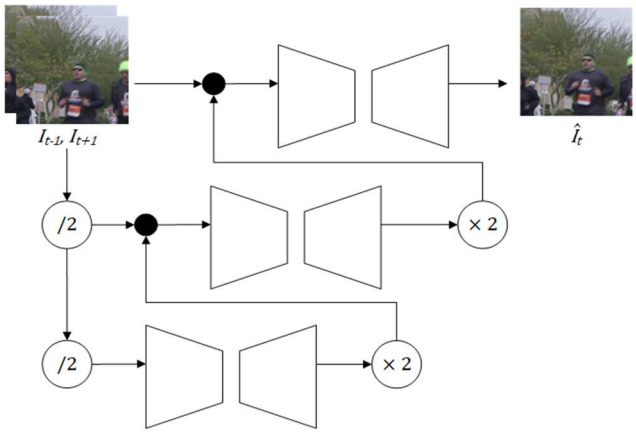


Fig. 2 The general architecture of the proposed generator.

S depicts the up-scale operation. There, the coarsest one captures the overall structure layout, and the finer one completes with more details. The finest frame $\hat{I}_t^0$ benefits from the previous inter-output $\hat{I}_t^1$ and $\hat{I}_t^2$.

All generator sub-modules share a similar backbone design following encoder-decoder architecture. Specifically, the encoder extracts features from input with a sequence of convolutional layers, batch normalizations, and Parametric Rectified Linear Unit (PreLU) [20] activation functions. Here, we use PReLU at the encoder side to reduce feature loss in the extraction process with convolution operations. Thus, the models become more robust in the deep learning environment. Furthermore, the generator is incorporated with skip connections for training efficiency enhancement and recovering spatial information by explicitly copying features between layers. Accordingly, the decoder synthesizes the output with transposed convolutional layers, batch normalizations, Dropout layers, and Rectified Linear Unit (ReLU) activation functions.

Spatial- and Temporal-Discriminators. A generated frame should contain corrected and consistent content with the adjacent ones. Therefore, we utilize two discriminators for classifying the output spatially and temporally. The first network evaluates frames individually (5), while the second considers a patch of three consecutive frames for temporal consistency (6). In this way, discrimination produces better feedback for the generation and facilitates the generation. Besides, this design improves the discrimination efficiency that helps balance the capacity and performance between generated and discriminated operations for more efficient training.

$$d_{spatial} = D_{spatial}(\hat{I}_t^0) \quad (5)$$

$$d_{temporal} = D_{temporal}(I_{t-1}^0, \hat{I}_t^0, I_{t+1}^0) \quad (6)$$

The two spatial and temporal discriminators share a similar backbone architecture (Fig. 3) that sequences convolutional layers, Batch normalizations, PReLU activation functions, and a sigmoid activation function.

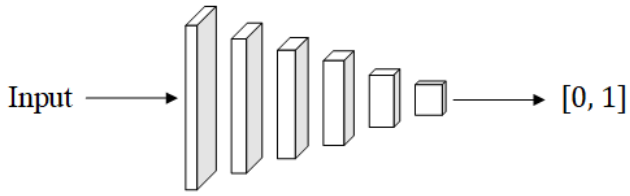


Fig. 3 The general architecture of the proposed discriminators.

B. Training Loss Function

Following the GAN-based design [8], we apply the minimax approach for training the proposed framework. As the below equation, the generator tries to minimize the classification accuracy, the term $(1 - D(x))$, with the generated data. Meanwhile, the discriminator attempts to maximize it.

$$\min_G \max_D E_{x \sim p_{data}(x)}[\log D(x)] + E_{x \sim p_g(x)}[\log(1 - D(x))] \quad (7)$$

We train the two discriminators by measuring classification accuracy with the Binary Cross Entropy (BCE) loss function. This function is also applied for measuring the video frame degradation as the adversarial loss term besides the reconstruction one, as (8), where λ_{adv} and λ_{rec} depict the non-negative weights. In practice, we apply 0.00005 for λ_{adv} and 1.0 for λ_{rec}.

$$\mathcal{L} = \lambda_{adv} \times \mathcal{L}_{adv} + \lambda_{rec} \times \mathcal{L}_{rec} \quad (8)$$

The adversarial loss term $\mathcal{L}_{adv}$ measures the feedback from the spatial and temporal discriminators, as (9), and allows the generator considers them for learning.

$$\mathcal{L}_{adv} = \mathcal{L}_{spatial} + \mathcal{L}_{temporal} \quad (9)$$

For reconstruction loss function, we refer to FI-DUSGAN [19], incorporating $\mathcal{L}_2$, multi-scale structural similarity (MS-SSIM) loss [21] $\mathcal{L}_{MS\text{-}SSIM}$, and gradient difference loss [22] $\mathcal{L}_{GDL}$ for pixel-wised generation correctness, low motion blur, and high friendliness in human vision. In this work, we add $\mathcal{L}_{census}$ to enhance the generation learning for balancing the performance between the generator and discriminators. This function measures the soft Hamming distance between ternary census transformed images with a patch size of 7×7 [23] to facilitate the generation with a more reliable constancy assumption based on the additive and multiplicative illumination changes compensation of the census transform. The proposed function is described below.

$$\mathcal{L}_{rec} = \mathcal{L}_2 + \mathcal{L}_{census} + \mathcal{L}_{MS-SSIM} + \mathcal{L}_{GDL} \quad (10)$$

IV. EXPERIMENTS

In this section, we first describe the experimental setup, including the dataset and parameters for training and implementation. Then, we evaluate the generation performance by an ablation study and a comparison with the state-of-the-art in terms of peak signal-to-noise ratio (PSNR) for pixel-by-pixel comparison and SSIM for perception-based similarity.

A. Experimental Setup

In this experiment, we train the proposed framework with the temporal frame interpolation subset of the Vimeo-90k dataset [24]. The dataset consists of several 480×256-frame patches containing three consecutive frames. Besides, we further pre-process the training dataset with random cropping and flipping operations. The target training dataset includes 51,313 128×128-frame-patches, and the testing dataset has 3,782 480×256-frame-ones.

We use PyTorch to develop the models and manipulate the training with the Adam optimizers with the recommendation parameters for implementation. Besides, we set the training batch size as 32 and the initial learning rate as 0.001 while decaying it after every 20 training epochs until reaching 10^{-7}.

We perform the experiment on the NVIDIA Tesla V100 GPU environment.

B. Ablation Study

In this ablation study, we compare various versions of the proposed framework to demonstrate the contributions of each component. As in Table I, the proposed design (5) produces the highest quality with 37.20dB in PSNR and 0.940 in SSIM. This version confirms the contribution of the modifications and the efficiency in balancing between generation and discrimination.

Table I. Ablation study on architecture and training loss function, in which the best version is highlighted in bold.

Version	Description	Loss	PSNR	SSIM
(1)	3-scale generator only	$\mathcal{L}_2$-based	36.85	0.940
(2)	2-scale-generator and seq-discriminator	$\mathcal{L}_2$-based	36.73	0.941
(3)	3-scale-generator and seq-discriminator	$\mathcal{L}_1$-based	35.75	0.928
(4)	3-scale-generator and seq-discriminator	$\mathcal{L}_2$-based	36.39	0.939
(5) Proposed	**3-scale-generator and 2 discriminators**	**$\mathcal{L}_2$-based**	**37.20**	**0.940**

Table II. Ablation study on weights of the training loss function, in which the proposed version is highlighted in bold.

Version	λ_{adv}	λ_{rec}	PSNR	SSIM
(1)	0.0001	0.5	36.84	0.938
(2)	**0.00005**	**1.0**	**37.20**	**0.940**
(3)	0.0	1.0	36.85	0.940

According to Table I, a good design GAN-based framework having efficient adversarial training will allow us to simplify the network, such as the two-scale and the three-scale one of (2) and (1), respectively. Versions (3) and (4) produce lower performance due to the imbalance between generation and discrimination. Here, the generator is more complex, and the discriminator does not have enough capability to keep the training effective. Therefore, the incorporated frame-based discriminator helps balance the generation and discrimination, enabling the proposed framework to achieve higher performance. From a different point of view, we consider replacing $\mathcal{L}_2$ with $\mathcal{L}_1$ for training the generator. However, the low-quality output of (3) shows that the mean square error-based loss function is more compatible with the proposed architecture and training loss function. In our observation, incorporating one more discriminator into (2) causes the vanishing gradients problem, and the generator fails to learn. The training problem confirms the suitability of our model designs with the applied training loss function. Besides, our study on weights of the training loss function (Table II) shows the effectiveness of the selected ones for the proposed function.

C. Comparison to the state-of-the-arts

In this section, we perform a comparison between DVF [4], SepConv [1], FI-MSAGAN [18], BMBC [7], FI-DUSGAN [19], and the proposed framework. For comparison, we retrain the compared models using a re-implemented version of FI-MSAGAN [25] and official implementations of others. Here, the proposed generator achieves the best PSNR and second-best SSIM values (Table III) with the Vimeo-90k dataset within a short runtime (9.1 ms). Overall, GAN-based methods can

Table III. Comparison between the frame interpolation methods, in which the best and second-best values are highlighted in bold and underlined, respectively.

Methods	PSNR (dB)	SSIM	Runtime (ms)
DVF	31.92	0.658	51
SepConv	35.97	0.913	115
FI-MSAGAN	35.99	0.893	26
BMBC	36.74	**0.948**	686
FI-DUSGAN	<u>37.02</u>	<u>0.940</u>	9
The proposed	**37.20**	<u>0.940</u>	9.1



(a) PSRN: 37.93dB; SSIM: 0.966



(b) PSRN: 38.40dB; SSIM: 0.952

Fig. 4 The generated output of the proposed framework.



Fig. 5 The generated output of the proposed framework (PSRN: 34.78dB; SSIM: 0.962).

achieve high-quality output with lower complexity owing to the efficient adversarial training. In our comparison, the proposed framework outperforms FI-DUSGAN by 0.18dB in PSNR and FI-MSAGAN by 1.21dB in PSNR and 0.47 in SSIM.

As shown in Fig. 4, the proposed framework synthesizes interpolated frames well in various scenarios, including frames with high contrast details with low to medium changes. On the other hand, it still fails to handle some challenging scenarios, particularly those with fast and sudden motions between frames, such as the moving cars in Fig. 5.

V. CONCLUSION

We introduced a frame interpolation framework utilizing three adversarial networks, including a three-scale generator and two discriminators. The devised generator captures coarse-to-fine visual characteristics by three-scale sub-modules with an adequate combination of loss functions. Besides a discriminator considering the quality of generated data, a temporal one is employed to guarantee consistency among adjacent frames. Thus, the proposed framework effectively balances the adversarial relationship to ensure optimal performance. Our ablation study confirmed the contribution of each designed element. The proposed design improves the PSNR by 0.35 dB without increasing complexity compared to the generator-only version. The proposed framework can derive high-quality frames in various scenarios, and it outperforms other GAN-based solutions, such as FI-MSAGAN (+3.4%) and FI-DUSGAN (+0.5%), with a similar complexity requirement. In future work, we will consider tackling more challenging scenarios, such as videos with sudden motions.

REFERENCES

[1] S. Niklaus, L. Mai and F. Liu, "Video Frame Interpolation via Adaptive Separable Convolution," in IEEE International Conference on Computer Vision (ICCV), Venice, 2017.

[2] T. Ding, L. Liang, Z. Zhu and I. Zharkov, "CDFI: Compression-Driven Network Design for Frame Interpolation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, 2021.

[3] X. Cheng and Z. Chen, "Multiple Video Frame Interpolation via Enhanced Deformable Separable Convolution," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 7029-7045, 2022.

[4] Z. Liu, R.A. Yeh, X. Tang, Y. Liu and A. Agarwala, "Video Frame Synthesis Using Deep Voxel Flow," in International Conference on Computer Vision (ICCV), Venice, 2017.

[5] H. Jiang, D. Sun, V. Jampani, M. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.

[6] W. Bao, W-S. Lai, C. Ma, X. Zhang, Z. Gao, M-H. Yang, "Depth-Aware Video Frame Interpolation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[7] J. Park, K. Ko, C. Lee and C-S. Kim, "BMBC: Bilateral Motion Estimation with Bilateral Cost Volume for Video Interpolation," in European Conference on Computer Vision, 2020.

[8] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in Neural Information Processing Systems (NIPS), 2014.

[9] M. Mirza and S. Osindero, Conditional Generative Adversarial Nets, arXiv preprint arXiv:1411.1784v1, 2014.

[10] A. Radford, L. Metz and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434v2, 2016.

[11] O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv preprint arXiv:1505.04597, 2015.

[12] W. Bao, W-S. Lai, X. Zhang, Z. Gao and M-H. Yang, "MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 933-948, 2021.

[13] S. Niklaus and F. Liu, "Softmax Splatting for Video Frame Interpolation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[14] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, Ch. Wang and J. Yang, IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation, arXiv prePrint arXiv:2205.14620, 2022.

[15] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-Time Intermediate Flow Estimation for Video Frame Interpolation," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, 2022.

[16] J.V. Amersfoort, W. Shi, A. Acosta, F. Massa, J. Totz, Z. Wang and J. Caballero, Frame Interpolation with Multi-Scale Deep Loss Functions and Generative Adversarial Networks, arXiv preprint arXiv:1711.06045, 2019.

[17] S. Wen, W. Liu, Y. Yang, T. Huang and Z. Zeng, "Generating Realistic Videos From Keyframes With Concatenated GANs," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 8, pp. 2337 - 2348, 2019.

[18] J. Xiao and X. Bi, "Multi-Scale Attention Generative Adversarial Networks for Video Frame Interpolation," IEEE Access, vol. 8, pp. 94842-94851, 2020.

[19] Q.N. Tran and S-H. Yang, "Video Frame Interpolation via Down-Up Scale Generative Adversarial Networks," Computer Vision and Image Understanding, vol. 220, 2022.

[20] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in IEEE International Conference on Computer Vision (ICCV), Santiago, 2015.

[21] H. Zhao, O. Gallo, I. Frosio and J. Kautz, "Loss Functions for Image Restoration With Neural Networks," IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47-57, 2017.

[22] M. Mathieu, C. Couprie and L. LeCun, "Deep multi-scale video prediction beyond mean square error," in International Conference on Learning Representations (ICLR), 2016.

[23] S. Meister, J. Hur and S. Roth, "UnFlow: Unsupervised Learning of Optical Flow With a Bidirectional Census Loss," in AAAI Conference on Artificial Intelligence, 2018.

[24] T. Xue, B. Chen, J. Wu, D. Wei and W.T. Freeman, "Video Enhancement with Task-Oriented Flow," International Journal of Computer Vision, vol. 127, no. 8, pp. 1106-1125, 2019.

[25] Q. N. Tran, "A Reimplementation of FI-MSGAN Using PyTorch," 2021. [Online]. Available: https://github.com/tnquang1416/FI-MSGA.

ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation

Hrishav Bakul Barua*†, Ganesh Krishnasamy*, KokSheik Wong*, Kalin Stefanov‡, and Abhinav Dhall‡§
* School of Information Technology, Monash University, Malaysia
† Robotics and Autonomous Systems Group, TCS Research, India
‡ Faculty of Information Technology, Monash University, Australia
§ Centre for Applied Research in Data Sciences, Indian Institute of Technology Ropar, India

***Abstract*—High Dynamic Range (HDR) content creation has become an important topic for modern media and entertainment sectors, gaming and Augmented/Virtual Reality industries. Many methods have been proposed to recreate the HDR counterparts of input Low Dynamic Range (LDR) images/videos given a single exposure or multi-exposure LDRs. The state-of-the-art methods focus primarily on the preservation of the reconstruction's structural similarity and the pixel-wise accuracy. However, these conventional approaches do not emphasize preserving the artistic intent of the images in terms of human visual perception, which is an essential element in media, entertainment and gaming. In this paper, we attempt to study and fill this gap. We propose an architecture called *ArtHDR-Net* based on a Convolutional Neural Network that uses multi-exposed LDR features as input. Experimental results show that *ArtHDR-Net* can achieve state-of-the-art performance in terms of the HDR-VDP-2 score (i.e., mean opinion score index) while reaching competitive performance in terms of PSNR and SSIM.**

I. Introduction

The High Dynamic Range (HDR) imaging is introduced to handle the scenario where the visual content has a high ratio between the brightest and the darkest pixels [1]. HDR allows the inclusion of multiple dynamic exposures in a single visual content. Hence, if an image is in HDR, it has a wide gamut of brightness levels, i.e., there is a significant variation in light levels within a scene with high details. In other words, the well-lid and dark regions in a scene are captured with equal clarity using HDR technology. The main goal of HDR imaging is to capture the real world lighting, greater luminance distribution, brightness, detailed shadows, and radiance in a manner which is as realistic as possible and can mimic the Human Visual System and its level of perception. HDR technology has many applications, e.g., in the fields of media and entertainment [1], Augmented/Virtual Reality (AR/VR) [2], [3], gaming and visualization [2], [4], and computational photography [5]. With the increase in HDR supported hardware devices such as HDR displays, AR/VR sets, mobile screens, and cameras, researchers have tried to find ways and methods to successfully retrieve the HDR counterparts from the conventional Low Dynamic Range (LDR) content, known as inverse tone mapping.

HDR content can be created using special hardware (e.g., expensive cameras and visual sensors) and/or software (e.g., virtual environments and renderers). However, converting the existing LDR content to HDR requires unique methods and algorithms [6], [7] to expand the color gamut and dynamic range, which is usually achieved through inverse tone mapping and image enhancement. In this paper, we attempt to take an LDR image as input and create an HDR image with perceptually realistic and visually accurate information in terms of the Human Visual System represented by the HDR-VDP-2 scores (see section IV-B). This work contrasts most state-of-the-art methods that attempt to solve the problem of structural similarity of the generated HDR with respect to the LDR input. The main contributions of this work are threefold:

- We propose a CNN-based network (Section III) for HDR image content creation drawing inspiration from the method in [8]. We leverage the features from three different exposure values (EV -2, 0, and +2) and create a combined feature map for a better representation of under/over-exposed areas in the LDRs.
- We perform a thorough experimental validation with quantitative and qualitative comparisons (Sections IV and V) to establish the efficacy of our contributions.
- We present a detailed ablation study (Section VI) on the different losses used to train the network and the skip connections of the CNN layers utilized for better representation of the low-level features.

II. Related Work

Researchers have employed various Machine Learning (ML) and Deep Learning (DL) methods [7] for HDR content creation. Below, we categorize the recent relevant literature into two major buckets, i.e., image- and video-based. These buckets can be further categorized based on single and multiple exposure LDR inputs.

A. Image-based Methods

Khan et al. [8] proposed a feedback-based Convolutional Neural Network (CNN) to create an HDR image from a single exposure LDR input image. Their method uses a dense feedback mechanism based on Recurrent Neural Network (RNN) on top of a conventional CNN to obtain both low-level and high-level features. This coarse-to-fine feature

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

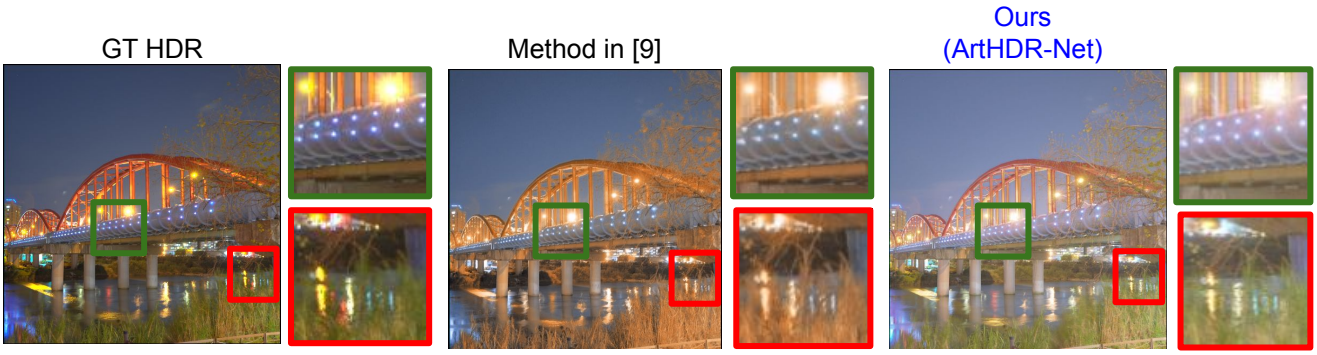


Fig. 1: Output produced by our method ArtHDR-Net compared to one of the closest state-of-the-art [9] methods in terms of HDR-VDP-2 score [10], [11]. Our method reconstructs the HDR content with better naturalness in terms of color contrast and luminance levels.

representation in various iterations leads to HDR creation at a higher quality. Phuoc-Hieu Le et al. [12] proposed a Neural Network (NN) architecture to learn and generate multiple exposures from a single image. The model can reconstruct pixel radiance and details of hallucination in different exposures by inverting the camera response. Yu-Lun Liu et al. [9] proposed a CNN architecture that learns the reverse process of camera pipeline, i.e., inverse tone mapping. The method employs three CNNs to reverse the three basic steps in camera pipeline, i.e., clipping of dynamic range, camera response function (CRF) based non-linear mapping, and quantization. Then the three sub-tasks are jointly fine-tuned to reduce the accumulated errors. Yet another interesting work [13] (although not in the DL domain) talks about an inverse tone mapping operator, which can preserve the artistic intent of the content creator in the generated HDR images i.e., the images look more realistic perceptually. Jinghui Li et al. [14] presented a novel multi-scale CNN with an attention mechanism that tackles the problems of color quantization and the loss of information in over/under-exposed regions. The method re-scales channel-wise features using a channel attention method adaptively. Another DL based method [15] proposed a novel architecture that can recover the saturated pixels of the LDR images with better quality leading to visually and perceptually better HDR images with well-preserved textures. The method employs a feature masking mechanism that limits the contribution of the features from the saturated regions of the images, which in turn reduces the checkerboard and halo artifacts of these regions. Gabriel Eilertsen et al. [16] proposed a CNN-based method that overcomes the challenges of predicting HDR pixels from complex LDR images with over/under exposed areas. Another interesting work [17] considered the low peak brightness issues and the missing of artistic intent of the existing methods. It proposed a mid-level automatic tone-mapping operator. Gaofeng Cau et al. [18] proposed a spatial and channel dimensional decoupled kernel prediction network that combines a simple module which produces preliminary HDR results and an encoder/decoder module to produce pixel-wise kernels. Finally, these two results are convolved to produce high-quality HDR images. The authors claim that the method can use the relevant information from under- and over-exposed images while avoiding pixel-wise degradation and information redundancy across channels. A work by Hanbyol Jang et al. [19] porposed a novel method of HDR generation from LDR based on histogram and color difference between HDR-LDR (multi-exposure) pairs.

B. *Video-based Methods*

Chen et al. [1] proposed a new architecture for reconstructing HDR videos from LDR counterparts based on a divide-and-conquer policy. The final framework has three components: adaptive global color mapping, local enhancement, and highlight generation. Yang et al. [20] proposed an interesting framework which is capable of utilizing two modalities of input, i.e., stacked events and the LDR frames to create a learnt latent space. The method is guided by the events to facilitate multi-modal learning for HDR creation. The method decribed in [21] proposed a Generative Adversarial Network (GAN) approach for reconstructing HDR videos from multiple exposed LDR sequences. The method starts with a denoising module to clean the LDR input and uses optical flow to align the neighbouring multi-exposed LDRs in the sequence.

C. *Positioning Our Method*

Almost all of the discussed state-of-the-art methods focus on improving the overall structural quality of the generated HDR images while preserving the original structure in the LDR inputs. However, only a handful (i.e., less than five) of the works try to tackle the problem of generating HDR images from the perspective of the Human Visual System and creating perceptually pleasing image for the human eye (see Figure 1). Most conventional methods also lack discussion of the preservation of the artistic intent [13] of the generated HDR content. In this work, we attempt to address these limitations in the current literature along with structural similarity preservation. We achieve better peak brightness with a deep CNN-based learning model along with a feedback block drawing inspiration from [8], [22] by utilizing features extracted from a single LDR image.

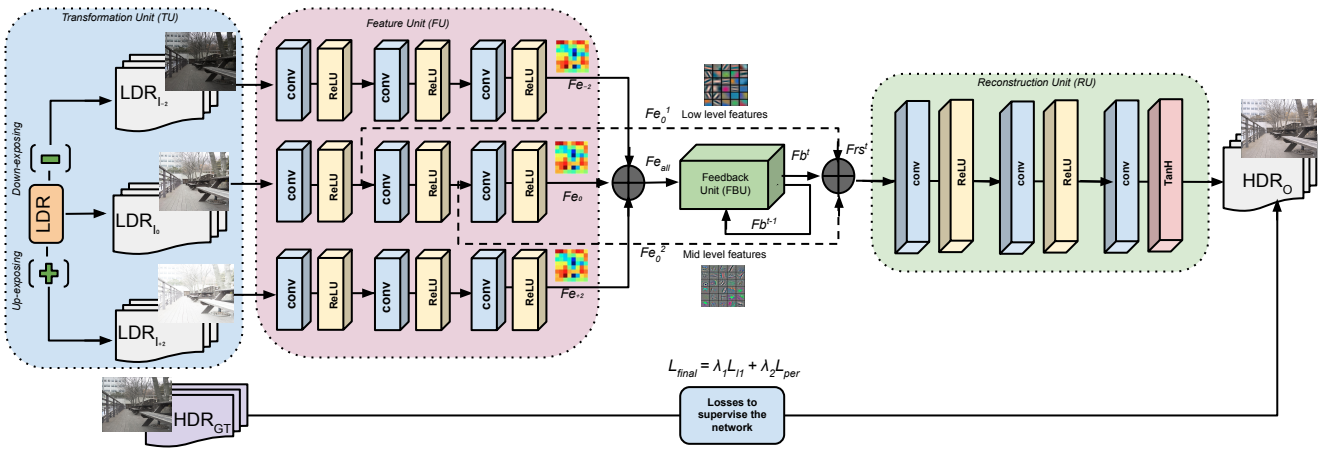


Fig. 2: Illustration of the proposed ArtHDR-Net architecture.

III. Proposed Method

This section describes our proposed architecture for HDR image creation using CNN and a feedback mechanism utilizing coarse and fine features extracted from multi-exposed versions of an input LDR image.

A. Architecture and Model

Our architecture, which we name ArtHDR-Net, comprises four basic units: a) Transformation unit (TU), b) Feature unit (FU), c) Feedback unit (FBU), and d) Reconstruction unit (RU). Figure 2 shows the overall architecture of the method.

TU is responsible for taking an LDR image as input and generating the over-exposed and under-exposed versions (EV ±2). We use an off-the-shelf method [12] to accomplish this task. Regardless of how the input image was captured, we assume that it has a middle exposure value, EV 0. The resulting multi-exposed LDRs (LDR_{I_0}, $LDR_{I_{+2}}$, and $LDR_{I_{-2}}$) are used by FU.

FU has three parallel branches for the three LDRs. Each branch has 3×3 kernels in each of the three 2D convolutional layers (conv), followed by rectified linear unit (ReLU) activation function. The three layers (in each branch) help extract three levels of features (i.e., low, mid, and high). We denote the process in FU as $func_{FU}$ and the output of FU (feature information in each branch) as Fe such that:

$$Fe_{-2} = func_{FU}(LDR_{I_{-2}}) \tag{1}$$

$$Fe_0 = func_{FU}(LDR_{I_0}) \tag{2}$$

$$Fe_{+2} = func_{FU}(LDR_{I_{+2}}). \tag{3}$$

The final feature map Fe_{all} is defined as follows:

$$Fe_{all} = Fe_{-2} + Fe_0 + Fe_{+2}. \tag{4}$$

Initially, FBU takes the Fe_{all} as input and subsequently considers the output of FBU from the previous iteration as shown in Figure 2. Similar to [23], [24], we adopt the feedback mechanism with three dilated dense blocks (having dilation rate of 3), which enhances the network's receptive field without increasing the number of parameters and computational time. This choice is made to utilize all the hierarchical features of the input LDRs. FBU also contains two 1×1 feature compression layers at the beginning and the end of each of the three dilated blocks. Each of the three dilated blocks contain four 3×3 convolutional layers with ReLU activation. This helps to reduce the number of parameters of the network and enhances the learning capability of the network. Finally, FBU starts with a 1×1 convolutional layer for compression and ends with a 3×3 convolutional layer for further processing. This supports global and local feedback mechanisms in the entire FBU as well as in the three dilated block levels. We denote the process of FBU by $func_{FBU}$ and its output by Fb such that a hidden state defined as:

$$Fb^t = func_{FBU}(Fe_{all}, Fb^{t-1}), \tag{5}$$

where t represents the iteration in the FBU.

When $t = 1$ and Fb is nil, we initialize the hidden state of the FBU with Fe_{all} only. Then we also consider the low-level features, i.e., the coarser (low and mid level) features from the first and second convolutional layers of the middle branch (see Figure 2) of FU which has LDR_{I_0}. We do this to guide the network with a set of two residual skip connections for better HDR feature representations in RU. This also helps in preserving the details and artistic intent of the generated HDR image. We denote the final output of the residual skip connections along with the output of FBU at every iteration by Frs. We represent this as follows:

$$Frs^t = Fe_0^1 + Fe_0^2 + Fb^t, \tag{6}$$

where Frs^t represents the final feature map (at iteration t) after the FBU operates on the feature output. Fe_0^1 and Fe_0^2 represent the low-level features from LDR_{I_0}, which we get

as output from the first and second convolutional layers of FU (see Figure 2).

Next, RU reconstruct the HDR image, where it takes the Frs as input in every iteration. RU is made of three 3×3 2D convolutional layers with ReLU activation (following the first two layers) and a hyperbolic tangent activation function (TanH) following the third layer (see Figure 2). We denote the function of RU by $func_{RU}$ and the output by HDR_O.

$$HDR_O^t = func_{RU}(Frs^t), \tag{7}$$

where Frs^t represents the final learned feature map at the t^{th} iteration and HDR_O^t is the final HDR image generated at iteration t.

B. Loss Functions

In this work, we calculate the losses between the tone-mapped versions of the generated HDR image ($HDR_{O(TM)}$) and ground truth (GT) HDR image ($HDR_{GT(TM)}$). This is done to avoid the high-intensity values to dominate the loss function values. We use the $\mu - law$ [25] to compress the HDRs. We employ two loss functions, namely, L1 (L_{l1}) and perceptual (L_{per}) losses, to supervise our model at every iteration t. The complete loss function is defined as

$$L_{final} = \lambda_1 L_{l1} + \lambda_2 L_{per}. \tag{8}$$

The L_{l1} loss is the Mean Absolute Error (MAE) calculated between each pixel of $HDR_{GT(TM)}$ and $HDR_{O(TM)}$. The L_{l1} loss is defined as:

$$L_{l1} = \frac{1}{n} \sum_{t=1}^{n} ||HDR_{O(TM)}^t - HDR_{GT(TM)}||. \tag{9}$$

The L_{per} loss is the perceptual loss between $HDR_{GT(TM)}$ and $HDR_{O(TM)}$ generated via the function $func_{per}$ inspired by [26]. $func_{per}$ is a VGG[1] based loss which tries to make the output closer to perceptual similarity. It is created using the ReLU activation layers of the famous 19 layer VGG network. The L_{per} loss is defined as:

$$L_{per} = \frac{1}{n} \sum_{t=1}^{n} func_{per}(HDR_{O(TM)}^t, HDR_{GT(TM)}). \tag{10}$$

We do not consider L2 loss because it is too sensitive towards outliers. We use a combination of the L_{l1} and L_{per} loss because using only the L1 loss produces visually degraded HDRs. λ_1 and λ_2 are set to 0.1 and 0.5 empirically.

IV. Experiments

We trained and tested our model on Ubuntu 20.04.6 LTS work station with Intel(R) Xeon(R) CPU E5-2687W v3 @ 3.10GHz (20 CPU cores), 126 GB RAM (with 2 GB Swap space) and 1.4 TB SSD for dataset storage. We used batch size of 10 and Adam optimizer [27] for training the model. The model was trained for 150 epochs. The learning rate of the training process was set to 2×10^{-4} at first and later adjusted to decay. We employed an iteration count of 4 for the feedback operation in FBU.

A. Datasets

We used three datasets for training and testing. Specifically, the City Scene dataset [5] (*ICCV 2017*) contains about 20K HDR-LDR pairs and ground truth HDR images. We also used the LDR-HDR pair dataset from [19] (*IEEE Access 2020*) which contains images from 450 different scenes captured using a Samsung NX3000 camera in three exposure levels (EV -2, 0, +2). This dataset contains images from indoor and outdoor environments as well as scenes containing landscapes, objects, and buildings. It also contains night scenes. Finally we used another dataset, HDR-Synth and HDR-Real from [9] (*CVPR 2020*) which contains real world LDR images and HDR counterparts (9785 in total) as well as synthetic pairs (around 500). We resized the images to 512×512 before feeding them into the model. We did an 80-20 split on the datasets for training and testing purposes.

B. Evaluation Metrics

We use three metrics to evaluate our method quantitatively. For evaluating on the basis of human level perception, we use the *High Dynamic Range Visual Differences Predictor* (HDR-VDP-2 or Q-score) [10], [11]. For structural similarity and accurate reconstruction evaluation, we use the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR). The HDR-VDP-2 score and SSIM are calculated on the actual GT and generated HDRs whereas the PSNR is calculated on the $\mu - law$ based tone-mapped GT and generated HDRs.

V. Results

A. Quantitative Evaluation

We evaluate our method ArtHDR-Net against three similar methods [8] (*GlobalSIP 2019*), [9] (*CVPR 2020*), and [12] (*WACV 2023*). We train and test the models on the datasets described in Section IV-A. Table I summarizes the quantitative results in terms of PSNR, SSIM, and HDR-VDP-2 (higher is better). The results confirm that our method is performing well compared to the considered state-of-the-art methods. In the case of City Scene dataset [5], ArtHDR-Net performs the best in terms of PSNR and HDR-VDP-2 scores and second best in terms of SSIM metric. On the other hand, [12] performs best in in terms of SSIM and second best in terms of PSNR scores, but the second best performer in terms of HDR-VDP-2 is [9]. For LDR-HDR pair dataset [19], once again ArtHDR-Net is best in terms of PSNR and HDR-VDP-2 scores. The method in [12] achieves top performance in terms of SSIM metric. Finally for the HDR-Synth and HDR-Real [9] dataset, ArtHDR-Net comes out best in terms of SSIM and HDR-VDP-2

[1] https://paperswithcode.com/method/vgg-loss

TABLE I: **Quantitative comparison** - PSNR, SSIM and HDR-VDP-2 attained by the proposed method and the state-of-the-art methods. The best results are in **bold**, while the second best results are underlined.

Method	City Scene dataset [5]			LDR-HDR pair dataset [19]			HDR-Synth and HDR-Real [9]		
	PSNR ↑	SSIM ↑	HDR-VDP-2 ↑	PSNR ↑	SSIM ↑	HDR-VDP-2 ↑	PSNR ↑	SSIM ↑	HDR-VDP-2 ↑
Method in [8] (2019)	32.5	0.90	67	15.2	0.66	64.5	17.1	0.71	66.7
Method in [9] (2020)	33.4	0.91	68.2	27.5	0.77	65.1	26	0.85	68
Method in [12] (2023)	34	**0.94**	68	29	**0.79**	66	**34**	0.87	67.5
Ours (ArtHDR-Net)	**35**	0.93	**69**	**30**	0.78	**66.5**	33.4	**0.88**	**68.3**

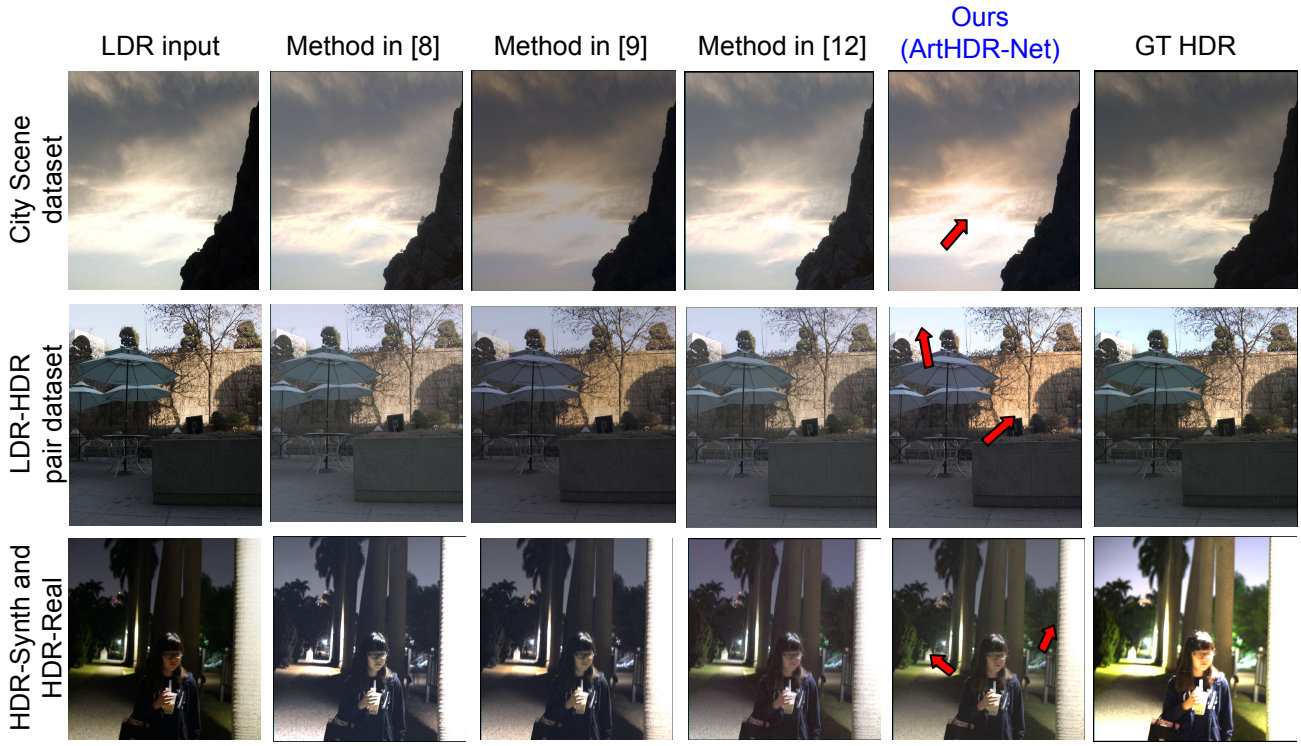


Fig. 3: **Qualitative evaluation** - Output HDR content produced by our method ArtHDR-Net and other state-of-the-art methods for the three datasets. The red arrows in our output highlight the naturalness in color and brightness, hence the artistic intent being preserved with respect to the input LDR image and the GT HDR image.

metrics whereas the method in [12] is best in terms of PSNR score. It is noteworthy that our method outperforms all the considered state-of-the-art methods and on all of the considered datasets in terms of HDR-VDP-2 score, which suggests that our method capable of reconstructing HDRs that capture the visual quality of real scenes and better represent all luminance conditions [10], [11]. The methods in [9], [12] are second best in this case. The structural similarity with respect to the input LDRs is better preserved by the method in [12].

B. Qualitative Comparison

Figure 3 depicts the qualitative comparison of our method with the considered conventional methods on the three datasets. The red arrows introduced to the images generated by our method show some sample regions which display naturalness in color and better luminance in comparison to the other methods. The methods in [8], [12] completely fail in this respect while the method in [9] gives similar results to our method, but it is more prone towards dark images. Hence, our method better represents the artistic intent of the images. Similarly, we also evaluate our method on arbitrary images retrieved from the Internet (see Figure 4). In the first row (under-exposed), we see that our method retains good color contrast and luminance compared to the other methods (yellow arrows). The methods in [8], [9] seem closer to ours but still lack in terms of luminance in some regions. In the second row (over-exposed), most methods [8], [12] fail to capture the details and artistic intent of the blue sky where the banding problem is also apparent. In contrast, whereas our method gives more natural and realistic output (red arrows). Although the method in [9] is able to give details of the blue sky to some extent, it still lacks in terms of luminance and color naturalness. The displayed GT as well as the generated HDRs from the various methods (including ours) have been tone-mapped using the algorithm in [28].

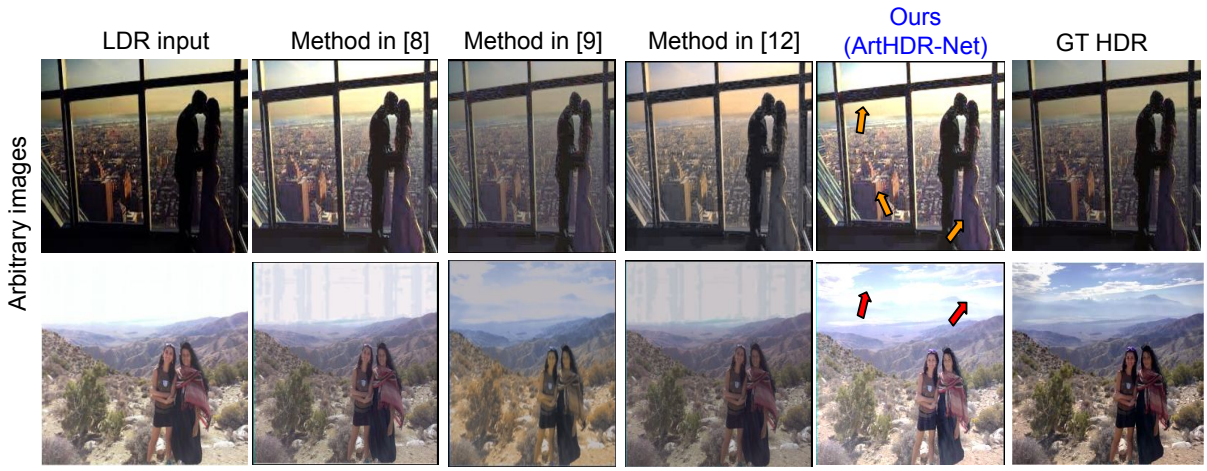


Fig. 4: **Qualitative comparison.** Comparison of our method with the considered state-of-the-art methods with respect to the input LDR and GT HDR (for images not in the mentioned datasets in section IV-A).

TABLE II: **Loss contribution** - The PSNR, SSIM, and HDR-VDP-2 achieved by using the different loss combinations. The best result is in **bold**, while the second best is underlined.

Loss	PSNR↑	SSIM↑	HDR-VDP-2 ↑
L_{l1}	36	0.931	63.5
L_{per}	35.4	0.928	66.3
$L_{l1} + L_{per}$	**36.5**	**0.942**	**67.1**

TABLE III: **Skip connections contribution** - The PSNR, SSIM, and HDR-VDP-2 achieved by using different combinations of features selected using skip connections to guide the network during the final stage in RU (see Figure 2). The best result is in **bold**, while the second best is underlined.

Fe_0^1	Fe_0^2	PSNR↑	SSIM↑	HDR-VDP-2 ↑
✗	✗	35.1	0.930	63.1
✗	✓	35.7	0.938	64.2
✓	✗	35.8	0.934	64
✓	✓	**36.5**	**0.942**	**67.1**

VI. Ablation Analysis

We experiment with the components and settings of our model to present the contributions made by of each of them in the achieved output. We create a subset of 5000 arbitrary images (collected from the three datasets described in Section IV-A) to perform the ablation experiments. Table II shows how each of the losses (see Section III-B) contributes to the final accuracy and quality of the HDR images reconstructed using our model. We see that L_{l1} contributes more towards better PSNR and SSIM values, whereas L_{per} have more impact towards the HDR-VDP-2 (Q-score). This suggests that the accuracy of the image reconstruction is mostly influenced by L_{l1}, whereas the perceptual quality is maintained by L_{per}. Finally, the combination of both losses leads to the best scores for all three metrics. Here, we also weight the losses with λ_1 and λ_2 (see Section III-B). Table III illustrates the contribution of each of the skip connections we use in our CNN architecture (see Figure 2). We observe that if we consider the skip connection Fe_0^1 only, the PSNR reaches better scores, whereas if we consider only Fe_0^2, SSIM and Q-score are better. However, if we consider both together, we get the best results for all three metrics. This implies that more low-level features improve the PSNR, while mid-level features boost up the SSIM as well as Q-score.

VII. Conclusions

In this paper, we attempt to improve the state-of-the-art HDR image generator with respect to the artistic quality of the generated HDRs from input LDRs. The quantitative results show the dominance of our method in terms of the HDR-VDP-2 which signifies that our method produces output that is visually more pleasing and appears more realistic than the other considered state-of-the-art methods. The qualitative results also support this argument. In addition, the PSNR and SSIM attained by our method are also comparable to the other state-of-the-art methods.

As future work, we intend to study the impact of multiple EV images (e.g., EV ±1, ±2, ±3 and ±4) on the feature map as well as the behaviour of the model on different EV inputs of the same image. Currently, we do not consider images with foreground motion or dynamic scenes [29]. We aim to upscale our method to videos [1], [20], [21] and dynamic scenes in the future [30]–[32]. Future work will also include the introduction of multiple tasks, such as denoising [33], deblurring, super-resolution [22], [26], and demosaicing into our framework. Finally, we plan to experiment with different loss functions and metrics [14], [15], [34], which will be more efficient in supervising the network in terms of visual quality and preservation of artistic intent.

Acknowledgement

This work is supported by Global Excellence and Mobility Scholarship (GEMS)[2], Monash University Malaysia.

[2]https://www.monash.edu.my/research/support-and-scholarships/gems-scholarship

REFERENCES

[1] X. Chen, Z. Zhang, J. S. Ren, L. Tian, Y. Qiao, and C. Dong, "A new journey from sdrtv to hdrtv," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4500–4509.

[2] N. Matsuda, Y. Zhao, A. Chapiro, C. Smith, and D. Lanman, "Hdr vr," in *ACM SIGGRAPH 2022 Emerging Technologies*, ser. SIGGRAPH '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3532721.3535566

[3] P. Satilmis and T. Bashford-Rogers, "Deep dynamic cloud lighting," *arXiv preprint arXiv:2304.09317*, 2023.

[4] X. Huang, Q. Zhang, Y. Feng, H. Li, X. Wang, and Q. Wang, "Hdr-nerf: High dynamic range neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 398–18 408.

[5] J. Zhang and J.-F. Lalonde, "Learning high dynamic range from outdoor panoramas," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4519–4528.

[6] P. Raipurkar, R. Pal, and S. Raman, "Hdr-cgan: single ldr to hdr image translation using conditional gan," in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, 2021, pp. 1–9.

[7] L. Wang and K.-J. Yoon, "Deep learning for hdr imaging: State-of-the-art and future trends," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8874–8895, 2021.

[8] Z. Khan, M. Khanna, and S. Raman, "Fhdr: Hdr image reconstruction from a single ldr image using feedback network," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[9] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image hdr reconstruction by learning to reverse the camera pipeline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1651–1660.

[10] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on graphics (TOG)*, vol. 30, no. 4, pp. 1–14, 2011.

[11] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 010 501–010 501, 2015.

[12] P.-H. Le, Q. Le, R. Nguyen, and B.-S. Hua, "Single-image hdr reconstruction by multi-exposure generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4063–4072.

[13] G. Luzardo, J. Aelterman, H. Luong, W. Philips, D. Ochoa, and S. Rousseaux, "Fully-automatic inverse tone mapping preserving the content creator's artistic intentions," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 199–203.

[14] J. Li and P. Fang, "Hdrnet: Single-image-based hdr reconstruction using channel attention cnn," in *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, 2019, pp. 119–124.

[15] M. S. Santos, T. I. Ren, and N. K. Kalantari, "Single image hdr reconstruction using a cnn with masked features and perceptual loss," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 80–1, 2020.

[16] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM transactions on graphics (TOG)*, vol. 36, no. 6, pp. 1–15, 2017.

[17] G. Luzardo, J. Aelterman, H. Luong, S. Rousseaux, D. Ochoa, and W. Philips, "Fully-automatic inverse tone mapping algorithm based on dynamic mid-level tone mapping," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e7, 2020.

[18] G. Cao, F. Zhou, K. Liu, A. Wang, and L. Fan, "A decoupled kernel prediction network guided by soft mask for single image hdr reconstruction," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–23, 2023.

[19] H. Jang, K. Bang, J. Jang, and D. Hwang, "Dynamic range expansion using cumulative histogram learning for high dynamic range image generation," *IEEE Access*, vol. 8, pp. 38 554–38 567, 2020.

[20] Y. Yang, J. Han, J. Liang, I. Sato, and B. Shi, "Learning event guided high dynamic range video reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 924–13 934.

[21] M. Anand, N. Harilal, C. Kumar, and S. Raman, "Hdrvideo-gan: deep generative hdr video reconstruction," in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, 2021, pp. 1–9.

[22] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3867–3876.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[24] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[25] T. Jinno, H. Kaida, X. Xue, N. Adami, and M. Okuda, "μ-law based hdr coding and its error analysis," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 94, no. 3, pp. 972–978, 2011.

[26] J. Johson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," Springer, Tech. Rep., 2016.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 661–670.

[29] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. Lau, "Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions," *IEEE Transactions on Image Processing*, vol. 30, pp. 3885–3896, 2021.

[30] N. K. Kalantari, R. Ramamoorthi *et al.*, "Deep high dynamic range imaging of dynamic scenes." *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144–1, 2017.

[31] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 117–132.

[32] Z. Pu, P. Guo, M. S. Asif, and Z. Ma, "Robust high dynamic range (hdr) imaging with complex motion and parallax," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[33] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, "Nerf in the dark: High dynamic range view synthesis from noisy raw images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 190–16 199.

[34] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

LSR++: An Efficient and Tiny Model for Image Super-Resolution

Wei Wang* Xuejing Lei* Yueru Chen† Ming-Siu Lee‡ and C.-C. Jay Kuo*
* University of Southern California, Los Angeles, CA, USA
† PengCheng Laboratory, Shenzhen, China
‡ National Taiwan University, Taipei, Taiwan

***Abstract*—A single-image lightweight super-resolution (LSR) method was presented in our earlier work. LSR predicts the residual image between the interpolated low-resolution (ILR) and high-resolution (HR) images using a self-supervised framework. A further enhanced version of LSR, called LSR++, is proposed in this work. The improvement of LSR++ over LSR primarily lies in a new patch alignment scheme, where the spatially geometrical property of patches is exploited. Patches can be categorized into two main classes through alignment. Then, one regressor is needed for each class. Compared with LSR, LSR++ lowers the computational complexity from 3.83K to 1.71K FLOPs/pixel and the model size from 770K to 200K parameters while preserving close visual performance measured by PSNR/SSIM. LSR++ offers an attractive super-resolution method for mobile applications.**

I. Introduction

Single image super-resolution (SISR) [1] targets recovering a high-resolution (HR) image from its low-resolution (LR) counterpart. It has many real-world applications, e.g., the digital zoom-in function in mobile phones. It has been intensively studied in the last two decades. Two popular approaches are taken to tackle this problem: 1) exemplar-based methods, and 2) deep-learning (DL) methods.

For the former, patches are conventionally presented by reference vectors (known as dictionary atoms) and partitioned into multiple groups. Mappings from LR-to-HR patches are developed for each individual group. Examples include [2]–[6]. Priors can be leveraged for patch partitioning [7]. This approach has several limitations. First, the LR-to-HR mapping is based on hand-crafted features. Second, the training of the LR-to-HR mapping is time-consuming [3], [4]. Third, the quality of obtained SR images is inferior to that achieved by DL methods.

For the latter, SRCNN [8] is among the earliest ones. Substantial advances have been made since then. Examples of DL methods include [9]–[17]. Most efforts ([8]–[13]) have focused on achieving higher visual quality (namely, better PSNR and SSIM). Implementational efficiency has been explored in recent years, e.g., [14], [15]. Other SISR problems have also been examined such as unknown degradation kernels [16] and magnification with a non-integer factor [17], etc.

Although DL methods offer significant performance breakthroughs, it is challenging to implement them in mobile/edge devices due to their high computational cost and memory requirement. Implementing lightweight DL methods on mobile devices has received attention from academia and industry. Some successful results are reported in [18]–[20]. They are accomplished through better neural network architectural search and pruning but their designs appear to be tedious and ad hoc. Furthermore, all DL methods share one common shortcoming, namely, lack of mathematical transparency.

Recently, based on the green learning paradigm [21]–[25], a lightweight super-resolution (LSR) method was developed in [26] and it is reviewed in Sec. II. Patch augmentation with a flipping and/or a rotation is commonly adopted by SISR methods to enlarge the training dataset. It contributes to better quality at the expense of a higher training cost. LSR adopts patch augmentation not only in the training stage but also in the inference stage to boost the regression performance of a target pixel. That is, LSR augments its corresponding patch through flipping and rotating, conducts regression for each patch variant, and assembles the estimates by averaging.

A further enhanced version of LSR, called LSR++, is proposed in this work. The improvement of LSR++ over LSR lies in the replacement of patch augmentation with a new patch alignment scheme. Intuitively speaking, we flip and rotate a patch so that it is aligned along a certain direction. By considering the patch's symmetry in the spatial domain, we align patches into two main classes. Then, one LR-to-HR regressor is developed for each class. The patch alignment scheme reduces training and inference complexities and the model size, which is the main contribution of this paper. It is worthwhile to emphasize that the patch alignment scheme is not only applicable to the SISR problem but also other applications that need patch-wise data augmentation. It is a powerful tool in visual-based machine learning.

We conduct experiments to demonstrate the advantages of LSR++. Compared with LSR, LSR++ lowers the computational complexity from 3.83K to 1.71K floating-point operations per pixel (FLOPs/pixel) and reduces the number of model parameters from 770K to 200K. At the same time, LSR++ offers quality performance close to LSR in terms of PSNR and SSIM metrics. Thus, LSR++ offers an even more attractive lightweight SR method for mobile applications.

The rest of the paper is organized as follows. A brief review of the LSR method is provided in Sec. II. The LSR++ method is proposed in Sec. III. Experimental results are shown in Sec. IV. Finally, concluding remarks are given in Sec. V.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Review of LSR Method

A lightweight super-resolution (LSR) method based on green learning [24] was introduced in [26]. It follows the conventional SISR pre-processing routine: 1) transforming RGB images to the YCbCr color space and only the luminance (Y) component is processed, and 2) applying the interpolation (Lanczos) to LR images to yield interpolated LR (ILR) images. LSR contains the following three modules in cascade.

- Module 1: unsupervised representation learning
  The objective of the first module is to generate a rich and diversified set of representations in the neighborhood of a target pixel. Five types of representations are generated.
- Module 2: supervised feature learning
  The total number of representations obtained from Module 1 is around 1500. A subset of the most relevant representations can be selected based on a mechanism called the Relevant Feature Test (RFT) [25].
- Module 3: supervised decision learning
  It predicts the residual of the target pixel based on selected features through the XGBoost regressor. In the training stage, data augmentation is performed (via 90-degree rotations and flipping) to enlarge the training sample size. Furthermore, "clustering" and "prediction fusion" are adopted to boost the regression performance via ensemble learning.

There are two versions of LSR, called LSR V1 and LSR V2, respectively. All five types of representations are used in LSR V1 while only one specific type of representations is used in LSR V2. On the one hand, LSR V1 has a larger model size and a higher computational complexity than LSR V2. On the other hand, LSR V1 has better SR quality (measured by PSNR/SSIM) than LSR V2. Nevertheless, their performance differences are minor. Compared with other SISR methods, LSR V1 and V2 offer better SR quality than exemplar-based methods and comparable performance with the entry-level DL solution, SRCNN. They demand much less computational complexity measured by the floating-point operations per pixel (FLOPs/pixel).

III. Proposed LSR++ Method

A. Motivation

LSR++ inherits the same system block diagram of LSR. It follows the same high-level processing strategy: 1) patch-to-pixel residual regression, and 2) adaptive regression based on partitioned patches. It adopts the same SISR pre-processing routine and contains three processing modules in cascade (i.e., unsupervised representation learning, supervised feature learning, and supervised decision learning) as described in Sec. II.

The main difference between LSR and LSR++ lies in one aspect. That is, LSR++ employs a patch alignment scheme to achieve lower computational complexity and a smaller model size, which is the main contribution of this paper. The amount and diversity of training samples are highly related to the performance of SR methods. It is typical to use data augmentation to increase the number of training samples. For example, given N training patches, one can obtain $8N$ training patches by flipping and rotating each of them. This technique is adopted by LSR. The training complexity increases with data augmentation. Furthermore, an ensemble technique is used by LSR to increase the performance in the inference stage. That is, an input patch is flipped and rotated to get multiple variant patches in testing. The regression for the central pixel residual is conducted independently for each variant, and the predicted values of all variants are averaged to obtain the ultimate prediction.

The data augmentation technique was exploited by LSR to trade training/testing complexities for higher visual quality. Here, we propose a new method, LSR++, to preserve SR performance without data augmentation. It has the benefits of a smaller model size and a lower computational complexity.

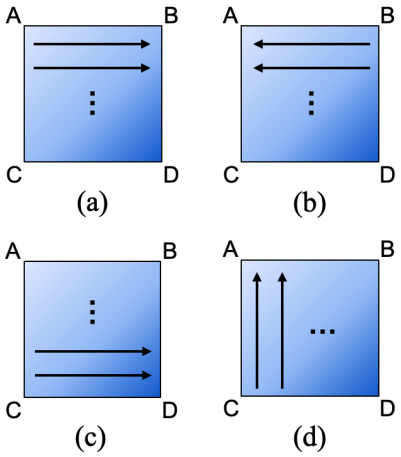


Fig. 1. Pixel access patterns on the original patch for equivalent augmented patch obtained by a) no operation, b) left-right flipping, c) top-bottom flipping, and d) clock-wise 90-degree rotation.

B. Patch Alignment

1) Implementation of Patch Flipping and Rotations: Flipping and rotating are standard ways to augment patches. They can be easily implemented by different pixel access patterns as shown in Fig. 1. "A", "B", "C", and "D" denote four corners of a patch in this figure. We can design various pixel access patterns as indicated by the arrow direction to attain an equivalent effect of data augmentation. Without loss of generality, we show adaptive pixel scanning order on the original patch to equivalently achieve three representative augmented patches via left-right flipping, top-bottom flipping, and clock-wise 90-degree rotation in Fig. 1 (b), (c), and (d), respectively. For example, if pixels are scanned from right to left in a row-wise fashion as shown in Fig. 1 (b), it is equivalent to the left-right flipping. We propose a patch alignment scheme through the adaptive pixel access patterns to achieve the same effect as patch augmentation yet without increasing the time or space consumption.

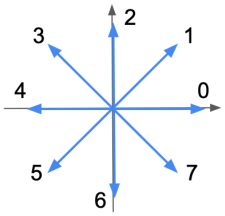


Fig. 2. Visualization of the eight reference directions for the patch gradient histogram calculation.

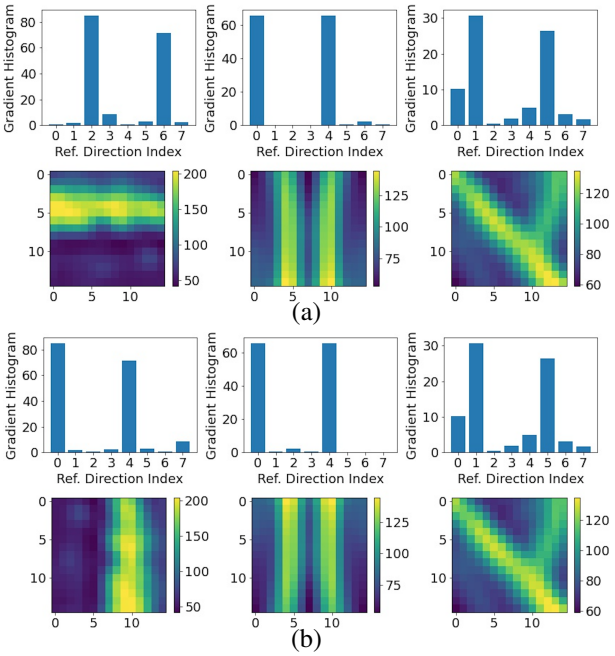


Fig. 3. Visualization of three exemplary patches (a) before and (b) after patch alignment.

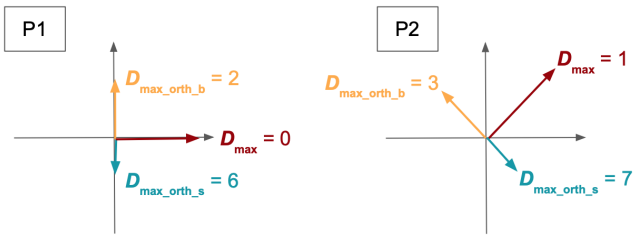


Fig. 4. Two standard templates for patch alignment: the template for P1 patches (left) and the template for P2 patches (right).

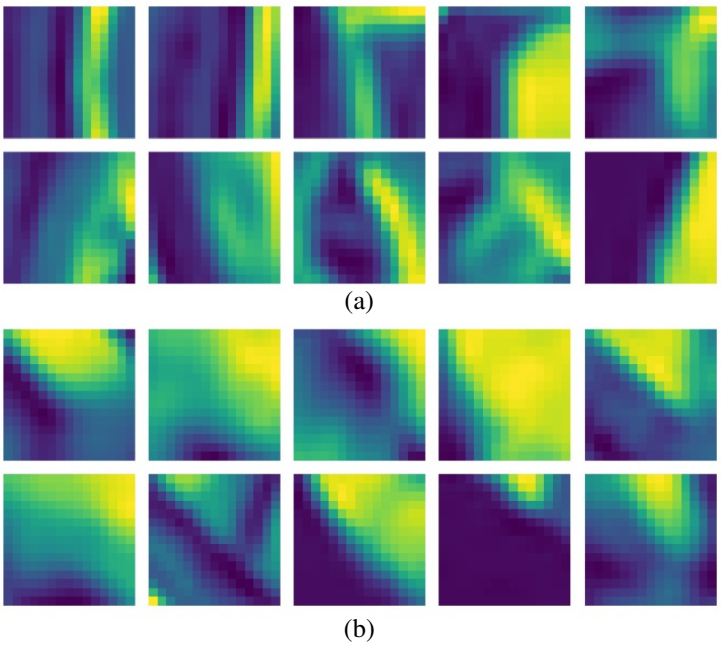


Fig. 5. Exemplary patches after patch alignment from (a) P1 patches and (b) P2 patches.

2) Patch Gradient Histogram Vector: LSR++ utilizes patch gradient histograms to achieve patch alignment. For an ILR patch of size 15×15, the horizontal and vertical gradients, G_x and G_y, are calculated by using horizontal and vertical Sobel filters of size 3×3 at each pixel. Afterwards, the gradient magnitude, r, and the gradient orientation angle, θ, are derived based on G_x and G_y. We further quantize the orientation angles into eight bins with their centers pointing to $45k$ degree, where $k = 0, 1, \cdots 7$, as the reference directions shown in Fig. 2. Given 13×13 interior points, the gradient magnitude r of pixels in their associated bins are accumulated to form an 8-D patch gradient histogram vector.

Each vector presents the quantized edge orientation statistics of a patch along these eight reference directions. Denote the gradient histogram vector dimension index by D_{old} and D_{new} for a certain gradient histogram bar before and after the rotation or flipping operation. These vectors have four properties:

- Patch gradient histogram vectors capture local edge orientation patterns. Small magnitudes in all eight dimensions indicates smooth patches. One or two larger magnitudes imply the existence of dominating edges. Otherwise, the associated patches are in a textured region.
- Patch rotation and flipping operations preserve the height and cyclic adjacency of patch gradient histogram bars as shown in Fig. 3.
- Rotation operation cyclically shifts the gradient histogram bars such that $D_{new} = (D_{old} - m)\ mod\ 8$, where m is the multiple of 45 degrees to clock-wise rotation.
- Up-down Flipping operation symmetrically shifts the gradient histogram bars such that $D_{new} = (8 - D_{old})\ mod\ 8$. The Left-right flipping can be obtained by up-down flipping followed by a 180-degree rotation.

3) Patch Alignment Algorithm: In general, patch alignment operation is performed via patch flipping and patch rotations by clock-wise 90, 180, or 270 degrees, so that patches are aligned to standard reference templates. Let D_{max} denote the dimension of the patch gradient histogram vector that has the

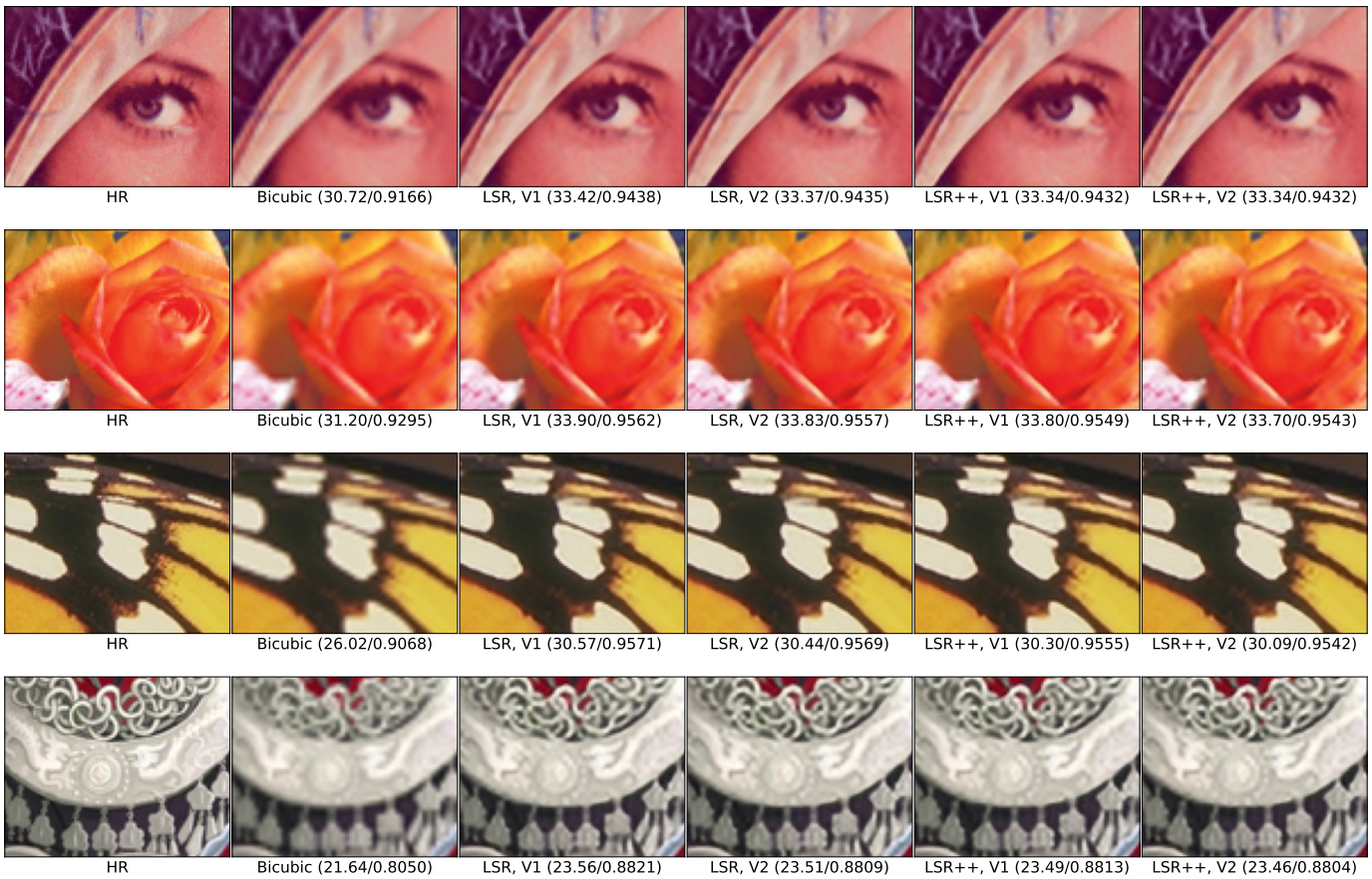


Fig. 6. Visual comparison of exemplary test SR images, where the PSNR/SSIM results are provided in brackets.

largest magnitude, and $D_{max_orth_b}$ and $D_{max_orth_s}$ represent its two orthogonal dimensions which have bigger and smaller magnitudes, respectively. Patches are first partitioned into two portions based on their D_{max}:

- Portion 1 (P1): samples with D_{max} on even indices, and
- Portion 2 (P2): samples with D_{max} on odd indices.

It is worth mentioning that patch alignment operation does not change the patch portion type. We then execute the patch alignment operation on each patch to match one of the standard reference templates in Fig. 4. Patches belonging to P1 and P2 use the left and right templates, respectively.

Patch gradient histogram properties described in Sec. III-B2 are critical to the complexity and efficiency of patch alignment. Associated re-indexing of patch gradient histogram dimensions effectively generates the updated gradient histogram after each single rotation/flipping operation. Therefore, instead of exhaustive attempts of the combination of rotation/flipping operations, one sample can be systematically aligned by the proper rotation/flipping operation sequence based on these patch gradient histogram properties. Consequently, patch alignment requires no additional complexity cost, and its high implementation efficiency is surely guaranteed.

4) Geometrical Interpretation: Geometrically, P1 and P2 patches are samples with significant gradient values along horizontal/vertical and diagonal/off-diagonal directions, respectively. Some observations are summarized as follows.

- The D_{max} direction characterizes the gradient information orthogonal to the dominant edge pattern. This is illustrated in Fig. 3 and Fig. 5.
- $D_{max_orth_b}$ and $D_{max_orth_s}$ characterize the gradient information along the dominating edges.
- The two templates shown in Fig. 4 constrain P1 and P2 samples to be aligned with vertical edges and northwest-to-southeast edges, respectively.

Through rotation and flipping operations on raw patches to satisfy the two standard reference templates, patch alignment productively summarizes the diversity of patch patterns, and therefore effectively reduces the geometrical complexity of raw patches. Accordingly, patch alignment is equivalent to data augmentation concerning the generic representation of patch patterns, while it exhibits more advanced complexity control simultaneously.

IV. Experiments

A. Experimental Setup

Only the Y channel is processed for super resolution quality evaluation. The Cb and Cr channels are obtained by the bicubic interpolation for final visualization. This is the standardized

TABLE I
COMPARISON OF AVERAGED PSNR/SSIM VALUES ON DATASETS SET5, SET14, B100 AND URBAN100, WHERE RED AND BLUE COLORS INDICATE THE BEST AND THE SECOND BEST PERFORMANCE FOR EACH DATASET.

PSNR/SSIM	SelfExSR	A+	SRCNN	LSR, V1	LSR, V2	LSR++, V1	LSR++, V2
Set5	36.49 / 0.9537	36.54 / 0.9544	**36.66** / 0.9542	**36.57** / **0.9595**	36.50 / **0.9593**	36.42 / 0.9589	36.28 / 0.9583
Set14	32.22 / 0.9034	32.28 / 0.9056	**32.42** / 0.9063	**32.30** / **0.9159**	32.23 / **0.9155**	32.24 / 0.9152	32.11 / 0.9142
BSD100	31.18 / 0.8855	31.21 / 0.8861	**31.36** / 0.8879	**31.37** / **0.8985**	31.32 / **0.8980**	31.31 / 0.8978	31.23 / 0.8968
Urban100	**29.54** / 0.8967	29.20 / 0.8938	29.50 / 0.8946	**29.51** / **0.9024**	29.42 / **0.9013**	29.37 / 0.9006	29.18 / 0.8978

TABLE II
COMPARISON OF INFERENCE COMPUTATIONAL COMPLEXITY (FLOPS/PIXEL) AND MODEL SIZES OF SEVERAL BENCHMARKING SR METHODS.

	FLOPs / pixel	Model Size
A+	15.7K(9.24X)	1.06M (18.59X)
SRCNN	114K (67.06X)	57.3K (1X)
VDSR	1.33M (782.35X)	665K (11.60X)
LSR, V1	9.28K (5.46X)	774K (13.51X)
LSR, V2	3.83K (2.25X)	770K (13.45X)
LSR++, V1	4.43K (2.61X)	207K (3.61X)
LSR++, V2	1.71K (1X)	200K (3.49X)

experimental protocol used by all benchmarking SR papers. The results with scaling factor of 2 are demonstrated here. The BSD200 dataset [27] is adopted to train the proposed LSR++ model. The tests are performed on four datasets: Set5 [28], Set14 [29], BSD100 [27], and Urban100 [5].

Our model uses ILR patches of size 15×15 to predict the residual value of the pixel at the patch center. LSR partitions data into hard and easy samples. There is no data augmentation, no clustering, or fusion strategy for easy samples. LSR++ follows the same practice. For hard data, LSR++ adopts patch alignment to replace data augmentation in the training stage and fusion in the inference stage. Representation filters and XGBoost regressor (500 trees with the maximum depth equal 6) are learned for each of the two portions of hard data after patch alignment.

The corresponding enhancements for LSR V1 and LSR V2 are called LSR++ V1 and LSR++ V2, respectively. As described in Sec. II, the differences between V1 and V2 lie in the usage of different representations.

B. Quality Performance Evaluation

Several benchmarking SR methods listed below are considered for quality comparison.

- Exemplar Methods: SelfExSR [5], A+ [4]
- Neural Network Method: SRCNN [8]
- Green Learning Methods: LSR V1/V2 and LSR++ V1/V2

The PSRN/SSIM quality values are demonstrated in Table I, where the best and the second best are marked in red and blue, respectively. Overall speaking, their PSNR/SSIM values are similar. Note that SRCNN is the simplest DL-based method. Here we do not include advanced DL-based solutions in the table since their model sizes and computational complexity are too high to be used on mobile/edge platforms.

The averaged PSNR and SSIM degradations over four testing datasets from LSR V1 to LSR++ V1 are only 0.1dB and 0.0009, respectively. The averaged PSNR and SSIM degradations from LSR V2 to LSR++ V2 are 0.17dB and 0.0018, respectively. Such minor differences are difficult for human eyes to tell. We show four exemplary images in each row of Fig 6. The images from left to right are: HR, bicubic interpolated LR, LSR V1, LSR V2, LSR++ V1, and LSR++ V2. The quality of bicubic interpolated LR is clearly the worst. The visual quality of SR images obtained by LSR V1, LSR V2, LSR++ V1, and LSR++ V2 is difficult to differentiate.

C. Complexity Analysis

1) Training Phase: As mentioned earlier, patch alignment is equivalent to data augmentation for raw P1 and P2 patches satisfying the two standard templates. No additional data augmentation is needed. The total training time is reduced by a factor of eight. The rotation and flip can be easily implemented by various pixel access patterns as explained in Sec. III-B1.

2) Inference Complexity: One important step in LSR is the fusion operation in the inference stage. For an input patch, we generate multiple variants by rotations and flips, and ensemble the predictions from all candidates (the original input plus other variants) to yield the final estimate. Both variant generation and fusion operations are eliminated by using patch alignment in LSR++, thus leading to a faster inference speed. The inference computational complexity comparison of several super-resolution algorithms in terms of FLOPs/pixel is given in Table II. LSR++ exhibits much lower inference FLOPs/pixel values against all benchmarking methods. The detailed computation of inference complexities of LSR V1/V2 and LSR++ V1/V2 are presented in Table III in Appendix.

3) Model Size: LSR partitions patches into eight clusters aiming at grouping them with eight orientations separated by every 45 degrees. LSR++ only partitions patches into two portions. In general, LSR++ reduces its model parameters (i.e. model size) to one quarter of LSR. The comparison of number of model size with other representative SISR methods is shown in Table II. The model size of LSR++ is smaller than all benchmarking methods except for SRCNN. However, the computational complexity of SRCNN is much higher. Detailed analysis on the model sizes of LSR V1/V2 and LSR++ V1/V2 are provided in Table III in Appendix.

V. CONCLUSION AND FUTURE WORK

An enhanced version of the LSR method, called LSR++, was proposed in this work. The improvement of LSR++ over LSR lies in a new patch alignment scheme, where the geometrical property of patches in the spatial domain is exploited. Compared with LSR, LSR++ lowers the computational

complexity from 3.83K to 1.71K FLOPs/pixel and the model size from 770K to 200K parameters. At the same time, it preserves similar SR quality. LSR++ provides an attractive super-resolution method for mobile and edge computing. The patch alignment proposed in LSR++ can serve as a generic regularization process on patch diversity, which can be applied in all patch-based algorithms for the purpose of complexity control.

As to future extensions, we would like to reduce the model size furthermore. This could be achieved by replacing the XGBoost regressors with linear least-squared regression. Another possible direction is to apply LSR++ iteratively for further quality improvement. Since LSR++ has a lower computational complexity, it is acceptable to apply it more times for better visual quality.

APPENDIX

A detailed computation of the inference complexity (in terms of FLOPs/pixel) and model sizes (in terms of model parameter numbers) for LSR (V1/V2) and LSR++ (V1/V2) is summarzied in Table III. The reason for larger model parameters in Modules 1&2 of LSR++ is due to the portion-wise representation learning, while global representation filters are shared in all clusters in LSR.

TABLE III
COMPARISON OF INFERENCE COMPLEXITY (FLOPS/PIXEL) AND MODEL SIZES FOR LSR (V1/V2) AND LSR++ (V1/V2).

LSR V1/V2 Summary

Complexity	FLOPs / pixel		Model Size	
Data Type	Easy	Hard	Easy	Hard
Module 1 Sub-total	153 / 153	13.0k / 612	81 / 81	3.3k / 162
Module 2 Sub-total	0 / 0	0 / 0	105 / 105	374 / 135
Module 3 Sub-total	300 / 300	7.5k / 7.5k	9.5k / 9.5k	760.3k / 760.3k
Post-process	1 / 1	1 / 1	0 / 0	0 / 0
Sub-Total	454 / 454	20.5k / 8.1k	9.7k / 9.7k	763.9k / 760.6k
Combination Weight	0.56	0.44	1	1
Total	9.3k / 3.8k		773.6k / 770.2k	

LSR++ V1/V2 Summary

Complexity	FLOPs / pixel		Model Size	
Data Type	Easy	Hard	Easy	Hard
Module 1 Sub-total	153 / 153	6.5k / 306	81 / 81	6.4k / 162
Module 2 Sub-total	0 / 0	0 / 0	105 / 105	748 / 270
Module 3 Sub-total	300 / 300	3k / 3k	9.5k / 9.5k	190.0k / 190.0k
Post-process	1 / 1	1 / 1	0 / 0	0 / 0
Sub-Total	454 / 454	9.5k / 3.3k	9.7k / 9.7k	197.2k / 190.4k
Combination Weight	0.56	0.44	1	1
Total	4.43k / 1.71k		206.8k / 200.1k	

ACKNOWLEDGMENT

The authors acknowledge the gift support from MediaTek Inc. as well as the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication. URL: https://carc.usc.edu.

REFERENCES

[1] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[2] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, IEEE, vol. 1, 2004, pp. I–I.

[3] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

[4] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian conference on computer vision*, Springer, 2014, pp. 111–126.

[5] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[6] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3791–3799.

[7] P. Sandeep and T. Jacob, "Single image super-resolution using a joint gmm method," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4233–4244, 2016.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[9] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[10] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[11] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 11 065–11 074.

[12] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5690–5699.

[13] H. Chen, Y. Wang, T. Guo, *et al.*, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[14] X. Kong, H. Zhao, Y. Qiao, and C. Dong, "Classsr: A general framework to accelerate super-resolution networks by data characteristic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 016–12 025.

[15] L. Wang, X. Dong, Y. Wang, *et al.*, "Exploring sparsity in image super-resolution for efficient inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4917–4926.

[16] R. Zhou and S. Susstrunk, "Kernel modeling super-resolution on real low-resolution images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2433–2443.

[17] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1575–1584.

[18] H. Wang, V. Bhaskara, A. Levinshtein, S. Tsogkas, and A. Jepson, "Efficient super-resolution using mobilenetv3," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, Springer, 2020, pp. 87–102.

[19] Y. Nie, K. Han, Z. Liu, C. Liu, and Y. Wang, "Ghostsr: Learning ghost features for efficient image super-resolution," *arXiv preprint arXiv:2101.08525*, 2021.

[20] Z. Zhan *et al.*, "Achieving on-mobile real-time super-resolution with neural architecture and pruning search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4821–4831.

[21] C.-C. J. Kuo, "Understanding convolutional neural networks with a mathematical model," *Journal of Visual Communication and Image Representation*, vol. 41, pp. 406–413, 2016.

[22] C.-C. J. Kuo, "The cnn as a guided multilayer recos transform [lecture notes]," *IEEE signal processing magazine*, vol. 34, no. 3, pp. 81–89, 2017.

[23] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable convolutional neural networks via feedforward design," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 346–359, 2019.

[24] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook," *arXiv preprint arXiv:2210.00965*, 2022.

[25] Y. Yang, W. Wang, H. Fu, and C.-C. J. Kuo, "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[26] W. Wang, X. Lei, Y. Chen, M.-S. Lee, and C.-C. J. Kuo, "LSR: A light-weight super-resolution method," *arXiv preprint arXiv:2302.13596*, 2023.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, IEEE, vol. 2, 2001, pp. 416–423.

[28] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[29] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, Springer, 2012, pp. 711–730.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

High-Quality Font Generation Based on StyleGAN2 and FSFont Font Generation Model

Yuki SHIMAMURA* and Michiharu NIIMI†
* Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka-shi, Fukuoka, 820-8502 Japan
E-mail: shimamura.yuki703@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology,
680-4, Kawazu, Iizuka-shi, Fukuoka, 820-8502 Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—This paper proposes a high-quality font generation method based on StyleGAN2 and FSFont font generation model. In our previous work, we proposed a system that expresses emotions by using font changes and call it an EmoEditor (Emotion-Expressive Editor). This system uses a deep neural network based on zi2zi that is able to generate a desired rendered character by giving a fixed rendered character and a font label which distinguishes between fonts. By changing the font label, the system generates the different font characters which corresponds to several emotions. However, the system was not able to reconstruct the expected ones for our collected fonts that is needed to express emotions. Considering the rapid development of DNN technology, in this paper, we propose an extension of the StyleGAN2 model for font generation. Specifically, we enhance StyleGAN2 by incorporating an Encoder which is come from FSFont generation model and a class label input structure. By adopting a class label input, the proposed model is able to generate indicated font by users. Experimental results show that the proposed method can generate better fonts qualitatively and quantitatively than previous studies on font generation.**

I. Introduction

Text is used in various situations as a means of conveying information. One of its applications is communication. Text is used not only for communication with close friends, but also for communication with people far away, such as in the comments sections of various websites and social networking services. On the other hand, it is difficult to convey emotions over text. To address this issue, we have proposed a system that conveys emotions through font changes, which is called an EmoEditor[1]. EmoEditor uses a deep neural network (DNN) that generates fonts called zi2zi [2]. Zi2zi is a model that takes as input two data, a character image called the source font and a class label, and produces as output a character image in which the source font is changed to a font corresponding to the class label. By mapping emotions to class labels, EmoEditor changes the font based on the input emotion. The EmoEditor system can also interpolate between two fonts by using the zi2zi function. This allows users to employ emotion-expressive fonts in text, enabling richer communication compared to the use of uniform fonts. However, the zi2zi structure is not sufficient to generate fonts that express emotions, and the expected reconstruction of fonts cannot be achieved.

In recent years, significant research has been conducted on font generation. Tang proposed FSFont font generation model[3] (referred to as FSFont), which utilizes two images - a content image and a reference image - as input. Each image is processed by two different Encoders. The Content Encoder extracts the textual meaning of the font as features from the content image, while the Reference Encoder extracts font style information as features from the reference image. Considering that fonts consist of various partial elements like radicals and strokes, Tang introduced a mechanism that employs Cross Attention in the DNN to learn the partial styles of the reference image and the spatial correspondence of the components in both the content and reference images. Yangchen proposed DGFont font generation model[4](referred to as DGFont) which is based on the same concept as [3], where the content and reference images are fed into the Content Encoder and Reference Encoder, respectively. Acknowledging that the shape of the lines and strokes that form a character varies depending on the font, Yangchen adopted Deformable-Convolution [5] for the network. Deformable-Convolution allows for geometric deformation of strokes. Chi proposed CFFont font generation model[6](referred to as CFFont) which is based on DGFont. This approach takes multiple types of fonts as input content images and merges them by assigning weights to the obtained features. The loss function used in this method is not the L1 Loss between each pixel in the generated image and the correct image. Instead, it employs the L1 Loss of the histogram of the projected image in one dimension across six different directions. This enables learning that focuses on the overall shape of the characters.

In this paper, we propose a method to generate high-quality fonts based on both StyleGAN2[7] and FSFont. Our method incorporates the Content Encoder in FSFont into the StyleGAN2 model, where the output of the Encoder serves as the initial input to the Generator in StyleGAN2. In addition, we assign class label for the font which is used in training and testing, and we give this label to the Mapping network in StyleGAN2. In StyleGAN2, the generated image depends on the noise giving to Mapping network. This means that you need to search the initial noise to generate the image you want to.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Therefore, the proposed our model is able to generate indicated fonts with giving adequate class label which is regarded as a noise without searching initial noise.

The contributions of this paper are two as follows.

- We develop a method for addressing the font generation problem by utilizing the powerful StyleGAN2 model known for its ability to generate high-quality images. Through our proposed approach, we have successfully achieved high-quality font generation by leveraging class labels.
- We don't use L1 Loss as the primary loss function. To do so, we can realize the robustness of the learning process to potential misalignment issues that may arise during the reconstruction of fonts.

In section 2, we describe the DNN proposed in this paper, and in section 3, we show the conditions and results of experiments in which fonts are trained. Finally, in section 4, we summarize the results and discuss future issues.

II. Proposed Method

The proposed Generator architecture consists of two main components: an Encoder and a Decoder. In our approach, we adopt the Encoder used in the FSFont as the Encoder component, while the Decoder component is based on the StyleGAN2 Generator. The Encoder plays an important role in capturing the content and meaning of the characters in the font to be generated. It takes the content character image as input and extracts the relevant features that represent the desired font character. These extracted features are then passed to the Decoder for further processing. On the other hand, the Decoder receives the output of the Encoder, which encodes the desired font style, along with a class label. The class label serves as a guiding signal to instruct the Generator on the specific font style to be generated. By combining the encoded features from the Encoder and the class label, the Decoder synthesizes an output image where the font of the content character image is transformed to match the corresponding font style dictated by the class label.

In this section, we provide a detailed explanation of both the FSFont model and the StyleGAN2 model, which form the foundational components of our proposed method. We then proceed to describe the integration of these models in our novel architecture, showcasing how they synergistically contribute to the generation of high-quality fonts with desired styles.

A. Useful two DNN models for font generation

1) StyleGAN2: StyleGAN2 is a type of Generative Adversarial Network (GAN) that has gained significant attention for its ability to generate high-quality images with controlled attributes. One of its distinguishing features is its attribute control, allowing for fine-grained manipulation of generated images. The network structure of StyleGAN2 is depicted in Fig.1(a), illustrating the flow of information within the model. Firstly, StyleGAN2 takes a noise vector, denoted as z, as input. The noise vector z is indeed mapped to the latent variable w through the Mapping Network in StyleGAN2. The Synthesis Network initially receives a 4x4 constant tensor as the first input（c_1）, which is then progressively up-sampled. During the up-sampling process, a latent variable called w is introduced, and it plays a crucial role in controlling the attributes of the generated image. By modifying the value of w during the up-sampling stages, the Synthesis Network produces an image that exhibits desired attribute variations based on the provided w.

2) FSFont font generation model: FSFont is a type of font generation model whose network structure is depicted in Fig.1(b). The network takes two types of images as input. The first type is the content image, which influences the output font character based on the characters present in the image.The second type of image is the reference image, which determines the output font style based on the style present in the image. One unique feature of this model is its utilization of Cross Attention to learn the spatial correspondence between the components of the content image and the reference image.

Specifically, Cross Attention is employed in the Style Aggregation Module, where the features of the content image serve as Query and the features of the reference image serve as Key and Value. Initially, the Query and Key are utilized to assess the similarity between the content and reference. For instance, when "盟" is used as the Query and "是" as the Key, the upper left portion "日" of "盟" is mapped to the upper part of "是," resulting in a matrix that emphasizes the pixels in the upper part of "是."

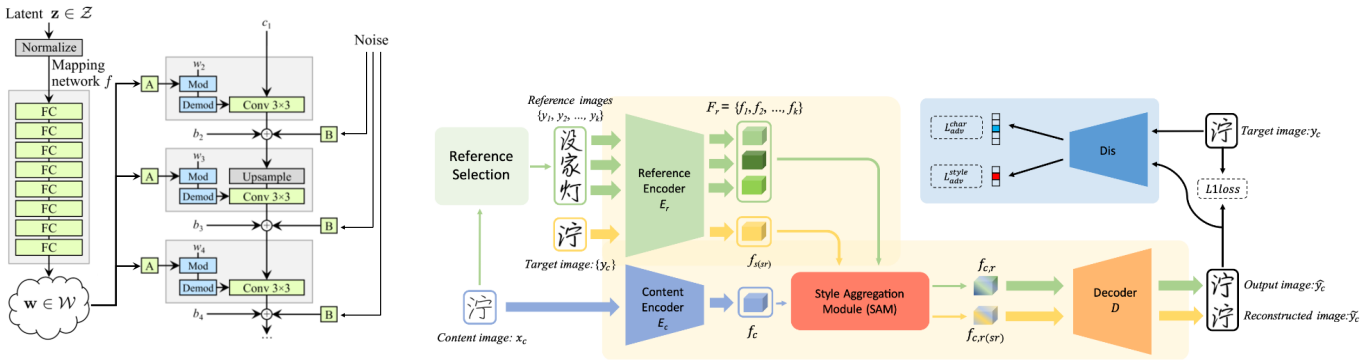
Once this matrix is obtained, the style of the part that is mapped to the Query is extracted from the corresponding Value. By extracting features from the reference image, specifically focusing on the style of the parts that exhibit similarity to the content, FSFont was able to generate a higher quality font compared to conventional methods.

B. Proposed Model

StyleGAN2 is one of powerful methods to generate high-quality images. However, to achieve the desired font character and style output, a pre-search for the character and corresponding latent variable is necessary. In the case of FSFont, in order to generate a font based on a specific content character, it is essential to pre-search for a reference character that corresponds to the content character and include it in the font to be generated.

The network architecture proposed in this paper is depicted in Fig.2. The Content Encoder used in FSFont is employed as the Encoder since it adequately reconstructs characters. In StyleGAN2, the input to the Synthesis Network was a constant value, but in this case, an Encoder-Decoder model structure was implemented by feeding the output of the Encoder into the input of the Synthesis Network.

The Mapping Network takes in noise that corresponds to class labels instead of random noise. This allows for the production of the intended font output. Additionally, we incorporated a StyleGAN2-ADA structure (Aug)[8] into our model, which aids in mitigating over-fitting by applying image augmentations to the inputs fed into the Discriminator. The



(a) The network structure of StyleGAN2 [7]

(b) The network structure of FSFont [3]

Fig. 1: Two DNN models we focused on

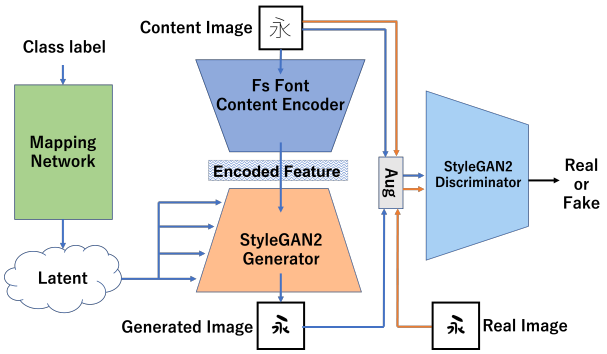


Fig. 2: Proposed Model

Discriminator receives either a combination of content images and a generator image (depicted by the blue line flow in Fig.2) or a combination of content images and a correct image (represented by the orange line flow in Fig. 2).

C. Loss Function

The loss functions utilized in this approach consist of the loss function from StyleGAN2 and the Content Consistent Loss from DGFont. The Content Consistent Loss is represented by equation (1).

$$L_{cnt} = \|f_c(I_c) - f_c(G(I_c, l))\|_1 \tag{1}$$

The f_c represents the Content Encoder, the I_c denotes the content image, the G indicates the Generator, and the l represents the class label. The equation (2) shows the entire loss function.

$$L = L_{StyleGAN2} + \lambda_{cnt} L_{cnt} \tag{2}$$

The λ_{cnt} denotes the weight of the Content consistent loss.

III. Experiments

In this section, we evaluate the model's performance in generating Japanese fonts. Firstly, we describe the training conditions, including the data set and hyper-parameters. Then, we report the training results in order of qualitative evaluation, quantitative evaluation, and font interpolation results.

A. training conditions

In this experiment, a total of ten fonts were used, consisting of one content font and nine style fonts. The font class labels and names, images are shown in Table I. For training data, a total of 2,284 hiragana, katakana, kanji, and alphabet characters were generated for each font. These images were then paired with the content font. Other training conditions were the same as those used in StyleGAN2. This includes a batch size of 32, Adam optimization method with beta1=0 and beta2=0.99, and a learning rate of 0.0025. The data set uses 128×128 gray-scale images. The output size of Content Encoder is 16×16, therefore, the StyleGAN2 network starts at 16×16.

B. Results

In order to compare our model to previous method, we selected two font generation models : zi2zi and DGFont. This is the reason why it is easy for not only our model but also these models to prepare the required data to train and test.

1) Qualitative evaluation: The comparisons of the generated fonts are shown in Table II.

From Table II(a), label 1 has blurred spikes in zi2zi and DGFont has distorted strokes in the upper part of the character, but the proposed model is able to reconstruct them well. Label 3, DGFont, has a loss of strokes, but the proposed model is able to reconstruct it. Label 7, zi2zi has blurred strokes, and DGFont cannot reconstruct triangular strokes, which is a characteristic of the font. The proposal improves on these failures. For label 8, the previous study failed to reconstruct the note shape of the top-right to bottom-left strokes, but the proposed model did. In Table II(b), zi2zi fails to reconstruct font features in labels 3, 6, 7, and 8, while DGFont fails to reconstruct strokes in labels 0 and 2, and font features in label 7. The proposed model has fewer missing strokes and reproduces font features correctly.

From on Table II(c), in label 6, the proposed model is able to reconstruct the upper left serif. In label 8, the previous study

TABLE I: The list of fonts

class label	font name	rendered character
Content	NotoSansJP Thin	永
0	851CHIKARA-YOWAKU	永
1	FZ IMOKENPI	永
2	hakidame	永
3	Onryo	永
4	RiiPopkkR	永
5	NotoSansJP Regular	永
6	NotoSerifJP Regular	永
7	TT ComicReggae-B	永
8	utamin=plus=s6	永



(a) Generated Fonts **(b)** Real Fonts

Fig. 3: Font generation results

cannot reconstruct the note type, but the proposed model can. In Table II(d), the proposed model is the most successful in reconstructing the spikes in label 1. The proposed model can reproduce the blurring of font strokes in label 3. The proposed model can reproduce serifs in the upper left corner of the font, while the previous study cannot reproduce serifs in label 6.

A list of the font generation results is presented in Fig.3. It is evident that the generated fonts are of comparable quality to the correct fonts included in the font data.

2) Quantitative evaluation: For quantitative evaluation, we employed the perceptual metric called Frechet Inception Distance (FID) [9]. FID was proposed by Martin et al. as a measure to assess the quality of images generated by a GAN. It calculates the distance between two sets of images to quantify their dissimilarity. In order to calculate the FID, we compare two sets of images: a data set of real character images and a set of character images generated by the DNNs of each method. The process involves the following steps:

1) Compute feature vectors for the images using the image classification model Inception-v3 [10].
2) Calculate the mean and covariance of the feature vectors for each image set.
3) Compare the mean and covariance values between the two sets to determine their dissimilarity.

Considering the recommended guideline of having 10,000 or more samples in an image set for accurate FID calculation, we conducted the FID calculation using images generated with all labels instead of calculating it on a per-label basis. This approach ensures a sufficiently large and diverse set of generated images for a robust evaluation of the model's performance. Table III presents the results of the FID calculation. It is evident that our font generation method has achieved perceptually superior results compared to previous studies.

3) Font Interpolation: The font shape is interpolated by linear interpolation of the two latent variables. The interpolation formula is represented by equation (3).

$$I_{interpolated} = f_{styleGAN2}(t * f_m(l_1) + (1-t) * f_m(l_2)) \quad (3)$$

The l_1 and l_2 represents the class labels, the f_m represents the mapping network, the t represent the interpolation percentage, the $f_{stylegan2}$ represent the StyleGAN2 Generator. The result of interpolating the fonts of label 1 and label 3 is shown in TableIV. In the case of zi2zi, font collapse does not occur, but the intermediate font does not seem to reflect the characteristics of either label 1 or label 3. On the other hand, in DGFont, font collapse occurs during the interpolation process. In the proposed model, if we focus on the roof radical "宀", we can see that the spikes, which are a feature of label 1, gradually appear during interpolation. This indicates that the interpolation is more successful compared to the previous methods, demonstrating improved performance in capturing the desired font characteristics.

4) Discussions: We conducted the comparison experiment between the character images generated by our proposed model and those generated by the previous methods, which are zi2zi and DGFont. The results clearly demonstrate that our proposed model excels in reconstructing fine font features more beautifully. Remarkably, it effectively captures intricate details such as spikes in fonts with label 1, blurred letters with label 3, and musical note shapes with label 8. This achievement can be attributed to the network structure of StyleGAN2, which is adept at generating high-quality fine details in human faces. Such a structure proves beneficial for the font generation problem.

Furthermore, we observed a reduction in stroke loss and blurring of the generated fonts compared to the previous method. The utilization of L1 Loss as the loss function in the previous study tended to lead to blurring, and the misalignment of strokes in fonts with distorted strokes (e.g., label 0) resulted in unnecessary loss. In conclusion, our proposed model demonstrates the ability to generate cleaner and more practical fonts compared to the previous method. However,

TABLE II: Comparisons of generated fonts

(a) character "永"

method		content	zi2zi	DGFont	Proposed	target
label	0	永				
	1	永				
	2	永				
	3	永				
	4	永				
	5	永				
	6	永				
	7	永				
	8	永				

(b) character "あ"

method		content	zi2zi	DGFont	Proposed	target
label	0	あ				
	1	あ				
	2	あ				
	3	あ				
	4	あ				
	5	あ				
	6	あ				
	7	あ				
	8	あ				

(c) character "ア"

method		content	zi2zi	DGFont	Proposed	target
label	0	ア				
	1	ア				
	2	ア				
	3	ア				
	4	ア				
	5	ア				
	6	ア				
	7	ア				
	8	ア				

(d) character "a"

method		content	zi2zi	DGFont	Proposed	target
label	0	a				
	1	a				
	2	a				
	3	a				
	4	a				
	5	a				
	6	a				
	7	a				
	8	a				

TABLE III: Comparisons of FID

model type	FID
zi2zi	117.9732
DGFont	93.9091
Proposed model	**66.1377**

as indicated in Table II(d), there are still some issues that need to be addressed, such as the presence of unnecessary strokes in label 3, differing reconstruction in label 6 compared to the target, and insufficient feature reconstruction in label 7 and 8. Future work entails adjusting the hyper-parameters and enhancing the network structure to resolve these challenges and further improve the performance of the model.

In our previous work[1], we reduced the network size of zi2zi for the purpose of shortening the font generation time. Considering practical applications, it is better to be finished font generation as fast as possible. We compared the computational time for 1000 rendered font character generation with NVIDIA GeForce RTX 3070 GPU, and show it in Table V.

While zi2zi has a simple network structure that repeats convolutions, StyleGAN2 has a complex network structure that transforms latent variables and adjusts the weights of convolutions accordingly. This is the reason why our model takes time longer than zi2zi. Compared to DGFont, the generation

TABLE IV: Results of interpolation

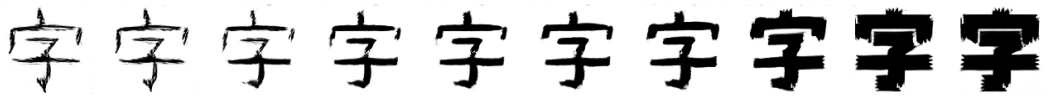
t	0	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	1
zi2zi										
DGFont										
Proposed model										

TABLE V: Computational time for 1000 rendered font characters generation

model type	Computational time [s]
zi2zi	10.51
DGFont	18.83
Proposed model	**17.08**

time is shorter. This is because the amount of calculation of DGFont is large due to the encoding of the reference image, which procedure is not adopted our model.

IV. Conclusion

In this paper, we proposed a label-based font generation system aimed at expressing emotions through fonts. The experimental results demonstrate that by applying the StyleGAN2 network, known for its ability to generate high-quality images, to font generation, we can produce cleaner fonts compared to previous studies. However, as observed "a" of Table II(d), there still exist challenges in adequately reproducing certain fonts and characters. Improving upon this limitation is one of our future challenges.

References

[1] Y. Shimamura and M. Niimi, "A proposal for emotion-expressive editor:emoeditor by font changing," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1008–1013. DOI: 10.23919/APSIPAASC55919.2022.9980240.

[2] Y. Tian, *Zi2zi: Master chinese calligraphy with conditional adversarial networks*, https://github.com/kaonashi-tyc/zi2zi, 2017.

[3] L. Tang, Y. Cai, J. Liu, *et al.*, "Few-shot font generation by learning fine-grained local styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 7895–7904.

[4] Y. Xie, X. Chen, L. sun, and Y. lu, *Dg-font: Deformable generative networks for unsupervised font generation*, 2021.

[5] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9300–9308, 2018.

[6] C. Wang, M. Zhou, T. Ge, Y. Jiang, and H. B. W. Xu, *Cf-font: Content fusion for few-shot font generation*, 2023.

[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8107–8116. DOI: 10.1109/CVPR42600.2020.00813.

[8] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 12 104–12 114.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhanced Residual Fourier Transformation Network for Lightweight Image Super-resolution

Yunming Yang* and Masaaki Ikehara†
* Keio University, Yokohama, Kanagawa 223-8522, Japan
E-mail: yunmingyang@keio.jp Tel/Fax: +81 080-4585-1915
† Keio University, Yokohama, Kanagawa 223-8522, Japan
E-mail: ikehara@tkhm.elec.keio.ac.jp Tel/Fax: 045-566-1530

***Abstract*—Residual blocks have been widely used in the single image super-resolution (SR), boosting the performance of SR results. The stacking of residual blocks helps to grab more information needed for the recovery of high-resolution (HR) images, but at the cost of larger model size and computational complexity. To address this problem, we propose a lightweight enhanced residual fourier transformation network (ERFTN). To ensure the network is lightweight enough while being capable of capturing both spatial and frequency information, we introduce a novel enhanced residual fourier transformation block (ERFTB) with an enhanced spatial attention (ESA) block. In addition, we adopt the contrastive loss for training acceleration without additional parameters. Extensive experiments demonstrate that our method can achieve SR performance superior to the state-of-the-art SR methods while reducing approximately half of the number of parameters (e.g., with 190K number of parameters, achieving 32.39dB in PSNR on Urban100 dataset).**

I. Introduction

Image super-resolution (SR) is frequently referred to as the process of recovering a high-resolution (HR) image from a given low-resolution (LR) image, which plays a vital role in the field of computer vision. SR has become a powerful tool in many real-world applications, such as medical image reconstruction and image quality improvement in surveillance and security fields. Given that not only the requirement of providing images with high fidelity is increasing, but also there is a growing demand of SR implementation in real-world scenarios, it is crucial to develop lightweight SR methods compatible with memory-constrained devices.

Currently, powerful deep learning techniques have been introduced into SR tasks to address the inherently ill-posed SR problems, boosting the performance of SR results from both the quantitative and qualitative views. As a pioneering work done by Dong et al. [1], the SRCNN first introduced a network structure with only three convolution layers and achieved better performance compared to traditional SR methods. Residual blocks (Resblocks), skip connections and attention mechanism had also been embedded into the network of SR models, like what EDSR [2] and RCAN [3] did, to further improve the SR performance. However, researchers noticed that for further improvements on SR performance, the network structure is not "the deeper the better", since the high-frequency details might be overlooked due to the learning bias towards low-frequency information. Meanwhile, the performance gain takes the cost of higher computational cost or memory usage, leading to limited application on low computing power devices.

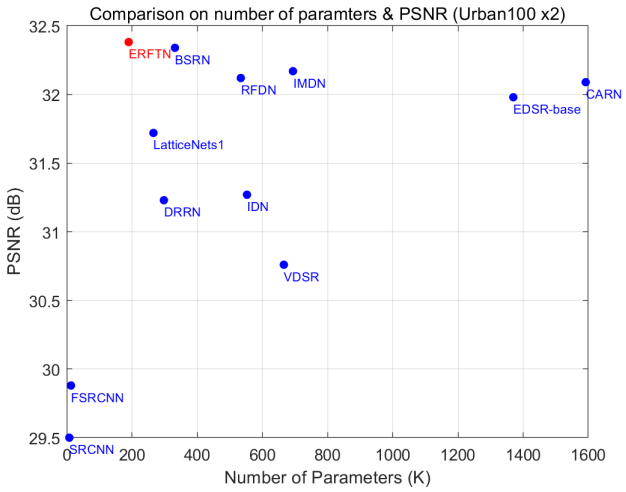


Fig. 1. Comparison on PSNR and number of parameters between the proposed and state-of-the-art SR methods on Urban100 dataset with scale factor 2.

To economize the number of parameters, deep recursive residual network[4] and distillation network [5]–[7] are introduced to SR models. However, we think that there is still space left for improving the accuracy of SR results while keeping the model lightweight with less FLOPs and fewer number of parameters.

Considering the wide use of residual blocks in the above-mentioned methods, we decide to propose a lightweight SR model, leveraging the merits of residual blocks. We design the enhanced residual fourier transformation network (ERFTN), with efficient plug-in residual blocks, enhanced residual fourier transformation block (ERFTB). The fast fourier transformation (FFT) branch in the residual blocks has shown its efficiency of capturing both high and low frequency information, as DeepRFT [8] has confirmed in the image deblurring tasks. Our proposed ERFTB combines the FFT branch with the normal convolution stream and an enhanced spatial attention (ESA) [9] mechanism, to make the SR model capable of capturing information of features in both spatial and frequency domains. Within the ERFTB, the ESA mechanism helps to enlarge the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

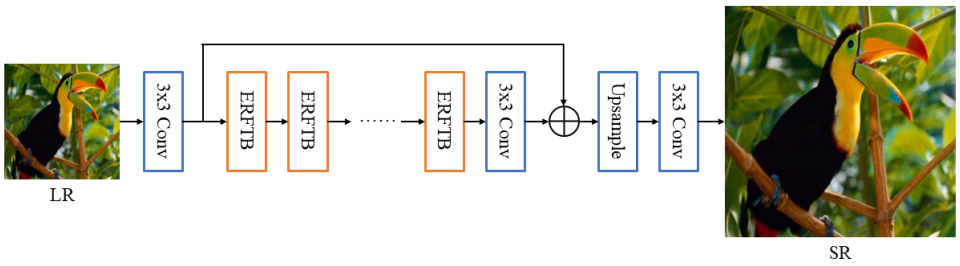


Fig. 2. The overall architecture of our proposed method ERFTN. ERFTB refers to the enhanced residual fourier transformation block.

receptive fields at the very beginning of the network without more parameters, and meanwhile, the low-frequency information is mined directly by the short skip connection structure. We also use depthwise over-parameterized convolution (DO-Conv) [10] to replace the traditional convolution layers in the residual blocks. DO-Conv could further shrink the model size and reduce the FLOPs. We additionally use a contrastive loss [11] for training acceleration, which can distinguish the positive and negative samples clearly during training.

As shown in Fig.1, the proposed method achieves better performance on PSNR while keeping the model lightweight compared to state-of-the-art methods.

II. Proposed Method

The overall architecture of our proposed method ERFTN is shown in Fig.2. We take EDSR [2] as the baseline and replace the residual blocks by our proposed ERFTB. With the simple but efficient structure, our method can achieve better SR results with fewer model storage consumption.

A. Enhanced Residual Fourier Transformation Block

We propose a new residual block Enhanced Residual Fourier Transformation Block (ERFTB), as shown in Fig.3, to replace the conventional residual blocks in the EDSR[2] baseline. ERFTB can integrate information from both spatial domain and frequency domain, supporting the SR result to be conformed to human perception and enjoy quantitative SR results improvements. ERFTB also has the merit of low computational cost, which supports the implementation of lightweight SR model.

1) Application of Fourier Transform:

Residual blocks have shown great success in computer vision problems from the low-level tasks to high-level tasks. The high-frequency information has always been treated carefully and make contributions to the SR recovering of features like edges and boundaries. However, if only the certain information is retaining, the learning bias will lead to severe distortion. Thus, it is important for the residual blocks to acquire a global learning skill.

Inspired by the success of Res FFT-Conv Block [8] in the image deblurring task and other adoption of frequency-aware network [12], we notice the ability of fast fourier transform

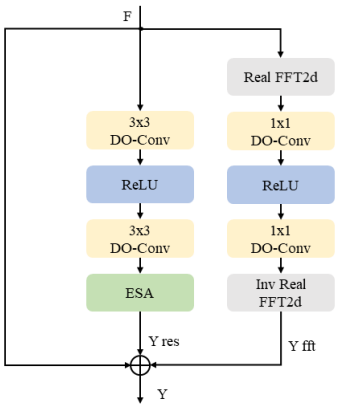


Fig. 3. Proposed ERFTB.

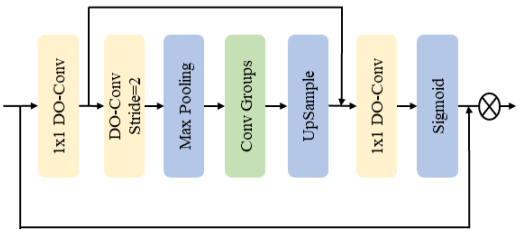


Fig. 4. Detail of ESA Block.

(FFT) [13] which can effectively integrate the frequency information without huge computations. We then produce the novel residual block ERFTB, as shown in Fig.3.

Based on the normal spatial residual stream of Conv-ReLu-Conv structure, we add the branch of FFT to compute the frequency information efficiently. Discrete Fourier Transform (DFT) is essential in the signal processing and Fast Fourier Transform (FFT) is a more efficient algorithm to count DFT according to the symmetric property of real array, and thus reduce the complexity. The 2D DFT is calculated as:

$$F(u,v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} f(x,y) \cdot e^{-j2\pi(\frac{ux}{H}+\frac{vy}{W})} \quad (1)$$

where j represents the imaginary unit; $H \times W$ denotes the image size; f is the input feature map and F is the output feature map. With the DFT transform, the output spectrum is

able to provide frequency from a global view.

The branch of FFT is processed as follows: First, applying the 2D FFT to the input feature $F_{input} \in \mathbb{R}^{H\times W\times C}$ to obtain the frequency spectrum $F_{fre} \in \mathbb{C}^{H\times W/2\times C}$. And then, since the real parts and imaginary parts can be separated in a complex tensor, we concatenate these two parts along the channel attention to acquire the representation of $\tilde{F} \in \mathbb{R}^{H\times W/2\times 2C}$. Subsequently, we apply Conv-ReLu-Conv to the $\tilde{F}$, and get the intermediate F_{inter} as:

$$F_{inter} = Conv(ReLU(Conv(\tilde{F}))) \tag{2}$$

where $Conv()$ denotes the 1×1 DO-Conv layers. Later, we reconstruct the complex tensor as $\widetilde{F_{fre}} \in \mathbb{C}^{H\times W/2\times C}$ by the processed real and imaginary parts as follows:

$$\widetilde{F_{fre}} = Real(F_{inter}) + jImg(F_{inter}). \tag{3}$$

Finally, we apply the inverse 2D FFT (IFFT) to $\widetilde{F_{fre}}$ to transform it back to the spatial domain, of which the output feature of FFT branch is $F_{out} \in \mathbb{R}^{H\times W\times C}$.

2) Application of Attention Mechanism:

Since the effectiveness of attention module has been proven [3], [6], [9], in this paper, we introduce the enhanced spatial attention (ESA) block to our approach. The detail of ESA Block is shown in Fig.4.

Compared to channel attention, we apply ESA Block to achieve a better performance on model complexity. ESA block is used for enlarging the receptive field quickly at the beginning of the network and ensuring the model being compact. The attention block is added to the end of each residual block, so that it should be lightweight enough, otherwise after the cascading of residual blocks, the computational cost would be overwhelming.

Note that we replace all the convolutional layers in ESA with DO-Conv [10] for a further decrease of model size. The main difference between ESA and CA is that ESA starts with a 1×1 convolutional layer, which largely reduce the channel dimensions at the very beginning, and thus keeping the whole module extremely compact. Then, a strided convolution with stride 2 and a strided max-pooling layer with large kernel size of 7×7 are used to enlarge the receptive field and to reduce the spatial size. Subsequently, the feature is processed by a group of convolutional layers, after which is an up-sampling layer and a 1×1 convolutional layer. Finally, HR features are combined directly through the skip connection when a sigmoid layer is then applied to generate the attention matrix.

The structure of ERFTB with inserted FFT branch and ESA Block, can not only reduce the number of parameters, but also bring the performance gain. Compared to ResBlocks in EDSR, the effectiveness of ERFTB is shown by the visualized feature maps in Fig.5. Our proposed method can obtain more frequency information effectively, ensuring that both high and low frequency information are adequately represented and not overlooked.

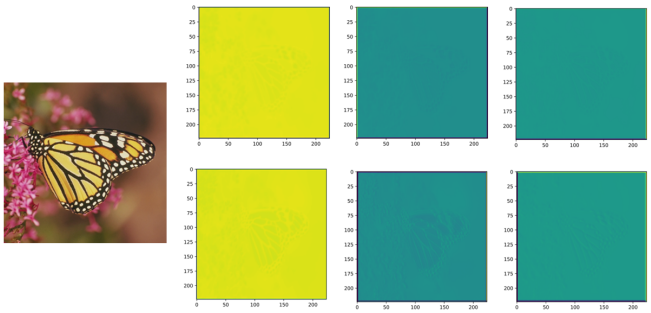


Fig. 5. Comparison of visualized feature map generated by layers 6,7 and 11 in the EDSR and ERFTN for image 'monarch' on Set 14. The upper row represents EDSR and the bottom comes from ERFTN.

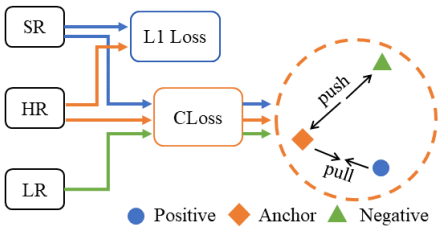


Fig. 6. The principle of contrastive loss.

B. Framework of enhanced residual fourier transformation network (ERFTN)

Learning from the prominent success of EDSR [2], we construct our framework based on the EDSR baseline for accurate and efficiency image super-resolution. As shown in Fig.2, we replace all the residual blocks by our proposed ERFTB. For a fair comparison to other state-of-the-art methods, we maintain the standard ERFTN model with 32 ERFTBs and the baseline model with 16 ERFTBs.

Depthwise over-parameterized convolution (DO-Conv)

DO-Conv [10] augment the convolutional layer with a depthwise convolution operation, which helps to achieve better performance and to accelerate the training process. To ensure the lightweight SR model could enjoy wider applications, we adopt DO-Conv without additional parameters.

Meanwhile, since the interface and receptive field of DO-Conv is the same as traditional convolutional layer, it can be easily applied in ERFTB, resulting in the same computational cost at the inference phase. We find that by employing DO-Conv, the number of parameters of our baseline with scale factor 2 can be reduced from 0.38MB to 0.19MB while achieveing slight improvements in PSNR.

C. Contrastive Loss

Normally, L1 Loss and MSE Loss are used to train SR model, of which the upper bound of the solution space is focused while the lower bound tends to be neglected. Considering contrastive loss (Closs) has already been proved to be efficient in many computer vision tasks, we introduce contrastive learning [11] into SR, which can separate the samples clearly and accelerate the training process.



Fig. 7. Visual comparison on Urban100 dataset with scale factor 4.

The specific representation of contrastive loss is shown in Fig.6. In SR, the positive sample is defined as the output of SR result, the anchor denotes the ground truth, which is the HR image, and the negative sample is defined as LR image, the Bicubic. We aim to pull the anchor closer to the positive sample while push the anchor far away from negative samples. Based on the Euclidean distance [11], Closs is calculated as:

$$L_{Closs}(W, Y, \vec{X_1}, \vec{X_2}) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2 \quad (4)$$

where D_W is the shorten notation of $D_W(\vec{X_1}, \vec{X_2})$, the Euclidean distance between two outputs; Y denotes the label, of which equals to 0 when two samples are similar and equals to 1 when two samples are dissimilar; $m > 0$ is the margin. We add a hyperparameter $\alpha = 0.1$ into Closs to control the distance between positive and negative samples in case the distance between positive samples and anchors keeps decreasing to a point where it fails to maintain the Closs to be positive. The overall loss function is as follows:

$$L = (n-t)L_1 + tL_{Closs} \quad (5)$$

where we set $n = 20$ and $t = 0.1$ based on the validation results.

III. Experiments

A. Datasets and Evaluation Metrics

Our model is trained on 800 images from DIV2K [14], which is commonly used in SR. Considering the diversity, there are various contents in DIV2K, such as people, environments, flora and fauna, handmade object, and natural scenery, etc. Four standard benchmark datasets, Set 5 [15], Set 14 [16], B100 [17] and Urban 100 [18] are used for testing. Set5 has five popular images in the field of computer vision. Set14 contains 14 commonly used images in the field of image processing. B100 contains more pictures related to real-world scenarios, like human and animals. Urban100 is a very typical dataset, which contains 100 clean images of urban environments, especially with buildings full of edges. The diversity of these four datasets helps to form an overview towards the effects of the SR model.

We compare the SR results from both qualitative and quantitative comparisons. For the quantitative part, we calculate PSNR and SSIM on the Y channel of YCbCr color space. The computational cost is evaluated by the number of parameters and FLOPs.

B. Implementation Details

In the training process, 16 images with patch size 48×48 are prepared as an input batch. Data augmentation includes random rotation by $90°$,$180°$, $270°$ and horizontal flipping. We use ADAM [19] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is set as 10^{-4} and reduced to $5 \cdot 10^{-5}$ after 200 epochs. The total training stops at 300 epochs. We set the ERFTB as 16 as the ERFTN baseline, and ERFTB as 32 as ERFTN+. The model in Pytorch framework is conducted on the NVIDIA GeForce RTX 3090 GPU.

C. Comparison with the State-of-the-arts

To demonstrate that our proposed ERFTN can achieve better balance between the accuracy and efficiency, we conduct extensive experiments and compare with 11 state-of-the-arts lightweight SR methods: SRCNN [1], FSRCNN [20], VDSR [21], DRRN [4], IDN [5], EDSR-Baseline [2], CARN [22], BSRN [7], IMDN [6], PAN [23], LatticeNet [24]. The results are shown in Table 1. Meanwhile, in Table 2, we compare our ERFTN to 3 latest prominent SR model, RCAN [3], NLSN [25], ENLCN [26] on the upscaling factor 4 to show that our method can acquire notable and even higher SR results with an extremely small model size.

The quantitative results in Table 1 demonstrate the effectiveness of our lightweight model, which can achieve higher

TABLE I

QUANTITATIVE COMPARISONS WITH OTHER LIGHTWEIGHT STATE-OF-THE-ARTS SR METHODS. RED TEXT INDICATES THE BEST PERFORMANCE AND THE BLUE TEXT INDICATES THE SECOND BEST PERFORMANCE.

Scale	Methods	Params(K)	Set5		Set14		B100		Urban100	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	SRCNN	8	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
	FSRCNN	13	37.00	0.9558	32.63	0.9088	31.53	0.8920	29.88	0.9020
	VDSR	666	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140
	DRRN	298	37.74	0.9591	33.23	0.9136	32.05	0.8973	31.23	0.9188
	IDN	553	37.83	0.9600	33.30	0.9148	32.08	0.8985	31.27	0.9196
	EDSR-Base	1370	37.99	0.9604	33.57	0.9175	32.16	0.8994	31.98	0.9272
	CARN	1592	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256
	BSRN	332	38.10	0.9610	33.74	0.9193	32.24	0.9006	32.34	0.9303
	IMDN	694	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283
	PAN	261	38.00	0.9605	33.59	0.9181	32.18	0.8997	32.01	0.9273
	LatticeNet	756	38.06	0.9607	33.70	0.9187	32.20	0.8999	32.25	0.9288
	ERFTN(Ours)	190	38.08	0.9608	33.65	0.9163	32.24	0.9001	32.39	0.9301
×3	SRCNN	8	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989
	FSRCNN	13	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080
	VDSR	666	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279
	DRRN	298	34.03	0.9244	29.96	0.8349	28.95	0.8004	27.53	0.8378
	IDN	553	34.11	0.9253	29.99	0.8354	28.95	0.8013	27.42	0.8359
	EDSR-Base	1555	34.37	0.9270	30.28	0.8417	29.09	0.8052	28.15	0.8527
	CARN	1592	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493
	BSRN	340	34.46	0.9277	30.47	0.8449	29.18	0.8068	28.39	0.8567
	IMDN	703	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519
	PAN	261	34.40	0.9271	30.36	0.8423	29.11	0.8050	28.11	0.8511
	LatticeNet	765	34.30	0.9272	30.32	0.8416	29.10	0.8049	28.19	0.8513
	ERFTN(Ours)	370	34.57	0.9282	30.49	0.8417	29.18	0.8071	28.54	0.8584
×4	SRCNN	8	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221
	FSCRNN	13	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280
	VDSR	666	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524
	DRRN	298	31.68	0.888	28.21	0.7720	27.38	0.7284	25.44	0.7638
	IDN	553	31.82	0.8903	28.25	0.7730	27.41	0.7297	25.41	0.7632
	EDSR-Base	1518	32.09	0.8938	28.58	0.7813	27.57	0.7357	26.04	0.7849
	CARN	1592	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837
	BSRN	352	32.35	0.8966	28.73	0.7847	27.65	0.7387	26.27	0.7908
	IMDN	715	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838
	PAN	272	32.13	0.8948	28.61	0.7822	27.59	0.7363	26.11	0.7854
	LatticeNet	777	32.18	0.8943	28.61	0.7812	27.57	0.7355	26.14	0.7844
	ERFTN(Ours)	340	32.41	0.8971	28.73	0.7810	27.67	0.7388	26.42	0.7942

TABLE II

QUANTITATIVE COMPARISONS OF FULL SR MODEL WITH PROMINENT STATE-OF-THE-ARTS. RED TEXT INDICATES THE BEST PERFORMANCE AND THE BLUE TEXT INDICATES THE SECOND BEST PERFORMANCE.

Scale	Methods	Params(M)	Set5		Set14		B100		Urban100	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×4	RCAN	16	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087
	NLSN	41	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109
	ENLCN	44	32.67	0.9004	28.94	0.7892	27.82	0.7452	27.12	0.8141
	ERFTN+(Ours)	6	32.69	0.9007	28.97	0.7888	27.83	0.7442	27.02	0.8116

TABLE III

COMPARISONS BETWEEN FLOPS AND PARAMETERS.

Methods	EDSR-Base	RCAN	ENLCN	IMDN	ERFTN
FLOPs (G)	50.60	1501.56	1519.68	25.42	7.11
Param (M)	1.37	40.73	41.22	0.69	0.19

TABLE IV

ABLATION STUDY ON ERFTB.

FFT	✓	✓	✓	✓	✓	✓	✓
ESA		✓		✓	✓	✓	✓
DO-Conv1			✓		✓	✓	✓
DO-Conv2				✓		✓	✓
CA	✓						
Closs							✓
PSNR	37.939	38.029	37.866	37.999	38.052	38.077	38.078
Param(M)	1.90	2.08	0.19	0.19	0.38	0.19	0.19

performance with compact model size especially in large scales. We adopt a further comparison with SwinIR [27], a very powerful latest SR model. On scale factor 4, our ERFTN only uses 340KB to achieve 27.67dB in B100 dataset, while lightweight SwinIR requires 897KB to achieve 27.69dB, which demonstrates that our model can keep lightweight while achieving competitive SR results. Our performance gain is particularly clear in dataset B100 and Urban100, which shows the promising application of our model in real-world scenarios.

Fig.7 shows the qualitative comparison results on Urban100. Our proposed method is able to reconstruct the SR images which are more faithful to the ground truth with more detailed information. Especially when conducting on lines and edges,

TABLE V
RESULTS OF DIFFERENT HYPERPARAMETER COMBINATIONS IN CLOSS. (PSNR RESULTS IN SET 5 DATASET $\times 2$)

t \ n	1	5	10	12	15	19	20	40	60	100
0.1	37.944	38.045	38.048	38.069	38.051	38.062	38.078	38.065	38.051	38.058

our method can produce lines in correct directions, while other methods may meet the problem of misleading results.

Table 3 compares the FLOPs and number of parameters cost by SR models with scale factor 2. It is obvious that our ERFTN largely reduces the FLOPs and uses the fewest number of parameters. The comparisons between other lightweight SR methods are shown in Fig.1. As a lightweight SR method, our model achieves a best trade-off between model accuracy and efficiency.

D. Ablation study

We conduct the ablation study of ERFTB on Set 5 dataset with scale factor 2 in Table 4. CA refers to the channel attention, which has commonly been used as the attention mechanism in SR models, like RCAN [3] and FADN [12]. The first and second column shows the effectiveness of applying ESA block. The adoption of DO-Conv on both FFT branch and ESA block can achieve the best results in PSNR and model size. The further implementation of Closs can introduce more performance gain without additional parameters, and what's more important, the training speed will decrease from 7.5 seconds to 5.2 seconds per epoch.

It should be noticed that the different combinations of hyperparameters in Closs will lead to different results, and sometimes even worse results. The experiments on Set 5 dataset $\times 2$ are shown in Table 5. As a result, we select n as 20 and t as 0.1 in our model. With reasonable selected parameters, the introduction of Closs could accelerate the training process and provide improvements in quantitative results.

IV. CONCLUSION

In this paper, we propose a novel enhanced residual fourier transformation network (ERFTN) for lightweight image super-resolution. The proposed enhanced residual fourier transformation block (ERFTB) with a FFT branch and an ESA attention mechanism can maintain more useful information in both frequency and spatial domain, leading to a better performance of SR results. The adoption of DO-Conv and contrastive loss can help to keep the model lightweight and accelerate the training process. We achieve satisfying trade-off between model accuracy and computational cost. Experiments on four datasets show that our method can provide promising SR results while having an extremely compact model size.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[3] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[4] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[5] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 723–731.

[6] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proceedings of the 27th acm international conference on multimedia*, 2019, pp. 2024–2032.

[7] Z. Li, Y. Liu, X. Chen, *et al.*, "Blueprint separable residual network for efficient image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 833–843.

[8] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual fourier transformation for single image deblurring," *arXiv preprint arXiv:2111.11745*, 2021.

[9] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2359–2368.

[10] J. Cao, Y. Li, M. Sun, *et al.*, "Do-conv: Depthwise over-parameterized convolutional layer," *IEEE Transactions on Image Processing*, vol. 31, pp. 3726–3736, 2022.

[11] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[12] W. Xie, D. Song, C. Xu, C. Xu, H. Zhang, and Y. Wang, "Learning frequency-aware dynamic network for efficient super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4308–4317.

[13] E. O. Brigham and R. Morrow, "The fast fourier transform," *IEEE spectrum*, vol. 4, no. 12, pp. 63–70, 1967.

[14] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 114–125.

[15] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[16] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, Springer, 2012, pp. 711–730.

[17] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, IEEE, vol. 2, 2001, pp. 416–423.

[18] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, Springer, 2016, pp. 391–407.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[22] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 252–268.

[23] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, Springer, 2020, pp. 56–72.

[24] X. Luo, Y. Qu, Y. Xie, Y. Zhang, C. Li, and Y. Fu, "Lattice network for lightweight image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[25] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3517–3526.

[26] B. Xia, Y. Hang, Y. Tian, W. Yang, Q. Liao, and J. Zhou, "Efficient non-local contrastive attention for image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2759–2767.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

ELEGANT: End-to-end Language Grounded Speech Denoiser for Efficient Generation of Talking Face

Ai-Fang Chai*†, Sailaja Rajanala*†, Arghya Pal*†, Raphaël C.-W. Phan†‡, and Chee-Ming Ting†
† CyPhi (ΨΦ) AI Lab, School of IT, Monash University, Malaysia campus
‡ Dept of Software Systems & Cybersecurity, Faculty of IT, Monash University, Clayton campus
* Equal contributions as first author
⋆ Corresponding Author's E-mail: sailaja.rajanala@monash.edu

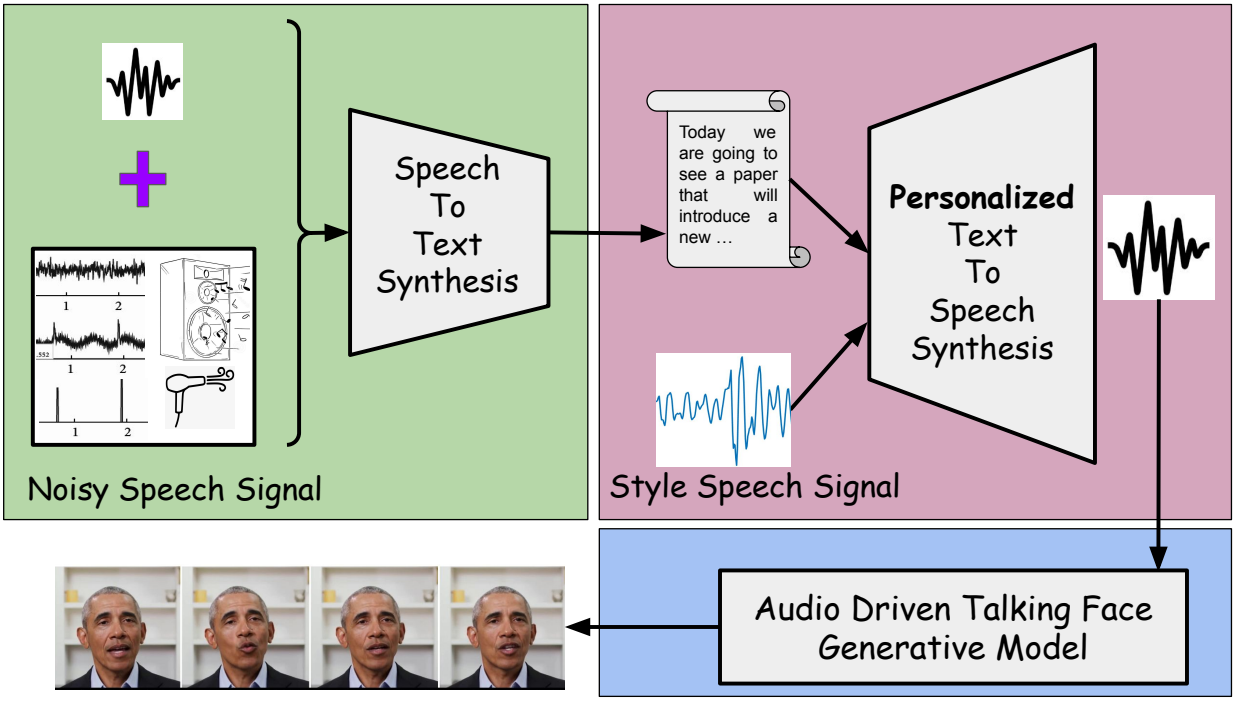


Fig. 1: ***ELEGANT*** *is a Speech2Face generative model (cyan-colored box), that generates talking faces, i.e. Obama's face in the image, given denoised personalized audio signals (pink-colored box). Different from other Speech2Face methods, the* ***ELEGANT*** *model performs a text grounding of the audio signal to mitigate any sources of noises (refer to § III), and to eliminate any possibility of hate speech to promote robustness and fairness in generative AI.*

***Abstract*—Existing speech driven talking face generation methods (*a.k.a* Speech2Face models) provide realistic-looking talking avatars. Yet, they are not appropriate if (1) the input speech signals are from the wild which could contain background noise; and (2) the input signal contains hate speech. In the presence of in-the-wild audio signals, the Speech2Face models do a bad lip-sync, unwanted facial movements, and sudden jitters on the head movements. While on the other hand, the Speech2Face models do not perform any reasoning on language understanding of input speech signal which could enable malicious users to translate hateful speech to a synthetic talking face promoting internet, social, and political threats. In this paper, we serve dual objectives on a single go. To the best of our knowledge, our method ELEGANT is the first Speech2Face generative model that performs a language grounding on the input speech that eliminates the transfer of spurious features originating from audio noise. Subsequently, the text embedding is associated with the speech style and passed on to a generative model with a view to learn the phoneme-viseme correspondence. In this way, our proposed ELEGANT model suppresses negative and hateful words using text embedding and also suppresses audio-specific noises using text embedding since noise-to-phoneme mapping would be random. Our experiments show that adopting the speech-denoising technique through text grounding eliminates the transfer of spurious features originating from audio noise to the vision domain. Consequently, a good phoneme-viseme correspondence leads to a comparable performances of SSIM and PSNR scores w.r.t state of the art methods.**

I. Introduction

Speech driven talking face video generation, i.e., Speech2Face, is a multi-modal generative model that translates speech signals to talking human faces with a realistic looking

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

facial expressions, head movement, eyebrow movement while lip-syncing the input speech to the generated talking face. However, we note that such Speech2Face models do not provide good generation if the input speech signals are from the wild. In the wild speech signals generally contain a lot of different types of background noises such as pure Gaussian noise, salt-and-pepper noise, etc., that a speech-capturing device such as a microphone might induce. While, surrounding noises such as wind blowing, baby crying, etc., are also considered as structured noise that comes in addition to the speakers' voices onto the in-the-wild audio. The generated talking face videos provided by the Speech2Face models do a bad lip-sync, unwanted facial movements, and sudden jitters on the head movements in the presence of noisy input speech signals. If we can still ignore such error provided by the Speech2Face generated model, any malicious user could provide harmful hate-speech audios and possibly get corresponding generated talking face. As a consequence, understanding the language of the audio should be a *must* for generating talking face videos otherwise it could harness serious ethical and societal concerns of responsible generative AI models.

To answer this question, we adopt two guiding principles:

Principle 1: ***The Speech2Face model must do away with any sources of background noises to generate talking face videos with realism and proper lip-sync without losing input audio quality*****.**
Principle 2: ***The Speech2Face model must understand the meaning of words of the spoken audio inputs to mitigate negative or hateful words that would negatively translate to talking face generation*****.**

We note that Principle 1 helps to mitigate any sources of in the wild noise [1]. Principle 2 promotes fairness in audio-driven talking face generation.

To this end, we propose a diffusion generative model based [2] Speech2Face model ELEGANT, aka End-to-end Language Grounded Speech Denoiser for Efficient Generation of Talking Face Videos, that aims to solve the challenge of talking face video generation, v, given a noisy input speech signal, s, to model $P(v|s)$. Different from the other Speech2Face models, we are first one to apply a comprehensive text grounding of the input in-the-wild audio signal, i.e., $P(v|s)P(\tilde{s}|t)P(t|\hat{s})$, where $\hat{s}$: noisy speech signal, $\tilde{s}$: reconstructed denoised speech signal through text grounding, and t: generated text from $P(t|\hat{s})$ later used by $P(\tilde{s}|t)$. Text grounding of the input signal is one of the main contributions of the paper and a novel contribution as far as we know.

We now describe each component of the $P(v|\tilde{s})P(\tilde{s}|t)P(t|\hat{s})$ model in below paragraphs:

Text Grounding of Input Audio Signal, $P(t|\hat{s})$: To do the diffusion based ELEGANT model audio to text grounding, i.e. , the methodology first gets the in-the-wild audio signal $\hat{s}$, encodes that to get the corresponding text. Hence, the ELEGANT considers the text grounding of the audio denoising as a language modeling task rather than continuous signal regression adopted by previous method [3]. The noisy speech signals are passed to the Silero based [4] speech-to-text model.

Personalized audio generation, $P(\tilde{s}|t)$: To model the $P(\tilde{s}|t)$, the generated text, t, is passed to a personalized text-to-speech decoder, $P(\cdot|t)$, that jointly trains a neural codec language model [5]. The neural codec language model provides an acoustic code matrix which is later passed to a LSTM based Grapheme-to-phoneme so as to get the phoneme sequence. Reference speech signal, i.e. the personalized audio signal, is provided to the text-to-speech decoder along with the generated text so that the decoder can learn the content and the speaker-specific styles, see Fig. [1].

Audio Driven Talking Face Generation, $P(v|\tilde{s})$: The denoised speech is then passed to the taking face generator module, i.e., $P(v|\tilde{s})$. Our methodology leverages the diffusion based "Sadtalker" [2] for the purpose of Speech2Face modeling. The Sadtalker model jointly trains identity-specific information such as pose and facial movements and a non-identity space such as texture, color, background, and facial structure information. The audio encoder of Sadtalker learns the 3D facial coefficients to represent the intermediate facial motion representation. The facial expression and head pose are generated from the audio which then are used to model final talking face video using a 3D-aware face render.

Surely, we are not the first to propose multimodal grounding. Our sister domains, such as [6] have shown promise by leveraging on the correspondence text to the image for captioning, clean audio-to-text for language modeling, and image-to-video for video understanding. However, we are the first to propose text grounding of noisy speech signals that not only denoises a noisy speech signal but retains meaningful language and emotion-specific information for a good quality audio-driven talking face generation. Our contributions are:

1. Text-Grounded Audio Driven Speech2Face Generation, to the best of our knowledge, we are the first one proposing text-grounded Speech2Face model. We show that the proposed Speech2Face model could mitigate any background noise using reasoning on language understanding of input speech.

2. Remove of Hate-speech to Promote Fairness, ELEGANT understands the language-specific information by learning a joint space of audio and text embedding. The learned joint space helps the ELEGANT model to suppress any form of negative, hateful, and biased words, thus promoting fairness.

3. High Quality Speech2Face Generation, using the 3D facial coefficients to represent the intermediate facial motion representation. The facial expression and head pose are generated from the audio which then are used to model final talking face video using a 3D-aware face render.

II. Related Work

Personalized Speech2Face models. The early research in Speech2Face is attributed to synthesizing-Obama [7] that uses more than 20 hrs of Obama's speeches. In recent years, the LSP has generated talking heads only from speech signals at over 30fps. Other works, such as [8], use audio signals to provide personalized Speech2Face generations using generative adversarial networks.

Speaker agnostic Speech2Face models. Speaker agnostic Speech2Face works such as Wav2lip [9], Audio2head [10], audio-to-face from arbitrary video generate talking faces by learning styles from a reference video and speech signals. Meanwhile, face re-enactment work [11] use a source video to style transfer the head motion to an identity.

Text grounding methods. Text grounding for audio signals [6] has shown promise by leveraging the correspondence between sound events and the captions provided. Similarly, the text grounding has shown significant improvement in video-grounded understanding [12]. Text grounding in image understanding [13] talks about the use of text descriptions that serves as an interpretation to qualitatively measure the rationality and model transparency. Similar to this, our work adapts text grounding involving semantic dependencies of text and speech signals to understand the interactions between text modality, speech, and video-specific dataset biases such as background noises, speaker head movements, non-agnostic user-specific styles, and stereotypes.

III. Methodology

Our goal is to learn the parameters of an audio-driven talking face generator, i.e.;

$$P_\theta(v|\hat{s}); v: \text{video}, \hat{s}: \text{noisy speech}, \theta: \text{parameter} \quad (1)$$

that is robust to noisy speech inputs.

Noisy Speech Signal, $\hat{s}$: We consider three major types of noisy speech, i.e. (1) background noises such as pure Gaussian noise, salt-and-pepper noise, etc., that a speech-capturing device such as a microphone might induce; (2) surrounding noises such as wind blowing, baby crying, etc., are also considered as structured noise that comes in addition to the speakers' voices onto the in-the-wild audio; and (3) the hate-speech provided by malicious user to the Speech2Face model.

The existing methods, that we discussed in Sec. II, encode both speech s and the noise ϵ, i.e., $\hat{s} = (s + \epsilon)$, that does not correctly translate the actual content of the speech and vaguely translates the noise component to the generated talking face. This produces poor-quality talking face videos with sudden head movements, bad lip sync, unnecessary eyebrow movements, and jitters in the background.

In this work, we leverages a speech denoiser $P_\psi(s|\hat{s})$ parameterized by ψ produces a cleaner version of the noisy image changing our Eqn. 1 to the following;

$$P_\theta(v|s)P_\psi(s|\hat{s}) \quad (2)$$

However, existing denoisers do not scale for in-the-wild noises or change, mask, or alter different language cues such as utterances, pauses, or emotions of spoken words. Our proposed methodology ELEGANT model considers an alternate text grounding way of denoising that poses speech denoising a language modeling task rather than a continuous speech signal regression task that provides high-quality talking face videos.

Without the loss of generality, the noisy speech signal $\hat{s}$ is first text grounded with an encoder $P_\beta(t|\hat{s})$ and then the generated text t is passed to the personalized text to speech decoder $P_\alpha(\tilde{s}|t)$. Thus, changing our Eqn. 1 to the following:

$$\begin{aligned} P_\theta(v|s)P_\alpha(\tilde{s}|t)P_\beta(t|\hat{s}) & \\ &\equiv P_\theta(v|s)P_\psi(s|\hat{s}) \\ &\equiv P_\theta(v|\hat{s}) \end{aligned} \quad (3)$$

The text-to-speech encoder $P_\beta(\cdot)$, quantizes noisy speech signals into discrete acoustic codes [4], i.e. $Encodes(\hat{s}) = C^{T\times 8}$, where C is the codec matrix, T is the downsampled uttrance length, and the 8 represents the eight codes for frame $t \in T$. The row vector of C represents eight codes of a frame, while the column vector represents the code sequence from the codebook, as shown in [5]. The discrete acoustic codes are then jointly trained with a phoneme-acoustic language model to get the text output.

The personalized text-to-speech decoder $P_\alpha(\cdot)$, on the other hand, gets the generated text from $t \sim P_\beta(\cdot)$ and provides the corresponding phoneme, i.e. $\mathbf{x} = \{x_0, x_1, \cdots, x_L\}$, using an LSTM based Grapheme-to-phoneme network [14]. Our personalized text-to-speech decoder $P_\alpha(\cdot)$ captures speaker-specific styles using clean, prompt audio of the user, see in Fig. [1]. The style prompt audio, i.e. s_{prompt}, is quantized into discrete acoustic codes, i.e. $Encodes(s_{prompt}) = \tilde{C}^{T'\times 8}$. The personalized text-to-speech decoder $P_\alpha(\cdot)$ generates the codec matrix C given the phoneme sequence $\mathbf{x}$ and the acoustic prompt matrix $\tilde{C}$, i.e.;

$$P_\alpha(C|\mathbf{x}, \tilde{C}) \quad (4)$$

The clean speech signal is generated back with a neural codec decoder that is able to reconstruct the waveform, i.e., $Decodes(C) = \tilde{y}$ where $C \sim P_\alpha(C|\mathbf{x}, \tilde{C})$. The speech modality maps syntactic and semantic representations to the text but fails to pick up audio-specific noises since noise-to-phoneme mapping is random.

Audio-driven talking face video generation, $P_\theta(v|s)$, model is inspired by the SadTalker model [2]. We assume to have a single image deep 3D reconstruction method [15] called 3D Morphable Models (3DMMs) to get the representation. In 3DMM, a face shape is modeled as a linear combination of average shape of the 3D face $F^{'}$, orthonormal basis of identity αF_{id}, and expression F_{exp}, i.e.,

$$F = F^{'} + \alpha F_{\text{id}} + \beta F_{\text{exp}} \quad (5)$$

The values of $\alpha \in R^{80}$ and $\beta \in R^{64}$ is taken from the LSFM [16] model. After that, the motion of the head pose and expression using the proposed PoseVAE and ExpNet

as described in [2]. Then, a 3D-aware Face Render [17] is used to render the 3D facial video. Similarly, one can use other Speech2Face model such as PC-AVS [18] method. The PC-AVS method modularizes the representation of human talking faces into speech content, identity space, and head pose by devising a compact pose code. This modularization is implicitly and complementarily learned within a construction-based framework using modulated convolution. Moreover, it is trained on five losses, (1) the generative adversarial loss $\mathcal{L}_{GAN}$ that provides a learning signal for overall frame by frame quality; (2) the identity loss $\mathcal{L}_i$ is a softmax cross entropy loss that preserves the identity of an actor; (3) the speech content loss $\mathcal{L}_c$ which is a contrastive loss that seeks the synchronization between audio and visual features by leveraging a negative and positive set of audio and visual feature pairs; (4) the $\mathcal{L}_{VGG}$ is the perceptual loss that relies on a pre-trained VGG network; (5) the $\mathcal{L}_1$ loss computes the $\mathcal{L}_1$ distance between the generated frames of the video and actual frames of a real video.

We note that our proposed text grounded Speech2Face ELEGANT model not only work well with PC-AVS but it can be incorporated with any other framework. In Sec. IV, we illustrate some ablation studies that shows the feasibility of using our text-grounded speech denoiser to other state-of-the-art methods.

IV. Experiments

Dataset Curation and Pre-processing: We leveraged our clean video dataset from the open source project page of [21], retaining the subjects in the video but only changed the video source. Our preprocessing involves cropping the videos in three-second clips. The noisy speech of the above mentioned videos are created following [22], wherein a noise from ten categories of Urbansound8K [23] dataset is overlaid on the clean speech signals. For each noise category of Urbansound8K, over 3000 recordings are downloaded and manual checking is done before Audacity is used to label the start and end times of every occurrence of the sound in each recording [23]. This is done with repetition to create 3 noise classes: Class0-Air conditioner noise, Class1-Car horn noise and Class2-Children playing. In addition, we have Class G-Unstructured Gaussian noise sampled using PyTorch. Unstructured noise do not require any dataset any could be generated using PyTorch packages [24]. We use Urbansound8K [23] that contains 27hrs of audio with 18.5 hrs of annotated sound event occurrences across 10 sound classes such as air conditioner, car horn, children playing, dog bark, drilling, gun shot, etc.

The Baselines Methods. For the purpose of understanding the effect of inducing noisy signals into speech signals on the talking face generation, we have broadly divided the baselines into methods that disentangle speech signals to obtain the content and speaker-specific signals and methods that do not. We run our experiments on the pre-trained models (code sourced from the respective repositories) on the videos in the wild collected as above.

- *Methods w/o Speech Disentanglement*
  1. **PC-AVS** [18] proposes pose-controllable talking face generations. The speech features and the visual features are embedded in the time-based feature space. Thus altering the speech signals will also affect the visual features.
  2. **Wav2Lip** [20] proposes an efficient lip-sync method for videos in the wild. During training, the speech encoder is tuned to generate realistic faces i.e. faces of good visual quality. Thus noisy speech features have the ability to hamper the quality of the generated faces.
  3. **Audio2Head** [10] makes use of keypoint-based dense motion fields to generate realistic and natural head movements for talking face generations. The acoustic features are used in generating the head poses and the keypoints of the facial movements.
- *Speech Disentanglement Based Methods*
  1. **Live Speech Portraits**(LSP) proposes a personalized talking face generative model. It uses APC (Autoregressive Predictive Coding) [25] to learn semantic speech embedding based on speakers' speech content. APC learns a speech based language model by trying to generate speech frames conditionally.
  2. **MakeitTalk** [19] proposes a speaker-independent Speech2Face model that provides Speech2Face videos understanding content and speaker specific features. The content specific audio features are only used to generate the lips and surrounding regions and the speaker specific features are used to generate other facial engagements.

Discussion on Quantitative Results: SSIM and PSNR are adopted to understand the generated video qualities and the results are in Table I. Four types of noises namely Noise-0: air conditioner noise, Noise-1: car honk, Noise-2: children playing, and Noise-G: Gaussian noises are overlaid on input speech signals. Our proposed text grounded speech denoiser provides comparable performaces when applied to PC-AVS, Wav2Lip, and Audio2Head, showing the efficacy of our method. Please note that the baseline denoisers are trained on custom noise categories but our method is generalizable across noises without further training on noise categories. We employ two video quality metrics, SSIM and PSNR to understand effects of noisy speech signal towards noisy video generation. Table I shows the results on four noise types. In the first three models we see that the text-grounding based denoising performs as close to a custom denoiser trained on the specific noise [22]. Interestingly, the denoisers are quite ineffective at denoising unstructured noise, as evidenced by the comparison between Denoised-G and TG-G. MakeitTalk and LSP are unaffected by noise induction as they have phoneme based speech disentanglement procedures. Methods not employing speech disentanglement use entire speech feature space to derive structural characteristics of the face. Thus small perturbation in the speech signal could be multiplied many folds translating into spurious audio-visual features. This results in unwanted jitters, and eye and hair movements.



Fig. 2: **Lip-sync qualitative comparison:** We use noisy speech signal, i.e. input with car horn noise, from Urbansound8K dataset. We note that the face generate by ELEGANTmodel is similar to ground truth. But, other methods such as **PC-AVS** [18], **MakeitTalk** [19], **Wav2Lip** [20], **Audio2Head** [10].

Model	Metric	Noise-0	Denoised-0	EG	Noise-1	Denoised-1	EG	Noise-2	Denoised-2	EG	Noise-G	Denoised-G	EG
PC-AVS [18]	SSIM	0.407	0.438	0.460	0.399	0.438	0.449	0.407	0.438	0.436	0.407	0.438	0.470
	PSNR	24.552	26.469	26.560	24.526	26.470	25.803	24.552	26.471	24.843	24.552	26.472	27.015
Wav2Lip [20]	SSIM	0.902	0.946	0.946	0.902	0.946	0.912	0.902	0.946	0.946	0.900	0.947	0.946
	PSNR	34.41	36.24	36.17	34.38	36.20	34.80	34.40	36.24	36.286	34.35	36.27	36.18
Audio2Head [10]	SSIM	0.431	0.490	0.465	0.437	0.489	0.435	0.434	0.489	0.469	0.434	0.492	0.528
	PSNR	22.91	25.68	24.75	22.93	25.67	23.20	22.91	25.67	24.75	22.91	25.68	28.11
MakeitTalk [19]	SSIM	0.514	0.513	0.513	0.513	0.515	0.517	0.515	0.512	0.522	0.514	0.521	0.515
	PSNR	28.12	28.13	28.13	28.14	28.14	28.14	28.13	28.13	28.14	28.11	28.15	28.16
LSP	SSIM	0.513	0.513	0.511	0.516	0.514	0.511	0.514	0.512	0.511	0.513	0.512	0.511
	PSNR	28.14	28.13	28.13	28.15	28.16	28.15	28.15	28.14	28.13	28.13	28.15	28.11
SadTalker	SSIM	0.513	0.513	0.511	0.516	0.514	0.511	0.514	0.512	0.511	0.513	0.512	0.511
	PSNR	28.14	28.13	28.13	28.15	28.16	28.15	28.15	28.14	28.13	28.13	28.15	28.11
ELEGANT (OUR METHOD)	**SSIM**	**0.972**	**0.961**	**0.965**	**0.956**	**0.974**	**0.971**	**0.949**	**0.952**	**0.951**	**0.953**	**0.952**	**0.951**
	PSNR	**38.14**	**38.13**	**38.73**	**39.15**	**39.16**	**38.15**	**38.15**	**39.14**	**39.13**	**38.13**	**38.15**	**38.11**

TABLE I: Quantitative comparison of the state-of-the-art methods. **Unaffected Models**: Since MakeitTalk [19] and LSP [21] employ speech disentanglement that dampens spurious audio-visual correlations, this results in good quality of generations. **Noise-0**: air conditioner noise, **Noise-1**: car honk, **Noise-2**: children playing, **Noise-G**: Gaussian noise, **EG**: Proposed ELEGANT model, **Denoised**: Custom denoisers trained on specific noise type [22].

between those who were _idiots_ and those who are not
between those who were _[MASK]_ and those who are not
between those who are _prospering_ and those who are not
between those who are not good and those who are not

TABLE II: **Fairness**: The language grounded speech denoiser of ELEGANTcould replace negative words, for example “idiots” either with other words or blank.

Discussion on Qualitative Results: The Fig. 2 shows a qualitative comparison among noisy audio, denoised audio, and text grounded audio (our method) of PC-AVS method. We note that other Speech2Face models have achieved bad SSIM and PSNR scores w.r.t our ELEGANTmethod. Also, the other state-of-the-art methods produce bad lip sync and eyebrow movements which are crucial for Speech2Face models.

V. Fairness of the ELEGANT model

The main job of the ELEGANT model is more than generating *only* high quality Speech2Face videos from noisy speech signals. To promote fairness in the Speech2Face domain, the ELEGANT model must understand speech signals to avoid denoising a speech signal that contains negative or hateful words; which would later be used for personalized Speech2Face videos. Similar to the work proposed in [26] that provides ways to mitigate hate words within a text document by remapping them through text embedding, ELEGANT could remap negative and hate words by leveraging the joint acoustic and text embedding. This is illustrated in Table II.

VI. Conclusion

In this work, we propose a Speech2Face model called ELEGANT that provides fair high quality audio-driven talking face videos given noisy speech or spoken signals that contains negative and hateful words. The ELEGANTconsiders understands the (noisy-)speech using a language model and suppresses any form of background noise, negative, or hateful words within the speech, thus promoting fairness in Speech2Face domain. Our experiments exhibit that adopting speech denoising technique provides high quality Speech2Face videos.

ACKNOWLEDGMENT

This research was supported in part by the Malaysia Ministry of Higher Education (MoHE)'s Fundamental Research Grant Scheme (FRGS) under the AdverSurreal project FRGS/1/2021/ICT02/MUSM/01/1. Experimentations were run on Montage, the Monash University Malaysia campus HPC platform. We acknowledge the reviewers whose comments and suggestions significantly enhanced this manuscript.

REFERENCES

[1] Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34:2556–2569, 2021.

[2] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, 2022.

[3] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

[4] Silero Team. Silero models: pre-trained enterprise-grade stt / tts models and benchmarks. https://github.com/snakers4/silero-models, 2021.

[5] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[6] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. Text-to-audio grounding: Building correspondence between captions and sound events. In *ICASSP 2021*, pages 606–610, 2021.

[7] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*, 2017.

[8] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38(4), jul 2019.

[9] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery.

[10] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

[11] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Audio-visual face reenactment. In *WACV*, pages 5178–5187, 2023.

[12] Hung Le and Steven C.H. Hoi. Video-grounded dialogues with pretrained generation language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5842–5848, Online, July 2020. Association for Computational Linguistics.

[13] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *CVPR*, 2020.

[14] Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. IEEE, 2015.

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[16] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[17] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[18] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021.

[19] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.

[20] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACMMM*, pages 484–492, 2020.

[21] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021.

[22] Madhav Mahesh Kashyap, Anuj Tambwekar, Krishnamoorthy Manohara, and S Natarajan. Speech denoising without clean training data: a noise2noise approach. *arXiv preprint arXiv:2104.03838*, 2021.

[23] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.

[24] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*, 2021.

[25] Yu-An Chung and James Glass. Generative pre-training for speech with autoregressive predictive coding. In *(ICASSP 2020)*, pages 3497–3501. IEEE, 2020.

[26] Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. Hate speech detection based on sentiment knowledge sharing. In *IJCNLP,2021*, pages 7158–7166, Online, August 2021. Association for Computational Linguistics.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Segmentation Enhancement for Iris Recognition Using Unit Gradient Vectors

Limhourlaurent Meam*, Suradej Duangpummet†, and Waree Kongprawechnon*
* Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
E-mail: m6522040614@g.siit.tu.ac.th, waree@siit.tu.ac.th
† NECTEC, National Science and Technology Development Agency, Pathum Thani, Thailand
E-mail: suradej.dua@nectec.or.th

***Abstract*—Iris recognition is a prominent technique in biometric identification due to its robustness and distinctive features. However, challenges in iris segmentation, including limitation of sensors, occlusion, and inconsistent lighting, significantly impact the accuracy and efficiency of the recognition. Therefore, this paper proposes improving iris segmentation by incorporating a technique known as Unit Gradient Vectors (UGVs) into the traditional preprocessing stage. The UGVs at each pixel provide edge direction, enabling the accentuation of areas with significant directional changes. Thus, a more robust and correct edge detection from inconsistent lighting images can be obtained. An experiment was conducted using the CASIA-Iris-Lamp, which has challenging lighting conditions. Results suggest that the proposed method achieves superior performance compared to the existing method. The accuracy, false acceptance and rejection rates were improved by 2.04%, 10.00%, and 7.33%, respectively. Specifically, the equal error rate was dramatically improved by 37.78%. These promising results can contribute to advancing security and efficiency in iris recognition systems.**

I. Introduction

Biometric recognition systems offer a crucial alternative to traditional security methods reliant on passwords or physical keys. These systems leverage unique physiological or behavioral characteristics, providing intrinsic and readily accessible identity verification. Among the various biometric modalities, irises have garnered significant attention due to their superior degree of uniqueness and stability throughout an individual's life [1]. Iris recognition uses intricate patterns within the iris, the ring-shaped region surrounding the pupil, to distinguish individuals. According to its reliable and robust identification, it has been reported as exceedingly low false match rates [2].

Despite the promise of iris recognition, challenges persist. The acquisition of high-quality iris images, for instance, is often a significant hurdle. If iris images are captured with low-quality sensors, the images may be noisy, blurred, or low contrast, which could lead to identification errors and false positives. Such challenges can adversely impact the security and efficiency of various sectors that rely on these identification mechanisms.

In addition, one critical aspect of iris recognition is the accurate segmentation of the iris within an image. Segmentation involves detecting the iris's inner and outer boundaries accurately. Precise segmentation is pivotal to eliminating noise and artifacts from the surrounding eye, such as the sclera and eyelids. Inaccurate segmentation can introduce spurious features, which can significantly reduce the overall performance of the iris recognition system. However, achieving accurate iris localization is far from straightforward. Factors such as occlusion by reflections, low resolution, noises, and unpredictable lighting conditions introduce considerable difficulties. The most challenges often arise when dealing with iris images under varying lighting conditions.

Numerous segmentation methods have been proposed to improve the precision of iris localization. Traditional methods such as Daugman's algorithm [3], circular Hough transform [4], and active contour [5] employed fitting methods based on circle or ellipse shapes to localize iris and pupil regions. Recently, many image processing techniques and machine learning algorithms have been proposed [6]–[8]. Notable contributions include the application of the Hough transform and Canny edge detector [7] and the convolutional neural networks (CNNs) used for iris segmentation [8].

In the pursuit of more precise edge detection mechanisms, the utilization of unit gradient vectors (UGVs) has been proposed for accurate edge detection [9]–[11]. UGVs offer an effective way to quantify the directional changes in an image, thereby enhancing edge detection capabilities. They capture and vectorise changes in pixel intensity, facilitating the handling of intricate patterns in iris textures.

Therefore, we aim to investigate the potential of utilizing unit gradient vectors to improve the process of iris segmentation under the ongoing challenge in the field, i.e., lighting inconsistency. As a result, this study focuses on enhancing the iris localization techniques in a traditional iris recognition system. We propose a more robust method for accurately localizing the iris, thereby improving the robustness and accuracy of iris recognition systems overall. Hence, this research contributes to developing a more efficient iris recognition system applicable to real-world scenarios.

This paper is organized as follows. The traditional iris recognition process is introduced in Section II. Section III describes the proposed method. Section IV presents the experiment and result. Lastly, section V and section VI discusses and summarizes key points of this study.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Iris Recognition

In this section, the conventional approach used in the iris recognition process is explored, starting with the fundamental step of image preprocessing. This stage is set as the foundation for subsequent procedures, playing an essential role in determining the subsequent process of recognition.

A. Image Preprocessing

Image preprocessing comprises three steps: reflection removal, image resizing, and target pixel identification.

1) Reflection Removal: A process called "inpainting" creates the masked areas based on the surrounding image content [12]. Firstly, the mean intensity value in a neighborhood around each pixel is computed, called *adaptive thresholding*. The pixel with an intensity higher than its mean with some constant is identified as a reflection. Secondly, each connected component in the mask is labelled, and its size is computed. The components outside the threshold are removed from the mask. Lastly, morphological dilation using a specific size circular structuring element is applied to the mask to expand the areas for covering all reflections.

2) Resize Image: The width and height of the input image are reduced to two-thirds of its original size. Then, an erosion operation is applied to enhance the iris details and eliminate the noise around the boundaries [13].

3) Find Target Pixel: The local minima, a low-intensity spot in an image, is identified as a target pixel. It is a pixel or group of pixels with lower intensity values than their surrounding pixels.

B. Iris Segmentation using Daugman's method

The traditional method lies in Daugman's algorithm, called *integral differential operator* [3]. The objective is to find a circular path within the image where the change in pixel intensity is the most dramatic. The locations of these abrupt changes in intensity typically correspond to the boundaries of the pupil or iris. The integral differential operator, the so-called Daugman values, is defined as:

$$\max(r, x_0, y_0) \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint_{r,x_0,y_0} \frac{I(x,y)}{2\pi r} ds \right|, \quad (1)$$

where $G_\sigma(r)$ denotes a Gaussian filter of scale σ. r represents the radius of a circular region. x_0 and y_0 represent the coordinates of the center of the circular region. These vectors store "scores" for each potential circular boundary of the pupil and iris. The score is computed based on pixel intensity and gradient along a particular circular path at the center of the circle. Thus, the higher the Daugman values for a given circle, the more likely that the circle aligns with the edge of the pupil or iris.

Subsequently, the algorithm identifies the circle with differences in Daugman's values between consecutive radii. This circle represents the location and size of the pupil or iris, given that the abrupt change in Daugman's values between consecutive radii signifies a stark change in pixel intensity — indicative of the boundary between the iris and the sclera. For example, Fig. 1 shows the segmented area of the iris from an image in the CASIA-Iris-Lamp dataset [14].

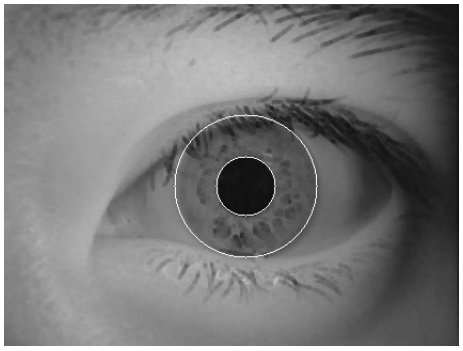
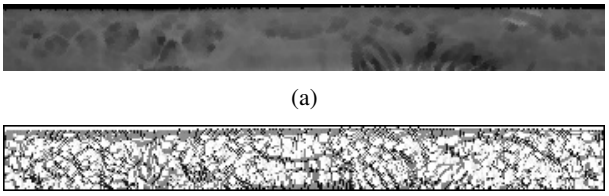


Fig. 1: An example of a segmented image from the CASIA-Iris-Lamp dataset[1] .



(a)

(b)

Fig. 2: A normalized image using Daugman rubber sheet model (a), and its feature using the local binary pattern (b).

C. Iris Normalization and Enhancement

Since to the elastic nature of the iris, pupil dilation and constriction are deformed due to changes in illumination, normalization is used to compensate for these deformations. The iris region is transformed into a fixed and consistent dimension. The size and shape variability of the iris can significantly impact the performance of the subsequent stages, i.e., feature extraction and matching.

A well-established technique is Daugman's rubber-sheet model [15]. It transforms the annular iris region from its original Cartesian coordinates into normalized polar coordinates. This transformation process, also known as "unwrapping", allows for the representation of the iris as a flexible rubber sheet, thus achieving a consistent representation regardless of pupil dilation or gaze direction.

During this unwrapping process, the coordinates of the inner (pupil boundary) and the outer (iris boundary) circles are calculated for each radial line originating from the pupil center to the iris. Each radial line is then re-sampled to a fixed number of points, creating a constant radial resolution.

The result of this unwrapping process is a rectangular block where one axis represents the change in angle around the iris, ranging from 0 to 2π, and the other axis corresponds to the radial distance from the pupil boundary to the iris boundary.

[1]Casia Iris Image Databases, Available: http://biometrics.idealtest.org

The pixel intensity values from the original image are then transferred to this new normalized representation based on their corresponding spatial location, as shown in Fig. 2a.

In addition, an image enhancement technique can be employed. For example, the histogram equalization [16] is used for adjusting the distribution of pixel intensities so that a more uniform histogram enhances the contrast of the image. It effectively reveals the complex details of the iris, making features more distinguishable and prominent.

D. Feature Extraction

Feature extraction can apply a local-binary pattern (LPB) technique [17] to generate an iris vector from the iris region. An LPB compares each pixel in the image with its neighboring pixels. A central pixel is compared with its eight surrounding pixels, forming a binary pattern. Depending on whether the central pixel's intensity is less than or equal to the neighboring pixel's intensity, a '1' or '0' is assigned, forming an 8-bit binary number (pattern) for each pixel in the iris region. The entire iris region is thus transformed into an LBP image that encapsulates the unique texture of the iris, as shown in Fig. 2b.

From the LBP image, a histogram is calculated to encapsulate the frequency of each binary pattern and encode the iris texture. The normalization of this histogram results in the formation of a feature vector, which is the input for matching.

E. Matching

The matching stage follows the feature extraction process and involves comparing the feature vectors that are derived from different iris images. The chi-square test statistic [18] is utilized for this comparison, measuring the similarity between the histograms of two feature vectors. This statistical test quantifies the difference between two histograms and provides the matching score. The lower the score, the higher the similarity between the two irises. By distinctly dividing the process into feature extraction and matching stages, the method proves to be an effective and robust solution for iris recognition.

III. Proposed Method

This section describes the proposed method based on the iris recognition processes. The proposed method is incorporated into a traditional segmentation technique which describes in Section II, to improve iris segmentation precision and to enhance the overall performance of the iris recognition system. The proposed iris recognition is shown in Fig. 3.

A. Unit Gradient Method

The preprocessing stage during the application of Daugman's algorithm involves the unit gradient method. The unit gradient vector of an image at a given pixel is a two-dimensional vector that points in the direction of the steepest increase in pixel intensity, and its length (or magnitude) is normalized to one. It is obtained by first calculating the gradients in the x and y directions (using a method such as the Sobel operator [16]), then combining it to get the gradient vector at each pixel, and finally normalizing these vectors to have a length of one.

The unit gradient vectors contain information about the direction of the edges in the image. Adding these vectors into the computation of Daugman values could help to emphasize the areas where there is a strong directional change, potentially improving the precision of the edge detection. This combined method enables a more accurate distinction between real edges and image noise, enhancing the overall edge detection results.

According to the proposed algorithm, the unit gradient vectors (UGVs) are vital to the iris recognition process. The computation of the UGVs begins with estimating gradients along the x and y axes using the Sobel operator. The gradients can be represented as:

$$G_x = \frac{\partial I}{\partial x}, \quad G_y = \frac{\partial I}{\partial y}. \tag{2}$$

In Eq. (2), G_x and G_y denote the image gradients along the x and y axes respectively, with I representing the image. The symbols $\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$ denote the partial derivatives of the image with respect to the x and y axes, respectively.

Thus, the gradient magnitude, $|G|$, can be calculated by using the Euclidean norm, shown in (3):

$$|G| = \sqrt{G_x^2 + G_y^2}. \tag{3}$$

The gradient magnitude $|G|$ measures the intensity change rate at a given pixel.

Finally, the UGVs are computed by normalizing the gradients. This is shown in (4):

$$\mathbf{U} = \frac{G}{|G|} = \left(\frac{G_x}{\sqrt{G_x^2 + G_y^2}}, \frac{G_y}{\sqrt{G_x^2 + G_y^2}} \right). \tag{4}$$

In Eq. (4), normalization ensures that each vector has a magnitude of one while maintaining the direction of the maximum change in intensity. Hence, the UGVs offer valuable edge direction information, enhancing iris segmentation's precision by adding them with the Daugman's values as shown in Eq. (5):

$$\begin{aligned} \text{Daughman_values} = \text{conv_2d}(image, x, y, mask) \\ +\text{unit_gradient_x}[y, x] \\ +\text{unit_ gradient_ y}[y, x], \end{aligned} \tag{5}$$

where the Daughman's values are the Daugman's representations at a specific polar coordinate (pre_loc_y, pre_loc_x) and radius (r − r_min). These values store "scores" for each potential circular boundary of the pupil and iris. Conv_2d is represents the result of a 2D convolution operation between the input iris image and a filter mask. Convolutions are used to extract features from images by applying a filter/mask and summing up the element-wise products. This operation is being applied to enhance the iris texture. Unit gradient represent the

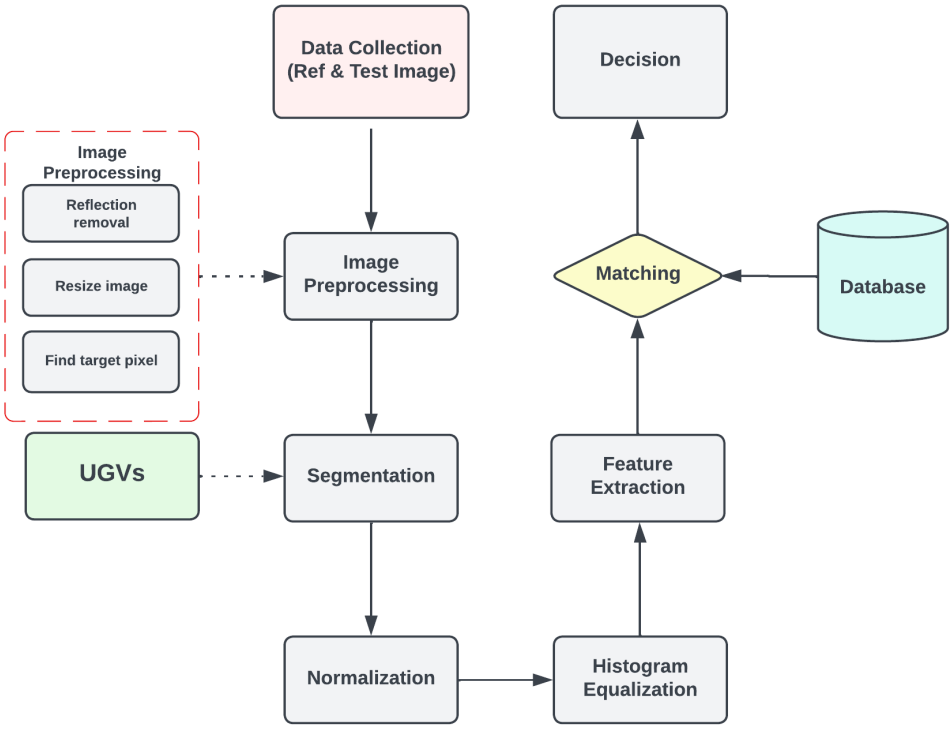


Fig. 3: A diagram of the proposed iris recognition using unit gradient vectors (UGVs).

unit gradient components in the x and y directions at pixel coordinates (x, y). The gradient in the x and y directions can provide information about the direction of intensity change at a given point. Finally by adding the score with the UGVs, we could emphasize more about the area which is the boundary of the circular irises.

B. Feature Extraction

The segmented image is successfully obtained by applying the proposed method in tandem with the traditional segmentation approach. The resulting segmented images, with improved precision, are subject to feature extraction. We leverage the principles of transfer learning to extract features from the normalized images. Since the ResNet50 model is trained on the extensive ImageNet database [19], it is highly regarded for its robust performance in image recognition tasks. Therefore, we utilize the prowess of this pre-trained model in this work [20].

Note that although we have chosen to use ResNet50, the other models, such as VGG16 [21], VGG19 [22], ResNet101 [23], InceptionV3 [24], and InceptionResNetV2, also offers remarkable capabilities for feature extraction in image data, having been pre-trained on ImageNet and demonstrated their effectiveness in various contexts. The choice of ResNet50 in our current approach was motivated by its specific strengths and suitability for our task.

We customize the application of these models to the unique needs of iris recognition. The model is loaded without its top layer comprising fully connected layers. This step customizes the architecture to our task and allows us to adjust the input shape to fit the dimensions of the normalized iris images, resized to 60×360 pixels.

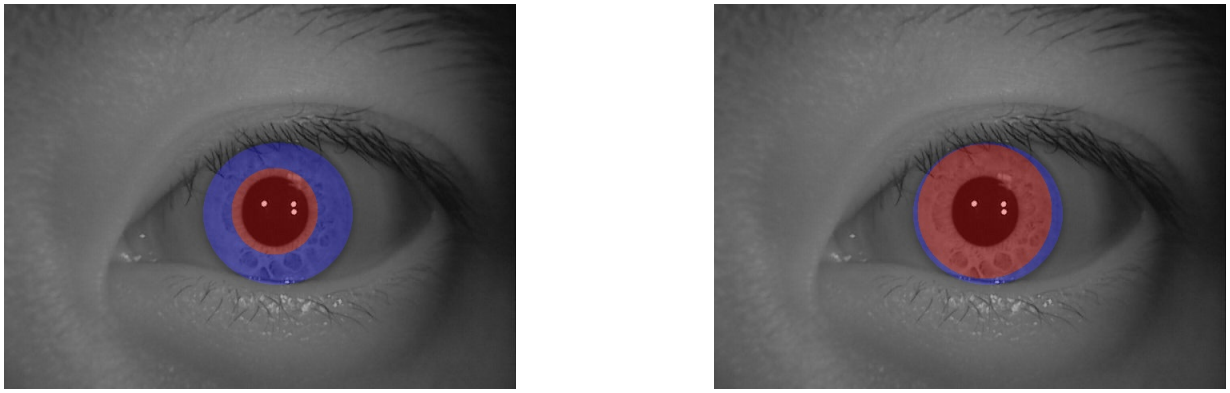
We freeze the weights of all layers in each base model to preserve the high-quality, pre-trained feature detectors in these models. This measure ensures the pre-existing learned features are not distorted during the training phase of our iris recognition task.

We enhance the base models by appending our custom layers, including a global average pooling 2D layer for dimensionality reduction, dense layers for learning task-specific features, and batch normalization and dropout layers for model regularization and overfitting prevention. These modifications allow the models to effectively capture intricate patterns within iris images, thereby contributing to the overall performance of our iris recognition system.

By harnessing the power of the pre-trained models, our feature extraction strategy is not only enables efficient recognition of iris patterns but also maximizes the potential of our iris recognition system.

C. Matching

The matching process compares features extracted from the previous state. For the identification task, the closet iris image is assigned to the target class against the feature of all iris images in the database. Here, the classification function is softmax. It is an adaptation of logistic regression for multiclass classification. The output layer of the transfer learning model utilized is a softmax layer, which produces



(a) (b)

Fig. 4: A comparative demonstration using an image from the CASIA-Iris-Lamp dataset: (a) illustrates the segmentation by the existing method (red circle) against the ground truth (blue circle) with IoU = 35.17%, (b) displays the segmentation by our proposed method (red circle), compared against the ground truth (blue circle) with IoU = 85.13%. The two panels highlight the differences in segmentation outcomes between the baseline and the proposed method.

a vector representing the probability distributions of each class in our database. Each element in this vector represents the probability that the input iris image belongs to the corresponding class.

For every input iris image, the class with the highest predicted probability is selected as the matching class. This selection process inherently provides a confidence score for the recognition process, offering insight into the reliability of the match.

The matching phase is deemed complete when all input iris images have been assigned a matching class from the database, effectively determining the individual's identity. The efficiency and accuracy of iris matching are significantly enhanced by using softmax classification, thereby contributing substantially to the performance of the iris recognition system.

IV. Experiment and Result

A comprehensive comparative analysis evaluates the impact of combining the UGVs with traditional segmentation techniques. The comparative framework was devised to clearly distinguish between the results achieved with and without applying UGVs.

The experiments were carried out by using the CASIA-Iris-Lamp dataset [14]. This dataset is a well-regarded benchmark in iris recognition research, and it provides a diverse range of iris images, which allows us to thoroughly test our approach under various conditions. The test set used for testing was only between $10 - 20$ % of the entire dataset.

Accuracy was the primary metric used to measure the performance of our iris recognition framework. The evaluation was carried out using the pre-trained model ResNet50 with transfer learning implemented for feature extraction. Experiments were conducted both with the standard application of these models and with the inclusion of the proposed UGVs framework in the segmentation stage.

TABLE I: Comparison between the proposed method using unit gradient vectors and the baseline for iris recognition.

Performance Metric	Baseline [3]	**Proposed method**
Accuracy	80.38	**82.02**
FAR	0.10	**0.09**
FRR	6.95	**6.44**
ERR	2.25	**1.40**

Table I shows the comparative results, providing valuable insights into the effectiveness of the proposed method. An observable improvement in accuracy is evident when the UGVs are incorporated into the segmentation process across all employed models. The baseline accuracy without the UGVs stands at 80.34%, while the proposed method using the UGVs achieves a higher accuracy of 82.02%, indicating a performance enhancement of 2.04%. A false acceptance rate (FAR), the proportion of invalid inputs incorrectly accepted, decreased marginally from 0.10% to 0.09% when the UGVs technique is employed. In the context of iris recognition, a lower FAR implies that the system is less likely to falsely identify someone, indicating that using the UGVs slightly improves the chance of rejecting an incorrect iris match by 10%. False Rejection Rate (FRR), the proportion of valid inputs erroneously rejected, reduced from 6.95% to 6.44%. This suggests that using the UGVs enhances the chance of accepting a correct iris match by 7.33%. A lower FRR means the system is less prone to reject an accurate identity. An equal error rate (ERR) is a point where both FAR and FRR are equal and serves as a convenient summary statistic. The ERR fell from 2.25% to 1.40%, indicating an overall performance improvement of 37.78% in the system when utilizing UGVs.

Simultaneously, the results revealed a reduction in table I when the UGV-enhanced segmentation method was employed.

This reduction indicates an increased model efficiency and a better generalization of unseen data, further underscoring the benefits of integrating UGVs with traditional segmentation techniques.

The enhancement was observed to extend beyond the performance metrics. Fig. 4, which illustrates a comparison of the intersection over union (IoU) with the ground truth and the segmentation, further highlights the enhancements from the proposed method. This comparison result clearly explains the effectiveness of the UGV-enhanced segmentation techniques.

Fig. 4b, where UGVs were used, the IoU score reached a high of 85.13%. This is a marked improvement compared to the score of 35.17% achieved without employing UGVs, as shown in Fig. 4a.

Our findings demonstrate that the combined approach enhances segmentation precision and provides a more robust base for subsequent feature extraction and matching processes. Improving the initial iris segmentation can significantly enhance the overall performance of iris recognition systems.

V. Discussion

Our research into integrating the use of the UGVs with Daugman's algorithm for iris segmentation has offered the potential to significantly enhance iris recognition system performance. However, as with any method, there are limitations that must be acknowledged.

One of the main limitations of our approach is observed in scenarios where the initial segmentation closely resembles the iris. In this case, our approach does not modify the circle, which could limit potential improvements. It's important to note that this behaviour is inherent to our method, which is primarily designed to improve less accurate segmentation.

Another potential limitation is observed when the approach expands the circle larger than the original size of the iris. In certain situations the circle closely resemble the iris, this enlargement could include non-iris regions, leading to a decrease in accuracy. Careful handling is required to ensure that the circle does not grow excessively, potentially including non-relevant areas in the segmentation.

Despite these limitations, our method is effective when the initial segmentation is not accurate as shown in Fig. 4a. Our approach can significantly increase the size of the circle since it provides edge direction information. If we compare the segmentation with the ground truth, the proposed method increases the intersection area as shown in Fig. 4b. This result improves segmentation accuracy and, consequently, enhances iris recognition performance.

These findings underscore the potential of our approach, despite its limitations. As with all research, understanding these limitations is key to future improvements.

VI. Conclusion

This paper presented the integration of the UGVs with a traditional segmentation technique to improve the performance of iris recognition systems. This work contributed to the iris recognition process with a more robust iris localization. The proposed method using the UGVs contained the direction information of the edges in the image. Thus, incorporating these vectors into the computation could help emphasize areas with a strong directional change, potentially improving the precision of edge detection. Specifically, our method enhances accuracy by 2.04%, reduces the FAR by 10%, and decreases the FRR by 7.33%. Furthermore, the ERR was decreased by 37.78%. These improvements contribute to more robust iris segmentation and thereby, the overall advancement of security and efficiency in iris recognition systems. However, some inaccuracies of the proposed method for iris segmentation still need refinement. In future work, we will explore a larger dataset, images from different sensors, and low-quality images.

Acknowledgment

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU), and National Research Council of Thailand (NRCT) under the TAIST Tokyo Tech Program.

References

[1] A. K. Jain, A. A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE transactions on information forensics and security*, vol. 1, no. 2, pp. 125–143, 2006.

[2] K. Bowyer and M. Burge, *Handbook of Iris Recognition*. Springer, 2016.

[3] D. A. Roy and U. S. Soni, "Iris segmentation using daughman's method," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 2668–2676.

[4] J. A. Ridha and J. H. Saud, "Iris segmentation approach based on adaptive threshold value and circular hough transform," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, 2020, pp. 32–37.

[5] J. Daugman, "New methods in iris recognition," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 37, pp. 1167–75, Nov. 2007.

[6] Q. Zhang, H. Li, Z. Sun, and T. Tan, "Deep feature fusion for iris and periocular biometrics on mobile devices," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2897–2912, 2018.

[7] C. Bhawna and S. Shukla, "Iris recognition system using canny edge detection for biometric identification," *International Journal of Engineering Science and Technology*, vol. 3, Jan. 2011.

[8] Y. Chen, W. Wang, Z. Zeng, and Y. Wang, "An adaptive CNNs technology for robust iris segmentation," *IEEE Access*, vol. 7, pp. 64 517–64 532, 2019. DOI: 10.1109/ACCESS.2019.2917153.

[9] T. Kondo and W. Kongprawechnon, "A matching technique using gradient orientation patterns," *Science & Technology Asia*, pp. 41–55, 2009.

[10] P. Boonsieng, T. Kondo, and W. Kongprawechnon, "Unit gradient vectors based motion estimation techniques," *ECTI Transactions on Electrical Engineering, Electronics, and Communications*, vol. 9, no. 2, pp. 246–254, 2011.

[11] V. Chhor and T. Kondo, "Illumination-invariant line detection with the gray-scale hough transform," in *2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, 2015, pp. 19–23. DOI: 10.1109/ICCIS.2015.7274541.

[12] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004.

[13] R. Haralick, S. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-9, pp. 532–550, Aug. 1987.

[14] C. I. I. D. S. Team, *CASIA iris image databases*, CAS Institute of Automation. [Online]. Available: http : / / biometrics.idealtest.org/.

[15] J. Daugman, "How iris recognition works," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 21–30, 2004.

[16] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Prentice Hall, 2007.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[18] W. Härdle, *Applied Multivariate Statistical Analysis*. Springer, 2007, p. 344, "Section title: 13.2 Chi-square Decomposition".

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[20] T. Le-Tien, H. Phan-Xuan, P. Nguyen-Duy, and L. Le-Ba, "Iris-based biometric recognition using modified convolutional neural network," in *2018 International Conference on Advanced Technologies for Communications (ATC)*, 2018, pp. 184–188.

[21] A. Singh, A. Pandey, M. Rakhra, D. Singh, G. Singh, and O. Dahiya, "An iris recognition system using cnn vgg16 technique," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2022, pp. 1–6.

[22] R. Nijhawan, N. Sule, M. Verma, B. Sharma, and I. Bansal, "Analysis of coloboma defected eyes using automated pre-trained cnn models," in *2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 2022, pp. 1–5.

[23] O. Elharrouss, Y. Akbari, N. Almaadeed, and S. Al-ma'adeed, "Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches," in Jun. 2022.

[24] K. Sonkar, R. Rani, and R. Dhir, "Iris recognition using transfer learning of inception v3," in *Applications of Machine intelligence in Engineering*, 1st ed. CRC Press, 2022, p. 10, ISBN: 9781003269793.

[25] Y. Zhuang, J. H. Chuah, C. O. Chow, and M. G. Lim, "Iris recognition using convolutional neural network," in *2020 IEEE 10th International Conference on System Engineering and Technology (ICSET)*, 2020, pp. 134–138.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *The 3rd International Conference on Learning Representations*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking

Tsun-Hin Cheung and Kin-Man Lam
Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong
E-mail: tsun-hin.cheung@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract* — Automatic fact-checking plays a crucial role in combating the spread of misinformation. Large Language Models (LLMs) and Instruction-Following variants, such as InstructGPT and Alpaca, have shown remarkable performance in various natural language processing tasks. However, their knowledge may not always be up-to-date or sufficient, potentially leading to inaccuracies in fact-checking. To address this limitation, we propose combining the power of instruction-following language models with external evidence retrieval to enhance fact-checking performance. Our approach involves leveraging search engines to retrieve relevant evidence for a given input claim. This external evidence serves as valuable supplementary information to augment the knowledge of the pretrained language model. Then, we instruct-tune an open-sourced language model, called LLaMA, using this evidence, enabling it to predict the veracity of the input claim more accurately. To evaluate our method, we conducted experiments on two widely used fact-checking datasets: RAWFC and LIAR. The results demonstrate that our approach achieves state-of-the-art performance in fact-checking tasks. By integrating external evidence, we bridge the gap between the model's knowledge and the most up-to-date and sufficient context available, leading to improved fact-checking outcomes. Our findings have implications for combating misinformation and promoting the dissemination of accurate information on online platforms. Our released materials are accessible at: https://thcheung.github.io/factllama.**

I. INTRODUCTION

With the continuous growth of social media and online communication, we are experiencing a deluge of information through these channels, raising the urgent need to distinguish facts from fiction. Unfortunately, these platforms provide almost everyone with the ability to spread misinformation or fake news that often goes viral, leading to the widespread dissemination of inaccurate or false information to a global audience. Misinformation can take different forms, ranging from fabricated news stories to manipulated images or videos aimed at swaying public opinion [1].

To counteract this alarming trend, the research community is continuously developing new and innovative approaches to tackle the problem of identifying and correcting misinformation. Artificial intelligence (AI)-assisted fact-checking is one such approach and has been gaining attention in recent years [2]. It has the potential to automate the laborious and time-consuming process of manually verifying facts, while enabling quick dissemination of accurate information. With the recent rise in popularity of Large Language Models (LLMs) [3], [4] in Natural Language Processing (NLP) tasks, such as machine translation, text classification, and data extraction, they have also been used for fake news detection [5].

Despite the impressive capabilities of these LLMs, a significant limitation of these language models is their reliance on pre-existing knowledge, which may not always be up-to-date or sufficient. In the context of fact-checking, the reliance of language models solely on their internal knowledge raises concerns about their ability to accurately assess the veracity of claims, especially when faced with rapidly evolving information [6]. To address this limitation, it becomes imperative to consider external knowledge sources that provide updated and reliable information in recent fact-checking algorithms [6]–[9].

This paper aims to enhance the fact-checking capabilities of instruction-following language models by leveraging external evidence. We propose a method that combines the power of pretrained language models with the retrieval of relevant external evidence from search engines. By integrating this external evidence during the instruct-tuning process, we aim to augment the knowledge of the language model, enabling it to make more accurate predictions. The contributions of our work are summarized as follows:

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

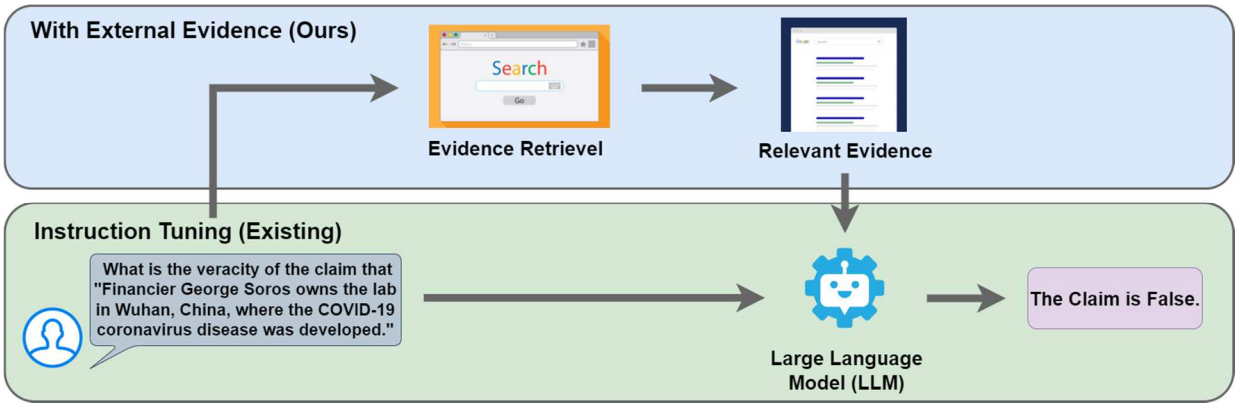


Fig. 1. Our approach for automatic fact-checking with external evidence retrieved from search engine.

- **Introducing instruct-tuned language models for fact-checking**: We propose the application of instruct-tuned language models, i.e., LLaMA, for automatic fact-checking tasks, expanding their scope beyond language generation.
- **Addressing the limitations of instruct-tuned models**: We identify the limitations of instruct-tuned language models in fact-checking due to outdated knowledge, and propose the integration of external evidence to enhance their accuracy and reliability.
- **Proposing a method for incorporating external knowledge**: We present a novel approach that combines pretrained language models with external evidence retrieval from search engines, augmenting the knowledge base for fact-checking.
- **Achieving state-of-the-art performance**: Through experiments on the RAWFC and LIAR datasets, we demonstrate that our method achieves state-of-the-art performance in fact-checking tasks.

II. Related Work

A. *Automatic Fact-Checking*

In recent years, the task of automatic fact-checking has gained significant attention due to the growing concern about the spread of false information. Many studies have been carried out using different machine learning paradigms. This section provides an overview of related works that focus mainly on machine learning-based, deep learning-based, and transformer-based approaches to automated fact-checking.

Machine learning algorithms have been widely used to build models for automated fact-checking. In traditional machine learning methods, the features are manually engineered and used to train a classifier. Several studies have explored and compared different machine learning-based methods for automated fact-checking. Hassan et al. [10] proposed a solution based on the Support Vector Machine algorithm that performs a multi-level fact-checking process on news claims. Their results indicate that the proposed solution outperforms several traditional methods in terms of accuracy.

Deep learning-based methods have been successful in many Natural Language Processing (NLP) tasks, including automated fact-checking [11]. Deep neural networks can automatically learn complex features from raw data, including word embeddings and network architectures that operate on these embeddings. In 2018, Popat et al. [12] proposed a neural network-based solution to verify claims by directly comparing them to the corresponding evidence. The model utilizes a hierarchical attention mechanism to extract relevant evidence and compares it to the given claim to make a prediction. Another study presented a fact-checking system that combines deep neural networks with natural language generation to generate explanations for the model's output [9].

The Transformer is a recent innovation in deep learning that has shown great potential in several natural language processing tasks, including automated fact-checking. Transformer-based models [13], [14] employs self-attention mechanisms that enable them to capture long-range dependencies between input tokens. Recently, several studies have demonstrated the effectiveness of Transformer-based models in automated fact-checking. For instance, Kotonya et al. [6] proposed a Transformer-based solution that operates on pairs of claims and evidence. Their model utilizes an attention mechanism that allows it to focus on the most relevant pieces of evidence when verifying a claim. The authors reported that their model achieved high accuracy on a publicly available fact-checking dataset. Another study utilized a Transformer model to predict the veracity of a claim by performing joint

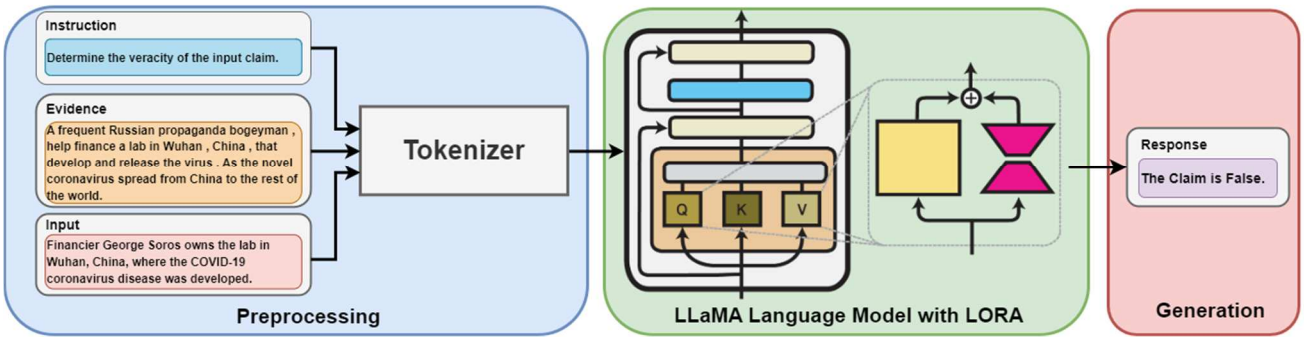


Fig. 2. Illustration of optimizing instruction-following models with external evidence using Low Rank Adaption Tuning (LORA) [22].

reasoning over both textual evidence and commonsense knowledge [15].

B. Instruction-Following Language Models

Instruction-following language models have emerged as a promising and novel approach for many NLP tasks [16]. These models can follow a set of instructions to perform a specific task, and perform classification and generation tasks.

InstructGPT [17] is an instruction-following language model that combines the power of the GPT architecture with a sequence of natural language instructions to perform a specific task. In the context of automated fact-checking, the model is presented with a claim. It then uses a set of instructions to perform verification for its output. The InstructGPT model has been evaluated on several publicly available commonsense reasoning datasets and has been shown to achieve high accuracy.

Self-Instruct [18] is an extension of the InstructGPT model that leverages self-supervised learning to improve its performance on a variety of natural language processing tasks. The model uses a simple heuristic algorithm that generates its own sequence of instructions and samples from the task pool. This self-supervised approach has been shown to improve the model's accuracy on many NLP tasks.

ChatGPT [19] is another instruction-following language model that can perform a wide range of natural language processing tasks, including automated fact-checking. The model leverages a conversational approach to fact-checking by generating questions that a human fact-checker would ask and answering them on its own. To achieve this, the model follows a set of instructions when interacting with input claims. New Bing from Microsoft[1] is a recently released industrial product that is powered by GPT and Bing search engine. However, both ChatGPT and New Bing are proprietary models, which cannot be tuned by the general public and researchers for downstream uses. Moreover, limited access to ChatGPT restricts the number of times that an account can perform fact-checking, especially for large-scale use.

Stanford Alpaca [20] is an open-sourced instruction-following language model based on the paradigm of open sequence-to-sequence model, called LLaMA [21], that is capable of performing a wide range of natural language processing tasks. In the context of fact-checking, the model can be used presented with a claim. The model derives a sequence of rules from the task specification and applies them to predict whether a given claim is factual or not. In our work, we aim to improve the fact-checking capability of LLaMA with the help of external knowledge.

C. Low-Rank Adaptation Tuning

Low-Rank Adaptation (LORA) [22] is a parameter-efficient fine-tuning method that focuses on identifying the most important neurons in a pretrained language model and updating only those neurons to improve performance on a target dataset. The "Low-Rank" component in LORA involves low-rank matrix factorization in the fine-tuning process, which compresses pretrained model parameters into a lower-dimensional space, while preserving as much of the original information as possible. The resulting lower-dimensional parameter space can be fine-tuned more efficiently and effectively using fewer target samples. In our work, we leverage the LORA tuning in instruct-tuning LLaMA, because LORA tuning can reduce GPU memory to 16 GB during tuning, when compared to 64GB during fine-tuning [20].

III. METHODOLOGY

In this section, we describe the methodology for instruct-tuning language models with external evidence for automatic fact-checking. Our proposed method consists of two key components: (1) the generation of instruction-evidence-input

[1] https://www.bing.com/new

claim samples and (2) instruct-tuning of a generative pretrained language model, i.e., LLaMA, using these samples.

A. Instruction-Evidence-Input Generation

Generation of instruction-evidence-input samples. In order to train the instruct-tuned language model, we first generate instruction-evidence-input samples. We combine the instruction, evidence, and input claim into a single sequence, with appropriate special tokens to differentiate them. The instruction provides guidance on how to incorporate the evidence for fact-checking, while the evidence consists of relevant information, retrieved from search engines, using the Google API. The input claim represents the claim that needs to be fact-checked. Similar to previous studies [8], [12], we filtered out those evidence that are from fact-checking sites or published before the claim is recorded in the fact-checking sites.

Evidence collection pipeline. Our evidence collection pipeline utilizes the Google API to retrieve relevant evidence given the textual inputs. We formulate queries, based on the input claim, and use the API to search for information from reputable sources. The retrieved evidence includes sentences or paragraphs that are likely to contain relevant information related to the claim. This external evidence serves as valuable supplementary information for the fact-checking process.

Converting factual classification to sequence-to-sequence problem. To convert the factual classification task into a sequence-to-sequence problem suitable for generative transformer models, we frame the task as text generation for automatic fact-checking. We treat the input claim as the source sequence and the fact-checking result as the target sequence. The instruct-tuned language model learns to generate a fact-checking response given the input claim and the relevant evidence. This approach allows the model to capture the nuances and context of the claim and produce more accurate fact-checking outcomes.

B. Tuning Pretrained Instruction-Following Models with External Knowledge

In this study, we instruct-tune the pretrained LLaMA [21] model using the LORA algorithm [22]. Our approach not only takes the text claim as input for factual classification, but also the retrieval evidence.

Optimization of language model. During instruct-tuning, we aim to optimize our LLaMA model's parameters θ to minimize a loss function that measures the difference between the predicted fact-check results and the ground truth of the training dataset. Suppose we have a set of input-output pairs (x, y) in the training set, where x is the instruction-evidence-input to be input to the LLaMA model, and y is the corresponding label for that claim. Suppose $f(x;\theta)$ denotes the output, i.e., predicted veracity of the input claim, of the LLaMA with parameters θ for input x. We define the loss function as follows:

$$L(\theta) = \sum(y - f(x;\theta))^2. \quad (1)$$

LORA tuning. Our goal is to find the optimal values for the parameters of the LLaMA model θ that minimize the loss function $L(\theta)$. To achieve this, we leverage the LORA algorithm, which involves a low-rank approximation of the parameter matrix θ. As a result, it reduces the number of trainable parameters and improves generalization, while preserving most of the information contained in the original parameters. Specifically, LORA compresses θ into a low-rank matrix product:

$$\Omega = UV, \quad (2)$$

where U and V are matrices of lower rank than θ. The LORA algorithm updates the parameters θ by adding a regularized factorization of θ to the existing parameters, as follows:

$$\theta' = UV, \quad (3)$$

$$\theta' = \theta + (\Omega - \theta), \quad (4)$$

where θ' is the updated parameter matrix.

IV. Experimental Setups and Results

A. Datasets

We use two publicly available datasets, namely the RAWFC [8] and LIAR [23] datasets, for evaluating our proposed approach. The RAWFC dataset is a collection of claims related to factual verification tasks, consisting of 2,012 claims with supporting evidence, labelled either true, false, and half-true, from Snopes.com. The LIAR dataset is a dataset of political statements fact-checked by PolitiFact.com, consisting of 12,836 short statements, each labelled as true, mostly true, half-true, barely true, false, or pants-on-fire. For fair comparison to other methods, we use the same split released in [8].

B. Experimental Setup

We run the experiments using the smallest LLaMA model, i.e., LLaMA-7B [21] model, in our experiment, as it is the

largest model that fits our hardware setup. We evaluate our proposed approach for automatic fact-checking using three standard evaluation metrics: Precision Scores, Recall Scores, and F1 Scores. To instruct-tune the LLaMA model, we trained the models for 3 epochs with a mini-batch size of 32. We employ the Adam optimizer with an initial learning rate of 10-4 and a linear learning decay from the initial value to 0. To avoid overfitting, we applied a dropout rate of 0.05. Our models were implemented in PyTorch [24] and HuggingFace [25]. All experiments are conducted on two GeForce RTX 2080 Ti GPUs.

C. Comparison to Other Methods

Table I and Table II provide the evaluation results of various methods on two different fact-checking datasets, i.e., RAWFC and LIAR. The methods are compared based on precision, recall, and F1-score, which are commonly used metrics to assess the performance of classification tasks. The baselines that we compared include SVM [26] , which utilizes bag-of-words features for fake news detection. CNN [23] incorporates metadata features to enhance representation learning. RNN [11] learns representations from word sequences without relying on external resources. DeClarE [12] combines word embeddings from the claim, report, and source to assess the credibility of the claim. dEFEND [7] employs a GRU-based model for veracity prediction, providing explanations. SentHAN [6] represents each sentence based on coherence and semantic conflicts with the claim. SBERT-FC [6] utilizes SentenceBERT (SBERT) for encoding and identifies fake news based on the top-ranked sentences. GenFE [9] and GenFE-MT [9] detect fake news independently or jointly with explanations in a multi-task setup. CofCED [8] is a Coarse-to-fine Cascaded Evidence-Distillation neural network for explainable fake news detection based on such raw reports, alleviating the dependency on fact-checked ones.

From Table I for the RAWFC dataset, it can be observed that traditional machine learning methods, like SVM, CNN, and RNN, achieve moderate results in terms of precision, recall, and F1-score. However, more advanced models, such as DeClarE, dEFEND, sentHAN, SBERT-FC, GenFE, GenFE-MT, and CofCED, outperform the traditional methods, particularly CofCED, which achieves the highest F1-score of 0.5107.

Interestingly, LLaMA without tuning, i.e., zero-shot prediction, performs relatively poorly compared to the other methods. However, when Instruct-tuning is applied, there is a significant improvement in performance, particularly when external knowledge is incorporated. Instruct-tuned LLaMA with external knowledge achieves the highest F1-score of 0.5565, surpassing all other methods and demonstrating the effectiveness of leveraging external evidence.

On the evaluation on the LIAR dataset, as shown in Table II, similar patterns can be observed. Traditional machine learning methods show relatively low performance, while more advanced models exhibit better results. CofCED achieves the highest F1-score of 0.2893, indicating its effectiveness in fact-checking on the LIAR dataset.

Methods	Precision	Recall	F1
SVM [26]	0.3233	0.3251	0.3171
CNN [23]	0.3880	0.3850	0.3859
RNN [11]	0.4135	0.4209	0.4039
DeClarE [12]	0.4339	0.4352	0.4218
dEFEND [7]	0.4493	0.4326	0.4407
sentHAN [27]	0.4566	0.4554	0.4425
SBERT-FC [6]	0.5106	0.4592	0.4551
GenFE [9]	0.4429	0.4474	0.4443
GenFE-MT [9]	0.4564	0.4527	0.4508
CofCED [8]	0.5299	0.5099	0.5107
LLaMA (w/o tuning) [20]	0.3350	0.3255	0.2643
FactLLaMA (Instruct-tuning w/o external knowledge)	0.5376	05400	0.5376
FactLLaMA (Instruct-tuning with external knowledge)	**0.5611**	**0.5550**	**0.5565**

Table I. Results on the RAWFC dataset.

Methods	Precision	Recall	F1
SVM [26]	0.1578	0.1592	0.1534
CNN [23]	0.2258	0.2239	0.2136
RNN [11]	0.2436	0.2120	0.2079
DeClarE [12]	0.2286	0.2055	0.1843
dEFEND [7]	0.2309	0.1856	0.1751
sentHAN [27]	0.2264	0.1996	0.1846
SBERT-FC [6]	0.2409	0.2207	0.2219
GenFE [9]	0.2801	0.2616	0.2649
GenFE-MT [9]	0.1855	0.1990	0.1515
CofCED [8]	0.2948	0.2955	0.2893
LLaMA (w/o tuning) [20]	0.1587	0.2069	0.1224
FactLLaMA (Instruct-tuning w/o external knowledge)	0.3232	0.3157	0.2998
FactLLaMA (Instruct-tuning with external knowledge)	**0.3246**	**0.3205**	**0.3044**

Table II. Results on the LIAR dataset.

Once again, LLaMA without tuning performs poorly, but instruct-tuning leads to substantial improvements.

Incorporating external knowledge in the instruct-tuning process further enhances the performance, with LLaMA, Instruct-tuning and external knowledge achieving the highest F1-score of 0.3044.

In summary, the evaluation results from both datasets highlight the superiority of advanced models over traditional machine learning methods in fact-checking tasks. The instruct-tuning approach, particularly when combined with external knowledge, consistently outperforms other methods, showcasing the value of leveraging external evidence for accurate fact-checking. These findings emphasize the importance of staying updated with the latest information and leveraging advanced techniques to effectively combat the spread of misinformation.

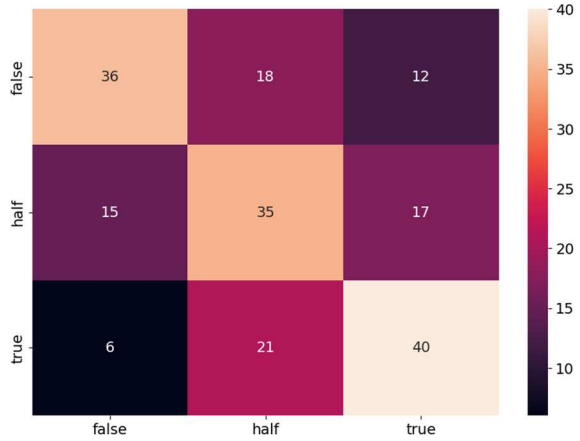


Fig.3 Confusion matrix on the RAWFC Dataset.



Fig.4 Confusion matrix on the LIAR Dataset.

D. Confusion Matrices

Figures 3 and 4 present the confusion matrices for the RAWFC and LIAR datasets, respectively. The rows and columns in the figures represent the ground-truth and predictions, respectively.

In Figure 3, it is evident that the model can effectively distinguish between the TRUE and FALSE labels. However, classifying the HALF-TRUE label proves to be more challenging for the model. This difficulty arises because both the HALF-TRUE and FALSE labels contain misinformation, albeit with differing degrees of accuracy. Moving to Figure 4, we observe that the model shows clear classification performance for the TRUE and PLANT-FIRES classes, compared to the other classes. However, it struggles to accurately classify the BARELY-TRUE, HALF-TRUE, and MOSTLY-TRUE classes. This difficulty arises from the fact that items in these classes contain a mixture of true and false information, making it a subjective task for both humans and machines to classify them accurately without specialized expertise.

V. CONCLUSION

In conclusion, this research highlights the crucial role of automatic fact-checking in combating the spread of misinformation online. While Large Language Models (LLMs) and Instruction-Following variants, like InstructGPT and Alpaca, have demonstrated remarkable performance in various natural language processing tasks, their potential lack of up-to-date or sufficient knowledge can lead to inaccuracies in fact-checking. To address this limitation, we proposed a method that combines pretrained language models with external evidence retrieval, resulting in enhanced fact-checking accuracy. By leveraging search engines to retrieve relevant evidence for a given claim, we successfully augmented the knowledge of the pretrained language model. Through instruct-tuning an open-source language model called LLaMA, with this external evidence, we achieved more accurate predictions regarding the veracity of input claims. Experimental evaluations on widely used fact-checking datasets, RAWFC and LIAR, showcased that our approach achieved state-of-the-art performance in fact-checking tasks. The integration of external evidence effectively bridged the knowledge gap between the model and the most up-to-date information available, leading to improved fact-checking outcomes. We believe our research has significant implications for combatting misinformation and promoting the dissemination of accurate information on online platforms. In our future work, we plan to generate explanations with these pretrained language models for more general use.

REFERENCES

[1] M. S. Raval, M. Roy, and M. Kuribayashi, "Survey on Vision based Fake News Detection and its Impact Analysis," in *Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2022*, 2022. doi: 10.23919/APSIPAASC55919.2022.9980089.

[2] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A Survey on Automated Fact-Checking," *Trans Assoc Comput Linguist*, vol. 10, 2022, doi: 10.1162/tacl_a_00454.

[3] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 2020.

[4] Google AI, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," *Google AI Blog*, 2022.

[5] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open*, vol. 3, 2022, doi: 10.1016/j.aiopen.2022.09.001.

[6] N. Kotonya and F. Toni, "Explainable automated fact-checking for public health claims," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.emnlp-main.623.

[7] L. Cui, K. Shu, S. Wang, D. Lee, and H. Liu, "dEFEND: A system for explainable fake news detection," in *International Conference on Information and Knowledge Management, Proceedings*, 2019. doi: 10.1145/3357384.3357862.

[8] Z. Yang, J. Ma, H. Chen, H. Lin, Z. Luo, and Y. Chang, "A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2608–2621. [Online]. Available: https://aclanthology.org/2022.coling-1.230

[9] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating fact checking explanations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.656.

[10] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. doi: 10.1145/3097983.3098131.

[11] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2017. doi: 10.18653/v1/d17-1317.

[12] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018. doi: 10.18653/v1/d18-1003.

[13] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.

[15] B. Zhu, X. Zhang, M. Gu, and Y. Deng, "Knowledge Enhanced Fact Checking and Verification," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, 2021, doi: 10.1109/TASLP.2021.3120636.

[16] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Cross-Task Generalization via Natural Language Crowdsourcing Instructions," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022. doi: 10.18653/v1/2022.acl-long.244.

[17] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 27730–27744. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[18] Y. Wang *et al.*, "Self-Instruct: Aligning Language Models with Self-Generated Instructions." 2023.

[19] OpenAI, "GPT-4 Technical Report." 2023.

[20] R. Taori *et al.*, "Stanford Alpaca: An Instruction-following LLaMA model," *GitHub repository*. GitHub, 2023.

[21] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] E. Hu *et al.*, "LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS," in *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[23] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017. doi: 10.18653/v1/P17-2067.

[24] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019.

[25] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[26] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[27] J. Ma, W. Gao, S. Joty, and K. F. Wong, "Sentence-level evidence embedding for claim verification with hierarchical attention networks," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1244.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Auditory Representation Effective for Estimating Vocal Tract Information

Toshio Irino* and Shintaro Doan*
* Faculty of Systems Engineering, Wakayama University, Japan
E-mail: irino@wakayama-u.ac.jp, doan.shintaro@g.wakayama-u.jp

***Abstract*—We can estimate the size of the speakers based on their speech sounds alone. We had proposed an auditory computational theory of the Stabilised Wavelet-Mellin Transform (SWMT), which segregates information about the size and shape of the vocal tract and glottal vibration, to explain this observation. It has been shown that the auditory representation or excitation pattern (EP) associated with a weighting function based on the SWMT, termed the "SSI weight," can account for the psychometric functions of size perception. In this study, we investigated whether EP with SSI weight can accurately estimate vocal tract lengths (VTLs) which were measured by magnetic resonance imaging (MRI) in male and female subjects. It was found that the use of SSI weight significantly improved the VTL estimation. Furthermore, the estimation errors in the EP with the SSI weight were significantly smaller than those in the commonly used spectra derived from the Fourier transform, Mel filterbank, and WORLD vocoder. It was also shown that the SSI weight can be easily introduced into these spectra to improve the performance.**

Index Terms: Speaker size perception, Auditory model, Vocal tract length (VTL), Glottal vibration, Size-shape image (SSI)

I. Introduction

We can recognize phonemes pronounced by children, women, and men despite the large differences in their heights. This indicates that our auditory system can extract and identify phonemes in which variations in the pattern of formant frequencies distinguish vowel types and the fundamental frequency determines the pitch. Speaker information can also be extracted simultaneously. Speech sounds contain information about vocal tract size, which is closely correlated with speaker size [1]. Many psychoacoustic studies have been conducted on size discrimination and phoneme perception from voiced and unvoiced speech sounds (e.g., [2]–[4]; see review in [5]).

Irino and Patterson [6] proposed the Stabilised Wavelet-Mellin Transform (SWMT) as a computational theory to explain how the auditory system estimates size and shape of vocal tract separately from glottal pulse information. As explained in Section II, size estimation from voiced sounds is more difficult than from unvoiced sounds. This is because voiced speech sounds contain information about vocal tract response (filter characteristics) and glottal vibration (source characteristics) as shown in the source filter theory of speech [7], [8]. Therefore, it is necessary to effectively separate this information. An auditory model based on SWMT was proposed to explain the experimental results on size discrimination of both unvoiced and voiced speech sounds[5], [9]. It was demonstrated that introduction of a simple weighting function, referred to as "size-shape image weight (SSI weight)" (see Section II), enabled to explain the results successfully. However, the domain of previous studies was restricted to the explanation of psychometric functions derived using synthetic speech sounds. For practical applications in signal processing, it is necessary to demonstrate that the auditory model can also extract vocal tract information from natural speech sounds.

In the current study, we focus on the estimation of vocal tract lengths (VTLs) of different speakers. This is because the VTLs of sustained vowels can be accurately measured from magnetic resonance imaging (MRI) data [10], [11] and thus provide "ground truth" usable for evaluation. VTL has sometimes been estimated by "auditory motivated" models, including Mel-Frequency Cepstrum Coefficient (MFCC) [12]–[16]. Recent DNN studies [17]–[19] have also used MFCC. Although the computational power has allowed for high VTL estimation accuracy, little question has been raised as to whether MFCC is really effective for vocal tract information estimation, including speech feature extraction. We consider an effective auditory representation to answer this question.

In this paper, we first describe the problem of VTL estimation using the auditory spectrum and our approach. Then, we describe the VTL estimation method and present the evaluation results. We compared the auditory spectrum with the commonly used Fourier spectrum, the Mel filterbank spectrum [12], and the WORLD vocoder spectrum [20]. We also investigated whether introducing the SSI weight into these spectra improves the estimation accuracy. If it is the case, this simple function, which does not require any training data as in DNN, could improve the performance of many speech processing tasks.

II. Auditory process to estimate size

A. Physics of vocal tract and estimation

The main difference between the vowels of males and females lies in the differences in the VTL and fundamental frequency F_o. When VTL is shortened by a factor of $1/\alpha$, the formant frequencies F_1 and F_2 move upward to αF_1 and αF_2. On the logarithmic frequency axis, $\log F_1$ and $\log F_2$ move up to $\{\log F_1 + \log \alpha\}$ and $\{\log F_2 + \log \alpha\}$, respectively. This implies that the logarithmic scale factor, $\log \alpha$, is a constant independent of the formant frequencies. Therefore, the VTL ratio can be estimated using the cross-correlation of the log spectra corresponding to the original and shorted vocal tracts. Auditory spectra derived by gammatone or gammachirp filterbanks [21], [22] are suitable for this purpose

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

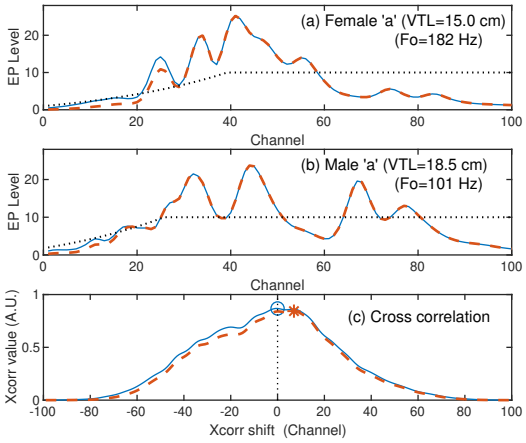


Fig. 1: Excitation patterns (EPs, i.e., GCFB outputs) of the vowel 'a' of a male (a) and a female (b) and their cross-correlation function (c). The horizontal axis for (a) and (b) is the number of GCFB channels. The vertical axis represents the EP level. In (c), the horizontal axis represents the amount of shift in terms of the channel and the vertical axis represents the correlation function value (arbitrary unit). Blue solid line: EP; black dotted line: SSI weight (see Section II-D); red dashed line: SSI-weighted EP. Circle (o) and asterisk (*) in (c) indicate the peaks in the cross correlations of the EPs and SSIweighted EPs, respectively.

because the frequency axis, $\mathrm{ERB}_{\mathrm{Nnumber}}$, is approximately a log frequency axis above 500 Hz [23]. This appears to be an easy task if the spectrum is calculated solely from the impulse response of the vocal tract. However, voiced sounds used in speech communication are derived from the convolution of the impulse response of the vocal tract and the waveform of glottal vibration [7]. This makes the estimation more difficult than expected.

B. Problem when using auditory spectrum

To be more specific, an actual female voice 'a' (VTL = 15.0 cm, F_o = 182 Hz) and a male voice 'a' (VTL = 18.5 cm, F_o = 101 Hz), drawn from a database described in Section III-A, were analyzed with a dynamic compressive gammachirp auditory filterbank (GCFB) [22], [24], [25]. The output level was averaged over a short period. This is commonly called an excitation pattern (EP) [23]. The solid blue lines in Fig. 1 show the EPs for the female voice (a) and male voice (b). The horizontal axis is the GCFB channel number, which is equally spaced from 100 to 8000 Hz on the $\mathrm{ERB}_{\mathrm{N}}\mathrm{number}$ axis [23]; this axis is effectively a logarithmic frequency axis above 500 Hz. In the female voice shown in Fig. 1 (a), prominent peaks are observed in channels 24, 32, and 41. Among these, the peak at channel 41 is associated with a formant, and the spectral shape at higher frequencies is important for VTL estimation. In contrast, the peaks at channels 24 and 32 are associated with the harmonics of F_o (182 Hz), which are resolved harmonics [23]. These additional peaks reduce the

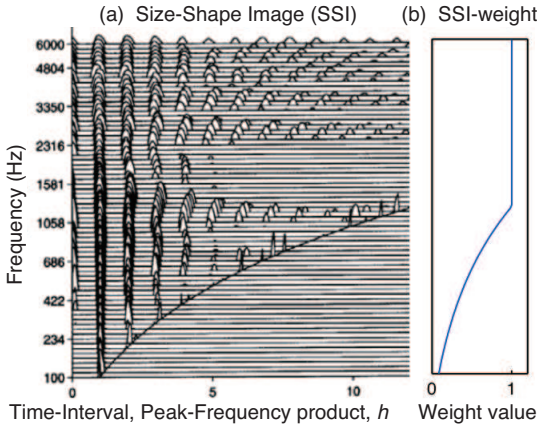


Fig. 2: Size-Shape Image and SSI weight (adapted from [9]). (a) SSI of the vowel 'o' [6]. The horizontal axis represents the product of the time interval and the peak frequency of the auditory filter. The vertical axis is the peak frequency of the auditory filter equally spaced on the $\mathrm{ERB}_{\mathrm{N}}\mathrm{number}$ axis. (b) Weighting function based on active region of SSI. (see also Fig. 8 of Appendix A).

accuracy of the VTL estimation. In contrast, for the male voice in Fig. 1(b), the peaks corresponding to the resolved harmonics are relatively small, and four formant peaks are clearly observed. The cross-correlation between the two EPs is shown by the blue solid line in Fig. 1(c). The peak (blue circle) is obtained at a shift of zero, which implies that the VTLs are the same. This estimation is obviously incorrect because the measured VTL ratio is 1.23 (=18.5/15.0). The problem lies in the fact that the resolved harmonics interfere with estimation. A similar and worse problem occured in the Mel spectrum.

C. Approach from auditory computational theory

We approached this problem using a computational theory in which the auditory system can segregate and extract information about the vocal tract and glottal vibration from speech sounds. Specifically, Irino and Patterson [6] proposed the "Stabilised Wavelet-Mellin Transform" (SWMT) which has been supported by several psychological experiments on size perception (e.g., [2], [5], [9], [26]). The SWMT process is briefly described here (see [6] for details and the upper path in Fig. 8 of Appendix A). In SWMT, the EP derived from GCFB is converted into a two-dimensional "Auditory Image (AI)" by "Strobe Temporal Integration (STI)," which is synchronized with the glottal pulse. The representation of vocal tract response is repeated at the glottal pulse rate (i.e., F_o). One cycle of the response is extracted from the AI to obtain a two-dimensional "Auditory Figure (AF)", which maximally represents information about a single pulse response of the vocal tract. This representation does not exactly correspond to the impulse response but is much closer than the usual spectrum representation obtained from repeated pulse excitation. Then, the AF is transformed into a "size-shape image (SSI)," as shown in Fig. 2(a). Here, the vertical axis represents the peak frequency of the auditory filter, f_p, and the horizontal axis is h, which is the product of the time interval and peak

frequency. SSI is a representation of the single-pulse response and eliminates the response of the adjacent pulse, which is located below the diagonal upright curve. Although it is a good method in principle, developing an STI algorithm to obtain stable images for various speech sounds is not easy. This is because estimating the strobe point or period synchrony information in the EP is a difficult task due to the variability of speech sounds (e.g., [27], [28]).

D. F_o adaptive weighting function, SSI weight

To solve this stability problem, we designed a weight function that corresponds to the area of the active part of the SSI (see also the middle block and the lower path in Fig. 8 of Appendix A). Figure 2 (b) shows a function named "SSI weight" [5], [9]. This one-dimensional function is directly applicable to EP via simple multiplication. The SSI weight (w_{SSI}) is defined as

$$w_{SSI}(f_p, F_o) \quad = \quad \min(\frac{f_p}{h_{max} \cdot F_o}, 1), \tag{1}$$

where h_{max} is the upper limit of h on the horizontal axis of the SSI and determines the area of information extracted from the SSI, as shown in Fig. 2(a). f_p is the peak frequency of the analysis filer. F_o is the fundamental frequency at the time of the analysis. When F_o is not determined, as in unvoiced sounds, setting $F_o \simeq 0$ will result in a w_{SSI} of unity across the peak frequency. The SSI weight was used to explain the results of the human size perception experiments [5], [9] and was introduced to a new objective speech intelligibility measure, GESI, to predict the simulated hearing loss sounds [29], [30]. Most importantly, the SSI weight is applicable to any spectral representation because it is a simple weighting function on the frequency axis.

E. Applying the SSI weight to EP

In Fig. 1, we apply the SSI weight (black dotted line) to the original spectrum (blue solid line) to derive the weighted spectrum (red dashed line). In particular, in Fig. 1(a), the first peak of the resolved harmonics of the female voice is effectively suppressed. The red dashed line in Fig. 1(c) shows the cross-correlation function between the SSI-weighted EPs. The peak (red asterisk) is at a shift of 6, which is approximately 1.2 times the frequency and coincides with the VTL ratio (=1.23) between the male and female. Thus, introducing the SSI weight is expected to reduce the difficulty of VTL estimation.

III. VTL ESTIMATION METHOD AND EVALUATION

A. Measured VTL data

The effectiveness of the SSI weight was evaluated using the "ATR vowel speech MRI data" [10], [11], which contains five vowel sounds, $v \{v|`a', `i', `u', `e', `o'\}$, accompanied by accurate VTLs measured from MRI images of sustained vocalizations. As it contained only 13 male prepared data points, we additionally derived six female data points [31], [32]. Therefore, we used data from 19 speakers.

B. Spectral analysis by GCFB

We used the GCFB [22], [24], [25] for VTL estimation, as described in Section II-C. It was set to process 100 channels with a sampling frequency of 48 kHz and a filter center frequency range of 100 – 8000 Hz. Upon audio input, the EP is output with a frame period of 0.5 ms, and the EP spectrogram is calculated. EPs in the range of $\pm$25 ms from the center of the speech data were averaged to obtain the spectral representation "Ep" of the vowel described above. The spectral representation "Ep_{SSI}" was also calculated by applying the SSI weight in Eq. 1 with the F_o estimated by WORLD [20].

C. VTL estimation algorithm

For each of the five vowels, the cross-correlation functions of the EPs were calculated for all combinations of the 19 speakers ($N = 19$). Between the i-th and j-th speakers, the shift in the peak position from the center, c_{ij}, was extracted using the cross-correlation function. Note that the EP was linearly interpolated in advance by a factor of 10 to make the resolution of the peak shift as a 0.1 channel. The resulting shift was assumed to be caused by the VTL difference. When the orders of i and j are swapped, the amount of shift is the same in absolute value and only the sign is reversed. Therefore, the matrix notation of all permutations is represented as

$$\begin{bmatrix} 0 & c_{12} & c_{13} & \cdots & c_{1N} \\ -c_{12} & 0 & c_{23} & \cdots & c_{2N} \\ -c_{13} & -c_{23} & 0 & \cdots & c_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -c_{1N} & -c_{2N} & -c_{3N} & \cdots & 0 \end{bmatrix}.$$

The relative shift for the i-th speaker, S_i, from the average of all speakers is calculated by taking the difference between the vertical and horizontal sums of this matrix.

$$C_j^{row} = \sum_{i=1}^{N} c_{ij}, \qquad C_i^{col} = \sum_{j=1}^{N} c_{ij} \tag{2}$$

$$S_i \quad = \quad (C_i^{row} - C_i^{col})/2N \tag{3}$$

Although this equation is simple, the equivalent value was obtained using an estimation method with a generalized inverse matrix [33], [34]. Since this value is a shift quantity on a logarithmic scale, its exponent yields the VTL, L_i as

$$L_i \quad = \quad \exp(qS_i) \cdot \bar{L} \tag{4}$$

where $\bar{L}$ is the measured VTL averaged across speakers. q is a conversion coefficient that depends on the spectral representation. In this study, q was determined to minimize the squared error between the regression line, calculated from the measured and estimated VTLs for all 19 speakers and five vowels, and the 1:1 identical line.

D. Estimation results

Figure 3 shows scatter plots between the measured and estimated VTLs when using Ep (Fig. 3a) and Ep_{SSI} with $h_{max} = 3.5$ (Fig. 3b). The correlation coefficients for Ep and Ep_{SSI} were calculated for all vowels and speakers. The values were 0.71 and 0.80, respectively. Vowel labels are generally

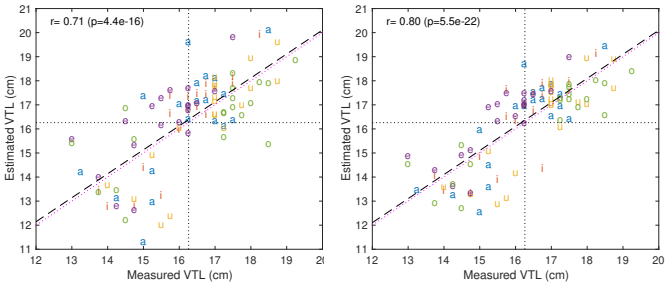


Fig. 3: Scatter plot of estimated VTL ratio against VTL measured from MRI data. The regression line was obtained using all the vowel and speaker data (dashed lines). Left: Ep, right: Ep_{SSI} with h_{max} = 3.5. Color coded for each vowel. Dotted line: 1:1 identical line.

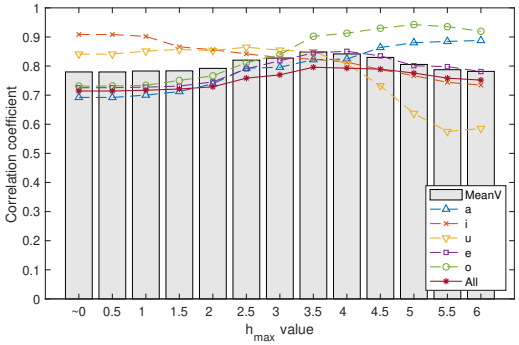


Fig. 4: Correlation coefficients between measured and estimated VTLs for various h_{max} values. Lines: Coefficients for the individual vowel ('a','i','u','e','o') and when using all vowels ("All"). Bar: mean value for the five vowels.

closer to the regression line in Ep_{SSI} than in Ep. This finding indicates that applying the SSI weight to the EP can improve the accuracy of the VTL estimation, as described in Section II-E.

1) The effect of h_{max} on VTL estimnation: Figure 4 shows the correlation coefficients between the measured and estimated VTLs calculated for h_{max} values between 0 and 6 in 0.5 steps. We assumed that the frequency response of the vocal tract filter may be best extracted by properly setting the h_{max} value to reduce the resolved harmonics in the EP shown in Fig. 4, although the h_{max} value was arbitrarily fixed at 5.0 when explaining the results of the human size perception experiments [5], [9]. It is also important to estimate vocal tract information equally well for all vowel types.

The correlation coefficients were calculated for each vowel, as shown by the lines in Fig. 4. When h_{max} was less than 3, the correlation coefficients for 'i' and 'u' (low first formant, F_1) were high, while those for 'a,' 'e,' and 'o' (high F_1) were low. This result implies that the F_o information from the resolved harmonics was not sufficiently suppressed in the low-frequency region. In contrast, when h_{max} exceeded 4, the correlation coefficients for 'a' and 'o' (high F_1) were high, and those for 'i' and 'u' (low F_1) were low. Therefore, F_1 and F_o information were excessively suppressed. The best results were obtained when h_{max} was 3.5, for which the difference between the five vowels was small, and the correlation coefficient obtained from all five vowels ('All') was the highest. In this case, information about the vocal tract and glottal vibration seemed to be properly separated regardless of vowel type.

TABLE I: Abbreviations of the compressed spectrum

spectrum	log	power, P	log+SSI	P+SSI
Fourier	F^{log}	F^{P}	F^{log}_{SSI}	F^{P}_{SSI}
Mel	M^{log}	M^{P}	M^{log}_{SSI}	M^{P}_{SSI}
WORLD	W^{log}	W^{P}	W^{log}_{SSI}	W^{P}_{SSI}

IV. Comparison with commonly used spectra

We compared the estimation performances when using the above auditory representation and commonly used spectral representations to evaluate their effectiveness. A comparison was made with a Fourier spectrum, "F," and a Mel-filterbank (MFB) spectrum [12],"M." We also included a WORLD spectrum [20],"W," because it reduces the effect of F_o by smoothing the frequency distribution and is commonly used in voice conversion as a successor of Tandem-STRAIGHT[35]. The estimation algorithm was identical to that described in Section III-C.

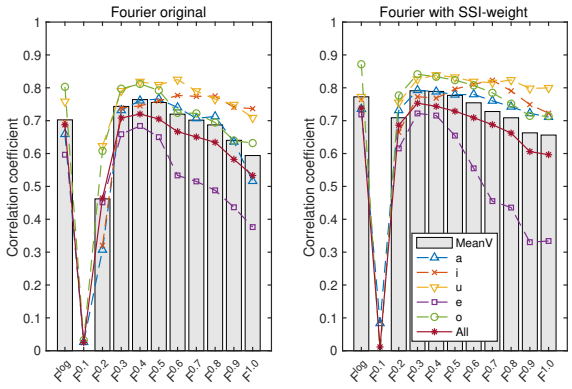
A. Calculation of the spectrum

The short-time Fourier spectrogram of the vowel sound was calculated with a frame length of 25 ms, a hamming window, and a frame shift of 5 ms. The Mel spectrogram was obtained from the STFT spectrogram with a Mel filterbank with 25 filters equally spaced on a Mel frequency axis corresponding to between 100 and 8000 Hz. The WORLD spectrogram was obtained with a default frame rate of 5 ms. Then the obtained amplitude spectrogram, $S(f_p, \tau)$, was subjected to logarithmic compression, $20\log_{10}\{S(f_p,\tau)\}$, and power compression, $S(f_p,\tau)^P \quad \{P \,|\, 0.1 \le P \le 1.0, \text{every } 0.1\}$. The compressed spectrogram in the range of $\pm$25 ms from the center was averaged to obtain the spectrum as calculated in EP (Section III-B). For the Fourier and WORLD spectra, the frequency axis was logarithmically transformed, and a spectrum with 100 channels equally spaced between $\log_{10}(100)$ and $\log_{10}(8000)$ was obtained by linear interpolation. The Mel spectrum with 100 channels was obtained from the 25-channel spectrum via linear interpolation.

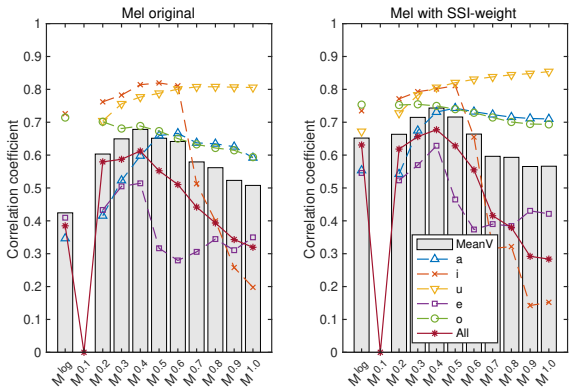
As decribed in section II-D, the SSI weight improved the estimation performance when applied to the EP and was applicable to any spectrum. Therefore, the SSI weight was also applied to these spectra to investigate its effect. Table I presents the abbreviations for each compressed spectrum.

B. Effect of compression on the estimation

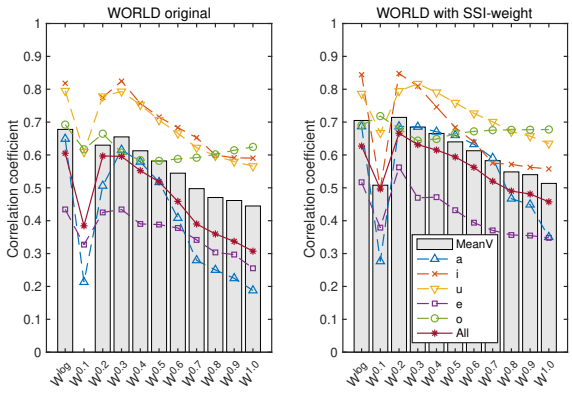
Initially, we sought to identify the log and exponential compressions that yielded the best VTL estimation in each spectrum. The resulting spectra are suitable for comparison with EP.



(a) Fourier spectrum



(b) Mel spectrum



(c) WORLD spectrum

Fig. 5: Correlation coefficients between the measured VTLs and the VTLs estimated from the log and power compressed spectrum. Left panel shows the result when using the original compressed spectrum and the right panel shows the result when using the spectrum with the SSI weight. Lines and bars are the same as in Fig. 4.

Figure 5a shows the correlation coefficients between the measured VTLs and the VTLs estimated from the various compressed Fourier spectra. The correlation coefficients were generally higher when using the spectrum with the SSI weight (left panel) than when using the original spectrum (right panel).The Fourier spectrum with the SSI weight F^{log}_{SSI} yielded the best correlation coefficient and a relatively small variability

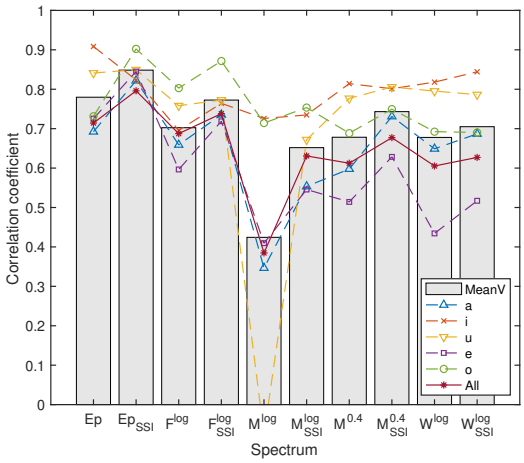


Fig. 6: Correlation coefficients derived from different spectra. Lines and bars are the same as in Fig. 4

between vowels. The F^{log} in the left panel showed high correlation coefficient although it was not the best. We used these data for comparison.

Figure 5b shows the results of the Mel spectrum. The correlation coefficients were generally smaller than those in the Fourier spectrum. The Mel spectra of $M^{0.4}$ and $M^{0.4}_{SSI}$ were the best for each panel. We also included M^{log}, which is the spectrum for the calculation of Mel-frequency cepstrum coefficient (MFCC) [12], and M^{log}_{SSI} in the comparison.

Figure 5c shows the results of the WORLD spectrum. We selected W^{log} and W^{log}_{SSI} for comparison, because they provided the best correlation coefficients on average.

C. Comparison between the EP and various spectra

Figure 6 shows the correlation coefficients for the EP and the spectra selected above. Ep_{SSI} had the highest correlation coefficient and smallest variability among the five vowels. In contrast, M^{log} had the lowest correlation coefficient. Notably, the correlation coefficient was always higher in any spectrum when introducing the SSI weight. Therefore, the SSI weight could improve the estimation of vocal tract information.

These differences were tested statistically. We performed the VTL estimation 10 times using data from 16 speakers after excluding two males and one female from the original 19 speakers at random. Subsequently, the RMS error between the measured and estimated VTLs was calculated. If the VTL estimation is sufficiently good, it should be stable and accurate even if three data points are randomly eliminated. Figure 7 shows the mean and standard deviation and the results of Tukey's HSD multiple-comparison test. The RMS error for Ep_{SSI} was approximately 1 cm, which was significantly smaller than those for the other spectra. In contrast, the RMS error for M^{log} was approximately 3 cm, which was significantly greater than that for the other spectra. More importantly, the results also demonstrate that introducing the SSI weight reduced the error in any spectrum. These reductions were statistically significant ($p < 0.05$), except for the WORLD

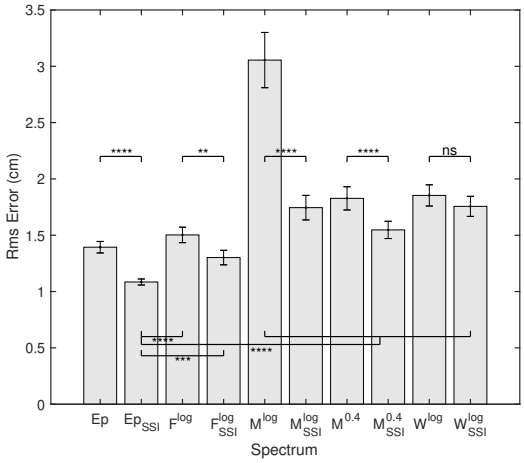


Fig. 7: Mean and standard deviation of the RMS errors in cm when the data of three speakers were randomly excluded 10 times. Tukey's HSD test: $* : p < 0.05; ** : p < 0.01; *** : p < 0.001; **** : p < 0.0001$; ns: not significant. Note that there are many other significant differences, but they are not shown here for simplicity.

spectrum. It is noteworthy that the difference between the errors of M^{log} and M^{log}_{SSI} is extremely large.

D. Some lessons from the results

Suggestions obtained from the results may serve as lessons for future speech signal processing.

1) Effective auditory representation: Many "auditory motivated" models have been proposed for various speech signal processing tasks. Most of them only introduced peripheral frequency analysis, such as the auditory and Mel filterbanks. The current results imply that such a frequency analysis is not sufficient for extracting vocal tract information. The introduction of the SSI weight improved the performance. As shown in Figs. 2 and 8, the SSI weight is a simplified version of the Size-Shape Image in the SWMT, which was proposed as a computational theory of the central auditory process. It is important to introduce knowledge of human auditory processing to derive effective auditory representation.

2) Effectiveness of MFCC: Mel-frequency cepstrum coefficient (MFCC) has been commonly used in many kinds of speech processing after its tremendous success in ASR [12]. MFCC has also been used in VTL estimation [13]–[16], even in recent DNN studies [17]–[19]. However, it has rarely been questioned whether the MFCC is an effective representation for this purpose.

The log Mel-spectrum M^{log} is a basic representation for calculating MFCC. The results in Fig. 7 imply that information about the vocal tract and glottal vibration is not sufficiently separated in M^{log}. As the discrete cosine transform (DCT) is applied across all frequency ranges, all cepstrum coefficients unavoidably contain both types of information. Therefore, the use of MFCC does not seem effective for VTL estimation, even if a state-of-the-art DNN method is used in the back-end. This is because the DNN is required to segregate both types of information embedded in the individual cepstral coefficients before estimating the VTL. Although it would be possible to use a large number of parameters to resolve this, interpretation of the internal representation could be difficult because of complexity. A modified version of the MFCC derived from the log Mel-spectrum associated with the SSI weight may improve the performance and interpretation.

3) Merit and usage of the SSI weight: the SSI weight is applicable to any type of commonly used spectra because it is a simple F_o adaptive function on a frequency axis (Eq. 1 and Fig. 2). It is not necessary to estimate F_o accurately because the SSI weight is less sensitive to the F_o value. the SSI weight can be easily implemented in any speech processing program by adding a few lines and an F_o estimation package such as WORLD [20]. Practically, it has been introduced into an objective speech intelligibility measure, GESI [29], [30], to improve the prediction of both male and female speech sounds.

V. Summary

In this study, we investigated auditory representations, which are effective for estimating vocal tract information. We proposed the use of the SSI weight which is derived from SWMT to segregate information about the vocal tract and glottal pulse from speech sounds. The auditory EP associated with the SSI weight improved the estimation of VTLs measured from the MRI data. Moreover, the estimation error was significantly smaller than when using the commonly used Fourier, Mel, and WORLD spectra. It was also demonstrated that the SSI weight can be easily introduced into these spectra to improve the performance.

Acknowledgments

This research was supported by JSPS KAKENHI Nos. 21H03468 and 21K19794. The authors would like to thank Prof. Kitamura for providing the female MRI-VTL data.

Appendix A. SWMT and auditory spectrogram

Figure 8 shows the relationship between the SWMT (Stabilised Wavelet-Mellin Transform) [6], the auditory spectrogram, and the SSI (Size-Shape Image) weight introduced in Section II-D (see Appendix A in [5] for more details).

The upper path in Fig. 8 shows the signal processing of the SWMT, which is supported by several experiments on size information processing in the auditory system (for example [2], [3], [5]). This is effective in theory but has a problem when applied to practical applications. The process of strobe temporal integration for stabilization and 2-dimensional conversion in SWMT is rather difficult to implement in a computational model. To resolve this problem, we developed the SSI weight (Eq. 1) which is applicable to any spectrographic representation.

The lower path in Fig. 8 shows the analysis using an auditory spectrogram. The auditory filterbank is the same as that in SWMT. The output is processed by windowing and averaging to obtain an excitation pattern (EP). This process is simple and

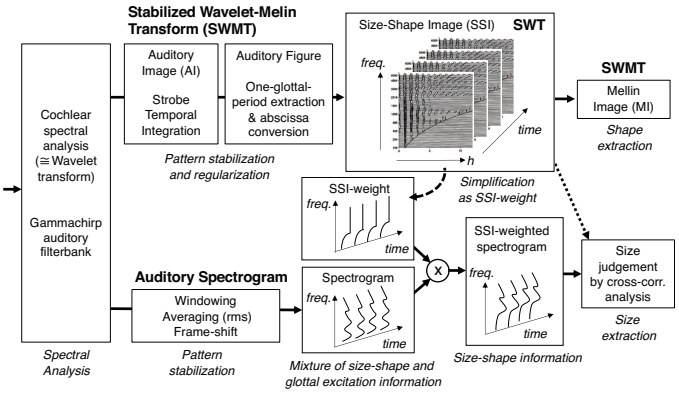


Fig. 8: Relationship between SWMT, auditory spectrogram, and SSI weight [5]. The upper path shows the SWMT process; the lower path shows the auditory spectrogram process along with the SSI weight shown in the middle block. See Appendix A.

produces a stable representation of the spectral information in the sound, although it discards the temporal fine structure that also plays an important role in the auditory perception [36]. An auditory spectrogram is a stream of EPs derived from each frame. As described in Section II, the product of the SSI weight (in the middle block) and the frame-based spectrum is suitable for extracting size information.

REFERENCES

[1] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999. DOI: 10.1121/1.427148.

[2] D. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *The Journal of the Acoustical Society of America*, vol. 117, no. 1, pp. 305–318, 2005.

[3] D. T. Ives, D. R. Smith, and R. D. Patterson, "Discrimination of speaker size from syllable phrases," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3816–3822, 2005.

[4] D. R. Smith, T. C. Walters, and R. D. Patterson, "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3628–3639, 2007.

[5] T. Matsui, T. Irino, R. Uemura, K. Yamamoto, H. Kawahara, and R. D. Patterson, "Modelling speaker-size discrimination with voiced and unvoiced speech sounds based on the effect of spectral lift," *Speech Communication*, vol. 136, pp. 23–41, 2022. DOI: 10.1016/j.specom.2021.10.006.

[6] T. Irino and R. D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-mellin transform," *Speech Commun.*, vol. 36, no. 3-4, pp. 181–203, 2002.

[7] G. Fant, *Acoustic theory of speech production*. 2nd ed., Mouton, 1970.

[8] G. Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 1, no. 1981, pp. 21–37, 1981.

[9] T. Irino, E. Takimoto, T. Matsui, and R. Patterson, "An auditory model of speaker size perception for voiced speech sounds," in *Proc. Interspeech 2017*, ISCA, 2017, pp. 1153–1157.

[10] S. Masaki, M. K. Tiede, K. Honda, *et al.*, "Mri-based speech production study using a synchronized sampling method," *J. Acoust. Soc. Jpn (E)*, vol. 20, no. 5, pp. 375–379, 1999.

[11] H. Takemoto, T. Kitamura, H. Nishimoto, and K. Honda, "A method of tooth superimposition on mri data for accurate measurement of vocal tract shape and dimensions," *Acoust. Sci. & Tech.*, vol. 25, no. 6, pp. 468–474, 2004.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980. DOI: 10.1109/TASSP.1980.1163420.

[13] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE ICASSP 1996*, vol. 1, 1996, pp. 353–356.

[14] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[15] D. R. Sanand, D. D. Kumar, and S. Umesh, "Linear transformation approach to vtln using dynamic frequency warping," in *Proc. Interspeech 2007*, 2007.

[16] A. K. Sarkar, S. Umesh, and S. P. Rath, "Text-independent speaker identification using vocal tract length normalization for building universal background model," in *Proc. Interspeech 2009*, 2009, pp. 2331–2334.

[17] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[18] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 135–140.

[19] Z.-H. Tan *et al.*, "Vocal tract length perturbation for text-dependent speaker verification with autoregressive prediction coding," *IEEE Signal Processing Letters*, vol. 28, pp. 364–368, 2021.

[20] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans Info. Sys.*, vol. 99, no. 7, pp. 1877–1884, 2016. DOI: 10.1587/transinf.2015EDP7457.

[21] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *JASA*, vol. 98, no. 4, pp. 1890–1894, 1995, ISSN: 00014966. DOI: 10.1121/1.414456.

[22] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank.," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2222–2232, 2006, ISSN: 1558-7916. DOI: 10.1109/TASL.2006.874669.

[23] B. C. J. Moore, *An introduction to the psychology of hearing*, 6th. Leiden, The Netherlands: Brill, 2013, ISBN: 9781780520384.

[24] T. Irino, "Hearing impairment simulator based on auditory excitation pattern playback: WHIS," *IEEE access*, vol. 11, pp. 78 419–78 430, 2023. DOI: 10.1109/ACCESS.2023.3298673.

[25] AMLAB, *Github AMLAB-Wakayama*, Accessed: Aug. 26, 2023, 2023. [Online]. Available: https://github.com/AMLAB-Wakayama/.

[26] T. Irino, Y. Aoki, H. Kawahara, and R. D. Patterson, "Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination," *Speech Communication*, vol. 54, no. 9, pp. 998–1013, 2012.

[27] T. Irino, R. D. Patterson, and H. Kawahara, "Speech segregation using an auditory vocoder with event-synchronous enhancements," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2212–2221, 2006.

[28] V. Hohmann, "The period-modulated harmonic locked loop (pm-hll): A low-effort algorithm for rapid time-domain multi-periodicity estimation," *Acta Acustica*, vol. 5, p. 56, 2021.

[29] T. Irino, H. Tamaru, and A. Yamamoto, "Speech intelligibility of simulated hearing loss sounds and its prediction using the Gammachirp Envelope Similarity Index (GESI)," in *Proc. Interspeech 2022*, 2022, pp.3929–3933. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-211.

[30] A. Yamamoto, T. Irino, F. Miyazaki, and H. Tamaru, "Intelligibility prediction of simulated hearing loss sound from individual subject's data using GESI," in *Proc. Spring meeting of Acoust Soc. Jp. (in Japanese)*, vol. 3-4P-7, 2023, pp. 527–530.

[31] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical science and technology*, vol. 26, no. 1, pp. 16–26, 2005.

[32] T. Kitamura, H. Takemoto, and K. Honda, "Acoustic characteristics of solid vocal tracts modeled from atr mri database of japanese vowel production," in *Abstract of Acoustical Society of America meeting*, vol. 123, 2008, p. 3734.

[33] T. Irino, E. Okamoto, R. Nisimura, and H. Kawahara, "Vocal tract length estimation for voiced and whispered speech using gammachirp filterbank," in *Proc. APSIPA ASC 2013*, 2013.

[34] H. Kawahara, T. Kitamura, H. Takemoto, R. Nisimura, and T. Irino, "Vocal tract length estimation based on vowels using a database consisting of 385 speakers and a database with mri-based vocal tract shape information," in *Proc. Interspeech 2014*, 2014.

[35] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. ICASSP 2008*, 2008, pp. 3933–3936.

[36] B. C. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing- impaired people," *Journal of the Association for Research in Otolaryngology*, vol. 9, pp. 399–406, 2008.

Accurate and Practical Query-by-Example Using Multiple Deep Learning Models and Frame Compression Methods

Hikaru Yamaga*, Kazuki Hatakeyama†, Kazunori Kojima*, Shi-wook Lee‡, Yoshiaki Itoh*
* Iwate Prefectural University, Japan, E-mail: y-itoh@iwate-pu.ac.jp
† Graduated from Iwate Prefectural University, Panasonic ITS Co. Ltd., Japan
‡National Institute of Advanced Industrial Science and Technology, Japan, E-mail: s.lee@aist.go.jp

***Abstract*—Recently, studies of spoken term detection (STD) and spoken query STD (SQ-STD), also known as query-by-example (QbE), have been actively pursued. A representative method of QbE is posteriorgram matching using outputs of deep neural networks. However, that method requires much retrieval time and memory size. To address this difficulty, we proposed a maximum likelihood state sequence method (MLSS) for retrieval time reduction. This paper presents a proposal of two methods named "blank-cut (b-cut)" and "frame de-duplication (FDD)" to compress posteriorgram frames, by which we reduce retrieval times and memory sizes. Multiple matching scores are obtained using multiple deep learning models and architectures in the proposed methods. Then they are integrated. We achieved state-of-the-art retrieval accuracy as shown by evaluation experiments using two open test sets of about 30 hr of speech data. Furthermore, the proposed method achieved a retrieval time of less than 1 s and a memory requirement of about 1 GB. These results demonstrated the effectiveness of the proposed method.**

Index Terms: blank-cut, frame de-duplication, query-by-example, maximum likelihood state sequence, spoken term detection

I. INTRODUCTION

Along with the increasing capacity of storage media, studies of spoken term detection (STD) have been actively conducted. STD is a task that seeks specific keywords or phrases consisting of one or more words among large amounts of speech data [1–3]. In fact, NIST STD evaluation [4], Spoken Web Search (SWS) [5], QUery by Example Search on Speech Task (QUESST) [6], and NTCIR Workshop [7, 8] have been held to evaluate STD systems.

In representative conventional STD systems, speech data are first transcribed into subword sequences of text data using an automatic speech recognizer (ASR). Then, the speech data sequences are compared with those of search words (queries) using dynamic time warping (DTW) continuously at a frame level. Some misrecognition might have occurred, thereby leading to deterioration of the retrieval accuracy.

Queries of the STD system are given by text. Spoken query STD (SQ-STD) or query-by-example (QbE), whose queries are given by audio or speech, have also been conducted because of the rapid spread of smartphones and other devices. Earlier studies have also evaluated QbE systems [9–13]. A common method used for QbE is posteriorgram matching method. Posteriorgrams, sequences of posterior probability vectors for an utterance, are speaker-independent features obtained by Deep Neural Networks (DNN). When inputting a feature vector of a single frame to DNN, posterior probabilities corresponding to the states of Hidden Markov Models (HMM) or phonemes and so on per frame are obtained. These output posterior probabilities are known as posterior probability vectors. In posteriorgram matching, posteriorgrams of speech data are generated in advance. A posteriorgram of spoken query is generated when it is given. Then frame-level matching between two posteriorgrams is performed using DTW continuously. The local distance in DTW matching is obtained by computing the inner product of each posterior probability vector, and then by taking the negative logarithm of it. Posteriorgram matching [14] shows the state-of-the-art retrieval accuracy, but it requires long times and large amounts of memory because of the large dimension of posterior probability vectors. To reduce the retrieval times and memory requirements, maximum likelihood state sequence (MLSS) for a spoken query or speech data has been proposed [15, 16]. This method compresses a 3,000-dimensional posteriorgram to one dimension and then eliminates the inner product calculation. Therefore, it reduces the retrieval time. However, the retrieval accuracy of MLSS is lower than that of posteriorgram matching because of the reduction of information. For another earlier study [17], two frame forced integration (2FFI) was proposed to reduce time and memory, as well as MLSS. It reduced the number of frames in the posteriorgram by averaging two frames of the posterior probability vectors in the posteriorgram to one frame.

As described herein, we propose two compression methods to reduce time and memory. The blank-cut method removes <blank> labels. We expect that removing them will improve the retrieval accuracy by highlighting only the valid information. The frame de-duplication (FDD) method compresses similar frames to one average frame. We first evaluated these methods using the output of multiple deep learning models and then obtained matching scores that are integrated to improve accuracy. We tested the effectiveness of the proposed method and the integration of multiple models using open test collections.

The remainder of this paper is organized as follows: Section II introduces earlier studies. Section III presents blank-cut method and FDD method, proposed herein. Section IV presents

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

results of experiments conducted using the proposed method. Finally, the conclusion is presented in Section V.

II. RELATED WORK

A. *Posteriorgram matching for QbE*

Recently, deep neural networks (DNNs) have been developed and used in various fields [18]. Furthermore, in speech recognition systems, acoustic models trained with DNN are often used. Even for unspecified speakers, QbE using DNN enables highly accurate retrieval by performing frame-level matching using posteriorgrams, which are frame-by-frame posterior probabilities obtained from DNN. By using this method, the speech data are first converted into features such as filter bank (FBANK) or Mel-Frequency Cepstrum Coefficients (MFCCs), which are then input to the DNN to produce a posteriorgram. The given spoken query is also converted into features and is input to the DNN to produce a posteriorgram. The local distance matrix is then produced using the posteriorgrams of the spoken query and the speech data. By computing the inner product, the similarity of vectors in each frame of the speech data and the speech query are obtained. Then DTW matching is performed using the distance matrix, which has been distanced by taking the negative logarithm after computing the inner product [18]. However, because the posterior probability vector consists of approximately 3,000 dimensions, the creation of the local distance matrix takes a high computational cost and engenders a long search time. Furthermore, because the posteriorgram of the speech data is retained, the memory requirement increases.

B. *Spoken query / Speech data maximum likelihood state sequence method (MLSS)*

A spoken query / speech data maximum likelihood state sequence (MLSS) method was proposed to reduce the memory requirement and long retrieval time of posteriorgram matching [15, 16].

In MLSS for spoken query, a posteriorgram of the spoken query is made. The state number with the largest posterior probability value in each frame is assigned to the maximum likelihood state. Each frame is associated with a maximum likelihood state number. This number converts a posteriorgram of approximately 3,000 dimensions into a one-dimensional maximum likelihood state number series. The probability values of the posteriorgrams of speech data are taken as a local distance matrix with negative logarithms. The matrix is referenced using the maximum likelihood state series of the spoken query.

In MLSS for speech data, the posteriorgrams of the speech data are converted to a maximum likelihood state series in advance for referencing and matching. Although the retrieval time and memory requirement were reduced, the retrieval accuracy was lower than that of posteriorgram matching. Fig. 1 depicts a flow of MLSS for speech data.

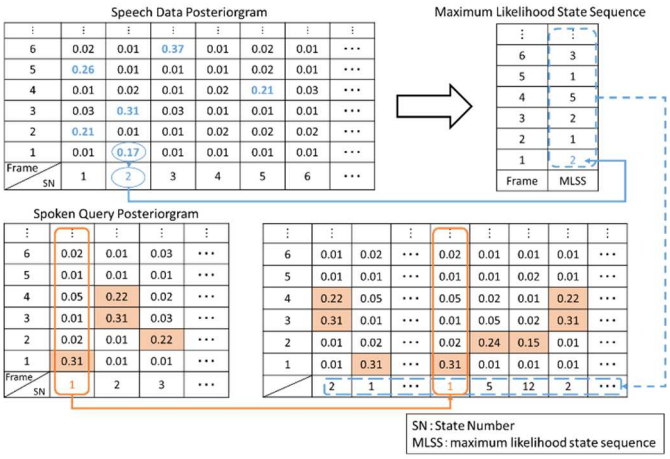


Fig. 1. Speech data maximum likelihood sequence method flow, adapted from an earlier report [16].

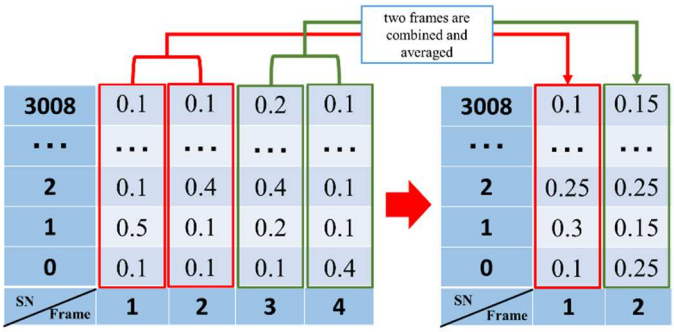


Fig. 2. Flow of two frame forced integration (2FFI) method.

C. *Two frame forced integration*

As described in an earlier report [17], a method was proposed to average the posterior probability vectors of two adjacent frames of a posteriorgram. As described in this paper, we designate it as two frame forced integration (2FFI). In this method, the posterior probability vectors of one frame and the next frame in the posteriorgram are combined and averaged. The resulting posterior probability vector is stored in a new posteriorgram. This process halves the number of frames, thereby reducing the retrieval time for matching. Fig. 2 depicts the flow of the 2FFI method.

D. *Bidirectional long short-term memory (BLSTM)*

Bidirectional Long Short-Term Memory (BLSTM) is a deep learning model that is an extension of a Recurrent Neural Network (RNN). Actually, RNN has a recursive structure by which the middle layer output returns to itself after a time delay. It is regarded as effective for learning time series data. For this reason, it is used widely in the fields of speech recognition and natural language processing.

Actually, LSTM is a model that can learn long-term dependencies by replacing the nodes in the middle layer of RNN with LSTM gates of three types. A BLSTM model uses the LSTM structure for forward propagation to the next step, makes a model for back propagation to the previous step, and integrates each output into the output layer. Compared with

unidirectional deep learning models such as simple feed forward neural networks, BLSTM can learn bidirectional features of time series data, and can thereby improve the retrieval accuracy [17].

E. ESPnet

In recent years, end-to-end learning models have been used in ASR. They reportedly provide high recognition accuracy [19]. End-to-end ASR requires no frame-by-frame correct labels, which are necessary for traditional hybrid ASR. Actually, ESPnet allows the flexible construction of models with different numbers of states corresponding to speech features such as words, subwords and other speech components. Therefore, we used ESPnet, which consists of Hybrid CTC/Attention [20] as an end-to-end speech recognizer. As a shared encoder, the same encoder is used for both CTC and Attention. The posteriorgram consisting of the posterior probability vector output from CTC is extracted and used because an earlier report [17] has described high retrieval accuracy.

F. Score integration

An important difficulty associated with MLSS method is that its retrieval accuracy is lower than that of the posteriorgram matching method because of the decreased amount of information. To resolve that point, matching with multiple deep learning models was performed; then score integration method was used for retrieval in an earlier study [17]. When using this method, the information is complemented by linear summation of multiple matching scores using Equation (1). The integration ratio of the linear summation is adjusted by weight α, for which the possible values are $0 \leq \alpha \leq 1$.

$$D_{new}(\mathrm{i}) = \sum_{n=1}^{N} \alpha_n D_n(i) \qquad (1)$$

III. PROPOSED METHOD

This section proposes blank-cut method (b-cut) and FDD to compress the posteriorgram frames used in matching. This method further reduces the retrieval time and memory requirements of posteriorgram matching and MLSS methods. Moreover, with scores obtained using the proposed method in the score integration, we propose a method that achieves high retrieval accuracy, and retrieval time of about 2 s. The memory requirement is about 2 GB for about 30 hr of speech data to be retrieved.

A. Blank-cut (b-cut)

Into the posteriorgram made from Connectionist Temporal Classification (CTC) of the Hybrid CTC/Attention architecture in ESPnet, <blank> labels are inserted to match the input and output series lengths. Assuming that <blank> labels have no valid information, we propose a method that removes <blank> labels and matches them. Flow of the b-cut method is depicted in Fig. 3. In the MLSS of posteriorgrams, removing frames

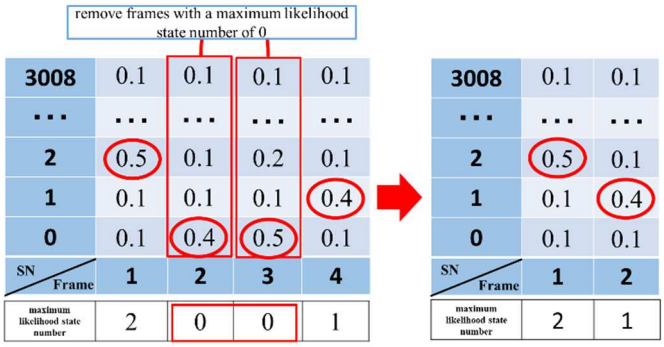


Fig. 3. Flow of blank-cut (b-cut) method.

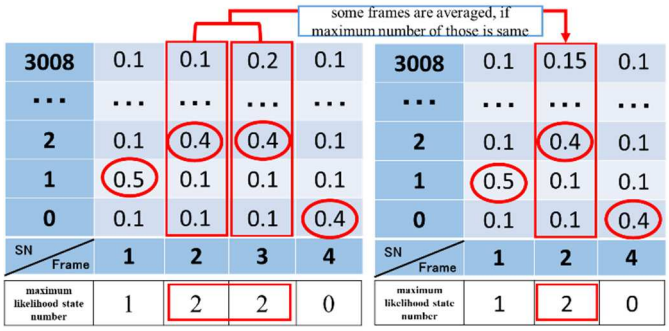


Fig. 4. Flow of FDD method.

with a maximum likelihood state number of 0, which correspond to <blank>, reduces the number of frames in the posteriorgrams, thereby reducing the retrieval time and memory. About half of the frames in the posteriorgrams obtained from CTC are <blank>. Therefore, removing them is expected to improve the retrieval accuracy by highlighting only valid information. The b-cut method is applied only to the posteriorgram obtained from ESPnet because <blank> exists only in the posteriorgram made from CTC.

B. Frame de-duplication (FDD)

The posteriorgram contains a series of frames with similar features and trends in the posterior probability vector. When local distances are made for matching, similar frames also show similar values. Therefore, even if similar frames are compressed to one average frame, the matching process is not affected markedly. The frames in the posteriorgram are reduced while maintaining the original retrieval accuracy, thereby reducing the retrieval time and memory requirement.

The FDD method flow is portrayed in Fig. 4. First, MLSS is made using the respective posteriorgrams of the spoken query and speech data. Next, each MLSS is referenced to ascertain the proportion of consecutive frames which have the same maximum likelihood state number. If frames of the same maximum likelihood state number are consecutive, then the posterior probability vector obtained by averaging multiple posterior probability vectors for those frames is stored as a new posteriorgram. If not, the posterior probability vectors of frames are stored as they are. Compressing multiple frames into a single frame reduces the overall number of frames in the posteriorgram, therefore engendering reduction of the time

necessary for local distance computation and DTW matching. Reduction in the number of retained posteriorgram frames also reduces the memory requirement. Furthermore, unlike the 2FFI method, which averages all frames by two frames, the FDD method, which compresses only frames with similar trends in the posterior probability vectors, can compress and match while preserving the features and trends of each frame's posterior probability vector.

C. Combination with score integration

Because the amount of information possessed by the posteriorgram is reduced when the proposed methods are applied, the retrieval accuracy might be lower than that of posteriorgram matching. To address this possible shortcoming, score integration using different deep learning models is applied to complement the information and to improve the retrieval accuracy. Memory requirements increase with the number of models used for score integration. However, the proposed methods reduce both retrieval time and memory during each matching. The methods are expected to maintain the memory requirement at a practical level even when using score integration.

IV. EXPERIMENTS

A. Experiment conditions

As described in this report, BLSTM and the model of ESPnet (Hybrid CTC/Attention) were trained and used for evaluation experiments. TABLE I presents the extraction conditions of speech features. The BLSTM model consists of 120 dimensional features (FBANK 40 dimensions, its ΔΔ features, and its ΔΔ features) concatenated in before and after 5 frames of one frame. Therefore, the input features are 1,320 (11 × 120)-dimensional. The input of Hybrid CTC/Attention was set to 83 dimensions, consisting of 80 FBANK dimensions and 3 pitch dimensions, as in an earlier study [17] that showed high retrieval accuracy.

The training data for each model are 2,702 lectures from the Corpus of Spontaneous Japanese (CSJ): speech of about 600 hr. The BLSTM consists of one input layer, five intermediate layers (256 dim.), and one output layer. The output layer corresponds to triphone. The number of states is 3,009 with state sharing. In Hybrid CTC/Attention, the output nodes correspond to character (3,242 dim.), syllable (261 dim.), and monophone (40 dim.), each of which has a <blank> label, an unknown state <unk> label, and a terminating <eos> label. Consequently, character has 3,245 dimensions, syllable has 264 dimensions, and monophone has 43 dimensions. For our experiments, we used a posteriorgram obtained from CTC, which has been shown to be highly accurate in an earlier study [17]. The retrieval time was measured using a personal computer: Intel Core i5-9400 CPU, NVIDIA GeForce RTX 2070SP GPU, and 16 GB RAM.

B. Test collections

TABLE II presents open test collections, called NTCIR-10 and NTCIR-12, used for the evaluation of the retrieval experiments. Because no spoken queries were provided in NTCIR-10, we recorded 100 queries uttered by 10 people. In NTCIR-10, the target speech data for retrieval were 104 lectures from the Spoken Document Processing Workshop (SDPWS), and in NTCIR-12, 98 lectures from the SDPWS. Correct answers were provided by the organizer for both NTCIR-10 and NTCIR-12. The mean average precision (MAP) was used to evaluate the retrieval accuracy.

TABLE I
Conditions for Feature Extraction

Model Architecture	BLSTM	Hybrid CTC / Attention
Feature parameter	120 dimensions FBANK (40 dims) + ΔFBANK (40 dims) + ΔΔFBANK (40 dims)	83 dimensions FBANK (80 dims) + Pitch (3 dims)
Window length	25 ms	
Frame shift	10 ms	

TABLE II
Two Open Test Collections

	NTCIR-10	NTCIR-12
Spoken documents	104 presentations, 29 hr, 40,746 utterances	98 presentations, 27.5 hr, 37,782 utterances
Query sets	Formal run: 100 (10 people)	Formal run: 113 (10 people)

C. Results obtained using proposed methods (single model)

To evaluate the proposed method, we first conducted matching experiments using the single model. In the matching experiment, we compare the normal condition with the 2FFI, b-cut, and FDD methods. Fig. 5 presents results of the matching experiments on NTCIR-10 using the BLSTM model. As described above, we use the posteriorgram matching method as the baseline because it allows detailed retrieval. Considering that it will be used together with the score integration method in later experiments, we conducted experiments on MLSS for spoken query and MLSS for speech data, which are expected to be of practical retrieval time in BLSTM. In the figure, the posteriorgram matching method is denoted as “Poste”, MLSS for spoken query as “Sq”, and MLSS for speech data as “Sd”.

In NTCIR-10, compared to Sq normal, the 2FFI method decreased MAP by about 1 pt, but reduced the retrieval time by about 2.6 s and the memory requirement by about half. The FDD method also reduced the time by about 2.6 s and memory by about half, and improved MAP by 0.2 pt. Compared to Sd normal, the 2FFI method decreased MAP by about 1.6 pt, but reduced the time by about 2.5 s and memory by about half. Similarly, the FDD method decreased MAP by about 2 pt, but reduced the time by about 2.5 s and memory by about half.

Experiments were also conducted using both the 2FFI and FDD methods. Compared to the results obtained using each method applied alone, the size of the posteriorgram can be reduced further. Then we aim to reduce the retrieval time and memory requirement further. Experiment results demonstrated that the 2FFI&FDD MAP outperformed the stand-alone MAP in both Sq and Sd, further reducing the retrieval time and

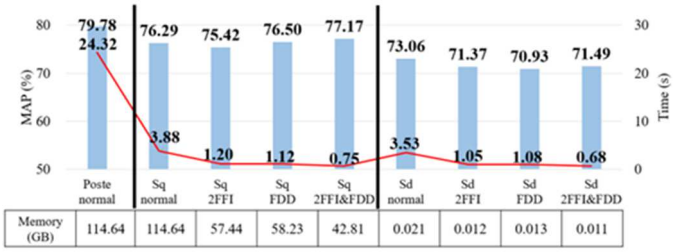


Fig. 5. Results of 2FFI method and FDD method (BLSTM: NTCIR-10).

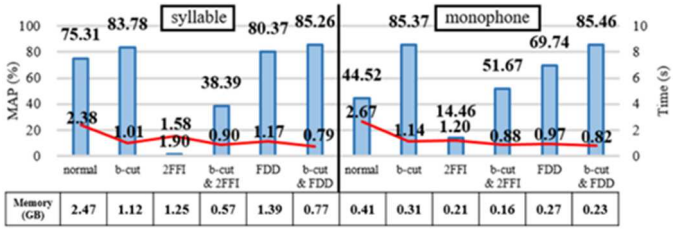


Fig. 6. Results of Posteriorgram matching using b-cut method, 2FFI method and FDD method (syllable, monophone: NTCIR-10).

memory requirements. A similar trend was observed from the NTCIR-12 experiment results.

Fig. 6 portrays results obtained from the experiments on NTCIR-10 using monophone and syllable. Monophone and syllable data have fewer dimensions in the posterior probability vector. Therefore, the retrieval time is expected to be practical even with posteriorgram matching. Therefore, we conduct experiments on the posteriorgram matching method.

In NTCIR-10 of monophone, the b-cut method improved MAP by about 41 pt, halved the retrieval time, and reduced the memory requirement by about 0.1 GB, compared to normal. The 2FFI method decreased MAP by about 30 pt, but halved both the time and memory. The FDD method improved MAP by 25 pt and halved both time and memory.

We also conducted experiments using the b-cut method in combination with the 2FFI and FDD methods. The b-cut&2FFI method decreased MAP by about 34 pt, but reduced the retrieval time by about 0.3 s and the memory requirement by half. The b-cut&FDD improved MAP by about 0.1 pt, reduced the time by about 0.3 s and memory by about 0.1 GB. Because MAP dropped markedly when the 2FFI method was applied, experiments using the 2FFI&FDD method were not conducted for monophone or syllable. The same tendency was observed from experiments with syllable in NTCIR-10 and with monophone and syllable in NTCIR-12.

The 2FFI method reduced the retrieval time and memory requirement while maintaining MAP in BLSTM, whereas MAP decreased considerably in monophone and syllable matching because the distribution of posterior probability values in ESPnet tends to be more extreme than that in BLSTM, with the posterior probability value of maximum likelihood being approximately 0.9 and the probability values of other states being almost zero. The 2FFI method buried high probability values by forcibly averaging them. In contrast, the FDD method compresses by averaging only similar frames

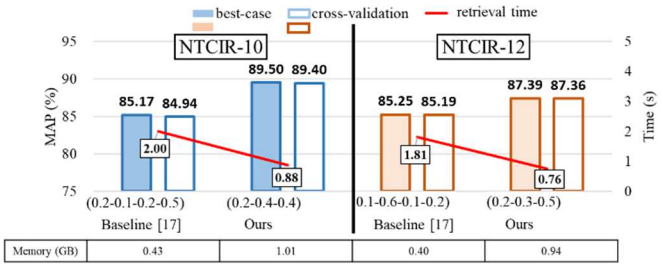


Fig. 7. Results in combination with the score integration method (NTCIR-10 and NTCIR-12).

with the same maximum likelihood state number, which reduces the previously described effects and which therefore prevents the MAP degradation. Regarding the b-cut method, we have concluded that removal of the blank part contributed to MAP improvement by coordinating the valid information. Therefore, the 2FFI&FDD method effectiveness was confirmed for BLSTM. That of the b-cut&FDD method was confirmed for monophone and syllable.

D. Results of score integration (multiple models)

Because the amount of information in the posteriorgram decreases when FDD method is applied, achieving clear improvement in retrieval accuracy compared to normal posteriorgram matching is difficult. Therefore, to complement the information of different tendencies and to improve the retrieval accuracy we apply score integration using multiple models.

Fig. 7 presents results of score integration for NTCIR-10 and NTCIR-12. The baseline is the highest MAP among the experimentally obtained results shown in an earlier study [17] that used score integration. The integration ratio is changed by 0.1. The integration ratio with the highest MAP is shown in the figure as "best-case" as a result of the experiment with all ratios. For cross-validation, results of another experiment using the highest integration ratio for one test set are shown alongside the best-case results as "cross-validation".

In NTCIR-10 at baseline, the highest MAP was found when a total of four types of matching scores were integrated. They are BLSTM (MLSS for speech data, 2FFI for only speech data), character (MLSS for speech data), syllable (MLSS for speech data) and monophone (MLSS for spoken query), in the integration ratio of (0.2 – 0.1 – 0.2 – 0.5). However, in this study, the MAP was highest when a total of matching scores of three types were integrated. They are BLSTM (MLSS for speech data, 2FFI&FDD), syllable (b-cut&FDD) and monophone (b-cut&FDD), in the integration ratio of (0.2 – 0.4 – 0.4). Compared to the baseline, the results obtained using the proposed method improved MAP by about 4.3 pt and reduced the retrieval time by about 1 s. The memory requirement increased by about 0.6 GB, but because it is only about 1 GB, it is a practical amount.

The same trend was observed for NTCIR-12. The highest MAP was observed when the baseline was integrated with the same matching scores as NTCIR-10 in the ratio of (0.1 – 0.6 – 0.1 – 0.2), whereas proposed method was integrated in the ratio

of (0.2 – 0.3 – 0.5). Compared to the baseline, our method improved MAP by about 2.2 pt and reduced the retrieval time by about 1 s. For both test sets, MAP during cross-validation was similar to that of the best case. Therefore, no bias was found among the test sets in terms of the integration ratio.

V. CONCLUSIONS

As described herein, we propose blank-cut method and FDD method to compress the posteriorgram frames. We applied 2FFI method and two proposed methods to three models and integrated their scores to obtain the best retrieval accuracy: 89.50% for NTCIR-10 and 87.39% for NTCIR-12. The retrieval time was 0.88 s for NTCIR-10 and 0.76 s for NTCIR-12, each of which was reduced by about 1 s. Although the amount of memory requirement increased from the baseline, it was only about 1 GB, which is within a practical range. The findings confirmed the effectiveness of the proposed method.

In future studies, we expect to achieve higher accuracy by introducing and integrating further multiple deep learning models.

ACKNOWLEDGMENT

This work was supported by JSPS (c), KAKENHI Grant Number 21K12611.

REFERENCES

[1] C. Auzanne, JS. Garofolo, JG. Fiscus, and WM Fisher, “Automatic Language Model Adaptation for Spoken Document Retrieval,” 2000TREC-9 SDR Track, 2000.

[2] A. Fujii, and K. itou, “Evaluating Speech-Driven IR in the NTCIR-3Web Retrieval Task,” Third NTCIR Workshop, 2003.

[3] P. Motlicek, F. Valente, and PN. Garner, “English Spoken Term Detection in Multilingual Recordings," in Proc. INTERSPEECH, pp.206-209, 2010.

[4] OpenKWS13 Keyword Search Evaluation Plan, 2013 http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-evalplan-v4.pdf

[5] The 2013 Spoken Web Search Task, MediaEval Benchmarking Initiative for Multimedia Evaluation, 2013, http://www.multimediaeval.org/mediaeval2013 /sws2013.

[6] Igor Szöke, Luis Javier, Rodriguez-Fuentes, Andi Buzo, Florian Metze, Xavier Anguera, “Query by Example Search on Speech Task (QUESST 2015),” http://www.slideshare.net/multimediaeval/mediaeval-2015-query-by-example-search-on-speech-task

[7] Tomoyosi Akiba et al, “Overview of the NTCIR-10 SpokenDoc Task,” Proceedings of the 10th NTCIR Conference, pp.573-587, 2013.

[8] Tomoyosi Akiba et al, “Overview of the NTCIR-11 SpokenQuery&Doc Task,” Proceedings of the 11th NTCIR Conference, pp.350-364, December 2014.

[9] Jonathan G. Fiscus et al, SIGIR Workshop Searching Spontaneous Conversational Speech. Results of the 2006 spoken term detection evaluation, pp. 45-50, 2007.

[10] Tomoyosi Akiba et al, “Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop,” NTCIR-9 Workshop Meeting, pp. 223-235, 2011.

[11] Tomoyosi Akiba et al., “Overview of the NTCIR-10 SpokenDoc-2 Task,” NTCIR-10 Workshop Meeting, pp. 573-587, 2013.

[12] Tomoyosi Akiba et al, “Overview of NTCIR-11 Spoken&Doc Task, NTCIR-11,” pp. 350-364, 2014.

[13] Tomoyoshi Akiba et al, “Overview of NTCIR-12 Spoken&Doc-2 Task,” NTCIR-12 Workshop Meeting, pp.167-179, 2016.

[14] Yoshiaki Itoh and Kazuyo Tanaka, “Frequent word section extraction in a presentation speech by an effective dynamic programming algorithm,” The Journal of the Acoustical Society of America, vol.116 No.2, pp. 1234-1243, 2004-8.

[15] Daisuke Kaneko, Ryota Konno, Kazunori Kojima, Kazuyo Tanaka , Shi-wook Lee and Yoshiaki Itoh, “Constructing Acoustic Distances between Subwords and States Obtained from a Deep Neural Network for Spoken Term Detection,” in Proc. INTERSPEECH, pp.2879-2883, 2017.

[16] Takashi Yokota , Kazunori Kojima , Shi-wook Lee and Yoshiaki Itoh, ”Reduction of Speech Data Posteriorgrams by Compressing Maximum-likelihood State Sequences in Query by Example,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020.

[17] Kazuki Hatakeyama, Masahiro Nishino, Kazunori Kojima, Shi-wook Lee, Yoshiaki Itoh, “Multiple Deep Learning Models and Architectures with Different Numbers of States Used to Improve Retrieval Accuracy of Query-by-Example,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021.

[18] Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” Nature, 521(7553):436-444, 2015.

[19] Shinji Watanabe et al, “ESPnet: End-to-End Speech Processing Toolkit,” in Proc. INTERSPEECH, pp.2207-2211, 2018.

[20] Shinji Watanabe et al, “Hybrid CTC/Attention Architecture for End-to-End Speech Recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Fundamental Frequency Estimation Based on Finite-Order Harmonic Constraint Differential Equation

Kenta Yamada*, Yoshiki Masuyama*, Kouei Yamaoka*, Nobutaka Ono*
*Tokyo Metropolitan University, Japan
E-mail: {yamada-kenta5@ed., masuyama-yoshiki@ed., yamaoka-kouei@ed., onono@}tmu.ac.jp
Tel/Fax: +81-42-585-8418

***Abstract*—In this paper, we present a short-time fundamental frequency estimation method based on a differential equation. Fundamental frequency estimation is a popular problem in audio analysis. Particularly, the estimation from short signals has been desired in various real-time processing. The sinusoidal constraint differential equation (SCDE)-based method for realizing high-temporal resolution has been proposed for the frequency estimation of a single sinusoid or multiple sinusoids. The SCDE-based method can efficiently and accurately estimate frequency even from a short-term signal. In this paper, as its extension towards the fundamental frequency estimation, we introduce the harmonic constraint to the SCDE-based method, and propose the finite-order harmonic constraint differential equation-based method. We first derive a differential equation that any harmonic signals with harmonics up to Nth order satisfy, where the fundamental frequency is included in coefficients. Then, the fundamental frequency is estimated by minimizing an objective function expressed as the short-time integral of the squared error of its differential equation. Our experimental results show that our method outperformed the existing methods for estimation with noisy signals of 5 ms.**

I. Introduction

The frequency estimation of sinusoids has been studied as a fundamental problem in audio signal processing with various applications [1]–[7]. The fundamental frequency estimation has also been studied as a popular problem [8]–[10]. Signals such as speech and music have a harmonic structure. The frequencies with the harmonic structure are integer multiples of a common low frequency, which is known as the "fundamental frequency" or "F0". We aim to track the fundamental frequency in real time. To this end, we extend a frequency estimation method that has performed well on a short signal with high-frequency resolution.

A popular method of frequency estimation is the maximum likelihood estimation [11]–[15]. Although it can accurately estimate frequencies, it requires an iterative optimization to solve a nonconvex optimization problem. Other methods for frequency estimation have used discrete Fourier transform (DFT) [16]–[21]. Parabolic interpolation [16], [17] and spectral reassignment [18], [19] improve the estimation accuracy by taking the spectral peak. However, their estimation accuracy depends on the frequency resolution. It is difficult to estimate high resolution in a short signal.

One promising method for frequency estimation is the sinusoidal constraint differential equation (SCDE)-based method [22]–[25]. The key to this method is the use of signal derivatives. In detail, the SCDE is a differential equation that holds for a single sinusoid. The equation is independent of the amplitude and phase of the sinusoid, allowing us to estimate the frequency in a closed form from a short signal. We expanded the SCDE-based method for an arbitrary number of sinusoids [26]. Our method can accurately estimate the frequencies of all the sinusoids simultaneously from a short observed signal.

An approach to performing the fundamental frequency estimation based on the SCDE method would be applying it directly to a harmonic signal as a multi-sinusoidal signal. In this case, the frequencies of the harmonics are estimated separately. Then, by post-processing them, the fundamental frequency could be estimated. However, we consider that a more effective performance could be achieved by incorporating the harmonic structure into the model in advance.

In this paper, based on such motivation, we present the finite-order harmonic constraint differential equation (FOHCDE)-based method that uses the SCDE with the harmonic structure of signals as a constraint. Firstly, we formulate a differential equation satisfied by any harmonic signal with harmonics up to the Nth order, with the fundamental frequency incorporated within the coefficients. After that, the fundamental frequency is estimated by minimizing an objective function defined as the short-time integral of the squared error from the differential equation. Our experimental results show that the FOHCDE-based method is effective in high resolution for a small length. The proposed method can achieve a higher performance than the SCDE-based method under noisy conditions.

II. Sinusoidal Constraint Differential Equation

We previously presented a method of estimating the frequencies of multiple sinusoids from an extremely short signal [26]. In this section, we briefly explain this method based on the SCDE.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

A. Signal model

Let us consider a signal modeled as

$$x(t) = \sum_{n=1}^{N} A_n \cos(\omega_n t + \phi_n), \tag{1}$$

where t denotes continuous time, $n = 1, \ldots, N$ denotes the sinusoid index, and A_n, ω_n, and ϕ_n denote the amplitude, frequency, and phase of the n-th sinusoid, respectively. The parameters A_n, ω_n, and ϕ_n are unknown, while the number of sinusoids (N) is assumed to be known in this paper.

B. Frequency estimation of SCDE-based method

We first introduce the SCDE-based method for a sinusoid referring to [22], [23]. The SCDE-based method for a sinusoid is the basis for the SCDE-based method for multiple sinusoids and the FOHCDE-based method. In this case, the following second-order differential equation holds for any t and ω_1:

$$\frac{d^2}{dt^2} x(t) + \omega_1^2 x(t) = 0. \tag{2}$$

This equation is called the SCDE [22], [23]. Since it does not contain A_1 or ϕ_1, the frequency estimation can be performed independently of these parameters. To obtain ω_1 satisfying (2), we minimize the following objective function:

$$J(\omega_1) = \int_\Gamma \left[\frac{d^2}{dt^2} x(t) + \omega_1^2 x(t) \right]^2 dt, \tag{3}$$

where Γ denotes the range of $t \in \mathbb{R}$ in the definite integral. It has also been extended to the case with multiple sinusoids by combining the DFT and the SCDE-based method in each sub-band [22], [23]. Since spectral peaks are not separable when the signal is very short, this method is not applicable for short-time signals.

C. Multiple frequency estimation by SCDE-based method

Next, let us consider a signal with N sinusoids. The SCDE with multiple sinusoids is defined as [26]

$$\begin{aligned} &\left(\frac{d^2}{dt^2} + \omega_1^2 \right) \cdots \left(\frac{d^2}{dt^2} + \omega_N^2 \right) x(t) = 0 \\ &\Leftrightarrow c_0 \frac{d^{2N}}{dt^{2N}} x(t) + c_1 \frac{d^{2(N-1)}}{dt^{2(N-1)}} x(t) \\ &\quad + \cdots + c_{N-1} \frac{d^2}{dt^2} x(t) + c_N x(t) = 0, \end{aligned} \tag{4}$$

where c_n denotes the coefficient of the differential equation, depending on ω_n for all n. We obtain ω_n from each c_n. To optimize c_n under the SCDE (4), we minimize the following objective function:

$$J(c_0, \ldots, c_N) = \int_\Gamma \left[\sum_{n=0}^{N} c_n \frac{d^{2(N-n)}}{dt^{2(N-n)}} x(t) \right]^2 dt. \tag{5}$$

We can then estimate frequencies by solving the simultaneous equations for c_n.

III. Finite-Order Harmonic Constraint Differential Equation

A. Motivation

The SCDE-based method is applicable to any signal, including that with the harmonic structure. The harmonic structure of sound is a unique feature of voice and instruments. By introducing the constraint of the harmonic structure into the SCDE-based method, we expect to realize the fundamental frequency estimation with good accuracy, even from a short signal.

B. Signal model

Let us consider a signal with a harmonic structure. We model the signal as

$$x(t) = \sum_{n=1}^{N} A_n \cos(\omega_n t + \phi_n), \tag{6}$$

$$\omega_n = n\omega_1, \tag{7}$$

where ω_1 denotes the fundamental frequency and N denotes the finite number of harmonics included in $x(t)$. As described in Section II-A, A_n, ω_n, and ϕ_n are unknown, while N is assumed to be given in advance.

C. Fundamental frequency estimation by FOHCDE-based method

First, each sinusoid in (6) satisfies the following SCDE as in (2):

$$\left(\frac{d^2}{dt^2} + \omega_n^2 \right) A_n \cos(\omega_n + \phi_n) = 0 \quad \forall n. \tag{8}$$

From (7), the following differential equation holds for $x(t)$ with any t, N, and ω_1:

$$\left(\frac{d^2}{dt^2} + \omega_1^2 \right) \cdots \left(\frac{d^2}{dt^2} + N^2 \omega_1^2 \right) x(t) = 0. \tag{9}$$

We call this equation the FOHCDE since any harmonic signals with harmonics up to Nth order satisfies this differential equation. To estimate ω_1 that satisfies constraint (9) as much as possible, we minimize the following objective function:

$$\begin{aligned} J(\Omega_1) &= \int_\Gamma \left[\sum_{n=0}^{N} \alpha_n \Omega_1^n \frac{d^{2(N-n)}}{dt^{2(N-n)}} x(t) \right]^2 dt \\ &= \sum_{n=0}^{2N} \beta_n \Omega_1^n, \end{aligned} \tag{10}$$

where $\Omega_1 = \omega_1^2$. The coefficients α_n and β_n respectively satisfy the equations below.

$$\sum_{n=0}^{N} \alpha_n \Omega_1^n = \prod_{n=0}^{N} (1 + n^2 \Omega_1) \tag{11}$$

$$\sum_{n=0}^{2N} \beta_n \Omega_1^n = \sum_{i=0}^{N} \sum_{j=0}^{N} \alpha_i \alpha_j S_{ij} \Omega_1^{i+j} = \boldsymbol{z}^\mathsf{T} \boldsymbol{S} \boldsymbol{z} \tag{12}$$

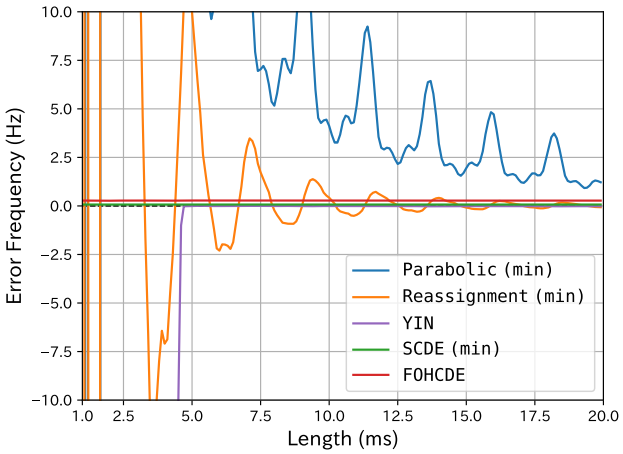


Fig. 1. Estimation error with respect to the signal length for harmonic signals whose frequencies are 440 Hz and 880 Hz.

Here,

$$\boldsymbol{z} = \left[\alpha_0 \Omega_1^0 \ \cdots \ \alpha_N \Omega_1^N\right], \tag{13}$$

$$\boldsymbol{S} = \begin{bmatrix} S_{00} & \cdots & S_{0N} \\ \vdots & \ddots & \vdots \\ S_{N0} & \cdots & S_{NN} \end{bmatrix}, \tag{14}$$

$$S_{ij} = \int_\Gamma \left[\frac{d^{2(N-i)}}{dt^{2(N-i)}} x(t) \frac{d^{2(N-j)}}{dt^{2(N-j)}} x(t)\right] dt. \tag{15}$$

Now, we find the stationary point by taking the derivative of $J(\Omega_1)$ with respect to Ω_1 and setting it to zero.

$$\frac{dJ(\Omega_1)}{d\Omega_1} = \sum_{n=1}^{2N} n\beta_n \Omega_1^{n-1} = 0 \tag{16}$$

Eq. (16) is the polynomial equation of Ω_1 with degree $2N-1$. This equation is identical to that in the original SCDE-based method [22], [23] when $N = 1$. When $N = 2$, the equation, which is a cubic polynomial, has a closed-form solution. In the case where $N = 3$ or more, (16) becomes a polynomial of degree five or higher and cannot be solved algebraically; it is still possible to compute the solution efficiently using an iterative algorithm. Finally, although any real root of the polynomial can be used as the estimate of Ω_1 that satisfies (16), we choose the one that minimizes the objective function (10) as the optimal solution.

IV. Experiments

A. Evaluation using signal length

We investigated the relationship between the signal length and the estimation error by various methods under a noiseless condition. We evaluated the following five conventional methods and the proposed method: The parabolic interpolation method [16], [17] estimates the frequency of a sinusoid by performing DFT. This method finds a frequency bin where the spectral peak is located, fits a quadratic function to the bin and its two side neighbors, and obtains the frequency estimate from the peak of the quadratic function instead of the spectral peak. The spectral reassignment method [18], [19] is also a DFT-based method, which calculates the phase derivative by using a time-differentiated window function. The number of DFT points for these methods was set to 65536 by zero padding, regardless of the signal length. The YIN method [8] is one of the methods for fundamental frequency estimation and utilizes the autocorrelation of the signal. This method utilizes the feature of a periodic signal and estimates the frequency by measuring the highest non-zero-lag peak of the autocorrelation. We also compared the SCDE-based method based on the differential equation [26]. The SCDE-based method is the one described in Section II. Since these methods estimated multiple frequencies, we took the minimum one as the fundamental frequency estimate. We used the centered difference scheme to compute the differential. We simulated a harmonic signal consisting of 440 Hz and 880 Hz, where no noise signal was added in this section. The sampling frequency was 44.1 kHz. In the YIN method, the fundamental frequency is given in advance as 40 Hz to 4000 Hz for estimation.

Fig. 1 shows the estimation error as a function of signal length. `Parabolic (min)`, `Reassignment (min)`, `YIN`, `SCDE (min)` and `FOHCDE` indicate the parabolic interpolation, spectral reassignment, YIN, SCDE-based, and the proposed method, respectively. As shown in this figure, the errors of the DFT-based methods, i.e., the `Parabolic (min)` and `Reassignment (min)` were large when the signal length was small because the frequency resolution was insufficient. The estimation accuracy of the `YIN` was significantly high, and the error was almost 0. Note that the current implementation of the `YIN` requires at least 5 ms of observation, and thus, no results are shown in Fig. 1. In contrast, the `SCDE (min)` method achieved highly accurate frequency estimation even if the signal length was very small, such as 1 ms. Since the `SCDE (min)` and the `FOHCDE` are performed in the time domain, the estimation accuracy is not affected by the frequency resolution, resulting in these remarkable results. The `SCDE (min)` was slightly better than the `FOHCDE`. One reason may be considered that the effect of errors in numerical derivatives and computations is inevitable for the `FOHCDE` because β_n in (10) is significantly large.

B. Evaluation under noisy condition

We evaluated the frequency estimation accuracy under a noisy condition. We used the root mean square error (RMSE) as a measure for evaluation. We performed 1000 trials. In this evaluation, we used the harmonic structure with two components whose frequencies were 440 Hz and 880 Hz. The signal length was 5 ms. We added the band-limited noise with bandwidths from 340 Hz to 540 Hz and from 780 Hz to 980 Hz by changing the SNR from -30 dB to 30 dB. Band-limited noise was generated by performing a DFT on a 10 s white noise signal, applying masks to specific frequency bins, and then performing an inverse discrete Fourier transform (IDFT) on the modified signal. Assuming a combination of the proposed method with a rough frequency decomposition, such as in [22]

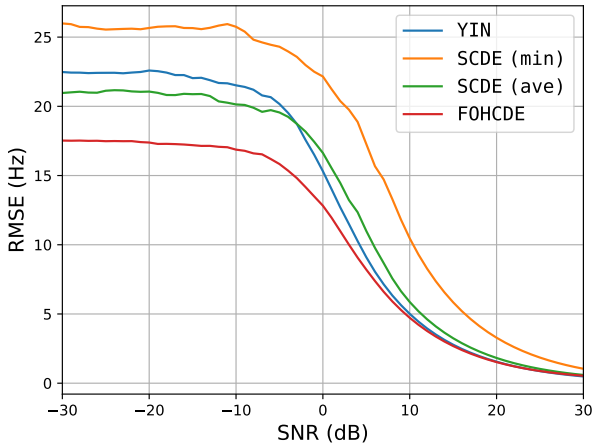


Fig. 2. RMSE for harmonic signals with two components whose frequencies 440 Hz and 880 Hz. The signal length was 5 ms.

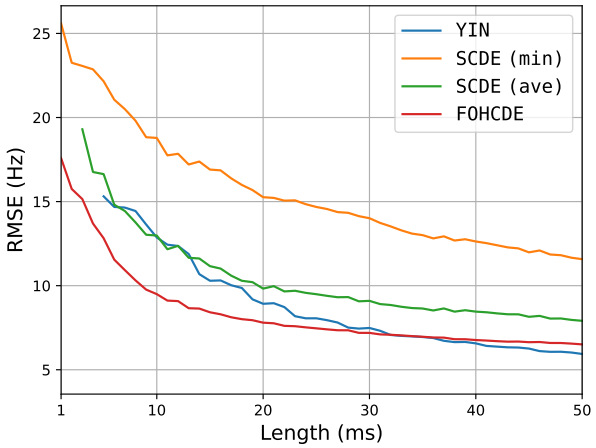


Fig. 3. RMSE of the signal length in 1000 trials. The SNR was 0 dB.

and [25], we employ band-limited noise. If the absolute error of the estimate exceeds 50 Hz, the result was excluded from the RMSE computation as gross error. Other experimental conditions were the same as described in Section IV-A.

Fig. 2 shows the RMSE as a function of SNR. We compared with the proposed method `FOHCDE` with three conventional methods: `YIN`, `SCDE (min)`, and `SCDE (ave)`. `SCDE (ave)` estimates the fundamental frequency by taking the mean of the two frequency estimates.

This figure shows that the estimation accuracy of all methods tended to improve as the SNR increased. In particular, their errors were close to 0 with high SNR. The `YIN` achieved the best performance. The `SCDE (ave)` had a higher accuracy than the `SCDE (min)`. It can be considered that the `SCDE (ave)` utilizes each sinusoid to estimate the fundamental frequency. However, the proposed method was effective for low SNR. The proposed method achieved estimation with noise robustness since the harmonic structure was added as a constraint. Gross errors occurred about 150 out of 1000 times in the `YIN` and about 15 out of 1000 times in the `FOHCDE`.

Next, we fixed SNR to 0 dB and changed the signal length. The signal length ranged from 5 ms to 50 ms. Fig. 3 shows the RMSE as a function of signal length. This figure does not have estimations in `YIN` and `SCDE (min)` for a signal of 5 ms or less since all estimates have errors greater than 50 Hz. This figure shows that all methods tended to improve in estimation accuracy as the signal length. Similarly to the result in Fig. 3, the `SCDE (ave)` had a higher accuracy than the `SCDE (min)`. The `YIN` was better than either, and its estimation accuracy was high when the signal length was sufficient. The `FOHCDE` has the smallest error for a signal of 30 ms or less.

As a result, the `FOHCDE` achieved high accuracy estimation with a small signal with noise robustness since the harmonic structure was added as a constraint. When the signal length is sufficiently large, or the SNR is high, the estimation accuracy of the `FOHCDE` is limited, which can be caused by errors in numerical derivatives. Hence, future work is needed to improve this issue by using more sophisticated methods for numerical derivatives.

V. Conclusion

In this paper, we proposed the FOHCDE-based method for estimating the fundamental frequency. We used the equation related to the differentiation of the signal and the harmonic structure as a constraint. Then, the fundamental frequency was estimated by minimizing the objective function of the FOHCDE-based method. Our experimental results confirmed that our method was more accurate than the SCDE-based method under noisy conditions, and it maintained its performance for a short signal even when the DFT-based methods did not work. In future work, we aim to avoid errors from higher-order numerical derivatives by utilizing the weighted integral method [22], [25].

Acknowledgment

This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

References

[1] D. J. Hirst, "Form and function in the representation of speech prosody," *Speech, Commun.*, vol. 46, no. 3–4, pp. 334–347, 2005.

[2] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE J. Sel. Top. Signal process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] J. Ventura, R. Sousa, and A. Ferreira, "Accurate analysis and visual feedback of vibrato in singing," in *Proc. IEEE Int. Symp. Commun., Control. Signal Process.*, 2012, pp. 1–6.

[4] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 20–30, 2018.

[5] Y. Masuyama, T. Kusano, K. Yatabe, and Y. Oikawa, "Modal decomposition of musical instrument sounds via optimization-based non-linear filtering," *Acoust. Sci. Technol.*, vol. 40, no. 3, pp. 186–197, 2019.

[6] K. Yatabe and Y. Oikawa, "Phase corrected total variation for audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 656–660.

[7] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Low-rankness of complex-valued spectrogram and its application to phase-aware audio processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 855–859.

[8] A. D. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[9] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, pp. 713–727, 2011.

[10] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals.," in *Proc. INTERSPEECH*, 2017, pp. 2321–2325.

[11] D. Rife and R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, 1974.

[12] P. Stoica, R. L. Moses, B. Friedlander, and T. Soderstrom, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Singal Process.*, vol. 37, no. 3, pp. 378–392, 1989.

[13] T. Abatzoglou, "A fast maximum likelihood algorithm for frequency estimation of a sinusoid based on newton's method," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 1, pp. 77–89, 1985.

[14] M. J. Turmon and M. I. Miller, "Maximum-likelihood estimation of complex sinusoids and toeplitz covariances," *IEEE Trans. Signal Process.*, vol. 42, no. 5, pp. 1074–1086, 1994.

[15] M. Ghogho, A. K. Nandi, and A. Swami, "Cramer-Rao bounds and maximum likelihood estimation for random amplitude phase-modulated signals," *IEEE Trans. Signal Process.*, vol. 47, no. 11, pp. 2905–2916, 1999.

[16] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Minimization or maximization of functions*. Cambridge University Press, 1992, pp. 402–405.

[17] M. Abe and J. O. Smith, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of fft magnitude peaks," *J. Audio Eng. Soc.*, 2004.

[18] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, no. 5, pp. 1068–1089, 1995.

[19] F. Auger, P. Flandrin, Y. Lin, *et al.*, "Time-frequency reassignment and synchrosqueezing: An overview," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 32–41, 2013.

[20] M. D. Macleod, "Fast nearly ml estimation of the parameters of real or complex single tones or resolved multiple tones," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 141–148, 1998.

[21] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. Digit. Audio Effect. Conf.*, 2002, pp. 51–58.

[22] S. Ando and T. Nara, "An exact direct method of sinusoidal parameter estimation derived from finite Fourier integral of differential equation," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3317–3329, 2009.

[23] S. Ando, "An exact direct method for short-time sinusoidal parameter estimation," in *Proc. IEICE Tech. Rep.*, vol. 107, 2007, 25–30, in Japanese.

[24] S. Ando, "An exact direct method for short-time sinusoidal parameter estimation by means of finite fourier transform," in *Proc. IEICE Tech. Rep., EA*, vol. 107, 2007, 37–42, in Japanese.

[25] S. Ando, "Frequency-domain prony method for autoregressive model identification and sinusoidal parameter estimation," *IEEE Trans. Signal Process.*, vol. 68, pp. 3461–3470, 2020.

[26] K. Yamada, Y. Masuyama, Y. Wakabayashi, and N. Ono, "Simultaneous frequency estimation for three or more sinusoids based on sinusoidal constraint differential equation," in *Proc. APSIPA ASC*, 2022, pp. 975–978.

Tone Labeling by Deep Learning-based Tone Recognizer for Mandarin Speech

Wu-Hao Li*, Chen-Yu Chiang and Te-hsin Liu
* National Taipei University, Taiwan
E-mail: hank12451@gmail.com, cychiang@mail.ntpu.edu.tw
National Taiwan University, Taiwan
E-mail: tehsinliu@ntu.edu.tw

***Abstract* — Tone labeling of tone sandhi and polyphones is crucial when preparing a high-quality speech corpus for constructing a Mandarin text-to-speech system. Correct tone labeling may ensure that the constructed text-to-speech system can generate a natural prosody. This paper proposes tone labeling using an iterative method with a deep learning-based tone recognizer. The experimental results showed that the proposed method could robustly label tones for syllables of tone sandhi and polyphones on a multi-speaking rate Mandarin speech corpus. Furthermore, this study found that syllables misrecognized as different tones from lexical tones may reflect the true tone realizations caused by coarticulation, location in a prosodic structure, and speaking rates. This study also provided a quantitative analysis of the relationship between labeled tones and prosodic structure to conform to the characteristics found in previous linguistic studies.**

I. INTRODUCTION

Labeling syllable tones in a speech corpus is essential for constructing a Mandarin text-to-speech (TTS) system. Correct tone labeling would help a trained TTS system to produce natural prosody, especially regarding tone sandhi and polyphone syllables. The simplest way to label syllable tones is to utilize text processors to predict proper tones from the input text. However, acoustic realizations of tone-sandhi and polyphone syllables would have some acceptable variations that cannot be predicted only from text information. Therefore, the results of tone labeling usually need to be checked manually to ensure labeling quality [1].

Although end-to-end (e2e) TTS development techniques can build high-quality TTS by constructing mapping from raw input text to spectrograms or waveforms, tone labeling for the corpora is still required because the collected speech corpora cannot guarantee sufficient text-speech training pairs to learn mappings from text to speech with correct prosody regarding tone sandhi and polyphones. Therefore, training corpora for constructing TTS systems are usually labeled with tones. The constructed e2e TTS systems operate with a front-end text processor to extract linguistic features from input text to predict *surface* tones rather than *lexical* tones [2-8].

Because tone labeling is time-consuming and labor-intensive, it is worth constructing a robust automatic tone-labeling method for preparing TTS speech corpora. In addition, tone recognition is essential for assessing the tones realized by Mandarin learners. This paper hence proposes tone labeling using a deep learning-based tone recognition method. The proposed method is examined on the Treebank multi-SR corpus - a multi-speaking rate Mandarin speech corpus uttered by a female professional announcer which has been used in the previous studies of prosody modelings [9-11].

Figure 1 illustrates the research methods used in this study. Section 2 uses the Speaking Rate-dependent Hierarchical Prosodic Model (SR-HPM) with local speaking rate modeling [9-11] to obtain prosody tags representing speech utterances' prosodic structure and the associated local speaking rates. The labeled prosody tags can help analyze the results of tone labeling.

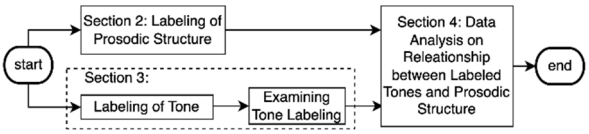


Figure 1: *Research method of this study.*

In Section 3, the tone of each syllable is labeled by a tone recognizer realized by a deep learning-based model (specifically, a transformer-based model) with input acoustic features, and a tone posteriorgram is generated by the recognizer, which is regarded as the likelihood of five Mandarin tones. We checked the tone labeled by the tone recognizer and found that the accuracy for tone labeling of tone sandhi and polyphones was quite high. In addition, we found

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

that some syllables in the corpus pronounced in different tones from lexical tones can be correctly identified by the proposed method. Finally, Section 4 provides a quantitative analysis of the relationship between the labeled tones and prosodic structure, indicating that syllables misrecognized as different tones from lexical tones may reflect the true tone realizations (i.e., surface tones) caused by coarticulation, location in a prosodic structure, and speaking rates.

The tone labeling in this study relies on the constructed deep learning-based tone recognizer. Many previous studies have constructed tone recognizers as components in computer-assisted language learning (CALL) system [12-14] or to provide auxiliary information for automatic speech recognition (ASR) [15,16]. Various deep learning models have been proposed to improve the performance of tone recognition, including DNN-HMM [17], CNN-DBLSTM [18] with an attention mechanism, CTC [19], an end-to-end method by combining CNN, Bi-GRU, Time Distributed Dense layer, and CTC [20], and an end-to-end tone classification method utilizing both a variant of Residual Neural Network (ResNet) with mel-spectrogram input and a multilayer perceptron model with context syllable segment features extracted from the force alignment results by ASR [21]. To develop a CALL system, the Goodness of Tone (GOT) inspired by the goodness of pronunciation (GOP) [22] and the soft-target tone model [14] can improve tone recognition accuracies or equal error rates for detecting mispronunciations. From the perspective of speech production, the target approximation model (TAM) [23], which extracts the underlying pitch targets for tone recognition from the surface F0 contour, was also used to improve tone recognition accuracies with BSLTM [24].

To the best of our knowledge, this study differs from the above studies in the following aspects:1) the tone recognizer functions as a tone labeler rather than a detector for a CALL or ASR system, and 2) with the prosodic structure labeled by SR-HPM [9-11], the relationship between labeled tone and prosodic structures can be investigated quantitatively. Note that the proposed transformer-based tone recognizer in this study (as will be illustrated in Subsection III.B) is new to the above-mentioned existing tone recognition methods and has not been tested with publicly available datasets. In order to check if the proposed transformer-based tone recognizer has a reasonable performance compared with the existing methods, the proposed tone recognizer was tested on a large-scale open-source Mandarin speech dataset – the AISHELL3 database [1]. The competitive tone recognition accuracies were obtained by the proposed transformer-based tone recognizer with respect to the performance reported in the previous study [21]

Although the findings related to tone labeling reported in this study are limited by the experimental results on the speech corpus uttered by only one speaker, the analyses in Section 4 may shed light on finding a tolerance for assessing tone realizations by Mandarin learners.

II. Labeling of Prosodic Structure

The experimental database in this study is the Treebank multi-SR corpus, composed of four parallel sub-corpora in four speaking rates read by a professional female announcer. The average utterance-wise speaking rates (SRs) for the four sub-corpora, from fast to slow, were 0.181, 0.198, 0.244, and 0.264 s/syllable, respectively. There are 1478 files with 203,746 syllables. The SR-HPM modeling [9] with local speaking rate estimation [10,11] jointly labels the corpus with a four-layer prosodic structure, as shown in Figure 2, and robustly estimates the speaking rate of each prosodic phrase. The four-layer prosodic structure consists of prosodic units of syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and prosodic phrase group (PG), from bottom to top. Break types with specified prosodic-acoustic characteristics delimit these layered prosodic units. It has been shown that the PG-initial and PG-final prosodic phrases (PPhs) have distinctive intonation patterns. The intonation of the PG-initial PPhs was marked by an F0 reset before a rapid decline. The PG-final PPhs also possess an F0 reset, and the contour then trails to an ending with final lengthening [25]. One may find the details of break types in previous studies [9-11].

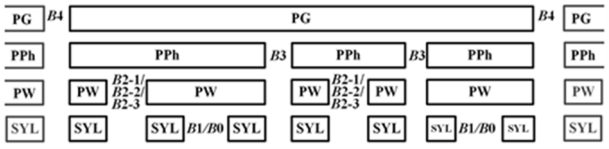


Figure 2: *Prosodic structure used in this study.*

III. Labeling of Tones

A. Iterative Tone Labeling

The experimental corpus was labeled with lexical tones and thus may contain incorrect tone labeling. Therefore, we developed a 5-step iterative tone labeling method to jointly label syllables with the correct realized tones (i.e.m surface tones) rather than lexical tones and constructed a robust tone recognizer.

Step 1 updates the target tones to train the deep learning-based tone recognizer. The target tones are set to be lexical tones for the 1st iteration, that is, iteration index *k*=1. For the other iterations, the target tones were set as the re-labeled tones.

Step 2 trains the tone recognition model that minimizes the cross entropy loss expressed by

$$\theta_k = \arg\min_{\theta} \sum_{n=1}^{N} \sum_{c=1}^{5} \hat{y}_{n,c}^{k} \log(y_{n,c}) \quad (1)$$

where *n* denotes the syllable index, c denotes the tone class, $y_n = f_\theta(x_n)$ represents the function of tone recognition with the trainable parameter set θ; x_n is the input acoustic feature for the *n*-th syllable, y_n is a 5 × 1 vector to express the output vector of the tone recognizer, $\hat{y}_n^k$ is the one-hot vector in which $\hat{y}_{n,t_n}^k = 1$ if the target tone of the *n*-th syllable is t_n.

Step 3 generates tone posteriorgram y with θ_k :

$$y = [y_1^T \quad \dots \quad y_N^T]^T \text{ and } y_n = f_{\theta_k}(x_n) \text{ for } n = 1 \sim N \quad (2)$$

Step 4 re-labels the tones of syllables by

$$t_n^* = \begin{cases} \arg\max_t y_{n,t}, \text{if } y_{n,t} \geq Th(s_n) \text{ and } s_n \in \mathcal{T} \\ t_n \quad , \text{otherwise} \end{cases} \quad (3)$$

where t_n^* is the re-labeled tone, $Th(s_n)$ denotes the likelihood threshold for re-labeling, s_n denotes the linguistic features of *n*-th syllable, $\mathcal{T}$ is the set of linguistic units for tone re-labeling.

Step 5 checks a convergence of tone re-labeling. If convergence is reached, that is, $t_n^* = t_n$ for all *n*, we stop the iteration or return to Step 1 to continue training and re-labeling.

B. Set of Linguistic Units for Tone Re-labeling

The set can be categorized into four exclusive subsets:

Set 1: Tone 3 sandhi syllables, i.e., s_n with $t_n = t_{n+1} =$ T3 and $Th(s_n) = 0$. The threshold $Th(s_n) = 0$ indicates that the tone predicted by the tone recognizer is the relabeled tone, without considering its corresponding degree of likelihood. A speaker may use various tone 3 sandhi realizations to manifest syntactic or semantic conditions properly.

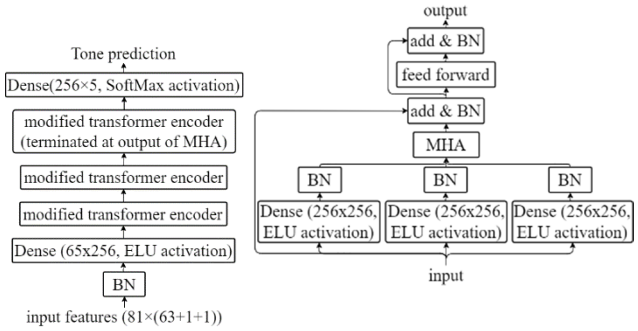
Set 2: Tone sandhi for $s_n \in \{$不(bù), 一(yī)$\}$ and $Th(s_n)$=0.7. The tone for *bù* could be tone 2 (T2) or tone 4 (T4), whereas the tone for yī could be tone 1 (T1), T2, or T4. Although the tones for {*bù*, yī} can be determined by their written contexts, we find that the speaker manifested much tone variation in acoustic features beyond the tone sandhi rules inferred from information extracted from the text. We thus set the threshold $Th(s_n)$=0.7 for a more reliable tone re-labeling.

Set 3: $s_n \in$ Polyphone, $s_n \notin$ Set 1, and $Th(s_n)$=0.7. We selected 26 frequent polyphone characters from a polyphone character table [26]. Each polyphone character sometimes has only one phonetic combination but may have different tones for different meanings. Parts of the polyphone character cases considered in Set 3 overlapped the cases in Set 1. For simplicity, we excluded cases with $t_n = t_{n+1} =$ T3. We also set the $Th(s_n)$=0.7 for a more reliable tone re-labeling.

Set 4: $s_n \notin$ {Set 1, Set 2, Set 3} and $Th(s_n) = 0.98$. We found that some syllables in the corpus may have different phonetic realizations than lexical tones. These syllables are usually located in specific positions of a prosodic structure, that is, sentence-initial and sentence-final positions. Parts of the tone change are caused by the idiomatic realization of tones on specific words or characters. We also found that the speaker made a few erroneous tones due to misreading or a creaky voice. Because the cases for Set 4 are rare, we set a high threshold, $Th(s_n) = 0.98$, to ensure that the re-labeling is robust.

C. Transformer-based Tone Recognizer

Figure 3(a) shows the structure of the proposed transformer-based tone recognizer implemented by Tensorflow 2.11.0 [27]. The structure consists of a batch normalization (BN) layer and a dense layer (65x512) with Exponential Linear Unit (ELU) activation, followed by a stack of three modified transformer encoder blocks. Figure 3(b) shows the structure of the modified transformer encoder block. The input of Multi-head Attention (MHA) is followed by a cascade of dense layers (256×256) with ELU activation and the BN layer. The remainder of the architecture is the same as that of a conventional transformer encoder [28]. The number of heads and key dimensions of the MHA are 2 and 256, respectively. The dimensions of the feed-forward were 512. The output is a 5 × 1 hot vector by SoftMax activation cascaded to the first dimension of the MHA in the top modified transformer encoder block, indicating the tone of the target syllable. The input of the recognizer considers syllable-based features extracted from a fixed-length speech segment with 0.81 seconds centered around the target syllable.



(a) (b)

Figure 3: *the structure of the proposed tone recognizer.*

The input features include:

1) 81×63 Mel-spectrogram: We take the 63-band Mel-spectrogram (without DC) that considers the frequency range from 0 to 2000 Hz with a 512-point FFT in a 10ms frame interval for a 16KHz sample rate. The number 81 indicates that we take the spectrogram information of 40 frames before and after the frame at the center of the target syllable.

2) 81x1 syllable boundary flag to indicate whether the frame belongs to the target syllable with binary values.

3) 81x1 frame index to mark the relative position of the frames, ranging from -40 to +40.

The proposed modified transformer encoder can achieve the same tone recognition rate as the original one with a shorter training duration, i.e., fewer training epochs to reach convergence for loss or accuracy. The training duration for the proposed modified transformer can reduce by 40% of the one taken by the original transformer encoder.

D. Performance of the Tone Labeling

We used all the Treebank multi-SR speech corpus in the tone recognizer's iterative re-labeling (or training) processing. The tone re-labeling process took two iterations to reach convergence, and each iteration took 80 epochs of approximately 85 min with Nvidia 2080ti. The Adam optimizer with a learning rate of 1.0e-5 was used to update the model parameters. In each training epoch, 10% of the corpus was randomly chosen as the validation set to avoid overtraining. We found that in both iterations, the accuracies for the training and validation sets in each epoch were similar, illustrating the robustness of the training process. The recognition accuracy for the 1st iteration with lexical tone targets was 96.1%. The accuracy reached 96.7% in converged iterations. Note that the true tone classes for the 1st iteration and 2nd iteration were lexical tones and re-labeled tones, respectively. The slight increase in accuracy was mainly attributed to the relabeling of tones for the four sets.

Table 1 presents the confusion matrix for tone recognition in the 1st and 2nd iterations. The numbers without brackets are the results of the model obtained in the 1st iteration with lexical tones as true classes. The results of the 2nd iteration that regarded the re-labeled tone as the training target are shown in the numbers within the round and square brackets. The numbers within round brackets count lexical tones as true classes, while those within square brackets count re-labeled tones as true classes. As shown in the *Count* column, the number for T2 increased from 48983 to 53845, while the number for T3 decreased from 37065 to 31766 because of tone re-labeling for tone 3 sandhi. The higher accuracies in the square brackets than those in the round brackets indicate that the proposed method can re-label the tones of syllables to obtain a more consistent or robust tone recognition model.

We then checked the performance of tone re-labeling for the speech tokens of the four sets by listening to these tokens. Two native Mandarin speakers were invited to participate in the check task. Because there are numerous syllables for tone re-labeling, we sampled 7%, 10%, 7%, and 10% of the datum for Sets 1 to 4 for the checking, respectively, with a uniform distribution for the four speaking rates. Table 2 shows the accuracies and corresponding 95% confidence intervals for labeling the four sets. The accuracies were generally high, except for polyphone and tone sandhi for "yi." We found that a few polyphone characters with low tone recognition accuracies and short durations of the syllable "yi" may degrade the performance of tone re-labeling.

Table 1: *Confusion matrix for the tone recognitions. Numbers within round brackets and without brackets are calculated when assuming lexical tones are true tone classes. Numbers in square brackets regard the re-labeled tones as true tone classes.*

		Predicted Class						
		T1	T2	T3	T4	T5	Count	Acc. %
True Class	T1	**37096** **(36576)** **[38128]**	362 (732) [520]	134 (189) [119]	413 (463) [459]	29 (74) [37]	38034 [39263]	97.5 (96.2) [97.1]
	T2	311 (906) [369]	**47544** **(46691)** **[52417]**	787 (707) [601]	235 (486) [359]	106 (193) [99]	**48983** **[53845]**	97.1 (95.3) [97.3]
	T3	154 (364) [77]	2256 (6419) [787]	**33755** **(28775)** **[30132]**	821 (1388) [657]	79 (119) [113]	**37065** **[31766]**	91.1 (77.2) [94.9]
	T4	433 (1494) [572]	271 (1158) [587]	897 (1702) [662]	**67353** **(64542)** **[66619]**	114 (172) [149]	69068 [68589]	97.5 (93.4) [96.5]
	T5	63 (91) [15]	140 (344) [180]	177 (475) [115]	151 (486) [103]	**10065** **(9200)** **[9870]**	10596 [10283]	95.0 (86.8) [96.0]

Table 2: *Numbers of the re-labeled tone tokens (token#), number of the examined samples (sample#), accuracies (Acc.), and 95% confidence intervals (CIs) of the accuracies.*

	Token#	Sample#	Acc. (%)	95% CI for Acc. (%)
Set 1: Tone 3 sandhi for consecutive *x* tone 3 syllable				
x=2	4317	200	94.5	90.4 ≤ Acc. ≤ 97.2
x=3	889	100	96.0	90.1 ≤ Acc. ≤ 98.9
x=4	159	40	95.0	83.1 ≤ Acc. ≤ 99.4
x=5	36	36	97.2	85.5 ≤ Acc. ≤ 99.9
x=6	12	12	91.7	61.5 ≤ Acc. ≤ 99.8
Set 2: Tone sandhi for yī and bù				
yī	2846	304	86.8	82.5 ≤Acc. ≤ 90.4
bù	2560	246	95.5	92.1 ≤ Acc. ≤ 97.7
Set 3: Polyphone				
	15225	669	86.4	83.6≤ Acc. ≤88.8
Set 4: misrecognized tones with high tone posteriorgram				
	537	54	96.3	87.3 ≤ Acc. ≤ 99.5

E. Performance comparison to the previous work

In order to check the performance of the proposed transformer-based tone recognizer, we conducted a tone recognition task on the AISHELL3 speech database [1] in comparison with the existing method reported in the literature [21]. The method proposed in [21] is similar to our proposed method in that both methods utilize information from Mel-spectrogram and syllable boundary as inputs. The two methods differ in that the proposed method in [21] uses the extra linguistic feature, i.e., base-syllable type, and syllable boundaries obtained by the force alignment results according to the ASR output. The proposed method, however, uses the force alignment results according to true syllable transcription. We may neglect the differences made by the force alignments with the ASR and the true syllable transcriptions because the reported syllable error rate [21] made by the ASR was low (6.29%). The most syllable errors would be substitutional errors and the syllable boundaries obtained by the force alignment with the ASR results would be very close to the one made by the true syllable transcription. Since the literation [21] did not clearly describe the settings for the training and test datasets, we follow the document released by AISHELL3 to extract 65.6 and 9.4 hours of speech from AISHELL3 as training and test datasets, respectively. The lengths for the training and test sets are the same as the ones shown in [21]. Table 3 displays the distribution of tones, which is close to the distribution in the [21].

Table 3: *Distribution of tones for AISHELL3*

	T1	T2	T3	T4	T5	Total sample#
training	0.210	0.235	0.161	0.332	0.061	774714
test	0.217	0.234	0.158	0.335	0.053	102742

The authors of [21] displayed various accuracies with different parameters. We select the configure "+SF+C2" as the object for comparison because its length of acoustic feature is closest to our setting, i.e., 81 frames with 10ms frame intervals. To make fair comparisons, the base-syllable type information is added to the input of our transformer-based tone recognizer. Table 4 shows the result of the comparisons. As can be seen from the table that our proposed tone recognizer achieved better accuracies for T1 and T4 (accounting for 55% of data), a slightly lower accuracy for T2 (accounting for 23% of data), and lower accuracies for T3, and T5 (accounting for 22% of data). We believe that our proposed tone recognition method is competitive with the one proposed in [21] according to the compared results shown in Table 4.

Table 4: *Our method vs. results reported by [21]*

	T1	T2	T3	T4	T5
Ours	**0.931**	0.869	0.818	**0.923**	0.862
[21]	0.910	**0.874**	**0.874**	0.875	**0.916**

In addition, we apply the mechanism of tone relabeling to the training set of AISHELL3. To simplify the processing of relabeling, as long as the probability of the predicting tone class exceeds 0.5, we change/relabel the tone class. Then, two transformer-based tone recognizers were trained with the original tone labeling (denoted as the *raw model*) and the tone relabeled by the proposed method (denoted as the *relabeling model*). Table 5 shows the performances of tone recognition on the test sets of AISHELL3 by the *raw model* and the *relabeling model.* As shown in Table 5, the relabeling mechanism increased the accuracy for T1, T4, and T5, but an overall slightly lower accuracy was obtained. This result is reasonable because AISHELL3 has been labeled with surface tones, and most of the labelings were manually checked. The tone relabeling would not help improve tone recognition accuracies if the surface tones have been correctly labeled, indicating the corpus is consistently labeled in high quality.

Table 5: *Performances of tone recognition on the test sets of AISHELL3 by the raw model and the relabeling model.*

	T1	T2	T3	T4	T5	accuracy
raw	0.913	**0.871**	**0.819**	0.909	0.765	**0.879**
relabeling	**0.915**	0.864	0.764	**0.916**	**0.781**	0.872

With the results shown in this subsection, we may conclude that the proposed transformer-based tone recognizer could achieve a competitive performance to the existing method.

IV. Data Analysis

This section provides a quantitative analysis of the relationship between the labeled tones and prosodic structures for the Treebank multi-SR corpus. Note that here we analyzed the cases except for the cases of tone sandhi and polyphones. Table 6 shows the percentages of prosodic boundary types preceding and following target syllables with combinations of true and predicted tone classes. The boundaries are categorized into three types:1) intra-PW boundary with B0 and B1, denoted as "intra," 2) PW boundary with B2-1/2/3, denoted as "PW," and 3) PPh boundary with B3, B4, and Be, denoted as "PPh." Numbers in square brackets denote the number of syllables with true-predicted tone pairs. The numbers in brackets are percentages of the three prosodic boundaries.

A. Pitch Reset at PW/PPh Initial Positions

It was found that the following misrecognized cases may be caused by F0 resets that overwrite syllable tone realizations at the beginning of PWs, PPhs, or PGs:

1) The T3 syllables misrecognized as tones with higher F0 registers, that is, T4s (698 samples), T2s (643 samples), and T1s (83 samples), were more likely to follow PW or PPh boundaries than the correctly recognized T3 cases (intra:PW:PPh=63:22:15). The misrecognized cases (698+643+83 samples) were contributed by 470, 354, 218, and 382 samples, respectively, from the fast to slow sub-corpora.
2) The T4 syllables misrecognized as T1s (528 samples) were more likely to follow PPh boundaries than correctly recognized T4 cases (intra:PW:PPh=64:25:11). The 528 samples consist of 228, 139, 71, and 90 samples from the fast to slow sub-corpora.

The above two cases also show that more overwritten tone realizations with F0 resets can be observed when the SR is faster.

Table 6: *Percentages of prosodic boundary types preceding* (B_{n-1}) *and following* (B_n) *the target syllables* (s_n) *with various combinations of the true* ($\hat{t}_n$) *and predicted* (t_n) *tone classes.*

			Predicted Class t_n									
			T1		T2		T3		T4		T5	
			B_{n-1}	B_n	B_{n-1}	B_n	B_{n-1}	B_n	B_{n-1}	B_n	B_{n-1}	B_n
True Class $\hat{t}_n$	T1	intra	**55**	**68**	62	**56**	71	**31**	61	67	**87**	12
		PW	**33**	**21**	10	**37**	10	**29**	21	26	**15**	17
		PPh	**12**	**11**	18	**7**	19	**40**	18	7	**0**	71
			[35190]		**[326]**		**[108]**		[262]		[52]	
	T2	intra	62	**89**	**65**	**66**	69	**68**	71	**81**	**85**	61
		PW	27	**10**	**23**	**23**	17	**12**	24	**10**	**15**	15
		PPh	11	**1**	**12**	**11**	14	**20**	5	**9**	**0**	24
			[260]		[44455]		**[309]**		[165]		[54]	
	T3	intra	19	78	**49**	76	**63**	**64**	**48**	68	**81**	56
		PW	39	21	**22**	18	**22**	**26**	**29**	22	**19**	29
		PPh	42	1	**29**	6	**15**	**10**	**23**	10	**0**	15
			[83]		**[643]**		[27240]		**[698]**		[99]	
	T4	intra	**53**	66	69	55	65	**39**	**64**	**59**	**92**	32
		PW	**26**	31	20	40	21	**42**	**25**	**27**	**8**	51
		PPh	**21**	3	11	5	14	**19**	**11**	**14**	**0**	17
			[528]		[281]		**[594]**		[62671]		[83]	
	T5	intra	69	0	97	32	98	12	95	23	**94**	**29**
		PW	31	16	3	58	2	48	5	54	**6**	**50**
		PPh	0	84	0	10	0	40	0	13	**0**	**21**
			[32]		[110]		[92]		[56]		[2351]	

B. Pitch Lowering at PW/PPh Final Positions

The following misrecognized cases occurred in the endings of PWs, PPhs, and PGs, manifesting the final F0 lowering:

1) The T4 syllables misrecognized as T3s (594 samples) were more likely to be followed by PW or PPh boundaries than the correctly recognized T4 cases (intra:PW:PPh=59:27:14). The 594 samples consisted of 136, 142, 97, and 219 samples, respectively, from the fast to slow sub-corpo
2) The T4 syllables misrecognized as T3s (594 samples) were more likely to be followed by PW or PPh boundaries than the correctly recognized T4 cases (intra:PW:PPh=59:27:14). The 594 samples consisted of 136, 142, 97, and 219 samples, respectively, from the fast to slow sub-corpora.
3) The T1 syllables misrecognized as tones with lower F0 registers, that is, T2s (326 samples) or T3s (108 samples), were more likely to be followed by PW or PPh boundaries than the correctly recognized T1 cases (intra:PW:PPh=68:21:11). The 326+108=434 samples were composed of 158, 104, 69, and 103, respectively, from the fast to slow sub-corpora.
4) The T2 syllables misrecognized as T3s (309 samples) were more likely to be followed by PPh boundaries than the correctly recognized T3 cases (intra:PW:PPh=66:23:11). The 309 samples were contributed by 92, 60, 63, and 94, respectively, from the fast to slow sub-corpora.

These three cases also show that F0 lowering at the end of a prosodic unit is not consistently correlated with SRs.

C. Coarticulation in PW/PPh Medial Positions

It has been observed that the magnitude of F0 is larger at edges than at medial positions, given that syllable edges have a longer duration [29]. The following misrecognized cases may have resulted from insufficient length and coarticulation.

1) The syllables with true-predicted tone pairs of T2→T1 (T2 misrecognized as T1, 260 samples) and T2→T4 (165 samples) were more likely to be followed by intra-PW boundaries that may be more likely to cause co-articulations than PW and PPh boundaries. The 260+165=425 samples were contributed by 222, 96, 49, and 58 samples, respectively, from the fast to slow sub-corpora, showing that syllables with faster SRs would be more likely to co-articulate with the following syllables.
2) Misrecognized syllables with T1→T5 (52 samples), T2→T5(54 samples), T3→T5(99 samples), and T4→T5 (83 samples) were more likely to follow the intra-PW boundaries. The 52+54+99+83=288 samples were contributed by 124, 97, 29, and 39 samples from the fast to slow sub-corpora, respectively, showing that the

syllables in faster SRs would be more easily pronounced as neutral tones (T5).

3) Articulatorily, laryngeal movements are at the physiological limit of maximum speed in medial positions; therefore, contour tones tend to be flattened. The misrecognition with T4→T1 can be explained by the insufficient length in the medial positions.

D. Small-Scale Perception Test

By investigating some tone-misrecognized samples with high likelihoods (>0.90) for nontarget tones, we suspect that they may be perceived as the predicted tones. Eighteen native Mandarin speakers without speech or hearing impairments participated in a small-scale perception test. The test was divided into three parts:1) PW/PPh initial with true-predicted tone pair T3→T4 (2 tokens), 2) PW/PPh final with true-predicted tone pair T3→T4 (1 token), and 3) PP medial with true-predicted tone pairs T2→T1 (3 tokens) and T4→T1 (4 tokens). The above 2+1+3+4=10 misrecognized tokens were used as the treatment group and randomized with the control group (20 tokens) composed of correctly predicted tokens in the same corresponding prosodic locations. Each token is an excerpted speech segment centered around the target syllable, with one preceding syllable and one following syllable. In each test of the three parts, the tones of tokens in the treatment and control groups were mutually exclusive. For example, the control group in the PW/PPh medial test comprised tokens with target syllables T1 and T3, while the treatment group consisted of T2 and T4 tokens. Eighteen subjects were asked to identify the tone of the target syllable when playing a token.

In the PW/PPh initial positions, the accuracy rates for T1, T2, and T4 were 83.3%, 88.9%, and 90.8%, respectively, whereas the accuracy rate for T3 was only 51.44%. T3 was identified with 29.36% of responses of T4. The accuracy rates for T1, T2, T3, and T4 in the PW/PPh final positions were 83.4%, 96.3%, 100%, and 94.5%, respectively. Finally, in the PW/PPh medial positions, the accuracy rates for T1 and T3 were 85.23% and 88.9%, respectively, whereas those for T2 and T4 were only 9.3% and 3.73%, respectively. T2 was identified with 72.3% of responses of T1, and T4 was identified with 72.3% of responses of T1. Put it differently, tokens with low accuracy rates in the perception test were also misrecognized by the tone recognizer. Meanwhile, the results of this small-scale perception test may indicate that the tones predicted by the proposed deep learning-based tone recognizer may reflect the actual phonetic realizations of the speaker; in corrected speech, the pitch of T3 was likely to be raised due to pitch reset in PP initial positions and was misheard as T4 by native speakers. In PP-medial positions, T2 and T4 were likely to be shortened because of insufficient length and coarticulation, such that they were misperceived as T1 by native speakers.

V. Conclusions and Future Works

This study reported the use of a deep learning-based tone recognition model for Mandarin tone labeling. Not only the labeling for tone sandhi and polyphones, but also the misrecognized cases associated with prosodic structures were investigated. In the future, conducting more tone labeling experiments on Mandarin or Chinese dialect corpora with more speaker variety and conducting more subjective tests to examine the tone identities of syllables in various prosodic contexts would be worthwhile.

VI. References

[1] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines." arXiv, Apr. 22, 2021. doi: 10.48550/arXiv.2010.11567.

[2] Y. Zhang, L. Deng, and Y. Wang, "Unified Mandarin TTS Front-end Based on Distilled BERT Model," 2020, doi: 10.48550/ARXIV.2012.15404.

[3] J. Pan *et al.*, "A Unified Sequence-to-Sequence Front-End Model for Mandarin Text-to-Speech Synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6689–6693. doi: 10.1109/ICASSP40776.2020.9053390.

[4] Z. Bai and B. Hu, "A Universal Bert-Based Front-End Model for Mandarin Text-To-Speech Synthesis," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 6074–6078. doi: 10.1109/ICASSP39728.2021.9414935.

[5] H. Zhang, "PDF: Polyphone Disambiguation in Chinese by Using FLAT," in *Interspeech 2021*, Aug. 2021, pp. 4099–4103. doi: 10.21437/Interspeech.2021-1087.

[6] S. Zhang, K. Zheng, X. Zhu, and B. Li, "A polyphone BERT for Polyphone Disambiguation in Mandarin Chinese," in *Interspeech 2022*, Sep. 2022, pp. 436–440. doi: 10.21437/Interspeech.2022-229.

[7] Y.-C. Chen, Y.-C. Chang, Y.-C. Chang, and Y.-R. Yeh, "g2pW: A Conditional Weighted Softmax BERT for Polyphone Disambiguation in Mandarin," 2022, doi: 10.48550/ARXIV.2203.10430.

[8] J. Cui *et al.*, “Efficient Text Analysis with Pre-Trained Neural Network Models,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, Jan. 2023, pp. 671–676. doi: 10.1109/SLT54892.2023.10022565.

[9] S.-H. Chen *et al.*, “Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 7, pp. 1158–1171, Jul. 2014, doi: 10.1109/TASLP.2014.2321482.

[10] G.-T. Liou, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, “An Exploration of Local Speaking Rate Variations in Mandarin Read Speech,” in *Interspeech 2018*, Sep. 2018, pp. 42–46. doi: 10.21437/Interspeech.2018-1214.

[11] C. Y. Chiang, G. T. Liou, Y. R. Wang, and S. H. Chen, “Method of generating estimated value of local inverse speaking rate (ISR) and device and method of generating predicted value of local ISR accordingly,” US20210035598A1, Feb. 04, 2021

[12] J.-L. Huang, “Mandarin Tone Pronunciation Error Detection – Computer Assisted Laguage Learning for Foreign Speaker,” National Taipei University of Technology, 2012.

[13] H. Hussein, H. Mixdorff, and R. Hoffmann, “Real-Time Tone Recognition in A Computer-Assisted Language Learning System for German Learners of Mandarin,” Dec. 2012.

[14] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, “Improving Mispronunciation Detection of Mandarin Tones for Non-Native Learners With Soft-Target Tone Labels and BLSTM-Based Deep Tone Models,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2012–2024, Dec. 2019, doi: 10.1109/TASLP.2019.2936755.

[15] C.-C. Chung, “A Mandarin Keyword Spotting System Assisted with Tone Recognition,” National Chiao Tung University, 2011.

[16] Y.-B. Wang, S.-W. Li, and L. Lee, “An Experimental Analysis on Integrating Multi-Stream Spectro-Temporal, Cepstral and Pitch Information for Mandarin Speech Recognition,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2006–2014, Oct. 2013, doi: 10.1109/TASL.2013.2263803.

[17] J. Lin, Y. Xie, Y. Gao, and J. Zhang, “Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features,” in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Tianjin, China, Oct. 2016, pp. 1–5. doi: 10.1109/ISCSLP.2016.7918472.

[18] L. Yang, Y. Xie, and J. Zhang, “Applying Deep Bidirectional Long Short-Term Memory to Mandarin Tone Recognition,” in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, Beijing, China, Aug. 2018, pp. 1124–1127. doi: 10.1109/ICSP.2018.8652486.

[19] L. Lugosch and V. S. Tomar, “Tone Recognition Using Lifters and CTC,” in *Interspeech 2018*, Sep. 2018, pp. 2305–2309. doi: 10.21437/Interspeech.2018-2293.

[20] L. Peng, W. Dai, D. Ke, and J. Zhang, “Multi-Scale Model for Mandarin Tone Recognition,” in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong, Jan. 2021, pp. 1–5. doi: 10.1109/ISCSLP49672.2021.9362063.

[21] J. Tang and M. Li, "End-to-End Mandarin Tone Classification with Short Term Context Information," *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Tokyo, Japan, 2021, pp. 878-883.

[22] R. Tong, N. F. Chen, B. Ma, and H. Li, “Goodness of tone (GOT) for non-native Mandarin tone recognition,” in *Interspeech 2015*, Sep. 2015, pp. 801–805. doi: 10.21437/Interspeech.2015-254.

[23] S. Prom-on, Y. Xu, and B. Thipakorn, “Modeling tone and intonation in Mandarin and English as a process of target approximation,” *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, Jan. 2009, doi: 10.1121/1.3037222.

[24] Y. Gao, X. Zhang, Y. Xu, J. Zhang, and P. Birkholz, “An Investigation of the Target Approximation Model for Tone Modeling and Recognition in Continuous Mandarin Speech,” in *Interspeech 2020*, Oct. 2020, pp. 1913–1917. doi: 10.21437/Interspeech.2020-2823.

[25] C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen, “Fluent speech prosody: Framework and modeling,” *Speech Communication*, vol. 46, no. 3–4, pp. 284–309, Jul. 2005, doi: 10.1016/j.specom.2005.03.015.

[26] “The Table of Chinese Polyphone Characters.” https://language.moe.gov.tw/result.aspx?classify_sn=42&subclassify_sn=443 (accessed Mar. 06, 2023).

[27] TensorFlow Developers, “TensorFlow.” Zenodo, Mar. 07, 2023. Accessed: Mar. 08, 2023. [Online]. Available: https://zenodo.org/record/4724125

[28] A. Vaswani *et al.*, “Attention Is All You Need,” 2017, doi: 10.48550/ARXIV.1706.03762.

[29] Y. Xu and M. Wang, “Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin,” *Journal of Phonetics*, vol. 37, no. 4, pp. 502–520, Oct. 2009, doi: 10.1016/j.wocn.2009.08.003.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Learning to Enhance the Position Embedding and Coherence

Ting-Jia Shu and Jen-Tzung Chien
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Transformer is known as a powerful learning machine for sequence-to-sequence learning. One weakness of this attention-based machine is that the attention mechanism is not sensitive to position information. The absolute position embedding lacks of the generalization to an unknown test sequence where the length is longer than those of training sequences. It becomes important to adopt the relative position embedding in transformer, but the computation load is significantly increased. This paper deals with this issue through a new relation-aware position embedding which properly characterizes the position information with the limited computation overhead where the property of position coherence is held. The merit of the proposed method is illustrated by the experiments on neural machine translation in terms of BLEU and computation time,**

I. Introduction

The attention-based model [1], [2], [3], [4], [5] has achieved a much better performance than traditional recurrent neural network (RNN) [6], [7] without attention in a variety of natural language processing (NLP) tasks [8], [9]. Transformer [1], [10], [11] has been a successful NLP machine which is based on the self-attention mechanism. The standard attention scheme lacks of a natural encoding method for word ordering. The sine and cosine positional embedding was accordingly introduced in the vanilla transformer in order to address this issue. In general, transformer adopts the so-called absolute position embedding (APE) [1], [12] which converts each position index into the corresponding position embedding. The number of word tokens in a sentence is seen as the length of position embeddings. Such a position embedding is then added to the token embedding to form the input to the model. This is an effective and simple method, but there is an insufficiency [13], [14], [15] by using this method. Namely, once there are some test sentences which are longer than training sentences or exceed the length of sequences that the trained model can handle. Figure 1 illustrates the BLEU value of test sentence versus the length of source sentence where the transformer is implemented for machine translation in the task of WMT14. To evaluate the capability of a model to face the mismatch in the lengths of training sentences, the transformer is optionally trained from the selected sentences where the length is shorter than 50 word tokens. The red curve shows the results by using the complete training sentences while the blue curve reports the BLEUs by only using the selected sentences. The gray area means the case that the test sentences are longer than the training sentences. It can be seen that the BLEU is almost the same when the sentence length is shorter than 50. But, there is a clear gap in the metric of BLEU between two models in the gray area. Obviously, the trained transformer [16], [11], [17] has an issue of generalization in representation of different lengths of sentences. To cope with this issue, there are some alternatives where the relative position embedding (RPE) [18], [19], [20], [21], [22] was proposed. RPE is basically based on the relative distance between two tokens to express the position information. The scheme of relative position embedding incorporates the position information in the self-attention mechanism [23], [24], [25]. Unfortunately, since the attention mechanism exists in each layer of transformers, this approach requires additional computational cost which is proportional to the number of stacked transformer layers [26], [27].

To enhance the relation-awareness in position embedding without too much computation overhead, this paper proposes a relation-aware position embedding by preserving the coherence in the relative position and regularizing the difference between the position embeddings for those of input tokens with the same relative distance. In the implementation, in addition to the addition of sinusoidal position embeddings and token embeddings, the inputs to the model further merge the learned relation-aware position embeddings. The proposed position embedding and coherence are evaluated by the experiments of using transformer for neural machine translation.

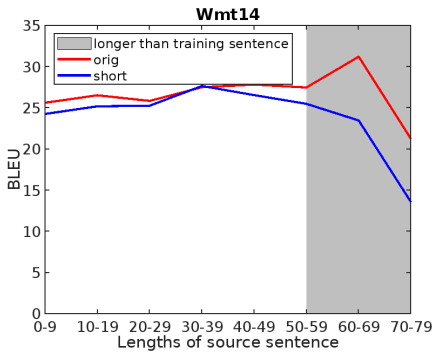


Fig. 1. BLEU value of test sentence versus the length of source sentence in a machine translation task where the transformer is implemented.

II. Position Embeddings

A. Absolute position embedding

Absolute position embedding (APE) is added with the token embedding as the input to learn the vanilla transformer. APE basically provides the information on word order in a sentence while the token embedding conveys the high-level information

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

on semantics. APE in transformer is seen as a fixed function which is composed of sine and cosine functions as follows

$$\mathrm{PE}(i,j) = \begin{cases} \sin\left(i/10000^{\frac{j}{d}}\right) & \text{if } j \text{ is even} \\ \cos\left(i/10000^{\frac{j}{d}}\right) & \text{else} \end{cases} \quad (1)$$

where i denotes the position of a token in a sequence and j denotes the dimension index for positional encoding vector.

B. Relative position embedding

Although sinusoidal position embedding does not require additional parameters and is easy to implement, the insufficiency of sinusoidal position embedding is caused when the mismatch in sentence lengths between training and test data happens. The work in [18] presented an approach based on the relative position embedding (RPE). A key difference of RPE compared with APE is made due to the circumstance that the position information of current token is encoded in a relation to those of the other tokens in the sentence. To fulfill this concept, two relative position representations are trained with the combination of self-attention [28] in a transformer. These two representations are described below.

1) Relative position representations: Basically, the edges between word tokens x_i and x_j could be used to represent the relative positional differences between two input tokens. The maximum of absolute value of a clipping distance of two tokens r is used to express the maximum of relative position in real situation. It is because that beyond a certain distance exacting the relative position information becomes useless. This method allows the model to generalize for sequence lengths which are unseen during the training phase. There are $2r+1$ unique edges $\{a_{ij}^K, a_{ij}^V\}$ which are given by [18]

$$\begin{aligned} a_{ij}^K &= w^K_{\mathrm{clip}(j-i,k)}, \quad a_{ij}^V = w^V_{\mathrm{clip}(j-i,k)} \\ \mathrm{clip}(x,k) &= \max(-k, \min(k,x)). \end{aligned} \quad (2)$$

where K and V denote the key and value in self-attention, respectively. The relative position representations $w^K = \left(w^K_{-k}, \ldots, w^K_k\right)$, $w^V = \left(w^V_{-k}, \ldots, w^V_k\right)$ are accordingly taken into account as shown in Figure 2 where $w_i^K, w_i^V \in \mathbb{R}^d$.

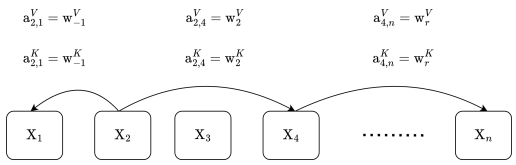


Fig. 2. Illustration for an example of the distinct edge representations between word tokens. The representations for each position relative within a clipping distance r can be learned.

2) Relation-aware self-attention: On the other hand, the self-attention in transformer was extended by considering the pairwise relations between input tokens. A relation-aware self-attention was proposed by learning two distinct edge representations a_{ij}^V and a_{ij}^K which are distinct for two individual tokens x_i and x_j. For an efficient implementation, each head for the edges for key and value $\{a_{ij}^K, a_{ij}^V\}$ is shared. After having the attention weights, the distance information is incorporated into self-attention in a form of

$$e_{ij} = \frac{x_i w^Q \left(x_j w^K + a_{ij}^K\right)^\top}{\sqrt{d}} \quad (3)$$

where w^Q is the relative position representation of query in self-attention. This distance measure is then used to calculate the attention weights for a sequence of length n via the following softmax function

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^n \exp e_{ik}}. \quad (4)$$

Unlike the standard self-attention, the way of calculating latent variable z_i is not only fulfilled by multiplying the attention weights by using value matrix, but also performed by considering the distance information. To do so, the edge a_{ij}^V and value are summed and then multiplied by attention weight α_{ij}. The self-attention is modified by considering the distinct edge representation by

$$z_i = \sum_{j=1}^n \alpha_{ij} \left(x_j w^V + a_{ij}^V\right). \quad (5)$$

This modification is essential for the tasks where the information about the edge types chosen by a specific attention head is helpful to enrich encoder or decoder layers for downstream tasks. Basically, the position embedding method in [18] was developed to enhance the model generalization, but this position embedding method suffered from a relatively high amount of parameters and calculations [29], [30], [31]. In order to cope with the problem of increasing number of parameters, it was proposed [18] that the same relative position embedding for each head was shared. For an n-length sequence with h attention heads and the dimensions of each head d, the memory complexity of relative position representation was reduced from $\mathcal{O}(hn^2d)$ to $\mathcal{O}(n^2d)$ through the sharing of position embedding across individual heads. However, the overhead of training procedure was still considerable. The reduction in training time for the experiments on machine translation compared with standard transformer was only reduced by 7% in steps per second.

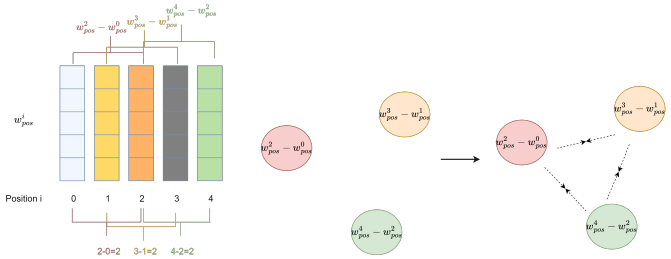


Fig. 3. Illustration of relation-aware position embedding which aims to minimize the difference between any two pairs of position embeddings with the same relative distance.

III. Relation-Aware Position Embedding

This study proposes a new relation-aware position embedding to deal with the issue of model generalization in presence of various sequence lengths without increasing computation overhead, The proposed relation-aware position embedding incorporates the relative position information into position embedding and pursues the coherence of embedding by reducing the difference between the position embeddings for those input tokens with the same relative distance. Figure 3 illustrates the key idea of this method. The relation-aware position embedding is implemented by minimizing the difference between any two pairs of position embeddings s and r with the same distance. An intuitive way to handle this issue is to introduce a learning objective

$$\sum_{s=0}^{\text{maxlen}-k} \sum_{r=0}^{\text{maxlen}-k} \left((w_{\text{pos}}^{s+k} - w_{\text{pos}}^{s}) - (w_{\text{pos}}^{r+k} - w_{\text{pos}}^{r})\right)^2 \quad (6)$$

where maxlen denotes the maximum length of sentence, k denotes the relative distance and $w_{\text{pos}} = \{w_{\text{pos}}^0, w_{\text{pos}}^1, \ldots, w_{\text{pos}}^{\text{maxlen}}\}$ denotes the relation-aware position embedding.

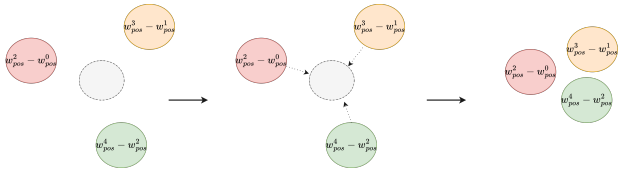


Fig. 4. Implementation of an efficient method to relation-aware position embedding with two stages. First stage is to calculate the mean of those differences between the pairs of position embeddings. Second stage is to subtract them from the distances of individual pairs.

A. Position embedding and regularization

However, such a way suffers from substantial computation overhead if the difference between two pairs of embeddings with the same distance is directly minimized. This study presents a new learning objective to pursue embedding coherence in relation-aware position embedding by first calculating the means of those differences between two position embeddings and then subtracting them from the corresponding distance between the pair of position embeddings as shown in Figure 4. The computation overhead can be considerably reduced. To learn for position embedding and coherence, the coherence objective is therefore formulated by

$$\sum_{s=0}^{\text{maxlen}-k} \left((w_{\text{pos}}^{s+k} - w_{\text{pos}}^{s}) - \frac{1}{\text{maxlen}-k} \sum_{j=0}^{\text{maxlen}-k} (w_{\text{pos}}^{j+k} - w_{\text{pos}}^{j})\right)^2 \quad (7)$$

where only one summation in two stages is used. Mean subtraction aims to promote the coherence of embedding difference. However, there will be an issue happening when calculating the average value. Considering the case of $k = 2$, the mean value is calculated by

$$\begin{aligned} \sum_{j=0}^{\text{maxlen}-2} (w_{\text{pos}}^{j+k} - w_{\text{pos}}^{j}) &= (w_{\text{pos}}^{1} - w_{\text{pos}}^{0}) \\ &+ (w_{\text{pos}}^{2} - w_{\text{pos}}^{1}) + \cdots + (w_{\text{pos}}^{\text{maxlen}} - w_{\text{pos}}^{\text{maxlen}-k}) \\ &= w_{\text{pos}}^{\text{maxlen}} - w_{\text{pos}}^{0}. \end{aligned} \quad (8)$$

It is found that the intermediate terms will be removed during the continuous addition. Some information is not counted. To prevent this insufficiency, this paper presents a learning objective by considering the square of the relation between the pair of position embeddings as follows

$$\sum_{s=0}^{\text{maxlen}-k} \left((w_{\text{pos}}^{s+k} - w_{\text{pos}}^{s})^2 - \frac{1}{\text{maxlen}-k} \sum_{j=0}^{\text{maxlen}-k} (w_{\text{pos}}^{j+k} - w_{\text{pos}}^{j})^2\right)^2 \quad (9)$$

where all difference terms are preserved in the calculation. The coherence of relative embedding is optimized to regularize the training of transformer.

In addition to formal design of learning objective, this study further extends the setting of the relative distance in the implementation as illustrated in Figure 5. Ideally, training the model with a relative distance of 1 (as shown in left subfigure) would allow the model to extrapolate the information from any pair tokens with the same relative distance. However, the actual situation may happen when the situation in right subfigure is met. We can see that although the distance between w_{pos}^2 and w_{pos}^1 is consistent with the distance between w_{pos}^2 and w_{pos}^3 but the distance to w_{pos}^1 is not double the distance to w_{pos}^3.

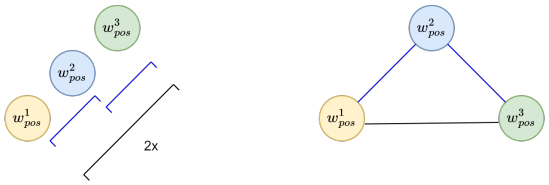


Fig. 5. (Left) An ideal case when model only needs to use the relative distance to be 1, (Right) The case that can be extrapolated to the other distances.

B. Learning objectives

The proposed method is composed of the trainable position embeddings and the transformer blocks in encoder and decoder for sequence-to-sequence learning. There are two learning objectives in the downstream task for machine translation where the parameters of position embeddings and transformer blocks are trained. In what follows, the mechanics of merging different objectives to form the training objective $\mathcal{L}$ are addressed. First of all, the cross entropy (CE) loss is formulated according to the summation of the negative conditional likelihood of output sequence given input sequence

$$\mathcal{L}_{ce} = -\sum_{x,y} \log p(y_t | y_{<t}, x) \quad (10)$$

where $x = \{x_t\}$ denotes the input string, $y = \{y_t\}$ denotes the output sequence, and $y_{<t}$ is the history of y_t at token t. Next, the second term in $\mathcal{L}$ is formed as the regularization term $\mathcal{L}_{\text{pos}}$ which is used to reflect the relative position information in position embeddings with dimension d. In Eq. (9), only a specific distance k is considered as the relative position information. Therefore, the regularization term is calculated by considering various values of k in relative position

$$\mathcal{L}_{\text{pos}} = \sum_{k} \sum_{s=0}^{\text{maxlen}-k} \Big((w_{\text{pos}}^{s+k} - w_{\text{pos}}^{s})^2 - \frac{1}{\text{maxlen} - k} \sum_{j=0}^{\text{maxlen}-k} (w_{\text{pos}}^{j+k} - w_{\text{pos}}^{j})^2 \Big)^2. \tag{11}$$

The final learning objective is constructed by using the cross entropy loss and the relation-aware position embedding loss

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{pos}} \tag{12}$$

where λ is a regularization parameter.

Algorithm 1: Training procedure for relation-aware position embedding in transformer

input: input and output sequences $\mathcal{D} = \{x, y\}$
learning rate α, activation function $f(\cdot)$
initial encoder and decoder parameters $\phi_{\text{enc}}^{(0)} = \phi_{\text{dec}}^{(0)} = 0$
initial relation-aware position embed. $w_{\text{pos}} = 0$
initial annealing ratio between sine position encoding and relation-aware position embed. $r_{\text{pos}} = 0$
repeat
  sample mini-batch from $\mathcal{D}$
  pos embed. = sine pos encoding $+ r_{\text{pos}} * w_{\text{pos}}$
  encoder input z_0^x = word embed $(x)+$ pos embed.
  decoder input z_0^y = word embed $(y)+$ pos embed.
  for *each encoder layer* $\ell = 1, \ldots, L$ **do**
    compute $a_{\text{enc}}^{(\ell)} = \phi_{\text{enc}}^{(\ell)} * z_{\text{enc}}^{(\ell-1)}$, $z_{\text{enc}}^{(\ell)} = f(a_{\text{enc}}^{(\ell)})$
  end
  for *each decoder layer* ℓ **do**
    compute $a_{\text{dec}}^{(\ell)} = \phi_{\text{dec}}^{(\ell)} * z_{\text{dec}}^{(\ell-1)}$, $z_{\text{dec}}^{(\ell)} = f(a_{\text{dec}}^{(\ell)})$
  end
  $z_{\text{out}} = \text{feedforward}(z_{\text{dec}}^{L})$
  $\hat{y} = \text{softmax}(z_{\text{out}})$
  calculate cross entropy loss $\mathcal{L}_{\text{ce}}$ in (10)
  calculate relation aware position loss $\mathcal{L}_{\text{pos}}$ in (11)
  calculate $\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{pos}}$
  update transformer parameter $\phi \leftarrow \phi - \alpha \frac{\partial \mathcal{L}}{\partial \phi}$
  update relation embedding $w_{\text{pos}} \leftarrow w_{\text{pos}} - \alpha \frac{\partial \mathcal{L}}{\partial w_{\text{pos}}}$
until *training converged*

C. Learning algorithm

Algorithm 1 shows the learning procedure for transformer encoder and decoder $\phi = \{\phi_{\text{enc}}, \phi_{\text{dec}}\}$ as well as relation-aware position embeddings. The proposed position embedding is calculated by combining the sinusoidal position embedding and the relation-aware position embedding with an annealing ratio parameter r_{pos}. Such embedding is then added by the word embedding and used to calculate the inputs $\{z_0^x, z_0^y\}$ to feed in encoder and decoder for finding the transformer output z_{out}. The objectives of training transformer parameters ϕ and relation-aware position embeddings w_{pos} based on $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{pos}}$, respectively, are then jointly optimized to fulfill the whole model for a downstream task of sequence-to-sequence learning for machine translation between English and German with the evaluation as shown in the experiments.

IV. Experiments

There are two datasets IWSLT'14 De-En (German to English) and WMT'14 En-De (English to German) [32] collected to evaluate the proposed method for sequence-to-sequence learning. The subsequent experiments were carried out by using the Fairseq toolkit [33]. For the task on WMT'14 En-De, the tokens of English sentences were first encoded into word embeddings which were then added by the mixed position embeddings from the sinusoidal position embedding and the proposed relation-aware embedding. Simultaneously, the tokens of German sentences were also encoded by using the same way so as to calculate the decoder inputs. After feeding the inputs into transformer, the encoding of English sentences as the hidden states have been decoded into those for German sentences. The overall model is then jointly optimized according to the cross-entropy loss regularized by the relation-aware position embedding loss.

A. Experimental settings

IWSLT'14 De-En translation task contained around 167K training sentence pairs and 7K testing sentence pairs. This corpus consisted of the TED talk transcriptions and their translations. In the experiments, the dataset was separated into training, validation, and test sets containing 160K, 7K, and 7K sentence pairs, respectively. On the other hand, WMT'14 En-De translation task contained roughly 3.9M, 39K, 3K of sentence pairs pairs in training, validation, and test sessions, respectively. The byte pair-encoding (BPE) [34] was applied to segment the words into subword units. In order to make the test data longer than the train data, this study removed the WMT'14 training data with the sentence length longer than 50 and the IWSLT'14 training data with the sentence length longer than 40. After removing, the WMT'14 training and test data were resized to 3.5M and 0.2K sentence pairs and IWSLT'14 training and test data were reduced to 137K and 0.8K sentence pairs, respectively. The performance of translation models were evaluated in terms of BLEU score [35]. The vanilla transformer was implemented as the baseline model. The word embedding size and position embedding size have been set as 512 for WMT'14 and 256 for IWSLT'14, respectively. The dimension of intermediate hidden layer of a feed forward network was set as 2048 for WMT'14 and 1024 for IWSLT'14. There were six encoder and decoder blocks in model configuration. To control the size of model parameters, the dimension of position embedding vector was set as 128,

and the vector was concatenated four times. The parameter r_{pos} was set as 2 and the relative distance was set from 1 to 15.

TABLE I
COMPARISON OF THE RESULTS ON WMT'14 EN-DE. THE RESULTS MARKED BY * ARE REFERRED FROM [36].

Model	Params	BLEU	Training time
Transformer [1]	61M	27.66*	–
RPR [18]	–	27.92*	–
CAPE [36]	–	27.89*	–
Transformer	66.5M	27.75	1.00
RA transformer	66.7M	**28.18**	1.009

TABLE II
COMPARISON OF THE RESULTS ON IWSLT'14 DE-EN.

Model	Params	BLEU
Transformer [1]	39.5M	34.54
RA Transformer	39.7M	**34.91**

B. Experimental results

This study implements the so-called relation-aware (RA) transformer with the proposed relation position embedding and compares the performance of machine translation using vanilla transformer [1] and RA transformer. First of all, the results of WMT'14 and IWSLT'14 tasks are shown for the case where there is no removal of training data. In the WMT'14 task as shown in Table I, the lower part shows the results of our implementation. The comparison is made in terms of the number of parameters, the BLEU and the training time relative to vanilla transformer. It is found that the proposed RA transformer obtains 0.43 absolute improvement in BLEU, spends only 0.3 $\%$ overhead of memory requirement for model parameters, and takes only 0.9$\%$ longer training time when compared with vanilla transformer. Relative to the related works including the methods called the relative position representation (RPR) [18] and the continuous augmented positional embeddings (CAPE) [36], it is found that the proposed RA transformer achieved 0.29 improvement in BLEU compared with state-of-the-art method CAPE. In the IWSLT'14 task as shown in Table II, RA transformer obtains an absolute improvement of 0.37 in BLEU score over standard transformer with very limited overhead in model size. The model size of WMT'14 is much larger than that of IWSLT'14. These results demonstrate that RA transformer considerably performs better than vanilla transformer in various tasks.

Next, the same evaluation is performed by considering the case that the pairs of long sentences are removed from training data. Total number of sentence pairs in training data is reduced. This set of experiments is to assess the generalization of the performance of BLEU when the mismatch in sentence length happens. As shown in the results on WMT'14 and IWSLT'14 as given in Table III, it is still found that the proposed RA transformer obtains higher BLEU scores than vanilla transformer in both English-to-German and German-to-English tasks. The robustness of RA transformer in presence of various lengths of sentences is illustrated. The BLEU values in Table III are lower than those in Tables I and II because the size of training data is reduced.

TABLE III
COMPARISON OF THE RESULTS ON WMT'14 EN-DE AND IWSLT'14 DE-EN WHERE THE PAIRS OF LONG SENTENCES ARE REMOVED FROM TRAINING DATA.

Model	Params	BLEU
Transformer (WMT)	66.0M	23.98
RA Transformer (WMT)	66.2M	**24.15**
Transformer (IWSLT)	39.3M	20.43
RA transformer (IWSLT)	39.5M	**20.71**

TABLE IV
COMPARISON OF THE RESULTS ON IWSLT'14 DE-EN IN PRESENCE OF DIFFERENT RANGES OF RELATIVE DISTANCES k.

Model	k	BLEU
Transformer	-	34.54
RA transformer	1 to 5	34.60
RA transformer	1 to 10	34.72
RA transformer	1 to 15	**34.91**
RA transformer	1 to 20	34.74

C. Ablation study

Following the setting of preserving all pairs of training sentences, an ablation study on IWSLT'14 DE-EN is further examined by evaluating the effect of adopting different ranges of relative distances k, i.e. considering multiple relative distances. As shown in Table IV. the highest BLEU score is achieved in case of the range of k from 1 to 15. Even only considering the range from 1 to 5, there is still slight improvement obtained by using RA transformer. These results using the proposed RA transformer are consistently better than those using standard transformer.

V. CONCLUSIONS

This paper presented a new relation-aware position embedding and combined it with traditional word embedding and sinusoidal position embedding. This new embedding brings relative position information to the transformer model where the cross-entropy loss and the relation-aware position information loss were jointly minimized. The experiments on machine translation showed that the BLEU values were increased by using the resulting relation-aware transformer compared with the vanilla transformer where the overheads in memory requirement and computation cost were relatively limited. The effect of relative distances was also evaluated. Future studies will be further extended for the other downstream tasks for sequntial learning based on transformer.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[5] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al., "Recipes for building an open-domain chatbot," in *Proc. of Conference of the European Chapter of the Association for Computational Linguistics*, 2021, pp. 300–325.

[6] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, pp. 1735–1780, 1997.

[7] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[8] Shinji Watanabe and Jen-Tzung Chien, *Bayesian Speech and Language Processing*, Cambridge University Press, 2015.

[9] Dong Yu, Geoffrey Hinton, Nelson Morgan, Jen-Tzung Chien, and Shigeki Sagayama, "Introduction to the special section on deep learning for speech and language processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 4–6, 2012.

[10] Chi-Hang Leong, Yu-Han Huang, and Jen-Tzung Chien, "Online Compressive Transformer for End-to-End Speech Recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2082–2086.

[11] Jen-Tzung Chien and Yu-Han Huang, "Bayesian Transformer Using Disentangled Mask Attention," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 1761–1765.

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, "Convolutional sequence to sequence learning," in *Proc. of International Conference on Machine Learning*, 2017, pp. 1243–1252.

[13] Dusan Varis and Ondřej Bojar, "Sequence length is a domain: Length-based overfitting in transformer models," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 8246–8257.

[14] Jan Rosendahl, Viet Anh Khoa Tran, Weiyue Wang, and Hermann Ney, "Analysis of positional encodings for neural machine translation," in *Proc. of International Conference on Spoken Language Translation*, 2019.

[15] Pan Zhou, Ruchao Fan, Wei Chen, and Jia Jia, "Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding," *arXiv preprint arXiv:1911.00203*, 2019.

[16] Jen-Tzung Chien and Chia-Wei Tsao, "Graph evolving and embedding in transformer," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 538–545.

[17] Hou Lio, Shang-En Li, and Jen-Tzung Chien, "Adversarial mask transformer for sequential learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4178–4182.

[18] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," in *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 464–468.

[19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[20] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck, "Music transformer: Generating music with long-term structure," in *Proc. of International Conference on Learning Representations*, 2018.

[21] Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang, "Improve transformer models with better relative position embeddings," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 3327–3335.

[22] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen, "On position embeddings in BERT," in *Proc. of International Conference on Learning Representations*, 2021.

[23] Jen-Tzung Chien and Yi-Hsiang Chen, "Continuous-time attention for sequential learning," *Proc. of AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 7116–7124, 2021.

[24] Jen-Tzung Chien and Yi-Hsiang Chen, "Learning continuous-time dynamics with attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1906–1918, 2023.

[25] Jen-Tzung Chien and Hsiu-Wei Tien, "Variational disentangled attention and regularization for visual dialog," in *Proc. of International Joint Conference on Neural Networks*, 2023, pp. 1–9.

[26] Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Févry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al., "Do transformer modifications transfer across implementations and applications?," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 5758–5773.

[27] Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui, "SHAPE: Shifted absolute position embedding for transformers," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3309–3321.

[28] Jen-Tzung Chien and Ting-An Lin, "Supportive and self attentions for image caption," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 1713–1718.

[29] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

[30] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya, "Reformer: The efficient transformer," in *Proc. of International Conference on Learning Representations*, 2020.

[31] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al., "Rethinking attention with performers," in *Proc. of International Conference on Learning Representations*, 2021.

[32] Jen-Tzung Chien and Wei-Hsiang Chang, "Dualformer: A unified bidirectional sequence-to-sequence learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7718–7722.

[33] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "FAIRSEQ: A fast, extensible toolkit for sequence modeling," in *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 48–53.

[34] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.

[36] Tatiana Likhomanenko, Qiantong Xu, Gabriel Synnaeve, Ronan Collobert, and Alex Rogozhnikov, "CAPE: Encoding relative positions with continuous augmented positional embeddings," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 16079–16092.

VLSI Design of Near-Lossless Image Compression using Improved LZW

Yao-Zhong Zhang[1],Chiung-An Chen*[1], Jia-Sheng Zhang[1],and Jia-Wen Wang[1], *Member, IEEE*

[1]Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City, Taiwan

An improved encoding method based on Lempel-Ziv-Welch (LZW) is proposed, utilizing the advantages of VLSI parallel processing to overcome the drawback of lengthy encoding time in LZW. The improved LZW encoding method enhances the encoding efficiency by 90% and simultaneously improves the compression ratio by 30%. The near - lossless image compression process in this paper incorporates integer DCT, Huffman coding, run-length encoding, and the improved LZW encoding. Compression experiments were conducted on a Kodak image dataset, resulting in an average bit rate of 7.13 bpp (Bits Per Pixel) and a PSNR of 44.5.

Keywords—Image Compression, LZW Algorithm, Parallel

I. INTRODUCTION

With the advancement of technology, the images captured by cameras are becoming increasingly sophisticated. When photos contain richer details and colors, the amount of data needed to store them will significantly increase. The inevitably affects the speed of uploading and other processes, making image compression an essential operation. Image compression reduces the data size of images, resulting in three advantages: shortened upload and download times, moreover, storage space occupation less, and lower power consumption [1-3].

Image compression is the processing of compressing image data into smaller storage space or narrower transmission bandwidth. Traditional image compression methods typically use either lossless compression or lossy compression [4]. Lossless compression methods compress image data into smaller storage space without losing image quality, while lossy compression methods can compress image data into smaller storage space but may affect image quality. Many image compression methods have been proposed, such as JPEG [5], PNG, and GIF, and there is an increasing demand for more efficient compression methods. Therefore, the motivation behind image compression research is to develop more effective image compression techniques to meet the needs of modern applications and improve efficiency in storing and transmitting image data .

This paper aims to design an improved LZW encoding compression hardware circuit for wireless transmission. The design are significant differences between software programming languages and hardware description languages [6-7]. The former offers flexibility, rich functionality, data abstraction, object-oriented programming, diverse databases, such as C, Java, Python, MATLAB, and so on, with the latter has advantages in sequential circuits and parallel processing, achieving simultaneous processing of multiple procedures through different timing executions to increase work efficiency. Overall, the computational performance of VLSI is higher than that of software.

II. IMPROVED LZW ENCODING

Traditional LZW encoding [8-9] consists of a prefix code (P), current character (C), and a dictionary. The encoding steps are as follows: check if PC exists in the dictionary, if not, output P, add PC to the dictionary, and update P as C. If PC already exists in the dictionary, update P to the code for PC in the dictionary. Repeat these steps until the encoding is complete. As each PC signal needs to be compared with the data in the dictionary to determine if it has been encoded, the comparison time increases as the dictionary grows larger. This leads to longer encoding times for traditional LZW encoding. Taking ACABAC as an example, following the traditional LZW method, the encoded result is ACAB1. Therefore, traditional LZW encoding has two issues: a large number of entries in the dictionary (dictionary occupancy) as shown in Table 1 Part A, and empty codes as shown in Table 1 Part C.

Table 1. Traditional and improved LZW encoding methods

Part. A	Traditional LZW encoding		Part. B	Improved LZW encoding	
	1	AC		1	AC
	2	CA		2	AB
	3	AB			
	4	BA			
Part. C	Example		ACABAC		
	Traditional LZW encoding		ACAB1		
	Improved LZW encoding		121		

This paper presents improvements to the Traditional LZW encoding, as shown in Figure 2. The enhancement lies in the approach: when LZW fails to find a matching PC value within the dictionary, it incorporates that PC value into the dictionary and promptly outputs the corresponding encoding, as shown in Step2 of the Table 2. Subsequently, the next two data entries are consecutively retrieved from the buffer and designated as P and C. This ensures the correct functioning of the subsequent LZW process, as elaborated in Steps 3 to 4 in Table 2 and Figure 1. This strategic method guarantees that each key within the dictionary is used at least once, mitigating the occurrence of encoded values that remain unused, as shown in Parts B and C of the Table 1. The Improved LZW encoding technique introduces two pivotal advantages: a more compact dictionary, exemplified in Parts A and B of the Table 1; and reduced output, as shown in Part C of the Table 1. These improvements significantly elevate the efficiency of the LZW encoding process.

Table 2. Encoding steps and dictionary generation

	(Prefix, Current)	Encoded string	Dictionary		Output
			Index	Key	
Step1	(, A)				
Step2	(A, C)	AC	1	AC	1
Step3	(, A)				
Step4	(A, B)	AB	2	AB	2
Step5	(, A)				
Step6	(A, C)	AC			1

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

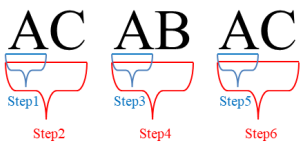


Figure 1. Steps of improved LZW encoding method using ACABAC

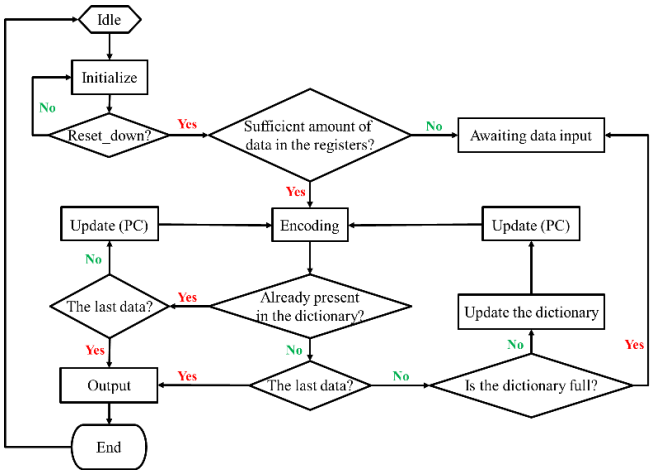


Figure 2. Flowchart of the improved LZW method

The improved VLSI design of the LZW encoding, as proposed in this paper, is illustrated in Figure 3. It consists of five modules: ***Buffer***, ***Found***, ***PC_data***, ***Code_out***, and ***LZW_Ctrl***. The *Buffer* module serves as a register, primarily designated for accessing the input data (*data_in*). The *Found* module encompasses a **Comparator**, **Counter**, **Dictionary**, and **Code_stage**, as shown in Figure 4. A comprehensive description of their functions is provided in Table 3. Notably, the signals *rd*, *we*, *en_coun*, and *en_camp* are generated by the *LZW_Ctrl* module, controlling the actions of dictionary read/write, comparator, and counter operations.

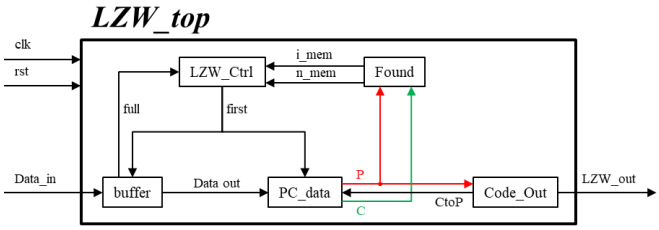


Figure 3. VLSI design of improved LZW encoding

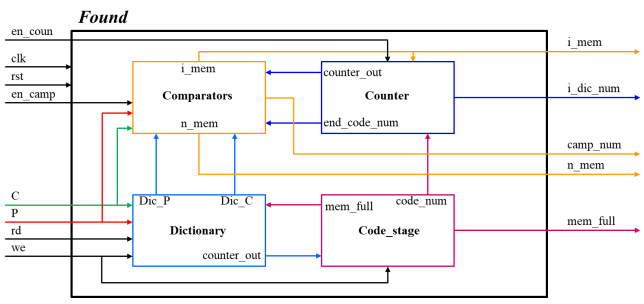


Figure 4. Block diagram of the ***Found*** module

Table 3. Description of the ***Found*** module

Found	Description
Comparators	By contrasting the current *P* and *C* values with the *Dic_P* and *Dic_C* in the dictionary, the encoding status of the PC pair is determined. This process generates the signals *i_mem* (indicating presence in the dictionary) and *n_mem* (indicating absence in the dictionary). These signals are utilized to inform the control module about the current status of whether the *PC* value is already present in the dictionary.
Counter	Based on the current number of encoded entries (*Code_num*), the selection of dictionaries for comparison by the comparator is determined. When the *i_mem* signal is high, the counter outputs the current count value, serving as the position of this *PC* pair within the dictionary, thereby generating the *i_dic_num* signal. On the other hand, *end_code_num* indicates to the comparator whether the dictionary entry being processed is the last one that has been encoded. This signal can enhance encoding efficiency when the dictionary is not fully populated.
Dictionary	The dictionary simultaneously outputs multiple pairs of *Dic_P* and *Dic_C* for comparison by the comparator, while *camp_num* indicates in which dictionary the current PC is located. The input signals *rd* and we determine whether to perform an output operation or write to the dictionary.
Code_stage	The *we* signal is used to update the current count of encoded elements. In LZW encoding, if the dictionary size is not restricted, the number of entries in the dictionary can continuously expand with the increase of data, leading to an escalation of required encoding bits. This might even surpass the size of the original data. Consequently, it is customary to set a limitation on the dictionary size (which can be tailored to specific requirements) for encoding purposes. Additionally, the state module includes a *mem_full* signal, indicating whether the dictionary module can continue writing entries to the dictionary.

The *PC_data* module, as shown in Figure 5, serves as the processing unit for *P* and *C* values in the context of LZW encoding. This module primarily makes decisions based on comparison results, determining whether to output the *P* value or transform *P* into an encoded value from the dictionary.

Within the module, *datain* represents the output value of the LZW register, and the first signal is generated by the *LZW_Ctrl* module, indicating that this is the first data entry that doesn't require encoding but only awaits the next data input. *P* and *C* signify the current *PC* values and are updated according to corresponding conditions. *in_memnum* is produced by the Found module as *camp_num*, informing us about the dictionary where this *PC* value exists. *dic_num* corresponds to *i_dic_num* in the Found module, denoting the address within that dictionary. When the comparison outcome is existence, these sets of information are used to update the *P* value.

When the *PC* value isn't present in the dictionary, the *Code_out* module outputs the current *P* value and generates the *CtoP* signal after the output completion. At this point, the *P* value is updated to *C*, while C is updated to the next data entry, creating a new *PC* value to proceed with encoding operations. Upon reaching the last data entry, the Found module similarly generates the *last_imem* and *last_nmem* signals. The former indicates the continued existence in the dictionary, while the latter signifies its

absence. When *last_imem* is at a high voltage level, the *P* value is updated to the encoded value, halting the encoding process. Meanwhile, when *last_nmem* is at a high voltage level, both *P* and *C* maintain their current states and are sequentially outputted. Subsequently, the encoding process concludes.

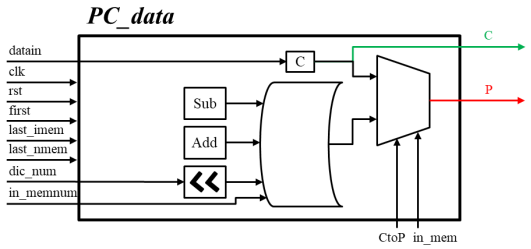


Figure 5. Block diagram of the ***PC_data*** module

The *Code_Out* module, as shown in Figure 6, encompasses several critical signals. Firstly, there's the *end_data* signal, originating from the LZW register. When this signal is set to a high voltage level, the current *P* value is outputted, and concurrently, the *end_code* signal is also set to a high voltage level, indicating the end of encoding. The we signal, on the other hand, is provided by the *LZW_Ctrl* module. When it transitions to a high voltage level, the current *PC* value is written into the dictionary, which is subsequently reset. During this time, the *P* value is also outputted to ensure error-free encoding. Simultaneously, the *CtoP* signal is raised, notifying both the *PC_data* module and the *LZW_Ctrl* module that the output has been completed, enabling them to proceed with reading the next set of data for encoding. The *rst_CtoP* signal, provided by the *LZW_Ctrl* module, resets the *CtoP* signal to prevent erroneous operations in the *PC_data* module during *PC* value updates, thus averting encoding errors.

Lastly, the *last_nmem* signal is also generated by the *LZW_Ctrl* module, signifying one of the outcomes of the final encoding. At this juncture, the *PC_data* module maintains the present state of the PC value, while the Output module concurrently outputs the *P* and *C* values. Simultaneously, the *end_code* signal is set to a high voltage level, culminating the encoding operation.

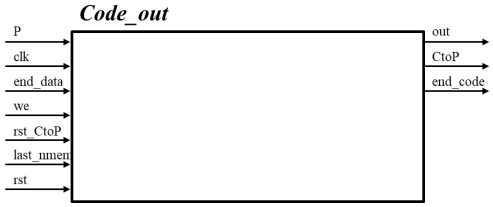


Figure 6. Block diagram of the ***Code_Out*** module

The *LZW_Ctrl* module, as shown in Figure 7, constitutes the central component of the entire control system. This module comprises ten distinct states, each with specific functionalities. To begin, when the system is in the ***IDLE*** state, it signifies an idle state. Only when the start signal is triggered does the system transition to the subsequent state, ***INIT_RAM***. In this state, the system generates the *rst_ram* signal, which serves to clear all registers. This action ensures that the results of the previous encoding do not influence the current encoding.

Following this, the system progresses to the ***CONFIRM_RAM*** state. Here, the state awaits the clearance of all registers. Once this is accomplished, the system proceeds to the *DATAIN* state. During the ***DATAIN*** state, the system awaits sufficient data accumulation in the LZW buffer (based on the full signal) to ensure an adequate amount of data for encoding. The system then advances to the ***ENCODE_START*** state. In this phase, the system unconditionally transitions to the *P_C_data* state and simultaneously issues the first signal. This signal informs the system that this is the first set of data, obviating the need for encoding and only necessitating the arrival of the second data set for input.

Within the ***P_C_data*** state, the system emits the *en_data* and *rst_CtoP* signals. The former controls the timing of LZW buffer data readouts, while the latter clears the *CtoP* signal within the Output module to prevent anomalies in the *PC_data* module during updates. Subsequently, the system unconditionally enters the ***DATA*** state and, upon the first trigger, returns to the *P_C_data* state to process the second set of data. When the *in_mem* signal surfaces, indicating the presence of a corresponding *PC* value in the dictionary, the system resets the comparator's enable signal and then proceeds to the *FOUND* state.

In all other circumstances, the system unconditionally enters the ***FOUND*** state. Within this state, the enable signals for the comparator, counter, and dictionary are activated and maintained until the comparison outcome materializes. Based on the post-comparison result, the system undertakes actions contingent on the existence of the *PC* value in the dictionary. If the *PC* value is present in the dictionary, the comparator's enable signal persists, and the *en_data* signal is set to a high voltage level, subsequently reverting to the DATA state. If the PC value is absent in the dictionary, the enable signal is deactivated, simultaneously triggering the we signal (the write signal for the dictionary module), and then progressing to the *OUTRAM* state.

During the ***OUTRAM*** state, the system awaits the completion of output from the Output module while concurrently resetting the *we* and *rst_CtoP* signals. It then reverts to the *P_C_data* state to continue with the next encoding until the LZW buffer emits the *end_data* signal. At this juncture, the final data is encoded, and regardless of the *PC*'s presence in the dictionary, the system enters the *OUTRAM* state. Depending on the scenario, either *last_imem* or *last_nmem* is emitted. The former indicates the final encoding exists in the dictionary, while the latter suggests otherwise. When the *last_imem* signal emerges, the *P* value is converted into the dictionary's address and directly outputted. Conversely, when the *last_nmem* signal is high, the *P* and *C* values maintain their present states and are sequentially outputted. Subsequently, the system enters the ***ENCODE_END*** state, wherein the *en_dicout* signal is transmitted to the dictionary module to output encoding data from the dictionary. This action primarily facilitates recovery operations by software. Upon the complete output of all dictionary data, the system returns to the *IDLE* state, awaiting the next encoding process.

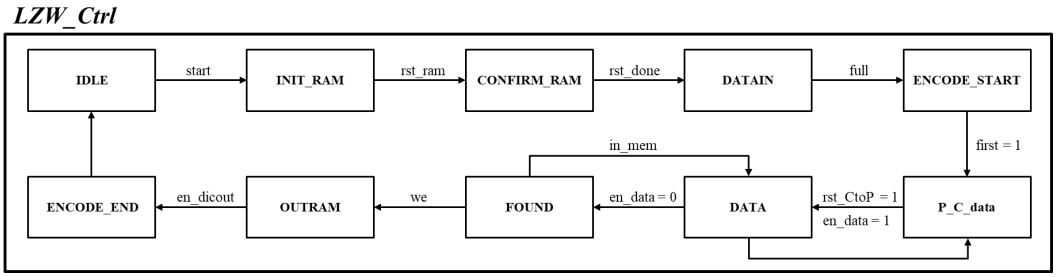


Figure 7. Block diagram of the ***LZW_Ctrl*** module

Table 4. BBP and PSNR data for compressing the Kodak image dataset using B. Koc et al. [10] and the improved LZW

Method	Kodak01		Kodak02		Kodak03		Kodak04		Kodak05	
	BPP	PSNR	BPP	PSNR	BPP	PSNR	BPP	PSNR	BPP	PSNR
B. Koc et al. [6]	9.28	51.3	8.07	51.0	7.12	51.2	8.15	51.1	9.92	51.2
Improved LZW	5.70	44.5	5.69	43.6	5.15	45.2	6.21	44.3	6.54	44.8

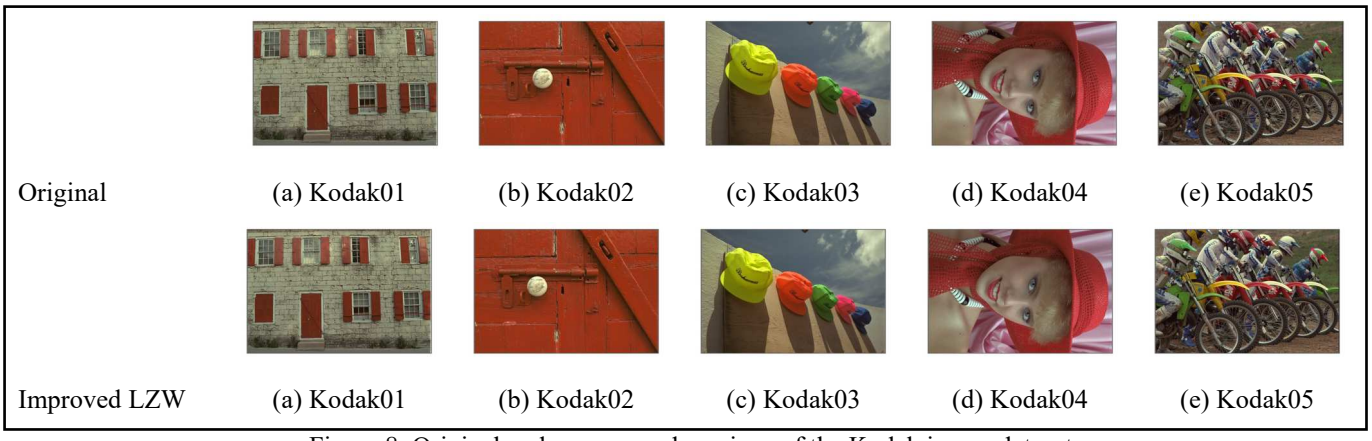


Figure 8. Original and compressed versions of the Kodak image dataset

III. RESULTS

In this experiment, both the traditional LZW algorithm [8-9] and the improved LZW algorithm were subjected to VLSI design for image compression. The test image used was an 8-bit color Lena image with a size of 256*256 pixels. The test data results are presented in Table 5. It is noteworthy that while the traditional LZW algorithm required 91.73 milliseconds to complete the image compression process, the improved LZW algorithm presented in this paper achieved image compression in just 0.91 milliseconds. This demonstrates a remarkable speed improvement of 100 times when compared to the traditional method. This underscores the effectiveness of splitting the dictionary into multiple smaller dictionaries and performing simultaneous comparisons in significantly mitigating the encoding drawbacks of LZW.

Table 5. Seconds for VLSI circuits of traditional and improved LZW methods in lena image compression

Method	Time (seconds)
Improved LZW	**0.91 m**
Traditional LZW	91.73 m

Table 4 presents the shift compression algorithm proposed by B. Koc, Z. Arnavut, et al. [10], using the Kodak image dataset. Table 4 shows our method achieves better compression rates compared to the approach proposed by B. Koc, Z. Arnavut, et al. The original images from the Kodak dataset and the compressed image dataset using our method (with the improved LZW encoding as the core) are shown in Figure 8. Although the PSNR value is slightly lower, it still remains at around 44dB. Therefore, our proposed compression algorithm achieves good compression rates and can be applied for faster image transmission in wireless communication.

Table 6 presents the gate count simulated by the Vivado software along with the clock frequency used in this experiment. The operation frequency employed in this experiment is 50MHz. The gate count typically ranges from LUT multiplied by 4 to 6. Consequently, the gate count in this experiment is around 17,000, which is not considered substantial in today's circuit designs. This gate count aligns with the original intent of this experiment, which is to apply it to portable electronic devices.

Table 6. VLSI design for image compression in this paper, simulated with Vivado software

Operation Frequency	LUT	Gate Count (k)
50MHz	2849	17

IV. Conclusion

This paper combines DCT, Huffman coding, run-length encoding, and the improved LZW encoding for near-lossless image compression. Due to the DCT encoding in the image compression process, it leads to a slightly lower PSNR value. However, the average PSNR value remains at 44.5. The focus of image compression is on the improved LZW encoding, which not only achieves fast encoding but also exhibits high compression efficiency. As shown in Table 5, our method demonstrates a rapid encoding rate, allowing for quick compression of each image. By effectively and swiftly compressing the images, data size is significantly reduced, as indicated in Table 4. This facilitates fast transmission in wireless communication, simultaneously reducing power consumption and extending battery life.

References

[1] P. W. Jones and M. Rabbani, Digital Image Compression Techniques. 7th ed., Washington USA: Spie Optical Engineering Press, 1991

[2] Byoung-Il Kim and S. G. Ziavras, "Low-power multiplierless DCT for image/video coders," 2009 IEEE 13th International Symposium on Consumer Electronics, Kyoto, 2009, pp. 133-136, doi: 10.1109/ISCE.2009.5156873.

[3] L. -H. Chen, C. G. Bampis, Z. Li, A. Norkin and A. C. Bovik, "ProxIQA: A Proxy Approach to Perceptual Optimization of Learned Image Compression," in IEEE Transactions on Image Processing, vol. 30, pp. 360-373, 2021, doi: 10.1109/TIP.2020.3036752.

[4] Rabbani, Majid., Jones, Paul W.. Digital Image Compression Techniques. USA: Spie Optical Engineering Press, 1991.

[5] JPEG: Still Image Data Compression Standard by William B. Pennebaker and Joan L. Mitchell, Kluwer Academic Publishers, ISBN: 0442012721

[6] H. Kagawa, Y. Ito and K. Nakano, "Throughput-Optimal Hardware Implementation of LZW Decompression on the FPGA," 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), Nagasaki, Japan, 2019, pp. 78-83, doi: 10.1109/CANDARW.2019.00022.

[7] Y. -L. Tsai and J. -J. Ding, "An Improved LZW Algorithm for Large Data Size and Low Bitwidth per Code," TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON), Auckland, New Zealand, 2021, pp. 203-208, doi: 10.1109/TENCON54134.2021.9707201.

[8] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," in IEEE Transactions on Information Theory, vol. 23, no. 3, pp. 337-343, May 1977, doi: 10.1109/TIT.1977.1055714.

[9] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," in IEEE Transactions on Information Theory, vol. 24, no. 5, pp. 530-536, September 1978, doi: 10.1109/TIT.1978.1055934.

[10] B. Koc, Z. Arnavut, S. Voronin and H. Koçak, "Near-lossless Image Compression with Parity Reduction," 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES), Reykjavík, Iceland, 2020, pp. 225-230, doi: 10.1109/INES49302.2020.9147124.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

The color demosaicing and image scaling based on improve Hamilton-Adams

Yu-Wen Peng [1], Chia-Yu Hu [1], Yen-Ju Chin [1], He-Sheng Chou [1], Yuan-Jin Lin [2], Yu-Lin Liu [1], Shih-Lun Chen [1]*
Tsung-Yi Chen [3], Kuo-Chen Li [4], and Chiung-An Chen [5]
*Department of Electronic Engineering, Chung Yuan Christian University, Chung Li City 32023, Taiwan;
E-mail: s10926313@cycu.edu.tw; s10926209@cycu.edu.tw; s10926204@cycu.edu.tw; g11176050@cycu.edu.tw; g10776018@cycu.edu.tw; chrischen@cycu.edu.tw; Tel: +886-3-265-4644
Department of Electrical Engineering, National Cheng Kung University, Taiwan;
E-mail: m26124069@gs.ncku.edu.tw; Tel: +886-3-265-4644
Department of Electronic Engineering, Feng Chia University, Taichung City 407102, Taiwan;
E-mail:tsungychen@fcu.edu.tw; Tel: +886-4-2451-7250#4910
Department of Information Management, Chung Yuan Christian University, Chung Li City 32023, Taiwan;
E-mail:kuochen@cycu.edu.tw; Tel: +886-3-265- 2050
Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 243303, Taiwan
E-mail:joannechen@mail.mcut.edu.tw; Tel: +886-2-2908-9899#4813

***Abstract* —The demosaicing process is the most critical step in image processing, this study presents a Hamilton-Adams-based algorithm for mosaic removal and image scaling. The algorithm consists of three steps: image pre-processing, image scaling, and improved Hamilton-Adams mosaic removal. The image pre-processing stage employs a mean filter to reduce high-frequency signals at the image edges, effectively mitigating the zipper effect caused by image scaling and improving image quality. The improved Hamilton-Adams method incorporates high-quality linear interpolation to achieve more accurate edge computations. In terms of image quality assessment, this paper increased 1.13 dB of PSNR, 1.19 dB of CPSNR, and 0.03 of SSIM values compared to high-quality linear interpolation with bicubic interpolation.**

KEYWORDS: Color filter array, Image pre-processing, Demosaicing, Hamilton-Adams, Image scaling

I. INTRODUCTION

Consumer electronic products commonly feature image signal processors, such as digital cameras and smartphones. This product internally incorporates a charge-coupled device (CCD) or a complementary metal-oxide-semiconductor and utilizes the color filter array (CFA) for image acquisition. Most of the CFAs use the Bayer pattern [1], as shown in Fig.1, where a group of 2×2 pixels consists of two green filters, one blue filter, and one red filter. These components selectively filter light to capture color information and achieve the reconstruction of full-color images. The Bayer CFA pattern is widely employed in consumer electronic devices due to its simplicity and effectiveness, contributing to enhancing the luminance information and overall image quality.

After capturing the image using the CFA, only one color (red, green, or blue) can be recorded for each pixel, resulting in a monochrome representation. By applying a demosaicing algorithm, such as bilinear interpolation (BI) [2], it is possible to reconstruct the RGB values for each pixel. However, this approach often leads to unnatural colors and the appearance of zipper artifacts. Other methods, such as High-quality linear interpolation (HQLI) [3] and Hamilton-Adams (HA) [4], have been proposed to mitigate zipper artifacts and provide more realistic color reproduction.

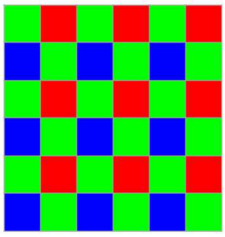


Fig. 1. Bayer format.

A. Linear interpolation demosaicing

HQLI introduces a high-quality linear interpolation demosaicing method for Bayer pattern color images. This method emphasizes the preservation and restoration of image details. Linear interpolation demosaicing is a prevalent technique that employs linear interpolation methods to interpolate the missing color values for each pixel in the image. Unlike some traditional methods that may suffer from edge blur and loss of fine details, this HQLI method can better preserve the details and edge structures of the image, providing a more realistic and clear visual effect.

This study compares the proposed method with other interpolation methods through experiments to validate its effectiveness. The experimental results demonstrate that this HQLI method produces superior results in the interpolation

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

process of Bayer pattern images, offering a more realistic and meticulous color restoration.

B. *Hamilton-Adams demosaicing*

HA demosaicing [4] describes a device for processing digital image signals acquired from a sensor. The sensor has color pixels aligned in rows and columns, where each pixel position generates at least three independent color values, but only one color value is available.

The device includes memory for storing digital image signals and a processor connected to the memory. The processor performs interpolation to obtain missing color values based on the different color values of nearby pixel positions. The processor also includes a structure for obtaining Laplacian values, gradient values, and color difference deviation values from neighboring pixels in the same column and row. Subsequently, a classifier is formulated by aggregating these values, and the preferred direction for interpolating the missing color values is determined based on the classifier. The device presents a technique for interpolating multiple designated color values within the proximity, ensuring conformity with the preferred direction.

This method presents a hardware-friendly implementation by utilizing chroma difference for blue and red channel calculation, thereby eliminating the requirement for complex multiplication or extensive memory. However, dealing with images containing intricate edges may produce slight zipper artifacts, which can be considered a limitation of the algorithm. Nonetheless, in terms of hardware implementation, it provides an efficient and straightforward approach.

The previous two methods only focus on demosaicing and cannot compress computational data, requiring extensive delays or memory storage in hardware implementation. Therefore, [5] introduces an innovative, low-complexity joint color demosaicing and digital zooming algorithm, providing an effective and high-quality solution to image processing for single-sensor digital cameras.

Traditionally, color demosaicing and digital zooming are separate steps performed using different techniques. However, the algorithm proposed in this paper adopts a novel approach by combining these two steps for improved results.

This algorithm utilizes the edge information extracted from the original sensor data to perform direct interpolation during the color demosaicing and digital zooming processes. This approach has several advantages. Firstly, it ensures the preservation of edge features in the output, thereby enhancing image clarity and details. Secondly, due to the elimination of separate information extraction processes, the algorithm exhibits lower computational complexity, enabling higher processing speeds. Most importantly, this joint processing method leverages the correlation between the two steps, resulting in more accurate results.

Method [5] demonstrates exceptional performance in generating scaled full-color images and scaled Bayer color filter array images, surpassing the conventional separate processing methods. However, a drawback of this method is its reliance on substantial computational resources during hardware implementation, which may limit its suitability for scenarios.

This study employs down-sampling, bicubic interpolation [6], and mosaic removal combined with image preprocessing to address the issue of coarse image data and ensure hardware feasibility. These techniques effectively tackle the problem of coarse image data. However, the resulting zipper effect caused by compression necessitates the use of image preprocessing to mitigate this issue. For hardware implementation convenience, this study utilizes an improved version of the HA demosaicing [4], followed by sequential steps of image down-sampling, image preprocessing, HA demosaicing [4], and bicubic interpolation [6], the flow chart is shown in Fig. 2.

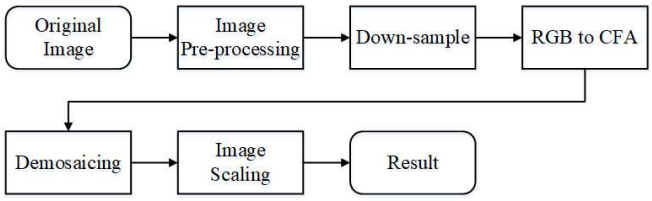


Fig. 2. Flow chart of the image processing.

II. PROPOSAL WORK

A. *Down sample*

This study utilized the Kodak dataset [7] with original image dimensions of 768×512. This study employed down-sampling to reduce the size of the images simply and efficiently. Down-sampling involves discarding every other row and column, resulting in a downsized image of 384×256, the size is one-fourth the size of the original image, and the concept is shown in Fig. 3.

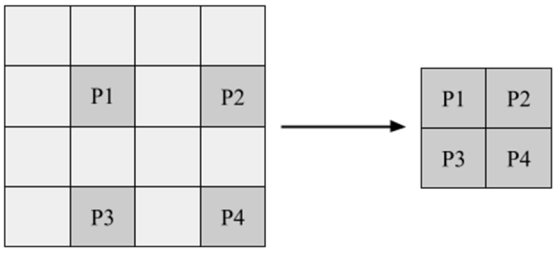


Fig. 3. The schematic diagram of down-sampling.

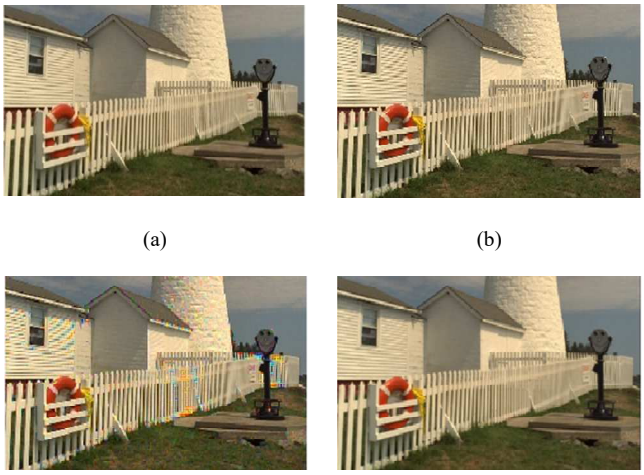
B. *Image preprocessing*

After the down-sampling process, the image loses a significant amount of pixel values, resulting in a jagged or wavy texture, as shown in Fig.4 (b). Due to the absence of a single texture direction, which is not horizontal or vertical, it becomes challenging to determine the texture direction

accurately along the demosaicing edges. This leads to the creation of zipper artifacts in the image, as illustrated in Fig.4 (c). To address this issue, smoothing the texture is beneficial for demosaicing direction determination. A low-pass filter is employed to eliminate high-frequency signals in the image, resulting in a smoother texture. The simplest low-pass filter is the mean filter (MF) [8], formula (1) calculates the average value of pixels within the convolution kernel's region and assigns this average value to the pixel at the center of the kernel. Let $g(s,t)$ represent the original image, and $f(x,y)$ denotes the image obtained after applying the mean filter. Formula (2) is M x N chosen convolutional kernel, this study uses a 3×3 size kernel:

$$f(x,y) = \frac{1}{mn} \sum_{(x,y) \in Sxy} g(s,t) \tag{1}$$

$$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{2}$$



(a) (b)

(c) (d)

Fig.4.(a)Original(b)Down sampling without preprocessing (c)Demosaicing without preprocessing (d)Down sampling with preprocessing.

C. *Improved Hamilton-Adams demosaicing*

This section presents the utilization of HA as the fundamental framework, augmented with algorithmic enhancements through HQLI. By integrating color difference information into HA, the computational process becomes more efficient and precise in determining edge directions. Moreover, HQLI excels in preserving fine details in smooth regions. Consequently, by combining these two methods, a three-step demosaicing approach is devised. The first step involves restoring the G pixels in the R (B) channels. To calculate the missing green pixels, the R and B channels are used. The texture on the edges is classified into vertical, horizontal, and no specific texture. According to the HA, the gradient formula for the horizontal and vertical directions is as (3):

$$\begin{cases} H = |G_2 - G_3| + |2 \times R_k - R_2 - R_3| \\ V = |G_0 - G_1| + |2 \times R_k - R_0 - R_1| \end{cases} \tag{3}$$

H represents the vertical gradient, and V represents the horizontal gradient. As illustrated in Fig.5 (a), the red denotes the central point and calculated the missing green pixel value. The direction is determined by comparing the gradient values, the pixel value is computed according to the texture direction. G_k represents the missing green pixel value for the center point G_k, the calculation formula for G_k is shown in (4):

$$\begin{cases} H > V & G_k = \frac{(G_0 - G_1)}{2} + \frac{(2 \times R_k - R_0 - R_1)}{2} \\ H < V & G_k = \frac{(G_2 - G_3)}{2} + \frac{(2 \times R_k - R_2 - R_3)}{2} \end{cases} \tag{4}$$

This study replaces the calculation parameters for the direction where H equals V and no specific texture, as shown in Fig.5 (b), with those from HQLI [3], which introduces the concept that edges have greater luminance variations than chromaticity. In comparison with the BI estimation of the central red value in Fig.5 (a) to its nearest red sample, if it differs from this estimated value, it implies a significant luminance variation at that pixel. To leverage the advantages of HQLI, estimated luminance variations are added to correct the green value of Bilinear Interpolation [2]. The correction formula is shown below (5):

$$G_K = \widehat{G_k} + \Delta\alpha \times \Delta R_k \tag{5}$$

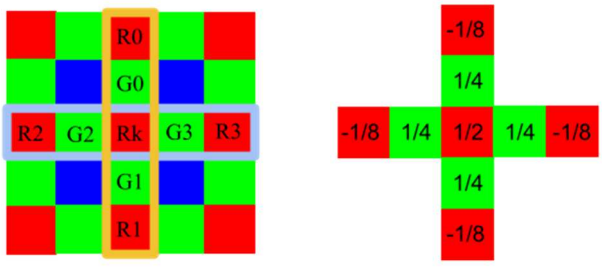
Where $\widehat{G_k}$ represents the estimated value obtained using BI, $\Delta\alpha$ is the parameter controlling the strength of this correction formula. HQLI proposes using the Wiener method to find the most suitable parameter value, and in the paper, it suggests $\Delta\alpha=1/2$ for better results. Meanwhile, ΔR_k represents the gradient of R_k at that location, used for estimating luminance changes, and its calculation formula is as follows (6):

$$\Delta R_k \triangleq R_k - \frac{1}{4}(R_0 + R_1 + R_2 + R_3) \tag{6}$$

After determining the strength parameter for the correction formula, the calculation formula for the HQLI can be expressed as follows (7):

$$G_k = \frac{1}{2} \times R_k + \frac{1}{4}(G_0 + G_1 + G_2 + G_3) - \frac{1}{8}(R_0 + R_1 + R_2 + R_3) \tag{7}$$

For the remaining pixels as shown in Fig.5 (a), the same procedure applies, but this time using blue as the center point instead of red to calculate the missing green pixel value.



(a) (b)

Fig.5. Using red color at the center point.

The second step involves restoring the missing red or blue

pixels in the G channel. Building upon the previous step, where the missing green pixels were already filled in, as indicated by the green triangles in Fig.6 (a), this step utilizes the parameters from HQLI to calculate the missing pixel values, as shown in Fig.6 (b). HQLI incorporates an estimated luminance change to correct the values obtained from BI. The correction formula is as follows (8):

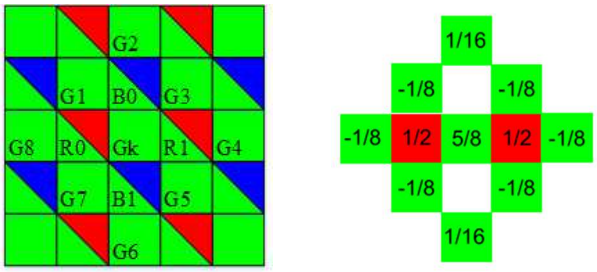
$$R_K = \widehat{R_k} + \Delta\beta \times \Delta G_k \quad \textbf{(8)}$$

Here, $\widehat{R_k}$ represents the estimated values obtained using BI, and HQLI suggests that $\Delta\beta$ = 5/8 yields optimal results. ΔG_k denotes the gradient at that position for G_k. After obtaining the strength parameter for the correction formula, the calculation formula for HQLI is organized as follows (9) to compute the missing red pixel value R_k for the center point G_k:

$$\mathrm{R}_k = \frac{5}{8} \times \mathrm{G}_k + \frac{1}{2} \times (\mathrm{R}_0 + \mathrm{R}_1) + \frac{1}{16}(\mathrm{G}_2 + \mathrm{G}_6) - \frac{1}{8} \times (\mathrm{G}_1 + \mathrm{G}_3 + \mathrm{G}_8 + \mathrm{G}_7 + \mathrm{G}_5 + \mathrm{G}_4) \quad (9)$$

Similarly, the center point G_k is also missing blue pixel values. This case employs the calculation method mentioned in HA. B_k represents the missing blue pixel for the center point, and GB_0 and GB_1 are the restored green pixels in the B channel obtained from the previous step. The formula for computing GB_0 is shown in (10):

$$\mathrm{B}_k = \frac{(B_0 + B_1)}{2} + \frac{(2 \times G_k - GB_0 - GB_1)}{2} \quad (10)$$



(a) (b)

Fig. 6. Using green color at the center point.

In another scenario, the pixels above and below G_k are red, and the ones to the left and right are blue. This case employs a reversed calculation method. The formula (11) can be utilized to reconstruct the missing red and blue pixels at the center point.

Furthermore, the last step is to restore the missing red or blue pixels in the B or R channel. In this step, the green channel pixels have already been restored, leaving only the red and blue channels with missing pixels, as shown in Fig. 7 (a). The transparent pixels represent points where all three channels have no missing values. The remaining texture directions for pixel points are at 45 degrees and 135 degrees, as mentioned in HA regarding the gradient calculation formula (11) for these directions:

$$\begin{cases} N = |\mathrm{R}_0 - \mathrm{R}_1| + |2 \times \mathrm{G}_k - RG_0 - RG_1| \\ P = |\mathrm{R}_2 - \mathrm{R}_3| + |2 \times \mathrm{G}_k - RG_2 - RG_3| \end{cases} \quad (11)$$

N represents the gradient in the 135-degree direction, and P represents the gradient in the 45-degree direction. RG_0 refers to the green pixel of R_0, and G_k represents the green pixel of the center point B_k. In this case, this study uses blue as the center point and calculates the missing red pixel based on the comparison of the gradient values. R_k represents the missing green pixel of the center point B_k, and the calculation formula for R_k is shown in (12):

$$\begin{cases} N > P & \mathrm{R}_k = \frac{(\mathrm{R}_2 - \mathrm{R}_3)}{2} + \frac{(2 \times \mathrm{G}_k - RG_0 - RG_1)}{2} \\ N < P & \mathrm{R}_k = \frac{(\mathrm{R}_2 - \mathrm{R}_3)}{2} + \frac{(2 \times \mathrm{G}_k - RG_2 - RG_3)}{2} \end{cases} \quad (12)$$

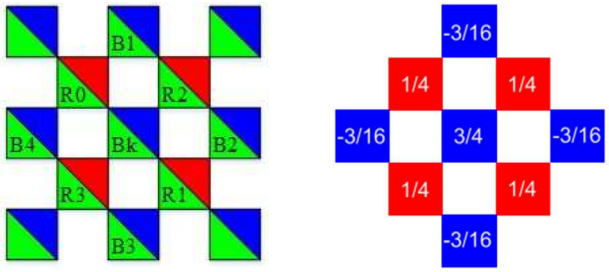
Instead of calculating specific directions for 135 degrees and 45 degrees, this study replaces them with the calculation parameters from the HQLI, as shown in Fig.7 (b). The HQLI incorporates an estimated luminance change to correct the values obtained from Bilinear Interpolation [2]. The correction formula is as follows (13):

$$R_K = \widehat{R_k} + \Delta\gamma \times \Delta B_k \quad (13)$$

Here, $\widehat{R_k}$ represents the estimated values obtained using BI, and HQLI suggests that $\Delta\gamma$ = 3/4 yields optimal results. ΔB_k denotes the gradient at that position for B_k. After obtaining the strength parameter for the correction formula, the calculation formula is organized as follows (14) to compute R_k:

$$\mathrm{R}_k = \frac{3}{4} \times \mathrm{B}_k + \frac{1}{4}(\mathrm{R}_0 + \mathrm{R}_1 + \mathrm{R}_2 + \mathrm{R}_3) - \frac{3}{16}(\mathrm{B}_0 + \mathrm{B}_1 + \mathrm{B}_2 + \mathrm{B}_3) \quad (14)$$

Similarly, for the remaining pixel points shown in Fig. 7. (a), the blue can be treated as the center point, and calculate the missing blue pixel values by swapping the red and blue and performing the respective calculations.



(a) (b)

Fig. 7. Using blue color at the center point.

D. *Image Scaling*

Common image enlargement techniques include bilinear interpolation [9] and bicubic interpolation (BICI) [6]. While bilinear interpolation [9] is simple to implement, it results in variations in image details. On the other hand, bicubic interpolation, although more complex in terms of algorithm, avoids the generation of jagged edges and is suitable for

handling intricate patterns. BICI calculates the pixel value of the target point based on the weighted average of the surrounding 16 pixels. The weight calculation formula is shown in (15):

$$W(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^3 + 1 & for\ |x| \le 1 \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a & for\ 1 \le |x| \le 2 \\ 0 & other \end{cases} \quad (15)$$

x represents the pixel distance between the target point and a neighboring pixel, and adjusts the smoothness of the function, with the optimal value being -0.5. After calculating the weight values for the 16 neighboring points of the target point, the original pixel values are multiplied by the corresponding weights and summed up as follows:

$$P(x,y) = \sum_{i=1}^{4}\sum_{j=4}^{4} a_{ij} x^i y^j \quad (16)$$

P represents the target pixel value, i denotes the row number, j denotes the column number, a_{ij} represents the pixel value of the neighboring point, x^i represents the weight in the X-axis direction, and y^j represents the weight in the Y-axis direction. This process allows for image enlargement that is clearer compared to the original bilinear interpolation [9].

III. RESULTS

This study conducted tests using the Kodak dataset [7] with dimensions of 768×512, which is shown in Table 1. After simulating the Bayer CFA pattern and applying it to the down-sample, and employed an improved HA demosaicking algorithm to restore colors. Finally, this study utilized BICI. On average, this method achieved a 1.13 dB difference in PSNR compared to HQLI combined with BICI, as shown in Table 2. The noticeable differences can be observed in Fig. 8. and Fig. 9., particularly in the fenced area of Fig. 9. In addition to the proposed method for quantitative comparison, this study also attempted other approaches. The first approach involved unimproved HQLI demosaicing with bilinear interpolation and no image pre-processing. The second approach utilized unimproved HQLI demosaicing with bilinear interpolation and image pre-processing. The third approach employed unimproved HA demosaicing with bicubic interpolation and no image pre-processing. The fourth approach applied unimproved HA demosaicing with bicubic interpolation and image pre-processing. The fifth approach employed an unimproved proposed method with bicubic interpolation and no image pre-processing. The sixth approach applied the unimproved proposed method with bicubic interpolation and image pre-processing.

Table 1. Kodak Lossless True Color Image Suite [7].





Table 2 Comparisons of PSNR, CPSNR, and SSIM results.

	PSNR	CPSNR	SSIM
HQLI[3] (Method A)	26.5182	26.4511	0.88555
HQLI[3]+MF[8] (Method B)	27.2546	27.2430	0.91098
HA[4] (Method C)	26.7441	26.7255	0.8856
HA[4]+MF[8] (Method D)	27.5759	27.5659	0.91509
Proposed method (Method E)	26.8447	26.8258	0.88801
Proposed method+MF[8] (Method F)	27.6530	27.6427	0.91638
Kuan-Yu Huang[5]	29.19	—	—

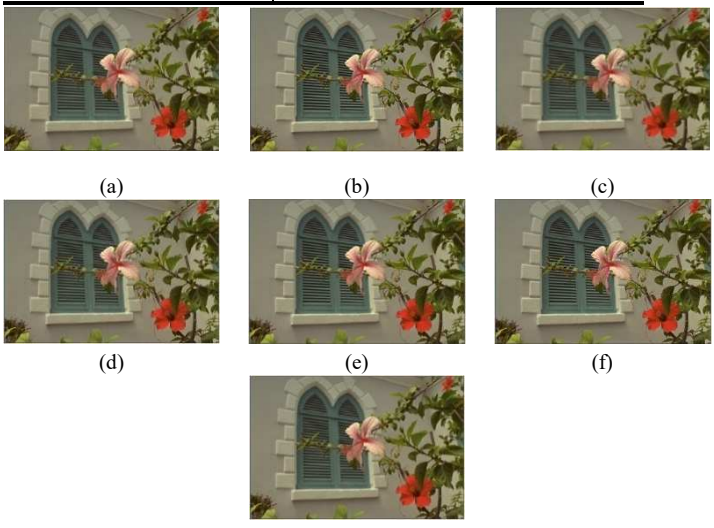


Fig. 8. (a) Testing image Kodak7 (b) Method A [3](c) Method B [3][8] (d) Method C [4] (e) Method D [4][8] (f) Method E (g) Method F[8].

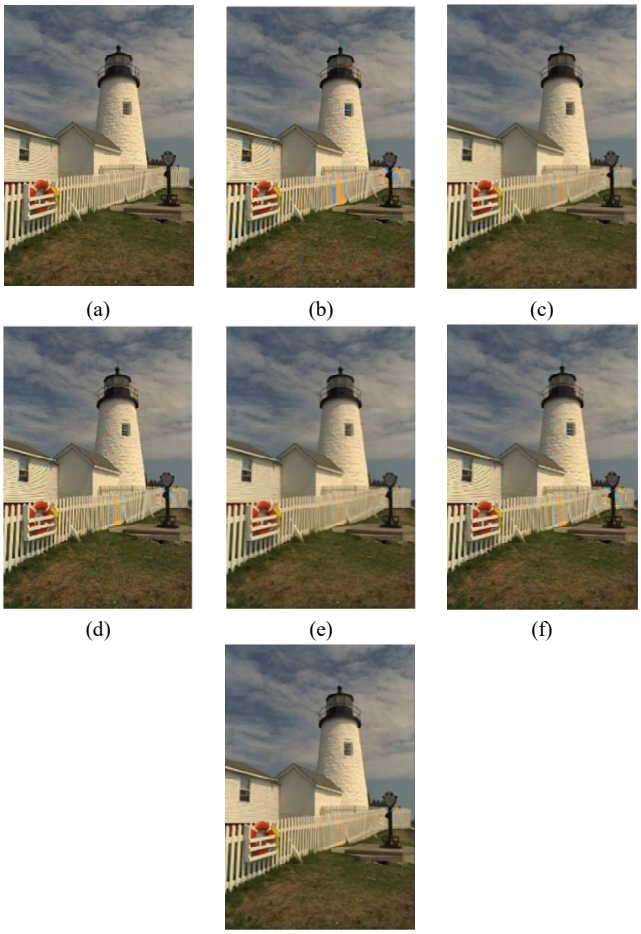


Fig. 9. (a)Testing image Kodak19 (b) Method A [3] (c) Method B [3][8] (d) Method C [4] (e) Method D [4][8] (f) Method E (g) Method F[8].

IV. CONCLUSIONS

This paper introduces a comprehensive image processing pipeline that integrates mean filtering [8], an improved adaptive mosaic algorithm, and bicubic interpolation for image upscaling. This procedure offers a simple yet effective approach to enhance image quality and improve visual effects. Initially, mean filtering is applied to smooth the image and reduce the zipper effect. Subsequently, the weights of parameters in the HA demosaicing [4] algorithm are adjusted using High-quality linear interpolation [3] to enhance image sharpness and color details. Finally, bicubic interpolation [6]is used for scaling, resulting in improved image quality and visual effects. This paper demonstrates excellent performance in terms of evaluation metrics while effectively addressing the zipper effect.

Regarding image quality assessment, this paper achieves a 1.13 dB increase in PSNR, 1.19 dB in CPSNR, and 0.03 in SSIM values compared to HQLI [3] combined with bicubic interpolation [6]. Although it slightly lags behind the approach proposed by Kuan-Yu Huang et al. [5] based on machine learning, the method presented in this study has greater potential for hardware implementation and can be applied to various low-complexity image processing tasks, such as image upscaling, denoising, and image enhancement.

REFERENCES

[1] B. E. Bayer, Color imaging array, U.S. Patent No. 3 971 065, to Eastman Kodak Company (Rochester, NY), 1976.

[2] P. Longère, X. Zhang, P. B. Delahunt, and D. H. Brainard, “Perceptual assessment of demosaicing algorithm performance,” Proc. IEEE, vol. 90, pp. 123–132, Jan. 2002.

[3] H. S. Malvar, Li-wei He and R. Cutler, "High-quality linear interpolation for demosaicing of Bayer-patterned color images," *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, 2004, pp. iii-485, doi: 10.1109/ICASSP.2004.1326587.

[4] J.F. Hamilton Jr. and J.E. Adams, Adaptive color plane interpolation in single color electronic camera, U.S. Patent 5 629 734, May 1997.

[5] K. -Y. Huang, P. -Y. Chen and C. -C. Huang, "A low-complexity learning-based algorithm for joint application of demosaicking and up-scaling," 2020 International Computer Symposium (ICS), Tainan, Taiwan, 2020, pp. 370-375, doi: 10.1109/ICS51289.2020.00080.

[6] R. Keys, "Cubic convolution interpolation for digital image processing," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153-1160, December 1981, doi:10.1109/TASSP.1981.1163711.

[7] The Kodak Color Image Dataset, Image Available [Online]. Available: http://www.site.uottawa.ca/~edubois/demosaicking/.

[8] Li Tan, *Digital Signal Processing: Fundamentals and Applications: Elsevier Science*. 2008.

[9] J. A. Parker, R. V. Kenyon and D. E. Troxel, "Comparison of Interpolating Methods for Image Resampling," in IEEE Transactions on Medical Imaging, vol. 2, no. 1, pp. 31-39, March 1983, doi: 10.1109/TMI.1983.4307610.

[10] K. Chung and Y. Chan, "A low-complexity joint color demosaicking and zooming algorithm for digital camera," IEEE Transactions on Image Processing, vol. 16, no. 7, pp. 1705-1715, Jul. 2007.

[11] K. Chung, W. Yang, P. Chen, W. Yan and C. Fuh, "New joint demosaicing and zooming algorithm for color filter array," IEEE Transactions on Consumer Electronics., vol. 55, no. 3, pp. 1477-1486, Aug. 2009.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Improving Regularization of Deep Learning Models in Fundus Analysis

Wei-Wen Hsu[1], Yao-Chung Chang[1], Wei-Min Lee[1], Yu-Chuan Huang[2], and Da-Wen Lu[*,3]

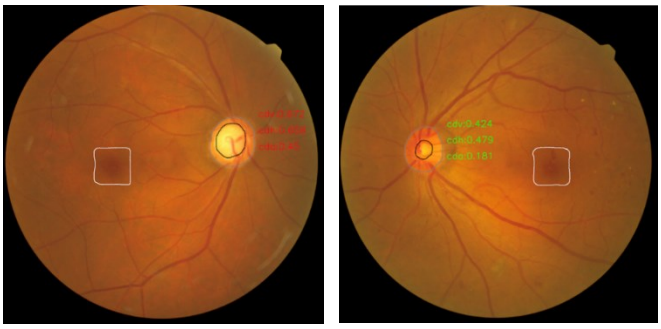
[1] Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan
[2] School of Pharmacy, National Defense Medical Center, Taipei, Taiwan
[3] Department of Ophthalmology, Tri-Service General Hospital and National Defense Medical Center, Taipei, Taiwan
[*]E-mail: p310849@ms23.hinet.net

***Abstract*—With the advancement of deep learning in many computer vision applications, the development of computer-aided diagnosis systems has been focused on recently. The deep learning approaches for glaucoma assessment through fundus analysis have shown extremely high performance, especially for early detection, in many studies. In addition, researchers also found that the deep learning models detect not only the subtle changes that reflect RNFL thinning but also the morphological features that are present outside the optic disc in fundus photography. It suggests that the deep learning models have great potential for extracting detailed morphologies which may not be aware by human's vision. However, the deep learning models also suffered from the problem of poor regularization when dealing with cross-institutional datasets. That is, the model's performance dropped dramatically on the input images from different institutions. The feature extraction by deep learning models can detect detailed characteristics for distinguishing different categories; nonetheless, these detailed features also cause the overfitting problem for cross-institutional data. Therefore, in this study, we aim to verify our proposed hypothesis by the application of fundus analysis for glaucoma assessment using deep learning approaches. The features with better regularization, called robust features, were selected to reduce the influences from the non-robust features that overfit the training dataset from a specific source. With feature selection, the performance of the glaucoma assessment cross-institutional dataset can be improved from 62.7% to 81% in accuracy.**

Keywords— Glaucoma Assessment, Deep Learning Approach, Model Regularization.

I. INTRODUCTION

Glaucoma is considered one of the leading causes of irreversible blindness worldwide. It is known as the silent thief of sight since it causes vision loss or even blindness due to the progressive damage of optic nerve. As it is asymptomatic, early screening of glaucoma and the corresponding treatment are very important for patients to prevent the progression of vision loss.



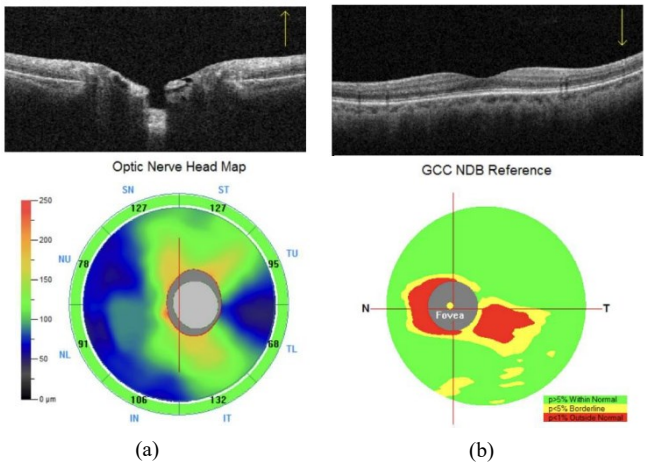
(a) (b)

Fig. 1. The model of Mask R-CNN was applied to perform instance segmentation for optic cup (the enclosed area in black), optic disk (the enclosed area in light blue), and fovea (bounded area in white) in macula. The measurements of cup-disk ratio with respect to vertical height, horizontal width, and areas were computed, denoted as cdv, cdh, cda, respectively. (a) A case of glaucoma with enlarged cup-to-disk ratios. (b) A case of non-glaucoma with normal cup-to-disk ratios.

Clinically, color fundus photography, as shown in Figure 1, is one of the mandatory imaging modalities in glaucoma assessment, characterized by its convenience and relative affordability. Pathologically, glaucoma leads to the elevation of patient's intraocular pressure, which compresses the optic nerve and causes the depression of the optic disc. Such nerve fiber damage results in an enlarged cup-to-disc ratio (CDR) as it can be observed from Fig. 1-(a). In fundus analysis, the measurement of CDR is one of the factors for glaucoma assessment; however, not every glaucomatous case follows such rule. Besides, the borderline measurements of CDR may also confuse ophthalmologists in diagnosis.

On the other hand, as it is shown in Figure 2, the non-invasive imaging modality of optical coherence tomography (OCT) with high resolution 3D reconstruction of retina can provide more detailed information to compute the thickness of the retinal nerve fiber layer (RNFL) and the ganglion cell complex (GCC), which are considered the useful biomarkers in diagnosing glaucoma clinically.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



(a) (b)

Fig. 2. The (a) thickness distribution maps of RNFL and (b) reference maps for GCC can be derived from OCT with high resolution 3D reconstruction of retina.

Recently, the task performance in image classification has significantly advanced as a result of the development of deep learning techniques. For fundus analysis, the deep learning models of convolutional neural network (CNN) have been widely adopted in fundus image analysis for the diseases of diabetic retinopathy (DR) [1], age-related macular degeneration (AMD) [2], and glaucomatous optic neuropathy (GON) [3-6]. These studies have shown the strength of the DCNN models to provide comparably accurate diagnostic predictions to human experts simply with the single modality of fundus images. In addition, researchers found that the deep learning models can even predict an individual's age, gender, and cardiovascular risk factors by assessing a fundus photo [7], which is extremely unlikely to be achieved by human's vision. Taking glaucoma assessment as an example, deep learning models detect not only the subtle changes that reflect RNFL thinning but also the morphological features that are present outside the optic disc in fundus photos [3]. Therefore, the main purpose of this study is to reveal the insights of the decision process by the deep learning models of DCNN in the diagnosis of glaucoma using a single modality of color fundus images.

However, researchers also found that the tasks of glaucoma assessment with the deep learning approaches suffer from the problem of poor model generalization for the cross-institutional imaging data [4-6]. As a result, even though the deep learning models have shown the extraordinary performance on a specific dataset in fundus analysis, the trained models may not be applicable on the fundus images which came from the other medical institutions due to its poor generalization. From the perspective of development, it becomes the top issue to be tackled when deploying the computer-aided diagnosis (CAD) systems.

In this study, the properties of feature extraction in deep learning approaches were mainly discussed through the tasks of fundus analysis for glaucoma assessment. We argue that the deep learning models are good at finding the detailed features which are statistically useful in classification but hard to be noticed by our vision. This kind of features is classified to the category of "non-robust features", which may be the key to success for the deep learning approaches to achieve outperformance than human experts. However, such non-robust features also overfit the dataset easily and result in the problem of poor generalization when dealing with the cross-institutional datasets. For our hypothesis, the features extracted by deep convolutional neural networks (DCNN) can be categorized into "robust features" and "non-robust features", where the "robust features" show better regularization while the "non-robust features" overfit the dataset easily for detailed characteristics [8]. Therefore, several experiments were designed to verify the abovementioned hypothesis in this study.

II. Materials and Methods

From the perspective of the development of AI-assisted systems, the AI models are supposed to maintain decent level of generalization to deal with the imaging data from different institutions or devices. Our goal is to improve the AI models' regularization property for cross-institutional data in the tasks of glaucoma assessment. In this study, we aimed to verify the poor regularization of DCNN models came from the influences of the overfitting features. Accordingly, feature selection for the robust ones can mitigate the overfitting problem successfully.

We argue that the performance of the DCNN models dropped dramatically on the cross-institutional data due to the influence of the overfitting features learned from the source dataset. Therefore, the robust features with better generalization were chosen after feature selection, mitigating the overfitting problem. To verify our hypothesis, two datasets of fundus images from different institutions were collected. The main dataset for models' training came from the Tris-Service General Hospital (TSGH). And another dataset came from the Taipei Veterans General Hospital (TPEVGH) for the purpose of evaluating the model's performance on the cross-institutional dataset. The fundus images in the main dataset were collected retrospectively from the ophthalmology clinic at TSGH. And the numbers of glaucomatous samples and non-glaucomatous samples for each dataset are listed in Table I. Another dataset from TPEVGH was randomly divided in half to become a validation set and a testing set, as listed in Table II.

Table I. Datasets of fundus images from the Tri-Service General Hospital (TSGH).

dataset	label	# fundus photos
training	Glaucoma	536
	Non-glaucoma	639
validation	Glaucoma	146
	Non-glaucoma	189
testing	Glaucoma	75
	Non-glaucoma	153

Table II. Datasets of fundus images from the Taipei Veterans General Hospital (TPEVGH).

dataset	label	# fundus photos
validation	Glaucoma	150
	Non-glaucoma	150
testing	Glaucoma	150
	Non-glaucoma	150

In the experiment, the DCNN model of ResNet50 [9] was chosen to train on the training dataset from TSGH. And the final model was determined by choosing the model's weights that reached the best performance on the validation dataset of TSGH. Then, the final model was applied on the testing sets of TSGH and TPEVGH for inference to evaluate its performance of generalization on the cross-institutional dataset.

For feature selection, the end-to-end DCNN model was decomposed into two parts: (1) convolutional layers for feature extraction and (2) fully connected layers for classification. Then, the convolutional layers and their weights were kept to perform feature extraction, and the fully connected layers were replaced by a random forest classifier for classification. For ResNet50, 2048 features were extracted from an input image through the convolutional layers. Here the classifier of random forest is to encode the relation of these 2048 features and the ground truth outputs and also estimate the feature importance. Subsequently, the validation set from TPEVGH was used to train a random forest classifier to rank all 2048 features by the feature importance. As a result, we know what features learned from the dataset of TSGH are also useful for the dataset of TPEVGH in distinguishing the glaucomatous cases, called the robust features. Finally, only some robust features were selected to train a random forest classifier on the TSGH dataset, screening out those overfitting features. And we need to check if the random forest classifier trained with the selected features can improve its performance on the cross-institutional datasets.

III. Experimental Results

From the previous two sections, the DCNN models have shown the outstanding performance in classification, capturing the detailed morphologies that surpass humans' perception. However, those detailed features that overfit the dataset may lead to poor generalization of model. In Table III, the end-to-end DCNN model trained on the TSGH dataset achieved 88.2% in accuracy on the testing dataset from the same institution, but the performance dropped to 63.3% on the cross-institutional testing set. The similar phenomenon also happened with the classifier of random forest. It means the extracted features are the bottleneck of performance. Therefore, the validation set from another institution (TPEVGH) was used to rank the feature importance for selecting the generalized features. After feature selection, even though the performance dropped about 2% on the same source dataset, it increased 18% on the cross-institutional dataset. As a result, better generalization of model can be achieved by screening out the overfitting features. Figure 3 shows the performance of the model on the cross-institutional dataset increased with more robust features selected, and it started to drop when more and more overfitting features were included.

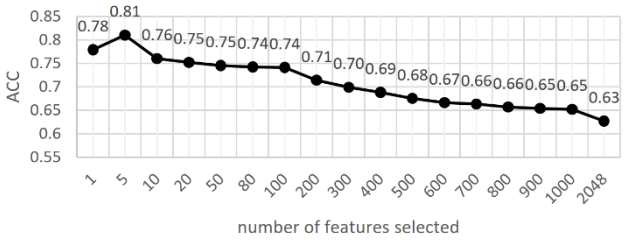


Fig. 3. The accuracy of the model on the cross-institutional dataset with respect to the number of features used in classification.

Table III. Performance comparison among different methods on the cross-institutional dataset.

Method	Data from TSGH	Data from TPEVGH
DCNN trained on TSGH data	0.882 (201/228)	0.663 (190/300)
CNL + RF (all 2048 features)	0.886 (202/228)	0.627 (188/300)
CNL + RF (5 selected features)	0.864 (197/228)	0.81 (243/300)

IV. CONCLUSIONS

In this study, we have shown the DCNN models are specialized in capturing the detailed morphologies, even those features are barely detected by our vision. On the other hand, the extracted features with detailed morphologies became the bottleneck of models' performance for the cross-institutional data, leading to poor generalization. After feature selection, the performance on the same source data may be sacrificed to trade better generalization from the perspective of CAD systems' development.

ACKNOWLEDGEMENTS

We would like to express our appreciation to the National Center for High-performance team for sharing TPEVGH fundus database in the project of MIAI-2022-M10-01.

REFERENCES

[1] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama,* vol. 316, no. 22, pp. 2402-2410, 2016.

[2] P. M. Burlina, N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong, and N. M. Bressler, "Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration," *JAMA ophthalmology,* vol. 136, no. 12, pp. 1359-1366, 2018.

[3] R. Hemelings, B. Elen, J. Barbosa-Breda, M. B. Blaschko, P. De Boever, and I. Stalmans, "Deep learning on fundus images detects glaucoma beyond the optic disc," *Scientific Reports,* vol. 11, no. 1, p. 20313, 2021.

[4] Y.-C. Ko *et al.*, "Deep learning assisted detection of glaucomatous optic neuropathy and potential designs for a generalizable model," *PLoS One,* vol. 15, no. 5, p. e0233079, 2020.

[5] A. Almazroa, R. Burman, K. Raahemifar, and V. Lakshminarayanan, "Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey," *Journal of ophthalmology,* vol. 2015, 2015.

[6] Y.-C. Ko *et al.*, "Widen the Applicability of a Convolutional Neural-Network-Assisted Glaucoma Detection Algorithm of Limited Training Images across Different Datasets," *Biomedicines,* vol. 10, no. 6, p. 1314, 2022.

[7] R. Poplin *et al.*, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature biomedical engineering,* vol. 2, no. 3, pp. 158-164, 2018.

[8] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems,* vol. 32, 2019.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

Design of Interactive System for Acupoint Analysis Based on Augmented Reality

Chung-Yen Wei, Bo-Yuan Xu, and Yu-Xiang Zhao*
National Quemoy University, Taiwan.
*E-mail: yxzhao@nqu.edu.tw

Abstract — **In this paper, we utilize the edge computing of Jetson Nano J1010 development board and employ MediaPipe technology and acupoint mapping algorithm to display acupoints on the screen. By using a custom deep neural network (DNN) hand recognition model, we achieve augmented reality (AR) interaction with acupoints. Users can view the corresponding acupoint locations on the body and acupoint descriptions through gestures. Additionally, we develop a mobile app that enables interaction with the system via Bluetooth technology. Users can select symptoms they wish to understand, and the system will display acupoints that can alleviate those symptoms. Our team also develops an AI-based diagnostic system using the open-source conversational AI Rasa, which utilizes natural language understanding (NLU) to determine the user's intents from input on the mobile app. When the intent is related to medical content, the information is passed to a BERT model for diagnosis analysis, determining possible diseases or symptoms and providing corresponding recommended acupoints for symptom relief. This aids users in making preliminary judgments about their own conditions. All AI diagnostic information is transmitted to the server-side, where physicians can view user information through a website. By visualizing the data through charts, physicians can quickly understand the user's condition and further save medical resources. This acupoint-based healthcare system offers a safe, convenient, and efficient solution for health management, particularly valuable for individuals lacking experience and medical knowledge.**

Keywords: edge computing, acupoints, AR, AI, NLU, BERT.

I. INTRODUCTION

One of the methods for maintaining good health is acupressure. There are many specific acupoints that have the ability to treat various ailments. These acupoints not only have the effect of healing specific diseases but also help in regulating overall body functions, promoting physical fitness. However, as acupoints are distributed throughout the human body, their precise locations often require guidance from books or others, especially for beginners. We aim to enable users to easily identify the locations and functions of acupoints on their hands and heads at a glance, allowing them to effectively utilize these acupoints to alleviate symptoms in a timely manner, ultimately preventing the onset of diseases.

This system utilizes the edge computing of Jetson Nano J1010 development board and incorporates MediaPipe technology to achieve augmented reality (AR) interaction with acupoints, allowing users to view the corresponding acupoint locations on the body. We have also developed a mobile app that connects to a smart mirror via Bluetooth. Users can select symptoms they wish to understand, and the screen will provide feedback on acupoints that can alleviate those symptoms. The system also includes AI diagnosis using Rasa and BERT technology, which analyzes user input to determine possible diseases or symptoms and provides recommended acupoints for further symptom relief.

This acupoint-based healthcare system offers a safe, convenient, and efficient solution for health management, particularly valuable for individuals lacking experience and medical knowledge. The innovation of this project lies in the integration of the mobile app, AI, and AR technology, which realizes a acupoint-based healthcare solution that is more convenient and efficient compared to traditional healthcare approaches.

II. RELATED WORK

Acupoints are essential elements in traditional Chinese healthcare, and in recent years, artificial intelligence (AI) technology has made significant progress in acupoint recognition. Early approaches utilized CNN models and mathematical calculations or edge detection to infer acupoint locations [1]-[2], while now models such as MediaPipe can obtain human body coordinate points. Previous studies have used 3DMM models to display facial acupoints [3], but their complex computational requirements limit real-time operation. MediaPipe Face Mesh is a real-time facial landmark detection

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

system that uses machine learning to predict the 3D positions of 468 facial keypoints. Compared to 3DMM models, MediaPipe performs better and is suitable for running on various devices.

Recent research has also used MediaPipe for facial acupoint display [4], but with limited functionality. This paper further develops on the foundation of MediaPipe, implementing not only facial acupoint display but also hand acupoints, acupoint classification, and gesture interaction, providing users with a more comprehensive experience. In recent years, artificial intelligence has made significant progress in various fields such as natural language processing, image recognition, visual retrieval, and medical image analysis [5]-[6]. Especially in the past few years, with the rapid improvement of computer specifications and the growth of talents in various domains, the power of artificial intelligence has been strengthening year by year. The following sections will introduce the relevant technologies used in this research.

A. MediaPipe

MediaPipe is an open-source, cross-platform multimedia processing framework developed by Google [7]. It provides a set of templates for image and video processing, enabling the construction of complex multimedia applications. MediaPipe supports multiple platforms, including Android, iOS, Linux, and Windows. It offers various built-in algorithms such as face detection, holistic detection, and hand tracking, while also allowing easy extension with custom algorithms. MediaPipe can be used in various environments, such as mobile and embedded devices. Within the scope of this project, MediaPipe serves as the means to collect body pose data, resulting in the implementation of two key features: the acupoint mapping and the gesture recognition facilitated by a custom DNN calculator [8].

B. PyQT

PyQt is a framework for developing graphical user interface (GUI) applications [9]. It is based on the Qt framework, which is a cross-platform application development framework used for desktop, mobile, and embedded systems. PyQt can run on multiple operating systems, including Windows, Mac OS X, Linux, and embedded Linux, making program development on Jetson Nano much more convenient. Due to its foundation on the Qt framework, PyQt offers excellent performance. It provides efficient drawing and event handling through optimization and design, enabling the handling of large amounts of data and complex operations. Additionally, PyQt offers rich configuration and customization options, allowing the creation of unique user interfaces.

C. Jetson Nano

The Jetson Nano J1010 development board is a high-performance single-board computer introduced by NVIDIA, designed specifically for machine learning and artificial intelligence applications. Jetson Nano J1010 features NVIDIA's quad-core ARM Cortex-A57 processor and a 128-core NVIDIA Maxwell GPU, providing powerful computational capabilities. This allows it to excel in handling complex machine learning and visual computing tasks, capable of real-time processing of high-resolution images and videos. The NVIDIA GPU on Jetson Nano J1010 provides additional computing resources for accelerating machine learning and deep learning tasks. This enables the execution of complex neural network models on Jetson Nano and facilitates real-time face detection and hand recognition tasks as proposed in this paper.

J1010 is also an optimized hardware platform for AI and machine learning applications. It supports popular deep learning frameworks such as TensorFlow, PyTorch, and Caffe, as well as GPU acceleration libraries like CUDA and cuDNN, enabling efficient utilization of custom models.

D. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model for natural language processing (NLP) developed by Google [10], extending from the Transformers [11] decoder. BERT is capable of bidirectional encoding of sentences, considering the contextual information of all words in the sentence. This allows BERT to exhibit excellent performance in various NLP tasks, including sentiment analysis, machine translation, keyword extraction, and question-answering systems.

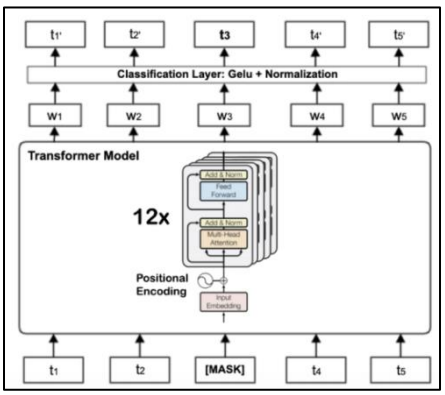


Fig. 1 BERT basic architecture diagram.

The architecture of BERT used in this project consists of 12 transformer layers (see Figure 1), which is a neural network architecture that effectively learns long-range dependencies. BERT employs a technique called adaptive learning during

training, enabling the model to dynamically adapt to different input lengths of sentences. Additionally, BERT utilizes residual connections and self-attention mechanisms [11], allowing the model to capture the relevance within sentences better. Through the BERT model, we aim to infer user diseases based on conversations and provide corresponding acupoints displayed on the human body.

E. MVC and State Machine

This paper adopts the Model-View-Controller (MVC) architecture on the software side to organize and separate different components of the application, aiming to achieve better code structure and maintainability. The Model represents the data structure, logic, and state of the application. It is responsible for handling operations, validation, persistence, and other data-related tasks. The Model does not directly interact with the user interface (View) or user interactions (Controller) but provides interfaces for them to interact.

The View is responsible for presenting the interface, allowing users to intuitively view and manipulate data. It typically represents user interface elements such as buttons, forms, charts, etc., and updates accordingly based on the state of the Model. The Controller serves as an intermediary between the Model and the View. It receives user inputs, processes user requests, and then passes the results to the Model or View as needed. It determines how to update the Model and View based on the logic of the application. In addition to the MVC architecture, this paper also utilizes a state machine to manage the different states and transitions within the application. The state machine helps define the behavior and flow of the application by representing various states and specifying the conditions and actions associated with state transitions. By adopting the MVC architecture and employing a state machine, this paper achieves a well-structured and maintainable software solution for the intended application.

III. APPROACH

A. System Architecture

In this paper, an interactive AR acupoint healthcare system is developed, which can recognize acupoints on the user's hands and head and assist users in learning and understanding acupoint knowledge. As shown in Figure 2. The system consists of four main components: the software side, the hardware side, the server side, and the mobile side. The hardware side utilizes the edge computing of Jetson Nano J1010 development board and interacts with the user through a webcam. The software side is developed using PyQt and utilizes MediaPipe to obtain the coordinates of the human body. With the help of algorithms, the obtained coordinate points from MediaPipe are mapped to the acupoints on the head and hands. This allows the display screen to show the acupoints, enabling users to learn and understand acupoint knowledge.

Additionally, the system includes an Android app that users can download. Through this app, users can conduct AI consultations and record their physical condition and related information. During the consultation, the app connects to the server side to perform AI consultations. The server side utilizes natural language understanding (NLU) to interpret user inputs and determine their physical condition. The user's information is stored on the server. Doctors can log in to the server through a web interface to quickly understand the user's physical condition and related information.

This design enables doctors to gain a faster understanding of the user's physical condition, improving diagnostic efficiency, and allows users to conveniently access medical assistance. Overall, the developed system provides an interactive AR acupoint healthcare solution, combining hardware, software, server, and mobile components to enhance acupoint learning, AI consultations, and efficient healthcare delivery.

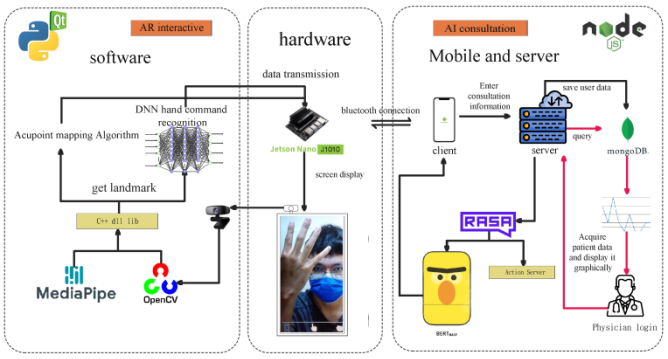


Fig. 2 System Architecture Diagram.

B. Produce Architecture

The product architecture is visually represented in Figure 3. Initially, we utilize a webcam to capture the user's hand data, which is subsequently transmitted to the Jetson Nano J1010 for processing.

In the Data Analysis phase, we leverage the hand and head landmarks provided by the Mediapipe holistic model to enable two essential features: acupoint mapping and gesture recognition. Acupoint mapping involves projecting these acupoints onto the human body, while Deep Neural Networks

(DNNs) are employed for hand gesture recognition, providing users with an intuitive means of controlling the entire system. This allows users to view the desired acupoint locations or issue commands effortlessly.

Additionally, we take measures to optimize this resource-intensive computation, resulting in reduced power consumption and lighter system load requirements. We empower users to select from our customized facial recognition models, and we implement asynchronous output for data analysis, ensuring a smoother visual interaction with their hand acupoints in augmented reality. Finally, in the front-end interaction, users can engage with and control the system with gesture interaction and mobile app.

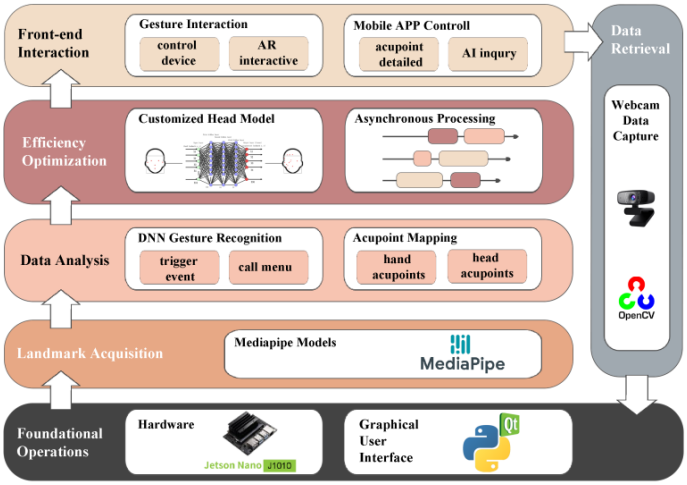


Fig. 3 Product Architecture Diagram.

C. *Acupoint Mapping*

Our hand and head localization utilize position data from a database, which includes reference MediaPipe coordinate positions, relative distances to the coordinate positions, and acupoint names. The steps for displaying acupoints can be seen in Figure 4.

We employ a custom hand acupoint mapping algorithm that utilizes various mathematical functions such as cross products, dot products. We employ the intuitive definition of the dot product in Euclidean space, which involves taking the dot product of two vectors, dividing it by the distance between the two vectors, and then applying the inverse cosine function to obtain Phi (Equation (1)), and generates transformation matrices based on the rotation amount of the base axis (Equation (2)). By comparing the relative positioning points from the database with the hand localization points provided by MediaPipe, the acupoints can be accurately displayed on the screen and scaled based on distance. The development of this custom algorithm ensures reliable acupoint mapping.

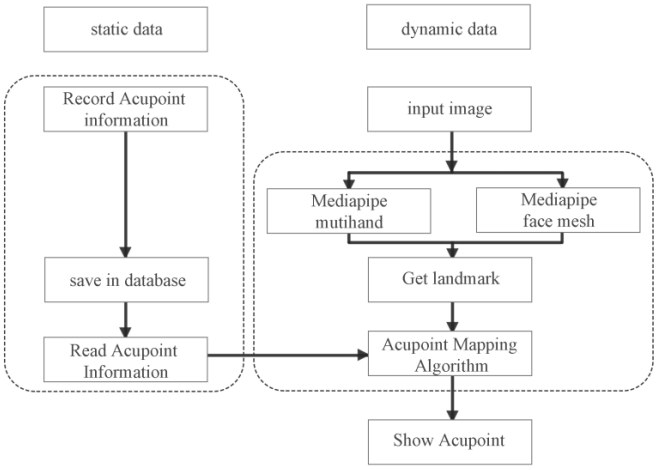


Fig. 4 Product Architecture Diagram.

$$\varphi = \cos^{-1}\frac{\overrightarrow{v_{0to9}} \cdot \overrightarrow{nv_{0to9}}}{\sqrt{(x_0-x_9)^2+(y_0-y_9)^2}\times\sqrt{(nx_0-nx_9)^2+(ny_0-ny_9)^2}} \quad (1)$$

$$where \quad \overrightarrow{v_{0to9}} = (x_0 - x_9, y_0 - y_9),$$
$$\overrightarrow{nv_{0to9}} = (nx_0 - nx_9, ny_0 - ny_9)$$

$$[\alpha, \beta] = [x,\ y] \cdot \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \times z \quad (2)$$

D. *DNN Hand Gesture Recognition*

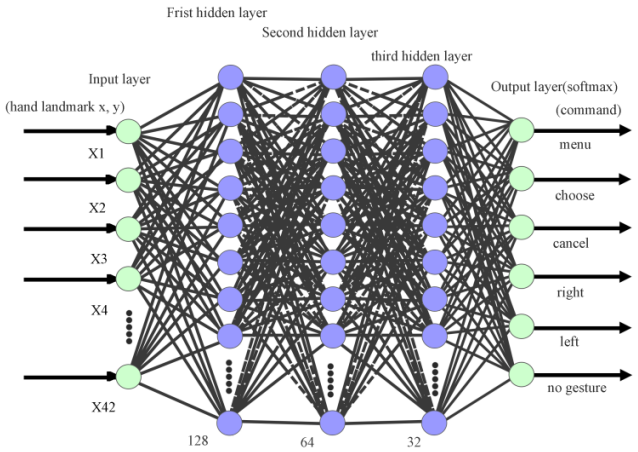


Fig. 5 DNN hand recognition model

HaGRID (Hand Gesture Recognition Image Dataset) [12] is a dataset for hand gesture recognition images. In this project, we utilize the data from HaGRID as well as custom hand data to train a DNN (Deep Neural Network) model. The dataset consists of a total of 28,304 samples. Figure 5 illustrates the DNN network. The input for this model consists

of 42 features, including the X and Y coordinates of 21 hand localization points returned by MediaPipe. Since the input features are coordinate positions, a simple three-layer DNN model is sufficient to quickly classify the current gesture into its corresponding type. The output consists of four classes: menu, choose, cancel, right, left, and no gesture. This hand gesture recognition functionality enables users to interact more conveniently with the smart mirror, enhancing the overall user experience of the product.

E. State Machine

Figure 6 illustrates the main operational logic of the state machine in the software code of this project. The purpose of this state machine is to perform a series of operations and decisions based on the landmark data returned by the hand tracking model of MediaPipe. A more detailed explanation of this operational logic is provided below.

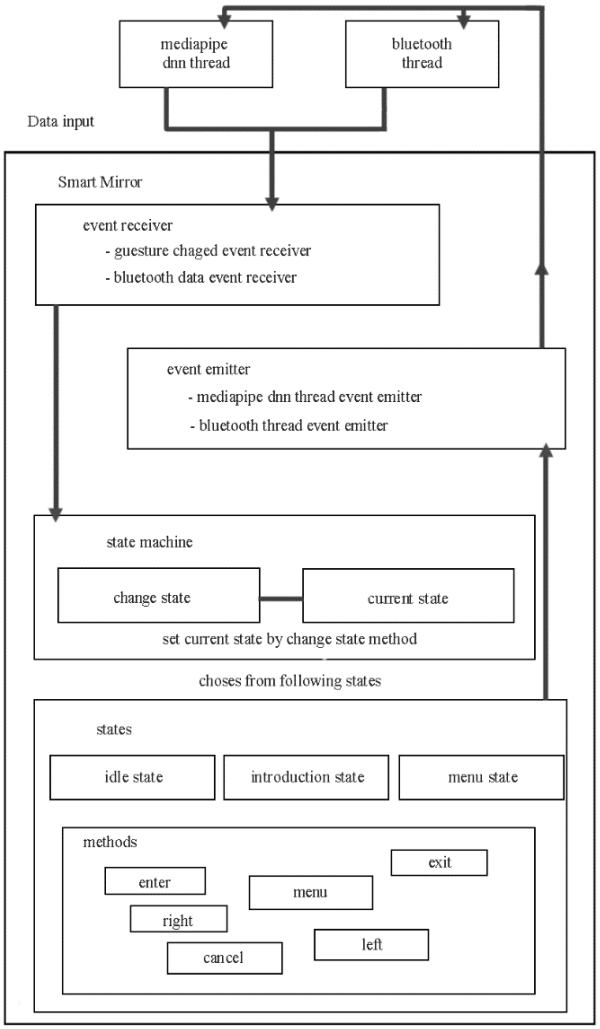


Fig. 6 Finite state machine architecture

Firstly, the system retrieves the landmark data as input from the hand tracking model of MediaPipe. Next, these input data are passed to a custom DNN model for processing. The output of the DNN model is transmitted to the main thread (smart mirror - main thread) through event functions. This design helps prevent race conditions and critical sections between different threads, ensuring the proper handling of data. Once the output of the DNN model reaches the main thread, the smart mirror evaluates the output. This evaluation may involve operations based on predefined logic or conditions. Based on the evaluation result, the state will execute the corresponding method. These methods may involve updating the display content of the mirror, adjusting mirror settings, or performing other relevant operations.

If specific transition conditions are met, the state instructs the state machine to perform a state transition. State transitions may involve entering another state or executing different operational flows. Additionally, the state can also notify the UI controller to switch UI interfaces. This allows the screen display to change based on the system's requirements and state variations. The state can also send events to other threads. These events can trigger specific operations in other threads or perform specific tasks. For example, these events can be used to control the model's activation, notifying relevant threads to execute model-related operations.

F. Mobile Bluetooth Pairing

We provide Bluetooth connectivity between the mobile device and the software side (as shown in Figure 7).

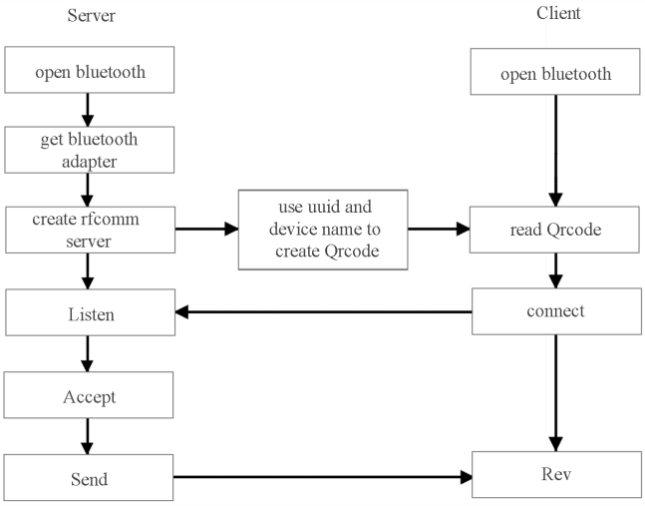


Fig. 7 Mobile phone bluetooth connection

Through Bluetooth pairing, users can conveniently view the corresponding acupoints for their symptoms and access

detailed information about the desired acupoints. When the user activates the Bluetooth pairing function, the mobile device acts as an RFCOMM server and generates a specific UUID for identification during the connection. This UUID, along with the device name, is encoded into a QR code and displayed on the mobile device. The user simply needs to place the QR code displayed on the mobile device within the field of view of the webcam. The Jetson Nano, equipped with built-in Bluetooth functionality, scans and decodes the QR code, thereby obtaining the device's MAC address and server UUID.

The Bluetooth pairing functionality transforms the mobile device into a convenient control interface, allowing users to effortlessly operate and access the desired acupoint information. This design not only provides a more intuitive and interactive interface but also enhances usability by increasing convenience.

G. *BERT AI Consultation*

BERT AI Figure 8 illustrates the AI consultation architecture of this project. We have set up our own server and built a simple chatbot on the server. We utilize Rasa to create the chatbot, and when the user's intent on the mobile device is to inquire about a disease or symptom, the BERT model trained in this project provides corresponding effective acupoints based on the user's input. The smart mirror then displays only these acupoints, allowing the user to further alleviate their symptoms.

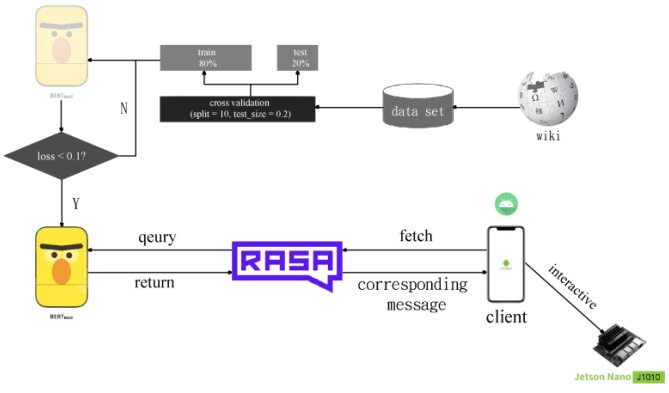


Fig. 8 Architecture Diagram of BERT AI Consultation

In this work, we employ the pre-trained BERT model, specifically the bert-base-chinese variant [13]. This model consists of 12 layers, 768 hidden units, and 12 attention heads, with a parameter size of 110MB. Before training the model, we initialize the token input of the BERT model and incorporate data collected from Wikipedia. Currently, our dataset includes dialogue information associated with 109 disease categories and symptoms related to these diseases, totaling 623 records. We use this data for fine-tuning the model, which involves connecting a new simple classifier to the last layer of BERT to identify the disease category based on the user's input.

IV. RESULT

A. *Acupoint Mapping Effect*

Our algorithm is based on the model proposed by Mediapipe and applied to hand and head acupoint mapping. Using the landmarks provided by Mediapipe, our algorithm achieves acupoint display accuracy consistent with the Mediapipe model [14], reaching up to 95.7%, when the hand is rotated or fully extended. The effectiveness of acupoint mapping can be observed in Figure 9. The hand acupoints adhere to fixed positions and do not shift due to hand rotation or other gestures. This demonstrates the accuracy and effectiveness of our acupoint mapping algorithm.

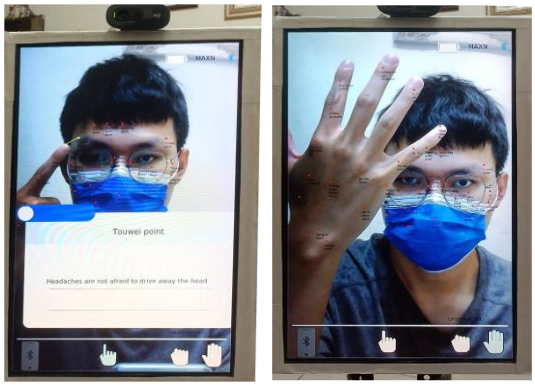


Fig. 9 Acupoint Mapping Effect Display

B. *DNN Performance Evaluation*

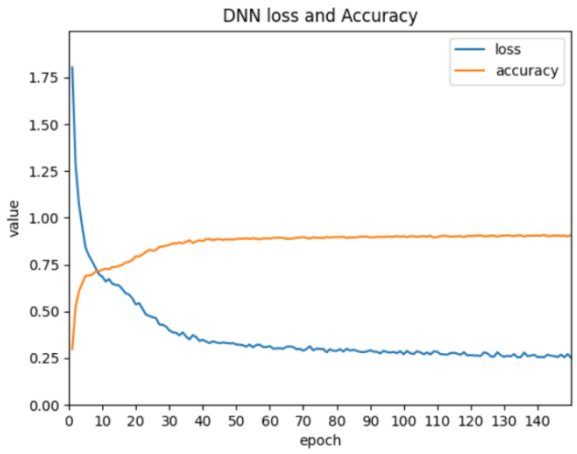


Fig. 10 DNN Hand Gesture Recognition Loss and Accuracy

$$\text{Loss} = -\sum_{i=1}^{output\ size} y_i \cdot \log \hat{y}_i \quad (3)$$

We employed DNN for hand gesture recognition with 150 epochs, a batch size of 16, and the categorical cross-entropy loss function (Equation 3). By observing the loss function during training, we can see a gradual decrease in loss and a corresponding increase in accuracy, indicating good learning performance of the model. The DNN hand gesture recognition model in this project achieved an impressive accuracy rate of up to 90%. This indicates that our model can accurately classify the user's gestures into six predefined types (menu, cancel, choose, right, left, no gesture). Such high accuracy provides reliable hand gesture recognition functionality for our application, enabling users to perform corresponding operations smoothly. Monitoring the loss function and accuracy ensures that the training process and results of the model meet the intended goals. Figure 10 intuitively illustrates the progress and effectiveness of our model training. This technology exhibits high accuracy in hand gesture recognition, providing a solid foundation for the successful implementation of this project.

C. BERT Performance Evaluation

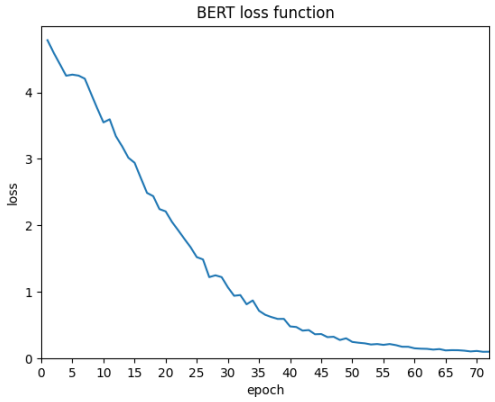


Fig. 11 BERT Loss Fuction

In this paper, we divided the dataset into an 80% training set and a 20% testing set for fine-tuning the BERT model. During the training process, we used 72 epochs and a batch size of 10. By observing the variation of the loss function in Figure 11, we can see a gradual decrease in the loss function, indicating an improvement in the model's learning performance. We terminated the training when the loss function was below 0.1. After each training session, we obtained four model metrics: Accuracy, Precision, Recall, and F1-Score. To ensure the credibility of the experiment, we performed ten-fold cross-validation, averaging the model metrics from the ten iterations as the final performance indicators. This approach helps reduce the impact of random variations in individual training sessions and provides a more comprehensive evaluation of the model's performance.

Based on our results, our model performed impressively in disease prediction, achieving an accuracy rate of nearly 80%. This means that our model can provide accurate predictions for the symptoms and disease categories entered by the user. The results in Figure 12 further validate the effectiveness and reliability of our trained BERT model in the AI chatbot for medical consultations.。

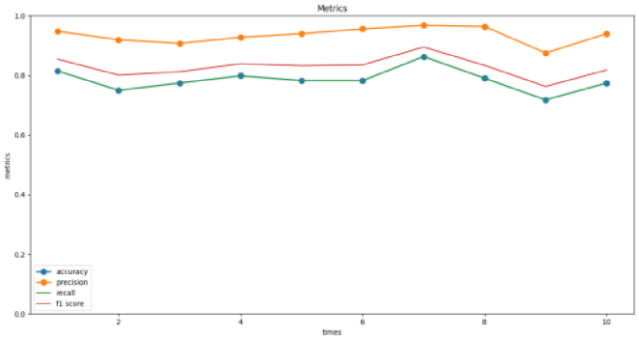


Fig. 12 Architecture Diagram of BERT AI Consultation

V. CONCLUSIONS

In conclusion, this paper has established algorithms for accurate mapping of acupuncture points, demonstrated augmented reality visualization using MEDIAPIPE, achieved hand gesture recognition using a simple neural network based on DNN, and proposed a disease analysis technique based on BERT. We have discussed the accuracy of disease inference in our study and successfully developed an embedded system, integrating the technology into JETSON NANO. This allows users to interactively explore and understand acupuncture knowledge by observing acupuncture points on the human body through our developed product.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology (MOST), Taiwan, R.O.C., under the Grant MOST 111-2221-E-507-006.

REFERENCES

[1] Gee-Sern Hsu, Hsiao-Chia Peng and Kai-Hsiang Chang, Landmark Based Facial Component Reconstruction for Recognition Across Pose, June 23-28 2014, pp. 34-39.

[2] Menglong Chang, Qing Zhu, Automatic location of facial acupuncture-point based on facial feature points positioning April 1 2017, pp. 545-549.

[3] Yi-Zhang Chen, Corky Maigre , Min-Chun Hu and Kun-chan Lan, Localization of Acupoints using Augmented Reality, June 20 2017, pp. 239-241.

[4] Menghe Zhang, Jürgen P. Schulze, Dong Zhang, FACEATLASAR: ATLAS OF FACIAL ACUPUNCTURE POINTS IN AUGMENTED REALITY, Mar 29 2022.

[5] Baohua Sun, Lin Yang, Wenhan Zhang, Patrick Dong, Charles Young, Jason Dong, Michael Lin, Demonstration of Applications in Computer Vision and NLP on Ultra Power-Efficient CNN Domain Specific Accelerator with 9.3TOPS/Watt, 2019, pp.1213-1223.

[6] Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, Constantinos S. Pattichis, AI in Medical Imaging Informatics: Current Challenges and Future Directions, 2020, pp.1-15.

[7] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann , MediaPipe: A Framework for Building Perception Pipelines,Jun 14 2019, pp.1-8.

[8] Google(n.d.), MediaPipe Holistic,Retrieved December 2 , 2022, from https://google.github.io/MediaPipe/solutions/holistic.html.

[9] Joshua Willman, Apress Berkeley, Hampton, USA, December 8, 2020.

[10] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Oct 11 2018 , pp.1~16.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, Jun 12 2017, pp.1-15.

[12] Alexander Kapitanov, Andrey Makhlyarchuk, Karina Kvanchiani, HaGRID -- HAnd Gesture Recognition Image Dataset, Jun 16 2022 [CrossRef][PubMed].

[13] Hugging Face team, “BERT model pre-trained for Chinese” , Hugging Face, https://huggingface.co/bert-base-chinese, Jun 2019(accessed Dec 2022)

[14] Google, “MediaPipe Hands Solution”, github, https://github.com/google/mediapipe/blob/master/docs/solutions/hands.md , Jun. 2020(accessed May. 2023).

Anomaly Detection of Rolling Element Bearings using Domain Adaptation

Hsin-Tse Lu*
* National Chengchi University, Taiwan
E-mail: owen.lu.academic@gmail.com

***Abstract*—To address manpower and equipment efficiency issues in the manufacturing industry, IoT and AI technologies are employed to collect and diagnose machine operation. Vibration analysis has been proven effective in monitoring rotating equipment and predicting abnormalities. However, when applied to large machines, a universal anomaly detection model becomes impractical, requiring numerous models for different areas, leading to management challenges.This study proposes field adaptation to align the features of two machines onto the same vector in a coordinate plane. This enables a single anomaly detection model to monitor both machines, achieving model generalization and reducing administrative costs. By adjusting the features, the same detection model can be applied to different machines, enhancing efficiency and simplifying model management in the monitoring process.**

I. Introduction

The rolling mill process is a vital component of metalworking, involving the reduction of metal stock thickness through the use of rollers[1]. To ensure uniform thinness, precise control is necessary, relying on a rolling theory-based mathematical model that considers key parameters such as rolling speed, roll gap, and rolling force. Among the critical components involved in these subprocesses are the rolling element bearings (REBs) within the rotating equipment. Monitoring the condition of REBs on the rolling mill production line is crucial to maintain product quality. However, implementing effective monitoring methods in factories faces obstacles related to label uncertainty and model unscalability. Class imbalance is a significant issue stemming from the division of REB conditions into abnormal and normal classes. This imbalance leads to costly deployments on production lines and exposes them to uncertainty. Classifiers operating under such conditions may yield misleading accuracy rates, failing to detect true positive faults and compromising the effectiveness of the monitoring solutions. Addressing label uncertainty and class imbalance is essential to enhance the implementation of monitoring methods in factory settings.

II. Methodology

The proposed approach aiming to align them by projecting them onto the same coordinate plane during training. The objective is to maximize the overlap between the two data sets. This training method requires complete data labels for the source data set, while the target data set can be unlabeled. The key process in this approach is Domain Adaptation[2], which focuses on aligning the data sets from the two fields on the same coordinate plane. The health score, defined as the density of anomalies over a fixed time period, will serve as the performance indicator in this study.

III. Results and Discussion

Before Domain Adaptation, we analyzing the model cross-validation results between Device 1 and Device 2. The obtained health score was 23.12 during normal condition and 28.23 during abnormal condition. This implies that the model fails to provide accurate alerts when Device 2 is in an abnormal condition. Conversely, when using the model trained on Device 2 to assess the normal operation of Device 1, a health score of 55.21 was obtained during normal condition and 32.45 during abnormal condition. Here, the issue of false alarms becomes apparent.

After Domain Adaptation, it was found that training the model using Device 1 as the source yielded an average health score of 1.33 during normal condition and a health score of 75.55 during abnormal condition. Similarly, training Model 1 with Device 2 as the target resulted in a health score of 2.12 during normal condition and 67.31 during abnormal condition. These outcomes indicate that both devices can derive satisfactory benefits when training the model with their respective data.

IV. Conclusions

The study utilizes domain adaptation to achieve overlap between the two sets of equipment data in the vector space. By aligning the distributions closely, a model is employed to detect the normalcy of the equipment, resulting in promising outcomes. This highlights the significance of field adaptation in achieving universality among data with similar characteristics.

Acknowledgment

This work was supported by National Chengchi University, Taiwan under grants 112KY01202.

References

[1] B. Li, M.-Y. Chow, Y. Tipsuwan, and J. C. Hung, "Neural-network-based motor rolling bearing fault diagnosis," *IEEE transactions on industrial electronics*, vol. 47, no. 5, pp. 1060–1069, 2000.

[2] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

Dental Positioning Medical Assistance System for BiteWing Radiograph Based on YOLOV4 and Image Enhancement

Mu-Feng Lin [1], Yi-Qian Li [1], Tsung-Yi Chen [1], Yu-Lin Liu [1], Yuan-Jin Lin [2], Mei-Ling Chan [3], Chiung-An Chen [4], Kuo-Chen Li [5], Chun-Wei Li [6], and Shih-Lun Chen [1,*]

* Department of Electronic Engineering, Chung Yuan Christian University, Chung Li City 32023, Taiwan;
E-mail: s10926229@cycu.edu.tw; s10926248@cycu.edu.tw; g10976016@cycu.edu.tw; g11176050@cycu.edu.tw; chischen@cycu.edu.tw; Tel: +886-3-265-4644

Department of Electrical Engineering, National Cheng Kung University, Taiwan
E-mail: m26124069@gs.ncku.edu.tw Tel: +886-3-265-4644

School of Physical Educational College, Jiaying University, Meizhou City 514000, Guangdong Province, China;
E-mail: Lynn202207017@jyu.edu.cn

Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 243303, Taiwan
E-mail: joannechen@mail.mcut.edu.tw; Tel: 886-2-2908-9899#4813

Department of Information Management, Chung Yuan Christian University, Chung Li City 32023, Taiwan;
E-mail: kuochen@cycu.edu.tw; Tel: +886-3-265- 2050

Department of General Dentistry, Chang Gung Memorial Hospital, Taoyuan City 33305, Taiwan;
E-mail: harrythebold@cgmh.org.tw

Abstract — In modern medicine, X-rays play a crucial role in diagnosis and maintaining patient records. In dentistry, bite-wing radiographs are commonly used to evaluate dental conditions. However, manually positioning teeth during diagnosis can be time-consuming and labor-intensive. To address this, this study utilizes artificial intelligence (AI) to assist dentists, speeding up the diagnosis process and providing rapid reference samples. We collaborated with five experienced dentists from Chang Gung Memorial Hospital, who provided bite-wing radiographs and corresponding tooth position data for training the model. Using the You Only Look Once (YOLO) V4 model as the foundation, this research applied image enhancement techniques inspired by existing literature, including bilateral filters and adaptive Gaussian binarization. These techniques reduced noise and improved tooth contour clarity, resulting in improved model performance, reduced processing time, and approximately 7% higher accuracy compared to existing techniques. Additionally, this research proposed image enhancement method increased the model recognition speed by 60.19% compared to using the original images. The study obtained ethical approval from the Research Institution Review Board (IRB) under application number 202002030B0. Integrating AI into tooth positioning enhances diagnostic efficiency and advances AI in dentistry, contributing to more effective and efficient dental care.

I. INTRODUCTION

The rapid advancements in medical technology have driven the search for more effective and accurate diagnostic methods. Precise and fine-tuning of teeth is the future of dental applications, especially the precise positioning of teeth. Current tooth positioning techniques often use various types of deep learning models and primarily focus on symptom detection, such as the study [1] which utilizes mask-RCNN for caries symptom detection with a maximum accuracy of 92%. However, there is less emphasis on dedicated tooth positioning. Yet, tooth positioning aids dentists in understanding the location of teeth by using the FDI tooth notation system. Therefore, there is a space for development, as the reference [2], where a CNN model based on Dental Panoramic Radiograph (DPR) imaging data achieves a tooth detection accuracy of 99.7%. However, compared to DPR, bite-wing radiographs are more commonly used for diagnosis. Therefore, the purpose of this paper is to explore a precise tooth position determination method based on the bitewing sheet and using a deep learning model.

According to the reference[3], image preprocessing plays a crucial role in enhancing model training effectiveness. In this approach, we employ image enhancement techniques to optimize the data. Firstly, bilateral filtering [4] is applied, effectively eliminating noise while preserving edge information within the image. Subsequently, adaptive Gaussian binarization [5] is utilized to establish pixel relationships with neighboring regions. This process converts the image into a binary black-and-white format, often used in conjunction with filtering methods. To facilitate tooth recognition, we employ the YOLOv4 model [6], a popular CNN model renowned for real-time object detection in images. Previous iterations of YOLO, such as YOLOv1 and YOLOv2, had limitations in accuracy and recall, particularly for detecting smaller objects. YOLOv3 improved upon these limitations but required substantial computational resources. Hence, our study adopts the YOLOv4 model, which not only provides higher accuracy and speed but also integrates various optimization techniques for improved resource efficiency and ease of use. By applying the YOLOv4 model to bite-wing radiographs, we can effectively detect and localize tooth positions, enabling the extraction of individual teeth for further analysis

In view of this, the aim of this research is to explore the integration of artificial intelligence techniques into tooth positioning in dentistry. By leveraging bite-wing radiographs and the YOLOv4 model, the study will investigate how AI can assist dentists in positioning teeth accurately and efficiently. The research will focus on developing image enhancement techniques, such as bilateral filters and adaptive Gaussian binarization, to improve tooth contour clarity and reduce processing time. The dissertation will assess the impact of AI-based tooth positioning on diagnostic accuracy and processing speed, enhancing the objectivity and accuracy of diagnoses.

II. METHODS

This study proposes a solution to automatically locate and extract teeth in bitewing radiographs and build up a medical database. In the past, this task was accomplished through manual operations. In recent years, specialized algorithms have been proposed, however, pixel-based algorithms have limited capabilities in handling complex images and are prone to interference from non-teeth objects. Therefore, this study utilizes the YOLOv4 model, based on the One-Stage object detection method, to replace the approach and achieve tooth localization and cropping. The method of this study, as depicted in Fig.1, can be divided into three main steps: (1) image processing, (2) YOLOv4 model training, and (3) tooth detection for tooth localization and image cropping upon completion of the model training.

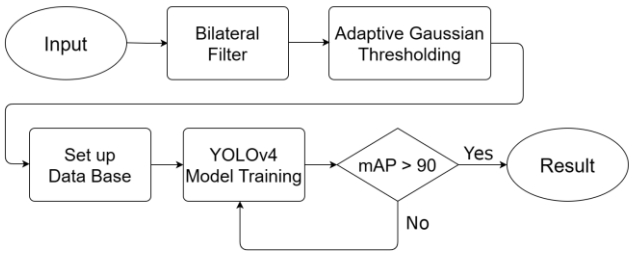


Fig. 1. The flowchart of the method of this study.

In this study, a total of 315 bitewing radiographs provided by Chang Gung Memorial Hospital in Taoyuan, were employed to execute the tasks of Training and Evaluating the object detection model. In addition, this study disregarded the irrecoverable factors such as the imaging range and photography parameters that may affect the quality of the bitewing radiographic images.

A. Image Enhancement

In the process of object detection, the localization boundaries of tooth objects are determined based on the contours of the teeth. Moreover, the brightness and shape of teeth on each individual bitewing radiograph can vary significantly. Therefore, in this study, bilateral filtering and adaptive Gaussian thresholding methods are employed to obtain a binary image with well-defined tooth contours from the bitewing radiographs. This binary image is then used for training the object detection model, enabling it to capture the characteristic contours in a visually informative manner.

1. Noise Reduction

The application of low-pass filtering techniques allows for effective blurring and denoising of images. In geometric space, common methods such as Gaussian filtering utilize the spatial relationship between pixels to apply a weighted filter based on

this relationship. Similarly, bilateral filtering takes into account both the spatial and intensity relationships of pixels. It applies a weighted filter considering the spatial distance and intensity similarity of pixels. As a result, bilateral filtering simultaneously achieves image blurring, noise reduction, and preservation of edge characteristics. This is particularly beneficial as pixels along edges often exhibit significant intensity differences. By leveraging these advantages, the utilization of bilateral filtering prior to image binarization can effectively address the issue of excessive noise being transformed into small black-and-white dots, which can potentially interfere with model learning. Moreover, this approach maximizes the preservation of edge features. The bilateral filtering process is mathematically represented by equation (2), where I(x,y) denotes the denoised pixel.

$$w(x,y,i,j) = \exp\left(-\frac{(x-i)^2+(y-j)^2}{2\sigma_s^2} - \frac{\|I(x,y)-I(i,j)\|^2}{2\sigma_r^2}\right) \quad (1)$$

$$I(x,y) = \frac{\sum_{i,j} I(i,j)w(x,y,i,j)}{\sum_{i,j} w(x,y,i,j)} \quad (2)$$

By incorporating bilateral filtering into the image preprocessing pipeline, we can enhance the quality of images for subsequent analysis, minimizing noise interference while retaining important edge information. Besides, a 13×13 kernel is used as the neighboring region for filtering in this study, and spatial parameter σ_s and range parameter σ_r are set to 3 and 100, respectively. Here, σ_s represents the standard deviation of the geometric distribution in the coordinate space, and σ_r represents the standard deviation of the photometric distribution in the color space. A larger σ_s results in a more pronounced blurring effect and an increased proportion of combined pixels from greater spatial distances. Similarly, a larger σ_r results in a more pronounced blurring effect and an increased proportion of combined pixels with greater intensity differences.

2. Image Binarization

Global thresholding, which applies a single threshold to the entire image for binarization, may result in the suboptimal conversion of details in high or low intensity regions. In contrast, adaptive thresholding calculates thresholds based on regional characteristics and performs binarization on pixels. This method helps avoid the issues mentioned above. Additionally, considering the non-original objects overlapped on bitewing radiographs, using global thresholding methods like Otsu's method [7] may introduce biases that affect the accuracy of the binarization. Adaptive thresholding, with its ability to adapt to local characteristics, can mitigate interference from external objects as well.

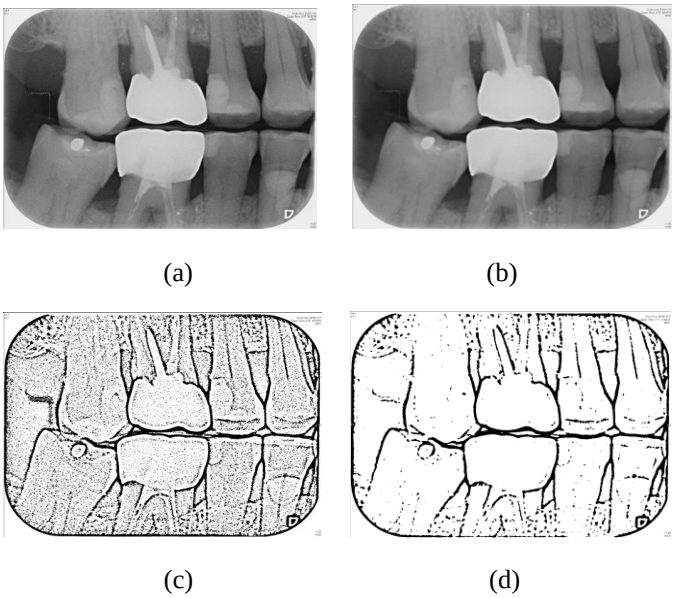
Adaptive mean thresholding [8][9] and adaptive Gaussian thresholding are both types of adaptive thresholding methods. The former uses the average pixel value of the neighboring pixels around the target pixel as the threshold, while the latter, like Gaussian filtering, considers the spatial relationship of pixels and uses a weighted average pixel value as the threshold as shown in equation (4).

$$g(x,y,i,j) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x-i)^2+(y-j)^2}{2\sigma^2}\right) \quad (3)$$

$$T(x,y) = \frac{\sum_{i,j} I(i,j)g(x,y,i,j)}{\sum_{i,j} g(x,y,i,j)} - C \quad (4)$$

$$I(x,y) = \begin{cases} 255, & I(x,y) > T(x,y) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

In this study, we utilized a 55×55 Gaussian kernel and a standard deviation σ of 8.6. This configuration ensures that the binarization effect highlights the contours of the teeth in the image. Fig 2 illustrates the original image and the enhanced image before and after the binarization process.



(a) (b)

(c) (d)

Fig. 2. Pre-cropping processing of BW image. (a) Original image. (b) Image denoising using Bilateral filtering. (c) Original image binarization using adaptive thresholding. (d) Binarize the denoised image using an adaptive threshold.

B. One-Stage Object Detection

Object detection methods play a pivotal role in accurately localizing and classifying teeth, providing significant potential for handling complex patterns in practical applications. Within the state-of-the-art techniques, two main approaches stand out: two-stage object detection and one-stage object detection. The former emphasizes detection accuracy and commonly utilizes models such as the R-CNN series. The latter prioritizes inference speed by combining recognition (classification) and localization into a single step, thereby enhancing execution efficiency. Notable models in this category include the YOLO series.

With the aim of achieving real-time performance, this study adopts the YOLOv4 architecture[10], a representative model from the one-stage object detection method, as the core technology for tooth detection and model development. By leveraging the strengths of YOLOv4, this research seeks to achieve efficient and accurate tooth localization within dental imaging applications.

1. Data Distribution

In this study, the image samples were allocated randomly, following a distribution of 60% for the training set, 20% for the validation set, and 20% for the testing set. Notably, the data from the testing set was entirely excluded from the model training process. This strict separation ensures that the testing set remains independent and is not utilized in any stage of model learning. The aim of this step is to ensure that the model possesses reliable and robust generalization capabilities. Furthermore, increasing the sample size enables a larger proportion of the training set to be utilized. This offers several advantages, including enhanced model performance and an improved ability to capture diverse patterns and variations in the data. In this study, by maximizing the sample size, the richness and complexity of the training set can be effectively utilized to enhance the learning and predictive capabilities of the model.The allocated training and validation samples were processed to create the datasets available for model learning. The data distribution used in this study is shown in Table 1.

Table. 1. Quantity of bitewing images in each dataset.

Quantity of Bitewing Image in each Dataset				
Dataset	Training	Validation	Testing	Total
Bitewings	189	21	105	315

2. Data Preprocessing

Prior to model training, the data undergoes additional preprocessing steps to optimize the learning process. Firstly, the bitewing radiographs and their corresponding annotated bounding boxes are resized and padded to align with the input size required by the model. This ensures consistency and compatibility between the data and the model architecture. Secondly, data augmentation techniques are applied to the images within the training set. Random operations such as vertical and horizontal flipping, as well as mosaic effects, are introduced. These augmentations introduce variability and diversity into the training data, enabling the model to generalize better and improve its ability to handle various scenarios. Finally, to mitigate any potential bias during training, the order of the data is randomized. This shuffling process ensures that the model receives a diverse range of samples in each training batch, preventing it from learning patterns specific to the sequential arrangement of the data.

By employing these preprocessing steps, we aim to optimize the data representation, enhance the model's adaptability, and minimize any unintended biases, thereby facilitating effective and robust training of the model.

3. Model Training & Tooth Detection

After the preprocessing stage, the training of the YOLOv4 model is initiated. The architecture of the YOLOv4 model is illustrated in Figure 3. The diagram showcases a section of the model, with an input size of 416x416x3. The model incorporates various components: CSPDarknet53 [11] serves as the Backbone, Spatial Pyramid Pooling (SPP)[12], and Path Aggregation Network (PAN) [13] functions as the Neck, and YOLO [14][15]acts as the Head. Additional hyperparameters can be found in Table 2. For further context, Table 3 provides

supplementary details regarding the software and hardware equipment employed in this study.

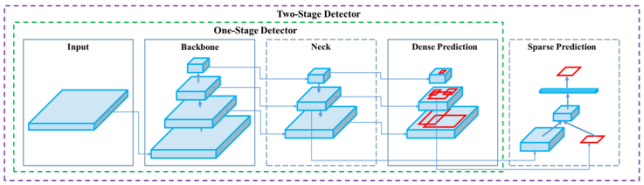


Fig. 3. Schematic diagram of YOLOv4 architecture [14].

Table. 2. Parameters of YOLOv4 model used in this study.

YOLOv4 Model	
Backbone	CSPDarknet53
Neck	SPP, PAN
Head	YOLO
Input Size	416×416×3
Optimizer	Adam
Learning Rate	0.001
Momentum	0.9
Batch Size	4
Epoch	400

Table. 3. Hardware devices and software applications

Hardware Devices	
CPU	Intel® Core™ i7-11800H
GPU	NVIDIA® GeForce RTX™ 3060
Memory	GDDR6 (6GB), DDR4 (40G)
Software Applications	
OS	Windows 11 Pro
Python	Ver. 3.8.16
VSC	Ver. 1.79.0

C. Tooth Localization & Image Cropping

Utilizing a trained model for tooth detection enables the extraction of tooth positions and boundaries from bitewing radiographs. In clinical settings, dental conditions are frequently assessed by focusing on half of a tooth as a unit. Consequently, this study adopts a methodology where the detected teeth are cropped using their localization boundaries and subsequently vertically sliced into halves. The resulting teeth are then assigned sequential numbers from left to right, distinguishing between upper and lower positions based on their relative X and Y coordinates. This systematic approach facilitates the automated construction of a comprehensive medical database.

III. Results

Table 4 presents the operational guidelines for the confusion matrix, while equations (6)-(9) outline the calculation methods for these indicators. Following the aforementioned approach, the mean average precision (mAP), precision, recall, and F1-score of the YOLOv4 model were evaluated based on its prediction results utilizing a confusion matrix. Figure 4 illustrates the training process of the YOLOv4 loss function. Additionally, Table 5 displays the changes in mAP at each stage during the training process.

$$\text{Accuracy: } \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{6}$$

$$\text{Precision: } \frac{TP}{(TP+FP)} \tag{7}$$

$$\text{Recall: } \frac{TP}{(TP+FN)} \tag{8}$$

$$\text{F1 Score: } \frac{2\times(Precision\times Recall)}{(Precision+Recall)} \tag{9}$$

Table. 4. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

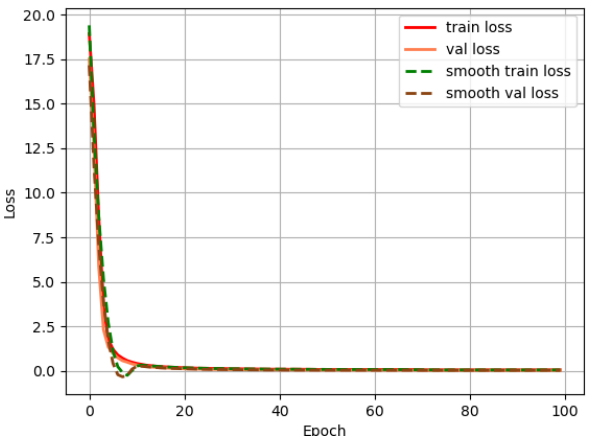


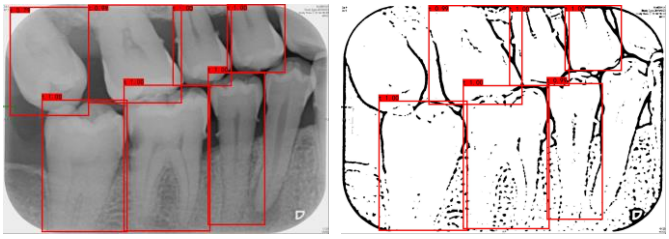
Fig. 4. Training Process for the YOLOv4 loss function.

Table. 5. The change of mAP value during the training process of this study.

Epoch	Origin Data	Processed Data
20	98.16%	97.62%
40	99.25%	98.91%
60	99.23%	98.82%

80	99.36%	98.87%
100	99.38%	99.10%

During the testing phase, we used a set of 105 untrained images as the test dataset. The mAP (mean Average Precision) value can be used to evaluate the performance of a model. When the training epoch is set to 100, using the original images as input data gets the better results. However, even when using enhancement images, the performance does not significantly decrease. The actual identification results are shown in Fig 5. According to the resulting photo, it can be observed that even though using original images captures more target objects compared to using enhancement images, there are still some instances of missing detections.



(a) (b)

Fig. 5. Recognition results through YOLOV4 (a) Original Image. (b)Enhancement Image

Based on the findings presented in Table 6 and Figure 6, it is evident that employing image augmentation techniques yields comparable performance to using original images across various metrics. Additionally, the prediction time for a single image is significantly reduced by over 61%, highlighting the efficiency and effectiveness of employing image processing techniques. Moreover, in comparison to the existing technology [16], as illustrated in Table 7, the Precision, Recall, and F1 Score exhibited enhancements of 4.85, 0.38, and 2.99, correspondingly. These experimental results demonstrate the enhanced effectiveness of the proposed technology in accurately identifying tooth positions. Furthermore, the utilization of image augmentation techniques not only maintains performance but also enhances the model's robustness by introducing variations in the training data. This ensures that the model can handle diverse dental conditions and scenarios. These advancements contribute to more reliable and effective tooth position identification in dental diagnostics.

Table. 6. Original image compared to Enhancement image.

	Original	Enhancement
mAP	99.38%	99.10%
Recall	98.43%	97.86%
Precision	98.01%	97.44%
F1 Score	98%	98%
Processing Time	0.134s	0.052s

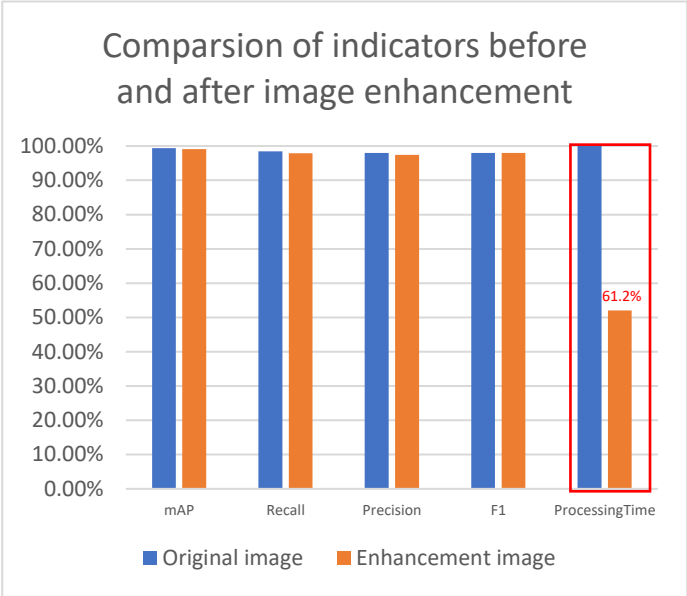


Fig. 6. Comparison of indicators before and after image enhancement

Table. 7. The comparison between the prior and this study

	This Work	Method in [16]
Precision	97.44%	92.93%
Recall	97.86%	97.48%
F1 Score	98%	95.15%

IV. DISCUSSIONS

In this study, we achieved a high mAP of 99.10% in tooth position detection using the YOLOv4 model, surpassing the performance of the Faster R-CNN model used in the existing method [16]. The YOLOv4 model demonstrated superior performance in comparison. The incorporation of image enhancement techniques, specifically bilateral filtering for noise removal and adaptive Gaussian thresholding for image binarization, played a crucial role in improving the results. Compared with the reference model [16], the model shows significant enhancement in various metrics. Precision increased

by 4.51%, recall increased by 0.38%, and F1 score improved by 2.85%. It is important to note that the reference model does not use image augmentation, which may lead to its lower mAP. Despite this minor decrease, our model maintained comparable performance, and the average processing time for predicting images was significantly reduced. The average processing time for predicting an unknown image was reduced by more than 50% compared to using original images. These results indicate that the integration of image enhancement techniques with the YOLOv4 model can effectively approximate the expertise of experienced dental professionals.

V. CONCLUSIONS

This dissertation aims to integrate artificial intelligence into bite-wing radiographs to automate the process of tooth positioning in dental diagnosis. Although the binarization process resulted in a lower level of clarity compared to the original images, it is evident that automation using AI techniques holds great potential for improving efficiency and reducing human errors in dental diagnosis. Future studies should focus on preserving more contours in binarized images while simultaneously optimizing processing time, which will contribute to the development of accurate AI models for dental condition diagnosis. The findings of this research provide a solid foundation for the advancement of deep learning models in dental diagnostics, highlighting the importance of advancing image enhancement techniques and their integration with state-of-the-art object detection models. Continued exploration in this field will lead to enhanced accuracy and effectiveness in dental diagnosis, benefiting both patients and dental professionals.

VI. REFERENCES

[1] U. Rashid et al., ‘A hybrid mask RCNN-based tool to localize dental cavities from real-time mixed photographic images’, PeerJ Comput Sci, vol. 8, p. e888, Feb. 2022, doi: 10.7717/peerj-cs.888.

[2] T. H. Bui, K. Hamamoto, and M. P. Paing, ‘Tooth Localization using YOLOv3 for Dental Diagnosis on Panoramic Radiographs’, 電気学会論文誌 C（電子・情報・システム部門誌）, vol. 142, no. 5, pp. 557–562, 2022, doi: 10.1541/ieejeiss.142.557.

[3] ‘The Application of Deep Learning in Image Processing is Studied Based on the Reel Neural Network Model - IOPscience’. https://iopscience.iop.org/article/10.1088/1742-6596/1881/3/032096/meta (accessed Jun. 21, 2023).

[4] ‘Bilateral filtering for gray and color images | IEEE Conference Publication | IEEE Xplore’. https://ieeexplore.ieee.org/document/710815 (accessed Jun. 21, 2023).

[5] ‘Enhancement of Degraded Manuscript Images using Adaptive Gaussian Thresholding | Semantic Scholar’. https://www.semanticscholar.org/paper/Enhancement-of-Degraded-Manuscript-Images-using-Khetan-Kejriwal/a8840403c2ecd83ef28db05302fbb992c45a6797 (accessed Jun. 21, 2023).

[6] O. Boudraa, W. K. Hidouci, and D. Michelucci, ‘Degraded Historical Documents Images Binarization Using a Combination of Enhanced Techniques’. arXiv, Jan. 27, 2019. doi: 10.48550/arXiv.1901.09425.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ‘YOLOv4: Optimal Speed and Accuracy of Object Detection’. arXiv, Apr. 22, 2020. doi: 10.48550/arXiv.2004.10934.

[8] N. Otsu, ‘A Threshold Selection Method from Gray-Level Histograms’, IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[9] ‘Adaptive document image binarization - ScienceDirect’. https://www.sciencedirect.com/science/article/abs/pii/S0031320399000552 (accessed Jun. 21, 2023).

[10] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, ‘CSPNet: A New Backbone that can Enhance Learning Capability of CNN’. arXiv, Nov. 26, 2019. doi: 10.48550/arXiv.1911.11929.

[11] K. He, X. Zhang, S. Ren, and J. Sun, ‘Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition’, 2014, pp. 346–361. doi: 10.1007/978-3-319-10578-9_23.

[12] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, 'Path Aggregation Network for Instance Segmentation'. arXiv, Sep. 18, 2018. doi: 10.48550/arXiv.1803.01534.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You Only Look Once: Unified, Real-Time Object Detection'. arXiv, May 09, 2016. doi: 10.48550/arXiv.1506.02640.

[14] J. Redmon and A. Farhadi, 'YOLO9000: Better, Faster, Stronger'. arXiv, Dec. 25, 2016. doi: 10.48550/arXiv.1612.08242.

[15] J. Redmon and A. Farhadi, 'YOLOv3: An Incremental Improvement'. arXiv, Apr. 08, 2018. doi: 10.48550/arXiv.1804.02767.

[16] Y. Yasa et al., 'An artificial intelligence proposal to automatic teeth detection and numbering in dental bite-wing radiographs', Acta Odontol Scand, vol. 79, no. 4, pp. 275–281, May 2021, doi: 10.1080/00016357.2020.1840624.

Asynchronous Federated Learning Method in the 6G Massive Multimedia Transmission System

Yen-Ching Chang, Sheng-Siang Liao and Chin-Feng Lai *
* National Cheng Kung University, Taiwan
E-mail: cinfon@ieee.org Tel: +886-2757575

Abstract — **In the past, machine learning required centralizing the training dataset on a server or data center for training. However, due to recent concerns regarding privacy, related legal regulations, and competition among peers, this approach of uploading data to servers is not suitable in certain domains. Therefore, Google proposed 'federated learning' in 2016, which distributes the training data across different client devices. The client devices train models and upload model parameters, which are eventually aggregated on the server to generate a global model, ensuring the privacy of the training data. Traditional federated learning often faces challenges due to significant disparities in computational capabilities among heterogeneous devices. As a result, asynchronous federated learning has been introduced to improve training efficiency and reliability. The deployment of a scalable federated learning framework becomes critical, especially with the emergence of 6G communication systems, which offer low latency and high coverage, enabling services for multimedia-related applications such as autonomous vehicles, smart industries, and augmented virtual reality. Handling multi-gigabyte multimedia data to enable learning at the edge devices is of paramount importance.**

I. INTRODUCTION

The application characteristic of 6G is high coverage, which is encountered in many federated learning scenarios where a large number of devices participate in training simultaneously. For such a large number of devices, asynchronous federated learning is one solution. Through an asynchronous approach, client devices can upload their models to the server for aggregation immediately after completing training, without waiting for other devices. Therefore, the support for asynchronous weight updates in the federated learning framework becomes even more important.

II. RELATED WORKS

Federated Learning is a novel approach that enables multiple entities to collaborate on machine learning model training without directly accessing local training data. First introduced by Google in 2016, federated learning serves as a machine learning solution that meets the requirements of data privacy and communication efficiency [6, 8]. The design objective of federated learning is to ensure individual data privacy and develop satisfactory machine learning models across multiple participants or computing nodes, under lawful and compliant conditions [7, 9]. Consequently, many papers adopt the federated learning framework, where the central server collects parameters of local models from multiple nodes before aggregating them into a global model in each training round [10].

When traditional federated learning is applied to resource-constrained devices, various challenges inevitably arise [5]: (1) Unreliable heterogeneous devices: The central server waits for selected endpoints to upload the latest local models, but these endpoints may unintentionally drop out of training due to uncontrollable factors. (2) Low training round efficiency: Due to device resource disparities (device heterogeneity) and variations in training data across devices (data heterogeneity), fast devices have to wait for outdated local models uploaded from slower devices in each training round. (3) Low resource utilization: Selected devices' quality varies due to inefficient node selection algorithms. (4) Security and privacy concerns: Several attacks, including poisoning and backdoor attacks, can compromise the security of traditional federated learning. Privacy concerns arise from potential data leakage during the training process. To overcome these challenges and achieve rapid aggregation of appropriate local models to the central server, asynchronous federated learning was proposed.

The primary challenge in asynchronous federated learning for heterogeneous systems lies in efficiently maximizing resource utilization across such systems. In a survey paper [1] that collected research papers on asynchronous federated learning from 2019 to 2022, the problem of device heterogeneity was categorized into six major classes: Node Selection, Weighted Aggregation, Gradient Compression,

Semi-Asynchronous, Cluster FL (Federated Learning), and Model Split.

Among these classes, we specifically focus on Node Selection. In traditional federated learning, all nodes contribute equally to the global model. However, in asynchronous federated learning, to avoid overfitting, we prefer selecting nodes with robustness and better computational capabilities. In [2], the authors proposed a heuristic greedy node selection strategy based on the local computing and communication resources of heterogeneous nodes. This strategy iteratively selects nodes with higher contributions to participate in global model aggregation.

III. Asynchronous Federated Learning Method

This paper presents the implementation of an asynchronous federated learning algorithm using Nvidia Flare and conducts related data experiments. Building upon [4], an examination of the functionality of the NVFlare federated learning framework is performed, with a focus on comparing it with the FML system view presented in Chapter 7 of [4]. The goal is to verify whether the NVFlare framework possesses a robust architecture for practical deployment in future 6G multimedia transmission systems.

The main focus of this paper is the implementation of asynchronous federated learning on nodes. The server-side is referred to as the "Server," while the nodes participating in the federated learning training are called "Nodes." The algorithm training process is as follows:

1. Weight Initialization: The Server initializes the Global model's weights and sends the Global model to all participating clients. Upon receiving the Global model, each client immediately starts training.
2. Node Selection: After a certain waiting period, the Server calculates the contribution of each client. The contribution includes two indicators: computational contribution and network stability. The computational contribution is calculated based on the current epoch and mini-batch size of each node, while network stability is evaluated using ACK packets. The Server receives the data returned by the clients, calculates their contributions, and sorts them in descending order. The clients with the highest contributions are selected until the Server's predetermined time is reached.
3. Model Update: The Server collects the models of the nodes selected in Step 2 and performs FedAvg aggregation. Nodes that were not selected continue training. The resulting aggregated model is sent back to the nodes as the updated Global model. The contribution scores of the nodes participating in this round of training are reset to zero, while the scores of the non-selected nodes continue accumulating. This process is repeated until the Global model reaches the desired accuracy level.

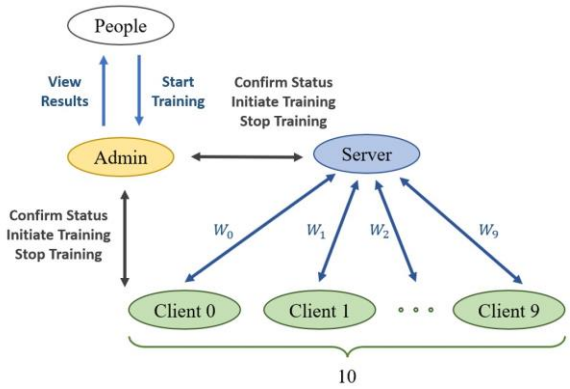


Fig. 1 Asynchronous Federated Learning Architecture.

The architecture is shown in Figure 1, and the experimental environment is set up as a POC single-machine simulation using Ubuntu Server 20.05LTS. In this study, the implementation will follow the aforementioned steps, with the network state evaluation omitted during the implementation phase. The selection of clients will be based solely on computing contributions. Figure 2 illustrates the process of the paper's implementation using NVFlare, and Figure 3 provides an explanation of the node selection process. The detailed steps of the Flare implementation are as follows:

1. The server creates a task.
2. The client requests the task.
3. The server applies the task.
4. The server waits for the client to train for 30 seconds. This time allows each client to perform its own training. In our simulation, we randomly introduce sleep functions in different clients to simulate varying computational capabilities.
5. The server uses the fire_fed_event API in Flare to send a request to all clients to stop training and upload local epoch and batch size.
6. The client uses the handle_event() API in Flare to receive the fire_fed_event and perform the corresponding actions, such as stopping training and sending parameters.
7. The client contributions are calculated, sorted, and stored as a list. The top two-thirds of the list are selected

as the chosen client nodes. For detailed steps of node selection, please refer to the next subsection.

8. The server uses the Broadcast() function to terminate the selected clients, retrieve their models, and perform model aggregation. After completion, the server broadcasts again to continue training.
9. Return to step 4 and repeat until the training completion threshold is reached.

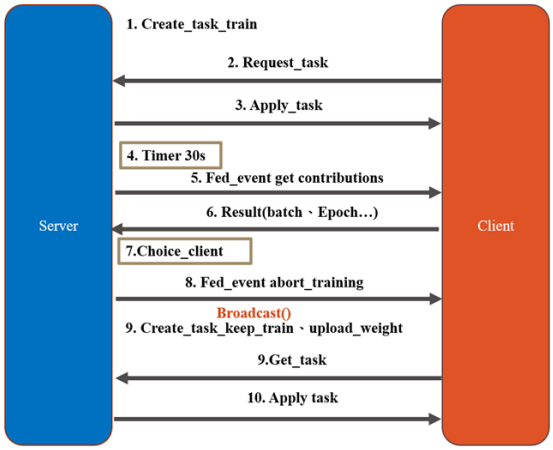


Fig. 2 Implementation Flowchart.

The detailed steps for node selection are as follows:

1. The server sets two timeouts (one for contribution and another for weights) and starts training.
2. After a certain period of local training, the client sends the local epoch count to the server for calculating the contribution. Then, it continues training. Different local epoch counts are implemented by setting different sleep times.
3. The server calculates the contribution λ of the clients received within the $timeout_{\lambda}$ period (calculated based on the local epoch count to simulate varying client computational performance). Only the clients with contributions in the top 2/3 are added to the selected list.
4. The server broadcasts a request for the selected clients to return their weights, denoted as W.
5. The server receives the W returned by the clients within the $timeout_{W}$ period.
6. The server aggregates the received W from the clients.
7. After aggregation, the server returns the updated weights, denoted as W_{new}, to the clients, completing one global round.

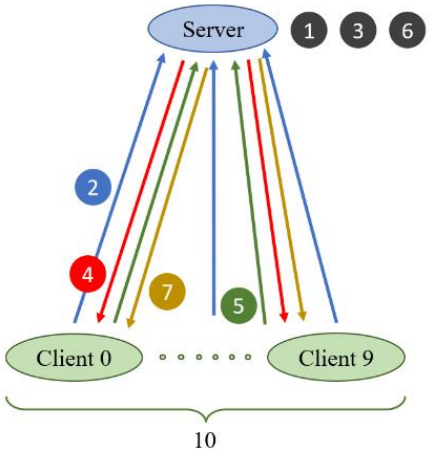


Fig. 3 Node Selection Flowchart.

To validate the approach mentioned in the reference paper, we first implemented the method described in the paper and used the provided algorithms along with a dataset for verification. For this purpose, we used the commonly used image dataset CIFAR10. CIFAR10 is a color image dataset consisting of 10 classes, with a total of 60,000 images. Among these, 50,000 images are used for training and 10,000 images are used for testing. Each image has a size of 32x32.

As described in [3], the non-IID data distribution in federated learning can be categorized into several types:

1. Non-IID on feature distribution (feature distribution skew): In domains like handwritten recognition, different individuals may have variations in stroke width, slant, etc., even when writing the same digit.
2. Non-IID on label distribution (label distribution skew): When clients are limited to specific geographic regions, the distribution of labels can vary among clients. For example, kangaroo images may only be present in clients from Australian zoos.
3. Non-IID with the same label but different feature distribution (same label, different feature): Even for the same label, the features can vary significantly across different time scales, such as day and night, seasonal variations, natural disasters, fashion trends, design trends, etc.
4. Non-IID with the same feature but different label distribution (same feature, different label): For instance, labels that reflect emotions or predict the next word can differ based on individual or regional differences.

5. Non-IID or data imbalance in quantity distribution (quantity skew or unbalance): Different clients may possess significantly varying amounts of data.

Based on the definitions of non-IID mentioned above, we chose two different types of non-IID for experimentation: non-IID in data quantity and non-IID in data labels. For non-IID in data quantity, we followed the methods described in most papers and used the Dirichlet distribution to generate the non-IID data. We adjusted the degree of non-IID using the parameter α, where larger values (>10000) indicate a more IID distribution, while smaller values (<1) represent a more non-IID distribution.

IV. Experimental Results

Parameter Settings and Explanations:

- Local training time: This refers to the duration for which each client waits to perform training in each global round.
- delay_big: It represents the waiting time for the client (simulating computational power disparity) and is directly proportional to the data size.
- delay_small: This indicates the waiting time for the client (simulating computational power disparity) and is inversely proportional to the data size.
- Dataset: CIFAR10

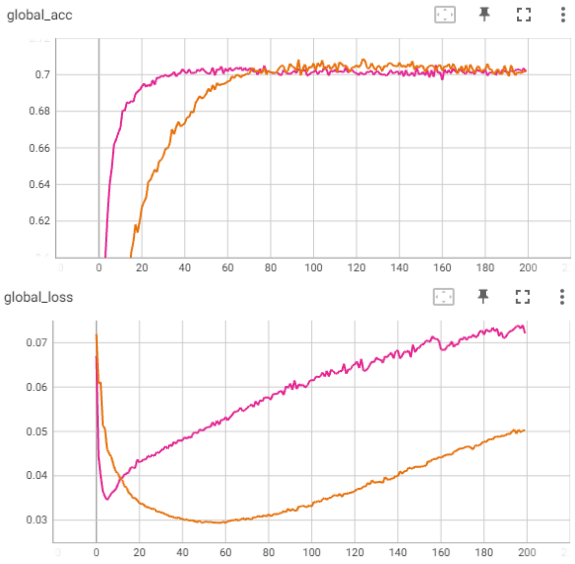


Fig. 4 Distribution: IID, local training time=5s, different local learning rate comparison..

Figures 4-7 present experimental comparisons under the IID data distribution.

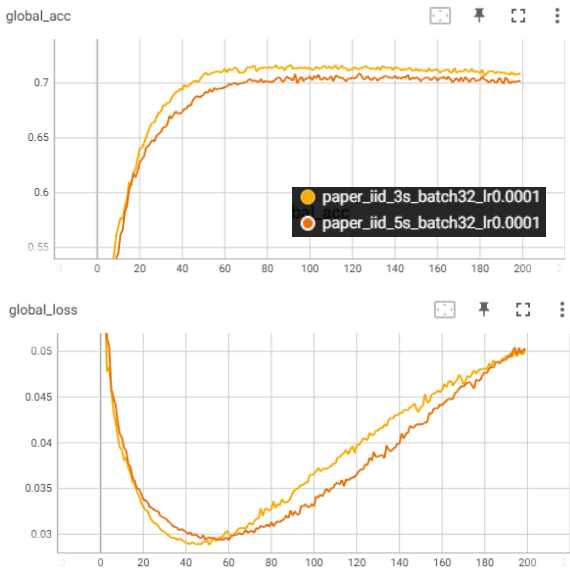


Fig. 5 IID, local learning rate = 0.0001, comparison of different local training times.

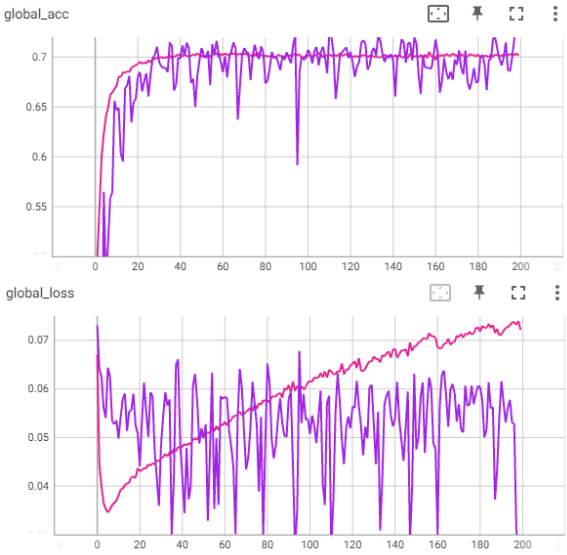


Fig. 6 IID, aggregation of the entire model approach compared to sending only model gradient.

When training machine learning models, convergence speed and accuracy are two crucial metrics. A faster convergence indicates that the model learns features and patterns more quickly, while higher accuracy implies better generalization ability. Therefore, it is essential to study and optimize these metrics.

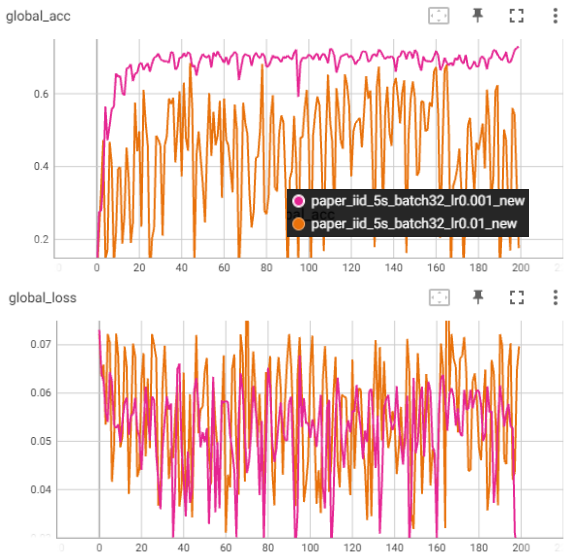


Fig. 7 IID, only send model gradient with different learning rate.

In Figure 4, we compare the performance of different learning rates under the same local training time. The results show that higher learning rates lead to faster convergence. However, when the learning rate is set to 0.001, the model exhibits severe overfitting. Overfitting occurs when the model performs well on the training data but poorly on the test data. Therefore, careful selection of the learning rate is necessary to avoid overfitting. Figure 5 compares the impact of different local training times on model performance. The results show that there is little difference between the two different local training times, and a longer local training time does not necessarily mean better performance. Therefore, we need to strike a balance between time and accuracy when choosing the appropriate local training time. In addition to the above factors such as hyperparameters and training time, we also compared different model aggregation methods. The conventional federated learning aggregation method involves returning the entire model and aggregating it. However, we explored an alternative approach of only returning the difference between the client's post-training model and pre-training model (i.e., sending only the model gradients). We aimed to reduce the global model discrepancy between different rounds using this approach.

From the experiments shown in Figures 6 and 7, we observe that under the IID data distribution, the approach of sending only the model gradients achieves similar accuracy compared to aggregating the entire model. This may be because, under the IID data distribution, the training data distributions across different clients are similar, resulting in smaller model discrepancies. However, we also found that the approach of sending only the model gradients leads to more significant fluctuations in model loss values. Therefore, the learning rate should not be too large, as it can cause excessive fluctuations during the training process, resulting in suboptimal performance.

Figures 8-11 depict the experimental results under the Non-IID (α=0.1) data distribution. Based on the results shown in Figure 8, we compare the influence of different local learning rates on model convergence under the same local training time. The results indicate that, compared to the IID data distribution, the Non-IID data distribution exhibits higher accuracy and more significant loss oscillations. This may be attributed to the unevenness of the Non-IID data distribution, which affects the global model aggregation across different rounds. Additionally, Figure 8 reveals that larger learning rates seem to facilitate model convergence under the Non-IID data distribution.

In the experiment depicted in Figure 9, we compare different model aggregation weight aggregation methods. One approach involves using the client's data volume as weights for weighted averaging of the models, while the other approach directly averages all received models. Typically, we expect that clients with larger data volumes will have better models, so using data volume weighting should yield higher accuracy. However, in the experiment shown in Figure 9, we find that using data volume weighting does not necessarily improve accuracy. On the contrary, directly averaging all received models appears to perform better.

In the experiment depicted in Figure 10, we compare the impact of different client computational capabilities on the convergence of federated learning. We simulate the performance differences among clients by making them wait for a certain period of time before starting training. In this experiment, we set three different waiting times: delay_big, delay_small, and no_suffix. Here, delay_big represents waiting time proportional to the client's data volume, delay_small represents waiting time inversely proportional to the client's data volume, and no_suffix represents all clients having the same waiting time.

From the experimental results, we can observe that delay_small leads to better accuracy. This suggests that when clients with larger data volumes upload model updates more frequently, the convergence of federated learning improves. Therefore, we can infer that allowing clients with larger data volumes to upload model updates more frequently can yield better results in federated learning. This finding will serve as a basis for our subsequent experiments.

In the experiment shown in Figure 11, we compare the aggregation methods of directly aggregating the entire model and aggregating only the gradients of the client models in the case of Non-IID data distribution. In this experiment, we find that aggregating the entire model achieves better accuracy under the Non-IID data distribution, rather than only transmitting the differences between the client models before and after training. This is different from the situation in the IID case and should be taken into account in practical applications.

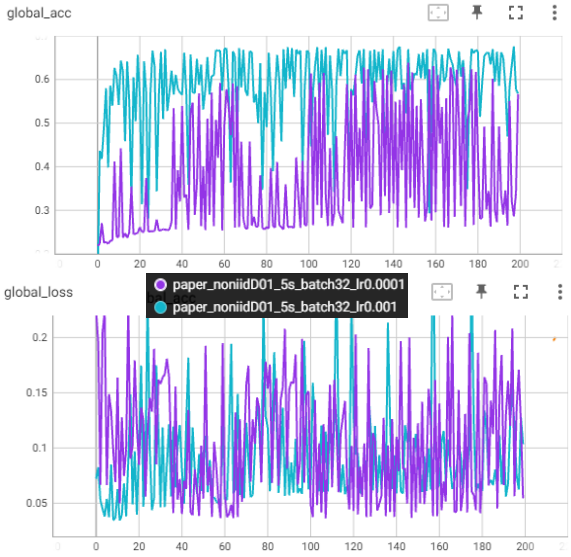


Fig. 8 NonIID, local training time=5s, different local learning rate comparison.

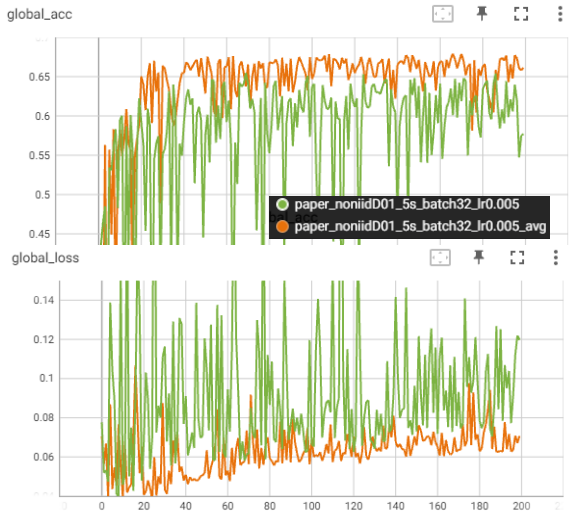


Fig. 9 Non-IID, local training time=5s, compare the aggregation results of different model aggregation weights.

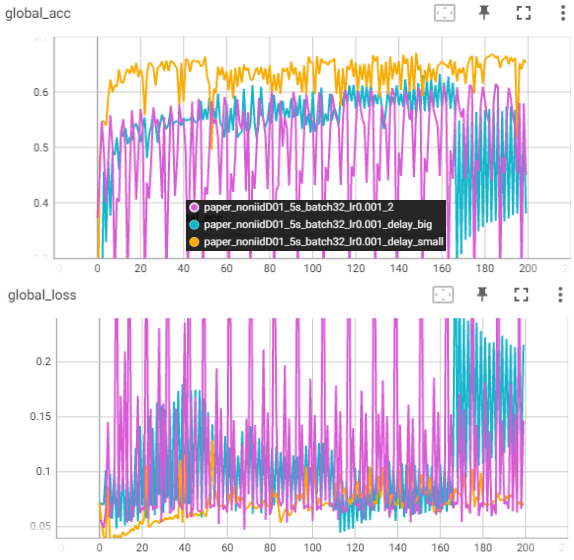


Fig. 10 Comparison of the waiting time per turn for different clients (simulating the difference in performance of different clients).

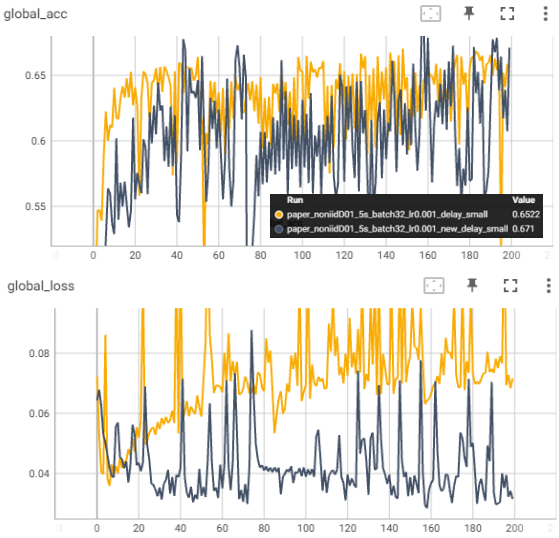


Fig. 11 Compare the performance of the direct aggregation model with that of the client-side gradient only.

V. CONCLUSIONS

In the case of asynchronous and Non-IID data distribution, we observed that the setting of the learning rate is crucial. Significant improvements were achieved by adjusting the learning rate to address the issues encountered. Both excessively large and excessively small learning rates yielded suboptimal results. Future research could explore the development of an adaptive learning rate tuner to achieve more stable training.

After fine-tuning the parameters over a period of time, we found that under well-adjusted settings, the asynchronous algorithm can achieve similar accuracy to the synchronous algorithm. This suggests that the proposed method in the paper can be effective in the context of Massive IoT and may be a promising area for future research.

From the aforementioned experimental results, it was observed that regardless of the IID or Non-IID data distribution, aggregating the entire model yields a better global model compared to aggregating only the client-side gradients. Additionally, in the investigation of using data volume as weights for model aggregation, there was no significant difference observed between the two aggregation methods in the IID data distribution. However, in the case of Non-IID data distribution, direct averaging of models can mitigate model discrepancies between different global rounds, resulting in a smoother training process.

REFERENCES

[1] C. Xu, et al. (2023) "Asynchronous federated learning on Heterogeneous Devices: A survey, " arXiv.org. Available at: https://arxiv.org/abs/2109.04269 (Accessed: April 11, 2023).

[2] Z. Chen, et al. (2021) "Towards asynchronous federated learning for heterogeneous edge-powered internet of things," Digital Communications and Networks, 7(3), pp. 317–326. Available at: https://doi.org/10.1016/j.dcan.2021.04.001.

[3] P. Kairouz et al., "Advances and open problems in federated learning," 2019. [Online]. Available: arXiv:1912.04977.

[4] "IEEE Guide for Architectural Framework and Application of Federated Machine Learning," in IEEE Std 3652.1-2020, vol., no., pp.1-69, 19 March 2021, doi: 10.1109/IEEESTD.2021.9382202.

[5] W. Wu, L. He, W. Lin, R. Mao, C. Maple and S. Jarvis, "SAFA: A Semi-Asynchronous Protocol for Fast Federated Learning With Low Overhead," in IEEE Transactions on Computers, vol. 70, no. 5, pp. 655-668, 1 May 2021, doi: 10.1109/TC.2020.2994391.

[6] J. Konečný, et al. (2017) "Federated learning: Strategies for improving communication efficiency," arXiv.org. Available at: https://arxiv.org/abs/1610.05492 (Accessed: April 11, 2023).

[7] W. Y. B. Lim et al., "Federated Learning in Mobile Edge Networks: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 2031-2063, thirdquarter 2020, doi: 10.1109/COMST.2020.2986024.

[8] J. Konečný, et al. (2016) "Federated Optimization: Distributed Machine Learning for On-Device Intelligence," arXiv.org. Available at: https://arxiv.org/abs/1610.02527 (Accessed: April 11, 2023).

[9] X. Yin, Y. Zhu and J. Hu, (2022) "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and Future Directions, " ACM Computing Surveys. Available at: https://dl.acm.org/doi/10.1145/3460427 (Accessed: April 11, 2023).

[10] O. A. Wahab, A. Mourad, H. Otrok and T. Taleb, "Federated Machine Learning: Survey, Multi-Level Classification, Desirable Criteria and Future Directions in Communication and Networking Systems, " in IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 1342-1397, Secondquarter 2021, doi: 10.1109/COMST.2021.3058573.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

The Development of an AI-assisted Diagnosis System for Adult Glioma Subtyping Prediction

Wei-Wen Hsu[1], Jia-Yi Lin[2], Hsin-Hung Lai[2], Wan-Lin Hsu[1], Jeng-Ting Jiang[1], Yao-Chung Chang[1], Yao-Feng Li[*,2]

[1] Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan
[2] Department of Pathology, Tri-Service General Hospital and National Defense Medical Center, Taipei, Taiwan
[*]E-mail: liyaofeng1109@gmail.com

***Abstract*— Primary brain tumors are among the ten most common causes of cancer-related death, and gliomas are the most prevalent type of adult brain tumor, accounting for 78% of malignant brain tumors. Accurate glioma subtype classification is critical to treating patients with brain tumors; however, developing an automated computer-aided diagnosis system for glioma is not trivial. This study introduces the development of an AI-assisted diagnosis system for adult glioma. There are four subtypes of glioma targeted, including Oligodendroglioma (O), Anaplastic Oligodendroglioma (AO), Astrocytoma (A), and Glioblastoma (G). Instead of directly classifying all whole-slide images into four categories, the hierarchical framework with a multi-view scheme was proposed. That is, two-stage predictions were performed. The first stage is to classify all slides into the Oligo-like and Astro-like classes by observing cells' morphologies with high magnification (x40) views by the patching approach. Subsequently, the second stage is to quantify the areas of normal vessels, hyperplasia vessels (vascular proliferation), and necrosis regions with views of lower magnification (x5, x10) by YOLO segmentation, predicting grading levels for each type. After false detection reduction, the experimental results show the quantification of vessels and necrosis regions are the valuable features to indicate grading levels, and the performance of the proposed two-stage hierarchical framework in the final 4-class predictions surpassed the approach of classifying all slides into 4 categories directly.**

Keywords*— *Glioma Diagnosis, AI-assisted Diagnosis System, Deep Learning Approaches.

I. INTRODUCTION

Adult brain tumors represent a widespread cancer classification, contributing to substantial morbidity and mortality worldwide. Data from the Central Brain Tumor Registry of the United States (CBTRUS) demonstrates that, upon histological examination, diffuse gliomas account for 68.5% of primary brain malignancies, with glioblastoma being the most fatal subtype [1]. Although advances have been made in precision diagnostics through the use of molecular data [2] and methylation status [3], histopathology continues to serve as an essential starting point. It helps to ensure more accurate precision diagnostics by reducing errors that may occur when relying solely on molecular testing methods. Consequently, our objective in this study is to establish an artificial intelligence-supported diagnostic system for the prediction of glioma histologic subtypes, such as oligodendroglioma, anaplastic oligodendroglioma, astrocytoma, and glioblastoma.

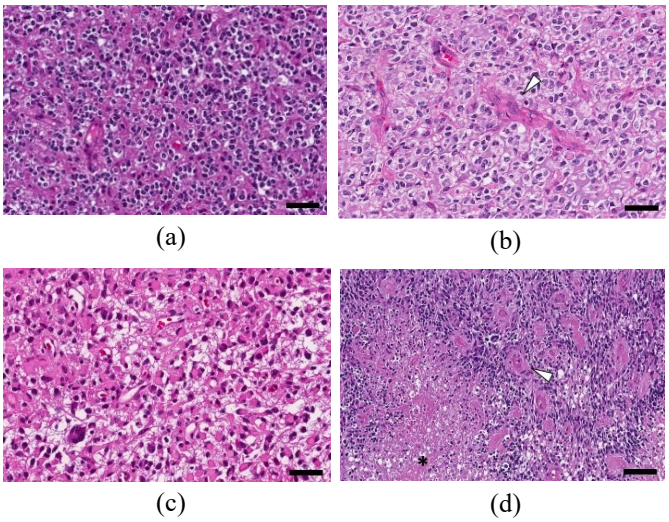


Fig. 1. The major histological subtypes of glioma found in adults encompass. (a) Oligodendroglioma, (b) Anaplastic Oligodendroglioma, (c) Astrocytoma, and (d) Glioblastoma. The arrowhead denotes endothelial proliferation, whereas the asterisk symbolizes tumor necrosis. The scale is 50 μm.

From the pathological perspective, the morphologies of glioma subtypes are somehow distinct. As can be observed in Figure 1-(a) and 1-(b), for oligodendroglioma and anaplastic oligodendroglioma, cells appear round shape nuclei with open chromatin and artifactual cytoplasmic retraction, leading to the "fried egg" appearance. On the other hand, astrocytoma and glioblastoma revealed hypercellularity with irregular and hyperchromatic nuclei in the fibrillary background, as shown in Fig. 1-(c) and 1-(d). To discriminate anaplastic oligodendroglioma from oligodendroglioma or distinguish glioblastoma from astrocytoma, the histopathologic features, like glomerular endothelial proliferation (pointed by a white arrow in Fig. 1-(c)) and tumor necrosis (the area marked with a star-shaped sign * in Fig. 1-(d)), are helpful in subtype classification. Then, the quantification of these histopathologic features can be achieved by AI technologies.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. MATERIALS AND METHODS

In this study, a total of 646 cases of whole-slide images (WSIs) with adult glioma were collected from the database of The Cancer Genome Altas (TCGA) program [4]. There are 4 subtypes of adult glioma targeted in our experiment, which are oligodendroglioma (denoted as **O**), Anaplastic Oligodendroglioma (indicated as **AO**), Astrocytoma (shown as **A**), and Glioblastoma (denoted as **G**). Then, all WSI cases in each glioma subtype were randomly split into the training dataset and testing dataset by the ratio of 7:3. And the numbers of WSI cases for training and testing in each glioma subtype are listed in Table I.

Table I. Numbers of WSI cases in each glioma subtype's training and testing datasets.

Glioma Subtype	# training cases	# testing cases	# total cases
Oligodendroglioma (**O**)	68	29	97
Anaplastic Oligodendroglioma (**AO**)	52	22	74
Astrocytoma (**A**)	171	73	244
Glioblastoma (**G**)	162	69	231

To deal with gigapixel whole-slide tissue images, the patching approach [5, 6] that collects several patch samples from the tissue areas at a specific magnification seems easier to implement. Figure 2 shows the baseline implementation to achieve 4-subtype prediction. For each category, hundreds of patches were randomly sampled from the tissue areas in each slide at a magnification of x40, collecting training samples to learn the features of cells' morphologies. Then, for slide inference, about the same number of sampling patches were collected from the testing slide at the same magnification rate and fed into the trained model for patch predictions. The majority vote scheme from the projections of those sampling patches can determine the final predicting subtype for the testing slide.

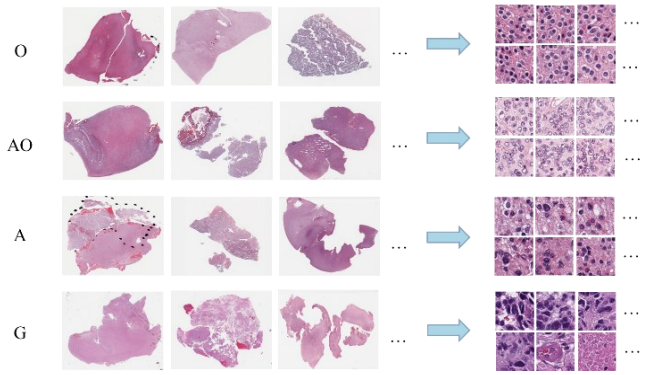


Fig. 2. The baseline implementation using the patching approach for 4-subtype prediction.

However, from the pathological perspective, features of cells' morphologies in the sampling patches at high magnification may be inadequate to evaluate the severity of cases in the Oligo-like class (including **O** and **AO**) or the Astro-like class (including **A** and **G**). Clinically, the existences of vascular proliferation (patches sampled at x10) and necrosis areas (patches sampled at x5) are crucial evidence to distinguish **AO** from **O** in the Oligo-like class or discriminate **G** from **A** in the Astro-like class. Therefore, the two-stage hierarchical framework with a multi-view scheme was proposed in this study to achieve glioma subtyping prediction, as shown in Figure 3.

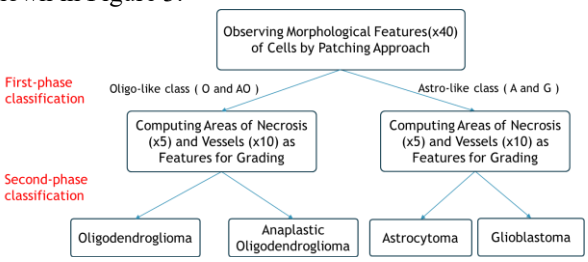
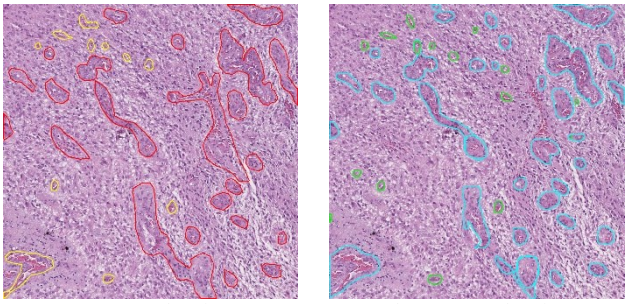


Fig. 3. The proposed framework for glioma subtyping prediction.

Instead of performing 4-subtype classification directly in the baseline approach, in the proposed framework, the first phase is to conduct binary classification, which categorizes all slides into Oligo-like class and Astro-like class based on cells' morphologies, using the model of ResNet50 [7]. It can be easily implemented for training data acquisition by merging the training samples of **O** and **AO** to become the training set for the Oligo-like class and merging the samples of **A** and **G** for the Astro-like class. In addition, optical density transformation [8] was applied on the sampling patches to reduce the H&E staining gaps for slides from different institutions for better regularization.

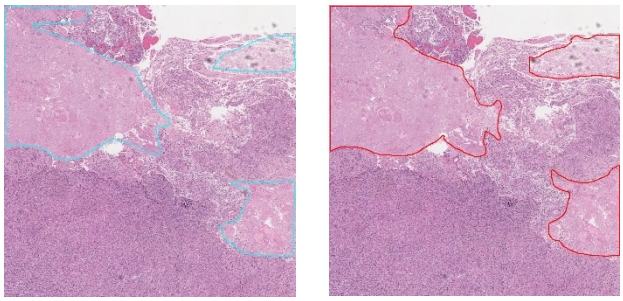
For the second phase of the proposed framework, normal vessels, vascular proliferation regions, and necrosis lesions were detected and segmented automatically by the deep learning model of YOLOv8 [8]. Patches were sampled at a magnification of x10 from the slides in the training dataset and annotated with the regions of normal vessels (in yellow) and vascular proliferation (in red), as shown in Figure 4-(a). The segmented results of normal vessels (in green) and regions of vascular proliferation (in cyan) by the YOLOv8 model are shown in Fig. 4-(b).



(a) annotated regions (b) segmented regions

Fig. 4. A patch with regions of normal vessels and vascular proliferation.

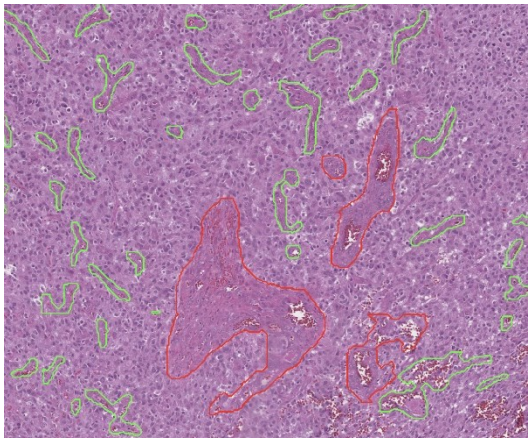
Similarly, for the segmentation of necrosis lesions, as shown in Figure 5-(a), patches were sampled from the slides at a magnification of x5 and annotated with the necrosis lesions (in cyan) to train the segmenting model. Fig. 5-(b) shows the segmented results of necrosis lesions (in red) by the trained model of YOLOv8.



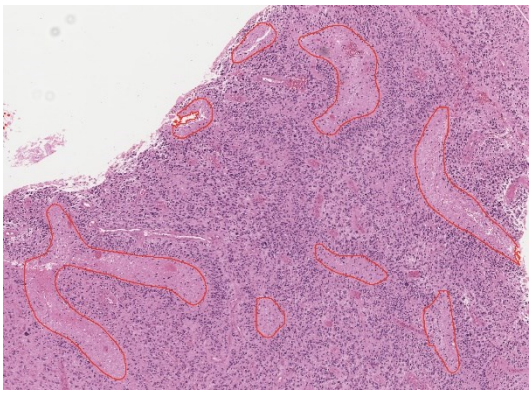
(a) annotated regions (b) segmented regions

Fig. 5. A patch with regions of necrosis lesions.

Subsequently, the segmentation models for vessel regions and necrosis lesions were applied on the WSIs to scan each tile to determine if there exists vascular proliferation or necrosis lesions for the evaluation of cases' severity in diagnosis. The segmentation results for different kinds of vessel regions and necrosis lesions on the WSIs are shown in Figures 6-(a) and 6-(b), respectively.



(a) Segmented normal vessels (in green) and regions of vascular proliferation (in red) on a WSI.



(b) Segmented necrosis lesions (in red) on a WSI.

Fig. 6. The segmentation results on a WSI.

However, to obtain a perfect segmentation model that runs on hundreds of tiles over a WSI without any false positive regions of vascular proliferation or necrosis lesions is almost impossible. Moreover, the area ratio between normal vessels and regions of vascular proliferation is statistically substantial to the evaluation of severity for a case [2]. Therefore, the areas of the segmented regions of normal vessels, vascular proliferation, and necrosis lesions were computed as the features to train the logistic regression models for classification, performing 4-subtype prediction in the second phase of the proposed framework.

III. EXPERIMENTAL RESULTS

For the baseline approach that directly predicts all cases into 4 subtypes, its corresponding confusion matrix of the final prediction results is listed in Table II, and the predicting accuracy is 66.84% (129/193).

Table II. The confusion matrix of the predicting results by the baseline approach.

GT \ Prediction	O	AO	A	G
O	**3**	15	8	3
AO	2	**15**	3	2
A	5	3	**56**	9
G	0	9	5	**55**

On the other hand, for the proposed framework, the confusion matrix of the binary classification results in the first phase, in which all cases were categorized into Oligo-like class and Astro-like class, is listed in Table III. And the accuracy is about 82.9%. (160/193)

Table III. The confusion matrix of the binary classification results in the first phase of the proposed framework, using the model of ResNet50.

GT \ Prediction	Oligo-like class (O and AO)	Astro-like class (A and G)
Oligo-like class (O and AO)	**32**	19
Astro-like class (A and G)	14	**128**

Afterward, in the second phase of the proposed framework, the performance metrics in the accuracy of the logistic regression models are 98.04% (50/51) for O vs. AO and 91.55% (130/142) for A vs. G. The confusion matrices are listed in Table IV and V, respectively. As a result, the final 4-subtype prediction can be achieved by applying the patch approach using the ResNet50 model for binary classification in the first phase and quantifying the segmented areas of vessels regions and necrosis lesions as the features for further subtype

classification using the logistic regression classifiers. The confusion matrix of the final predicting results by the proposed methods is listed in Table VI, and the predicting accuracy reached 76.17% (147/193), surpassing the baseline approach's performance with 66.84% (129/193).

Table IV. The confusion matrix of the binary classification results in the second phase of the proposed framework for cases in the **Oligo-like class**, using the logistic regression classifier.

GT \ Prediction	O	AO
O	**29**	0
AO	1	**21**

Table V. The confusion matrix of the binary classification results in the second phase of the proposed framework for cases in the **Astro-like class**, using the logistic regression classifier.

GT \ Prediction	A	G
A	**71**	2
G	10	**59**

Table VI. The confusion matrix of the predicting results by the proposed framework.

GT \ Prediction	O	AO	A	G
O	**18**	0	11	0
AO	1	**13**	0	8
A	10	1	**60**	2
G	0	3	10	**56**

IV. CONCLUSIONS

This study proposed a hierarchical framework with a multi-view scheme to achieve adult glioma subtyping prediction. The experimental results implied that the areas of normal vessels, regions of vascular proliferation, and necrosis lesions computed from each slide are useful features to judge the case's severity for further subtype classification. In addition, since the over-detection of vascular proliferation and necrosis lesions is unavoidable, the classification models of logistic regression classifier trained in the second phase can alleviate the impact of those regions of false positive detection, reducing the misinterpretation cases successfully. However, a case will be misinterpreted once misclassified in the first phase. Therefore, the performance bottleneck of the proposed framework to tackle in the future is to boost the performance of binary classification for the Oligo-like class vs. Astro-like class in the first phase.

REFERENCES

[1] Q. T. Ostrom *et al.*, "CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016," *Neuro-oncology,* vol. 21, no. Supplement_5, pp. v1-v100, 2019.

[2] D. N. Louis *et al.*, "The 2021 WHO classification of tumors of the central nervous system: a summary," *Neuro-oncology,* vol. 23, no. 8, pp. 1231-1251, 2021.

[3] D. Capper *et al.*, "DNA methylation-based classification of central nervous system tumours," *Nature,* vol. 555, no. 7697, pp. 469-474, 2018.

[4] J. N. Weinstein *et al.*, "Cancer Genome Atlas Research Network," *Nat Genet,* vol. 45, no. 10, pp. 1113-1120, 2013.

[5] Y. Liu *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442,* 2017.

[6] W.-W. Hsu *et al.*, "A weakly supervised deep learning-based method for glioma subtype classification using WSI and mpMRIs," *Scientific Reports,* vol. 12, no. 1, p. 6111, 2022.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[8] G. Jocher, A. Chaurasia, and J. Qiu. "YOLO by Ultralytics." https://github.com/ultralytics/ (accessed.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Poisoning Attacks against Gait-based Identity Recognition

Jianmin Dong*, Da-tian Peng†, Guanxiong Pei*, and Taihao Li*‡
* Department of Artificial Intelligence at Zhejiang Lab, Hangzhou, Zhejiang 311121, China
E-mail: dongjm@zhejianglab.edu.cn, pgx@zhejianglab.com, lith@zhejianglab.com
† School of Cybersecurity, Northwestern Polytechnical University, Xi'an, Shaanxi, China
E-mail: pengdatian@nwpu.edu.cn
‡ Corresponding author

***Abstract*—Investigating the endogenous vulnerabilities of gait-based identity recognition (GBIR) is extremely important for protecting the privacy information of users stored in smart devices. In this paper, we first elaborate the poisoning attacks against the GBIR to evaluate its robustness, in which the model training process can be deliberately corrupted to manipulate its identity prediction intentionally, e.g., the illegitimate user *Alice* is misidentified as the targeted legitimate user *Bob*. To this end, we formulate a bilevel optimization problem to characterize the sequential decision-making interaction process between the attacker and the gait recognizer that is formalized by a sparse representation optimization. Further, we reformulate this NP-hard bilevel problem into a computable mixed integer program to purse a maximum-effectiveness and minimum-cost attack strategy. Finally, we perform experiments on the single- and multi-source datasets to verify the validity of our poisoning attacks in inducing the identity misrecognition.**

I. Introduction

Currently, a lot of sensitive and commercial-in-confidence data including chat records, shopping information, medical records, bank flow and so on are recorded by various smart devices [1], e.g., smartphones and wearable devices. To avoid the owners' privacy information leakage, it is critical to develop the built-in identity recognition module for implementing the identity authentication and security assurance. Due to the widespread popularity of smart devices, it is thoroughly easy to collect the user's gait data by the embedded motion sensors such as accelerometer and gyroscope. Consequently, the emerging gait-based identity recognition (GBIR) has attracted much research attention [2]–[4]. Compared with some traditional biometric information, i.e., face [5], fingerprint [6], iris [7], and voice [8], the gait has the unique advantages of that is, it can be authenticated implicitly and untouchably by more high accuracy and it is difficult to imitate explicitly.

Actually, the GBIR can be treated as a machine-learning classification task, in which the motion pattern of each individual is distinguishable and its belonging identity is recognizable intuitively. However, extensive researches [9]–[12] indicate that the GBIR is vulnerable to these adversarial impersonation attacks by the intentional impostors who can mimic the walking style of target users and even resort to the extra assistance with the similar physical shape and the speed-controlled treadmill. Furthermore, there exist a major challenge for these existing attacks to investigate the endogenous vulnerabilities of GBIR, because these strategies are heuristic in an evasion manner where only external testing samples are perturbed and the internal trained model is undisturbed, not only lacking of the strong attack effectiveness, but also taking the high attack cost. Therefore, these previous studies are limited to inspire the key secure insights to protect the user's privacy information.

To evaluate the robustness of GBIR deeply and understand its endogenous vulnerabilities well, we are motivated to elaborate the effective poisoning attacks that can cover the limitations of these aforementioned attacks. Main contributions are included as follows.

(i) We from the attacker's perspective explore the poisoning feasibility via deliberately corrupting the model training process of GBIR and intentionally manipulate its identity prediction, resulting that the illegitimate user *Alice* is misidentified as the specific legitimate user *Bob*.
(ii) We treat such a poisoning behavior as the leader-follower game between attacker and recognizer and then formulate a bilevel optimization problem to model their sequential decision-making interactions and seek for the final Nash equilibrium.
(iii) We run extensive numerical experiments on the single- and multi-source datasets to validate our proposed maximum-effectiveness and minimum-cost poisoning attacks in inducing the identity misrecognition.

II. Related Work

The brief reviews of related work are summarized below from two aspects: the advancements of existing impersonation attacks against the GBIR [9]–[12] and the applications of novel poisoning attacks in the artificial intelligent security domain [13]–[16].

1) Impersonation Attacks: The implementation strategies of impersonation attacks have evolved over the past decade [9]–[12]. Gafurov et al. [9] investigated the security of gait under two types of deliberate spoofing scenarios: friendly and hostile. Under the friendly scenario, the imitators did not make any deliberate attempt to copy the genuine user, whereas under the hostile scenario, individuals with similar physical characteristics attempted to imitate each other. The authors concluded that the imitator with similar physical characteristics or of the same gender might have a better chance of getting

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

accepted. In a follow-up, many works in Ref. [10]–[12] used sensors embedded in smart devices to capture gait acceleration and achieved gait imitation. The results are not acceptable, the reason may be that gait mimicry is a difficult task in practice and that a meticulous impersonation strategy hardly increases the chances of success of an impostor, even though they were closely physiologically paired with the victims and given live audio and visual feedback.

2) Poisoning Attacks: Besides the works mentioned earlier, poisoning attack [13]–[16] is another kind of adversarial attacks, which poisons the training templates to induce a wrongly trained model. Luis et. al [14] extended data poisoning attack approach to multi-class DNN, in which a fraction of training data are manipulated to subvert learning process. Zhu et al. [15] presented a transferable clean-label poisoning attack approach that can attack image classification system successfully without any knowledge of victim DNN via designing poison images to surround the targeted image in feature space. Huang et al. [16] devised a data poisoning attack approach that can manipulate the DNN based recommendation system to recommend attacker-chosen target items to as many customers as possible. The poisoning attack strategy can threaten the security of pattern recognition models seriously and it inspires us to absorb into our optimal attack strategy.

Till now, there is few studies considerd adversarial attacks on GBIR system. In this paper, from the perspective of attackers, we propose a poisoning attack strategy to explore the endogenous vulnerabilities in existing GBIR systems.

III. Problem Formulation

In this section, we first introduce the sparse representation optimization on the GBIR. Then, we describe the intentions of poisoning attacks and formalize its bilevel optimization model. Finally, we give the problem solving method.

A. Sparse Representation Optimization on GBIR

In general, the distinctive gait pattern can be recognized to identify its owner. Thus, the GBIR is treated as a classification task, which can be chracterized by sparse representation optimization. This includes two parts.

1) Dictionary Construction: For all c users (i.e., classes), we use their d training samples to construct a dictionary matrix D as follows.

$$D = [D_1, \cdots, D_k, \cdots, D_c] \in \mathcal{R}^{K \times d} \tag{1}$$

where $D_k = [\boldsymbol{d}_{k1}, \cdots, \boldsymbol{d}_{k\overline{m}_k}]$ is constructed by the training samples of k^{th} user, $k = 1, 2, \cdots, c$, and $\overline{m}_k$ is total number of training samples of k^{th} user, implying that $(m_k - \overline{m}_k)$ is total number of testing samples of k^{th} user, $\sum_{k=1}^{c} \overline{m}_k = d$. Often, $K \ll d$ is known.

2) L_1 Minimization: Suppose $\boldsymbol{y}$ be one of testing samples of k^{th} user. Due to the fact that both $\boldsymbol{y}$ and D_k are from the same user with the identical-distribution samples, they have high positive correlations. So, we can use a linear combination to describe the correlations between $\boldsymbol{y}$ and D_k as follows.

$$\begin{aligned} \boldsymbol{y} &= D_i \boldsymbol{x}_i \\ &= [D_1, \cdots, D_i, \cdots, D_c][\mathbf{0}_1^{\mathrm{T}}, \cdots, \boldsymbol{x}_i^{\mathrm{T}}, \cdots, \mathbf{0}_c^{\mathrm{T}}]^{\mathrm{T}} \\ &= D\boldsymbol{x} \end{aligned} \tag{2}$$

where $\boldsymbol{x} \in \mathcal{R}^d$ is a sparse vector with the maximum sparsity $\overline{m}_i$ and its non-zero elements is the coefficient $\boldsymbol{x}_i$.

Due to $K \ll d$, this is a under-determined equation system so that it exists infinite solutions. Only the sparsest solution is desired to locate the unique class of user identity. To model this sparse representation, we adopt the L_1 minimization optimization as follows.

$$\min ||\boldsymbol{x}||_1 \tag{3}$$

$$\text{s.t. } \mathbf{0} \preceq \boldsymbol{x} \preceq \mathbf{1} \tag{4}$$

$$||\boldsymbol{y} - D\boldsymbol{x}||_2 \leq \epsilon \tag{5}$$

Remark: (i) The sparse representation optimization has been widely used in the classification tasks such as voice, face recognition [17]–[19], because this only requires to construct the complete dictionary matrix rather than the time-consuming training process. (ii) A common trick is to relax L_0 minimization by L_1 in this sparse representation optimization, when considering the solving challenge that the L_0 minimization is an NP-hard problem. (iii) The expression (2) is changed into (5), due to the existence of measurement error ϵ, and thus this optimization problem is a second-order cone program (SOCP).

B. Poisioning Attack

1) Attack Intention: An example is taken to treat *Alice* as a illegitimate and *Bob* as the targeted legitimate. Keep in mind that the attack intention is the illegitimate *Alice* is misidentified as the targeted legitimate *Bob*. Specifically, the goal of poisoning attack is to inject perturbations into the training samples in order to misunderstood the model classification. Without loss of generality, D' represents the corrupted dictionary matrix and ΔD the model bias, which can be expressed as follows.

$$D' = D + \Delta D \tag{6}$$

This means that the training templates are tampered to implement the key of poisoning attacks.

2) Bilevel Optimization Model: Intrinsically, a two-player game exists between the malicious attackers and the gait recognizer. Initially, the adversary actively injects small perturbations to the training dataset to induce the identity misidentification of a GBIR system. Subsequently, the gait recognizer passively responds these false data injections. Eventually, going on many-round attack-defense competitions, two players achieve the game equilibrium point in which they have the balance of their own objectives. In brief, such a sequentially adversarial decision-making interaction process is a Stackelberg leader-follower game. Therefore, our proposed poisoning attacks can be described as an two-player adversarial game.

Generally, the bilevel optimization is an adequate framework to model such a Stackelberg leader-follower game [20]–[22].

Based on the mentioned expressions (3)-(6), we formulate a bilevel optimization problem to characterize our poisoning attacks as follows.

$$\min_{\Delta \boldsymbol{D}} \|\boldsymbol{x}_B - \boldsymbol{x}'_A\|_1 - \|\boldsymbol{x}_A - \boldsymbol{x}'_A\|_1 \tag{7}$$

$$\text{s.t. } \|\Delta \boldsymbol{D}\|_0 \leq \beta K \tag{8}$$

$$\|\Delta \boldsymbol{D}\|_2 \leq \beta \|\boldsymbol{D}\|_2 \tag{9}$$

$$\min ||\boldsymbol{x}||_1 \tag{10}$$

$$\text{s.t. } \|\boldsymbol{y}_A - (D + \Delta \boldsymbol{D})\boldsymbol{x}'_A\|_2 \leq \epsilon \tag{11}$$

where $\boldsymbol{y}_A$ is the testing sample from *Alice* and β is the threshold factor.

We note that $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are the specified sparse vectors used to accurately locate the identity index of *Alice* and *Bob* without any attack, respectively. We use the corrupted sparse vector $\boldsymbol{x}'_A$ to locate the *Alice*'s identity index under poisoning attack. In this leader-follower game, the upper-level leader is the attacker who focuses on designing optimal attack strategy to induce the identity misidentification, and the lower-level follower is the gait recognizer who aims at the accurate identity identification.

C. Problem Solving

The bilevel optimization problem is NP-hard, because the bilevel architecture raises a extremely huge solution space to cause the intractable computation, as investigated in our publication [23]. To solve our bilevel problem efficiently, we use Lagrangian dual of SOCP in [24] to solve the lower-level optimization problem. Finally, we adopt these mentioned duality theory to reformulate a single-level SOCP as follows.

$$\begin{aligned}
&\min_{\Delta \boldsymbol{D}} \|\boldsymbol{x}_B - \boldsymbol{x}'_A\|_1 - \|\boldsymbol{x}_A - \boldsymbol{x}'_A\|_1 \\
&\text{s.t.} \|\Delta \boldsymbol{D}\|_0 \leq \beta K \\
&\quad \|\Delta \boldsymbol{D}\|_2 \leq \beta \|\boldsymbol{D}\|_2 \\
&\quad \|\boldsymbol{y}_A - (D + \Delta \boldsymbol{D})\boldsymbol{x}'_A\|_2 \leq \epsilon \\
&\quad \mathbf{1} + D^{\mathrm{T}}\boldsymbol{\nu} - \boldsymbol{\eta} = \mathbf{0} \\
&\quad\quad ||\boldsymbol{\nu}||_2 \leq \lambda \\
&\quad\quad \boldsymbol{\eta} \succeq \mathbf{0} \\
&\quad \mathbf{1}^{\mathrm{T}}\boldsymbol{x}'_A \leq -\epsilon\lambda - (\boldsymbol{y}_A + \Delta \boldsymbol{y}_A)^{\mathrm{T}}\boldsymbol{\nu}
\end{aligned} \tag{12}$$

We notice that (12) can be run in a powerful mathematical program solver Gurobi [25], which can obtain a near-globally optimal solution to accurately recognize the user identity.

IV. Experiments and Evaluation

A. Experiment Settings

To verify the validity of our model in different gait data, we use two public datasets as experiment data. Concretely, PubData1 is from [26], including single-source gait signals with 50 subjects; and PubData2 is from [27], including multi-source gait signals with 56 subjects. After the gait data signal reprocessing, we extract 126- and 264-dimension gait signal features from PubData1 and PubData2 to produce the dictionary matrices and testing examples, respectively. These details are referred to our publication [27].

B. Evaluation Metric

In this work, we present seven evaluation criteria to quantify the perturbation magnitude level as follows.

(i) Average attack success rate (AASR): the average successful attack in the 100-round experiments;

(ii) Average attack confidence rate (AACR) of successful attack: $\frac{\sum_l \max\{\boldsymbol{x}_A\}_l}{100}$, which is to evaluate the average attack confidence of successful attack;

(iii) L_0-relative cost (RC): $\frac{\|\Delta \boldsymbol{D}\|_0}{m}$, which is to measure the sparsity of perturbations;

(iv) L_1-absolute cost (AC): $\frac{\|\Delta \boldsymbol{D}\|_1}{\|\boldsymbol{D}\|_1}$, which is to measure the absolute cost of perturbations using L_1-norm;

(v) L_1-RC: $\frac{\|\Delta \boldsymbol{D}\|_1}{\|\boldsymbol{D}_A - \boldsymbol{D}_B\|_1}$, which is to measure the relative cost of perturbations using L_1-norm;

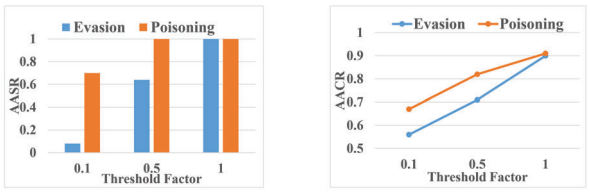
(vi) L_2-AC: $\frac{\|\Delta \boldsymbol{D}\|_2}{\|\boldsymbol{D}\|_2}$, which is to measure the absolute cost of perturbations using L_2-norm.

(vii) L_2-RC: $\frac{\|\Delta \boldsymbol{D}\|_2}{\|\boldsymbol{D}_A - \boldsymbol{D}_B\|_2}$, which is to measure the relative cost of perturbations using L_2-norm;

C. Evaluation Experiments

1) Attack Experiment on PubData1: In this experiment, we evaluate the performance of poisoning attack strategy on PubData1 with different threshold factor β. Each experiment is repeated for 100 times. The evasion attack approach is compared with our method. The average results are shown in Fig.1 and Table I.

Fig. 1 shows the changes of AASR and average AACR for the testing samples of *Alice* under the evasion and poisoning attack strategies against PubData1. We can see that the AASR under these two attack strategies present the increasing trending till 100% and the AACR rises up to 90%, as the threshold factor β changes from 0.1 to 1. These results illustrate that the gait recognizer of GBIR system is really susceptible to these adversarial attacks to misidentify *Alice* as *Bob*. Also, we can find that the AASR and AACR of poisoning attack strategy are higher than that of evasion attack strategy. Even though given the slight attack resources, that is, the threshold factor β is assigned 0.1, the AASR remains above 70% and the AACR is over 68% under the poisoning attack strategy. These results indicate that poisoning attack strategy is more effective than evasion attack strategy.



(a) AASR (b) AACR

Fig. 1. Experiment Results of AASR and AACR with PubData1.

Table I mainly lists the results of L_0-RC, L_1-RC, L_2-RC, L_1-AC, and L_2-AC for the testing samples of *Alice* under

TABLE I
THE RESULTS OF ATTACK COST FOR *Alice* ON PUBDATA1

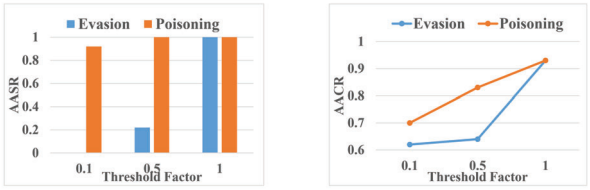
Threshold factor	Strategy	Original Label	Present Label	Variable	L_0-RC	L_1-RC	L_2-RC	L_1-AC	L_2-AC
0.1	Evasion	Alice	Alice/Bob	$\Delta \boldsymbol{y}_A$	9.99%	7.75%	9.04%	3.02%	4.67%
		Bob	Bob						
0.1	Poisoning	Alice	Alice/Bob	$\Delta \boldsymbol{D}_A$	6.66%	5.07%	6.35%	2.29%	3.01%
		Bob	Bob	$\Delta \boldsymbol{D}_B$	9.99%	7.61%	9.11%	2.81%	3.69%
0.5	Evasion	Alice	Alice/Bob	$\Delta \boldsymbol{y}_A$	36.63%	42.91%	46.72%	16.72%	18.89%
		Bob	Bob						
0.5	Poisoning	Alice	Bob	$\Delta \boldsymbol{D}_A$	9.99%	43.23%	47.17%	16.65%	18.79%
		Bob	Bob	$\Delta \boldsymbol{D}_B$	49.95%	47.22%	49.02%	17.64%	19.79%
1	Evasion	Alice	Bob	$\Delta \boldsymbol{y}_A$	16.65%	78.56%	85.17%	35.88%	39.34%
		Bob	Bob						
1	Poisoning	Alice	Bob	$\Delta \boldsymbol{D}_A$	13.32%	83.30%	91.72%	34.72%	39.93%
		Bob	Bob	$\Delta \boldsymbol{D}_B$	69.93%	86.10%	88.51%	36.73%	39.76%

the evasion and poisoning attack strategies against PubData1. Table I also shows the label change before and after attacks. If the original label *Alice* is identified as the present label *Alice/Bob* (marked in blue color), this means that only partial attacks are successful for 100-time testing experiments. The red color *Bob* means *Alice* is identified as the present label *Bob* with 100%.

In Table I, we observe that the possibility of successful attack (marked in red color) increase as the threshold factor increases from 0.1 to 1. This confirms that the more attack resources are consumed, the higher attack cost is spent. When the threshold factor is 0.1, poisoning attack strategy can achieve higher attack successful rate shown in Fig. 1 with small attack cost, compared with evasion attack. When the threshold factor is 0.5, the attack successful rate of poisoning attack can achieve 100% with only modifying less than half of the features.

2) Attack Experiment on PubData2: Fig. 2 shows the results of AASR and AAC for the testing samples of *Alice* under the evasion and poisoning attack strategies against multi-source PubData2. We observe that the values of AASR and AACR under these two strategies increase gradually, as the threshold factors are assigned from 0.1 to 1. At the same time, we also see that the AASR and AACR of poisoning attack strategy are higher than that of evasion attack strategy. Even though the threshold factors are assigned 0.1 to allow an imperceptible attack resource, the AASR remains above 90% and the AACR is over 70% under poisoning attack strategy. These results indicate that the gait recognizer is readily deceived to cause the identity misidentification and poisoning attack strategy is the more effective than evasion attack strategy, which are also verified in the testing experiment against single-source PubData1.

Table II mainly lists the results for the testing samples of *Alice* under the evasion and poisoning attack strategies against PubData2. Compared with PubData1, similar results are achieved in PubData2. The poisoning attack is more efficient than evasion attack with different threshold factors. When the threshold factor is 0.1, poisoning attack strategy can achieve higher attack successful rate of 90% shown in Fig. 2 with small attack cost, while the attack successful rate of evasion attack is 0. These results illustrate that poisoning attack



(a) AASR (b) AACR

Fig. 2. Experiment Results of AASR and AACR with PubData2.

which attacks the models is relatively more serious than that of attacking the testing samples. Compared with the single-source PubData1 (totally 132 features), the multi-source PubData2 (totally 264 features) is easier to be attacked, the reason may be that in PubData2, more invalid features are extracted from different sources that may affect the robustness of the model. In the future, we will consider some strategies such as feature selection, to increase the robustness of GBIR model.

V. CONCLUSIONS

In this paper, we propose a generalized bilevel optimization framework to model the poisoning attacks against the GBIR to study how *Alice* is misidentified as the legitimate user *Bob*. Numerical experiments in the single- and multi-source datasets verify that poisoning attack is the more efficient than evasion attack. In the future, we will make more efforts on the study of how to defense the poisoning attacks.

VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China [grant number T2241018], the National Science and Technology Major Project of the Ministry of Science and Technology of China [grant number 2021ZD0114303].

REFERENCES

[1] C. J. D'Orazio, K.-K. R. Choo, and L. T. Yang, "Data exfiltration from internet of things devices: ios devices as case studies," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 524–535, 2016.

[2] F. Loula, S. Prasad, K. Harber, and M. Shiffrar, "Recognizing people from their movement." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 1, p. 210, 2005.

TABLE II
THE RESULTS OF ATTACK COST FOR *Alice* IN PUBDATA2

Threshold factor	Strategy	Original Label	Present Label	Variable	L_0-RC	L_1-RC	L_2-RC	L_1-AC	L_2-AC
0.1	Evasion	Alice	Alice	$\Delta \boldsymbol{y}_A$	10.00%	5.57%	8.99%	1.92%	3.72%
		Bob	Bob						
0.1	Poisoning	Alice	Alice/Bob	$\Delta \boldsymbol{D}_A$	3.33%	2.50%	4.58%	0.58%	2.33%
		Bob	Bob	$\Delta \boldsymbol{D}_B$	10.00%	5.58%	9.08%	2.75%	3.67%
0.5	Evasion	Alice	Alice/Bob	$\Delta \boldsymbol{y}_A$	45.00%	43.86%	45.71%	19.54%	21.99%
		Bob	Bob						
0.5	Poisoning	Alice	Bob	$\Delta \boldsymbol{D}_A$	15.00%	34.88%	45.71%	18.71%	22.14%
		Bob	Bob	$\Delta \boldsymbol{D}_B$	50.00%	48.04%	49.32%	21.15%	23.47%
1	Evasion	Alice	Bob	$\Delta \boldsymbol{y}_A$	33.33%	85.25%	95.88%	30.96%	34.23%
		Bob	Bob						
1	Poisoning	Alice	Bob	$\Delta \boldsymbol{D}_A$	20.00%	55.36%	76.11%	19.80%	27.98%
		Bob	Bob	$\Delta \boldsymbol{D}_B$	86.67%	82.23%	85.72%	29.13%	30.57%

[3] F. Sun, W. Zang, R. Gravina, G. Fortino, and Y. Li, "Gait-based identification for elderly users in wearable healthcare systems," *Information fusion*, vol. 53, pp. 134–144, 2020.

[4] W. Sheng and X. Li, "Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network," *Pattern Recognition*, vol. 114, p. 107868, 2021.

[5] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, 1995.

[6] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[7] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, "Image understanding for iris biometrics: A survey," *Computer vision and image understanding*, vol. 110, no. 2, pp. 281–307, 2008.

[8] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[9] D. Gafurov, E. Snekkenes, and P. Bours, "Spoof attacks on gait authentication system," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 491–502, 2007.

[10] M. Muaaz and R. Mayrhofer, "Smartphone-based gait recognition: From authentication to imitation," *IEEE Transactions on Mobile Computing*, vol. 16, no. 11, pp. 3209–3221, 2017.

[11] B. B. Mjaaland, P. Bours, and D. Gligoroski, "Walk the walk: Attacking gait biometrics by imitation," in *International Conference on Information Security*. Springer, 2010, pp. 361–380.

[12] R. Kumar, I. Can, and V. Phoha, "Treadmill assisted gait spoofing (tags): An emerging threat to wearable sensor-based gait authentication," *ACM Journal of Digital Threats: Research and Practice*, vol. 00, no. 0, pp. 1–17, 2020.

[13] O. Suciu, R. Marginean, Y. Kaya, H. Daume III, and T. Dumitras, "When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1299–1316.

[14] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 27–38.

[15] C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7614–7623.

[16] H. Huang, J. Mu, N. Z. Gong, Q. Li, B. Liu, and M. Xu, "Data poisoning attacks to deep learning based recommender systems," in *Proceedings of Network and Distributed System Security Symposium*. Internet Society, 2021, pp. 1–17.

[17] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2368–2378, 2014.

[18] Y. Xu, Z. Zhong, J. Yang, J. You, and D. Zhang, "A new discriminative sparse representation method for robust face recognition via l_{2} regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2233–2242, 2017.

[19] Q. Zhang and M. D. Levine, "Robust multi-focus image fusion using multi-task sparse representation and spatial context," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2045–2058, 2016.

[20] Y. Chu and F. You, "Integrated scheduling and dynamic optimization by stackelberg game: Bilevel model formulation and efficient solution algorithm," *Industrial & Engineering Chemistry Research*, vol. 53, no. 13, pp. 5564–5581, 2014.

[21] B. Zeng, "A practical scheme to compute the pessimistic bilevel optimization problem," *INFORMS Journal on Computing*, vol. 32, no. 4, pp. 1128–1142, 2020.

[22] D.-T. Peng, J. Dong, and Q. Peng, "Overloaded branch chains induced by false data injection attack in smart grid," *IEEE Signal Processing Letters*, vol. 27, pp. 426–430, 2020.

[23] D.-T. Peng, J. Dong, J. Yang, and Q. Peng, "Dynamical failures driven by false load injection attacks against smart grid," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2213–2226, 2022.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, NY: Cambridge University Press, 2004, pp. 215–287.

[25] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2019. [Online]. Available: http://www.gurobi.com

[26] G. Vavoulas, C. Chatzaki, T. Malliotakis, M. Pediaditis, and M. Tsiknakis, "The mobiact dataset: Recognition of activities of daily living using smartphones." in *ICT4AgeingWell*, 2016, pp. 143–151.

[27] J. Dong, Y. Du, and Z. Cai, "Gender recognition using motion data from multiple smart devices," *Expert Systems with Applications*, vol. 147, p. 113195, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

STrack: Velocity Estimation Using Single Antenna WiFi Devices

Jian Xu[1], *Dongheng Zhang[1,3] , Jiamu Li[1], Qibin Sun[1], Yan Chen[1,2,3]
[1]School of Cyber Science and Technology, University of Science and Technology of China, Hefei, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[3] Institute of Dataspace, Hefei Comprehensive National Science Center, Hefei, China
Email: xujian0710@mail.ustc.edu.cn, dongheng@ustc.edu.cn, lijiamu@mail.ustc.edu.cn, {qibinsun,eecyan}@ustc.edu.cn

***Abstract*—As the Internet of Things (IoT) continues to evolve, WiFi devices are becoming increasingly ubiquitous in our everyday lives. This proliferation of devices has paved the way for innovative technologies such as device-free passive indoor tracking using WiFi devices. For device-free passive WiFi sensing, the Doppler Frequency Shift (DFS) extracted from fine-grained Channel State Information (CSI) phase has been widely adopted for estimating target motion velocity. Nevertheless, due to the presence of random phase offsets in CSI and the fact that many IoT devices have only one WiFi antenna, utilizing CSI phase for calculating DFS is usually unfeasible. In this paper, we present a novel Single Antenna WiFi Tracking (STrack) system that estimates DFS using CSI amplitude, eliminating the need for DFS symbols to directly participate in velocity calculation. Extensive experimental results indicate that STrack outperforms state-of-the-art indoor tracking systems, with a median velocity error of 1.49cm/s, and a median direction error of 12.41°, a median positioning error of 48.529cm.**

I. Introduction

Passive human tracking technology has gained significant attention in recent years due to its potential applications in various fields. It can monitor activities of elderly individuals who are at risk of falls or other health issues with little privacy concerns [1]. It is also applied for in-home security systems, where it can efficiently detect and track intruders. Existing devices that utilized for passive human tracking can be generally categorized to WiFi devices [2], [3], cameras [4], [5], and Ultra Wide Band (UWB) devices [6], etc. Compared with other devices, WiFi-based ones are cost-effective and privacy-friendly, which can be deployed widely in smart homes [7].

In device-free passive WiFi indoor tracking technology, the Doppler Frequency Shift (DFS) extracted from fine-grained Channel State Information (CSI) is a widely adopted technique for estimating target motion velocity [8]. Channel State Information (CSI) is obtained from the physical layer of the device [9], with each antenna capable of capturing the corresponding CSI data. However, the CSI obtained by each antenna contains a random phase offset, which hinders the direct extraction of Doppler Frequency Shift (DFS) information from the CSI. One common approach to eliminate random phase offsets is to employ conjugate multiplication [10] or division [11] of the CSI data of the two antennas. This method requires that the WiFi device has at least two antennas operating in the same frequency band, while not all WiFi devices meet this requirement. In the realm of the Internet of Things, single-antenna WiFi devices [12], [13] are employed to meet the requirement for low-cost hardware.

In this paper, we delve into passive human tracking using IoT devices equipped with only one antenna. Our core idea revolves estimating the DFS utilizing the CSI amplitude. This approach enables us to achieve target tracking while avoiding the unsolvable CSI phase offset problem. Nonetheless, since the amplitude of CSI is a set of real numbers, the utilization of amplitude-based estimation for DFS presents a challenge in determining the DFS symbols, which can only take positive values in the calculation. This will lead to wrong estimation of velocity direction. In previous work, Widar [2] exploits the continuity of motion to determine a unique direction of motion. This judgment method relies on the last speed estimation, which is not reliable for random motion. According to the principle of Fresnel zone [14], WiDir [15] uses the frequency difference of different subcarriers to analyze the timing of all subcarriers to obtain the symbol information of DFS. However, this method also exhibits certain limitations. Specifically, in instances where the angle between the direction of motion and the transmitter-receiver link is small, the accurate estimation of the DFS symbol becomes difficult.

In order to tackle the challenge of losing DFS symbols during estimation, we propose a new velocity estimation framework for single-antenna devices, termed STrack, which capitalizes on the observation that smart homes frequently accommodate multiple IoT devices within a single room. In this framework, initially, we compute unsigned DFS from the CSI amplitude collected by each WiFi device. Then, these unsigned DFSs are employed to calculate the complete set of feasible velocities. Finally, the actual velocity is extracted from the array of potential velocities using both the K-means algorithm and a novel weighting algorithm. Our main contributions can be summarized as follows:

1) We establish a model relating CSI amplitude and motion velocity by quantifying the length variation of the dynamic path, which can be used to measure and analyze the changes in CSI amplitude corresponding to different velocities.
2) We use multiple unsigned DFS to jointly estimate the

*Dongheng Zhang is the corresponding author.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

velocity of the target, and design a weighted estimation algorithm to judge the authenticity of the velocity.

3) We developed a single antenna WiFi tracking system (STrack) that accurately tracks the trajectory of a target. Through extensive experimentation, we have demonstrated that STrack outperforms other state-of-the-art algorithms, with a median velocity error of 1.49cm/s, and a median direction error of 12.41°, a median positioning error of 48.529cm.

II. Velocity Estimation Model

In this section, we present the model between CSI amplitude variation and target motion velocity.

A. CSI-Velocity Model

In device-free passive WiFi tracking, the position of receiver and transmitter is usually fixed, so the signal path from transmitter to receiver can be divided into two parts: dynamic path and static path. Static paths include direct paths (Line-of-Sight) and paths reflected by stationary objects. The dynamic path is introduced by the moving target in the environment. We define the single tone signal sent by the transmitter as $X(f_0, t_0) = ae^{j2\pi f_0 t_0}$ at time t_0, where f_0 is carrier frequency. The signal received by the receiver is $Y(f_0, t_0) = a'e^{j2\pi f_0(t_0-\tau_s)}$, where τ_s represents the time interval for the wave to travel from the transmitter to the receiver. τ_s is a fixed value because the distance between transmitter and receiver is fixed. The measured CSI with only static path can be expressed as

$$H_s(f, t_0) = \frac{Y(f, t_0)}{X(f, t_0)} = \frac{a'e^{j2\pi f(t_0-\tau_s)}}{ae^{j2\pi f t_0}} = Ae^{-j2\pi f\tau_s}, \quad (1)$$

where $A = a/a'$ is amplitude of CSI.

In the dynamic path, the length varies depending on the motion of the target. To establish the relationship between the time interval τ_d and the target's velocity, we aim to determine how changes in the target's velocity impact the duration τ_d.

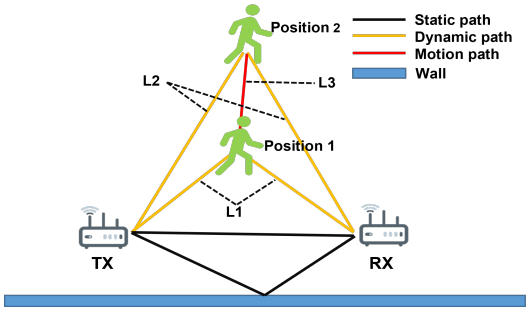


Fig. 1. Schematic diagram of static path and dynamic path generated by target motion. L_1 represents the reflected path length of the wave when the target is at position 1, and L_2 represents the reflected path length of the wave when the target is at position 2. L_3 represents the distance of the target movement

We assume that the magnitude and direction of the target's velocity are constant in a short period of time $\triangle t$. As shown in Fig. 1, the target starts from position 1 and moves to position 2 after $\triangle t$ seconds. When the target is in position 1, the CSI of the receiving end can be expressed as $Y(f_0, t_0) = a'e^{j2\pi f_0(t_0-\tau_{d1})}$, where $\tau_{d1} = L_1/c$, c is the speed of light, and L_1 represents the reflected path length of the wave when the target is at position 1. Similarly, when the target is in position 2, the CSI of the receiving end can be expressed as $Y(f_0, t_0) = a'e^{j2\pi f_0(t_0-\tau_{d2})}$, where $\tau_{d2} = L_2/c$. Now, we need to find the relationship between L_1, L_2, V (velocity magnitude) and θ (velocity direction). As shown in Fig. 2, we

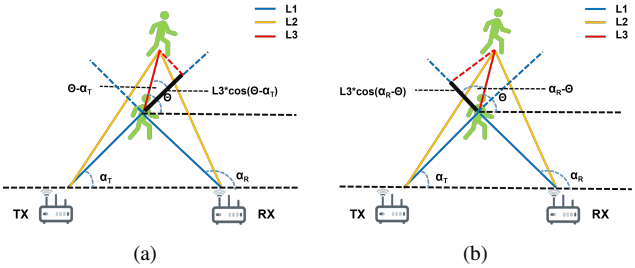


Fig. 2. Schematic diagram of motion displacement decomposition. Figure (a) represents the component of the target motion displacement (L3) on the TX-Target vector. Figure (b) represents the component of the target motion displacement (L3) on the Target-RX vector.

use the sum of the components of the target displacement length in the Target-TX and Target-RX directions to approximately replace the change in the signal propagation path length. The target displacement length can be expressed as : $L_3 = V * \triangle t$. We get the following formula:

$$L_2 = L_1 + V * \triangle t * (cos(\theta - \alpha_T) + cos(\theta - \alpha_R)). \quad (2)$$

α_T is the Angle of departure (AoD) and α_R is the Angle of arrival (AoA). Thus, when the target is in position 2, the signal of the receiving end can be rewritten as:

$$\begin{aligned} Y(f, t_0) &= a'e^{j2\pi f(t_0-L_2/c)} \\ &= a'e^{j2\pi f(t_0-(L_1+V*\triangle t*(cos(\theta-\alpha_T)+cos(\theta-\alpha_R))/c)}, \end{aligned} \quad (3)$$

the CSI of a dynamic path can be expressed as:

$$H_d(f, t_0) = Ae^{-j2\pi f((L_1+V*\triangle t*(cos(\theta-\alpha_T)+cos(\theta-\alpha_R))/c)}. \quad (4)$$

It should be noted that in the derivation of [11], DFS can be expressed as:

$$f_D = -f * V * (cos(\theta - \alpha_T) + cos(\theta - \alpha_R))/c, \quad (5)$$

and we assume that L_1 is constant over a short estimation period. So we can get that:

$$H_d(f, t) = Ae^{-j2\pi(f*L_1/c-f_D*t)}. \quad (6)$$

In practical indoor environments, CSI is the superposition of channel responses of static path and dynamic path, which can be expressed as:

$$H(f, t) = \sum_{i\in P_s} A_i e^{-j2\pi f\tau_s^i} + \sum_{j\in P_d} A_j e^{-j2\pi(f*L_j/c-f_D*t)}. \quad (7)$$

Where P_s is the set of all static paths and P_d is the set of all dynamic paths.

B. CSI Amplitude-Velocity Model

In commodity WiFi devices, since the time of transmitter and receiver is not completely synchronized, there will be a time-varying phase offset $e^{-j\phi_{offset}}$ in every CSI measurement [16], [17]. So, a complete CSI can be expressed as:

$$H(f,t) = e^{-j\phi_{offset}(t)}(H_s(f) + \sum_{j \in P_d} A_j e^{-j2\pi(f*L_j/c - f_D*t)}). \tag{8}$$

This time-varying phase offset has a severe effect on estimating DFS. Fortunately, we find that the CSI amplitude is not affected by random phase offset. According to [17], CFR power can be expressed as:

$$\begin{aligned} \mid H(f,t) \mid^2 = & \sum_{j \in P_d} 2 \mid H_s(f) A_j \mid cos(2\pi(\mid f_{D_j} \mid *t + f * L_j/c) \\ & + \phi_{sj}) + \sum_{\substack{n,m \in P_d \\ n \neq m}} 2 \mid A_n A_m \mid cos(2\pi((f_{D_n} - f_{D_m})t + f(L_n - \\ & L_m)/c) + \phi_{nm}) + \sum_{j \in P_s} \mid A_j \mid^2 + \mid H_s(f) \mid^2 . \end{aligned} \tag{9}$$

According to Eq. 9, the amplitude of CSI is the sum of a set of cosines and constant offset values, where the frequency of cosines is the absolute value of DFS. Therefore, we only need to calculate the frequency of the CSI amplitude in a short time to get the absolute value of the DFS caused by the target motion. When the target moves in a fixed direction away from the transmitter-receiver link, the variation of the CSI amplitude is depicted in Fig. 3. It was observed that the CSI amplitude waveform resembles a cosine waveform, where the variation decrease with the increase of dynamic path length.

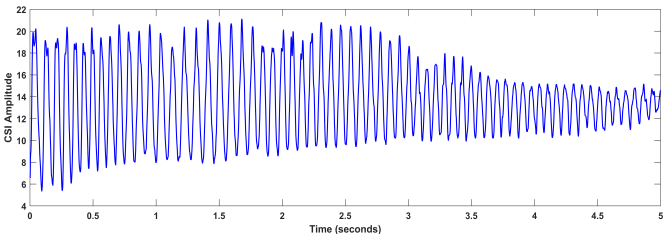


Fig. 3. CSI amplitude waveform resembles a sum of cosine and constant offset values. As the distance of the dynamic path increases, the impact of the dynamic path on the CSI amplitude decreases.

III. Joint Estimation of Motion Velocity Using Multiple Unsigned DFS

The utilization of CSI amplitude for DFS calculation results in the inevitable loss of DFS symbols, which are crucial in accurately computing motion velocity. In this section, we present a novel method to calculate motion velocity using unsigned DFS. This approach effectively prevents velocity estimation errors resulting from the loss of DFS symbols.

As we rely on the CSI amplitude for DFS estimation, the actual symbol of DFS is not directly available to us. Therefore,

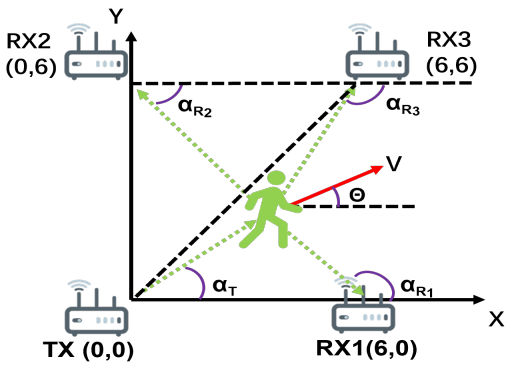
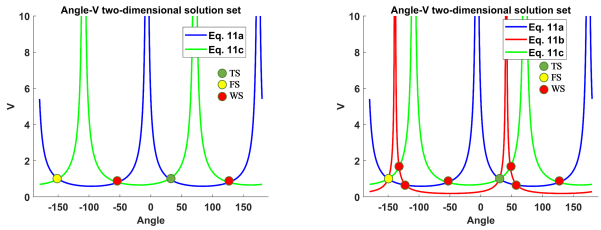


Fig. 4. Multiple decives (1 transmitter, 3 receivers) jointly estimate the moving velocity of the target



(a) (b)

Fig. 5. In Figure (a), the blue line represents all solutions satisfying Eq. 11a. The green lines represent all solutions satisfying Eq. 11c. In Figure (b), green, yellow, and red thick dots represent True solution (TS), Flase solution (FS), and Wrong solution (WS), respectively.

the relationship between DFS extracted from CSI amplitude and target velocity is as follows:

$$\mid f_D \mid = \mid f * V * (cos(\theta - \alpha_T) + cos(\theta - \alpha_R))/c \mid . \tag{10}$$

According to Eq. 10, DFS is not only determined by the velocity of target movement, but also determined by the position of transmitter, receiver and target. As shown in Fig. 4, combing the DFS measured from multiple different receivers, we have the follwing equations

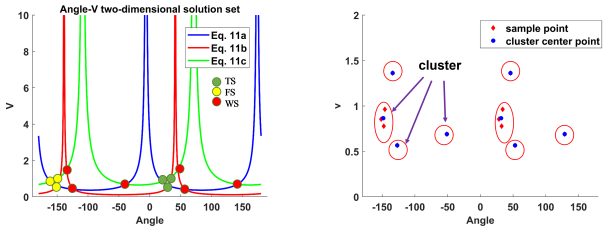
$$\mid f_{D_1} \mid = \mid f * V * (cos(\theta - \alpha_T) + cos(\theta - \alpha_{R_1}))/c \mid \tag{11a}$$
$$\mid f_{D_2} \mid = \mid f * V * (cos(\theta - \alpha_T) + cos(\theta - \alpha_{R_2}))/c \mid \tag{11b}$$
$$\mid f_{D_3} \mid = \mid f * V * (cos(\theta - \alpha_T) + cos(\theta - \alpha_{R_3}))/c \mid \tag{11c}$$

When combining any two equations in Eq. 11, we can get 4 different velocities as shown in Fig. 5(a), where the intersection of the two curves are the estimated velocity.

As shown in in Fig. 5(b), the True Solution (TS) could satisfy the constraint of all equations, which is the intersection of all three curves. Meanwhile, since the DFS symbol has been lost, there is another False Solution (FS) which has the same absolute value but opposite direction compared with TS. The rest of the intersections are Wrong Solutions (WS). To achieve accurate velocity estimation, we utilize the distribution difference between TS and WS and the direction disparity between TS and FS to determine which intersection is the actual velocity.

A. Differentiate between TS and WS

To distinguish TS and WS, we have noted that since TS can satisfy all equations, a natural approach is to solve Eq. 11. However, due to the estimation error of DFS, it is difficult to directly estimate the solution of TS from all equations in Eq. 11. To resolve this problem, we have noted that by combing any two equations in Eq. 11, we can obtain all possible solutions as shown in Fig. 6, where the possible solution of TS forms clusters while WS distribute randomly. Hence, we can adopt K-Means algorithm to find the largest cluster as the estimation of TS.



(a) (b)

Fig. 6. In Figure (a), TS and FS are gathered together, WS is scattered. In Figure (b), The red dots represent all the solutions computed by Eq. 11. The blue dots represent the center points of the clusters.

K-means is a commonly used unsupervised clustering algorithm [18]. Its basic idea is to find a scheme to divide K clusters through iteration, so that the loss function corresponding to the clustering result is minimized. The loss function can be defined as the sum of squared errors of each sample from the center point of the cluster to which it belongs:

$$Loss(c,u) = \sum_{i=1}^{M} \parallel x_i - u_{c_i} \parallel^2, \tag{12}$$

where M is the total number of samples, x_i is the i-th sample, c_i is the cluster to which x_i belongs, and u_{c_i} is the center point of cluster c_i.

As depicted in Fig. 6(b), the K-Means algorithm utilizes all points within the velocity set as sample points, and subsequently divides points with closer distances into the corresponding clusters. However, due to the aggregation of all TS, and the scattering of WS, all TS are consolidated into a single cluster, while WS is divided into multiple distinct clusters. Similarly, FS is assigned to its own cluster. The determination of the cluster that divides TS or FS is based on considering the cluster with the highest count of sample points. Hence, the number of clusters, K, in the K-means algorithm is a important hyperparameter.

In a scenario where there is one transmitter and N receivers, we can derive N equations similar to Eq. 10. By combining any two equations, we obtain $M = C_N^2 = N*(N-1)/2$ combinations, each of which yields 4 distinct solutions. Consequently, there are a total of $L = 4M = N*(N-1)*2$ solutions. Theoretically, TS manifests as a set of identical solutions, which constitutes 1/4 of the original solution set, and so does FS. On the other hand, WS manifests as a set of distinct solutions, accounting for 1/2 of the original solution set. Therefore, the number of clusters should be $K = L/2+2 = N*(N-1)+2$. After selecting the cluster with the largest number of sample points, we find the sample point closest to the cluster center as the velocity to be distinguished, denoted as (V^*, θ^*). Note that since the size of the cluster for TS and FS are the same, we cannot distinguish TS and FS based on the number of sample points in the cluster.

B. Differentiate between TS and FS

To distinguish TS and FS, we develop a weighted discrimination algorithm based on Widir [15] to effectively identify and select the desired TS. According to the Fresnel Zone principle, for two different sub-carriers, the wavelengths are λ_1, λ_2 respectively, When the movement of the target is close to the transmitter-receiver link, the fluctuation first appears in the subcarrier with longer wavelength, and then appears in the subcarrier with shorter wavelength. On the contrary, when the target is far away from the transmitter-receiver link, fluctuations first appear on the subcarriers with shorter wavelengths and then on the subcarriers with longer wavelengths. By analyzing the time delays on different subcarriers, we can assess the temporal order of fluctuations across the subcarriers [19]. Based on this sequence, we can estimate the DFS symbol. However, the accuracy of this method is limited, particularly when the angle between the motion direction and the transmitter-receiver link is small. Fortunately, we find out that for the velocity obtained through clustering, whether it corresponds to TS or FS, maintains a consistent angle with all transmitter-receiver links which can be employed to mitigate the loss of precision. Consequently, we can assign weights on the estimated DFS from different RXs based on the angle, which could improve the accuracy for the DFS symbol.

Thus, we develop a weighted discrimination algorithm. Initially, We utilize time delay analysis [15] to ascertain an initial DFS symbol for each RXs. Subsequently, the symbols estimated by each RXs are weighted and combined to obtain the final estimated DFS symbol. The weight is determined by the angle between the direction of velocity and the transmitter-receiver link. Finally, we leverage the estimated DFS symbol to differentiate between TS and FS.

IV. Tracking System Implementation

In this section, we introduce the implementation details of the proposed system. As shown in Fig. 7, the system is composed of three modules: CSI Acquisition and Processing, Velocity Estimation, Target Location Tracking.

CSI Acquisition and Processing: We utilize one WiFi transmitter and at least three receivers for CSI acquisition. In addition, in order to ensure the fairness of the comparison with other algorithms, all transmitting and receiving devices use mini-computers with Intel 5300 network cards. We set the working frequency band of the WiFi signal to 5G Hz, where the DFS generated by human body movement is in the range of 0-75 Hz. The mini-computer equipped with a 5300 network card has three antennas, and we only use the CSI data obtained

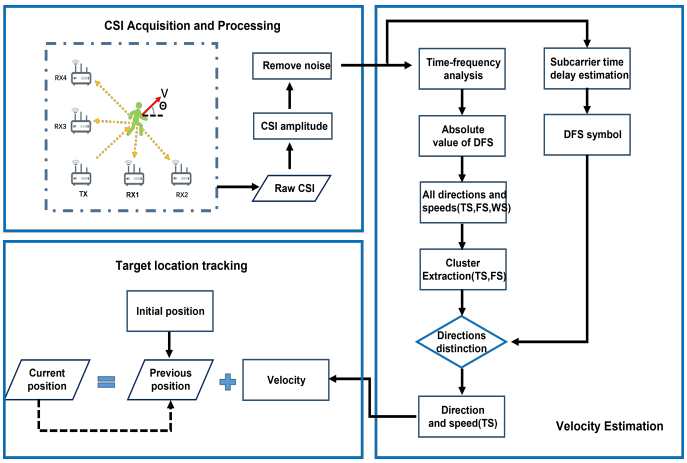


Fig. 7. System overview of STrack.

by the first antenna to complete subsequent calculations. The CSI data obtained by the remaining antennas is utilized for the calculation of other algorithms. we use the CSI data of the first and second antennas to implement PITrack [11]. We use all antennas to implement Widar2.0 [20]. To suppress random noise, we pass the signal through S-G filter [21].

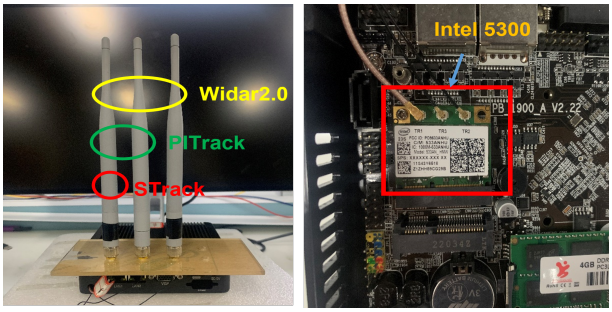


Fig. 8. Devices for experiments.

Velocity Estimation: We use the CWT algorithm [22] to extract the time-frequency spectrum from the CSI amplitude of all RXs. We then estimate the DFS from the time-frequency spectrum. Finally, we utilize the proposed algorithm to achieve velocity estimation.

Target Location Tracking: It can be seen from Eq. 5 that DFS is closely related to the speed of target motion and the position of the target at the previous timestamp. This requires us to know the initial position of the target motion in advance. With the initial position, we then calculate new positions based on the velocity estimation.

V. Experiments

In this section, we conduct extensive experiments under various settings and compare the results with state-of-the-art algorithms (PITrack [11] and Widar2.0 [20]). To assess the performance of STrack, we design experiments to evaluate its capabilities in three key aspects: velocity direction estimation, velocity magnitude estimation, and localization accuracy. We also conduct experiments in different environments to verify the robustness of the proposed framework.

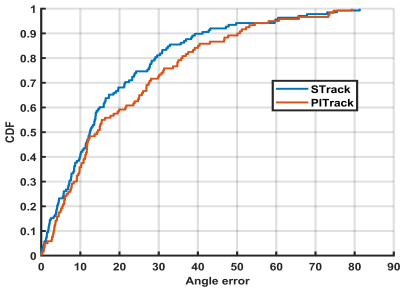


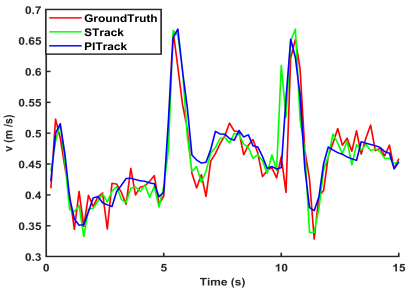
Fig. 9. CDF of velocity direction estimation error.

Velocity Direction Estimation: Direction estimation is crucial for device-free passive indoor tracking. The volunteer starts from a fixed position and move towards different directions, and conducts 13 sets of experiments along different directions, and each set of experiments is repeated 3 times. As shown in Fig. 9, the accuracy is slightly better than that of PITrack. The STrack direction estimation has a median error of 12.41° and an average error of 17.93°. PITrack has a median error of 14.22° and an average error of 21.31°. It worth noting that while the performance of STrack and PITrack is close, STrack can be deployed on IoT devices with single antenna, while PITrack requires 2 antennas to suppress random phase offset.

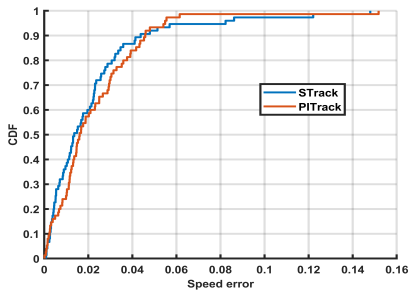
Velocity Magnitude Estimation: In experiments for direction estimation, we let the subject moves in a straight line at a fixed angle to ensure a constant direction of motion. However, maintaining a constant velocity magnitude during the exercise is challenging, necessitating a tool to collect motion velocity in real-time. To achieve this, we installed a Intel RealSense Tracking Camera T265 on the mobile platform, which can obtain a three-dimensional velocity information every 0.1s. In the experiment, we control the mobile platform to move in a fixed direction, record the velocity of movement in real-time, and then convert the three-dimensional velocity into a one-dimensional velocity to obtain the actual velocity required. As shown in Fig. 10(a), the velocity magnitude estimation with STrack is consistent with ground-truth with better accuracy than PITrack. As shown in Fig. 10(b), when the mobile platform moves at an angle of 90°, the median errors of STrack and PITrack are 1.49cm/s, 1.67cm/s, and the average errors are 2.16cm/s, 2.26cm/s, respectively.

Location Estimation: As shown in Fig. 11, the median positioning errors of STrack, PITrack, and Widar2.0 are 61.325cm, 66.114cm, and 135.47cm, respectively. In order to show the tracking performance of our system more vividly, we require volunteers to move according to the specified trajectory ('N', Triangle, Rectangle), as shown in Fig. 12, the red dotted line is the ground truth, and the black solid line is the trajectory calculated by the system , through comparison, we can find that the trajectory estimated by the system has a high degree of coincidence with the ground truth, which demonstrates the effectiveness of the proposed system.

REFERENCES



(a)



(b)

Fig. 10. Figure (a) shows an example of speed estimation. Figure (b) shows the CDF comparison of different methods

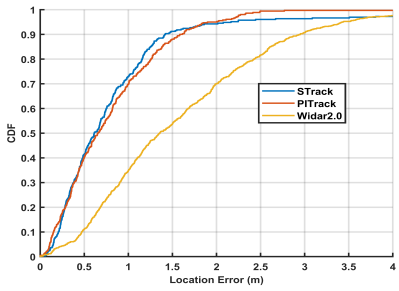


Fig. 11. CDF of location error

Impact of Environments: The above-mentioned experiments were all carried out in an open room with few reflection paths. To verify the impact of environments, we conduct experiments in two environments (Corridor and Meeting room). The corridor is 8m long, 2m wide, and the sensing range is 2m*6m. The meeting room is 12m long, 8m wide, and the sensing range is 5m*5m. Volunteers are also asked to walk according to the prescribed trajectory. As shown in Fig. 13 , whether in the corridor or in the meeting room, the trajectory

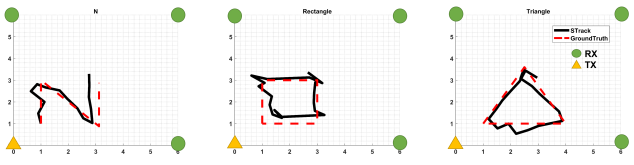
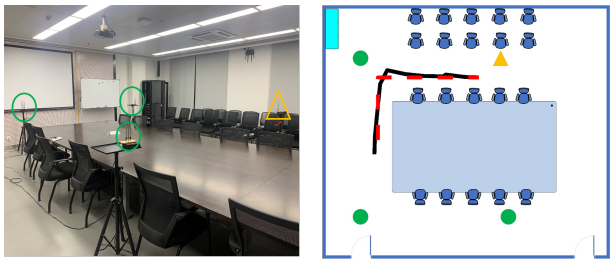
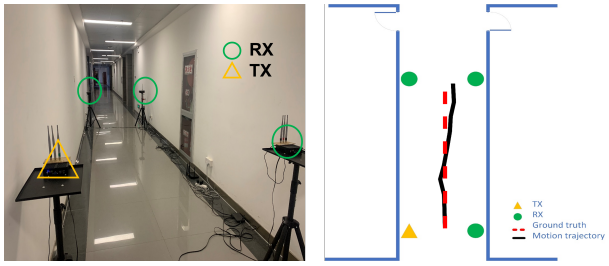


Fig. 12. Three different motion trajectories

estimated by the system is consistent with the ground truth.



(a)



(b)

Fig. 13. In corridor and meeting room scenarios, STrack can still work effectively.

VI. CONCLUSION

In this paper, we present STrack, a new Single Antenna WiFi Tracking system that estimates DFS using CSI amplitude. In this framework, we first compute unsigned DFS from the CSI amplitude collected by each WiFi device. Then, these unsigned DFSs are employed to calculate the complete set of feasible velocities. Finally, the actual velocity is extracted from the array of potential velocities using both the K-means algorithm and a novel weighting algorithm. Extensive experimental results indicate that STrack outperforms state-of-the-art indoor tracking systems, with a median velocity error of 1.49cm/s, and a median direction error of 12.41°, a median positioning error of 48.529cm.

ACKNOWLEDGEMENT

This work was supported by National Key R&D Programmes under Grant 2022YFC0869800 and 2022YFC2503405, National Natural Science Foundation of China under Grant 62201542 and 62172381, fellowship of China Postdoctoral Science Foundation under grant 2022M723069, and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] D. Zhang, Y. Hu, and Y. Chen, "Mtrack: Tracking multiperson moving trajectories and vital signs with radio signals," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3904–3914, 2020.

[2] K. Qian, C. Wu, Z. Yang, C. Yang, and Y. Liu, "Decimeter level passive tracking with wifi," in *Proceedings of the 3rd Workshop on Hot Topics in Wireless*, 2016, pp. 44–48.

[3] Y. Xie, J. Xiong, M. Li, and K. Jamieson, "Xd-track: Leveraging multi-dimensional information for passive wi-fi tracking," in *Proceedings of the 3rd Workshop on Hot Topics in Wireless*, 2016, pp. 39–43.

[4] M. R. U. Saputra, G. D. Putra, P. I. Santosa, *et al.*, "Indoor human tracking application using multiple depth-cameras," in *2012 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, IEEE, 2012, pp. 307–312.

[5] H. Kavalionak, E. Carlini, A. Lulli, *et al.*, "A prediction-based distributed tracking protocol for video surveillance," in *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*, IEEE, 2017, pp. 140–145.

[6] C. Li, E. Tanghe, J. Fontaine, *et al.*, "Multistatic uwb radar-based passive human tracking using cots devices," *IEEE Antennas and Wireless Propagation Letters*, vol. 21, no. 4, pp. 695–699, 2022.

[7] Y. He, Y. Chen, Y. Hu, and B. Zeng, "Wifi vision: Sensing, recognition, and detection with commodity mimo-ofdm wifi," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8296–8317, 2020.

[8] Y. Chen, H. Deng, D. Zhang, and Y. Hu, "Speednet: Indoor speed estimation with radio signals," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2762–2774, 2020.

[9] Z. Yang, Z. Zhou, and Y. Liu, "From rssi to csi: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–32, 2013.

[10] X. Li, D. Zhang, Q. Lv, *et al.*, "Indotrack: Device-free indoor human tracking with commodity wi-fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–22, 2017.

[11] K. Niu, X. Wang, F. Zhang, R. Zheng, Z. Yao, and D. Zhang, "Rethinking doppler effect for accurate velocity estimation with commodity wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 7, pp. 2164–2178, 2022.

[12] M. Atif, S. Muralidharan, H. Ko, and B. Yoo, "Wi-esp—a tool for csi-based device-free wi-fi sensing (dfws)," *Journal of Computational Design and Engineering*, vol. 7, no. 5, pp. 644–656, 2020.

[13] Espressif, https : / / www . espressif . com . cn / zh - hans / products/socs/esp32/, Last accessed on 2023-5-03, 2023.

[14] T. S. Rappaport, "Wireless communications–principles and practice, (the book end)," *Microwave Journal*, vol. 45, no. 12, pp. 128–129, 2002.

[15] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, "Widir: Walking direction estimation using wireless signals," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 2016, pp. 351–362.

[16] J. Xiong, K. Sundaresan, and K. Jamieson, "Tonetrack: Leveraging frequency-agile radios for time-based indoor wireless localization," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 537–549.

[17] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015, pp. 65–76.

[18] D. Polykovskiy and A. Novikov, "Bayesian methods for machine learning," *Coursera and National Research University Higher School of Economics*, 2017.

[19] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[20] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2. 0: Passive human tracking with a single wi-fi link," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018, pp. 350–361.

[21] R. W. Schafer, "What is a savitzky-golay filter?[lecture notes]," *IEEE Signal processing magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[22] J. M. Lilly and S. C. Olhede, "Generalized morse wavelets as a superfamily of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6036–6041, 2012.

SEformer: Dual-Path Conformer Neural Network is a Good Speech Denoiser

Kai Wang and Dimitrios Hatzinakos
Department of Electrical and Computer Engineering, University of Toronto, ON, Canada
E-mail: kaikai.wang@mail.utoronto.ca, dimitris@comm.utoronto.ca

***Abstract*—In this paper, we propose the SEformer, an efficient dual-path conformer neural network for speech enhancement. The proposed SEformer is comprised of an encoder, a decoder and dual-path conformer blocks in between. First, the encoder receives the noisy speech waveform to generate the speech features. Then, each proposed dual-path conformer block employs a temporal and frequency conformer in parallel to simultaneously extract temporal and frequency features of speech sequences, which are then fused by a transformer block for contextual representation. Finally, the decoder transforms the contextual speech features into a mask to filter out noise-related information of noisy speech, generating the estimated clean speech. Experimental results on the benchmark dataset exhibit that our SEformer yields a competitive performance to existing state-of-the-art methods while containing the fewest model parameters (about 590k).**

I. Introduction

Speech enhancement is a task aiming to improve the quality of corrupted speech by removing the noise-related information, which has become an indispensable stage in various applications like automatic speech recognition (ASR) and hearing aids. Recently, deep neural networks (DNNs) emerged as mainstream approaches to efficiently improve the performance of speech enhancement systems. Depending on the types of input signal, DNN-based speech enhancement can be categorized into Time-Frequency (T-F) and Time domain methods. In T-F domain, DNNs receive the image-like spectrum features of noisy speech as input to estimate the clean one, where the spectrum is obtained by implementing the short-time Fourier transform (STFT) on the speech waveform. For example, some DNN-based models such as neural networks (CNNs) [1] and recurrent neural network (RNNs) [2], exclusively estimate the clean magnitude from the noisy one, and reuse the phase information for reconstruction stage. However, it is sub-optimal to neglect to denoise the phase of noisy speech since it can provide extra information for better speech quality. Afterwards, various works focus on modeling both magnitude and phase information for improving the performance of speech enhancement [3]–[5]. For instance, the imaginary and real part of speech spectrum is first packed together to avoid the shortage of magnitude and phase information, and then passes through DNN-based models for predicting the clean imaginary and real spectrogram [3], respectively. Instead, the complex-value CNN is proposed to simultaneously enhance magnitude and phase of noisy speech by using more complicated structure [4].

Differing from T-F domain methods, the time-domain methods directly extract the clean speech from noisy one, which operates on the one-dimensional speech sequences. For the purpose of extracting the contextual information of long-range speech sequences, some works utilize encoder-decoder frameworks to map the inputs into the compressed features via an encoder, and then recover features back to enhanced speech through a decoder [6], [7]. To further extract the long-range dependencies of speech sequences, various DNN-based modules are inserted between the encoder and the decoder, including dilated CNN module [8], RNN module [7], and attention-based modules [9], [10]. Unlike T-F domain methods implementing on the spectrum with harmonic characteristics, time-domain methods generally involve more model or computational complexity to extract the contextual features of speech waveforms for remarkable performance.

To extract the contextual information of long-range speech sequences, some recent works adopt transformer neural network as a backbone or a component of speech enhancement system, where the transformer is originally proposed in [11] to address the long-dependency problem with self-attention mechanisms, and has been widely used in natural language processing (NLP) [12], computer vision (CV) [13], and speech processing tasks [14]–[16]. Authors of [14] introduced the Gaussian-weighted self-attention into transformer to pay more attention to the closer speech frames and less attention to the distant ones, significantly improving the performance of speech enhancement. Authors of [15] and [16] adopted two transformers to successively perform the local and global feature extraction for contextual representation, efficiently improving the performance of speech separation and enhancement. Furthermore, a convolution-augmented transformer named as Conformer [17] is proposed to efficiently leverage the local property of CNNs and the global characteristics of transformers for learning the contextual representation in speech recognition. Inspired by this, authors of [18] proposed a SE-Conformer for speech enhancement in the time domain, where the conformer is followed by the encoder for further extracting contextual features which are then passed through a decoder for speech reconstruction. Different from [18], authors of [19] adopted a two-stage conformer collaborating with generative adversarial network into speech spectrum, achieving an impressive performance of speech enhancement.

Despite various methods have been proposed to improve the performance of speech enhancement, it is still challenging to efficiently extract the long-range dependencies of speech

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

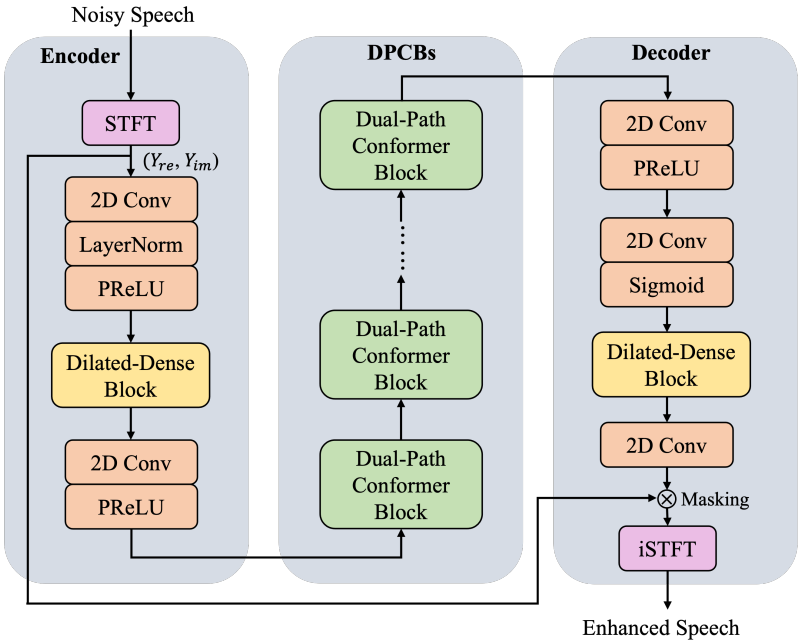


Fig. 1: Overview of proposed dual-path conformer neural network (SEformer)

sequences. By leveraging the facts that the two-stage extraction efficiently reduces the complexity of processing long-range speech sequences [20], and the conformer effectively learns the contextual information of input signals [17], we here propose a lightweight, simple but efficient SEformer for T-F domain speech enhancement. Our experiments on the benchmark dataset show that the proposed SEformer yields competitive performance to state-of-the-art denoising systems while employing the lowest model parameters.

II. Proposed Architecture

In this section, we propose a dual-path conformer, called SEformer, for speech enhancement in the T-F domain. As shown in Fig. 1, SEformer includes an encoder, a decoder and dual-path conformer blocks in between.

A. Encoder

The encoder is used to transform the input noisy speech waveform into speech features via an STFT operator, an input convolution block, a dilated-dense block, and an output convolution block. First, the noisy speech waveform s undergoes the STFT operator to obtain the spectrum with plural representation $Y = Y_{re} + jY_{im}$, where $Y_{re} \in \mathbb{R}^{B\times T\times F}$ and $Y_{im} \in \mathbb{R}^{B\times T\times F}$ denote the real and imaginary components of speech spectrum, respectively, and T and F mean the the dimension of time and frequency, respectively, and B is the batch size. Next, the real and imaginary parts are stacked together along the new channel dimension, constructing the input of model $S \in \mathbb{R}^{B\times 2\times T\times F}$. Then, the channel number of S is increased from 2 into 64 by the input convolution block using 64 $1*1$ convolution kernels attached with layer normalization and PReLU non-linearity activation. In the following, to obtain the rich feature representation, a dilated-dense block [16] is used to achieve the feature propagation and increase the receptive field with reduced parameters by using 4 depth-wise separable convolutions [21] with dense connection and dilated operation. Finally, the output convolution block compresses the channel number of the features from dilated-dense block to form the input of conformer blocks $H \in \mathbb{R}^{B\times 32\times T\times F}$ via 32 kernels of size $(1, 1)$ followed by PReLU activation operation, which aims to avoid the overflow of memory caused by the computational complexity.

B. Dual-path Conformer Block

The proposed dual-path conformer blocks consist of conformers and transformer-based fusors to extract the contextual information of speech features. The vanilla transformer has three important components such as positional encoding (PE), multi-head self-attention (MHSA), and feed-forward network (FFN). Compared with transformer, the conformer structure contains two macaron-like FFN layers sandwiching the convolution block and the MHSA [17], where the convolution block is inserted between MHSA and the second FFN for learning local information. More specifically, in the beginning, the input is processed by a half-step FFN consisting of two fully-connected (FC) layers for passing through MHSA. In MHSA, the input is first transformed into three different feature representations including Queries (Q), Keys (K) and Values (V) by three learnable linear matrixes with multiple times. Then, each query is multiplied by all keys in the dot-product manner, generating the attention weights after the softmax operation. Next, attention weights are multiplied with values to get the attention of each head. Finally, the attentions of all heads are packed together to form the output of MHSA

by a linear projection. In the following, a convolution module is employed to learn the local relationship of speech features, which is comprised of point-wise and depth-wise convolutions, normalization, and non-linear activation operations as shown in Fig. 2. Finally, the outputs from MHSA are processed by a second half-step FFN to extract abundant information. Due to the fact that the self-attention neglects the order information of sequential signals, the PE is adopted in the beginning of MHSA for adding position information.

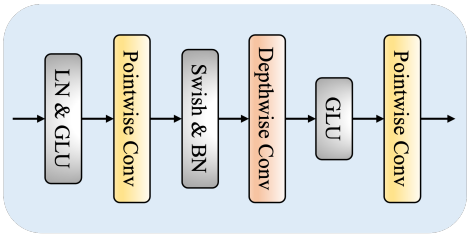


Fig. 2: Convolution module in conformer

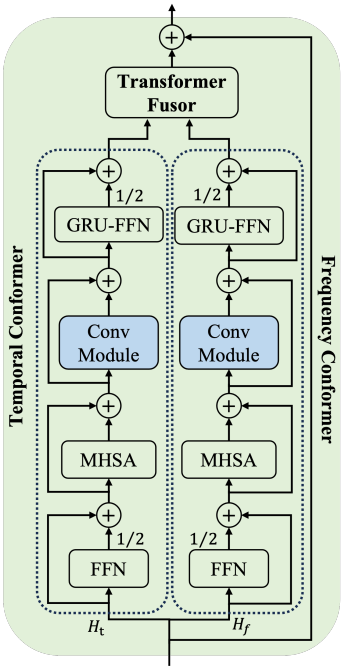


Fig. 3: Proposed dual-path conformer block (DPCB)

Each dual-path conformer block (DPCB) is comprised of a temporal conformer, a frequency conformer, and a transformer fusor as shown in Fig. 3, where the temporal and frequency conformers are employed to simultaneously extract temporal and frequency information of speech, and the transformer fusor is used to merge the generated temporal and frequency features for better contextual representation. Different from vanilla conformer and transformer, we here modify these two structures in some aspects: first, we omit the positional encoding module commonly used in conformer and transformer; Second, to learn the positional information, we employ the GRU-augmented FFN after MHSA in both conformer and transformer, where the first FC layer is substitute by a bi-directional GRU layer to track the positional information of speech signals.

During each DPCB procedure, given the compressed features $H \in \mathbb{R}^{B\times 32\times T\times F}$ generated from the encoder, the permutation first operates on the dimension of features to yield two types of features $H_t \in \mathbb{R}^{(B\times F)\times T\times 32}$ and $H_f \in \mathbb{R}^{(B\times T)\times F\times 32}$ with different dimension shape. Next, the frequency conformer operates on the H_f along the dimension F to extract the frequency dependency of each independent speech frame. Simultaneously, the temporal conformer performs on the H_t along the dimension T to aggregate the temporal dependency of various speech frames. Subsequently, the transformer fusor is used to merge the generated features from the temporal and frequency conformers by considering frequency features as queries, and temporal features as keys and values, yielding the contextual representation of speech features. The whole procedure in each DPCB can be denoted as follows:

$$Mid_{cf}(H_f) = Conv_{cf}(MHSA_{cf}(MLP_{cf}(H_f))) \quad (1)$$
$$Conf_f(H_f) = GRU\text{-}FFN_{cf}(Mid_{cf}((H_f))) \quad (2)$$
$$Mid_{ct}(H_t) = Conv_{ct}(MHSA_{ct}(MLP_{ct}(H_t))) \quad (3)$$
$$Conf_t(H_t) = GRU\text{-}FFN_{ct}(Mid_{ct}(H_t)) \quad (4)$$
$$H_{context} = Transformer(Conf_f(H_f), Conf_t(H_t)) \quad (5)$$

where $Conf_f$ and $Conf_t$ denote the frequency and temporal conformer, $GRU\text{-}FFN_{cf(ct)}$, $Conv_{cf(ct)}$, $MHSA_{cf(ct)}$ and $MLP_{cf(ct)}$ are the GRU-augmented FFN, convolution block, multi-head self-attention and vanilla FFN in frequency or temporal conformer, and $H_{context}$ is the contextual features from DPCB. Note that, in our implementation, the group normalization is added after each conformer for better training, and the residual connection is utilized in each DPCB to mitigate the gradient vanishing issue.

C. Decoder

The decoder receives the contextual features $H_{context} \in \mathbb{R}^{B\times 32\times T\times F}$ from DPCBs to estimate a mask which is used to filter out the noise-related information of noisy speech features. First, 64 1^*1 convolution kernels followed by PReLU activation are used to recover the feature channel of $H_{context}$ back to be the same as the one of H. Then, the values of features are then squeezed between 0 and 1 by adopting a gated convolution [22] consisting of a 1^*1 convolutions followed by a $Sigmoid$ activation function. Next, a dilated-dense block followed by 1^*1 convolution are employed to decrease the feature channel of $H_{context}$ to 2, creating a mask $M = [M_{re}, M_{im}] \in \mathbb{R}^{B\times 2\times T\times F}$, whose complex-valued

TABLE I: Comparison results of our proposed model with other methods on the dataset [23]

Model	Domain	PESQ	STOI(%)	CSIG	CBAK	COVL	Para.(Million)	MACs(G)
Noisy	-	1.97	91	3.34	2.44	2.63	-	-
SEGAN, 2017 [24]	T	2.16	93	2.48	2.94	2.80	97.47	9.47
Wave U-Net, 2018 [6]	T	2.40	-	3.52	3.24	2.96	10	2.45
DCUNet-16, 2019 [25]	T-F	2.93	-	4.10	3.77	3.52	2.3	21.01
PHASEN, 2019 [5]	T-F	2.99	-	4.21	3.55	3.62	-	-
NAAGN, 2020 [26]	T-F	2.90	-	4.13	3.50	3.51	-	-
DEMUCS, 2020 [7]	T	3.07	95	4.31	3.40	3.63	33.5	17.27
T-GSA, 2020 [14]	T-F	3.06	-	4.18	3.59	3.62	63	-
PFPL, 2020 [3]	T-F	3.15	95	4.18	3.60	3.67	-	-
TSTNN, 2021 [16]	T	2.96	95	4.33	3.53	3.67	0.92	75.52
CAUNet, 2021 [27]	T	2.96	95	4.22	3.53	3.60	0.92	21.43
MetricGAN+, 2021 [28]	T-F	3.15	-	4.14	3.16	3.64	-	-
SE-Conformer, 2021 [18]	T	3.13	95	4.45	3.55	3.82	-	-
CPTNN, 2022 [29]	T	3.07	95	4.40	3.59	3.76	0.76	-
FSCNet, 2022 [30]	T-F	2.94	-	4.26	3.57	3.68	0.90	-
SADNUNet, 2022 [31]	T	2.82	95	4.18	3.47	3.51	2.63	-
CMGAN, 2022 [19]	T-F	**3.41**	**96**	**4.63**	**3.94**	**4.12**	1.83	-
SEformer (ours)	T-F	3.19	95	4.42	3.62	3.84	**0.59**	28.94

representation is $M = M_{re} + jM_{im}$. In the following, the produced mask will be applied into the input $Y = [Y_{re}, Y_{im}]$ to obtain the enhanced real and imaginary spectrograms, which are eventually proposed by iSTFT operator for reconstruction of clean speech waveform. The overall masking procedure can be denoted as follows:

$$Y \cdot M = (Y_{re}M_{re} - Y_{im}M_{im}) + j(Y_{re}M_{im} + Y_{im}M_{re}) \quad (6)$$

$$E_{re} = Y_{re}M_{re} - Y_{im}M_{im} \quad (7)$$

$$E_{im} = Y_{re}M_{im} + Y_{im}M_{re} \quad (8)$$

$$x = iSTFT(E_{re}, E_{im}) \quad (9)$$

where M_{re} and M_{im} denote the masks for real and imaginary spectrograms, respectively, and E_{re} and E_{im} mean the real and imaginary of enhanced spectrogram, respectively, and $iSTFT$ is the inverse STFT operator, and x is obtained the enhanced speech waveform.

III. Experiments

A. Dataset

Our experiments are conducted on the VoiceBank-Demand dataset [23], a benchmark dataset commonly used for speech enhancement task, including 11572 utterances from 14 females and 14 male speakers for training, and 824 utterances from one male and one female speakers for testing. Both trainset and testset are comprised of pairs of noisy and clean speech, where the noisy speech of trainset is the mixture of clean speech of trainset and 10 types of noises (8 types of noises chosen from DEMAND dataset [32] and 2 types of artificial noises) at 0 dB, 5 dB, 10 dB, and 15 dB SNR levels, and the acquisition of noisy speech in testset is achieved by corrupting clean speech from testset with 5 types of noises not used in trainset at SNR levels of 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB.

B. Evaluation Metrics

To assess the quality of enhanced speech with respect to clean speech, some evaluation measurements are used to evaluate the model performance by operating on the enhanced and clean speech, including perceptual evaluation of speech quality (PESQ) [33] ranging from -0.5 to 4.5, and short-time objective intelligibility (STOI) [34] constrained between 0 to 1; In addition, three mean opinion scores (MOSs) [35] fluctuated between 1 and 5 are employed for subjective evaluation, like CSIG for signal distortion, CBAK for noise disturbance, and COVL for overall quality of enhanced speech.

C. Experimental Setup

All speech utterances are re-sampled to 16 KHz in both the trainset and testset. We randomly choose a 3-second speech clip from each utterance in the trainset, and maintain the length of testing utterances unchanged. To obtain the speech spectrum, a 320-point STFT operator applies a Hann window with a 20ms duration and a 10ms stride into each speech clip, resulting in the 161 frequency bins in the spectrum. Considering the hardware resource, we use 3 dual-path conformer blocks, where the input feature dimension is 32, and 4 attention heads and 64 hidden units are used in each multi-head attention block and GRU-augmented FFN block, respectively. By following the training strategy of [16], the Adam optimizer [36] is adopted to update model parameters for 100 epochs, in which the maximum L2 gradients is constrained within 5 for alleviating the explosion of gradient through gradient clipping. The learning rate is first warmed up to $4e^{-4}$, and then multiplied by 0.98 every two training epochs. For loss functions, we adopt a loss composed of the time and T-F domain to expect the proposed model to learn abundant knowledge from different aspects [16], which can be formulated as follows:

TABLE II: Evaluation results among reference models with various conformer blocks

Model	Domain	PESQ	STOI(%)	CSIG	CBAK	COVL	Para.(Million)	MACs(G)
SEformer (ours)	T-F	**3.19**	**95**	4.42	**3.62**	**3.84**	0.59	28.94
SEformer_A	T-F	3.13	**95**	4.42	3.61	3.80	**0.58**	27.89
SEformer_C	T-F	3.13	**95**	4.42	3.61	3.81	**0.58**	27.89
SEformer_V1	T-F	3.14	**95**	**4.43**	**3.62**	3.81	0.59	28.94
SEformer_D	T-F	3.01	**95**	4.30	3.53	3.67	0.73	**18.95**

$$Loss = \alpha * Loss_{time} + (1 - \alpha) * Loss_{frequency} \quad (10)$$

where $Loss_{time}$ is the mean square error (MSE) between the clean and enhanced speech waveforms in time domain, and $Loss_{frequency}$ is the frequency-domain loss to improve the speech intelligibility and perceptual quality, which calculates the distance between the clean and enhanced speech spectrum, α is a tuneable parameter to control the weight of two losses, which is configured as 0.6 in our all experiments.

D. Experimental Results

Table I exhibits the comparison results of our SEformer and some existing methods conducted on the VoiceBank-Demand dataset. Generally, our SEformer achieves a competitive performance to existing models in all evaluation measurements while involving the fewest model parameters (around 0.59 million). First of all, compared with all existing time-domain methods, our SEformer attains a superior PESQ improvement from 1.97 to 3.19. Besides, SEformer further has a 0.23 PESQ improvement based on TSTNN using two-stage feature extraction with cascaded structure while involves fewer model parameters and lower multiply-accumulate operations (MACs), presenting the superiority of proposed dual-path conformer with inner fusion mechanism. Compared with the SE-Conformer using vanilla conformer, the proposed SEformer yields better performance than most evaluation scores except for CSIG. In addition, SEformer achieves the highest STOI (95%) and MOSs scores compared with all time-domain methods.

Moreover, SEformer achieves competitive performance compared with state-of-the-art methods in T-F domain. More specifically, our SEformer achieves better evaluation results than T-GSA using the transformer with modified self-attention mechanism. Despite CMGAN achieves the better enhancement performance than SEformer, it adopts adversarial learning and possesses three times more parameters than SEformer. Differing from other methods, our SEformer is a simple but efficient enough architecture which is free from data augmentation technologies or complicated design.

Through observing the comparison results between proposed models and existing speech enhancement systems shown in Table I, our SEformer is efficient in modeling long-range speech sequences while having the lowest model parameters. This implies that our proposed dual-path conformer block with inner fusion mechanism can effectively capture the temporal and frequency information of speech sequences, and achieves the interaction between temporal and frequency features for attaining strong contextual information.

E. Ablation Study

We present some comparison models to demonstrate the effectiveness of various components of our model. First, to further explore the efficiency of proposed DPCBs, we construct two different conformer blocks, where the first one exclusively substitutes the transformer fusor with average operation named as DPCB_Avg, the second one uses the temporal and frequency conformers with cascaded structure named as DPCB_cascade. Second, to probe the sensitivity of transformer fusor on location of temporal and frequency features, we make temporal features as queries, and frequency features as keys and values for undergoing the transformer fusor, named as DPCB_V1. To make a fair comparison, we use 4 DPCB_Avg, 4 DPCB_cascade and 3 DPCB_V1 to separately replace the original DPCBs, leading to three reference models with the similar parameters to SEformer, namely SEformer_A, SEformer_C, and SEformer_V1, respectively, as shown in Table II. Finally, to investigate the effect of speech features with decreased spatial size on the performance, we introduce a down-sampling layer and an up-sampling layer into encoder and decoder, respectively, and adopt 4 DPCBs in middle, yielding the reference model called SEformer_D.

As shown in Table II, our SEformer using proposed DPCBs attains the best results in most evaluation metrics. Especially, both SEformer_C using cascaded structure and SEformer_A using averaging operation obtain the inferior performance than SEformer, indicating that dual-path extraction with conformer can efficiently extract temporal and frequency information and dynamically fuse them for better contextual representation of long-range speech sequences. In addition, by comparing with SEformer_V1, proposed SEformer achieves the similar evaluation scores, showing that the denoising performance of SEformer would not be affected by the position of temporal and frequency features in passing through the transformer fusor. Finally, despite the SEformer_D with down-sampling operation involves the fewer model parameters, its performance shows a significant drop in all evaluation metrics, which implies that adopting sufficient spatial resolution of speech features is crucial for conformer blocks to learn the contextual relationship among speech sequences.

IV. Conclusions

In short, we proposed an efficient SEformer system for T-F domain speech enhancement, which adopts the dual-path conformers with inner fusion mechanism to simultaneously extract temporal and frequency features for better contextual representation of speech sequences. The comparison results with recent state-of-art methods indicate that our SEformer achieves competitive performance and contains the fewest model parameters. In future work, we continue to modify the self-attention mechanism of conformer and transformer with reduced computations like linear-complexity attention [37], achieving the trade-off between the denoising performance and computational complexity. Besides, it is worth evaluating the performance and robustness of our proposed SEformer on more challenging datasets with low SNR environments.

References

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[2] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[3] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement," *arXiv preprint arXiv:2010.15174*, 2020.

[4] Y. Hu, Y. Liu, S. Lv, *et al.*, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[5] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[6] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[7] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[8] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6875–6879.

[9] L. Zhou, Y. Gao, Z. Wang, J. Li, and W. Zhang, "Complex spectral mapping with attention based convolution recurrent neural network for speech enhancement," *arXiv preprint arXiv:2104.05267*, 2021.

[10] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 181–185.

[11] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6649–6653.

[15] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[16] K. Wang, B. He, and W.-P. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7098–7102.

[17] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[18] E. Kim and H. Seo, "Se-conformer: Time-domain speech enhancement using conformer.," in *Interspeech*, 2021, pp. 2736–2740.

[19] R. Cao, S. Abdulatif, and B. Yang, "Cmgan: Conformer-based metric gan for speech enhancement," *arXiv preprint arXiv:2203.15149*, 2022.

[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[22] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, PMLR, 2017, pp. 933–941.

[23] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, 2016, pp. 146–152.

[24] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[25] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2019.

[26] F. Deng, T. Jiang, X. Wang, C. Zhang, and Y. Li, "Naagn: Noise-aware attention-gated network for speech enhancement.," in *Interspeech*, 2020, pp. 2457–2461.

[27] K. Wang, B. He, and W.-P. Zhu, "Caunet: Context-aware u-net for speech enhancement in time domain," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2021, pp. 1–5.

[28] S.-W. Fu, C. Yu, T.-A. Hsieh, *et al.*, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[29] K. Wang, B. He, and W.-P. Zhu, "Cptnn: Cross-parallel transformer neural network for time-domain speech enhancement," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[30] L. Cheng, J. Li, and Y. Yan, "Fscnet: Feature-specific convolution neural network for real-time speech enhancement," *IEEE Signal Processing Letters*, vol. 28, pp. 1958–1962, 2021.

[31] X. Xiang, X. Zhang, and H. Chen, "A nested u-net with self-attention and dense connectivity for monaural speech enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 105–109, 2021.

[32] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, Acoustical Society of America, vol. 19, 2013, p. 035 081.

[33] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Complex Feature Information Enhanced Speech Emotion Recognition

Pengcheng Yue, Shukai Zheng and Taihao Li †
Zhejiang Lab, China
E-mail: {pcyue, zhengsk, lith}@zhejianglab.com
† Corresponding author

***Abstract*—Speech emotion recognition (SER) helps machines behave and communicate in a more human-like way, thus has attracted extensive attention from industry and academia. Traditional SER algorithms usually discard the phase information in the speech spectrum. However, phase information plays an important role in many speech-related tasks according to recent speech-processing researches. In the SER area, the pre-trained Wav2Vec 2.0 model reached SOTA performance, still, no attention was paid to the complex information in speech. In this paper, we propose a novel framework utilizing both the real and the imaginary parts of the complex feature obtained by speech short-time Fourier transform to improve the SER performance. Specifically, we construct a deep complex neural network (DCNN) to handle the information in the speech complex feature, and then the DCNN is fused with the Wav2Vec 2.0 model to enhance the SER performance. Experiment results show that the proposed framework benefits from the compensation of DCNN and obtains nearly 1 percent performance improvement over model Wav2Vec 2.0 in terms of accuracy and F1.**

I. Introduction

Speech emotion recognition (SER) is an essential research branch of artificial intelligence. With SER, human-computer interaction becomes smoother. Besides, SER is also helpful for emergency discovery, fraud detection, psychological comfort, etc.

Speech in digital signal view is a complex wave composed of a bunch of sine waves with different frequencies and amplitudes. Therefore, short-time Fourier transform (STFT) was usually adopted to analyze the frequency domain feature information of the speech, which was proved useful fed to an emotion classifier [1][2][3][4]. There were many speech spectrum related features in SER research literature, where different spectral representations of speech were used as the input for deep learning models, including Mel-frequency Cepstral Coefficients (MFCCs), Mel-scaled spectrogram, Chromagram, Spectral contrast feature, logMel spectrogram and Tonnetz representation[5] [6] [7], etc. All these spectrum-related features could be extracted via Librosa audio library[8]. The spectrum feature based SER algorithms often used multi-time-scale method or multi-scale area attention with varied granularities to enhance the flexibility towards temporal variations[6][7]. In [9], Gammatone Frequency Cepstral Coefficients (GFCC) was proposed as a potentially better representation of speech signals for emotion recognition, which was also adopted in [10]. The key point for the success of GFCC was the usage of additional voice features. Many works [11][12][13][14][2] used MFCC as the input feature of a deep neural network to recognize the speech emotion and achieved good performance. In [15], short-time log frequency power coefficients (LFPC) was used to represent the speech signals. In previous SER works, most classifiers took features coming from the frequency spectrogram as the input, which contained only the amplitude part of the outcome of STFT, however, the phase information was neglected.

To compensate for the phase information ignored by frequency spectral features in audio-related tasks, deep complex neural network (DCNN) considering both the real and imaginary part of the speech complex feature was proposed in [16], where the building blocks and architectures for deep learning were based on complex-valued operations and representations. Based on this work, many researchers studied the advantage of complex deep neural networks in speech, music, or even biomedical signal processing tasks. In [17][18][19], models incorporating well-defined complex-valued building blocks were proposed to deal with complex-valued spectrograms to solve the difficulty of estimating the phase of clean speech during speech enhancement. In [20], several complex-valued activation functions were implemented for magnetic resonance imaging (MRI) reconstruction and that helped to achieve better performance than real-valued networks. Though complex deep neural networks are now extensively used in speech-related tasks, to the best of our knowledge, it has not been explored for the recognition of speech emotion. According to the experience of speech reconstruction, when a speech segment is modified with phase information, one can apparently hear the difference, which may influence the emotion of the speech. Inspired by this phenomenon, we investigate whether complex feature containing phase information is helpful for speech emotion recognition in this paper.

Recently, pre-trained speech models such as Wav2Vec 2.0 provided very effective acoustic feature extraction methods and were proven to have sound effects in many speech-related tasks. Wav2Vec was first proposed in [21], which trained the model on large amounts of unlabeled audio data and the resulting representations were then used to improve acoustic model training. Then Wav2Vec 2.0 was proposed in [22] which used a self-supervised learning approach to learn audio representations and trained both continuous context representations and discrete quantized representations. Wav2vec and Wav2vec 2.0 both provided outstanding pre-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

training methods for speech, but Wav2vec 2.0 was proved superior to Wav2vec in TIMIT Phoneme recognition and other related tasks. Besides, Wav2Vec 2.0 was also helpful in SER[23][24]. Particularly, [25] improved the performance of SER on IEMOCAP to 70.83 percent in terms of UAR, which we take as the benchmark of our SER study.

In this work, we use the DCNN model which takes all the speech complex features into account as a supplement to Wav2vec 2.0 in the task of speech emotion recognition. The contribution of the work is concluded as follows: 1) This work is the first to emphasize and use the speech complex feature information in the task of SER to the best of our knowledge; 2) The proposed SER framework reaches the SOTA performance on IEMOCAP and RAVDESS and brings nearly 1 percentage performance improvement in terms of accuracy and F1.

II. Method

In this section, the SER model based on a deep complex neural network is firstly introduced. We then review Vanilla fine-tuning Wav2Vec 2.0[25], which is the baseline of our work and serves as one of the basic models in the proposed model. Finally, we introduce the details of the proposed SER framework, namely an enhanced Wav2Vec 2.0 model, which inherits the advantages of high-quality speech representation of Wav2Vec 2.0 and the full use of phase information of the deep complex neural network.

A. Deep Complex Neural Network (DCNN)

In [16], the deep complex neural network was proposed to process speech signals, and was proved beneficial for tasks such as music processing, speech enhancement, etc. In this work, we try to implement DCNN in speech emotion recognition, which has not been investigated in previous SER works.

In the DCNN of this work, a deep complex convolution network is implemented to extract the speech features and a linear layer is followed to classify the emotion category. The detailed architecture is pictured in Fig. 1. Firstly, a short-time Fourier transform (STFT) module is used to deal with 20 ms speech fragments and obtain the complex spectral feature which contains the real part and the imaginary part. Then the same complex convolution block modules are concatenated 5 times to extract more expressive features from the output of the STFT module. Then follows a global average pooling module to fuse the information coming from different speech fragments into a specified length, which is 256 in our work. Finally, a linear module is taken as the classifier, of which the output length is consistent with the number of emotion categories. Concretely, the complex convolution block contains three function layers: convolution2d, LeakyReLU, and batch normalization layer, which follow the rules of complex-valued calculation and are presented below.

B. Complex-valued Convolution

Suppose we are convolving a complex filter matrix $\mathbf{M} = \mathbf{X}+i\mathbf{Y}$ by a complex vector $\mathbf{v} = \mathbf{a}+i\mathbf{b}$, where $\mathbf{X}$, $\mathbf{Y}$ are real-valued matrix and $\mathbf{a}$, $\mathbf{b}$ are real-valued vectors. The outcome of the complex convolution is

$$\mathbf{M} * \mathbf{v} = (\mathbf{X} * \mathbf{a} - \mathbf{Y} * \mathbf{b}) + i(\mathbf{Y} * \mathbf{a} + \mathbf{X} * \mathbf{b}). \quad (1)$$

This equation gives an important role to associate real convolution with complex convolution. The implementation details are presented in Fig. 2.

C. Complex Global Averaging

Global averaging is an operation to reduce the dimension of the extracted features to a more expressive vector, which also helps to unify the speech feature size to the channel size no matter what length the incoming speech is. In real-valued global averaging, it is simply calculating the average value of each channel. But for complex-valued global averaging, the average values of both real and imaginary parts are calculated separately. Besides, the complex global averaging in our work is followed by the linear classifier, of which the input should be a real-valued vector. So, the complex global averaging module is where we combine the real part and the imaginary part of the complex feature. After global averaging calculation, we calculate the real-valued power of the complex value of each channel.

The other modules of the complex-valued neural network such as complex LeakyReLU are of simple calculation, which is performed separately on the real and imaginary parts of the complex value.

D. Vanilla Fine-tuning Wav2Vec 2.0 on SER

In [25], the authors implemented Vanilla fine-tuning Wav2Vec 2.0 and reached the SOTA performance in the field of SER. The Wav2Vec 2.0 model extracts unsupervised speech features by multi-layer convolutional neural network taking raw speech signals as the input. Under the default setup in Wav2Vec 2.0 encoder, the CNN contains 7 layers of which the step sizes are (5,2,2,2,2,2) and the convolution kernel widths are (10,3,3,3,2,2). As the number of output channels of the last convolution in the encoder is set to 512, a 512-dimensional speech feature is generated for every 20ms speech. To utilize Wav2Vec representation to speech emotion recognition, Vanilla fine-tuning Wav2Vec 2.0 processes the final contextualized embedding generalized by Wav2Vec 2.0 using a global average pooling across the time dimension, and then classifies the emotion using the ReLU activation followed by a full connection layer. Concretely, the global average pooling unifies the length of the speech feature embedding extracted by the Wav2Vec 2.0 model to 768 regardless of the input speech length.

E. Complex feature information enhanced Wav2Vec 2.0

Fig. 1 presents the framework of the proposed model, which is composed of model Wav2Vec 2.0 and model DCNN. The two basic models both take a speech signal as the input and the recognized emotion category probability as the output. Though Wav2Vec 2.0 is of strong representation ability, it does not pay attention to the phase information of the speech, which is exactly what the DCNN model can compensate for.

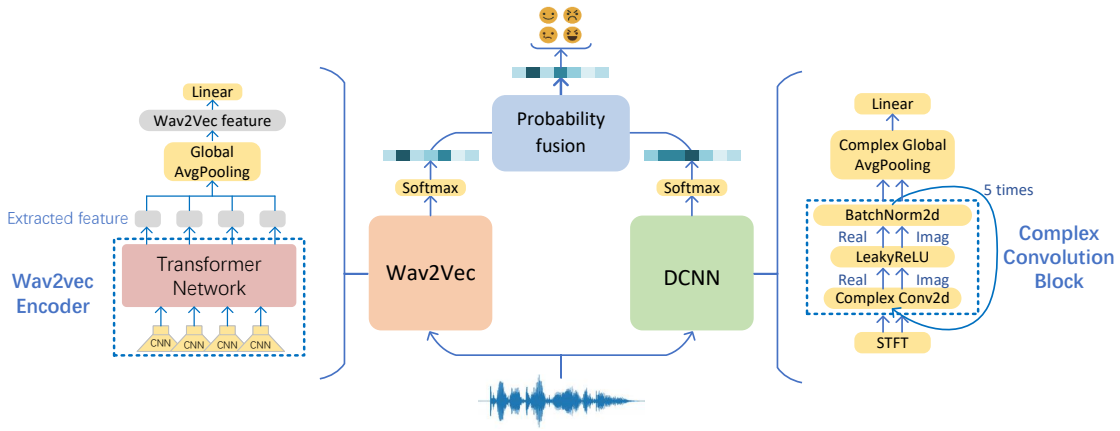


Fig. 1. Framework of the proposed complex feature information enhanced Wav2Vec 2.0

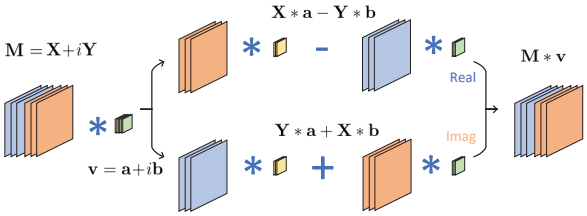


Fig. 2. Complex convolution

Specifically, for a certain speech signal input, the DCNN model provides the probability of the emotion category, $Prob_{DCNN}$, which is obtained by a softmax layer after the output of the linear classifier, $output_{DCNN}$. The calculation is as follows:

$$Prob_{DCNN} = softmax(output_{DCNN})$$

The outcome probability of DCNN $Prob_{DCNN}$ is then fused with the outcome probability of Wav2Vec 2.0, $Prob_{w2v}$ in the proportion of α:

$$Prob_{Fusion} = Prob_{w2v} + \alpha Prob_{DCNN}$$

The compensated outcome probability of the model $Prob_{Fusion}$ is then used to decide on the emotion category, Emo:

$$\begin{aligned} Emo &= argmax(Prob_{Fusion}) \\ &= argmax(Prob_{w2v} + \alpha Prob_{DCNN}) \end{aligned}$$

Therefore, the proposed enhanced Wav2Vec 2.0 model can better consider the phase information corresponding with both the real and imaginary parts of the speech complex feature when recognizing speech emotion.

III. Experiments and Results Discussion

This section introduces the datasets used, and then the experimental setup is elaborated, after that the experiments and the results are presented and discussed.

A. Dataset

Two datasets in the field of speech emotion recognition, IEMOCAP and RAVDESS, were considered to prove the advance of our proposed method. The two datasets are described below.

IEMOCAP. Interactive Emotional Dyadic Motion Capture is a popular dataset for evaluating SER systems. It contains five recording sessions, each with one male speaker and one female speaker. Following the popular setup with previous SER works, we used the default labels provided by IEMOCAP, except that the "excited" category was merged with "happy" due to its sparsity in the dataset. Finally, only four emotion categories were considered: neutral, sad, angry, and happy. After that, the amounts of examples labeled 'neutral', 'happy', 'anger', 'sad' were 1708, 1636, 1103, 1084, respectively. The total amount of speech in IEMOCAP is about 7 hours. In this work, we only used the speech modality.

RAVDESS. Ryerson Audio-Visual Database of Emotional Speech and Song dataset is a validated multi-modal dataset of emotional speech and song. This gender-balanced dataset consists of 24 professional actors, each performing 104 unique vocalizations with emotions that include: happy, sad, angry, fearful, surprise, disgust, calm, and neutral[26]. Each actor enacted 2 statements for each emotion: "Kids are talking by the door" and "Dogs are sitting by the door." These statements were also recorded in two different emotional intensities, normal and strong, for each emotion, except for neutral (normal only). Actors repeated each vocalization twice. There are a total of 1440 speech utterances and 1012 song utterances. In this work, we only used the speech part.

B. Experimental setup

We cross-validated the models on the two datasets. As for IEMOCAP, we split the dataset by leaving one session out as the test set, and the remaining four sessions were used for training. As for RAVDESS, we split the dataset into six equal subsets according to the speaker ID and chose one subset as the test set while the rest subsets as training data.

Under this setup, we performed every experiment with multi-fold cross-validation. The experimental results presented the average value of cross-validation. Four metrics were used to evaluate the performance, including unweighted accuracy rate (UAR), weighted accuracy rate (WAR), micro-F1, and macro-F1.

We carried out three experiments in this work. Experiment 1 compared several models, including deep complex neural network (DCNN), Wav2Vec 2.0 (w2v) and the enhanced w2v model (EnW) by the compensation of DCNN, which utilized both the real and the imaginary parts of the speech complex feature. The structure of model EnW was presented in Fig. 1, and the fusion proportion α in EnW was set to 1 in our experiments. As for model w2v, we followed the Vanilla fine-tuning Wav2Vec 2.0 method described in [25]. Experiment 2 was the ablation experiment carried out to check the SER performance when only using the real part or the imaginary part information of the speech complex feature. For convenience, DCNN model using only the real part or the imaginary part information of the complex feature was respectively represented by DCNN_R and DCNN_I, and EnW_R and EnW_I was short for the enhanced w2v model by the compensation of DCNN_R and DCNN_I. Model DCNN_R and DCNN_I were similar in structure with DCNN, but DCNN_R and DCNN_I did not need any complex operation and only extracted the speech embedding from the real or imaginary part of the speech complex feature using convolutional neural network. In experiment 3, we also compared the proposed complex feature information enhanced model EnW with the SOTA models in SER, including WavLM Large, HuBERT Large, and data2vec Large of which the performance and model size all came from SUPERB [27]. The models compared in one table were under the same experimental setup.

C. Experimental results and discussion

Experiment 1 explored the SER performance of DCNN and w2v, as well as the proposed model EnW in order to see the auxiliary effect of speech complex feature on emotion recognition. Then in experiment 2, we carried out ablation experiments where half of the speech complex feature was neglected to reflect the importance of using all speech complex information in SER. In experiment 3, we compared the proposed complex feature information enhanced SER model, EnW with the SOTA models in terms of UAR and parameter quantity used.

Table I and Table II show the speech emotion recognition results of models using speech complex feature on IEMOCAP and RAVDESS respectively. One can observe that the w2v model pre-trained on very large speech datasets outperforms the DCNN model which is only trained on the limited training set of certain datasets. Nevertheless, the proposed model EnW achieves more superior performance than w2v model. This reveals that although the DCNN model is not so desirable in SER, the complex feature information extracted by DCNN is still helpful to improve the SER performance.

Table III and Table IV show the results of ablation experiments only using half part of the complex feature on IEMOCAP and RAVDESS. One can observe that: 1) The model DCNN_R or DCNN_I using only the real or the imaginary part of the complex feature information could not perform better than DCNN which uses all the speech complex feature; 2) EnW_R or EnW_I outperforms w2v; 3) Models using half part of the complex feature have similar performance. Specifically, DCNN_R and DCNN_I have similar performance, so are EnW_R and EnW_I; 4) EnW_R or EnW_I neglecting half part of the speech complex feature still fails to meet the performance of EnW. The results reveal that the real part and the imaginary part of the speech complex feature contain equally important information and both real and imaginary part are necessary to achieve superior performance for the SER model.

TABLE I
SPEECH EMOTION RECOGNITION RESULTS OF MODELS USING SPEECH COMPLEX FEATURE ON IEMOCAP

Model	UAR	WAR	micro-F1	macro-F1
DCNN	0.5494	0.5225	0.5557	0.5094
w2v	0.7083	0.6945	0.7055	0.6953
EnW	0.7146	0.7011	0.7124	0.7021

TABLE II
SPEECH EMOTION RECOGNITION RESULTS OF MODELS USING SPEECH COMPLEX FEATURE ON RAVDESS

Model	UAR	WAR	micro-F1	macro-F1
DCNN	0.3501	0.3806	0.3017	0.2707
w2v	0.7624	0.7639	0.7727	0.7609
EnW	0.7721	0.7729	0.7809	0.7704

It should be noted that although the real part and the imaginary part of the speech complex feature are different, the model using the real part of the complex information has a similar performance with the one using the imaginary part. This can be explained by the similarity of the two parts in a large-scale receptive field. Fig. 3 shows the comparison of the real part and the imaginary part of the STFT result of a speech. It can be observed that the two parts are different in detail, but are similar on a larger scale. In fact, the real part and the imaginary part are both correlated with the spectrum, which means a large number in the real part is likely to correspond to a large imaginary number, and vice versa.

The superior performance of the proposed model EnW in experiment 1 and experiment 2 are consistent on the two datasets, which shows that EnW has certain generalization performance, and complex feature containing both the real part and the imaginary part really has an auxiliary effect on SER.

Table V shows the comparison of our proposed model EnW with the existing SOTA pre-trained model in SER. The results come from [27] where the experimental setup is the same as in our work. The comparison terms are limited because the experiment in [27] was only carried out on IEMOCAP in terms of UAR. It can be observed that the accuracy of SER of our model EnW outperforms the other SOTA models and uses much fewer parameters at the same time. The superior performance is inseparable from the use of the whole speech

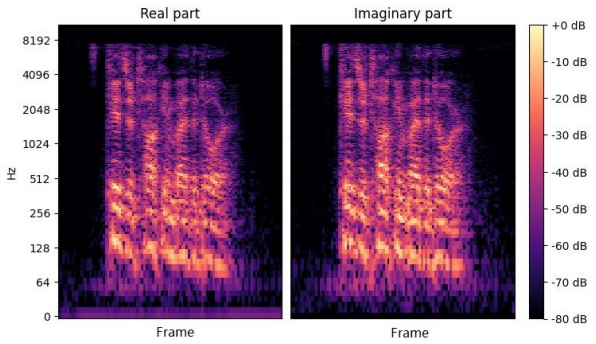


Fig. 3. Comparision of the real part and the imaginary part of the speech complex feature

TABLE III
SPEECH EMOTION RECOGNITION RESULTS OF MODELS ONLY USING HALF PART OF THE COMPLEX FEATURE ON IEMOCAP

Model	UAR	WAR	micro-F1	macro-F1
$DCNN_R$	0.4853	0.4327	0.4604	0.4119
$DCNN_I$	0.4847	0.4307	0.4595	0.4107
DCNN	0.5494	0.5225	0.5557	0.5094
EnW_R	0.7112	0.6963	0.7084	0.6963
EnW_I	0.7108	0.6957	0.708	0.6959
EnW	0.7146	0.7011	0.7124	0.7021

complex feature.

IV. CONCLUSION

Traditional SER works ignored the complex feature information in speech and that could lead to insufficient use of information. This paper proposes a method using complex feature information to enhance SER performance, where the deep complex neural network is constructed to compensate Wav2Vec 2.0 model. Experiment results demonstrate that the proposed method achieves the best performance compared with other SOTA methods, which proves the auxiliary effect of the complex feature in SER. Besides, the ablation experiment results indicate that the real part and imaginary part of the speech complex feature are equally necessary to achieve a superior SER performance.

In future work, we will explore other information fusion methods to enhance Wav2Vec 2.0 using speech the complex feature in view of the good SER performance of the proposed model in this paper.

TABLE IV
SPEECH EMOTION RECOGNITION RESULTS OF MODELS ONLY USING HALF PART OF THE COMPLEX FEATURE ON RAVDESS

Model	UAR	WAR	micro-F1	macro-F1
$DCNN_R$	0.3098	0.3306	0.2811	0.2382
$DCNN_I$	0.3101	0.331	0.2814	0.2385
DCNN	0.3501	0.3806	0.3017	0.2707
EnW_R	0.7656	0.7674	0.7760	0.7645
EnW_I	0.766	0.7678	0.7763	0.7649
EnW	0.7721	0.7729	0.7809	0.7704

TABLE V
SPEECH EMOTION RECOGNITION COMPARISON OF DIFFERENT PRE-TRAINED MODELS ON IEMOCAP

Model	UAR	Params
WavLM Large	0.7062	316.6M
HuBERT Large	0.6762	316.6M
data2vec Large	0.6631	314.3M
w2v	0.7083	96.5M
EnW	0.7146	97.4M

ACKNOWLEDGMENT

The work is supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2021ZD0114303), and Grant No. 2020KB0AC01 from the Scientific Project of Zhejiang Lab.

REFERENCES

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[2] S. Kuchibhotla, H. D. Vankayalapati, R. Vaddi, and K. R. Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 401–408, 2014.

[3] E. Wong and S. Sridharan, "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*. IEEE, 2001, pp. 95–98.

[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Detection of stress and emotion in speech using traditional and fft based log energy features," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, vol. 3. IEEE, 2003, pp. 1619–1623.

[5] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[6] E. Guizzo, T. Weyde, and J. B. Leveson, "Multi-time-scale convolution for emotion recognition from speech audio signals," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6489–6493.

[7] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6319–6323.

[8] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[9] G. K. Liu, "Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech," *arXiv preprint arXiv:1806.09010*, 2018.

[10] H. Patni, A. Jagtap, V. Bhoyar, and A. Gupta, "Speech emotion recognition using mfcc, gfcc, chromagram and rmse features," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2021, pp. 892–897.

[11] K. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using mfcc and wavelet features," in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 842–847.

[12] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using mfcc," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017, pp. 2257–2260.

[13] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using mfcc and support vector machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 2016, pp. 1080–1084.

[14] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[15] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[16] T. Chiheb, B. Olexa, S. Dmitriy, S. Sandeep, F. S. João, M. Soroush, R. Negar, B. Yoshua, and J. P. Christopher, "Deep complex networks," in *International Conference on Learning Representations(ICLR)*, 2018.

[17] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[18] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[19] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6885–6889.

[20] E. Cole, J. Cheng, J. Pauly, and S. Vasanawala, "Analysis of deep complex-valued convolutional neural networks for mri reconstruction and phase-focused applications," *Magnetic resonance in medicine*, vol. 86, no. 2, pp. 1093–1109, 2021.

[21] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[24] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[25] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," *arXiv preprint arXiv:2110.06309*, 2021.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[27] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Incorporating Pinyin into Pipeline Named Entity Recognition from Chinese Speech

Min Zhang, Xiaosong Qiao, Yanqing Zhao, Chang Su, Yinglu Li, Ming Zhu,
Junhao Zhu, Yuang Li, Xiaofeng Zhao, Yilun Liu, Wenbing Ma, Mengyao Piao, Jiawei Yu, Xinglin Lv,
Song Peng, Shimin Tao, Hao Yang and Yanfei Jiang
Huawei Translation Services Center, Beijing, China
E-mail: {zhangmin186, qiaoxiaosong, zhaoyanqing, suchang, liyinglu, zhuming47, zhujunhao, lyuang3, zhaoxiaofeng, liuyilun3, mawenbing, piaomengyao1, yujiawei19, lvxinglin, pengsong2, taoshimin, yanghao30, jiangyanfei}@huawei.com

***Abstract*—Named Entity Recognition (NER) from speech is usually implemented through a two-step pipeline that consists of (1) processing the audio using an Automatic Speech Recognition (ASR) system and (2) applying an NER tagger to the ASR output. In this paper, we incorporate *pinyin*[1] — spelled sounds of Chinese characters — into the pipeline NER from Chinese speech, aiming to improve the NER performance through two steps. First, we take the pretrained model ChineseBERT to embed pinyin, as well as glyph and char, as the input of NER tagger. Second, we introduce homophone noises into training data for NER tagger, as homophone errors most likely exist in ASR output for Chinese speech. Using the two-step pipeline method with pinyin incorporated into the NER tagger, the $F1$ score is improved by nearly 1% absolute points in the experiment on the AISHELL-NER dataset, which is a significant improvement in the field of NER. And the $F1$ score outperforms the current state-of-the-art (SOTA) result on the AISHELL-NER dataset by 0.4% absolute points, despite the slightly worse Character Error Rate (CER) of our ASR.**

I. Introduction

Named Entity Recognition (NER) is a fundamental information extraction task and plays an essential role in Natural Language Processing (NLP) applications such as information retrieval, question and answering, automatic text summarization, machine translation and knowledge graphs [1]. The goal of NER is to extract predefined specific entities from sentences and identify their correct types, such as *person* (PER), *location* (LOC), *organization* (ORG). NER has been well studied in the context of written languages [2].

Recently, NER from speech has been actively studied in the Spoken Language Understanding (SLU) field [3]. As illustrated in Fig. 1, it is usually implemented through a two-step pipeline that consists of (1) processing the audio using an Automatic Speech Recognition (ASR) system and (2) applying an NER tagger to the ASR output. Although the works in [4]–[6] have shown the capability of the End-to-End (E2E) approach for NER from French, English and Chinese speech, the two-step pipeline NER approach achieves the best $F1$ score on the Chinese speech dataset AISHELL-NER [6].

In Chinese, its characters are a written form of an ideographic symbol, and pinyin is a romanization system that uses the Latin alphabet to spell Chinese characters phonetically. For example, 妈 (means mum) is a Chinese character, and `ma1` is its pinyin, where the number 1 denotes the first tone among the four tones used in pinyin.

In this paper, we incorporate *pinyin* of Chinese characters into pipeline NER from Chinese speech. First, in order to embed the pinyin as part of the input of NER taggers, we use the off-the-shelf pretrained model ChineseBERT [7] as the encoder. ChineseBERT incorporates *pinyin*, as well as *glyph* and *char*, of Chinese characters during language model pretraining on large-scale unlabeled Chinese corpus, and achieves state-of-the-art (SOTA) performances on a wide range of Chinese NLP tasks. Second, as the training data for the NER tagger is almost error free while the ASR output contains errors or

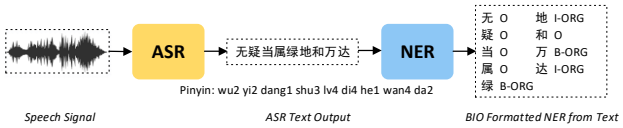


Fig. 1. Two-step pipeline NER from Chinese speech. The text of this ASR output has 9 Chinese characters with 2 named entities (绿地 for Greenland and 万达 for Wanda, which are the names of two Chinese companies), and the text means "Undoubtedly Greenland and Wanda".

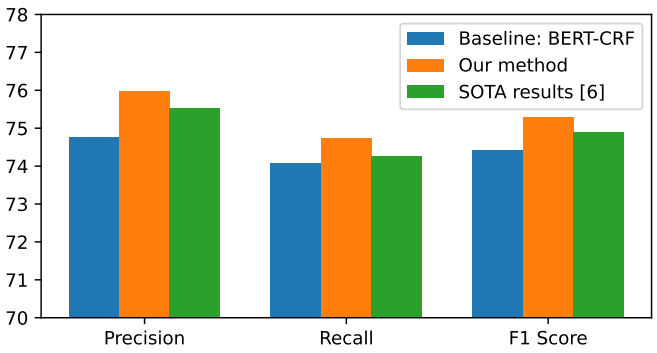


Fig. 2. Precision, Recall and $F1$ scores of the baseline pipeline method, our method and the SOTA results on the test dataset of AISHELL-NER [6].

[1]https://en.wikipedia.org/wiki/Pinyin

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

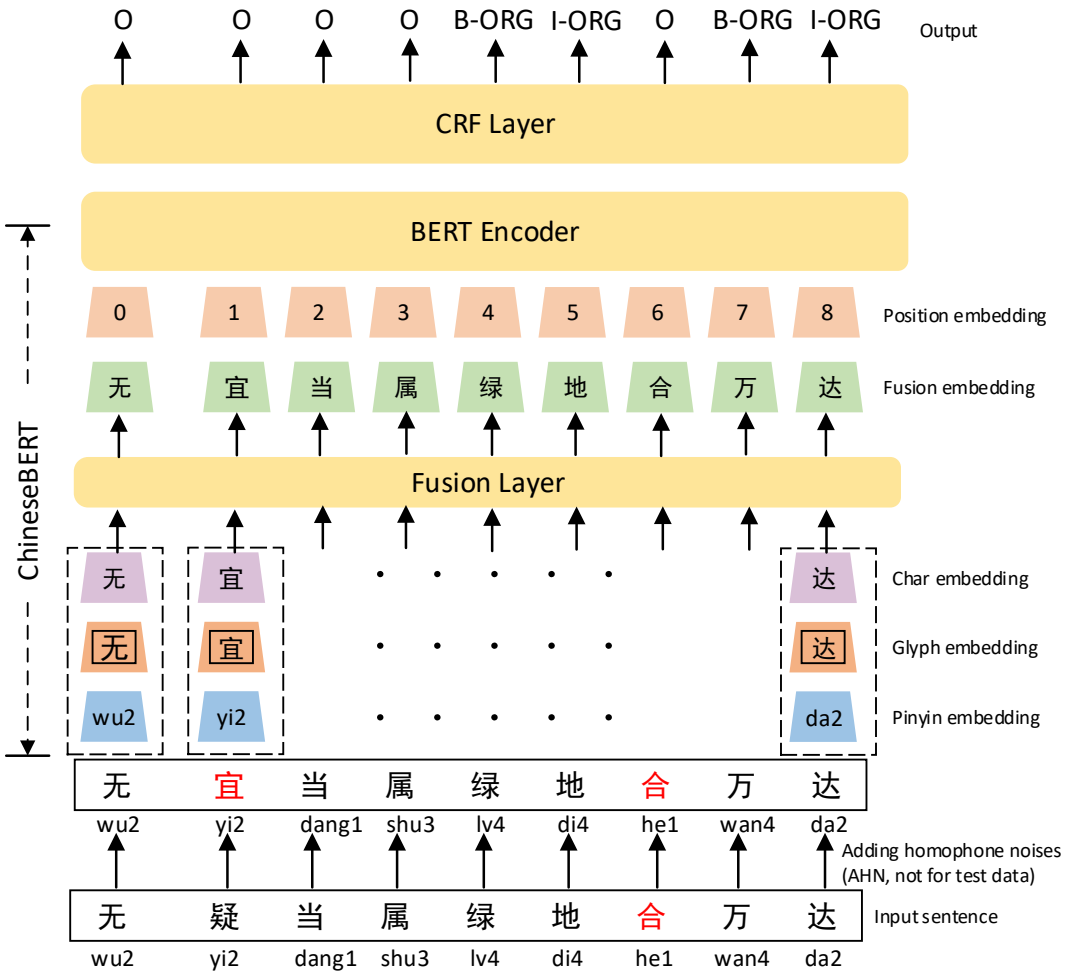


Fig. 3. NER tagger model architecture with three components AHN (adding homophone noises), ChineseBERT and CRF.

noises, we introduce homophone noises[2] into the training data to improve the robustness of the NER tagger.

To verify the effectiveness of our method, we adopt the classical ASR architecture Conformer [8] and NER architecture BERT-CRF where BERT is replaced by ChineseBERT, and test the two-step pipeline method by adding homophone noises (AHN) on the AISHELL-NER dataset [6], which is publicly available for NER from Chinese speech with three Named Entity (NE) categories (PER, LOC and ORG). Experimental results demonstrate that the $F1$ score is improved by nearly 1% absolute points (a significant improvement in the field of NER) with *pinyin* and outperforms the SOTA result in [6] by 0.4% absolute points (despite the slightly worse CER of our ASR outputs), as illustrated in Fig. 2.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose to use *pinyin* for NER from Chinese speech.
- Besides the off-the-shelf pinyin embedding method, we propose to introduce the homophone noises into the NER training data to make better use of pinyin.
- By using pinyin information, our two-step pipeline method outperforms the best result on the dataset AISHELL-NER by 0.4% absolute points in $F1$ score.

II. Related Work

NER from speech could be divided into two categories [4]–[6], i.e., two-step pipeline approach and E2E approach.

The two-step pipeline approach consists of processing the given audio on an ASR system and then feeding the transcribed ASR outputs to the NER tagger, as illustrated in Fig. 1. Although ASR and NER have been well studied in their respective fields [8]–[12], there is little research on how to better combine ASR and NER for better results.

[2]Homophone errors are very common in Chinese ASR output because a great number of Chinese characters have the same pronunciation but different written forms.

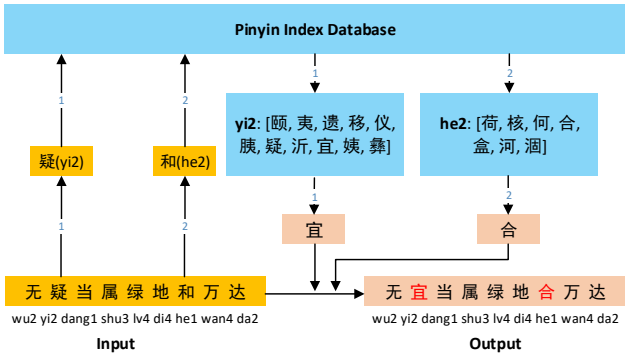


Fig. 4. Example of adding homophone noises for a Chinese sentence: the two characters 宜 (yi2) and 帝 (di4) are used to replace the two selected characters 疑 (yi2) and 地 (di4) to add homophone noises.

Lately, the E2E approach for NER from French [4], English [5] and Chinese [6], [12] speech has been proposed. Entity-aware ASR (EA-ASR) is adopted to predict NEs during the decoding process, where special tokens are added to the ASR vocabulary list to annotate NEs in the transcriptions, such as [] for PER, () for LOC, and <> for ORG. However, this method requires annotating new datasets where new NE types or new languages are introduced, and cannot directly use the existing annotation data for ASR and NER. More importantly, it is pointed out in [6] that the two-step pipeline approach gets better results than the E2E approach for NER from Chinese speech.

III. Methodology

As illustrated in Fig. 1, the two-step pipeline method (ASR + NER Tagger) is used in this paper, and the details are described as follows.

A. ASR

E2E ASR systems based on neural networks have seen large improvements in recent years. Since Transformer and Convolution Neural Network (CNN)-based model are more promising than RNN/LSMT for ASR on various corpora [13], [14], we build the ASR components with convolution-augmented Transformer-based encoder-decoder architecture [8], which has achieved SOTA accuracy in a list of benchmark datasets such as LibriSpeech. The encoder is enhanced with Convolution Neural Network (CNN), namely, Conformer [8], which maps the input sequence of audio features to a sequence of hidden states by a convolution subsampling layer and a number of conformer blocks. A conformer block is composed of four modules stacked together, i.e, a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module at the end. The decoding process is enhanced with the joint CTC/attention approach and the Transformer-based Language Model (LM).

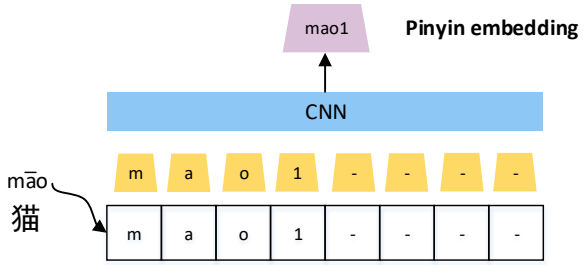


Fig. 5. Example for inducing the pinyin embedding of Chinese character 猫 (cat) with pinyin mao1 (a CNN with width 2 is applied to the sequence of Romanized pinyin letters, followed by max-pooling to derive the final pinyin embedding) [7].

B. NER Tagger

The NER tagger model architecture is illustrated in Fig. 3, which consists of three components: AHN (adding homophone noises), ChineseBERT, and CRF.

1) AHN: First, we generate a pinyin index database for 3,500 common Chinese characters[3] with the open-source pypinyin[4] package. For each entry in the database, its key is a pinyin, and its value is a list of Chinese characters with the same pinyin as the key. Fig. 4 shows two such entries with key values yi2[5] and he2[5] and with the corresponding value lists [颐, 夷, 遗, 移, 仪, 胰, 疑, 沂, 宜, 姨, 彝] and [荷, 核, 何, 合, 盒, 河, 涸]. The pinyin for the characters in the first list is yi2, and that for those in the second list is he2.

Second, we randomly select characters from a given sentence with a specified probability, and search the pinyin index database with the pinyin of the selected characters. Then, for each selected character, it will be replaced by a random selection from the searched entry value. So we achieve the addition of homophone noises in the given sentence. Fig. 4 provides an example for this process: the two characters 疑 and 和 are randomly selected from the given sentence, and are replaced by the two characters 宜 and 合 from the two entries with key values yi2 and he2. Since the two Chinese characters (宜 and 和) have the same pinyin as but different written forms from the two characters (疑 and 合), the homophone noises are successfully added for the given sentence, i.e., from input 无疑当属绿地和万达 to output 无宜当属绿地合万达.

It should be pointed out that the characters in NEs of the given sentence need not be selected for AHN, since homophonic NEs are not considered correct. And it should be noted that AHN is only used for training and development data, not for test data.

2) ChineseBERT: ChineseBERT is a pretrained Chinese language model which incorporates both glyph and pinyin embeddings of Chinese characters into model training [7]. The

[3]https://github.com/kaienfr/Font/tree/master/learnfiles
[4]https://pypi.org/project/pypinyin/
[5]2 denotes the second tone used in pinyin.

TABLE I
STATISTICS OF AISHELL-NER

	# Sentence	w/ NE	# PER	# LOC	# ORG
Train	120,098	40,839	15,842	20,693	21,455
Dev	14,326	5,173	1,902	2,588	2,731
Test	7,176	2,376	898	1,330	1,165

TABLE II
CER (%) OF CONFORMER ASR

CER (%)	Dev	Test
Conformer ASR (our)	4.59	4.99
Conformer ASR [6]	4.42	4.75

TABLE III
CER (%) ON THE SENTENCES WITH AND WITHOUT NES

	Number	CER (%)
Sentences with NEs	2,376	6.73
Sentences without NEs	4,800	4.02

glyph embedding is obtained based on different fonts of a Chinese character, and the pinyin embedding is obtained by a CNN model with width 2 on the pinyin sequence of a Chinese character. Fig. 5 provides a pinyin embedding example for Chinese character 猫 (cat) with pinyin `mao1`.

As illustrated in Fig. 3, for each Chinese character, its char embedding, glyph embedding and pinyin embedding are first concatenated, and then mapped to a D-dimensional embedding through a fully connected layer to form the fusion embedding. The fusion embedding is then added with the position embedding, which is fed as input to the BERT model.

As it is common that a character in Chinese ASR outputs is wrong but its pinyin is correct, we use ChineseBERT as the encoder for our NER tagger to improve its robustness.

3) CRF: Conditional Random Field (CRF) [15] is very effective for NER in both statistical machine learning and deep learning. Therefore, after obtaining all the character representations from ChineseBERT for a given sentence, we feed the representations into a CRF output layer, which can be seen as a char-level classifier over the NER label set.

IV. EXPERIMENTS

In this section, we evaluate the performance of our two-step pipeline method for NER from Chinese speech.

A. Experimental Setup

All the experiments are carried out on AISHELL-NER [6] dataset[6], which is built upon the open-source Chinese speech corpus AISHELL-1 [16] (150 hours training set, 10 hours development set and 5 hours test set). Table I gives the numbers of sentences that contain NEs and the detailed statistics of NE numbers for each category in AISHELL-NER.

The Conformer ASR model is built with WeNet [17] and makes use of both the joint CTC/attention decoding and LM. The parameter settings are same with [6]. The encoders have 4 attention heads with 2048 linear units. They stack 12 layers of encoder blocks, and their output size is 256. The decoders use the same configuration, except that they have only 6 layers. SpecAugment is used to randomly apply time and frequency masking blocks to log mel-filterbank features. Speed perturbation with factors of 0.9, 1.0 and 1.1 is also applied to Conformer, resulting in 3-fold data augmentation. The CNN module of the Conformer encoder uses a kernel size of 15. These ASR models are trained with the Adam optimizer, with the learning rate set to 0.0005 and a linear warmup scheduler. During decoding, the width of the beam search is set to 5.

TABLE IV
NER RESULTS FOR AISHELL-NER

		Precision	Recall	$F1$
Ground-Truth Text	BERT-CRF	**94.45**	**93.75**	**94.10**
Best result in [6]		75.54	74.27	74.90
Conformer ASR	BERT-CRF	74.77	74.08	74.42
Conformer ASR	+ChineseBERT	74.54	**74.76**	74.65
Conformer ASR	+AHN	**75.97**	74.64	**75.30**

The NER tagger ChineseBERT-CRF is trained on the BIO formatted text with AHN (the probability of AHN for each character is set to 0.05). ChineseBERT-base[7] (12 layers of 12 attention heads and the hidden size of 768) is selected for experiment. We also train two NER taggers ChineseBERT-CRF and BERT-base-CRF without AHN for comparison. The three NER taggers are optimized by AdamW [18], with the learning rate set to 0.00003 and a linear warmup scheduler. The training epoch number is set to 5, and the batch size is set to 64.

In addition, the results of the pipeline approach in [6] are chosen for comparison with our method.

Similar to the works [4]–[6], we choose the commonly used metrics Character Error Rate (CER) [19] and $F1$ score [20] to evaluate ASR and NER respectively.

B. ASR Results

The results of CER of our ASR model Conformer on dev and test datasets are reported in Table II, and the results from [6] are listed for comparison. From Table II, it could be seen that under the same model and parameter configuration, our results are very close (only slightly worse) to [6]. Therefore, the output of our ASR is reliable for the NER tagger.

In addition, we report the CER of the sentences with and without NEs from the test dataset in Table III. It could be seen that the CER of sentences without NEs are much smaller than that of sentences with NEs (4.02 vs. 6.73), which indicates that it is more difficult to correctly identify NEs in ASR and entity correction is very necessary after ASR.

C. NER Results

For the two-step pipeline method, the ASR model outputs the speech transcriptions, and then the NER tagger predicts the NEs. Table IV provides the results (*precision*, *recall*, and $F1$ *score*) of the three NER taggers (BERT-CRF, ChineseBERT-CRF with or without AHN) on the outputs of our Conformer

[6]https://github.com/Alibaba-NLP/AISHELL-NER

[7]https://github.com/ShannonAI/ChineseBert

TABLE V
MAXIMUM RECALL RATES OF DIFFERENT NE CATEGORIES FOR NER MODELS FROM OUR ASR RESULTS

NE Categories	Total	Correct	Maximum Recall
PER	898	501	55.79
LOC	1,330	1,175	88.35
ORG	1,165	978	83.95
ALL	3,393	2,654	78.22

TABLE VI
RESULTS IN DIFFERENT NE CATEGORIES

		Precision	Recall	$F1$
Ground-Truth Text	PER	95.61	94.65	95.13
+	LOC	94.88	96.17	95.52
BERT-CRF	ORG	93.02	90.30	91.64
Conformer ASR	PER	56.80	53.51	55.11
+	LOC	83.16	85.25	84.19
ChineseBERT-CRF-AHN	ORG	81.79	78.97	80.35

ASR. Meanwhile, the results of BERT-CRF on ground-truth text and the best results in [6] are also included.

In Table IV, the result of the NER tagger BERT-CRF on the ground-truth text is much better than that on the ASR output. This is because the errors in ASR output affect the NER. And it could be seen that ChineseBERT is slightly better ($F1$ score 74.65 vs. 74.42) than BERT for NER from the ASR output, which indicates that pinyin embedding is useful for NER. And when AHN is introduced to train the NER tagger, the improvement gets a very significant boost ($F1$ score from 74.65 to 75.30). Therefore, by incorporating *pinyin* into the NER tagger, the $F1$ score could be improved by nearly 1% absolute points ($0.88\% = 75.30\% - 74.42\%$), which is a significant improvement in the field of NER. And when compared with the current SOTA results in [6], although the CER of our ASR is slightly worse than [6] (CER 4.99 vs. 4.75), our NER tagger could still get a better result ($F1$ score 75.30 vs. 74.90).

In a word, from the above comparison results, pinyin is indeed beneficial to NER from Chinese speech.

In addition, we report the maximum recall rates of different NE categories from our ASR outputs of the test dataset in Table V, where "Correct" represents the number of correctly identified NEs in ASR outputs and "Total" represents the number of NEs in ground-truth text. The recall rates are calculated by "Correct/Total", which are the theoretical maximum recall for NER models. From Table V, the recall rate for PER is the lowest (55.79%). This is because there are a lot of homophones and polyphones in Chinese, especially for person names. For example, the PER NE 姚明 (`yao2 ming2`) has the same pronunciation and same pinyin as the PER NE 姚鸣 (`yao2 ming2`), which makes ASR really hard to recognize the correct PER NE from Chinese speech.

We also report the results of different NE categories for NER models on ground-truth text and ASR outputs in Table VI. Although the results on ASR outputs significantly lag behind those on ground-truth text, especially in the PER category, their recall rates are close to the recall rates in Table V (53.51% vs. 55.79% for PER, 85.25% vs. 88.35% for LOC, and 78.97% vs. 83.95% for ORG). This shows that our pipeline method is very effective for NER from Chinese speech and further improvement of NER requires entity correction to break the recall bottlenecks.

TABLE VII
NER RESULTS FOR DIFFERENT AHN PROBABILITIES

AHN Probability	Precision	Recall	$F1$
0.00	74.54	**74.76**	74.65
0.05	**75.97**	74.64	**75.30**
0.10	74.93	74.11	74.52
0.15	74.92	73.55	74.23

D. Parameter Study

In this section, we study the probability parameter of AHN for each character. Four NER taggers with the architecture ChineseBERT-CRF are trained with the AHN probability varying from 0 to 0.15 (with step 0.05). The results of the four NER taggers on the same ASR output of AISHELL-NER test data are shown in Table VII.

From Table VII, it could be seen that the $F1$ score is the best when the probability of AHN takes the value 0.05, which is very close to the CER of ASR (4.99%). We guess that the error distributions of the training and test datasets for NER taggers are very consistent in this probability setting, so that the results of NER are the best. Therefore, it could be a good choice if the AHN probability for NER tagger takes the CER value of ASR.

E. Discussion

Essentially, incorporating *pinyin* into pipeline NER from Chinese speech is a data argumentation method for NER tagger, which tries to make training data and test data consistent in error distribution by introducing AHN.

A straightforward method is to train or fine-tune NER tagger by using the text of ASR output, which could also make the NER tagger robust. However, when this way is adopted, additional annotated NER data (namely, NE-tagged ASR outputs) is required to train or fine-tune the NER tagger so as to keep the CER consistent with that obtained using testing data. Furthermore, when the ASR model is upgraded, new training data needs to be annotated to retrain the NER model, which is costly.

In contrast, when homophone noises (independent of ASR outputs) are used, the training data for the ASR model can be used to train the NER model, not requiring additional annotated NER data. And when the ASR model is upgraded, we only need to tune the noise hyper-parameter according to the CER of the ASR model to retrain the NER model, during which additional annotated NER data is not required either.

In a word, using homograph noises is much more economic than using the text of ASR output.

V. CONCLUSION

In this paper, an important information *pinyin* of Chinese character is introduced to the two-step pipeline method (ASR

+ NER Tagger) for NER from Chinese speech. The pretrained model ChineseBERT is used to embed the pinyin to the input of an NER tagger, and the homophone noises are introduced to the training data of NER tagger. From the comparison results on the AISHELL-NER dataset for NER from Chinese speech, the $F1$ score of the NER tagger with architecture BERT-CRF could be improved by nearly 1% absolute points (a significant improvement in the field of NER), and the $F1$ score of our method outperforms the current SOTA result by 0.4% absolute points despite the slightly worse CER of our ASR. In addition, the choice for the AHN probability is discussed.

REFERENCES

[1] P. Liu, Y. Guo, F. Wang, and G. Li, "Chinese named entity recognition: The state of the art," *Neurocomput.*, vol. 473, no. C, pp. 37–53, Feb. 2022, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.10.101. [Online]. Available: https://doi.org/10.1016/j.neucom.2021.10.101.

[2] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2145–2158. [Online]. Available: https://aclanthology.org/C18-1182.

[3] A. Caubrière, S. Rosset, Y. Estève, A. Laurent, and E. Morin, "Where are we in named entity recognition from speech?" English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 4514–4520, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.556.

[4] S. Ghannay, A. Caubrière, Y. Estève, *et al.*, "End-to-end named entity and semantic concept extraction from speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 692–699. DOI: 10.1109/SLT.2018.8639513.

[5] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, "End-to-end named entity recognition from english speech," in *INTERSPEECH*, 2020, pp. 4268–4272.

[6] B. Chen, G. Xu, X. Wang, P. Xie, M. Zhang, and F. Huang, "AISHELL-NER: named entity recognition from chinese speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, IEEE, 2022, pp. 8352–8356. DOI: 10.1109/ICASSP43922.2022.9746955. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9746955.

[7] Z. Sun, X. Li, X. Sun, *et al.*, "ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 2065–2075. DOI: 10.18653/v1/2021.acl-long.161. [Online]. Available: https://aclanthology.org/2021.acl-long.161.

[8] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[9] Y. zhao, J. Li, X. Wang, and Y. Li, "The speechtransformer for large-scale mandarin chinese speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7095–7099. DOI: 10.1109/ICASSP.2019.8682586.

[10] J. Fu, X. Huang, and P. Liu, "SpanNER: Named entity re-/recognition as span prediction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 7183–7195. DOI: 10.18653/v1/2021.acl-long.558. [Online]. Available: https://aclanthology.org/2021.acl-long.558.

[11] X. Wang, Y. Jiang, N. Bach, *et al.*, "Automated concatenation of embeddings for structured prediction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 2643–2660. DOI: 10.18653/v1/2021.acl-long.206. [Online]. Available: https://aclanthology.org/2021.acl-long.206.

[12] H. Yang, M. Zhang, S. Tao, M. Ma, and Y. Qin, "Chinese asr and ner improvement based on whisper fine-tuning," in *2023 25th International Conference on Advanced Communication Technology (ICACT)*, 2023, pp. 213–217. DOI: 10.23919/ICACT56868.2023.10079686.

[13] Q. Zhang, H. Lu, H. Sak, *et al.*, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7829–7833, 2020.

[14] W. Han, Z. Zhang, Y. Zhang, *et al.*, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. X. 0011, and T. F. Zheng, Eds., ISCA, 2020, pp. 3610–3614. DOI: 10.21437/Interspeech.2020-2059. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2059.

[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for

segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289, ISBN: 1558607781.

[16] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5. DOI: 10.1109/ICSDA.2017.8384449.

[17] Z. Yao, D. Wu, X. Wang, *et al.*, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Interspeech*, 2021, pp. 4054–4058.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[19] A. C. Morris, V. Maier, and P. D. Green, "From wer and ril to mer and wil: Improved evaluation measures for connected speech recognition," in *Interspeech*, 2004.

[20] Y. Sasaki, "The truth of the F-measure," 2007.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Learning Semantic Information from Machine Translation to Improve Speech-to-Text Translation

Pan Deng*, Jie Zhang*, Xinyuan Zhou†, Zhongyi Ye†, Weitai Zhang*, Jianwei Cui*, Lirong Dai*
* NERC-SLIP, University of Science and Technology of China (USTC), Hefei, China
E-mail: {pdeng, zwt2021, jwcui}@mail.ustc.edu.cn {jzhang6, lrdai}@ustc.edu.cn
† iFlytek Research, Hefei, China
E-mail: {xyzhou15, zyye7}@iflytek.com

***Abstract*—End-to-end speech translation (ST) directly translates the source speech to the target text, following a typical encoder-decoder framework. However, it has shown that the conventional ST encoder is mainly used to extract long but locally attentive acoustic features, which may lead to a lack of global semantic features. In this work, we therefore propose to integrate a semantic decoder into the speech translation (SD-ST) model, where the semantic decoder can generate text-like features with more global semantic information analogously to the machine translation system (MT). We also investigate different strategies to ensure length consistency between text-like features and text sequences. Experimental results show that the proposed SD-ST model achieves the best BLEU score on the 40-hour subset of the Fisher Spanish English dataset and a comparable BLEU score on the MuST-C dataset. Furthermore, it is shown that the SD-ST model can even perform zero-shot ST.**

Index Terms: End-to-end speech translation, semantic information, encoder-decoder, speech recognition.

I. Introduction

Speech translation (ST) aims to translate speech in the source language to text in the target language. Conventional ST systems usually follow a cascaded paradigm consisting of two subtasks: automatic speech recognition (ASR) and machine translation (MT) [1]. This cascaded model in principle suffers from error propagation and latency problems [2]. Recently, many end-to-end ST approaches have been proposed to tackle these problems [3]–[5], which can be considered as a joint task of ASR and MT. However, the end-to-end ST methods suffer from the scarcity of paired speech and text data [6]. Traditional pre-training methods use a large amount of paired data to train ASR and MT models, then separately initialize the ST encoder and decoder and finally fine-tune the ST model with a small amount of parallel data [7]–[9]. However, this encoder-decoder structure cannot learn sufficient semantic information or reorder feature sequences, which are required for practical translation applications [10], [11].

This study defines semantic information as the relationship between words and the meanings of sentences. Therefore, the semantic information is mainly conveyed through text and learned by the MT model. To improve the modelling capacity of global semantic information, the MT encoder can be integrated into the ST model, such that the ST is decomposed into two subtasks: ASR and MT. For example, in [11], [12], the MT encoder is placed between the ST encoder and decoder, where [11] uses an adapter and [12] shares weights between the CTC classification layer and the embedding layer to maintain semantic consistency. In [13]–[15], an external model is additionally used to extract semantic and contextual information. Although the MT-integrated ST methods show the superiority over conventional encoder-decoder based counterparts, there exist two major limitations. 1) **Limited semantic information**: The learned semantic information is rather limited even if the acoustic features are further processed by the pre-trained MT encoder. This is because the desired input to the MT encoder are text sequences that are significantly different from the acoustic feature sequences, leading to a bias in the acquisition of semantic features. On the other hand, the semantic modeling module is insufficiently constrained when fine-tuning, which might forget or lose the original semantic information. 2) **Length inconsistency**: The acoustic feature sequences contain too much redundant information, and the sequence length is much longer than that of the text sequence [16]. This inconsistency makes it difficult to learn the alignment relation between the pre-trained decoder and encoder.

In this work, we propose to use a semantic decoder to generate text-like feature sequences, which are shown to be more fruitful in terms of semantic information and can ensure the consistency between feature and text lengths. We first train an MT model and regard the hidden-layer features therein as text features. Then, we build a semantic decoder that uses the attention mechanism to search for similar pronunciation units and aggregates acoustic features based on the text features. Further, we compute the L_1 loss at the token level between the aggregated acoustic and text features to generate text-like features. Finally, the text-like features are constructed as a sequence, which turns out to match the text sequence from the perspective of length. Experimental analysis shows that the generated text-like feature sequences can not only preserve sufficient acoustic information for ASR, but also contain more semantic information that is beneficial for ST. Our experimental results show that the proposed model achieves the best BLEU score on the 40-hour subset of the Fisher Spanish English dataset. On the MuST-C En-De dataset, the proposed model achieves a BLEU score comparable to previous baselines and it also shows the applicability to zero-shot speech translation.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

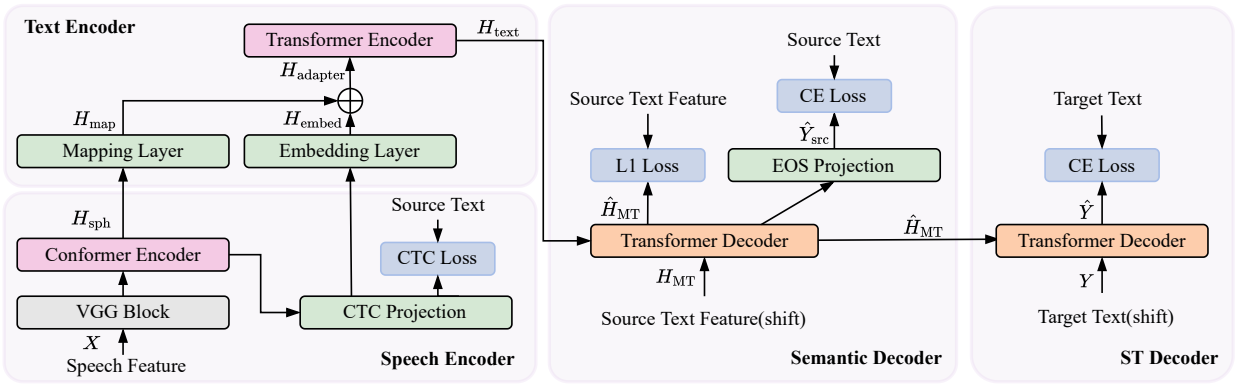


Fig. 1. The overview of the proposed SD-ST model, which reduces to the SATE model with a removal of the semantic decoder and becomes identical to the E2E-ST model by removing both the text encoder and semantic decoder.

II. Preliminaries

In this section, we will present two classic ST models: E2E-ST and SATE, which are the basis of the proposed method.

A. End-to-end speech translation (E2E-ST)

The conventional end-to-end ST model can be built by combining the speech encoder and ST decoder modules. The speech encoder extracts the acoustic feature $H_{\text{sph}} = \{h_1, h_2, ..., h_N\}$ from the raw speech feature sequence $X = \{x_1, x_2, ..., x_U\}$, where U and N denote the lengths of the speech feature and extracted acoustic feature, respectively, i.e.,

$$H_{\text{sph}} = \text{Enc}_{\text{sph}}(X), \tag{1}$$

Based on the hidden feature H_{sph} and the text sequence $Y = \{y_1, y_2, ..., y_M\}$ in the target language, the ST decoder generates the translation result $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_M\}$, where M denotes the length of the text sequence in target language. The training objective can be the cross entropy between the translation result and the true text, given by

$$L_{\text{E2E-ST}} = -\log P(\hat{Y}|Y, H_{\text{sph}}, \theta^{\text{sph}}_{\text{enc}}, \theta^{\text{ST}}_{\text{dec}}), \tag{2}$$

where $\theta^{\text{sph}}_{\text{enc}}$ and $\theta^{\text{ST}}_{\text{dec}}$ are trainable parameters of the speech encoder and ST decoder, respectively. From the viewpoint of multi-task learning, ASR can be included as an auxiliary training task, e.g., by adding an additional ASR decoder to minimize the cross entropy loss [17], [18] or the CTC loss L_{CTC} [19]. The resulting training loss function for joint ASR and ST can thus be written as

$$L = \lambda \cdot L_{\text{CTC}} + (1 - \lambda) \cdot L_{\text{E2E-ST}}, \tag{3}$$

where λ is the balancing weight.

B. Stacked acoustic-and-textual encoding (SATE)

In order to address the lack of semantic learning capability of the E2E-ST model, the SATE [11] model stacks the speech encoder and text encoder to encode acoustic and semantic information from raw speech features. The SATE model uses an adapter architecture to fit the output of the speech encoder to the expected input of the text encoder, which is given by

$$H_{\text{adapter}} = H_{\text{map}} + H_{\text{embed}}, \tag{4}$$

where

$$H_{\text{map}} = \text{ReLU}(W^{\text{map}} \cdot H_{\text{sph}} + b^{\text{map}}), \tag{5}$$

$$H_{\text{embed}} = W^{\text{embed}} \cdot P_{\text{ctc}}, \tag{6}$$

$$P_{\text{ctc}} = \text{Softmax}(W^{\text{ctc}} \cdot H_{\text{sph}} + b^{\text{ctc}}), \tag{7}$$

where W^{map}, W^{ctc}, W^{embed}, b^{map}, b^{ctc} are the parameters for the mapping layer, the CTC projection layer, and the embedding layer, respectively. Then the text encoder is used to generate a representation for translation, i.e.,

$$H_{\text{text}} = \text{Enc}_{\text{text}}(H_{\text{adapter}}), \tag{8}$$

and the training objective functions for the entire model are similar to E2E-ST. The parameters of the text encoder Enc_{text} can be initialized using off-the-shelf pre-trained MT encoders, which enables the text encoder to generate features containing semantic information. However, the semantic modeling capability would be rather limited due to the catastrophic forgetting in the absence of external supervision. Moreover, there might exist an inconsistency between the sequence lengths of the features generated by the text encoder and the desired feature for the ST decoder.

III. Semantic decoder based E2E-ST

In order to enhance the semantic modeling capability for ST, in this section we propose a semantic decoder based E2E-ST model as shown in Fig.1 consisting of four modules, i.e., speech encoder, text encoder, semantic decoder and ST decoder.

A. Semantic decoder module

The purpose of the semantic decoder module is to refine the semantic information and remove redundant noisy information. As shown in Fig.1, the semantic decoder is composed of several transformer layers and an optional projection layer

which is used to predict the end of the sentence. During training, the input to the semantic decoder is the text feature H_{MT}, which is encoded by the pre-trained MT model as

$$H_{\mathrm{MT}} = \mathrm{Enc}_{\mathrm{MT}}(Y^{\mathrm{src}}), \tag{9}$$

where $Y^{\mathrm{src}} = \{y_1^{\mathrm{src}}, y_2^{\mathrm{src}}, ..., y_V^{\mathrm{src}}\}$ is the text in the source language at a length of V and $H_{\mathrm{MT}} = \{h_1^{\mathrm{mt}}, h_2^{\mathrm{mt}}, ..., h_V^{\mathrm{mt}}\}$. The semantic decoder predicts the text-like features using H_{MT} and H_{text} and then computes the L_1 loss at the token level as

$$\hat{H}_{\mathrm{MT}} = \mathrm{Dec}_{\mathrm{semantic}}(H_{\mathrm{MT}}, H_{\mathrm{text}}), \tag{10}$$

$$L_{\mathrm{semantic}} = \frac{1}{V}\sum_{i}^{V}\left|\hat{h}_i^{\mathrm{mt}} - h_i^{\mathrm{mt}}\right|, \tag{11}$$

where $\hat{H}_{\mathrm{MT}}$ is generated in a teacher-forcing fashion when training and $\hat{H}_{\mathrm{MT}} = \{\hat{h}_1^{\mathrm{mt}}, \hat{h}_2^{\mathrm{mt}}, ..., \hat{h}_V^{\mathrm{mt}}\}$.

B. EOS prediction

The semantic decoder generates text-like features in an autoregressive way when decoding, so the exact position of the end of sentence (EOS) needs to be predicted. An inaccurate EOS position would cause the loss of useful features or redundant information contained in the generated features. For this, we consider two EOS prediction methods in this section.

Direct prediction: Let at step t of decoding the feature predicted by the semantic decoder be represented by $\hat{h}_t^{\mathrm{mt}}$. We directly normalize $\hat{h}_t^{\mathrm{mt}}$ and determine the index corresponding to the largest value as the EOS position, i.e.,

$$EOS_{\mathrm{pos}} = \mathrm{argmax}(\mathrm{Softmax}(\hat{h}_i^{mt})). \tag{12}$$

Projection network prediction: Inspired by the text-to-speech method [20], we use a single-layer projection network to project the text-like features into the dictionary space of the source language. The projection network is shown on the top of the semantic decoder module in Fig.1. We believe that it helps the text-like feature retain links to the source text. Specifically, the decoding steps of the semantic decoder will be completed in case the text-like feature is projected as an EOS token, i.e.,

$$P_t^{\mathrm{src}} = \mathrm{Softmax}(\mathrm{Proj}_{\mathrm{eos}}(\hat{h}_t^{\mathrm{mt}})), \tag{13}$$

$$EOS_{\mathrm{pos}} = \mathrm{argmax}(P_t^{\mathrm{src}}), \tag{14}$$

where $\mathrm{Proj}_{\mathrm{eos}}$ is the EOS projection network and P_t^{src} is the probability distribution at step t of decoding. The training objective is to minimize the cross-entropy loss between the predicted token and the text in the source language, given by

$$L_{\mathrm{eos}} = -\log(P_t^{\mathrm{src}}|\theta_{\mathrm{dec}}^{\mathrm{semantic}}). \tag{15}$$

C. Loss function

The proposed SD-ST model generates translation sequences using the text-like features $\hat{H}_{\mathrm{MT}}$ and is optimized by minimizing the cross-entropy loss, which is given by

$$L_{\mathrm{SD-ST}} = -\log P(\hat{Y}|Y, \hat{H}_{\mathrm{MT}}, \theta_{\mathrm{enc}}^{\mathrm{ST}}, \theta_{\mathrm{dec}}^{\mathrm{semantic}}, \theta_{\mathrm{dec}}^{\mathrm{ST}}), \tag{16}$$

where $\theta_{\mathrm{enc}}^{\mathrm{ST}}$ represents the parameter set for the speech and text encoders. The proposed SD-ST model is trained in a multi-task form, where the objective function can be formulated similarly to (3) as

$$\begin{aligned} L = \lambda \cdot L_{\mathrm{CTC}} + \alpha \cdot L_{\mathrm{semantic}} + \beta \cdot L_{\mathrm{eos}} \\ + (1 - \lambda - \alpha - \beta) \cdot L_{\mathrm{SD-ST}}, \end{aligned} \tag{17}$$

where λ, α, β are set to tradeoff the ASR-related CTC, semantic decoder and EOS prediction losses.

IV. Experiments

A. Experimental setup

The experiments are conducted on two often-used ST datasets, e.g., the Fisher Spanish-English (Es-En) dataset[1] [21] and MuST-C dataset [22]. For the Fisher Es-En dataset, we select the same 40 hours of speech as in [23]. For the MuST-C dataset, we select the English-German (En-De) language pair, which contains 408 hours of speech. We add external 18M sentence pairs from Opensubtitle2018[2] En-De translation dataset to the MuST-C experiment. For evaluation, we use the standard Fisher dev, Fisher dev2 and Fisher test sets in Fisher Es-En language pair, and use tst-COMMON set in MuST-C En-De language pair. The data pre-processing and model configuration keep the same for both datasets.

The ESPnet toolkit [24] is used to pre-process the data. We consider the 80-dimensional log-Mel filterbank coefficients and 3-dimensional pitches as the speech features, which are normalized using the mean and variance. The utterances that have more than 3000 frames or 400 words are removed. The speech data are augmented by applying speed perturbations with factors of 0.9 and 1.1 and the SpecAugment technique [25]. For text data, we tokenize the bilingual text and use Sentencepeice[3] to learn a joint source and target subword vocabulary. The vocabulary size is 1000 for Fisher Es-En language pair and 8000 for MuST-C En-De language pair.

The model configuration and hyper-parameters keep the default settings of "fisher callhome Spanish" recipe in ESPnet [24] (version 202205). The speech encoder is a 12-layer conformer with 2 additional VGG convolution blocks, and the text encoder, semantic decoder, and ST decoder each contain a 6-layer transformer, respectively. The self-attention module comprises 256 hidden units, 4 attention heads, and the FFN module has 2048 hidden units. The dropout and label-smoothing rates are set to be 0.1. The models are optimized using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and the inverse square root learning rate schedule.

The model training involves two phases: 1) training the speech encoder, the text encoder and the semantic decoder modules, where the speech and text encoders are initialized by ASR and MT models, respectively, and 2) training the entire SD-ST model, where the model parameters from the first stage are used for initialization except that the ST decoder is initialized by the MT model. During inference, the semantic

[1]The LDC Catalog of Fisher Spanish Speech and Fisher Spanish Transcripts are LDC2010S01 and LDC2010T04 respectively.

[2]http://opus.nlpl.eu/OpenSubtitles-v2018.php

[3]https://github.com/google/sentencepiece

TABLE I
THE BLEU SCORES ON THE FISHER ES-EN DATASET.

Model	Params(M)	dev	dev2	test
MT	18.2	53.6	54.4	51.9
Phone[26]	12.3	34.5	-	33
E2E-ST-MTL[23]	52.6	31.2	32.6	31.5
E2E-ST[24]	42.3	34.3	35.2	34.1
E2E-ST-Large[24]	57.8	35.2	35.9	35
SATE[11]	50.5	36.4	37.7	36.3
SD-ST	60.5	**38**	**39.2**	**38.4**

TABLE II
THE BLEU SCORES ON THE MUST-C EN-DE TST-COMMON SET. ALL MODELS EXCEPT MULTI-DECODER MODEL USE ADDITIONAL MT DATA AND DO NOT APPLY KNOWLEDGE DISTILLATION FROM EXTERNAL MODELS. BESIDES, CHIMERA AND SATE MODEL USE ADDITIONAL SPEECH DATA.

Model	Speech data	Params(M)	tst-COMMON
MT	-	23.6	31.1
Zero-shot SD-ST	-	73.4	**21.5**
E2E-ST-Large[24]	-	74	25.6
Multi-Decoder[27]	-	135	26.4
JT-S-MT[15]	-	76	26.8
Chimera[28]	✓	165	27.1
SATE[11]	✓	135.6	27.3
SD-ST	-	73.4	**27.2**
SD-ST-Large	-	173	**27.9**

TABLE III
THE BLEU SCORES ON FISHER DEV/DEV2/TEST SETS WITHOUT THE EOS PROJECTION NETWORK OR THE SEMANTIC DECODER MODULE.

Model	Params(M)	dev	dev2	test
SD-ST	60.5	38	39.2	38.4
- EOS projection network	60.3	37.2	38.4	37.4
- Semantic Decoder	50.5	36.4	37.7	36.3

decoder generates text-like features in an autoregressive way and predicts the EOS position. Then, the ST decoder generates translation results based on text-like features with a beam size of 10. We report 4-references case-insensitive detokenized BLEU score on Fisher dev/dev2/test sets and 1-reference case-sensitive detokenized BLEU score on MuST-C En-De tst-COMMON set.

B. Experimental results

First, we show the ST performance in Fig. 2 in terms of hyper-parameters, where $\lambda \in \{0.1, 0.2, 0.3\}$, $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\beta = 0.1$. The parameter ranges are chosen based on pre-experiments. It is clear that the choice of $\{\lambda = 0.2, \alpha = 0.1, \beta = 0.1\}$ can result in a better BLEU score, which is thus used for the following experiments on both Fisher Es-En and MuST-C En-De datasets. Table I shows the

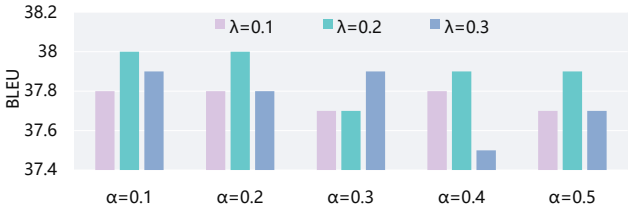


Fig. 2. The BLEU score on Fisher dev set in terms of $\{\lambda, \alpha\}$.

TABLE IV
THE WERS AND LENGTH GAP ON FISHER DEV, WHERE WER OF ASR-CTC IS CALCULATED BY REMOVING THE BLANK AND REPEATED SYMBOLS IN THE PREDICTED SEQUENCES.

Model	Beam	WER	Length gap
ASR-ED[24]	1	32.9	1.31
ASR-ED[24]	5	32.1	1.31
ASR-CTC[24]	-	34.3	1.52
SD-ST	1	33	1.34

translation performance of the SD-ST model on Fisher Es-En dataset in comparison to other baseline methods. It is clear that the proposed SD-ST model outperforms the best published approach (i.e., Phone [26]) and improves the BLEU scores by 3.5 and 5.4 on Fisher dev and test sets, respectively. Note that the Phone method [26] uses the LSTM network as the backbone, resulting in less parameters. To be fair, the E2E-ST, E2E-ST-Large[4] and SATE models are also compared, which are trained using the same experimental configuration and have a comparable model size as the SD-ST model. The results show that our method still outperforms the E2E-ST, E2E-ST-Large and SATE models and achieves the best performance on the Fisher Es-En dataset.

Table II shows the translation performance of the SD-ST model on MuST-C En-De dataset in comparison to previous methods that can enhance the semantic information for ST. It is clear that our SD-ST model achieves a BLEU score of 27.2 which is comparable to previous methods, while we use fewer model parameters and speech data. Further, our SD-ST-Large[5] model achieves 27.9 BLEU score. Moreover, we try to use only the paired speech and text data of the source language for one-stage model training mentioned in Section IV-A, and then initialize the ST decoder of the SD-ST model with pre-trained MT decoder, i.e., zero-shot inference. As shown in Table II, our zero-shot SD-ST model achieves a BLEU score of 21.5, implying a potential in the zero-shot case.

In order to analyze the impact of the semantic decoder and the EOS projection network, we evaluate the performance of the proposed SD-ST model with a removal of each module in Table III. We can see that both the semantic decoder and the EOS projection network are indispensable functional modules that can effectively improve the BLEU scores. For

[4]The E2E-ST-Large model has an 18-layer Conformer module.
[5]The SD-ST-Large model has hidden size of 512 and 8 heads in attention module.

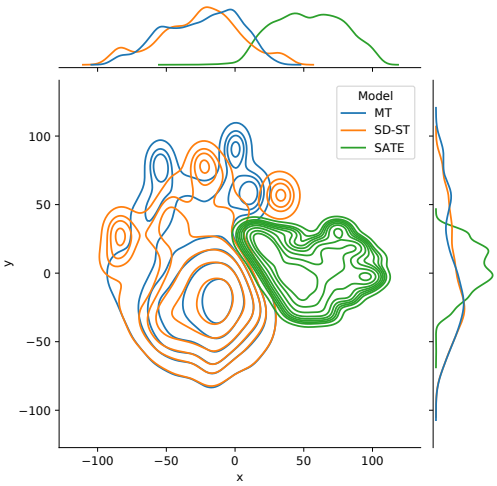


Fig. 3. Bivariate KDE contour plot of features generated by MT, SD-ST and SATE (SD-ST without the semantic decoder).

the prediction of EOS positions, the direct prediction (i.e., removing the EOS projection network) performs worse than the EOS projection network, which reveals the necessity of the proposed EOS projection method.

Further, in order to analyze the impact of the proposed semantic decoder module on the issues of limited semantic information and inconsistent length, similarly to [29] we use t-SNE[30] to compress the high-dimensional features into a 2-dimensional space. Then we show the bivariate kernel density estimates [31] (KDE) contours to visualize different features in Fig.3, where the average-pooling over the dimension of time is applied on the features. It is clear that the features generated by the MT encoder overlap with those generated by the semantic decoder of SD-ST, while they are further away from those generated by the text encoder of SATE. This implies that the semantic decoder is beneficial for generating text-like features that are rich in semantic information. In addition, with the EOS projection network in the semantic decoder, we can predict the text sequence in the source language and then compute the WER. Table IV shows that the WER calculated using the output of the semantic decoder module is comparable to the WER of the general ASR models that are trained based on default settings of ESPnet. Moreover, we calculate the length gap which measures the difference between the length of the text sequence and that of the feature or token sequence generated by each model (the length gap is averaged at sentence level). We can see that length gap of the SD-ST model is only around 1 token (subword) and comparable to ASR models that are used for the cascaded ST. We can therefore conclude that the text-like features generated by the semantic decoder not only retain sufficient acoustic and semantic information, but also maintain length consistency with the text sequence.

V. Conclusion

In this paper, we proposed a speech translation model called SD-ST, which uses a semantic decoder to generate text-like features that are rich in semantic information. The text-like features can be converted to a textual sequence as well, which is comparable to the ASR models in terms of WER. Consequently, ASR can be seen as a by-product of the proposed model. It was shown that the text-like features can preserve sufficient acoustic/semantic information and maintain length consistency with the text sequences. The semantic decoder proposed in this paper can be further investigated in future work to enhance the robustness of its text-like features to improve the translation performance in different scenarios.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62101523), Hefei Municipal Natural Science Foundation (2022012), Fundamental Research Funds for the Central Universities and the Leading Plan of CAS (XDC08010200).

References

[1] H. Ney, "Speech translation: Coupling of recognition and translation," in *ICASSP*, 1999, pp. 517–520.

[2] N. Ruiz and M. Federico, "Assessing the impact of speech recognition errors on machine translation quality," in *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, 2014, pp. 261–274.

[3] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *ICASSP*, 2018, pp. 6224–6228.

[4] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *ISCA Interspeech*, 2017, pp. 2625–2629.

[5] R. Fukuda, K. Sudoh, and S. Nakamura, "Speech Segmentation Optimization using Segmented Bilingual Speech Corpus for End-to-end Speech Translation," in *Proc. Interspeech 2022*, 2022, pp. 121–125. DOI: 10.21437/Interspeech.2022-11382.

[6] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Attention-passing models for robust and data-efficient end-to-end speech translation," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 313–325, 2019.

[7] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 58–68.

[8] Z. Zhang and J. Ao, "The YiTrans speech translation system for IWSLT 2022 offline shared task," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Association for Computational Linguistics, 2022, pp. 158–168.

[9] J. Zhao, H. Yang, G. Haffari, and E. Shareghi, "M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation," in *ISCA Interspeech*, 2022, pp. 111–115.

[10] J. Chen, M. Ma, R. Zheng, and L. Huang, "Specrec: An alternative solution for improving end-to-end speech-to-text translation via spectrogram reconstruction.," in *ISCA Interspeech*, 2021, pp. 2232–2236.

[11] C. Xu, B. Hu, Y. Li, *et al.*, "Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2619–2630.

[12] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 9161–9168.

[13] Y. Liu, J. Zhu, J. Zhang, and C. Zong, "Bridging the modality gap for speech-to-text translation," *arXiv preprint arXiv:2010.14920*, 2020.

[14] Q. Dong, R. Ye, M. Wang, *et al.*, "Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 12 749–12 759.

[15] Y. Tang, J. Pino, X. Li, C. Wang, and D. Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4252–4261.

[16] M. Gaido, M. Cettolo, M. Negri, and M. Turchi, "CTC-based compression for direct speech translation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 690–696.

[17] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 82–91.

[18] Y. Ko, K. Sudoh, S. Sakti, and S. Nakamura, "ASR Posterior-Based Loss for Multi-Task End-to-End Speech Translation," in *ISCA Interspeech*, 2021, pp. 2272–2276.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[20] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[21] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus," in *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, 2013.

[22] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: A Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2012–2017.

[23] Y.-F. Cheng, H.-S. Lee, and H.-M. Wang, "Allost: Low-resource speech translation without source transcription.," in *ISCA Interspeech*, 2021, pp. 2252–2256.

[24] H. Inaguma, S. Kiyono, K. Duh, *et al.*, "Espnet-st: All-in-one speech translation toolkit," *arXiv preprint arXiv:2004.10234*, 2020.

[25] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *ISCA Interspeech*, 2019, pp. 2613–2617.

[26] E. Salesky and A. W. Black, "Phone features improve speech translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2388–2397.

[27] S. Dalmia, B. Yan, V. Raunak, F. Metze, and S. Watanabe, "Searchable hidden intermediates for end-to-end models of decomposable sequence tasks," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2021, pp. 1882–1896.

[28] C. Han, M. Wang, H. Ji, and L. Li, "Learning shared semantic space for speech-to-text translation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2214–2225.

[29] R. Ye, M. Wang, and L. Li, "Cross-modal contrastive learning for speech translation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5099–5113.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[31] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Effective Fine-tuning Method for Tibetan Low-resource Dialect Speech Recognition

Jiahao Yang*, Jianguo Wei†, Kuntharrgyal Khysru, Junhai Xu, Wenhuan Lu†, Wenjun Ke, Xiaokang Yang
* Tianjin International Engineering Institute, Tianjin University, Tianjin, China
† College of Intelligence and Computing, Tianjin University, Tianjin, China
E-mail: {jianguo,wenhuan}@tju.edu.cn

***Abstract*—Tibetan is a distinctive and culturally rich language spoken by millions of people across the Tibetan Plateau and surrounding regions. Exploring the application of speech recognition technology to Tibetan has special significance for preserving language diversity and fostering cultural integration. Moreover, Tibetan comprises a multitude of distinct dialects, which present a hurdle for reusing speech recognition models. In low-resource dialect tasks, conventional approaches endeavor to transfer well-trained models from linguistically akin languages to the target. However, recent studies have shown that an indiscriminate fine-tuning of all parameters may disrupt the feature extractor of the pre-trained model, leading to catastrophic forgetting. This paper introduces an innovative fine-tuning method grounded in model adaptation. Aimed at training automatic speech recognition (ASR) models within the constraints of limited training data and cross-dialect transfer, our novel approach refines a select group of language-specific parameters, leading to robust performance. These parameters, signified by a sparse binary mask identical to the model, circumvent the need for additional parameters. Experiments conducted on two downstream low-resource Tibetan languages show that our proposed methodology outperforms the traditional fine-tuning and adapter based fine-tuning.**

I. Introduction

Influenced by history and geography, Tibetan has gradually evolved various dialects and accents, and it is widely believed that there are more than twenty different spoken forms. Generally speaking, there are three main dialects of Tibetan language: Amdo Tibetan, Khampa Tibetan, and U-Gtsang Tibetan. Due to the linguistic diversity of Tibetan and the problem of limited data, the application of Tibetan speech recognition technology trails behind [1], [2].

Automatic Speech Recognition (ASR) has made remarkable progress with the development of deep learning. Transformer-based [3]–[5] models greatly benefit from a large number of training parameters. Despite this, in scenarios where data resources are restricted, the potential for enhancements remains somewhat hampered. For some low-resource languages, fine-tuning speech recognition models trained on other dialects within the same language family has become a more effective paradigm [6]–[9]. Meanwhile, Choosing an appropriate fine-tuning strategy holds significance in bolstering the model's generalization ability [10], [11].

The advantages of neural networks in transferring knowledge between linguistically related languages are evident, primarily due to the efficiency of adaptation depends on various factors, including the distribution differences between the source and target languages [12]. In the presence of insufficient target data that fails to capture the acoustic diversity, the conventional fine-tuning approach may suffer from issues such as low generalizability and catastrophic forgetting. These challenges arise due to the high complexity and data sensitivity of the Transformer-based architecture [13], [14].

To mitigate these limitations, researchers have explored alternative strategies to alleviate the issue of knowledge forgetting during the process of transfer learning. For instance, one approach is to use L2 regularization to constrain the divergence between the model's weights and the pre-trained weights [15]. Furthermore, substituting some model parameters stochastically with their corresponding pre-trained parameters can also produce effects similar to regularization [16]. Another prevalent technique hinges on adapters, which can add additional intermediate layer parameters and fine-tuned while keeping all the pre-trained parameters frozen [17]–[19]. Nonetheless, the disadvantages of adapters remain distinctly noticeable. On the one hand, adding adapters for each task will lead to a sharp increase in model size. On the other hand, the adapter module itself requires a large amount of target language data for training, rendering the incorporation of intricate network configurations superfluous in scenarios of data paucity. Recent endeavors within the realm of Natural Language Processing have turned to sparse fine-tuning techniques as a surrogate to adapters, abstaining from the introduction of any additional structure [20], [21]. This not only curtails computational expenditure but also retains the performance supremacy of pre-trained models.

To better adapt pre-trained model to target low-resource Tibetan dialects, we proposed an effective fine-tuning method inspired by the Lottery Ticket Hypothesis (LTH) [22], which only fine tune the most task-related parameters within a subnetwork of the full model. The crux of the matter currently resides in discerning these precise parameters. In particular, there are two variations of this method: the first one discovers the subnetwork by measuring the differences between the upstream and downstream tasks, while the second one selects the most important parameters for the downstream task based on their gradient information. In comparison to the original idea, our approach diverges in two main facets. Firstly, LTH primarily focuses on the compression of large models, while

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

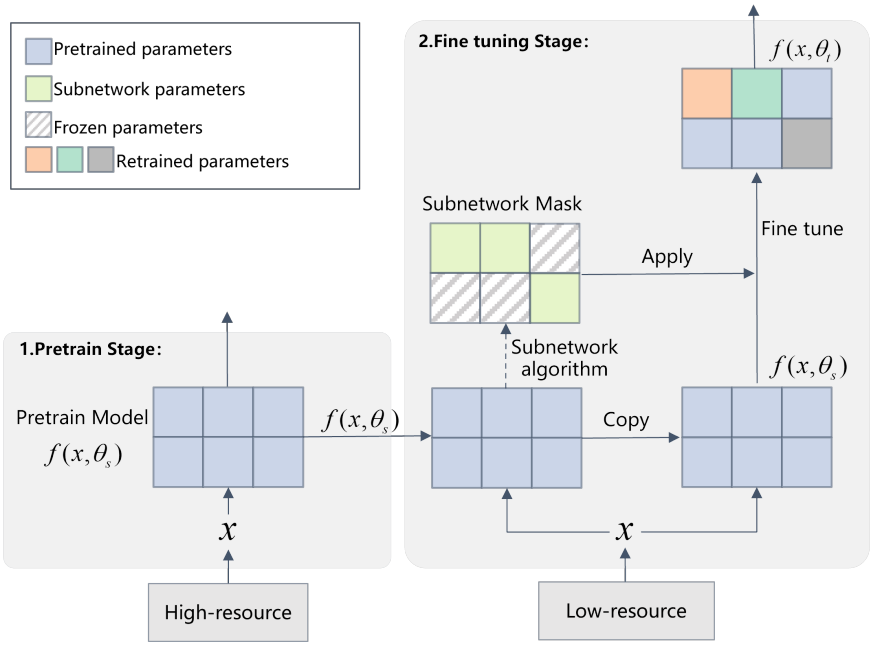


Fig. 1. Overview of the fine-tuning algorithm proposed in this work: Firstly, we train the model on high-resource corpus, and the pre-trained model is represented as function $f(x, \theta)$. Then, we obtain a mask to represent the corresponding language-specific subnetwork, where the green blocks symbolize trainable neurons, and the white blocks denote frozen ones. Finally, retrain the network using mask.

our proposed approach preserves all parameters of the network and avoids any pruning that may affect model accuracy. Furthermore, the parameters of the subnetwork extracted by the original LTH are determined by their significance in the pre-training task, which may not be applicable in transfer learning due to the potential mismatch between upstream and downstream tasks.

The rest of this paper is organized as follows. In Section 2, the proposed two fine-tuning strategies are introduced. Section 3 shows the experimental settings and evaluation results. The conclusions are in Section 4.

II. Proposed Methods

In this section, we mainly introduce our proposed cross-lingual transfer-learning methods based on language specific sparse fine-tuning, as shown in Figure 1.

A. Lottery Ticket Hypothesis

The Lottery Ticket Hypothesis (LTH) illustrates that a sparse subnetwork within a fully dense network is capable of being trained in isolation from initialization to achieve high accuracy, as shown in Figure 2.

Recent works have leveraged unstructured pruning technology to compress large pre-trained models based on the Lottery Ticket Hypothesis [22]. These methods only need to update a specific sparse sub-network and prune the remaining weights, achieving high sparsity with minimal performance degradation [23]–[25]. In different tasks or models, it is necessary to conduct multiple experimental iterations to identify an optimal "winning tickets" subnetwork. For instance, during the fine-tuning stage, a mask is typically applied, consisting of absolute values such as a set of ones and zeros. Usually, the subnetwork is identified by training a model on the source language and pruning the weights with the smallest magnitudes. This iterative pruning and re-training process may need to be executed numerous times to achieve an optimal compressed model.

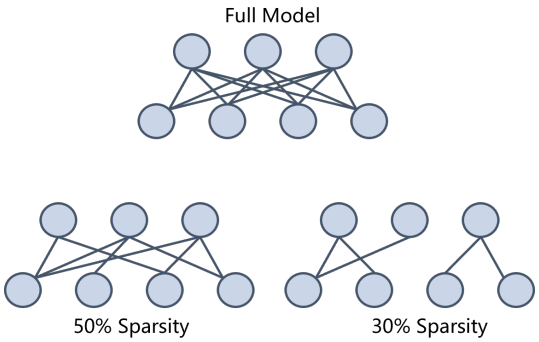


Fig. 2. Compared to a full model, a sparse subnetwork has only a few parameters trained.

B. Sparse Fine-tuning Method

This work investigates the generalization capability of sparse fine-tuning techniques in cross-dialect transfer learning, specifically in the domain of speech recognition for languages with limited resources. Consider a generic neural network parameterized by function $f(x, \theta)$ with initial parameters $\theta = \theta_0$ and input x. The output from the forward pass is $y = f(x, \theta)$. The parameters in full fine-tuning strategy would be updated

as follow:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \tag{1}$$

Where $\mathcal{L}(\theta_t)$ is the loss function. $\frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t}$ represents the gradients corresponding to the model parameters θ_t, and η is the learning rate. Our objective is to determine a subnetwork with a sparsity $S \in [0, 1]$. To achieve this, we consider a pruning mask consisting of a series of ones representing the subnetwork and replace the rest with zeros. This is illustrated as follows:

$$M_t^i = \begin{cases} 1 & \text{if } \theta^i \in P^t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

P^t represents a series of weight parameters considered to be activated at iteration t. i is the i-th weight, and M is the subnetwork mask used to update parameters at each backpropagation. Usually, P^t is obtained from the top-K proportion of weights θ_t based on their magnitude. This process can be described as follows:

$$P^t = \left\{ i \mid \theta_t^i \in \mathrm{TopK}(\theta_t) \right\} \tag{3}$$

Then we update the parameters with the following equation:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \odot M_t \tag{4}$$

C. Language Specific Transfer

The selection of subnetworks is often crucial for sparse fine-tuning methods. In this paper, we proposed two variants methods for obtaining a set of target language-specific parameters from the backbone network, as shown in Algorithm 1. One method is based on weight amplitude difference, and the other is based on the gradient of the neural parameter.

Algorithm 1 Framework of our fine-tuning method.

Require:

$\mathcal{L}()$: learning objective of model;
θ_s : the pretrain neural model weights;
S : Sparsity of subnetworks;
subnetwork($TopK(|\theta_s - \theta_t|), S$) : different strategies of subnetwork extraction amied at downstream language are implemented (if using fisher method, the first input parameter is $TopK(\mathbf{H})$;

Ensure:

Initialize model, Start fine-tuning stage:
while in epochs **do**
  if Calculate mask **then**
    M_t= subnetwork($TopK(|\theta_s - \theta_t|), S$).
  end if
  Update θ_t using Mask.
  $\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \odot M_t$
end while

Weight-Difference Fine-tuning(differ): First, the pre-trained model is thoroughly fine-tuned on the downstream language. θ_s on the source language and θ_t on the target language represent the neural network parameters before and after fine-tuning. The parameters are sorted by importance according to specific rules, which can be calculated by the greatest absolute difference $|\theta_t^i - \theta_s^i|$. i is the i-th weight. Those parameters with the largest difference in the above formula are trained in the next step on downstream language, and the remaining parameters will be frozen. We use a mask M to distinguish the two types of parameters. Precisely, the K in Top-K function corresponds to the sparsity setting S at the beginning.

Second, the model re-fine-tuning on the target language, and this time, the mask M given by the previous step is used. Parameter updating is similar to equation (4). Notice that the L1 regularization is applied in this work to encourage more robust parameter updating.

Gradient based Fine-tuning(fisher): Fisher Information [26] determines the update direction in natural gradient descent, which is a nonlinear function of the weights and data [27]. As described below, the Fisher is usually given by a hessian matrix:

$$\mathbf{H} = E_X[J^T J], J = \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t}, \tag{5}$$

where H is Hessian Matrix, and J denote the gradient matrix of the model. Given the task-specific training data D, we use the diagonal elements of the empirical Fisher Information Matrix (FIM) to estimate the task-related importance of the parameters , as described in [21]. We derive the Fisher information for the i-th parameter as shown in (6):

$$\mathbf{H} = \sum_{j=1}^{|D|} \left(\frac{\partial \log p(y_j \mid \mathbf{x}_j; \theta)}{\partial \theta^{(i)}} \right)^2 \tag{6}$$

We rank all parameters based on their Fisher Information, assuming that those with higher Fisher Information are more important for the target task. Then, we use the Top-K function to select the highly relevant parameters to form the subnetwork. During the back propagation process, only the parameters in the subnetwork are updated by gradients, while the other non-subnetwork parameters are frozen.

TABLE I
STATISTICS OF ANNOTATED LANGUAGE DATASETS.

Dialect	Size (Hours)			Speakers	Utterances
	Train	Dev	Test	Total	Total
Am-do	133.1	5.8	2.1	114	113643
U-Gtsang	4.6	2.2	2.3	8	7892
Khampa	5.1	2.5	2.4	8	7996

III. TASK DESCRIPTION AND BASELINE SYSTEMS

A. Dataset

We employ a comprehensive Tibetan dataset, including Amdo, U-Gtsang, and Khampa. The corpus is recorded in a serene environment, and sentences are employed in everyday discourse. The data acquisition process employed a monophonic

recording setup with a sampling rate of 16 kHz and 16-bit quantization. The speakers who participated in the recordings were youthful denizens hailing from Tibetan regions, aged between 18 and 23. The duration of each audio is generally within ten seconds. Subsequently, the datasets were segregated into 'train', 'dev' and 'test' splits for training and validation correspondingly for all datasets. The size details of the dataset are presented in Table I.

B. Baseline system description

We utilize Transformer[3] based ASR model implemented in SpeechBrain [28]. In order to better evaluate the performance of our proposed method on models with different numbers of parameters, we designed two Transformer models with different model size. For the transformer-small, the encoder has 12 layers and the decoder has 6 layers with width 256. For the transformer-large, the encoder has 16 layers and the decoder has 8 layers with width 512, and other parameters are exactly the same. Each layer includes multi-head attention and fully connected layer. We extract 80 dimensions fbank features (windows with 25ms size and 10ms shift). Except for the mentioned parameters, all other settings are exactly the same as in this work.

IV. Experimental results

TABLE II
WORD ERROR RATE (WER) FOR CROSS-DIALECT FINE-TUNING FROM AM-DO TO U-GTSANG AND KHAMPA.

Model	Params (M)	Method	Khampa		U-Gtsang		Avg
			Dev	Test	Dev	Test	
Transformer (Small)	27.3	vallina	14.08	14.22	23.60	24.46	19.09
		fisher	14.18	14.43	24.27	24.86	19.43
		diff	14.02	14.20	22.14	22.47	**18.20**
Transformer (Large)	87.5	vallina	12.62	13.30	23.45	24.91	18.57
		fisher	10.30	11.65	21.20	22.28	16.35
		diff	10.26	10.93	19.76	20.02	**15.24**

We compare our proposed two fine-tuning strategies: diff and fisher, to vanilla fine-tuning (update all parameters). As shown in Table II, our proposed methods outperform vanila fine-tuning, achieving a relative reduction in WER ranging from 5% - 18%. In particular, the 'diff' method exhibits superior efficacy on U-Gtsang and Khampa compared to 'fisher'. Upon comparing the experimental results of the 'Transformer (Small)' model with those of the 'Transformer (Large)' model, , it can be observed that as the model's parameter size increases (from 27.3M to 87.5M), the advantages of the sparse fine-tuning method become more pronounced.

TABLE III
COMPARED TO OTHER SPARSE FINE-TUNING METHODS

Method	Khampa		U-Gtsang		Avg
	Dev	Test	Dev	Test	
vallina	12.62	13.30	23.45	24.91	18.57
topk	13.37	13.92	23.38	24.26	18.73
random	20.17	21.45	31.34	37.64	27.65
adapter	18.04	19.11	28.66	30.58	24.09
diff	10.26	10.93	19.76	20.02	**15.24**

We further compare our proposed method with other sparse fine-tuning techniques, such as topk (selecting subnetwork from the pre-trained model), random (randomly selecting parameters as subnetwork), and adapter-based methods, employing Transformer (Large) as the foundational model. As exhibited in Table III, the experimental results indicate 'diff' method is superior to other sparse fine-tuning methods. Moreover, it suggests that obtaining task-specific sparse subnetworks from the target domain enables a more precise depiction of the target domain's characteristics, thereby avoiding the necessity for extensive data transformation or alignment between the source and target domain.

TABLE IV
THE EXPERIMENTAL RESULTS OF FUSING KNOWLEDGE REPRESENTED BY VARIOUS METHODOLOGIES.

Method	Combination Type	Khampa		U-Gtsang		Avg
		Dev	Test	Dev	Test	
vallina	-	12.62	13.30	23.45	24.91	18.57
fisher	-	10.30	11.65	21.20	22.28	16.35
diff	-	10.26	10.93	19.76	20.02	**15.24**
diff + fisher	union	10.41	11.09	20.05	20.53	15.52
	intersection	10.40	11.23	20.51	21.38	15.88
diff(30%) + topk(10%)	union	10.02	10.27	19.82	20.13	**15.06**
	intersection	10.73	11.49	21.14	21.68	16.26

To analyze whether there is a correlation between the knowledge encapsulated in the subnetworks extracted using different sparse fine-tuning methods, we conducted experiments utilizing distinct combination techniques (union and intersection) and employ Transformer (Large) as the foundational model. As shown in Table IV, compared to using the diff method alone, there is a slight degradation in performance when combining the two variants of the method proposed by us (diff and fisher). However, by conjoining the knowledge derived from the source domain ('topk') and the target domain ('diff'), the efficacy of the model can be further enhanced.

As Figure 3, to investigate how diverse levels of sparsity may impact the fine-tuning efficacy, we varied our target sparsity rate to span from 10% to 100%. Notice that the sparsity denotes the trainable weights compared to the entire model. We analyze the effect of sparsity on both low-resource datasets and found that lower degrees of sparsity result in greater enhancements compared to higher levels of sparsity. Meanwhile, the model achieved its highest accuracy in low-

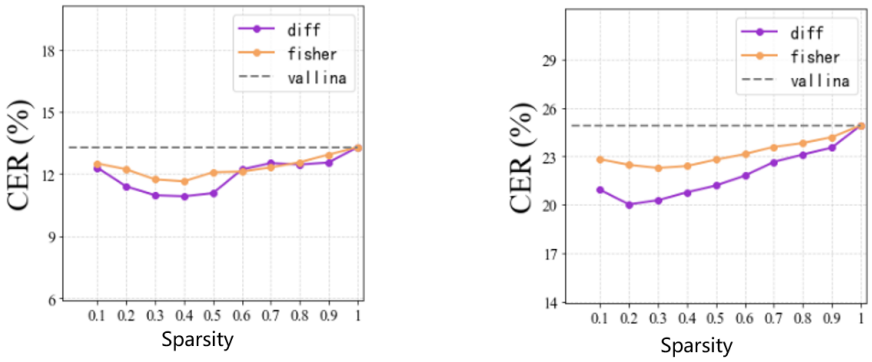


Fig. 3. Performance across two dialects of limited resources, contemplating the range of sparsity from 0.1 to 1. Sparsity denotes the remaining weights compared to the full model.

resource languages when the sparsity ranged from 10% to 30%.

V. Conclusions

Conventional methodologies of transfer learning often suffer from the issue of knowledge forgetting, leading to suboptimal performance in Tibetan dialect transfer learning. In this paper, we propose two sparse fine-tuning techniques based on the lottery ticket hypothesis, which evaluates the importance of each parameter for the target low-resource dialect. By refining the language-specific subnetwork and solidifying the remaining parameters, we mitigate knowledge loss, thereby achieving a more potent transfer of knowledge among Tibetan dialects.

Acknowledgment

Thanks to the National Natural Science Foundation of China (No. 62261045), National Key R&D Program of China (No. 2020YFC2004103) and National Natural Science Foundation of China (No. 61876131, U1936102).

References

[1] L. Pan, S. Li, L. Wang, and J. Dang, "Effective training end-to-end asr systems for low-resource lhasa dialect of tibetan language," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2019, pp. 1152–1156.

[2] S. Qin, L. Wang, S. Li, J. Dang, and L. Pan, "Improving low-resource tibetan end-to-end asr by multilingual and multilevel unit modeling," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–10, 2022.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "A neural network for large vocabulary conversational speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, 2016.

[5] X. Song, Z. Wu, Y.Huang, C. W, D. Su, and H. Meng, "Non-autoregressive transformer asr with ctc-enhanced decoder input," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5894–5898, 2021.

[6] A. Kuznetsova, A. Kumar, J. D. Fox, and F. M, "Curriculum optimization for low-resource speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8187–8191, 2022.

[7] S. Ueno, T. Moriya, M. Mimura, *et al.*, "Encoder transfer for attention-based acoustic-to-word speech recognition," *INTERSPEECH*, pp. 2424–2428, 2018.

[8] T. Moriya, R. Masumura, T. Asami, *et al.*, "Progressive neural network-based knowledge transfer in acoustic models," *Asia-Pacifific Signal and Information Processing Association Annual Summit and Confer ence (APSIPA ASC)*, pp. 998–1002, 2018.

[9] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov, "Emerging cross-lingual structure in pretrained language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6022–6034.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[11] F. Zhuang, Z. Qi, K. Duan, *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[12] V. Joshi, R. Zhao, R. R. Mehta, K. Kumar, and J. Li, "Transfer learning approaches for streaming end-to-end speech recognition system," *arXiv preprint arXiv:2008.05086*, 2020.

[13] Z. Wang, Z. C. Lipton, and Y. Tsvetkov, "On negative interference in multilingual models: Findings and a meta-learning treatment," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4438–4450.

[14] A. Aghajanyan, A. Shrivastava, A. Gupta, N. Goyal, L. Zettlemoyer, and S. Gupta, "Better fine-tuning by reducing representational collapse," *arXiv preprint arXiv:2008.03156*, 2020.

[15] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7870–7881.

[16] C. Lee, K. Cho, and W. Kang, "Mixout: Effective regularization to finetune large-scale pretrained language models," in *International Conference on Learning Representations*, 2020.

[17] A. Ansell, E. M. Ponti, J. Pfeiffer, *et al.*, "Mad-g: Multilingual adapter generation for efficient cross-lingual transfer," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4762–4781.

[18] A. Kannan, A. Datta, T. N. Sainath, *et al.*, "Large-scale multilingual speech recognition with a streaming end-to-end model," *arXiv preprint arXiv:1909.05330*, 2019.

[19] W. Hou, H. Zhu, Y. Wang, *et al.*, "Exploiting adapters for cross-lingual low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2021.

[20] A. Ansell, E. Ponti, A. Korhonen, and I. Vulić, "Composable sparse fine-tuning for cross-lingual transfer," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1778–1796.

[21] R.Xu, F.Luo, Z.Zhang, *et al.*, "Raise a child in large language model: Towards effective and generalizable fine-tuning," *Association for Computational Linguistics(ACL)*, pp. 9514–9528, 2021.

[22] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.

[23] Z. Wu, D. Zhao, Q. Liang, J. Yu, A. Gulati, and R. Pang, "Dynamic sparsity neural networks for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6014–6018.

[24] Z.Lin, L. Wu, M.Wang, and L. Li, "Learning language specific sub-network for multilingual machine translation," *Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pp. 293–305, 2021.

[25] D. Guo, A. Rush, and Y. Kim, "Parameter-efficient transfer learning with diff pruning," *Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pp. 4884–4896, 2021.

[26] J.Pennington and P.Worah, "The spectrum of the fisher information matrix of a single-hidden-layer neural network," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5410–5419, 2018.

[27] D. G. S. B. Hassibi and G. J. Wolff, "Optimal brain surgeon and general network pruning," *IEEE International Conference on Neural Networks*, pp. 293–299, 1993.

[28] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624. arXiv: 2106.04624.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multi-task Piano Transcription with Local Relative Time Attention

Qi Wang, Mingkuan Liu, Xianhong Chen, Mengwen Xiong
Beijing University of Technology, China
E-mail: {wangqi91, chenxianhong, wenmeng.xiong}@bjut.edu.cn, liumkuan@emails.bjut.edu.cn

***Abstract*—Automatic music transcription (AMT) is to transcribe music audios into musical symbol representations. Recently, the Transformer-based transcription systems have shown superiority on modeling note-wise sequences. For the frame-wise transcription targets in the AMT, the attention needs to focus more on the neighboring frames instead of notes in context. In this work, we propose a multi-task transcription system with a self-attention mechanism. The designed relative positional self-attention aims to model frame-wise short-term dependencies in audio and transcribe music of variable length. Adding the learnable attention mask on multiple attention head, the network can obtain different multi-scale attention distances for each subtask. Experiments on the MAESTRO dataset show the proposed system with the local relative time attention mechanism achieves state-of-the-art transcription performance on both frame and note metrics (frame F1 93.40%, note with offset F1 88.50%).**

I. Introduction

As one of the key technologies in music information retrieval (MIR), automatic music transcription (AMT) aims to process music audio signals to music symbols [1]. A large number of tasks can benefit from AMT, such as automatic music generation [2], music education [3], and musicology research [4].

Specifically, the purpose of the piano transcription is to transform piano music to note events with pitch, onset, offset and other music symbols. Some early piano transcription methods were based on the discriminative models such as support vector machines (SVM) [5], [6] and hidden Markov models (HMM) [7]. Non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) were also proposed to decompose polyphony piano music spectra into note spectrum templates and activations [8], [9]. Due to the harmonic overlap of concurrent notes in the frequency domain, it is a challenging task for traditional hand-crafted transcription method.

With the development of deep learning, the neural networks (NNs) have been widely used in AMT systems. The multi-task framework achieved superior performance in transcription, such as the onsets and frames (OAF) model [10], [11]. Recently, the Transformer showed great success in the sequence-to-sequence transcription methods [12], [13]. These Transformers modeled note-wise sequence context information with absolute position embedding method. The Transformer's potential was also explored on the multi-task AMT system [14]. By modeling the correlation of note-wise velocity sequence, the Transformer achieved an improvement on velocity estimation subtask. Due to the global absolute position embedding, the self-attention mechanism in these Transformers connected all the frames on time steps. In this scenario, not only the transcription flexibility on variable audio length is limited by the absolute position embedding, but the global self-attention also causes confusion in the network by mixing up on several note event features.

In this work, we implement a multi-task piano transcription system with local relative time self-attention. The self-attention mechanism is improved using the local relative position embedding to model frame-wise short-term dependencies on different subtasks. In addition, a learnable attention mask is adopted to obtain the proper attention distance for each head. To aggregate the inter-head information across the multi-scale self-attention heads in the transcription subtasks, a head-wise gating mechanism is utilized to emphasize the informative features among these multi-scale attention heads. The experiments show that the proposed system achieves state-of-the-art performance on MAESTRO dataset [11] compared to previous transcription systems.

II. Related work

Multi-task automatic piano transcription systems have been deeply developed recently. These transcription systems consist of multiple subtasks, such as onset and offset detection, pitch and velocity estimation. Hawthorne et al. [10] proposed the OAF transcription system to extract the pitch features of notes attack stage, which could model the frame-level targets and the note-level targets at the same time. Kelz et al. [15] proposed a multi-task framework with an extra note offset detection subtask and decomposed a note to attack, decay, sustain and release (ADSR) phases. Kim et al. [16] improved the OAF system with adversarial learning method and Kong et al. [17] designed the high-resolution note targets for the multi-task transcription system. Wei et al. [18] used harmonic dilated convolution to capture the harmonic structures and frequency grouped recurrent neural network to model the pitch-invariance over time. With multiple modifications, the multi-task framework has became a typical piano transcription method.

Attention mechanism [19] was also introduced into the piano AMT tasks to explore the temporal relationship of the audio frames [20]. Further, the generic Transformer was directly utilized to accomplish the end-to-end piano transcription [12],

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

[13]. In these previous studies, a standard absolute position embedding method was utilized to represent the positional relationship of each audio frame in the attention mechanism. The embedded position vectors were summed to frame spectra before all the encoders and decoders. These researches indicated that the attention mechanism was promising in the the piano transcription tasks.

III. Method

A. System architecture

The architecture of the proposed piano transcription system is shown in Fig. 1(a), which we term as Local Relative Time Attention (LRTA) transcription system. There are two acoustic models and three FG-LSTMs [18] to model three transcription subtasks, which include onset detection, pitch estimation and offset detection. We use a joint binary cross entropy (BCE) loss of all subtasks for model optimization. Specifically, the loss functions are:

$$l_{\text{BCE}}(y,p) = -wy\log p - (1-y)\log(1-p) \tag{1}$$

$$l_{\text{onset}} = \sum_{k=1}^{K}\sum_{t=1}^{T} l_{\text{BCE}}(I_{\text{onset}}(t,k), P_{\text{onset}}(t,k)) \tag{2}$$

$$l_{\text{frame}} = \sum_{k=1}^{K}\sum_{t=1}^{T} l_{\text{BCE}}(I_{\text{frame}}(t,k), P_{\text{frame}}(t,k)) \tag{3}$$

$$l_{\text{offset}} = \sum_{k=1}^{K}\sum_{t=1}^{T} l_{\text{BCE}}(I_{\text{offset}}(t,k), P_{\text{offset}}(t,k)) \tag{4}$$

where all the losses are computed on frame aligned frame-wise piano-roll-liked ground truth I and model output P, K refers to all the pitches and T is the number of frames in each sample, w in (1) is a positive weight. $w = 2$ only in onset loss and $w = 1$ in other subtasks. The joint BCE loss is given as the following function:

$$l_{\text{note}} = l_{\text{onset}} + l_{\text{frame}} + l_{\text{offset}} \tag{5}$$

The acoustic model consists of stacked standard convolution layers and temporal attention blocks. Firstly, the local features are captured by the input standard convolutions. Then we use the temporal dynamic convolution (TDY-Conv) [21] and local relative time attention with residuals to extract notes local time-variant attributes from the former hidden features, which these attributes can be benefit from more on neighboring frames. Here, the local relative time attention is devised to extract short-term correlations among the frame-wise features. After the acoustic model, for each transcription subtask, FG-LSTMs are adopted to model the temporal relationship on individual pitches.

B. Local relative time attention

Inspired by Music Transformer [2], we propose the self-attention with a local relative position embedding method to extend the transcription length flexibility of the attention mechanism. Different from absolute position embedding method, the relative position embedding method focuses more on the relative distance between the elements in sequence. At each tap of the multi-head self-attention calculation, where the head dimension is d_h, there is:

$$\text{RelativeAttention} = \text{softmax}(\frac{QK^T + S}{\sqrt{d_h}})V \tag{6}$$

where Q, K, V represent query, key, and value vectors mapped from input, S is the relative position weights. For a sequence of length L, the shape of S is (L, L). Equation (6) was firstly proposed in [22]. Then the implementation of relative attention was improved by reducing its intermediate memory requirement in [2] and [23]. We further modify the relative position embedding to local situation, where each frame in the audio sequence interacts with the forward and backward terms in a limited scope. In this case, the length of the audio being transcribed can be arbitrary, as long as it is larger than the maximum attention range.

During the evolution of a piano note, the energy has a rapid increase at onset. For the onset branch, each frame is more relevant to neighboring frames in the audio sequence. Therefore, the attention distance N in different subtasks on each head may be different. To get the proper N, the adaptive attention mask [24] is simplified and adopted in the bi-directional relative position embedding method:

$$m_z(x) = \text{min}[\text{max}(z - |x|, 0), 1] \tag{7}$$

where $m_z(x)$ is the attention mask add on each S and $m_z(x) \in [0, 1]$. $x \in [-L + 1, L - 1]$ represents the directed relative distance in sequence L. $z \in [0, L - 1]$ is the learnable parameter controlling relative attention distance N. Here, the N is represented as $N = 2z + 1$. Specifically, we apply attention masks on multi-heads in different branches, which enables the network to model multi-scale temporal dependencies in different subtasks.

We use the Squeeze-and-Excitation (SE) [25] network to help the LRTA extract useful ones across these multi-scale attention heads. The SE network applied on multi-head self-attention can be regarded as a superior attention mechanism of the basic attention across different attention scales. It allows the network to aggregate the inter-head information and boost temporal feature discriminability.

As illustrated in Fig. 2, the h-heads relative position weights S in different branches are masked by the designed multi-scale attention masks. After the relative positions are skewed to the corresponding absolute positions, the multi-scale head-wise information are aggregated through the SE network.

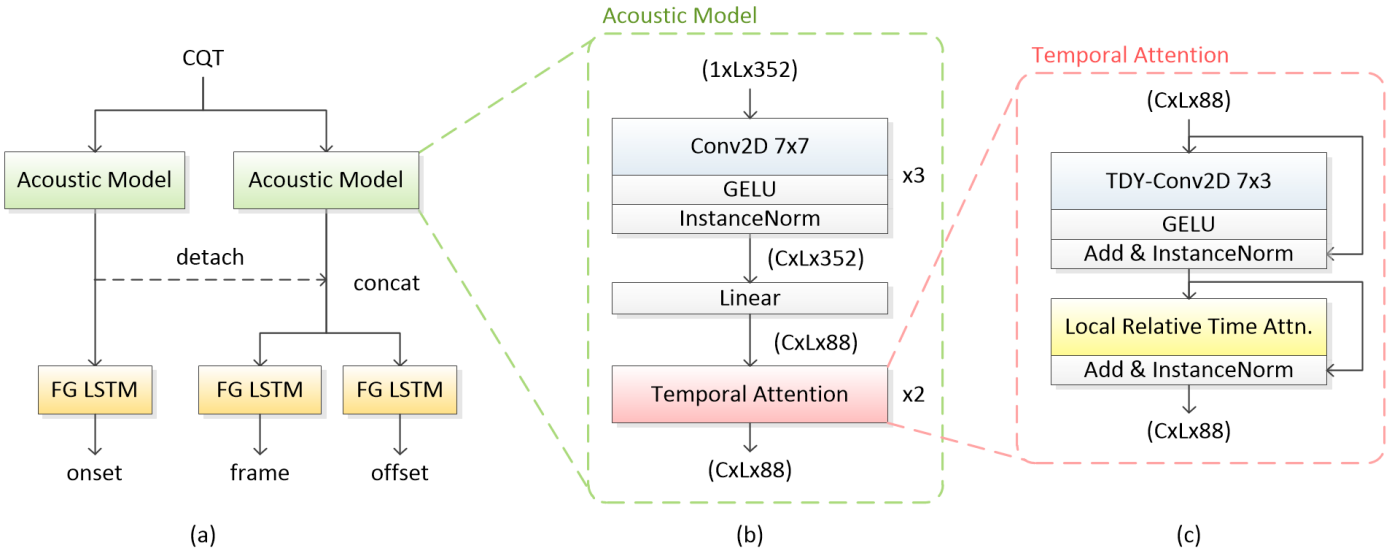


Fig. 1. Architecture of the proposed AMT system. (a) The entire architecture of the transcription system. The "detach" operation means the gradient along this path will not be backpropagated. (b) The architecture of the acoustic model. (c) The architecture of the temporal attention block.

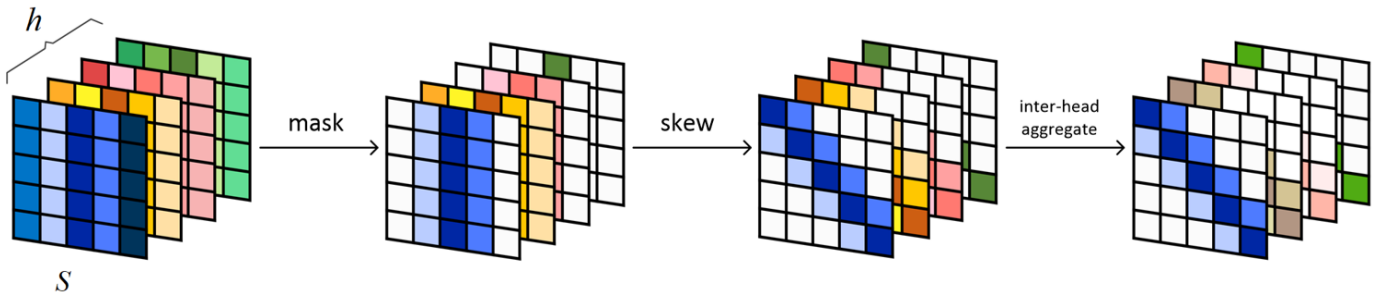


Fig. 2. Schematic diagram of the operations on the relative position weights S of h-heads.

IV. Experiments

A. Dataset and metrics

The experiments are conducted on the MAESTRO V3.0.0 dataset [11], which contains 200 hours of high-quality realistic recordings. We use the provided train, validation, and test splits in the MAESTRO.

We use the frame-level [26] and the note-level evaluation metrics in the *mir_eval* [27] library to get the precision, recall and F1 score. Following the tolerance range settings in previous researches [10], [17], we use $\pm$50ms as the onset tolerance and use offset tolerance of $\pm$50ms or 20% of the total note duration. The velocity estimation includes a tolerance within 0.1 of the prediction and ground truth after scaling the velocity values to $[0, 1]$.

B. Preprocessing

We use the *librosa* [28] library to preprocess the audio in the MAESTRO dataset. The raw audio is downsampled to 16kHz mono soundtrack. The frame length is set to 128ms, and the hop length is 20ms. Then the log amplitude CQT [29] spectrum of each sample is calculated with a Hanning window of 2048 with minimum frequency of 27.5Hz (A0) and bins per octave of 48, which the total frequency bins number is 352. At last, the dynamic range of the spectrum is rescaled to $[-40\text{dB}, +40\text{dB}]$. Each training sample contains 600 frames, which is corresponding to 12 seconds of the raw piano audio. During the evaluation, the length of each test sample is 19200 frames, which contains 384 seconds of the piano audio.

For the label preprocessing, when a note is played and the sustain pedal value is no less than 64, the note will be extend until the key is re-pressed or the sustain pedal is off [10].

C. Model configuration

As Fig. 1(a) shown, the proposed AMT system consists of acoustic models and FG-LSTMs. The acoustic model, as Fig. 1(b) shown, are composed of a input batch normalization [31], stacked standard CNNs, a linear layer and temporal attention blocks. The input standard CNNs are composed of three convolution blocks, and each contains a convolution kernel with a size of 7×7, a GELU layer, a instance-norm

TABLE I
TRANSCRIPTION EVALUATION RESULTS ON THE MAESTRO DATASET

Model	FRAME			NOTE			NOTE W/OFFSET			NOTE W/OFF. & VEL.		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
	MAESTRO V1.0.0											
OAF [11]	92.11	88.41	90.15	98.27	92.61	95.32	82.95	78.24	80.50	79.89	75.37	77.54
GAN [16]	93.10	89.80	91.40	98.10	93.20	95.60	83.50	79.30	81.30	82.30	78.20	80.20
Transformer [12]	-	-	-	98.62	93.46	95.95	85.77	81.31	83.46	84.45	80.07	82.18
LRTA(proposed)	94.79	93.59	**94.17**	98.55	96.18	**97.34**	90.08	87.94	**88.98**	88.51	86.42	**87.45**
	MAESTRO V2.0.0											
HPT [17]	88.71	90.73	89.62	98.17	95.35	96.72	83.68	81.32	82.47	82.10	79.80	80.92
HPPNet [18]	92.36	93.46	92.86	98.31	96.18	**97.21**	85.36	83.54	84.41	83.85	82.08	82.93
Semi-CRFs [30]	93.85	88.72	91.11	98.66	94.50	96.51	90.68	86.89	**88.72**	89.68	85.96	**87.75**
LRTA(proposed)	93.99	92.18	**93.04**	97.47	95.68	96.47	88.76	87.07	87.83	86.89	85.18	85.95
	MAESTRO V3.0.0											
HPT-T [14]	-	-	90.09	97.88	95.72	96.77	84.13	82.31	83.20	82.85	81.07	81.90
Transformer [12]	-	-	-	98.61	93.60	96.01	86.19	81.86	83.94	84.95	80.70	82.75
HPPNet [18]	92.79	93.59	93.15	98.45	95.95	**97.18**	84.88	82.76	83.80	83.29	81.24	82.24
Semi-CRFs [30]	93.79	88.36	90.75	98.69	93.96	96.11	90.79	86.46	88.42	89.78	85.51	**87.44**
LRTA(proposed)	94.40	92.47	**93.40**	98.37	95.70	97.00	89.73	87.33	**88.50**	88.02	85.68	86.82

layer, and channel expansion from 1 to 16 and $C = 64$. After that, a linear layer is applied to map the frequency bins from 352 to 88.

For the temporal attention block, the TDY-Conv layer is consisted of 4 temporal dynamic convolution kernels, which each kernel has larger reception filed along the time axis and has the shape of 7×3. The reduction factor in the convolution channels squeeze is set to 8. The local relative time attention is consisted of 64 heads, which each head is corresponded to each channel in the previous TDY-Conv layer. We use 8 learnable attention mask parameter z to adjust every 8 heads' attention range. We set $z = 49$ as the maximum attention range, where the maximum relative distance $N = 99$. In addition, the inter-head information are aggregated by the SE network with the reduction factor of 8.

For the FG-LSTM, we follow the configuration proposed in [18]. There are 64 units in each direction of a one layer bidirectional LSTM. At last, the outputs of all the subtasks are $T \times 88$ piano rolls, where T refers to the frame number of the audio sample and 88 indicates the number of piano keys. All the outputs are rescaled by a sigmoid layer.

We use an extra acoustic and FG-LSTM for note velocity estimation branch. The velocity branch has the same configuration as described above, except for the local relative time attention, which we found that replacing the attention layer with a linear layer resulted in better performance. The onset hidden feature is also concatenated to the velocity branch before the FG-LSTM. The velocity estimation module is optimized with the multi-task transcription model, and its mean square error loss is only computed at onset valid points.

The entire model contains 3.2M trainable parameters. We set batch size of 8 for training, and use the Adam [32] optimizer with learning rate of 0.0005, and reduce the learning rate to 98% after every 5 epochs. The complete model is trained on two GeForce RTX 3090 for about 70 hours, with a total of 35 epochs of 210K steps.

In post-processing, we use 0.4 and 0.4 as the onset and the frame output posterior probabilities binarization thresholds. Notice that the offset is only used in training. To get complete note events, we start each note with its velocity at onset valid point, and end the note at frame invalid point or re-onset point.

V. RESULTS

A. Final system

We compare the proposed LRTA system with other state-of-the-art systems in Table I. Among the existing AMT systems, the HPPNet [18] is the best model in the frame F1 score and the note F1 score, the Semi-CRFs [30] used Semi-Markov Conditional Random Fields to directly predict note events and is the best model in the note with offset and velocity F1 scores.

The LRTA model achieves state-of-the-art in both frame F1 score and note with offset F1 score in the MAESTRO V3.0.0 dataset, which improves 0.25% and 0.08%, reaching 93.40% and 88.50% respectively. We also evaluate our LRTA transcription model on the test splits of MAESTRO V1.0.0 and V2.0.0. Most of the performances in different versions of the MAESTRO datasets are the same. While there are additional performances in v2 and v3 comparing with v1, and several erroneously included recordings were removed in v3. It is remarkable that the LRTA model outperforms all other models in the frame F1 scores, which indicates the superior performance of its designed local relative time attention on frame-wise targets. With a total of 3.2M parameters, the LRTA model's effectiveness makes it an ideal transcription model for various tasks.

TABLE II
ABLATION RESULTS ON THE MAESTRO DATASET

Model	FRAME			NOTE			NOTE W/OFFSET			NOTE W/OFF. & VEL.		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
	MAESTRO V3.0.0											
LRTA	94.40	92.47	**93.40**	98.37	95.70	**97.00**	89.73	87.33	**88.50**	88.02	85.68	**86.82**
w/o attention	94.63	91.45	92.97	98.86	94.92	96.83	89.66	86.12	87.84	87.72	84.28	85.95

B. Ablation study

To demonstrate the contribution of the local relative time attention, we conduct an ablation study on the MAESTRO dataset. We replace the self-attention layer with a single linear layer, which the linear layer has the 88 input and output features. The results are shown in Table II. These results indicate that the the local relative time attention layer is crucial in all transcription tasks, where all the F1 scores decrease after the self-attention layer is ablated.

VI. CONCLUSIONS

In this paper, we propose a multi-task automatic piano transcription system, which utilizes the adaptive local relative position embedding self-attention to model frame-wise transcription targets. The relative position embedding makes it possible to transcribe arbitrary length of piano music. The local self-attention mechanism helps the network to focus more on the temporal features across adjacent audio frames. We also apply learnable attention masks and a head-wise gate mechanism to improve and aggregate multi-scale attention ranges. The experimental results show that the designed local relative time attention is effective on modeling frame-wise short-term denpendencies, and the proposed system achieves state-of-the-art performance both in frame and note transcriptions. For the future work, it is worth exploring a reasonable self-attention method for the velocity estimation subtask.

VII. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China No.62001012.

REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic Music Transcription: An Overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[2] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, *et al.*, "Music transformer: Generating music with long-term structure," in *7th International Conference on Learning Representations, ICLR 2019, May 6, 2019 - May 9, 2019*, 2019.

[3] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music information retrieval meets music education," in *Dagstuhl Follow-Ups*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, vol. 3, 2012.

[4] Q. Kong, B. Li, J. Chen, and Y. Wang, "Giantmidi-piano: A large-scale midi dataset for classical piano music," *arXiv preprint arXiv:2010.07061*, 2020.

[5] G. Costantini, M. Todisco, R. Perfetti, R. Basili, and D. Casali, "SVM based transcription system with short-term memory oriented to polyphonic piano music," in *15th IEEE Mediterranean Electrotechnical Conference, MELECON 2010, April 25, 2010 - April 28, 2010*, 2010, pp. 196–201.

[6] G. Costantini, M. Todisco, and R. Perfetti, "On the use of memory for detecting musical notes in polyphonic piano music," in *ECCTD 2009 - European Conference on Circuit Theory and Design Conference Program, August 23, 2009 - August 27, 2009*, 2009, pp. 806–809.

[7] V. Emiya, R. Badeau, and B. David, "Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches," in *16th European Signal Processing Conference, EUSIPCO 2008, August 25, 2008 - August 29, 2008*, 2008, Canton de Vaud, Interactive Multimodal Information Management (IM), SIMILAR–Network of Excellence, Swiss National Science Foundation (SNSF.

[8] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, 2015, pp. 701–707.

[9] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *11th International Society for Music Information Retrieval Conference, ISMIR 2010, August 9, 2010 - August 13, 2010*, 2010, pp. 489–494.

[10] C. Hawthorne, E. Elsen, J. Song, *et al.*, "Onsets and frames: Dual-objective piano transcription," in *19th International Society for Music Information Retrieval Conference, ISMIR 2018, September 23, 2018 - September 27, 2018*, 2018, pp. 50–57.

[11] C. Hawthorne, A. Stasyuk, A. Roberts, *et al.*, "Enabling factorized piano music modeling and generation with the Maestro dataset," in *7th International Conference on Learning Representations, ICLR 2019, May 6, 2019 - May 9, 2019*, 2019.

[12] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-Sequence Piano Transcription

with Transformers," in *International Society for Music Information Retrieval Conference*, 2021.

[13] J. P. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-Task Multitrack Music Transcription," in *International Conference on Learning Representations*, 2022.

[14] L. Ou, Z. Guo, E. Benetos, J. Han, and Y. Wang, "EXPLORING TRANSFORMER'S POTENTIAL ON AUTOMATIC PIANO TRANSCRIPTION," in *47th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022, May 23, 2022 - May 27, 2022*, vol. 2022-May, 2022, pp. 776–780.

[15] R. Kelz, S. Bock, and G. Widmer, "Deep Polyphonic ADSR Piano Note Transcription," in *44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019, May 12, 2019 - May 17, 2019*, vol. 2019-May, 2019, pp. 246–250.

[16] J. W. Kim and J. P. Bello, "Adversarial learning for improved onsets and frames music transcription," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, 2019, pp. 670–677.

[17] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-Resolution Piano Transcription with Pedals by Regressing Onset and Offset Times," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[18] W. Wei, P. Li, Y. Yu, and W. Li, "HPPNet: Modeling the Harmonic Structure and Pitch Invariance in Piano Transcription," in *Proceedings of the First MiniCon Conference*, Dec. 2022. (visited on 04/03/2023).

[19] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *31st Annual Conference on Neural Information Processing Systems, NIPS 2017, December 4, 2017 - December 9, 2017*, vol. 2017-December, 2017, pp. 5999–6009.

[20] K. W. Cheuk, Y.-J. Luo, E. Benetos, and D. Herremans, "Revisiting the Onsets and Frames Model with Additive Attention," in *2021 International Joint Conference on Neural Networks, IJCNN 2021, July 18, 2021 - July 22, 2021*, vol. 2021-July, 2021, IEEE Computational Intelligence Society, International Neural Network Society.

[21] S.-H. Kim, H. Nam, and Y.-H. Park, "Temporal Dynamic Convolutional Neural Network for Text-Independent Speaker Verification and Phonemic Analysis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6742–6746.

[22] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018, June 1, 2018 - June 6, 2018*, vol. 2, 2018, pp. 464–468.

[23] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, July 28, 2019 - August 2, 2019*, 2020, pp. 2978–2988.

[24] S. Sukhbaatar, E. Grave, P. Bojanowski, and A. Joulin, "Adaptive attention span in transformers," in *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, July 28, 2019 - August 2, 2019*, 2020, pp. 331–335.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141.

[26] M. Bay, A. F. Ehmann, and J. Stephen Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *10th International Society for Music Information Retrieval Conference, ISMIR 2009, October 26, 2009 - October 30, 2009*, 2009, pp. 315–320.

[27] C. Raffel, B. McFee, E. J. Humphrey, *et al.*, "Mir_eval: A transparent implementation of common MIR metrics," in *15th International Society for Music Information Retrieval Conference, ISMIR 2014, October 27, 2014 - October 31, 2014*, 2014, pp. 367–372.

[28] B. McFee, C. Raffel, D. Liang, *et al.*, "Librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, K. Huff and J. Bergstra, Eds., 2015, pp. 18–24.

[29] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[30] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the Frame-Level: Event-Based Piano TranscriptionWith Neural Semi-CRFs," in *35th Conference on Neural Information Processing Systems, NeurIPS 2021, December 6, 2021 - December 14, 2021*, vol. 25, 2021, pp. 20 583–20 595.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015, July 6, 2015 - July 11, 2015*, vol. 1, 2015, pp. 448–456.

[32] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Real and imaginary part interaction network for monaural speech enhancement and de-reverberation

Zehua Zhang, Changjun He, Shiyun Xu and Mingjiang Wang
Harbin Institute of Technology, Shenzhen
21B952001@stu.hit.edu.cn, hehelita@163.com, 21S052011@stu.hit.edu.cn, mjwang@hit.edu.cn

***Abstract*—Speech enhancement and de-reverberation are the key points of speech signal front-end processing. Direct mapping of speech spectrum is a standard method in speech enhancement and de-reverberation. In previous studies, the real and imaginary parts of the spectrum are estimated separately, which would make the error of direct estimation more significant. This paper proposes a real and imaginary parts interactive network (RIINet) in which the branches of real and imaginary parts interact through a linear complex attention mechanism. In addition, this paper proposes a time-frequency analysis module to improve the performance of local time-frequency feature extraction effectively. RIINet achieved state-of-the-art performance in the Deep Noise Suppression Challenge using minimal model parameters. Excellent performance has also been achieved in the noise-reverberation speech enhancement test set.**

I. Introduction

In natural environments, the presence of noise and reverberation can significantly disrupt the clarity of speech. While traditional speech enhancement methods [1], [2] have been developed over many years, they primarily address stationary noise and may struggle to handle non-stationary noise effectively. However, numerous highly effective speech enhancement methods have been proposed with the emergence of deep neural networks (DNN). Because of the nonlinear mapping capabilities of DNNs, DNN-based methods mitigate non-stationary noise compared to traditional methods. DNN-based speech enhancement methods can be broadly categorized into two groups: time-frequency domain speech enhancement methods and time-domain speech enhancement methods.

In the time domain, DNN-based speech enhancement methods such as CleanUNet [3] and DenseCNN [4] often utilize end-to-end waveform recovery. In these methods, the input consists of the noisy waveform and the enhanced waveform is directly predicted, encompassing the enhancement of both magnitude and phase during this process. Time-domain enhancement offers the advantage of circumventing the need for signal conversion from the time domain to the frequency domain. It also enables the utilization of smaller frame lengths and frameshifts, thereby reducing the model's real-time processing delay. However, it is essential to note that when the frame length is set to a minimal value, additional calculations are required to ensure the model's performance remains uncompromised.

In the time-frequency domain, the primary training targets in DNN-based speech enhancement methods involve spectral mapping methods and spectral masking methods. Owning to the powerful nonlinear mapping capability of deep neural networks, the spectrum mapping method can directly estimate the clean speech spectrum. There are magnitude spectral mapping methods [5] and complex spectral mapping methods [6]. Furthermore, spectral mapping methods encompass magnitude spectral masking methods such as ideal binary masking (IBM) [7], [8] and ideal ratio masking (IRM) [9], [10], as well as complex spectral masking methods like complex ideal ratio mask (cIRM) [11], [12].

In previous spectral mapping speech enhancement methods [6], the real and imaginary parts of the spectrum are estimated by independent branches. In complex operations, the real and imaginary parts are closely related. Estimating the real and imaginary parts separately may lead to large errors. This paper proposes a real and imaginary part interactive network (RIINet) for speech enhancement and de-reverberation. In this paper, we propose that the time-frequency analysis module can capture local time-frequency information and specific long-term context dependence. We propose a linearly complex attention module for the interaction of real and imaginary parts. RIINet outperforms state-of-the-art models on the blind test set of the Deep Noise Suppression Challenge [13]. Excellent denoising and de-reverberation performance are demonstrated on the noisy-reverberant verification set.

II. Real and imaginary part interaction network

The overall structure of RIINet is shown in Figure1. k ,d, s denote kernel size, dilation rate, and stride, respectively. The model adopts the classical convolutional recurrent neural network (CRN), which has been proven to be a very effective speech enhancement structure. Different from CRN, we use two branch decoders to estimate the real and imaginary parts of the complex spectrum, respectively. The detailed structure of each module is described in the following subsections.

A. Gated feature downsampling and upsampling

Convolutional recurrent neural network [6], [14] is a classical network structure in speech enhancement tasks. Convolution and deconvolution can change the spectral feature dimension to enhance the ability of the model to extract embedding. Context dependencies in speech are captured by recurrent neural networks (RNN). In this paper, Gated feature downsampling and upsampling are used. We use gated 2-D convolution for feature downsampling to reduce the feature dimensionality. The gated 2-D convolution consists of two 2-D

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

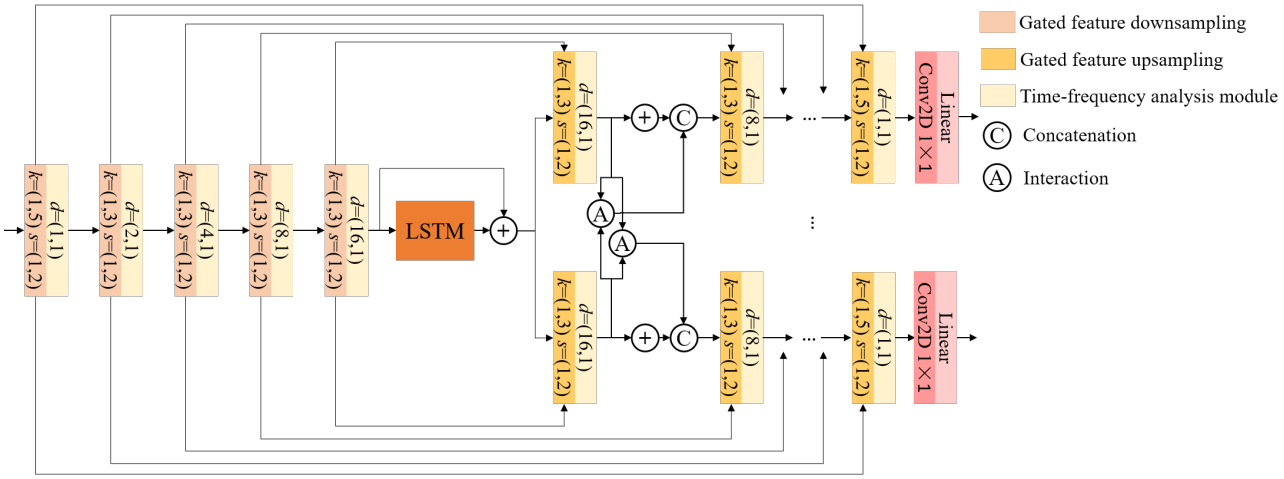


Fig. 1: Illustration of the proposed RIINet.

convolutions with the same structure, where the output of one is multiplied by the output of the other after a sigmoid activation function to generate the final output. The stride of the 2-D convolution is (1, 2) to achieve the feature dimension halving. Instance normalization and PReLU are used after gated 2-D convolution. The structure of feature upsampling is similar to feature downsampling, except that 2-D deconvolutions replace 2-D convolutions.

B. Time-frequency analysis module

To further analyze the spectral features of noisy speech, each encoder and decoder use a time-frequency analysis module (TFAM) to extract local features. The structure of TFAM is shown in Fig.2. Instance normalization and PReLU activation functions are used after pointwise convolution. After causal padding, a depthwise 2-D convolution [15] with a kernel size of (5,5) and a dilation rate of d is applied to extract local features. Instance normalization, PReLU activation function, and pointwise convolution are applied to the extracted features. The output of TFAM is generated by adding the output of the pointwise convolution with the original input.

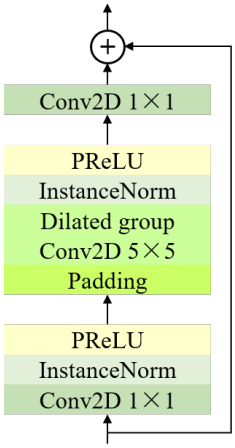


Fig. 2: Illustration of the time-frequency analysis module.

C. Real and imaginary part interaction module

To avoid the distortion of the reconstructed speech due to the separate decoding of the real and imaginary part, we add a real and imaginary part interaction module (RIIM) between the two branches. In classical attention algorithms [16], the Softmax activation function and the multiplication of matrices make attention inevitably introduce quadratic complexity. The squared complexity causes the computational cost to increase significantly as the speech duration grows. We extend Flowformer [17] to propose multi-channel flow attention for RIIM.

According to the flow network theory, the information flow in the attention mechanism can be expressed as converging from the sources (V) to the sinks (R) based on the learned flow capacity (corresponding attention weights calculated from Q and K), as shown in Fig.3. The amount of inflow information to the i-th sink I_i and outflow information from the j-th

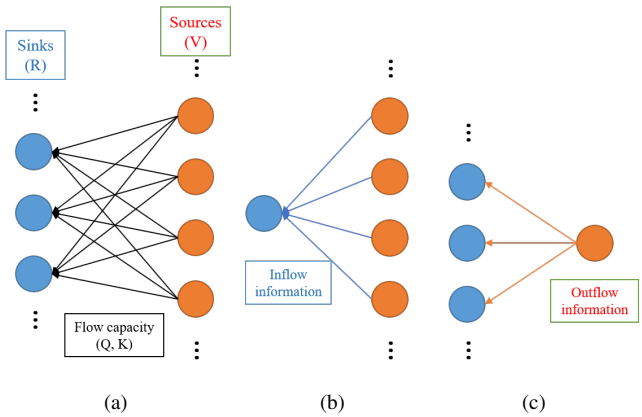


Fig. 3: (a)Sources-Sinks information flow. (b)Inflow information of each sink. (c)Outflow information of each source.

source O_j can be calculated as Eq.1. To keep the flow capacity positive, non-negative functions $\phi(\cdot)$ are used.

$$I_i = \phi(Q_i)\sum_{j=1}^{m}\phi(K_j)^T, O_j = \phi(K_j)\sum_{i=1}^{n}\phi(Q_i)^T \quad (1)$$

To utilize the conservation property of the flow network theory, the conservation of sources outflow and sink inflow are used through normalization.

$$\hat{I} = \phi(Q)\sum_{j=1}^{m}\frac{\phi(K_j)^T}{O_j}, \widehat{O} = \phi(K)\sum_{i=1}^{n}\frac{\phi(Q_i)^T}{I_i} \quad (2)$$

where $\hat{I} \in R^{n\times 1}$ indicates the amount of information obtained by each sink after the competition when the amount of source outflow information is fixed. $\widehat{O} \in R^{m\times 1}$ indicates the amount of the information supplied by each source when the amount of inflow information is fixed. Finally, the attention mechanism based on flow network theory can be expressed as the Eq.3:

$$\begin{aligned} \text{Competition}: \widehat{V} &= \text{Softmax}(\widehat{\text{O}}) \odot V \\ \text{Aggregation}: A &= \frac{\phi(Q)}{I}\left(\phi(K)^T\hat{V}\right) \\ \text{Allocation}: R &= \text{Sigmoid}(\hat{I}) \odot A \end{aligned} \quad (3)$$

where $\odot$ indicates the element-wise multiplication. $\hat{V} \in R^{m\times d}$ represents the source after the competition, which is reweighted according to the inflow conservation. $\hat{A} \in R^{m\times d}$ is the aggregated information source, which is calculated by the associativity of matrix multiplication to reduce the computational complexity. $\hat{R} \in R^{n\times d}$ represents the results of this new attention mechanism. The whole process of the SDCAN based on flow network theory is shown in Fig.4.

In calculating the real part to imaginary part attention, the real part is used as Q, and the imaginary part is used as K and V, and vice versa. By this method, we achieve to compute the attention value between real part to imaginary part with linear complexity.

D. Loss function

We use a combined mean error loss function because both magnitude and phase enhancement are essential for speech quality improvement. Research [18] shows that the complex compressed spectrum is closer to subjective human hearing than the complex spectrum. The combined mean square error loss function is demonstrated in Eq.4

$$\mathcal{L}_{cMSE} = (1-\beta)\cdot\|\hat{S}_{cRI} - S_{cRI}\|^2 + \beta\cdot\|\hat{S}_{cMag} - S_{cMag}\|^2 \quad (4)$$

where $\hat{S}_{cRI}$ and $\hat{S}_{cMag}$ denote the enhanced complex compression spectrum and the enhanced magnitude compression spectrum, respectively. In this paper, β takes 0.7.

III. Experiments and Discussions

A. Experiment setup

We have created two training datasets for the purpose of comparing advanced models and conducting the ablation study. To construct the first dataset, we have extracted clean speech and noise samples from the 1st Deep Noise Suppression Challenge [13]. This dataset comprises 1000 hours of training data, which we use to evaluate the speech enhancement performance in comparison to other state-of-the-art (SOTA) models. The noisy speech in this dataset has a signal-to-noise ratio (SNR) ranging from -5 dB to 15 dB, with approximately 75% of the samples mixed with reverberation. The reverberation time (RT60) varies between 0.3 seconds and 1.3 seconds.

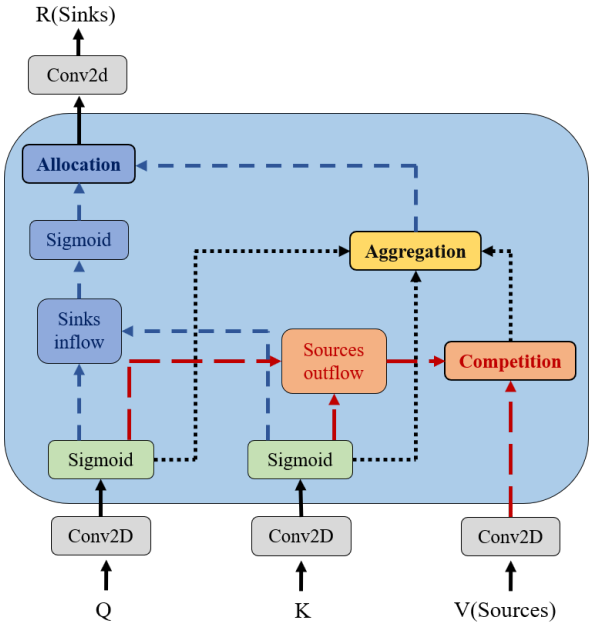


Fig. 4: Illustration of the real and imaginary part interaction module.

For the ablation study and comparison of noisy-reverberant speech enhancement performance, we also utilize the clean speech and noise samples from the 1st Deep Noise Suppression Challenge. However, this dataset has a duration of 500 hours, and the SNR of the noisy speech ranges from -5 dB to 20 dB. All speech samples in this dataset are mixed with simulated room impulse responses, where the RT60 values span from 0.3 seconds to 1.3 seconds.

The duration of each frame is set to 20 ms, with a frameshift of 50%. Prior to the Short-Time Fourier Transform (STFT), Hamming windows are applied. Moreover, all model hyper-parameters are initialized randomly and optimized using the AdamW optimizer [24]. The initial learning rate is set to 1×10^{-3}, and it exponentially decays by 0.98 as the training epoch progresses. If the loss no longer decreases, we reduce the learning rate by half. The training process continues until the learning rate reaches 1×10^{-4}.

For assessing the speech enhancement performance against state-of-the-art (SOTA) models, we employ the following evaluation metrics: Wideband Perceptual Evaluation of Speech Quality (WB-PESQ) [25], Narrowband PESQ (NB-PESQ) [26], Short-Time Objective Intelligibility (STOI) [27], and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [28]. Ad-

TABLE I: Speech enhancement performance comparison with SOTA models.

Model	Year	#Para.(M)	Caus.	With Reverb				Without Reverb			
				WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy	-	-	-	1.822	2.753	86.62	9.033	1.582	2.454	91.52	9.07
DCCRN-E[11]	2020	3.67	T	-	3.077	-	-	-	3.266	-	-
DCCRN+[19]	2021	4.71	T	-	3.30	-	-	-	3.33	-	-
CTS-Net[20]	2021	4.35	T	3.02	3.47	92.70	15.58	2.94	3.42	96.66	17.99
FullSubNet[21]	2021	5.64	T	3.057	3.584	92.11	16.04	2.882	3.428	96.32	17.30
HGCN[22]	2022	-	T	-	-	-	-	2.883	-	96.50	18.14
FullSubNet+[12]	2022	8.67	T	3.177	3.648	93.64	16.44	3.002	3.503	96.67	18.00
FS-CANet[23]	2022	4.21	T	3.218	**3.665**	**93.93**	16.82	3.017	3.513	96.74	18.08
RIINet	2023	3.66	T	**3.277**	3.633	93.77	**17.02**	**3.146**	**3.555**	**96.94**	**18.68**

TABLE II: Speech enhancement and de-reverberation performance comparsion.

Model	#Para.(M)	WB-PESQ	STOI(%)	SI-SDR	DNSMOS	CD	LLR	SNR_{fw}
Noisy	-	1.29	78.66	4.03	1.38	4.91	0.790	9.81
Uformer[31]	9.46	2.16	88.52	9.37	**2.85**	3.03	0.420	10.82
RIINet	3.13	**2.20**	**88.35**	**9.54**	**2.85**	**2.85**	0.394	**11.19**
-RIIM	3.11	2.13	87.93	9.42	2.79	2.88	0.391	11.18
-TFAM	2.87	2.04	86.94	8.97	2.73	2.93	**0.390**	11.10

ditionally, when comparing denoising and de-reverberation performance, we incorporate the following evaluation metrics: Deep Noise Suppression Mean Opinion Score (DNSMOS) [29], Composite Distortion (CD) [30], Log-Likelihood Ratio (LLR) [30], and Signal-to-Noise Ratio with Frequency Weights (SNR_{fw}) [30].

B. Speech enhancement performance comparison with SOTA models

The results of the blind test set for the Deep Noise Suppression Challenge are presented in Table I. The evaluation includes reverberation to assess the models' robustness, although reverberation is not the primary target for suppression. The terms '#Para.' and 'Caus.' indicate the model parameters and whether the model operates causally, respectively. Our proposed RIINet exhibits significant performance enhancements compared to state-of-the-art (SOTA) models in terms of the WB-PSEQ evaluation metric, both with and without reverberation. Furthermore, RIINet performs competitively with FS-CANet regarding the NB-PESQ, STOI and SI-SDR evaluation metrics. It is worth noting that RIINet achieves these advanced performance levels while utilizing minimal model parameters.

C. Speech enhancement and de-reverberation performance comparison

In this subsection, we evaluate the denoising and de-reverberation capabilities of RIINet, and the corresponding results are presented in Table II. The performance of RIINet surpasses that of Uformer, with RIINet's model parameters being only 32% of Uformer's.

We have conducted experiments to assess the impact of our proposed modules on model performance by removing the RIIM and TFAM modules from RIINet, which serves as the backbone network. The RIIM, despite utilizing only an additional 0.02M hyperparameters, demonstrates a significant influence on model performance. This module validates the effectiveness of the proposed interaction between the real and imaginary parts and provides a cost-effective solution with low computational requirements. On the other hand, the TFAM module exhibits strong performance at the expense of approximately 0.24M hyperparameters. After removing TFAM, there is a decrease of 0.07 in average PESQ, a 0.99% decrease in STOI, a 0.45 decrease in SI-SDR, and a 0.06 decrease in DNSMOS. These findings highlight the importance of the TFAM module.

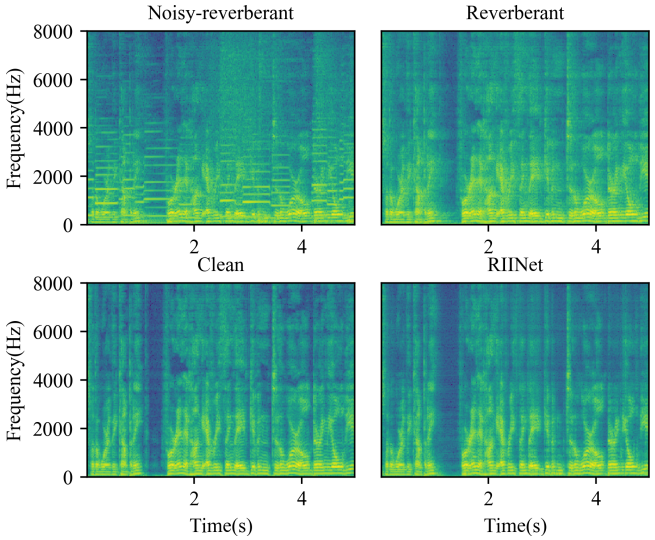


Fig. 5: Noisy-reverberat, reverberat, clean and enhanced speech spectrums

IV. Conclusions

In this study, we introduce a novel single-channel speech enhancement network architecture known as the Real and Imaginary Part Interactive Network (RIINet). Our network incorporates a two-branch interaction module with small parameters and low computational requirements, facilitating effective interaction between the real and imaginary partial branches. Furthermore, we incorporate a time-frequency analysis module that efficiently extracts local time-frequency feature information and captures long-term dependencies. The experimental results demonstrate that RIINet outperforms other state-of-the-art models in denoising and de-reverberation performance, showcasing its excellent capabilities in speech enhancement.

Acknowledgment

Thanks to the Shenzhen key project of technological research (JSGG20200831092401003) and the Joint Fund of Basic and Applied Basic Research Fund of Guangdong Province (2020A1515110498).

References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979. DOI: 10.1109/PROC.1979.11540.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7867–7871.

[4] A. Pandey and D. Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, 2021.

[5] T. Lan, J. Li, Y. Feng, *et al.*, "Multi-scale informative perceptual network for monaural speech enhancement," *APPLIED ACOUSTICS*, vol. 195, Jun. 2022.

[6] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[7] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. IV-921-IV–924.

[8] G. Kim, Y. Lu, Y. Hu, and Y. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, Jul. 2009.

[9] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, Nov. 2006.

[10] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7092–7096.

[11] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020, pp. 2472–2476.

[12] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP*, 2022, pp. 7857–7861.

[13] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech*, 2020.

[14] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Interspeech*, 2018, pp. 3229–3233.

[15] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving cnn efficiency with hierarchical filter groups," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5977–5986.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[17] H. Wu, J. Wu, J. Xu, J. Wang, and M. Long, "Flowformer: Linearizing transformers with conservation flows," in *International Council for Machinery Lubrication (ICML)*, 2022.

[18] S. Braun and H. Gamper, "Effect of noise suppression losses on speech distortion and asr performance," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 996–1000.

[19] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *Interspeech*, 2021, pp. 2816–2820.

[20] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[21] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.

[22] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "Hgcn: Harmonic gated compensation network for

speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 371–375.

[23] J. Chen, W. Rao, Z. Wang, *et al.*, "Speech Enhancement with Fullband-Subband Cross-Attention Network," in *Interspeech*, 2022, pp. 976–980.

[24] M. S. Zeke Xie Issei Sato, "Stable weight decay regularization.," in *arXiv*, 2020.

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[26] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union*, 2005.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, 2011.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[30] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, pp. 3387–405, 2009.

[31] Y. Fu, Y. Liu, J. Li, *et al.*, "Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7417–7421.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Progressive Multi-scale Self-supervised Learning for Speech Recognition

Genshun Wan[†], Hang Chen[†*], Tan Liu[‡], Chenxi Wang[†], Jia Pan[‡], Zhongfu Ye[†],
[†] University of Science and Technology of China, China
E-mail: gswan@mail.ustc.edu.cn,ch199703@mail.ustc.edu.cn,yezf@ustc.edu.cn
[‡] iFLYTEK Research, iFLYTEK Co. Ltd., China
E-mail: tanliu2@iflytek.com,jiapan@iflytek.com

***Abstract*—Self-supervised learning has shown great potential in improving automatic speech recognition (ASR) systems. However, further improvements in recognition performance could be achieved if models focus on audio content information learning. In this paper, we propose a progressive multi-scale self-supervised learning method that reinforces the learning process from easy to difficult. Our progressive strategy utilizes fine-grained target sets to compute self-supervised learning loss at the top layer while using coarse-grained target sets at intermediate layers. Additionally, to match the difficulty of the learning process, we introduce a multi-scale structure into the multi-head self-attention module. We evaluate our method on the Librispeech dataset and demonstrate its effectiveness. Our proposed method achieves a relative word error rate (WER) reduction of 13.7% and 12.7% on the test_other evaluation subsets, respectively, when fine-tuned on 10-hour and 100-hour subsets, outperforming HuBERT.**

I. Introduction

In recent years, deep neural networks have become the backbone of automatic speech recognition (ASR) systems [1], [2], yielding impressive advances in reliability and practicality. However, the reliance of these models on large amounts of supervised speech data poses a fundamental limitation to their widespread application. Therefore, exploring methods to leverage unlabeled data for ASR model training has become a significant research topic, with the potential to further enhance performance. Self-supervised learning, which uses unsupervised objectives to learn representations from raw audio, has shown promising results in other domains and has recently emerged as a promising approach for improving ASR performance without requiring human-labeled data.

In recent years, self-supervised learning (SSL) has emerged as a promising approach to learn general data representations from unlabeled speech examples. Several speech SSL methods have demonstrated impressive results for downstream tasks [3]–[9], such as speaker identification, emotion recognition, and automatic speech recognition (ASR). However, for the ASR task, it is essential to improve the generalization ability of semantic content. While some SSL methods have been successful in learning representations that correlate with phonetic units, not all layers have such high correlations. For instance, the intermediate layers in the HuBERT model have been found to have lower mutual information scores with force-aligned phonetic transcripts in [6]. To address this limitation, some recent SSL approaches, such as [10], have computed SSL loss on both intermediate and top layers using the same target labels.

As the number of layers in a self-supervised learning model increases, the hidden representations gradually adjust to the final constrained objective. We hypothesize that these hidden representations at different layers respond to different levels of information granularity. Specifically, the higher layers may focus on high-level linguistic information that is correlated with fine-grained phonetic units, while the lower layers may capture coarse-grained phonetic information. By following a learning process that progresses from simple to complex and from rough to subtle, self-supervised learning can reinforce the practicality and applicability of the learned representations for ASR tasks.

Based on the insights gained from previous research, we propose a novel approach in this paper to further enhance the self-supervised learning (SSL) of ASR models. Our method, called Progressive Multi-Scale SSL (PMS-SSL), is designed to enable the learning process to reinforce itself from easy to difficult, with a focus on the audio content information level. We adopt the two-iteration pretraining process with masked prediction loss in HuBERT as the basis for our method. The main contributions of this paper are as follows:

1) In the second iteration, we perform multiple k-means clustering on the learned representations to generate the final target sets with different granularity. The fine-grained target sets have large cluster numbers, while the coarse-grained target sets have small cluster numbers. We then compute the SSL loss on both intermediate layers and the top layer using multi-task learning, with fine-grained target sets used at the top layer and coarse-grained target sets used at intermediate layers.
2) We introduce a multi-scale structure into the self-attention module to match the target granularity. This is achieved by restricting the attention area of the self-attention module at each layer to a different scope, which is more suited for the progressive learning strategy. This approach is closely related to the diversity and discrimination of local and global contexts, which in turn affects the information granularity.

Our proposed PMS-SSL method has been evaluated on the Librispeech dataset. Following standard HuBERT, 960 hours of

*Corresponding author

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

audio are used for pre-training and two subsets with different size are used for fine-tuning. When progressive target sets and multi-scale self-attention are both introduced into pre-training, the model fine-tuned with 10 hours and 100 hours subsets can achieve 13.7% adn 12.7% word error rate reduction respectively on test_other subset compared with HuBERT. The progressive multi-scale self-supervised learning method follows the principle from easy to difficult and makes speech representation more suitable for ASR task at the audio content information level.

II. RELATED WORK

The basic idea of SSL is to learn the superior universal representations effectively based on unlabeled data by reconstructing or predicting itself with the context information. As a representative SSL model, HuBERT follows the wav2vec 2.0 architecture [5], consisting of a convolutional waveform encoder, a BERT encoder [11], a projection layer and a code embedding layer.

During pre-training, $p\%$ of the timesteps are randomly selected as start indices, and spans of l timesteps are masked. Let $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \cdots \tilde{x}_T]$ denote the corrupted sequence of T frames. The Transformer encoder module consumes $\tilde{X}$ and outputs a hidden state sequence $O = [o_1, o_2, \cdots o_T]$. The distribution over codewords is parameterized with:

$$p_f(c \mid \tilde{X}, t) = \frac{\exp\left(\mathrm{sim}\left(Ao_t, e_c\right)/\tau\right)}{\sum_{c'=1}^{C} \exp\left(\mathrm{sim}\left(Ao_t, e_{c'}\right)/\tau\right)} \tag{1}$$

where A is the projection matrix, $e_c$is the embedding for codeword c, $sim()$ computes the cosine similarity between two vectors, and τ scales the logit. And an important factor in the implementation of HuBERT is the design of objective function, which applys the prediction loss over the masked regions only. Realizing the mixture learning of acoustic and linguistic representation is very important to obtain different aspects of the speech content. We denote the cross-entropy loss computed over masked timesteps as L_m, which is defined as:

$$L_m = \sum_{t \in M} \log p_f(c \mid \tilde{X}, t) \tag{2}$$

where M denotes the set of timestep indices to be masked.

Following the idea that iterative refinement target labels can generate better pseudo-labels for the next iteration of training [12], [13], HuBERT conducts two iterations of pre-training. During the first iteration, HuBERT runs k-means clustering on MFCCs to generate targets for speech features. Then HuBERT re-clusters the learned latent representations and re-train the model with the newly discovered units for second iteration.

HuBERT is able to learn the speech representations which are highly correlated with phonetic units. However, not all layers have such a high correlation with phonetic units. In order to force more layers to learn more content information,[10] simultaneously computes the SSL loss on both the intermediate layers and the top layer. Therefore, the total loss can be formulated as:

$$L = \sum_{l \in K} L_m^l \tag{3}$$

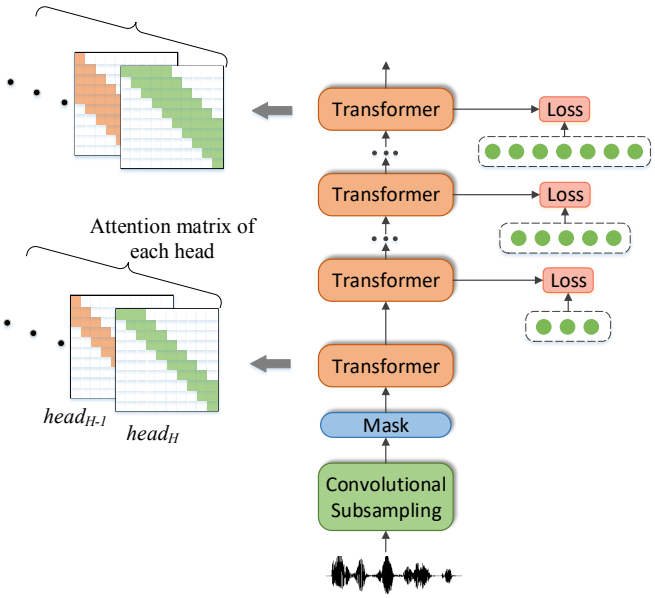


Fig. 1. Model architecture

where K denotes the set of selected layers for computing SSL loss. L_m^l denotes the predictive loss at l-th layer, which is calculated by Eq.(2).

III. PROPOSED METHOD

To improve the suitability of self-supervised learning representations for the ASR task, it is essential to prioritize context learning following the principle of progressing from easy to difficult. In this section, we present the Progressive Multi-Scale Self-Supervised Learning (PMS-SSL) method, which assigns different-granularity target sets for various layers and incorporates multi-scale structures into multi-head self-attention. The architecture of our model is shown in Fig. 1.

A. Progressive target sets

As previously discussed, prior work such as [10] has used the same target sets for both intermediate and top layers when computing the SSL loss. However, the receptive fields and learning difficulties of different layers may correspond to phonetic units of different granularities. Using the same target labels for all layers may result in overlooking the learning pattern of the different layers, and therefore, we propose a different approach.

In order to effectively guide the model to learn from easy to difficult levels, we propose the Progressive Multi-Scale Self-Supervised Learning (PMS-SSL) method, which assigns different-granularity target sets to different layers. We use the same clustering and pre-training process as HuBERT for the first iteration, given the uncertainty of target sets from MFCC features. During the second iteration, the learned representations can achieve better matching precision for different target labels. To generate multiple target sets, we perform multiple k-means clustering on the learned representations with different

cluster numbers. Coarse-grained target sets are considered to be easier to learn compared to fine-grained target sets. We then select a set of layers K as the supervised layers set to match the different-granularity target sets. Specifically, the larger target sets are used at the top layer and the smaller target sets are used at the intermediate layers. The size of different target sets is closely tied to the degree of difficulty in model learning. By assigning different-granularity target sets to different layers, our proposed PMS-SSL method reinforces the learning process from easy to difficult and helps the model learn more effectively. Therefore, the distribution defined in Eq.(1) can be refined as:

$$p_f^l(c \mid \tilde{X}, t) = \frac{\exp\left(\mathrm{sim}\left(Ao_t, e_c^l\right)/\tau\right)}{\sum_{c'=1}^{c^l} \exp\left(\mathrm{sim}\left(Ao_t, e_{c'}^l\right)/\tau\right)} \tag{4}$$

Where e_c^l is the embedding of c-th label used on the l-th layer.

We adopt the predictive loss in HuBERT which only computes SSL loss over masked regions as the object, therefore p_f in Eq.(2) is replaced by p_f^l and the total loss of all the supervised layers is also computed by Eq.(3).

B. Multi-scale multi-head self-attention

In order to improve the suitability of self-supervised learning representations for the ASR task, we propose a progressive learning strategy that leverages multi-scale structures in multi-head self-attention (MSA) [14] to emphasize the diversity of information and determine the classification target. With this approach, we bundle and deploy the receptive field of different layers as part of the progressive model.

By restricting the working area of each layer into various scopes through multi-scale attention, we enable higher layers to have larger attention areas, while lower layers have smaller attention areas. As the layer number increases, the gap in the receptive field widens. Compared to regular self-attention, which has a fixed scale of sequence length, multi-scale self-attention enables models to capture patterns at different scales and extract more robust features[15], [16].

Given a sequence of features $X \in R^{N\times D}$, where N is the length of sequence and D is the dimension of features, regular MSA first projects X into three matrices: the query $\boldsymbol{Q}$, the key $\boldsymbol{K}$ and the value $\boldsymbol{V}$, and then decompose the three matrices into H sub-spaces which corresponds to H heads. MSA computes the scaled dot-product attention in each head and integrates the results of all heads, which can be formulated as:

$$\begin{aligned} \mathrm{head}_i &= \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i \\ \boldsymbol{O} &= [\ \mathrm{head}_{\ 1},\ \mathrm{head}_{\ 2}, \cdots \mathrm{head}_{\ H}]\, W^o \end{aligned} \tag{5}$$

where i indicates i-th head, d indicates the dimension of each head, and W^o are parameter matrices.

To incorporate both local and global context information into our model, we adopt multi-scale multi-head self-attention. Specifically, we restrict the attention area of only two heads, while the other heads retain the original global attention scale. This design allows the model to learn representations that capture both fine-grained and coarse-grained information, enabling it to extract robust features from speech data. A restricted attention head can be computed as:

$$\mathrm{head}(X, w)_{i,j} = \mathrm{softmax}\left(\frac{Q_{i,j} C_{i,j}(K, w)^T}{\sqrt{d}} C_{i,j}(V, w)\right) \tag{6}$$

where j indicates j-th position, $C()$ is the function to restrict the attention area with a context window of length w . Specially, one of the two restricted heads is used to learn history context information, thus $C()$ is defined as:

$$C_{i,j}(K, w) = [K_{i,j-w}, \cdots, K_{i,j}] \tag{7}$$

Another restricted heads is used to learn future context information, in which case $C()$ is formulated as:

$$C_{i,j}(K, w) = [K_{i,j}, \cdots, K_{i,j+w}] \tag{8}$$

The history window size and future window size at each layer are the same, while the size of window grows up progressively from bottom layer to top layer. After pre-training, We use the connectionist temporal classification (CTC) loss [17] for ASR fine-tuning. Whole model weights except the convolutional audio encoder remains frozen and the projection layer is removed and replaced with a randomly initialized softmax layer.

IV. EXPERIMENTAL DETAILS

A. Data

We evaluated the performance of the PMS-SSL nethod on LibriSpeech task [18] . To keep the tasks comparable with the base HuBERT model, the full 960 hours of LibriSpeech audio are used for pre-training. For fine-tuning stage, we consider two different partitions: 10 hours and 100 hours. Each model is tested on the standard LibriSpeech test_other sets and test_clean sets respectively.

B. Experiment Setup

We followed the pre-training process and model configuration used in HuBERT to train our model. To generate the target labels for pre-training stage, we ran k-means clustering with 100 clusters on 39-dimensional MFCC features over the full 960 hours of training audio sets for the first iteration. For the second iteration, we ran multiple k-means clustering with different cluster numbers $\{50, 100, 300, 500\}$ on the latent features extracted from the 6th transformer layer of the pre-trained model to obtain different granularity target sets. The SSL loss was computed on both the top layers and intermediate layers with different target sets. Since the extracted feature dimension was much higher than the MFCC features, we randomly sampled 10% of the extracted features for clustering to reduce memory costs. The model was trained on 32 GPUs, with a batch size of 87.5 seconds of audio per GPU. The first iteration was trained for 250k steps, while the second iteration was trained for 400k steps.

As for the masking configurations, $p = 8\%$ of the waveform encoder output frames were randomly selected as mask start

and the mask span is set as $l = 10$. We also used Adam optimizer with $\beta = (0.9, 0.98)$ to train the model, and the learning rate ramped up linearly from 0 to the peak learning rate for the first 8% of the training steps, and then decayed linearly back to zero. The peak learning rate was $5e^{-4}$.

During the supervised fine-tuning, the model was fine-tuned on 8 GPUs with a batch size of 200 seconds of audio per GPU. In addition, the parameters of the waveform encoder were fixed, and only the new softmax matrix was trained. We used Adam optimizer with tri-stage schedule to fine-tune the model and adopted the word error rate (WER) on the dev-other subset as the criterion for model selection.

Finally, we used wav2letter++ [19] beam search decoder with beam size 1500 for 4-gram language model fused decoding, which optimizes:

$$\log p_{CTC}(\boldsymbol{y} \mid \boldsymbol{x}) + w_1 \log p_{LM}(\boldsymbol{y}) + w_2|\boldsymbol{y}| \tag{9}$$

where the hyperparameter settings are the same as those used in HuBERT method.

C. Results and analysis

Table 1 and Table 2 presents the performances of different pre-trained models in the w/o language model setting , all of which are pre-trained on 960h subset and fine-tuned on 10-hour and 100-hour subsets. We compare the proposed PMS-SSL method with several competitive self-supervised approaches in the literature, including wav2vec 2.0 [5], HuBERT[6] and ILS-SSL [10].It should be noted especially that ILS-SSL adds a bucket relative position embedding to the model, which further improves the performance. For comparison with ILS-SSL, we select same supervised layer set $\{4, 12\}$ and target set $\{500, 500\}$ to pre-trained models without bucket relative position embedding. As for PMS-SSL, we set the supervised layer set K as $\{6, 12\}$, and the target sets with size $\{300, 500\}$ are used for computing self-supervised learning loss. What's more, we set w (attention context window size) at layers 1-6 as 80 and set w at layers 7-12 as 160 respectively.

The results shown both in Table 1 and Table 2 suggest that the proposed mthhod are practical strategies that work, whether including 4-gram language model or not. In order to prove the practicability and generalization of the proposed pre-trained model, the results on different self-supervised approaches with 4-gram language model are taken as the main reference as described in Table 2.

In the ultra-low resource setup with 10 hours of labeled data, PMS-SSL method obtained the WERs of 4.11% / 8.27% on test_clean / test_other set respectively, which achieves a relative word error rate reduction of 4.6% / 13.7% compared with standard HuBERT, and 3.9% / 5.8% compared with ISL-SSL mthod. By further increasing the amount of labeled data from 10 hours to 100 hours, PMS-SSL method can also get stable improvement, with a relative word error rate reduction of 5.9% / 12.7% compared with HuBERT, and 3.3% / 5.3% compared with ISL-SSL. Overall, progressive learning paradigm makes self-supervised approaches obtain more competitive advantage.

TABLE I
MODEL COMPARISONS WITHOUT LM ON 10-HOUR AND 100-HOUR LABELED DATA.

Model	LM	Test_clean	Test_other
10-hour labeled			
wav2vec 2.0 Base	None	11.1	17.6
HuBERT Base	None	10.1	16.8
ILS-SSL Base	None	10.79	16.53
PMS-SSL	None	9.43	15.02
100-hour labeled			
wav2vec 2.0 Base	None	6.1	13.3
HuBERT Base	None	6.3	13.2
ILS-SSL Base	None	6.06	11.89
PMS-SSL	None	5.52	11.21

TABLE II
MODEL COMPARISONS WITH 4-GRAM LM ON 10-HOUR AND 100-HOUR LABELED DATA.

Model	LM	Test_clean	Test_other
10-hour labeled			
wav2vec 2.0 Base	4-gram	4.3	9.5
HuBERT Base	4-gram	4.3	9.4
ILS-SSL Base	4-gram	4.27	8.75
PMS-SSL	4-gram	4.11	8.27
100-hour labeled			
wav2vec 2.0 Base	4-gram	3.4	8.0
HuBERT Base	4-gram	3.4	8.1
ILS-SSL Base	4-gram	3.32	7.59
PMS-SSL	4-gram	3.21	7.19

D. Ablation Studies

To explore how the selected supervised layer set K and the target sets influence the performance of pre-trained models, we perform some ablation experiments without multi-scale self-attention. We select different sizes of target sets for the selected supervised layer. And all ablation models are finetuned with 10-hour subset and the results evaluated on test_other set are listed in Table 3, where target_layer denotes K and target_label denotes the size of selected target sets. As shown in Table 3, supervised layer set $\{6, 12\}$ paired with target set $\{300, 500\}$ achieves the best performance for two target layers. What's more, both the two sets: $\{4, 12\}$,$\{6, 12\}$ outperform the set $\{8, 12\}$, which indicates that small-size target sets are more suitable for lower layers.

In addition, more supervised layers such as $\{3, 6, 12\}$ is introduced to prove the motivation. Compared with the super-

TABLE III
RESULTS OF THE MODELS WITH DIFFERENT SUPERVISED LAYER.

Target_layer	Target_label	Test_other
4,12	300,500	8.76
6,12	300,500	**8.55**
8,12	300,500	9.04
3, 6, 12	100, 300, 500	8.71
3, 6, 12	50, 300, 500	**8.51**

TABLE IV
RESULTS OF THE MODELS WITH DIFFERENT ATTENTION SCALE.

layer	Window_size	Test_other
base		8.55
1-12	80	8.38
1-6,7-12	80,160	8.27

vised layer set $\{3, 6, 12\}$ paired with target set $\{100, 300, 500\}$, target set $\{50, 300, 500\}$ achieved better performance because learning difficulty has such a high correlation with different target granularity. Meanwhile, supervised layer set $\{3, 6, 12\}$ paired with target set $\{50, 300, 500\}$ can also guarantee better performance than the supervised layer set$\{6, 12\}$.

Furthermore, we build some ablation models with different attention window size w to explore the effect of different attention scales. As mentioned in section 3.2, the receptive field of different layers has a correlation with the progressive target numbers. The attention areas of only two heads are restricted. Based on the model with supervised layer set $\{6, 12\}$, we restrict the attention area of each layer into variable scopes. The results of different attention scale are listed in Table 4. The setting at second line restricts the attention area of each layer into an 80-timestep context window, which achieves 2.0% WER reduction on test_other. The setting at third line enables the attentions of layer 1-6 work on an 80-timestep scope and enables the attentions of layers 7-12 work on a 160-timestep scope, which obtains 3.4% WER reduction.

V. Conclusion

In this paper, we proposed a novel approach to self-supervised learning for speech recognition, called Progressive Multi-Scale Self-Supervised Learning (PMS-SSL). Our method introduces a progressive learning strategy by assigning different granularity target sets for different layers, with the target set size increasing progressively from the bottom to the top supervised layer. In addition, we incorporated a multi-scale structure into the multi-head self-attention mechanism to restrict the attention area into a small scope at lower layers and a larger scope at higher layers. Experimental results demonstrate that PMS-SSL outperforms the base HuBERT model, indicating the effectiveness of our proposed approach.

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, PMLR, 2016, pp. 173–182.

[2] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[3] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7694–7698.

[4] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[8] S. Pascual, M. Ravanelli, J. Serra, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[9] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[10] C. Wang, Y. Wu, S. Chen, *et al.*, "Self-supervised learning for speech recognition with intermediate layer supervision," *arXiv preprint arXiv:2112.08778*, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *arXiv preprint arXiv:2005.09267*, 2020.

[13] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "Slimipl: Language-model-free iterative pseudo-labeling," *arXiv preprint arXiv:2010.11524*, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[16] J. Chung, S. Ahn, and Y. Bengio, "Hierarchical multiscale recurrent neural networks," *arXiv preprint arXiv:1609.01704*, 2016.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[19] A. H. QXJCJKGSVLRC Vineel Pratap, "Wav2letter++: The fastest open-source speech recognition system," *CoRR, vol. abs/1812.07625*, 2018.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Improved Data2vec with Soft Supervised Hidden Unit for Mandarin Speech Recognition

Genshun Wan[†], Hang Chen[†*], Pengcheng Li[‡], Jia Pan[‡], Zhongfu Ye[†],
[†] University of Science and Technology of China, China
E-mail: gswan@mail.ustc.edu.cn,ch199703@mail.ustc.edu.cn,yezf@ustc.edu.cn
[‡] iFLYTEK Research, iFLYTEK Co. Ltd., China
E-mail: pcli2@iflytek.com,jiapan@iflytek.com

***Abstract*—Speech pre-training method has shown great success in learning useful and general latent representations from large-scale unlabeled data. In order to further improve the performance of self-supervised learning method for specific downstream tasks, an improved approach based on data2vec framework with soft supervised hidden unit is proposed. To take full advantage of the labeled data from downstream task, a supervised model is firstly trained to extract supervised hidden unit. And then based on data2vec, an extra Bert-like prediction task with soft cluster distance is introduced to match the downstream task and avoid unnecessary information loss. The proposed method can form a virtuous circulating utilization pattern for the downstream labeld data. Experiments on the small open-source Mandarin speech corpus AISHELL-2 and large private-source Mandarin speech corpus TRANS-L tasks show that our method can achieve relative character error rate reductions of 13.2% and 5.2% respectively when pre-trained on AISHELL-2 and TRANS-L corpus compared with data2vec framework.**

I. Introduction

With the rapid development of deep learning, intelligent speech technology has been applied in a wide range of fields. However, automatic speech recognition (ASR) system is still significantly dependent on labeled data, which may limit its application and popularization in low resource scenarios.

In order to better use the unlabeled speech data, self-training and self-supervised learning (SSL) methods have been adopted to improve the recognition performance. Self-training method, which applies the pseudo-labels on unsupervised speech data to augment training effect, may depend a great deal on the quality of pseudo-labels [1]–[5]. Self-supervised learning for speech data aims to learn better representations by predicting targets that are derived from the input speech data itself [6]–[10]. These typical methods include prediction of the future frames on past frames or reconstruction of the masked input frames, and so on.

As one of the representative self-supervised learning method for speech filed, HuBERT carries out the clustering operation to obtain the target labels for a BERT-like prediction loss [8]. Compared with HuBERT, data2vec provides a more general paradigm by predicting the contextualized latent representations of the full input data [10]. Unlike the limitation on the number of target units for HuBERT, there are little or no constraints for data2vec. However, similar to other self-supervised learning method, HuBERT and data2vec tend to generate a more universal speech representation which may be suitable for multiple downstream tasks. Obviously, using ensemble techniques to combine multiple variations of self-supervised learning method can achieve even higher performance [11].

So as to better grasp the subdivision scenarios and improve the interpretability of the hidden layer representation, specific downstream tasks can be closely bound up with the self-supervised learning method. In [12], an end-to-end joint unsupervised and supervised training method is proposed to select a pre-trained checkpoint which is better suited to the downstream ASR task. In [13], embeddings trained both under supervised and unsupervised settings are explored to improve robustness to accents. However, the existing methods do not fully exploit the advantages of downstream tasks, especially the labeled data which contain abundant information. In terms of ASR task, we argue that the performance could be further improved if the self-supervised network focus on learning content information from downstream implementation tasks. Meanwhile, multiple iteration pattern such as HuBERT may affect the generalization performance. Obviously, better initial target labels from supervised model can accelerate convergence and improve the representation while most methods do not fully exploit these advantages in low resource application tasks.

Motivated by the above, an improved self-supervised multi-task learning method which is more suited for the downstream task is proposed. Take ASR task for example, the labeled data are not only used to optimize the final model during the fine-tuning stage in the usual manner, but also to guide the generation of target labels for pre-training stage. The two training stages including pre-training and fine-tuning can form a virtuous cycle. Firstly, a supervised ASR model is trained with a certain amount of labeled speech data from downstream task. Then all unlabeled audio are fed into the supervised model to extract the latent features, and a quantization procedure similar to HuBERT is applied on these feature to get the target labels for self-supervised stage. Based on the data2vec self-supervised framework, an extra BERT-like prediction task is introduced to match the downstream task. In order to avoid unnecessary information loss, the BERT-like prediction task use soft cluster distance. To evaluate the effectiveness of the proposed method, we conducted experiments on both the small open-source Mandarin speech corpus AISHELL-2 and large private-source Mandarin speech corpus TRANS-L.

*Corresponding author

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

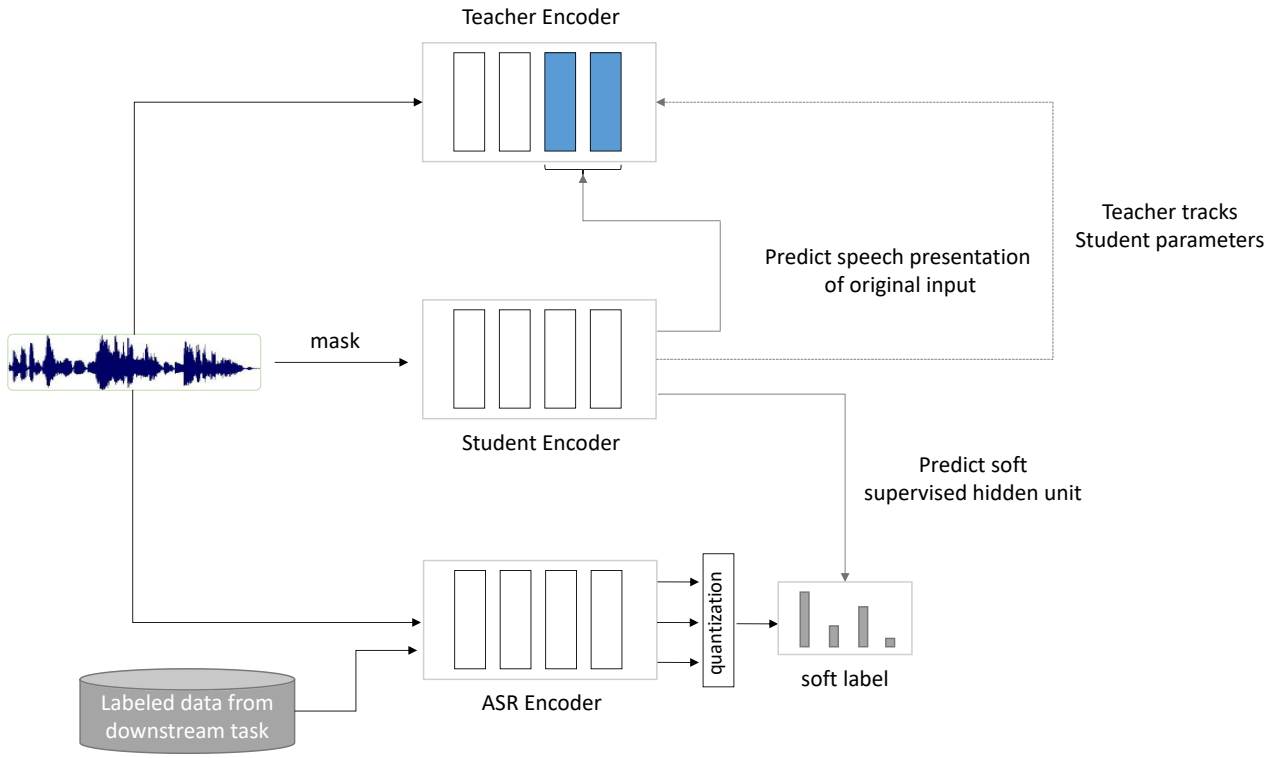


Fig. 1. The framework of proposed data2vec with soft supervised hidden units.

Compared with the baseline data2vec, our approach fintuned on AISHELL-2 achieved relative CER reductions of 13.2% and 5.2% respectively when pre-trained on AISHELL-2 and TRANS-L corpus.

II. RELATED WORK

As the mainstream theory, self-supervised pre-training paradigm has develop various types of method. Both HuBERT and data2vec have been proved to be effective for downstream task. In HuBERT, an offline clustering step is utilized to generate aligned target labels for a BERT-like pre-training [8], [10], [14]. Different from wav2vec 2.0, HuBERT benefits from the offline clustering step due to the consistency of the targets [7], not just their correctness. What is more, the prediction loss is only applied over the masked regions, forcing the model to learn a combined acoustic and language model over the continuous inputs. HuBERT pre-training adopts an iterative re-clustering and re-training process. As a more general pre-training paradigm, data2cvec predicts latent representations of the full input data based on a masked view of the input in a self-distillation setup using a standard Transformer architecture.

Multi-task learning in self-supervised pre-training method aims to introduce some extra task to improve the representation of unlabeled data or better match downstream tasks. Some research work has focused on applying multi self-supervised method [15], [16] while others has combined self-supervised learning and supervised learning to improve downstream task performance [17].

III. PROPOSED METHOD

In order to take advantage of different self-supervised pre-training method and combine more guidance information from downstream task, an improved self-supervised learning method based on data2vec with soft supervised hidden unit is proposed. The architecture of the proposed method is illustrated as figure 1. A supervised ASR model with a certain amount of labeled speech data from the downstream tasks is firstly trained to extracted latent feature. And then simple discrete latent variable models such as k-means are used to generate target units with the deep representations. Finally the data2vec framework is trained alongside with the soft supervised instruction loss which is more suitable for the downstream task. The two training stages including pre-training and fine-tuning forms a mutually reinforcing mechanism.

A. Extraction of supervised latent feature

For most self-supervised paradigm, the labeled information from downstream task is insufficiently exploited. In order to combine some guidance information to Data2vec framework, an offline clustering step to provide aligned target labels for a BERT-like prediction loss is introduced into the framework.

To obtain better supervised representation, the first step of the proposed method is using the labeled data from the specific downstream task to train a supervised instructive

model. Compared with standard HuBERT pre-training process, the discrete targets are assigned by clustering on the supervised model instead of MFCC features to strengthen the content presentation and reduce the training cost.

To simplify the training procedure, the model architecture used in supervised training is same as data2vec pre-training. A classification layer and a CTC loss is added to the baseline architecture, and the model is trained until converges. After supervised training, the supervised latent feature can be extract from last transformer layer, denoted as $H = [h_1, h_2, ..., h_t]$.

B. Quantization of supervised feature

After supervised training, the extracted latent feature is expected to instruct the audio pre-training. Motivated by HuBERT, the latent feature is quantized to build the supervised hidden unit. Feature quantization can improve the consistency of training target, where all frames are clustered into limited classes. In addition, by using different codebook size, the granularity of target labels can be more easily controlled.

A mini batch K-Means algorithm implemented in scikit-learn is used to trained the k-means model [18]. The cluster centers C can be denoted as $C = [c_1, c_2, ..., c_i]$, where c_i represents i-th cluster center. The assigned center of each frame represents supervised hidden unit, which can be used as training target like HuBERT. The training object can be wrote as follow:

$$\mathcal{L}_{Sup} = -\sum_{i} logP_f(i) \tag{1}$$

where P_f is the parameterized distribution over codebooks:

$$P_f(i) = \frac{exp(sim(Ao_t, e_i)/\tau)}{\sum_j exp(sim(Ao_t, e_j)/\tau)} \tag{2}$$

where $O = [o_1, o_2, ..., o_t]$ is the masked version of feature, A is the project matrix, e_i is the embedding of cluster center i, and τ is a logit scalar set to 0.1.

C. Soft supervised instruction loss

However, the progress of the quantization step will also cause information loss, because different frames may have different distances to codebooks even though they are assigned to the same one. On the other hand, the distance itself also contains the information of acoustic property. To tackle the information loss caused by quantization, we define $P_t(i)$ as the negative softmax of t-th frame to measure the distances to center i:

$$P_t(i) = \frac{exp(-||h_t - c_i||^2)}{\sum_j exp(-||h_t - c_j||^2)} \tag{3}$$

On this basis, the training objective can be rewrote by intorducing the distances to codebooks as:

$$\mathcal{L}_{Sup} = -\sum_{i} logP_f(i)P_t(i) \tag{4}$$

D. Model pre-training and fine-tuning

The pre-training and fine-tuning details follow the scheme of data2vec. The masked version of waveform can be fed into the main framework, followed by a feature extractor composed by convolution layers, and transformer layers are used on extracting high-level feature. The teacher network is fed with unmasked waveform, and updated by exponential mean average. The training target is about mean square error, which can be computed between student and teacher:

$$\mathcal{L}_{MSE} = (y_t - f_t(x))^2 \tag{5}$$

where y_t is the output of teacher network, $f(x)$ is the prediction of student network.

The total training objective can be defined as follow:

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda\mathcal{L}_{Sup} \tag{6}$$

where λ is used as a hyper-parameter to balance the two tasks.

During pre-training, the model is joint optimized by $\mathcal{L}_{MSE}$ and $\mathcal{L}_{Sup}$. While in fine-tuning procedure, a linear layer and a CTC loss is added on the model, and the model is trained with labeled data for specific downstream task.

IV. Experimental Evaluation

A. Data

We evaluated the performance of our proposed approach on both the small open-source speech corpus AISHELL-2 and the sizeable private-source speech corpus TRANS-L; the details of the two speech recognition corpus are illustrated as follows.

AISHELL-2 dataset: The AISHELL-2 task is a benchmark for large-Mandarin-vocabulary continuous speech recognition, so we choose it as the main task. As an elaborately prepared public datase, AISHELL-2 contains 1000 hours speech data. The dev-other subset is used to select the final model. For pre-training stage, all 1000 hours speech data are used to conduct self-supervised learning. In addition, 100 hours and 1000 hours speech data are used as different experimental configuration both on first supervised training stage and downstream fine-tuning stage respectively.

TRANS-L dataset: TRANS-M is a private-source Mandarin dataset with 50K hours speech data which combines multiple data styles, including lessons, conferences, interviews, and television programs. Unlike AISHELL-2, TRANS-L turns to a more colloquial and dialogic linguistic style. We use TRANS-L dataset during the pre-training stage, while all AISHELL-2 are used on first supervised training stage and downstream fine-tuning stage.

B. Model configuration

For AISHELL-2 dataset, the model configuration is same as data2vec. The architecture of the model is composed of 12-layer transformer blocks with 1024 hidden dimension. For TRANS-L dataset, the feature extractor is removed and the waveform input is replaced by 40-dimensional filterbank. These modifications are mainly used to improve the training efficiency.

TABLE I
CER (%) RESULTS OF DIFFERENT AMOUNT OF TRAINING DATA ON AISHELL-2.

Model	data for first supervised training(h)	data for pre-traing (h)	data for downstream task(h)	dev	test
data2vec	-	1000	100	9.6	9.8
our method	100	1000	100	9.1	9.6
data2vec	-	1000	1000	7.2	7.6
our method	100	1000	1000	6.6	7.1
our method	1000	1000	1000	6.2	6.6

C. training configuration

The AISHELL-2 audios are fed into the main framework with 16KHZ waveform. For AISHELL-2 pre-training stage, we decrease the updates to 20K while keep other configurations the same as wav2vec 2.0. For AISHELL-2 fine-tuning stage, the updates of 100 hours is increased to 120K, while the configuration of 1000 hours keeps the same as standard configurations. The configurations of supervised training is similar as fine-tuning stage. The only difference is that the updates is increased into 350K and 500K for 100 hours and 1000 hours labeled data respectively. The standard data2vec result was reproduced with fairseq tool [19]. For TRANS-L dataset, the training configurations are same as AISHELL-2.

D. Results on AISHELL-2 dataset

The results with different amount of training data are listed in Table I. The training data were used both in the first supervised training stage and downstream fintuning stage. Compared with standard data2vec, the proposed method can get significant recognition performance improvement on different amount of training data. Especially as the amount of training data increases into 1000 hours, the CER of development set dropped from 7.2% to 6.2% while test set dropped from 7.6% to 6.6%, which indicated that the proposed method achieve a considerable improvement of over 13%. Obviously, the quality of the first supervised model affects the effectiveness of the target units used for pre-training stage. That is to say, with a well-trained supervised model suited for downstream task, the pretraining stage can achieve better performance.

E. Results on TRANS-L dataset

The results on TRANS-L dataset are shown in table II. After replacing the feature extractor and training with more unlabeled data, the CER on AISHELL-2 test set is also significantly improved because of the better representation learning. The 5% relative increase demonstrates that our proposed method is still effective on large-scale industrial audio dataset. Unintentionally, the increase in data size and the mismatch between pre-training and downstream task data to some extent reduced the improvement of our method.

F. Ablation studies

To evaluate the effectiveness of the proposed soft supervised instruction label, we replace the soft label with hard label. The experiment is conducted on AISHELL-2 while 1000 hours data can be used for first supervised training stage, pre-training stage and downstream fine-tuning stage.

TABLE II
CER (%) RESULTS OF DIFFERENT PRE-TRAINING METHOD WHICH FINTUNED ON ALL AISHELL2 DATA.

Model	data for pre-traing(h)	test
data2vec	50k	5.8
our method	50k	5.5

TABLE III
CER (%) RESULTS OF DIFFERENT LABELS USED IN OUR METHOD.

Model	dev	test
Hard lables	6.4	6.9
Soft lables	6.2	6.6

The results in Table III shows that the CER rises from 6.2% to 6.4% on AISHELL-2 development set, and rises from 6.6% to 6.9% on AISHELL-2 test set. The results further proves the involving of distance information leads to better pre-training effectiveness. To further illustrate the information loss of hard label, we choose some frames from latent feature of supervised model and depict their $P_t(i)$ in figure 2. Some frames are in unimodal distribution while the other are in multimodal distribution. For the multimodal distribution, applying hard target will result in information loss.

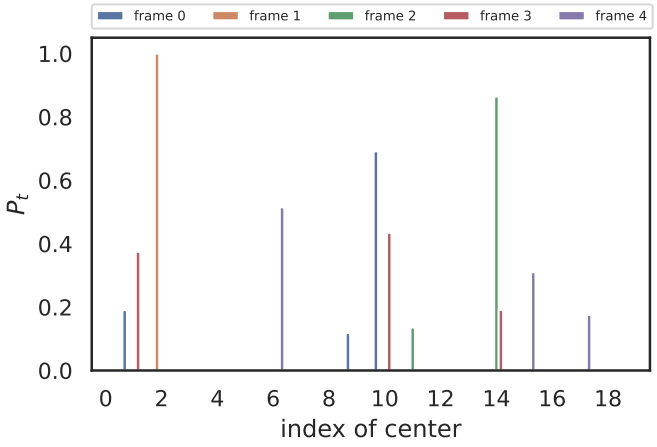


Fig. 2. P_f of extracted supervised latent features, and the frames are distinguished by color.

V. Conclusion

In this work, an improved self-supervised learning method based on data2vec with soft supervised hidden unit is proposed, which is more suited for downstream task. The experiments on both AISHELL-2 and TRANS-L achieve better performance compared with the baseline data2vec. We believe that this approach has great potential for improving the performance of downstream task which can form a virtous circulatin utilization pattern for the labeled data.

References

[1] W. Han, Z. Zhang, Y. Zhang, *et al.*, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv preprint arXiv:2005.03191*, 2020.

[2] S. H. K. Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6670–6674.

[3] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7084–7088.

[4] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *arXiv preprint arXiv:2005.09267*, 2020.

[5] D. S. Park, Y. Zhang, Y. Jia, *et al.*, "Improved noisy student training for automatic speech recognition," *arXiv preprint arXiv:2005.09629*, 2020.

[6] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[10] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[11] T. Hallmen, S. Mertes, D. Schiller, and E. André, *An efficient multitask learning architecture for affective vocal burst analysis*, 2022. arXiv: 2209.13914 [cs.SD].

[12] J. Bai, B. Li, Y. Zhang, *et al.*, *Joint unsupervised and supervised training for multilingual asr*, 2021.

[13] J. Li, V. Manohar, P. Chitkara, *et al.*, *Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings*, 2021.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] A. Bonafonte, J Serrà, M. Ravanelli, S. Pascual, and Y. Bengio, *Learning problem-agnostic speech representations from multiple self-supervised tasks*, 2019.

[16] M. Ravanelli, J. Zhong, S. Pascual, *et al.*, *Multi-task self-supervised learning for robust speech recognition*, 2020.

[17] C. Wang, Y. Wu, Y. Qian, *et al.*, *Unispeech: Unified speech representation learning with labeled and unlabeled data*, 2021.

[18] A. Swami and R. Jain, *Scikit-learn: Machine learning in python*, 2013.

[19] M. Ott, S. Edunov, A. Baevski, *et al.*, *Fairseq: A fast, extensible toolkit for sequence modeling*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Investigation of Ensemble of Self-Supervised Models for Speech Emotion Recognition

Yanfeng Wu†, Pengcheng Yue†, Cuiping Cheng and Taihao Li*
Zhejiang Lab, Hangzhou, China
E-mail: {wuyanfeng, pcyue, lith}@zhejianglab.com, chengcuipingswu@163.com
* Corresponding Author
† Equal Contribution

***Abstract*—Traditional deep learning based speech emotion recognition (SER) methods are restricted to the lack of emotional speeches. Recently, the self-supervised learning (SSL) speech representation model is emerging and has become the state-of-the-art method for SER because of its ability to learn a more generalized emotion representation. Some SSL models adopt a similar network architecture, but utilize various training strategies or pre-trained speech data to learn speech representations from different perspectives, thus containing complementary information for the SER task. In this paper, we conduct an investigation of ensemble learning with six SSL models for SER using three kinds of ensemble methods: decision ensemble, late feature ensemble, and early feature ensemble. The six SSL models adopted in this paper are Wav2vec 2.0 Base, HuBERT Base, WavLM Base, WavLM Base+, UniSpeech-SAT Base, and UniSpeech-SAT Base+. Our experimental results on three datasets show that: 1) all three kinds of ensemble learning methods can significantly improve the SER performance, demonstrating that there may exist a large amount of complementary information between different SSL models; 2) in terms of the decision ensemble methods, the evidence theory (ET) based fusion performs better than the average-based fusion because of its ability to deal with the uncertainty of different models; 3) the ET-based decision ensemble is the best in comparative ensemble methods because of its superior performance and lower computational cost.**

I. Introduction

Speech emotion recognition (SER) is the task of identifying the emotional state of the speaker from speech, which plays a critical role in the field of human-computer interaction. Since it may be difficult for humans to perceive others' emotions in some scenarios, understanding humans' emotions is a more challenging task for computers. Considerable efforts have been devoted to improving the performance of SER in the research community. Traditional machine learning (ML) approaches rely on handcrafted features (e.g. OpenSmile [1]) or feature engineering techniques, resulting in low generalization on various datasets. With the rapid development of deep learning (DL), the DL-based SER approaches [2]–[4] become popular and achieve superior performance than ML-based methods. However, the improvement of DL-based algorithms is restricted to the small size of the SER dataset. Compared with the large-scaled datasets on other speech-related tasks (e.g., LibriSpeech [5] for speech recognition and VoxCeleb [6], [7] for speaker recognition), the commonly used SER datasets (e.g. IEMOCAP [8], Emo-DB [9] and RAVDESS [10]) are not enough for DL models to learn a generalized representation.

Recently, several self-supervised learning (SSL) speech representation models [11]–[17] are emerging and perform well on various speech-related downstream tasks. Those SSL models are pre-trained to learn a generalized and useful representation with a large number of unlabeled speeches. Fine-tuning the SSL model becomes the state-of-the-art SER method and shows higher accuracy on many SER datasets [18]–[20]. Ref. [21] evaluates nineteen SSL models on five SER datasets and we observe that four kinds of SSL models, i.e., Wav2vec 2.0 [12], HuBERT [13], WavLM [14] and UniSpeech-SAT [15], perform significantly better than other models. The four SSL works adopt a similar network architecture but utilize different training strategies and objective functions to learn speech representations from different perspectives. In addition, these models use various kinds of speech data during the pre-training and learn various knowledge. Based on above observations, we argue that these SSL models contain complementary information for SER tasks, which inspires us to exploit ensemble learning methods of SSL models.

In this paper, we conduct an investigation about ensemble learning of SSL models for SER using three kinds of ensemble methods: decision ensemble, late feature ensemble, and early feature ensemble. The decision ensemble method fuses the probabilities produced from each fine-tuned SSL model via the average-based fusion or the evidence theory (ET) based fusion. The late feature ensemble approach trains a simple classification model with two fully-connected (FC) layers to combine the features from fine-tuned SSL models. The early feature ensemble method trains a well-performed backend model called ECAPA-TDNN [22] based on several kinds of pre-trained SSL features. Six SSL models are utilized for the ensemble including Wav2vec 2.0 Base, HuBERT Base, WavLM Base, WavLM Base+, UniSpeech-SAT Base, and UniSpeech-SAT Base+.

Our experiments are conducted on three commonly used SER datasets, IEMOCAP [8], RAVDESS [10] and MSP-IMPROV [23]. The experimental results show that: 1) all three kinds of ensemble learning methods can significantly improve the SER performance, demonstrating that there may exist a large amount of complementary information between different SSL models; 2) in terms of the decision ensemble methods, the ET-based fusion performs better than the average-based fusion because of its ability to deal with the uncertainty of

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I
A SUMMARY OF THE SIX SSL MODELS USED IN THIS PAPER.

Model	Size	Pre-trained dataset	Key Techniques
Wav2vec 2.0	Base	960 hrs LibriSpeech	Masking latent representations; solving a contrastive task over quantized speech representations.
HuBERT	Base	960 hrs LibriSpeech	Generating aligned target labels with a K-Means cluster approach; applying the prediction loss over masked regions.
WavLM	Base	960 hrs LibriSpeech	Employing Gated relative position bias; Jointly Learning masked speech prediction and denoising.
	Base+	60k hrs Libri-Light & 10k hrs GigaSpeech & 24k hrs VoxPopuli	
UniSpeech-SAT	Base	960hrs LibriSpeech	Integrating utterance-wise contrastive loss; presenting an utterance mixing strategy for data augmentation.
	Base+	60k hrs Libri-Light & 10k hrs GigaSpeech & 24k hrs VoxPopuli	

different models; 3) the ET-based decision ensemble is the best in comparative ensemble methods because of its superior performance and lower computational cost. We hope that our observations can provide the research community with a new perspective to improve the effectiveness of SSL models on the SER task.

II. Related Work

A. Self-supervised speech representation learning

Self-supervised speech representation learning provides a single universal model which can benefit a wide variety of speech-related tasks [24]. With the help of the release of more and more large-scale speech databases, many SSL models have emerged in recent years. Among them, Wav2vec 2.0 [12], HuBERT [13], WavLM [14] and UniSpeech-SAT [15] are four widely used models. They adopt similar network architectures which consist of a CNN-based feature extractor module, a feature projection module, and a Transformer-based [25] feature encoder module. The main differences between them are the employed training approaches and the amount of training data. Table I gives a summary of the six SSL models used in this paper. The four kinds of SSL works utilize various training objectives and data augmentation techniques. In addition, WavLM and UniSpeech-SAT provide the Base+ model which extends the training data with the GigaSpeech [26] and the VoxPopuli [27] dataset. Thanks to the diversity of training objectives and pre-trained speech data, the six SSL models can generate speech representations from different aspects with the same model size. Hence, we select the six SSL models as the baseline model to evaluate all ensemble methods in this paper.

B. Fine-tuning SSL models for SER

Fine-tuning SSL models has become the state-of-the-art method for many downstream tasks. Ref. [28] applied the Wav2vec 2.0 framework to speaker recognition and achieved a comparable EER of 1.88% on the VoxCeleb dataset. Ref. [20] presented a simple framework, namely SUPERB, to benchmark the capability of SSL models on downstream speech processing tasks. Ref. [29] explored fine-tuning on Wav2vec 2.0 and HuBERT for three speech tasks.

In terms of the SER task, due to the insufficiency of large-scale well-labeled emotional data, it is hard for current DL architectures to learn a generalized and robust emotion representation. Since performing well with a small amount of fine-tuned data is a main advantage of SSL models, many researchers utilize fine-tuned SSL models or pre-trained SSL features for SER. Ref. [19] combined the different layers in Wav2vec 2.0 model and then extracted pre-trained features for SER. Ref. [18] developed a novel method for fine-tuning Wav2vec 2.0 and yielded better performance on the IEMOCAP dataset. Ref. [21] evaluated the SER performance of nineteen SSL models on five datasets. However, there is a main limitation in [21] that cross-validation is not performed. Ref. [30] conducted a comprehensive analysis of fine-tuning Wav2vec 2.0 and HuBERT for SER. The main difference between our work and prior works is that we mainly focus on the ensemble learning of the SSL models and conduct extensive experiments on six SSL models to demonstrate the effectiveness of the adopted methods.

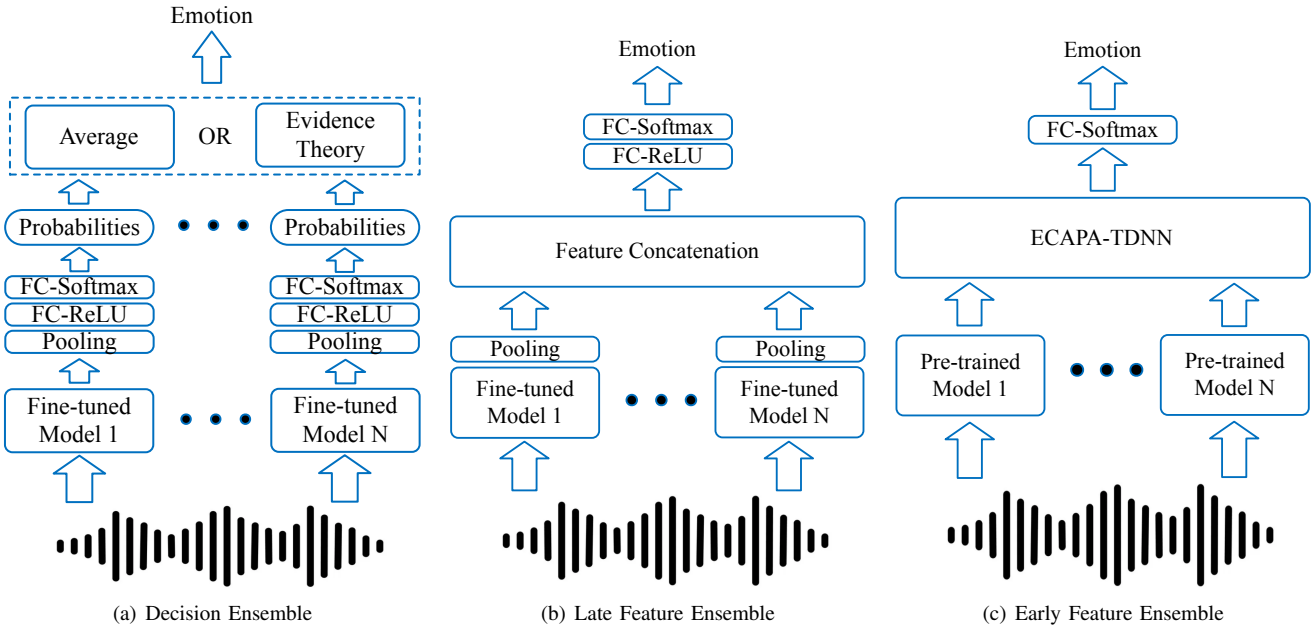
C. Ensemble learning

Ensemble learning aims to exploit multiple classifiers or models to yield better performance than any single one, which is widely applied in speech-related tasks. Ref. [31] proposed an ensemble of seven different models to predict the mean opinion score. Ref. [32] proposed to utilize an ensemble of many SSL representations and models for the speech recognition task. Ref. [33] proposed an ensemble of various pre-trained SSL speech models to train a perceptual loss for speech denoising. In this paper, we adopt three kinds of ensemble methods to combine six different SSL models for SER. To the best of our knowledge, our work is the first to apply an ensemble of SSL models to the SER task.

III. Approach

A. Fine-tuning SSL models

Fig. 1 shows the three kinds of used ensemble methods in this paper: decision ensemble, late feature ensemble, and early feature ensemble. Both the decision ensemble method and late feature ensemble method need fine-tuned SSL models. Hence, the first step is fine-tuning six SSL models in Table I with emotion labels, respectively. The detailed architecture is shown in 1(a). The output of the last layer in the feature encoder module is extracted. Then the average pooling operation is



(a) Decision Ensemble (b) Late Feature Ensemble (c) Early Feature Ensemble

Fig. 1. Three kinds of ensemble methods of the SSL models. The "Fine-tuned Model" represents the model fine-tuned with emotion labels, and the "Pre-trained Model" means it is the pre-trained version without fine-tuning.

performed to acquire a 768-dimensional embedding. Finally, the embedding is fed into two FC layers to generate the predicted emotions. We exploit the weighted cross-entropy loss to train all models. In the fine-tuning step, only the feature encoder module is trained, while the parameters of the feature extractor module and the feature projection module are frozen. A modified version of SpecAugment [34] is exploited for SSL models to obtain better generalization.

B. Decision ensemble

The decision ensemble method fuses the probabilities from each fine-tuned SSL model, as shown in Fig. 1(a). Since the average and the evidence theory are both sample-level fusion methods, we take a sample as an example. Assuming that the probability produced by the SSL model n for this sample is

$$P_n = [p_1^n, p_2^n, ..., p_M^n], \forall n \in \{1, 2, ..., N\}, \tag{1}$$

where M is the number of the emotion class and N is the number of SSL models.

1) Average: The first approach is a simple average fusion, which averages the probabilities of each class:

$$P_{Avg} = \frac{\sum_{n=1}^{N} P_n}{N}. \tag{2}$$

2) Evidence theory: The evidence theory approach [35]–[38] is a reasoning theory that deals with uncertainty. The ensemble decision based on evidence theory is the fusion of the decision beliefs produced by several SSL models through Dempster's combination rule [35] and can be presented as:

$$P_{ET} = \underbrace{P_1 \oplus P_2 \oplus \cdots \oplus P_N}_{N-1 \text{ times}} \tag{3}$$

where $\oplus$ represents the Dempster's rule of combination. Taking the emotion recognition result produced by certain SSL model P_n as one piece of "evidence" in evidence theory, two pieces of evidence P_1 and P_2 can be combined with Dempster's fusion rule as follows:

$$P(A) = \frac{\sum_{B \cap C = A} P_1(B) P_2(C)}{1 - K} \tag{4}$$

where

$$K = \sum_{B \cap C = \emptyset} P_1(B) P_2(C) \tag{5}$$

where A, B and C represent the emotion classes of the SSL models.

C. Late feature ensemble

The late feature ensemble method takes all features extracted from fine-tuned SSL models as input and trains another network including two FC layers, as shown in Fig. 1(b). The emotion embeddings from all fine-tuned SSL models are combined by a feature concatenation operation. Note that all parameters of fine-tuned models are frozen and only the parameters of top FC layers are optimized in this training step. This method is inspired by the findings in [39], which shows that the two-step training can greatly improve SER performance. On the other hand, it is difficult to jointly fine-tune the parameters of several SSL models because of their large model sizes. The late feature ensemble method can be regarded as an alternative method.

D. Early feature ensemble

The early feature ensemble method utilizes the pre-trained speech representations from multiple SSL models and trains a small network as an emotion classifier. Note that the SSL

TABLE II
THE LABEL DISTRIBUTION OF THE DATASETS USED IN THIS PAPER.

Dataset	Angry	Happy	Neutral	Sad	Calm	Surprised	Fear	Disgust	Total
IEMOCAP	1,103	1,636	1,708	1,084	-	-	-	-	5,531
RAVDESS	192	192	96	192	192	192	192	192	1,440
MSP-IMPROV	792	2,644	3,477	885	-	-	-	-	7,798

models in this approach are not fine-tuned with emotion labels. In this paper, we choose the ECAPA-TDNN [22] as the backbone network. ECAPA-TDNN is the state-of-the-art network in the speaker recognition tasks and also performs well on SER [17]. The architecture of ECAPA-TDNN is composed of two one-dimensional convolutional layers, three SE-Res2Blocks, an attentive statistics pooling layer, and an FC layer. The input of ECAPA-TDNN in the original paper is 80-dimensional MFCCs. In our implementation, we concatenate the output of SSL models on feature dimension. The output of ECAPA-TDNN is a 192-dimensional vector, followed by an FC layer to predict the emotion class. Like the late feature ensemble approach, the parameters of pre-trained SSL models are also frozen during the training.

IV. EXPERIMENT

A. Datasets

TABLE III
THE HYPER-PARAMETERS USED IN THE EXPERIMENTS.

Method	Hyper-Parameter	IEMOCAP	RAVDESS	MSP-IMPROV
Fine-Tuning	Max Length	8s	5s	8s
	Training epochs	100[a]	100	100
	Batch size	48	48	48
	Optimizer	Adam	Adam	Adam
	Learning rate	1e-4	1e-4	1e-4
	Scheduler	linear warm-up, 0.1		
Late Feature Ensemble	Training epochs	10	10	10
	Batch size	64	64	64
	Optimizer	Adam	Adam	Adam
	Learning rate	1e-3	1e-3	1e-3
	Scheduler	polynomial decay		
Early Feature Ensemble	Training epochs	50	50	50
	Batch size	64	64	64
	Optimizer	AdamW	AdamW	AdamW
	Learning rate	1e-3	1e-3	1e-3
	Scheduler	polynomial decay		

[a] Early stopping with the patience of 50 epochs.

All experiments are conducted on three public datasets, IEMOCAP, RAVDESS, and MSP-IMPROV. The label distribution of the datasets used in this paper is listed in Table II.

a) IEMOCAP: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [8] contains approximately twelve hours of data labeled with nine different emotions. Each utterance includes three kinds of modalities: audio, video, and text. Following the conventional evaluation protocol, we only adopt four categories of utterances: *angry*, *happy*, *neutral* and *sad*, and merge the *excited* label into the *happy* label. As a result, there are a total of 5,531 utterances. We conduct five-fold leave-one-session-out cross-validation to evaluate all models.

b) RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database [10] is an audio-visual database consisting of emotional speech and song acted by 24 professional actors. We select the speech recordings partition of the RAVDESS, which contains 1,440 utterances labeled with eight emotions: *angry*, *calm*, *surprised*, *fear*, *happy*, *sad*, *neutral* and *disgust*. We perform six-fold leave-four-speakers-out cross-validation. The number of male and female speakers in the training set and test set is equal for a fair comparison.

c) MSP-IMPROV: The MSP-IMPROV [23] is a multimodal emotional database containing 8,438 improvised, read, and natural emotional utterances. We select a subset of 7,798 utterances with four emotion labels: *angry*, *happy*, *neutral*, and *sad*. The MSP-IMPROV dataset is divided into six sessions, so we perform six-fold leave-one-session-out cross-validation.

B. Implementation details

The experiments are conducted using six pre-trained SSL models on the Huggingface website: Wav2vec 2.0 Base[1], HuBERT Base[2], WavLM Base[3], WavLM Base+[4], UniSpeech-SAT Base[5], UniSpeech-SAT Base+[6]. Our implementation is based on an existing Pytorch framework [18]. Table III lists the settings of hyper-parameters used in the experiments. Considering that the average duration of utterances in the three datasets is quite different, the max sequence length is set to 8s for IEMOCAP and MSP-IMPROV, and 5s for RAVDESS. If the utterance is shorter than the max length, we pad it to the max length and exploit a mask to let the network ignore the padding part. We find that the learning rate and the scheduler play a critical role in the fine-tuning of SSL models. In our experiments, the learning rate is set to 1e-4 for all three datasets. We employ a linear decay scheduler after a warm-up period in the first 10 epochs. All experiments are carried out on an NVIDIA Tesla V100 32GB GPU.

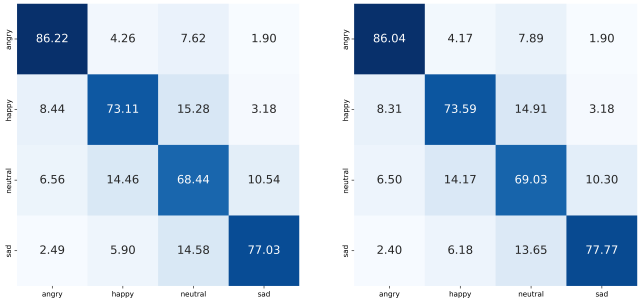
C. Experimental results

Table IV lists the SER results of the fine-tuned SSL models as well as four ensemble methods on three datasets. The unweighted accuracy (UA) and the weighted accuracy (WA) are reported as evaluation metrics.

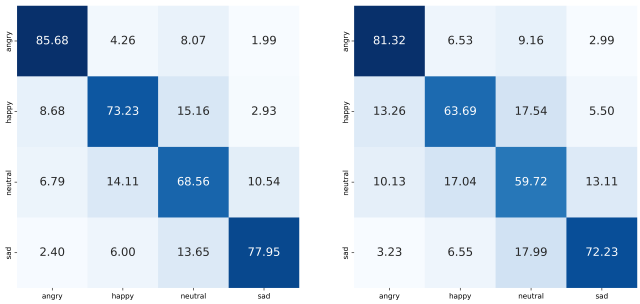
[1]https://huggingface.co/facebook/wav2vec2-base
[2]https://huggingface.co/facebook/hubert-base-ls960
[3]https://huggingface.co/microsoft/wavlm-base
[4]https://huggingface.co/microsoft/wavlm-base-plus
[5]https://huggingface.co/microsoft/unispeech-sat-base
[6]https://huggingface.co/microsoft/unispeech-sat-base-plus

TABLE IV
THE SER RESULTS OVER THREE EMOTIONAL SPEECH DATASETS.

Type	Method	Param. (Training)	IEMOCAP		RAVDESS		MSP-IMPROV	
			UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)
Single Model	Wav2vec 2.0 Base	94.5M (90.3M)	73.46	72.00	76.43	76.81	62.22	67.82
	HuBERT Base	94.5M (90.3M)	72.94	71.49	77.34	78.06	62.87	68.65
	WavLM Base	94.5M (90.3M)	**74.38**	72.73	**79.30**	**80.07**	63.63	69.28
	WavLM Base+	94.5M (90.3M)	73.60	**73.18**	77.99	78.75	**64.84**	**70.20**
	UniSpeech-SAT Base	94.5M (90.3M)	72.88	71.78	78.12	78.47	62.63	68.78
	UniSpeech-SAT Base+	94.5M (90.3M)	73.73	72.90	77.67	78.33	62.76	69.50
	Fixed HuBERT + ECAPA-TDNN	102M (8.0M)	68.66	67.95	72.14	71.74	58.13	63.28
Ensemble Method	Decision Ensemble (Average)	566M (0)	76.30	75.14	80.86	81.25	65.08	71.61
	Decision Ensemble (Evidence Theory)	566M (0)	**76.74**	**75.60**	**81.18**	**81.67**	**65.21**	**71.78**
	Late Feature Ensemble	567M (0.6M)	76.43	75.30	81.12	81.46	64.41	71.40
	Early Feature Ensemble	584M (17.8M)	69.25	67.73	73.31	72.85	60.73	64.93



(a) Average (b) Evidence Theory

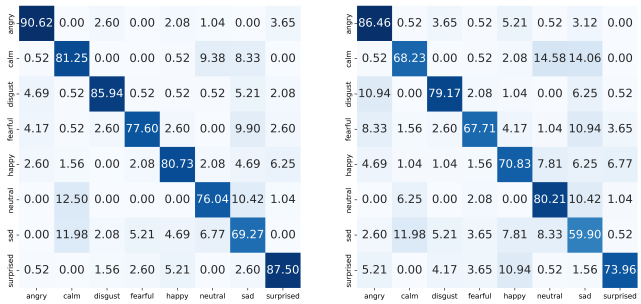


(c) Late Feature Ensemble (d) Early Feature Ensemble

Fig. 2. The confusion matrix of four evaluated ensemble methods for the IEMOCAP dataset.



(a) Average (b) Evidence Theory



(c) Late Feature Ensemble (d) Early Feature Ensemble

Fig. 3. The confusion matrix of four evaluated ensemble methods for the RAVDESS dataset.

We first compare the performance of the single SSL models. It can be observed that the best model is different in various datasets. For the IEMOCAP dataset, the WavLM Base obtains the best UA while the WavLM Base+ achieves the best WA. WavLM Base achieves the highest UA and WA on the RAVDESS dataset, and the WavLM Base+ yields the highest UA and WA on the MSP-IMPROV dataset. In terms of the four Base models, the WavLM Base performs best and others obtain similar performance, which is different from [21]. A possible reason is that we utilize the SpecAugment technique and the cross-validation which [21] does not use. In addition, compared with the Base models, the Base+ models of the WavLM and the UniSpeech-SAT perform better on the IEMOCAP and MSP-IMPROV, but perform worse on the RAVDESS.

We then evaluate the performance of four ensemble methods. It can be seen that the average, the evidence theory, and the late feature ensemble method can significantly improve the UA and WA on three datasets compared to the single SSL model, demonstrating that the ensemble of various SSL models is effective on the SER task. The early feature ensemble method is worse than the fine-tuned SSL model because it only employs six kinds of pre-trained SSL features without fine-tuning. However, it outperforms the "Fixed HuBERT + ECAPA-TDNN" model which only uses the pre-trained HuBERT features and the ECAPA-TDNN model, suggesting that the ensemble of pre-trained SSL features can also improve the SER performance. Overall, all four ensemble methods



(a) Average (b) Evidence Theory

(c) Late Feature Ensemble (d) Early Feature Ensemble

Fig. 4. The confusion matrix of four evaluated ensemble methods for the MSP-IMPROV dataset.

can notably improve the performance on three SER datasets compared to the single fine-tuned or pre-trained SSL model. This suggests that there may exist a large of complementary information between different SSL models and combining multiple SSL models or SSL features is an effective method to improve the performance of SER.

In terms of the decision ensemble method, it can be observed from Table IV that the ET-based fusion method performs better than the average-based fusion method on all three datasets. The reason may be that the evidence theory has the ability to deal with the uncertainty of different SSL models, thus can make full use of multi-source information to improve the reliability and accuracy of the decision ensemble. The experimental result suggests that the ET-based fusion is a better method for the decision ensemble in the SER task.

In addition, it can be seen from Table IV that the ET-based decision ensemble method obtains higher UA and WA on all three datasets, compared with the late feature ensemble method. The average-based decision ensemble method also yields competitive performance. A possible reason is that the late feature ensemble method only integrates the output feature of the last layer in the SSL models. The late feature ensemble and the decision ensemble are both in top level of the model, and thus yield similar performance. However, the late feature ensemble method needs additional training with much computational cost, while the decision ensemble methods only need to fuse the probabilities produced from multiple SSL models. Hence, it can be concluded that the ET-based decision ensemble method is the best in four comparative ensemble methods to ensemble various SSL models for SER because of its superior performance and lower computational cost.

TABLE V
COMPARISON WITH PRIOR WORKS ON IEMOCAP.

Method	Test Set	Accuracy(%)[a]
WavLM Large [21]	The fifth session	72.60
P-TAPT [18]	Five-fold CV	74.3
Co-attention [40]	Ten-fold CV	72.70
SMW_CAT [41]	Ten-fold CV	74.25
WavLM Base (Ours)	Five-fold CV	**74.38**
ET-Based Decision Ensemble (Ours)	Five-fold CV	**76.74**

[a] We only report the highest of UA or WA as the accuracy in all comparative works.

To verify the performance of ensemble methods on each emotion class, we report their prediction confusion matrix for each dataset, which is shown in Fig. 2, Fig. 3 and Fig. 4, respectively. It can be seen from Fig. 2 that the *angry* emotion is the best-performing class. The *neutral* class and *happy* class are hard to distinguish in the IEMOCAP. In the RAVDESS dataset, the best-performing class is also *angry*, while the worst-performing class is *sad*. The two classes that are hardest to distinguish for models are *neutral* and *calm*. In the MSP-IMPROV, the *sad* class is the worst performing class, while the *happy* class and *neutral* class are easily recognized. The reason is that the MSP-IMPROV is an unbalanced dataset where the samples of the *happy* and *neutral* are obviously more, making it difficult for models to recognize other emotions.

To verify the effectiveness of our work, we further compare our systems with recent state-of-the-art works. Since these works usually take the IEMOCAP as the dataset, we list the experimental results of different works on the IEMOCAP, as shown in Table V. Our implemented WavLM Base model achieves the highest accuracy of 74.38% in all comparative works, and the ET-based decision ensemble method using six SSL models further improves the accuracy to 76.74%. The results indicate the effectiveness of our introduced ensemble methods.

V. CONCLUSIONS

In this paper, we conduct an investigation of ensemble learning of SSL models for SER. We utilize three kinds of ensemble methods (i.e., decision ensemble, late feature ensemble, and early feature ensemble) to combine six SSL models (i.e., Wav2vec 2.0 Base, HuBERT Base, WavLM Base, WavLM Base+, UniSpeech-SAT Base, and UniSpeech-SAT Base+). The experiments are conducted on three datasets, IEMOCAP, RAVDESS and MSP-IMPROV. The results demonstrate the effectiveness of the ensemble method of multiple SSL models on the SER task. We also observe that the ET-based decision ensemble method performs best in all comparative ensemble methods. In the future, we plan to study the multi-view SER based on multiple pre-trained SSL features.

ACKNOWLEDGMENT

The work is supported by Grant No. 2020KB0AC01 from the Scientific Project of Zhejiang Lab, and the National Sci-

ence and Technology Major Project of the Ministry of Science and Technology of China (No. 2021ZD0114303).

REFERENCES

[1] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462.

[2] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[3] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[4] Mustaqeem and S. Kwon, "Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, p. 114177, 2021.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[7] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec 2008.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 1517–1520.

[10] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.

[11] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[15] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6152–6156.

[16] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[17] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6922–6926.

[18] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[19] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[20] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[21] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[23] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[24] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[26] G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 3670–3674.

[27] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," in *ACL 2021 - 59th Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, Aug. 2021.

[28] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7967–7971.

[29] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," 2022.

[30] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[31] M. Kunešová, J. Matoušek, J. Lehečka, J. Švec, J. Michálek, D. Tihelka, M. Bulín, Z. Hanzlíček, and M. Řezáčková, "Ensemble of deep neural network models for mos prediction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[32] A. Arunkumar, V. Nileshkumar Sukhadia, and S. Umesh, "Investigation of Ensemble features of Self-Supervised Pretrained Models for Automatic Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 5145–5149.

[33] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7118–7122.

[34] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[35] D. AP, "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[36] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[37] D. Yong, S. WenKang, Z. ZhenFu, and L. Qi, "Combining belief functions based on distance of evidence," *Decision Support Systems*, vol. 38, no. 3, pp. 489–493, 2004.

[38] C. Cheng, X. Zhang, and T. Li, "Distance-based dynamic weight: A novel framework for multi-source information fusion," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 378–385.

[39] X. Xu, D. Li, Y. Zhou, and Z. Wang, "Multi-type features separating fusion learning for speech emotion recognition," *Applied Soft Computing*, vol. 130, p. 109648, 2022.

[40] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.

[41] Y. He, N. Minematsu, and D. Saito, "Multiple acoustic features speech emotion recognition using cross-attention transformer," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

Single Source Zone Detection in the Spherical Harmonic Domain for Multisource Localization

Liang Tao *, Maoshen Jia †, Bing Bu ♦, Dingding Yao ⍲
* † ♦ Faculty of Information Technology, Beijing University of Technology, Beijing China
* E-mail: liangtaobjut@163.com Tel: +86-10-67391642
† E-mail: jiamaoshen@bjut.edu.cn Tel: +86-10-67391642
♦ E-mail: bing.bu@outlook.com Tel: +86-10-67391642
⍲ Institute of Acoustics, Chinese Academy of Sciences, Beijing China
⍲ E-mail: yaodingding@hccl.ioa.ac.cn Tel: +86-10-82547570

***Abstract*— Estimating the direction of arrival (DOA) is an important topic in the array signal processing. This paper addresses the issue of multisource localization in a closed environment. We propose a single source zone (SSZ) detection method based on first-order relative harmonic coefficient (RHC), and designs a dynamic SSZ detection rule. Finally, 2-D kernel density estimation (KDE) and peak search are used to achieve multisource DOA estimation. The proposed method is evaluated by simulation experiment and compared with the reference methods, and the effectiveness of the proposed method is verified.**

I. INTRODUCTION

The purpose of multi-source localization is to estimate the Direction of Arrival (DOA) of each sound source. DOA estimation is widely used in various application scenarios, such as sound event detection [1], fault detection, sound source target tracking [2], speech separation [3], and human-computer interaction [4].

In the early days, the main DOA estimation methods were based on the time difference of arrival (TDOA) [5]. The cross-correlation was adopted in these methods to estimate the relative delay of sources to different microphones [6]. And the estimated relative delay is combined with microphone array geometry to obtain the DOA information. Some improved methods have been proposed to by changing the array structure and enhancing the reliability of geometric mapping. However, the TDOA method is mainly used in single sound source scenarios. Another type of method is multiple signal classification (MUSIC) [7], which uses the orthogonality of signal and noise subspace to obtain the spatial spectrum, and then estimates the direction information of sources through spectral peak search [8]. However, the priori known number of sources is necessary in these methods. Beamforming technology is also an important approach to estimate the multi-source DOAs. The principle is to realize DOA estimation by minimum variance distortionless response beamformer [9]. Recently, deep learning has been introduced into DOA estimation [10]. The mapping between the extracted feature of recording signal and the direction information of sources is constructed to complete DOA estimation of the sound source. Nevertheless, if the number of sources increases, the number of categories of network output increases significantly.

Due to the sparsity of speech signal in the time-frequency (TF) domain, sparse component analysis (SCA) based multi-source DOA estimation methods are proposed using the W-Disjoint Orthogonality (WDO) [11]. The WDO reveals the phenomenon that most of the components of different sources are disjoint in the TF domain. In a word, only one source dominates for most TF points. These points are considered as single-source points (SSPs). Multi-source localization can be transformed into multiple single-source localization by using SSPs [12]. Several SSP detection methods have been proposed for different microphone arrays, such as spherical array [13] and B-format array [14,15]. Hu et al. proposed a new feature named relative harmonic coefficient (RHC) in [16-18], which is used to detect SSPs to achieve multi-source localization. The localization performance of SSP-based methods usually depends on the accuracy and number of detected SSPs. Nevertheless, it is difficult to detect a sufficient number of SSPs for localization as the recording environment deteriorates.

In order to obtain more TF points which are conducive to localization, some DOA estimation methods were proposed by detecting the single source dominant components at TF zone level. If most of the TF components in a TF zone (consist of contiguous and adjacent TF points) belong to the same source, this zone is called a single source zone (SSZ) [19]. The

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

correlation between different channel have been introduce for SSZ detection. Several correlation-based SSZ detection methods have been designed for different arrays, such as circular array [20], linear array [21], acoustic vector sensor [22,23].

In addition, a SSZ detection method was proposed based on the angle distribution of TF points. However, in these SSZ detection methods, the setting of the TF zone is fixed. In fact, single source dominant components are randomly distributed. Therefore, the existing SSZ detection methods cannot detect these components as much as possible. In order to solve this issue, this paper designs a dynamic SSZ detection method. Meanwhile, a SSZ detection method is proposed based on the first-order RHC.

The reminder of the paper is organized as follows: the spherical harmonic decomposition of sound pressure and RHC are introduced in Section II. Next, the proposed DOA estimation method is presents in Section III. Then, the performance evaluation through various experiments is given in Section IV. Finally, some conclusion are discuss in Section V.

II. PRELIMINARY KNOWLEDGE

This section introduces some preparatory knowledge about spherical harmonic decomposition and gives the definition of RHC to prepare for the proposed method.

A. Spherical Harmonic Decomposition of Sound Pressure

Assume there are Q simultaneous active sound sources in a closed room with reverberation, and a spherical microphone array consists of J sensors is used to pick up the sound pressure, the sound pressure measured by the j-th microphone in the TF domain is given as,

$$X_j(l,k) = \sum_{q=1}^{Q} S_q(l,k) H_{q,j}(l,k) + V_j(l,k), \quad (1)$$

where l and k denote the indices of frame and frequency, respectively. $S_q(l,k)$, $H_{q,j}(l,k)$ and $V_j(l,k)$ denote the STFT coefficients of the q-th source, acoustic transfer function from the q-th source to the j-th microphone and white Gaussian noise attached to the j-th microphone, respectively. In remainder of this paper, the frame index l will be omitted for simplicity unless otherwise specified. Further, the sound pressure can be decomposed by a complete set of orthogonal spherical harmonic basis functions [16],

$$X_j(k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k) \beta_n(\bar{k}r) Y_{nm}(\vartheta_j, \varphi_j), \quad (2)$$

where $N = \lceil \bar{k}r \rceil$ denotes the spherical harmonic order of truncation, $\bar{k} = 2\pi f / c$ is the wave number, where f and c are the frequency and the speed of sound, respectively, r is the array radius. n and m are the spherical harmonic expansion order and mode, respectively. $Y_{nm}(\vartheta_j, \varphi_j)$ is the spherical harmonic function,

$$Y_{nm}(\vartheta_j, \varphi_j) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_{nm}(cos\vartheta_j) e^{im\varphi_j}, \quad (3)$$

where ϑ_j and φ_j are the elevation and azimuth, respectively, of the j-th microphone. $\alpha_{nm}(k)$ in (2) is the spherical harmonic coefficient of sound field，which can be obtained using the weighted sum of the sound pressure measured by each microphone,

$$\alpha_{nm}(k) = \frac{1}{\beta_n(\bar{k}r)} \sum_{j=1}^{J} w_j X_j(k) Y_{nm}^*(\vartheta_j, \varphi_j), \quad (4)$$

where w_j is the weight of the corresponding microphone on the array. $\beta_n(\bar{k}r)$is a function related to array configuration and wave number, whose definition is,

$$\beta_n(\bar{k}r) = \begin{cases} j_n(\bar{k}r), & \text{open array} \\ j_n(\bar{k}r) - \dfrac{j_n'(\bar{k}r)}{h_n'(\bar{k}r)} h_n(\bar{k}r), & \text{rigid array} \end{cases} \quad (5)$$

where $j_n'(\cdot)$ and $h_n'(\cdot)$ are the first partial derivative of the first kind of n-order spherical Bessel and Hanke functions, respectively.

B. RHC Definition

In the far-field hypothesis, the recording received by the microphone follows the plane wave propagation model. According to the Green's function, hence, the spherical harmonic coefficient $\alpha_{nm}(k)$ for the TF point dominated by the q-th source is [17],

$$\alpha_{nm}(k) = S_q(k) 4\pi i^n Y_{nm}^*(\vartheta_q, \varphi_q), \quad (6)$$

where $(\cdot)^*$ denotes the conjugate operation, ϑ_q and φ_q are the elevation and azimuth of the q-th source, respectively. The definition of RHC is the ratio between $\alpha_{nm}(k)$ and $\alpha_{00}(k)$, i.e.,

$$\lambda_{nm}(k) = \frac{\alpha_{nm}(k)}{\alpha_{00}(k)} = 2\sqrt{\pi} i^n Y_{nm}^*(\vartheta_q, \varphi_q). \quad (7)$$

Therefore, a first-order RHC vector is given as,

$$\boldsymbol{\lambda}(k) = \left[\lambda_{1,-1}(k), \lambda_{1,0}(k), \lambda_{1,1}(k)\right]^T, \tag{8}$$

where each element is frequency independent and only related to the azimuth and elevation of the source. Further, $\boldsymbol{\lambda}(k)$ can be rewritten according to the expression of spherical harmonic function,

$$\boldsymbol{\lambda}(k) = \left[\sqrt{\frac{3}{2}}\sin(\vartheta_q)\, e^{i\left(\varphi_q+\frac{\pi}{2}\right)}, \sqrt{3}\cos(\vartheta_q) e^{i\frac{\pi}{2}}, \sqrt{\frac{3}{2}}\sin(\vartheta_q)\, e^{-i\left(\varphi_q+\frac{\pi}{2}\right)}\right]. \tag{9}$$

III. THE PROPOSED METHOD

In this section, we propose an SSZ detection method based on first-order RHC. Then, the DOA estimates of TF points are obtained using the relationship between the elements of first-order RHC vector. Finally, 2D kernel density estimation (KDE) and peak search are used to achieve multi-source localization.

A. SSZ Detection

For a single source TF point, we can find that the magnitude of $\lambda_{1,-1}(k)$ is the same as that of $\lambda_{1,1}(k)$ from (9). Assume that there are M TF points in the g-th zone of certain time frame, the construction of this zone is defined as,

$$\boldsymbol{\mathcal{H}}_g = \{k_g, k_g+1, \cdots, k_g+M-1\}, \tag{10}$$

where k_g denotes the first frequency point in the g-th TF zone. Thus, we can construct two magnitude vectors of this zone about $\lambda_{1,-1}(k)$ and $\lambda_{1,1}(k)$,

$$\boldsymbol{\mathcal{H}}_g^{\mathrm{N}} = \left[\left|\lambda_{1,-1}(k_g)\right|, \left|\lambda_{1,-1}(k_g+1)\right|, \cdots, \left|\lambda_{1,-1}(k_g+M-1)\right|\right]^T, \tag{11}$$

$$\boldsymbol{\mathcal{H}}_g^{\mathrm{P}} = \left[\left|\lambda_{1,1}(k_g)\right|, \left|\lambda_{1,1}(k_g+1)\right|, \cdots, \left|\lambda_{1,1}(k_g+M-1)\right|\right]^T. \tag{12}$$

Assume the zone $\boldsymbol{\mathcal{H}}_g$ is an ideal SSZ, where all TF points are SSPs, there is,

$$\left|\lambda_{1,-1}(k)\right| = \left|\lambda_{1,1}(k)\right|. \forall k \in \boldsymbol{\mathcal{H}}_g. \tag{13}$$

Hence, these two vectors are the same under interference-free condition, i.e.,

$$\boldsymbol{\mathcal{H}}_g^{\mathrm{N}} = \boldsymbol{\mathcal{H}}_g^{\mathrm{P}}. \tag{14}$$

Considering that the actual recording environment is usually affected by reverberation, noise and intersource interference, the corresponding elements of $\boldsymbol{\mathcal{H}}_g^{\mathrm{N}}$ and $\boldsymbol{\mathcal{H}}_g^{\mathrm{P}}$ are approximately equal rather than absolute. So, there will be a strong correlation between $\boldsymbol{\mathcal{H}}_g^{\mathrm{N}}$ and $\boldsymbol{\mathcal{H}}_g^{\mathrm{P}}$ when the zone $\boldsymbol{\mathcal{H}}_g$ is an SSZ.

However, for TF zones dominated by multi-source mixed components and/or significantly interfered by reverberation components, $|\lambda_{1,-1}(k)|$ and $|\lambda_{1,1}(k)|$ are usually quite different. Consequently, the correlation between $\boldsymbol{\mathcal{H}}_g^{\mathrm{N}}$ and $\boldsymbol{\mathcal{H}}_g^{\mathrm{P}}$ is weak.

According to the above analysis, we propose to use normalized cross-correlation function to judge the similarity between these two vectors, i.e.,

$$r_{\mathrm{NP}}(g) = \frac{C_{\mathrm{NP}}(g)}{\sqrt{C_{\mathrm{NN}}(g) \cdot C_{\mathrm{PP}}(g)}}, \tag{15}$$

where $C_{\mathrm{NP}}(g)$ denotes the cross-correlation coefficient, it is defined as,

$$C_{\mathrm{NP}}(g) = \left(\boldsymbol{\mathcal{H}}_g^{\mathrm{N}}\right)^T \cdot \boldsymbol{\mathcal{H}}_g^{\mathrm{P}}. \tag{16}$$

Therefore, if TF zone $\boldsymbol{\mathcal{H}}_g$ satisfies,

$$r_{\mathrm{NP}}(g) > 1-\xi, \tag{17}$$

it will be identified as an SSZ, where ξ is a pre-defined SSZ detection threshold, a small ξ indicates that the stronger the correlation between the two vectors and this zone is an SSZ with high probability.

Traditional SSZ detection methods usually divide the frequency-band into several zones of fixed length, and then detect whether these zones are SSZ in turn. We know that when most of the points in a TF zone are SSPs, this zone can be regarded as an SSZ. In fact, SSPs are randomly distributed in the TF domain. The fixed length of zone may lead to the partition of some continuously distributed SSPs into two adjacent TF zones. In this way, these two zones are identified as non-SSZs because of the small number of SSPs. Therefore, many SSZs will not be identified by traditional SSZ detection methods, which results in the loss of localization accuracy. Aiming at this problem, we propose to detect SSZs dynamically. To be specific, if the TF zone $\boldsymbol{\mathcal{H}}_g$ do not satisfy the SSZ detection rule, the traditional detection algorithm commands that the first frequency point of the next zone to be detected is k_g+M, while the proposed detection rule requires moving frequency point by frequency point until a zone satisfying the detection rules is identified. That is, the first frequency point of the next zone to be detected is k_g+1. The detail is shown in **Algorithm 1**. Note that the cost of improving localization accuracy is the increase of computational complexity.

B. DOA Estimation

Benefited from the first-order RHC containing the azimuth and elevation information of source, [17] proposes a direction

vector to achieve single source localization. The definition of direction vector equation is given as,

$$\begin{cases} \mathcal{T}_1(k) = \sqrt{\frac{1}{6}} Im\{\gamma_{1,-1}(k) - \gamma_{1,1}(k)\} = sin(\vartheta)\, cos(\varphi) \\ \mathcal{T}_2(k) = -\sqrt{\frac{1}{6}} Re\{\gamma_{1,-1}(k) - \gamma_{1,1}(k)\} = sin(\vartheta)\, sin(\varphi) \\ \mathcal{T}_3(k) = \sqrt{\frac{1}{3}} Im\{\gamma_{1,0}(k)\} = cos(\vartheta) \end{cases}, \quad (18)$$

where $Re\{\cdot\}$ and $Im\{\cdot\}$ denote the real and imaginary parts of a complex number, respectively. Thus, we can estimate the azimuth and elevation from above relation, respectively,

$$\varphi(k) = arctan\left(\frac{\mathcal{T}_2(k)}{\mathcal{T}_1(k)}\right), \quad (19)$$

$$\vartheta(k) = arctan\left(\frac{\sqrt{\mathcal{T}_1{}^2(k) + \mathcal{T}_2{}^2(k)}}{\mathcal{T}_3(k)}\right). \quad (20)$$

For a detected TF point DOA set, the 2-D angle histogram of can be obtained. Some existing probability density estimation methods such as Gaussian Mixture Model (GMM) or KDE can be used to smooth the histogram. In this paper, we employ 2-D KDE as a post-processing tool. Therefore, the estimated probability density function is given as,

$$\mathcal{F}(\varphi, \vartheta) = \frac{1}{\mathcal{C}} \sum_{k \in \boldsymbol{\mathcal{H}}_{\mathrm{SSZ}}} \frac{1}{\rho_\varphi \rho_\vartheta} \mathcal{K}\left(\frac{\varphi - \tilde{\varphi}(k)}{\rho_\varphi}, \frac{\vartheta - \tilde{\vartheta}(k)}{\rho_\vartheta}\right), \quad (21)$$

where C denotes the number of elements in $\boldsymbol{\mathcal{H}}_{SSZ}$, and $\mathcal{K}(\cdot)$ is a kernel function. Gaussian kernel function is selected in this work,

$$\mathcal{K}\left(\frac{\varphi - \tilde{\varphi}(k)}{\rho_\varphi}, \frac{\vartheta - \tilde{\vartheta}(k)}{\rho_\vartheta}\right) = \frac{1}{2\pi} e^{-\frac{1}{2}\left[\frac{\left(\varphi - \tilde{\varphi}(k)\right)^2}{\rho_\varphi{}^2} + \frac{\left(\vartheta - \tilde{\vartheta}(k)\right)^2}{\rho_\vartheta{}^2}\right]}, \quad (22)$$

where ρ_φ and ρ_ϑ are the corresponding bandwidth parameter. Subsequently, the peak search algorithm is chosen to estimate the DOAs of sources. The proposed DOA estimation method is presented in **Algorithm1**.

Algorithm 1: The proposed DOA estimation method

Input: $X_j(l,k), j = 1,2,\cdots J$ // the sound pressure

Output: $\{\vartheta_q, \varphi_q\}_{q=1}^{Q}$ // the estimated DOAs

---

$l = 1$ // the index of the frame
$\lambda_{nm}(l,k)$ is obtained based on (7)
while $l \le L$ **do** // dynamic SSZ detection in the l-th frame
  $g = 1$ // the index of the detected SSZ
  $k = 1$ // the index of frequency point
  while $k \le K$ **do**
    $k_g = k$
    $\boldsymbol{\mathcal{H}}_g^{\mathrm{N}}$ and $\boldsymbol{\mathcal{H}}_g^{\mathrm{P}}$ are obtained based on (11) and (12)
    $r_{\mathrm{NP}}(g)$ is obtained based on (15)
    if $r_{\mathrm{NP}}(g) > 1 - \xi$, **do**
      $k = k + M$
      $g++$
    end if
    $k++$
  end
  $g++$
end

$$\boldsymbol{\mathcal{H}}_{\mathrm{SSZ}} = \bigcup_{l=1}^{L} \bigcup_{g=1}^{l_g} \boldsymbol{\mathcal{H}}_{(l,g)}$$

// TF points set of the detected SSZs
// l_g denotes the number of the detected SSZs in the l-th frame
$\{\vartheta_q, \varphi_q\}_{q=1}^{Q}$ are estimated by 2-D KDE and peak search

IV. Experimental Evaluation

This section used several simulated experimental data to evaluate the performance of the proposed method. Room impulse response (RIR) is generated using a famous mirror sound source model [24]. The experiment was conducted in a closed room with size of $[6m, 4m, 3m]$, where an open spherical microphone array is simulated and is placed in the center, with a distance of $1m$ between the sound source and the array. Note that the array consists of only four omnidirectional microphones located on non-adjacent vertices of a cube. The speech source is taken from the Nippon Telegraph and Telephone (NTT) database, and the sampling frequency is 16kHz. When the time domain signal is converted into the TF domain, 2048-point discrete Fourier transform (DFT) is done for the sample points with a window length of 0.032s, where the overlap rate between adjacent frames is 50%. Besides, frequency points in the frequency range of [800Hz, 3200Hz] in each frame are selected. The length of the detected TF zone was set to 10 and the SSZ detection threshold was set to $\xi = 0.04$. We use the mean absolute estimated error (MAEE) to evaluate the localization performance,

$$\mathrm{MAEE} = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{2Q} \sum_{q=1}^{Q} \left| \tilde{\varphi}_{t,q} - \varphi_{t,q} \right| + \left| \tilde{\vartheta}_{t,q} - \vartheta_{t,q} \right| \right), \quad (23)$$

where t and q denote the indices of the t-th experiment and the q-th source, T denotes the number of experiments performed.

Fig.1 demonstrates the MAEE results of the proposed methods and the other two reference methods under different reverberation times {0.2 s, 0.3 s, 0.4 s, 0.5 s} with SNR = 25 dB, where (a) and (b) denote the two- and three-source scenarios, respectively and source separation is set to 60°. The method without any process (TF points in the frequency band of the selected frequency range are used) is labeled as "Baseline" and single source confidence measure-based method is represented using the abbreviation "SSCM" [21]. The method based on the proposed SSZ detection rule with fix length of TF zone is represented using the abbreviation "Proposed SSZ detection non-dynamically". And the method based on the proposed dynamic SSZ detection is represented as "Proposed SSZ detection dynamically". It is obvious that these four methods can all exhibit excellent localization performance under low reverberation environment. In comparison with two-source scenario, the DOA estimation accuracy is lower in the three-source scenario. Besides, with the increase of reverberation time, their localization accuracy deteriorates significantly due to the directional interference from early reverberation components and contamination from late reverberation components. Generally, the MAEE result of the DOA estimation method based on the proposed dynamic SSZ detection achieves lower error compared with the other three methods, which also verifies the effectiveness of the proposed method.

It is well-known that the angular separation between sources also influence the localization accuracy, which is similar to the principle of reverberation interfere localization performance, a smaller separation usually increases the risk of contamination of SSPs. Fig. 2 compares the MAEEs of these four methods at different angular separations with reverberation time of 0.4 s and SNR = 25 dB. When the angular separation is 90° and 120°, the localization accuracy of these methods is high. While the angular separation is set as 60° and 40°, the performance of the reference methods gradually increases, but the DOA estimation method based on the proposed dynamic SSZ detection can achieve a good localization performance.

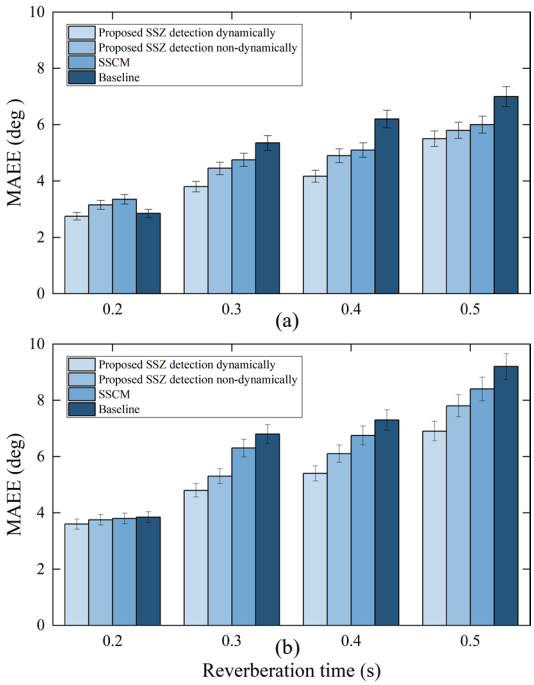


Fig. 1. MAEE results with different reverberation times in (a) two- and (b) three-source scenario.

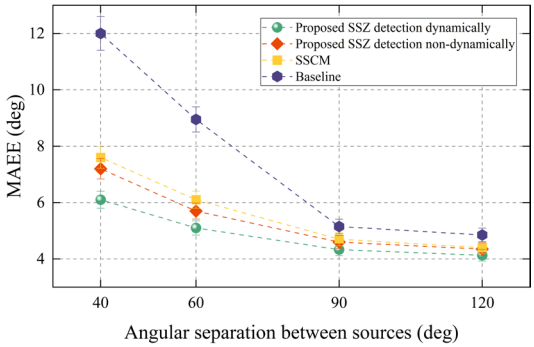


Fig. 2. MAEE results with different angular separation between sources in two-source scenario.

V. CONCLUSION

This paper proposed a multisource localization method based on SSZ detection in spherical harmonic domain. We use the consistency between the magnitude of elements with mode -1 and 1 in first-order RHC to identify TF zones where only one source dominants. In addition, different from the traditional SSZ detection rule, we propose to detect the SSZs dynamically, which can improve the localization accuracy. Then, 2-D KDE and peak search are used as post-processing to realize the multisource localization task. In the future, we focus on analyzing SSP or SSZ characteristics in more complex scenario to achieve multisource localization with high precision and robustness and apply this feature to deep learning.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants (No. 61971015), Beijing Natural Science Foundation (No. L233032, L223033).

REFERENCES

[1] X. Chen, W. Wang, Y. Wang, X. Zhong, and A. Alinaghi, "Reverberant speech separation with probabilistic time-frequency masking for b-format recordings," *Speech Commun.*, vol. 68, pp. 41–54, 2015.

[2] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 520–529, Sep. 2004.

[3] Belouchrani A, Amin M, Thirion-Moreau N, et al, "Source separation and localization using time-frequency distributions," IEEE Signal Processing Magazine, vol.30, no.6, pp.97-107, 2013.

[4] T. Latif, E. Whitmire, T. Novak, and A. Bozkurt, "Sound localization sensors for search and rescue biobots," *IEEE Sensors J*, vol. 16, no. 10, pp. 3444–3453, May 2016.

[5] Chen, J., Benesty, J., Huang, Y., "Time delay estimation in room acoustic environments: an overview," *EURASIP J. Appl. Signal Process*, 26503, pp.1–19, 2006.

[6] Knapp, C., Carter, G., "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process*, vol. ASSP-24, no.4, pp.320–327, 1976.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986

[8] Hu, Y., Abhayapala, T.D., Samarasinghe, P.N., "Multiple source direction of arrival estimations using relative sound pressure based MUSIC," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol.29, pp.253–264, 2021.

[9] Daniele Salvati, Carlo Drioli, Gian Luca Foresti, "A Low-Complexity Robust Beamforming Using Diagonal Unloading for Acoustic Source Localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol.26, no.3, pp.609-622, 2018.

[10] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," *IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, 2017: 2217-2221.

[11] Yılmaz, O., Rickard, S., "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol.52, no.7, pp.1830–1847, 2004.

[12] Jia, M., et al., "Multiple-to-Single sound source localization by applying single-source bins detection," *Appl. Acoust.*, vol.138, pp.28–38, 2018.

[13] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using relative harmonic coefficients," *IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, 2020, pp. 571–575.

[14] K. Wu, V. G. Reju, and A. W. H. Khong, "Multi-source DOA estimation in a reverberant environment using a single acoustic vector sensor," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1848–1859, Oct. 2018.

[15] Jia, M., et al., "Multi-source DOA estimation in reverberant environments by jointing detection and modeling of time-frequency points," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol.29, pp.379–392, 2021.

[16] Hu Y, Samarasinghe P N, Gannot S, et al., "Semi-Supervised Multiple Source Localization Using Relative Harmonic Coefficients Under Noisy and Reverberant Environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol.28, pp.3108-3123, 2020.

[17] Y. Hu, S. Gannot, "Closed-form single source direction-of-arrival estimator using first-order relative harmonic coefficients," *IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, 2022, pp. 726-730.

[18] Hu, Y., et al. "Decoupled Multiple Speaker Direction-of-Arrival Estimator Under Reverberant Environments." *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol.30, pp.3120-3133, 2022.

[19] M. Jia, J. Sun, and C. Bao, "Real-time multiple sound source localization and counting using a soundfield microphone," *J. Ambient Intell. Humaniz. Comput.*, vol. 8, no. 6, pp. 829–844, 2017.

[20] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[21] D. Ying, R. Zhou, J. Li, and Y. Yan, "Window-dominant signal subspace methods for multiple short-Term speech source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 731–744, Apr. 2017.

[22] K. Wu, V. G. Reju, and A. W. H. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Queensland, Australia, 2015, pp. 444–448.

[23] Zou, Y., et al., "Multi-source DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 4011–4015.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Robust Representation Learning for Speech Emotion Recognition with Moment Exchange

Yunrui Cai*[1], Changhe Song*[1], Boshi Tang*, Dongyang Dai*, Zhiyong Wu*†✉, Helen Meng*†
* Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
E-mail: {caiyr22, songch22, tbs22, ddy17}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn✉
† Department of Systems Engineering and Engineering Management,
The Chinese Unviersity of Hong Kong, Hong Kong SAR, China
E-mail: hmmeng@se.cuhk.edu.hk

***Abstract*—Speech emotion recognition (SER) is essential for the machine to understand human intentions. Various deep neural network (DNN) based models are proposed but still suffer from overfitting on small training dataset, especially the identical distribution assumption is usually violated in practical datasets. Data augmentation with normalization for single instances is a common approach to train models more stably, where the moments (a.k.a. mean and variance) of the augmented data representation are removed as noise. To further improve the robustness, a label-perturbing data augmentation method Moment Exchange (MoEx) is introduced in this paper, forcing DNN models to pay attention to the difference in moments extracted from individual instances of each emotion class. Experiments on several classical model structures demonstrate the promising performance improvement with an extremely simple but effective implementation of MoEx.**

I. Introduction

Emotions are an important factor in enabling machines to understand human intentions during human-computer interaction. Speech is one of the most natural ways for humans to express emotions in real life [1]. Nowadays, machines have become proficient at processing speech signals, and have achieved high performance in areas such as automatic speech recognition (ASR) and speaker recognition. Speech emotion recognition (SER) has also recently gained attention, as it can make human-computer interactions more humane and harmonious.

A typical SER system extracts emotion features from the original speech signals and then classifies them with machine learning models into specific emotion categories [2].Prosodic features like energy, pitch, and duration are proved to be useful in emotion prediction tasks. Researchers also use Teager Energy Operator (TEO) features [3] and some hand-crafted features [4] to recognize emotions. Recently, spectral features have been widely used and performed relatively better in SER task [5].

With the rapid development of deep learning, deep neural network (DNN) has brought a lot of breakthroughs in many fields, which includes the SER task. A convolutional neural network (CNN) was firstly proposed to learn local feature information in spectrograms [5], which outperforms traditional machine learning methods on multiple metrics. Recurrent neural networks (RNN) are also used to learn timing information in speech signals [6]. CNN and Long Short-term Memory (LSTM) are fused into one model [7], in which CNN can process spectrogram features, and then LSTM can predict emotions from timing CNN outputs. More models based on CNN+RNN are proposed because of its excellent performance. The RNN model is replaced by the Bi-RNN model in [8], which is proved better in audio tasks [9] because it is able to take the longer bidirectional features of input vectors into account. A new SER model based on non-linear Radial Basis Function Network (RBFN) [10] is proposed with CNN and deep BiLSTM [11], in which ResNet [12] was further applied in the CNN model. Transformer model [13], a multi-head attention mechanism, allows for parallelization and jointly attends to the feature information. It has also been widely used in SER [14] recently. Conformer model [15] is proposed to combine convolutional layers and a transformer model for speech recognition. Conformer blocks are also used in SER task to learn local information from extracted emotional features [16], which is the state-of-the-art work.

However, one of the challenges that the SER task is facing now is that the existing datasets are relatively small, while the robustness of emotional representations extracted from speech signals is particularly limited to the training dataset [17]. The performance of complex deep networks usually plummets under the situation where the data distribution between the training dataset and the testing dataset is not identical in practice [18]. A piece of evidence is that categories can be separated on the training set while still being mixed together on the testing set. Commonly, data augmentation is used to solve the above problems [19], and it has been proved useful in image [20] and text [21] tasks, which can be divided into two categories: label-preserving or label-perturbing. Compared with label-preserving methods, label-perturbing methods process the labels of augmented data more smoothly and achieve better performance. For example, the corresponding labels of two inputs are linearly interpolated in MixUp [22].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Normalization operations including LayerNorm (LN) [23], InstanceNorm (IN) [24] or PositionalNorm (PONO) [25] on the label-perturbing augmented single instances are usually needed to make sure that training is stable. Among these methods, the first moment (a.k.a. mean) and the second moment (a.k.a. variance/standard deviation) are removed in intra-sample normalization, which means, simply subtracting the mean and dividing by the variance. After normalization, an underlying representation of single instances with more robustness is captured. Instead of simply removing the moments, Moment Exchange (MoEx) [26] is further proposed to improve the stability of training by randomly exchanging the feature moments across samples.

In this paper, we apply MoEx to the previously mentioned classical model structures of SER task and achieve a significant improvement in performance with extremely simple implementation. More specifically, the mean and variance of representation from a certain selected layer of every sample are computed and then re-injected into target representation from the same layer of other randomly-selected samples among one batch during training, which is figuratively called a Moment Exchange operation. Meanwhile, the corresponding labels involved in each exchange are interpolated by a hyper-parameter when calculating the loss. The experiments show that the classification performance and robustness of the previously mentioned classical model structures are improved by MoEx because the representation from the selected layer has more discriminative power by forcing each model to pay special attention to the variation of moments extracted from different intra-samples. Among these structures, Conformer with MoEx is verified to get the best performance. The sensitivity of parameters is also fully verified.

II. Methods

The framework of the proposed model is depicted in Figure 1, which includes the structure of the SER model and the structure of Moment Exchange.

SER works generally first extract speech features, and then use a series of deep learning networks to learn these features in the hidden layers. However, there are enormous challenges in the robustness of current SER works, for which we introduce Moment Exchange into the common SER models. Moment Exchange is applied to the hidden layers' output to learn normalized features of the original label and moment information. The last linear and softmax layers are used for emotion classification.

A. SER Model details

We extract the Mel-spectrogram features F of the speech signal as input. We set $F \in \mathbf{R}^{T \times M}$, where T is variable based on the length of the audio and M is the number of Mel bands.

In the field of speech emotion recognition (SER), several deep learning architectures have been developed to learn speech features. Common deep learning model structures such as CNN, RNN,and Transformer are used as hidden layers respectively. Additionally, we use the mask to compress the variable length of the inputs to a fixed-length output vector, resulting from T.

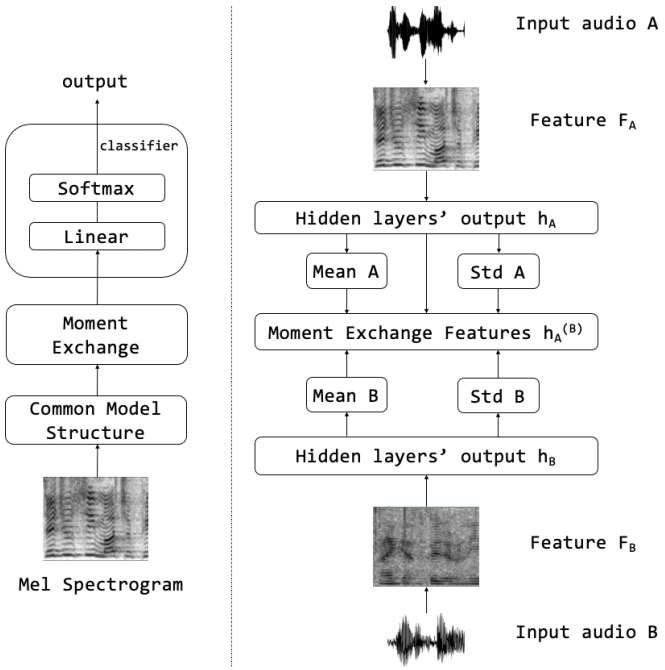


Fig. 1. Framework of the proposed model. The SER model (left) takes mel spectrograms as input and learns discriminative features using deep learning and Moment Exchange. The Moment Exchange (right) injects the moments of hidden layers' output from randomly chosen audio B to audio A.

Representative deep learning models employed in the SER domain often encounter problems with unstable training and overfitting, likely due to small datasets and complex feature dimensions. To counter these issues and improve overall model robustness and classification performance, data augmentation is a widely used technique. While more complex methods, such as collecting noisy data and utilizing adversarial attacks have been employed [27], we propose using Moment Exchange directly within the model as a simple and effective data augmentation method.

The Moment Exchange is performed within each batch for the final hidden output h of the deep learning model and does not change the dimensions, which will be elaborated on in Section 2.2. After that is an emotion classifier, which contains a linear layer and a softmax layer. The output of the linear layer indicates the expected posterior probabilities for the relevant emotion categories and is used for calculating softmax cross-entropy loss. For convenience, we define the function of the classifier here as $c(h_i)$ and softmax cross-entropy loss is $l\big(c(h_i), y_i\big)$, where h_i is hidden layers' output and y_i is the ground truth label of the i-th input F_i.

B. Moment Exchange

Moment Exchange is applied to audio A and audio B as shown in Figure 1, where A and B are two randomly chosen samples. Specifically speaking, a random full permutation

array p is generated according to the length b of the batch and then $B = p[A]$.

PONO [25] is used to compute the first moment (mean) μ_i and the second moment (variance) σ_i of hidden features h_i:

$$\mu_i = \frac{1}{H}\sum_{j=0}^{H-1} h_i[j] \tag{1}$$

$$\sigma_i = \sqrt{\frac{1}{H}\sum_{j=0}^{H-1}(h_i[j] - \mu_i)^2} \tag{2}$$

H in formula (1)(2) is the size of the hidden layers' output. To motivate the model to make use of the moments information, the moments of hidden layers' output from audio B are injected into the normalized features of audio A:

$$h_A^{(B)} = \sigma_B \frac{h_A - \mu_A}{\sigma_A} + \mu_B \tag{3}$$

The new Moment Exchange features $h_A^{(B)}$ contain both the normalized features of audio A and the moments of audio B. Therefore, we need to modify the loss function to let the model notice the label information of y_A and also y_B at the same time. Here we directly introduce a new hyper-parameter λ to propose a straight-forward combination function:

$$\mathcal{L} = \lambda * l(c(h_A^{(B)}), y_A) + (1-\lambda) * l(c(h_A^{(B)}), y_B) \tag{4}$$

The hyper-parameter λ and $1-\lambda$ is the target interpolation weight of label y_A and y_B. And l represents the loss function of the emotion classifier mentioned above.

III. Experiments

A. Dataset and Feature extraction

We evaluated the effectiveness of introducing Moment Exchange into different model structures on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [28] dataset. This dataset is widely used for studying English speech emotions in the SER task and was collected by the University of Southern California (USC) Speech Analysis and Interpretation Laboratory (SAIL). The dataset contains approximately 12 hours of audio data of ten speakers expressing different emotions in specific scenarios. The original dataset includes nine different emotions, but to better stimulate emotional responses, it makes more sense to group certain similar emotions together. We combined similar emotions to create four categories: happiness, anger, neutral, and sadness (as shown in Table I).

TABLE I
THE DISTRIBUTION OF EMOTIONS IN IEMOCAP

Emotion	Utterance	Participation(%)
Happiness	1636	29.6
Anger	1103	19.9
Neutral	1708	30.9
Sadness	1084	19.6

Log scale Mel-spectrogram is used as the emotion feature. To generate this, we resample the original audio at 16kHz and only use the middle 14 seconds based on statistics of the audio duration. Spectrograms are calculated with a DFT length of 2048, 128 Mel bands, and a Hamming window with a window shift of 10 milliseconds and window size of 40 milliseconds.

B. Experiment setup

We implement our model on the Pytorch deep learning library using Python. Training is done on a machine equipped with a GeForce RTX 2080Ti using Adam optimizer. A learning rate of 0.001 and a drop rate of 0.05 are set during training.

To evaluate our model, we utilize a 5-fold cross-validation (CV) and report the average results. During each fold, 20% of the data are used for testing, while the remaining 80% is randomly split into 80% for training and 20% for validation. Additionally, to account for category imbalance in the IEMOCAP dataset, we use both unweighted accuracy (UA) and weighted accuracy (WA) to compare model performance in our experiments.

To demonstrate the effectiveness of the Moment Exchange operation, we introduce it into five classical model structures. We then compare the performances of the original models and the models with Moment Exchange. The five model structures include the representative DNN baseline models of CNN [29], CNN+unidirectional RNN [7], CNN+bidirectional RNN [30], LSTM+Transformer [14] and Conformer [15]. The performance of Moment Exchange is verified in experiment 1. In experiment 2, we test the sensitivity of the parameters to further evaluate the stability and effectiveness of Moment Exchange.

C. Experimental Results

Main experiment For this experiment, we uniformly set λ to 0.95 and used a mini-batch size of 56. The experimental results of the five deep learning models, with and without Moment Exchange, are presented in Table II. The results show that the Conformer model with Moment Exchange outperformed the other baseline models by achieving 71.6% and 70.5% on the UA and WA metrics, respectively. Furthermore, after implementing Moment Exchange to all five models, both UA and WA values increase in different levels.

TABLE II
THE UA AND WA ON DIFFERENT MODELS WITH AND WITHOUT MOMENT EXCHANGE

Model	UA(%)	WA(%)
CNN	53.3	53.3
+Moment Exchange	54.7	54.6
CNN+RNN	57.9	57.9
+Moment Exchange	60.9	59.9
CNN+Bi-RNN	61.8	60.4
+Moment Exchange	65.5	64.8
LSTM+Transformer	67.2	66.2
+Moment Exchange	70.1	69.5
Conformer	68.4	67.1
+Moment Exchange	**71.6**	**70.5**

Compared to CNN and RNN models, Transformer-based models offer better global information learning and model

expression capabilities, for which they can achieve higher performance. The Conformer model enhances the transformer structure by incorporating a convolution module to learn local features from mel spectrograms. This approach combines both global and local information, resulting in optimal performance. In addition, Moment Exchange helps the model to further learn the robust expression of speech features, so Conformer model with Moment Exchange achieves the highest metrics among all the models in our experiments.

Confusion matrix The confusion matrix of the Conformer model with and without Moment Exchange are shown in Table 3. The results indicate that the accuracy of all classes has been improved to some extent following the introduction of Moment Exchange. This outcome serves as evidence of the capability of the Moment Exchange technique to enhance model classification ability by exchanging information between different samples and categories. Furthermore, Moment Exchange can be widely applied to various deep learning models and is effective across all emotional categories. These results underscore the role of Moment Exchange in improving the robustness of the original model's representation.

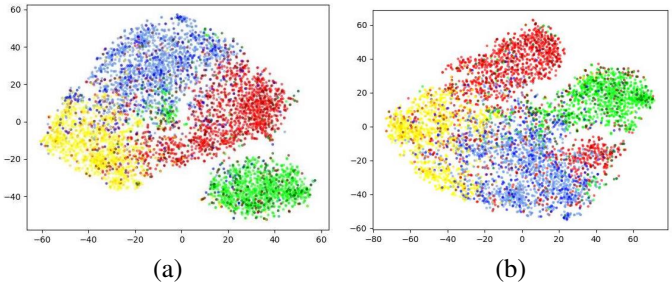
TABLE III
CONFUSION MATRIX (%) WITH AND WITHOUT MOMENT EXCHANGE

with/without	hap	ang	neu	sad
hap	**65.0**/58.9	8.8/10.4	17.2/20.6	9.0/10.1
ang	13.7/19.5	**70.7**/64.3	12.7/11.0	2.9/5.2
neu	17.0/16.7	4.5/5.5	**68.2**/63.0	10.3/14.8
sad	6.1/6.9	3.1/2.6	8.2/8.4	**82.6**/82.1

Robustness representaion To demonstrate the impact of Moment Exchange, we utilize Principal Component Analysis (PCA) to embed learned features before softmax layer in both the Conformer model with and without Moment Exchange. For visual analysis, we use the same fold of cross-validation.

Figure 2 presents the characteristics of the train set and the test set on a graph, with (Figure 2-(a)) and without (Figure 2-(b)) Moment Exchange. Different colors represent distinct types, while dark colors within the same color refer to features in the training set. As shown in the figure, we find that the different feature categories are more clearly distinguished and less confused in the model with Moment Exchange. To clarify, let's utilize green dots to demonstrate the emotion of **anger** in this passage. Green dots are more off-center and more clustered internally in the (Figure 2-(a)) than that in Figure 2-(b)), which has a significant impact on the model's classification. In the confusion matrix, we can alse see that the accuracy of anger emotion with Moment Exchange is 70.7%, higher than that without Moment Exchange at 64.3%, which confirms the analysis brought by the visual results. Similar conclusions can be drawn from the visualization results and the numerical results of the confusion matrix on other categories.

Figure 3 presents the characteristics of the model with and without Moment Exchange on the train set (Figure 3-(a)) and the test set (Figure 3-(b)). Similar to the previous figure, different colors represent different types, and dark colors of

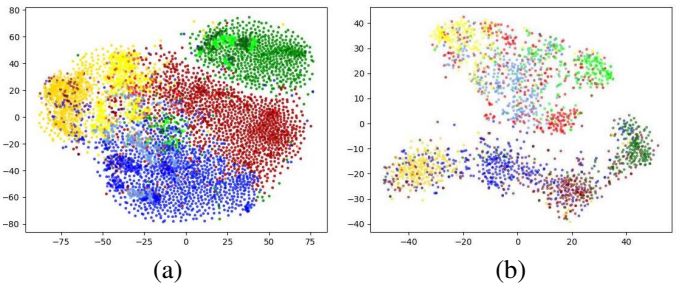


(a) (b)

Fig. 2. PCA embedding of features including train and test set from (a) with Moment Exchange and (b) without Moment Exchange. Dark colors represent the train set. **The model with Moment Exchange can separate various emotions more clearly.**

the same color represent features from the model with Moment Exchange. We observe that with Moment Exchange, the dark points tend to move further outward, indicating better distinction between different categories during training. Furthermore, in the test set, dark points with Moment Exchange exhibit clearer distinction, implying better classification performance. As an example, let's also consider the emotion of anger. The dark green dots are located further out in the training set and are more concentrated and distant from the other emotions in the test set. Despite the unequal distribution of green points within the two sets, model with Moment Exchange is proficient in classifying them effectively based on their distinct regions.

Consequently, we can conclude that Moment Exchange enables the original model to amplify the differences between each category in the process of information exchange, so that can better distinguish between different classes when learning speech features during training. Simultaneously, Moment Exchange enhances the model's ability to discriminate during the final test. Moment Exchange enables a model to exchange information, improving the features of a specific class, like the anger class. This strengthens internal features of green points and differentiates them from other color points. Therefore, this method is an effective way to augment data.



(a) (b)

Fig. 3. PCA embedding of features including with/without Moment Exchange from (a) train set and (b) test set. Dark colors represent the model with Moment Exchange. **The darker dots are more scattered around the periphery in (a); The division of dark points is more obvious in (b). Moment Exchange enables the model to enhance the classification ability at different stages**

D. Parameter sensitivity

In order to investigate the stability of Moment Exchange, we test the Conformer and CNN+Bi-RNN model on different hyper-parameters.

TABLE IV
COMPARISON AMONG DIFFERENT λ

	Conformer		CNN+Bi-RNN	
λ	UA(%)	WA(%)	UA(%)	WA(%)
0.3	64.2	64.8	57.8	55.8
0.5	67.9	67.5	61.0	59.2
0.7	68.9	68.1	63.6	62.2
0.9	69.2	68.4	63.5	63.2
0.95	**71.6**	**70.5**	**65.5**	**64.8**
0.97	69.1	68.3	63.1	62.1
0.99	68.5	68.1	62.1	60.8

The weights λ and $1-\lambda$ are used to interpolate between labels y_A and y_B. To demonstrate how λ impacts model performance, we trained our model on the IEMOCAP dataset using values of λ ranging from 0.3 to 0.99. Our results, shown in Table IV, indicate that both UA and WA are maximized when $\lambda = 0.95$. We found that higher values of λ generally lead to increased UA and WA, likely because the normalized features of the original label contain more critical information than the moments, as noted in [26]. However, performance begins to decrease once λ exceeds 0.95, possibly due to the fact that the moments must also contribute to the final result and that λ has a marginal utility.

IV. CONCLUSIONS

This paper presents an approach to enhance DNN training stability and improve classification accuracy and robustness for SER using Moment Exchange. Results indicate that Moment Exchange with appropriate hyper-parameters can improve the model's SER performance, as demonstrated by both experimental numerical comparison and visual analysis. In future work, we plan to explore more robust normalization methods and moment injection techniques. We also aim to conduct experiments on different layers of various DNN models to enhance the model's discrimination power and further improve SER performance.

ACKNOWLEDGMENT

This work was partially supported by National Natural Science Foundation of China (62076144) and Shenzhen Science and Technology Program (WDZC20200818121348001, WDZC20220816140515001).

REFERENCES

[1] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1334–1345, 2007.

[2] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, 2022.

[3] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.

[4] B. W. Schuller, S. Steidl, A. Batliner, *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), Makuhari, Chiba, Japan, September 26-30, 2010*, ISCA, 2010, pp. 2794–2797.

[5] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *2017 ACM on Multimedia Conference (MM 2017), Mountain View, CA, USA, October 23-27, 2017*, ACM, 2017, pp. 478–484.

[6] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *CoRR*, vol. abs/1701.08071, 2017. arXiv: 1701.08071.

[7] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, April 15-20, 2018*, IEEE, 2018, pp. 5089–5093.

[8] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Stockholm, Sweden, August 20-24, 2017*, ISCA, 2017, pp. 1089–1093.

[9] A. Ogawa and T. Hori, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks," *Speech Commun.*, vol. 89, pp. 70–83, 2017.

[10] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," *Knowl. Based Syst.*, vol. 117, pp. 56–69, 2017.

[11] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 2016, pp. 770–778.

[13] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NeurIPS 2017), December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

[14] F. Andayani, L. B. Theng, M. T. K. Tsun, and C. Chua, "Hybrid LSTM-Transformer model for emotion recog-

nition from speech audio files," *IEEE Access*, vol. 10, pp. 36 018–36 027, 2022.

[15] A. Gulati, J. Qin, C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Virtual Event, Shanghai, China, 25-29 October 2020*, ISCA, 2020, pp. 5036–5040.

[16] P. Zhao, F. Liu, and X. Zhuang, "Speech sentiment analysis using hierarchical Conformer networks," *Applied Sciences*, vol. 12, no. 16, p. 8076, 2022.

[17] S. Prasanth, M. R. Thanka, E. B. Edwin, and V. Nagaraj, "Speech emotion recognition based on machine learning tactics and algorithms," *Materials Today: Proceedings*, 2021.

[18] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India, 2-6 September 2018*, ISCA, 2018, pp. 3087–3091.

[19] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: When to warp?" In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2016, pp. 1–6.

[20] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[21] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–34, 2021.

[22] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[25] B. Li, F. Wu, K. Q. Weinberger, and S. J. Belongie, "Positional normalization," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 1620–1632.

[26] B. Li, F. Wu, S. Lim, S. J. Belongie, and K. Q. Weinberger, "On feature normalization and data augmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021), virtual, June 19-25, 2021*, Computer Vision Foundation / IEEE, 2021, pp. 12 383–12 392.

[27] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020), Virtual Event, Shanghai, China, 25-29 October 2020*, ISCA, 2020, pp. 2327–2331.

[28] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[30] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), Brighton, United Kingdom, May 12-17, 2019*, IEEE, 2019, pp. 7405–7409.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Few Shot Learning Guided by Emotion Distance for Cross-corpus Speech Emotion Recognition

Pengcheng Yue, Yanfeng Wu, Leyuan Qu, Shukai Zheng, Shuyuan Zhao and Taihao Li †
Zhejiang Lab, China
E-mail: {pcyue, wuyanfeng, leyuan.qu, zhengsk, zhaosy, lith}@zhejianglab.com
† Corresponding author

***Abstract*—Cross-corpus speech emotion recognition (SER) is important for building robust and practical SER systems that can adapt to various real-world scenarios and acoustic conditions. Cross-corpus SER faces the problem of low accuracy partly due to the scarcity and inconsistency of labeled data in the target corpus, in which situation prior knowledge about the relationship between emotion categories may be helpful. Previous studies have suggested that discrete emotion categories and continuous emotion space can complement each other in describing emotions. In this paper, we propose to use the distance between emotion categories derived from their distribution in the continuous emotion space as prior knowledge for cross-corpus speech emotion recognition. We hypothesize that this prior knowledge can help the SER system to learn more meaningful and generalizable representations of emotions that are consistent across domains. Experiment results show that the proposed few-shot learning method based on metric learning leveraging the prior knowledge of emotion distance achieves good performance.**

I. Introduction

Cross-corpus speech emotion recognition (SER) is useful for building robust and practical SER systems and can be adapted to many potential applications in human-computer interaction, health care, education, entertainment, security, etc [1][2].

There have been many studies on cross-corpus SER, and most of them focused on the mismatch and gap between the source and target datasets [3][4][5]. However, the relationship between speech emotion categories was ignored.

Two main types of emotion classification systems have been widely used in emotion research and applications, namely discrete emotion categories and continuous emotion. Discrete emotion theories propose that there are a number of emotion categories, such as anger, fear, happiness, sadness, disgust and surprise, that are biologically based and universally recognizable by facial expressions and physiological responses [6][7][8]. These emotion categories are thought to be triggered by specific appraisals of stimuli or events that have implications for one's goals or survival. Dimensional models of emotion suggest that emotions can be described along a few continuous dimensions, such as valence (positive-negative), arousal (high-low) and dominance (active-passive) [9][10][11]. These dimensions are assumed to reflect underlying neurophysiological processes that modulate emotional states. Dimensional models can account for the gradation and mixture of emotions, as well as the similarities and differences among them. In recent years, there have been articles verifying that discrete emotion labeling and continuous emotion labeling can complement each other, and better emotion recognition performance can be achieved through multi-task learning[12][13][14][15]. The logic behind these articles is that both discrete and continuous emotion annotation information can provide a basic emotion feature extraction capability for emotion recognition models, and the information provided by the two is complementary. Especially, continuous emotion annotations can provide useful information for discrete emotion recognition.

Another problem with traditional cross-corpus SER methods is their weak ability in speech representation because they use traditional speech features mostly coming from openSMILE toolkit or low-level speech descriptor instead of state-of-the-art unsupervised pre-trained speech features. [3] considered deep feature extracted by ResNet50 from the Mel spectrograms, but the speech representation ability of this deep feature still needs to be improved compared with the features obtained by unsupervised pre-training, such as Wav2vec 2.0 [16][17][18], Hubert [19][20], etc.

Inspired by the above series of studies on discrete versus continuous emotions, this paper quantifies the distance between emotion categories through their distribution in continuous emotion space, and uses this as prior knowledge in discrete emotion category learning. The coarse-grained distribution of emotions is shown in Figure 1, where the location relationship of the class centers of the four basic emotional categories in the continuous emotional space is presented. For instance, the Euclidean distance between the two emotions of happy and sad is the farthest, because they are relatively far apart in the two dimensions of valence and arousal. According to the distribution, one can divide the distance between pairs of emotion categories into several levels.

In this paper, we propose to use the distance between emotion categories derived from their distribution in the continuous emotion space as prior knowledge for cross-corpus SER. The emotion distance guides the metric loss construction in few-shot learning to learn more meaningful and generalizable representations of emotions that are consistent across domains. The contributions of this paper are as follows: 1) This paper introduces emotion distance for the first time to the cross-corpus SER task. 2) This paper adopts a self-supervised speech feature extractor instead of traditional features for cross-corpus SER. 3) The proposed few-shot learning guided by emotion distance method achieves good cross-corpus SER performance.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

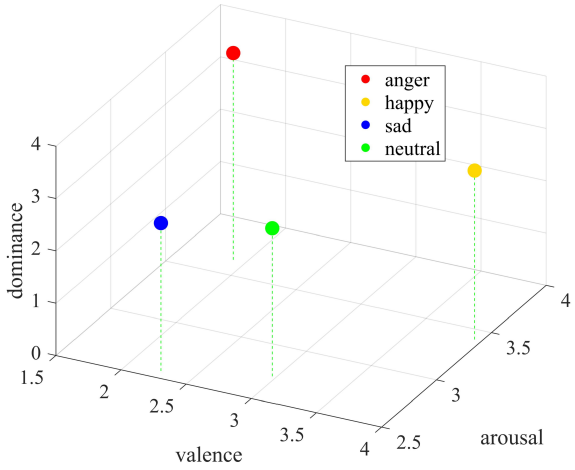


Fig. 1. Emotion distribution in valence-arousal-domination axes

II. Method

In this section, we introduce the baseline transfer learning based on few-shot learning and our proposed method, i.e., few-shot learning guided by emotion distance.

Figure 2 shows the fine-tuning and testing stage of the baseline and emotion distance-guided few-shot learning, from which one can also see how the emotional distance knowledge is used in the construction of the metric loss.

Some general notation assumptions of this paper are as follows. $\mathbf{x}$ represents a speech sample and $\mathbf{F}$ represents the speech emotion feature extractor. $\mathbf{F}(\mathbf{x})$ is the emotion embedding of sample $\mathbf{x}$, and $\mathbf{cen}$ represents the shot average embedding of a certain emotion in the support set. The parameters of the SER model can be represented as $\mathbf{W}$ and the loss function is $L(\mathbf{W})$.

A. Baseline transfer learning based on few-shot learning

A common approach to address cross-corpus speech emotion recognition is to use transfer learning, which leverages the knowledge learned from a source corpus to improve the performance on a target corpus. Besides, few-shot learning that aims to learn from a small number of labeled examples has been proven to be effective in cross-corpus transfer learning.

The baseline method consists of three steps: 1) Training a feature extractor on a large source corpus using a self-supervised contrastive learning objective. 2) Fine-tuning a classifier on a few labeled examples from the target corpus using the extracted features. 3) Testing the SER model on the target corpus. During the fine-tuning stage on the target corpus, the training loss of basic transfer learning (BTL) only contains the CE loss related to emotion labels, while the few-shot learning based transfer learning (FSTL) also takes into account the relationship between categories, as shown in the 'metric loss' in Figure 2.

We employ the framework of the siamese neural network (SNN) for few-shot learning process, which uses the same weights while working on a pair of input vectors to compute comparable output vectors. The SNN is trained by creating positive and negative pairs of inputs, where positive pairs belong to the same class and negative pairs belong to different classes. The network uses a contrastive loss to minimize the distance between similar inputs and maximize the distance between dissimilar inputs:

$$L_{FSTL}(\mathbf{W}) = \sum_{c} \sum_{\mathbf{x_i},\mathbf{x_j} \in \mathcal{X}'_c} d\left(\mathbf{F}\left(\mathbf{x_i}\right), \mathbf{F}\left(\mathbf{x_j}\right)\right) - \kappa \sum_{c \neq c'} \sum_{\substack{\mathbf{x_i} \in \mathcal{X}_c \\ \mathbf{x_j} \in \mathcal{X}_{c'}}} d\left(\mathbf{F}\left(\mathbf{x_i}\right), \mathbf{F}\left(\mathbf{x_j}\right)\right) \tag{1}$$

where $\mathcal{X}_c$ is the set of data belonging to class c, $\mathcal{X}_{c'}$ is the set of data belonging to class c' different than c, d is the Euclidean distance between sample pair embeddings and κ determines the trade-off between penalizing dissimilarity between samples belonging to the same class against similarity between samples belonging to different classes.

B. Few-shot learning guided by emotion distance

Few-shot learning-based transfer learning in cross-corpus SER makes good use of the limited data in the target corpus and the relationship between different emotional categories, but this relationship is only a coarse-grained distinction. Inspired by the distribution of emotion categories in the valence-arousal continuous emotion space, this study proposes few-shot learning based on a fine-grained emotion relationship metric.

As shown in Figure 1, in the continuous emotion space, the {happy, sad} sample pair has the farthest Euclidean distance. In addition, in the learning of speech emotion representation, the representation ability of emotion representation in the arousal dimension is stronger than that of the valence dimension, so when we consider the distance between emotion categories, the distance in the valence dimension is compressed, which leads to the emotion distance between {anger, sadness} sample pair being as far as that between {happy, sad} sample pair. These two types of emotion sample pairs are considered to be pure negative (PN) pairs. For pairs of samples with the same emotion category, we believe that the distance between them should be as close as possible and they are pure positive (PP) pairs. For other emotion sample pairs, such as {happy, neutral} and {happy, angry}, we consider the distance between them to be at an intermediate level, not as far away as PN pairs, and we call them medium negative (MN) sample pairs.

Following the above discussion of emotion distance, we construct a fine-grained metric loss function for few-shot learning guided by emotion distance (EMOFSL):

$$\begin{aligned} L_{EMOFSL}(\mathbf{W}) = & \sum_{c} \sum_{\{\mathbf{x_i},\mathbf{x_j}\} \in \mathbf{PP}} d\left(\mathbf{F}\left(\mathbf{x_i}\right), \mathbf{F}\left(\mathbf{x_j}\right)\right) + \\ & \kappa_1 \sum_{c \neq c'} \sum_{\{\mathbf{x_i},\mathbf{x_j}\} \in \mathbf{PN}} max(\alpha - d\left(\mathbf{F}\left(\mathbf{x_i}\right), \mathbf{F}\left(\mathbf{x_j}\right)\right), 0) + \\ & \kappa_2 \sum_{c \neq c'} \sum_{\{\mathbf{x_i},\mathbf{x_j}\} \in \mathbf{MN}} max(\beta - d\left(\mathbf{F}\left(\mathbf{x_i}\right), \mathbf{F}\left(\mathbf{x_j}\right)\right), 0) \end{aligned} \tag{2}$$

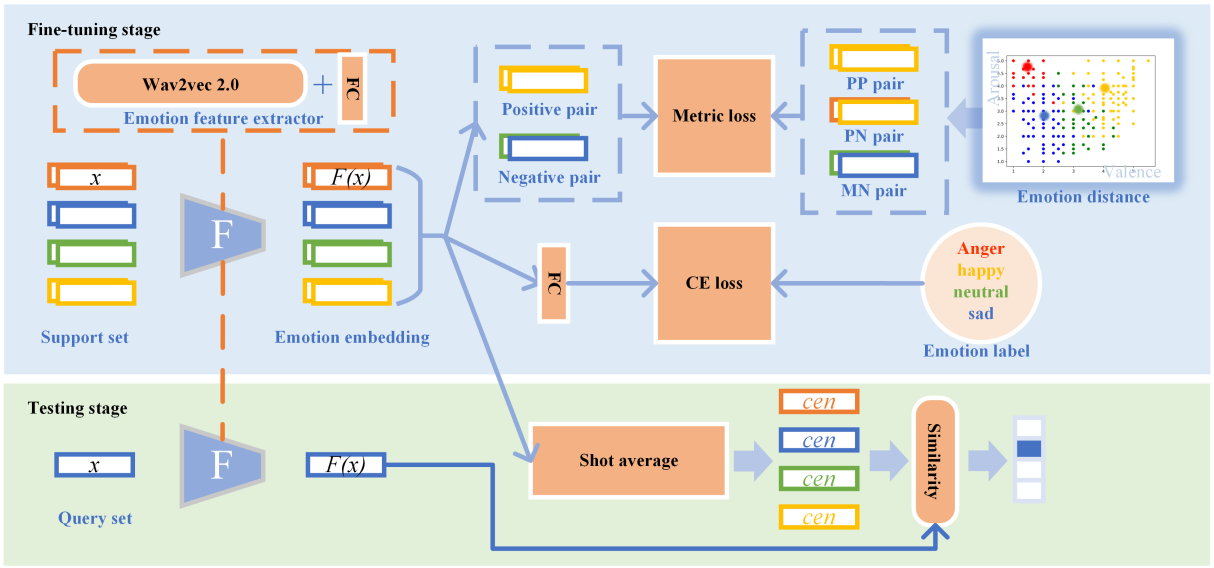


Fig. 2. The fine-tuning and testing stage of the basic and emotion distance-guided few-shot learning

where $\alpha > \beta$ and represent the restricted distance margin of the PN pairs and the MN pairs, respectively, and κ_1 and κ_2 determine the trade-off among different parts of the loss.

III. Experimental Results and Discussions

A. Experimental setup

In this section, we evaluate the proposed framework for cross-corpus SER. Three most common emotion corpora containing English speech, including IEMOCAP (I) [21], RAVDESS (R) [22], and MSP-IMPROV (M) [23], are employed in our experiments. Two of the above corpora are randomly selected as the source and target corpora respectively, and six groups of cross-corpus SER tasks (source corpus-target corpus) I-R, I-M, R-M, R-I, M-I, M-R, are conducted. We select four common emotional categories, i.e., anger, neutral, happy, and sad, in our experiments.

In the transfer learning process, cross-validation (CV) was conducted with different settings for each target corpus to ensure the reliability of the experimental results. For IEMOCAP, we split the dataset by leaving one session out as the test set, and the remaining four sessions were used for training. For RAVDESS, we split the dataset into six equal subsets according to the speaker ID, and chose one subset as test set while the rest subsets as training data. For MSP, 6-fold leave-one-speaker-out CV was conducted according to the session division.

For few-shot learning based transfer learning, 5 random fixed samples of each emotion in the training set of the target corpus are selected for fine-tuning, and all of the samples in the testing set are used for testing. The training set and the testing set are obtained from the cross-validation settings.

We compared the proposed few-shot learning method guided by emotion distance (EMOFSL) with basic transfer learning (BTL) and few-shot learning based transfer learning (FSTL) in terms of unweighted accuracy assessing the SER performance.

For emotional features extraction, we adopt the Wav2Vec 2.0 model that extracts unsupervised speech features by a multi-layer convolutional neural network taking raw speech signals as the input. The output embedding dimension of Wav2vec 2.0 is 768. Then follows a fully connected layer whose input and output dimensions are 768 and 256 respectively.

The common experimental hyperparameters in the finetuning stage are listed in Table I. Note that the finetuning used an exponentially decaying learning rate, which can be calculated as follows:

$$lr = lr_{init} * (r_{\text{decay}})^{\frac{Step_{\text{curr}}}{Step_{\text{decay}}}},$$

where lr_{init} denotes the initial learning rate, $Step_{\text{curr}}$ is the current number of training steps, and $Step_{\text{decay}}$ is the decay period.

TABLE I
Common Hyperparameters in the Finetuning Stage

Hyperparameters	Value
initial learning rate	0.0001
decay rate of learning rate	0.98
max speech length	10 (s)
max epochs	10
precision	32
optimizer	Adam
α	1
β	0.6
κ	1
κ_1	0.6
κ_2	0.3

B. Results and discussions

Table II shows the cross-corpus SER performance of the compared methods, in which 5 samples from each of the 4

TABLE II
CROSS-CORPUS SER PERFORMANCE OF 4-WAY 5-SHOT FEW-SHOT LEARNING

UAR(source-target)	BTL	FSTL	EMOFSL
I-R	49.17	48.39	51.43
I-M	47.76	48.4	48.48
M-I	42.77	46.55	46.6
M-R	47.5	48.39	49.76
R-I	44.58	44.52	45.19
R-M	37.15	37.04	37.93



Fig. 3. Confusion matrices of the proposed EMOFSL method in six cross-corpus tasks. (R: RAVDESS, M: MSP-IMPROV, I: IEMOCAP) The horizontal axis represents the true labels, and the vertical axis represents the predicted labels.

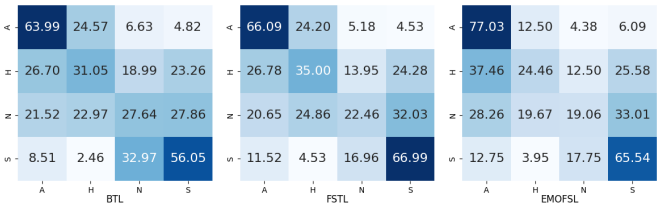


Fig. 4. Confusion matrices of the compared methods in the cross-corpus task when the source and target corpora are RAVDESS and IEMOCAP, respectively. The horizontal axis represents the true labels, and the vertical axis represents the predicted labels.

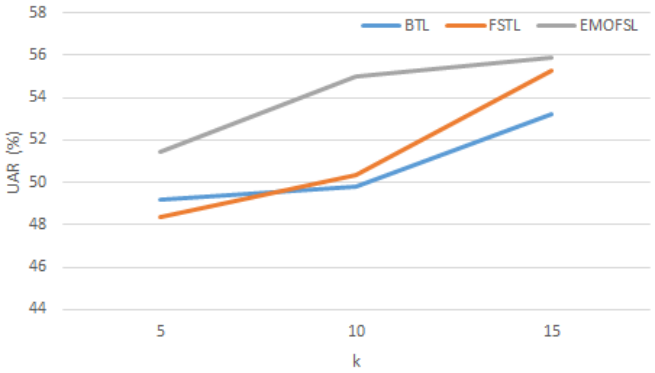


Fig. 5. Performance comparison of BTL, FSTL, and EMOFSL when the size of support set k is 5, 10, and 15, respectively. The source corpus of the cross-corpus task is RAVDESS, and the target corpus is IEMOCAP.

emotions are selected to form the support set in the fine-tuning process. Few-shot learning method guided by emotion distance (EMOFSL) is compared with basic transfer learning (BTL) and few-shot learning based transfer learning (FSTL) in terms of unweighted accuracy rate (UAR).

It can be observed that our proposed EMOFSL performs better than the other two baseline methods, which proves the effectiveness of emotion category distance prior knowledge for cross-corpus SER task. It can be also observed that in most cross-corpus tasks, the few-shot learning-based transfer learning performs better than the basic transfer learning, which verifies that metric loss obtained from the positive and negative sample pairs helps the learning of speech emotion on the target corpus.

Figure 3 presents the confusion matrices of the proposed EMOFSL method for six cross-corpus tasks. One can observe that the recognition results of anger and sad in all tasks are generally better than other emotions, indicating that these two emotions in speech are easier to be recognized, which is partly related to the prior knowledge of the emotional distance we added to the metric loss.

Figure 4 shows the confusion matrix comparison between the proposed method and the baseline methods. We can see that with metric loss representing the distance relationship of different emotion categories, FSTL and EMOFSL achieve better recognition on anger and sad.

Figure 3 and 4 also indicate that happy is often misidentified as anger and neutral is often misidentified as sad, which is consistent with our previous understanding in the continuous emotion space that arousal is easier to distinguish than valence in speech emotion expression.

Figure 5 shows the performance comparison of BTL, FSTL, and EMOFSL when the size of support set k is 5, 10, and 15, respectively. The source corpus of the cross-corpus task is RAVDESS, and the target corpus is IEMOCAP. One can observe that: 1) As the number of training samples in the support set increases, the SER performance of the three methods improves accordingly. 2) In the case of different support set sizes, the performance of EMOFSL is better than the other comparison methods. 3) The performance of FSTL is better than that of BTL in most cases.

There have been several papers on the cross-corpus SER, but most of them are not suitable for comparison because of different configurations for the databases and the utilization of the target domain data for validation. Among the approaches, [5] reached a 43.8% UAR when the source and target corpora are IEMOCAP and MSP-IMPROV, respectively. Our proposed EMOFSL method achieves a higher UAR of 48.48% using fewer shots. The two main reasons for the better performance of the proposed EMOFSL are: 1) A more advanced speech

feature extractor is used. 2) The prior knowledge of emotion distance is introduced as a guide in the cross-corpus speech emotion representation learning.

IV. Conclusion

Cross-corpus speech emotion recognition faces the problem of sparse target set data. To solve this problem, this study introduces the prior knowledge of emotion distance to guide the few-shot learning process of cross-corpus emotion recognition, thereby alleviating the problem of insufficient emotional information to a certain extent. Experimental results show that the proposed method performs better than traditional few-shot learning verifying the efficiency of the emotion distance prior knowledge. In addition, this paper introduces a self-supervised pre-training model with stronger speech representation ability than traditional features as a feature extractor, which also improves the performance of cross-corpus speech emotion recognition.

Acknowledgment

The work is supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2021ZD0114303), and Grant No. 2020KB0AC01 from the Scientific Project of Zhejiang Lab.

References

[1] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[2] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition." in *Interspeech*, 2019, pp. 1656–1660.

[3] S. Li, P. Song, L. Ji, Y. Jin, and W. Zheng, "A generalized subspace distribution adaptation framework for cross-corpus speech emotion recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] K. Feng and T. Chaspari, "Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation," *IEEE Transactions on Affective Computing*, 2021.

[5] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[6] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[7] C. E. Izard, "The many meanings/aspects of emotion: Definitions, functions, activation, and regulation," *Emotion Review*, vol. 2, no. 4, pp. 363–370, 2010.

[8] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[9] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[11] L. Feldman Barrett and J. A. Russell, "Independence and bipolarity in the structure of current affect." *Journal of personality and social psychology*, vol. 74, no. 4, p. 967, 1998.

[12] R. Cai, K. Guo, B. Xu, X. Yang, and Z. Zhang, "Meta multi-task learning for speech emotion recognition." in *INTERSPEECH*, 2020, pp. 3336–3340.

[13] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, and D. Jiang, "Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models," *Knowledge-Based Systems*, vol. 235, p. 107598, 2022.

[14] Z. Huijuan, Y. Ning, and W. Ruchuan, "Coarse-to-fine speech emotion recognition based on multi-task learning," *Journal of Signal Processing Systems*, vol. 93, pp. 299–308, 2021.

[15] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on affective computing*, vol. 8, no. 1, pp. 3–14, 2015.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[17] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[20] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[21] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[22] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[23] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

Speech Emotion Recognition by Late Fusion of Linguistic and Acoustic Features using Deep Learning Models

Kiyohide Sato*, Keita Kishi* and Tetsuo Kosaka*
* Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
E-mail: t223797d@st.yamagata-u.ac.jp Tel: +81-238-263368

***Abstract* — In this study, we investigated speech emotion recognition using both linguistic and acoustic features contained in emotional speech. Speech recognition is necessary to extract linguistic features in this system. In this case, speech recognition accuracy affects emotion recognition accuracy; therefore, it is important to improve the accuracy. In particular, we use a language model adapted to a large amount of Twitter data to model colloquial expressions that frequently appear in emotional speech. Consequently, a word recognition accuracy of 82.2% was obtained. Finally, we obtained 51.5% of emotion recognition accuracy from linguistic features, 77.25% from acoustic features, and 82.25% from the fusion of both. Furthermore, according to the analysis of the recognition results, the error tendency of linguistic and acoustic features differed, and we found that complementary effects could be obtained when fusing both features. In this corpus, a human was reported to have an emotion recognition rate of 75.5%; we surpassed that performance. This demonstrated the effectiveness of the proposed method for speech emotion recognition.**

I. INTRODUCTION

As with human interaction, the use of nonverbal information, such as emotions, as a means of communication plays an important role in the smoothness of spoken dialogue systems. Therefore, the importance of emotion recognition has recently increased.

Considerable research has been conducted using acoustic features in this field. Linguistic features obtained from speech can also be used for emotion recognition. Therefore, methods using both linguistic and acoustic features have been studied [1-3]. Speech recognition for emotional speech (emotion speech recognition) is challenging and is necessary for linguistic features extraction. Previously, speech recognition performance was considered insufficient [3].

One reason for the difficulty in emotion speech recognition is the large variation in acoustic features. Furthermore, it is difficult to construct a language model for emotion speech. Colloquial expressions are frequently observed in emotional speech. However, few language corpora include colloquial expressions. For this reason, many studies have used only acoustic features [4-6], even when linguistic features are used, transcripts are often used instead of speech recognition results [1,2,7]. Some studies have attempted to use speech recognition but with poor results [3].

Therefore, we first attempted to improve the speech recognition performance for emotional speech. Consequently, it has been shown that a language model for emotional speech recognition can be effectively created using a large amount of tweet data [8], where the combination of acoustic and language model adaptations improved the recognition rate from 63.9% to 82.2%.

In this paper, we extract linguistic features using the above speech recognition results, and try to recognize emotions using both linguistic and acoustic features. In particular, we compare the recognition results of linguistic and acoustic features and clarify the effect that can be obtained in emotion recognition when they are used together.

In recent years, many open-source frameworks, such as end-to-end models, have been released, and by using them, it has become possible to test speech recognition easily. In particular, methods using model structures, such as wav2vec [21] and huBERT [22], can yield highly accurate recognition results [9-11]. However, when these models are used, the recognition process becomes a black box. Therefore, when linguistic and acoustic features are combined, the role each plays is unclear. In addition, there are reports on the effect of ASR on emotion recognition in [19] and [20], however, the influence of each emotion on linguistic features and acoustic features is not clear.

In our system, emotion speech recognition is performed using

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

uttered speech and linguistic features are extracted from the speech recognition results to perform emotion recognition. In addition, acoustic information is extracted from speech and used for emotion recognition based on acoustic features. Subsequently, the emotion recognition results of the linguistic and acoustic features are fused to obtain the final emotion recognition results. Owing to this procedure, it is possible to clarify why fusion is effective by comparing the recognition results for each feature.

In [12], linguistic information was obtained through speech recognition, and a method for emotion recognition using both acoustic and linguistic information was proposed. In this study, emotion recognition was performed by late fusion using the weighted addition of three types of recognition outputs: linguistic features, acoustic feature statistics from whole utterances, and acoustic features by low-level-descriptor (LLD). However, fusing by the weights had a practical problem that it required data to be recognized for determining the weights, because those were task dependent. In addition, the recognition performance using linguistic features was insufficient.

In this study, the following methods are used to improve the performance. We use BERT [13] for emotion recognition of linguistic features and a deep learning model for late fusion. We demonstrate that a recognition performance equal to or higher than that of humans can be obtained using the above methods.

II. Emotion Recognition by Late Fusion

Conventionally, the LLD is extracted as a time series representing emotional features, and various acoustic statistics representing the features of an entire utterance are calculated. However, because these features represent the characteristics of the entire utterance, it is not possible to extract local emotional fluctuations. Therefore, in the proposed method, emotion recognition based on acoustic features uses both the recognition result from LLD, which is a time-series feature, and the recognition result from the statistics of the entire utterance [12].

Recognition based on time-series features uses bidirectional LSTM (BiLSTM) and bidirectional GRU (BiGRU) as classifiers, whereas recognition based on statistics uses a DNN as a classifier. In emotion recognition based on linguistic features, speech recognition is first performed using acoustic and linguistic models, and emotion is estimated by BERT using the speech recognition results. Finally, emotion recognition was performed by fusion of the recognition results of acoustic features and language features using DNN, and four emotions, "anger (ang)," "joy (joy)," "sadness (sad)," and "neutral (neu)," were recognized. Figure 1 shows a schematic diagram of the proposed emotion recognition system. The left side of the figure shows the recognition unit based on linguistic features, and the right side shows the recognition unit based on acoustic features.

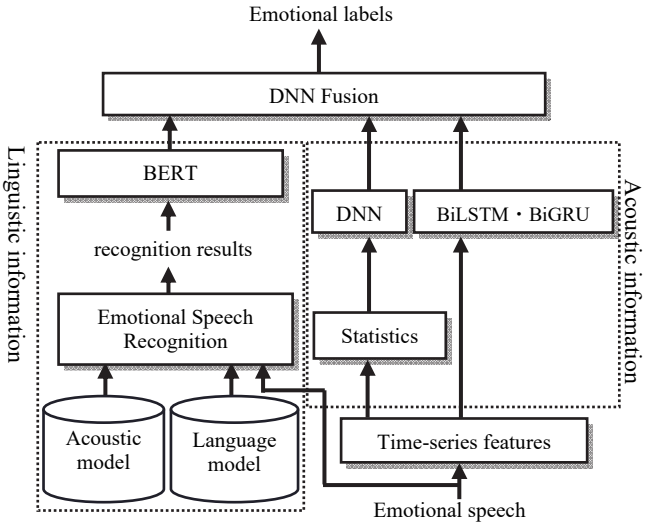


Fig. 1 Block diagram of the proposed system

III. Emotion Recognition Method

A. Emotion recognition using linguistic features

Speech recognition was first performed for emotion recognition based on linguistic features. The speech recognition method presented in [8] was used. A two-pass decoder was used for decoding. In the first pass, word graphs are generated from beam searches using triphones and bigrams, and in the second pass, recognition results are obtained by the trigram rescores of the generated word graphs. The DMM-HMM was used as the acoustic model. In this recognition system, we first generate a baseline acoustic model (AM) and a language model (LM) for emotional speech using a Corpus of Spoken Japanese as training data. Subsequently, model adaptations of AM and LM were performed. In AM adaptation, emotional speech data are used as adaptation data, and adaptation is performed using the back-propagation method. The number of epochs was automatically determined by dividing the adaptation data into a 9:1 ratio and using early stopping by cross-validation using the frame recognition rate. The weighted mixture method was used for the LM adaptation [14]. For LM adaptation, a large amount of tweet data of 25.86 million words collected from Twitter was used as the adaptation data, and the mixed n-gram method was used [14]. In this approach, we generated

the mixed n-gram by combining the n-gram count derived from the adaptation data (n_i^{adapt}) and that derived from baseline data (n_i^{base}). N^{adapt} and N^{base} denote the total n-gram counts of the adaptation and baseline data, respectively. We estimated the occurrence probability of word w_i in the adapted LM by the following formula:

$$P(w_i) = \frac{w \cdot n_i^{adapt} + n_i^{base}}{w \cdot N^{adapt} + N^{base}} \quad (1)$$

where w represents a weighting factor designed to address the imbalance between adaptation and baseline data. The baseline word recognition accuracy was 63.9%, but the recognition accuracy after AM and LM adaptation was 82.2%.

Using these speech recognition results, we performed emotion recognition based on linguistic features. In this study, we used BERT as the discriminator. A pretrained model trained with a large amount of unlabeled data and a fine-tuning method using a small amount of labeled data for the target task were used. For the pretrained model, we used a model trained using Japanese Wikipedia texts [15]. Using this pretrained model, fine-tuning was performed using emotion-labeled data to train an emotion classifier for the four emotions. A total of 1,963 sentences from the Japanese Twitter-based emotional speech (JTES) were used for fine-tuning, 10% of which were used as validation data.

B. Emotion recognition using acoustic features

As the acoustic features for emotion recognition, the LLD was extracted as a time series representing emotional features, and various statistics representing the features of the entire utterance were calculated from the LLD. First, the input speech was divided into frames using a window function, and LLD, such as the fundamental frequency and MFCCs, were extracted. Here, the large openSMILE emotion feature set [16] was used as the LLD, and 168-dimensional feature vectors were used per frame (10 ms). Next, 39 types of statistics were calculated from the 168-dimensional LLD, and 6,552-dimensional features ($= 39 \times 168$) were input into the DNN for statistical features. The structure of the DNN consists of three hidden layers of 2,048 units. The emotion recognition rate was only 69.25% using the feature parameters by the statistics mentioned above.

For the LLD, which is a time-series feature, it is possible to use a model suitable for changes over time. Therefore, in this study, we used an LSTM model as the discriminator. Table I presents the emotion recognition results obtained by examining the combinations of LSTMs, GRUs, bidirectional LSTMs, and bidirectional GRUs. Based on the results, we used a combination of bidirectional LSTM and bidirectional GRU, which achieved the highest accuracy (75.75%).

The network structure in this model consists of one BiLSTM layer, one BiGRU layer, and three fully connected layers (fc1–fc3) from the input side. The BiLSTM and BiGRU had 100 hidden layers. The fully connected layers fc1, fc2, and fc3 contained 512, 256, and 64 units, respectively. The learning rate was set to 5.0×10^{-4}. Adam was used as the optimization method, and cross-entropy was used as a loss function.

Table I

Emotion recognition results by LLD (%)

Conditions	Accuracy
LSTM×2 layers	72.50
BiLSTM×2 layers	73.00
GRU×2 layers	74.25
BiGRU×2 layers	74.50
BiLSTM+BiGRU	75.75

C. Late fusion

Next, late-stage fusion was performed using a DNN-based fusion method for the linguistic and acoustic features. The final result is determined using the fusion method of the DNN using the recognition output of linguistic features based on emotional speech recognition results, statistics for the whole utterance, and time-series features of acoustic characteristics. Using this method, it is possible to automatically estimate emotions without using manually determined weights. The network structure of the DNN consists of 128, 32, and 32 units in the fully connected layers fc1, fc2, and fc3, respectively. The learning rate was 4.0×10^{-5}, the optimization method was Adam, and the loss function was cross-entropy.

For comparison with the DNN fusion, late-stage fusion was performed using the weighted fusion method for linguistic and acoustic features according to the following formula:

$$emo = \alpha Le + (1-\alpha)Ae \quad (2)$$

where emo referred to the final emotion score, Le is the emotion likelihood of a linguistic feature, and Ae is the emotion likelihood of acoustic features. α is a coefficient

with the weight that takes values from 0 to 1. Furthermore, we combined LLD and statistics by the same method.

IV. Emotional Speech Corpus

In this study, we used the JTES database as the speech corpus [17]. In this corpus, colloquial sentences containing emotional expressions were selected from tweets on *Twitter*, considering the balance of phonemes and prosody. A total of 20,000 utterances were prepared for 100 speakers (50 males and 50 females) and 50 sentences for each of the four emotions, ``anger,'' ``joy,'' ``sadness,'' and ``neutral.'' The speaker was instructed to ``convey the intended emotion to the robot (machine)". Anger is divided into ``hot anger" and ``cold anger." They are instructed to be conscious of ``hot anger" in some cases.

V. Emotion Recognition Experiment

A. Experimental conditions

A total of 14,400 utterances (40 sentences × 4 emotions × (45 male speakers + 45 female speakers)) in JTES were used as training data, and 400 utterances (10 sentences × 4 emotions × (5 male speakers + 5 female speakers)) were used as evaluation data. In the late fusion of the weighted fusion method and the fusion method using the DNN described in Section III-C. A total of 12 dimensions of the 4-dimensional LLD, 4-dimensional statistics, and 4-dimensional BERT output were used as the input of the late fusion. We compared the following three types of emotion recognition with different linguistic features extraction methods.

Baseline ASR we use the automatic speech recognizer (ASR) result of 63.9% word recognition rate before adaptation.

Adapted ASR we use the ASR result of 82.2% word recognition rate after adaptation.

Oracle we used the correct text assuming a word recognition rate of 100% for comparison.

In the late fusion of linguistic and acoustic features using the weighted fusion method, the weight α in Eq. (1) was experimentally set to 0.35 for the baseline ASR, 0.25 for adapted ASR, and 0.35 for oracle.

B. Experimental results

For emotion recognition in acoustic features, the recognition rate of late fusion between LLD and statistics using the weighted fusion method was 77.25%. For emotion recognition in linguistic features, the recognition results were 45.5% for baseline ASR, 51.5% for adapted ASR, and 62.5% for Oracle. Figure 2 shows the results of the late fusion of the acoustic and language features using the weighted fusion method.

The emotion recognition rate under optimal conditions was 80.50% with α of 0.35 in baseline ASR, 80.25% with α of 0.25 in adapted ASR, and 83.00% with α of 0.35 in Oracle. Compared with the emotion recognition rate of 77.25% based on acoustic features only, the recognition performance of various linguistic features is low; however, the performance improves when the fusion method is used.

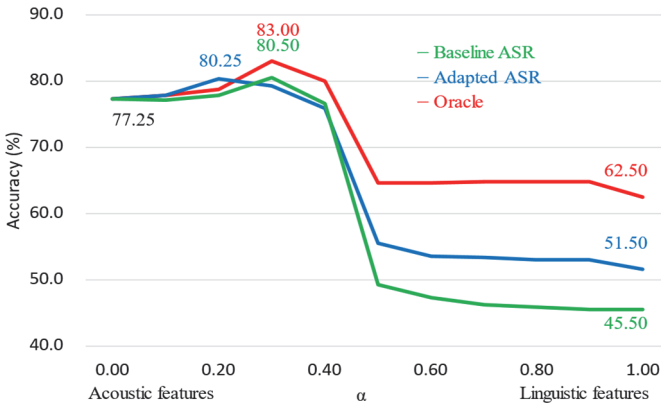


Fig.2 Relationship between weight and emotion recognition accuracy in weighted late fusion

Next, late fusion of linguistic and acoustic features by the fusion method using a DNN was performed. Table II summarizes the emotion recognition results for each fusion condition. In the comparison between the three kinds of linguistic features, Oracle showed the highest recognition rate, but the difference with the other two was not large. Thus, even if recognition accuracy based on linguistic features is somewhat low, it is useful for improving the performance of emotion recognition. In addition, the DNN fusion method achieved a higher performance than the weighted fusion method, as shown in Figure 2.

Next, we present a confusion matrix for the recognition results. Table III shows the results of the acoustic features using the weighted fusion method, Table IV shows the results of the two types of linguistic features only, and Table V shows the results of DNN fusion. In these tables, the vertical axis represents the correct answer, and the horizontal axis represents the predicted result.

The confusion between ‘ang’ and ‘joy’ in the case of acoustic features only in Table III is considered to be due to acoustic similarities such as the strength of the utterance and the fluctuation of the fundamental frequency being larger than that of other emotions.

Table II

Speech recognition and emotion recognition results for each fusion condition (%)

Types＼Conditions	Baseline ASR	Adapted ASR	Oracle
Word recognition rate	63.9	82.2	100
Emotion recognition rate by linguistic features	45.5	51.5	62.5
Emotion recognition rate by acoustic features		77.25	
Emotion recognition rate by DNN fusion	82.00	82.25	85.50

Table III

Confusion matrix for acoustic features only (%). (average recognition rate: 77.25%)

	ang	joy	neu	sad
ang	85	14	1	0
joy	38	57	2	3
neu	1	6	85	8
sad	1	2	15	82

Table IV

Confusion matrix for linguistic features only (%).

	Baseline ASR (average: 45.5%)				Adapted ASR (average: 51.5%)			
	ang	joy	neu	sad	ang	joy	neu	sad
ang	41	11	28	20	52	4	20	24
joy	33	36	3	28	12	55	6	27
neu	15	7	58	20	10	23	46	21
sad	22	11	20	47	20	25	2	53

Table V

Confusion matrix for acoustic and linguistic features (%).

	Baseline ASR (average: 82.0%)				Adapted ASR (average: 82.5%)			
	ang	joy	neu	sad	ang	joy	neu	sad
ang	86	10	4	0	86	8	6	0
joy	16	78	2	4	23	72	2	3
neu	1	5	87	7	2	3	90	5
sad	3	3	17	77	1	4	14	81

When only linguistic features were used (Table IV), there was a tendency to confuse ‘neu’ with ‘joy’ or ‘sad’, and ‘sad’ with ‘ang’ or ‘joy’. These results indicate that it is difficult to discriminate between the emotions ‘neu’ and ‘sad’ using only linguistic features.

Looking at the results of the late fusion of acoustic and language features shown in Table V, it can be confirmed that these recognition errors are greatly reduced, and the accuracy of emotion recognition is improved. Thus, it was shown that the proposed method is effective because it obtains the complementary effects of acoustic and linguistic features.

VI. Conclusions

In this paper, we proposed a speech emotion recognition method based on late fusion using a DNN fusion method for linguistic and acoustic features. The results of the experiments confirmed that highly accurate recognition is possible by compensating for each part of the low recognition rate through the fusion of acoustic and linguistic features. Furthermore, it was confirmed that acoustic and linguistic features could complement each other without being affected by the superiority or inferiority of the emotional speech recognition results when the DNN fusion method was used. In addition, the effectiveness of this system was demonstrated by obtaining a recognition accuracy that surpassed that of the weighted fusion method.

It has been reported that the accuracy of human emotion recognition in the JTES corpus is 75.5% (however, this evaluation focused only on acoustic features) [18]. Because we have obtained the recognition performance sufficient for practical use in the closed task using this corpus, we want to consider methods that target open tasks in the future. In addition, we would like to compare the proposed method with other methods such as wav2vec and HuBERT as ASR.

Acknowledgment

This study was supported in part by a Grant-in-Aid for Scientific Research (KAKENHI 22K12087) from the Japan

Society for the Promotion of Science. We thank Dr. Takashi Nose of Tohoku University for providing the Japanese Twitter-based emotional speech (JTES). Ms. Misaki Sakurai (NEC Nets SI) and Mr. Kazuki Susuta (Advanced Media) participated in the experiment.

REFERENCES

[1] Q. Jin, C. Li, S. Chen and H. Wu, “Speech emotion recognition with acoustic and lexical features,” *Proc. of ICASSP2015*, 2015, pp. 4749-4753.

[2] R. Nagase, T. Fukumori and Y. Yamashita, “Speech Emotion Recognition with Fusion of Acoustic and Linguistic-Feature-Based Decisions,” *Proc. of APSIPA ASC2021*, 2021, pp. 725-730.

[3] V. Rozgic et al., “Emotion recognition using acoustic and lexical features,” *Proc. of INTERSPEECH2012*, 2012, pp. 366-369.

[4] X. Li, T. Guo, X. Hu, X. Xu, J. Dang, M. Akagi, “Hierarchical Prosody Analysis Improves Categorical and Dimensional Emotion Recognition,” *Proc. of APSIPA ASC2021*, 2021, pp. 700-704.

[5] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, B. Schuller, “Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition,” *IEEE Access*, vol. 7, pp. 97515-97525, 2019.

[6] K. Zvarevashe, O. Olugbara, “Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition,” *Algorithms*, vol. 13, no. 70, p.1-24, 2020.

[7] G. Mendels, S. Levitan, K. Lee, J. Hirschberg, “Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection,” *Proc. of INTERSPEECH2017*, 2017, pp. 1472-1476.

[8] K. Saeki, M. Kato, T. Kosaka, “Language Model Adaptation for Emotional Speech Recognition using Tweet data,” *Proc. of APSIPA ASC2020*, 2020, pp. 371 - 375.

[9] X. Cai, J. Yuan, R. Zheng, L. Huang, K. Church, “Speech Emotion Recognition with Multi-task Learning,” *Proc. of INTERSPEECH2021*, 2021, pp. 4508 - 4512.

[10] L. Pepino, P. Riera, L. Ferrer, “Emotion Recognition from Speech UsingWav2vec 2.0 Embeddings,” *Proc. of INTERSPEECH2021*, 2021, pp. 3400 - 3404.

[11] I. Gat, H. Aronowitz, W. Zhu, E. Morais, R. Hoory, “Speaker Normalization for Self-supervised Speech Emotion Recognition,” *Proc. of ICASSP2022*, 2022, pp. 7342 - 7346.

[12] M. Sakurai and T. Kosaka, “Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results,” *Proc. of GCCE2021*, 2021, pp. 889-892.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proc. of NAACL2019*, 2019, pp. 4171 - 4186.

[14] A. Ito and M. Kohda, “Evaluation of task adaptation using N-gram count mixture,” *IEICE Trans.* volJ83-D-II, No. 11, pp. 2418-2427, 2000 (in Japanese).

[15] Tohoku University Inui-Suzuki Laboratory, Pretrained Japanese BERT models, "https://github.com/cl-tohoku/bert-japanese".

[16] F. Eyben et al., "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc. of the 9th ACM international conference on Multimedia*, 2010, pp.1459-1462.

[17] E. Takeishi, T. Nose, Y. Chiba, A. Ito, “Construction and analysis of phonetically and prosodically balanced emotional speech database,” *Proc. of O-COCOSDA2016*, 2016, pp. 16–21.

[18] Y. Chiba, T. Nose, A. Ito, “Multi-stream Attention-based BLSTM with Feature Segmentation for Speech Emotion Recognition,” *Proc. of INTERSPEECH2020*, 2020, pp. 3301-3305.

[19] F. Han, S. Ueno, T. Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," *Proc. of INTERSPEECH2020*, 2020, pp. 501-505.

[20] S. Bhosale, R. Chakraborty, and S. K. Kopparapu. "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," *Proc. of ICASSP2020*, 2020, pp. 7184-7188.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0:A Framework for Self-Supervised Learning of Speech Representations,” *arXiv:2006.11477*, 2020.

[22] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *arXiv:2106.07447*, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multilingual, Cross-lingual, and Monolingual Speech Emotion Recognition on EmoFilm Dataset

Bagus Tris Atmaja* and Akira Sasou*
* National Institute of Advanced Industrial Science and Technology, Japan
E-mail: b-atmaja@aist.go.jp, a-sasou@aist.go.jp

***Abstract*—Research on speech emotion recognition has been actively conducted; most are in monolingual settings. Considering that emotion expressed in speech is universal, it is noteworthy to conduct multilingual emotion recognition across different cultures. This paper contributes to evaluating multilingual, cross-lingual, and monolingual automatic speech emotion recognition (SER) on the same EmoFilm dataset. We first evaluated these three scenarios on a fixed training/test split. For multilingual emotion recognition, we then expanded the evaluation with cross-validation. The results show that the multilingual SER gained the highest performance with 74.86% of balanced accuracy for five categorical emotions, followed by 72.17% and 58.03% for monolingual and cross-lingual evaluations. We reduced the number of training samples to observe its impact and found that the monolingual setting is superior among others on the same number of samples. The results of this study could suggest the potential use of multilingual SER over cross-lingual and monolingual SER in future speech technologies.**

***Index Terms*—speech emotion recognition, pre-trained model, multilingual emotion recognition, EmoFilm**

I. Introduction

There is some universality in emotion perception by humans, particularly from vocal cues. A study by Scherer et al. [1] reported that the accuracy of vocal expression across cultures was 66%. The study was conducted in nine countries in Europe, the United States, and Asia. The vocal emotion is acted in the voices of anger, sadness, fear, joy, and neutrality. Another study [2], a meta-analysis study by Elfenbein and Ambaday, revealed that emotions were universally recognized at better-than-chance levels. The study also examined studies that included both a majority and minority group in the design. Majority group members were poorer at judging minority group members than the reverse.

Scherer's study [1] suggested that language differences between encoders (speakers) and decoders (listeners) may play a major role in recognizing emotion through voices. The difference is significant for Germany and Indonesia but marginally different for the German and Indo-European languages. In Elfenbein and Ambady [2], the accuracy of emotion recognition was higher when emotions were both expressed and recognized by members of the same national, ethnic, or regional group. This result suggests an in-group advantage; members in the same culture have higher recognition than members from different cultures. Both studies suggest that although emotion recognition among humans is universal, there are differences between cultures.

The commonness and differences in human perceptual emotion recognition have been studied (in [3], Figs. 6-7) for building automatic emotion recognition by computers. As in human perception of emotion through voice, computers should be able to distinguish human emotion regardless of the culture. The authors found that the direction of neutral speech to other emotional states is similar for Japanese and Chinese. The difference includes the neutral position, which is slightly different among Japanese, Chinese, and German by Japanese, Chinese, and Vietnamese listeners. Since there is commonness in the perception of emotion from speech, it is possible to build a system that can recognize emotion from speech multiculturally or cross-culturally, operationalized by different countries through languages.

The first study for building a cross-cultural automatic emotion recognition dataset for machines perhaps is the work of Scherer in 2000 [4]. He claimed to build the first large-scale dataset to obtain empirical data on cross-cultural emotion recognition from voices in nine countries on three different continents. The results, still evaluated by human listeners, showed recognition rates that were much better than chance accuracy. Scherer also found that in affective speech, voices other than vocal bursts, segmental and suprasegmental aspects of language affect the encoding and decoding of emotion.

The research of cross-cultural and multicultural emotion then grew with its debates. Other researchers [5] claimed that six basic (facial expressions) emotions are not universal. The claim is based on the study among Westeners and Easteners on Ekman's six basic emotions (happy, surprise, fear, disgust, anger, and sad). The Westeners show a unique pattern of facial movement for each basic emotion, while the Easteners do not (but show unique eye movement). On the other hand, Narimatsu et al. [6] have found some commonalities and differences between English and Japanese visual art labels by neural speakers. It can be argued that aside from emotions being expressed differently among cultures, the perceiver across cultures still could recognize the emotion with some cultural differences.

This paper contributes to the evaluation of multilingual, cross-lingual, and monolingual automatic speech emotion recognition (SER) on the same dataset, i.e., EmoFilm Dataset [7]. To the best of our knowledge, this is the first study to evaluate multilingual, cross-lingual, and monolingual SER simultaneously on this dataset. Given the argument that emo-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

tional speech is universal with cultural differences, the performance of all three evaluations (multilingual, cross-lingual, and monolingual) should be comparable under the same test set. The results of this study could suggest the readability of which SER technologies, from these three evaluations, have the potential to be used in future speech technologies.

II. Previous studies

In the previous section, we introduced the classical approach to multilingual emotion recognition. This section discusses recent research on cross-lingual and multilingual speech emotion recognition in chronological order.

Cross-lingual SER

Elbarougy and Akagi [8] in 2013 investigate whether emotions can be estimated cross-lingually. The study proposed three-layer models with semantic primitives in between categorical emotional and acoustic features. To experiment with that proposal, the authors evaluated the Japanese dataset (Fujitsu database) and the German dataset (EmoDB). The results, based on a system with a fuzzy inference system, show small errors in cross-lingual emotion recognition. Both Japanese and German datasets are small and clean (140 and 200 samples); whether the results can be generalized to other languages is questionable.

Neuman and Vu [9] in 2018 conducted experiments with monolingual, cross-lingual, and multilingual SER in English (IEMOCAP dataset) and French (Recola dataset). The best score for French was achieved using a cross-lingual scenario with finetuning on arousal detection. Other dimensions are obtained their best with monolingual (IEMOCAP-valence and Recola-valence) and multilingual (IEMOCAP-arousal). This study is the closest to our work with different datasets and approaches.

Latif et al. [10] in 2019 proposed unsupervised adversarial domain adaptation for cross-lingual SER. The method was evaluated on four datasets: EmoDB, SAVEE (English), EMOVO (Italian), and URDU. They used EmoDB, SAVEE, and EMOVO as source domains and URDU as the target domain and vice versa. The former achieved 65% of UA while the latter achieved 61% of UA (URDU $\rightarrow$ EmoDB). As a comparison, using a multilingual approach achieved 67.3% on URDU and 65.2% on EmoDB (German).

Cai et al. [11] in 2021 proposed a domain adversarial neural network for cross-lingual SER on IEMOCAP and Recola datasets (IEMOCAP $\rightarrow$ Recola and Recola $\rightarrow$ IEMOCAP). Their method outperformed the baseline method but not the monolingual method. The main advantage of their method is no requirement for the target data. No evaluation of multilingual SER is reported, which may obtain better results than the cross-lingual SER.

Multilingual SER

Li and Akagi [12] in 2016 expanded the previous cross-lingual study [8] with a Chinese dataset for multilingual SER. With an SVM as a classifier, the authors evaluated similar three-layer models with the same semantic primitives and acoustic features. The results show that there are small degradations in classification rates from monolingual to multilingual emotion recognition. As in the previous research, the new Casia dataset is also small and clean (198 selected samples), which makes it difficult to draw robust conclusions from the results. The authors then improved their results in 2019 [13] for Japanese and German by 3-4% of UA by combining several acoustic features. The new result, instead, decreases the performance in the Chinese dataset.

Lee [14] in 2019 evaluated two architectures of multitask learning optimized for emotion recognition, three softmax layers and one softmax layer. The first architecture is to predict gender, emotion, and language. The second architecture is to predict language and emotion. The results revealed a better generalization of the second architecture to predict emotion categories with common normalization. The common normalization was performed by normalized features by variances and means from two databases of two languages (English and Japanese).

Zang et al. [15] in 2021 mixed datasets of IEMOCAP, TESS (English), EMA (English), EmoDB, and EMOVO (Italian) and created their training/set partitions based on the newly merged data. The ratio of training:test is set to 4:1. Using wav2vec 2.0 XLSR version (cross-lingual speech representations) with hierarchical grained and feature model (HGFM) for multilingual SER outperformed the baseline with wav2vec 2.0 and HGFM.

Wang et al. [16] in 2022 proposed a multi-gating mechanism with neural architecture search on English (IEMOCAP), German (EmoDB), and French (Cafe) databases. Their proposed methods outperformed the baseline methods for German and French on monolingual evaluations. The result for multilingual evaluations is very similar to the monolingual methods, highlighting the effectiveness of their proposed gating mechanism across different language corpus.

There are several other papers that include "multilingual" terms, but the definition is for the evaluation of several mono languages (e.g., [17], [18]), not a mixed language, nor a multilingual SER model evaluated on mono languages. We exclude these paper since it does not fall in our scope. Furthermore, we provided a fixed split and cross-validation setting in our project repository that other researchers can benchmark in the future.

III. EmoFilm dataset and its partition

The EmoFilm dataset [7] is the focus of this study on evaluating a multilingual speech emotion recognition system, along with its comparison to cross-lingual and monolingual analyses. The emotional speech corpus consists of 1115 samples from English (EN), Italian (IT), and Spanish (SP), extracted from 43 films and 207 speakers. The Italian and Spanish are the dubbed versions of English. The distribution of samples for each language can be seen in Fig. 1, in which Italian has the most portion, followed by English and Spanish.

The dataset includes five categorical emotions: fear, contempt, happiness, anger, and sadness (chance level of 20%).

The average utterance per emotion is 34.3, 41.3, and 35.9 for English, Italian, and Spanish. The distribution of each emotion category is not balanced since the same sentence in the original English is not always considered as emotional as in Italian and Spanish. Table I shows the distribution of each emotion category for each language. Note that the number (including the number in Fig. 1) might be different from the original paper [7]. The counting is based on the latest version (v3) of the EmoFilm dataset available at https://zenodo.org/record/7665999 (accessed on 29 March 2023).

We performed two evaluations for partitioning the dataset into training and test split. Both evaluations are speaker-independent. First, we create a near-balanced fixed split for training and test partition. We evaluated multilingual, cross-lingual, and monolingual emotion recognition with this first evaluation. Second, we evaluated multilingual-only emotion recognition using five-fold cross-validation. The number of samples for each fold is 223 samples (I), which makes it impossible to create a language-balanced partition for each fold due to the nature of the dataset. For the multilingual with the first fixed split, we evaluated the multilingual model on both multilingual samples (a mixed language, 181 samples in the test set) and three monolingual samples (50, 65, and 66 samples in the test set for English, Italian, and Spanish, respectively). The former represents a multilingual model for multilingual samples, while the latter refers to a multilingual model for monolingual samples.

Finally, we downsampled the original samples with 48 kHz to 16 kHz WAV files. This downsampling is needed since the wav2vec2-based model ([19], [20]) to extract acoustic embedding requires 16 kHz WAV files. We provided a Python file to downsample the dataset's files using SoX in the project repository. The repository also contains a CSV file listing files allocated for fixed training/test split and the code to evaluate the cross-validation.

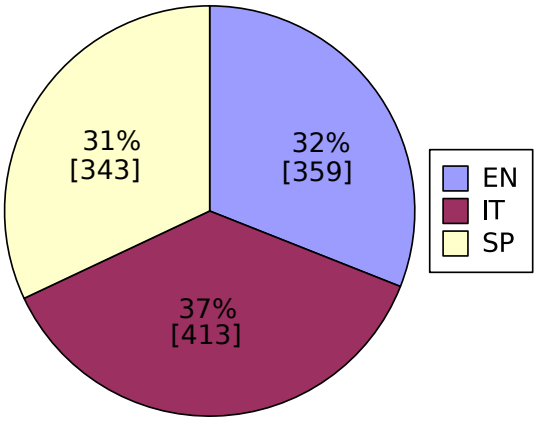


Fig. 1. Percentage [number of samples] of language distribution in EmoFilm dataset; EN: English; IT: Italian; SP: Spanish.

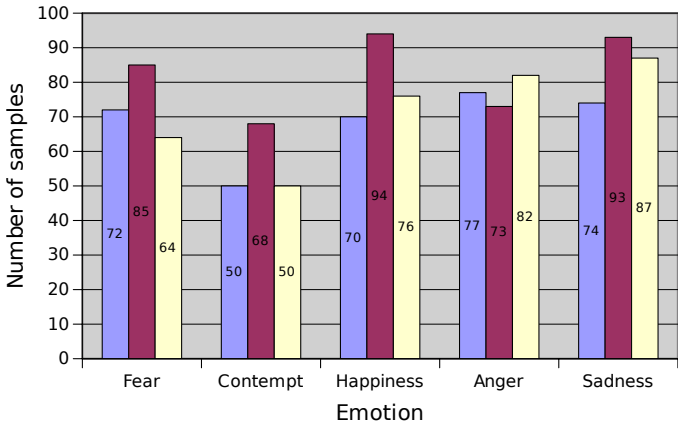


Fig. 2. Distribuion of samples on each emotion category; see Fig. 1 for color code.

TABLE I
DISTRIBUTION OF SAMPLES IN EMOFILM DATASET IN FIXED (TRAIN/TEST) SPLIT AND CROSS-VALIDATION (CV) SPLIT.

Partition	EN	IT	SP	Total
Fixed split				
Train	293	347	294	934
Test	50	66	65	181
Cross-validation				
Fold 1	49	122	52	223
Fold 2	100	86	37	223
Fold 3	64	75	84	223
Fold 4	83	72	68	223
Fold 5	47	58	118	223

IV. MULTILINGUAL, CROSS-LINGUAL, AND MONOLINGUAL SER

In this section, we defined the terms multilingual, cross-lingual, and monolingual emotion recognition as shown in Fig. 3. Multilingual emotion recognition is the process of recognizing emotions in multiple languages (e.g., languages A+B+C), either in a set of mixed languages (language A+B) or sets of several mono languages (Fig. 3 left). In this study, we mixed English, Italian, and Spanish to train the speech-emotion recognition (SER) system and tested the system in a set containing sentences from these languages (both in a mixed language or three mono languages). Cross-lingual emotion recognition is the process of recognizing emotions in different languages (language A+B $\rightarrow$ language C or language B $\rightarrow$ language C). Monolingual emotion recognition is the process of recognizing emotions in a single language (e.g., language B $\rightarrow$ language B). The term multicultural and cross-cultural emotion recognition may be better to replace multilingual and cross-lingual since emotion is more related to culture than language. Since different countries with their specific languages operationalize the cultures, we use the term multilingual and cross-lingual in this study following the previous studies.

V. ACOUSTIC FEATURE AND CLASSIFIER

Recent self-supervised learning models have shown promising results in speech emotion recognition [21], [22], [23],

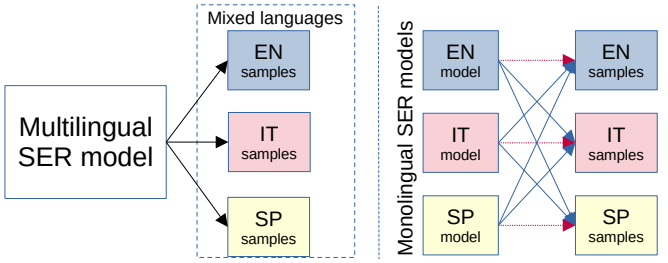


Fig. 3. The difference between multilingual (left), cross-lingual (right: solid blue line), and monolingual SER (right: magenta dot line) in experiments; the models are built using an SVM classifier.

[24]. Following this trend, we evaluated a pre-trained model finetuned on an affective speech dataset [20], [19], which is based on wav2vec2-large-robust model [25] trained on MSP-Podcasts dataset [26]. Models based on wav2vec 2.0 are known to be effective for SER and vocal bursts [27]. The speech embedding is the hidden state of the last layer before the output layer (1024-dim).

We evaluated the support vector machine (SVM) for classification (hence, support vector classification or SVC) for all experiments. This choice is based on the effectiveness of SVM in the previous research on small data [28], [29]. We optimized the C parameter in SVC with a loop in a range [0, 100] with step 0.1. The kernel is RBF with gamma "scale". Other parameters are left as default. The SVC is implemented in scikit-learn [30] with version `1.0.2`.

The evaluations are evaluated by both unbalanced accuracy (overall accuracy or weighted accuracy, WA) and balanced accuracy (unweighted average recall or unweighted accuracy, UA). WA is suitable for balanced datasets, and UA is suitable for unbalanced datasets. UA is perhaps more suitable in this study since the data is not balanced (both in categorical emotions and languages).

All information to reproduce the experiments is available in the project repository: [https://https://github.com/aistairc/emofilm-mser]. The Python codes are light and can be run on a notebook PC. We used the Iiyama laptop model L140MU (i7-1165G7, 24 GB RAM) with Ubuntu 20.04 OS for experiments. The experiments run for about three minutes for extracting acoustic embedding and within a minute for the classification result. We replicate the methods on the other two PCs to check the consistency.

VI. Results and Discussion

We present our results in two parts. The first part is the multilingual SER evaluation on both a multilingual model for multilingual/mixed data and a multilingual model for monolingual data. The second part is evaluations of cross-lingual SER and monolingual SER.

A. Multilingual results

Table II shows the result of WA and UA of a multilingual model on multilingual data or mixed languages (multi2multi). The first row refers to the fixed training/test split, while the rest are cross-validation results. The results of WA in both cases are generally higher than UA. In this unbalanced data, the UA scores are more reliable since they are the average score of each emotion category. The highest score from cross-validation was achieved on Fold 1 as test data, in which the Italian samples occupy the majority of that test data (Table I). In both cases, the cross-validation attained a lower score (average of five folds) and should be considered as the more reliable score since the data is shifted five times.

TABLE II
Results of Multilingual SER model on Mixed Languages

Split	% WA	% UA
Fixed split	78.45	75.77
Cross-validation		
Fold 1	79.37	78.89
Fold 2	68.61	68.27
Fold 3	69.06	66.04
Fold 4	77.13	77.13
Fold 5	77.13	77.13
Average	72.11	71.91

Table III exhibits the results of a multilingual model on monolingual data (multi2mono). This is similar to Table II but with a different test set for each language. In Table II, we mixed all samples from three languages; in Table III, we separated samples for each language. This evaluation used a fixed-split partition. This result is comparable to Table II, highlighting the robustness of the model. Italian samples attained the highest score in WA. Surprisingly, the highest score for UA was obtained in the Spanish language. A possible assumption for this finding is given as follows. Spanish may be more similar to English than Italian (perhaps because English has borrowed many words from it), and the pre-trained model is trained in English. Although the model for extracting acoustic embedding was trained in English, the portion of English samples in the training set is the lowest among the three languages. These explanations lead to Spanish being the highest score. However, the assumption is rejected since Spanish obtained the lowest scores on language-balanced evaluations (Table V), remaining unclear explanations on why Spanish is gaining the highest score in this multilingual to monolingual evaluation. Nevertheless, the differences in recognition rate occur mainly due to differences in the number of samples for each language. Future research could investigate this phenomenon, particularly by comparing the effect of languages in strict conditions (e.g., same number of samples, same speakers, same emotions, etc.).

TABLE III
Results of Multilingual SER model on Mono Languages (Fixed Split)

Language	% WA	% UA
EN	78.00	74.52
IT	81.81	72.34
SP	76.92	77.73
Average	78.91	74.86

B. Cross-lingual and monolingual results

Table IV depicts the results of cross-lingual and monolingual speech emotion recognition. Notice that monolingual evaluation can be seen as a special case of cross-lingual evaluation where the source and target are the same language. Hence, the configuration for both is the same. We combined both cases in one table for a better comparison. The diagonal cells are the monolingual SER results. The cross-lingual SER results are in the other cells. The results are in the form of WA / UA. The partition of data follows the fixed split evaluation. If we look at the score of UA, the preferred metric over WA, almost all scores showed low-performance scores ($< 70\%$) for cross-lingual emotion recognition. Only the Italian model tested on the Spanish model gained 70.52% of UA, perhaps due to the similarity between both languages. Monolingual evaluations scored higher than cross-lingual evaluations, particularly on the UA scores. The only exception for this case (monolingual vs. cross-lingual) is the score of WA for the Spanish model, in which the score of WA for Italian data (70.52%) is higher than its own Spanish data (69.15%). Among the three languages in monolingual evaluations, Italian scored the highest (78.41%), perhaps because of the number of samples (on both training and test) and the expressiveness of the Italian language. Although cross-lingual emotion recognition is more challenging than monolingual emotion recognition, one can merge multilingual data to build a multilingual model, which previously performed better than cross-lingual emotion recognition.

TABLE IV
RESULTS (% WA / % UA) OF CROSS-LINGUAL (ALL ROWS EXCEPT THE DIAGONAL MAGENTA) AND MONOLINGUAL (DIAGONAL MAGENTA) SER MODELS; PARTITION: FIXED SPLIT

Language	EN	IT	SP
EN	72.00 / 68.96	62.12 / 52.52	50.76 / 54.86
IT	62.00 / 61.14	77.27 / 78.41	67.69 / 70.52
SP	62.00 / 53.41	77.27 / 65.05	70.76 / 69.15

C. Discussion

The evaluations of multilingual, cross-lingual, and monolingual emotion recognition show the potential benefit of multilingual emotion recognition systems. First, it achieved comparable results to monolingual emotion recognition. The score of average UA was 71.91% in cross-validation multilingual SER (mixed languages) compared to 72.17% in monolingual SER (average of three languages). For a multilingual model to mono languages, the UA was higher, i.e., 74.86%, which is an average score for a multilingual model on three languages. The multilingual model to mono language evaluations may gain this higher score due to more samples in the training data. Cross-lingual SER performed worst since it is the most challenging task. Given the higher score of a multilingual model on mono languages, we found no benefit in using the monolingual model for SER in this study, evidenced by the scores evaluated on the same test set (fixed split). In this fair comparison, multilingual SER gained an average UA of 74.86%, followed by monolingual (average UA 72.21%) and cross-lingual (average UA 58.03%, still higher than chance level 20%). This result is in line with the previous study, in which multilingual SER outperformed cross-lingual and monolingual SER [9] but within different dataset evaluations.

To check that the high score of multilingual SER is due to the number of data, we reduced the number of training samples from the previous 934 samples to 420 samples (160 samples per language), 360 samples (120 samples per language), and 300 samples (100 samples per language). The last number of samples is to match the least sample number (EN 293 samples). Table V shows the result. It can be shown that by reducing the number of samples, the performance of both WA and UA decreased. Interestingly, using only 360 language-balanced samples, we achieved comparable results with 934 language-unbalanced samples. Furthermore, using 420 balanced samples, less than half of the original training samples, we achieved a new higher score of multilingual SER to monolingual samples (multi2mono) with a UA of 75.52%. Another interesting finding in Table V is that Italian SER still obtained the highest scores under language-balanced evaluations. Again, it highlights the expressiveness of Italian culture among English and Spanish.

Table V also shows the inconsistency ranking of languages in the multilingual model to monolingual evaluations. In 420 samples and 360 samples, the best-performing language is English, followed by Italian and Spanish. In this case, the benefit of pre-training wav2vec2 on the English dataset (MSP-Podcasts) as an acoustic feature extractor may be the cause. In the smallest number of samples setting, with 300 samples, Italian gained the highest score, followed by Spanish and English. In this small number of samples, the effect of the classifier may be more dominant than the acoustic feature extractor (perhaps due to the expressiveness of the Italian languages than others).

TABLE V
EFFECT OF REDUCING THE NUMBER OF TRAINING SAMPLES ON MULTILINGUAL SER (OVERLINES SHOW MEAN SCORES)

# Samples	Multi2mono				Multi2multi	
	WA	UA	$\overline{WA}$	$\overline{UA}$	WA	UA
420 samples						
EN	78.00	80.08				
IT	80.30	77.76	74.82	**75.52**	74.59	73.13
SP	66.15	68.71				
360 samples						
EN	74.00	77.39				
IT	78.79	76.33	72.98	74.24	72.83	72.51
SP	66.15	69.00				
300 samples						
EN	66.00	66.52				
IT	75.76	72.37	68.79	68.85	69.06	67.47
SP	64.62	67.67				

There is a tendency that the more samples for a language in the test set, the higher the score. An example is the Italian language which has the most samples in the Fold 1 cross-validation (mixed languages, Table II) and in the fixed split mono language evaluations (Table III). In both cases,

the performances were the highest among others. These also can be caused by Italian data having the largest samples in the training set. Future research could be directed to build a language-balanced multilingual speech emotion dataset for observing the effect of language and samples in multilingual SER.

As an extension of this study, we expand our experiments to find insights into the usefulness of the extracted acoustic embedding. We plot the t-SNE visualization in Fig. 4 to show the groups of features according to their labels. As shown in that figure, there is a strong separation between anger and sadness; however, there is confusion between contempt and fear. Happiness is located in between since its features occupy both a specific location (left-top) and other locations (spread). We checked the t-SNE for each language, and the results were similar to the mixed language. This result supports the universality of emotional information from speech.

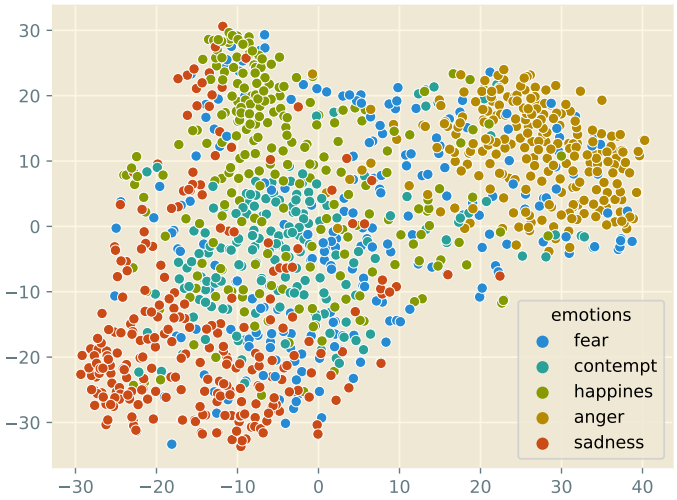


Fig. 4. t-SNE visualization of acoustic embedding from all samples (mixed languages)

VII. Conclusions

The research on multilingual speech emotion recognition has progressively evolved; this study evaluated different cases of speech emotion recognition in speaker-independent multilingual scenarios. It is believed that emotion, including those expressed in speech, is universal across cultures (operated by languages). We support this belief by showing the results of cross-lingual SER, which is higher than the chance level. Multilingual SER achieved better than cross-lingual and monolingual emotion recognition on the same test set. We achieved a score of 74.86% in the multilingual SER model on mono languages and 71.91% in the cross-validation multilingual SER model. The former score is higher than monolingual and cross-lingual evaluations in the same test set (but not on the same number of training sets). The high score of multilingual emotion recognition may be due to its natural benefit from merging all samples from all languages. We experimented with reduced numbers of training samples to prove this hypothesis. We also found a new high score using 420 language-balanced samples for the multilingual model to the monolingual with 75.52% balanced accuracy. However, the fair comparison using a similar number of training samples shows the superiority of monolingual SER among multilingual and cross-lingual.

Nevertheless, the previous scores (multilingual $>$ monolingual $>$ cross-lingual) may also show a fair comparison since all evaluations are done in the same test set. An ideal performance (unbalanced accuracy) would be over 90% for application in real-world scenarios. Hence, a direct implementation of multilingual speech emotion recognition into technology is far from being reliable. The results of this study are promising, but there is still a long way to go before we can achieve a reliable speech emotion recognition system. One way to mitigate this issue is by reducing the number of emotion categories to those that show strong correlations with the extracted acoustic feature/embedding. In this multilingual speech emotion recognition study, we showed that anger and happiness are separable from each other in t-SNE visualization.

Future research could also be directed to investigate the effectiveness of a large number of multilingual SER data by combining several datasets. It also discovered the possibility of training a model simultaneously to predict transcription, Language, age, gender, and emotion from speech [31], [22], [32]. This multitask learning may also help multilingual SER in the future. In this research, we showed that a monolingual setting is still the best-performing configuration under a similar number of training samples. Given the benefit of multilingual data that is available abundantly by nature, it is worth investigating the effectiveness of multilingual over monolingual SER on specific levels (e.g., benchmarking multilingual models to a monolingual dataset which have a large number of samples like English by tracking the performance improvements with the increase of training samples in the multilingual model).

Acknowledgment

This paper is partly based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan. The authors would like to thank the authors of the EmoFilm dataset [7] for sharing their data and Prof. Akagi of JAIST for the valuable discussion.

Ethical Impact Statement

This study explored the feasibility of multilingual speech emotion recognition (language A+B+C $\rightarrow$ language A+B+C) compared to cross-lingual (language A $\rightarrow$ language B) and monolingual (language A $\rightarrow$ language A) speech emotion recognition. One of the evaluations was performed on the same test set (a fixed train/test split) to compare the three evaluations fairly. The study shows the potential benefit of multilingual speech emotion recognition over monolingual and cross-lingual emotion recognition. The results support the previous finding of similar evaluations on the different datasets (English and French) [9]. This study's results are limited to the EmoFilm dataset (English, Italian, and Spanish) and should not

be generalized (to other datasets). Instead, it could be used as a reference for future studies. A CSV file for splitting samples into training/test sets will be given in the open repository to enable future benchmarking, along with cross-validation and other settings. The model generated by this study should not be used in real-case scenarios.

REFERENCES

[1] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *J. Cross. Cult. Psychol.*, vol. 32, no. 1, pp. 76–92, 2001.

[2] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, 2002.

[3] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Emotional Speech Recognition and Synthesis in Multiple Languages toward Affective Speech-to-Speech Translation System," in *2014 Tenth Int. Conf. Intell. Inf. Hiding Multimed. Signal Process.* IEEE, aug 2014, pp. 574–577. [Online]. Available: http://ieeexplore.ieee.org/document/6998394/http://ieeexplore.ieee.org/document/7041623/

[4] K. R. Scherer, "A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology," *6th Int. Conf. Spok. Lang. Process. ICSLP 2000*, no. Icslp, 2000.

[5] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 19, pp. 7241–7244, 2012.

[6] H. Narimatsu, R. Ueda, and S. Kumano, "Cross-Linguistic Study on Affective Impression and Language for Visual Art Using Neural Speaker," 2022.

[7] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs Dimensional Perception of Italian Emotional Speech," in *Interspeech 2018*, vol. 2018-Septe, no. September. ISCA: ISCA, sep 2018, pp. 3638–3642.

[8] R. Elbarougy and M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception," in *2013 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.* Kaohsiung: IEEE, oct 2013, pp. 1–10.

[9] M. Neumann and N. G. Thang Vu, "Cross-lingual and Multilingual Speech Emotion Recognition on English and French," in *2018 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2018-April. IEEE, apr 2018, pp. 5769–5773. [Online]. Available: https://ieeexplore.ieee.org/document/8462162/

[10] S. Latif, J. Qadir, and M. Bilal, "Unsupervised Adversarial Domain Adaptation for Cross-Lingual Speech Emotion Recognition," in *2019 8th Int. Conf. Affect. Comput. Intell. Interact. ACII 2019*, 2019.

[11] X. Cai, Z. Wu, K. Zhong, B. Su, D. Dai, and H. Meng, "Unsupervised Cross-Lingual Speech Emotion Recognition Using Domain Adversarial Neural Network," in *2021 12th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2021*, 2021, pp. 3–7.

[12] X. Li and M. Akagi, "Multilingual Speech Emotion Recognition system based on a three-layer model," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 3608–3612, 2016.

[13] ——, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Commun.*, vol. 110, pp. 1–12, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.04.004

[14] S.-w. Lee, "The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages," in *ICASSP 2019 - 2019 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2019, pp. 5881–5885.

[15] Z. Zhang, X. Zhang, M. Guo, W.-q. Zhang, K. Li, and Y. Huang, "A Multilingual Framework Based on Pre-training Model for Speech Emotion Recognition," in *APSIPA Annu. Summit Conf.*, no. December, 2021, pp. 750–755.

[16] Z. Wang, Q. Meng, H. Lan, X. Zhang, K. Guo, and A. Gupta, "Multilingual Speech Emotion Recognition With Multi-Gating Mechanism and Neural Architecture Search," in *SLT*, oct 2022. [Online]. Available: http://arxiv.org/abs/2211.08237

[17] V. Scotti, F. Galati, L. Sbattella, and R. Tedesco, "Combining deep and unsupervised features for multilingual speech emotion recognition," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12662 LNCS. Springer Science and Business Media Deutschland GmbH, 2021, pp. 114–128.

[18] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "SERAB: A Multi-Lingual Benchmark for Speech Emotion Recognition," in *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2022, pp. 7697–7701. [Online]. Available: https://ieeexplore.ieee.org/document/9747348/

[19] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–13, mar 2023.

[20] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0 (1.1.0)," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6221127

[21] B. T. Atmaja and A. Sasou, "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9964237/

[22] B. T. Atmaja, Zanjabila, and A. Sasou, "Jointly Predicting Emotion, Age, and Country Using Pre-Trained Acoustic Embedding," in *2022 10th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos*. IEEE, oct 2022, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10085991/

[23] M. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation," in *IberSPEECH 2022*, no. November. ISCA: ISCA, nov 2022, pp. 76–80.

[24] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1. ISCA: ISCA, aug 2021, pp. 521–525.

[25] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Interspeech 2021*, vol. 3. ISCA: ISCA, aug 2021, pp. 721–725.

[26] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, 2019.

[27] B. T. Atmaja and A. Sasou, "Evaluating Variants of wav2vec 2.0 on Affective Vocal Burst Tasks," in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, jun 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096552/

[28] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," in *2018 Int. Conf. Front. Inf. Technol.* IEEE, dec 2018, pp. 88–93.

[29] B. T. Atmaja, Zanjabila, and A. Sasou, "On The Optimal Classifier For Affective Vocal Bursts And Stuttering Predictions Based On Pre-Trained Acoustic Embedding," in *2022 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. IEEE, nov 2022, pp. 1690–1695.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825—-2830, 2011.

[31] Z. Kons, H. Aronowitz, E. Morais, M. Damasceno, H.-K. Kuo, S. Thomas, and G. Saon, "Extending RNN-T-based speech recognition systems with emotion and language classification," in *Interspeech 2022*, no. September. ISCA: ISCA, sep 2022, pp. 546–549.

[32] H. Zhao, N. Ye, and R. Wang, "Transferring Age and Gender Attributes for Dimensional Emotion Prediction from Big Speech Data Using Hierarchical Deep Learning," in *2018 IEEE 4th Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. IEEE Int. Conf. Intell. Data Secur.* IEEE, may 2018, pp. 20–24.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Ensembling Multilingual Pre-Trained Models for Predicting Multi-Label Regression Emotion Share from Speech

Bagus Tris Atmaja* and Akira Sasou*
* National Institute of Advanced Industrial Science and Technology, Japan
E-mail: b-atmaja@aist.go.jp, a-sasou@aist.go.jp

***Abstract*—Speech emotion recognition has evolved from research to practical applications. Previous studies of emotion recognition from speech have focused on developing models on certain datasets like IEMOCAP. The lack of data in the domain of emotion modeling emerges as a challenge to evaluate models in the other dataset, as well as to evaluate speech emotion recognition models that work in a multilingual setting. This paper proposes an ensemble learning to fuse results of pre-trained models for emotion share recognition from speech. The models were chosen to accommodate multilingual data from English and Spanish. The results show that ensemble learning can improve the performance of the baseline model with a single model and the previous best model from the late fusion. The performance is measured using the Spearman rank correlation coefficient since the task is a regression problem with ranking values. A Spearman rank correlation coefficient of 0.537 is reported for the test set, while for the development set, the score is 0.524. These scores are higher than the previous study of a fusion method from monolingual data, which achieved scores of 0.476 for the test and 0.470 for the development.**

***Index Terms*—speech emotion recognition, ensemble learning, multilingual speech processing, pre-trained model, emotion share**

I. Introduction

Research on speech emotion recognition (SER) has been focused mainly on recognizing categorial emotions like happy, angry, sad, or neutral. There are attempts to predict dimensional emotions like valence, arousal, and dominance; however, their application is still limited. The reason is that the dimensional emotion is not as straightforward as the categorical emotion. Nevertheless, the previous studies on categorial emotion recognition are designed to choose a single class of emotions in multiclass settings. In contrast to categorial emotions, the emotions, particularly conveyed by vocal bursts, are bridged by smooth gradients (instead of clear separation between discreet emotions) with continuously varying meanings [1]. In that study, the authors found 24 categorial emotions, which may not have been limited to vocal bursts only but also to general emotion perception or speech perception (since vocal bursts are commonly found within speech).

An attempt to analyze emotion fine-grained emotion (i.e., as a regression task) instead of coarse-grained emotion (i.e., by classification task) has been conducted from a linguistic point of view [2]. The authors mapped the degrees of valence and arousal given the sentence into six basic categorical emotions via several lexicons and corpus. The top model (using WKB lexicon) obtained correlation scores of 0.65 for dimensional emotions (valence, arousal, dominance) and 0.416 for categorial emotions. Besides only using linguistic information, the evaluation was conducted on English data only.

Recently, the multilingual approach to tackle the limitation of speech technology has been proposed in the parent tasks of speech processing, automatic speech recognition, speech synthesis, and speech translation. Models like Whisper [3] and MMS [4] can precisely recognize speech in multiple languages. Stirred by this direction, research on speech emotion recognition has been directed to tackle its limitation under multilingual settings. Assuming that emotion is a universal language (i.e., the pattern of angry speech is the same across languages or cultures), the multilingual approach is expected to work also on speech-emotion recognition with diverse language samples.

Multilingual SER attempts to recognize emotion regardless of the language. One can build a single model (including a model trained on multilingual data) for each language with language detection as pre-processing; another can build a single model for all languages. This study is aimed at the latter approach. The previous studies utilized transfer learning for cross-lingual SER [5], distribution adaption [6], and implicit distribution alignment [7]. The last two studies were developed for cross-corpus SER with different languages.

The use of multiple models for SER has shown to be effective for monolingual data. The technique of ensemble was usually adopted to fuse several models from different modalities and/or features. Examples are acoustic and linguistic for dimensional SER [8], [9], acoustic and facial expression [10], and multi-feature fusion [11]. In the previous study [12], ensemble learning was also utilized to tackle the limitation of skewed training data in monolingual SER by maximizing, averaging, and log averaging of the output probabilities of the component classifiers.

This paper contributes to the evaluation of ensemble learning, combining several models via late fusion methods for predicting emotion share from speech in multilingual settings. The fusion was performed by averaging predictions of different

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

models. We evaluated eight multilingual pre-trained models since the dataset contains multilingual data from English and Spanish. The pre-trained models are chosen from those languages. We showed that by combining the models via ensemble learning with the late fusion method, we could improve the performance of the baseline model with a single model. The performance is measured using the Spearman rank correlation coefficient since the task is a regression problem with ranking values.

II. Dataset and Task

The dataset used in this study is obtained from the dataset for The ACM Multimedia 2023 Computational Paralinguistic Challenge (ComParE) Emotion Request Sub-Challange [13], which is provided by Hume AI [14]. There are nine categorical emotions in the dataset: 'Anger', 'Boredom', 'Calmness', 'Concentration', 'Determination', 'Excitement', 'Interest', 'Sadness', and 'Tiredness'. The choice of these nine emotions is based on their more balanced distribution in the valence-arousal space. The basis for the dataset itself is more than 5,000 'seed' samples.

Seeds consist of various emotional expressions, which were gathered from openly available datasets, including MELD [15] and VENEC [16]. The seed samples were mimicked by speakers recruited via Amazon Mechanical Turk by the provider of the dataset [14]. The dataset consists of 51,881 'mimic' samples (total of 41:48:55 h of data, mean 2.9 s, range 1.2 - 7.98 s) from 1,004 speakers aged from 20 to 66 years old. The data were collected from 3 countries with broadly differing cultures: the United States (English/EN), South Africa (English/EN), and Venezuela (Spanish/SP). Each seed sample was rated by the individual who imitated it using a 'select-all-that-apply' method. Seeds were assigned a mean of 2.03 emotions per rater (max: 7.11, min: 1.00), with a standard deviation of 1.33 emotions. The proportion of times a given seed sample was rated with each emotion was then applied to all mimics of that seed sample. This method results in the share per emotion assigned by the speakers. The label then is normalized by dividing all (floating point) scores with their highest score per sample; hence, each sample has a maximum score of 1.0 for one emotion category and other scores less than 1.0 for other emotion categories. The baseline results for this dataset is a Spearman score of 0.500 for the development set and 0.514 for the test set (using wav2vec finetuned of affective speech dataset with SVM method). The labels of the test set are not open to the public; hence, only performance on the development/validation data can be calculated directly.

Table I shows the characteristics of the dataset. The ratio for training/test split follows the practical machine/deep learning practice, i.e., 80/20. About 18.3% of the samples are allocated for the test (without labels), while the rest are for training (including validation/development). The ratio of female/male is about 63.5/36.5, with the number of female samples being higher than male. The splitting of training/test follows speaker-independent criteria with a total of 1004 speakers; 20% of them is allocated for the test set while the rest is for training and development.

TABLE I
Summary of the dataset for Emotion Share Recognition, derived from [13]

	Train	Dev	Test	Σ
Sample no.	30,133	12,241	9,507	51,881
Sample %	58.1	23.6	18.3	100.0
Speaker no.	600	202	202	1,004
Gender (f:m)	379 : 221	117 : 85	141 : 61	637 : 367
Gender %	62.9 : 37.1	57.9 : 42.1	69.8 : 30.2	63.5 : 36.5

TABLE II
Pre-trained models used in this study

Model Name	Hugging Face ID/Name
wav2vec 2.0	audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
XLS-R 53	faceboook/wav2vec2-large-xlsr-53
XLS-R 53 EN	jonatasgrosman/wav2vec2-large-xlsr-53-english
XLS-R 53 SP	jonatasgrosman/wav2vec2-large-xlsr-53-spanish
XLSR-300M	facebook/wav2vec2-xls-r-300m
XLSR-1B	facebook/wav2vec2-xls-r-1b
XLSR-1B EN	jonatasgrosman/wav2vec2-xls-r-1b-english
XLSR-1B SP	jonatasgrosman/wav2vec2-xls-r-1b-spanish
XLSR-2B	facebook/wav2vec2-xls-r-2b

III. Pre-Trained Models

We evaluated nine pre-trained models as acoustic embedding feature extractors and fused the results of these models for ensemble learning. Ensemble learning leverages the advantages of multiple models to achieve a better performance score. In this study, we evaluated a single late fusion method by averaging the emotion share predictions of different models. The baseline model is the robust version of on wav2vec 2.0 [17] finetuned the affective dataset [18]. For multilingual pre-trained models, we employed XLS-R 53 [19] and its variants, XLSR variants [20] (XLSR-300M, XLSR-1B, and XLSR-2B). For the XLS-R 53 and XLSR-1B, we also evaluated the finetuned version of this model on the English (EN) and Spanish (SP) datasets. Finetuning for these languages is not available on other XLSR variants. The complete pre-trained models are listed in Table II.

IV. Classifier and Hyperparameter Search Space

We employed a support vector machine (SVM) classifier for regression (SVR, support vector regression). The type of kernel for SVR is linear (LinearSVR in the scikit-learn toolkit). The optimal parameters for this SVR were searched using the Grid Search algorithm, i.e., the regularization parameter C and the algorithm to solve either the 'dual' or 'primal' optimization problem. Other parameters like scoring were also searched. The scoring is either using negative mean absolute error (NMAE) or negative mean squared error (NMSE). The maximum iteration is fixed at 5000. The value range for these parameters is shown in Table III.

For the fusion of all nine models, we employed the average of the predicted values (arithmetic mean). The predicted values are the continuous values in either development or test for each of the nine emotion categories. The ordinal values are then

TABLE III
VALUE RANGE FOR HYPERPARAMETER GRID SEARCH IN LINEARSVR

Hyperparameter	Value range
Scoring	[NMAE, NMSE]
Scaler	[StandardScaler, MinMaxScaler]
C	$[10^{-2} - 10^{-6}]$
dual	[True, False]
max_iter	[50000]

used to calculate performance metrics. The final score is an average of nine emotional categories.

The performance for evaluating the models is measured as the Spearman correlation coefficient (ρ) [21]. This Spearman's ρ metric is similar to the Pearson correlation coefficient but for ranked data. This metric is chosen since the annotation of the dataset is ordinal (based on rank) and for consistency with the baseline model [13]. Spearman correlation coefficient is calculated as follows:

$$\rho = \frac{\text{cov}(\text{R}(X), \text{R}(Y))}{\sigma_{\text{R}(X)}\sigma_{\text{R}(Y)}}, \quad (1)$$

where $\text{cov}(\text{R}(X), \text{R}(Y))$ denotes the covariance of the ranked data $\text{R}(X)$ and $\text{R}(Y)$, and $\sigma_{\text{R}(X)}$ and $\sigma_{\text{R}(Y)}$ denote the standard deviation of the ranked data $\text{R}(X)$ and $\text{R}(Y)$, respectively.

A repository is created to ensure the reproducibility of the experiments [1]. The repository contains the experiment settings, including the hyperparameter search space, and the results of the experiments. The requirements to run the experiments are also provided in the repository with the exact version of the libraries used at the time of the experiment.

V. EXPERIMENT RESULTS AND DISCUSSIONS

We divided our results into validation or development (Table IV) and test results (Table V). In the development stage, we ensure that the result of ensemble learning (Fusion all in Table IV) is higher than baseline results, either from a single model or late fusion. We submitted our prediction of the test result to the provider of the dataset [13] to obtain the performance on the test set.

Table IV shows our result in the development stage using nine pre-trained models and a fusion of all these models. On a single model evaluation, it is shown that using pre-trained models (such as XLS-R 53 and XLSR variants) achieves higher accuracies than conventional acoustic feature extractors like auDeep and ComParE. For instance, XLS-R 53 (the smallest multilingual model) gained a Spearman score of 0.4328, while ComParE achieved 0.359. For the fusion models, our fusion of all nine pre-trained models also overcomes the previous late fusion of wav2vec2, auDeep, DeepSpectrum, and ComParE.

Having good results in the development stage, we evaluated the same model on the test set. Table V shows the result of the test set. The result shows that the fusion of all models achieves the highest Spearman score of 0.537 (compared to the 0.476 of the late fusion from the previous study [13]).

[1]https://github.com/bagustris/ComParE2023

TABLE IV
SPEARMAN CORRELATION COEFFICIENT (ρ) ON THE DEVELOPMENT SET OF THE EVALUATED PRE-TRAINED MODELS

No.	SSL model	NMSE	NMAE
1	wav2vec 2.0	0.5022	0.5001
2	XLS-R 53	0.4328	0.4341
3	XLS-R 53 EN	0.4732	0.4739
4	XLS-R 53 SP	0.4687	0.4691
5	XLSR-300M	0.4555	0.4561
6	XLSR-1B	0.4694	0.4679
7	XLSR-1B EN	0.4559	0.4533
8	XLSR-1B SP	0.4451	0.4440
9	XLSR-2B	0.4813	0.4835
	Fusion all (of 9)	**0.5236**	**0.5266**

TABLE V
COMPARISON OF SPEARMAN CORRELATION COEFFICIENT (ρ) ON THE DEV(EVELOPMENT) AND TEST SETS OF THE BASELINES (FIRST FIVE ROWS) AND THE PROPOSED FUSION METHOD (THE LAST ROW)

No.	Model	Dev	Test
1	wav2vec2 [18]	0.500	0.514
2	auDeep [22]	0.347	0.357
3	DeepSpectrum [23]	0.335	0.331
4	ComParE [24], [25]	0.359	0.365
	Late Fusion (of 4) [13]	0.470	0.476
	Fusion all (of 9)	**0.524**	**0.537**

This result is also higher than the previous best result of 0.500 from wav2vec2 [18]. We assume ensemble learning works by leveraging information across different languages and pooling maximum information for multilingual SER.

Similar trends between test and development sets are observed where the performance of the test set is slightly higher than the development set. An exception only applies to the DeepSpectrum result where the test result is 0.004 lower than the development set. This trend indicates the generalization of the evaluated method (including the baseline methods that use a similar SVM classifier) to unseen data and shows that, perhaps, the distribution of the test set is similar to the development set.

Finally, Table VI breaks down the average Spearman's ρ in the test set ($\rho = 0.537$) into each emotion category. The result is from 'Fusion all' of nine pre-trained models on the test set. The result shows that the model performs best in the 'Calmness' emotion category (0.6061) and worst in the 'Interest' emotion category (0.4238). The result is similar to the previous study [13] where the 'Calmness' emotion category is the highest (0.559), and the 'Anger' emotion category is the lowest (0.428). The previous study's result was achieved with wav2vec 2.0 but with a different setup from this study (particularly on the scoring method).

VI. CONCLUSIONS

In this study, we evaluated nine pre-trained speech models and a fusion of all these nine models for emotion share recognition from speech. The idea was to collect multilingual models from different languages and apply the fusion of these models to multilingual SER data. A dataset from three countries with English and Spanish languages was selected. First, the fusion

TABLE VI
SPEARMAN CORRELATION COEFFICIENT (ρ) ON EACH EMOTION CATEGORIES FOR THE TEST SET

Emotion	Spearman's ρ
Anger	0.4894
Boredom	0.6032
Calmness	0.6061
Concentration	0.5855
Determination	0.5589
Excitement	0.4669
Interest	0.4238
Sadness	0.5123
Tiredness	0.5867

of nine pre-trained models defeats the previously reported best result of a single model (in which the performance of this single model is higher than the late fusion of four methods). Second, the results of any single model in this study generally are also higher than the single models/methods in the previous study. Third, there is a trend in which the performance of the test set is slightly higher than the development set. Future studies could be directed to explore more about the dataset and improve the performance using more advanced methods with more recent speech embeddings.

ACKNOWLEDGMENT

This paper is partly based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

REFERENCES

[1] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization." *Am. Psychol.*, vol. 74, no. 6, pp. 698–712, sep 2019.

[2] S. Buechel and U. Hahn, "Emotion analysis as a regression problem-dimensional models and their implications on Emotion representation and metrical evaluation," *Front. Artif. Intell. Appl.*, vol. 285, pp. 1114–1122, 2016.

[3] A. Radford, J. Wook, K. Tao, X. Greg, B. Christine, and M. Ilya, "Robust Speech Recognition via Large-Scale Weak Supervision," *openai.com*, 2021.

[4] V. Pratap, M. Auli, and M. Ai, "Scaling Speech Technology to 1 , 000 + Languages."

[5] S. G. Upadhyay, L. Martinez-Lucas, B.-h. Su, W.-c. Lin, W.-s. Chien, Y.-t. Wu, W. Katz, C. Busso, and C.-c. Lee, "Phonetic Anchor-Based Transfer Learning to Facilitate Unsupervised Cross-Lingual Speech Emotion Recognition," in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, jun 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10095250/

[6] S. Li, P. Song, L. Ji, Y. Jin, and W. Zheng, "A Generalized Subspace Distribution Adaptation Framework for Cross-Corpus Speech Emotion Recognition," in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, jun 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10097258/

[7] Y. Zhao, J. Wang, Y. Zong, W. Zheng, H. Lian, and L. Zhao, "DEEP IMPLICIT DISTRIBUTION ALIGNMENT NETWORKS FOR CROSS-CORPUS SPEECH EMOTION RECOGNITION," *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.*

[8] B. T. Atmaja, Y. Hamada, and M. Akagi, "Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion," in *2020 IEEE Reg. 10 Conf.* IEEE, nov 2020, pp. 1081–1085.

[9] B. T. Atmaja and M. Akagi, "Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information," in *Proc. 2020 23rd Conf. Orient. COCOSDA Int. Comm. Co-ord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2020*. IEEE, nov 2020, pp. 166–171.

[10] J. Zhao and S. Chen, "Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions," in *Cross-cultural Emot. Sub-challenge AVEC'18*, 2018, pp. 65–72.

[11] B. T. Atmaja and M. Akagi, "The Effect of Silence Feature in Dimensional Speech Emotion Recognition," in *10th Int. Conf. Speech Prosody 2020*, no. May. ISCA, may 2020, pp. 26–30.

[12] P.-Y. Shih, C.-P. Chen, and C.-H. Wu, "Speech Emotion Recognition With Ensemble Learning Methods," in *IEEE Int. Conf. Acoust. Speech, Signal Process. 2017*, 2017, pp. 2756–2760.

[13] B. Schuller, A. Batliner, S. Amiriparian, C. Bergler, M. Gerczuk, N. Holz, P. Larrouy-Maestri, S. Bayerl, K. Riedhammer, A. Mallol-Ragolta, M. Pateraki, H. Coppock, I. Kiskin, M. Sinka, and S. Roberts, "The ACM Multimedia 2023 Computational Paralinguistics Challenge: Emotion Share and Requests," in *Proc. 30th ACM Int. Conf. Multimed.* New York, NY, USA: ACM, oct 2023, pp. 7120–7124. [Online]. Available: https://dl.acm.org/doi/10.1145/3503161.3551591

[14] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nat. Hum. Behav.*, vol. 3, no. 4, pp. 369–382, apr 2019. [Online]. Available: http://www.nature.com/articles/s41562-019-0533-6

[15] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 527–536.

[16] P. Laukka, H. A. Elfenbein, W. Chui, N. S. Thingujam, F. K. Iraki, T. Rockstuhl, and J. Althoff, "Presenting the VENEC Corpus: Development of a Cross-Cultural Corpus of Vocal Emotion Expressions and a Novel Method of Annotating Emotion Appraisals," *Lr. 2010 - Seventh Int. Conf. Lang. Resour. Eval.*, vol. 7, pp. 53–57, 2010.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, 2020.

[18] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–13, mar 2023.

[19] X. Cai, Z. Wu, K. Zhong, B. Su, D. Dai, and H. Meng, "Unsupervised Cross-Lingual Speech Emotion Recognition Using Domain Adversarial Neural Network," in *2021 12th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2021*, 2021, pp. 3–7.

[20] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*, vol. 2022-Septe. ISCA: ISCA, sep 2022, pp. 2278–2282. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2022/babu22_interspeech.html

[21] C. Spearman, "The Proof and Measurement of Association between Two Things," *Am. J. Psychol.*, vol. 100, no. 3/4, p. 441, 1987. [Online]. Available: https://www.jstor.org/stable/1422689?origin=crossref

[22] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, 2018.

[23] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore Sound Classification Using Image-Based Deep Spectrum Features," in *Interspeech 2017*. ISCA: ISCA, aug 2017, pp. 3512–3516.

[24] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimed. - MM '13*. New York, New York, USA: ACM Press, 2013, pp. 835–838.

[25] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Interspeech 2013*, no. August. ISCA: ISCA, aug 2013, pp. 148–152.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

An Automatic Pipeline For Building Emotional Speech Dataset

Ngoc-Anh Nguyen Thi, Bao Thang Ta, Nhat Minh Le, Van Hai Do
Viettel Cyberspace Center, Viettel Group, Vietnam
Email:{anhntn51,thangtb3,minhln2,haidv21}@viettel.com.vn

***Abstract*—Emotional speech synthesis has gained attention for enhancing the naturalness and expressiveness of synthesized speech. However, developing emotional speech synthesis presents challenges due to the lack of high-quality emotional speech datasets, especially for low-resource languages. This paper proposes an automated process for building datasets of emotional speech that reduces dataset generation time by 92%. The pipeline combines trained models and leverages publicly available data. The effectiveness of the proposed pipeline is examined and demonstrated using a Vietnamese case study.**

I. Introduction

In recent years, Text-to-Speech (TTS) technology has achieved widespread popularity for its time-saving and accessibility benefits in communication [1], [2]. One promising area of this research is emotional speech synthesis, which aims to generate speech that effectively conveys emotional nuances through the modulation of prosody and other vocal cues. By making human-computer interaction more natural and human-like, emotional speech synthesis has the potential to become a revolutionary technology.

However, the development of emotional speech synthesis comes with several technical challenges. One significant problem is the requirement for the quantity of training data [3]. Achieving human-like and emotional speech requires a large amount of data, and the process of collecting and classifying that data can be both time-consuming and expensive. Most publicly available emotional speech datasets are in popular languages, such as IEMOCAP [4], SAVEE [5], RAVDESS [6] for English; EMODB [7] for German; MES [7], CHEAVD [8] for Chinese, while efforts in other languages are less common [9]. To solve this problem, some papers mention generating data in specific language. Most emotional speech datasets are built by recording data from prepared scripts and specified emotion labels, such as [10]–[13]. With collected speech data like podcasts, the data are classified based on votes [14], [15]. Agreement voting is used to resolve disagreements among annotators. The three annotators rely on a majority vote when there is disagreement. The utterance is deleted if there is still no consensus among the annotators. However, these approaches lack uniformity, pose challenges when applied to different languages, and often rely on manual processes, which results in expenses for both money and labor.

In a low-resource language like Vietnamese, there has been some progress in the development of expressive speech synthesis, although it remains in the early stages. One of the primary challenges faced by researchers in Vietnam is the lack of high-quality speech datasets, which hampers the training of authentic emotion speech synthesis models. In an effort to address this, studies such as [16], [17] have constructed a 90-minute audio-visual expressive speech corpus for Vietnamese through recording methods, while [18] proposes a method to generate synthetic emotional speech using data from text-to-speech tools. Despite these constraints, there is still room for improvement in this domain, emphasizing the need to obtain a large number of datasets specifically designed for studying emotional speech synthesis.

To meet this demand, an automatic pipeline for quickly building an emotional speech dataset using available resources is presented. We collect data from the Internet and automatically process the steps of audio separation for target speakers, text transcription, emotion labeling, and audio enhancement. This method saves 92% of data construction time compared to the manual method. Additionally, to prove the effectiveness of the proposed pipeline, a Vietnamese emotional speech dataset was built, has been released, and is freely downloadable at https://github.com/ngocanh2162/Vietnamese_emotional_dataset.

The rest of the paper is organized as follows: Section II presents the design, followed by the description of the dataset and analysis of results in Section III. Finally, the conclusions are drawn in Section IV.

II. The automatic emotional speech corpus design

This work proposes a pipeline with data collection, four automatic processing steps, and manual verification (optional). The resulting dataset will contain speech sounds, corresponding text, and emotion labels.

A. Data Collection

The first step in creating a corpus is to collect audio data. This can be obtained from various sources, such as recorded speeches, phone conversations, or other audio files. The quality of the audio data is an important factor that affects the quality of the TTS model.

In general, there are three methods for creating an emotional speech dataset, depending on their intended application: acted (simulated), elicited (evoked, induced), and natural speech (spontaneous) [9]. The majority of the databases record forced (simulated) emotional speech using professional artists, drama students, or voice talents [12]. In this work, the natural speech

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

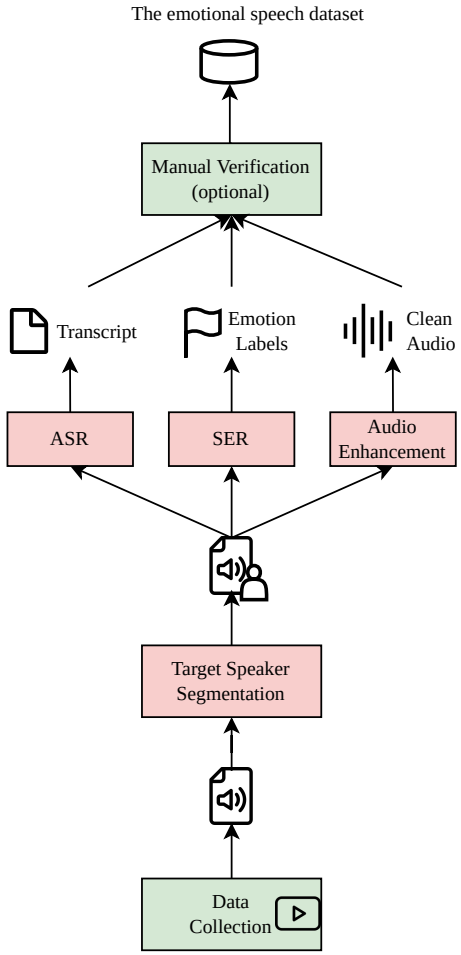


Figure 1. Automatic pipeline for building emotional speech dataset.

approach is preferred for creating a low-resource emotional speech dataset due to its benefits. It allows for the retention of natural emotions, and the necessary data is easily accessible and immediately available while helping to save time and reduce human costs.

To design a speech corpus for speech synthesis, specific criteria must be defined. The corpus will be used to study prosody and pronunciation and build expressive speech synthesis models. The following requirements have been determined:

- Large quantity of data from one speaker.
- Good sound quality.
- Varied discourse styles and literary genres.
- The emotion conveyed in speech.
- Consistent timbre and speed.

With the above criteria, many data sources can be selected, such as movies, talk shows, and audiobooks. In the Vietnamese case study, one female drama actress was chosen. All her speech data on the Internet was collected from 2008 to 2013. This ensures consistent data without too many variations in speech changes that could affect analysis and interpretation.

Data crawling collects the highest-quality video clips available. The audio is then extracted to a WAV file using the "SoX" package[1]. Due to the diversity of the sources from which the data is collected, there are numerous variations in acoustic conditions. Therefore, all audio is reduced to a sampling rate of 22,050 Hz and 16 bits. The next step is to preprocess the acquired data.

B. Target Speaker Segmentation

Once audio with multiple speakers, silence, and background noise has been collected, the audio must be extracted into segments containing only the speech of the target speaker.

Normally, to get an audio clip containing only the target speaker, the listener would have to watch the whole video and manually select it, adjusting the timestamp to avoid clipping into the audio containing the voice or other speakers. For this work, audio segments of the target speaker are obtained using Speaker Diarization and Speaker Verification. First, Speaker Diarization divides a conversation with multiple speakers into segments spoken by the same speaker. Then, Speaker Verification verifies a speaker's identity based on their speech characteristics.

The first step in the pipeline involves segmenting the 5-45 minute-long audio file into smaller, more manageable 1-15 second-long utterances. This is achieved through the use of a pre-trained Speaker Diarization model, which is capable of isolating human speech from other background noises. The resulting audio clips contain only the human speech that is of interest for further processing.

In the second step, these audio clips are passed through a pre-trained Speaker Verification model, which is used to identify the target speaker's audio segments. This is achieved by manually extracting 10- to 120-second audio segments containing only the target speaker's speech, which are then passed through the Speaker Verification model. This model is trained to recognize the unique characteristics of the target speaker's speech and can identify their speech segments with a high degree of accuracy.

In this step, depending on the confidence threshold set, the model can select more segments with lower accuracy or fewer segments with higher accuracy. A high confidence threshold to limit errors in the data obtained. If crawl data is a small quantity, a lower threshold can be selected to capture as much positive data as possible.

In our study case, a Vietnamese pre-trained model [19] was used for this step, but any language pre-trained model can be used because it is language-independent.

Then, the audio segments containing the target speaker's speech were obtained. To ensure data quality, files containing the speech of another speaker are manually found and excluded from further processing.

C. Text Transcribing

In this step, with the advanced development of Automatic Speech Recognition (ASR) in languages, a same-language ASR model is used to convert spoken language into a text

[1]https://sox.sourceforge.net/

transcript instead of using the labeler to listen to the audio and then write down the text. For Vietnamese, a state-of-the-art pre-trained ASR model [20] with 4.17% Syllable-Error-Rate (SyER) has been decided to be used.

After completing this step, the text file will contain transcripts for each audio file created. The text file may have recognition mistakes due to noise, distortion, accents, and out-of-vocabulary (OOV) words.

The text in the dataset, extracted from the ASR model, does not require segmentation or normalization, only occasional correction of OOV words or inconsistencies caused by ASR errors. Our experiments showed many long periods of silence in audio without any corresponding prosodic punctuation, negatively impacting TTS performance. To address this issue, punctuation was inserted into the transcript at the positions of corresponding silences. Using the aligned time stamps provided by the ASR system, the duration of internal silence is calculated, and prosodic punctuation is inserted into the transcript at the corresponding silence positions. The silence duration threshold for inserting punctuation is set to 0.35 seconds.

D. Emotional Labeling

The first thing to determine is which emotion labels will be included in the dataset. The basic emotions in emotional speech research include neutral, anger, disgust, fear, happiness, sadness, surprise, and others [21]. These are also prevalent in the majority of the world's datasets. In the Vietnamese case, the emotions chosen are neutral, happy, sad, and angry. By choosing these emotions, the classification process can be simplified, and the dataset can be representative of the most common emotions.

Manual labeling of emotional speech data is a time-consuming and expensive task that requires human annotators to listen to and label each utterance based on the expressed emotion. With manual labeling, it requires a large number of annotators, which is time- and labor-intensive, not to mention the disagreement in label voting.

However, in recent years, there has been significant progress in developing automatic methods for labeling emotional speech data. These methods can significantly reduce the cost and time required for labeling. Most of the research on Speech Emotion Recognition (SER) has been conducted in English, with state-of-the-art (SOTA) results reaching 75-80% [22]. However, this limits the applicability of SER in low-resource languages such as Vietnamese. The dataset was inferred for emotion labeling with some pre-trained models [23]–[25] that were available to the public. However, only 35% of the predicted labels matched the manual labels. In our experience, this result can be attributed to (1) the differences in emotional expression between languages and (2) the overfitting of the models to the training dataset.

To overcome these challenges, techniques such as transfer learning, data augmentation, and unsupervised learning have been developed. In this work, transfer learning was used to build emotion labels. An SER model [26] was trained using corpora that included several other emotional speech datasets, such as IEMOCAP, CremaD, RAVDESS, TESS, ShEMO, EmoDB, EmoVo, SAVEE, and then fine-tuned with one Vietnamese emotional speech dataset, which we manually created.

Although there has been significant progress in automatically labeling emotional speech data, there are still some challenges and limitations. One of the main challenges is the lack of a standardized emotional speech dataset. The emotional speech datasets that are currently available are often small and not diverse enough to cover all emotional states.

Table I
COMPARE MANUAL PIPELINE AND PROPOSED PIPELINE PROCESSING TIME

Step	Data Duration (hours)	Proposed Processing (hours)	Manually Processing (hours)
Speaker Segmentation	32.5	-	65
Text Transcribing	4.8	-	15
Emotional Labeling	4.8	-	21
Error checking	4.8	8	-
Processing Time	**3.9**	**8**	**101**

E. Audio Enhancement

To process the audio data, silence trims were performed initially. For this, a simple script was created to trim the audio based on the threshold of sound amplitude.

The dataset consists of spontaneous data captured in many environments at varying volumes. As a result, the data contains a lot of noise and non-valuable signals. Using a pre-trained Speech Enhancement model [27], the signal was enhanced and the noise was removed to improve the data quality. This process involved eliminating background music and enhancing the audio waveform's clarity.

To handle volume differences between distinct utterances, the peak amplitude of all audio clips was normalized. The silence portions at the start and finish of utterances were adjusted to 0.2 seconds.

After the aforementioned four steps, the dataset contains corresponding audio files, text transcripts, and emotion labels. However, to be used effectively in speech research, the data must undergo further processing to achieve optimal quality.

III. RESULT

A. Analysis of Pipeline Errors

An experiment was conducted to compare the manual data processing times using the traditional approach with those using the recommended pipeline.

According to the data presented in Table I, when employing manual processing techniques, a total of 32.5 hours of collected data needs 65 hours of processing time to obtain 4.8 hours of target speaker audio. Additionally, it takes an additional 15 hours to convert this audio into text using a single labeler. The subsequent step of labeling the emotions of each labeler requires at least 7 hours. When considering a scenario with

Table II
PIPELINE ERRORS IN THE DATASET

Error Types	%
Segmentation	5.17
Another speakers	9.13
Mispronounced words	8.69
Wrong emotion label	14.15
Total	**19.26**

*Some utterances may have multiple errors.

the number of labelers set to 3, this step takes a total of 21 hours.

In contrast, the proposed pipeline demonstrates significantly improved efficiency compared to manual processing methods. Without error checking, the pipeline achieves a processing time of zero. If adding the correction and error removal steps, the processing of 4.8 hours of data requires approximately 8 hours to produce 3.9 hours of data. This remarkable reduction of 92.08% in processing time surpasses the manual data processing duration, which amounts to 101 hours with three reviewers.

Overall, the proposed pipeline significantly enhances processing efficiency, leading to substantial time savings when compared to the traditional manual approach.

To evaluate the utility of the proposed automated data build pipeline, the error statistics of the dataset were analyzed. The results are shown in Table II. These errors can be seen as corresponding errors in the data processing steps: segmentation error and another speaker corresponding to the target speaker segmentation step; mispronounced words with the text transcribing step; and wrong emotion label with the emotional labeling step. The emotion labeling step has a high error rate across all types of errors. By applying this pipeline in the future, it becomes possible to generate more datasets, contributing to enhanced accuracy in SER and reducing this error.

As a result, the proposed pipeline can be utilized to quickly produce an emotional speech dataset with an error rate of approximately 20%, or to perform an additional step of error correction while still saving approximately 76% of the processing time to produce datasets with the same quality as human data production.

B. Released Dataset

The Vietnamese emotional speech dataset has been standardized with good quality to synthesize artificial speech. All audio recordings were saved in WAV format at 22,050 Hz sampling rate and stored at 16 bits per sample. All transcripts are stored as TXT files in UTF-8 encoding. The audio and the corresponding transcript filenames are identical except for the extension. The released dataset has removed the errors found. Creating an error-free speech dataset is not an easy task. Even after manual checks and data validation, some mistakes may remain. According to my knowledge, our work is the first study to publish a Vietnamese emotional speech dataset.

Table III
DATASET SUMMARY

Dataset	Duration (hours)	# Utterances	# Words
Neutral	2.98	7,204	48,558
Happy	0.23	625	3,463
Sad	0.22	521	3,368
Angry	0.45	1,255	7,583
Total	3.88	9,605	62,088

Table IV
THE MOS SCORE

Emotion	Ground truth speech	Synthetic speech
Neutral	4.463 ± 0.736	4.167 ± 0.741
Happy	4.458 ± 0.632	3.458 ± 0.785
Sad	4.250 ± 0.833	4.000 ± 0.667
Angry	4.528 ± 0.656	3.944 ± 0.846

Table III provides statistics for the dataset. The public dataset contains approximately 3.88 hours of audio with 9,605 utterances. The neutral label accounts for 75% of the dataset, followed by angry, while happy and sad labels have equal proportions.

One of the goals of building the emotional speech dataset is to create good-quality synthetic voices. Hence, the quality of the synthetic speech produced from emotional speech data is also assessed. In order to evaluate the data quality without being biased by the model, a fine-tuning approach [28] is employed: The acoustic model is first trained with a neutral sub-dataset of the target speaker to establish a base architecture. Then, for each remaining emotion (happy, sad, and angry), fine-tuning is done with the corresponding emotional sub-dataset. In this case, the acoustic model is FastSpeech2 [29] and the vocoder is HifiGAN [30] V1. To evaluate the naturalness of the synthesized speech, a group of 25 listeners was assigned mean opinion scores (MOS) on a scale of 1 to 5. This evaluation process aims to create good-quality synthetic speeches and ensure the overall fluency and authenticity of the generated speech.

As shown in Table IV, the MOS scores of the synthesized speech generally convey emotions well, although slightly lower than the ground truth score. It shows that the speech effectively expresses emotions, demonstrating its capability to generate speech that is emotionally expressive and understandable. There is room for improvement, but these scores indicate its potential in various emotional contexts.

IV. CONCLUSIONS

An automatic pipeline was proposed to efficiently generate a low-language emotional speech dataset, and the emotional speech dataset was released to facilitate research on speech emotion recognition and speech synthesis. In our pipeline, it is possible to annotate speeches at half-time without suffering a loss in quality.

ACKNOWLEDGMENT

This work is funded by Viettel Cyberspace Center, Viettel Group, Hanoi, Vietnam.

REFERENCES

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Z. Mu, X. Yang, and Y. Dong, "Review of end-to-end speech synthesis technology based on deep learning," *ArXiv*, vol. abs/2104.09995, 2021.

[3] E. Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology., 2016.

[4] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[5] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[6] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, pp. 1–35, May 2018.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," vol. 5, Sep. 2005, pp. 1517–1520.

[8] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: A chinese natural emotional audio–visual database," *Journal of ambient intelligence and humanized computing*, vol. 8, no. 6, pp. 913–924, 2017, ISSN: 1868-5137.

[9] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[10] K. Kaur and P. Singh, "Punjabi emotional speech database: Design, recording and verification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 9, no. 4, pp. 205–208, 2021.

[11] J. X. Koh, A. Mislan, K. Khoo, *et al.*, "Building the Singapore English National Speech Corpus," in *Proc. Interspeech 2019*, 2019, pp. 321–325. DOI: 10.21437/Interspeech.2019-1525.

[12] R. Rajan, U. Haritha, A. Sujitha, and T. Rejisha, "Design and development of a multi-lingual speech corpora (tamar-emodb) for emotion analysis.," in *INTERSPEECH*, 2019, pp. 3267–3271.

[13] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: Design and collection of a multilingual emotional speech database," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), 2002.

[14] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data.," in *Interspeech*, 2019, pp. 4435–4439.

[15] W. Fan, X. Xu, B. Cai, and X. Xing, "Isnet: Individual standardization network for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1803–1814, 2022. DOI: 10.1109/TASLP.2022.3171965.

[16] D.-K. Mac, V. Aubergé, A. Rilliard, and E. Castelli, "Audio-visual prosody of social attitudes in Vietnamese: Building and evaluating a tones balanced corpus," in *INTERSPEECH 2009*, International Speech Communication Association (ISCA), 2009, pp. 2263–2266.

[17] D.-K. Mac, E. Castelli, and V. Aubergé, "Modeling the prosody of Vietnamese attitudes for expressive speech synthesis," in *Proc. 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2012)*, 2012, pp. 114–118.

[18] Q. H. N. Thanh X. Le An T. Le, "Emotional vietnamese speech synthesis using style-transfer learning," *Computer Systems Science and Engineering*, vol. 44, no. 2, pp. 1263–1278, 2023.

[19] T. L. Nguyen, B. T. Ta, V. H. Do, T. A. X. Tran, and N. M. Le, "Speaker diarization for vietnamese conversations using deep neural network embeddings," in *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, 2022, pp. 300–305. DOI: 10.1109/ICCE55644.2022.9852042.

[20] T. B. Thang, D. D. Son, L. D. Linh, D. X. Vuong, and D. Q. Tien, "ASR - VLSP 2021: Automatic speech recognition with blank label re-weighting," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 38, no. 1, 2022.

[21] M. Schröder and R. Cowie, "Developing a consistent view on emotion-oriented computing," in *Machine Learning for Multimodal Interaction*. Springer Berlin Heidelberg, 2006, pp. 194–205.

[22] W. T. Majid, G. T. Surya, Q. S. A. Ahmad, K. Mira, and A. Eliathamby, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[23] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: A large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 641–645.

[24] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martınez, "A proposal for multimodal emotion recognition using aural trans-

formers and action units on RAVDESS dataset," *Applied Sciences*, vol. 12, no. 1, p. 327, 2021.

[26] B. T. Ta, T. L. Nguyen, D. S. Dang, N. M. Le, and V. H. Do, "Improving speech emotion recognition via fine-tuning ASR with speaker information," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1–6.

[27] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7867–7871.

[28] N. Tits, K. El Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional tts," in *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 1*, Springer, 2020, pp. 52–60.

[29] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2022.

[30] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Analysis of Emotions in Speech using AESDD

S. Uthiraa and Hemant Patil
Speech Research Lab, DA-IICT, Gandhinagar Gujarat
E-mail: {uthiraa_s, hemant_patil}@daiict.ac.in

***Abstract*—Speech Emotion Recognition (SER) has been in focus for some time now. With increasing technological advancements in Artificial Intelligence (AI), machines are becoming an integral part of human life, thus expanding the scope and motivation for emotion recognition. In this paper, we use a relatively new (2018) dataset, namely, Acted Emotional Speech for Dynamic Database (AESDD) for Greek language, for acoustic analysis of emotions in speech, as this dataset has not been explored much in the literature. We focus on some prosodic features, such as energy, loudness, pauses/ silences, and narrowband spectrograms, to analyze five emotions, namely, anger, fear, disgust, happiness, and sadness. This is done by studying the narrowband spectrograms and squared energy function (i.e., l^2 norm of a signal). Further, as hearing is the process of detecting energy, a non-linear energy tracking operator, namely, Teager Energy Operator (TEO) profile of the audio sample, is taken. Results showed some distinguishing parameters among emotions and also between male *vs.* female speakers uttering the same sentence. The effect of cognition on emotions is briefly discussed to strengthen the results obtained.**

***Index Terms*—Speech Emotion Recognition, AESDD, Narrowband Spectrogram, Teager Energy Operator, Cognition.**

I. Introduction

Emotion, a word so common in day-to-day life, yet we still have not found scientific consensus on its exact definition. Emotion is " a strong feeling, such as love or anger, or feelings in general". Further, " emotions are mental states brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure". It is also a "complex experience of consciousness, bodily sensation, and behavior that reflects the personal significance of a thing, an event, or a state of affairs". Emotion is the most significant attribute that distinguishes us (humans) from robots. In this era, where Artificial Intelligence (AI) and other technologies are advancing, we are more and more dependent on machines, and for them to understand us in a meaningful way, emotion recognition is essential. Researchers after analyzing emotion, its source and effects have classified it in a ***4***-D model, where each dimension is independent from one another, namely, duration, quality, intensity, and pleasure [1].

The literal meaning of emotion recognition would be to identify the emotions. It is known that the production of speech is not just because of the vocal tract system, larynx, and lungs, however, it is the cognitive processes that are going from mind to mouth, i.e., transforming a non-linguistic concept with communicative intent to linguistically well formed utterance, which involves linguistic as well as physiological level of speech communication between speakers and the listeners [2], [3], in particular, transmission of electric signals from the brain via motor nerves to activate the muscles of larynx and vocal tract system. This implies that if the speech, which carries the emotion has cognitive characteristics, the emotion recognition refers to the attributes of emotional states based on the observation of ***visual*** and ***auditory*** non-verbal cues. Eventhough emotion recognition can be done using speech and facial expressions, this paper is solely focused on the analysis of emotions using speech.

As it is already stated that there is a huge gap between man and machine-based processing of audio and visual stimuli, i.e., a machine is unable to identify the emotional state of a speaker. This problem has opened a new research field called Speech Emotion Recognition (SER) [4]. Applications of emotion recognition are vast, some of which include - man-machine interaction, drivers safety, mobile communications, military applications, educational software [4], [5], and incorporating emotions in Text-To-Speech (TTS) voice. Emotion recognition also plays a vital role in recognizing the emotions of people with disabilities, such as Cerebral Palsy, Parkinson's Disease, and Autism Spectrum Disorder; which would enable us to bring them into the mainstream (i.e., blend well) with society [6]. However, to develop a good emotion recognition model, it is very important to understand the acoustics of various emotions. In particular, humans have a remarkable ability to convey and perceive emotions in speech using various prosodic cues, such as loudness (amplitude), pitch, duration of speech sound units, intonation (i.e., derivative of F_0 (fundamental frequency) contour), and energy of speech wave.

In this paper, we analyze five emotions from the dataset, namely, anger, fear, disgust, happiness, and sadness using narrowband spectrograms, their energies, and their Teager Energy Operator (TEO) profile. Observations prove that both speech features and cognition are important to analyze emotions.

The organization of the rest of the paper is as follows: Section II discusses the proposed work in detail. Section III gives the details of the experimental setup used for this study and discusses the results and inferences made during the analysis. Section IV concludes the paper along with potential future research directions.

II. Proposed Work

For designing any system capable of emotion recognition, selection of suitable features and classifiers is essential. In

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

this paper, we will only focus on features. The local and global features of Speech Emotion Recognition (SER) are analyzed in the following categories [7] - 1. Prosodic Features 2. Spectral Features 3. Voice Quality Features 4. Teager Energy Operator (TEO)-based features. In prosodic features, the 4 levels of computing prosody are - (a) linguistic intention-level, (b) articulatory-level, (c) acoustic realization-level, and (d) perceptual-level [4]. They include accent, stress, rhythm, tone, pitch, intonation, and loudness. Therefore, in this paper, we use prosodic features, in particular, loudness, energy, and pauses to analyse emotions. So far, for AESDD corpus not much work is done in analysing the features of audio samples. Mostly, this database has been used for comparative analysis. To that effect, in this paper, we performed energy-based analysis w.r.t. three techniques, namely, narrowband spectrogram analysis, energy (via $\boldsymbol{l}^2$ norm) of audio data, and TEO.

A. *Spectrographic Analysis*

Spectrogram is a visual representation of an acoustic signal, which gives time (X- axis), frequency (Y-axis), and amplitude (brightness) measure all in one graph. Energy distribution and pauses are also measured from this plot depending on the size of analysis window (a function used for Fourier-based spectral analysis), different levels of frequency/time resolution is achieved, i.e., a long window (atleast 2-3 pitch periods: 10-30 ms) gives good frequency resolution as harmonic lines (seen as horizontal striations in time-frequency domain) are resolved, this is also known as ***narrowband*** spectrogram. Similarly, a short window (duration less than a single pitch period: 4 ms) shows the formants of the vocal tract system with individual pitch periods appearing as vertical lines (striations), called ***wideband*** spectrogram. In this paper, we obtained the narrowband spectrograms for male and female speakers in 5 emotions using *wavesurfer tool*. Fig.1 and Fig.2 depicts the spectrograms obtained for a sentence uttered by a male and a female speaker, respectively. Narrowband spectrograms are employed rather than their wideband counterparts as they represent pitch source harmonics clearly in the form of horizontal striations. Moreover, as discussed in Section I, the pitch and its harmonics (along with its dynamic implementation) are important correlates of speech prosody to convey a particular emotion in speech. Pauses or silences between the words, energy content, and gaps between the horizontal striations are studied to identify the difference between each emotion. Another observation from Fig.1 and Fig.2 is that the gap between pitch source harmonics is higher in Fig.2 than in Fig.1 as females have higher F_0 (fundamental frequency) than males.

B. *Energy of Emotions*

The energy measurements are obtained using the standard discrete-time energy formula. In particular, for a discrete-time signal $x(n)$, energy is computed as E= $\sum_{n=\infty}^{-\infty} |x(n)|^2$. It should be noted that this energy, E is conserved in frequency domain (i.e., Parseval's energy equivalence) and also in the Short Time Fourier Transform (STFT) framework (proof given in Appendix A1, for the continuous-time version of STFT).

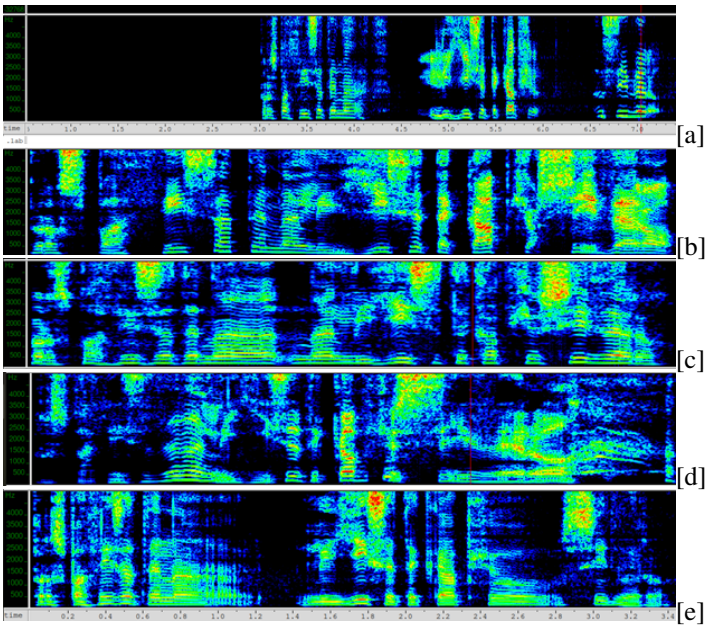


Fig. 1. Narrowband spectrograms of 5 emotions of a male speaker: (a) disgust, (b) anger, (c) fear, (d) happiness, and (e) sadness.

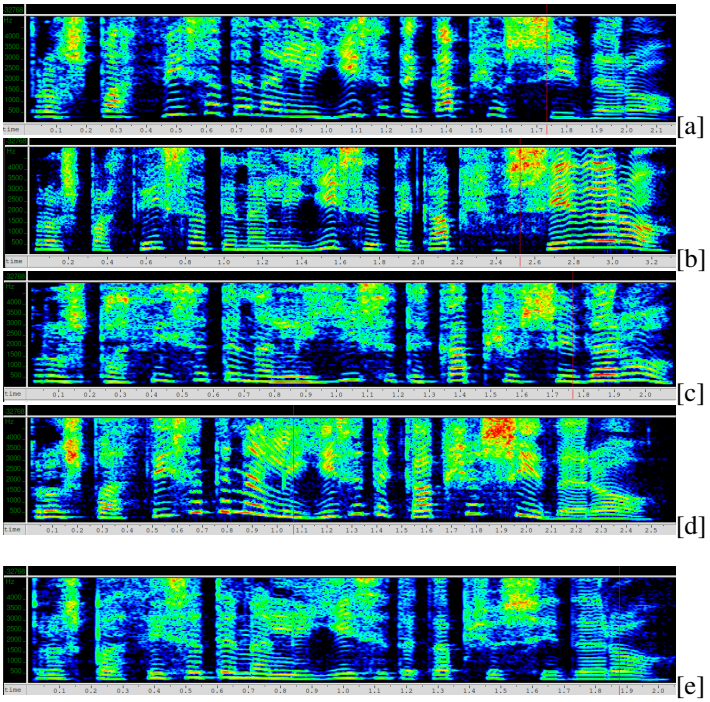


Fig. 2. Narrowband spectrogram of 5 emotions of a female speaker: (a) anger, (b) disgust, (c) fear, (d) happiness, and (e) sadness.

The sampling rate of the audio files is 44100 Hz. The audio sample is then frame-blocked and windowed (Hann window), so that the acoustic events in the speech sample can be represented as a compact set of speech parameters. Hop size is taken to be 10ms and Fast Fourier Transform (FFT) size is 2048 samples. Observations on how energy changes for different emotions and between gender is analysed. Fig 3-7 shows the plots of 5 speakers (3 females and 2 males) uttering the same sentence in 5 emotions, namely, anger, disgust, happiness, fear, and sadness, respectively. Energy is then represented in the

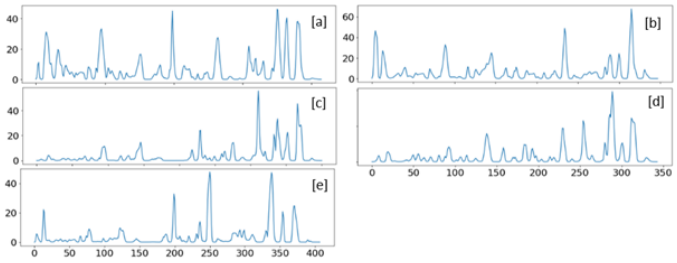


Fig. 3. l^2 Energy of ***anger*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) male1, (d) male2, and (e) female3).

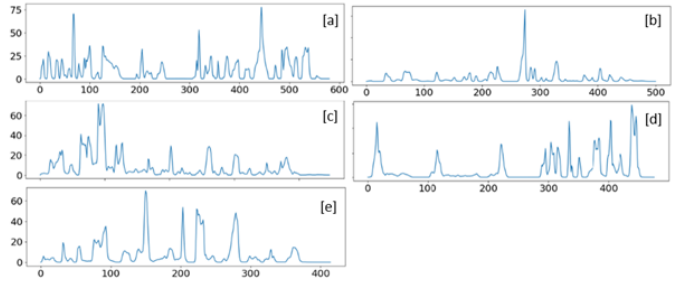


Fig. 4. l^2 Energy of ***disgust*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) male1, (d) male2, and (e) female3).

form of boxplot. Fig.8[a] describes how a male speaker (1 male speaker chosen of 2 male speakers) utters a sentence in 5 emotions (disgust, anger, fear, happiness, and sadness, respectively) and Fig.8[b], provides the same information for a female speaker (plot of 1 female speaker chosen from 3 female speakers). Boxplot provides a visual summary of data thereby enabling us to identify the average score, skewness, dispersion, and outliers in the dataset. Historgram is plotted for a male and female speaker (Fig.9[a] and Fig.9[b], respectively) to find the quantitative distribution of energy.

C. TEO for Analysis of Emotions

This non-linear feature or operator was introduced in 1990s by Teager and Kaiser [8]. Speech is produced by non-linear, vortex-airflow interaction in the vocal tract system. Stressful situation affects the muscle tension of the speaker that results in an alteration of the airflow during the production of the sound [9]. This is captured ***via*** TEO, in particular, Ψx(n)=

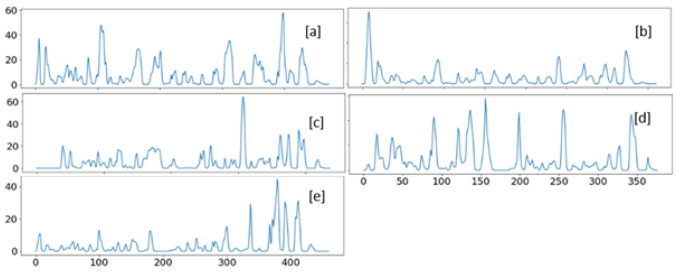


Fig. 5. l^2 Energy of ***happines*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) female3, (d) male1, (e) male2).

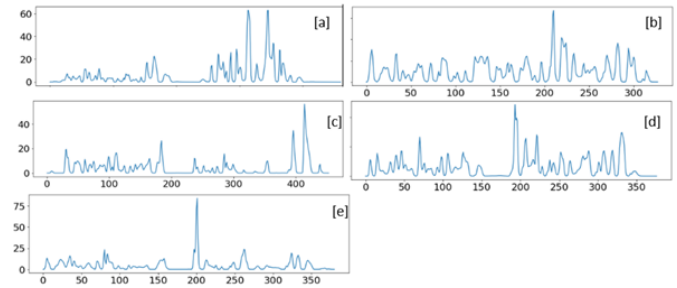


Fig. 6. l^2 Energy of ***fear*** by 5 speakers uttering the same sentence ((a) male1, (b) female2, (c) female3, (d) male2, and (e) female1).

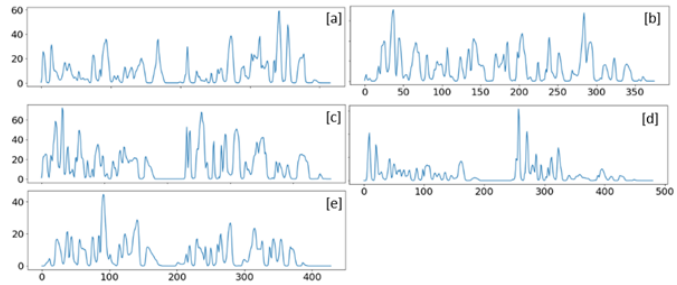


Fig. 7. l^2 Energy of ***sadness*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) female3, (d) male2, and (e) male1).

$x^2(n) - x(n+1)x(n-1)$, where Ψ is the Teager Energy Operator and $x(n)$ is the discrete time signal.

In Fig.10 and Fig.11, x-axis representing the frames and y-axis the amplitude, are the TEO plots obtained for a male speaker speaking a sentence in 5 emotions and a female speaker uttering the same sentence (as male) in 5 emotions respectively. The energy distribution, silence or pauses were observed to distinguish between the emotions.

III. Results and Inferences

A. Experimental Setup: Dataset Used

For this study, we have used AESDD corpus, developed in 2018, a publicly available greek dataset for SER [10]. This dataset contains utterances of ***acted*** emotional speech in Greek language [11]. The dataset contains 500 utterances of emotional speech for 5 actors (3 females and 2 males) X 5 emotions (anger, fear, disgust, happiness, and sadness) X 20 utterances [12]. This dataset is chosen as it is a relatively newer dataset with not much work reported using it. Another important reason for selecting this dataset is that each speaker has spoken the same sentence (total of 20 sentences) in different emotions and thus, providing a common base for analysis. In addition, analysis of emotions in male ***vs.*** female is also possible.

The results analyzed here takes atleast 5 sentences of each emotion and only some of them are plotted in this paper. The spectrographic analyzed from Fig.1 and Fig.2 had some distinct differences for each emotion, as shown in Fig.12. We also noted that in sadness, there is relatively least distance

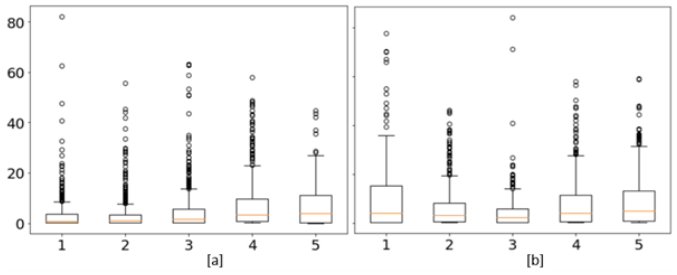


Fig. 8. l^2 Energy boxplot of [a] showing male-1 speaker, and [b] showing female-1 speaking the same sentence in 5 emotions- (1) disgust, (2) anger, (3) fear, (4) happy, and (5) sadness.

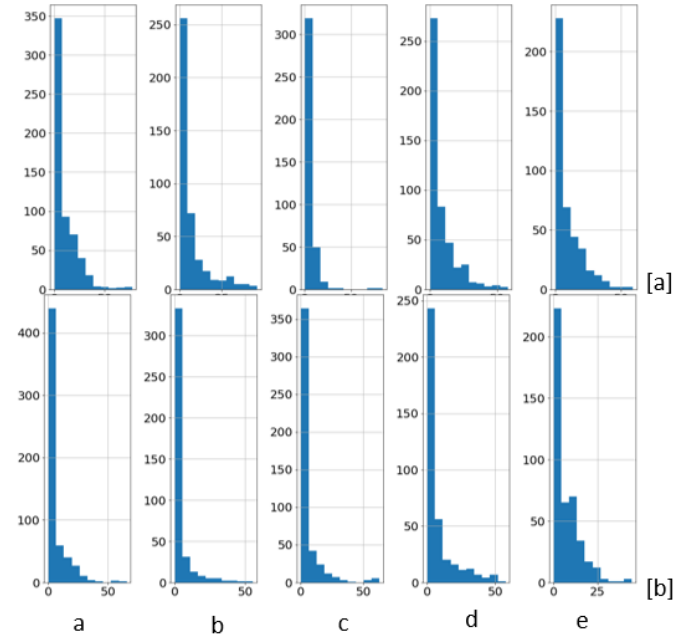


Fig. 9. l^2 Energy histogram of [a] showing male-1 speaker, and [b] showing female-1 speaking the same sentence in 5 emotions- (a) disgust, (b) anger, (c) fear, (d) happy, and (e) sadness.

between the harmonic striations, i.e., pitch source harmonics in both male and female. Hence, from Fig.12, it can be inferred that the highest energy content is available in anger and least in sadness. It should be noted that short pauses are highest in anger. This result is linked to the fact that when one gets angry, breathing is released in shorter and quicker puffs and the force with which air is released from the lungs to vocal tract system is also high and thus, showing these characteristics.

Energy plots in Fig.(3 to 7) providing the amplitude of audio signal is shown in Table II. It should be noted that apart from

TABLE I
SUMMARY OF SPEECH CORPORA USED IN THE EXPERIMENTS
(F - FEMALE, M - MALE)

Database	Speakers	Emotions
AESDD	5 (3F, 2M)	5 (anger, happy, disgust, fear, sadness)

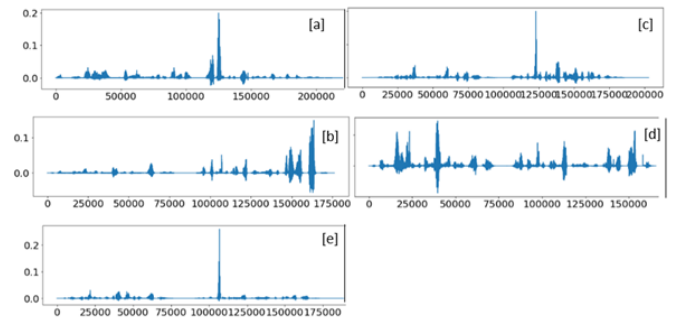


Fig. 10. TEO profile of male 1 uttering same sentence in 5 emotions- (a) disgust, (b) anger, (c) fear, (d) happy, and (e) sad.

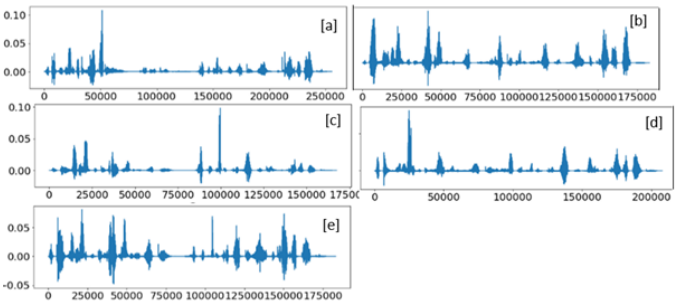


Fig. 11. TEO profile of female 1 uttering same sentence in 5 emotions- (a) disgust, (b) anger, (c) fear, (d) happy, and (e) sad.

the disgust emotion, females have higher amplitude peaks. This states that in females, loudness is seen more than males, while expressing emotions. We also see that fear and disgust have the highest amplitude in females and males, respectively. Surprisingly, we observed that anger has the least amplitude of all and we believe that this needs a further investigation. Boxplots of energy is shown in Fig.8(a) and Fig.8(b) for male and female, respectively. It can be observed that the median is lowest for disgust and highest for sadness in males,

Features	Anger	Disgust	Fear	Happy	Sadness
Pauses	Females: Highest (Short and Multiple)	Females: Lesser than anger	Females: Lesser than anger, but more than disgust	Females: Very less	Females: Clearly distinct
	Males: Highest (short and multiple)	Males: 2 huge gaps observed	Males: Least pauses found among males	Males: Very less	Males: High
Energy Concentration	Females: High throughout the sentence	Females: High concentration only at high frequencies	Females: Lesser compared to anger and disgust	Females: High but reduces towards end of the sentence	Females: Very less
	Males: High, particularly at higher frequencies	Males: High towards the end of the sentence	Males: High energy concentration at higher frequencies	Males: High throughout the sentence	Males: Very less

Fig. 12. Spectrographic analysis of emotions.

TABLE II
HIGHEST AMPLITUDES REACHED FOR EACH EMOTION BY MALE AND FEMALE SPEAKERS

Amplitude	Anger	Disgust	Fear	Happy	Sad
Females	60 dB	75 dB	80 dB	70 dB	65 dB
Males	55 dB	80 dB	60 dB	60 dB	60 dB

whereas, in females, fear has the lowest median with the other emotions having similar median values. All the emotions are positively skewed, implying most values are towards the lower bound. Sadness emotion has the most even spread among all the emotions and thus, having the least outliers. Spread/ dispersion is least in fear for females and anger for males. These findings of ours are in agreement with the original study reported in [13]. This implies, sadness and happiness has greater variability in speaking than the other emotions.

Histogram in Fig.9, is a right skewed distribution, i.e., data values fall at the lower range (as seen in the boxplots). Most of the values occur within the interval of 0-10 dB for all the emotions, and fear has the least range of 0-20 dB, within which almost all its values are covered, proving why fear has the least median in females.

TEO plots in (Fig.10 and Fig.11) shows that for male speaker 1, happiness followed by anger has the greatest energy profiles, and sadness has the lowest. However, for female speaker 1, the results obtained were surprisingly different from the other analysis, as sadness and anger has the greatest energy profiles and fear has the lowest energy profile. This result may be because of cognitive psychology [14]. Studies prove that emotions are not always instantaneous rather, it is a build up of feelings that are acquired over the period of time [14], [15]. This makes sense, as sometimes, a small trigger is enough to make a person emotional and thus, giving extreme reactions. Also, the dataset (AESDD) has its own errors [12]. All these cognitive factors affect the results obtained and we believe it requires much deeper investigations and remains as an open research problem.

IV. SUMMARY AND CONCLUSIONS

This study investigated the analysis of five emotions: anger, fear, disgust, happiness, and sadness. This is carried out by using prosodic features of speech, in particular, loudness (amplitude measure), energy and pauses or silences, and narrowband spectrograms. We also use the novel TEO on AESDD dataset to get energy profiles. It is found that anger and happiness show similar spectrographic characteristics, however, it can be distinguished by plotting their energy spectrum. Sadness has the most spread and balanced data values among the other emotions. Characteristics of disgust seemed very similar to anger. Distinct differences between male and female emotions can also be seen (boxplot- Fig.8) by acquiring their energy values. Along with speech features, the impact of cognition are equally important for recognizing a particular emotion. Future work can explore these features in pattern classifiers, developing a feature set that can distinguish anger and happiness better and extending this work to recognize emotions for disabled people. The investigations presented in this study are based on our current humble understanding of speech prosody in the literature, in particular, we believe that loudness, pitch, energy, intonation, speech spectrum dynamics, etc, all contribute ***implicitly*** to speech prosody. Since the definition of speech prosody is evolving everyday and authors believe that definition of speech emotion also continues to evolve (as it is a function of prosody), which may open interesting future research directions.

APPENDIX

A.1. Energy Conservation in STFT

The energy conservation in Short Time Fourier Transform (STFT) $f(t) \in L^2(R)$ is expressed as [16] :

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sf(u,\zeta)|^2 d\zeta du, \quad (1)$$

Here, u and ζ shows the time-frequency indices that vary across R and hence, covers the entire time-frequency plane. The reconstruction of the signal is given by

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Sf(u,\zeta) g(t-u) e^{i\zeta t} d\zeta du. \quad (2)$$

Applying Parseval's formula to eq. (2) w.r.t. to the integration in u, we get

$$Sf(u,\zeta) = e^{-iu\zeta} f * g_\zeta(u), \quad (3)$$

Here, $g_\zeta(t) = g(t)e^{i\zeta t}$. Hence, Fourier transform of $Sf(u,\zeta)$ is $\hat{f}(\omega_\zeta)\hat{g}(\omega)$. Furthermore, after applying the Plancherel's formula to eq. (1) gives

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sf(u,\zeta)|^2 du d\zeta =$$

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega+\zeta)\hat{g}(\omega)|^2 d\omega d\zeta. \quad (4)$$

Finally, the Plancheral formula and the Fubini theorem result in $\frac{1}{2\pi}\int_{-\infty}^{+\infty} |\hat{f}(\omega+\zeta)|^2 d\zeta = ||f||^2$, which validates STFT's energy conservation as demonstrated in eq. (1). It explains why the overall signal energy is the same as the time-frequency sum of the STFT.

ACKNOWLEDGEMENTS

The authors are thankful to the Ministry of Electronics and Information Technology (MeitY), New Delhi, Government of India, for sponsoring the project, "National Language Translation Mission (NLTM) : BHASHINI with the objective of Building Assistive Speech Technologies for the Challenged (Grant ID: 11(1)2022-HCC (TDIL)).

References

[1] M. Cabanac, "What is emotion?" *Behavioural processes*, vol. 60, no. 2, pp. 69–83, 2002.

[2] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[3] H. Schriefers and G. Vigliocco, "Speech production, psychology of [repr.]," in *International Encyclopedia of the Social & Behavioral Sciences (2nd ed) Vol. 23*, Elsevier, 2015, pp. 255–258.

[4] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *International Journal of Speech Technology (IJST)*, vol. 21, no. 1, pp. 93–120, 2018.

[5] A. A. Varghese, J. P. Cherian, and J. J. Kizhakkethottam, "Overview on emotion recognition system," in *International Conference on Soft-computing and Networks Security (ICSNS)*, IEEE, 2015, pp. 1–5.

[6] L. Alhinti, S. Cunningham, and H. Christensen, "Recognising emotions in dysarthric speech using typical speech data," in *INTERSPEECH 2020*, Shanghai, China, 2020, pp. 4821–4825.

[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[8] J. F. Kaiser, "On a simple algorithm to calculate the'energy'of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, USA, 1990, pp. 381–384.

[9] S. B. Alex, L. Mary, and B. P. Babu, "Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features," *Circuits, Systems, and Signal Processing*, vol. 39, no. 11, pp. 5681–5709, 2020.

[10] s. ntalampiras stavros, "Toward language-agnostic speech emotion recognition," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 7–13, Jan. 2020. DOI: https://doi.org/10.17743/jaes.2019.0045.

[11] *Multidisciplinary Media Mediated Communication, url : : http : // m3c . web . auth . gr / research / aesdd - speech - emotion-recognition/, Last Accessed : 2022-08-13*.

[12] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 457–467, 2018.

[13] S. Yildirim, M. Bulut, C. M. Lee, *et al.*, "An acoustic study of emotions expressed in speech," in *Eighth International Conference on Spoken Language Processing*, 2004.

[14] K. M. Galotti, *Cognitive Psychology In and Out of the Laboratory*. Sage Publications, 2017.

[15] S. Maitlis and H. Ozcelik, "Toxic decision processes: A study of emotion and organizational decision making," *Organization Science*, vol. 15, no. 4, pp. 375–393, 2004.

[16] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Modified Parametric Multichannel Wiener Filter for Low-latency Enhancement of Speech Mixtures with Unknown Number of Speakers

Ning Guo[1,*,†], Tomohiro Nakatani*, Shoko Araki*, and Takehiro Moriya*
* NTT Corporation, Japan
E-mail: tnak@ieee.org Tel: +81-774935134
† International Audio Laboratories Erlangen, Germany
E-mail: ning.guo@audiolabs-erlangen.de

***Abstract*—This paper introduces a novel low-latency online beamforming (BF) algorithm called the Modified Parametric Multichannel Wiener Filter (Mod-PMWF), which enhances speech mixtures with unknown and varying number of speakers. Although such conventional BFs as Linearly Constrained Minimum Variance BFs (LCMV BFs) can enhance a speech mixture, they typically require such speech mixture attributes as the number of speakers and the acoustic transfer functions (ATFs) from the speakers to the microphones. When mixture attributes are unavailable, estimating them by low-latency processing is challenging, hindering the application of BFs to the problem. In this paper, we overcome this problem by modifying a conventional Parametric Multichannel Wiener Filter (PMWF). The proposed Mod-PMWF can adaptively form a directivity pattern that enhances all the speakers in the mixture without explicitly estimating these attributes. Our experiments show the proposed BF's effectiveness in interference reduction ratios and subjective listening tests.**

I. Introduction

Microphone array processing has grown in importance due to the increased need for remote communication. Speech signals captured by microphones often suffer from degradation caused by background noise and reverberation. Microphone array processing improves the quality of speech signals by reducing noise and reverberation, thus enhancing speech perception and the performance of applications such as automatic speech recognition (ASR).

Researchers have successfully applied adaptive beamformers (BFs) [1]–[3] to enhance a single speech signal in a captured signal by low-latency processing. The speech's direction-of-arrival (DOA) is typically used to estimate the acoustic transfer function (ATF) from the speaker to the microphones (the steering vector) based on a plane-wave assumption, and the estimated ATF is used to optimize the BF [3], [4]. A method based on generalized eigenvalue decomposition (GEV) was developed to accurately estimate the ATF in reverberation (i.e., multipath environments) using the spatial covariance matrices (SCMs) of the signals [5], [6]. Next, a Parametric Multichannel Wiener Filter (PMWF) was proposed that allows the direct optimization of a BF from the SCMs without estimating the ATF [7], [8]. Furthermore, a PMWF can flexibly control the speech distortion and the noise reduction tradeoff.

In contrast, enhancing a mixture of speech signals, such as a conversation, with low-latency beamforming remains challenging. This is because the mixture may contain an unknown and varying number of speakers whose ATFs are unknown in general, and two or more speakers may talk simultaneously. In such a scenario, a BF needs to quickly and accurately adapt its directivity pattern to the unknown and varying number of speaker locations. If the number of speakers and the ATFs are given, a Linearly Constrained Minimum Variance (LCMV) BF can form a directivity pattern that enhances all the speakers in a mixture [1], [6], [9]–[11]. Similarly, a blind source separation (BSS) technique, e.g., Independent Vector Extraction (IVE) [12]–[14], can enhance a mixture by first extracting all the speakers and remixing them when the number is known. However, the requirements by these methods on the number of speakers and their ATFs are not acceptable in the assumed scenario. Alternatively, an Unknown Reference Multichannel Wiener Filter (UR-MWF) [15] can enhance a mixture without knowing the number of speakers or the ATFs. However, our experiments show that its denoising and dereverberation performance is limited.

To enable low-latency speech mixture enhancement, we propose a modified version of a PMWF, Mod-PMWF. A Mod-PMWF is defined by replacing the spatial covariance matrix (SCM) of a single target source in a conventional PMWF with the SCM of the mixture of the desired signals. Our mathematical analysis reveals that a Mod-PMWF has several desirable characteristics for low-latency speech mixture enhancement:

1) A Mod-PMWF achieves a weighted sum of the Minimum-Variance Distortionless Response (MVDR) BFs, each of which enhances each source in the mixture. It quickly adapts the weights of the MVDR BFs to enhance the sources with larger powers at each time-frequency (TF) point. Thus, a Mod-PMWF forms a directivity pattern that enhances the active sources at each TF point in the mixture.
2) We can reliably estimate a Mod-PMWF without knowing the number of sources in the mixture using a practical approximation technique. In addition, similar to a con-

[1]This work was supported by a JSPS summer program fellowship (No. SP21313)

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

ventional PMWF, we do not need to know the individual sources' ATFs for the estimation.

3) In terms of processing latency and computational efficiency, a Mod-PMWF is comparable to such conventional adaptive BFs as a PMWF.

In out experiments, we selected three conventional BFs, a GEV-based MVDR (GEV-MVDR) BF [6], Independent Single Component Extraction (ISCE) [13], and an UR-MWF since we believe that might perform mixture enhancement in certain senses without the number of sources or their ATFs. We also chose Independent Multiple Component Extraction (IMCE) [12] as an approach that can perform mixture enhancement with the prior knowledge of the number of sources. A comparison of our proposed Mod-PMWF with these conventional BFs using noisy reverberant speech mixtures confirms that Mod-PMWF outperformed the other BFs in terms of interference reduction ratios and MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) subjective listening tests [16].

In the following, we introduce a signal model and present the conventional PMWF in Sections II and III. Section IV presents the proposed Mod-PMWF, and Section V describes other conventional BFs that might be used for mixture enhancement. Section VI provides the experiment setup and results. Conclusions are drawn in Section VII.

II. Signal Model and Enhancement Goal

Suppose N speech signals are captured by M microphones with reverberation and diffuse noise. Let $x_{m,t,f}$ be the captured signal at the mth microphone and a TF point (t, f) in the short-time Fourier transform domain for $1 \leq t \leq T$ and $1 \leq f \leq F$, where T and F are the numbers of time frames and frequency bins, and let $(\cdot)^\top$ denote a non-conjugate transpose. Then the captured signal at all the microphones, $\mathbf{x}_{t,f} = [x_{1,t,f}, \ldots, x_{M,t,f}]^\top \in \mathbb{C}^M$, is modeled:

$$\mathbf{x}_{t,f} = \mathbf{d}_{t,f} + \mathbf{v}_{t,f}, \tag{1}$$

$$\mathbf{d}_{t,f} = \sum_{n=1}^{N} \mathbf{d}_{t,f}^{(n)} = \sum_{n=1}^{N} \mathbf{h}_f^{(n)} s_{t,f}^{(n)}, \tag{2}$$

where $\mathbf{d}_{t,f} \in \mathbb{C}^M$ is a mixture of speech signals $\mathbf{d}_{t,f}^{(n)}$ for all n. We assume that each speech signal, $\mathbf{d}_{t,f}^{(n)}$, which comprises a direct signal plus early reflections of the nth source, can be modeled by a product of a time-invariant ATF $\mathbf{h}_f^{(n)} \in \mathbb{C}^M$ and the nth clean source signal $s_{t,f}^{(n)} \in \mathbb{C}$. $\mathbf{v}_{t,f}$ is the sum of all the other signals, comprising the late reverberation of all the sources and the additive diffuse noise. In this paper, we deal with $\mathbf{d}_{t,f}$ as a desired signal to be obtained and $\mathbf{v}_{t,f}$ as an interference signal to be reduced. In addition, we assume that $\mathbf{d}_{t,f}^{(n)}$ for all n and $\mathbf{v}_{t,f}$ are mutually uncorrelated.

This paper assumes that both the late reverberation and the additive diffuse noise share the same spatial characteristics. We collectively model their sum as interference signal $\mathbf{v}_{t,f}$ that follows a stationary complex Gaussian distribution with mean vector $\mathbf{0}$ and SCM $\Phi_{\mathbf{v},f} \in \mathbb{C}^{M\times M}$:

$$p(\mathbf{v}_{t,f}) = \mathcal{N}(\mathbf{v}; \mathbf{0}, \Phi_{\mathbf{v},f}) \quad \text{where} \quad \Phi_{\mathbf{v},f} = E\{\mathbf{v}_{t,f}\mathbf{v}_{t,f}^{\mathsf{H}}\}, \tag{3}$$

where $E\{\cdot\}$ is an expectation function, and $(\cdot)^{\mathsf{H}}$ is a conjugate transpose.

Our goal in this paper is to estimate BF $\mathbf{W}_{t,f} \in \mathbb{C}^{M\times M}$ at each TF point that does not change speech mixture $\mathbf{d}_{t,f}$ and minimizes $\mathbf{v}_{t,f}$. With the BF, we obtain the following enhanced speech mixture, $\mathbf{y}_{t,f}\ (\approx \mathbf{d}_{t,f}) \in \mathbb{C}^M$:

$$\mathbf{y}_{t,f} = \mathbf{W}_{t,f}^{\mathsf{H}} \mathbf{x}_{t,f}. \tag{4}$$

III. Conventional BF for Single Source Extraction

This section briefly overviews a conventional BF (a PMWF) for single source extraction that we modify in the next section into a Mod-PMWF for mixture enhancement. Hereafter, we omit index f in the symbols since we apply the same processing separately in each frequency.

A. PMWF

When there is just one source in the observed signal (i.e., $N = 1$), PMWF $\mathbf{w}_{m,t} \in \mathbb{C}^M$ [7], which enhances $\mathbf{d}_t^{(1)}$ $(= \mathbf{h}^{(1)} s_t^{(1)})$ in Eq. (2) at the mth microphone, is defined as the one that minimizes the interference while constraining the speech distortion without exceeding allowable level σ:

$$\mathbf{w}_{m,t} = \arg\min_{\mathbf{w}} E\{|\mathbf{w}^{\mathsf{H}}\mathbf{v}_t|^2\} \quad \text{s.t.} \quad E\{|\mathbf{w}^{\mathsf{H}}\mathbf{d}_t^{(1)} - d_{m,t}^{(1)}|^2\} < \sigma.$$

Letting $\Phi_{\mathbf{d},t}^{(1)} = E\{\mathbf{d}_t^{(1)}(\mathbf{d}_t^{(1)})^{\mathsf{H}}\}$ be the SCM of $\mathbf{d}_t^{(1)}$ at t, solution $\mathbf{W}_t = [\mathbf{w}_{1,t}, \ldots, \mathbf{w}_{M,t}]$ is obtained at all the microphones:

$$\mathbf{W}_t = \left(\Phi_{\mathbf{d},t}^{(1)} + \gamma\Phi_{\mathbf{v}}\right)^{-1} \Phi_{\mathbf{d},t}^{(1)}. \tag{5}$$

Here $\gamma\ (\geq 0)$ is a weight that controls the tradeoff between the noise reduction and the speech distortion. This solution is also known as the Speech Distortion-Weighted MWF (SDW-MWF) [17]. Next, considering that $\Phi_{\mathbf{d},t}^{(1)}$ in Eq. (5) is rank-1 because it can be rewritten as $\Phi_{\mathbf{d},t}^{(1)} = E\{|s_t^{(1)}|^2\}\mathbf{h}^{(1)}(\mathbf{h}^{(1)})^{\mathsf{H}}$, we obtain a PMWF [7]:

$$\mathbf{W}_t^{\text{PMWF}} = \frac{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{d},t}^{(1)}}{\gamma + \lambda_{\mathbf{d},t}} \quad \text{where} \quad \lambda_{\mathbf{d},t} = \text{tr}\{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{d},t}^{(1)}\}, \tag{6}$$

and $\text{tr}\{\Phi\}$ is a matrix trace. Here, we can use γ as a parameter that controls the noise reduction level. For example, a PMWF with $\gamma = 0$ becomes an MVDR BF:

$$\mathbf{W}_t^{\text{MVDR}} = \frac{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{d},t}^{(1)}}{\lambda_{\mathbf{d},t}}. \tag{7}$$

A PMWF with $\gamma = 1$ also becomes a MWF.

An important advantage of a PMWF is that we can estimate it without estimating the ATF from the source to the microphones, i.e., $\mathbf{h}^{(1)}$, once we obtain $\Phi_{\mathbf{v}}$ and $\Phi_{\mathbf{d},t}^{(1)}$. Techniques for estimating $\Phi_{\mathbf{v}}$ and $\Phi_{\mathbf{d},t}^{(1)}$ have been proposed using voice activity detection [7] and TF masks estimated by a neural network [8].

IV. Proposed BF for Mixture Enhancement

This section presents a modified version of PMWF (Mod-PMWF) for mixture enhancement. We first define Mod-PMWF and analyze its characteristics focusing on how it performs mixture enhancement. Then we describe its implementation using a practical approximation technique.

A. Definition of Mod-PMWF

We define the Mod-PMWF by replacing the SCM of the target signal of a PMWF, i.e., $\Phi^{(1)}_{\mathbf{d},t}$ in Eq. (6), with the SCM of the desired mixture, i.e., $\Phi_{\mathbf{d},t} = E\{\mathbf{d}_t\mathbf{d}_t^{\mathsf{H}}\}$, at each time frame. The following is the Mod-PMWF formula:

$$\mathbf{W}_t^{\text{Mod-PMWF}} = \frac{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{d},t}}{\gamma + \lambda_{\mathbf{d},t}} \quad \text{where} \quad \lambda_{\mathbf{d},t} \quad = \text{tr}\{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{d},t}\}. \tag{8}$$

This BF is different from a PMWF in that $\Phi_{\mathbf{d},t}$ contains the characteristics not only of a single target source but of all the sources included in the mixture.

Below we explain how a Mod-PMWF can perform mixture enhancement. Then, in Sections IV-B and IV-C, we describe how we can reliably estimate a Mod-PMWF using an approximation technique.

1) Analysis of Mod-PMWF: Based on the assumption that $\mathbf{d}_t^{(n)}$ for all n are mutually uncorrelated, SCM $\Phi_{\mathbf{d},t}$ can be expanded:

$$\Phi_{\mathbf{d},t} = \sum_{n=1}^{N} \Phi^{(n)}_{\mathbf{d},t}, \tag{9}$$

where $\Phi^{(n)}_{\mathbf{d},t} = E\{\mathbf{d}_t^{(n)}(\mathbf{d}_t^{(n)})^{\mathsf{H}}\}$ is an unknown SCM of $\mathbf{d}_t^{(n)}$. Substituting Eq. (9) into Eq. (8) followed by a certain mathematical manipulation yields:

$$\mathbf{W}_t^{\text{Mod-PMWF}} = \sum_{n=1}^{N} \mu_{n,t}\mathbf{W}_t^{\text{MVDR}(n)}. \tag{10}$$

where $\mathbf{W}_t^{\text{MVDR}(n)}$ is an MVDR BF that enhances the nth source $\mathbf{d}_t^{(n)}$ by reducing interference $\mathbf{v}_t$. Each MVDR BF is defined based on a PMWF with $\gamma = 0$ in Eq. (7):

$$\mathbf{W}_t^{\text{MVDR}(n)} = \frac{\Phi_{\mathbf{v}}^{-1}\Phi^{(n)}_{\mathbf{d},t}}{\lambda^{(n)}_{\mathbf{d},t}} \quad \text{where} \quad \lambda^{(n)}_{\mathbf{d},t} = \text{tr}\{\Phi_{\mathbf{v}}^{-1}\Phi^{(n)}_{\mathbf{d},t}\}. \tag{11}$$

$\mu_{n,t}$ in Eq. (10) is a time-varying weight:

$$\mu_{n,t} = \frac{\lambda^{(n)}_{\mathbf{d},t}}{\gamma + \lambda_{\mathbf{d},t}} \quad \text{where} \quad \lambda_{\mathbf{d},t} = \sum_{n=1}^{N} \lambda^{(n)}_{\mathbf{d},t}. \tag{12}$$

According to Eq. (10), a Mod-PMWF is a weighted sum of the MVDR BFs, each of which is designed to enhance each source in the mixture. The weights are determined by $\lambda^{(n)}_{\mathbf{d},t}$ for all n, which are roughly proportional to the power of $\mathbf{d}_t^{(n)}$.

In the following, we look into more details of the BF characteristics for the cases with $\gamma = 0$ and $\gamma > 0$.

2) Behavior with $\gamma = 0$: When $\gamma = 0$, the Mod-PMWF has the following characteristics.

- Since the weights sum to one according to Eqs. (12), Eq. (10) becomes a weighted average of the MVDR BFs. Because all the MVDR BFs reduce $\mathbf{v}_t$, their weighted average, i.e., the Mod-PMWF, should also reduce $\mathbf{v}_t$.
- The MVDR BFs that compose the weighted average cover all the sources in $\mathbf{d}_t$ given $\Phi_{\mathbf{d},t}$, even without explicitly specifying the number of sources.
- Each MVDR BF follows the unknown ATF of each source included in $\mathbf{d}_t$. Thus, it can preserve its corresponding source without distortion.
- Since the weights of the BFs are roughly proportional to the powers of the respective sources, the Mod-PMWF can quickly adapt the weights to focus on active speakers at each TF point, i.e., sources with larger powers.

In summary, each MVDR BF that composes the Mod-PMWF preserves each source in the mixture and reduces the interference. The Mod-PMWF rapidly controls the weights of the MVDR BFs depending on the change in the relative powers of the sources. In this sense, a Mod-PMWF with $\gamma = 0$ can achieve mixture enhancement.

3) Behavior with $\gamma > 0$: With $\gamma > 0$, a Mod-PMWF is equivalent to multiplying a factor η_t to the output of the Mod-PMWF with $\gamma = 0$:

$$\eta_t = \frac{\lambda_{\mathbf{d},t}}{\gamma + \lambda_{\mathbf{d},t}}. \tag{13}$$

Based on analogy from the conventional PMWF, we expect this factor to work as a single channel Wiener filter that will further reduce the remaining interference in the output of the Mod-PMWF with $\gamma = 0$. We also confirmed such behavior of the Mod-PMWF with $\gamma > 0$ based on our preliminary experiments. A more thorough analysis of this behavior will be included in future work.

B. Practical approximation for Mod-PMWF

Although accurate estimation of $\Phi_{\mathbf{d},t}$ is crucial for Mod-PMWF, it is challenging under general recording conditions. To avoid this difficulty, this paper approximates $\Phi_{\mathbf{d},t}$ by the SCM of captured signal $\Phi_{\mathbf{x},t} = E\{\mathbf{x}_t\mathbf{x}_t^H\}$, disregarding $\Phi_{\mathbf{v}}$ in $\Phi_{\mathbf{x},t}$:

$$\Phi_{\mathbf{d},t} \simeq \Phi_{\mathbf{x},t} (= \Phi_{\mathbf{d},t} + \Phi_{\mathbf{v}}). \tag{14}$$

Unlike $\Phi_{\mathbf{d},t}$, $\Phi_{\mathbf{x},t}$ can be obtained easily and accurately from the given $\mathbf{x}_t$. In addition, this approximation does not significantly affect the Mod-PMWF performance, as discussed below.

With this approximation, the Mod-PMWF in Eq. (8) is rewritten:

$$\mathbf{W}_t^{\text{Approx-Mod-PMWF}} = \frac{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{x},t}}{\gamma + \lambda_{\mathbf{x},t}} \quad \text{where} \quad \lambda_{\mathbf{x},t} = \text{tr}\{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{x},t}\}. \tag{15}$$

Similar to Eq. (10), the approximated Mod-PMWF can be expanded using $\Phi_{\mathbf{x},t} = \Phi_{\mathbf{d},t} + \Phi_{\mathbf{v}}$:

$$\mathbf{W}_t^{\text{Approx-Mod-PMWF}} = \mu'_{\mathbf{d},t}\mathbf{W}_t^{\text{Mod-PMWF}} + \mu'_{\mathbf{v},t}\frac{\mathbf{I}_M}{M}, \tag{16}$$

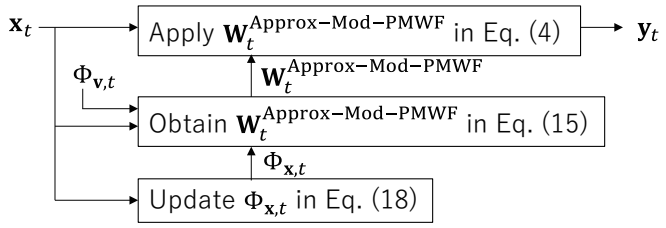


Fig. 1: Processing flow of Mod-PMWF at a time frame t and a frequency. Interference SCM $\Phi_{\mathbf{v},t}$ is assumed to be given in advance or estimated by online processing (Section VI-E).

where $\mathbf{I}_M \in \mathbb{R}^{M\times M}$ is an identity matrix. $\mu'_{\mathbf{d},t}$ and $\mu'_{\mathbf{v},t}$, which are time-varying weights roughly proportional to the power of $\mathbf{d}_t$ and $\mathbf{v}_t$, are defined as

$$\mu'_{\mathbf{d},t} = \frac{\gamma + \lambda_{\mathbf{d},t}}{\gamma + \lambda_{\mathbf{d},t} + \lambda_{\mathbf{v}}} \quad \text{and} \quad \mu'_{\mathbf{v},t} = \frac{\lambda_{\mathbf{v}}}{\gamma + \lambda_{\mathbf{d},t} + \lambda_{\mathbf{v}}}. \quad (17)$$

Here $\lambda_{\mathbf{v}} = M(= \text{tr}\{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{v}}\})$.

Equation (16) is a weighted average of the Mod-PMWF (without approximation) and a filter that reduces the gain of the captured signal by a factor of M. Note that $\mathbf{W}_t^{\text{Mod-PMWF}}$ in Eq. (16) is accurate because it is based on an accurate $\Phi_{\mathbf{d},t}$, which is included in $\Phi_{\mathbf{x},t}$. This means that using the approximated Mod-PMWF in Eq. (15) is equivalent to applying an accurate Mod-PMWF and adding the captured signal with a reduced level. Although this approximation slightly reduces the interference reduction effect of Mod-PMWF, it can precisely preserve the speech mixture. We use this approximation throughout out experiments.

C. Implementation of Mod-PMWF

Figure 1 shows the processing flow of Mod-PMWF. To estimate a Mod-PMWF with the approximation, we only need to estimate $\Phi_{\mathbf{x},t}$ and $\Phi_{\mathbf{v},t}$ and calculate Eq. (15) at each TF point. First, according to the convention of low-latency online adaptive BFs, we can estimate $\Phi_{\mathbf{x},t}$ by online processing as the time average of $\mathbf{x}_t\mathbf{x}_t^{\mathsf{H}}$ using forgetting factor β ($0 < \beta \leq 1$). $\Phi_{\mathbf{x},t}$ can be recursively updated:

$$\Phi_{\mathbf{x},t} = \beta\Phi_{\mathbf{x},t-1} + (1-\beta)\mathbf{x}_t\mathbf{x}_t^{\mathsf{H}}. \quad (18)$$

Note that active sources may differ at different time frames, and sources that are not included in current frame t may be included in previously captured signals in Eq. (18). Their influence remains in $\Phi_{\mathbf{x},t-1}$. Using a small β minimizes such influence.

The interference SCM $\Phi_{\mathbf{v},t}$ can be estimated, e.g., from noise recorded separately in advance or from a captured signal during speech absent periods. We can also use the same techniques proposed for conventional PMWFs [7], [8].

V. Potentially Applicable Conventional BFs to Mixture Enhancement

This section describes several conventional BFs that we believe may be used for mixture enhancement in certain senses. We also compare the computational time complexity and latency of the BFs with our proposed Mod-PMWF. In the next section, we experimentally compare the BFs with our proposed Mod-PMWF.

A. GEV-MVDR BF

The first conventional BF is a GEV-MVDR BF, which we present as a technique for single source enhancement and explain how to use for mixture enhancement.

When the number of sources in the observed signal is one (i.e., $N = 1$), an ATF at each time frame $\mathbf{h}^{(1)}$ for the source can be estimated based on GEV [6]:

$$\mathbf{h}_t^{(1)} = \Phi_{\mathbf{v}}\mathbf{u}_t \quad \text{where} \quad \mathbf{u}_t = \text{MaxEig}\{\Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{x},t}\}, \quad (19)$$

where $\text{MaxEig}\{\Phi\}$ extracts an eigenvector of Φ corresponding to the maximum eigenvalue. Considering $\Phi_{\mathbf{d},t}^{(1)} = E\{|s_t^{(1)}|^2\}\mathbf{h}_t^{(1)}(\mathbf{h}_t^{(1)})^{\mathsf{H}}$ in Eq. (7) and substituting Eq. (19) into Eq. (7), we obtain a GEV-MVDR BF:

$$\mathbf{W}_t^{\text{GEV-MVDR}} = \frac{\mathbf{u}_t\mathbf{u}_t^{\mathsf{H}}\Phi_{\mathbf{v}}}{\text{tr}\{\mathbf{u}_t\mathbf{u}_t^{\mathsf{H}}\Phi_{\mathbf{v}}\}}. \quad (20)$$

For low-latency processing, we can calculate $\Phi_{\mathbf{x},t}$ in Eq. (19) using Eq. (18) and obtain $\mathbf{u}_t$ in Eq. (20) in a computationally efficient way using a power method for calculating $\text{MaxEig}\{\cdot\}$ in Eq. (19) [18], [19]. This iterates the following update after first setting $\mathbf{u}_t = \mathbf{u}_{t-1}$ at each time frame:

$$\mathbf{u}_t \leftarrow \Phi_{\mathbf{v}}^{-1}\Phi_{\mathbf{x},t}\mathbf{u}_t/|\mathbf{u}_t|. \quad (21)$$

In general, the power method converges very quickly, and one iteration was sufficient in our experiments.

1) Application of GEV-MVDR BF to mixture enhancement: As shown in a previous work [20], a GEV-MVDR BF can also be viewed as a realization of a Maximum Signal-to-Noise Ratio (MaxSNR) BF [2], [21], [22], a property that can be used for mixture enhancement.

Regardless whether the captured signal contains a single source or multiple sources, we can define MaxSNR BF $\mathbf{w}_t^{\text{MaxSNR}} \in \mathbb{C}^M$ as one that maximizes the signal's SNR, i.e., the power ratio of desired signal $\mathbf{d}_t$ to interference $\mathbf{v}_t$:

$$\mathbf{w}_t^{\text{MaxSNR}} = \arg\max_{\mathbf{w}} \frac{E\{|\mathbf{w}^{\mathsf{H}}\mathbf{d}_t|^2\}}{E\{|\mathbf{w}^{\mathsf{H}}\mathbf{v}_t|^2\}}, \quad (22)$$

$$= \arg\max_{\mathbf{w}} \frac{E\{|\mathbf{w}^{\mathsf{H}}\mathbf{x}_t|^2\}}{E\{|\mathbf{w}^{\mathsf{H}}\mathbf{v}_t|^2\}}, \quad (23)$$

where we used $E\{|\mathbf{w}^{\mathsf{H}}\mathbf{x}_t|^2\} = E\{|\mathbf{w}^{\mathsf{H}}\mathbf{d}_t|^2\} + E\{|\mathbf{w}^{\mathsf{H}}\mathbf{v}_t|^2\}$ to obtain the above second line. The solution to Eq. (23) is obtained based on GEV: $\mathbf{w}_t^{\text{MaxSNR}} = b^*\mathbf{u}_t$, where b is an indefinite constant, $(\cdot)^*$ is a complex conjugate, and $\mathbf{u}_t$ is from Eq. (19). Let $\mathbf{b} \in \mathbb{C}^M$ be a vector composed of M different realizations of b and let $\mathbf{W}_t = \mathbf{u}_t\mathbf{b}^{\mathsf{H}}$ be a matrix realization of a MaxSNR BF based on $\mathbf{b}$. Then the MaxSNR BF with $\mathbf{b}$ defined below is identical to the GEV-MVDR BF in Eq. (20):

$$\mathbf{W}_t^{\text{MaxSNR}} = \mathbf{u}_t\mathbf{b}^{\mathsf{H}} \quad \text{where} \quad \mathbf{b} = \frac{\Phi_{\mathbf{v}}\mathbf{u}_t}{\text{tr}\{\mathbf{u}_t\mathbf{u}_t^{\mathsf{H}}\Phi_{\mathbf{v}}\}}. \quad (24)$$

The above analysis indicates that the GEV-MVDR BF in Eq. (20) works as a MaxSNR BF even when the observed

signal contains two or more sources. In other words, the GEV-MVDR BF may enhance the mixture of the speech signals, i.e., $\mathbf{d}_t$ in Eq. (1), in the MaxSNR sense regardless of the number of sources.

B. Two versions of Independent Component Extraction (ICE): ISCE and IMCE

Next, we chose Independent Component Extraction (ICE) since it is applicable for mixture enhancement. ICE is a variation of a popular BSS technique, IVE [12]–[14]. Even without prior knowledge of the ATFs of the sources, both IVE and ICE can extract a given number ($N \geq 1$) of sources based on an assumption that they are mutually independent. We here adopt ICE because our preliminary experiments identified its superiority over IVE for mixture enhancement. The difference between ICE and IVE is their source models. Although IVE's model specifies the source's spectral characteristics over all the frequencies, ICE's model separately specifies the characteristics at each frequency. We defined the latter model at each frequency by a complex Gaussian with a mean of 0 and a time-varying variance. The optimization algorithm for an online ICE can be easily obtained from that of an online IVE [14].

We below present two different usages of ICE for mixture enhancement, referred to as ISCE and IMCE, respectively.

1) Application of ISCE to mixture enhancement: When we set the number of sources to be extracted as $N = 1$ for ICE, we call it Independent Single Component Extraction (ISCE), which this paper uses for mixture enhancement by utilizing speech sparseness. With speech sparseness, we assume that the speakers are not simultaneously active at each TF point. With this assumption, which is based on the spectral characteristics of speech and has proven effective for source separation purposes [23]–[25], we can simplify a mixture enhancement task to a single source extraction task at each TF point. Thus, we can use ISCE for mixture enhancement by simply applying it to a mixture with a quick adaptation over time frames.

2) Application of IMCE to mixture enhancement: When we assume that the number of sources is given ($N \geq 2$) and use ICE to extract all of them from a mixture, we call it Independent Multiple Component Extraction (IMCE). By remixing the extracted sources, we can obtain an enhanced mixture. Unlike the other methods, this method requires the prior knowledge of the number of sources; we just use IMCE for reference.

C. UR-MWF

We can also use a version of a conventional MWF called UR-MWF [15] for mixture enhancement. MWF is defined as BF $\mathbf{W}_t^{\text{MWF}}$ that gives the minimum mean square error (MMSE) estimate of desired mixture $\mathbf{d}_t$:

$$\mathbf{W}_t^{\text{MWF}} = \arg\min_{\mathbf{W}} E\{\|\mathbf{d}_t - \mathbf{W}^{\mathsf{H}}\mathbf{x}_t\|_2^2\}. \tag{25}$$

Assuming $\Phi_{\mathbf{x},t} = E\{\mathbf{d}_t\mathbf{d}_t^{\mathsf{H}}\} + \Phi_{\mathbf{v}}$, we obtain UR-MWF:

$$\mathbf{W}_t^{\text{UR-MWF}} = \Phi_{\mathbf{x},t}^{-1}(\Phi_{\mathbf{x},t} - \Phi_{\mathbf{v}}). \tag{26}$$

Based on Eq. (26), UR-MWF can enhance a mixture given $\Phi_{\mathbf{x},t}$ and $\Phi_{\mathbf{v}}$ without the number of sources or their ATFs.

TABLE I: Computational time complexity for adapting and applying a BF to each TF point: Complexity of Mod-PMWF, GEV-MVDR, and UR-MWF is estimated assuming that we recursively update $\Phi_{\mathbf{v}}$ and its inverse, e.g., during speech absent periods, as in a previous work [19].

Mod-PMWF	GEV-MVDR	ISCE	IMCE	UR-MWF
$O(M^2)$	$O(M^2)$	$O(M^2)$	$O(NM^2)$	$O(M^2)$

D. Comparison of computational time complexity and latency

Table I shows the computational time complexity of the BFs and the Mod-PMWF. All the BFs (except for IMCE) have identical complexity $O(M^2)$. The complexity increases to $O(NM^2)$ for IMCE to extract N sources.

All the BFs update their coefficients at each time frame using current and past time frames. Thus, their algorithmic delay (or latency) corresponds to the window length minus the window shift of the overlap-add synthesis.

VI. Experiments

We evaluated our proposed Mod-PMWF in comparisons with other conventional BFs using signal level metrics and a MUSHRA listening test.

A. Dataset

For the evaluation, we prepared a simulated noisy reverberant mixture dataset containing 100 samples of two speakers from ATR digital speech database set B [26]. To reverberate each speaker's utterances, we used room impulse responses (RIRs) measured with a cubic array of eight microphones in a room with a T60 of 380 ms. Each edge of the cubic array was 4 cm, and a microphone was equipped at each vertex of the array. We used 5 of 8 microphones in the experiments. The two speakers were located 1 m from the array on opposite sides. We then mixed two speech signals starting simultaneously with babble noise recorded by the same array in the same room. We created two versions of the dataset by setting the Reverberant speech Mixture to Noise Ratio (RMNR) to 20 dB and 10 dB. The sampling frequency was set at 16 kHz.

B. Analysis condition

Throughout the experiments, we set $\gamma = 0$ for Mod-PMWF to evaluate its MVDR behavior. Forgetting factor β for updating $\Phi_{\mathbf{x},t}$ in Eq. (18) decides how quickly the algorithm adapts to the change in the signals' spatial characteristics. We chose $\beta = 0.99$ for UR-MWF, 0.96 for ISCE and IMCE, and 0.85 for Mod-PMWF and GEV-MVDR based on preliminary experiments. We set the window length and shift at 20 ms and 5 ms.

In Sections VI-C and VI-D, we assume that interference SCMs $\Phi_{\mathbf{v}}$ can be estimated in advance and estimated them from a part of the measured babble noise not used for generating the mixtures. Then we evaluated the Mod-PMWF with no prior knowledge of $\Phi_{\mathbf{v}}$ in Section VI-E, where we estimated $\Phi_{\mathbf{v}}$ by online processing from observed signals [19] using the estimated speech presence probability [27].

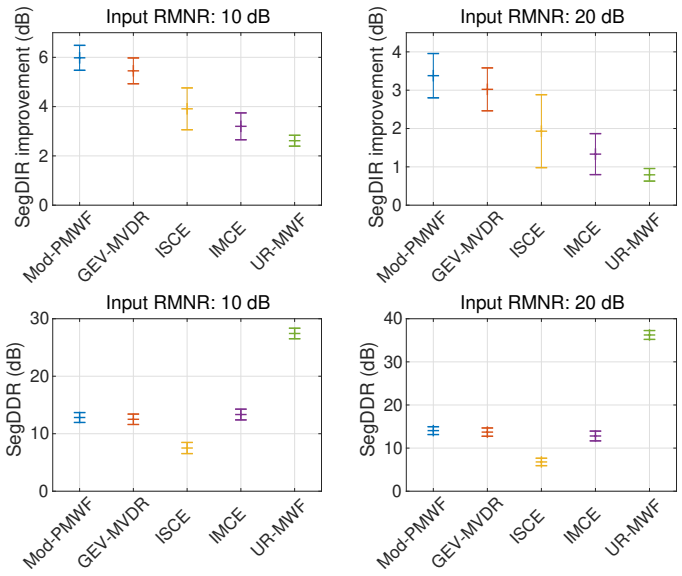


Fig. 2: SegDIR improvements and SegDDRs averaged over 100 samples: Error bars indicate standard deviation. Mod-MPDR, Gev-MVDR, and UR-MWF used interference SCMs separately estimated in advance.

C. Evaluation results using signal level metrics

We defined two signal level metrics for the evaluation [28], [29]: the Segmental Desired to Interference Ratio (SegDIR) and the Segmental Desired to Distortion Ratio (SegDDR). The former measured the degree of interference reduction; the latter measured the degree of signal distortion introduced to the enhanced speech. Let $d_m(i)$ and $v_m(i)$ be the ith samples of the desired mixture and the interference signal in the time domain at the mth microphone, and let $\hat{d}_m(i)$ and $\hat{v}_m(i)$ be samples obtained by applying a BF to $d_m(i)$ and $v_m(i)$. To determine desired signal $d_m(i)$, we set the duration of the early reflections at 50 ms. We defined the metrics:

$$\text{SegDIR} = \frac{10}{|\mathcal{A}|}\sum_{\tau\in\mathcal{A}}\log_{10}\frac{\sum_{m=1}^{M}\sum_{i=\tau\delta+1}^{\tau\delta+\delta}|\hat{d}_m(i)|^2}{\sum_{m=1}^{M}\sum_{i=\tau\delta+1}^{\tau\delta+\delta}|\hat{v}_m(i)|^2},$$

$$\text{SegDDR} = \frac{10}{|\mathcal{A}|}\sum_{\tau\in\mathcal{A}}\log_{10}\frac{\sum_{m=1}^{M}\sum_{i=\tau\delta+1}^{\tau\delta+\delta}|d_m(i)|^2}{\sum_{m=1}^{M}\sum_{i=\tau\delta+1}^{\tau\delta+\delta}|\hat{d}_m(i)-d_m(i)|^2},$$

where $\delta = 800$ samples ($= 50$ ms) is the segment length, τ is a segment index, $\mathcal{A}$ is a set of segments where at least one speech signal is active, and $|\mathcal{A}|$ is the number of segments in $\mathcal{A}$.

Figure 2 shows the evaluation results. We show SegDIR as its relative improvement from the captured signal and SegDDR as its absolute value.

First, when we compare the SegDIRs obtained by the BFs, the proposed Mod-PMWF had the best performance. In particular, Mod-PMWF significantly outperformed ISCE, IMCE, and UR-MWF. When we compare the SegDDRs obtained by the BFs, Mod-PMWF outperformed ISCM and was comparable with GEV-MVDR and IMCE. In contrast, UR-MWF achieved significantly better SegDDR results than all the others, although its SegDIR improvement was minimal.

We draw the following implications from these results.

1) Although Mod-PMWF and GEV-MVDR did not use prior knowledge of the number of sources, they achieved comparable SegDDRs with IMCE, which does use the knowledge of the number of sources. This suggests that Mod-PMWF and GEV-MVDR effectively enhanced the mixture based on their unique mechanisms described in Sections IV-A1 and V-A1.
2) For improving SegDIR, the sparseness assumption used for ISCE was effective, to some extent, in comparison with IMCE, but insufficient in comparison with Mod-PMWF and GEV-MVDR. The assumption also introduced more speech distortion than the other BFs.
3) The inferior SegDIR obtained by IMCE is probably caused by remixing the two separated signals. The remixing combines the interference that remained in each separated signal. In contrast, Mod-PMWF effectively reduced the interference while enhancing the mixture probably because it adaptively controls the BF weights to enhance only the active sources at each TF point.

In contrast, comparing UR-MWF's effectiveness with the other BFs is difficult based only on the above results.

To more reliably compare the proposed BF with the other BFs, including UR-MWF, we conducted a subjective listening test.

D. Evaluation results with MUSHRA listening test

We conducted a MUSHRA listening test [16] to evaluate the perceived speech quality. We selected two speech mixtures composed of either two male or two female speakers and tested them under each RMNR condition. Each sample had seven signals to rate, including the desired mixture as the (hidden) reference and the noisy reverberant observation as the anchor. Ten experienced listeners participated in the test and concentrated on the audio's overall quality including both the noise level and the speech degradation.

Figure 3 shows the result. The two sub-figures show the results under input RMNRs of 20 dB and 10 dB and both exhibit almost the same tendency. First, the participants rated the high-quality reference as 100 points and the test conditions between 40-80 points, confirming our choice of MUSHRA as appropriate based on ITU-R recommendation BS. 1534. In addition, the improvements obtained by all the BFs were statistically significant from the anchor, i.e., the noisy observation received the lowest scores. Thus our anchor indicates how the methods compare to known audio quality levels.

Next, when comparing the results among BFs, the proposed Mod-PMWF significantly outperformed all the other BFs, including GEV-MVDR and UR-MWF. The result was better correlated with SegDIR than with SegDDR. One reason might be that the speech distortion is, to some extent, masked by active speech and noise residual and thus less annoying to the listeners than the noise residual.

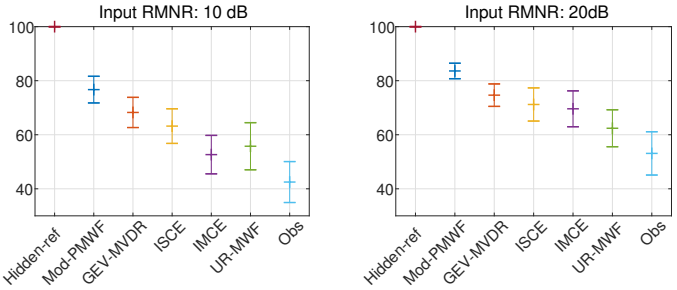


Fig. 3: MUSHRA test results: Error bars indicate 95% confidence based on t-distribution test. Mod-MPDR, GEV-MVDR, and UR-MWF used interference SCMs separately estimated in advance.

E. Evaluation with no prior knowledge of interference SCM

Finally, we evaluated Mod-PMWF without any prior knowledge of the interference SCM. We recursively updated interference SCM $\Phi_{\mathbf{v},t,f}$ at each TF point:

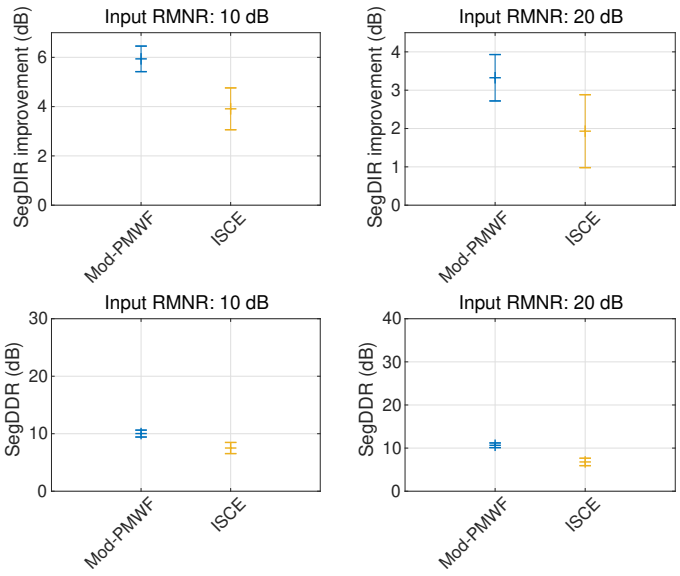
$$\Phi_{\mathbf{v},t,f} = \alpha' \Phi_{\mathbf{v},t-1,f} + (1-\alpha')\mathbf{x}_{t,f}\mathbf{x}_{t,f}^H, \tag{28}$$

$$\alpha' = \alpha + (1-\alpha)q_{t,f}, \tag{29}$$

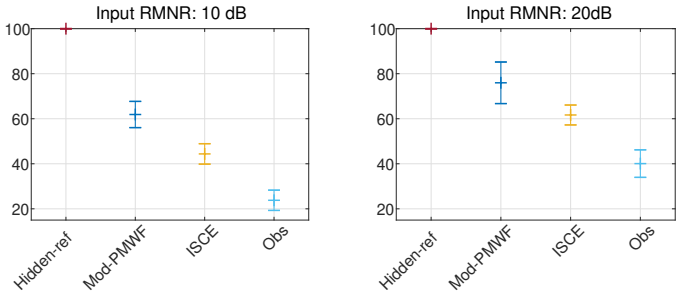
where $\alpha = 0.9998$ is a forgetting factor and $q_{t,f}$ is a speech presence probability estimated recursively by maximum likelihood estimation [27].

Figures 4 (a) and (b) show the SegDIR improvements, the SegDDRs, and the MUSHRA test results. We selected two methods, a Mod-PMWF based on online interference SCM estimation and an ISCE based on BSS, mainly to elaborate the reliability of the MUSHRA test. For it, we also added a mixture sample composed of male and female speakers. Totally we had three samples for each RMNR condition. Eight experienced listeners participated in this test.

Mod-PMWF significantly outperformed ISCE in terms of all the metrics. The overall tendency of the results is almost identical to that in Figs. 2 and 3 except that the SegDDRs obtained by Mod-PMWF were slightly lower than Fig. 2. These results show that Mod-PMWF effectively performed mixture enhancement even when no prior knowledge on the interference SCMs was available.



(a) SegDIR improvements and SegDDRs averaged over 100 samples: Error bars indicate standard deviation.



(b) MUSHRA test results: Error bars indicate 95% confidence based on t-distribution test.

Fig. 4: Evaluation without prior knowledge of interference SCM. The interference SCM was estimated by online processing for Mod-PMWF.

VII. Conclusion

We proposed a Mod-PMWF for low-latency online speech mixture enhancement in noisy reverberant environments. We showed mathematically that Mod-PMWF is equivalent to a weighted sum of MVDR BFs, each of which can preserve each source included in the captured signal and the time-varying weights quickly adapt at each time frame to put larger weights on sources with larger powers. We estimated the Mod-PMWF by low-latency online processing without relying on the prior knowledge of the number of sources or the ATFs from the sources to the microphones. We discussed that approximating the SCM of the desired mixture by the captured signal's SCM for Mod-PMWF is advantageous for accurate preservation of the speech mixture although it slightly reduces the interference reduction effect. We verified the effectiveness of the Mod-PMWF for mixture enhancement in terms of the interference reduction ratios and the subjective speech quality tests. The Mod-PMWF outperformed several conventional BFs, including GEV-MVDR BF that performs mixture enhancement in a MaxSNR sense, a BSS technique ISCE that utilizes speech sparseness, another BSS technique IMCE with prior knowledge of the number of sources, and an UR-MWF that gives the MMSE estimates of the desired mixtures.

References

[1] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[2] H. L. V. Trees, *Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory*. New York: Wiley-Interscience, 2002.

[3] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel signal ennhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[4] J. Xiao, Z.-Q. Luo, I. Merks, and T. Zhang, "A robust adaptive binaural beamformer for hearing devices," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 2017.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non-stationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[6] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[7] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2007.

[8] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. International conference on machine learning*, 2017, pp. 2632–2641.

[9] M. Er and A. Cantoni, "Derivative constraints for broadband element space antenna array processors," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 31, no. 6, pp. 1378–1393, 1983.

[10] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multispeaker LCMV beamformer and postfilter for source separation and noise reduction," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 5, 2017.

[11] A. Herzog and E. A. P. Habets, "Direction and reverberation preserving noise reduction of ambisonics signals," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 28, 2020.

[12] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE WASPAA*, 2019.

[13] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. Signal Processing*, vol. 69, pp. 3252–3267, 2021.

[14] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online source separation and noise reduction based on joint optimization with dereverberation," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1000–1004. DOI: 10.23919/EUSIPCO54536.2021.9616119.

[15] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 487–503, Jul. 2005.

[16] *ITU-R recommendation BS.1534*.

[17] S. Doclo, A. Spriet, M. Moonen, and J. Wouters, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, pp. 636–656, 2007.

[18] A. Krueger, E. Warsits, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratio estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, 2011.

[19] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," in *Proc. Interspeech 2019*, 2019, pp. 111–115. DOI: 10.21437/Interspeech.2019-1286.

[20] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, 2007.

[21] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformer," in *Proc. IEEE ICASSP*, 2007, pp. 41–44.

[22] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, 2016, pp. 196–200.

[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[24] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, 2005, pp. 181–197.

[25] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[26] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[27] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani, "Noise power spectral density tracking: A maximum likelihood perspective," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 495–498, 2012.

[28] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. International conference on spoken language processing*, 1998, pp. 2819–2822.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Blind Source Separation Using Independent Low-Rank Matrix Analysis with Spectrogram-Consistency Regularization

Sota Misawa*, Norihiro Takamune*, Kohei Yatabe†, Daichi Kitamura‡, and Hiroshi Saruwatari*
* The University of Tokyo, Japan
† Tokyo University of Agriculture and Technology, Japan
‡ National Institute of Technology, Kagawa College, Japan

***Abstract*—Block permutation is a problem that occurs in the separated signals obtained by independent low-rank matrix analysis. To deal with this problem, a new method that induces a low-rank property on the projection of the separated spectrogram to the so-called consistent spectrogram for short-time Fourier transform has recently been proposed. In contrast to conventional methods, which disregard how far the separated spectrogram is from the projected spectrogram, we design a regularization term that induces the separated spectrogram itself to be consistent. For the newly designed optimization problem, we derive the update rule based on the majorization-minimization algorithm and the vectorwise coordinate descent method. We verify the effectiveness of the proposed method experimentally.**

I. Introduction

Blind source separation (BSS) is a technique for estimating original source signals from mixed multichannel signals in situations where the characteristics and spatial arrangements of the sources and microphones are unknown. As a known instance of BSS, frequency-domain independent component analysis (FDICA) has been proposed [1]. FDICA estimates separation filters for each frequency bin under the overdetermined condition (where the number of microphones M is equal to or greater than the number of sources N). In FDICA, there is the well-known problem called the permutation problem: for each frequency bin, the order of the sources in the separated signals is not determined. Many methods to solve this problem have been researched [2]–[7]. This indicates that research on solving the permutation problem is one of the critical themes in BSS.

Among the solutions to the permutation problem, post-processing methods that solve the permutation problem by estimating the source's direction of arrival [2] or the correlation between the frequencies [3] have been proposed. Although these are two-step methods, the methods for estimating the separation filters and dealing with the permutation problem simultaneously by modeling the time-frequency structure of the separated signals have been proposed [4], [5]. Independent vector analysis (IVA) [4] is a method in which a distribution with higher-order correlations between frequency bins is assumed, which results in capturing the synchronicity in the time direction. Independent low-rank matrix analysis (ILRMA) [5] assumes low-rankness in the power spectrogram of the separated signals by introducing a nonnegative matrix factorization (NMF) model [8]. Inducing the time-frequency structure for the source model allows the permutation problem to be solved automatically within the optimization problem. Among these approaches, it has been reported that ILRMA achieves a higher separation performance than do FDICA and IVA [5]. Although ILRMA implicitly solves the permutation problem by capturing the time-frequency structure, the high degree of freedom in the NMF model can cause a block permutation problem, where inconsistencies occur in a neighboring frequency band [9].

In recent years, to deal with permutation and block permutation problems, new approaches based on the spectrogram consistency of complex spectrograms [10]–[13] have been proposed [6], [7]. Here, a spectrogram is said to be *spectrogram-consistent* (or simply *consistent*) if there exists a corresponding time-domain signal through short-time Fourier transformation (STFT) and inverse STFT (ISTFT). Since STFT is a linear transformation that converts time-domain signals to complex spectrograms in the time-frequency domain and the dimension of the spectrogram space is generally larger than that of the time-domain-signal space, consistent spectrograms form a linear subspace in the entire spectrogram space that we call *consistent space*. Some signal-processing methods in the time-frequency domain may cause the consistent spectrogram to deviate from the consistent space. Therefore, when converting such an inconsistent spectrogram to the time domain using ISTFT, unintended changes in the time-frequency structure may occur [10]. As an example, consider performing ISTFT on a separated signal with a block permutation in ILRMA and then applying STFT. In this case, a phenomenon that the frequency band appears to bleed into adjacent bands often occurs near the block-permutated frequency bands in spectrograms before performing ISTFT. This phenomenon causes the collapse of low-rankness and the loss of independence between the sources in the bled-out bands. There is also a method that considers consistency in ILRMA [7]. As described in [7], the spectrogram of the separated time-domain signals, which is projected from possibly inconsistent signals onto consistent space via STFT after ISTFT, is modeled using NMF for low-rankness. By inducing low-rankness in the time-frequency structure of the actual output time-domain signals,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

block permutation is expected to be alleviated. It has been reported that the separation performance of this method is equivalent to or better than that of ILRMA in source separation experiments [7]. Unfortunately, such a projection-based approach is based on heuristics and is not formulated as a strict optimization problem.

The conventional consistency-based approach disregards how far the original separated spectrogram is from the consistent space. In contrast, we are motivated to induce the separated spectrogram itself to be near the consistent space. In this study, we design a regularization term based on the distance between the separated spectrogram and consistent space, and formulate the regularized optimization problem with it. We derive the update rule for this regularized optimization problem on the basis of the majorization-minimization (MM) algorithm [14] and vectorwise coordinate descent (VCD) method [15]. We also confirm the effectiveness of the proposed method through a simulated experiment.

II. Preliminaries

A. MM Algorithm [14]

The MM algorithm is a technique for optimizing iteratively the objective function using an auxiliary function called the upper-bound function. The MM algorithm is often effective when it is difficult to optimize the objective function directly by finding the stationary points and there is an appropriate auxiliary function.

Let $\mathcal{J}(\Theta)$ be the objective function to be minimized with respect to the objective variable Θ. In the MM algorithm, we design an auxiliary function $\mathcal{J}^{+}(\Theta,\Omega)$ that satisfies the following two properties.

$$\mathcal{J}(\Theta) \leq \mathcal{J}^{+}(\Theta,\Omega), \qquad (\forall\Theta,\forall\Omega) \tag{1}$$

$$\mathcal{J}(\Theta) = \min_{\Omega} \mathcal{J}^{+}(\Theta,\Omega), \qquad (\forall\Theta) \tag{2}$$

where Ω is called the auxiliary variable. In the MM algorithm, the optimal solution is obtained by iterating the following steps (i) and (ii).

(i) Update the auxiliary variable from the latest objective variable

$$\tilde{\Omega} \leftarrow \arg\min_{\Omega} \mathcal{J}^{+}(\tilde{\Theta},\Omega). \tag{3}$$

By definition of the auxiliary function, $\tilde{\Omega}$ becomes the value that satisfies the equality in inequality (1).

(ii) Update the objective variable from the new auxiliary variable

$$\Theta \leftarrow \arg\min_{\hat{\Theta}} \mathcal{J}^{+}(\hat{\Theta},\tilde{\Omega}). \tag{4}$$

The update by the MM algorithm guarantees the monotonic nonincreasing property of the objective function.

It is necessary to design an auxiliary function $\mathcal{J}^{+}(\Theta,\Omega)$ for the MM algorithm so that the optimization of $\mathcal{J}^{+}(\Theta,\Omega)$ becomes possible to optimize (4). Furthermore, when the gap between the auxiliary function and the objective function is large, the step size of the update tends to be small, so it is desirable to have an auxiliary function close to the objective function.

B. Representation Matrix of STFT

Consider a real-valued signal of length L, denoted by $\tilde{\boldsymbol{\psi}} \in \mathbb{R}^{L}$. The corresponding vectorized complex-valued spectrogram is represented by $\boldsymbol{\Psi} \in \mathbb{C}^{I_{\mathrm{S}}J}$, where I_{S} and J represent the window length and the number of time frames in STFT, respectively. In this paper, we assume that I_{S} is an even integer for simplicity. Let τ be the shift length of STFT, and $\boldsymbol{h} \in \mathbb{R}^{I_{\mathrm{S}}}$ be the window function. $L^{(\mathrm{pre})}$ and $L^{(\mathrm{post})}$ denote the length of zero-padding at the beginning and end of the time signal, respectively. We denote the length of the signal after zero-padding as $L' = L + L^{(\mathrm{pre})} + L^{(\mathrm{post})}$. Hereafter, we consider the practical case where $I_{\mathrm{S}}J > L$ holds. The $L' \times L$ matrix representing zero-padding is given by $Z = (O_{L\times L^{(\mathrm{pre})}}, E_L, O_{L\times L^{(\mathrm{post})}})^{\mathsf{T}}$, where $O_{q\times q'}$ and E_q denote the $q \times q'$ zero matrix and $q \times q$ identity matrix, respectively. Here, $^{\mathsf{T}}$ represents the transpose operation. Note that $L^{(\mathrm{pre})}$, $L^{(\mathrm{post})}$, and J are defined as $L' = \tau(J-1) + I_{\mathrm{S}}$.

The representation matrix of the normalized discrete Fourier transform (DFT) is denoted by $F \in \mathbb{C}^{I_{\mathrm{S}}\times I_{\mathrm{S}}}$, and the representation matrix of the inverse DFT is given by F^{H}, where $^{\mathsf{H}}$ represents the Hermitian transpose. These matrices are formed by arranging the orthonormal Fourier bases. The $((j-1)I_{\mathrm{S}}+1)$th to jI_{S}th elements of $\boldsymbol{\Psi}$, which correspond to the jth frame of the spectrogram, are obtained by multiplying the $(\tau(j-1)+1)$th to $(\tau(j-1)+I_{\mathrm{S}})$th samples of the zero-padded time signal by the window function and applying DFT. We define $\Upsilon \in \mathbb{C}^{I_{\mathrm{S}}J\times L'}$ as

$$\Upsilon = \begin{pmatrix} F\Delta'_1 \\ \vdots \\ F\Delta'_J \end{pmatrix}, \tag{5}$$

where $\Delta'_j \in \mathbb{R}^{I_{\mathrm{S}}\times L'}$ $(j = 1,\ldots,J)$ is given by

$$\Delta'_j = (O_{I_{\mathrm{S}}\times\tau(j-1)}, \mathrm{diag}(\boldsymbol{h}), O_{I_{\mathrm{S}}\times(L'-\tau(j-1)-I_{\mathrm{S}})}). \tag{6}$$

Here, $\mathrm{diag}(\cdot)$ represents the diagonal matrix whose qth diagonal element is the qth element of the input complex vector. Then, $\Psi = \Upsilon Z\tilde{\boldsymbol{\psi}}$ holds. In the following, the representation matrix of STFT is given by $\mathcal{F} = \Upsilon Z$.

In this paper, we define ISTFT in the least-squares sense, meaning the representation matrix for ISTFT becomes the Moore–Penrose pseudo-inverse matrix corresponding to that for STFT. Therefore, the representation matrix of ISTFT $\mathcal{F}^{\dagger}$ is given by

$$\mathcal{F}^{\dagger} = (\mathcal{F}^{\mathsf{H}}\mathcal{F})^{-1}\mathcal{F}^{\mathsf{H}} = \left(\sum_j \Delta_j^{\mathsf{H}}\Delta_j\right)^{-1}\mathcal{F}^{\mathsf{H}}, \tag{7}$$

where $\Delta_j = \Delta'_j Z$. By definition, $\mathcal{F}^{\dagger}\mathcal{F} = E_L$ holds, and $\mathcal{F}\mathcal{F}^{\dagger}$ becomes an orthogonal projection matrix.

III. Conventional Methods

A. ILRMA [5]

Let $\boldsymbol{s}_{ij} = (s_{ij1}, \ldots, s_{ijN})^{\mathsf{T}}$, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijM})^{\mathsf{T}}$, and $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijN})^{\mathsf{T}}$ represent the spectrograms of the source signals, separated signals, and observed signals, respectively, where $i = 1, \ldots, I_{\mathrm{N}}$; $j = 1, \ldots, J$; $n = 1, \ldots, N$; and $m = 1, \ldots, M$ denote the indices of frequency bins, time frames, sources, and microphones, respectively. I_{N} is the index of the frequency bin corresponding to the Nyquist frequency. Assuming that each source is a point source and the reverberation time is sufficiently shorter than the length of the STFT window, the mixing system can be approximated as instantaneous mixing in the frequency domain: $\boldsymbol{x}_{ij} = A_i \boldsymbol{s}_{ij}$, where $A_i \in \mathbb{C}^{M \times N}$ is called the mixing matrix. In this study, we assume that $N = M$ and A_i is regular. Under these assumptions, the separated signals are obtained by estimating the inverse matrix of A_i, called the separation matrix, $W_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iN})^{\mathsf{H}} \in \mathbb{C}^{N \times M}$, as follows:

$$\boldsymbol{y}_{ij} = W_i \boldsymbol{x}_{ij}. \tag{8}$$

In ILRMA, the separated signals are assumed to follow the following zero-centered univariate complex Gaussian distribution with variance r_{ijn}^2:

$$p(y_{ijn}) = \frac{1}{\pi r_{ijn}^2} \exp\Big(- \frac{|y_{ijn}|^2}{r_{ijn}^2} \Big). \tag{9}$$

Furthermore, to model the low-rank structure of source signals, the $I_{\mathrm{N}} \times J$ matrix whose (i, j)th element is r_{ijn}^2 is modeled by NMF for each n as

$$r_{ijn}^2 = \sum_k t_{ikn} v_{kjn}, \tag{10}$$

where $t_{ikn} \geq 0$ and $v_{kjn} \geq 0$ are elements of the basis matrix and activation matrix of NMF, respectively, and $k = 1, \ldots, K$ denotes the index of the basis. In ILRMA, the following negative log-likelihood is used as the objective function:

$$\begin{aligned}
&\mathcal{J}_{\mathrm{I}}(W_i, \{t_{ikn}\}, \{v_{kjn}\}) \\
&= - \log \prod_{i=1}^{I_{\mathrm{N}}} \prod_{j,n} p(x_{ijn}) && (11) \\
&= - \log \prod_{i=1}^{I_{\mathrm{N}}} \prod_{j,n} p(y_{ijn}) - J \sum_i \log |\det W_i|^2 && (12) \\
&= \sum_{i=1}^{I_{\mathrm{N}}} \sum_{j,n} \left(\frac{|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\sum_k t_{ikn} v_{kjn}} + \log \sum_k t_{ikn} v_{kjn} \right) \\
&\quad - J \sum_{i=1}^{I_{\mathrm{N}}} \log |\det W_i|^2 + \mathrm{const.}, && (13)
\end{aligned}$$

where const. represents the terms independent of W_i, $\{t_{ikn}\}$, and $\{v_{kjn}\}$.

We focus (13) on the separation matrix W_i at first because we update W_i, $\{t_{ikn}\}$, and $\{v_{kjn}\}$ alternately. By rearranging (13) with respect to W_i, we obtain

$$\begin{aligned}
\mathcal{J}_{\mathrm{I}}(\{W_i\}) =& J \sum_{i=1}^{I_{\mathrm{N}}} \left[\sum_n \boldsymbol{w}_{in}^{\mathsf{H}} D_{in} \boldsymbol{w}_{in} - \log |\det W_i|^2 \right] \\
&+ \mathrm{const.}, && (14)
\end{aligned}$$

where $D_{in} = \sum_j \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}} / (J \sum_k t_{ikn} v_{kjn})$ and const. denotes terms independent of $\{W_i\}$. It is possible to minimize (14) by iterative projection (IP) [16], as described in [5]. The update rule for the separation filter $\boldsymbol{w}_{in}$ using IP are

$$\boldsymbol{w}_{in} \leftarrow (W_i D_{in})^{-1} \boldsymbol{e}_n, \tag{15}$$

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathsf{H}} D_{in} \boldsymbol{w}_{in}}}, \tag{16}$$

where $\boldsymbol{e}_n \in \mathbb{C}^N$ represents a vector whose nth element is one and others are zero. For the NMF variables $\{t_{ikn}\}$ and $\{v_{kjn}\}$, the following update rules have been derived on the basis of the MM algorithm (see details in [5]):

$$t_{ikn} \leftarrow t_{ikn} \sqrt{\frac{\sum_j |y_{ijn}|^2 v_{kjn} (\sum_{k'} t_{ik'n} v_{k'jn})^{-2}}{\sum_j v_{kjn} (\sum_{k'} t_{ik'n} v_{k'jn})^{-1}}}, \tag{17}$$

$$v_{kjn} \leftarrow v_{kjn} \sqrt{\frac{\sum_i |y_{ijn}|^2 t_{ikn} (\sum_{k'} t_{ik'n} v_{k'jn})^{-2}}{\sum_i t_{ikn} (\sum_{k'} t_{ik'n} v_{k'jn})^{-1}}}. \tag{18}$$

The output of ILRMA involves uncertainty at the scale of the separated signals. Therefore, the separated signals are converted into signals at the microphone positions by back-projection [17] as follows:

$$y_{ijn} \leftarrow \boldsymbol{e}_{m_{\mathrm{ref}}}^{\mathsf{H}} W_i^{-1} \boldsymbol{e}_n y_{ijn}, \tag{19}$$

where m_{ref} represents the index of the reference channel. To modify the variables in ILRMA while ensuring that the separated signal becomes (19) and the objective function (13) remains unchanged, we update $\boldsymbol{w}_{in}$ and t_{ikn} as follows:

$$\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in} \lambda_{inm_{\mathrm{ref}}}, \tag{20}$$

$$t_{ikn} \leftarrow t_{ikn} |\lambda_{inm_{\mathrm{ref}}}|^2, \tag{21}$$

where $\lambda_{inm_{\mathrm{ref}}} = \boldsymbol{e}_{m_{\mathrm{ref}}}^{\mathsf{H}} W_i^{-1} \boldsymbol{e}_n$.

B. ILRMA Based on Projection to Consistent Space [7]

Recently, methods that address spectrogram consistency have been proposed [6], [7]. The spectrogram is said to be consistent if and only if the spectrogram is in the image of STFT. Thus, the consistent space, which is spanned by consistent spectrograms, is denoted as $\{\mathcal{F}\tilde{\psi} | \tilde{\psi} \in \mathbb{R}^L\}$. Since the consistent space is the subspace of the entire spectrogram space under $I_{\mathrm{S}} J > L$, applying a naive signal processing in the time-frequency domain to the consistent spectrogram may result in deviation from the consistent space. To measure this deviation, we define the measure

$$\mathcal{E}(\boldsymbol{\Psi}) = \left\| \mathcal{F}\mathcal{F}^{\dagger} \boldsymbol{\Psi} - \boldsymbol{\Psi} \right\|_2^2 \tag{22}$$

for given $\boldsymbol{\Psi} \in \mathbb{C}^{I_S J}$ because $\mathcal{F}\mathcal{F}^{\dagger}$ is the orthogonal projection onto the consistent space. Here, the spectrogram $\boldsymbol{\Psi}$ is consistent if and only if $\mathcal{E}(\boldsymbol{\Psi}) = 0$ holds.

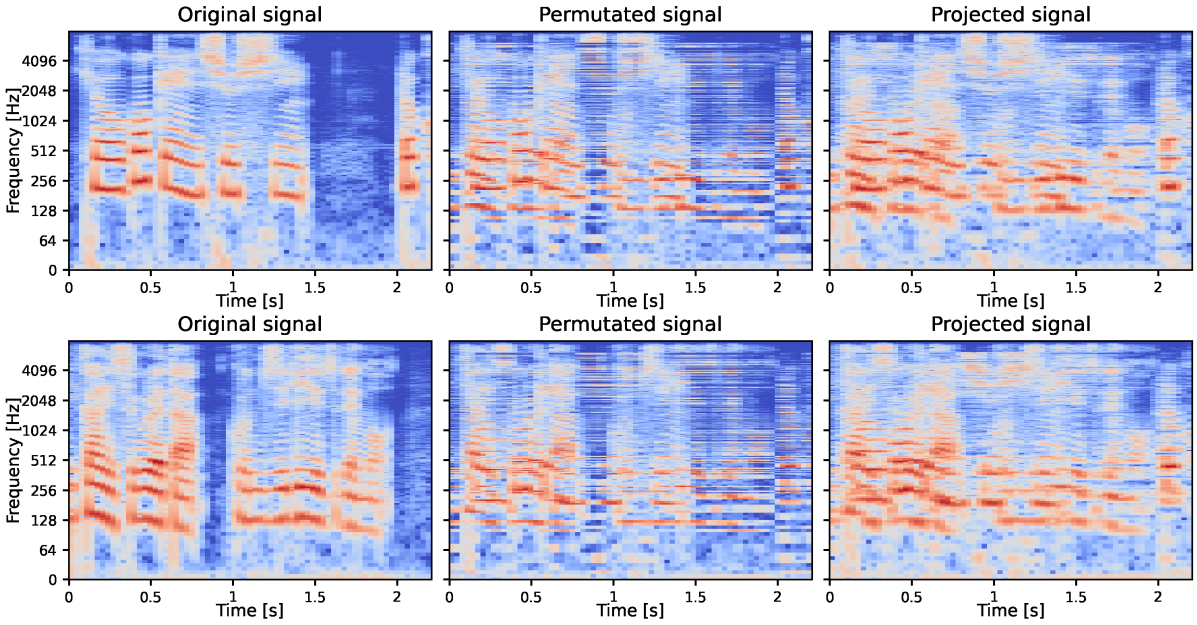


Fig. 1. Spectrograms whose frequency bins are intentionally swapped (middle) from two original spectrograms of speech (left) and projected spectrograms onto spectrogram-consistent space (right).

In ILRMA, it is found that the block permutation may cause a large deviation. When a block permutation occurs in the separated signals obtained by ILRMA, the separated signals projected onto the consistent space may exhibit the phenomenon of bleed-out into adjacent frequency bins in the block-permuted band. As an example, Fig. 1 represents the spectrograms projected onto the consistent space from inconsistent spectrograms obtained by artificially swapping frequency bins for two consistent spectrograms. It is easily seen that the time-frequency structure changes as a result of projection; mutual diffusion occurs in the discontinuous frequency bands. Therefore, independence between sources or low-rank properties of the power spectrograms may be lost upon projecting the block-permuted spectrogram onto the consistent space. To focus on this, in [7], it was hypothesized that block permutation is alleviated by updating parameters to induce low-rankness in the projected separated spectrograms. In detail, this method updates the power spectrogram modeled by NMF so that it becomes close not to Y_n but to $\mathcal{F}\mathcal{F}^{\dagger}Y_n$, where Y_n is the vectorized spectrogram whose $((j-1)I_{\mathrm{S}}+i)$th element is y_{ijn} for each n. The algorithm of ILRMA based on projection to consistent space is outlined in Algorithm 1.

Algorithm 1 Projection-based ILRMA

Initialize $\{W_i\}$, $\{t_{ikn}\}$, and $\{v_{kjn}\}$.
for iter $= 1, \ldots,$ maxIter **do**
  $y_{ij} \leftarrow \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}$
  $Y_n \leftarrow \mathcal{F}\mathcal{F}^{\dagger}Y_n$
  Update $\{t_{ikn}\}$ and $\{v_{kjn}\}$ using (17) and (18).
  Update $\{W_i\}$ using (15) and (16).
  Execute back-projection (19).
  Update variables using (20) and (21).
end for
return $\{y_{ijn}\}$

IV. Proposed Method

A. Motivation and Approach

The conventional method described in Section III-B is unfortunately a heuristic method that has not been rigorously formulated as an optimization problem. Moreover, since it only discusses the projected separated signals, it does not address how far the original separated signals are from the consistent space.

In this work, we aim to construct an optimization problem that induces the separated signals to be consistent by designing a regularization term based on the deviation from the consistent space. First, we design a regularization term related to the separation filters W_i from the deviation measure $\mathcal{E}(Y_n)$ and take it into account in the optimization problem. We find out that this new optimization problem is difficult to solve using conventional optimization methods. Therefore, on the basis of the MM algorithm, we design an auxiliary function to which VCD [15] can be applied, and derive a new update rule.

B. Designing the Regularization Term for Consistency

In this section, we define a regularization term inducing consistency for W_i, which consists of the separated spectrogram $\boldsymbol{y}_{ij}$ shown by (8).

For the separation filter W_i, we denote $\vec{\boldsymbol{w}}_{nm} = (w_{1nm}^*, \ldots, w_{I_{\mathrm{S}}nm}^*)^{\mathsf{T}}$, $\mathcal{W}_{nm} = \mathrm{diag}(\vec{\boldsymbol{w}}_{nm})$, and let $\tilde{x}_m$ be the time-domain signal of the mth observed signal, where * denotes the complex conjugate. The jth frame of the

separated signal corresponding to the nth source is given by $(y_{1jn},\dots,y_{I_{\mathrm{S}}jn})^{\mathsf{T}} = \sum_m \mathcal{W}_{nm}F\Delta_j\tilde{\boldsymbol{x}}_m$. From (5) and (7), we deform the deviation of the separated spectrogram from the consistent space as

$$\sum_n \mathcal{E}(Y_n) = \sum_{n,j}\Big\| \sum_m (F\Delta_j\Phi_{nm} - \mathcal{W}_{nm}F\Delta_j)\tilde{\boldsymbol{x}}_m \Big\|_2^2, \quad (23)$$

where $\Phi_{nm} = \big(\mathcal{F}^{\mathsf{H}}\mathcal{F}\big)^{-1}\Big(\sum_{j'}\Delta_{j'}^{\mathsf{H}}F^{\mathsf{H}}\mathcal{W}_{nm}F\Delta_{j'}\Big)$.

To induce consistency for arbitrary observed signals, we consider the case where $\tilde{\boldsymbol{x}}_m$ is a white Gaussian signal and the expectation of (23) is calculated as

$$\mathbb{E}\Big[\sum_n \mathcal{E}(Y_n)\Big] = \sum_{n,m,j}\|F\Delta_j\Phi_{nm} - \mathcal{W}_{nm}F\Delta_j\|_{\mathrm{F}}^2. \quad (24)$$

We use this expression as the regularization term $\mathcal{C}(W_i)$ and deform it as follows:

$$\begin{aligned}\mathcal{C}(W_i) =& \sum_{n,m}\Big\|\Big(\sum_j \Delta_j\Delta_j^{\mathsf{H}}\Big)^{\frac{1}{2}}U_{nm}\Big\|_{\mathrm{F}}^2 \\ &- \sum_{n,m}\Big\|\Big(\sum_j \Delta_j^{\mathsf{H}}\Delta_j\Big)^{-\frac{1}{2}}\Big(\sum_j \Delta_j^{\mathsf{H}}U_{nm}\Delta_j\Big)\Big\|_{\mathrm{F}}^2, \quad (25)\end{aligned}$$

where we define $U_{nm} = F^{\mathsf{H}}\mathcal{W}_{nm}F$. Because of the Fourier transform on a diagonal matrix, U_{nm} becomes a circulant Toeplitz matrix, which is formed by cyclically shifting its first column vector $\boldsymbol{u}_{nm}$. Note that there exists the following relationship between $\boldsymbol{u}_{nm}$ and $\vec{\boldsymbol{w}}_{nm}$:

$$\boldsymbol{u}_{nm} = \frac{1}{\sqrt{I_{\mathrm{S}}}}F^{\mathsf{H}}\vec{\boldsymbol{w}}_{nm}. \quad (26)$$

Now, let us define $\breve{U}_{nm} = \mathrm{diag}(\boldsymbol{h})^{\mathsf{H}}U_{nm}\mathrm{diag}(\boldsymbol{h})$, $\mathcal{I}(l) = \big\{j\big|-L^{(\mathrm{pre})} < l-\tau(j-1) \le I_{\mathrm{S}} - L^{(\mathrm{pre})}\big\}$, and $\sigma(l,j) = l - \tau(j-1) + L^{(\mathrm{pre})}$. Then, we have

$$\Big[\sum_j \Delta_j^{\mathsf{H}}\Delta_j\Big]_{ll'} = \begin{cases}\sum_{j\in\mathcal{I}(l)}[\boldsymbol{h}]_{\sigma(l,j)}^2 & \text{if } l=l' \\ 0 & \text{otherwise}\end{cases}, \quad (27)$$

$$\Big[\sum_j \Delta_j^{\mathsf{H}}U_{nm}\Delta_j\Big]_{ll'} = \sum_{j\in\mathcal{I}(l)\cap\mathcal{I}(l')}\Big[\breve{U}_{nm}\Big]_{\sigma(l,j),\sigma(l',j)}, \quad (28)$$

where $[\cdot]_{qq'}$ denotes the (q,q')th component of the matrix. According to (27) and (28), the second term of (25) becomes

$$\begin{aligned}&\sum_{n,m}\Big\|\Big(\sum_j \Delta_j^{\mathsf{H}}\Delta_j\Big)^{-\frac{1}{2}}\Big(\sum_j \Delta_j^{\mathsf{H}}U_{nm}\Delta_j\Big)\Big\|_{\mathrm{F}}^2 \\ &= \sum_{n,m,l,l'}\frac{\Big(\sum_{j\in\mathcal{I}(l)\cap\mathcal{I}(l')}\Big[\breve{U}_{nm}\Big]_{\sigma(l,j),\sigma(l',j)}\Big)^2}{\sum_{j\in\mathcal{I}(l)}[\boldsymbol{h}]_{\sigma(l,j)}^2}. \quad (29)\end{aligned}$$

In accordance with the definition of $\breve{U}_{nm}$, (29) becomes a quadratic form with respect to $\boldsymbol{u}_{nm}$, expressed as $\sum_{n,m}\boldsymbol{u}_{nm}^{\mathsf{H}}\breve{G}\boldsymbol{u}_{nm}$. In this case, the weight matrix $\breve{G}$ is diagonal and independent of n and m. The ith diagonal element of $\breve{G}$ is given by

$$\sum_l \sum_{l'\in\{l-i+1,l+I_{\mathrm{S}}-i+1\}}\frac{\Big(\sum_{j\in\mathcal{I}(l)\cap\mathcal{I}(l')}[\boldsymbol{h}]_{\sigma(l,j)}[\boldsymbol{h}]_{\sigma(l',j)}\Big)^2}{\sum_{j\in\mathcal{I}(l)}[\boldsymbol{h}]_{\sigma(l,j)}^2}. \quad (30)$$

Additionally, in the first term of (25), since $\sum_j \Delta_j\Delta_j^{\mathsf{H}}$ is a diagonal matrix and U_{nm} is the circulant Toeplitz, the following equation holds:

$$\sum_{n,m}\Big\|\Big(\sum_j \Delta_j\Delta_j^{\mathsf{H}}\Big)^{\frac{1}{2}}U_{nm}\Big\|_{\mathrm{F}}^2 = \sum_{n,m,l}\sum_{j\in\mathcal{I}(l)}[\boldsymbol{h}]_{\sigma(l,j)}^2\|\boldsymbol{u}_{nm}\|_2^2. \quad (31)$$

Thus, we obtain

$$\mathcal{C}(W_i) = \sum_{n,m}\boldsymbol{u}_{nm}^{\mathsf{H}}\bar{G}\boldsymbol{u}_{nm}, \quad (32)$$

where $\bar{G} = \Big(\sum_l\sum_{j\in\mathcal{I}(l)}[\boldsymbol{h}]_{\sigma(l,j)}^2\Big)E_{I_{\mathrm{S}}} - \breve{G}$. Consequently, from (26), the consistency regularization term is deformed to be a quadratic form of $\vec{\boldsymbol{w}}_{nm}$:

$$\mathcal{C}(W_i) = \frac{1}{I_{\mathrm{S}}}\sum_{n,m}\vec{\boldsymbol{w}}_{nm}^{\mathsf{H}}\big(F\bar{G}F^{\mathsf{H}}\big)\vec{\boldsymbol{w}}_{nm}. \quad (33)$$

In the following, we use $G = F\bar{G}F^{\mathsf{H}}/I_{\mathrm{S}}$.

C. Formulation of the Optimization Problem and Derivation of the Update Rule

For updating the NMF variables $\{t_{ikn}\},\{v_{kjn}\}$, since the regularization term is independent of the NMF variables, the updates are performed using (17) and (18) in the same manner as in ILRMA. Therefore, we hereafter discuss the optimization problem with respect to W_i. We formulate the objective function in ILRMA with the consistency regularization term (33) as follows:

$$\begin{aligned}\hat{\mathcal{J}}(W_i) =& J\sum_{i=1}^{I_{\mathrm{N}}}\Big[\sum_n \boldsymbol{w}_{in}^{\mathsf{H}}D_{in}\boldsymbol{w}_{in} - \log|\det W_i|^2\Big] \\ &+ \gamma\sum_{n,m}\vec{\boldsymbol{w}}_{nm}^{\mathsf{H}}G\vec{\boldsymbol{w}}_{nm} + \mathrm{const.}, \quad (34)\end{aligned}$$

where $\gamma > 0$ represents the weight of the regularization term and const. is the term independent of $\{W_i\}$. Here, we define $\tilde{i}$ by $I_{\mathrm{S}}+2-i$ for $2 < i < I_{\mathrm{N}}$, which is frequency bin folding by the Nyquist frequency. It is important that w_{inm} and $w_{\tilde{i}nm}$ have a complex conjugate relationship. With this in mind, (34) can be rewritten for $\boldsymbol{w}_{in}$ at a certain frequency i as follows:

$$\begin{aligned}\hat{\mathcal{J}}(W_i) =& J\Big[\sum_n \boldsymbol{w}_{in}^{\mathsf{H}}D_{in}\boldsymbol{w}_{in} - \log|\det W_i|^2\Big] \\ &+ \gamma\sum_{n,m}\Bigg(\begin{pmatrix}w_{inm}\\ w_{inm}^*\end{pmatrix}^{\mathsf{H}}\hat{G}_i\begin{pmatrix}w_{inm}\\ w_{inm}^*\end{pmatrix} \\ &\quad + \begin{pmatrix}w_{inm}\\ w_{inm}^*\end{pmatrix}^{\mathsf{H}}\boldsymbol{b}_{inm} + \boldsymbol{b}_{inm}^{\mathsf{H}}\begin{pmatrix}w_{inm}\\ w_{inm}^*\end{pmatrix}\Bigg) \\ &+ \mathrm{const.}, \quad (35)\end{aligned}$$

where const. is a term that is independent of $\boldsymbol{w}_{in}$ and $\boldsymbol{w}_{\tilde{i}n} = \boldsymbol{w}_{in}^*$, and we define

$$\hat{G}_i = \begin{pmatrix} [G]_{\tilde{i}\tilde{i}} & [G]_{\tilde{i}i} \\ [G]_{i\tilde{i}} & [G]_{ii} \end{pmatrix}, \tag{36}$$

$$\boldsymbol{b}_{inm} = \begin{pmatrix} \sum_{i' \neq i,\tilde{i}} [G]_{i'i} w_{i'nm} \\ \sum_{i' \neq i,\tilde{i}} [G]_{ii'} w_{i'nm}^* \end{pmatrix}. \tag{37}$$

Regarding the quadratic form of the second term in (35), the off-diagonal elements of the coefficient matrix cause w_{inm}^2 and $(w_{inm}^*)^2$ terms to appear, making it difficult to directly apply conventional IP [16] or VCD [15]. If we can bound the second term of (35) by a quadratic function with a diagonal coefficient matrix for the MM algorithm, it becomes possible to apply VCD. To achieve this, we use the following Claim 1 for designing an auxiliary function.

Claim 1: For any $\boldsymbol{w} \in \mathbb{C}^q$ and $\boldsymbol{\alpha} \in \mathbb{C}^q$ and any positive semidefinite Hermitian matrix $\hat{G} \in \mathbb{C}^{q \times q}$, we have

$$\begin{aligned} \boldsymbol{w}^{\mathsf{H}} \hat{G} \boldsymbol{w} \leq & \boldsymbol{w}^{\mathsf{H}} \hat{G}^{+} \boldsymbol{w} - \boldsymbol{\alpha}^{\mathsf{H}} \Big(\hat{G}^{+} - \hat{G} \Big) \boldsymbol{w} - \boldsymbol{w}^{\mathsf{H}} \Big(\hat{G}^{+} - \hat{G} \Big) \boldsymbol{\alpha} \\ & + \boldsymbol{\alpha}^{\mathsf{H}} \Big(\hat{G}^{+} - \hat{G} \Big) \boldsymbol{\alpha}, \end{aligned} \tag{38}$$

where $\hat{G}^{+}$ is a positive semidefinite Hermitian matrix satisfying $\hat{G}^{+} - \hat{G} \succeq O$. Equality holds when $\boldsymbol{\alpha} = \boldsymbol{w}$.

Proof: Since $\hat{G}^{+} - \hat{G} \succeq O$, for any $\boldsymbol{w} \in \mathbb{C}^q$ and $\boldsymbol{\alpha} \in \mathbb{C}^q$,

$$(\boldsymbol{w} - \boldsymbol{\alpha})^{\mathsf{H}} (\hat{G}^{+} - \hat{G})(\boldsymbol{w} - \boldsymbol{\alpha}) \geq 0 \tag{39}$$

always holds. Therefore, by expanding and rearranging (39), we obtain the desired inequality (38). Equality holds in (38) when equality holds in (39), i.e., when $\boldsymbol{\alpha} = \boldsymbol{w}$. ■

In Claim 1, we set $\boldsymbol{w} = (w_{inm}, w_{inm}^*)^{\mathsf{T}}$, $\boldsymbol{\alpha} = (\alpha_{inm}, \alpha_{inm}^*)^{\mathsf{T}}$, $\hat{G} = \hat{G}_i$, and $\hat{G}^{+} = \hat{G}_i^{+}$. To find a weight matrix $\hat{G}_i^{+}$ such that the auxiliary function becomes close to the original function and has a diagonal structure, we consider the following optimization problem for $\hat{G}_i^{+}$:

$$\begin{aligned} \min_{\hat{G}_i^{+}} \quad & \mathrm{tr}\Big(\hat{G}_i^{+} - \hat{G}_i \Big) \\ \text{s.t.} \quad & \hat{G}_i^{+} - \hat{G}_i \succeq O \\ & \hat{G}_i^{+} \text{ is diagonal.} \end{aligned} \tag{40}$$

Solving (40) yields $\hat{G}_i^{+} = \mathrm{diag}((|[G]_{\tilde{i}\tilde{i}}| + |[G]_{\tilde{i}i}|, |[G]_{ii}| + |[G]_{i\tilde{i}}|)^{\mathsf{T}})$.

By applying Claim 1 to (35) using $\hat{G}_i^{+}$, we obtain the auxiliary function of the objective function as

$$\begin{aligned} \hat{\mathcal{J}}^{+}(W_i) = & J \Big[\boldsymbol{w}_{in}^{\mathsf{H}} \mathcal{D}_{in} \boldsymbol{w}_{in} + \boldsymbol{w}_{in}^{\mathsf{H}} \tilde{\boldsymbol{b}}_{in} + \tilde{\boldsymbol{b}}_{in}^{\mathsf{H}} \boldsymbol{w}_{in} - \log |\det W_i|^2 \Big] \\ & + \text{const.}, \end{aligned} \tag{41}$$

where const. is a term independent of $\boldsymbol{w}_{in}$ and

$$\mathcal{D}_{in} = D_{in} + 2\frac{\gamma}{J}(|[G]_{\tilde{i}\tilde{i}}| + |[G]_{\tilde{i}i}|) E_M, \tag{42}$$

$$\Big[\tilde{\boldsymbol{b}}_{in} \Big]_m = 2\frac{\gamma}{J} \Big([\boldsymbol{b}_{inm}]_1 - |[G]_{i\tilde{i}}| \alpha_{inm} + [G]_{\tilde{i}i} \alpha_{inm}^* \Big). \tag{43}$$

Equality holds when $\alpha_{inm} = w_{inm}$.

Algorithm 2 Proposed method

1: Initialize $\{W_i\}$, $\{t_{ikn}\}$, and $\{v_{kjn}\}$
2: **for** iter $= 1, \ldots,$ maxIter **do**
3: Update $\{\sigma_{izn}\}$ and $\{\nu_{zjn}\}$ using (17) and (18)
4: Update the auxiliary variables by $\alpha_{inm} \leftarrow w_{inm}$
5: Update $\{W_i\}$ using (44)-(48)
6: **end for**
7: Perform back-projection using (19)
8: **return** $\{y_{ijn}\}$

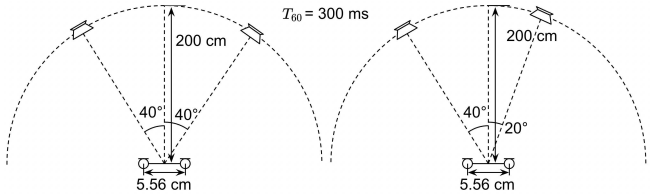


Fig. 2. Recording conditions of impulse response.

In (41), since there is no term related to w_{inm}^2 or $(w_{inm}^*)^2$, VCD can be used to obtain the following update rule:

$$\boldsymbol{\zeta}_{in} \leftarrow (W_i \mathcal{D}_{in})^{-1} \boldsymbol{e}_n, \tag{44}$$

$$\hat{\boldsymbol{\zeta}}_{in} \leftarrow \mathcal{D}_{in}^{-1} \tilde{\boldsymbol{b}}_{in}, \tag{45}$$

$$\chi_{in} \leftarrow \boldsymbol{\zeta}_{in}^{\mathsf{H}} \mathcal{D}_{in} \boldsymbol{\zeta}_{in}, \tag{46}$$

$$\hat{\chi}_{in} \leftarrow \boldsymbol{\zeta}_{in}^{\mathsf{H}} \mathcal{D}_{in} \hat{\boldsymbol{\zeta}}_{in}, \tag{47}$$

$$\boldsymbol{w}_{in} \leftarrow \begin{cases} \frac{\boldsymbol{\zeta}_{in}}{\sqrt{\chi_{in}}} - \hat{\boldsymbol{\zeta}}_{in}, & \text{if } \hat{\chi}_{in} = 0 \\ \frac{\hat{\chi}_{in}}{2\chi_{in}} \Big(1 - \sqrt{1 + \frac{4\chi_{in}}{|\hat{\chi}_{in}|^2}} \Big) \boldsymbol{\zeta}_{in} - \hat{\boldsymbol{\zeta}}_{in}. & \text{otherwise} \end{cases} \tag{48}$$

This update rule guarantees the monotonic nonincrease of the objective function.

The algorithm of the proposed method is outlined in Algorithm 2.

V. Experiment

A. Experimental Conditions

To verify the performance of the proposed method, we conducted a simulated experiment using stereo signals. The dry sources were taken from the JNAS speech corpus [18] and convolved by impulse responses under two recording conditions, as shown in Fig. 2. We used six combinations of mixture: Two mixtures are created for each combination of (male, female), (male, male), and (female, female). The sampling frequency was set to 16 kHz, and we used a 128 ms Hamming window with a 64 ms shift in STFT.

For comparison, we used ILRMA (naive ILRMA) [5], ILRMA based on projection to consistent space (projection-based ILRMA) [7], and the proposed method. In the proposed method, we used $\gamma = 0.01$, 0.1, 1, 10, and 100 as the regularization weights. We set the initial value of the separation matrix W_i in naive ILRMA as Ξ_i^{H}, where we denote $(\rho_{i1}, \ldots, \rho_{iM}) \in \mathbb{R}^M$ and $\Xi_i \in \mathbb{C}^{M \times M}$ as the eigenvalues

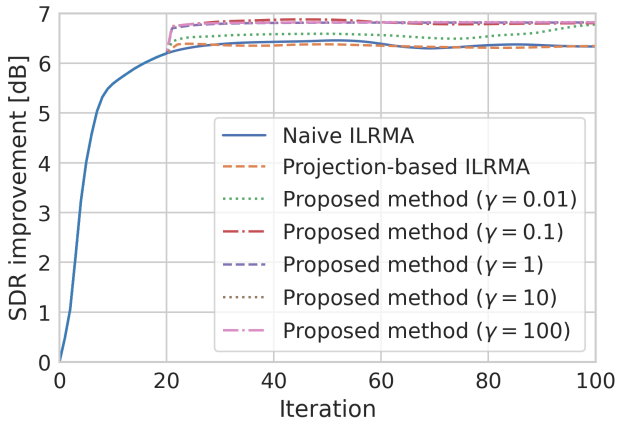


Fig. 3. Behavior of SDR improvement [dB] for each iteration.

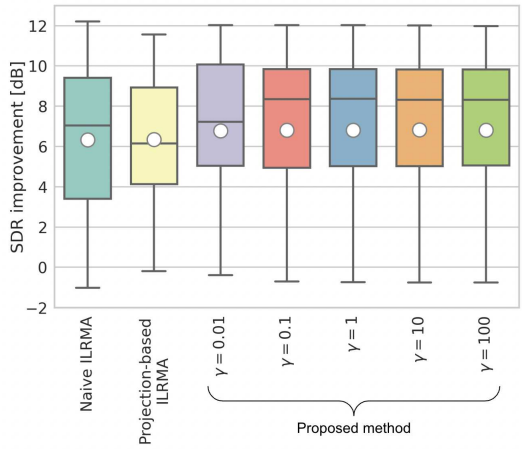


Fig. 4. Box plot of SDR improvement [dB] after 100 iterations for each method. Boxes represent quartiles of SDR improvement's distributions and circles denote averages of SDR improvements.

and the orthonormal matrix composed of the corresponding eigenvectors of the empirical spatial covariance matrix of the observed signals $\sum_j \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}/J$, respectively. The initial values of the basis t_{ikn} and the activation v_{kjn} in NMF were uniformly distributed random numbers on the intervals $[0.1 * \rho_{in}, \rho_{in}]$ and $[0.1, 1.0]$, respectively. For ILRMA, ten different random seeds were prepared for each condition. We conducted experiments under $6 \times 2 \times 10 = 120$ conditions. The number of iterations in naive ILRMA was 100. In projection-based ILRMA and the proposed method, the separation matrix and NMF variables were each taken from the 20-iteration point of ILRMA, and an additional 80 iterations were performed.

Throughout this experiment, we used the source-to-distortion ratio (SDR) improvement [19] as the performance metric for source separation.

B. Result

Fig. 3 shows the average SDR improvement for each iteration. In Fig. 3, naive ILRMA and projection-based ILRMA showed similar performances. The proposed method exhibited an improved performance from that of the ILRMA-initialized value with $\gamma = 0.01$ and achieved a 0.5 dB improvement in one iteration when γ was 0.1 or higher. The box plot of the SDR improvement after 100 iterations is shown in Fig. 4. Fig. 4 shows that the proposed method exhibited improved performances in the first and third quartiles compared with naive ILRMA. The median performance of the proposed method when γ was 0.1 or higher remained almost unchanged and was higher than those of naive ILRMA and projection-based ILRMA. Fig. 5 shows the SDR improvements of projection-based ILRMA and the proposed method with $\gamma = 100$ against that of naive ILRMA. In Fig. 5, projection-based ILRMA and the proposed method showed higher performance under most cases in the low-performance region of naive ILRMA, between 0 dB and 1 dB. On the other hand, in regions where the performance of naive ILRMA was 5 dB or higher, there were instances where the performance of projection-based ILRMA was both better and worse than that of naive ILRMA. However, the proposed method stably showed higher performance than that of naive ILRMA. In particular, when the performance of naive ILRMA was 7 dB or higher, the proposed method often achieved performances above 11 dB.

VI. Conclusion

The conventional method has been proposed to focus on spectrogram consistency in the separated spectrogram in order to address the block permutation problem in ILRMA. This method updates its variables so that the power spectrogram modeled by NMF becomes close to the projection of the separated spectrogram onto the consistent space instead of the original separated spectrogram. Unfortunately, the conventional method is not rigorously formulated as an optimization problem. In this study, we aimed to formulate an optimization problem by designing a regularization term that induces consistency in the separated spectrogram itself. On the basis of the MM algorithm, we derived the monotonic nonincreasing update rule of W_i by designing an auxiliary function to which VCD can be applied, because it is difficult to directly optimize the new objective function added to the regularization term by traditional IP or VCD. The effectiveness of the proposed method was confirmed through the simulated experiment.

Acknowledgment

This work was supported by JST, Moonshot R&D Grant Number JPMJMS2011 (for algorithm development), and Tateisi Science and Technology Foundation (for practical experiment).

References

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

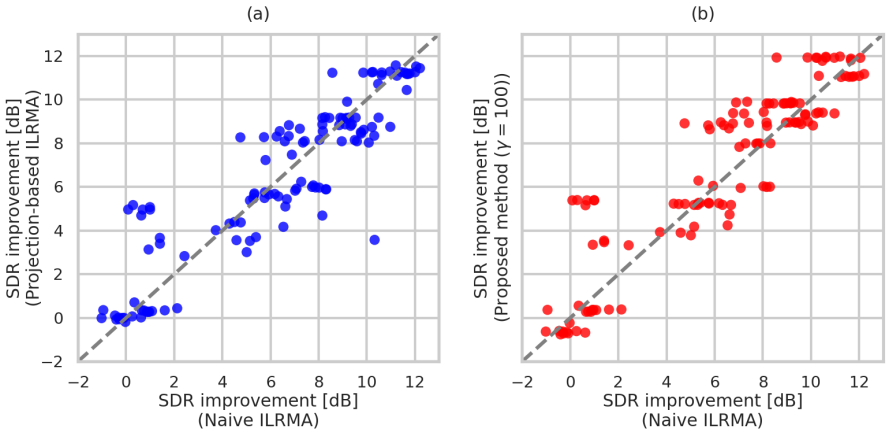


Fig. 5. Pairplot for SDR improvements of naive ILRMA with (a) projection-based ILRMA and (b) proposed method ($\gamma = 100$).

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.

[3] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[4] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2006.

[5] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[6] K. Yatabe, "Consistent ICA: Determined BSS meets spectrogram consistency," *IEEE SP Lett.*, vol. 27, pp. 870–874, 2020.

[7] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP JASP*, vol. 2020, no. 1, pp. 1–35, 2020.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[9] Y. Liang, S. M. Naqvi, and J. Chambers, "Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm," *Electron. Lett.*, vol. 48, no. 8, pp. 460–462, 2012.

[10] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. DAFx*, vol. 10, 2010, pp. 397–403.

[11] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE SP Lett.*, vol. 20, no. 3, pp. 217–220, 2012.

[12] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin–Lim algorithm," in *Proc. WASPAA*, 2013, pp. 1–4.

[13] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, "Representation of complex spectrogram via phase conversion," *AST*, vol. 40, no. 3, pp. 170–177, 2019.

[14] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Stat.*, vol. 9, no. 1, pp. 60–77, 2000.

[15] N. Makishima, Y. Mitsui, N. Takamune, *et al.*, "Independent deeply learned matrix analysis with automatic selection of stable microphone-wise update and fast sourcewise update of demixing matrix," *Signal Process.*, vol. 178, no. 107753, 2021.

[16] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[17] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. SICE*, vol. 4, 2002, pp. 2138–2143.

[18] K. Itou, M. Yamamoto, K. Takeda, *et al.*, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. ASJ (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Moving Interference Speaker removal using Geometrically Constrained Independent Vector Analysis

Shinya Furunaga*, Tetsuya Ueda*, and Shoji Makino*
* Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan

***Abstract*—In this paper, we propose Multiple Geometrically Constrained Auxiliary function-based Independent Vector Analysis with Iterative Source Steering (MGC-AuxIVA-ISS), an offline geometrically constrained source separation method that removes the moving interference signal. GC-AuxIVA is a method that combines AuxIVA, a Blind Source Separation method, with geometrical constraints so that it achieves high separation performance while obtaining the target speech at the desired output channel. Online processing for GC-AuxIVA has also been proposed and can handle the movement of sound sources. However, it requires the exact positions of the target and interference sound sources for each time frame to make an accurate geometrical constraint. On the other hand, it is relatively easier to obtain a range of the moving speaker. Thus, we propose overdetermined offline-GC-AuxIVA to remove the range of the moving speaker by providing multiple geometrical constraints. We conducted simulation experiments to confirm that the proposed method can remove the interference signal by giving geometrical constraints at equal intervals in the range of the movement.**

I. Introduction

In recent years, speech processing applications have become widespread. However, there is a problem in that the speech quality is decreased by diffuse noises and directional interferences. Therefore, there is a need for techniques that can extract the target speech from the recorded speech mixture. One of the useful techniques to enhance the target speech is Blind Source Separation (BSS), which estimates individual source signals from microphone-observed signals without such prior information as training data [1]. BSS methods for the determined condition (where the number of sound sources is equal to the number of microphones) include Independent Component Analysis (ICA) [2], which estimates a demixing matrix that maximizes the independence between source signals, Independent Vector Analysis (IVA) [3], [4], which solves both the source separation problem and the frequency permutation problem by modeling the entire frequency components as multivariate variables following spherical multivariate distribution, and Auxiliary function-based IVA (AuxIVA) [5], which is more stable and fast update rule using the auxiliary function method. However, they do not uniquely determine the output order of the separated signals, which we call global permutation problem.

Geometrically Constrained ICA (GC-ICA) and its extensions have been proposed to solve the global permutation problem simultaneously and the source separation problem [6], [7], [8], [9], [10], [11]. They can guide the demixing matrix to obtain a signal from a desired direction by exploiting spatial information such as the Direction of Arrival (DOA) of the signals and the microphone positions. Among them, Geometrically Constrained Auxiliary function-based IVA with Iterative Source Steering (GC-AuxIVA-ISS) [11] solves the global permutation problem and achieves better stability and faster update through an update algorithm using Iterative Source Steering (ISS) [12] without requiring inverse matrix operations. Its extension to online processing has also been developed to separate signals in real-time (which we call online-GC-AuxIVA-ISS [13]).

This paper focuses on the problem when applying GC-AuxIVA-ISS to moving source separation. One way to remove the moving interference signal is to estimate a time-variant demixing matrix that removes the interference signals according to the interference sound sources' position. There are several online source separation methods [8], [14], [15], and online-GC-AuxIVA-ISS [13] can estimate time-variant demixing matrix recursively. However, when the interference source moves quickly, it is difficult to remove the interference signal in real time because we obtain less spatial information about the interference signal than when the interference is fixed. Also, in online-GC-AuxIVA-ISS, we require time-variant spatial information for geometric constraints because the position of the interference source is changing. However, it is difficult to know the exact position of the interference source in a real environment. Although previous research has proposed a method that estimates spatial information from the current demixing matrix, preliminary experiments have confirmed that its source separation performance is limited when the sound source moves quickly.

Although it is difficult to obtain the positional information of the moving sound source at each time, it is relatively easy to obtain information on its range. Therefore, we propose a source separation method that removes the signal in the range of the interference source's movement. In concrete, we assume offline processing and an overdetermined condition where the number of microphones is more than the number of sources. Then we add several geometric constraints to guide the demixing matrix to remove the range of the interference source movement. We call this method Multiple GCAuxIVA-ISS (MGC-AuxIVA-ISS) and evaluate its effectiveness using

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

a simulation experiment. As a result, it can be expected to remove the interference signal without giving information on how the interference sound moves at each time.

II. Problem formulation

We consider a determined or an overdetermined situation where J sound source signals are mixed and captured by I ($\geq J$) microphones. The Short-Time Fourier Transform (STFT) coefficients of the j-th estimated source and i-th microphone signals are denoted by

$$\boldsymbol{y}_{fn} = [y_{1fn}, \ldots, y_{Jfn}]^{\mathsf{T}} \in \mathbb{C}^J, \tag{1}$$

$$\boldsymbol{x}_{fn} = [x_{1fn}, \ldots, x_{Ifn}]^{\mathsf{T}} \in \mathbb{C}^I. \tag{2}$$

Here, $f = 1, \ldots, F$ and $n = 1, \ldots, N$ are the indices of the frequency and frame, respectively and $(\cdot)^{\mathsf{T}}$ denotes the transpose.

We consider an instantaneous mixture model in the time-frequency domain where the STFT window length is sufficiently longer than the impulse response between the sound source and microphones. Then, the relationship between the observed signals $\boldsymbol{x}_{fn}$ and estimated sources $\boldsymbol{y}_{fn}$ can be expressed as

$$\boldsymbol{y}_{fn} = \boldsymbol{W}_{fn}\boldsymbol{x}_{fn}, \tag{3}$$

where, $\boldsymbol{W}_{fn} = [\boldsymbol{w}_{1fn}, \ldots, \boldsymbol{w}_{Jfn}]^{\mathsf{H}}$ is a time-variant demixing matrix containing demixing filters $\boldsymbol{w}_{jfn} \in \mathbb{C}^I$, and $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. Our goal is to estimate the source signal $\boldsymbol{y}_{fn} = [y_{1fn}, \ldots, y_{Jfn}]^{\mathsf{T}}$ and solve the global permutation problem.

III. Conventional method

In this section, we introduce two conventional methods: Offline-GC-AuxIVA-ISS and Online-GC-AuxIVA-ISS. Then, we describe the drawback of the conventional methods in Section III-C.

A. Offline-GC-AuxIVA-ISS [11]

Offline-GC-AuxIVA estimates an time-invariant demixing matrix $\boldsymbol{W}_f = [\boldsymbol{w}_{1f}, \ldots, \boldsymbol{w}_{Jf}]^{\mathsf{H}}$. The demixing matrix $\mathcal{W} = \{\boldsymbol{W}_f\}_f$ can be estimated by minimizing the following negative log-likelihood function:

$$\mathcal{L}_{\text{IVA}}(\mathcal{W}) = \sum_{j=1}^{I} \mathbb{E}[G(\boldsymbol{y}_{jn})] - \sum_{f=1}^{F} \log|\det \boldsymbol{W}_f|, \tag{4}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator and $\boldsymbol{y}_{jn}$ is the vector representation of the estimated sources:

$$\boldsymbol{y}_{jn} = [y_{j1n}, \ldots, y_{jFn}]^{\mathsf{T}} \in \mathbb{C}^F. \tag{5}$$

Here, $G(\boldsymbol{y}_{jn}) = -\log p(\boldsymbol{y}_{jn})$ is the contrast function where $p(\boldsymbol{y}_{jn})$ represents a multivariate probability density function at the j-th source. One typical choice of the contrast function is to use a spherical contract function [3], [4], [5], which is expressed as

$$G(\boldsymbol{y}_{jn}) = G_R(r_{jn}), \tag{6}$$

$$r_{jn} = ||\boldsymbol{y}_{jn}||_2 = \sqrt{\sum_f |y_{jfn}|^2} = \sqrt{\sum_f |\boldsymbol{w}_{jf}^{\mathsf{H}}\boldsymbol{x}_{fn}|^2}. \tag{7}$$

Here, $G_R(r)$ is a function of a real-valued scalar variable r, and $||\cdot||_2$ denotes the L_2 norm of a vector. Adopting the auxiliary function approach [5], the following equation is optimized instead of (4):

$$\begin{aligned}\mathcal{L}_{\text{IVA}}(\mathcal{W}) &\leq \mathcal{L}_{\text{AuxIVA}}(\Sigma, \mathcal{W}) \\ &= \frac{1}{2}\sum_{f=1}^{F}\sum_{j=1}^{J} \boldsymbol{w}_{jf}^{\mathsf{H}}\boldsymbol{\Sigma}_{jf}\boldsymbol{w}_{jf} - \sum_{f=1}^{F}\log|\det \boldsymbol{W}_f|,\end{aligned} \tag{8}$$

where $\Sigma = \{\boldsymbol{\Sigma}_{jf}\}_{jf}$ is the weighted spatial covariance matrix calculated as

$$\boldsymbol{\Sigma}_{jf} = \sum_n \varphi(r_{jn})\boldsymbol{x}_{fn}\boldsymbol{x}_{fn}^{\mathsf{H}}. \tag{9}$$

Here, $\varphi(r_{jn}) = G_R^{'}(r_{jn})/r_{jn}$ and $(\cdot)'$ denotes the derivative operator.

To avoid the global permutation problem, GC-AuxIVA-ISS uses geometric constraints [6] that controls the j-th demixing filter to respond a value $c_{j\theta}$ toward θ direction. The regularization term of the geometric constraint is expressed as

$$\mathcal{L}_{\text{GC}}(\mathcal{W}) = \sum_{f=1}^{F}\sum_{j=1}^{J}\sum_{\theta\in\Theta} \lambda_{j\theta}|\boldsymbol{w}_{jf}^{\mathsf{H}}\boldsymbol{d}_{f\theta} - c_{j\theta}|^2. \tag{10}$$

Here, $\Theta = \{\theta\}$ represents a set including all directions to be considered, $\boldsymbol{d}_{f\theta}$ is a steering vector pointing to the direction θ, $c_{j\theta}$ is a nonnegative value set for all frequency bins as constraints, and $\lambda_{j\theta} \geq 0$ is a parameter that weighs the importance of the constraint. When $c_{j\theta} = 1$, the jth demixing filter $\boldsymbol{w}_{jf}$ is guided to preserve the signals arriving from the θ direction. On the other hand, when the value of $c_{j\theta}$ is close to 0, the jth demixing filter $\boldsymbol{w}_{jf}$ is forced to create a spatial null to the θ direction. Thus, the objective function to be optimized in GC-AuxIVA is given by combining (8) and (10):

$$\mathcal{L}(\Sigma, \mathcal{W}) = \mathcal{L}_{\text{AuxIVA}}(\Sigma, \mathcal{W}) + \mathcal{L}_{\text{GC}}(\mathcal{W}). \tag{11}$$

Because there is no closed update rule to minimize the objective function in (11), GC-AuxIVA-ISS uses update an iterative update algorithm, ISS [11], [12]. In ISS, the demixing matrix is updated using the auxiliary variable $\boldsymbol{v}_{jf} = [v_{1jf}, \ldots, v_{Ijf}]^{\mathsf{T}} \in \mathbb{C}^I$ as follows:

$$\boldsymbol{W}_f \leftarrow \boldsymbol{W}_f - \boldsymbol{v}_{jf}\boldsymbol{w}_{jf}^{\mathsf{H}}. \tag{12}$$

Substituting (12) into the objective function in (11), we have a new objective function to be minimized. Then, we can obtain the update rule of the variable $\boldsymbol{v}_{jf}$ by calculating the partial derivative of the function and equating it to zero. Since the detailed update rule has been discussed in [11] and is not directly related to our main discussion, we skip explaining it.

B. Online-GC-AuxIVA-ISS [13]

Online-GC-AuxIVA-ISS uses time-variant demixing matrices $\boldsymbol{W}_{fn}$ and estimates the matrices in each time frame n.

In offline-GC-AuxIVA-ISS, spatial covariance matrix $\boldsymbol{\Sigma}_{jf}$ is calculated using all time frames of the observed signal. On the other hand in online-GC-AuxIVA-ISS, the spatial covariance matrix is updated in an autoregressive manner:

$$\boldsymbol{\Sigma}_{jfn} = \alpha \boldsymbol{\Sigma}_{jf(n-1)} + (1-\alpha)\varphi(r_{jn})\boldsymbol{x}_{fn}\boldsymbol{x}_{fn}^{\mathsf{H}}. \quad (13)$$

Here, $0 < \alpha < 1$ denotes a forgetting factor. The ISS update can be straightforwardly applied to online-GC-AuxIVA by replacing the update rule of the covariance matrices with (13). The detailed update rule of online-GC-AuxIVA-ISS is also written in [13].

We also need to consider how to model the geometrical constraints in online processing. In this paper, we use the following time-variant GC term for online-GC-AuxIVA-ISS:

$$\mathcal{L}_{\mathrm{GC}}(\boldsymbol{W}_{fn}) = \sum_{f=1}^{F}\sum_{j=1}^{J}\sum_{\theta_n \in \Theta_n} \lambda_{j\theta_n} |\boldsymbol{w}_{jfn}^{\mathsf{H}} \boldsymbol{d}_{f\theta_n} - c_{j\theta_n}|^2. \quad (14)$$

As discussed in the previous research [13], we obtain the steering vector $\boldsymbol{d}_{f\theta_n}$ in each frame n using a DOA estimation method, multiple signal classification (MUSIC) [16]. MUSIC is a subspace-based method that decomposes the spatial covariance matrix of the observed multichannel signals to obtain subspaces of signals and noise that are orthogonal to each other. The operation of MUSIC in this system is described in [13], [17]. We used [18] for implementing the MUSIC method.

C. Drawback of the conventional method in moving source separation

Let us consider a situation of the moving source separation where one interference source is moving as shown in Fig. 1. In this case, offline-GC-AuxIVA-ISS described in Section III-A estimates a time-invariant demixing filter using the signals of all time frames, and thus cannot follow changes in the environment such as the movement of sound source.

On the other hand, online-GC-AuxIVA-ISS estimates a time-variant demixing matrix using only the observed signals up to the present frame. Therefore, it can handle the movement of sound sources. However, when the interference source moves quickly, it is difficult to remove the interference signal in real time because we obtain less spatial information about the interference signal than when the interference is fixed. Moreover, it is not necessarily easy to prepare the accurate direction θ_n in each time frame n to provide geometrical constraints. Although there is a research to estimate the DOAs for online-GC-AuxIVA-ISS, the accuracy of the DOA estimation is limited because the estimation is conducted by using the observed mixture signals up to the present.

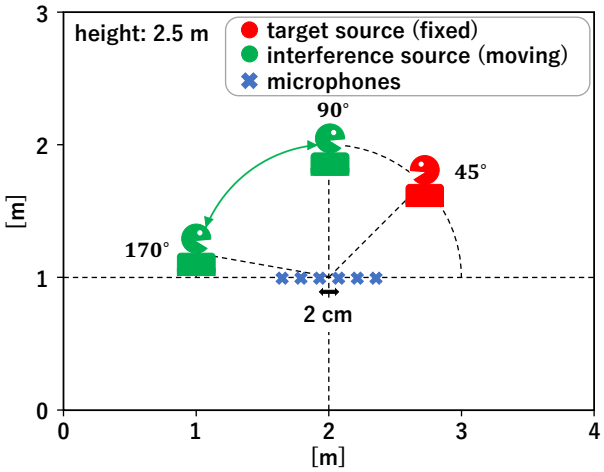


Fig. 1: Layout of a situation of the moving sourece separation. The target speaker is at 45°. The interference speaker moves on an arc from 90° to 170° for the first 10 seconds and from 170° to 90° for the next 10 seconds.

IV. Proposed Method: MGC-AuxIVA-ISS

In this section, we propose a method to solve the issues described in Section III-C. Although it is difficult to get the positions of each source at each time, it is easier to obtain the range of the moving sound source. Thus, we propose to remove the range of the moving source by utilizing geometric constraints. In concrete, we extend offline-GC-AuxIVA-ISS to an overdetermined condition and provide null constraints in multiple directions to remove the moving interference signal. We call this proposed method Multiple GC-AuxIVA-ISS (MGC-AuxIVA-ISS). Here, we assume to obtain the range of the moving interference with a range ϕ. For example in Fig. **??**, when the interference source move in the range in 90° to 170° ($\phi = \{\psi \mid 90° \leq \psi \leq 170°\}$) and we have 6 microphones, we set $\Theta = \{90, 110, \ldots, 170\}$, and apply the geometric constraint $c_{1\theta} = 0$ for all $\theta \in \Theta$. Thus, we expect that our MGC-AuxIVA-ISS can remove the entire moving range of the interference signal, regardless of the sound source's movement. This sets up $I - 1$ null constraints equally spaced for ϕ. In this study, we assume the direction range ϕ is known, and let its estimation be a future issue. The cost function for MGC-AuxIVA-ISS is equivalent to that of the Offline-GC-AuxIVA-ISS, and the parameter optimization algorithm is similar.

V. Experimental Evaluation

A. Experimental conditions

We conducted simulation experiments to evaluate the separation performance of MGC-AuxIVA-ISS in the situation of moving source separation. In this experiment, we used speech signals for 6 speakers (3 male and 3 female) selected randomly from a total of 503 sentences in Set B of the ATR Digital Japanese Speech Database [19]. To generate the

mixture signals, we randomly selected speech signals from two different speakers and created a total of 50 patterns of mixture signals whose length is 20 seconds. We used `signal generator` [20] to create the mixture signals. We show the layout of sound sources and microphones using Fig. 1. The DOA of the target signal was fixed at $45°$ for all 20 seconds and the interference source moved on an arc from $90°$ to $170°$ for the first 10 seconds and moved on an arc from $170°$ to $90°$ for the next 10 seconds. We used 6 microphones and set the distance between adjacent microphones to 2 cm. The distance between sound sources and microphones was set to 1 m. We set the reverberation times (RT_{60}) to 200 ms. We conducted the convolution at a sampling frequency of 48 kHz and then resampled the convoluted signals at 16 kHz to separate the mixtures.

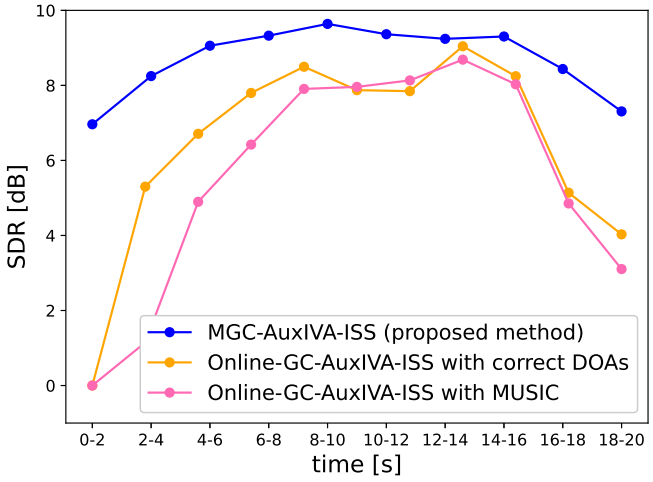
In this experiment, we used the following three methods, MGC-AuxIVA-ISS, online-GC-AuxIVA-ISS with correct DOAs and online-GC-AuxIVA-ISS with MUSIC. We used 6 microphones for MGC-AuxIVA-ISS, and 2 microphones for online-GC-AuxIVA-ISS. We gave information on the moving range of the interference source for MGC-AuxIVA-ISS. On the other hand, we gave the DOA of the interference signal at each time frame to online-GC-AuxIVA-ISS with correct DOAs. Online-GC-AuxIVA-ISS with MUSIC estimated the interference's DOA using MUSIC in each frame and applied the DOA into geometrical constraints. In this experiment, we used null constraints where $c_{j\theta} = 0$ and spatial nulls are formed in the direction of θ. The STFT was computed using a Hanning window, whose length and shift were set at 2048 samples (128 ms) and 1024 samples (64 ms), respectively. The number of iterations is set to 50 for MGC-AuxIVA-ISS and 2 in each frame for online-GC-AuxIVA-ISS. In online-GC-AuxIVA-ISS, the forgetting factor α is set to 0.99, and the spatial covariance matrix $\mathbf{\Sigma}_{jf0}$ is initialized as an identity matrix.

The following two objective metrics were used to evaluate the separation performance: Source-to-Distortions Ratio (SDR) and Source-to-Interferences Ratio (SIR) [21]. Higher values of SDR and SIR indicate better separation performance. The values shown in the experimental results are obtained by tuning the parameter $\lambda_{j\theta}$ so that the accuracy of output signal order is 100% and SDR is the highest value.

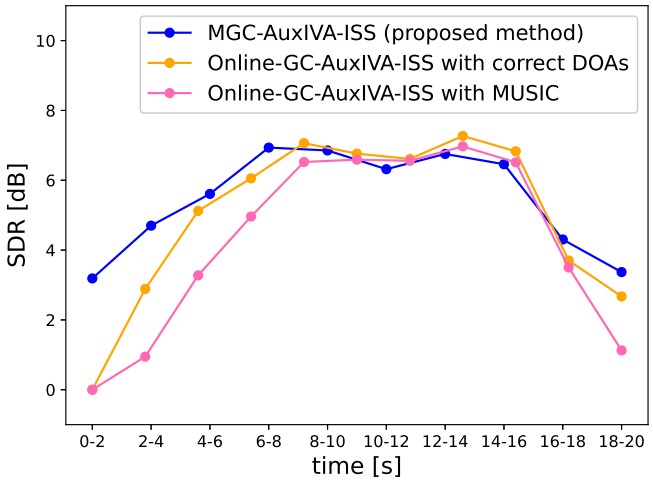
B. Results

We compared the SDR and SIR improvement of MGC-AuxIVA-ISS and the conventional methods using Fig. 2 and Table I. Figure 2 shows the average SDR for the emphasized target signal every 2 seconds obtained by each method. Table I shows the average SDR and SIR of the target signal enhanced by each method in every 2 seconds. To evaluate the SDRs with and without background noise, we conducted two experiments, one without background noise and the other with a noise added by adjusting Source-to-Noise Ratio (SNR) = 25 dB.

We first show the result without background noise. In this experiment, we gave null constraints towards



(a) No background noise

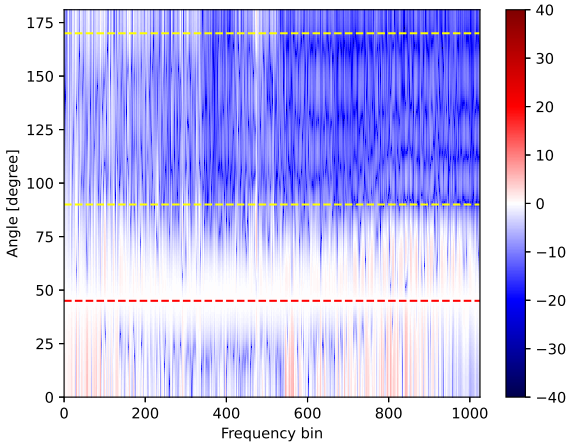


(b) SNR = 25 dB

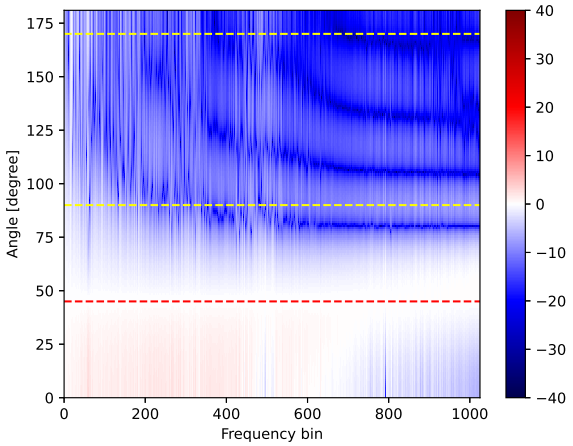
Fig. 2: Average SDR of the target signal enhanced by each method in every 2 s.

$\{90°, 110°, 130°, 150°, 170°\}$ and set $\lambda_{j\theta} = 100$ for MGC-AuxIVA-ISS. We set $\lambda_{j\theta} = 300$ for online-GC-AuxIVA-ISS with correct DOAs, and $\lambda_{j\theta} = 10$ for online-GC-AuxIVA-ISS with MUSIC. From Fig. 2 (a), MGC-AuxIVA-ISS shows the highest SDR overall times. Similarly, Table. I (a) shows that MGC-AuxIVA-ISS achieves the average SDR improvement of 1.64 dB and the average SIR improvement of 5.88 dB compared to online-GC-AuxIVA-ISS with correct DOAs. These results indicate that MGC-AuxIVA-ISS can remove the moving interference signal better than the conventional online-GC-AuxIVA-ISS.

Next, we show the result in a noisy environment. In this experiment, we gave null constraints towards $\{80°, 105°, 130°, 155°, 180°\}$ and set $\lambda_{j\theta} = 3590$ for MGC-AuxIVA-ISS. We set $\lambda_{j\theta} = 30$ for online-GC-AuxIVA-ISS with correct DOAs, and $\lambda_{j\theta} = 20$ for online-GC-AuxIVA-ISS



(a) No background noise



(b) SNR = 25 dB

Fig. 3: Examples of directivity patterns obtained by MGC-AuxIVA-ISS. Red line shows the position of the target source. Yellow lines show the range of the interference source.

with MUSIC. From Fig. 2 (b), MGC-AuxIVA-ISS shows the highest SDR before 4 seconds and after 16 seconds. However, the SDR of MGC-AuxIVA-ISS was close to that of online-GC-AuxIVA-ISS with correct DOAs between 4 to 16 seconds. Table I (b) also shows that MGC-AuxIVA-ISS performs equivalent SDR and higher SIR than online-GC-AuxIVA-ISS with correct DOAs. On the other hand, when compared to online-GC-AuxIVA-ISS with MUSIC, MGC-AuxIVA-ISS achieves the average SDR improvement of 0.77 dB. This result means that when we do not know the correct DOA but know the range of moving interference, using several null constraints toward the moving range is a good way, even in a noisy environment.

Finally, we show the directivity pattern using Fig. 3. The preferred result in Fig. 3 is that the demixing filter removes the steering vector arriving from the moving range of the

TABLE I: Average SDR [dB] and SIR [dB] of the target signal enhanced by each method. In these results, the accuracy of output order was 100%.

(a) No background noise

method	SDR [dB]	SIR [dB]
MGC-AuxIVA-ISS (proposed method)	8.69	16.01
Online-GC-AuxIVA-ISS with correct DOAs	7.05	10.13
Online-GC-AuxIVA-ISS with MUSIC	6.11	10.00

(b) SNR = 25 dB

method	SDR [dB]	SIR [dB]
MGC-AuxIVA-ISS (proposed method)	5.47	10.12
Online-GC-AuxIVA-ISS with correct DOAs	5.50	9.87
Online-GC-AuxIVA-ISS with MUSIC	4.70	9.69

interference source while remaining that from $45°$. So, we normalized the response in each frequency by the response from $45°$. This result shows that MGC-AuxIVA-ISS removes the moving range between $90°$ and $170°$.

VI. Conclusions

In this paper, we proposed MGC-AixIVA-ISS to remove the moving interference speaker. Online-GC-AuxIVA-ISS, which is effective for the movement of the interference source, requires accurate DOAs of the interference signal for each time frame. On the other hand, MGC-AuxIVA-ISS extends offline-GC-AuxIVA-ISS to the overdetermined condition and provides equally spaced null constraints within the moving range of the interference source. To investigate the effectiveness of MGC-AuxIVA-ISS, we conducted speech enhancement experiments to remove the moving interference signal. As a result, we confirmed that the separation performance of MGC-AuxIVA-ISS is more than that of online-GC-AuxIVA-ISS.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23H03423.

References

[1] S. Makino, T.-W. Lee, and H. Sewada, *Blind Speech Separation*, Springer, 2007.
[2] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[3] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.
[4] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.
[5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[6] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.
[7] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. ICASSP*, 2018, pp. 746–750.
[8] A. H. Khan, M. Taseska, and E. A. P. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.
[9] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. ASLP*, vol. 15, no. 2, pp. 715–726, 2007.
[10] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. ICASSP*, 2020, pp. 846–850.
[11] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering," in *Proc. EUSIPICO*, 2022, pp. 757–761.
[12] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. ICASSP*, 2020, pp. 236–240.
[13] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Accelerating online algorithm using geometrically constrained independent vector analysis with iterative source steering," in *Proc. APSIPA*, 2022, pp. 755–760.
[14] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, 2014, pp. 107–111.
[15] L. Li, K. Koishida, and S. makino, "Online directional speech enhancement using geometrically constrained independent vector analysis," in *Proc. Interspeech*, 2020, pp. 61–65.
[16] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.
[17] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Investigation of partial update and source localization in geometrically constrained independent vector analysis with auxiliary function approach for moving source," in *Proc. NCSP*, 2023, pp. 426–429.
[18] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulations and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.
[19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.
[20] "Signal generator," *https://www.audiolabs-erlangen.de/fau/professor/ habets/software/signalgene*.
[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Dual-Channel Three-Stage Model for DoA and Speech Enhancement

Meng-Hsuan Wu*†, Yih-Liang Shen*, Hsuan-Cheng Chou*, Bo-Wun Shih† and Tai-Shih Chi*
* Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan
† Realtek Semiconductor Corp., Hsinchu, Taiwan
luckily75368@gmail.com, dennis831209@gmail.com, azqlmpgghh@gmail.com, bowuns@realtek.com, tschi@nycu.edu.tw

***Abstract*—During the pandemic, teleconferencing becomes a necessity to our daily lives. It drives the demand for an integrated system which is not only able to effectively enhance speech sounds, but also to localize the speaker for video enhancement. In this paper, we propose a neural network based composite system which integrates a DoA estimator and a neural beamformer for dual-channel speech enhancement. The proposed system can accomplish two tasks at the same time by using sound signals received from dual microphones. The estimated DoA is converted into a spatial angle related feature, which provides complementary information to boost performance of the neural beamformer. The proposed system is evaluated in simulated far-field conditions with reverberations and noise. Simulation results demonstrate the proposed system outperforms stand-alone baseline systems in either one of the two tasks and achieves comparable results to the best stand-alone models in either one of the two tasks.**

I. Introduction

During the COVID pandemic, online teleconferencing has become a part of people's daily lives. To enhance voice quality, many single-channel and multi-channel speech enhancement methods have been proposed to reduce reverberation and ambient noise. Conventionally, for the multi-microphone scenario, the direction of arrival (DoA) estimation and the beamforming technology were combined to receive the speech signal coming from a specific direction while suppressing interference in other directions. For DoA estimation, the generalized cross correlation with phase transform (GCC-PHAT) [1], the steered-response power with phase transform (SRP-PHAT) [2], and the multiple signal classification (MUSIC) [3] methods are often used. As for beamforming, the delay and sum beamforming (DSB), the minimum variance distortionless response (MVDR) [4], and the linearly constrained minimum variance (LCMV) [5] are popular methods.

In recent years, with the development of deep learning, neural-network (NN) based methods have made substantial progress in both single-channel [6][7][8][9] and multi-channel speech enhancement. In multi-channel speech enhancement, some approaches use NNs to aid traditional beamforming method [10][11][12][13], some use NNs to directly simulate the beamforming mechanism [14][15][16]. For example, [15] directly estimates the weights of a filter-and-sum beamformer using the U-net, whose inputs are those signals received from microphones, and produces the best results in ConferencingSpeech 2021 Challenge [17]. Meanwhile, for robust DoA estimation, [18][19] use NNs to find the time-frequency (TF) bins dominated by the direct sound for calculating weighted GCC-PHAT to estimate DoA. Another type of approach is to directly estimate DoA based on the original spectrograms [20] or the masked spectrograms [21] of microphone signals.

Above mentioned studies treated DoA and beamforming/enhancement separately, i.e., the loss for training was either on DoA or speech enhancement. However, in real world applications, functions of DoA and beamforming are often required at the same time and their performance are intertwined, i.e., a good DoA estimate is beneficial to beamforming while enhanced speech signals help estimate DoA. Intuitively, both tasks can be co-optimized in an iterative fashion. Therefore, in this paper, we propose a unified 3-stage enhancement-DoA-enhancement system using both DoA and speech enhancement loss. The 3-stage system integrates NN-based mask predictors, a NN-based DoA estimator, and a neural beamformer to produce robust performance on both DoA estimation and dual-channel speech enhancement. In the system, the spatial angle related feature is derived from the estimated DoA [22] after single-channel enhancement. The spatial feature is then used as a complementary feature to the neural beamformer for further boosting performance of dual-channel speech enhancement. Not like [12], the proposed system is fully composed of neural networks and all modules can be trained simultaneously. Evaluation results demonstrate the proposed composite system outperforms stand-alone baseline systems in either one of the two tasks. Results also show the proposed system achieves comparable performance to the best stand-alone DoA estimation system and the best stand-alone neural beamformer in either one of the tasks. Under the mild reverberant conditions, the proposed system even slightly outperforms the best stand-alone neural beamformer.

The rest of the paper is organized as follows. In Section 2, we introduce the signal model and the proposed system architecture. In Section 3, the datasets, experimental settings and results are presented and discussed. Finally, we conclude this paper in Section 4.

II. Proposed System

A. Signal model

Fig.1 shows the simulation setup. The distance between the two microphones is 7 cm, and the microphone array is placed 1 meter in height and at least 50 cm from the wall. To simulate a real-world condition, the environment contains diffuse noise,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

directional noise, reverberation, and microphone device noise. The problem formulation is as follows:

$$y_i = s * d_i + s * h_i^{reverb} + n_i^{total} + n_i^{device},\ i = 1,\ 2 \quad (1)$$

where $*$ is the convolution, y_i is the signal received by the i-th microphone, and s is the target source signal. d_i and h_i^{reverb} represent the direct path, and combinations of early and late reflection paths from the target source to the microphone. n_i^{total} is composed of diffuse noise n_i^{dif} and directional noise n_i^{dir} as follows:

$$n_i^{total} = n_i^{dif} + n_i^{dir} \quad (2)$$

$$n_i^{dif} = \sum_{k=1}^{37} n_{dif}^{(k)} * h_{k,i}^{dif},\ i = 1,\ 2 \quad (3)$$

$$n_i^{dir} = n_{dir} * h_i^{dir},\ i = 1,\ 2 \quad (4)$$

The diffuse noise is simulated by placing 37 different segments of a noise signal, $n_{dif}^{(k)}$, at 37 locations in the step of 5° from 0° to 180° on the red semicircle, whose diameter is the length of the room. $h_{k,i}^{dif}$ represents the room impulse response (RIR) from the location k to the i^{th} microphone. The directional noise is used to simulate the outdoor noise leaking through the window. In our experiments, we place the directional noise source at random locations on the three walls in front of the microphone array. h_i^{dir} represents the RIR from the directional noise source to the i-th microphone. Finally, we add device thermal noise n_i^{device} (white noise, SNR = 50 dB) to the two microphone signals. All sound sources are placed on the same horizontal plane as the microphone array.

Above equations are all in the time domain. In simulations, we transformed signals from the time domain to the time-frequency (TF) domain using the short-time Fourier transform (STFT) with the frame length of 512 and the hop size of 256. In this paper, capital letters are used to represent the spectrograms of signals in the TF domain.

B. Proposed system architecture

As shown in Fig.2, the proposed 3-stage system is composed of a magnitude mask predictor, a DoA estimator, and a neural beamformer.

The system is trained in three stages. The first stage is to train the magnitude mask predictor, which gives the DoA estimator and the neural beamformer clues about which TF bins are dominated by the target direct sound. The mask predictor is a monaural system which takes magnitude of the spectrogram as input, and the ideal ratio mask (IRM) between the target direct sound and the received signal as the training target.

In the second stage, we fix the mask predictor and train the DoA estimator with the time difference of arrival (TDOA) as the training target. The DoA estimator takes the summation of masked GCC-PHAT along the frequency axis as the input.

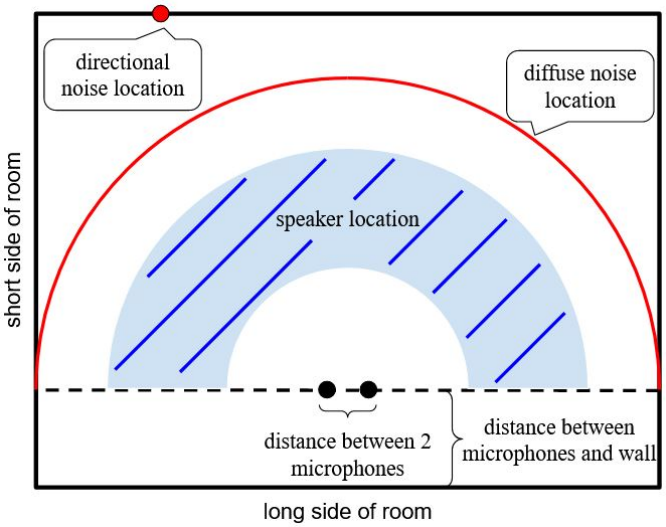


Fig. 1. Schematic diagram of the simulation environment.

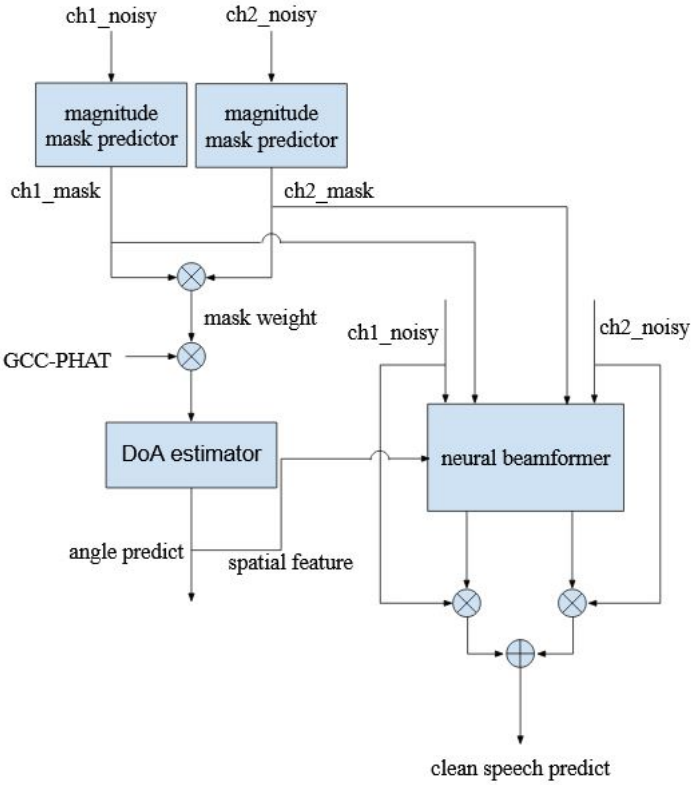


Fig. 2. Architecture of the proposed 3-stage system.

For convenience, the GCC in following equations is the abbreviation of GCC-PHAT:

$$DoA_{input}(\tau,\ t) = \sum_f GCC_{masked}\ (\tau,\ t,\ f) \quad (5)$$

$$GCC_{masked}\ (\tau,\ t,\ f) = \eta(t,\ f) GCC(\tau,\ t,\ f) \quad (6)$$

where $DoA_{input}(\tau,\ t)$ is the input of the DoA estimator and

$GCC_{masked}\ (\tau,\ t,\ f)$ is the TF-weighted version of GCC-PHAT. The weight $\eta(t,\ f)$ is the element-wise multiplication of two predicted masks from the magnitude mask predictor as:

$$\eta(t,\ f) = \hat{M}_1(t,\ f) \times \hat{M}_2(t,\ f) \tag{7}$$

The GCC-PHAT is calculated using received signals of the two microphones and the TDOA parameter τ between two microphones as:

$$\begin{aligned} GCC(\tau,\ t,\ f) &= Real\frac{Y_1(t,\ f)Y_2^H(t,\ f)}{|Y_1(t,\ f)||Y_2^H(t,\ f)|}e^{-j2\pi\frac{f}{N}f_s\tau} \\ &= cos(\angle Y_1(t,\ f) - \angle Y_2(t,\ f) - 2\pi\frac{f}{N}f_s\tau) \end{aligned} \tag{8}$$

where $\angle Y_1(t,\ f)$ and $\angle Y_2(t,\ f)$ represent the phase of the spectrograms of received signals, f_s is the sampling rate in Hz, N is the number of DFT points, f is the frequency bin index, and τ is considered from 37 possible locations (in the step of 5° from 0° to 180°).

Finally, the neural beamformer is added for the third stage training, and the parameters in the magnitude mask predictor and the DoA estimator are fine-tuned simultaneously and optimized during this stage of training. The neural beamformer takes the complex spectrograms of the two microphone signals, two predicted masks, and the estimated spatial angle related feature (SF) as input. The SF is derived as follows:

$$SF(t,\ f) = cos(\angle Y_1(t,\ f) - \angle Y_2(t,\ f) - 2\pi\frac{f}{N}f_s\hat{\tau}) \tag{9}$$

where $\hat{\tau}$ is the estimated TDOA from the DoA estimator.

C. Models and loss functions

The magnitude mask predictor is based on the popular convolutional recurrent neural network (CRN) [23]. It is composed of an encoder, a recurrent module and a decoder. The size of the convolution kernel and stride are respectively set to [5, 2] and [2, 1], and the kernel numbers are set to [16, 32, 64, 128, 256, 256] for the layers in the encoder. Each layer of the decoder mirrors the corresponding layer of the encoder except the transposed convolution layer is used. The middle LSTM recurrent module is constructed with 2 layers of 256 hidden units.

The DoA estimator is constructed with a 3-layer LSTM with 91 hidden units in each layer, followed by a fully connected layer. As for the neural beamformer, each of the input features (two complex spectrograms, two predicted masks, and the estimated spatial angle related feature) is first transformed by a 2-layer fully connected layer, and then concatenated for analysis by a 4-layer LSTM with 512 hidden units in each layer. At the end, the output of the LSTM is connected to four sets of 3-layer FC layer to predict four masks of real and imaginary parts of two complex spectrograms.

In the first stage of training, we use mean absolute error (MAE) as the loss function to calculate the error between the estimated mask and the IRM as in (10). In the second stage of training, MAE is used to calculate the error between the estimated and the true TDOA as in (11). In the third stage of training, we optimize the overall system with the DoA loss and the beamformer loss as in (12) and (13).

$$Loss_{stage1} = \frac{1}{T \times F}\|IRM_i - \hat{M}_i\|_1 \tag{10}$$

$$Loss_{stage2} = \|\tau - \hat{\tau}\|_1 \tag{11}$$

$$Loss_{stage3} = Loss_{stage2} + \alpha \times Loss_{beam} \tag{12}$$

$$\begin{aligned} Loss_{beam} = \frac{1}{T \times F}(&\|Real\{S\} - Real\{\hat{S}\}\|_1 \\ &+\|Imag\{S\} - Imag\{\hat{S}\}\|_1 + \||S| - |\hat{S}|\|_1) \end{aligned} \tag{13}$$

where α is a time-varying parameter to roughly normalize the DoA loss and the beamformer loss epoch by epoch.

III. EXPERIMENTAL RESULTS

A. Dataset preparation and training parameters

We selected the TIMIT database [24] as the source of clean speech. TIMIT contains speech samples of 630 American English speakers from different dialect regions of the United States. We used all 4620 sentences in the TRAIN category for training and randomly selected 48 and 30 sentences in the TEST category for validation and testing. When building the validation and test sets, we excluded the 'sa1' and 'sa2' sentences because these two sentences were spoken by all speakers. Meanwhile, we selected the PNL 100 nonspeech sounds database [25] as the source of noise to make diffuse noise and directional noise in training and validation data. In the test data, we selected 10 stationary noise from Noise92 to make diffuse noise, and other non-stationary noise to make directional noise.

To generate reverberant speech, we used the RIR generator [26] for synthesizing RIRs. Three different room sizes ([10, 7, 3], [15, 12, 3], [20, 17, 3] m) were used in the training set, two different room sizes ([12, 10, 3], [17, 15, 3] m) were used in the validation set, and two different room sizes ([10, 8, 3], [18, 15, 3] m) were used in the test set. The reverberation time T60 of 0.2, 0.4, and 0.6 s were used in the training set, 0.3 and 0.5 s were used in the validation set, and 0.25, 0.35, 0.45, and 0.55 s were used in the test set. SNR values of 0, 5, 10, and 15 dB were used in the training set, and 0, 10, and 20 dB were used in the validation and test sets. Note that our main goal is to solve the teleconferencing problem and the noise level in such venues seldom drops to negative SNR. The distance between the talker and the microphone array was randomly set from 2.5 to 5 meters, and the distance between the microphone array to the back wall was randomly set from 0.5 to 2.5 meters. In the end, about 124 hours of training data, 2.6 hours of validation data, and 13 hours of test data were generated.

In the three training stages, we used Adam optimizer and trained 100 epochs. Batch size was set to 128 in the first training stage and 32 in the second and third stages. Learning rate was set to 0.001 in the first training stage and 0.0001

TABLE I
RESULTS OF DOA ESTIMATION AND SPEECH ENHANCEMENT ON 17280 TEST SENTENCES (SAMPLING RATE = 16 KHZ). SCORES WERE AVERAGED OVER ALL TESTED CONDITIONS.

	DoA	Speech Enhancement			
	avg. error / accuracy	PESQ	CSIG	CBAK	COVL
Noisy	16.41 / 0.41	1.22	2.24	1.37	1.64
(DoA - baseline [19])	4.06 / 0.82	-	-	-	-
(DoA - A)	2.64 / 0.90	-	-	-	-
(beamformer - baseline [15])	-	1.71	2.98	2.25	2.29
(beamformer - A)	-	1.94	3.21	2.42	2.53
(beamformer - B)	-	1.94	**3.25**	2.41	**2.55**
(proposed)	**2.60 / 0.91**	1.92	3.19	2.41	2.51
(proposed-SF)	2.64 / **0.91**	**1.95**	3.23	**2.43**	2.54

in the second and third stages. All of the network weights were initialized by the method described in [27], the default initialization method of Pytorch [28].

B. Results

For evaluations, we compared the following systems:

- Noisy: The neural beamformer is turned off. The DoA estimator finds the $\hat{\tau}$ that maximizes the GCC-PHAT (8) by exhaustive search and outputs the corresponding angle.
- (DoA - baseline [19]): It refers to the subsystem with the fixed magnitude mask predictor and the DoA estimator. It is obtained after the second stage of training.
- (DoA - A): It refers to the subsystem with the adjustable magnitude mask predictor and the DoA estimator. It is obtained from (DoA - baseline) by jointly training the mask predictor and the DoA estimator.
- (beamformer - baseline): It refers to the neural beamformer trained only using microphone signals. It is similar to the system proposed in [15], except it uses LSTM and FC layers rather than U-net.
- (beamformer - A): It refers to the neural beamformer trained using outputs of the fixed magnitude mask predictor and microphone signals.
- (beamformer - B): It refers to the neural beamformer trained using outputs of the adjustable magnitude mask predictor and microphone signals. It is obtained from (beamformer - A) by jointly training the mask predictor and the neural beamformer.
- (proposed): It refers to the proposed 3-stage system. However, the spatial feature SF produced by the DoA estimator is not sent to the neural beamformer. Not like (beamformer - B), the mask predictor is now optimized for both DoA and beamforming tasks.
- (proposed-SF): It is identical to (proposed) except SF now being sent to the neural beamformer.

We used 17280 test sentences to evaluate the proposed system. We used mean absolute error and gross accuracy as criteria for DoA estimation, and four objective metrics, PESQ [29], CSIG, CBAK, and COVL [30], as criteria for dual-channel speech enhancement. In calculating gross accuracy, the case of the absolute error within 5° was considered correct.

The experimental results are shown in Table I. For speech enhancement, by comparing (beamformer - baseline) with (beamformer - A), we observe that adding the mask predictor to the system makes a significant improvement on system performance. The comparison between (beamformer - A) and (beamformer - B) shows that allowing the weights of the mask predictor to be modified during training of the neural beamformer makes a slight improvement on system performance. In contrast, by comparing (DoA - baseline) with (DoA - A), we observe that allowing the weights of the mask predictor to be modified during training of the DoA estimator makes a significant improvement on its estimation performance. These results indicate the joint training in (DoA - A) and (beamformer - B) has the mask predictor produce a more suitable mask for either task.

TABLE II
RESULTS OF PESQ SCORES UNDER DIFFERENT T60. SCORES WERE AVERAGED OVER ALL TESTED CONDITIONS INCLUDING ALL SNRS, RTS AND SPATIAL DISTANCES.

T60 (s)	0.25	0.35	0.45	0.55
Noisy	1.31	1.22	1.18	1.15
(beamformer - B)	2.14	1.98	**1.87**	**1.77**
(proposed-SF)	**2.19**	**1.99**	1.86	1.75

Next, by comparing (proposed) with (beamformer - B), we can see that when the mask predictor is simultaneously optimized for both tasks as in (proposed), it provides less benefit than the stand-alone mask predictor in (beamformer-B). However, when we combine the spatial feature SF from the DoA estimator into the neural beamformer, the proposed (proposed-SF) system gains back lost performance to achieve comparable performance with the stand-alone (beamformer-B) system. We rearrange PESQ scores based on T60 in Table II. It shows that (proposed-SF) system performs slightly better in less reverberant environments while the stand-alone (beamformer-B) system performs slightly better in more reverberant environments.

C. Limitation of the Study

In this study, all RIRs were synthesized using the RIR generator [26]. No real-world RIRs were measured to train the proposed 3-stage model. Hence, the performance of the model will inevitably deteriorate in real-world environments. Using a small amount of real-world data to fine-tune the model is necessary for product development.

IV. Conclusion

In this paper, we propose an integrated 3-stage system containing the magnitude mask predictor, the DoA estimator and the neural beamforemer to simultaneously tackle two common tasks in teleconferencing. Simulation results demonstrate the performance of the neural beamformer is significantly improved when adding the predicted magnitude masks as beamformer's input. In addition, by adding the spatial angle related feature, the neural beamformer achieves comparable results to the best stand-alone beamformer's model. As for DoA estimation, the proposed system also outperforms the best stand-alone DoA model. In the future, we will combine the proposed system with a dual-channel AEC system to constitute a complete system for teleconferencing.

References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*, pp. 157–180. Springer, 2001.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] Byung-Chul Kim and I-Tai Lu, "High resolution broadband beamforming based on the mvdr method," in *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings*, 2000, vol. 3, pp. 1673–1676.

[5] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[6] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy, "Attention wave-u-net for speech enhancement," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 249–253.

[7] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[8] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," *arXiv preprint arXiv:2008.04259*, 2020.

[9] Hao Shi, Longbiao Wang, Sheng Li, Jianwu Dang, and Tatsuya Kawahara, "Monaural speech enhancement based on spectrogram decomposition for convolutional neural network-sensitive feature extraction," *Proc. Interspeech 2022*, pp. 221–225, 2022.

[10] Soumitro Chakrabarty and Emanuël A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[11] Takuya Higuchi, Keisuke Kinoshita, Nobutaka Ito, Shigeki Karita, and Tomohiro Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *ICASSP*, 2018, pp. 531–535.

[12] Zhong-Qiu Wang and DeLiang Wang, "All-neural multi-channel speech enhancement.," in *Interspeech*, 2018, pp. 3234–3238.

[13] Ali Aroudi and Sebastian Braun, "Dbnet: Doa-driven beamforming network for end-to-end reverberant sound source separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 211–215.

[14] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, Ananya Misra, and Chanwoo Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[15] Xinlei Ren, Xu Zhang, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu, "A causal u-net based neural beamforming network for real-time multi-channel speech enhancement," in *Interspeech*, 2021, pp. 1832–1836.

[16] Andong Li, Wenzhe Liu, Chengshi Zheng, and Xiaodong Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6487–6491.

[17] Wei Rao, Yihui Fu, Yanxin Hu, Xin Xu, Yvkai Jv, Jiangyu Han, Zhongjie Jiang, Lei Xie, Yannan Wang, Shinji Watanabe, Zheng-Hua Tan, Hui Bu, Tao Yu, and Shidong Shang, "Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *ASRU*, 2021, pp. 679–686.

[18] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, "Robust tdoa estimation based on time-frequency masking and deep neural networks.," in *Interspeech*, 2018, pp. 322–326.

[19] Pasi Pertilä and Mikko Parviainen, "Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking," in *ICASSP*, 2019, pp. 436–440.

[20] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.

[21] Wangyou Zhang, Ying Zhou, and Yanmin Qian, "Robust doa estimation based on convolutional neural network and time-frequency masking.," in *INTERSPEECH*, 2019, pp. 2703–2707.

[22] Pasi Pertilä and Joonas Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[23] Ke Tan and DeLiang Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, 2018, vol. 2018, pp. 3229–3233.

[24] John S Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[25] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[26] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[29] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[30] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Weighted Binary Cross-Entropy for Sound Event Representation Learning and Few-Shot Classification

Zhongxin Bai*, Chao Pan*, Gong Chen*, Jingdong Chen*, and Jacob Benesty†
*CIAIC Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
†INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada

***Abstract*—The performance of sound event classification, including detection or tagging, depends heavily on the number of training samples and the quality of the training data. This paper presents an approach to improving sound event classification performance for events with limited training samples through using a weighted binary cross-entropy loss function. This function aims to constrain the representation space to have lower intra-class variance and higher inter-class differences by mining difficult samples and applying stricter penalties. Experiments demonstrate that the proposed method outperforms the existing ones, and the improvement is particularly significant in scenarios with limited training samples.**

I. Introduction

Sound event classification, including detection and tagging [1], is a process of analyzing audio signals to identify and categorize events present in the audio recordings. It has diverse applications in areas such as security, surveillance, environmental monitoring, and media content analysis, etc. Deep neural network (DNN)-based sound event classifiers can achieve impressive accuracy but typically require a large amount of training data. Due to the high cost of acquiring training data with fine-grained sound event annotations (i.e., specifying the type, onset time, and offset time of every event), many existing datasets, such as AudioSet [2], FSD50K [3], and SONYC-UST-V2 [4], opt for imprecise or coarse-grained labeling, i.e., only specifying the event types without temporal information. The fine-grained and coarse-grained annotations are also, respectively, referred to as strong-labeled and weak-labeled data [5].

Regarding different tasks of sound event classification, tagging focuses on determining the presence or absence of specific events in a given clip, while detection goes further by determining the onset and offset times of the events, typically operating at the frame-level. Consequently, multi-label learning on weakly labeled dataset is a natural choice for training sound event tagging DNN [6]. In contrast, for sound event detection, due to the scarcity of strongly labeled datasets, training models using weakly supervised learning techniques, such as multiple instance learning [1], [7], [8], serves as significant means to alleviate the burden of labeling efforts. In addition, there has also been increasing attention on semi-supervised learning to mitigate the reliance on labeled data [9], [10], [11], as seen in the Task 4 of the DCASE (Detection and Classification of Acoustic Scenes and Events) challenges[1]. While those methods are effective in reducing the burden of labeling efforts, the majority of those operate under the assumption that all training data belongs to events of the current task. Consequently, they encounter difficulties when confronted with few-shot tasks that possess minimal training data, regardless of whether it is labeled or unlabeled. A promising approach to improving performance in this scenario is to transfer knowledge from unrelated datasets. This approach often involves training a pre-trained representation network using techniques such as weakly supervised learning [6] or unsupervised approaches like contrastive learning [12]. Subsequently, the network is fine-tuned using the limited amount of training data available for the target events.

Large-scale weakly labeled datasets, such as Audioset [2], hold promise for training improved representation networks through multi-label classification. However, these datasets also come with challenges, such as label errors and class imbalance. Traditionally, the binary cross-entropy loss function is widely used in both pre-training and fine-tuning stages. Nevertheless, it falls short in addressing the class imbalance issue during pre-training and fails to achieve notable performance improvements during fine-tuning, especially when the available data for fine-tuning is extremely limited. To address this limitation, we propose to learn sound event representations on a hypersphere, and introduce a novel weighted binary cross-entropy loss function based on our previous work [13]. This weighted binary cross-entropy loss function is specifically designed to mitigate the class imbalance problem. Moreover, by incorporating a margin hyper-parameter, it not only promotes a more compact representation but also enhances the fine-tuning results, particularly in scenarios with very few training examples.

II. Preliminaries

Assume that the backbone of a sound event classification DNN is represented by $\hbar_\theta(\cdot)$, and its classification heads consist of a weight matrix $\mathbf{W} = [\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_J]$ and a bias vector $\mathbf{b} = [b_1 \ b_2 \ \cdots \ b_J]^T$, where θ represents the parameter of the backbone, J denotes the total number of event types, and the superscript ^{T} is the transpose operator. In the prediction stage, a given nth audio signal $\mathbf{x}_n$ is first embedded into a d-dimensional vector space by $\mathbf{z}_n = \hbar_\theta(\mathbf{x}_n) \in \mathbb{R}^d$ (learned representation). Then, the logits associating to the jth event is calculated as

$$s_{n,j} = \mathbf{w}_j^T \mathbf{z}_n + b_j \tag{1}$$

[1]https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

and the posterior probability is given by

$$p_{n,j} = \frac{1}{1 + e^{-s_{n,j}}}. \tag{2}$$

Comparing $p_{n,j}$ with a predefined threshold η, e.g., 0.5, one can predict whether the jth event is presented in the nth signal $\mathbf{x}_n$ or not. Note that the prediction corresponds to clip-level representations (e.g., 10 seconds) for the tagging task and frame-level representations (e.g., 25 milliseconds) for the detection task. This paper focuses on the clip-level representation learning.

In the training stage, let us denote $\mathcal{B} = \{(\mathbf{x}_n, \mathbf{y}_n) | n = 1, 2, \ldots, N\}$ as a min-batch of training representations, where the ground-truth $\mathbf{y}_n \in \{0, 1\}^J$ is

$$\mathbf{y}_n = \begin{bmatrix} y_{n,1} & y_{n,2} & \cdots & y_{n,J} \end{bmatrix}^T, \tag{3}$$

with $y_{n,j} \in \{0, 1\}$ marking the presence or absence of the jth event. The multi-label classification DNN is often trained by the following binary cross-entropy (BCE):

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{NJ} \sum_{n=1}^{N} \mathbf{y}_n^T \ln(\mathbf{p}_n) + (\mathbf{1} - \mathbf{y}_n)^T \ln(\mathbf{1} - \mathbf{p}_n), \tag{4}$$

where $\mathbf{p}_n = \begin{bmatrix} p_{n,1} & p_{n,2} & \cdots & p_{n,J} \end{bmatrix}^T$ consists of the posterior probability of different events presented in the nth signal, which is defined by (2), and $\mathbf{1}$ is a vector of one's.

III. Multi-label weighted binary cross entropy for sound event classification

Due to the number of negative samples ($y_{n,j} = 0$) being significantly larger than that of positive samples ($y_{n,j} = 1$) in the training set, the dominance of the second term in (4) may lead to poor performance. To tackle this challenging issue, this section discusses loss functions that can balance the distribution of positive and negative samples by minimizing category-insensitive partial AUC. Specifically, in Section III-A, we first introduce a weighted BCE loss based on partial AUC optimization for the general binary classification problem. Then, in Sections III-B and III-C, two different loss functions are derived for multi-label classification tasks.

A. Weighted BCE Based on Partial AUC Optimization

Let us first consider the special case of $J = 1$, i.e., we have only one category to detect. Since the dimensions of $\mathbf{y}_n$ and $\mathbf{s}_n$ are both equal to one, we write them as y_n and s_n, respectively, in this subsection. For convenience, we define two index sets of positive and negative samples, $\mathcal{I}^+ = \{n | y_n = 1\}$ and $\mathcal{I}^- = \{n | y_n = 0\}$, respectively, and further define the number of samples in each sets as N^+ and N^-. By definition, we have $N = N^+ + N^-$. Given a threshold η, we can calculate a false positive rate (FPR) of the prediction as

$$f(\eta) = \frac{1}{N^-} \sum_{n \in \mathcal{I}^-} \mathbb{I}(s_n \geq \eta) \tag{5}$$

and a true positive rate (TPR) as

$$t(\eta) = \frac{1}{N^+} \sum_{n \in \mathcal{I}^+} \mathbb{I}(s_n \geq \eta). \tag{6}$$

Then, a ROC curve can be obtained by changing the value of η, where the x-axis represents $f(\eta)$ and the y-axis represents $t(\eta)$. Calculating the partial area under the ROC curve for the range between α and β, we can obtain the partial AUC (pAUC), i.e.,

$$\text{pAUC} = \int_{\alpha}^{\beta} t(\eta) df(\eta). \tag{7}$$

Empirically, this integral form of pAUC can be estimated according to

$$\text{pAUC} \approx \frac{1}{N^+ N^-_{\text{sub}}} \sum_{n \in \mathcal{I}^+} \sum_{m \in \mathcal{I}^-_{\text{sub}}} \left[\mathbb{I}(s_m < s_n) + \frac{1}{2} \mathbb{I}(s_m = s_n) \right], \tag{8}$$

where $\mathcal{I}^-_{\text{sub}}$ is a subset of the index set $\mathcal{I}^-$, and N^-_{sub} denotes the number of samples in $\mathcal{I}^-_{\text{sub}}$. It is calculated by first sorting all logits associated to $\mathcal{I}^-$ in descending order, and then choosing all indices of logits that located between positions $\lfloor \alpha \times N^- \rfloor + 1$ and $\lceil \beta \times N^- \rceil$.

According to [13], the maximization of (8) can be relaxed to minimization of the following weighted binary cross-entropy loss function:

$$\begin{aligned} \mathcal{L}_{\text{pAUC}} = &-\frac{1}{N^+} \sum_{n \in \mathcal{I}^+} \varpi_n \cdot \ln[\varepsilon(s_n)] \\ &-\frac{1}{N^-_{\text{sub}}} \sum_{m \in \mathcal{I}^-_{\text{sub}}} \varpi'_m \cdot \ln[\zeta(s_m)], \end{aligned} \tag{9}$$

where

$$\varepsilon(s_n) = \frac{1}{1 + e^{-s_n + \gamma\delta}}, \tag{10}$$

$$\zeta(s_m) = 1 - \frac{1}{1 + e^{-s_m - (1-\gamma)\delta}}, \tag{11}$$

$$\varpi_n = \frac{1}{N^-_{\text{sub}}} \sum_{m \in \mathcal{I}^-_{\text{sub}}} \mathbb{I}(s_m > s_n - \delta), \tag{12}$$

$$\varpi'_m = \frac{1}{N^+} \sum_{n \in \mathcal{I}^+} \mathbb{I}(s_n < s_m + \delta), \tag{13}$$

with γ being a hyperparameter for balancing the strength of the margin added to positive and negative samples[2], δ being a margin hyparameter introduced in the relaxation process, the positive weight ϖ_n representing the FPR with threshold $s_n - \delta$ for set $\mathcal{I}^-_{\text{sub}}$, and the negative weight ϖ'_m representing the complement of the true positive rate ($1 - \text{TPR}$) with threshold $s_m + \delta$ for set $\mathcal{I}^+$. For simplicity, this paper sets γ to 0.5.

[2]Unlike [13], which only adds the margin to the positive samples, this paper introduces a margin to both the positive and negative samples.

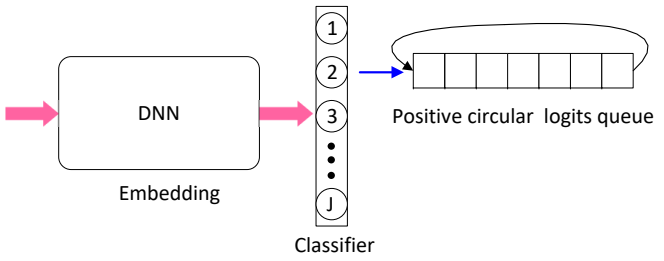


Fig. 1. Illustration of the sound event representation leaning and classification based on $\mathcal{L}_{\text{mpAUC}}$, where the queue is used for cyclically storing positve logits of corresponding sound events of previous iterations, and the computational graphs of these logits are not retained.

B. Sound Event Classification Based on Mean pAUC

As a multi-label sound event classifier can be divided into J binary classifiers, a natural definition of multi-label pAUC is to calculate the mean of the binary pAUC for all the events, i.e.,

$$\text{mpAUC} = \frac{1}{J}\sum_{j=1}^{J}\int_{\alpha_j}^{\beta_j} t_j(\eta)df_j(\eta). \tag{14}$$

where α_j and β_j are hyperparameters of the j-th event. Therefore, the loss function to optimize the multi-label pAUC can be obtained by averaging (9) for all the events.

Traditionally, the logits of a DNN are obtained by computing the inner product between the representations and weights, as shown in (1). Inspired from the results in speaker recognition [13] and face recognition [14], we perform in this work representation learning on a hypersphere with a fixed radius to obtain more effective representation backbone $\hbar_\theta(\cdot)$. Specifically, the logit of a given embedding $\mathbf{z}_n$ in terms of the j-th event is computed as

$$\begin{aligned}\widetilde{s}_{n,j} =& a \times \frac{\mathbf{w}_j^T\mathbf{z}_n}{\|\mathbf{w}_j\|\|\mathbf{z}_n\|} + b_j \\ =& a\cos\vartheta_{n,j} + b_j,\end{aligned} \tag{15}$$

where a is a tunable radius hyperparameter, b_j is a learnable bias corresponding to the jth event, $\vartheta_{n,j}$ is the angle between $\mathbf{z}_n$ and $\mathbf{w}_j$, and $\|\cdot\|$ is the Euclidean norm.

Given a mini-batch training data $\mathcal{B} = \{(\mathbf{x}_n, \mathbf{y}_n)|n = 1, 2, \ldots, N\}$, we have the positive and negative index sets corresponding to the jth event as $\mathcal{I}_j^+ = \{n|y_{n,j} = 1\}$ and $\mathcal{I}_j^- = \{n|y_{n,j} = 0\}$. For convenience, we denote the number of samples in each sets as N_j^+ and N_j^-, and further define a subset of the $\mathcal{I}_j^-$ as $\mathcal{I}_{\text{sub},j}^-$ with $N_{\text{sub},j}^-$ being the number of samples.

In practical situations, the number of sound events J could be comparable to or even larger than the batch size N. In this case, number of the positive samples in each batch is often very small, or even zero, so the approximation in (8) fails. To deal with this issue, as shown in Fig. 1, we first assign a circular logits queue to each kind of sound event to store positive logits from previous iteration. Those stored logits are re-indexed as $\mathcal{I}_{\text{circ},j}^+ = \{N+1, N+2, \cdots, N+L\}$, where L is the length of the queue. When computing the negative weights, we use $\mathcal{I}_j^+$ plus $\mathcal{I}_{\text{circ},j}^+$, i.e.,

$$\hat{\mathcal{I}}_j^+ \leftarrow \mathcal{I}_j^+ \cup \mathcal{I}_{\text{circ},j}^+, \tag{16}$$

to guarantee the number of positive samples in each mini-batch. Taking a subset of negative samples from $\mathcal{I}_j^-$, we form $\mathcal{I}_{\text{sub},j}^-$ for each sound event. The loss function used for optimizing mpAUC can then be expressed as

$$\begin{aligned}\mathcal{L}_{\text{mpAUC}} = -\frac{1}{J}\sum_{j=1}^{J}\Bigg\{&\frac{1}{N_j^+}\sum_{n\in\mathcal{I}_j^+}\varpi_{n,j}\cdot\ln\left[\varepsilon(\widetilde{s}_{n,j})\right] \\ &+\frac{1}{N_{\text{sub},j}^-}\sum_{m\in\mathcal{I}_{\text{sub},j}^-}\varpi'_{m,j}\cdot\ln\left[\zeta(\widetilde{s}_{m,j})\right]\Bigg\},\end{aligned} \tag{17}$$

where

$$\varpi_{n,j} = \frac{1}{N_{\text{sub},j}^-}\sum_{m\in\mathcal{I}_{\text{sub},j}^-}\mathbb{I}(\widetilde{s}_{m,j} > \widetilde{s}_{n,j} - \delta), \tag{18}$$

$$\varpi'_{m,j} = \frac{1}{N_j^+ + L}\sum_{n\in\hat{\mathcal{I}}_j^+}\mathbb{I}(\widetilde{s}_{n,j} < \widetilde{s}_{m,j} + \delta), \tag{19}$$

and the first term within the braces will disappear if $N_j^+ = 0$. Recall that $\vartheta_{n,j}$ is a function of $\mathbf{w}_j$ and $\mathbf{z}_n$, and the representation $\mathbf{z}_n$ is calculated according to $\mathbf{z}_n = \hbar_\theta(\mathbf{x}_n)$. Now, one can use the back-propagation method to learn a proper $\hbar_\theta(\cdot)$, $\mathbf{W}$ and $\mathbf{b}$ by minimizing the loss function $\mathcal{L}_{\text{mpAUC}}$.

C. Sound Event Classification Based on All pAUC

Another way for computing the pAUC of a multi-label sound event classifier is by aggregating all the categories together. We denote this type of multi-label pAUC as opAUC, which is defined as

$$\text{opAUC} = \int_{\alpha}^{\beta}\breve{t}(\eta)d\breve{f}(\eta), \tag{20}$$

where $\breve{t}(\eta)$ and $\breve{f}(\eta)$ are the overall TPR and FPR by considering all the J events. Like what in computing pAUC and mpAUC, we need the positive and negative subsets to compute opAUC. By aggregating the subsets of all the categories, we have $\breve{\mathcal{I}}^+ = \{(n,j)|y_{n,j} = 1\}$ and $\breve{\mathcal{I}}^- = \{(n,j)|y_{n,j} = 0\}$, and further denote the subset of $\breve{\mathcal{I}}^-$ as $\breve{\mathcal{I}}_{\text{sub}}^-$. The number of samples in the three sets are denoted as $\breve{N}^+$, $\breve{N}^-$, and $\breve{N}_{\text{sub}}^-$, respectively. With the introduced variables, we can write the loss function to maximize opAUC as

$$\begin{aligned}\mathcal{L}_{\text{opAUC}} = &-\frac{1}{\breve{N}^+}\sum_{(n,i)\in\breve{\mathcal{I}}^+}\breve{\varpi}_{n,i}\cdot\ln\left[\varepsilon(\widetilde{s}_{n,i})\right] \\ &-\frac{1}{\breve{N}_{\text{sub}}^-}\sum_{(m,j)\in\breve{\mathcal{I}}_{\text{sub}}^-}\breve{\varpi}'_{m,j}\cdot\ln\left[\zeta(\widetilde{s}_{m,j})\right],\end{aligned} \tag{21}$$

TABLE I
CNN STRUCTURE FOR SOUND EVENT REPRESENTATION LEARNING AND CLASSIFICATION [6], WHERE N AND T DENOTE BATCH-SIZE AND NUMBER OF FRAMES, RESPECTIVELY. THE THIRD DIMENSION OF THE OUTPUT SIZE IS AN APPROXIMATION VALUE.

Name	Structure	Output size
Input	—	N clips
Acoustic feature	Log-mel spectrogram (64 mel bins)	$N \times 1 \times T \times 64$
	BN	$N \times 1 \times T \times 64$
Backbone	$\begin{pmatrix} 3\times 3 & \text{Chan} = 64 \\ \text{BN} & \text{ReLU} \end{pmatrix} \times 2$, Stride=(1,1)	$N \times 64 \times T \times 64$
	Average Pooling 2×2	$N \times 64 \times \frac{T}{2} \times 32$
	$\begin{pmatrix} 3\times 3 & \text{Chan} = 128 \\ \text{BN} & \text{ReLU} \end{pmatrix} \times 2$, Stride=(1,1)	$N \times 128 \times \frac{T}{2} \times 32$
	Average Pooling 2×2	$N \times 128 \times \frac{T}{4} \times 16$
	$\begin{pmatrix} 3\times 3 & \text{Chan} = 256 \\ \text{BN} & \text{ReLU} \end{pmatrix} \times 2$, Stride=(1,1)	$N \times 256 \times \frac{T}{4} \times 16$
	Average Pooling 2×2	$N \times 256 \times \frac{T}{8} \times 8$
	$\begin{pmatrix} 3\times 3 & \text{Chan} = 512 \\ \text{BN} & \text{ReLU} \end{pmatrix} \times 2$, Stride=(1,1)	$N \times 512 \times \frac{T}{8} \times 8$
	Global Pooling 2×2	$N \times 512$
Dense layer	512×512 FC, **ReLU** (With / Without)	$N \times 512$
Classifier	$512 \times J$ Weight matrix and $J \times 1$ Bias vector	$N \times J$

where

$$\breve{\varpi}_{n,i} = \frac{1}{\breve{N}^{-}_{\text{sub}}} \sum_{(m,j)\in\breve{\mathcal{I}}^{-}_{\text{sub}}} \mathbb{I}(\widetilde{s}_{m,j} > \widetilde{s}_{n,i} - \delta), \tag{22}$$

$$\breve{\varpi}'_{m,j} = \frac{1}{\breve{N}^{+}} \sum_{(n,i)\in\breve{\mathcal{I}}^{+}} \mathbb{I}(\widetilde{s}_{n,i} < \widetilde{s}_{m,j} + \delta). \tag{23}$$

In practical situations, the process of computing opAUC is the same as that of computing the classical pAUC once we have the subsets of positive and negative samples, i.e., $\breve{\mathcal{I}}^{+}$ and $\breve{\mathcal{I}}^{-}$.

IV. EXPERIMENTS

A. Experimental Setup

We validate the effectiveness of the proposed loss functions for sound event representation learning and few-shot classification using three datasets: AudioSet [2], ESC [15], and FSD50K [3], where the first dataset is used to train the representation backbone, and the last two are used for the few-shot classification experiment. AudioSet consists of 527 sound classes, with the audio clips taken from YouTube videos. It is divided into three parts, i.e., "unbalanced train," "balanced train," and "evaluation," and we get, respectively, 1934187, 20550, and 18887 audio clips [6]. The duration of every clip is 10 seconds. It should be noted that all these three parts suffer from class imbalance. ESC consists of 2000 audio clips of 50 classes with 40 audio clips for each class. Each clip contains a single event and has a duration of 5 seconds. FSD50K is an open dataset that consist of over 51000 manually labeled audio clips, using 200 classes derived from the AudioSet Ontology.

In our experiments, all audio clips are resampled to a 32-kHz sampling rate. Following the work in [6], we use a DNN architecture as described in Table I. We employ a 64-dimensional log mel spectrogram as the input of the DNN. A batch size of 128 and the Adam optimizer with an initial learning rate of $1e^{-3}$ are used for training. During the representation learning phase, 100000 iterations were performed, and the learning rate is decreased to $1e^{-4}$ and $1e^{-5}$ at the 80000th and 90000th iteration, respectively. For the few-shot classification on ESC, we train 200 epochs, and the learning rate is decreased to $1e^{-4}$ and $1e^{-5}$ at the 100th and 150th epoch, respectively. On FSD50K, we train 150 epochs (for few shot classification) or 60 epochs (using all training data), and learning rate is decreased to $1e^{-4}$ and $1e^{-5}$ at the last 20th and 10th epochs from the end.

The loss function $\mathcal{L}_{\text{mpAUC}}$ in (17) involves six hyperparameters, i.e., the margin δ, the margin balance γ, the buffer length L, the radius a, and the two pAUC constraints α_j and β_j. For simplicity, we empirically set the value of γ to 0.5, the value of α to 0, and the size of the buffer $\mathcal{I}^{+}_{\text{circ},j}$ to 64 and 10 for, respectively, the representation leaning and few-shot fine-tuning experiments. We initialize β_j to 1, and decrease it according to

$$\beta_j^{(k)} = \max(\min(1 - \text{pAUC}_j, \beta_j^{(k-1)}), \beta_{\min}) \tag{24}$$

for every 500 (for representation learning) or 10 (for fine-tuning) iterations, where k is the index of the training iteration, pAUC_j is an exponentially smoothed value with momentum 0.99 (for representation learning) or 0.9 (for fine-tuning), and $\beta_{\min}$ is a tunable hyperparameter. Finally, we have three tunable hyperparameters and perform a grid search based on the following values: $a \in \{10, 20, 30\}$, $\frac{\delta}{a} \in \{0.2, 0.4\}$ and $\beta_{\min} \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. The loss function $\mathcal{L}_{\text{opAUC}}$ in (21) follows the same process as setting the hyperparameters.

The mean average precision (mAP) and mean area under the ROC curve (mAUC) are used as evaluation metrics for sound event classification [6]. Moreover, the equal error rate (EER) [13] is used to assess the quality of the learned representations.

B. Representation Learning Over AudioSet

To evaluate the quality of the learned representation backbone, we combine the 2000 audio clips of ESC in a pairwise manner, resulting in a total number of 1999000 trials. We then calculate the cosine similarity of their representations and evaluate it using EER. The results of representation learning on AudioSet are presented in Table II. Note that the results in Table II are the averaged results over iterations $\{92000, 94000, 96000, 98000, 100000\}$.

From Table II, one can see that when removing the ReLU activation function after the Dense layer, the proposed loss function yields a significantly lower EER in comparison with the classical BCE, with a relative reduction of more than 20%. It also demonstrates certain advantages in performance on the AudioSet evaluation dataset. Furthermore, comparing the results in the 5th and 6th rows, one can see that $\mathcal{L}_{\text{opAUC}}$ leads to a lower EER though both $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{opAUC}}$ have similar values of mAP on AudioSet-Eval.

TABLE II
REPRESENTATION LEARNING EXPERIMENT ON THE AUDIOSET.

Dense layer ReLU	Augmentation	Loss	Balance sampling	AudioSet-Eval mAP [%]	AudioSet-Eval mAUC [%]	ESC EER [%]
Without	Without	$\mathcal{L}_{\text{BCE}}$	With	36.17	96.53	9.00
		$\mathcal{L}_{\text{mpAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.1)$	With	36.97	96.70	7.74
		$\mathcal{L}_{\text{opAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.1)$	With	**38.19**	97.11	7.08
		$\mathcal{L}_{\text{opAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.4)$	With	35.41	**97.18**	**6.95**
	SpecAug	$\mathcal{L}_{\text{BCE}}$	With	**35.93**	96.58	8.86
		$\mathcal{L}_{\text{opAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.4)$	With	35.56	**97.21**	**6.65**
With	SpecAug	$\mathcal{L}_{\text{BCE}}$	With	**39.74**	97.26	6.99
		$\mathcal{L}_{\text{opAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.1)$	With	38.33	96.77	**6.40**
		$\mathcal{L}_{\text{opAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.4)$	With	35.68	97.26	7.56
		$\mathcal{L}_{\text{BCE}}$	Without	29.23	**95.80**	9.70
		$\mathcal{L}_{\text{opAUC}}$ $(a=10, \frac{a}{\delta}=0.2, \beta_{\min}=0.1)$	Without	**30.51**	95.71	**8.71**

TABLE III
GIRD SEARCH OF HYPERPARAMETERS FOR WEIGHTED BCE BASED ON THE FEW-SHOT SOUND EVENT CLASSIFICATION TASK OF ESC [%]. THE RELU ACTIVATION FUNCTION AFTER THE DENSE LAYER IS REMOVED.

	$(a, \frac{\delta}{a})$	mAP	mAUC	mAP	mAUC	mAP	mAUC	mAP	mAUC
		$\beta_{\min}=0.1$		$\beta_{\min}=0.2$		$\beta_{\min}=0.4$		$\beta_{\min}=0.8$	
$\mathcal{L}_{\text{mpAUC}}$	(10, 0.2)	89.48	99.00	88.74	**99.08**	87.58	99.03	85.72	98.99
	(10, 0.4)	**89.68**	98.49	89.09	98.30	—	—	—	—
	(10, 0.6)	87.74	97.76	—	—	—	—	—	—
	(20, 0.2)	89.02	99.07	—	—	—	—	—	—
	(30, 0.2)	88.60	98.94	—	—	—	—	—	—
		$\beta_{\min}=0.01$		$\beta_{\min}=0.1$		$\beta_{\min}=0.2$		$\beta_{\min}=0.4$	
$\mathcal{L}_{\text{opAUC}}$	(10, 0.2)	**90.84**	98.89	90.63	**99.06**	88.96	99.02	87.44	98.97
	(10, 0.4)	—	—	90.05	98.39	—	—	—	—
	(10, 0.6)	—	—	89.34	97.72	—	—	—	—

With ReLU after the Dense layer, we observe a significant improvement in the performance of BCE, while the proposed loss function shows only a modest improvement. In spite of this, the $\mathcal{L}_{\text{opAUC}}$ still maintains a lower EER. Moreover, if balanced sampling on the training data is not performed, one can see that $\mathcal{L}_{\text{opAUC}}$ achieves a relatively significant performance improvement in comparison with $\mathcal{L}_{\text{BCE}}$, indicating that BCE is more sensitive to the imbalance problem.

Overall, the proposed representation learning loss functions demonstrate significant advantages. Further validation is in progress for other network structures such as those performing better without adding ReLU after the Dense layer.

C. Few-Shot Classification Over ESC and FSD50K

This subsection validates the effectiveness of the proposed loss functions in few-shot classification tasks. Specifically, on the ESC dataset, we randomly chose 10 or 20 clips of each class of sound events as training data, with the rest being used as testing data. For simplicity, we refer to this as 10-shot and 20-shot, respectively. Table III presents the grid search results of hyperparameters on ESC, where the network is trained from scratch under 20-shot condition. From the table, it is observed that the $\beta_{\min}$ parameter has a significant impact on the performance, and the optimal values for a and δ/a are 10 and 0.2, respectively. The $\mathcal{L}_{\text{opAUC}}$ loss yields a better performance than $\mathcal{L}_{\text{mpAUC}}$.

Table IV presents the comparison between the performance of the proposed loss functions and that of the traditional BCE for few-shot classification. The experiment consists of three situations: training from scratch, only fine-tuning the classifier, and fine-tuning the entire network. During fine tuning, the network is initialized by parameters pretrained on AudioSet with the corresponding loss functions. The following two observations can be made from Table IV. 1) $\mathcal{L}_{\text{mpAUC}}$ outperforms $\mathcal{L}_{\text{BCE}}$ in terms of mAP, especially with limited training data like 10-shot. In comparison with $\mathcal{L}_{\text{BCE}}$, fine-tuning the classification head with $\mathcal{L}_{\text{mpAUC}}$ shows a relative mAP improvement of over 40%. 2) A well-initialized network leads to better performance, indicating that pretraining enhances significantly few-shot sound event classification on this dataset.

Due to redundancy among the 200 event classes in FSD50K, we remove all the intermediate nodes to ensure mutual exclusivity between events and retain only the leaf node events, resulting in a total of 135 event classes. We discard audio samples from the training, validation, and test sets that do not include the selected events, and the remaining samples constitute the final training, validation, and test sets. Moreover, during training, we vary the number of training clips per event to observe performance changes and the results as shown in Table V.

From Table V, it is seen that $\mathcal{L}_{\text{opAUC}}$ yields a great improvement in mAP as compared to $\mathcal{L}_{\text{BCE}}$, the absolute improvement in mAP ranges from 3% to 6% in the studied few-shot cases.

TABLE IV
PERFORMANCE [%] OF FEW-SHOT SOUND EVENT CLASSIFICATION ON ESC.THE HYPERPARAMETERS $(a, \frac{\delta}{a}, \beta_{\min})$ ARE SET TO $(10, 0.2, 0.1)$. THE RELU ACTIVATION FUNCTION AFTER THE DENSE LAYER IS ADDED.

		Scratch		Tuning classifier		Tuning all	
	Loss	mAP	mAUC	mAP	mAUC	mAP	mAUC
20-shot	$\mathcal{L}_{\text{BCE}}$	89.38	99.13	90.51	99.62	95.43	99.63
	$\mathcal{L}_{\text{mpAUC}}$	**91.19**	**99.26**	94.01	99.69	96.52	99.44
	$\mathcal{L}_{\text{opAUC}}$	90.69	99.06	**94.31**	**99.69**	**96.98**	**99.68**
10-shot	$\mathcal{L}_{\text{BCE}}$	72.37	97.47	84.73	99.18	88.86	99.16
	$\mathcal{L}_{\text{mpAUC}}$	78.10	97.65	**91.23**	99.33	93.20	99.29
	$\mathcal{L}_{\text{opAUC}}$	**78.97**	**97.73**	90.75	**99.38**	**93.92**	**99.32**

TABLE V
MAP [%] OF FEW-SHOT SOUND EVENT CLASSIFICATION ON FSD50K. THE HYPERPARAMETERS $(a, \frac{\delta}{a}, \beta_{\min})$ ARE SET TO $(10, 0.2, 0.1)$. THE RELU ACTIVATION FUNCTION AFTER THE DENSE LAYER IS ADDED.

		Loss	5-shot	10-shot	20-shot	30-shot	All
Scratch	Test	$\mathcal{L}_{\text{BCE}}$	15.82	22.30	32.96	39.36	53.74
		$\mathcal{L}_{\text{opAUC}}$	**18.53**	**28.60**	**37.87**	**42.39**	**54.26**
	Valid	$\mathcal{L}_{\text{BCE}}$	16.44	23.97	33.97	41.19	55.74
		$\mathcal{L}_{\text{opAUC}}$	**19.26**	**29.77**	**39.36**	**44.26**	**56.60**
Tuning all	Test	$\mathcal{L}_{\text{BCE}}$	14.14	23.58	33.63	40.65	53.54
		$\mathcal{L}_{\text{opAUC}}$	**18.19**	**27.23**	**37.72**	**41.68**	**54.11**
	Valid	$\mathcal{L}_{\text{BCE}}$	14.97	25.13	35.02	41.36	56.34
		$\mathcal{L}_{\text{opAUC}}$	**18.78**	**29.19**	**38.89**	**43.66**	**56.42**

It is also seen from the FSD50K dataset that pre-training is helpful for BCE but not for $\mathcal{L}_{\text{opAUC}}$. The reason for this is under investigation. Another interesting observation is that the mAUC of $\mathcal{L}_{\text{opAUC}}$ is not as good as $\mathcal{L}_{\text{BCE}}$. For example, in the 10-shot scratch training scenario, the mAUC of $\mathcal{L}_{\text{BCE}}$ on the test set is 91.52%, while the mAUC of $\mathcal{L}_{\text{opAUC}}$ is 89.64%. This may be due to the decreased value of β.

V. SUMMARY

In this paper, we proposed two weighted binary cross entropy loss functions based on pAUC optimization for sound event representation learning and few-shot classification. Experimental results showed that the DNN trained with the proposed loss functions improves significantly both the representation and few-shot classification performance. Note that we found that the network structure, such as the ReLU activation function before the classification head, affects significantly the relative performance improvement of the proposed weighted BCEs as compared to the traditional BCE. The reason for this is not clear and currently under investigation. Also, in few-shot classification, the performance of the proposed loss functions in terms of mAUC is not as good as BCE. This is related to the hyperparameter $\beta_{\min}$, which will also be studied in the near future.

VI. ACKNOWLEDGEMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0201502, in part by the NSFCs under Grants 62171373 and 61831019, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102021HHZY030001.

REFERENCES

[1] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[2] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[3] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[4] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon *et al.*, "Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context," *arXiv preprint arXiv:2009.05188*, 2020.

[5] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[6] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNS: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[7] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1038–1047.

[8] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[9] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[10] N. Shao, E. Loweimi, and X. Li, "RCT: Random consistency training for semi-supervised sound event detection," *arXiv preprint arXiv:2110.11144*, 2021.

[11] J. Yan, Y. Song, L.-R. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 326–330.

[12] L. Wang and A. v. d. Oord, "Multi-format contrastive learning of audio representations," *arXiv preprint arXiv:2103.06508*, 2021.

[13] Z. Bai, J. Wang, X.-L. Zhang, and J. Chen, "End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1330–1344, 2022.

[14] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1041–1049.

[15] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

A Reconfigurable Hardware Architecture for Graph Convolution Network in Action Recognition

Tsung-Han Tsai
Dept. of Electrical Engineering
National Central University
Taoyuan, Taiwan
han@ee.ncu.edu.tw

Tzu-Chieh Chen
Dept. of Electrical Engineering
National Central University
Taoyuan, Taiwan
zillion@dsp.ee.ncu.edu.tw

***Abstract*—Graph convolution neural network (GCN) has gained more attention in recent years. Data types like human skeleton joints which are in non-Euclidean space are suitable for GCN. However, due to the high computational complexity and data sparsity of GCN, it is common to have high latency or low power efficiency in CPU or GPU platforms. Therefore, dedicated hardware accelerators are critical for such tasks. In this paper, we proposed a highly parallelized and flexible architecture for the GCN layers of the Spatial-Temporal Graph Convolutional Networks (ST-GCN) model, which is a classic model and widely used in human action recognition. The accelerator also has high scalability due to our proposed method. Compared with the hardware implementation on ST-GCN, the proposed method reduces the latency by up to 39.5% and improves 1.46x on power efficiency.**

Keywords—graph convolutional neural network, hardware, accelerator, reconfigurable architecture, FPGA

I. INTRODUCTION

Human action recognition (HAR) has been used in different kinds of applications, such as medical health care [1], surveillance system [2], and VR gaming [3]. In the early stages, RGB images are the most common data type in HAR. Lots of deep learning methods have been proposed, such as 2D Convolutional Neural Network (CNN) [4], 3D CNN [5], and Recurrent Neural Network (RNN) [6]. However, models which use plain images as their input usually have limited performance due to variation backgrounds, illumination differences, and will be affected by irrelevant objects.

Skeleton data had gained more interest since it only focuses on human actions and is independent of background noises. The skeleton data can be obtained by post estimation algorithms on RGB videos or by depth camera. Even though there are still methods focusing on using CNN or RNN to extract features from skeleton data, it is clear that the data type between RGB image and skeleton data is quite different, while RGB image is categorized as Euclidean data, skeleton data is non-Euclidean space. For Euclidean data which is neatly aligned, it is easy and suitable to apply convolution operation. However, the model will lose the spatiality of the skeleton data if we apply the same convolution operation.

Therefore, Ref [7] proposed a GCN method which is called ST-GCN to handle the irregularity of skeleton data. Fig. 1 shows the graph structure of the skeleton sequence for ST-GCN. The blue dots represent the natural body joints and the lines between the dots within a frame are connected according to the human body structure. By applying GCN to the graph structure, the model can obtain spatial information about the human skeleton structure. ST-GCN also takes temporal information into account by connecting joints between different frames, and the results show that it can outperform other works which use CNN or RNN methods.

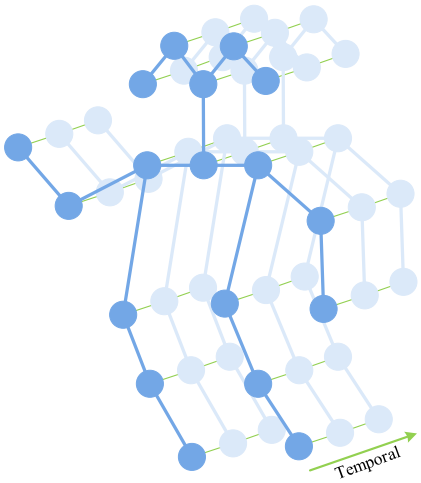


Fig. 1. The graph structure used in ST-GCN

Despite the success of GCN, the model still faces problems like large network architectures and is highly computationally intensive which takes a long time to inferencing on CPU or GPU platforms. Several papers have been presented to accelerate GCN on FPGA platforms. In [8], it proposed a fixed structure that concatenates the convolution engine and sparse matrix-vector multiplication (SpMxV) engine. However, with the design of using three parallel SpMxV engines based on the ST-GCN model which uses three adjacency matrices, it can cause more latency if the skeleton data contains more joints. STARS [9] proposed an architecture using high-level synthesis (HLS) to optimize the resources used on the FPGA. Nevertheless, different types of computation use different computational units and are independent of one another. This causes low resource utilization because when a computing kernel is operating, others remain inactive due to most parts of the ST-GCN process layer by layer. In this paper, we propose a reconfigurable hardware architecture to process the GCN layer. The main contributions in this paper are as follows:

- We design an architecture that can allocate various amounts of PEs for the aggregation engine to fit different kinds of skeleton structures with no delay increase.
- We design a general PE for all kinds of operations used in the GCN layer and also form a pipeline

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

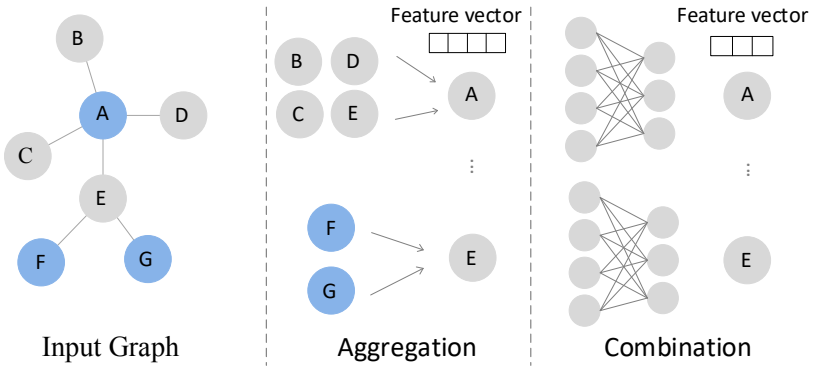


Fig. 2. GCN computation.

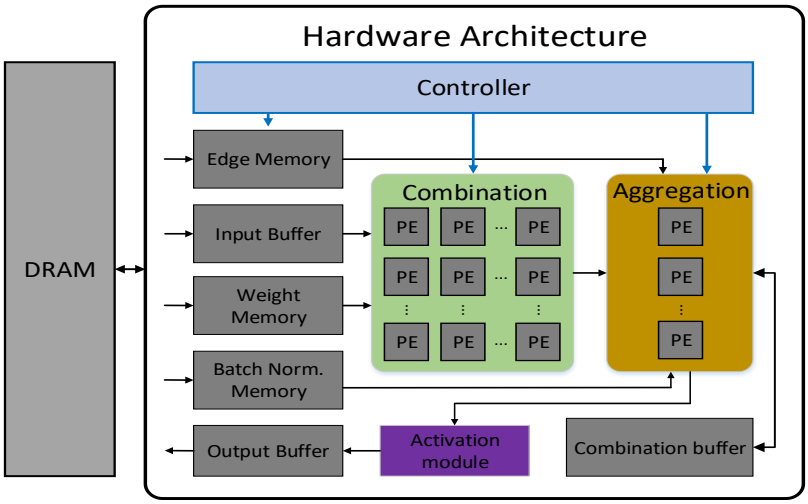


Fig. 3. Overall hardware architecture.

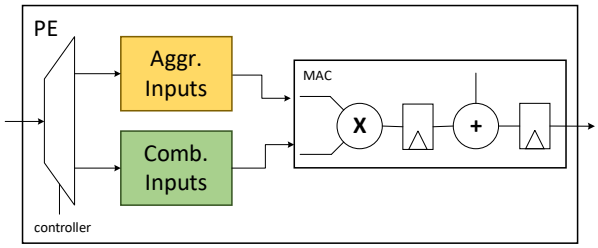


Fig. 4. PE architecture.

structure with the combination engine and aggregation engine to further improve hardware utilization.

- We implement our design on the FPGA platform and achieve better performance compared to the state-of-the-art.

The rest of the paper is organized as follows. Section II briefly explains how the GCN layer computes. Section III shows our overall architecture while section IV presents the experiment result and compares it with state-of-the-art papers. Finally, the paper is concluded in section V.

II. GCN Computation

ST-GCN is a GCN model which aims to exploit the spatiality and temporality of the skeleton joints. It constructs an undirected spatial-temporal graph $G = \{V, E\}$ on a skeleton sequence with N joints and T frames featuring both intra-frame and inter-frame connections. The node set $V = \{v_{ti} | t = 1, \dots, T, i = 1, \dots, N\}$ represent all skeleton joints contained in the action sequence. The edge set is composed of two parts: $E_s = \{v_{ti}v_{tj} | (i,j) \in H\}$ indicates the intra-skeleton connection in each frame, where H is the set of naturally connected human body joints, while $E_F = \{v_{ti}v_{(t+1)i}\}$ represents the inter-frame edges which are simply connected between two same joints in consecutive frames. The model can also support a different number of skeleton joints. For example, the Kinetics dataset [10] uses 18 joints as input, while the NTU-RGB+D dataset [11] contains 25 skeleton joints.

When applying spatial convolution to the input data, the formula can be expressed as follows:

$$f_{out} = \Lambda^{-\frac{1}{2}}(\mathrm{A} + \mathrm{I})\Lambda^{-\frac{1}{2}} f_{in} W \qquad (1)$$

Matrix A is an adjacency matrix that represents intra-body connections and matrix I represents self-connecting nodes to also takes its node information into account. The formula normalized the adjacency matrix by $\Lambda^{-\frac{1}{2}}(\mathrm{A} + \mathrm{I})\Lambda^{-\frac{1}{2}}$ to alleviate numerical instabilities and exploding/vanishing gradients when training a deep neural network model, where $\Lambda = \sum_j(\mathrm{A}_{ij} + \mathrm{I}_{ij})$. The normalized adjacency matrix can be computed in advance since the graph structure remains unchanged. Therefore, the formula can be simplified as $f_{out} = \mathrm{A}' f_{in} W$, which contains two different computations:

TABLE I. RESOURCE UTILIZATION COMPARISION

	ASICON 2019[8]	STARS[9]	OURS
LUT	44457/ 1182240	32995/ 53200	53431/ 230400
FF	34667 /2364480	36663/ 106400	49849/ 460800
BRAM	151/2160	87.5/280	183/312
DSP	228/6840	100/220	288/1728

TABLE II. LAYER BY LAYER PROCESSING TIME COMPARISON

ST-GCN layer latency(ms)	Input/Output Channels	ASICON 2019[8]	OURS
Layer1	3/64	2.61	4.38
Layer2-4	64/64	4.05	4.38
Layer5	64/128	8.03	8.76
Layer6-7	128/128	11.2	8.78
Layer8	128/256	22.3	13.17
Layer9-10	256/256	35.2	13.21

$A'f_{in}$ is a matrix-vector multiplication and $f_{in}W$ is known as a convolutional computation with the kernel size set as (1, 1). Fig. 2 shows the illustration of GCN computation, where the combination process represents convolutional computation and the aggregation process indicates matrix-vector multiplication.

Ref [7] also introduced different partition strategies for skeleton joints. With the traditional uni-labeling strategy, the model can lose the local differential properties due to the operation being equivalent to the inner product between the average feature vector of all neighboring nodes and the weight vector. ST-GCN model uses a spatial configuration partitioning strategy according to the motion of a body can be broadly categorized as concentric and eccentric motions. The strategy can separate the neighboring nodes into three parts: the root node, the centripetal group, and the centrifugal group. It can be expressed as follows:

$$l_{ti}(v_{tj}) = \begin{cases} 0 & if\ r_j = r_i \\ 1 & if\ r_j < r_i \\ 2 & if\ r_j > r_i \end{cases} \qquad (2)$$

Where r_i represents the distance between the skeleton gravity center and the root node. Therefore, group 0 contains all root nodes, group 1 includes all centripetal nodes which have shorter distances to the gravity center while group 2 has longer distances and will be categorized as the centrifugal group.

III. HARDWARE ARCHITECTURE

Fig. 3 shows our overall architecture. The design contains several SRAMs to store weights, edge information, and batch normalization parameters. Our computation unit uses a unified PE design but can be separated into two parts: combination PEs and aggregation PEs. First, the controller will receive the layer information and allocate a different amount of PEs for the combination and aggregation process,

TABLE III. OVERALL PERFORMANCE COMPARISON

	ASICON 2019[8]	STARS[9]	OURS
Quantification	Fixed-16	Fixed-16	Fixed-16
Clock (Hz)	188M	150M	158M
Power (W)	-	1.96	2.378
Latency (ms)	137.89	89.55	83.44
Peak Performance (GOPs)	46	73.4	91
Average Performance (GOPs)	23.505	36.192	64.269
Power Efficiency (GOPs/W)	-	18.47	27.03

which is shown in Fig. 4. Then, the weight data, edge information, and batch normalization parameters are transmitted from off-chip DRAM. Finally, the input feature maps will send into the architecture and compute the results at the same time. Before delivering the output feature maps to the off-chip DRAM, the data will go through the activation module which is set as ReLU in the ST-GCN network.

A. Combination Engine

The combination process takes input feature maps and weights from the input buffer and weight memory to generate temporary combination results. Each PE processes a skeleton joint feature one at a time. In our design, we use weight stationary dataflow due to the weight data in combination process is more likely to fit into the entire on-chip memory, therefore the accelerator will only read input feature map data and weight data once from DRAM, and the accelerator can operate while the input feature map data continuously being read in the design. For some layers of GCN in which its entire weight data cannot fit in the on-chip memory, the controller will split the layer into several processes, while the accelerator will use the maximum amount of PEs that are assigned for combination engine for maximum efficiency. Since the combination process forms a 1 x 1 convolution, the partial sum will always be stored in the local memory of its PE, which is also output stationary. With multiple PEs, it can generate different output channels of combination results.

It is worth noticing that generated output channels must be a multiple of three for aggregation engine to fully use all temporary results. In addition, the combination process will take up most of the processing time of a layer due to large input channels. Hence, the temporary combination result is sent to aggregation PEs to add bias.

B. Aggregation Engine

The aggregation engine can compute three processes, with the aggregation process being the main task, while adding bias and batch normalization process are also computed in its spare time to maximize the hardware efficiency. The aggregation process takes combination results and edges information from the combination buffer and edge memory to generate aggregation results. The aggregation process takes three input channels of combination results to operate matrix multiplication with the edge information due to the design of ST-GCN. To exploit high sparsity of the human skeleton

graph, the system only inputs non-zero values and their corresponding index of the edge information.

Each PE is assigned to process a skeleton joint aggregation. With different kinds of skeleton representations, different amounts of PEs can be assigned for the aggregation process. To minimize the latency of the overall process, we add bias in the aggregation engine. The aggregation engine also processes batch normalization operations due to the long processing time of the combination engine. Finally, the aggregation result is sent to the post process engine to perform the activation function.

C. Cooperation of Combination and Aggregation

When processing GCN layers, it is important to compute combination and aggregation processes at the same time to reduce the latency. First of all, the combination engine continuously generates temporary results for the aggregation engine to add biases. Then, the aggregation engine will send the combination results to the fmap tile for data reorganization. The fmap tile will reallocate the values to the corresponding joint memory for future aggregation computation.

To form an aggregation computation, all joint's feature vector information should be stored in the combination buffer. Due to the long operation time in the combination process, it is important to not have any kinds of data blockage which add even more time to the overall processing latency. However, the aggregation engine takes all joints but only three channels of information when operating the aggregation process, while the combination engine outputs all channels but only one joint of the feature vector. This causes new joint information will overwrite older joint information when the combination engine is processing at full speed. Hence, the combination buffer uses ping-pong RAM to solve the problem and elevate the overall processing speed of the accelerator.

With a simple graph representation of the human skeleton, a PE only needs 8 cycles to process aggregation and batch normalization process. In each combination processing time, the aggregation engine can compute 4 aggregation processes. For NTU-RGB+D and Kinetic-skeleton dataset, all aggregation processes can be done while the combination engine processes at full speed. The accelerator can process different kinds of human skeleton structures by reallocating different numbers of PEs to combination and aggregation engines, therefore, it can maximize PE utilization.

IV. Experiment Result

We implemented the proposed hardware architecture on ZCU104, and Table I shows the hardware utilization on the FPGA board. Although the resource usage of our work is higher than other works, STARS uses HLS to implement their work to optimize the hardware utilization but will potentially cause lower performance and power efficiency.

Table II analyzes the individual performance of each GCN layer. When comparing with ASICON 2019 [8], it shows that our work has a disadvantage when processing layer 1 since the input channel of the layer is only three, which causes the combination engine cannot work at full speed. Our accelerator shows an advantage at higher input channels due to the hardware architecture that can optimize the allocation of the combination PEs and split the layer into several processes properly.

Table III shows the overall comparison with other works which also realize the GCN part of the ST-GCN model for human action recognition. Our accelerator provides 1.65x and 1.07x speedup on latency, which is counted as the sum of the 10 layers of the GCN layer. Although our peak performance also outperforms other works, with limited clock frequency on the FPGA platform, peak performance is nearly proportional to the number of PE being used, which cannot reflect the hardware efficiency. Therefore, average performance is a fairer comparison to PE utilization. With the combination engine PEs continuously operating and aggregation engine performs not only aggregation but also bias and batch normalization operation, our average performance is 2.73x and 1.78x higher than other works. Our accelerator also improved by 1.46x on power efficiency than STARS.

V. Conclusion

GCN has become more popular since it can handle a wider range of data types, and it is proven to have higher accuracy in many tasks. In this paper, we proposed a reconfigurable hardware architecture for GCN to optimize the operations involved. The architecture separates PEs into two parts, each processes combination and aggregation respectively but also cooperates for maximum efficiency. The experiment results show that our accelerator has the best performance out of other hardware works with 1.46x power efficiency and up to 2.73x average performance. In the future, we will further optimize our architecture and complete other operations in the ST-GCN layer.

References

[1] Yin, Jun, et al. "A skeleton-based action recognition system for medical condition detection." 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2019.

[2] Lin, Weiyao, et al. "Human activity recognition for video surveillance." 2008 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2008.

[3] Ma, Zaosheng. "Human action recognition in smart cultural tourism based on fusion techniques of virtual reality and SOM neural network." Computational Intelligence and Neuroscience 2021 (2021).

[4] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems 27 (2014).

[5] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

[6] Zhang, Pengfei, et al. "EleAtt-RNN: Adding attentiveness to neurons in recurrent neural networks." IEEE Transactions on Image Processing 29 (2019): 1061-1073.

[7] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Thirty-second AAAI conference on artificial intelligence. 2018.

[8] Ding, Luchang, Zhize Huang, and Gengsheng Chen. "An FPGA implementation of GCN with sparse adjacency matrix." 2019 IEEE 13th international conference on ASIC (ASICON). IEEE, 2019.

[9] Pei, Songwen, et al. "STARS: Spatial Temporal Graph Convolution Network for Action Recognition System on FPGAs." 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2021.

[10] Kay, Will, et al. "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).

[11] Shahroudy, Amir, et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Automated Carina Detection in Chest X-ray Images Using Non-Overlapping and Cross-Squeeze Convolutional Neural Networks

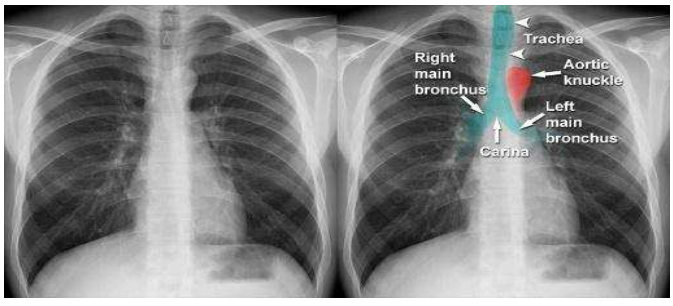
Chung-Chian Hsu[1,2], Chi-Yuan Chen[1], S. M. Salahuddin Morsalin[1,*], Arthur Chang[1] and Wen-Lin Fan[3]
[1]Department of Information Management, National Yunlin University of Science and Technology, Yunlin, 640, Taiwan
[2]International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Yunlin, 640, Taiwan
E-mail: hsucc@yuntech.edu.tw
E-mail: m11023024@yuntech.edu.tw
E-mail: D10813004@yuntech.edu.tw
E-mail: changart @yuntech.edu.tw
[3]Division of Surgical Intensive Care Unit, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Chiayi, 62247, Taiwan
E-mail: chaio@tzuchi.com.tw

***Abstract*— A chest X-ray test is necessary to verify the correct positioning of several medical devices and to check for any issues that may have resulted from improper placement. The trachea, carina, and bronchi localization in chest radiography is a crucial test for patients in the Intensive Care Unit (ICU). Most of the images become affected by the machine or the shooting camera's position, so a physician is required to manually adjust the images to locate Carina. Therefore, this research aim is to automate Carina detection in chest X-ray images without manual adjustment. We proposed a Non-Overlapping Max Pooling (NOMP) connection and Cross-Squeeze Convolutional Attention (CSCA) to make up the design architecture. In addition, a convexity defect technique is combined with the post-processing algorithm to determine Carina's final placement. The experimental results show that the overall performance for average error distance is 0.3097 cm, the accuracy rate is 81.5% when the error falls within 0.5 cm, and the accuracy rate is 97% when the error falls within 1 cm.**

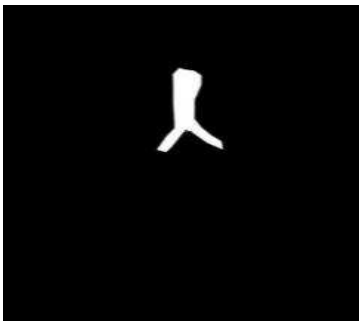
I. INTRODUCTION

Endotracheal intubation is a well-known life-sustaining medical procedure in the Intensive Care Unit (ICU) when patients cannot breathe independently. The appropriate position of the endotracheal tube is 5 ± 2 cm from the bifurcation point, and improper intubation may cause dangerous consequences. For example, Gupta et al. [1] explain in detail such as lung hyperinflation, lung collapse, or damage to the trachea. The method of confirming the correct position of the tube is most commonly performed on a chest X-ray (CXR), usually obtained with a portable X-ray image in the intensive care unit. The compact X-ray images are easily accessible throughout the hospital and also affordable. Chest X-rays proposed by Brunel et al. [2] are the traditional and frequently used method for intubation measurements due to their ability to be scaled on-screen. However, portable X-ray images are easily affected by external factors, resulting in unstable image quality, and it is difficult to judge the position of intubation appropriately. Therefore, in recent years, many researchers have paid more attention to Computer-Aided Diagnosis (CAD) methods, trying to improve the accuracy of tracheal intubation detection, thereby reducing the cost to physicians.

Deep learning has begun to be widely used in medical imaging fields, such as human pose improvement, brain activity, skin lesions, and the chest [3],[4],[5]. In recent years, many studies have used Convolutional Neural Networks (CNN) [7] to evaluate chest X-ray images for intubation. The model is being trained by fusing two pooling layers and adding a connection module for the center node. However, if many modules are included in the architecture, it causes an increase in data preprocessing cost, and parameters. The bifurcation endpoints and Carina prediction in of X-ray image are shown in Figures 1 (a) and (b).



(a)



(b)

Fig. 1. (a) The location of the tracheal bifurcation points in chest X-ray images and (b) the bifurcation points.

We have developed a deep learning algorithm to automatically detect tracheal bifurcation endpoints and Carina prediction on chest X-ray images. The proposed architecture improves the ability to extract model context information and retains useful feature information.

II. Related Works

Kara S. et al. [8] proposed an endotracheal tube identification and localization method that deleted 10% of the upper neck and 40% of the lower abdomen by cropping the Region of Interest (ROI). In addition, the retainer of the middle quarter of the horizontal axis of the X-ray image detection. Finally, adjust the image quality by enhancing contrast and morphology to find the location of the bronchial bifurcation. Jung et al. [9] used a Cascaded Neural Network (CNN) with a loss function and regression properties to find the position of the Endotracheal Tube Tip (ETT) and bronchial bifurcation point (Carina). Roy et al. created the model's detachable tracheal tube and bronchi and cut through the bounding box using concurrent spatial and channel squeeze & and excitation fully convolutional networks to create a two-stage model [10] design. Ultimately, the network is tasked with classifying the clipped data to decide whether the intubation position is adequate. Along with the U2-Net model, add the MA Block and the intermediate node layer. Then set the upper and lower bounds to find the position where the pixel contained for each horizontal axis, at least 0 in the range was the bronchial bifurcation point. Although they can concurrently locate the position of the endotracheal tube tip and the bronchial bifurcation point, the approaches rely on the configuration of hyperparameters to employ a multitasking strategy. The part about the location of the bifurcation point is still imprecise.

A variety of attention techniques have been employed in the past to address problems with picture vision, many of which were influenced by the attention strategy of Natural Language Processing (NLP). They create weights for each feature map value and instruct it to focus on particular areas, such as channel attention, spatial attention, and various attentional variants [11]. To address the issue of low interpretation in the fully connected layer, Hu et al. [12] proposed a Squeeze-and-Excitation Block (SE Block) through Global Average Pooling. As a result, Weighting is only carried out by feature maps, negating the need for parameter optimization and preventing overfitting. Lau et al. [13] used U-Net with a pre-trained ResNet-34 backbone network and added an SC-SE attention module in the decoder to improve the performance and segmentation accuracy of the model. Woo et al. [14] proposed a Convolutional Block Attention Module (CBAM) by combining two pooling techniques to obtain more accurate prediction results than adding an SE Block.

The fully convolutional neural network (Fully Connected Network, FCN) is the ancestor of the U-Net design [15]. It employs the skip connection to combine features and contracts the encoder's (Encoder) image through the contraction component before expanding it through the expansion part. As a result, even with tiny data sets, the U-shaped structure model [16] maintains high-quality feature maps and good performance for denoising image results.

The U-Net++ architecture [17] is a development of the U-Net model that uses several dense short connections in place of the skip connection used in U-Net for long-distance communication and the use of multiple U-Net models for merging. Combining feature maps with various up-sampling options improves the performance of the U-Net++ model. In addition, U-Net++ makes the model architecture more flexible by replacing the encoder with other backbones and can adjust the model depth according to experimental needs to improve the overall execution efficiency.

Qin et al. propose the U2-Net architecture [18] to compare the past module design methods and make a difference to the Feature Pyramid Networks (FPN) and earlier U-Net versions. The collection of residual U-blocks, also known as (RSU blocks), is included in the U2-Net using two layers. The RSU module builds a small U-Net based on the design of the feature pyramid network that can preserve high-resolution feature maps while enhancing the capacity to separate data at various scales through connections. We have adopted U2-Net as the backbone architecture for feature map classification by including the same modules. The proposed module reduces the overall computational cost by ignoring the unnecessary feature maps in segmentation tasks and highlighting the essential features.

III. Proposed Method

A. The Image Preprocessing

It is difficult to train the model with a limited number of datasets, making the model vulnerable to overfitting problems. We have used image data improvement methods to expand the data set. The data enhancements employed include picture random horizontal flip (Horizontal Flip), random zoom rotation (Shift Scale Rotate), and random contrast adaptive histogram equalization (CLAHE). The situation method enhances the simulation of reality. The scaling translation augments the unfixing angle of the chest X-ray images. The rotation and horizontal flip increase the diversity of patient shooting angles. The contrast adjustment simulates the unstable quality of portable chest X-ray images.

B. The NOMP-U2-Net

We have developed the NOMP-U2-Net model, as shown in Figure 2. It mainly consists of four sections: (1) each module uses the RSU-block, (2) the non-overlapping max pooling module, (3) The Cross -Squeeze Convolutional Attention, (4) The feature map fusion stage of each layer of Decoder.

1. The RSU-L module [19], where L is the number of layers of each module, has in the overall architecture. The first through fourth layers of the encoder contain the RSU-4 to RSU-7 module. In contrast, the fifth and sixth layers use the RSU-4F module in the encoder. These techniques utilize the dilation rates in the RSU Dilated convolution operation, which helps the model to get a broader receptive field when confronted with low-resolution feature maps, hence minimizing the loss of feature information.
2. The U2-Net model utilizes the RSU module to gather additional feature data. Deep semantic information is needed for this exertion to segment the entire bronchus. The NOMP convolutions between the encoder and decoder deal with shallow features. Figure 3 shows how the encoder combines after being pooled. The model performs better bronchial semantic segmentation while taking multi-scale input.
3. We have processed Cross-Spatial and Channel Residual (CSCR) attention. This simulation considers the original feature map and weighted feature map without adding additional information to the convolutional attention module. The Cross-Context Attention Mechanism (CCAM) enhances the model's capacity and gathers context information [20] to distinguish the foreground and background. Finally, the outcomes of cross-space and cross-channel residual attention are fed into the CCAM to generate Cross-Squeeze Convolutional Attention (CSCA), as shown in Figure 4.
4. Non-overlapping maximum pooling fuses the encoder's feature maps, resulting in the decoder getting more feature maps as input. The RSU module of each decoder layer will output the 64 feature maps. After integrating the decoders from each layer, the signal will pass through a convolutional layer and a single-row Sigmoid activation function.

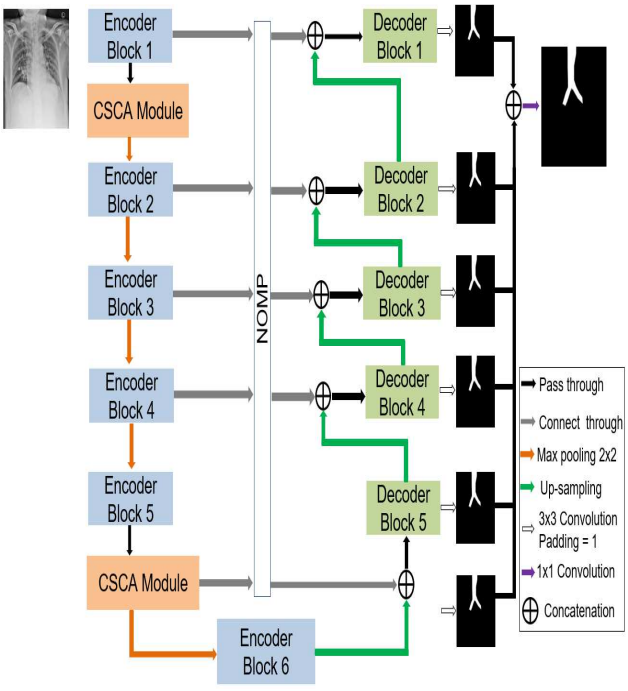


Fig. 2. NOMP U2-Net Structure Diagram.

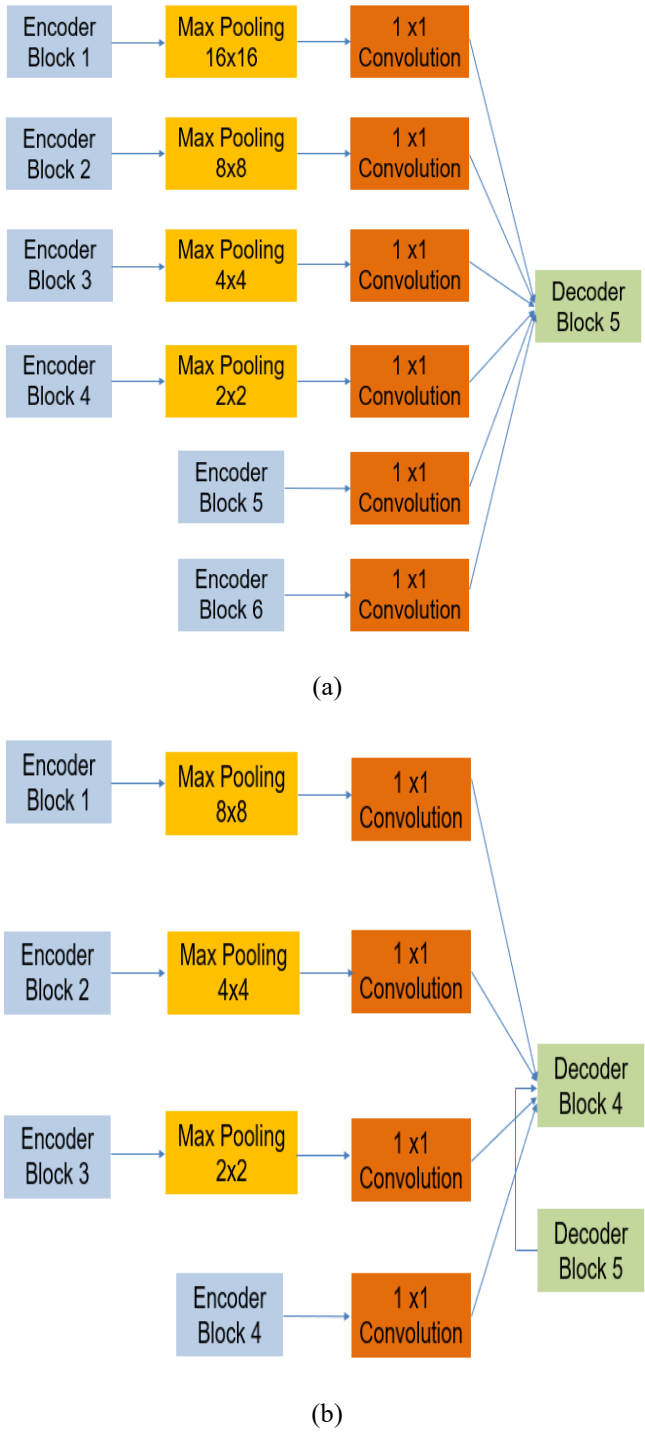


Fig. 3. (a) NOMP Decoder 5 link, (b) NOMP Decoder 4 link.

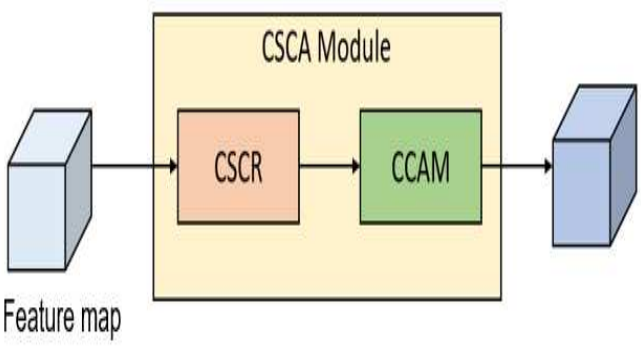


Fig. 4. CSCA module.

C. The Loss Function

The loss function clarifies how the actual value differs from the expected value. The deep learning training method involves minimizing the loss function. The projected value of the model approaches the actual value with decreasing loss functions. This work investigates Dice loss and BCE loss function to evaluate the model's performances [21]. Dice loss is often used in medical image segmentation when the background and foreground are unbalanced. The X and Y represent the pixel sets of prediction (Predict) and mask (Mask) consequently, and the range is between [0,1]. The dice loss equation (1) shows as follows:

$$\text{Dice loss} = 1 - \frac{2\,|X \cap Y|}{|X|+|Y|} \tag{1}$$

Binary Cross Entropy (BCE) cooperates with the Sigmoid function to calculate the loss of a single category. In bronchial segmentation, compare the predicted and masked bronchial parts, and adjust the model through the difference between the two values. In equation 2, N represents the pixel set of enter images, the y_i represents the mask, and $(\hat{y}_i)$ represents the prediction part, the range is between [0,2].

$$\text{BCE loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i * \log(\hat{y}_i) + (1-y_i) * \log(1-\hat{y}_i) \right) \tag{2}$$

Dice loss and BCE loss functions are combined to provide Dice_BCE loss shown in equation (3). The model is inclined to large gradient fluctuations, leading to unstable training. The model skews in favor of the background when BCE encounters an imbalance between the image's foreground and background. To address Dice loss and BCE loss functions and enhance the model's stability during training, two loss functions are combined in this work.

$$\text{Dice_BCE loss} = \text{Dice loss} + \text{BCE loss} \tag{3}$$

D. The Post-processing

Initially, the model predicts the bronchial segment as a contour of the bronchial mask and then counts the number of contours. The bifurcation locations and segmental bronchi are invisible without a contour. In the third step, concave points and edges are produced using convexity detection. The fourth part calculates the horizontal distance of each shape edge and selects the concave points corresponding to the two sides with the long distance as candidate points. The bronchial bifurcation point is the depression point that corresponds to the vertical gap between two sides of a massive space. The contour retains the largest gap effectively and solves the problem of noise interference, as shown in Figure 5.

The red color point is the bronchial bifurcation point obtained without setting the candidate point, and the green point is the setting candidate point, as shown in Figure 6. The resulting bronchial bifurcation points and blue dots are correctly labeled bronchial bifurcation points.

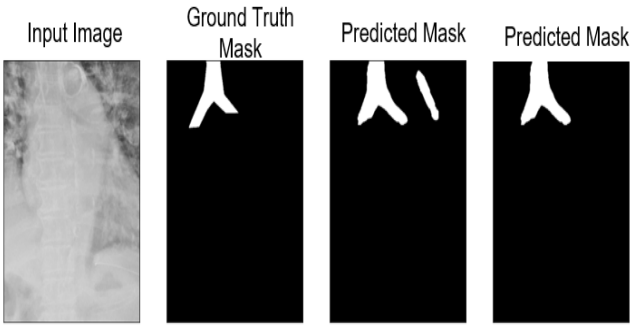


Fig. 5. Segmentation contours with large area preserved.

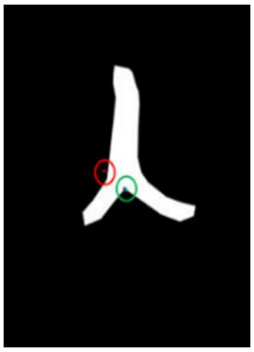


Fig. 6. Display of marked bronchial bifurcation points.

E. The Evaluation Criteria

This work uses four indicators to measure performance; Intersection over Union (IoU), average error distance (cm), accuracy within 0.5 cm, and 1 cm. The TP of IoU represents the total number of pixels that correctly predict the bronchi and are marked as positive. The FP is the total number of pixels that incorrectly predicts bronchi but is marked as positive. The FN is the total number of pixels wrongly forecasted as the background, marked as a negative value. The N represents the number of pictures in the test set, Y_i represents the mask, $(\hat{Y}_i)$ represents the prediction part, and P is 72 representing the total number of pixels converted to 1 cm.

$$\text{IoU score} = \frac{TP}{TP+FP+FN} \tag{4}$$

$$E_{_avg}\ (cm) = \frac{\frac{1}{N}\sum_{i=1}^{N}|Y_i-\hat{Y}_i|}{P} \tag{5}$$

$$E_{_0.5}\ (cm) = \frac{1}{N}\sum_{i=1}^{N} I_{0.5}\left(\left|Y_i-\hat{Y}_i\right| \le 0.5P\right) * 100\% \tag{6}$$

$$E_{1.0}\,(cm) = \frac{1}{N}\sum_{i=1}^{N} I_1\left(\left|Y_i - \hat{Y}_i\right| \le P\right) * 100\% \quad (7)$$

IV. Experimental Results

A. Dataset Rearrange

We obtained a real-case picture collection from a hospital for our research endeavor. There were 237 chest X-ray images. We have separated the data set into five folders, using three parts as training and two parts as validation and test datasets. In addition, did five times cross-validation to solve the overfitting problem caused by the scarcity of medical imaging samples.

B. Experimental Platform

We have used Pytorch 1.11 framework, and Adam optimizer for stochastic gradient descent to train the model. The number of iterations is 60, and the starting learning rate is 10^{-4}, so we decided to keep a batch size of 4. As a point of clarification, all testing has been performed on the Nvidia GeForce RTX 3080 GPU platform.

C. Experimental Results Analysis

The proposed NOMP U2-Net performs well in decreasing the error distance of the bronchial bifurcation point, with an average prediction error value of 0.3097 cm shown in Table 1. Besides, the internal accuracy is 0.815 percent for an error distance of 0.5 cm, and the accuracy is 0.970 cm for an error distance of 1.0 cm. The accuracy rate of error of 0.5 cm and 1 cm is enhanced after employing the proposed NOMP technique. The model concentrates on the segmentation of the bronchial component by adding the CSCA module. However, the cross-crowding method combines the channel's residual attention (CSCR) and cross-context attention mechanism (CCAM) in a Convolutional attention module CSCA, producing the best results for estimating the error distance of bronchial bifurcation points.

Table 1: Attention module ablation experiments

Attention	Dice_BCE	IoU	E_{avg}	$E_{0.5}$	$E_{1.0}$
CSCR	1.5267	0.707	0.421	0.780	0.940
CCAM	1.5262	0.713	0.405	0.801	0.957
CCAM & CSCR (Parallel)	1.5262	0.715	0.501	0.763	0.944
CCAM + CSCR (Series)	1.5259	0.717	0.415	0.741	0.953
CSCA Module	**1.5263**	**0.714**	**0.309**	**0.815**	**0.970**

Table 2 shows the result of the CSCA module connected to the intermediate layer of the encoder. The CSCA module restricted unnecessary features in the middle layers to perform well. The outcome of the first layer encoder using the attention module is utilized directly as the connection of the non-overlapping maximum pooling. The cross-squeezed convolutional attention module reduces the average error distance. The CSCA modules are incorporated to offer the maximum performance for forecasting the error distance of the bronchial bifurcation point in the middle layer, the fifth layer encoder, and the first layer.

Table 2: NOMP-U^2-Net Encoder Ablation Comparison

Structure	E_{avg}	$E_{0.5}$	$E_{1.0}$
NOMP U^2-Net	0.4079	0.8098	0.9618
U^2-Net + CSCA (**only en_1** add attention)	0.3634	0.7844	0.9491
U^2-Net + CSCA (**en_1 and en_1 connect** add attention)	0.4605	0.7335	0.9405
U^2-Net + CSCA (**only en_5** add attention)	0.3555	0.7755	0.9618
U^2-Net + CSCA **en_5 and en_5 connect** add attention)	0.3595	0.7715	0.966
U^2-Net + CSCA (**en_1 and en_5** add)	0.3848	0.7715	0.9364
U^2-Net + CSCA (**en_1 and en_5 and en_5 connect** add)	**0.3097**	**0.8157**	**0.9703**

V. Conclusions

We have adopted the U^2-Net model and modified it through Non-Overlapping Max Pooling (NOMP) operation to automatically detect the location of bifurcation points in chest X-ray images. In addition, we have developed a Cross-Squeeze Convolutional Attention (CSCA) link to make up the design of a neural network architecture. The proposed architecture retains essential and favorable feature information to achieve the best performance for predicting the error distance of bronchial bifurcation points. The result shows that the predicted average error distance of the bronchial bifurcation point is 0.3097 cm. Besides, the accuracy rate of detection error within 0.5 cm is 81.57%, and 1 cm is 97.03%. In the future, we will try to improve the structure of the model and try different attention mechanisms to see if we may achieve better performance by combining the Transformer modules to find the best model architecture.

References

[1] P. K. Gupta, K. Gupta, M. Jain, T. Garg, "Postprocedural Chest Radiograph: Impact On The Management In Critical Care Unit," in *Anesth Essays Res.* 2014 May-Aug;8(2):139-44. doi: 10.4103/0259-1162.134481.

[2] W. Brunel, D. L Coleman, D. E. Schwartz, E. Peper, N. H. Cohen, "Assessment Of Routine Chest Roentgenograms and The Physical Examination To Confirm Endotracheal Tube Positio," in *Chest.* 1989 Nov;96(5):1043-5. doi: 10.1378/chest.96.5.1043.

[3] M. -H. Sheu, S. M. S. Morsalin, C. -C. Hsu, S. -C. Lai, S. -H. Wang and C. -Y. Chang, "Improvement of Human Pose Estimation and Processing with the Intensive Feature Consistency Network," in *IEEE Access*, vol. 11, pp. 28045-28059, 2023, doi: 10.1109/ACCESS.2023.3258417.

[4] S. Soffer, A. Ben-Cohen, O. Shimon, M. M. Amitai, H. Greenspan, E. Klang, "Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide," in *Radiology* 2019 Mar; 290 (3):590-606. doi: 10.1148/radiol.2018180547.

[5] J. Wei *et al*., "Genetic U-Net: Automatically Designed Deep Networks for Retinal Vessel Segmentation Using a Genetic Algorithm," in *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 292-307, Feb. 2022, doi: 10.1109/TMI.2021.3111679.

[6] P. Lakhani, "Deep Convolutional Neural Networks For Endotracheal Tube Position And X-Ray Image Classification: Challenges And Opportunities," in *Journal of Digital Imaging*, *30*(4), 460-468 (20170. https://doi.org/10.1007/s10278-017-9980-7.

[7] M. -H. Huang, Z. -Y. Ting and S. -M. Guo, "Carina Detection on Chest Radiography," in *Proc. of the 2013 1st International Conference on Orange Technologies (ICOT)*, Tainan, Taiwan, 2013, pp. 127-130, doi: 10.1109/ICOT.2013.6521174.

[8] S. Kara, J.Y. Akers, P. D. Chang, "Identification and Localization of Endotracheal Tube on Chest Radiographs Using a Cascaded Convolutional Neural Network Approach," in *Journal of Digital Imaging*, 2021 Aug;34(4):898-904. doi: 10.1007/s10278-021-00463-0.

[9] H.C. Jung, C. Kim, J. Oh, T.H. Kim, B. Kim, J. Lee, J.H. Chung, H. Byun, M.S. Yoon, D. K. Lee, "Position Classification of the Endotracheal Tube with Automatic Segmentation of the Trachea and the Tube on Plain Chest Radiography Using Deep Convolutional Neural Network," in *Journal of Personalized Medicine* **2022**, *12*, 1363, https://doi.org/10.3390/jpm12091363.

[10] A. G. Roy, N. Navab, C. Wachinger, "Concurrent Spatial and Channel 'Squeeze & Excitation'in Fully Convolutional Networks," in *Proc. of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2018,* 2018, Volume 11070, pp 421–429, https://doi.org/10.1007/978-3-030-00928-1_48.

[11] L. Mou, Y. Zhao, L. Chen, J. Cheng, Z. Gu, H. Hao, H. Qi, Y. Zheng, A. Frangi, J. Liu, "CS-Net: channel and spatial attention network for curvilinear structure segmentation," in *Proc. of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019,* vol 11764, https://doi.org/10.1007/978-3-030-32239-7_80.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[13] S. L. H. Lau, E. K. P. Chong, X. Yang, and X. Wang, "Automated Pavement Crack Segmentation Using U-Net-Based Convolutional Neural Network," in *IEEE Access*, vol. 8, pp. 114892-114899, 2020, doi: 10.1109/ACCESS.2020.3003638.

[14] S. Woo, J. Park, J.-Y. Lee, & I. S. Kweon, "CBAM: Convolutional Block Attention Module." in *Proc. of the European Conference on Computer Vision (ECCV)*, vol 11211, https://doi.org/10.1007/978-3-030-01234-2_1.

[15] O. Ronneberger, P. Fischer, & T. Brox, "U-net: Convolutional Networks For Biomedical Image Segmentation," in *Proc. of the International Conference on Medical Image Computing and Computer-assisted Intervention-MICCAI 2015*, vol 9351, https://doi.org/10.1007/978-3-319-24574-4_28.

[16] M. -H. Sheu, S. M. S. Morsalin, S. -H. Wang, Y. -T. Shen, S. -C. Hsia and C. -Y. Chang, "FIBS-Unet: Feature Integration and Block Smoothing Network for Single Image Dehazing," in *IEEE Access*, vol. 10, pp. 71764-71776, 2022, doi: 10.1109/ACCESS.2022.3188860.

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, & J. Liang, "Unet++: A Nested U-Net Architecture For Medical Image Segmentation," in *Proc. of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support- DLMIA ML-CDS 2018* vol 11045, https://doi.org/10.1007/978-3-030-00889-5_1.

[18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, & M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," in *Pattern Recognition*, vol.106, 107404. https://doi.org/10.1016/j.patcog.2020.107404.

[19] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *Proc. of the* European Conference on Computer Vision, pages 483–499. Springer, 2016.

[20] W. Shi et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in *Proc. of the* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1874-1883, doi: 10.1109/CVPR.2016.207.

[21] H. Wei and P. Peng, "The Segmentation of Retinal Layer and Fluid in SD-OCT Images Using Mutex Dice Loss Based Fully Convolutional Networks," in IEEE Access, vol. 8, pp. 60929-60939, 2020, doi: 10.1109/ACCESS.2020.2983818.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Identifying the Style of Chatting

Manman Zhang*, Yuchen Ma*, Ge Luo†, Sheng Li, Zhenxing Qian, Xinpeng Zhang
School of Computer Science, Fudan University
* mmzhang21@m.fudan.edu.cn, ycma21@m.fudan.edu.cn
† gluo18@fudan.edu.cn
lisheng@fudan.edu.cn, zxqian@fudan.edu.cn, zhangxinpeng@fudan.edu.cn

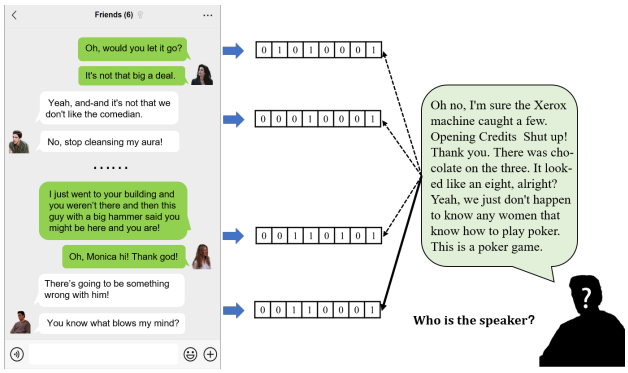


Fig. 1. An illustration of the process of identifying speaker with GCSH in group chat.

***Abstract*—Group chat brings great convenience to people's online social interaction. When identifying users in online social networks, individual speaking style of the member in group chat particularly plays an important role. In this paper, we propose a novel and efficient group chat hashing framework termed Group Chat Style Hashing(GCSH), which is the first work to utilize personal hashes in group chats to identify the speakers. GCSH combines a supervised VAE and a discriminator to extract accurate identity-related features and applies a new sample aggregation hash algorithm to generate an exclusive identity style hash for each one in the group chat. Our model fuses more context information to directly get the representative hashes, achieving an end-to-end identification model and sharply reducing the matching cost compared with previous text hashing methods. We conduct extensive experiments to prove the effectiveness and the generalization ability of our method across datasets.**

I. Introduction

With the internet fast developing, group chat scene has become a kind of conversation scene that people must come into contact with in the online social network. Numerous multi-person collaborative tasks and multi-person information sharing are realized through the group chat function of social software. Group chat involves rich multimodal information interaction. The members of group chat are able to send text, pictures, voice short videos and other forms of information to communicate with each other. Different from blogs and forums, contextual messages in group chat are more closely related. Group chat has the advantages of asynchronous communication and keeping users aware of the ongoing or past group agenda through traceable chat records. These features enable comfortable online group communication among members in different locations without disturbing their daily lives. With such rich and diverse interactive multimodal information, group chat is valuable for various research tasks. For example, it can help multimodal semantic comprehension tasks and multimodal content generation tasks.

In the group chat scene, user identification of group chat members has become a necessary research topic, which is also called authorship attribution in some studies. User identification is a task of identifying authors of anonymous content, according to the given examples of the writing of a predefined set of candidate authors [1]. In the task, we use given authors and content to analyze the features of each author, and distinguish the differences among the authors. In the test, the features extracted from each anonymous content are compared with those of known authors one by one. Among them, the author with the highest similarity to the anonymous content is most likely to be the author of the anonymous content. Therefore, in group chats, regardless of whether group chat members speak anonymously or not, we can target them easily. User identification now has been increasingly used in the fields of literature, intelligence, criminal law, civil law and computer forensics [2]. In addition, it also plays an important role in many areas such as information retrieval, information extraction and question answering.

Among the solutions to the user identification, text hashing has received widespread attention from researchers. Text hashing is a method that we can map the data points from the original space into the Hamming space of binary hash codes where the similarity in the original space is preserved in the Hamming space. By using binary hash codes to represent the original data, the storage cost can be dramatically reduced. Compared with using the classifier to solve the user identification problem, it is easier and faster to get high retrieval accuracy and high computational efficiency in the Hamming space by using text hashing.

Most of the existing methods using text hashing to solve the problem of user identification basically operate on traditional large-length texts such as news and literature. Few of them are suitable for short texts, especially in group chat scene because of a series of new challenges [3]. Compared with traditional scenes, what group chat members said are generally short in length and their personalities are not obvious [4]. The messages are mostly unstructured, informal and even grammatically incorrect, which are incomparable to elaborate literature and poetry in nature [5]. Only using the traditional

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

text hashing method can not solve the problem in group chat very well, their accuracy and performance are relatively poor in group chat scene.

In this paper, considering the particularity of short texts and the characteristics of group chat scene, we believe that the long-term speaking style of members in group chats and the personalities reflected in their words will have an important impact on the binary hash codes generated to distinguish different people. Therefore, we propose the Group Chat Style Hashing(GCSH), which is first work to generate group chat hashes to identify the speaker when someone speaks anonymously in online social networks. Figure 1 is an example, which illustrates the process of identifying speaker with our model in group chat. Our model contains a supervised Variational Auto-Encoder(VAE) [6], [7] and a discriminator. The encoder of VAE is responsible for extracting identity-related vectors and latent private vectors from the input message sentences. The decoder of VAE uses the above vectors to reconstruct the texts. The discriminator is used to form a Generative Adversarial Network(GAN) [8] framework with the decoder, which aims to help the decoder to get better performance. We also design a new sample aggregation hash algorithm. We aggregate the identity-related vectors of all of the messages belonging to one person together. So we can get the specific identity-related vectors of everyone to achieve high computational efficiency during identification.

The main contributions of this paper are summarized below:

- We propose the application of identifying the anonymous speakers in group chats with hashing, which can get accurate personal hashes from the messages of group chat members. To the best of our knowledge, this is the first work to generate personal hashes in group chats.
- We propose a novel supervised framework for group chat: Group Chat Style Hashing. We use the supervised VAE and the discriminator to extract accurate identity-related information of each member in group chats. We also design a new sample aggregation hash algorithm which averages the identity-related vectors of each person's messages so that we can get the exclusive personal style hashes for group chat members. With the supervised framework and the sample aggregation hash algorithm, we achieve an end-to-end generation of binary hash codes in group chats.
- A large number of experiments verify the advantages of our proposed model. We achieve better performance in group chats and significantly reduce the number of comparisons during the identification. Compared with previous text hashing methods, our model have higher identification accuracy and get better performance in terms of time consumption.

The rest of the paper is organized as follows. Section 2 reviews the related works of text hashing and group chat. Section 3 describes the proposed method in detail and experimental results are presented in Section 4. Finally Section 5 draws the conclusion of this paper.

II. Related Work

A. Group Chat

Group chat has attracted more and more attention from researchers in recent years. In [9], Soomin *et al.* proposed a bot named GroupfeedBot in group chat, which can facilitate group chat discussion more efficient by managing the discussion time, encouraging members to participate an organizing members' opinions. And another work [10] used emotion regulation and text analysis technology to propose a robot to guide the emotional enthusiasm of group chat. During the COVID-19 pandemic, some researchers analyzed the information of the group members from five dimension: participation, sentiment, negative emotions, topics, and message types [11]. They used this method to judge the opinions of the people and whether rumors had been spread in the telegram group. Tilda [12], a prototype system, was proposed to enable users more easily to make sense of the dialogues in group chat they missed. It enriched the representation of chat conversations, using techniques such as tagging and summary.

Some researches about response generation in group chats are also related. Ouchi *et al.* [13] proposes a framework called Dynamic-RNN to fulfil the response generation task. On this basis, SI-RNN [14] is raised to process group chat members in a role-sensitive manner, with which the performance of response generation is improved. But both of them are retrieval-based approaches and cannot generate diverse responses. Another work called GSN [15] designs a graph network to represent group chat members' interaction and generates responses. ICRED [16] models group chat members according to different roles, and then integrates the member information into the response generation. Although many valuable researches existed, there are still many research gaps in group chats presently.

B. Text Hashing

There have been many valuable research achievements in the field of text hashing. Locality Sensitive Hashing (LSH) [17] is one of the most popular unsupervised hashing methods, which simply used random linear projections to construct hash functions leading to better performance. Self Taught Hashing (STH) [18] is an excellent hashing method which excluded document features and only considered the document connectivity. In [19], Wang *et al.* proposed Semantic Hashing using Tags and Topic Model (SHTTM) to incorporate tags to obtain more effective hashing codes via a matrix factorization formulation.

Another great achievement in this field is semantic hashing [20]. It built a stack of restricted Boltzmann machines(RBMs) [21] to discover hidden binary units which could model input text data, i.e., word-count vectors. After learning a RBM through pre-training and tuning on a collection of documents, the hash code of the document could be acquired by the threshold of the output of the layers. The work [22] exploited convolutional neural network for text hashing, which embed the keyword features into compact binary code with

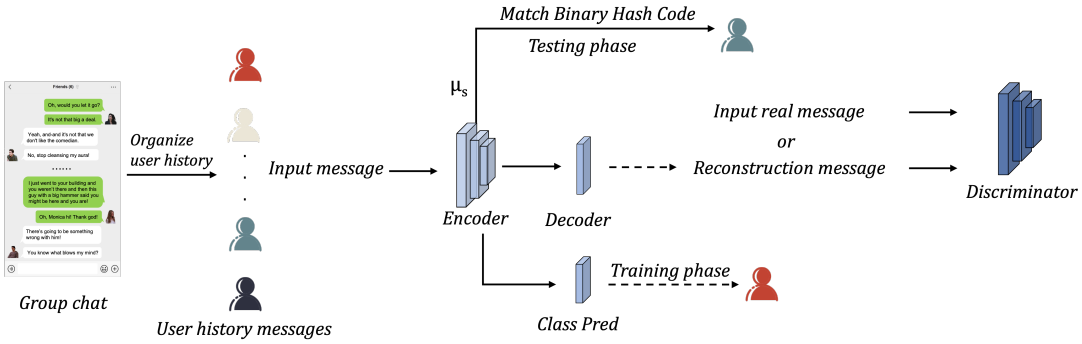


Fig. 2. Overview of the proposed model. We collated each user's historical chat data from user social group chats into the model. The encoder extracts the latent vectors and decoder reconstructs the message. The discriminator penalizes the framework for better reconstruction quality and more representative vectors. In the training phase, adding prediction branches helps the model learn better feature vectors. In the testing phase, the hash code is generated through the vector $\boldsymbol{\mu_s}$ learned by the encoder for direct user matching.

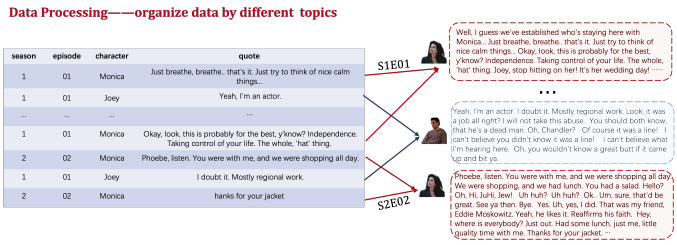


Fig. 3. An example of our data processing with the group chat messages in the sitcom "Friends". We categorize all the quote in the sitcom "Friends" by different episodes and different chat topics.

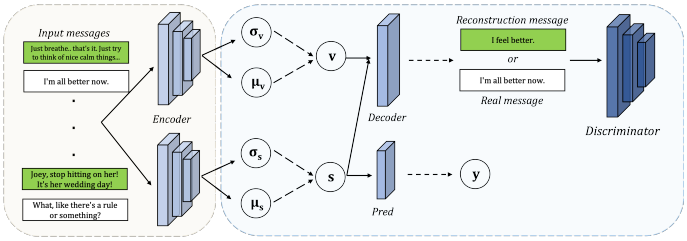


Fig. 4. a) The encoder extracts the identity-level vector $\boldsymbol{s}$ and private variables $\boldsymbol{v}$. b) The decoder reconstructs the message by the vector $\boldsymbol{s}$ and $\boldsymbol{v}$. c) The discriminator enables the framework to generate better reconstruction messages and more representative $\boldsymbol{s}$ and $\boldsymbol{v}$.

a locality preserving constraint. Meanwhile it combined the implicit features learned from word features and position features with the explicit features to fit the pre-trained binary code.

Deep generative models have made impressive progress with the introduction of the Variational Auto-Encoder(VAE) [6], [7] and Generative Adversarial Network(GAN) [8]. Some researchers applied these great achievements to text hashing. For example, Suthee *et al.* proposed the Variational Deep Semantic Hashing(VDSH) [23], an unsupervised model, which applied VAE into text hashing. At the same time, they also proposed a series of supervised models, i.e., VDSH-S and VDSH-SP. They generated hash codes by using the latent semantic vectors that generated from the encoders. And the supervised models enhanced the learning performance through using the labels in the decoders. In [24], with the use of BM25 as a weak annotator, Suthee *et al.* proposed a weak supervisied method based on the model in [23].

III. Group Chat Style Hashing

The proposed model is a group chat style hashing framework based on supervised VAE and the discriminator as shown in Figure 2, generating unique hashes for people for identification. Details will be shown in the following.

A. Group Chat Messages Organization

Messages in group chats are generally short in length, even unstructured and grammatically incorrect, which are very different from traditional texts. The length and content of any single message is not enough to obtain the style hash of a member in group chat. Due to the particularity of group chat messages, we need to process group chat messages differently than before. Figure 3 shows our data processing with an example.

The previous methods deal with data very little. In general, it is directly calculated with a piece of text. The data can be expressed by the formula as follows:

$$T = \{t_1, t_2, t_3, \cdots, t_{r-1}, t_r\} \tag{1}$$

where T represents the collection of all texts and t represents a piece of text in the collection. Such processing is obviously not suitable for group chat scenarios. In this paper, we collect all the historical conversations of group chat members first, which can be seen in the following equation:

$$C = \{C_{p_1}, C_{p_2}, C_{p_3}, \cdots, C_{p_{n-1}}, C_{p_n}\} \tag{2}$$

where C represents the total collection of all the messages spoken by everyone in the group chat. It contains all the messages in historical record. We categorize it by group chat members and we can get each person's message collection,

which is marked as $C_{p_i}, i \in [1, n]$, n is the number of the members in the group chat.

Then we classify all the chat messages of one member in group chat according to the topic and the dialogue scene. The dialogue scene mentioned here refers to a group chat conversation about one topic and the context is closely related. The following equation describes the classification result.

$$C_{p_i} = \left\{ C_{p_i}^{s_1}, C_{p_i}^{s_2}, C_{p_i}^{s_3}, \cdots, C_{p_i}^{s_{m-1}}, C_{p_i}^{s_m} \right\}, i \in [1, n] \quad (3)$$

where $C_{p_i}^{s_j}, i \in [1, n], j \in [1, m]$ is the message collection of the group chat member p_i in the dialogue scene s_j. m is the number of the topics or dialogue scenes that the group chat member p_i has participated in.

As for the collections $C_{p_i}^{s_j}, i \in [1, n], j \in [1, m]$, each of them contains all the messages of a group chat member in that scenario, and we put them together. It is shown as the following equation:

$$C_{p_i}^{s_j} = \{ s_1, s_2, s_3, \cdots, s_{q-1}, s_q \}, i \in [1, n], j \in [1, m] \quad (4)$$

where s represents a piece of message. The q in this formula is an indeterminate value, which is determined according to the amount of messages of the group chat member in the actual scene.

B. Hashing Architecture

The details of the model are shown in Figure 4. Let $\boldsymbol{x}$ be the bag-of-words representation of the messages of somebody in the group chat, and the latent vectors of the input message are obtained through the encoder. In the ordinary VAE framework, we usually reconstruct the texts by latent vectors without supervised labels.

However, in the group chat scenario, even if different people speak similar words, their identity-related high-level features such as personal speaking style are not the same. In other words, input text with similar content in the original space may not be similar in the identity-related feature space. Therefore, we need to use the person label information in the group chat to generate identity-related high-level features $\boldsymbol{s}$ for each person. There is another problem. Not in all cases, messages are always related to speaker characteristics. For example, someone who says "hi", "thank you", which have little relationship with their personalities. So we cannot extract unique identity-related features based on this. Inspired by VDSH-SP, we apply a supervised VAE with an extra latent private vector $\boldsymbol{v}$. Specifically, identity-related vector $\boldsymbol{s}$ extracts identity-related high-level features from input texts and label information while $\boldsymbol{v}$ only contains the input-specific information which is irrelevant to the identity label. Then, $\boldsymbol{s}$ and $\boldsymbol{v}$ jointly through the decoder to reconstruct the text. The class prediction is obtained from the $\boldsymbol{s}$ variable through the decoder and softmax function. In this way, we can ensure to the greatest extent that $\boldsymbol{s}$ contains valid identity features without other interference information. It is important for the subsequent generation of specific hashes for each person.

To enable the encoder-decoder structure to reconstruct more realistic texts and make the latent variables more in line with the original data distribution, we add a discriminator to the basic model to form a GAN framework with the decoder. The discriminator hopes to distinguish the reconstructed message from the original message as much as possible. While the generator, which means the decoder, expects the reconstructed text as similar as possible to the original text, thus fooling the discriminator. Finally, through the GCSH model, we can get an identity-related vector that is more consistent with the input to compute the binary hash code.

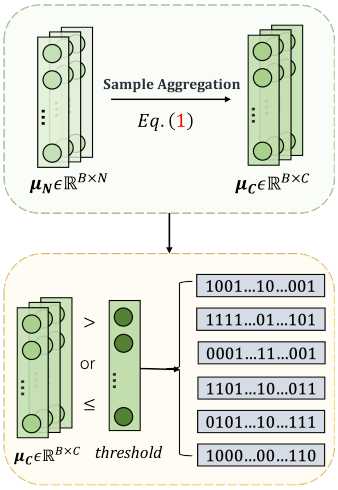


Fig. 5. The sample aggregation hash algorithm aggregates $\boldsymbol{\mu_N}$ to $\boldsymbol{\mu_c}$, then generate hash for each person by comparing $\boldsymbol{\mu_c}$ with threshold.

C. Sample Aggregation and Binary Hash Code

Different from traditional long texts such as news and articles, messages in group chats are usually short and informal. Due to the limited information and lack of rigorous text structure, it is difficult for us to obtain good contextual features. In addition, traditional retrieval or identification requires matching test data with all samples of all categories in the database, which is complicated and inefficient.

To address these issues, we propose a new sample aggregation hash algorithm to generate a unique hash for each person in the group chat, achieving an end-to-end identification framework. Specifically, we aggregate the identity-related variables $\boldsymbol{s}$ of samples belonging to the same person to directly obtain specific identity-related vectors of different people, as shown in the following equation and Figure 5.

$$\boldsymbol{\mu_c} = \frac{\sum_{k=1}^{N} \boldsymbol{\mu_k}}{N} \quad (5)$$

where N represents the total number of samples in each class, and we average each sample to their class to obtain the specific $\boldsymbol{\mu_c}$ value of each person. The expectation of $\boldsymbol{s}$ in the standard Gaussian distribution is $\boldsymbol{\mu}$, so the average of $\boldsymbol{s}$ is equal to the average of $\boldsymbol{\mu}$ in actual calculations.

Then, we take the median of $\boldsymbol{\mu_c}$ of all people as the threshold for obtaining the binary hash code [23]. If the m^{th} bit of $\boldsymbol{\mu_c}$ of someone is larger than the m^{th} bit of the threshold, the binary hash code of someone is set to 1, otherwise, it is 0.

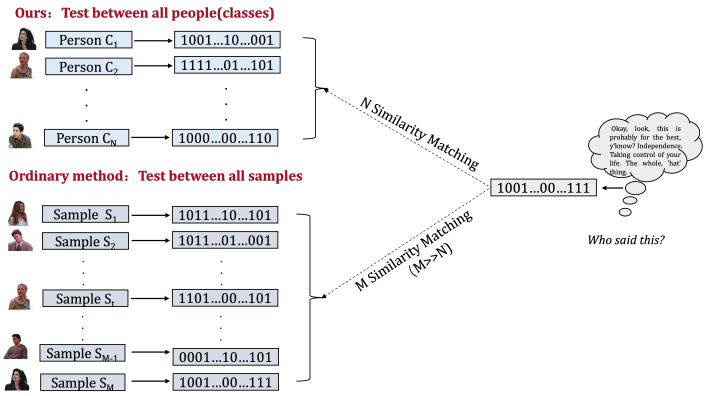


Fig. 6. Only need to compare the test data with each person's hash code instead of matching enormous samples.

It means, we generate the unique hash for each person in the group chat.

Through this strategy, we can not only directly obtain the representative style hash for each person, but also fuse more context information of multiple samples to obtain more accurate feature vectors. Instead of comparing all samples and then determining the final attribution of the test data based on the matched sample's label, we only need to compare the test data with each person's hash code during identification as shown in Figure 6. Therefore our method achieves an end-to-end identification framework, which is a complete process rather than multiple stages. Accordingly, the computational complexity during identification is reduced from the number of samples to the number of classes, which greatly reduces the magnitude of calculations.

D. Object Functions

The GAN structure composed of the decoder and discriminator uses the least-squares GANs[25] to optimize the GAN framework for more accurate identity-related vector $\boldsymbol{s}$ and improved quality of reconstructed texts, which can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[D\left(\boldsymbol{x}\right)-1\right]^2+\frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{x}}}\left[D(\hat{\boldsymbol{x}})\right]^2 \\ &+ \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[D(G(\boldsymbol{z})\right]^2 \end{aligned} \tag{6}$$

where $\boldsymbol{x}$ represents the input text, E is the encoder, G means the decoder and also generator, and D is the discriminator. The reconstructed text $\hat{\boldsymbol{x}}=G(E(\boldsymbol{x}))$. And $\boldsymbol{z}$ represents the random variable obtained from the standard normal distribution, used for the training of the discriminator.

The loss of the VAE is divided into three parts. The first part is the basic loss part, including the KL loss of $\boldsymbol{s}$ and $\boldsymbol{v}$ and the class prediction loss.

$$\mathcal{L}_{kl}=D_{KL}(Q(\boldsymbol{s})||P(\boldsymbol{s}))+D_{KL}(Q(\boldsymbol{v})||P(\boldsymbol{v})) \tag{7}$$

$$\mathcal{L}_{pred}=\ell_{CE}(\hat{\boldsymbol{y}},\boldsymbol{y}) \tag{8}$$

where the $Q(\boldsymbol{s})$ and $Q(\boldsymbol{v})$ represent the distribution of $\boldsymbol{s}$ and $\boldsymbol{v}$. It is worth noting that we usually assume that $\boldsymbol{s}$ and $\boldsymbol{v}$ obey the standard Gaussian distribution $P(\boldsymbol{s})$ and $P(\boldsymbol{v})$. The second part is the reconstruction loss of the decoder and the discrimination error of the discriminator.

$$\mathcal{L}_{re}=\frac{1}{2}(\alpha||\boldsymbol{x}-\hat{\boldsymbol{x}}||^2+||D(\boldsymbol{x})-D(\hat{\boldsymbol{x}})||^2) \tag{9}$$

where α denotes the weight of the reconstruction loss of the decoder and we set it to 0.01 in our experiment. The third part is the discrimination error of the random data.

$$\mathcal{L}_{dis}=\frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[D(G(\boldsymbol{z})-1\right]^2 \tag{10}$$

Finally, the loss of VAE can be expressed as:

$$\mathcal{L}=\mathcal{L}_{kl}+\mathcal{L}_{pred}+\beta\mathcal{L}_{re}+\mathcal{L}_{dis} \tag{11}$$

where β is the weight of $\mathcal{L}_{re}$ and usually be seted to 0.01.

IV. Experiment

We collect two datasets by ourselves and we conduct our experiments on them. We will extract the group chat style hashes of members in group chats during the training stage, and then in the testing phase we will test on different datasets to identify the group chat members. In this section, we first introduce the datasets we bulit and the experiment setup. Then we present the experiment results and generalization analysis of our model.

A. Datasets

For our experiments, We build two group chat dialogue datasets named "Friends" and "HIMYM" respectively, which contain intensive chat conversations among the main characters in different thematic scenarios. In this part, we first introduce the data collection procedure and then present the properties of these two datasets.

1) Data Collection: Due to the lack of high-quality group chat data and the copyright issues of private chat messages, we build two datasets of group chat dialogues in TV sitcoms to create a real group chat environment as much as possible. We choose the dialogue between the main characters in two sitcoms "Friends" and "How I Met Your Mother"(HIMYM) as the datasets. We believe that the scene in the sitcom "Friends" is more similar to the group chat scene, even the topics and the dialogue content in the sitcom "Friends" are richer. So we collect the Friends dataset. As for the sitcom "How I Met Your Mother", we believe that the protagonists in it have similar personalities to the protagonists in "Friends", and the dialogue content and chat themes are equally rich. Therefore, we collect the HIMYM dataset for cross-dataset test in the experiments.

After filtering out low quality and incomplete data, we finally collect 58,911 messages from lines in 236 episodes of the sitcom "Friends" and collect 34,589 messages from lines in 133 episodes of the sitcom "How I Met Your Mother". We will continue to expand the datasets in the future.

TABLE I

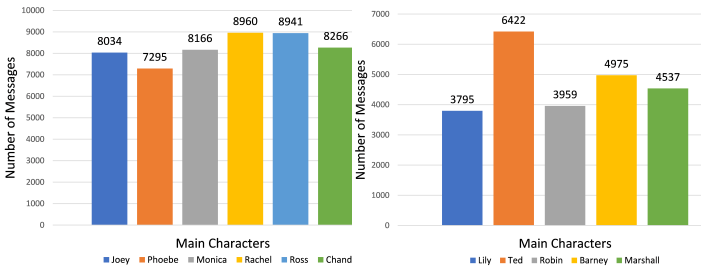
ACCURACY ON THE FRIENDS DATASET(TOP100/1089 MEANS THE PRECISION OF THE TOP 100 RETRIEVED TEXTS AMONG ALL SAMPLES. TOP1/6 MEANS THE PRECISION OF THE TOP 1 MATCHED AMONG SIX PEOPLE. PRECISION IS EQUIVALENT TO THE ACCURACY IN THIS TOP1 SITUATION)

Method	Top100/1089					Top1/6(Sample Aggregation)				
	8 bits	16 bits	32 bits	64 bits	128 bits	8 bits	16 bits	32 bits	64 bits	128 bits
VDSH-S	0.5270	0.5650	0.5501	0.5892	0.6054	0.6324	0.6397	0.6544	0.6618	0.6912
VDSH-SP	0.5982	0.6025	0.6607	0.6535	0.7001	0.6250	0.6471	0.7132	0.6838	0.7279
GCSH	**0.6409**	**0.6517**	**0.6911**	**0.6998**	**0.7190**	**0.6544**	**0.6838**	**0.7132**	**0.7353**	**0.7647**

TABLE II

THE COMPARISONS OF DIFFERENT RETRIEVED SCOPES AND ABLATION STUDIES WITH 128-BIT HASH

Method	Top100/1089	Top50/1089	Top10/1089	Top1/1089	Top1/6
VDSH-SP	0.7001	0.7076	**0.7184**	0.7147	0.7279
GCSH	**0.7190**	**0.7206**	0.7162	**0.7426**	**0.7647**



(a) Friends Dataset. (b) HIMYM Dataset.

Fig. 7. Statistics of main characters messages in Friends and HIMYM Dataset

2) Data Statistics: The Friends dataset contains 58,911 messages spoken by all characters, from which the dialogues of six main characters are selected. Figure 7(a) illustrates the number of messages from six main characters. The number of messages from the five main characters exceeded 8,000, except for Phoebe, only 7,295. Categorized by different chat topics and different chat scenarios, we divide what they said into many pieces, each of which contains all the messages of one person in one scenarios, totaling 1362 pieces.

The HIMYM dataset contains 34,589 messages spoken by all characters. With filtering all the messages, we remove a series of low-quality dialogues and some dialogues of unimportant characters. Finally, we get the dialogues of the five main characters, which is shown in Figure 7(b) in detail. Because this dataset is mainly used for cross-dataset test, it is not as large as the Friends dataset. The same processing method as the Friends dataset is used for the HIMYM dataset. We divide the messages according to the dialogue scenes, and obtain a total of 663 pieces, which consist of the five main characters dialogues in different chat scenarios.

The Friends dataset was split into three subsets with roughly 80% for training, 10% for validation, and 10% for test. The HIMYM dataset is used to test the model's generalization ability across datasets in the experiments.

B. Implementation Details

There are three components in GCSH, i.e., encoder, decoder, and discriminator. We apply ReLU, Tanh, and Leaky ReLU activation functions to the encoder, decoder(generator), and discriminator respectively.

In addition, GCSH is implemented via PyTorch and optimized by Adam with the learning rate of 0.001. We trained all models with a batch size of 16 on NVIDIA GeForce GTX 1080 Ti GPU with 11GB V-RAM.

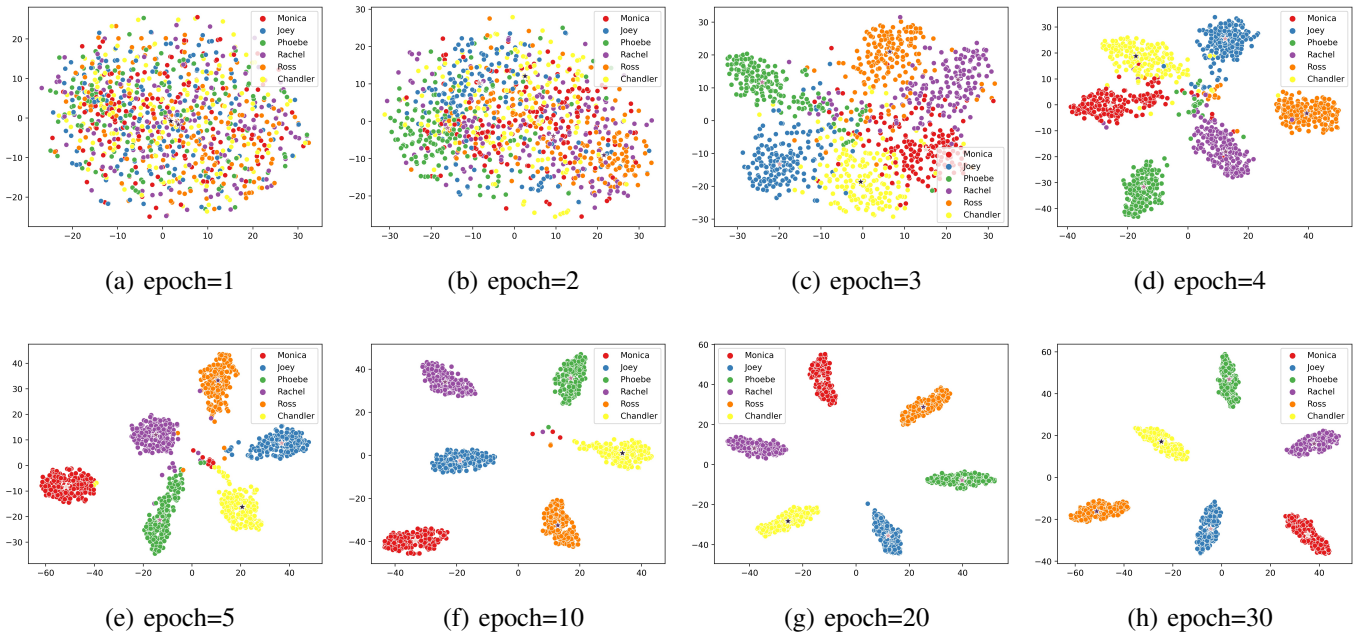
C. Experimental results and Ablation Study

We evaluate the proposed model on the Friends dataset together with the compared models, VDSH-S and VDSH-SP, proposed in [23] as shown in Table I. Obviously, our model gets better results than the baseline in all cases. The best accuracy is achieved to 0.7647 by our model using the new sample aggregation hash algorithm. The introduction of the discriminator allows us to get more realistic hashes than the baseline. We can also find that the best performance of all models is achieved when we use the 128-bit hash. The reason is that the hash of more bits can contain more information, which helps us to distinguish the different people better.

It is noteworthy that our sample aggregation hash algorithm plays an important role. The algorithm can increase the accuracy by 1%-5% and reduce the number of comparisons from 1089 to 6. The effect is especially obvious when using the 128-bit hash. We perform ablation studies to show the advantages of our algorithm more clearly as shown in Table II. As the retrieved samples decrease, the precision or accuracy(when Top1) is on the upward trend. Compared to the baseline VDSH-SP, the accuracy has increased from 0.7147 to 0.7426 when we utilize the discriminator. After the sample aggregation algorithm is adopted, the number of comparisons is reduced from 1089 to 6, but the accuracy does not decrease. Instead, the accuracy increases from 0.7426 to 0.7647, which proves the effectiveness of the algorithm. Figure 8 visualizes the different training phases of our proposed model. Each point in the figure represents a message someone said, and different colors refer to different people. As we can see in (a), the points are very messy at first. With the training of the model, the points gradually cluster together and the distances among clusters increase. In (h), different classes points have been completely separated.

D. Cross-Dataset Similarity Test and Generalization

To further validate GCSH can generate similar hashes for similar people, we calculate the Hamming distance to represent the similarity between different hashes. The smaller the distance, the greater the similarity. We set the similarity thresholds and divide the similarity among characters into three



(a) epoch=1 (b) epoch=2 (c) epoch=3 (d) epoch=4

(e) epoch=5 (f) epoch=10 (g) epoch=20 (h) epoch=30

Fig. 8. Visualization of the 16-dimensional identity-related vector $\boldsymbol{\mu}$ after dimensionality reduction.

TABLE III
THE SIMILARITY TEST RESULT ON THE FRIENDS DATASET (WE REPORTED THE DISTANCE BETWEEN HASHES RESPECTIVELY OBTAINED FROM THE TRAINING SET AND TEST SET FOR THE SAME PERSON, WHERE THE SMALL DISTANCE MEANS HIGH SIMILARITY.)

nbits	Monica	Joey	Phoebe	Rachel	Ross	Chandler
128 bits	22	26	32	**17**	36	33
64 bits	12	19	14	**3**	18	19
32 bits	9	10	10	**7**	9	9
16 bits	3	4	**1**	4	6	5
8 bits	2	**1**	2	2	2	**1**

TABLE IV
SIMILARITY TEST BETWEEN DIFFERENT CHARACTERS IN FRIENDS AND HIMYM WITH 128-BIT HASH

Character	Monica	Joey	Phoebe	Rachel	Ross	Chandler
Lily	59	65	76	63	64	**57**
Barney	78	**48**	73	70	67	48
Marshall	76	52	71	70	65	**50**
Robin	67	65	70	69	60	**53**
Ted	76	60	73	74	59	**42**

intervals, which are 0%-50%, 50%-70%, and 70%-100%. They represent the dissimilar, similar, and almost same respectively. We firstly conduct the similarity test on the Friends dataset, which calculates the similarity between the hash obtained from the training set and the hash obtained from the test set for the same person under different hash lengths. We find that almost all of the results in Table III are in the range of 70%-100% of their corresponding hash lengths(e.g. $1 - 17/128 \approx 87\%$). This means that our model can effectively generate almost the same exclusive hash for the same person with different test data. Very few results do not meet the conclusions because some words don't have personal characteristics, for example, hello, thanks. It is difficult to distinguish them accurately in the test.

To verify the generalization of our proposed model, we conduct a cross-dataset test between the Friends and HIMYM datasets. We want to know whether people with similar personalities in different datasets have similar hash codes, that is, how generalized the model is. Table IV indicates the distance between the characters in Friends and HIMYM under 128-bit length. According to our acknowledgment, there are two pairs of characters with similar personalities in these two sitcoms, namely Chandler and Marshall, Joey and Barney. The similarity of Barney-Joey is 62.5% and Marshall-Chandler is 60.9% which are in the range of 50%-70%. It meets the same conclusion as the actual personalities of characters in the sitcoms. This proves our model has great generalization ability that can generate similar hashes for similar people. For most people with different personalities, the results are almost all less than 50%. But there are also few abnormal results all related to Chandler. The reason is that Chandler has similarities with other characters and his character may not be clearly reflected in the lines, perhaps through tone and behavior.

V. CONCLUSION

In this paper, we proposed a novel group chat hashing model, which is called Group Chat Style Hashing(GCSH). It can generate personal style hashes of group chat members and achieve end-to-end identification in group chats efficiently. We designed a new sample aggregation hash algorithm, which helps us get the exclusive personal style hashes for group chat members and leads to a sharp reduction of the comparison number during the period of identification. Extensive experi-

ments demonstrate the superiority and good generalization of our model in the group chat scene.

REFERENCES

[1] C. Zhang, X. Wu, Z. Niu, and W. Ding, "Authorship identification from unstructured texts," *Knowledge-Based Systems*, vol. 66, pp. 99–111, 2014.

[2] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in *International conference on artificial intelligence: Methodology, systems, and applications*, Springer, 2006, pp. 77–86.

[3] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," in *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, IEEE, 2013, pp. 1–6.

[4] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel, "Authorship attribution of micro-messages," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1880–1891.

[5] M. Bhargava, P. Mehndiratta, and K. Asawa, "Stylometric analysis for authorship attribution on twitter," in *International Conference on Big Data Analytics*, Springer, 2013, pp. 37–47.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*, PMLR, 2014, pp. 1278–1286.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[9] S. Kim, J. Eun, C. Oh, B. Suh, and J. Lee, "Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[10] Z. Peng, T. Kim, and X. Ma, "Gremobot: Exploring emotion regulation in group chat," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 2019, pp. 335–340.

[11] L. H. X. Ng and J. Y. Loke, "Analyzing public opinion and misinformation in a covid-19 telegram group chat," *IEEE Internet Computing*, vol. 25, no. 2, pp. 84–91, 2020.

[12] A. X. Zhang and J. Cranshaw, "Making sense of group chat through collaborative tagging and summarization," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–27, 2018.

[13] H. Ouchi and Y. Tsuboi, "Addressee and response selection for multi-party conversation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2133–2143.

[14] R. Zhang, H. Lee, L. Polymenakos, and D. Radev, "Addressee and response selection in multi-party conversations with speaker interaction rnns," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[15] W. Hu, Z. Chan, B. Liu, D. Zhao, J. Ma, and R. Yan, "Gsn: A graph-structured network for multi-party dialogues," *arXiv preprint arXiv:1905.13637*, 2019.

[16] C. Liu, K. Liu, S. He, Z. Nie, and J. Zhao, "Incorporating interlocutor-aware context into response generation on multi-party chatbots," *arXiv preprint arXiv:1910.13106*, 2019.

[17] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, 2004, pp. 253–262.

[18] D. Zhang, J. Wang, D. Cai, and J. Lu, "Self-taught hashing for fast similarity search," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 18–25.

[19] Q. Wang, D. Zhang, and L. Si, "Semantic hashing using tags and topic modeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 213–222.

[20] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[21] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade*, Springer, 2012, pp. 599–619.

[22] J. Xu, P. Wang, G. Tian, *et al.*, "Convolutional neural networks for text hashing," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[23] S. Chaidaroon and Y. Fang, "Variational deep semantic hashing for text documents," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 75–84.

[24] S. Chaidaroon, T. Ebesu, and Y. Fang, "Deep semantic text hashing with weak supervision," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1109–1112.

[25] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

Pose-Based Visual Servoing with Lightweight Deep-Learning Binarization for Autonomous Mobile Robot Application

Chian C. Ho* and Cian-Duo Lin
* Department of Electrical Engineering, National Yunlin University of Science and Technology, Taiwan
E-mail:futureho@yuntech.edu.tw Tel: +886-5-5524613
Department of Electrical Engineering, National Yunlin University of Science and Technology, Taiwan
E-mail: m10812042@yuntech.edu.tw Tel: +886-5-5524613

Abstract — **Pose-based visual servoing (PBVS) can complement the frequent drift issue of light detection and ranging (LiDAR) coordinate of LiDAR-based simultaneous localization and mapping (SLAM), navigation, and servoing technology, especially when autonomous mobile robot (AMR) works on the automatic docking alignment missions for automatic pallet engaging or automatic battery charging whose alignment precision requirement is extremely stricter. But PBVS occasionally suffers from the poor detected image quality of ARTag landmark to cause the reading drift or error. This paper proposes the lightweight deep-learning binarization method to preprocess the image quality of ARTag landmarks so that PBVS can evaluate the distance and pose between the ARTag landmark and the camera sensor more accurately, promptly, and steadily, for better feasibility of PBVS on the automatic docking alignment missions. Experimental results show, against conventional image-processing-based binarization, conventional computer-vision-based binarization, and conventional deep-learning-based binarization, the proposed lightweight deep-learning binarization not only raises the accuracy and reliability of ARTag's reading, but also apparently improves the effectiveness and efficiency of PBVS's operation, especially under environmental conditions of shadow occlusion, image blurring, low contrast, uneven illumination, or complex background.**

I. INTRODUCTION

With the ever-increasing demand of automation transportation technologies in intelligent factories or intelligent warehouses, automatic mobile robot (AMR) without occupying a significant amount of fixed space to facilitate automatic material handling is the best choice [1], [2]. Consequently, research on AMRs has become more and more attractive in both academic and industrial fields, and has further extended its application to more scenes. In general, AMRs can be divided into two main categories: tracked AMRs and trackless AMRs. Tracked AMRs require the deployment of magnetic tracks, color tapes, or QR codes, and they can only move along these tracks. Parking, acceleration, deceleration, docking and rotation totally rely on the infrastructure of laid tracks on the floor. Moreover, the track redeployment is more time-consuming and cost-expensive to redesign the hardware and software of infrastructure. Tracked AMRs has poor flexibility and adaptability to complex or specific scenes, while it also has high maintenance costs. Trackless AMRs primarily utilize light detection and ranging (LiDAR) for positioning and navigation. Trackless AMRs can detect the contours of objects in intelligent factories or intelligent warehouses, build up the depth point cloud maps of the working area, and plan navigation routes without any infrastructure deployment of laid tracks. However, the frequent drift error of LiDAR coordinate adopted by trackless AMRs for positioning and navigation is essentially 5-10 cm at minimum if no other optimization firmware or preprocessing algorithm is further applied. Besides, since LiDAR sensors use infrared light as their light source, there are certain physical limitations. For example, infrared light can penetrate glass, infrared light is extraordinarily reflected by the metal plates of machinery equipment, and infrared light is completely absorbed by some materials like fabric and paper. Therefore, in specific situations, positioning failures and navigation errors are unavoidable.

However, on the automatic docking alignment missions of automatic pallet engaging or automatic battery charging whose alignment precision requirement is extremely stricter, trackless AMRs with the frequent drift error of LiDAR coordinate of 5-10 cm at minimum is not eligible to work, even not safe to work. As a result, pose-based visual servoing (PBVS) is the simplest

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



Fig. 1. Illustration of ARTag.

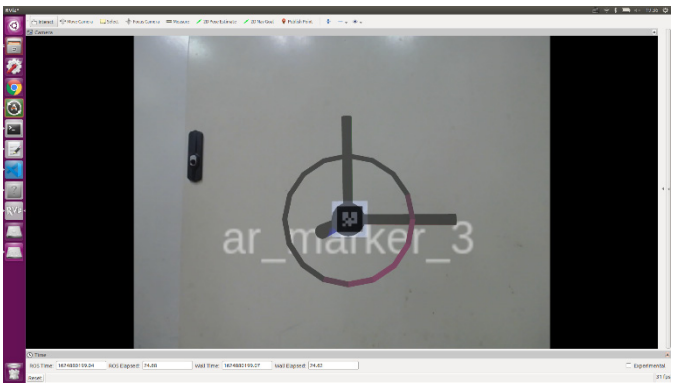


Fig. 2. Detection of the distance and pose between the ARTag landmark and the camera sensor.

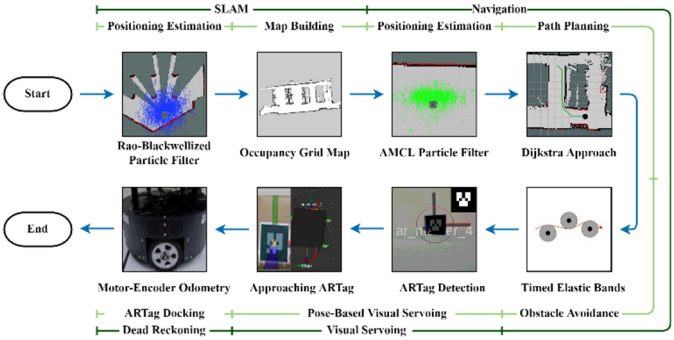


Fig. 3. Flowchart of conventional trackless AMRs.

but practical solution for trackless AMRs [3], [4]. PBVS can estimate the distance and pose between the tag landmark on the docking destination and the camera sensor on the trackless AMR, and then control the motion and rotation of trackless AMR to approach and align the docking destination precisely. With a few wallpapered and radioless tag landmarks, PBVS can offer additional simultaneous localization and mapping (SLAM) information to LiDAR-based trackless AMR for precise alignment and immediate relocalization.

This paper adopts ARTag shown in Fig. 1 as the tag landmark for PBVS functionality of trackless AMRs. Through the PBVS functionality working onto ARTag landmarks [5], trackless AMRs can obtain the location of ARTag landmarks on 3D image (camera viewing) coordinate, specifically referring to the return value of "distance" and "pose" between the ARTag landmark on the docking destination and the camera sensor on the trackless AMR. The ARTag landmark is a marker system proposed by "ARToolKit" that is an open-source computer tracking library originally for creation of augmented reality applications on the real world, and recently for facility of trackless AMR applications on the automatic docking alignment missions. The ARTag landmark makes good use of a dual-tone planar pattern, and each pattern contains a unique ID number. Apart from QR code, the graphical simplicity of the ARTag makes itself detected at a farther distance and has itself favored in the PBVS field. Since the ARTag landmark has a simple square planar pattern and a black-and-white design with black tags and a white outer frame, the distance and pose of the detected corner features are relatively stable through the intersection points of the edge lines. Therefore, the open-source project *Ar_track_alvar* [6] is used to obtain the location of ARTag landmarks on 3D image (camera viewing) coordinate, as shown in Fig. 2. Under the known size of the ARTag landmark, the four corner features and four edges of the ARTag landmark can be used to estimate the distance and pose between the ARTag landmark on the docking destination and the camera sensor on the trackless AMR, through Homography (perspective) transformation.

Fig. 3 shows the overall flowchart of conventional trackless AMRs featuring with 4 operating modes of SLAM, navigation, PBVS, and dead reckoning. In general, the overall flowchart of conventional trackless AMRs, as shown in Fig. 3, can be divided 4 stages: 1) SLAM stage usually adopts Rao-Blackwellized Particle Filter (RBPF) [7] at the step of positioning estimation for SLAM and employs Occupancy Grid Mapping [8] at the step of map building. SLAM stage maps the scan of point cloud of 2D LiDAR sensor onto the real-world environment for construction of incremental map. 2) Navigation stage adopts Adaptive Monte Carlo Localization (AMCL) [9] at the step of positioning estimation for navigation, utilizes Dijkstra algorithm [10] at the step of path planning to obtain the global shortest or fastest path to the destination, and employs Timed Elastic Bands (TEB) [11] at the step of obstacle avoidance for local path planning of optimal collision avoidance trajectories. 3) Visual servoing stage involves both LiDAR-based visual servoing and pose-based visual servoing with ARTag landmarks. The 3D image (camera viewing) coordinate system built by the camera sensor (at origin) and ARTag landmark is used to estimate the distance and pose between the ARTag landmark on the docking destination and

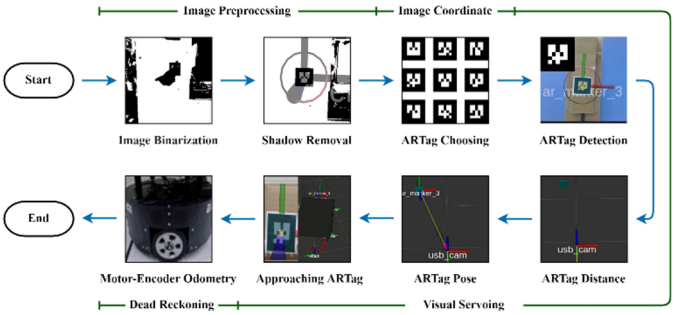


Fig. 4. Flowchart of conventional PBVS technology.



(a)

(b)

Fig. 5. Comparison between (a) high-quality image binarization and (b) poor-quality image binarization on ARTag landmark.

the camera sensor on the trackless AMR at the step of ARTag detection. Then, at the step of approaching ARTag, the trackless AMR is navigated by PBVS close to the ARTag landmark on the docking destination in the 3D image (camera viewing) coordinate system. 4) Dead reckoning stage leads the trackless AMR to the final destination inertially based on the wheel odometry [12], [13], according to the return value of "distance" and "pose" from the ARTag landmark at the last moment before the ARTag landmark gets too close to the camera sensor. After the ARTag landmark gets too close to the camera sensor, the ARTag landmark will get too large to be completely detected by the camera sensor, that is, ARTag landmark is out of view field of the camera sensor.

Among 4 stages of the overall flowchart of conventional trackless AMRs in Fig. 3, the accuracy, promptness, and steadiness of automatic docking alignment missions depends heavily upon the performance of PBVS at the stage of visual servoing. Fig. 4 shows the overall flowchart of conventional PBVS. It can be divided 4 stages: 1) Image preprocessing stage transforms the input image into a binary image at the step of image binarization. Then, shadow removal is performed on the image of ARTag landmarks to highlight the contours of ARTag landmarks for stable detection of the camera sensor at the step of shadow removal. 2) Image coordinate stage selects the ARTag landmark of the targeted ID of the automatic docking alignment and designates the location of the targeted ARTag landmark on 3D image (camera viewing) coordinate. 3) Visual servoing stage detects the targeted ARTag landmark, and estimates the distance and pose between the ARTag landmark on the docking destination and the camera sensor on the trackless AMR. Referring to the return value of "distance" and "pose" from the ARTag landmark, the trackless AMR is navigated by PBVS to approach the location of the targeted ARTag landmark on 3D image (camera viewing) coordinate. 4) Dead reckoning stage must be activated when the ARTag landmark becomes beyond the view field of the camera sensor, because the camera sensor on the trackless AMR approaches too close to the ARTag landmark on the docking destination. Finally, the automatic docking alignment of the trackless AMR to the final destination is achieved by the wheel odometry estimation and the inertial motion control.

Among 4 stages of the overall flowchart of conventional PBVS technology in Fig. 4, the accuracy, promptness, and steadiness of automatic docking alignment missions depends heavily upon the detected image quality at the stage of image preprocessing. The poor detected image quality sometimes causes the drift error on the return value of "distance" and "pose" from the ARTag landmark. It not only reduces the alignment precision of automatic docking alignment, but also decreases the docking efficiency (i.e., increases the docking iteration) of automatic docking alignment, even results in the failure or accident. Therefore, the step of image binarization at the stage of image preprocessing plays a crucial role in stabilizing the reading value of "distance" and "pose" between the ARTag landmark on the docking destination and the camera sensor on the trackless AMR. High-quality image binarization not only simplifies the background complexity but also diminishes the illumination interference, as shown in Fig. 5 (a). However, if shadow occlusion, image blurring, low contrast, uneven illumination, or complex backgrounds are happened with the ARTag landmark, Poor-quality or plain image binarization methods can not handle well with these difficult environmental conditions, probably fail to improve the clearness and contrast of the return value of "distance" and "pose" from the ARTag landmark, as shown in Fig. 5 (b).

II. Conventional Binarization Methods

In general, conventional and common image binarization methods can be divided into three main categories: 1) Image-processing-based binarization methods, including threshold

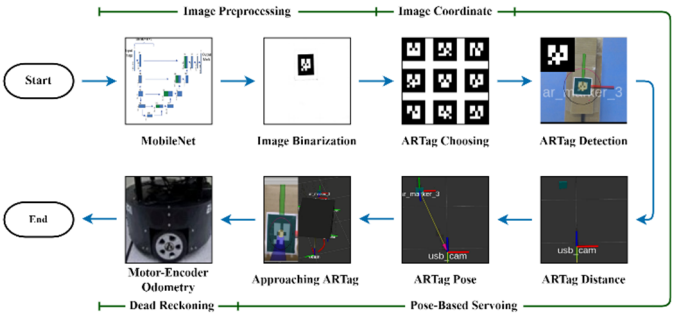


Fig. 6. Flowchart of proposed PBVS technology with lightweight deep-learning binarization.

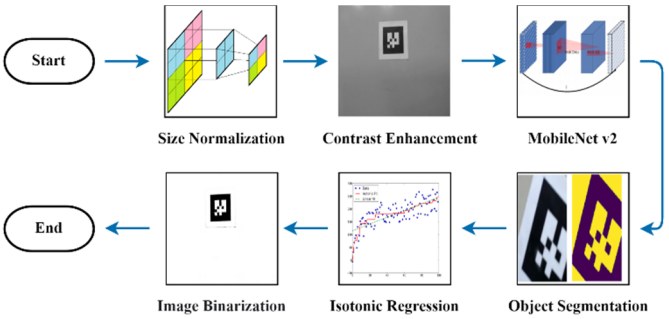


Fig. 7. Flowchart of proposed lightweight deep-learning binarization based on modified MobileNet v2.

from average of overall gray level [14] and dynamic threshold from maximum and minimum gray level [15]. 2) Computer-vision-based binarization methods, including threshold from gray-level histogram and Otsu optimization [16] and adaptive threshold from local histogram [17]. 3) Deep-learning-based binarization methods, including binarization with selectional auto-encoder (SAE) [18] and binarization with fully convolutional neural network [19].

However, image-processing-based binarization methods are not effective to generate proper binary images of ARTag landmarks under environmental conditions of shadow occlusion, image blurring, low contrast, uneven illumination, or complex background. Although image-processing-based binarization methods are efficient and straightforward in terms of computational complexity, they are not practical and feasible to real-world ARTag landmark detection issues. As for computer-vision-based binarization methods, they have better adaptability under environmental conditions of shadow occlusion, image blurring, low contrast, uneven illumination, or complex background. They can generate better binary images of ARTag landmarks while maintaining appropriate execution speed and computational efficiency. But, in some special real-world environmental conditions, they are still not good enough to completely eliminate adverse factors, especially resulting in accompanying noise in the binary images of ARTag landmarks. Finally, deep-learning-based binarization methods, as its name to make good use of deep learning techniques, exhibit high adaptability to various real-world environmental conditions and great capability to complex real-world environmental conditions. They can achieve the best binary images of ARTag landmarks in diverse situations. But, they require a large amount of computational time and it is contradictory for the stage of image preprocessing to consume too much computational time. In addition, deep-learning-based binarization methods may have poorer performance in binarizing low-resolution images of ARTag landmarks.

III. Proposed Lightweight Deep-Learning Binarization

This paper focuses on conventional deep-learning-based binarization methods and improves on its accuracy, promptness, and steadiness for ARTag landmark detection equivalently, against conventional image-processing-based binarization methods, conventional computer-vision-based binarization methods, and conventional deep-learning-based binarization methods. This paper develops a novel and practical method of lightweight deep-learning binarization at the stage of image preprocessing, as shown in Fig. 6, to enable PBVS to read the return value of “distance” and “pose” from the ARTag landmark exactly, quickly, and consistently.

In Fig. 6, at the stage of image preprocessing, a lightweight model modified from well-known MobileNet v2 [20] is proposed to address the issues of high computational requirements and poor binarization performance on low-resolution images in conventional deep-learning-based binarization methods. The detailed flowchart of proposed lightweight deep-learning binarization at the stage of image preprocessing of Fig. 6 is illustrated in Fig. 7.

In Fig. 7, the proposed lightweight deep-learning binarization can significantly reduce the architecture complexity of well-known MobileNet v2 network and further decrease the parameter amount of well-known MobileNet v2 network, without sacrifice of segmentation accuracy. Because the proposed lightweight deep-learning binarization simplifies depthwise separable convolution of MobileNet v2 network with adaptive kernel map to further reduce the redundancy in the convolutional kernel representation. Simplified depthwise separable convolution with adaptive kernel map during the inference process also can further reduce both memory consumption and computation time for predictions.

After the step of object segmentation, isotonic regression is also adopted to converge the training results, so higher

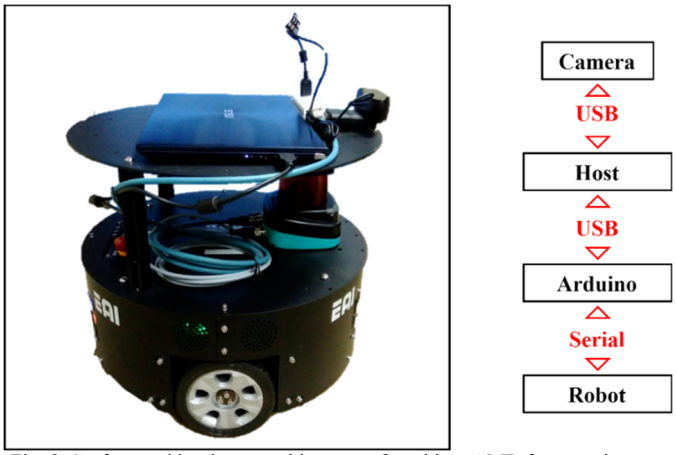
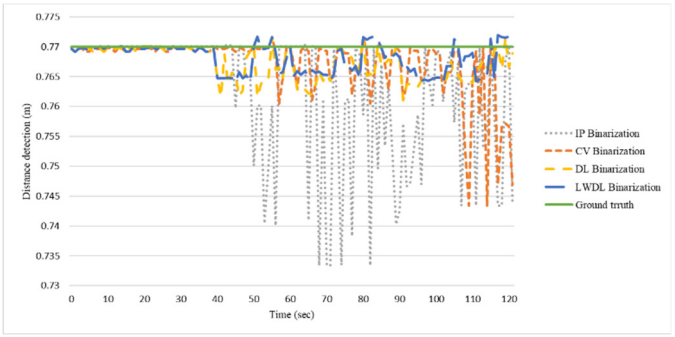


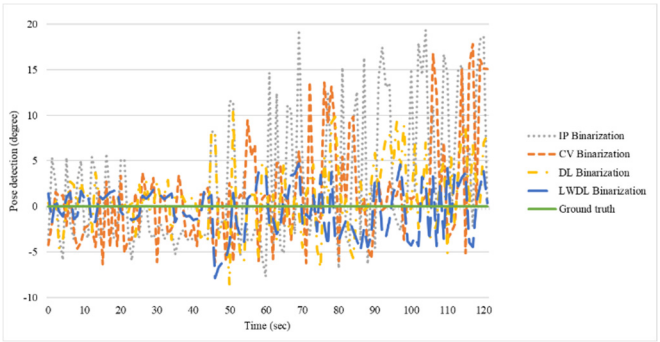
Fig. 8. Surface and hardware architecture of trackless AMR for experiments.

Table I. HW/SW specification of trackless AMR for experiments.

Hardware Specification	
Sensor	Logitech HD Pro Webcam C930e
Host	Intel Core i7-9750H 2.60GHz
Controller	Arduino Mega2560
Robot	EAI DashgoD1
Software Specification	
Operation System	Ubuntu v16.04
Computer Vision Library	OpenCV v3.1.0
ROS Framework	Kinetic
Deep Learning Framework	TensorFlow 2.5.0



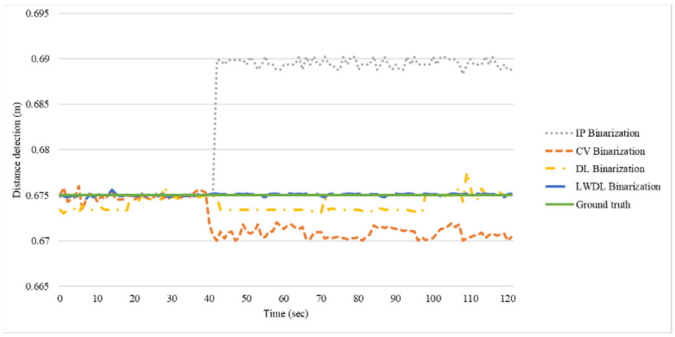
(a)



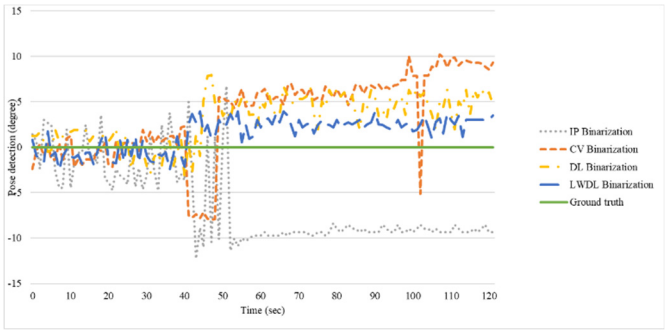
(b)

Fig. 9. (a) Distance detection and (b) pose detection comparison between 4 various binarization methods under condition of shadow occlusion.

accuracy and recall rates can be achieved. Isotonic regression algorithm minimizes the mean squared error of the training data while finding a non-decreasing approximation of the function.

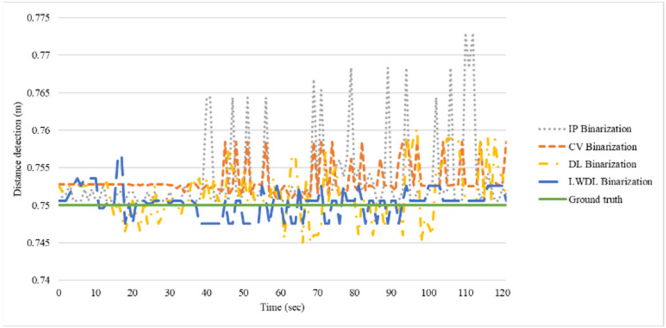


(a)

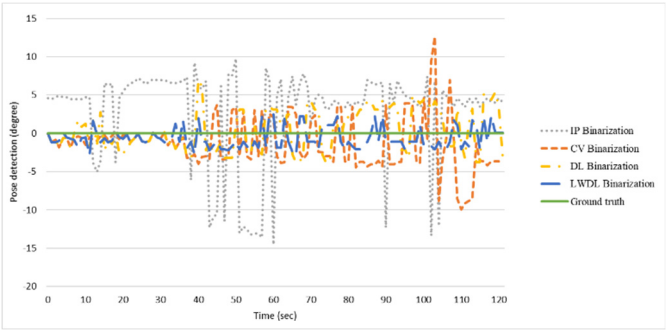


(b)

Fig. 10. (a) Distance detection and (b) pose detection comparison between 4 various binarization methods under condition of image blurring.



(a)

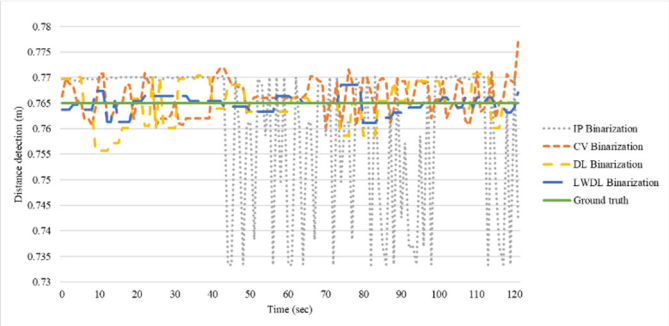


(b)

Fig. 11. (a) Distance detection and (b) pose detection comparison between 4 various binarization methods under condition of low contrast.

IV. Experimental Results

The appearance and hardware architecture of trackless AMR for experimental comparison is shown in Fig. 8 (a), and the



(a)

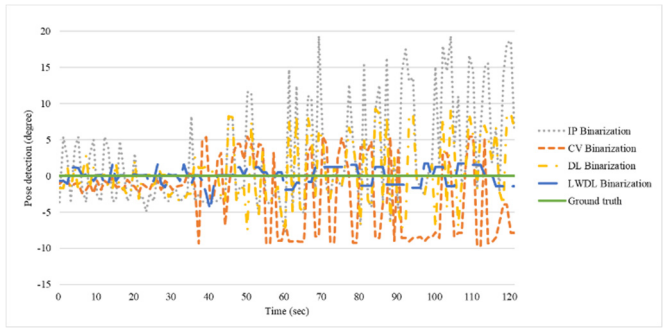
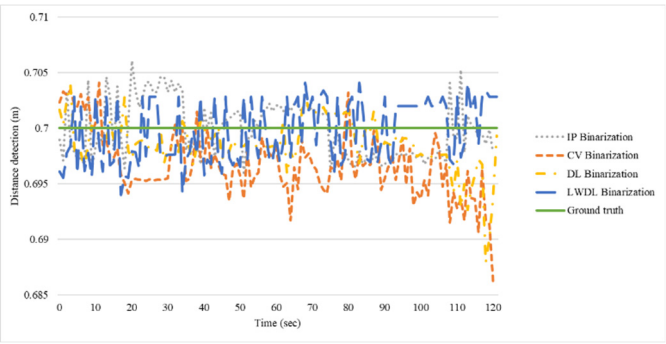


Fig. 12. (a) Distance detection and (b) pose detection comparison between 4 various binarization methods under condition of uneven illumination.



(a)

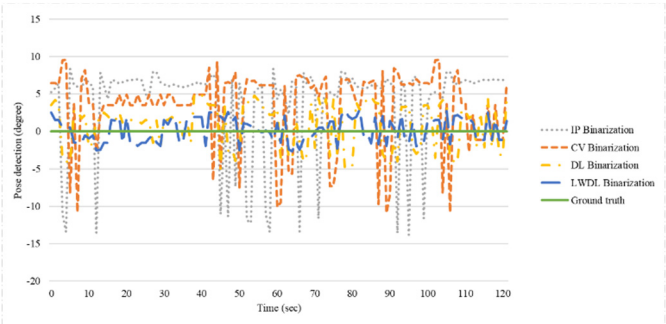


Fig. 13. (a) Distance detection and (b) pose detection comparison between 4 various binarization methods under condition of complex background.

HW/SW specification of trackless AMR for experimental comparison is listed in Table I. The experimental comparison between conventional image-processing-based binarization, conventional computer-vision-based binarization, conventional deep-learning-based binarization, and proposed lightweight deep-learning binarization, is performed under 5 various environmental conditions: 1) shadow occlusion, 2) image blurring, 3) low contrast, 4) uneven illumination, and 5) complex background. The experimental comparison makes use ARTag landmark with size dimensions of 5 centimeters. The working range between the ARTag landmark on the docking destination and the camera sensor on the trackless AMR is within 1.5 meters. Figs. 9-13 show the return value of "distance" and "pose" of the ARTag landmark detected by the camera sensor with 4 various aforementioned image binarization methods, under 5 various environmental conditions: : 1) shadow occlusion, 2) image blurring, 3) low contrast, 4) uneven illumination, and 5) complex background, respectively. The ground truth of "distance" and "pose" between the ARTag landmark on the docking destination and the camera sensor on the trackless AMR is measured by the open-source project *Ar_track_alvar* and determined by the stationary mean of the return value. From Figs. 9-13, it is verified that the accuracy and reliability of the return value of "distance" and "pose" of the ARTag landmark detected by the camera sensor with proposed lightweight deep-learning binarization is the best under 5 various environmental conditions.

V. Conclusions

The proposed lightweight deep-learning binarization not only raises the accuracy and reliability of ARTag's reading, but also apparently improves the effectiveness and efficiency of PBVS's operation. Consequently, PBVS becomes more precise and reliable. Future work will involve implementing the proposed lightweight deep-learning binarization onto numerous types of trackless AMRs for versatile indoor or outdoor automatic docking alignment missions.

Acknowledgment

This work was financially supported by the "Intelligent Recognition Industry Service Center" from The Featured Areas Research Center Program within the framework of Higher Education Sprout Project, Ministry of Education (MOE), Taiwan.

References

[1] K. Azadeh, R. Koster, and D. Roy. "Robotized and automated warehouse systems: Review and recent developments," *Transportation Science*, vol. 53, no. 4, pp. 917–945, Jun. 2019.

[2] S. G. M. Hossain, H. Jamil, M. Y. Ali, and M. Z. Haq, "Automated guided vehicles for industrial logistics - Development of intelligent prototypes using appropriate

technology," *Proceedings of International Conference on Computer and Automation Engineering (ICCAE)*, pp. 237–241, Apr. 2010.

[3] L. Zhao and Y. Ding, "Design of AGV Low-Delay Visual Servo System," *Proceedings of International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 6–10, Sep. 2018.

[4] A. Kelly, B. Nagy, D. Stager, and R. Unnikrishnan, "Field and service applications - An infrastructure-free automated guided vehicle based on computer vision - An Effort to Make an Industrial Robot Vehicle that Can Operate without Supporting Infrastructure," *IEEE Robotics & Automation Magazine*, vol. 14, no. 3, pp. 24–34, Sep. 2007.

[5] M. Fiala, "ARTag, a fiducial marker system using digital techniques," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 590–596, 2005.

[6] S. Niekum, "ar_track_alvar" *ROS Wiki*, [Online]. Available: http://www.ros.org/wiki/ar_track_alvar

[7] H. Wang, S. Wei, and Y. Chen, "An Improved Rao-Blackwellized Particle Filter for SLAM," *Proceedings of International Symposium on Intelligent Information Technology Application Workshops,* pp. 515–518, Dec. 2008.

[8] T. Collins, J. J. Collins, and D. Ryan, "Occupancy grid mapping: An empirical evaluation," *Proceedings of Mediterranean Conference on Control and Automation*, pp. 1–6, Jun. 2007.

[9] R. Hanten, et al., "Vector-AMCL: Vector based adaptive monte carlo localization for indoor maps," *Intelligent Autonomous Systems 14*, pp. 403–416, Feb. 2017.

[10] Y. Deng, et al., "Fuzzy Dijkstra algorithm for shortest path problem under uncertain environment," *Applied Soft Computing*, pp. 1231–1237, Mar. 2012.

[11] M. Keller, et al., "Planning of optimal collision avoidance trajectories with timed elastic bands," *IFAC Proceedings Volumes*, vol. 47, pp. 9822–9827, 2014.

[12] R. M. Nor, et al., "Mobile Robot Stable-Target Navigation Control via Encoder Data Feedback," *Proceedings of International Malaysia-Ireland Joint Symposium on Engineering, Science and Business*, Jun. 2012.

[13] K. Wang, Y. Liu, and L. Li, "Visual servoing trajectory tracking of nonholonomic mobile robots without direct position measurement," *IEEE Transactions on Robotics*, pp. 1026–1035, May 2014.

[14] S. S. Al-amri, N. V. Kalyankar, and S. D. Khamitkar, "Image Segmentation by Using Threshold Techniques," *Journal of Computing*, vol. 2, May 2010.

[15] C. Eyupoglu, "Implementation of Bernsen's Locally Adaptive Binarization Method for Gray Scale Images," *The Online Journal of Science and Technology*, vol. 7, pp. 68–72, Apr. 2017.

[16] D.-Y. Huang and C.-H. Wang, "Optimal multi-level thresholding using a two-stage Otsu optimization approach," *Pattern Recognition Letters*, vol. 30, pp. 275–284, Feb. 2009.

[17] S.-f. Zhou, et al., "An improved adaptive document image binarization method," *Proceedings of International Congress on Image and Signal Processing*, pp. 1–5, Oct. 2009.

[18] J. Calvo-Zaragoza and A.-J. Gallego "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, vol. 86, pp. 37–47, Feb. 2019.

[19] C. Tensmeyer and T. Martinez, "Document Image Binarization with Fully Convolutional Neural Networks," *Proceedings of IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 99–104, Nov. 2017.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, Jun. 2018.

Integration of Contextual Information with YOLOv5 for Detecting Tracheobronchial Foreign Objects

Chuan-Yu Chang[1,*], *Yu-Shan Peng* [1], *Chun-Liang Lai* [2], Roshana Mukhtar [3]
1, *Department of Computer Science and Information Engineering
National Yunlin University of Science and Technology, Douliu, Taiwan
E-mail: chuanyu@yuntech.edu.tw
[2]Division of Pulmonology and Critical Care, Department of Internal Medicine
Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Chiayi, Taiwan
E-mail: laicl.dalin@gmail.com
[3]Department of Computer Science and Information Engineering
National Yunlin University of Science and Technology, Douliu, Taiwan
E-mail: D11110209@yuntech.edu.tw

***Abstract* — Inhalation of foreign bodies in the airway is a rare event. Most cases of foreign body inhalation are eliminated through protective mechanisms of the body, such as coughing. However, there are some instances where the foreign body remains in the airway, even without the patient noticing. The symptoms of foreign body inhalation vary depending on the size and shape of the object. If it does not cause complete obstruction, it is often overlooked in image interpretation. In this study, the YOLOv5 is first used to locate the respiratory region in high-resolution computed tomography (HRCT), and this serves as the initial point for the 3D region growing. After identifying the airway using 3D region growing, the images are then analyzed using SEResNet18 to determine if any foreign bodies are present. The experimental results showed that using the primary YOLOv5's object detection method, an accuracy of 87.3% was achieved. After incorporating the SE channel attention mechanism into the YOLOv5 framework, the accuracy improved to 92.4%. In the foreign bodies classification, an accuracy of 85% was obtained.**

Keywords: High-resolution Computed Tomography, Object Detection, YOLO, SEResNet18

I. INTRODUCTION

When fish bones or thin bone fragments enter the airway, they most often get lodged in the narrower sections, especially at the bifurcation or carina of the bronchus. Due to their shape, it is difficult to cough them out. This can also cause physicians to mistake the fishbone or thin bone fragment as a ridge or prominence of the bronchus when observing high-resolution computed tomography images. There were many researchers focused on how to accurately segment the airway (including the trachea and finer bronchi) [1-3]. However, this paper focuses on detecting lesions within the airway, which still needs to be addressed. Therefore, this paper proposes a method that combines YOLOv5 [4] and 3D region growing locate airway region, and proposes a SEResNet18 for foreign object classification.

II. METHOD

Lesions in the airway may include secretions, foreign bodies, tumors, etc., which are referred to as respiratory foreign bodies in this paper. We use YOLOv5 combined with 3D region growing for airway detection and then using a SEResNet18 to detect airway foreign bodies. The system flowchart is shown in Figure 1. Table I shows the airway detection result with three different YOLOv5 backbones. The CSPNet achieved the highest accuracy of 87.3%. Figure 1 shows a case of foreign object detection result. The foreign objects classification results by SEResNet18 is 85%.

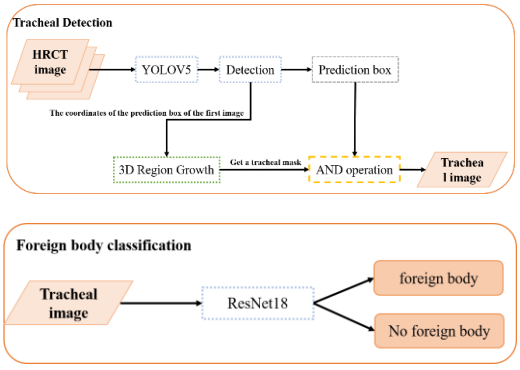
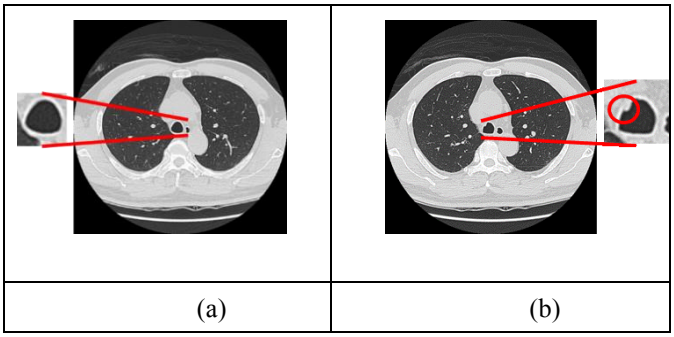


Fig. 1. The system flowchart.

TABLE I. AIRWAY DETECTION WITH DIFFERENT YOLO BACKBONES

Model	Backbone	mAP
YOLOv5	CSPNet	87.3%
YOLOv5	MobileNet V3	80.7%
YOLOv5	ConvNeXt	76.3%



(a) (b)

Fig. 2 foreign object classification result.

REFERENCES

[1] Antonio Garcia-Uceda, Raghavendra Selvan, Zaigham Saghir, Harm A. W. M. Tiddens and Marleen de Bruijne, "Automatic airway segmentation from computed tomography using robust and efficient 3-D convolutional neural networks," Scientific Reports, vol. 11, no. 1, pp. 1-15, Aug. 2021.

[2] Hoang Nhut Huynh1, My Duyen Nguyen, Thai Hong Truong, Quoc Tuan Nguyen Diep, Anh Tu Tran and Trung Nghia Tran, "Segmentation Trachea and Bronchial Branches in Chest Computed Tomography Image by Deep Learning-preliminary results," Kalpa Publications in Engineering, vol. 4, pp. 109-115, 2022.

[3] Wenjun Tan, Pan Liu, Xiaoshuo Li, Shaoxun Xu, Yufei Chen, Jinzhu Yang, "Segmentation of lung airways based on deep learning methods," IET Image Processing, vol. 16, no. 5, pp. 1444-1456, Apr. 2022.

[4] Glenn J, Alex S and Jirka B, ultralytics/yolov5:v6.0 - YOLOv5 Object Detection. (2021). GitHub. Retrieved [2023-03] from https://github.com/ultralytics/yolov5

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Real-Time Noise Suppression Using Harmonic/Percussive Separation with Morphological Operations for Hammering Test

Ryugo Uchiyama* and Nari Tanabe†
* Suwa University of Science, Nagano, Japan
E-mail: GH23504@ed.sus.ac.jp
† Suwa University of Science, Nagano, Japan
E-mail: nari@rs.sus.ac.jp

***Abstract*—This paper proposes a method to suppress noise in real-time using harmonic/percussive separation with morphological operations for hammering tests. The feature of the proposed method is that it focuses on the structure of the spectra of the hammering sound and noise obtained by analyzing the spectrogram of the signal observed in the hammering test and performs morphological operations using structuring elements suitable for each shape. This enables more appropriate extraction of the hammering sound than conventional methods. Although only the spectra of past time frames are used to achieve real-time noise suppression, the noise suppression performance is better than that of previous methods. The effectiveness of the proposed method is demonstrated by subjective and objective evaluations.**

I. Introduction

The walls of concrete structures such as bridges, highways, and tunnels are damaged daily by age-related deterioration, erosion caused by raindrops, and disasters such as earthquakes. The damage to walls varies from cracking, delamination, and hollowing. In particular, hammering test is used to inspect for internal damage such as delamination and hollowing, which are difficult to confirm visually, without destroying the object. In the hammering inspection method, the object is lightly struck with a hammer so as not to destroy it, and the inspector listens to the sound generated and examines abnormalities according to the difference in the sound.

In recent years, an accident occurred in Japan in which ceiling panels attached to ventilation ducts in tunnels have fallen [1]. This accident may have been caused by the fact that the damage to the wall surface had been confirmed visually and that the interior of the wall had not been inspected adequately by hammering inspection. In fact, there are quite a few deteriorations that cannot be detected simply by checking the concrete wall surface, and they are the cause of anchor bolt dropout and loosening of the ceiling panels. This incident led to a review of abnormality inspection methods of structural objects and reaffirmed the importance of hammering inspections. In this context, it is known that there are problems with the current hammering inspection method.

The first problem is the decrease in the volume of the hammering sound that accompanies the change in inspection hammers. In the conventional hammering test, inspectors lightly shook their wrists and struck the wall surface with an ordinary hammer. However, it is difficult to make accurate judgments because it is difficult inspect in high or narrow places, and the hammering sound that can be observed changes depending on the situation due to differences in inspectors, hammering speed, force, and surface of the structures. A new hammer was proposed to solve these problems. This new hammer generates sound by rolling a polygonal tangential surface while adjusting the angle and length, making it possible to efficiently observe hammering sound that are evenly spaced with even force in high or narrow locations. However, because of the specification of generating sound by rolling, there is a disadvantage in that the volume of the observed hammering sound is low.

The second problem is that environmental noise is superimposed on the hammering sound. In most cases, traffic control is imposed when conducting a hammering test, but it is impossible to close the road to traffic. This means that vehicles will be driving near the inspection site, so the sound of vehicles and wind noise will always be heard. In particular, inside the tunnel, the sound of ventilation and these sounds reverberate and become a complex noise, superimposed on the hammering sound whose volume was reduced in the first problem, making it extremely difficult to distinguish between the delicate sounds of the healthy and damaged areas.

To address these problems, it is necessary to extract the desired signal with noise suppression. Noise suppression methods include the spectral subtraction method, which suppresses noise by estimating it, and the method using Kalman filter algorithm with colored driving source [2]-[4]. These noise suppression methods are based on the assumption that the SNR is not that small, and that the power of the noise is lower than that of the hammering sound. Therefore, they are not suitable for wall inspections in low SNR environments such as inside tunnels. In addition, as the vehicle approaches the inspection position, the traveling sound causes a Doppler effect, and a sudden change in noise is observed. A noise suppression method that enables the desired hammering signal to be heard even in such a poor inspection environment is desirable.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

This paper proposes a noise suppression method suitable for hammering tests inside tunnels, especially those with low SNR. By using a filter that acts directly on the spectrogram, it is robust to changes in sound in the time and frequency directions and can suppress noise by taking into account the structural features of the hammering sound that appear in the spectrogram. It can also process the signal input from the microphone to the device in real time and output the desired signal to the earphone. In hammering test, the condition of the wall surface must be distinguished "on the spot" in order to accurately record abnormalities on the wall surface at the inspection site. Therefore, when developing a system to support hammering test in actual tunnels, the hammering sound generated at the inspection site must be noise suppressed in real time and delivered to the operator's ears without delay.

II. Previous Method

Noise suppression methods for hammering tests have been studied as follows. In [2], hammering tests were conducted on an actual bridge and the effect of the presence or absence of defects inside the walls on the hammering sound was discussed. In this study, it was confirmed that the use of Self Organizing Maps can solve the problem of difficulty in discriminating between healthy and damaged areas because the frequency spectrum of the observed signal varies depending on the scale and condition of the defect. This experiment assumes a hammering inspection of an actual bridge. However, because it does not take into account the condition of noise superimposition, it is expected that errors will occur in discriminating healthy and damage areas.

In contrast, reference [3] proposes noise suppression using a Kalman filter algorithm with colored driving source. In the method, the minimum statistics method [4] is used as a noise estimation method. The minimum statistics method can track noise spectrum smaller than the hammering spectrum. but it cannot estimate the noise spectrum well for noise spectrum larger than the hammering spectrum, implying its insufficient noise suppression. In [3], based on the results of frequency analysis of actual hammering and noise signals, the authors found that noise signals are distributed mostly in the low-frequency band and used a Butterworth type high-pass filter (BW-type HPF) to suppress spectrum of low frequency before estimating noise spectrum using the minimum statistics method. This means that the large noise spectrum distributed at low frequencies is made smaller than the hammering spectrum, so that the noise estimation by the minimum statistics works.

However, such digital filters require a predefined cutoff frequency. Since the frequency of noise differs depends on the inspection environment and the frequency of hammering sound depending on the material and condition, parameters must be adjusted for each inspection. In most case, the low-frequency component of hammering sound is included in the low-frequency band where much of the noise is distributed in places where there is a lot of peeling, such as the interior walls of tunnels. Therefore, if the frequency band where noise is distributed is simply suppressed, the sound quality of the hammering sound will deteriorate, making it difficult to distinguish between healthy and damaged areas.

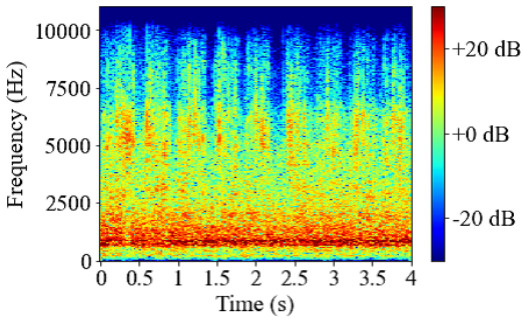


Fig. 1. Observed signal spectrum of the damaged area.

To address these problems, it is necessary to find "common features" that are independent of the test location. In this study, observing the spectrogram of the signal observed in a real environment brought attention to the spectral structure of the hammering signals and noise signals. Fig 1 shows that the spectral structure of the hammering signal in the spectrogram rises in the frequency axis direction at the moment (time) when the new hammer is rolled, while the spectral structure of the noise signal is continuous in the time axis direction. This structural feature is common and independent of the environment in general.

Such a method of separating each spectral structure has been widely studied in the field of source separation. As described in [5], source separation is an algorithm that separates the spectrogram into harmonic and percussive components, and is a technique used mainly for automatic musical notation and drum detection in musical components. Reference [6] uses a median filter to extract the harmonic and percussive components, and a Wiener filter to extract each component. On the other hand, reference [7] achieves better results in the extraction process of each component by applying a morphological filter [8], which is one of the nonlinear filters used in the field of image analysis processing. The method in [7] uses a property of mathematical morphology, i.e., the operation of extracting the characteristic structure of a figure in an image. This property is considered to be useful in sound source separation when the characteristic structure of the sound to be extracted can be represented as figures. As described in [9], there is a method of applying morphological filters to audio spectrograms, but it has been used as an adjunct to spectral subtraction methods.

In this study, the sound source separation technique used to extract specific sound source from a piece of music is considered to be able to extract the hammering signal by applying it to the observed signal of the hammering test. Since real-time implementation is being considered, it is desirable to achieve good accuracy with a relatively small amount of computation. Therefore, the reference [7] is modified to function in real time by adjusting it for hammering test, enabling noise suppression which retains some of the characteristics of the hammering

sound.

III. Mathematical morphology

In this section, we describe Morphology that is the fundamental principal of a Morphological filter used in our proposed method.

A. Basic morphological operations

Morphology is generally developed as a set theory in N-dimensional space, and in this section, we consider the case $N = 2$. Let A be a set in a 2-dimensional space and let a be a component of A, then the pixel value of A is 1 and the background pixel value is 0. The basic operation of morphology is Minkowski set theory, which consists of Minkowski sum and Minkowski difference, and using the set of structuring elements B and $b \in B$, the following equation is obtained.

$$A \oplus B = \bigcup_{b \in B} (A)_b, \tag{1}$$

$$A \ominus B = \bigcap_{b \in B} (A)_b, \tag{2}$$

where $(A)_b$ is the set obtained by moving the set A by b. If A^C and B^C are the complements of A and B, the following relation holds :

$$A \oplus B = (A^C \ominus B)^C, \tag{3}$$

$$A \ominus B = (A^C \oplus B)^C. \tag{4}$$

Thus, there is a duality between the Minkowski sum and the Minkowski difference.

Then applying the target set $B^S = -b : b \in B$ of B to (1) and (2), the following equation is given by

$$A \oplus B^S = \bigcup_{b \in B} (A)_{-b}, \tag{5}$$

$$A \ominus B^S = \bigcap_{b \in B} (A)_{-b}. \tag{6}$$

Equation (5) is dilation, (6) is erosion, which are basic operations in morphology. A morphological filter is defined by combining these basic operations

$$A \circ B = (A \ominus B^S) \oplus B, \tag{7}$$

$$A \bullet B = (A \oplus B^S) \ominus B, \tag{8}$$

where (7) is "Opening" and (8) is "Closing". "Opening" erodes the target set A and then dilates it, which has the effect of shaving off small protuberances on the outside of the target set A. "Closing" is process of dilating and then eroding the target set A, which has the effect of closing small holes or covers on the inside of the target set A.

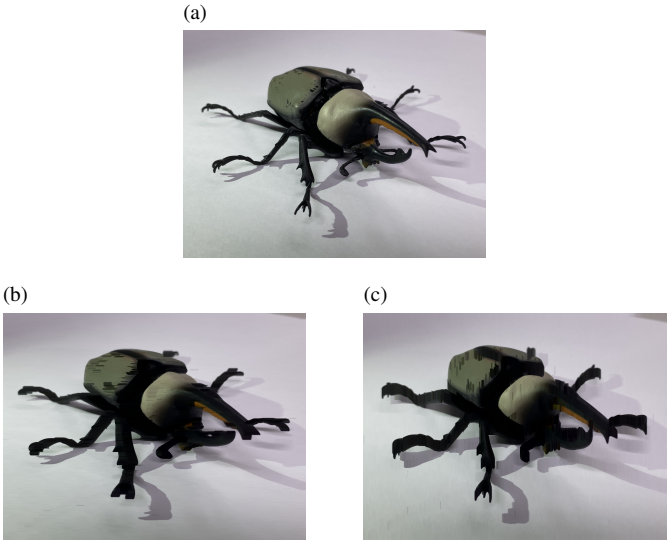


Fig. 2. (a) Original image, (b) Image erosion by a horizon line, (c) Image erosion by a vertical line.

B. Expansion to multi-valued functions

So far, the target to process has been discussed as a binary image. However, morphological operations can be extended to images with shading, i.e., functions that take various values.

First, give a two-dimensional vector $\boldsymbol{x}$ a third dimension y representing the pixel value $f(\boldsymbol{x})$, the following equation is given by

$$U(f) = \{(\boldsymbol{x}, y) : y \leq f(\boldsymbol{x})\}. \tag{9}$$

In this case, $U(f)$ is a shade. The shade can be calculated by the following equation :

$$f(\boldsymbol{x}) = \sup\{y \in \mathbb{R} : (\boldsymbol{x}, y) \in U(f)\}. \tag{10}$$

Then, since the shade is a binary image, the operations (5) and (6) are possible. The multi-valued Minkowski operation by the structuring function g of the function f is as follows :

$$(f \oplus g)(t) = \max_{\substack{t-u \in F \\ u \in G}} \{f(t-u) + g(u)\}, \tag{11}$$

$$(f \ominus g)(t) = \min_{\substack{t-u \in F \\ u \in G}} \{f(t-u) - g(u)\}, \tag{12}$$

where F and G are the domain of the functions f and g, respectively, and $-\infty$ is excluded from the domain. (11) and (12) are dilation and erosion in multi-valued functions, respectively. The detailed derivation of the morphological filter is described in [8].

Fig 2 shows a multi-valued image with erosion applied by two different structuring elements. Horizontal erosion can be confirmed by comparing (a) and (b), and vertical erosion can be confirmed by comparing (a) and (c). From this comparison, it can be seen that the shape of the structuring element greatly affects the distortion of the multi-valued function.

IV. Proposed Method

A. Spectral analysis and scaling

First, since the hammering sound is recorded as a stereo audio signal at the site, two-channel signals (left and right) are observed. Using the hammering signal $s_{L/R}(t)$ and the noise signal $n_{L/R}(t)$, the observed signal $x_{L/R}(t)$ at time t is expressed as follows :

$$x_{L/R}(t) = s_{L/R}(t) + n_{L/R}(t). \tag{13}$$

Next, this input signal is computed in the frequency domain. The observed signal $x_{L/R}(t)$ is divided into short time (frame) intervals, and its short-time Fourier transform (STFT) is calculated at each interval. By STFT, the observed signal spectrum $X_{L/R}(\lambda, k)$ is as follows :

$$X_{L/R}(\lambda, k) = \sum_{\mu=0}^{W-1} x_{L/R}(\lambda N + \mu) \cdot h(\mu) \cdot \exp\left(-j\frac{2\pi\mu k}{W}\right), \tag{14}$$

where λ is the frame number, k is the discrete Fourier transform (DFT) frame length ($W = 2048$), and N is the frame shift length ($N = W/2 = 1024$). In this case, a Hanning window is used as the window function, and the window functions overlap each other by half. Then, the observed signal spectrum $X_{L/R}(\lambda, k)$ can be expressed as follows using the hammering signal spectrum $S_{L/R}(\lambda, k)$ and the noise signal spectrum $N_{L/R}(\lambda, k)$:

$$X_{L/R}(\lambda, k) = S_{L/R}(\lambda, k) + N_{L/R}(\lambda, k). \tag{15}$$

It is inconvenient to apply a filter used for image processing with the obtained spectrum data as it is. Therefore, the obtained spectrum is converted to the logarithmic representation $\hat{X}_{L/R}(\lambda, k)$ by the following equation :

$$\hat{X}_{L/R}(\lambda, k) = 10\log(|X_{L/R}(\lambda, k)|^2). \tag{16}$$

Furthermore, based on the gray-scale spectrogram, expressing the log-display spectrum from -40 to $+40[dB]$ in 256 gradations from 0 to 255, the scaled spectrum $\tilde{X}_{L/R}(\lambda, k)$ is

$$\tilde{X}_{L/R}(\lambda, k) = \frac{255}{80}(\hat{X}_{L/R}(\lambda, k) + 40). \tag{17}$$

B. Harmonic/percussive component enhancement by morphological filter

In the spectrogram of the observed signal, since the features of the hammering component appear along the frequency axis (vertical direction), it is considered to be the percussive component, and since the features of the noise component are continuous along the time axis (horizontal direction), it is considered to be the harmonic component. Therefore, applying "Opening" to the spectrogram with a long vertical structuring element and a long horizontal structuring element, enhances each component.

Reference [7] prepared not only $n \times 1$ structuring elements but also several complex shapes, because the characteristic

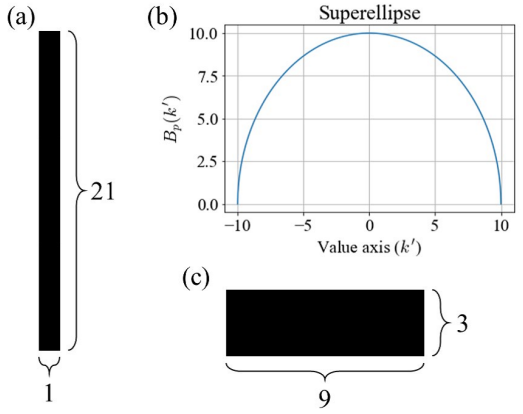


Fig. 3. (a) Shape of structuring element for masking percussive components, (b) Structuring function for masking percussive components, (c) Shape of structuring element for masking harmonic components.

percussive components appearing in the frequency direction are also spread a little horizontally at low frequencies. As a result, the separation performance was higher when $n \times 1$ structuring elements were used. However, it has been shown that the separation performance was lower when the structuring elements with the complex shapes are large. The hammering sound of the observed signal has characteristics that appear sparsely in the middle range, in an elliptical shape. To guarantee high separation accuracy even for signals with these characteristics, the proposed method considers the characteristics of noise components whose pitch varies with time (harmonic components), such as the Doppler effect due to moving noise sources.

Since the structuring element in the frequency axis direction is a single variable function of the frequency bin number k, the following superelliptic equation was used for the structure function $B_p(k')$ to better enhance the characteristics of the hammering sound.

$$B_p(k') = \begin{cases} b\left(1 - \left|\frac{k'}{a}\right|^p\right)^{\frac{1}{p}} & (-a \le k' \le a) \\ -\infty & (otherwise) \end{cases}. \tag{18}$$

The shape of the structuring function of the super-ellipse is determined by three parameters : the value axis b, the window width $2a$, and p. The horizontal structuring element is set to $k' = 0$, making it a flat structuring function. The shapes of the structuring elements are shown in Fig 3.

"Opening" is applied to the scaled spectrum $\tilde{X}_{L/R}(\lambda, k)$ for one frame by applying a structuring element for masking percussive components to enhance the vertical percussive component :

$$\tilde{P}_{L/R}(\lambda, k) = \tilde{X}_{L/R}(\lambda, k) \circ B_p(k'), \tag{19}$$

where $\tilde{P}_{L/R}(\lambda, k)$ is a scaled spectrum that enhances the percussive component.

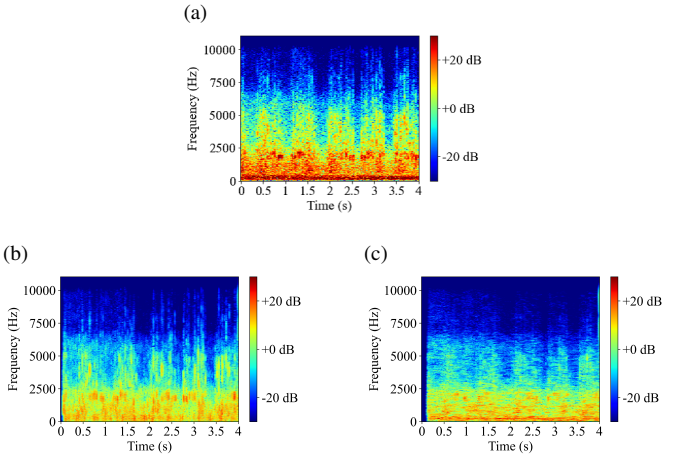


Fig. 4. (a) Observed signal spectrogram, (b) Spectrogram with enhanced percussive components, (c) Spectrogram with enhanced harmonic components.

Next, to enhance the harmonic component, scaled spectra for the past few frames are stored and "Opening" is applied with structuring elements for masking harmonic components.

$$\tilde{H}_{L/R}(\lambda, k) = \tilde{X}_{L/R}(\lambda, k) \circ B_h(\lambda', k'), \tag{20}$$

where $\tilde{H}_{L/R}(\lambda, k)$ is the scaled spectrum enhanced harmonic component and $B_h(\lambda', k')$ is the horizontal structuring element.

Each enhanced component $\tilde{P}_{L/R}(\lambda, k)$, $\tilde{H}_{L/R}(\lambda, k)$ can be output as a spectrogram, as shown in Fig 4.

C. Extraction of hammering components by Wiener filter

Next, a Wiener filter for masking the percussive component is applied to the enhanced scaled spectrum to calculate the percussive component. Before that, the scaled spectra $\tilde{P}_{L/R}(\lambda, k)$ and $\tilde{H}_{L/R}(\lambda, k)$ with each component highlighted are converted to Spectral values using (16) and (17). Then, according to [10], the Spectrogram can be considered to contain a residual component in addition to the harmonic and percussive components. If there is a sound that has the characteristics of both harmonic and percussive components, the harmonic and percussive components can be clearly separated by assigning the residual component to this sound. Therefore, the Wiener filter for takes the residual component into account is given by

$$M^P_{L/R}(\lambda, k) = \frac{\left(P_{L/R}(\lambda, k)\right)^2}{\left(P_{L/R}(\lambda, k)\right)^2 + \alpha \left(H_{L/R}(\lambda, k)\right)^2}, \tag{21}$$

where $H_{L/R}(\lambda, k)$ and $P_{L/R}(\lambda, k)$ are the harmonic and percussive component enhanced spectrum. The range of α is $1.0 \leq \alpha \leq 10.0$, and the larger the value, the larger the residual component. In this method, $\alpha = 2.0$ was used to take into account the existence of the residual component. This value was set experimentally. The separated percussive component $\tilde{S}_{L/R}(\lambda, k)$ is calculated by the following :

$$\tilde{S}_{L/R}(\lambda, k) = M^P_{L/R}(\lambda, k) \cdot X_{L/R}(\lambda, k). \tag{22}$$

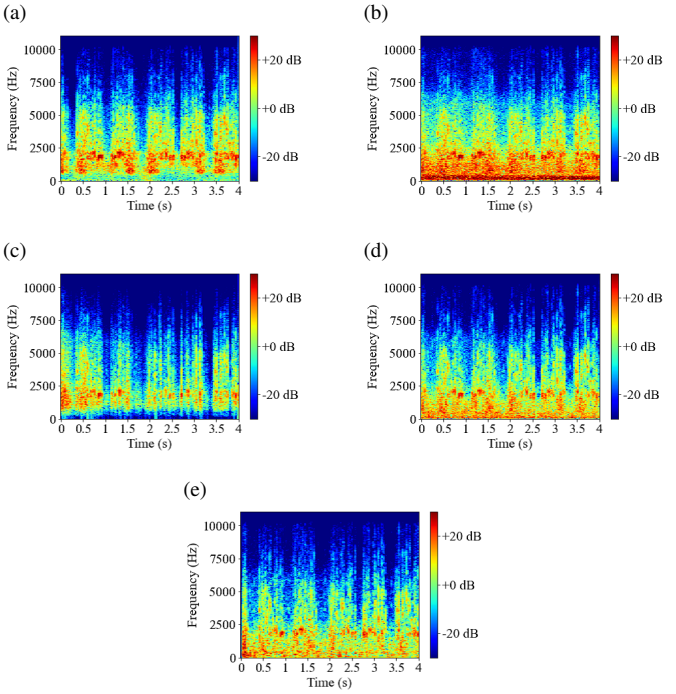


Fig. 5. (a) Spectrogram of clean hammering signal, (b) Spectrogram of observed signal, (c) Noise suppressed spectrogram by [3], (d) Noise suppressed spectrogram by [6], (e) Noise suppressed spectrogram by proposed method.

Finally, the separated hammering signal can be obtained by inverse short-time Fourier transform (ISTFT) as shown below :

$$\tilde{s}_{L/R}(t) = ISTFT\left(\tilde{S}_{L/R}(\lambda, k) e^{i\angle\theta}\right), \tag{23}$$

where θ is the phase spectrum of $X_{L/R}(\lambda, k)$.

V. Simulations

In this study, the observed signals used in the simulation are based on sounds collected at the actual site of a hammering test. For example, there is a method of testing by using the sound of a wall (mock-up specimen) that has a cavity inside to simulate a deteriorated condition. Using this method, it is possible to obtain a clean, noise-free hammering sound. However, this sound is noise-free, therefore making the hammering sound very different from the actual noisy environment. By using the actual sound as much as possible, the behavior of the noise suppressor can be confirmed to be similar to that of the on-site inspection. Thus, the observed signals used in this study were recorded in a relatively quieter environment than the tunnel, and then the loud noises in the tunnel were added. In detail, stereo audio signals with a sampling frequency f_s of 44100 Hz and SNRs of -5.62 dB and -4.79 dB for the left and right channels, respectively, were used.

A. Subjective evaluation

The spectrograms of the noise suppressed signals using the [3] and [6] are shown in Fig 5. First, a comparison of (a) and (b) shows that continuous noise is added in the low frequency time axis direction. It can also be seen that

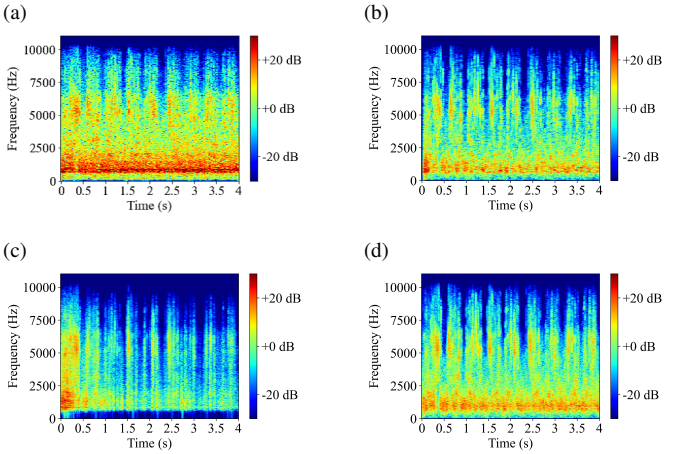


Fig. 6. (a) Spectrogram of actual observed signal, (b) Noise suppressed spectrogram by proposed method, (c) Noise suppressed spectrogram by [3], (d) Noise suppressed spectrogram by [6].

non-integer harmonic of the noise and reflected noise are spread to higher frequencies. The spectrogram (c) of [3] suppresses the hammering sound too much and sounds lighter. The spectrogram (d) of [6] retains the characteristics of the hammering sound well, but the harmonic components are not well extracted, and it is clear that the noise is not fully suppressed. On the other hand, the spectrogram of the proposed method, (e), shows that the noise component is suppressed better than spectrograms of [3] and [6]. Furthermore, the low frequencies are not completely suppressed but are moderately preserved, and the low pitched sounds of hammering sound are also successfully preserved. These results indicate that the proposed method is capable of capturing the structural features of sounds appearing in the spectrogram.

In addition, Fig 6 shows a similar experiment on an observation signal recorded during an actual tunnel hammering test, in which noise was convolved with the hammering sound. The spectrogram (c) of [3] shows that the sound is too much suppressed even in the actual observation signal. The spectrogram (d) of [6] shows that the sound is uniformly suppressed, so that noise remains and the hammering sound is muffled. The spectrogram (b) of the proposed method is able to suppress noise while preserving the structure of low and mid-frequency hammering sounds better than the conventional method. These results show that the proposed method works well in the actual tunnel hammering test and can extract features better than the conventional method.

B. Objective evaluation

In this section, we evaluate the performance using specific indices. First, we used three indices proposed in [11]. SI-SDR (scale-invariant signal-to-distortion ratio) is the ratio of the estimated separated sound to the residual noise, SI-SIR (scale-invariant source to interference ratio) is the ratio of the separated sound to the source-derived interference noise, and SI-SAR (scale-invariant source to artifact ratio) is defined as the ratio of the separated sound to the algorithm-derived noise.

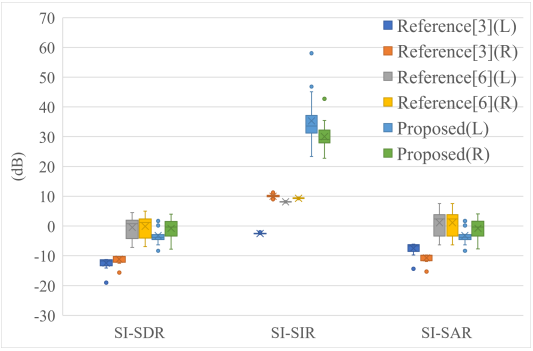


Fig. 7. SDR / SIR / SAR comparison.

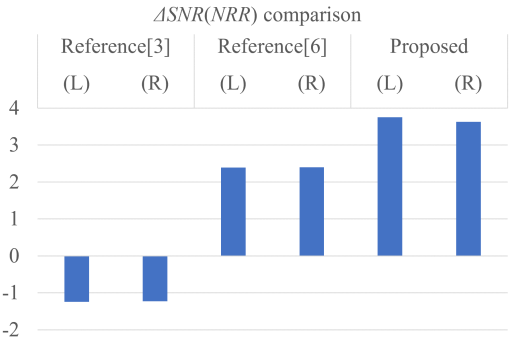


Fig. 8. Δ SNR comparison.

The larger the value of these indices, the better the performance. By comparing SI-SIR and SI-SAR, it is possible to determine whether source-derived noise or algorithm-derived noise dominates.

As shown in the box-and-whisker plot in Fig 7, the SI-SIR of the proposed method is extremely higher than that of any of the other methods. This indicates that the proposed method has less source-derived noise than any of the methods. Overall, the proposed method outperforms [3] in all indices, and its SI-SDR and SI-SAR are as good as [6].

Next, the SNR improvement rates (Δ SNR) were compared and evaluated. This value is expressed as the difference between the output signal SNR, and the input signal SNR, and is equivalent to the noise reduction ratio (NRR). In other words, the degree of noise suppression can be evaluated by Δ SNR and the larger the value, the better the performance. Since this index indicates noise reduction ratio, it can be compared and evaluated with [3] and [6].

The results are shown in Fig 8. As the graphs show, the proposed method performs good noise suppression.

As a general comment on the objective evaluation, the proposed method distorts the extracted hammering sound, but its high accuracy in classifying harmonic and percussive components indicates that it has good noise suppression accuracy.

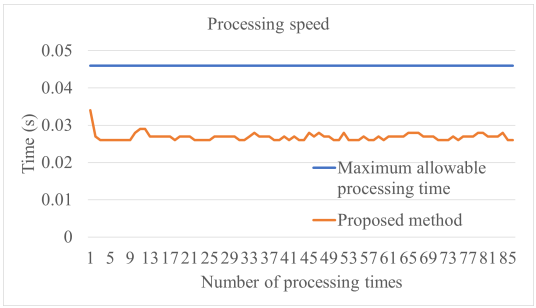


Fig. 9. Processing speed evaluation.

C. Processing speed evaluation

Finally, the processing speed is evaluated to confirm the real-time performance of the proposed method. Fig 9 shows the processing time per frame. The time per frame τ can be obtained by W/f_s, where $W = 2048$ and $f_s = 44100$ Hz, $\tau = 0.046$ s. To extract the desired signal in real-time, the processing of one DFT frame must be completed within the time of one frame. As shown in Fig 7, the processing time of the proposed method is below the maximum allowable processing time, indicating that real-time processing is possible.

VI. Conclusions

This paper proposed a real-time noise suppression method using harmonic/Percussive separation by morphological operations for hammering test. The main two features of the proposed method are as follows. First is that the "Opening" process is applied vertically and horizontally according to the distinctive features in the structures of the spectrograms of hammering sound and noise to enhance the hammering and noise components. Second is the masking of the hammering spectrum with a Wiener filter that considers a residual component, resulting in a high separation performance. Also, the effectiveness of the proposed method was demonstrated by comparing (i) spectrograms after noise suppression in a subjective evaluation, (ii) performance with specific numerical values in an objective evaluation, and (iii) processing speeds. Through the evaluation (iii), the proposed method guarantees real-time performance.

The proposed method does not aim at off-line noise suppression, such as estimating hitting sound by solving machine learning algorithms or mathematical optimization. Since the purpose of the proposed method is to provide the desired hammering sound to the inspector in real time at a hammering inspection site, it is considered to be an effective method when the desired signal must be extracted with high accuracy at low cost from an observed signal whose desired signal is degraded by noise or other factors.

One of the future works include developing a noise suppression system that can extract buried sounds (evidence of a damaged part) in highly noisy environments, such as tunnels, without sacrificing the quality of the sound.

References

[1] Research and Review Committee, "Report of the investigation and review committee on the falling tunnel ceiling place accident," 2013.
[2] Y. Sonoda, K. Kawabata, M. Beppu, Y. Fukui, "A fundamental study on the evaluation method of inner defect of concrete using hammering test data," *Structural engineering papers,* Vol.57A, 2011.
[3] K. Hori, N. Tanabe, M. Fujisawa, "Nonlinear SVM-Type automatic decision algorithm in noisy environment for hammering test system," *APSIPA ASC. Tokyo. Japan,* December. 2021.
[4] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Processing Conf.,* pp. 1182-1185, 1994.
[5] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of the EUSIPCO 2008 European Signal Processing Conference.,* Aug 2008.
[6] D. Fitzgerald, "Harmonic/Percussive separation using median filtering," *Proc. of the 13$^{\text{th}}$ Int. Conference on Digital Audio Effects (DAFx-10),* Sept 2010.
[7] A. Gkiokas, V. Papavassiliou, V. Katsouros, G. Carayannis, "Deploying nonlinear image filters to spectrogram for harmonic/percussive separation," *Proc. of the 15$^{\text{th}}$ Int. Conference on Digital Audio Effects (DAFx-12),* Sept 2012.
[8] J. Serra, "Image analysis and mathematical morphology," *Academic Press,* 1982.
[9] N. W. D. Evans, J. S. Manson, M. J. Roach, "Noise compensation using spectrogram Morphological Filtering," *Proc. 4th IASTED International Conf. on Signal and Image Processing,* pp.157-161, 2002.
[10] J. Driedger, M. Müller, S. Disch, "Extending harmonic-percussive separation of audio signals," *15th International Society for Music Information Retrieval Conference (ISMIR 2014),* Taipei, Taiwan, 2014.
[11] J. L. Roux, S. Wisdom, H. Erdogan, J. R. Hershey, "SDR - half-baked or well done?," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on.* IEEE, 2019, pp.626-630.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

$\Delta\Sigma$ Modulators for Discrete-time Closed Loop Control Systems with Quantization and Saturation

Shuichi Ohno*, Shenjian Wang* and Kiyotsugu Takaba †
Osaka Metropolitan University, Sugimoto 3-3-138, Sumiyoshi-ku, Osaka, 558-8585, JAPAN
* E-mail: ohno@omu.ac.jp /sn22857m@st.omu.ac.jp
† Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, JAPAN
E-mail: ktakaba@fc.ritsumei.ac.jp

***Abstract*—$\Delta\Sigma$ modulators have been utilized for analog-to-digital converters (ADCs). The performance of a $\Delta\Sigma$ modulator can be enhanced if it is designed by taking into account the characteristics of the system that uses the $\Delta\Sigma$ modulator. This paper considers $\Delta\Sigma$ modulators for discrete-time closed loop control systems. In the control system, quantization may degrade control performance and leads to oscillations of the system outputs, whereas saturation of the control input may cause windup phenomena such as overshoots of the system outputs and instability of the control system. This paper jointly mitigates the effects of quantization and saturation in a discrete-time closed loop control system, by minimizing the norm of the quantization error at the system output under a stability condition on the closed-loop system. Numerical examples are provided to see the effectiveness of our proposed design.**

I. Introduction

Digital communication technologies have been developed, and many devices have been connected to the Internet. To employ digital communication, analog signals have to be quantized into digital signals.

$\Delta\Sigma$ modulators have been utilized as efficient analog-to-digital converters (ADCs). (See, e.g., [1] and references therein). A $\Delta\Sigma$ modulator consists of a conventional (usually uniform) quantizer and a filter that feeds back the quantization error of the conventional quantizer to the input to the conventional quantizer. The feedback filter can be optimized to minimize the effect of the quantization by exploiting the system's characteristics [2], [3]. In this paper, we design a $\Delta\Sigma$ modulator for a discrete-time closed loop control system with quantization and saturation.

For a control system, if ideal actuators capable of realizing any desired controller outputs are available, one could design any kind of controller. However, in practice, there are always limitations on the actuator outputs, and the acceptable input range to the actuator is restricted. If the control input exceeds this range, input saturation is occurred. Therefore, for a system with input saturation, careful input design becomes necessary. Neglecting input saturation when designing a controller can lead to severe deterioration of control performance, such as overshoots of system outputs or instability in closed-loop systems, known as windup phenomena.

Input saturation can be modeled as a saturator followed by the controller. In most anti-windup controllers, the input and output of the saturator are fed back to the controller to mitigate the windup phenomena. (See, e.g., [4] and references therein.) [5] utilize a linear condition to design anti-windup controllers.

To reduce the effect of quantization errors on control systems, *dynamic quantizers* have been employed, where the quantizer parameters are designed using linear programming (LP) [6] or convex optimization [7]. However, most of these methods do not consider stability in the presence of input saturation. This paper proposes a method to jointly mitigate the effects of quantization and saturation.

We propose a design method of the $\Delta\Sigma$ modulator to reduce the l_2 norm of the quantization error at the system output under a stability condition on the closed-loop system. The l_2 norm is evaluated by the H_∞ norm of the transfer function from the quantization error to the system output, whereas the stability condition is described by the circle criterion obtained by [8]. Both conditions are expressed as bilinear matrix inequalities (BMI) that involve multiplications of design variables. Allowing to share a common Lyapunov matrix, our design problem is converted into a convex optimization, which can be solved numerically. The effectiveness of proposed method is verified through numerical examples.

II. Control Systems with Input Quantization and Saturation

Let us consider a discretized feedback system as shown in Fig. 1. For the sake of simplicity, we assume that the plant and the controller are single-input and single-output linear time-invariant systems, whose transfer functions are denoted as $P[\mathrm{z}]$ and $C[\mathrm{z}]$, respectively. The orders of $P[\mathrm{z}]$ and $C[\mathrm{z}]$ are denoted as n_p and n_c. We express the z-transform of a causal discrete-time signal denoted by a lowercase letter as its corresponding uppercase letter.

In Fig. 1, $u_r(k) \in \mathbb{R}$ is the reference signal at time k, $e(k) \in \mathbb{R}$ is the control deviation, $y_c(k) \in \mathbb{R}$ is the output of the controller, $u(k) \in \mathbb{R}$ is the input of a plant, and $y(k) \in \mathbb{R}$ is the control output.

We express the state space representations of the plant and the controller as

$$x_p(k+1) = A_p x_p(k) + B_p u(k) \tag{1}$$
$$y(k) = C_p x_p(k) \tag{2}$$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

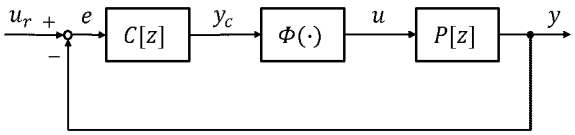


Fig. 1. Feedback control system with a constraint on the range of the control input.

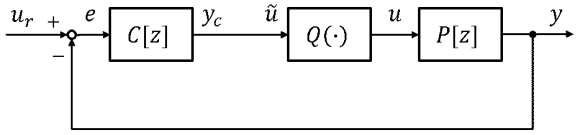


Fig. 2. Feedback control system with quantization and saturation.

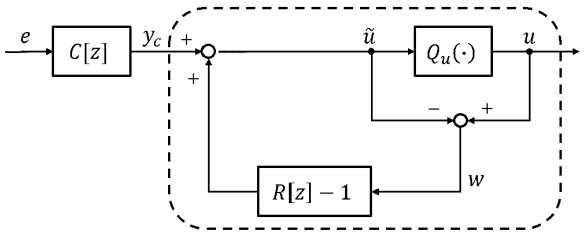


Fig. 3. $\Delta\Sigma$ modulator.

and

$$x_c(k+1) = A_c x_c(k) + B_c e(k) \tag{3}$$
$$y_c(k) = C_c x_c(k) \tag{4}$$

with $A_p \in \mathbb{R}^{n_p \times n_p}$, $B_p \in \mathbb{R}^{n_c \times 1}$, $C_p \in \mathbb{R}^{1 \times n_c}$, $A_c \in \mathbb{R}^{n_c \times n_c}$, $B_c \in \mathbb{R}^{n_c \times 1}$, and $C_c \in \mathbb{R}^{1 \times n_c}$, where $x_p(k) \in \mathbb{R}^{n_p}$ and $x_c(k) \in \mathbb{R}^{n_c}$ are state vectors of the plant and the controller.

For a real number $L > 0$, the range of the input of the plant is assumed to be limited between $[-L, L]$. (The range is assumed to be symmetric around 0 just for simplicity of presentation.) When the output of the controller exceeds the range of the control input, that is, $|y_c(k)| > L$, the input $u(k)$ to the plant is saturated, which is described by using the saturation element $\Phi(\cdot)$ in Fig. 1 defined as

$$\Phi(y_c(k)) = \begin{cases} L\text{sgn}(y_c(k)), & |y_c(k)| > L \\ y_c(k), & |y_c(k)| \le L \end{cases} \tag{5}$$

where sgn$(\cdot)$ is the sign function.

The difference between the output of the controller and the input to the plant is expressed as

$$w_a(k) := u(k) - y_c(k). \tag{6}$$

The value of $w_a(k)$ is non-zero only when $|y_c(k)| > L$. If $w_a(k)$ remains large for a certain period of time, the control deviation increases because the feedback is not functioning properly due to saturation. Consequently, the signals in the controller with integrators become large, leading to phenomena such as overshoot of the control output and system instability.

Now, let's consider that the controller output is transmitted to the plant through a digital communication channel. When transmitting an analog signal through a digital communication channel, the analog signal needs to be converted into a digital signal through quantization. If the channel has limited capacity, even the digital signal may be quantized into a low-resolution form.

Let the saturation levels of the quantizer be the same as the saturation levels of the plant. Thus, the feedback control system with input saturation and quantization can be illustrated as shown in Fig. 2, where $Q(\cdot)$ is the quantizer that satisfies

$$Q(x) = L\text{sgn}(x), \quad |x| > L \tag{7}$$

for a scalar input x.

To mitigate the quantization error and address the windup phenomena, we employ a $\Delta\Sigma$ modulator for our quantizer. The block diagram of a $\Delta\Sigma$ modulator is depicted inside the broken line in Fig. 3. Here, $w(k)$ represents the quantization error at time k, defined as

$$w(k) = u(k) - \tilde{u}(k) \tag{8}$$

and $R[\text{z}]$ is an n_r-order filter, known as the noise transfer function (NTF) or noise shaping filter (NSF), satisfying $R[\infty] = 1$. Note that the first term of the impulse response of $R[\text{z}] - 1$ is zero, ensuring there is no algebraic loop in the $\Delta\Sigma$ modulator. Moreover, $Q_u(\cdot)$ represents a static uniform quantizer. It is worth noting that when $R[\text{z}] = 1$, there is no feedback, and the $\Delta\Sigma$ modulator simplifies to a static quantizer.

III. $\Delta\Sigma$ MODULATOR FOR QUANTIZATION AND SATURATION

Let us design a $\Delta\Sigma$ modulator to reduce the effects of the quantization error under a stability condition to mitigate the windup phenomena.

The transfer function from the quantization error of the $\Delta\Sigma$ modulator to the system output is given by $P[\text{z}]/(1+P[\text{z}]C[\text{z}])$. Thus, the z-transform $Y[\text{z}]$ of the system output can be expressed by

$$Y[\text{z}] = \frac{P[\text{z}]C[\text{z}]}{1 + P[\text{z}]C[\text{z}]} U_r[\text{z}] + \frac{P[\text{z}]R[\text{z}]}{1 + P[\text{z}]C[\text{z}]} W[\text{z}]. \tag{9}$$

Here, $U_r[\text{z}]$ and $W[\text{z}]$ represent the z-transforms of the reference signal $u_r(k)$ and the quantization error $w(k)$ of the static quantizer, respectively.

The first term on the right-hand side of (9) corresponds to the system output without quantization, while the second term represents the error signal at the system output, which needs to be minimized.

Our design variable is the NTF $R[\text{z}]$ of the $\Delta\Sigma$ modulator, or equivalently, the feedback filter $R[\text{z}] - 1$. Let us express the state space representation of the filter $R[\text{z}] - 1$ as

$$x_r(k+1) = A_r x_r(k) + B_r w(k) \tag{10}$$
$$y_r(k) = C_r x_r(k) \tag{11}$$

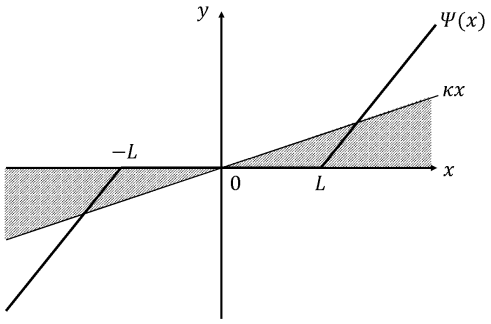


Fig. 4. Dead zone element and sector area.

with $x_r(k) \in \mathbb{R}^{n_r}$ the state vector, $A_r \in \mathbb{R}^{n_r \times n_r}$, $B_r \in \mathbb{R}^{n_r \times 1}$, and $C_r \in \mathbb{R}^{1 \times n_r}$.

A. Mitigation of Quantization effects

In equation (9), the transfer function from the quantization error $w(k)$ of the static quantizer to the system output $y(k)$ is denoted as $G_{yw}[\mathrm{z}]$, given by

$$G_{yw}[\mathrm{z}] = \frac{P[\mathrm{z}]R[\mathrm{z}]}{1 + P[\mathrm{z}]C[\mathrm{z}]}. \tag{12}$$

To mitigate the effect of the quantization error $w(k)$ on the system output, we impose a constraint on the norm of $G_{yw}[\mathrm{z}]$. Specifically, we consider the performance condition:

$$\|G_{yw}[\mathrm{z}]\|_\infty < \epsilon \tag{13}$$

where $\|G_{yw}[\mathrm{z}]\|_\infty$ is the H_∞ norm of $G_{yw}[\mathrm{z}]$ defined as

$$\|G_{yw}[\mathrm{z}]\|_\infty = \max_{|\omega| \le \pi} |G_{yw}[e^{j\omega}]| \tag{14}$$

and ϵ is a positive parameter. It should be noted that alternative norms, such as the H_2 norm, can be used to achieve similar results as discussed below.

From the definition of the augmented state equation, the system output can be expresses as

$$y(k) = C_{yw}x(k) \tag{15}$$

with $C_{yw} = [C_c, 0, 0]$.

It is known that (13) is satisfied if and only if there exists a positive definite matrix of $P_1 \succ 0$ that satisfies

$$\begin{bmatrix} P_1 & AP_1 & B & \mathbf{0} \\ P_1A^T & P_1 & \mathbf{0} & P_1C_{yw}^T \\ B^T & \mathbf{0} & I_n & \mathbf{0} \\ \mathbf{0} & C_{yw}P_1 & \mathbf{0} & \epsilon \end{bmatrix} \succ 0 \tag{16}$$

with $n = n_c + n_p + n_r$.

B. Condition for Stability

We utilize a traditional condition for the stability of a closed-loop non-linear systems. For simplicity, we temporarily assume that there is no quantization error in the range $|\tilde{u}(k)| \le L$ in this subsection.

We introduce a dead zone element $\Psi(\cdot)$ defined as

$$\Psi(x) = x - Q_u(x). \tag{17}$$

Let us denote the output of the dead zone element as $\tilde{w}(k) = -w(k)$. Fig. 4 shows the input and output relationship of the dead zone element. The gray zone corresponds to the sector area which is surrounded by $y = 0$ and $y = \kappa x$ for $0 \le \kappa \le 1$. It is easy to see that $\Psi(x)$ exists in this sector area if and only if the following sector condition is satisfied:

$$\Psi(x)[\Psi(x) - \kappa x] \le 0. \tag{18}$$

It follows from the small gain theorem that if the saturation error remains the sector area, then the system is l_2-stable. The condition on the saturation error leads to the well-known circle criterion: Let $G_{\tilde{u}\tilde{w}}[\mathrm{z}]$ be the transfer function from $\tilde{w}(k)$ to $\tilde{u}(k)$. Then, it follows from the small gain theorem that: When

$$(1 - \kappa G_{\tilde{u}\tilde{w}}(e^{j\theta})) + (1 - \kappa G_{\tilde{u}\tilde{w}}(e^{j\theta}))^* > 0, \ (\theta \in [0, 2\pi)) \tag{19}$$

the system is l_2-stable if $\tilde{w}(k)$ for $k \ge 0$ are in the sector area.

Let us define a augmented state vector from the state vectors $x_c(k)$, $x_p(k)$, and $x_r(k)$ of the controller, the plant, and the NTF as

$$x(k) = \begin{bmatrix} x_c(k+1) \\ x_p(k+1) \\ x_r(k+1) \end{bmatrix}. \tag{20}$$

The state equation is given by

$$x(k+1) = Ax(k) + B\tilde{w}(k) \tag{21}$$

where

$$A = \begin{bmatrix} \tilde{A} & \tilde{B}C_r \\ \mathbf{0} & A_r \end{bmatrix}, \quad B = \begin{bmatrix} \tilde{B} \\ B_r \end{bmatrix} \tag{22}$$

with $\mathbf{0}$ a zero matrix of an appropriate dimension,

$$\tilde{A} = \begin{bmatrix} A_c & -B_cC_p \\ B_pC_c & A_p \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} \mathbf{0} \\ -B_p \end{bmatrix}. \tag{23}$$

Then, we obtain

$$\tilde{u}(k) = C_{\tilde{u}\tilde{w}}x(k) \tag{24}$$

where

$$C_{\tilde{u}\tilde{w}} = \begin{bmatrix} C_c & \mathbf{0} & -C_r \end{bmatrix}. \tag{25}$$

Using the KYP lemma [9], we can show that the condition (19) is equivalent to the existence of a positive definite matrix $P_2 \succ 0$ that satisfies the following matrix inequality:

$$\begin{bmatrix} -P_2 & -B & AP_2 \\ -B^T & -2I_n & \kappa C_{\tilde{u}\tilde{w}}P_2 \\ P_2A^T & \kappa P_2C_{\tilde{u}\tilde{w}}^T & -P_2 \end{bmatrix} \prec 0. \tag{26}$$

C. Design of $\Delta\Sigma$ modulators

The value for ϵ should be decreased to reduce the effect of the quantization, whereas the value for κ should be increased to enhance the stability. In general, there is a trade-off between reducing the effect of quantization and enhancing stability. Let us consider the minimization of ϵ for a given κ.

Formally, our problem can be stated as follows:

$$\min_{P_1,P_2,A_r,B_r,C_r} \epsilon \tag{27}$$

subject to (16) and (26). Unfortunately, since the conditions are BMIs, our problem is of NP-hard, which is difficult to be solved globally.

To address this problem, we adopt the Lyapunov sharing paradigm, as suggested by [10]. By setting the two Lyapunov matrices to be identical, we introduce an additional constraint on the Lyapunov matrices. However, it should be noted that the global solution cannot be generally obtained due to this constraint.

If we set the order of the NTF to be

$$n_r = n_c + n_p := m, \tag{28}$$

then, by using the change of variables proposed independently in [10] and [11], the condition $P_2 = P_1$ leads to the following convex optimization problem with respect to $m \times m$ real positive definite matrices P_f and P_g , $W_f \in \mathbb{R}^{1\times m}$, $W_g \in \mathbb{R}^{m\times 1}$, and $H \in \mathbb{R}^{m\times m}$:

$$\min_{P_f,P_g,W_f,W_g,H} \epsilon \tag{29}$$

subject to

$$\begin{bmatrix} M_P & M_A & M_B & \mathbf{0} \\ M_A^T & M_P & \mathbf{0} & M_{C_2}^T \\ M_B^T & \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & M_{C_2} & \mathbf{0} & \epsilon \end{bmatrix} \succ 0 \tag{30}$$

$$\begin{bmatrix} -M_P & -M_B & M_A \\ -M_B^T & -2I & \kappa M_{C_1} \\ M_A^T & \kappa M_{C_1}^T & -M_P \end{bmatrix} \prec 0, \tag{31}$$

where matrices $\{M_P, M_A, M_B, M_{C_1}, M_{C_2}\}$ are defined as

$$M_P = \begin{bmatrix} P_f & I \\ I & P_g \end{bmatrix} \tag{32}$$

$$M_A = \begin{bmatrix} \tilde{A}P_f + \tilde{B}W_f & \tilde{A} \\ H & P_g\tilde{A} \end{bmatrix}, \quad M_B = \begin{bmatrix} \tilde{B} \\ W_g \end{bmatrix} \tag{33}$$

$$M_{C_1} = \begin{bmatrix} [C_c & \mathbf{0}]P_f - W_f[C_c & \mathbf{0}] \end{bmatrix} \tag{34}$$

$$M_{C_2} = \begin{bmatrix} [C_c & \mathbf{0}]P_f & [C_c & \mathbf{0}] \end{bmatrix}. \tag{35}$$

Since the objective in (29) is linear, and (30) and (31) are LMIs that are convex, the minimization problem is a convex optimization, which can be solved numerically.

Once $\{P_f, P_g, W_f, W_g, H\}$ are obtained, $\{A_r, B_r, C_r\}$ are given by

$$A_r = [\tilde{B}W_f - P_g^{-1}(H - P_g\tilde{A}P_f)](P_f - P_g^{-1})^{-1} \tag{36}$$

$$B_r = \tilde{B} - P_g^{-1}W_g \tag{37}$$

$$C_r = W_f(P_f - P_g^{-1})^{-1}. \tag{38}$$

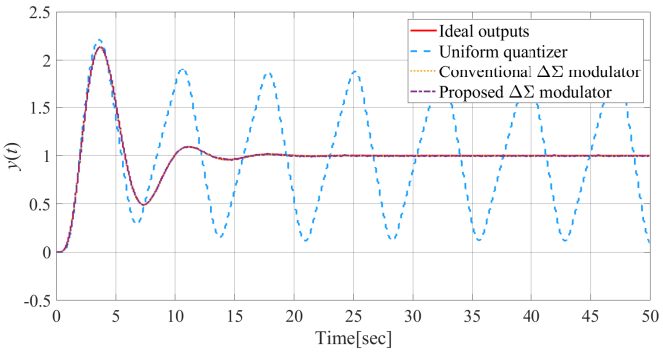


Fig. 5. System outputs $y(t)$ of plant ($d = 1, L = 2$).

IV. Numerical Examples

We consider a control system composed of a continuous-time plant P_c and controller C_c whose (A, B, C) matrices are given by

$$\begin{bmatrix} -0.01 & 1 \\ 0 & -0.01 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{39}$$

and

$$\begin{bmatrix} 0 & 1.000 & -2.414 & 2.414 \\ -2.414 & -2.414 & -2.000 & 1.000 \\ 1.000 & 0 & -2.414 & 2.414 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$
$$\begin{bmatrix} 2.414 & 2.414 & 1.000 & 0 \end{bmatrix},$$

respectively. The continuous-time system is discretized using the zero-order hold method and sampled with a period of $T_s = 0.1$ to obtain its equivalent discrete-time model, as described in [12].

The coefficient κ, which determines the sector area, is set to be $\kappa = 0.87$. We numerically solve our optimization problems using CVX, a package for specifying and solving convex programs, as introduced in [13], a package for specifying and solving convex programs.

We employ a mid-tread uniform quantizer whose quantization width is d that produces

$$Q_u(x) = id, \text{ for } x \in \left[\left(i - \frac{1}{2}\right)d, \left(i + \frac{1}{2}\right)d\right), |x| \le L \tag{40}$$

for x a scalar input with i an integer.

For the reference signal $u_r(t) = 1$ for $t \ge 0$, we compare the following systems: the ideal system without saturation and quantization, the system with a conventional uniform quantizer instead of the $\Delta\Sigma$ modulator, the system with the $\Delta\Sigma$ modulator proposed by [3], and our proposed system. In figures, their legends are *Ideal output*, *Uniform quantizer*, *Conventional $\Delta\Sigma$ modulator*, and *Proposed $\Delta\Sigma$ modulator*, respectively. They are plotted using red, dashed blue, dotted cyan, and dashed-dotted magenta curves, respectively.

Fig. 5 displays the system outputs $y(t)$ of the plant for the case where $d = 1$ and $L = 2$. We use a large quantization

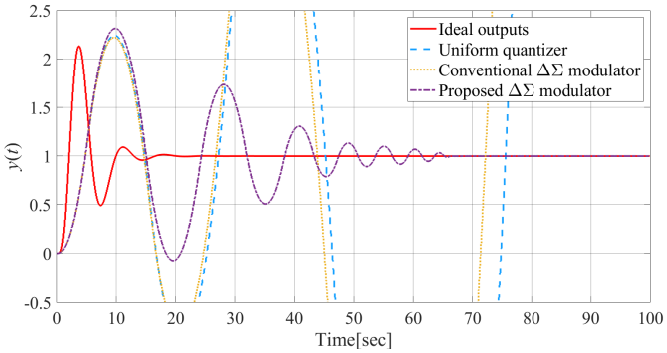


Fig. 6. System outputs $y(t)$ of plant ($d = 0.05, L = 0.1$).

width and a sufficiently large saturation level to observe the performance against quantization errors. Since the quantization level L is large enough, no windup phenomenon can be observed. The system outputs with the conventional $\Delta\Sigma$ modulator and our proposed $\Delta\Sigma$ modulator are almost identical to the ideal system output. However, the system output with the conventional uniform quantizer fluctuates due to quantization errors. This demonstrates that our proposed $\Delta\Sigma$ modulator exhibits similar robustness to quantization errors as the $\Delta\Sigma$ modulator proposed by [3].

In the next case, we examine the performance against saturation. Contrary to the previous case, we choose a small quantization width d and a smaller saturation level L, setting $d = 0.05$ and $L = 0.1$.

Figure 6 compares the system outputs of the four systems. As the saturation level is small, the system outputs with the uniform quantizer and the $\Delta\Sigma$ modulator of [3] diverge, exhibiting windup phenomena. On the other hand, the system output with our proposed $\Delta\Sigma$ modulator converges, although it fluctuates more than the ideal system output. This verifies that our design is robust to saturation.

V. Conclusions

In this study, we have proposed a design method for $\Delta\Sigma$ modulators that addresses the combined effects of quantization and saturation. Our approach aims to minimize the l_2 norm of the quantization error at the system output while satisfying a stability condition for the closed-loop system. By allowing the sharing of a common Lyapunov matrix, we have transformed our design problem into a convex optimization, which can be solved numerically. We have provided numerical examples to demonstrate the effectiveness of our proposed design approach.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number 22K04090.

References

[1] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2004.

[2] M. R. Tariq and S. Ohno, "Unified LMI based design of $\Delta\Sigma$ modulators," *EURASIP Journal on Advances in Signal Processing,*, no. 2016:29, 2016. DOI: 10.1186/s13634-016-0326-2.

[3] S. Ohno and M. R. Tariq, "Optimization of noise shaping filter for quantizer with error feedback," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 4, pp. 918–930, Apr. 2017, ISSN: 1549-8328. DOI: 10.1109/TCSI.2016.2627031.

[4] S. Galeani, S. Tarbouriech, M. Turner, and L. Zaccarian, "A tutorial on modern anti-windup design," *European Journal of Control*, vol. 15, no. 3, pp. 418–440, 2009, ISSN: 0947-3580. DOI: https://doi.org/10.3166/ejc.15.418-440.

[5] P. Weston and I. Postlethwaite, "Linear conditioning schemes for systems containing saturating actuators," *IFAC Proceedings Volumes*, vol. 31, no. 17, pp. 675–680, 1998, 4th IFAC Symposium on Nonlinear Control Systems Design 1998 (NOLCOS'98), Enschede, The Netherlands, 1-3 July, ISSN: 1474-6670. DOI: https://doi.org/10.1016/S1474-6670(17)40415-0.

[6] S. Azuma and T. Sugie, "Synthesis of optimal dynamic quantizers for discrete-valued input control," *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2064–2075, Oct. 2008, ISSN: 0018-9286. DOI: 10.1109/TAC.2008.929400.

[7] K. Sawada and S. Shin, "Dynamic quantizer synthesis based on invariant set analysis for SISO systems with discrete-valued input," in *the 19th International Symposium on Mathematical Theory of Networks and Systems*, 2010, pp. 1385–1390.

[8] W. M. Haddad and V. Chellaboina, *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton University Press, 2008, ISBN: 9780691133294.

[9] S. Boyd, L. E. Ghaoul, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Society for Industrial and Applied Mathematics, 1997.

[10] C. Scherer, P. Gahinet, and M. Chilali, "Multiobjective output-feedback control via LMI optimization," *IEEE Transactions on Automatic Control*, vol. 42, no. 7, pp. 896–911, Jul. 1997, ISSN: 0018-9286. DOI: 10.1109/9.599969.

[11] I. Masubuchi, A. Ohara, and N. Suda, "LMI-based controller synthesis: A unified formulation and solution," *International Journal of Robust and Nonlinear Control*, vol. 8, no. 8, pp. 669–686, Jul. 1998.

[12] G. F. Franklin, J. D. Powell, and M. L. Workman, *Digital control of dynamic systems*, 3rd ed. Addison-wesley Menlo Park, CA, 1998.

[13] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*, Mar. 2014. [Online]. Available: http://cvxr.com/cvx.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Asymptotic Estimation Performance of Linear Regression Model with Sparse Bayesian Learning as Both Samples and Signals Approach Infinity

Kazuaki Murayama*
* Department of Computer and Network Engineering, Graduate School of Informatics and Engineering,
The University of Electro-Communications 1–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585 Japan
E-mail: m2141009@edu.cc.uec.ac.jp

***Abstract*—This paper considers sparse Bayesian learning (SBL) in the linear regression model used in signal processing. An estimation performance of the method is analyzed in the asymptotic case where the sample size and the number of regression coefficients both increase. This subject has two cases: one is the dependence of performance on the hyperparameters in prior distribution and the other is how the performance depends on the sample size and the number of nonzero components in regression coefficients. To address this subject, we employ methods of statistical mechanics in physics. An equivalent of thermodynamic potential is calculated, following some similarities between Bayes inference and statistical mechanics. It is known in physics that the potential provides physical quantities. In a similar way, the equivalent of potential provides us with the asymptotic evaluation of estimation performance instead of physical quantities. The estimation performance in the two cases appears to have sharply divided regions in which the estimation succeeds and fails. These results are compared with related studies.**

I. Introduction

This paper deals with sparse estimation problem in signal processing through Bayesian linear regression model. The sparse estimation promotes zero estimation of variables which are the target of estimation, expecting variables have some zero components. In particular, the sparse estimation of regression coefficients in linear regression models has a wide range of applications. For example, it arises in the signal processing problems such as basis selection [1], genome-wide association studies [2], and compressed sensing [3], [4].

In order to perform sparse estimation, various methods have been developed. For example, lasso [5] is influential in this field of sparse estimation from a frequentist point of view. The minimization of L1 norm under certain constraint is pivotal as well in the context of compressed sensing [3]. Further we can also perform the sparse estimation in the framework of Bayesian inference such as Bayesian Lasso [6]. This paper concentrates our attention to a method called sparse Bayesian learning (SBL) [7] and its estimation performance in linear regression model among those methods. This issue has two aspects. One side is the dependence of performance on the hyperparameters in prior distribution. Another aspect is how the performance depends on the sample size n, the number of signal components p, and that of nonzero components in signal. The estimation performance of SBL was previously investigated using certain data with fixed n and p, and the effectiveness of SBL in basis selection and compressed sensing was numerically demonstrated [1], [4]. Additionally, hyperparameters dependence of estimation performance was numerically examined in specific data analysis on diffuse optical tomography using SBL with variational inference [8], [9]. These previous studies [8], [9] observed that the dependence has two regions, and then the estimation performance appears to be robust to variations in hyperparameters within each region.

Given the numerical demonstration of effectiveness of SBL in the case of finite n and p, we shall analyze its estimation performance in linear regression model in the asymptotic case where the number of samples n and that of signal components p both increase at once. This asymptotic evaluation is performed without specific factors like variational inference and diffuse optical tomography data included in previous studies. Such an analysis allows us to observe the asymptotic estimation performance of SBL without the influence of the specific factors and the finiteness of n and p, which contributes to understanding of further properties of SBL and its effectiveness in signal processing.

We explain the reason to focus on the asymptotic evaluation that takes the limit of both n and p, rather than the standard asymptotic evaluation that takes the limit of n with fixing p. The evaluation based on the former asymptotics is associated with the emergence of data in which p is comparable to n [10], [11], [12], [13], [14], [15]. With the emergence of such data, analyzing asymptotic properties of statistical models based on the asymptotics taking limit of both n and p has gained attention [10], [11], [12], [13], [14]. Following this direction, this paper analyzes the asymptotic estimation performance of SBL which is the important method in the field of signal processing.

We analyze the estimation performance using methods in statistical mechanics which is a branch of physics. There are some similarities between statistical mechanics and Bayes inference, and issues in statistics have been analyzed using statistical mechanics through these similarities [16]. For example, the estimation performance in the asymptotic case taking limit of both n and p was analyzed using the replica method

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of statistical mechanics in the context of several methods for signal reconstruction [17], [18]. Given these analysis cases, we employ the replica method and analyze the performance of SBL with it.

II. Bayes model and population distribution of data

A. Bayes model of sparse Bayesian learning

Let $\{(y_i, \boldsymbol{x}_i); i = 1, \cdots, n\}$ be data sets where y_i and $\boldsymbol{x}_i \in \mathbb{R}^p$ are response variable and explanatory variable, respectively. Here we define $\boldsymbol{y} \coloneqq (y_i, \cdots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\boldsymbol{X} \coloneqq (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times p}$ as observation vector and design matrix, respectively.

In linear regression model, an structure connecting $\boldsymbol{y}$ and $\boldsymbol{X}$ is modeled using the regression coefficient or signal $\boldsymbol{w} \in \mathbb{R}^p$ as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^{\mathrm{T}}$ is an error vector whose components are frequently assumed to independently, identically distributed according to normal distribution with zero mean and σ^2 as variance. In this situation, the likelihood function is given as

$$P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2\right), \tag{2}$$

where $\|\cdot\|$ is Euclidean norm.

To perform Bayes inference for the linear regression model, we shall equip the regression coefficient $\boldsymbol{w}$ in (1) with prior distribution. The framework of sparse Bayesian learning requires hierarchical prior distribution [7]. It is assumed that prior distribution for $\boldsymbol{w}$ is component-wise normal distribution as follows:

$$P(\boldsymbol{w}|\boldsymbol{\tau}) = \prod_{j=1}^{p} \left(\frac{\tau_j}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau_j}{2} w_j^2\right), \tag{3}$$

where $\boldsymbol{\tau} \coloneqq (\tau_1, \cdots, \tau_p)^{\mathrm{T}}$ is additional parameter distributed like component-wise Gamma distribution, i.e.,

$$P(\boldsymbol{\tau}) = \prod_{j=1}^{p} \frac{b^a}{\Gamma(a)} \tau_j^{a-1} e^{-b\tau_j}. \tag{4}$$

The hyperparameters $a > 0$ and $b > 0$ in (4) are shape and rate parameters, respectively. Marginalizing the product of $P(\boldsymbol{w}|\boldsymbol{\tau})$ and $P(\boldsymbol{\tau})$ with respect to $\boldsymbol{\tau}$ gives the prior distribution of $\boldsymbol{w}$ alone as follows:

$$P(\boldsymbol{w}) = \prod_{j=1}^{p} \frac{\Gamma(a + 1/2) b^a}{\sqrt{2\pi}\Gamma(a)} \left(b + \frac{w_j^2}{2}\right)^{-(a+1/2)}. \tag{5}$$

The likelihood function (2) and prior distribution (5) lead us to make the posterior distribution of $\boldsymbol{w}$ as follows:

$$P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) = \frac{1}{Z} P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) P(\boldsymbol{w}), \tag{6}$$

where

$$Z = \int d\boldsymbol{w}\, P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})\, P(\boldsymbol{w}) \tag{7}$$

is the marginal likelihood or evidence as normalization factor.

B. population distribution of data

To analyze estimation performance we assume that $\boldsymbol{y}$ and $\boldsymbol{X}$ are truly linearly connected. In addition, details of distribution for data $(\boldsymbol{y}, \boldsymbol{X})$ together with true signal are set.

We assume that the true regression coefficiet or signal $\boldsymbol{v} \in \mathbb{R}^p$ is distributed according to spike and slab prior as follows:

$$r(\boldsymbol{v}) = \prod_{j=1}^{p} \left\{(1 - \rho)\delta(v_j) + \rho \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} v_j^2\right)\right\}, \tag{8}$$

where $\delta(v_j)$ represents Kronecker delta and ρ takes values in the range of $[0, 1] \subset \mathbb{R}$ meaning the ratio of the number of nonzero components. Further design matrix $\boldsymbol{X}$ is assumed to be random matrix whose components are distributed like normal distribution with zero mean and $1/p$ as variance, i.e.,

$$r(\boldsymbol{X}) = \prod_{i=1}^{n} \prod_{j=1}^{p} \left(\frac{p}{2\pi}\right)^{1/2} \exp\left(-\frac{p}{2} x_{ij}^2\right). \tag{9}$$

We consider that $\boldsymbol{y}$ is obtained through (1) with $\boldsymbol{v}$ and $\boldsymbol{X}$ generated from (8) and (9), respectively. In this case, substituting (2) for $r(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{v}\|^2/(2\sigma^2)\right)$, the distribution of data together with $\boldsymbol{v}$ is given as

$$r(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v}) = r(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v})\, r(\boldsymbol{v})\, r(\boldsymbol{X}). \tag{10}$$

III. Statistical physical analysis of estimation performance

A. an equivalent of thermodynamic potential

We use statistical mechanics to analyze the asymptotic estimation performance. Based on the similarities between the marginal likelihood (7) in Bayesian inference and the partition function in statistical mechanics, which have been proposed so far, the analog of thermodynamic potential calculated in this paper is given as follows:

$$\phi(a, b, \alpha, \rho) = \lim_{n,p \to \infty} \frac{1}{p} \mathbb{E}_{r(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})}\big[\ln Z\big], \tag{11}$$

where Z is (7) and asymptotic limit associated with n and p is incorporated into the right-hand side. The limit of both n and p is taken so that $n/p \to \alpha \in (0, \infty) \subset \mathbb{R}$. We note that $\mathbb{E}_{r(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})}[\cdot]$ in (11) means the calculation of expected value with respect to (10). This potential, (11), is thought of as a function of (a, b, α, ρ) where (a, b) are hyperparameters in prior distribution ((4) and (5)) and ρ is the ratio of nonzero components in the true signal appeared in (8). In addition α represents converging value of n/p as both n and p approach infinity.

B. calculation of potential

Equation (11) is calculated using the replica method of statistical physics [19], [17], [18], [20] together with replica symmetry. To investigate the noise-free estimation performance, we consider (11) in the case of $\sigma^2 \to 0$ in (2) and $r(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{v}\|^2/(2\sigma^2)\right)$ in (10). The overview of calculation in (11) is concisely explained in

the later appendix. This section mainly describes the results. The calculation result of (11) is

$$\phi(a,b,\alpha,\rho) = -\frac{\alpha}{2}\left\{\ln[2\pi\,(Q-q)] + \frac{\rho-2m+q}{Q-q}\right\} + \frac{1}{2}\tilde{Q}Q + \frac{1}{2}\tilde{q}q - \tilde{m}m + \mathbb{E}_{r(\theta,v)}[\ln z\,(\theta,v)], \quad (12)$$

where

$$r(\theta,v) = r(\theta)r(v) = \frac{\exp(-\theta^2/2)}{\sqrt{2\pi}}\left\{(1-\rho)\delta(v) + \frac{\rho}{\sqrt{2\pi}}\exp\left(-\frac{v^2}{2}\right)\right\}. \quad (13)$$

In addition, $z\,(\theta,v)$ in (12) is

$$z\,(\theta,v) = \int dw\;\psi(w,\theta,v), \quad (14)$$

where the integrand is expressed as

$$\psi(w,\theta,v) = \left(b+\frac{w^2}{2}\right)^{-a-1/2} \times \exp\left\{-\frac{1}{2}(\tilde{Q}+\tilde{q})w^2 + \left(\sqrt{\tilde{q}}\theta + \tilde{m}v\right)w\right\}. \quad (15)$$

Here v in (12)-(15) is any one univariate in $\boldsymbol{v} \in \mathbb{R}^p$ without subscript $j \in \{1,\cdots,p\}$ for v_j. In the same way, w in (14) and (15) is any one univariate in $\boldsymbol{w} \in \mathbb{R}^p$ without subscript. Besides, $Q, q, m, \tilde{Q}, \tilde{q}, \tilde{m}$ in (12)-(15) satisfy following equations:

$$\tilde{Q} = \alpha\left\{\frac{1}{Q-q} - \frac{\rho-2m+q}{(Q-q)^2}\right\}, \quad (16)$$

$$\tilde{q} = \alpha\left\{-\frac{1}{Q-q} + \frac{\rho-2m+Q}{(Q-q)^2}\right\}, \quad (17)$$

$$\tilde{m} = \frac{\alpha}{Q-q}, \quad (18)$$

$$Q = \mathbb{E}_{r(\theta,v)}\left[\left\langle w^2\right\rangle\right], \quad (19)$$

$$q = \mathbb{E}_{r(\theta,v)}\left[\left\langle w\right\rangle^2\right], \quad (20)$$

$$m = \mathbb{E}_{r(\theta,v)}\left[v\left\langle w\right\rangle\right], \quad (21)$$

where $\left\langle w^k\right\rangle$ with $k=1$ or 2 is expressed as

$$\left\langle w^k\right\rangle = \frac{1}{z\,(\theta,v)}\int dw\;w^k\;\psi(w,\theta,v). \quad (22)$$

We note that these (14)-(21) are also thought of as the functions of (a,b,α,ρ) like $\phi(a,b,\alpha,\rho)$ in (11) and (12). The notation that $z(\theta,v)$, $\psi(w,\theta,v)$, $\tilde{Q}$, $\tilde{q}$, $\tilde{m}$, Q, q, and m are the functions of (a,b,α,ρ) has been omitted for simplicity.

C. measure of estimation performance

The asymptotic mean square error (MSE) which is the measure of estimation performance is given as

$$\mathrm{MSE}(a,b,\alpha,\rho) = \lim_{n,p\to\infty}\frac{1}{p}\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}\left[\left\|\mathbb{E}_{P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}] - \boldsymbol{v}\right\|^2\right] = \rho - 2m + q, \quad (23)$$

using m and q required in the process of calculating analog of thermodynamic potential (12) together with ρ in (8).

To analyze the estimation performance based on statistical physics, it is required to calculate the analog of thermodynamic potential (11). Here, we have obtained the form of (12) for (11) using the method of statistical physics such as the replica method. Since (12) includes $\tilde{Q}$, $\tilde{q}$, $\tilde{m}$, Q, q, and m that satisfy (16)-(21), we are required to evaluate them. On this occasion, the asymptotic estimation performance is obtained from (23). However the nonlinear equations for (16)-(21) cannot be solved explicitly. Therefore we evaluate $\tilde{Q}$, $\tilde{q}$, $\tilde{m}$, Q, q, and m through iterative calculation for (16)-(21). In the iterative calculation, we are confronted with the integrals with respect to (22) and $\mathbb{E}_{r(\theta,v)}[\cdot]$ in (19)-(21) that cannot be calculated analytically. The former and latter integrals are treated as numerical integral of Gaussian quadrature and that of Simpson's rule [21], respectively.

To sum up, examining the value of MSE (23) takes further the three steps following the statistical physical calculation from (11) to (12) as follows: First the specific value of (a,b,α,ρ) is fixed where (a,b), ρ, and α are hyperparameters in prior distribution, the ratio of nonzero components in the true vector (8), and the converging value of n/p as $n,p\to\infty$, respectively. Second the specific values of $\tilde{Q}$, $\tilde{q}$, $\tilde{m}$, Q, q, and m are determined through the iterative calculation for (16)-(21). Finally we observe the value of MSE substituting the obtained m and q into (23) together with ρ specified in the first step.

IV. ESTIMATION PERFORMANCE

A. hyperparameters dependence of estimation performance

To investigate the dependence of estimation performance on hyperparameters (a,b) in (4) and (5), we fix the remaining variables as $(\alpha,\rho) = (0.85, 0.5)$ as a representative case. This setting means the task of estimating a signal containing $p\times 0.5$ zero components using SBL under $n \sim 0.85p(<p)$ measurements. We obtain MSE for various values of (a,b) from the aforementioned procedures in Sec. III-C. The result is plotted in Fig. 1. In the current work, MSE could not be obtained in the upper right and lower right regions around $(a, a/b) \sim (10^2, 10^{-5})$ and $(a, a/b) \sim (10^2, 10^5)$. This issue will be discussed in the later section V.

The previous studies [8], [9] plotted a measure of estimation performance on $(a, a/b)$ plane. To compare Fig. 1 with the previous studies, MSE is plotted on $(a, a/b)$ plane instead of, for instance, (a,b) plane. The results of previous researches [8], [9] included the specific factors such as variational inference and diffuse optical tomography data. In addition, those were

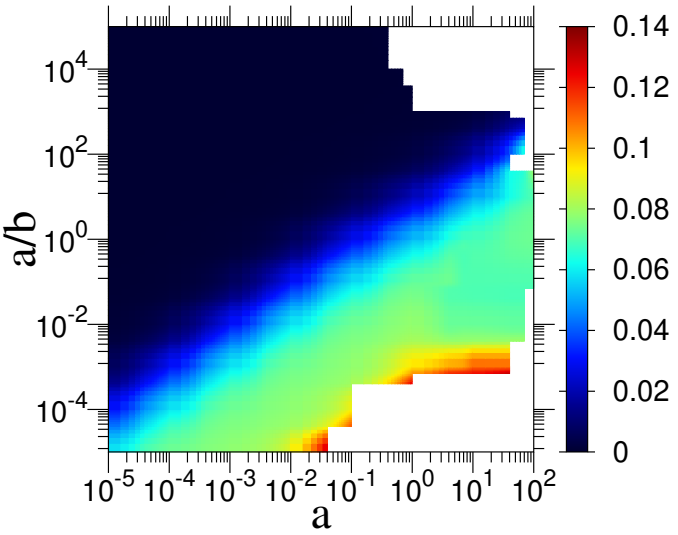


Fig. 1. Mean squared error ($\mathrm{MSE}(a, b, \alpha = 0.85, \rho = 0.5)$) is plotted in heat map format on $(a, a/b)$ with a logarithmic scale. The range of shape parameter is $a \in \{\xi \times 10^{\eta}; \xi = 1, 4, 7, \quad \eta = -5, -4, -3, -2, -1, 0, 1, 2\}$ while that of the ratio is $a/b \in \{\xi \times 10^{\eta}; \xi = 1, 4, 7, \quad \eta = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$. We note that a/b means mean parameter of Gamma distribution in (4). To facilitate comparison with previous studies [8], [9], MSE is plotted on $(a, a/b)$ plane instead of, for instance, (a, b) plane. The case of $\mathrm{MSE} \sim 0$ in the region where a is small and a/b is large represents success of estimation, while that of $\mathrm{MSE} \nsim 0$ represents failure.

not the results of asymptotic analysis. Therefore our attention should be paid to overall appearance comparison rather than detailed and quantitative comparison. Overall, as shown in Fig. 1, the MSE values are clearly divided between the small-a, large-a/b region and the large-a, small-a/b region. This overall picture of divided estimation performance is roughly in agreement with the observation by [8], [9], apart from in which regions the error is smaller. The result of this paper may differ from those of [8], [9] in terms of in which region the error is smaller. Reference [8] examined the dependence of the error on shape parameter a by analyzing the iterative formula of variational inference in a simplified one-dimensional case of $n = p = 1$ together with $a/b \to \infty$. According to that result of Fig. 7 in [8], the estimation error is smaller in the region where shape parameter is smaller. Therefore, our Fig. 1 that MSE is smaller in the small-a, large-a/b region does not contradict the result of Fig. 7 in [8]. However, Fig. 1 may contradict the heat maps of Fig. 5 in [8] and Fig. 7 in [9]. In those heat maps, at least as far as the author has seen, it seems that the error become large in the small-a, large-a/b region and it is small in the large-a, small-a/b region. If this is true, Fig. 1 is not consistent with those heat maps provided by [8], [9]. Although the cause of this discrepancy is not clear at this time, it may be due to the differences in the analysis settings. The settings used in [8], [9] are as follows: the specific finite size data for diffuse optical tomography, the noisy case of $\sigma^2 \neq 0$, the non-asymptotic evaluation with approximation of posterior distribution using variational inference. On the other hand, the settings in this paper are as follows: the data according to population distribution explained in II-B, the noise-free case of $\sigma^2 \to 0$, the asymptotic evaluation without approximation of posterior distribution. We consider that these differences in settings may make the inconsistency between our Fig. 1 and the heat maps of Fig. 5 in [8] and Fig. 7 in [9].

We shall turn our attention to the results in Fig. 1 itself from the comparison with previous studies. In the upper left region around $(a, a/b) \sim (10^{-5}, 10^{5})$, τ_j in prior distribution (4) behaves as $1/\tau_j \sim 0$ in the sense of average, because the mean of τ_j, a/b, becomes large ($a/b \sim 10^{5}$). This case also represents $w_j \sim 0$ because $1/\tau_j (\sim 0)$ plays the role of variance of (3) whose mean is zero per component. Therefore we consider zero components in signal to be easily estimated. Moreover, since the variance of τ_j in (4), a/b^2, becomes large ($a/b^2 \sim 10^{15}$), the mass of prior density is also distributed to $1/\tau_j \nsim 0$ which corresponds to the estimation of nonzero components in signal. As a result of these two aspects, we interpret that the MSE takes the values close to zero in Fig. 1, which suggests the success in estimating signal in the upper left region around $(a, a/b) \sim (10^{-5}, 10^{5})$. We also recognize, in the upper left region, that the prior distribution (4) is close to Jeffreys noninformative prior $P(\tau_j) \propto 1/\tau_j$. Thus Fig. 1 shows that the choice of the noninformative prior results in a low-error estimation, at least within the settings in Sec. II. In contrast to the upper left region around $(a, a/b) \sim (10^{-5}, 10^{5})$, both the mean (a/b) and variance (a/b^2) of τ_j in prior distribution (4) are small in the lower right region with $\mathrm{MSE} \nsim 0$. Such a prior is considered to estimate all of the components in signal to be nonzero, failing sparse estimation.

B. (α, ρ) dependence of estimation performance

Continuing from the previous subsection IV-A we examine the dependence of estimation performance on (α, ρ) instead of (a, b). Recall that α represents the convergence of sample size (n) per the number of signals (p), and ρ is the ratio of nonzero components in the true signal in (8). The hyperparameters in (4) and (5) are fixed as $(a, b) = (10^{-5}, 10^{-5}) \leftrightarrow (a, a/b) = (10^{-5}, 1)$ as the representative case. This selection has two implications. This point belongs to the region with $\mathrm{MSE} \sim 0$ in Fig. 1. Although Fig. 1 is the result of fixing (α, ρ) to $(0.85, 0.5)$, we would expect the selection as $(a, b) = (10^{-5}, 10^{-5})$ to be non-singular situation in the present context of (α, ρ) dependence on MSE in the sense that $(a, a/b) = (10^{-5}, 1)$ is not close to the boundary between $\mathrm{MSE} \sim 0$ and $\mathrm{MSE} \nsim 0$ in Fig. 1. Additionally, $(a, b) = (10^{-5}, 10^{-5})$ is close to the setting of $(a, b) = (10^{-6}, 10^{-6})$ which was used as noninformative prior in terms of Jeffreys prior in the context of relevance vector machine [22]. Although we do not calculate the case of just $(a, b) = (10^{-6}, 10^{-6})$ in this study, the difference from $(a, b) = (10^{-5}, 10^{-5})$ is only one digit. In this sense, it is considered that there is no significant difference in the result within this study. Therefore we regard $(a, b) = (10^{-5}, 10^{-5})$ as the canonical setting close to the noninformative prior. Values of MSE are obtained for various

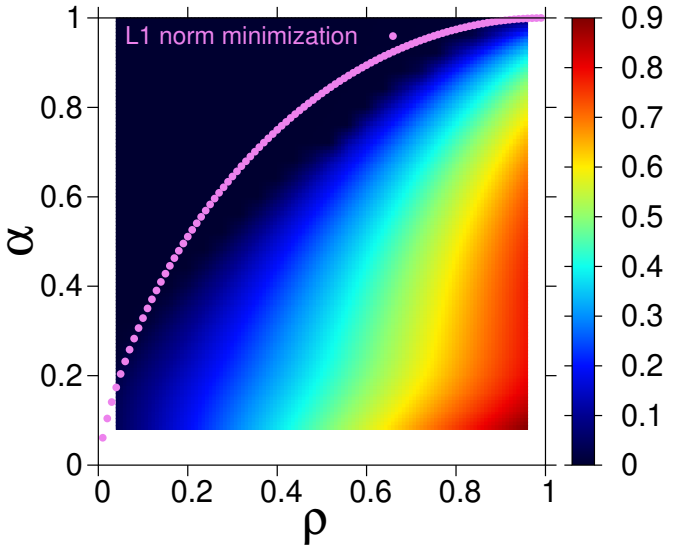


Fig. 2. Mean squared error (MSE($a = 10^{-5}, b = 10^{-5}, \alpha, \rho$)) is plotted in heat map format on (ρ, α). The range of percentage of nonzero components in (8) is $\rho \in \{0.04 \times \xi; \xi = 1, 2, 3, 4, 5, \cdots 20, 21, 22, 23, 24\}$. The range of convergence of sample size (n) per the number of signals (p) is $n/p \to \alpha \in \{0.04 \times \xi; \xi = 2, 3, 4, 5, 6, \cdots, 21, 22, 23, 24, 25\}$. The boundary between zero and nonzero of MSE is also plotted in the case of L1 norm minimization [17]. It was confirmed $\mathrm{MSE} = 0$ in the region above this boundary, while $\mathrm{MSE} \neq 0$ in the region below the boundary [17]. The case of $\mathrm{MSE} \sim 0$ of SBL in the upper left region around $(\alpha, \rho) \sim (1, 0)$ represents success of estimation. The case of $\mathrm{MSE} \nsim 0$ of SBL in the lower right region around $(\alpha, \rho) \sim (0, 1)$ means failure in estimating the signal.

values of (α, ρ) from the procedure described in Sec III-C. The result is plotted in Fig. 2.

The boundary between zero and nonzero of MSE by L1 norm minimization [17] is also plotted in Fig. 2 for comparison. This boundary was previously provided in [17]. In the region above this boundary, MSE with L1 norm minimization takes zero while it takes nonzero values in the region below the boundary. Both the boundary of L1 minimization and that of SBL exhibit similar behavior in the sense that the boundaries between $\mathrm{MSE} \sim 0$ and $\mathrm{MSE} \nsim 0$ appear to be located in an arc from $(\alpha, \rho) = (0, 0)$ to $(1, 1)$. The position of SBL boundary is not as clear as that of L1, however below the L1 boundary, a small region of $\mathrm{MSE} \sim 0$ is observed in the SBL heat map of Fig. (2). In such a region, the MSE of L1 minimization takes nonzero values while that of SBL could be close to zero. In other words, SBL could reduce the MSE to near zero with fewer observations than L1 minimization method, recalling that α is the (convergence of) measurements n per signals p. In this sense, Fig. 2 suggests that SBL could have a potential to perform well than L1 minimization method.

V. DISCUSSION AND CONCLUSION

This paper has analyzed the estimation performance of sparse Bayesian learning (SBL) in the linear regression model in the asymptotic case of both the sample size and the number of regression coefficients or signal components. We have calculated the analog of thermodynamic potential (11), importing the methods and concepts of physics into this subject. The analog of potential has provided us with two estimation performances. One is the dependence of the performance on hyperparameters in prior distribution as shown in Fig. 1. The other is how the performance depends on the sample size and the number of nonzero components in signal as shown in Fig. 2. We have observed that both the estimation performances have sharply divided regions in which the estimation succeeds and fails.

We discuss the numerical integrations of (22) and $\mathbb{E}_{r(\theta,v)}[\cdot]$ in (19)-(21). We have faced these numerical integrations in the iterative calculation of (16)-(21) to obtain the MSE. These have required a large amount of computation and computational time. Once this problem is overcome, we would examine the estimation performance in more detail such as the cases of $(a, b, \alpha \neq 0.85, \rho \neq 0.5)$ in Fig. 1 and $(a \neq 10^{-5}, b \neq 10^{-5}, \alpha, \rho)$ in Fig. 2. Another problem is that we have encountered the situations where the numerical integrals cannot be calculated at some points of (a, b, α, ρ), meaning that MSE also cannot be calculated. For example, these cases are the upper right and lower right regions around $(a, a/b) \sim (10^2, 10^{-5})$ and $(a, a/b) \sim (10^2, 10^5)$ in Fig. 1. The cause of this problem is not clear at present, however the prior distribution becoming sharp or flat in these regions might be related to the failure of numerical integrations. Although there is room for improvement in this study, we hope that the results obtained in this study are shared.

We mention directions for future research. The first is an analysis of estimation performance in the presence of noise. In this paper, we have analyzed the performance in the idealized noise-free situation of $\sigma^2 \to 0$ in (2) and $r(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{v}\|^2/(2\sigma^2)\right)$ in (10). Further analysis of performance in the noisy case of $\sigma^2 \neq 0$ is expected to provide additional insights into estimation performance in realistic situations. The second direction is to investigate the estimation performance when the posterior distribution of signal $\boldsymbol{w}$ and its expected value is determined by practical approximation algorithms. In this paper, we have obtained the results when no approximation algorithms have been applied to the posterior distribution. However, Bayesian sparse estimations of signal have often been performed with efficient approximation algorithms. For example, the variational Bayesian inference and Type-II algorithm based on evidence maximization were applied to the Bayes model of SBL [22], [1], [4]. The practicality of Type-I algorithm based on a maximum a posteriori and Type-II algorithm was further discussed by [23] in a framework that encompasses related methods for sparse estimation problem. Reference [23] proposed the sparse estimation using the power exponential scale mixture family, and showed that Type-II algorithm has better approximation performance than Type-I algorithm. These previous studies [22], [1], [4], [23] showing the affinity between Bayesian sparse estimations and the approximation algorithms motivate us to analyze the estimation performance when the approximation algorithms are applied to the posterior distribution.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP21J22128,JP22KJ1358 (Grant-in-Aid for Research Fellows of Japan Society for the Promotion of Science under Grant Numbers JP21J22128 and JP22KJ1358). The author is grateful to Yukiyasu Ozeki, Shuichi Kawano, Jun Suzuki, and the reviewers.

APPENDIX: OVERVIEW OF SPECIFIC CALCULATION OF POTENTIAL

Based on advice from the reviewers, we concisely explain the overview of the calculation to obtain (12)-(23) from (11). We refer to [19], [17], [18], [20] for the calculation. Please note that the replica method described below may have a problem with its mathematical rigor depending on certain situation. This appendix does not discuss this mathematical subtlety for the sake of simplicity. See [19], [17], [18], [20] for the subtlety.

step1: evidence replication and expectation with population distribution

It is difficult to directly calculate the expected value with $r(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})$ in (11). Therefore, in the replica method, (11) is transformed as follows:

$$\phi(a,b,\alpha,\rho) = \lim_{c\to 0}\frac{\partial}{\partial c}\lim_{n,p\to\infty}\frac{1}{p}\ln \mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}\big[Z^c\big], \tag{24}$$

where c is interpreted as an integer ($c \in \mathbb{Z}_{>0}$). For any $k \in \{1,\cdots,c\}$, one see that the signal $\boldsymbol{w}$ and evidence Z are replicated as follows:

$$Z_k = \int d\boldsymbol{w}^k\, P\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}^k\right) P\big(\boldsymbol{w}^k\big), \tag{25}$$

$$\boldsymbol{w}^k = \left(w_1^1,\cdots,w_j^k,\cdots,w_p^m\right)^{\mathrm{T}} \in \mathbb{R}^p. \tag{26}$$

Please imagine that the Bayesian model is replicated with index $k \in \{1,\cdots,c\}$, and the total number of copies is c. In this situation, Z^c in (24) is expressed as the product of copies as follows:

$$Z^c = \prod_{k=1}^{c} Z_k = \int \prod_{k=1}^{c}\Big\{d\boldsymbol{w}^k\, P\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}^k\right) P\big(\boldsymbol{w}^k\big)\Big\}. \tag{27}$$

Calculating the expectation of integrand with respect to $r(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})$, we obtain the following equation using a certain function $\Phi(\cdot)$:

$$\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}\big[Z^c\big] = \int \prod_{k,l=1}^{c} dq_{kl} \int \prod_{k=1}^{c} dm_k \exp\big\{p\times\Phi\left(\boldsymbol{Q},\boldsymbol{m}\right)\big\}, \tag{28}$$

where the details of components of matrix $\boldsymbol{Q} = (q_{kl}) \in \mathbb{R}^{c\times c}$ and vector $\boldsymbol{m} = (m_k) \in \mathbb{R}^c$ are

$$q_{kl} = \frac{1}{p}\left(\boldsymbol{w}^k\right)^{\mathrm{T}}\boldsymbol{w}^l, \quad \forall k,l \in \{1,\cdots,c\} \tag{29}$$

and

$$m_k = \frac{1}{p}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}^k, \quad \forall k \in \{1,\cdots,c\}, \tag{30}$$

respectively.

step2: integral calculation with Laplace's method

In (24), the limit of $n,p\to\infty$ is waiting for the next step. Therefore, one treat the integral in (28) using Laplace's method and obtain the following equation:

$$\phi(a,b,\alpha,\rho) = \max_{\boldsymbol{Q},\boldsymbol{m}}\Big\{\lim_{c\to 0}\frac{\partial}{\partial c}\Phi\left(\boldsymbol{Q},\boldsymbol{m}\right)\Big\}, \tag{31}$$

where an exchange of $\lim_{c\to 0}\partial_c$ and the maximization coming from Laplace's method is assumed.

step3: employing replica symmetry

Since c in (24) has been regarded as the integer at the beginning of step 1, $\lim_{c\to 0}\partial_c$ in (31) is not performed as it is. A concept called replica symmetry in replica method is the prescription for this issue. The replica symmetry expresses the property related to the exchange of subscripts in q_{kl} and m_k. Specifically, $\boldsymbol{Q} = (q_{kl}) \in \mathbb{R}^{c\times c}$ and $\boldsymbol{m} = (m_k) \in \mathbb{R}^c$ which have replica symmetry are expressed as follows: for any $k,l \in \{1,\cdots,c\}$

$$q_{kl}^{\mathrm{RS}} = \begin{cases} Q & \cdots k = l \\ q & \cdots k \neq l \end{cases} \tag{32}$$

and

$$m_k^{\mathrm{RS}} = m. \tag{33}$$

The calculation of $\lim_{c\to 0}\partial_c$ in (31) is performed by substituting $\boldsymbol{Q}^{\mathrm{RS}} = (q_{kl}^{\mathrm{RS}})$ and $\boldsymbol{m}^{\mathrm{RS}} = (m_k^{\mathrm{RS}})$ into $\Phi\left(\boldsymbol{Q},\boldsymbol{m}\right)$. We obtain

$$\phi(a,b,\alpha,\rho) = \max_{Q,q,m,\tilde{Q},\tilde{q},\tilde{m}}\Big\{\Phi(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})\Big\}, \tag{34}$$

where $(\tilde{Q},\tilde{q},\tilde{m})$ are auxiliary variables introduced in the process of calculation. The detail of $\Phi(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})$ is formally the same as the right side of (12) with (13)-(15). The equations for maximizing $\Phi(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})$ are (16)-(21). In (12), we have reexpressed the detail of $\Phi(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})$ as $\phi(a,b,\alpha,\rho)$, imposing the maximizing condition (16)-(21) on the detail of $\Phi(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})$.

From the knowledge of statistical mechanics, Q, q, and m are expected to be

$$Q = \lim_{n,p\to\infty}\frac{1}{p}\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}\Big[\mathbb{E}_{P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}]\Big], \tag{35}$$

$$q = \lim_{n,p\to\infty}\frac{1}{p}\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}\Big[\mathbb{E}_{P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}]^{\mathrm{T}}\mathbb{E}_{P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}]\Big], \tag{36}$$

and

$$m = \lim_{n,p\to\infty}\frac{1}{p}\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}\Big[\boldsymbol{v}^{\mathrm{T}}\mathbb{E}_{P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}]\Big], \tag{37}$$

respectively. Furthermore, Q, q, and m are reduced from the right side of (35)-(37) to the univariate expressions such as the right side of (19)-(21) through the calculation of replica method explained in this appendix. Therefore, MSE is expressed as the right most side of (23) ($\rho - 2m + q$).

REFERENCES

[1] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[2] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature genetics*, vol. 37, no. 4, pp. 413–417, 2005.

[3] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[4] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[7] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[8] A. Miyamoto, K. Watanabe, K. Ikeda, and M.-A. Sato, "Variational inference with ard prior for nirs diffuse optical tomography," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1109–1114, 2015.

[9] A. Miyamoto, K. Watanabe, K. Ikeda, and M.-a. Sato, "Phase diagrams of a variational bayesian approach with ard prior in nirs-dot," in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 1230–1236.

[10] V. I. Serdobolskii, *Multivariate statistical analysis: A high-dimensional approach*. Springer Science & Business Media, 2000, vol. 41.

[11] N. El Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *The Annals of Statistics*, vol. 36, no. 6, pp. 2757–2790, 2008.

[12] ——, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *The Annals of Statistics*, vol. 36, no. 6, pp. 2717–2756, 2008.

[13] Š. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former soviet union literature," *Journal of Multivariate Analysis*, vol. 89, no. 1, pp. 1–35, 2004.

[14] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4238–4255, 2011.

[15] A. Nishimura and M. A. Suchard, "Prior-preconditioned conjugate gradient method for accelerated gibbs sampling in "large n, large p" bayesian sparse regression," *Journal of the American Statistical Association*, pp. 1–14, 2022.

[16] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[17] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on L_p-norm minimization," *J. Stat. Mech.: Theory Exp.*, vol. 2009, no. 09, p. L09003, 2009.

[18] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 08, p. P08009, 2012.

[19] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. World Scientific Publishing Company, 1987, vol. 9.

[20] A. Sakata and T. Obuchi, "Perfect reconstruction of sparse signals with piecewise continuous nonconvex penalties and nonconvexity control," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 9, p. 093401, 2021.

[21] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[22] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proc. 6th conf. Uncert. Artif. Intell.* Morgan Kaufmann Publishers Inc., 2000, pp. 46–53.

[23] R. Giri and B. Rao, "Type i and type ii bayesian methods for sparse signal recovery using scale mixtures," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3418–3428, 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Convolutional Multidimensional Amplitude Spectrum Nuclear Norm for Frequency-domain Robust Principal Component Analysis

Ryoya Harashima*, Ryunosuke Eguchi†, and Seisuke Kyochi*
* Kogakuin University, Japan
E-mail: kyochi@cc.kogakuin.ac.jp
† The University of Kitakyushu, Japan

***Abstract*—In this paper, we propose convolutional multidimensional amplitude spectrum nuclear norm (CMASNN) for the task of frequency-domain robust principal component analysis (FRPCA). Principal component extraction (PCE) from observed signals, particularly in the presence of outliers, has long been recognized as a fundamental challenge in the field of signal processing. Although RPCA has demonstrated its efficacy as a technique, it suffers significant limitations when confronted with misaligned principal components due to its reliance on the low-rankness assumption. To address this issue, we initially introduced ASNN, which leverages the invariant property of amplitude spectra under (circular) shifts, enabling the robust extraction of principal components from observed signals by incorporating the ASNN into RPCA. Building upon this foundation, the CMASNN further enhances the ASNN by unifying multiple signals into a higher dimensional representation. This enhancement enables us to discriminate between true principal components and outliers more effectively. By substituting the CMASNN for the ASNN in FRPCA, our experimental results show that the proposed CMASNN-based FRPCA outperforms the performance of its conventional counterpart.**

I. Introduction

Convex optimization employing sparse and low-rank modeling has proven to be highly effective in various domains of modern signal processing and machine learning, and so on [1]–[7]. Among these applications, a particularly noteworthy one is principal component extraction (PCE) from observed signals in the presence of outliers [8]–[16]. The fundamental approach is robust principal component analysis (RPCA), which aims to extract principal components in a low-dimensional subspace by utilizing low-rank modeling with nuclear norm regularization (see Fig. 1(a)). However, a limitation of RPCA lies in its sensitivity to shifts or misalignments in the target components, which increases the rank of principal components (Fig. 1(b)).

In order to address this issue, we introduced amplitude spectrum nuclear norm (ASNN) in our previous study [17]. Fundamentally, a signal and its shifted counterpart exhibit identical amplitude spectra, indicating that both possess low-rank amplitude spectra in the frequency domain. Taking this into consideration, we substituted the ASNN, which evaluates the nuclear norm of signal amplitude spectra, with NN in RPCA. This modified approach, known as frequency-domain RPCA (FRPCA), robustly extracts principal components from a data matrix even in the presence of misalignment and outliers.

In this paper, we enhance the robustness of principal component extraction by customizing the ASNN. An inherent limitation of the ASNN arises from its treatment of each observation as a one-dimensional signal, where a signal containing outliers may resemble a shifted version of another signal (Fig. 1(b)), leading to nearly identical amplitude spectra. Consequently, the precise separation of principal components and outliers becomes challenging. To address this issue, we propose a modified version of the ASNN called convolutional multidimensional ASNN (CMASNN). The CMASNN combines multiple observations into a multidimensional signal in a convolutional fashion and evaluates the low-rankness of its amplitude spectra (Fig. 1(c)). By unifying multidimensional signals that incorporate shifted principal components and outliers, the probability of signal matching under circular shifts is reduced. Experimental results demonstrate that replacing the ASNN with the CMASNN enables more accurate extraction of principal components.

The rest of this paper is organized as follows. Sec. II reviews the fundamentals of convex optimization and the ASNN/FRPCA. Sec. III introduces the proposed OMdASNN. The proposed methods are evaluated in Sec. IV. Finally, this paper concludes in Sec. V.

II. Preliminaries

A. Notations

Bold-faced lower-case letters and upper-case letters are vectors and matrices, respectively. The other mathematical notations are summarized in Table I.

B. Convex Function

A function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be a convex (scalar-valued) function [18] if and only if, for ${}^\forall\mathbf{x},\ \mathbf{y} \in \mathbb{R}^N$ and ${}^\forall\alpha \in [0,1]$, $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$.

C. Epigraph

Epigraph of a function $f \in \Gamma_0(\mathbb{R}^N)$ ($\Gamma_0(\mathbb{R}^N)$ is the set of proper lower semicontinuous convex functions on $\mathbb{R}^N$) is a subset of the product space $\mathrm{epi}_f \subset \mathbb{R}^N \times \mathbb{R}$ defined by $\mathrm{epi}_f := \{(\mathbf{x}, \xi) \in \mathbb{R}^N \times \mathbb{R} \mid f(\mathbf{x}) \leq \xi\}$ [19].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

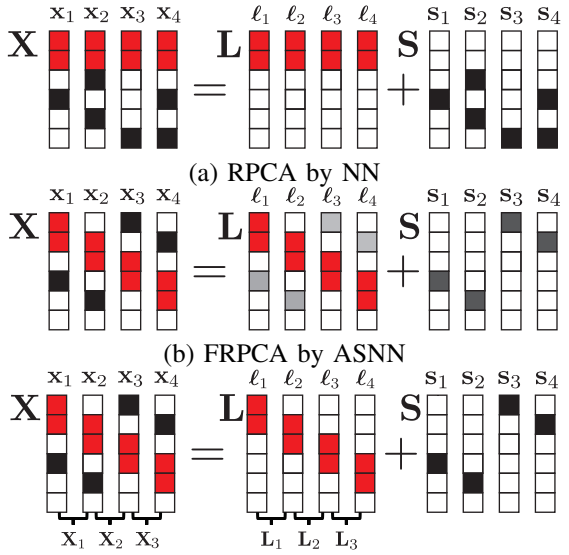


Fig. 1: Conventional and proposed framework of PCE (**L**, red) from observation (**X**) with outliers (**S**, black).

TABLE I: Basic notations

Notation	Terminology
A^N and $A^{M\times N}$ $(A \subset \mathbb{C})$	N- and $M \times N$-dimensional vectors/matrices with elements in A
$\mathbf{I}$, $\mathbf{J}$, $\mathbf{O}$	Identity, reversal identity, and zero matrix
$\mathbf{X}^\top$	Transpose of $\mathbf{X}$
$\otimes$	Kronecker product
x_n and $[\mathbf{x}]_n$	n-th element of a vector $\mathbf{x}$
$X_{m,n}$ and $[\mathbf{X}]_{m,n}$	(m,n)-th element of a matrix $\mathbf{X}$
$\lvert\mathbf{x}\rvert \in \mathbb{R}^N$, $\lvert\mathbf{X}\rvert \in \mathbb{R}^{M\times N}$ $(\mathbf{x} \in \mathbb{C}^N, \mathbf{X} \in \mathbb{C}^{M\times N})$	Element-wise absolute value, $[\lvert\mathbf{x}\rvert]_n = \lvert x_n\rvert$, $[\lvert\mathbf{X}\rvert]_{m,n} = \lvert X_{m.n}\rvert$
$\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{MN}$	Vectorization of $\mathbf{X} \in \mathbb{R}^{M\times N}$, $x_{Mn+m} = X_{m,n}$
$\mathrm{vec}^{-1}(\mathbf{x}) \in \mathbb{R}^{M\times N}$	Matricization of $\mathbf{x} \in \mathbb{R}^{MN}$, $X_{m,n} = x_{Mn+m}$
$\lVert\mathbf{x}\rVert_p$ $(p \in [1,\infty))$	ℓ_p-norm, $\lVert\mathbf{x}\rVert_p = \left(\sum_{n=1}^N \lvert x_n\rvert^p\right)^{\frac{1}{p}}$
$\mathcal{B}_p(\mathbf{y},\epsilon)$ $p \in [1,\infty]$	$\mathbf{y}$-centered ℓ_p-norm ball with the radius of ϵ, $\mathcal{B}_p(\mathbf{y},\epsilon) = \{\mathbf{x} \in \mathbb{R}^N \mid \lVert\mathbf{x}-\mathbf{y}\rVert_p \le \epsilon\}$
$\lVert\mathbf{X}\rVert_*$ or $\lVert\mathbf{x}\rVert_*$	Nuclear norm, $\lVert\mathbf{X}\rVert_* = \sum_{r=1}^R \sigma_r(\mathbf{X})$ or $\lVert\mathbf{x}\rVert_* = \lVert\mathrm{vec}^{-1}(\mathbf{x})\rVert_*$ ($\sigma_r(\mathbf{X})$: r-th singular value of $\mathbf{X}$)
$\lVert\mathbf{X}\rVert_{\mathrm{op}}$	Induced norm (Schatten ∞-matrix norm), $\lVert\mathbf{X}\rVert_{\mathrm{op}} = \lVert\mathbf{X}\rVert_{S_\infty} = \max_{\lVert\mathbf{v}\rVert_2\le 1} \lVert\mathbf{X}\mathbf{v}\rVert_2$
$\mathbf{x} \le \mathbf{y}$ $(\mathbf{x},\mathbf{y} \in \mathbb{R}^N)$	$x_n \le y_n$ $(1 \le {}^\forall n \le N)$.

D. Indicator Function

ι_C is the indicator function of C defined by $\iota_C(\mathbf{x}) := 0$ $(\mathbf{x} \in C)$, $\iota_C(\mathbf{x}) := \infty$ $(\mathbf{x} \notin C)$.

E. Proximity Operator and Projection onto Convex Set

The proximity operator, denoted as $\mathrm{prox}_{\gamma f} : \mathbb{R}^N \to \mathbb{R}^N$, is defined for a function $f \in \Gamma_0(\mathbb{R}^N)$ and an index $\gamma \in (0,\infty)$ by $\mathrm{prox}_{\gamma f}(\mathbf{x}) := \underset{\mathbf{y}\in\mathbb{R}^N}{\mathrm{argmin}}\, \gamma f(\mathbf{y}) + \frac{1}{2}\lVert\mathbf{x}-\mathbf{y}\rVert_2^2$.

The projection onto a convex set $C \subset \mathbb{R}^N$, denoted as $\mathcal{P}_C(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}^N$, is one the special cases of the proximity operator, which can be represented as $\mathrm{prox}_{\gamma\iota_C}(\mathbf{x}) := \underset{\mathbf{y}\in\mathbb{R}^N}{\mathrm{argmin}}\, \gamma\iota_C(\mathbf{y}) + \frac{1}{2}\lVert\mathbf{x}-\mathbf{y}\rVert_2^2 = \underset{\mathbf{y}\in C}{\mathrm{argmin}}\, \lVert\mathbf{x}-\mathbf{y}\rVert_2^2 =: \mathcal{P}_C(\mathbf{x})$.

TABLE II: Proximity operator and computational complexity

Function f	Proximity operator $\widetilde{\mathbf{x}} = \mathrm{prox}_{\gamma f}(\mathbf{x})$ and computational complexity for N- and $M\times N$-dimensional vectors/matrices
$\lVert\cdot\rVert_1$	Soft-thresholding $\widetilde{\mathbf{x}} = \mathcal{T}_\gamma(\mathbf{x})$ [19], $\mathcal{O}(N)$ $(\widetilde{\mathbf{x}}_n = [\mathcal{T}_\gamma(\mathbf{x})]_n = \mathrm{sign}(x_n)\max\{\lvert x_n\rvert - \gamma, 0\}$)
$\iota_{\mathcal{B}_1(\mathbf{y},\epsilon)}(\cdot)$	Using projection onto the simplex [21], $\mathcal{O}(N\log N)$.
$\lVert\mathbf{X}\rVert_*$	Soft-thresholding on singular values of $\mathbf{X} \in \mathbb{R}^{M\times N}$ [19], $\mathcal{O}(MN\min\{M,N\})$ $\boldsymbol{\sigma} = [\sigma_1 \ \dots \ \sigma_r]^\top$ $(\mathbf{X} = \mathbf{U}\mathrm{diag}(\boldsymbol{\sigma})\mathbf{V}^\top)$, $\widetilde{\mathbf{X}} = \mathbf{U}\mathrm{diag}(\mathcal{T}_\gamma(\boldsymbol{\sigma}))\mathbf{V}^\top$.

F. Projection onto the Epigraph of ℓ_2-norm

For ${}^\forall(\mathbf{x},\xi) \in \mathbb{R}^N \times \mathbb{R}$, the projection onto the epigraph of the (scaled) ℓ_2-norm ($\tau\lVert\cdot\rVert_2$, $\tau > 0$) is expressed as [20]:

$$\mathcal{P}_{\mathrm{epi}_{\tau\lVert\cdot\rVert_2}}(\mathbf{x},\xi) = \begin{cases} (\mathbf{x},\xi) & (\tau\lVert\mathbf{x}\rVert_2 \le \xi) \\ (\mathbf{0},0) & (\lVert\mathbf{x}\rVert_2 < -\tau\xi) \\ \alpha\,(\mathbf{x}, \tau\lVert\mathbf{x}\rVert_2) & (\text{otherwise}) \end{cases}, \quad (1)$$

where $\alpha = \frac{1}{1+\tau^2}\left(1 + \frac{\tau\xi}{\lVert\mathbf{x}\rVert_2}\right)$.

G. Fenchel-Rockafellar conjugate function and its proximity operator

For ${}^\forall f \in \Gamma_0(\mathbb{R}^p)$, the conjugate function f^* of f is defined as $f^*(\boldsymbol{\xi}) := \sup_{\mathbf{x}\in\mathbb{R}^N} \langle\mathbf{x},\boldsymbol{\xi}\rangle - f(\mathbf{x})$, and the proximity operator of the conjugate function is calculated as: $\mathrm{prox}_{\gamma f^*}(\mathbf{x}) = \mathbf{x} - \gamma\mathrm{prox}_{\frac{1}{\gamma}f}\left(\frac{1}{\gamma}\mathbf{x}\right)$ [22].

H. Primal-dual Splitting

Consider the following convex optimization problem to find

$$\mathbf{x}^\star \in \underset{\mathbf{x}\in\mathbb{R}^N}{\mathrm{argmin}}\, G(\mathbf{x}) + H(\mathbf{A}\mathbf{x}) \ (\neq \emptyset), \quad (2)$$

where $G \in \Gamma_0(\mathbb{R}^N)$, $H \in \Gamma_0(\mathbb{R}^M)$, and $\mathbf{A} \in \mathbb{R}^{M\times N}$, respectively. We assume that the solution set of (2) is nonempty. Then the PDS algorithm [23]–[27] for solving (2) is given as follows:

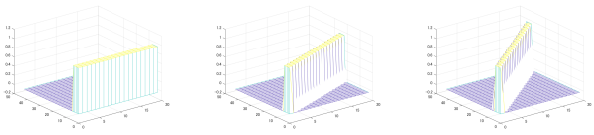
$$\begin{cases} \mathbf{x}^{(n+1)} = \mathrm{prox}_{\gamma_1 G}[\mathbf{x}^{(n)} - \gamma_1\mathbf{A}^\top\mathbf{z}^{(n)}] \\ \mathbf{z}^{(n+1)} = \mathrm{prox}_{\gamma_2 H^*}[\mathbf{z}^{(n)} + \gamma_2\mathbf{A}(2\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)})] \end{cases}, \quad (3)$$

where prox denotes the proximity operator and H^* is the conjugate function of H.

According to Theorem 3.3 in [24], for given $\gamma_1 > 0$ and $\gamma_2 > 0$ satisfying $\gamma_1\gamma_2\lVert\mathbf{A}\rVert_{\mathrm{op}} \le 1$ (the definition of the induced norm, see Table I), the sequence of $\{\mathbf{x}^{(n)}\}_{n=0}^\infty$ in (3) converges to a minimizer $\mathbf{x}^\star$ of (2). Thanks to the convexity nature, the initial value of $\mathbf{x}^{(0)}$ and $\mathbf{z}^{(0)}$ can be chosen arbitrarily.

I. Amplitude Spectrum Nuclear Norm and Its Application to Frequency-domain RPCA

This section reviews the ASNN for PCE [17]. Low-rank modeling is one of the most effective approaches for PCE [8]–[16], and robust PCA (RPCA) [8], [9], [14], [16] is known as the standard technique that can extract principal components lying in a low-dimensional subspace from an observed matrix in the presence of outliers. An example of 1-dimensional (1D)



(a) Shift:0, $\widehat{\mathbf{L}}_0$ (b) Shift:1, $\widehat{\mathbf{L}}_1$ (c) Shift:2, $\widehat{\mathbf{L}}_2$

Fig. 2: Examples of target signals.

principal components $\widehat{\mathbf{L}}_0 \in \mathbb{R}^{25\times 20}$ is shown in Fig. 2(a), where m-th column of $\widehat{\mathbf{L}}_s$ is set as

$$[\widehat{\mathbf{L}}_s]_{m,n} = \begin{cases} 1 & (sn \le m \le sn+4) \\ 0 & (\text{otherwise}) \end{cases}. \tag{4}$$

The problem formulation for RPCA is given as follows [9]:

$$\{\mathbf{L}^\star,\ \mathbf{S}^\star\} \in \underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} \|\mathbf{L}\|_* + \phi(\mathbf{S}) \ \text{ s.t. } \ \mathbf{X} = \mathbf{L} + \mathbf{S}, \tag{5}$$

where ϕ is a sparsity-promoting function, e.g., the ℓ_1-norm and the indicator function of the ℓ_1-norm ball constraint[1].

One of the drawback of RPCA is the sensitivity against the shift or the misalignment of target components. Fig. 2(b) and (c) show target components shifted every one and two sample $\widehat{\mathbf{L}}_1,\ \widehat{\mathbf{L}}_2 \in \mathbb{R}^{25\times 20}$, respectively. When target components are not aligned exactly due the shift (Fig. 2(b) and (c)), they are not lying low-dimensional subspace any more. Consequently, RPCA cannot extract the principal components precisely.

To resolve this drawback, we utilize the fundamental signal processing theory: the amplitude spectra of an original and its (circularly) shifted signals are the same. For example, for given $\widehat{\mathbf{L}}_s = \begin{bmatrix}\widehat{\boldsymbol{\ell}}_{s,1} & \cdots & \widehat{\boldsymbol{\ell}}_{s,N}\end{bmatrix} \in \mathbb{R}^{M\times N}$, $|\mathbf{W}_M \widehat{\boldsymbol{\ell}}_{s,n_1}| = |\mathbf{W}_M \widehat{\boldsymbol{\ell}}_{s,n_2}|$, where $\mathbf{W}_M \in \mathbb{C}^{M\times M}$ denotes the nomalized DFT matrix ($[\mathbf{W}_M]_{m_1,m_2} = \frac{1}{\sqrt{M}}\exp\left(-\mathrm{j}\frac{2\pi}{M}m_1 m_2\right)$) and the definition $|\cdot|$ for vectors and matrices is given in Table I. According to this nature, we design the ASNN $\|\cdot\|_{\mathrm{ASNN}} : \mathbb{R}^{M\times N} \to \mathbb{R}$ for a set of 1D signals $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_N\end{bmatrix}$ as:

$$\|\mathbf{X}\|_{\mathrm{ASNN}} := \||\mathbf{W}_M \mathbf{X}|\|_*, \tag{6}$$

where $|\mathbf{W}_M\mathbf{X}|$ denotes the amplitude spectra of input 1D signals $\mathbf{X}$. The ASNN takes small value when columns in $\mathbf{X}$ have an isomorphic structure even in the presence of shift. The ASNN can be considered as a 2-layered composite function as follows:

$$\begin{aligned} \||\mathbf{W}_M\mathbf{X}|\|_* &= \||(\mathbf{C}_M - \mathrm{j}\mathbf{S}_M)\mathbf{X}|\|_* \\ &= \left\|\mathbf{P}^{(1)}\begin{bmatrix}\mathbf{C}_M^\top & -\mathbf{S}_M^\top\end{bmatrix}^\top \mathbf{X}\right\|_{2,*} \\ &= \|\mathbf{T}\mathbf{L}\|_{2,*}, \end{aligned} \tag{7}$$

where $[\mathbf{C}_M]_{m_1,m_2} = \frac{1}{\sqrt{M}}\cos\left(\frac{2\pi}{M}m_1m_2\right)$, $[\mathbf{S}_M]_{m_1,m_2} = \frac{1}{\sqrt{M}}\sin\left(\frac{2\pi}{M}m_1m_2\right)$, $\mathbf{P}^{(1)}$ is a permutation matrix, and $\|\cdot\|_{2,*} : \mathbb{R}^{2M\times N} \to \mathbb{R}$ is defined as:

$$\begin{aligned} \mathbf{X} &:= \begin{bmatrix}\mathbf{x}_{1,1} & \cdots & \mathbf{x}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{M,1} & \cdots & \mathbf{x}_{M,N}\end{bmatrix}, \ (\mathbf{x}_{m,n} \in \mathbb{R}^2) \\ \|\mathbf{X}\|_{2,*} &:= \|g_{\|\cdot\|_2}(\mathbf{X})\|_*, \\ g_{\|\cdot\|_2}(\mathbf{X}) &:= \begin{bmatrix}\|\mathbf{x}_{1,1}\|_2 & \cdots & \|\mathbf{x}_{1,N}\|_2 \\ \vdots & \ddots & \vdots \\ \|\mathbf{x}_{M,1}\|_2 & \cdots & \|\mathbf{x}_{M,N}\|_2\end{bmatrix}. \end{aligned} \tag{8}$$

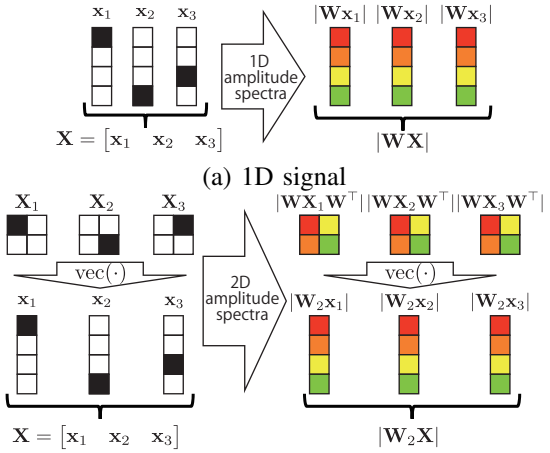
As shown in [17], since the nuclear norm is not a non-decreasing function on $\mathbb{R}_+^{M\times N}$, $\|\cdot\|_{2,*}$ is not even a convex function.

The ASNN for 1D signals can be generalized to the one for MD signals $\{\mathcal{X}_n\}_{n=1}^N \subset \mathbb{R}^{M_1\times M_2\times\cdots\times M_D}$, which is denoted as multidimensional ASNN (MASNN) [17] with multidimensional separable DFT $\mathbf{W}^{[D]}_{\{M_d\}_{d=D}^1}$, as in the following:

$$\begin{aligned} \|\mathbf{X}\|_{\mathrm{MASNN}} &:= \||\mathbf{W}^{[D]}_{\{M_d\}_{d=D}^1}\mathbf{X}|\|_* = \|\mathbf{T}\mathbf{X}\|_{2,*}, \\ \mathbf{X} &= \begin{bmatrix}\mathrm{vec}(\mathcal{X}_1) & \cdots & \mathrm{vec}(\mathcal{X}_N)\end{bmatrix} \in \mathbb{R}^{\Pi_{d=1}^D M_d\times N}, \\ \mathbf{W}^{[D]}_{\{M_d\}_{d=D}^1} &= \mathbf{W}_{M_D}\otimes\cdots\otimes\mathbf{W}_{M_1}, \\ \mathbf{T} &= \mathbf{P}^{(2)}\begin{bmatrix}\mathbf{T}_r^\top & \mathbf{T}_i^\top\end{bmatrix}^\top, \end{aligned} \tag{9}$$

where $\mathbf{P}^{(2)}$ is a permutation matrix and $\mathbf{T}_r$ and $\mathbf{T}_i$ are the combination of the Kronecker products of $\mathbf{C}_{M_d}$ and $\mathbf{S}_{M_d}$. For example $D = 2$, the construction of $|\mathbf{W}^{[2]}_{\{M_2,M_1\}}\mathbf{X}|$ is illustrated in Fig. 3(b), where

$$\begin{aligned} \mathbf{W}^{[2]}_{\{M_2,M_1\}} &= \mathbf{W}_{M_2}\otimes\mathbf{W}_{M_1} \\ &= (\mathbf{C}_{M_2} - j\mathbf{S}_{M_2})\otimes(\mathbf{C}_{M_1} - j\mathbf{S}_{M_1}), \\ \mathbf{T}_r &= (\mathbf{C}_{M_2}\otimes\mathbf{C}_{M_1}) - (\mathbf{S}_{M_2}\otimes\mathbf{S}_{M_1}), \\ \mathbf{T}_i &= -((\mathbf{C}_{M_2}\otimes\mathbf{S}_{M_1}) + (\mathbf{S}_{M_2}\otimes\mathbf{C}_{M_1})). \end{aligned} \tag{10}$$

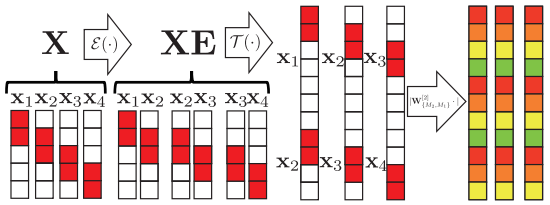


(a) 1D signal

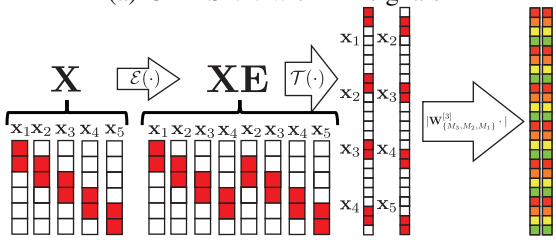
(b) 2D signal

Fig. 3: ASNN for multidimensional signals.

[1]Although the original RPCA [9] employs the ℓ_1-norm as ϕ, we use the ℓ-norm ball constraint for experiments in Sec. IV for fair performance comparison between nuclear norm, the ASNN, and the CMASNN.



(a) CMASNN with 2D signals



(b) CMASNN with 3D signals

Fig. 4: CMASNN

III. Convolutional Multidimensional Amplitude Spectrum Nuclear Norm for Frequency-domain Robust Principal Component Analysis

A. Definition of Overlapped Multidimensional Amplitude Spectrum Nuclear Norm

In this section, we extend the conventional MASNN to the CMASNN for more robust PCE. As mentioned in Sec. I, a limitation of FRPCA when using the conventional ASNN is its tendency to struggle in distinguishing principal components from outliers. This occurs when observed principal components, along with outliers, closely resemble shifted versions of other observations, as depicted in Fig. 1. To enhance the accuracy of PCE, we unify multiple observations into a higher-dimensional signal with overlapping and evaluate the low-rankness of its amplitude spectra using the ASNN. Essentially, when two high-dimensional signals are shifted versions of one another, their amplitude spectra exhibit identical patterns. Furthermore, the introduction of a higher-dimensional signal is expected to mitigate the possibility of principal components with outliers aligning under circular shifts. With this in mind, we introduce the CMASNN in the following way:

$$\begin{aligned}\|\{\mathcal{X}_n\}_{n=1}^N\|_{\mathrm{CMASNN}} &:= \||\mathbf{W}^{[D]}_{\{M_d\}_{d=D}^1}\mathcal{T}(\mathcal{E}(\{\mathcal{X}_n\}_{n=1}^N)|\|_* \\ &= \left\|\mathbf{T}\mathcal{T}(\mathcal{E}(\{\mathcal{X}_n\}_{n=1}^N))\right\|_{2,*}, \end{aligned} \tag{11}$$

where $\mathbf{W}^{[D]}_{\{M_d\}_{d=D}^1}$ and $\mathbf{T}$ are the same matrices in Sec. II-I and the operator $\mathcal{E}$ duplicates signals in $\{\mathcal{X}_n\}$ and $\mathcal{T}$ is an unifying operator that unifies several signals into a (vectorized) higher-dimensional tensor. Fig. 4 shows an example of the CMASNN which embeds several 1D signals into a 2D or 3D signal in a convolutional fashion and calculates the 2D or 3D amplitude spectra, where $\{\mathcal{X}_n\}_{n=1}^N$ is simply denoted as the matrix $\mathbf{X}$ and $\mathcal{E}(\mathbf{X}) = \mathbf{XE}$ ($\mathbf{E}$ duplicates columns of $\mathbf{X}$).

B. Algotithm for Convolutional Multidimensional Amplitude Spectrum Nuclear Norm Minimization with PDS

Now we modify the cost function of RPCA by using the CMASNN, which is termed as frequency-domain RPCA ($\mathcal{F}$RPCA). For simplicity, we consider the case of PCE from 1D observations using CMASNN with multidimensional tensors as:

$$\begin{aligned}\{\mathbf{L}^\star,\ \mathbf{S}^\star\} \in \underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} &\|\mathbf{L}\|_{\mathrm{CMASNN}} + \phi(\mathbf{S}) \ \text{ s.t. } \ \mathbf{X} = \mathbf{L} + \mathbf{S} \\ = \underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} &\||\mathbf{W}^{[D]}_{\{M_d\}_{d=D}^1}\mathcal{T}(\mathbf{LE})|\|_* \\ &+ \ \phi(\mathbf{S}) + \iota_{\{\mathbf{X}\}}(\mathbf{L}+\mathbf{S}) \\ = \underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} &\|\mathbf{T}\mathcal{T}(\mathbf{LE})\|_{2,*} + \widetilde{\phi}(\mathbf{L},\mathbf{S},\mathbf{X}) \end{aligned} \tag{12}$$

where $\widetilde{\phi}(\mathbf{L},\mathbf{S},\mathbf{X}) = \phi(\mathbf{S}) + \iota_{\{\mathbf{X}\}}(\mathbf{L}+\mathbf{S})$. Unfortunately the above optimization problem is nonconvex due to nonconvexity of $\|\cdot\|_{2,*}$. Furthermore, the closed-form proximity operator of $\|\cdot\|_{2,*}$ cannot be calculated trivially. Thus, we find an approximated solution for the problem (12) by using epigraphical relaxation (ERx) [17] which decouple a non-proximable composite function into a outermost norm and an epigrahical constraint as follows:

$$\begin{aligned}&\underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} \|\mathbf{Z}\|_* + \widetilde{\phi}(\mathbf{L},\mathbf{S},\mathbf{X}) \text{ s.t. } g_{\|\cdot\|_2}(\mathbf{T}\mathcal{T}(\mathbf{LE})) = \mathbf{Z}, \\ \xrightarrow{\text{ERx}}\ &\underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} \|\mathbf{Z}\|_* + \widetilde{\phi}(\mathbf{L},\mathbf{S},\mathbf{X}) \text{ s.t. } g_{\|\cdot\|_2}(\mathbf{T}\mathcal{T}(\mathbf{LE})) \le \mathbf{Z}, \\ =\ &\underset{\mathbf{L},\ \mathbf{S}}{\operatorname{argmin}} \|\mathbf{Z}\|_* + \widetilde{\phi}(\mathbf{L},\mathbf{S},\mathbf{X}) + \iota_{\mathrm{epi}_{\|\cdot\|_2}}(\mathbf{T}\mathcal{T}(\mathbf{LE}),\mathbf{Z}), \end{aligned} \tag{13}$$

Finally we can apply PDS by reformulating the above optimization with vector notations as:

$$\{\widetilde{\boldsymbol{\ell}},\ \widetilde{\mathbf{s}}\} \in \underset{\boldsymbol{\ell},\ \mathbf{s}}{\operatorname{argmin}} \|\mathbf{z}\|_* + \widetilde{\phi}(\boldsymbol{\ell},\mathbf{s},\mathbf{x}) + \iota_{\mathrm{epi}_{\|\cdot\|_2}}\left(\widehat{\mathbf{T}}\boldsymbol{\ell},\mathbf{z}\right), \tag{14}$$

where $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, $\boldsymbol{\ell} = \mathrm{vec}(\mathbf{L})$, $\mathbf{s} = \mathrm{vec}(\mathbf{S})$, and $\widehat{\mathbf{T}} = \mathbf{I}\otimes\mathbf{T}\widetilde{\mathbf{E}}$. The matrix $\widetilde{\mathbf{E}}$ duplicates subvectors in the vectorized data $\boldsymbol{\ell}$. For example, let $\mathbf{L} = \begin{bmatrix}\boldsymbol{\ell}_1 & \boldsymbol{\ell}_2 & \boldsymbol{\ell}_3 & \boldsymbol{\ell}_4\end{bmatrix}$ and consider the case shown in Fig. 4(a). Then, the matrix $\widetilde{\mathbf{E}}$ maps $\boldsymbol{\ell} = \mathrm{vec}(\mathbf{L}) = \begin{bmatrix}\boldsymbol{\ell}_1^\top & \cdots & \boldsymbol{\ell}_4^\top\end{bmatrix}^\top$ to $\begin{bmatrix}\boldsymbol{\ell}_1^\top & \boldsymbol{\ell}_2^\top & \boldsymbol{\ell}_2^\top & \boldsymbol{\ell}_3^\top & \boldsymbol{\ell}_3^\top & \boldsymbol{\ell}_4^\top\end{bmatrix}^\top$.

Since the objective function in the last line of (14) is convex, the global minimizer can be found by the PDS algorithm. Note that the problem (14) is obtained by epigraphical relaxation, the minimizer for (13) is not the one for (12). The minimization problem can be solved by the PDS algorithm as

$$\begin{aligned}&\mathbf{p} = \begin{bmatrix}\boldsymbol{\ell}^\top & \mathbf{s}^\top & \mathbf{z}^\top\end{bmatrix}^\top,\ G(\mathbf{p}) = \mathbf{0}, \\ &H(\mathbf{q}) = \|\mathbf{q}_1\|_* + \phi(\mathbf{q}_2) + \iota_{\{\mathbf{x}\}}(\mathbf{q}_3) + \iota_{\mathrm{epi}_{\|\cdot\|_2}}(\mathbf{q}_4,\mathbf{q}_5), \\ &\mathbf{q} = \mathbf{A}\mathbf{p},\ \mathbf{A} = \begin{bmatrix}\mathbf{O} & \mathbf{O} & \mathbf{I} & \widehat{\mathbf{T}}^\top & \mathbf{O} \\ \mathbf{O} & \mathbf{I} & \mathbf{I} & \mathbf{O} & \mathbf{O} \\ \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{I}\end{bmatrix}^\top. \end{aligned} \tag{15}$$

The detailed algorithm is summarized in Algorithm 1, where the proximity operator of the nuclear norm $\|\cdot\|_*$ is in Table II and that of $\iota_{\mathrm{epi}_{\|\cdot\|_2}}$ is computed by (1).

Algorithm 1 Solver for (13)

set $n = 0$ and choose initial parameters for $\mathbf{p}^{(0)}$, $\mathbf{q}^{(0)}$.
while $\|\mathbf{p}^{(n)} - \mathbf{p}^{(n-1)}\|_2 > \epsilon_{\rm stop}$ **do**
  $\mathbf{p}^{(n+1)} = \mathbf{p}^{(n)} - \gamma_1 \mathbf{A}^\top \mathbf{q}^{(n)}$
  $\mathbf{t}^{(n)} = \mathbf{q}^{(n)} + \gamma_2 \mathbf{A}(2\mathbf{p}^{(n+1)} - \mathbf{p}^{(n)})$
  $\mathbf{q}_1^{(n+1)} = \mathbf{t}_1^{(n)} - \gamma_2 \mathrm{prox}_{\frac{1}{\gamma_2}\|\cdot\|_*}\left(\frac{1}{\gamma_2}\mathbf{t}_1^{(n)}\right)$.
  $\mathbf{q}_2^{(n+1)} = \mathbf{t}_2^{(n)} - \gamma_2 \mathrm{prox}_{\frac{1}{\gamma_2}\phi}\left(\frac{1}{\gamma_2}\mathbf{t}_2^{(n)}\right)$.
  $\mathbf{q}_3^{(n+1)} = \mathbf{t}_3^{(n)} - \gamma_2 \mathrm{prox}_{\frac{1}{\gamma_2}\iota_{\{\mathbf{x}\}}}\left(\frac{1}{\gamma_2}\mathbf{t}_3^{(n)}\right)$.
  $(\mathbf{q}_4^{(n+1)}, \mathbf{q}_5^{(n+1)}) = (\mathbf{t}_4^{(n)}, \mathbf{t}_5^{(n)}) - \gamma_2 \mathrm{prox}_{\frac{1}{\gamma_2}\mathrm{epi}\|\cdot\|_2}\left(\frac{1}{\gamma_2}(\mathbf{t}_4^{(n)}, \mathbf{t}_5^{(n)})\right)$.
  $n = n + 1$.
end while
Output $\mathbf{x}^{(n)}$.

IV. Experimental Results

This section compares the performance of FRPCA by the CMASNN presented in Sec. III with RPCA by the NN and FRPCA by the ASNN. Two types of the proposed method, one is the CMASNN with 2D signals (Fig. 4(a)) and the other is the CMASNN with 3D signals (Fig. 4(b)) are evaluated (both methods are denoted as CMASNN-2D and CMASNN-3D, respectively). In this experiment, we extract target components with an isomorphic structure $\widehat{\mathbf{L}}_s$ from an observed signal corrupted with sparse noise as $\mathbf{X} = \widehat{\mathbf{L}}_s + \widehat{\mathbf{S}}_s$. The target components used in this experiment are $\widehat{\mathbf{L}}_0$, $\widehat{\mathbf{L}}_1$, and $\widehat{\mathbf{L}}_2$ given in Fig. 2. The sparse noise $\widehat{\mathbf{S}}_s \in \{0,1\}^{43\times 20}$ contains 1s in the region of $\mathcal{N}_s = \{(m,n) \in \{1,2,\ldots,43\}\times\{1,2,\ldots,20\} \mid [L_s]_{m,n} = 0\}$ with the probability of $p = 0.05, 0.1, 0.15$. We apply RPCA, FRPCA-ASNN, FRPCA-CMASNN-2D, and FRPCA-MASNN-3D to extract the target component $\widehat{\mathbf{L}}_s$. ϕ is set to the ℓ_1-ball constraint as $\phi(\mathbf{s}) = \mathcal{B}_1(\mathbf{s}, \epsilon)$ where ϵ is set as the oracle value, i.e., $\epsilon = \|\widehat{\mathbf{S}}_s\|_1$. The stop criteria in Algorithm 1 is $\epsilon_{\rm stop} = 10^{-5}$.

Fig. 5 shows the estimated components $\mathbf{L}_s^\star$ and $\mathbf{S}_s^\star$ ($p = 0.1$) obtained from the proposed and conventional methods, respectively. Note that, in Fig. 5, the estimates are flipped as $\mathbf{J}\widehat{\mathbf{L}}_0$ and $\mathbf{J}\widehat{\mathbf{S}}_0$ for clearer visualization. In the case of the shift of 0, the target components $\widehat{\mathbf{L}}_0$ are lying in 1-dimensional subspace. In this case, RPCA decomposes the target components and the noise components almost perfectly by minimizing the nuclear norm. On the other hand, the results from FRPCA-CMASNN-2D/FRPCA-CMASNN-3D are worse than RPCA but outperforms FRPCA-ASNN because the CMASNN involving multidimensional ASNN enables us to distinguish principal components and outliners more robustly. In the case of the shift of 1 and 2, FRPCA-CMASNN-2D/FRPCA-CMASNN-3D can estimate the original component $\widehat{\mathbf{L}}_s$ much better than RPCA/FRPCA-ASNN due to the enhancement of low-rankness promoting for amplitude spectrum. As well as subjective quality, FRPCA-CMASNN-2D or FRPCA-CMASNN-3D achieved higher PSNR than RPCA and FRPCA-ASNN as shown in Table III.

TABLE III: Numerical results (PSNR [dB])

		$p = 0.05$	$p = 0.1$	$p = 0.15$
Shift:0	RPCA	**40.75**	**30.19**	**30.39**
	FRPCA-ASNN	29.16	22.96	20.54
	FRPCA-CMASNN-2D	34.74	25.20	23.01
	FRPCA-CMASNN-3D	37.45	24.65	23.07
Shift:1	RPCA	16.71	14.89	13.98
	FRPCA-ASNN	25.96	21.24	20.19
	FRPCA-CMASNN-2D	**29.72**	**23.24**	**23.75**
	FRPCA-CMASNN-3D	28.10	21.70	21.29
Shift:2	RPCA	15.83	12.98	11.52
	FRPCA-ASNN	29.71	23.87	17.70
	FRPCA-CMASNN-2D	**34.05**	**27.04**	**19.31**
	FRPCA-CMASNN-3D	25.59	21.05	18.20

V. Conclusions

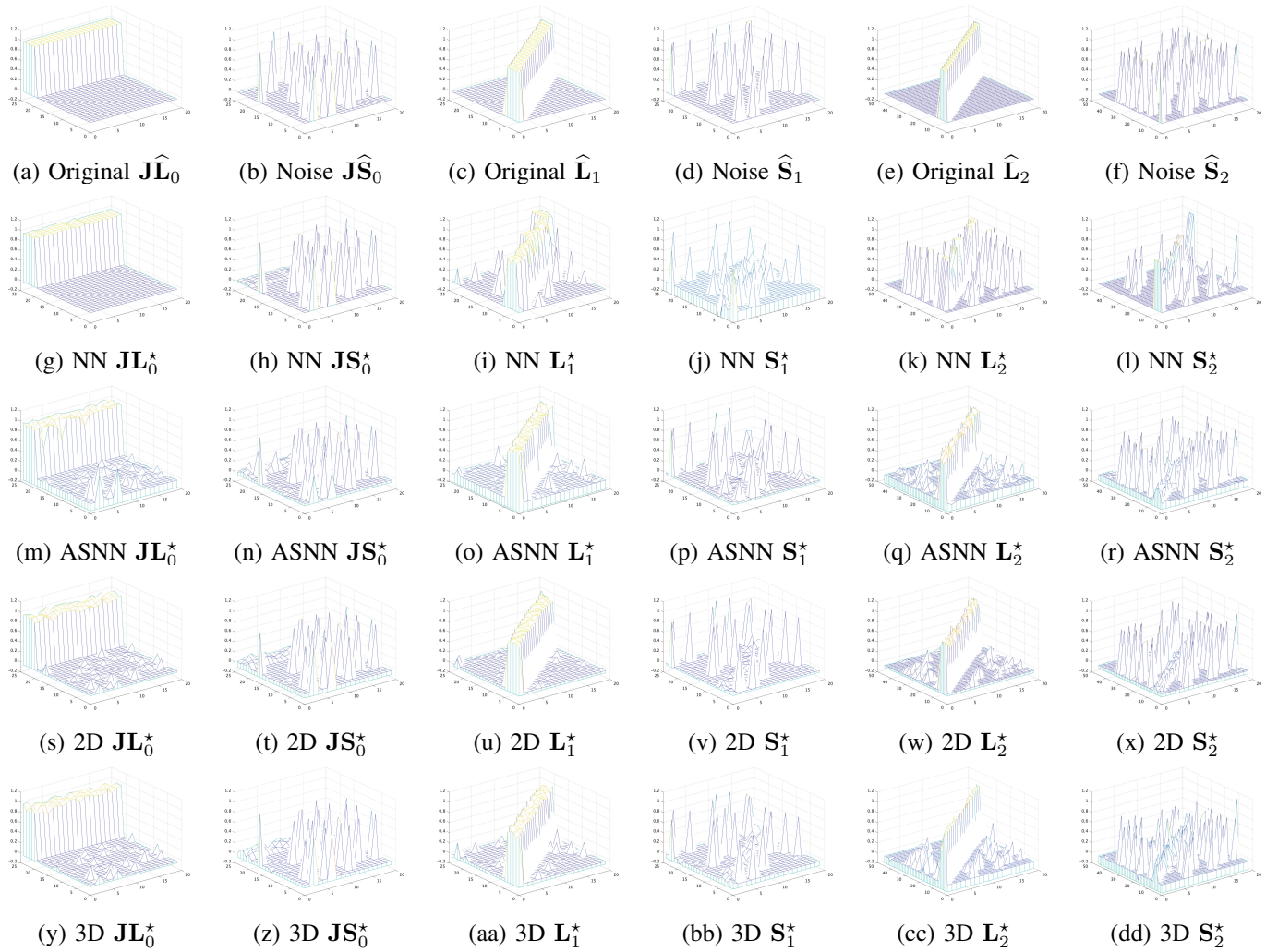
In this manuscript, we propose the use of the CMASNN for robust PCE. The conventional ASNN tends to fail in distinguishing principal components from outliers. This occurs when observed principal components, along with outliers, closely resemble shifted versions of other observations. To enhance the accuracy of PCE, we unified multiple observations into a higher-dimensional signal in a convolutional way and evaluate the low-rankness of its amplitude spectra using the ASNN. This customization aids in the discrimination between true principal components and outliers more efficiently. By substituting the CMASNN for the ASNN in FRPCA, experimental results demonstrate that the proposed CMASNN-based FRPCA achieves superior performance compared to the conventional approaches.

Acknowledgment

This work was supported in part by JSPS Grants-in-Aid (21K04045) and ROIS NII Open Collaborative Research 2023 (23S0202).

References

[1] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 1st ed. Orlando, FL, USA: Academic Press, Inc., 2015.
[2] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*. The MIT Press, 2011.
[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.
[4] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Problems*, vol. 24, no. 6, p. 065014, Nov. 2008.
[5] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 323–343, Apr. 2009.
[6] N. Pustelnik, C. Chaux, and J. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2450–2462, Sep. 2011.
[7] M. V. Afonso, J. M. B.-Dias, and M. A. T. Figueiredo, "An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, Mar. 2011.
[8] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 2080–2088.
[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, Jun. 2011.



(a) Original $\mathbf{J}\widehat{\mathbf{L}}_0$ (b) Noise $\mathbf{J}\widehat{\mathbf{S}}_0$ (c) Original $\widehat{\mathbf{L}}_1$ (d) Noise $\widehat{\mathbf{S}}_1$ (e) Original $\widehat{\mathbf{L}}_2$ (f) Noise $\widehat{\mathbf{S}}_2$

(g) NN $\mathbf{JL}_0^\star$ (h) NN $\mathbf{JS}_0^\star$ (i) NN $\mathbf{L}_1^\star$ (j) NN $\mathbf{S}_1^\star$ (k) NN $\mathbf{L}_2^\star$ (l) NN $\mathbf{S}_2^\star$

(m) ASNN $\mathbf{JL}_0^\star$ (n) ASNN $\mathbf{JS}_0^\star$ (o) ASNN $\mathbf{L}_1^\star$ (p) ASNN $\mathbf{S}_1^\star$ (q) ASNN $\mathbf{L}_2^\star$ (r) ASNN $\mathbf{S}_2^\star$

(s) 2D $\mathbf{JL}_0^\star$ (t) 2D $\mathbf{JS}_0^\star$ (u) 2D $\mathbf{L}_1^\star$ (v) 2D $\mathbf{S}_1^\star$ (w) 2D $\mathbf{L}_2^\star$ (x) 2D $\mathbf{S}_2^\star$

(y) 3D $\mathbf{JL}_0^\star$ (z) 3D $\mathbf{JS}_0^\star$ (aa) 3D $\mathbf{L}_1^\star$ (bb) 3D $\mathbf{S}_1^\star$ (cc) 3D $\mathbf{L}_2^\star$ (dd) 3D $\mathbf{S}_2^\star$

Fig. 5: Original/Noise ($p = 0.1$) (a)–(f) and estimated components obtained by RPCA-NN (g)–(l), FRPCA-ASNN (m)–(r), FRPCA-CMASNN-2D (s)–(x), and FRPCA-CMASNN-3D (y)–(dd).

[10] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013.

[11] S. Ono, T. Miyata, and I. Yamada, "Cartoon-texture image decomposition using blockwise low-rank texture characterization," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1128–1142, Mar. 2014.

[12] S. Ono and I. Yamada, "Color-line regularization for color artifact removal," *IEEE Trans. Comput. Imag.*, vol. 2, no. 3, pp. 204–217, Sep. 2016.

[13] S. Ono, K. Shirai, and M. Okuda, "Vectorial total variation based on arranged structure tensor for multichannel image restoration," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2016, pp. 4528–4532.

[14] X. Song, D. Xiang, K. Zhou, and Y. Su, "Improving RPCA-based clutter suppression in gpr detection of antipersonnel mines," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1338–1342, 2017.

[15] J. Yin, C. Unal, M. Schleiss, and H. Russchenberg, "Radar target and moving clutter separation based on the low-rank matrix optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4765–4780, 2018.

[16] F. H. C. Tivive, A. Bouzerdoum, and C. Abeynayake, "GPR target detection by joint sparse and low-rank matrix decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2583–2595, 2019.

[17] S. Kyochi, S. Ono, and I. Selesnick, "Epigraphical relaxation for minimizing layered mixed norms," *IEEE Trans. on Signal Process.*, vol. 69, pp. 2923–2938, 2021.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. USA: Cambridge University Press, 2004.

[19] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017.

[20] G. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, "Epigraphical projection and proximal tools for solving constrained convex optimization problems," *Signal, Image and Video Processing*, vol. 9, no. 8, pp. 1737–1749, Nov. 2015.

[21] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the L1-ball for learning in high dimensions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 272–279.

[22] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York, NY, USA: Springer-Verlag, 2011.

[23] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, Dec. 2010.

[24] L. Condat, "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, Aug. 2013.

[25] B. C. Vu, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Adv. Comput. Math.*, vol. 38, no. 3, pp. 667–681, Nov. 2011.

[26] N. Komodakis and J. C. Pesquet, "Playing with duality: An overview

of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 31–54, Nov. 2015.

[27] R. I. Boţ and E. R. Csetnek, "On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems," *Optimization*, vol. 64, no. 1, pp. 5–23, 2015.

Moreau Envelope ADMM for Decentralized Weakly Convex Optimization

Reza Mirzaeifard[†], Naveen K. D. Venkategowda[§], Alexander Jung[⋆], Stefan Werner[†]
[†]Department of Electronic Systems, Norwegian University of Science and Technology, Norway
[§]Department of Science and Technology, Linköping University, Sweden
[⋆]Department of Computer Science, Aalto University, Finland
E-mails: {reza.mirzaeifard, stefan.werner}@ntnu.no, naveen.venkategowda@liu.se, alex.jung@aalto.fi

***Abstract*—This paper proposes a proximal variant of the alternating direction method of multipliers (ADMM) for distributed optimization. Although the current versions of ADMM algorithm provide promising numerical results in producing solutions that are close to optimal for many convex and non-convex optimization problems, it remains unclear if they can converge to a stationary point for weakly convex and locally non-smooth functions. Through our analysis using the Moreau envelope function, we demonstrate that MADM can indeed converge to a stationary point under mild conditions. Our analysis also includes computing the bounds on the amount of change in the dual variable update step by relating the gradient of the Moreau envelope function to the proximal function. Furthermore, the results of our numerical experiments indicate that our method is faster and more robust than widely-used approaches.**

***Index Terms*—Distributed optimization, non-convex and non-smooth optimization, weakly convex functions, ADMM, Moreau envelope.**

I. Introduction

Many systems, like the internet-of-things (IoT) and cyber-physical systems, comprise distributed devices and sensors that gather data for inference and decision-making. Building distributed models in such systems without data transfer to a central hub calls for distributed optimization methods involving peer-to-peer interactions. In addition, these methods allow for coping with resource constraints, e.g., computational resources, battery power, communication bandwidth, and privacy protection [1]–[3].

There is a large body of work on distributed optimization methods from different perspectives. The most direct approach to the design of distribution optimization methods is via message-passing implementations of subgradient computation within subgradient descent methods [4]–[6]. Gradient methods are generalized to solve structured optimization problems using proximal methods [7], [8]. Additionally, variational methods for probabilistic models lend themselves naturally to optimization algorithms, such as variants of belief propagation [9]. The subgradient method is well-known for its ease of implementation, wherein a subgradient is taken at each step, followed by an average with neighbors. On the downside, subgradient descent has a sublinear convergence rate and requires tuning of the step size. Meanwhile, ADMM performs fast and accurately in many practical convex and non-convex optimization problems. For convex objective functions, subgradient methods and ADMM are guaranteed to converge to a global optimum under suitable parameter choices [6], [10], [11]. However, the analysis of these methods for non-convex problems is challenging due to the potential ill-behavior of the objective function. The convergence analysis for ADMM in non-convex problems is particularly more challenging as it requires analyzing the convergence of multiple sub-problems with different structures and assumptions.

One important family of non-convex optimization problems is weakly convex problems. Several such problems arise in machine learning, including robust phase retrieval [12], blind deconvolution [13], biconvex compressive sensing [14], and dictionary learning [15]. Smooth functions, or functions with Lipschitz continuous gradients, are weakly convex functions. Several non-convex optimization algorithms are proposed based on the smoothness assumption (e.g., [16], [17]); however, weakly convex functions are not restricted to smooth functions, and non-smooth functions can also be weakly convex [18]. Existing work on distributed optimization of weakly convex functions includes [19]. However, in the subgradient-based algorithm in [19], the local functions must satisfy the bounded subgradient assumption, and the accuracy of the estimation depends on the step size. Although ADMM is a powerful algorithm applicable to many problems, it is not currently used to solve weakly convex problems. As seen in [20]–[23], several ADMM-based works study non-convex optimization; however, these works require a smooth objective function. It is still necessary to provide a distributed ADMM-based algorithm that could work in the weakly convex setting without having any Lipschitz differentiability condition.

We propose a Moreau envelope-based ADMM (MADM), suitable for distributed optimization where local objectives are weakly convex and not necessarily smooth. We chose the Moreau envelope-based ADMM approach because it allows us to guarantee a decrease in each primal update step and bound the amount of change in the dual update step by primary variables. This is achieved by incorporating the relationship between the Moreau envelope function and proximal function. Therefore, by selecting appropriate penalty parameters under mild conditions, including weakly convexity of each

This work was supported by the Research Council of Norway.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

local function, a connected network, and the boundness of augmented Lagrangian function, we can ensure that the algorithm converges to a stationary point. The Moreau envelope-based ADMM approach stands out from other penalty-based ADMM algorithms due to its superior theoretical convergence properties. We conduct illustrative numerical experiments to verify the convergence properties of the proposed method. The experiments demonstrate the robustness of the proposed algorithm when we fix the penalty parameters and step size, and the problem structure remains the same. Unlike subgradient-based methods, the MADM approach ensures faster and more reliable convergence in this setting.

Mathematical Notations: Scalars, column vectors, and matrices are respectively denoted by lowercase, bold lowercase, and bold uppercase letters. The operator $(\cdot)^{\mathrm{T}}$ denotes transpose of a matrix, and the jth column of matrix $\mathbf{A}$ is denoted by $\mathbf{A}_j$. The set $\{1, \cdots, L\}$ is denoted by $[L]$. For a function $h : \mathbb{R}^n \to \mathbb{R}$ and penalty parameter $\gamma > 0$, $\mathcal{M}_h(\mathbf{w};\gamma) = \min_{\mathbf{x}} \left\{h(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{w}\|_2^2\right\}$ is the Moreau envelope function [24], and $\mathrm{Prox}_h(\mathbf{w};\gamma) = \arg\min_{\mathbf{x}} \left\{h(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{w}\|^2\right\}$ is its associated proximal operator.

II. Problem formulation

Suppose L agents solve the following problem:

$$\min_{\mathbf{x}} \sum_{i=1}^{L} f_i(\mathbf{x}), \tag{1}$$

where $f_i(\cdot) : \mathbb{R}^N \to \mathbb{R}$ represents the local objective function that is only known to agent i. Additionally, each agent may exchange information with its neighbors through the underlying undirected communication network, which can be modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [L]$ represents the set of agents and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ represents the set of edges. In other words, the existence of $e_{i,j} \in \mathcal{E}$ indicates that i and j can exchange information. Due to the fact that both $e_{i,j}$ and $e_{j,i}$ denote the same edge, we merely use the expressions $e_{i,j}$ (if $i < j$) or $e_{j,i}$ (if $j < i$) to avoid repetition. Additionally, $E = |\mathcal{E}|$ is the total number of edges, and $|\mathcal{N}_i|$ is the number of neighbors for node i in which $\mathcal{N}_i$ is its set of neighbors. In order to apply ADMM, one can rewrite (1) in the form of an edge consensus problem as follows:

$$\begin{aligned} \min_{\{\mathbf{x}_1,\cdots,\mathbf{x}_L,\mathbf{Z}\}} \quad & \sum_{i=1}^{L} f_i(\mathbf{x}_i) + g(\mathbf{Z}) \\ \text{subject to} \quad & \mathbf{x}_i = \mathbf{z}_{i,j}, \mathbf{x}_j = \mathbf{z}_{i,j}, \quad \forall e_{i,j} \in \mathcal{E} \end{aligned} \tag{2}$$

where each $\mathbf{Z} = \{\{\mathbf{z}_{i,j}\}_{j\in\mathcal{N}_i, j>i}\}_{i=1}^{L}$ are auxiliary variables, and $g(\cdot) = 0$. The augmented Lagrangian of (2) is:

$$\begin{aligned} \mathcal{L}_{\rho_\lambda}(\mathbf{X},\mathbf{Z},\boldsymbol{\lambda}) = \sum_{i=1}^{L} f_i(\mathbf{x}_i) + \sum_{e_{i,j}\in\mathcal{E}} \Big(& (\boldsymbol{\lambda}_{i,j}^{i})^{\mathrm{T}}(\mathbf{x}_i - \mathbf{z}_{i,j}) \\ + (\boldsymbol{\lambda}_{i,j}^{j})^{\mathrm{T}}(\mathbf{x}_j - \mathbf{z}_{i,j}) + \frac{\rho_\lambda}{2}\|\mathbf{x}_j - \mathbf{z}_{i,j}\|^2 + \frac{\rho_\lambda}{2}\|\mathbf{x}_i - \mathbf{z}_{i,j}\|^2 \Big), \end{aligned} \tag{3}$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_L]$, $\boldsymbol{\lambda} = \{\{\boldsymbol{\lambda}_{i,j}^{i}, \boldsymbol{\lambda}_{i,j}^{j}\}_{j\in\mathcal{N}_i, j>i}\}_{i=1}^{L}$ are dual variables, and ρ_λ is a penalty parameter. Distributed ADMMs are iterative procedures that involve three steps at each iteration. The first step is to minimize $\mathcal{L}_{\rho_\lambda}$ with respect to $\mathbf{X}$. Afterward, $\mathcal{L}_{\rho_\lambda}$ is minimized based on $\mathbf{Z}$. In the last step, a dual gradient-ascent iteration is used to update $\boldsymbol{\lambda}$.

Definition 1. *A function $f(x)$ is $\rho-$weakly convex ($\rho > 0$) if there exists a convex function $h(x)$ such that $h(x) = f(x) + \rho\|x\|^2$.*

Weakly convex local functions pose a challenge to distributed ADMM convergence because it can both be non-convex and non-smooth. In the absence of Lipschitz differentiability, which is coming from smoothness, and convexity of the objective function, existing distributed ADMM-based approaches cannot guarantee convergence. In the following section, we present an ADMM-based algorithm can deal with weakly convex functions, regardless of whether they are smooth.

III. Moreau envelope ADMM

The proximal augmented Lagrangian of (2) can be derived as:

$$\Psi_{\rho_\lambda,\rho_\beta}(\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\lambda}) = \mathcal{L}_{\rho_\lambda}(\mathbf{X},\mathbf{Z},\boldsymbol{\lambda}) + \frac{\rho_\beta}{2}\|\mathbf{Z}-\boldsymbol{\beta}\|_F^2, \tag{4}$$

where $\boldsymbol{\beta} = \{\{\boldsymbol{\beta}_{i,j}\}_{j\in\mathcal{N}_i, j>i}\}_{i=1}^{L}$ are auxiliary variables, and ρ_β is a penalty parameter. In (4), The proximal term plays a crucial role in obtaining convergent results. It regulates the behavior of the algorithm in both the $\mathbf{Z}$-update step and indirectly in the $\boldsymbol{\lambda}$-update step by incorporating the Moreau envelope function, resulting in provable convergence. According to our proposed proximal ADMM algorithm, the $(k+1)$th iteration is as follows:

$$\mathbf{X}^{(k+1)} = \arg\min_{\mathbf{x}} \Psi_{\rho_\lambda,\rho_\beta}\left(\mathbf{X}, \mathbf{Z}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right), \tag{5a}$$

$$\mathbf{Z}^{(k+1)} = \arg\min_{\mathbf{Z}} \Psi_{\rho_\lambda,\rho_\beta}\left(\mathbf{X}^{(k+1)}, \mathbf{Z}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right), \tag{5b}$$

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \eta\left(\boldsymbol{\beta}^{(k)} - \boldsymbol{Z}^{(k+1)}\right), \tag{5c}$$

$$\boldsymbol{\lambda}_{i,j}^{i,(k+1)} = \boldsymbol{\lambda}_{i,j}^{i,(k)} + \rho_\lambda\left(\mathbf{x}_i^{(k+1)} - \mathbf{z}_{i,j}^{(k+1)}\right), \forall e_{i,j} \in \mathcal{E} \tag{5d}$$

$$\boldsymbol{\lambda}_{i,j}^{j,(k+1)} = \boldsymbol{\lambda}_{i,j}^{j,(k)} + \rho_\lambda\left(\mathbf{x}_j^{(k+1)} - \mathbf{z}_{i,j}^{(k+1)}\right), \forall e_{i,j} \in \mathcal{E} \tag{5e}$$

where $\eta \in (0, 2)$.

More precisely, each $\mathbf{x}_i$ can be updated individually in update-step $\mathbf{X}$, which after several simplifications, can be stated as follows:

$$\begin{aligned} \mathbf{x}_i^{(k+1)} = \mathrm{Prox}_{f_i}\Bigg(& \frac{\sum_{j\in\mathcal{N}_i, j>i} \mathbf{z}_{i,j}^{(k)} - \frac{\lambda_{i,j}^{i,(k)}}{\rho_\lambda}}{|\mathcal{N}_i|} \\ & + \frac{\sum_{j\in\mathcal{N}_i, j<i} \mathbf{z}_{j,i}^{(k)} - \frac{\lambda_{j,i}^{i,(k)}}{\rho_\lambda}}{|\mathcal{N}_i|}; \frac{1}{\rho_\lambda|\mathcal{N}_i|}\Bigg) \end{aligned} \tag{6}$$

Algorithm 1: Moreau envelope ADMM for distributed optimization (MADM)

Initialize $\mathbf{X}^{(0)}$, $\mathbf{Z}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, ρ_β, ρ_λ and $\eta \in (0,2)$;
repeat
  for $i \in [L]$ **do**
    Update $\mathbf{x}_i$ as:
    $$\mathbf{x}_i^{k+1} = \text{Prox}_{f_i}\left(\sum_{j\in\mathcal{N}_i} \frac{\mathbf{z}_{i,j}^{(k)} - \frac{\lambda_{i,j}^{i,(k)}}{\rho_\lambda}}{|\mathcal{N}_i|}; \frac{1}{\rho_\lambda|\mathcal{N}_i|}\right);$$
  end
  Each agent sends its local vector $\mathbf{x}_i^{k+1}$ to neighboring agents;
  for $i \in [L]$ **do**
    for $j \in \mathcal{N}_i$ **do**
      Update $\mathbf{z}_{i,j}$ as: $\mathbf{z}_{i,j}^{(k+1)} = \frac{\rho_\lambda\left(\mathbf{x}_j^{(k+1)}+\mathbf{x}_i^{(k+1)}\right)+\rho_\beta\boldsymbol{\beta}_{i,j}^{(k)}+\lambda_{i,j}^{i,(k)}+\lambda_{i,j}^{j,(k)}}{2\rho_\lambda+\rho_\beta}$;
      Update $\boldsymbol{\beta}_{i,j}$ as: $\boldsymbol{\beta}_{i,j}^{(k)} - \eta\left(\boldsymbol{\beta}_{i,j}^{(k)} - \boldsymbol{z}_{i,j}^{(k+1)}\right)$;
      Update $\boldsymbol{\lambda}_{i,j}^{i}$ as:
      $\boldsymbol{\lambda}_{i,j}^{i,(k+1)} = \boldsymbol{\lambda}_{i,j}^{i,(k)} + \rho_\lambda\left(\mathbf{x}_i^{(k+1)} - \mathbf{z}_{i,j}^{(k+1)}\right)$;
      Update $\boldsymbol{\lambda}_{i,j}^{j}$ as:
      $\boldsymbol{\lambda}_{i,j}^{j,(k+1)} = \boldsymbol{\lambda}_{i,j}^{j,(k)} + \rho_\lambda\left(\mathbf{x}_j^{(k+1)} - \mathbf{z}_{i,j}^{(k+1)}\right)$;
    end
  end
until *the convergence criterion is met*;

Also, in the $\mathbf{Z}$-update step, $\mathbf{z}_{i,j}$ can be updated separately as:

$$\begin{aligned}\mathbf{z}_{i,j}^{(k+1)} = \arg\min_{\mathbf{z}_{i,j}} \Big(&(\lambda_{i,j}^{i,(k)})^{\mathrm{T}}(\mathbf{x}_i^{(k+1)} - \mathbf{z}_{i,j}) \\ &+ (\lambda_{i,j}^{j,(k)})^{\mathrm{T}}(\mathbf{x}_j^{(k+1)} - \mathbf{z}_{i,j}) + \frac{\rho_\lambda}{2}\|\mathbf{x}_j^{(k+1)} - \mathbf{z}_{i,j}\|^2 \\ &+ \frac{\rho_\lambda}{2}\|\mathbf{x}_i^{(k+1)} - \mathbf{z}_{i,j}\|^2 + \frac{\rho_\beta}{2}\|\mathbf{z}_{i,j} - \boldsymbol{\beta}_{i,j}^{(k)}\|^2\Big), \quad (7)\end{aligned}$$

which can be simplified as follows:

$$\mathbf{z}_{i,j}^{(k+1)} = \frac{\rho_\lambda\left(\mathbf{x}_j^{(k+1)} + \mathbf{x}_i^{(k+1)}\right) + \rho_\beta\boldsymbol{\beta}_{i,j}^{(k)} + \lambda_{i,j}^{i,(k)} + \lambda_{i,j}^{j,(k)}}{2\rho_\lambda + \rho_\beta}. \quad (8)$$

Moreover, for each $\boldsymbol{\beta}_{i,j}$ we have:

$$\boldsymbol{\beta}_{i,j}^{(k+1)} = \boldsymbol{\beta}_{i,j}^{(k)} - \eta\left(\boldsymbol{\beta}_{i,j}^{(k)} - \boldsymbol{z}_{i,j}^{(k+1)}\right). \quad (9)$$

By introducing $\boldsymbol{\lambda}_{i,j}^{i}$, $\boldsymbol{\lambda}_{i,j}^{j}$, $\mathbf{z}_{i,j}$, and $\boldsymbol{\beta}_{i,j}$ to represent $\boldsymbol{\lambda}_{j,i}^{i}$, $\boldsymbol{\lambda}_{j,i}^{j}$, $\mathbf{z}_{j,i}$, and $\boldsymbol{\beta}_{j,i}$, respectively in each agent i, the proposed method is simplified and summarized in Algorithm 1.

IV. Convergence proof

This section presents the convergence analysis for Algorithm 1. Several conventional assumptions are made to build our convergence proof.

Assumption 1. *The undirected graph $\mathcal{G}$ is connected.*

Assumption 2. *$\Psi_{\rho_\lambda,\rho_\beta}\left(\mathbf{X}^{(k)}, \mathbf{Z}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right)$ is lower bounded, and $\left(\mathbf{X}^{(k)}, \mathbf{Z}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right)$ are bounded, in each iteration k.*

Remark. *It can be shown that for coercive functions*[1] *Assumption 2 holds true.*

Assumption 3. *Local objectives $f_i(\cdot), \forall i \in [L]$, are continuous, and weakly convex by parameter ρ_f.*

The proof of convergence relies on a canonical methodology as described in [25, Theorem 2.9]. Each algorithm iteration has only one increasing step, which is the $\boldsymbol{\lambda}$-update step. As the gradient of the Moreau envelope is related to the proximal function, the amount of increase in the $\boldsymbol{\lambda}$-update step ($\rho_\lambda^{-1}\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\|^2$), is bounded based on the primal and auxiliary variables. Thus, tuning the parameters lets us guarantee the *sufficient decrease condition* of [25, Theorem 2.9]. In addition, the subgradient of the proximal augmented Lagrangian based on each of its inputs can easily be shown to be bound in each iteration, which is sufficient to validate the *bounded subgradient condition* of the [25, Theorem 2.9]. Finally, by having the boundedness assumption and knowing that the proximal augmented Lagrangian is continuous based on each of its inputs, it can be shown that the *continuity condition* of [25, Theorem 2.9] holds.

Lemma 1. *Function $g(\cdot)$, $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, satisfies condition:*

$$\|\nabla\mathcal{M}_{g(\cdot)}(\mathbf{u},\gamma) - \nabla\mathcal{M}_{g(\cdot)}(\mathbf{v},\gamma)\| = 0. \quad (10)$$

Lemma 2. *If Assumption 1 is held, for any $m \geq 1$, the following inequality is held:*

$$\begin{aligned}&\|\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}\|_F^2 \leq \\ &\qquad \rho_\beta^2\left(\|\mathbf{Z}^{(k+1)} - \mathbf{Z}^{(k)}\|_F^2 + \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_F^2\right) \quad (11)\end{aligned}$$

Proof. The lemma is proved by combining Lemma 1 with [26, Lemma 4]. □

Lemma 3. *Assuming Assumptions 1, 2, and 3 and $\rho_\lambda|\mathcal{N}_i| > \rho_f, \forall i \in [L]$, the following inequality holds:*

$$\begin{aligned}&\Psi_{\rho_\lambda,\rho_\beta}\left(\mathbf{X}^{(k)}, \mathbf{Z}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right) - \\ &\qquad \Psi_{\rho_\lambda,\rho_\beta}\left(\mathbf{X}^{(k+1)}, \mathbf{Z}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}\right) \geq \quad (12a)\\ &C(\rho_\lambda)\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F^2 + \left(\frac{\rho_\lambda}{2} + \frac{\rho_\beta}{2}\right)\|\mathbf{Z}^{(k+1)} - \mathbf{Z}^{(k)}\|_F^2 \quad (12b)\\ &\qquad + \frac{\rho_\beta}{2}\left(\frac{2}{\eta} - 1\right)\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_F^2 \quad (12c)\\ &-\rho_\lambda^{-1}\rho_\beta^2\left(\|\mathbf{Z}^{(k+1)} - \mathbf{Z}^{(k)}\|_F^2 + \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_F^2\right), \quad (12d)\end{aligned}$$

where $C(\cdot)$ is a function with positive value for ρ_λ.

Proof. Equation (12b) is derived based on the weak convexity of local functions, and $g(\cdot)$, while (12c) is the result of

[1] A function $f(\cdot)$ is coercive if $f(\mathbf{x})\mathbf{x} \to \infty$ as $\|\mathbf{x}\| \to \infty$

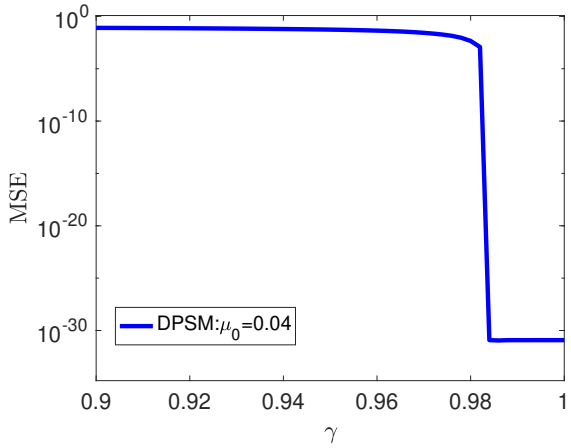


Fig. 1: MSE versus γ

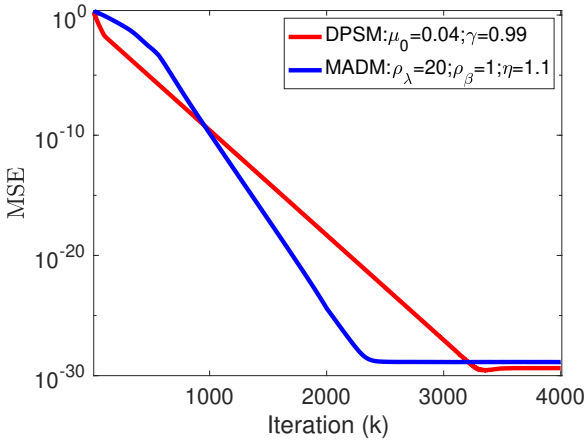


Fig. 2: MSE versus iteration

expanding the amount of change from the $\boldsymbol{\beta}$-update step. Further, (12d) derives from multiplying the bound obtained from Lemma 2 with $-\rho_\lambda^{-1}$, which gives an upper bound for the amount of change in the $\boldsymbol{\lambda}$-update step. □

Theorem 1. *By having* $\rho_\lambda|\mathcal{N}_i| > \rho_f, \forall i \in [L]$, $\frac{1}{\eta} \geq \frac{1}{2} + \rho_\lambda^{-1}\rho_\beta$, *and* $\rho_\lambda \geq \frac{(2\sqrt{2}-1)}{2}\rho_\beta$, *if Assumptions 1,2, and 3 hold, the algorithm 1 converges to a stationary point.*

Proof. The *sufficient decrease condition* of [25, Theorem 2.9] holds when ρ_λ, ρ_β, and ρ_β satisfy the condition of Theorem 1, by Lemma 3. The same results are obtained for the *bounded subgradient condition* of [25, Theorem 2.9] when it depends on the norm of the successive difference of the variables. Finally, employing Assumption 2 and knowing that the proximal augmented Lagrangian is continuous for each of its inputs, we can prove the *continuity condition* of [25, Theorem 2.9], which completes the proof. □

V. Simulation Results

In this section, we evaluate the performance of the MADM by conducting simulations of distributed robust phase retrieval with the objective function:

$$\hat{\mathbf{x}} = \min_{\mathbf{x}} \frac{1}{m} \sum_{i=1}^{L} |\langle \mathbf{a}_i, \mathbf{x}\rangle^2 - b_i^2|, \tag{13}$$

where $\mathbf{x}$ is the target signal, $\mathbf{a}_i$ is the measurement and b_i is the observation in node i. We assume that each node i has one measurement and observation, $\mathbf{a}_i \in \mathbb{R}^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{1}), \forall i \in [L]$, and $\mathbf{x} \in \mathbb{R}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. For simplicity, we assume a noiseless setting with $b_i = \langle \mathbf{a}_i, \mathbf{x}\rangle, \forall i \in [L]$. All simulations are performed by averaging over 50 trials, and in each case, an Erdös-Rényi graph consisting of $L = 50$ nodes was generated as the communication network. To evaluate the performance of this method, we also simulated the distributed projected subgradient method (DPSM) proposed in [19]. The mean square error (MSE) $:= \frac{\sum_{i=1}^{L} ||\hat{\mathbf{x}}_i - \mathbf{x}||_2^2}{L}$ was utilized as the performance measure. Moreover, $\mathbf{z}_{i,j}^{(0)}, \forall e_{i,j} \in \mathcal{E}$ in MADM and $\mathbf{x}_i^{(0)}, \forall i \in \mathcal{V}$ in DPSM were initialized based on the procedure proposed in [27, Sec. 4.2].

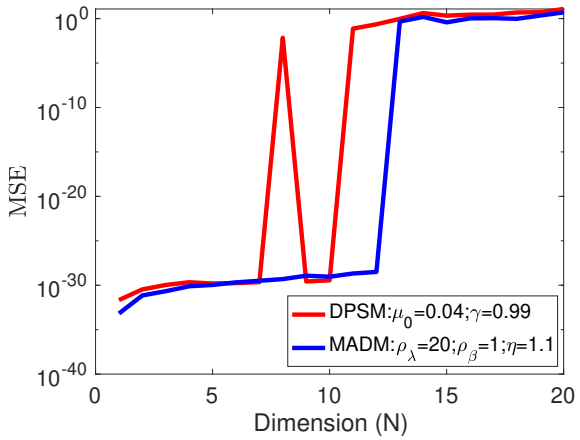


Fig. 3: MSE versus dimension

We first compare the convergence rate and efficiency of the two algorithms. The dimension of the target signal was $N = 10$, and for the DPSM, μ_0 was 0.04, while γ was chosen through a grid search to achieve a minimum and stable error with fast convergence. Fig. 1 illustrates the results of the grid search. We see that the DPSM highly depends on the choice of γ. In our algorithm we set $\rho_\lambda = 20$, $\rho_\beta = 1$, and $\eta = 1.1$. These values satisfy the conditions in Theorem 1. Fig. 2 shows that MADM can achieve a faster convergence rate than DPSM while maintaining a similar MSE.

Next, we study the robustness of the algorithms as a function of the dimension of the target signal for fixed parameters; N ranged from 1 to 20. A comparison of MADM and DPSM behavior under different dimensions is illustrated in Fig. 3. We see that MADM is more stable than DPSM when parameters are fixed. Although Fig. 1 indicates $\gamma = 0.99$ is in the safe zone for $N = 10$, it fails for $N = 8$ and $N > 10$.

VI. Conclusions

This paper presented a new proximal variant of the ADMM algorithm, named MADM, for solving distributed optimization problems. Our analysis demonstrated that the proposed method could be applied to weakly convex functions under mild conditions. In particular, we derived a bound on the change in the dual variable update step by leveraging the relationship between the gradient of the Moreau envelope function and the proximal function. This allowed us to ensure convergence to a stationary point. The simulation results showed that MADM outperforms subgradient methods in terms of speed and robustness. These findings suggest that MADM can be a promising tool for solving a wide range of distributed optimization problems in practice.

References

[1] V. C. Gogineni, S. Werner, Y.-F. Huang, and A. Kuh, "Communication-efficient online federated learning framework for nonlinear regression," in *Proc. - ICASSP IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 5228–5232.

[2] N. K. Venkategowda and S. Werner, "Privacy-preserving distributed maximum consensus," *IEEE Signal Process. Lett.*, vol. 27, pp. 1839–1843, Oct. 2020.

[3] Y. Wang, "Privacy-preserving average consensus via state decomposition," *IEEE Trans. Automat. Contr.*, vol. 64, no. 11, pp. 4711–4716, Mar. 2019.

[4] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automat. Contr.*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[5] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2597–2633, Dec. 2017.

[6] A. Makhdoumi and A. Ozdaglar, "Convergence rate of distributed ADMM over networks," *IEEE Trans. Automat. Contr.*, vol. 62, no. 10, pp. 5082–5095, Oct. 2017.

[7] N. Parikh and S. Boyd, *Proximal algorithms*. Now Publishers, Inc., 2014, vol. 1, no. 3.

[8] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[9] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers, Inc., 2008, vol. 1, no. 1–2.

[10] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Automat. Contr.*, vol. 56, no. 6, pp. 1291–1306, Nov. 2010.

[11] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 102–113, May 2020.

[12] C. Qian, X. Fu, N. D. Sidiropoulos, L. Huang, and J. Xie, "Inexact alternating optimization for phase retrieval in the presence of outliers," *IEEE Trans.Signal Process.*, vol. 65, no. 22, pp. 6069–6082, Nov. 2017.

[13] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding blind deconvolution algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2354–2367, Dec. 2011.

[14] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, Sep. 2015.

[15] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM J. Optim.*, vol. 29, no. 1, pp. 207–239, Jan. 2019.

[16] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Sublinear and linear convergence of modified ADMM for distributed nonconvex optimization," *IEEE Trans. Automat. Contr.*, June 2022.

[17] M. Hong, D. Hajinezhad, and M.-M. Zhao, "Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proc. Mach. Learn. Res.*, 2017, pp. 1529–1538.

[18] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529, Sep. 2019.

[19] S. Chen, A. Garcia, and S. Shahrampour, "On distributed nonconvex optimization: Projected subgradient method for weakly convex problems in networks," *IEEE Trans. Automat. Contr.*, vol. 67, no. 2, pp. 662–675, Feb. 2021.

[20] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J Sci. Comput.*, vol. 78, no. 1, pp. 29–63, Jan. 2019.

[21] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, Jan. 2016.

[22] M. Yashtini, "Convergence and rate analysis of a proximal linearized ADMM for nonconvex nonsmooth optimization," *J. Glob. Optim.*, pp. 1–27, May 2022.

[23] A. Themelis and P. Patrinos, "Douglas–rachford splitting and ADMM for nonconvex optimization: Tight convergence results," *SIAM J. Optim.*, vol. 30, no. 1, pp. 149–181, Jan. 2020.

[24] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[25] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, Feb. 2013.

[26] J. Zeng, W. Yin, and D.-X. Zhou, "Moreau envelope augmented lagrangian method for nonconvex optimization with linear constraints," *J. Sci. Comput.*, vol. 91, no. 2, pp. 1–36, Apr. 2022.

[27] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

An Audio-Visual Speech Enhancement System Based on 3D Image Features: An Application in Hearing Aids

Yu-Ching Chung, Ji-Yan Han, Bo-Sin Wang, Wei-Zhong Zheng, Kung-Yao Shen and Ying-Hui Lai*
Department of Biomedical Engineering, National Yang-Ming University
E-mail: s610316.be09@nycu.edu.tw Tel: +886-901059489
E-mail: jy.han.be08@nycu.edu.tw Tel: +886-918264316
E-mail: sin110529@gmail.com Tel: +886-28267021
E-mail: s1010654.be07@nycu.edu.tw Tel: +886-28267021
E-mail: roger20415.be08@nycu.edu.tw Tel: +886-937030676
E-mail: yh.lai@nycu.edu.tw Tel: +886-228267021

Abstract — **Previous research has shown that auditory and visual inputs are not asynchronous in the human brain, and that visual cues can enhance attention in the hearing process. Therefore, this study proposes audio-visual speech enhancement (SE) with 3D image features (AV-3D-SE) that imitates the auditory process of humans to elevate listening quality. More specifically, AV-3D-SE uses the FlowNet3D model to predict temporal facial motion from the recorded 3D image combining with features for SE applications. The evaluation results showed that the average scores of perceptual evaluation of speech quality and short-time objective intelligibility in 3 dB signal-to-noise ratio increased to 3.229 and 0.914, respectively, while the average hearing aid speech quality index significantly outperformed baseline SE systems (audio-only and audio-visual-2D) in seven typical types of hearing loss with high hearing aid speech perception index. In conclusion, the proposed AV-3D-SE enhances the effectiveness of the SE system and can increase the listening satisfaction of hearing aid users.**

Keywords: deep learning, point cloud, scene flow, hearing aid and speech enhancement.

I. INTRODUCTION

Hearing is one of the most important human senses. It provides the ability to communicate with people, detect danger, and appreciate aesthetics. However, the human ear may be unable to detect sounds below 35 decibels (dB) due to various reasons — such as aging, trauma, viral infection, or long-term exposure to noisy environments — called hearing loss [1,2]. According to the World Health Organization, more than 430 million people worldwide suffer from hearing loss. Furthermore, it is estimated that the proportion of the global population with hearing loss will reach 10% by 2050; that is, more than 700 million people will have problems associated with hearing loss [2]. Consequently, it is crucial to help patients with hearing loss strengthen their semantic comprehension and improve their speech quality. To solve these problems, hearing-assistive devices, such as hearing aids, are mostly used to help people with hearing loss communicate better.

Previous studies have pointed out that patients with hearing loss can significantly improve their speech comprehension ability when using hearing aids in a quiet environment; however, 66% of users indicated that they still struggle to listen clearly amidst noisy surroundings [3]. In other words, there may still be limitations to hearing aids in environments with background noise. Speech enhancement (SE) technology has been proposed to alleviate these problems and improve the speech intelligibility of hearing aid users operating in noisy environments. The purpose of SE is to remove background noise and avoid speech distortion after processing [4,5]. Well-known SE systems have been developed over the past few decades, including the log minimum mean square error (MMSE) [6], Wiener filter [7], Karhunen-Loeve transform [8], and the Generalized Maximum a Posteriori Spectral Amplitude [9]. Although the above SE technologies provide a suitable performance for users in stationary noise scenarios, there is still room for improvement for those in challenging listening conditions (e.g., nonstationary and competing noise) [10].

Recently, deep-learning technology has been used to design SE systems and has been proven to provide better performance than the conventional SE methods mentioned above under challenging listening conditions. Wang et al. [11] have proposed a SE system using a deep neural network (DNN) and a support vector machine-based classification approach to estimate an ideal binary mask. The study indicated that the signal-to-noise ratio (SNR) increased by 2.5 dB compared with SE using MMSE. In 2018, Luo et al. [12] proposed the Conv-TasNet system that uses an end-to-end DNN architecture to implement speech separation technology. Using this model, the target speaker can be predicted in a noisy environment with two or three speakers. Lai at el. [13] proposed a SE system called a "multi-objective learning-based deep denoising autoencoder" (MDDAE). The MDDAE-SE system provides higher scores on the hearing aid speech quality index (HASQI) [14] and hearing

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

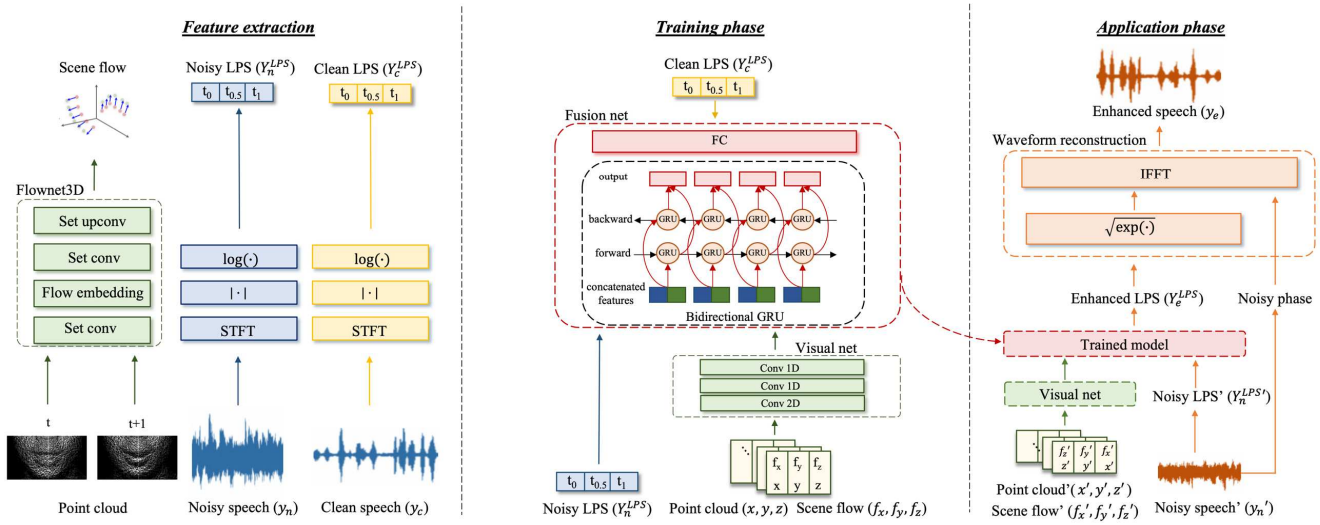


Fig. 1 Flowchart of the proposed AV-3D-SE system.

aid speech perception index (HASPI) [15] than classical SE approaches. Although the audio-only (AO) model (the input feature of the SE system only uses audio features) provides suitable performance for users, there is still room for improvement in the effect of AO-SE due to the similarity between background noise and the target voice or the complexity of the background noise type [16].

When communicating with people, we depend not only on auditory information, but on visual information as well [17-20]. In a 2013 study conducted by Golumbic et al. [21], the results show that an increase in visual information resulted in better speech comprehension than under AO information conditions. Furthermore, some scholars have speculated that, because visual information is transmitted to the brain faster than sound, vision has a predictive auxiliary effect on the auditory system, improving the brain's concentration on sound [22-24]. Based on the above physiological research results, a model combining visual information with voice information, called the audio-visual (AV) model, has been proposed in recent years and has proven to have better results than AO models. For instance, Hou et al. [25] combined 2D image features and acoustic features of the lips of the face to build an AV deep convolutional neural network SE model (AV-2D-SE in this study), showing that, when compared with the AO-SE system, the AV-2D-SE model has made significant progress in both objective and subjective evaluations. This result shows that audio features combined with visual information can help the SE system perform better than AO information.

At present, all AV models use 2D plane images as inputs. However, the movements of facial muscles and joints when humans speak are 3D and undergo continuous changes [26]. Planar images may lose some facial depth change features. In other words, if we can obtain the 3D features of human facial movements while speaking, more accurate information can be provided to obtain a better AV-SE model. One method can obtain 3D features of human facial movement, such as a time-of-flight camera [27], while scene flow [28] technology can extract key features from the recorded 3D image (the so-called point cloud). The scene flow proposed by Vedula et al. [29, 30] describes the change in each point in two adjacent frames in a 3D space, and which can extract useful features for different applications. For example, Menze et al. [31] used scene flow combined with deep learning to conduct related research on automatic driving image recognition. The results indicate a 36% reduction in the root-mean-square error when comparing the 3D scene flow method with the 2D flow method. Recently, a novel deep-learning-based 3D feature extraction model — FlowNet3D [32] — was proposed, and which not only achieves a significantly lower 3D endpoint error [33] but also outperforms conventional methods [34]. Following the success of the FlowNet3D technology, this study proposes an AV-SE system called Audio Visual-3D-SE (AV-3D-SE) to enhance noisy speech for hearing aid users, whereby audio and 3D point-cloud image features are used. The FlowNet3D model is used as a feature extractor for the proposed system. Several objective evaluation metrics for hearing aids were used to ensure the benefits of the proposed AV-3D-SE system.

II. Proposed System

A flowchart of the proposed AV-3D-SE system is shown in Fig. 1; it is divided into three stages: feature extraction, training phase, and application phase. In the feature extraction stage, we extracted the scene flow from the point cloud data using FlowNet3D and the noisy log power spectrum (LPS) [35] Y_n^{LPS} from the noisy speech y_n using Short-time Fourier transform (STFT). In the training phase, the point cloud and scene flow were used as inputs for the visual net. Second, the outputs of the visual net and Y_n^{LPS} were used as the inputs of the fusion net to obtain an enhanced LPS, Y_e^{LPS}. Finally, in the application phase, the enhanced speech y_e is reconstructed using an

inverse fast Fourier transform (IFFT), with the phase derived from the y_n' and Y_e^{LPS}. The details of each stage are as follows:

A. Feature extraction stage

1. 3D image feature

Because the region of facial deformation during speech is around the lips, we first captured a point cloud image around this area during the preprocessing stage. We then applied it to the FlowNet3D model, which is pretrained using the FlyingThings3D dataset, for scene flow feature extraction. The FlowNet3D model architecture (Fig. 2) can be divided into three main processing layers: (1) set conv, (2) flow embedding, and (3) set upconv layers. First, two sets of conv layers take the point clouds of consecutive frames. The set conv layer uses the farthest point sampling to sample the representative key points of the point cloud image and groups these sampling points with radius, *r,* using a ball query. Next, they are connected to PointNet [36], which learns the local features of each group of points using a multilayer perceptron. Finally, the most representative features from each point group were selected via the max-pooling layer. Subsequently, the flow-embedding layer learns the feature similarities and spatial relationships between two adjacent point-cloud frames to mix them. More precisely, the flow embedding layer uses the point cloud at frame *t* to find the nearest k points within radius *r* at frame *t*+1 and then uses a PointNet to learn the flow of the area in the current and next frame features. Finally, the set upconv layer is used to upsample according to the point cloud and features of the flow embedding, and to predict the flow direction of all original points in the last layer.

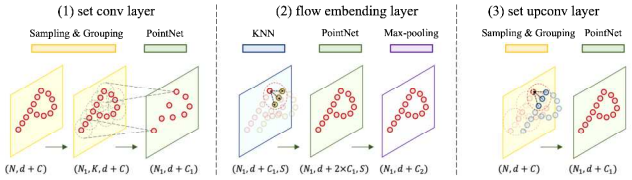


Fig. 2 The FlowNet3D model architecture: (1) set conv layer, (2) flow embedding layer, and (3) set upconv layer.

2. Audio features

In this study, the LPS feature was used as the auditory feature input for the multimodal SE system. First, we computed the STFT for each frame of noisy speech and target speech to obtain its energy and phase. The formula used is as follows:

$$\mathrm{Y}(k) = \sum_{l=0}^{L-1} y(l)h(l)e^{-\frac{j2\pi kl}{L}} \quad k = 0,1,\dots,L-1 \tag{1}$$

where $y(l)$ is the input speech of the frame, k is the frequency bin index, and $h(l)$ is the Hamming window function. The LPS feature of voice $\mathrm{Y}(k)$ can then be obtained by taking the logarithm of the energy spectrum, as follows:

$$\mathrm{Y}^{LPS}(k) = \log|\mathrm{Y}(k)|^2 \quad k = 0,1,\dots,L/2 \tag{2}$$

B. Model training and application phases

After obtaining Y_n^{LPS} from y_n, a fusion net was used to obtain the Y_e^{LPS}. The point cloud data (*x,y,z*) and their corresponding scene flow features (f_x, f_y, f_z) at time *t* were concatenated and then input into the three visual nets. The visual net contains a 2D convolution layer and two 1D convolution layers to identify the relationship between the adjacent point cloud and scene flow. The outputs of the three visual nets are superimposed with the audio signal LPS at times *t*, *t*+0.5, and *t*+1 and input to the bidirectional gated recurrent unit (BGRU) [37], which prevents the vanishing gradient problem and reduces the computing time. The design of the gated recurrent unit module was based on long short-term memory (LSTM), which has a simpler structure. It only needs to reset and update the gate to achieve the same benefits as the LSTM with a small number of parameters. Finally, we obtain Y_e^{LPS} by integrating the feature information through a fully connected (FC) network. We calculated the loss between the predicted LPS, Y_e^{LPS}, and the clean LPS, Y_c^{LPS}, using the MMSE loss function with Adam optimizer and learning rate of 5×10^{-5}. The loss function is expressed as follows:

$$\theta = \min\left(\frac{1}{K}\sum_{i=1}^{K}(\mathrm{Y}_e^{LPS} - \mathrm{Y}_c^{LPS})^2\right) \quad i = 1,\dots,\mathrm{K} \tag{3}$$

where K denotes the total frame number, and i is the frame index.

In the application phase, we can convert noisy speech y_n' into y_e with a visual net, a trained model, and waveform reconstruction. First, we input the visual information (point clouds and scene flows) into the visual net. The output of the visual net and its corresponding $\mathrm{Y}_n^{LPS\prime}$ can then be fed into the trained model to obtain Y_e^{LPS}. Additionally, by incorporating the phase of y_n' during waveform reconstruction, we can obtain y_e.

III. Experiment Design and Results

A. Material

This research recorded audio and visual signals with a Focus Vision RL100 Pro sensor, shooting at 50 fps, and with a

microphone to simultaneously capture 2D images, 3D images, and audio with 20 ms frame size and 10 ms frame shift, respectively. The frame size and shift can be adjusted to reduce group delays in the hearing aid application scenario. Furthermore, to ensure that the number of points in the 3D image remained the same, 1,050 points were randomly selected from the original 3D images to represent 3D features. The 2D images were scaled to 66 pixels × 85 pixels so as to ensure that the size of each dataset was the same. The distance between the microphone and camera was approximately 25–30 cm, and video and audio were recorded by the same Taiwanese man. In the training phase, three sets of 80 sentences from the Taiwan Mandarin hearing-in-noise test were used, with 15 types of noise at a 3 dB SNR level. Hence, 3,600 noisy speech sentences and videos were generated to train the model. In the testing phase, the remaining 50 sentences were mixed with 11 types of noise (not included in the training phase) at two SNR levels (0 and 3 dB).

B. Procedure

To verify the effectiveness of the proposed AV-3D-SE system, the AO-SE and AV-2D-SE models were used as baseline systems. The AO-SE model directly inputs the Y_n^{LPS} features into a single BGRU layer and then connects it to a single FC layer to output the Y_e^{LPS}. For the AV-2D-SE system, we refer to the visual net in [25], which extracts the image features through three 2D convolutional layers and then stacks the features with the Y_n^{LPS}. These features were used as inputs to a single BGRU layer and an FC layer to obtain Y_e^{LPS}. Detailed settings of the AV-2D-SE models can be found in [25]. To compare the three systems' performance, we used four common evaluation metrics — perceptual evaluation of speech quality (PESQ) [38], short-time objective intelligibility (STOI) [39], HASQI [14], and HASPI [15] — to compare the benefits of SE between the proposed system and two baseline systems. PESQ is an automatic and objective evaluation of speech quality standards which is used to quantify the difference between the original and processed speech in mean option score with numbers from -0.5 to 4.5. The higher the number, the better the quality. The STOI is an objective speech comprehension evaluation method that is highly correlated with the comprehension of time-frequency-weighted noisy speech. STOI scores range from 0 to 1, with higher values indicating better speech comprehension. HASPI and HASQI are indices used to evaluate the performance of hearing aids in improving speech comprehension and quality in people with hearing loss. Table 1 shows the seven types of typical hearing loss types [40] used in this study. Both HASPI and HASQI provide scores ranging from 0 to 1, with higher scores indicating better speech perception and quality.

Table 1. The seven typical hearing loss audiograms [40].

	Frequency (kHz) in dB HL					
Audiogram	**0.25**	**0.5**	**1**	**2**	**4**	**6**
1	60	60	60	60	60	60
2	70	70	70	50	10	10
3	40	40	50	60	65	65
4	0	0	0	60	80	90
5	0	15	30	60	80	85
6	14	14	11	14	24	39
7	24	24	25	31	46	60

C. Results of objective evaluation metrics

Tables 2 and 3 present the average PESQ and STOI scores for 50 sentences in 11 types of noise situations for each approach, including the original noisy speech, AO-SE, AV-2D-SE, and proposed AV-3D-SE systems at SNRs of 0 and 3 dB.

Table 2. Average PESQ scores of the proposed and baseline systems at SNR 0 and 3 dB.

SNR level	**Noisy**	**AO-SE**	**AV-2D-SE**	**AV-3D-SE**
0 dB	1.687	2.334	2.485	**3.084**
3 dB	1.826	2.519	2.581	**3.229**

Table 3. Average STOI scores of the proposed and baseline systems at SNR 0 and 3 dB.

SNR level	**Noisy**	**AO-SE**	**AV-2D-SE**	**AV-3D-SE**
0 dB	0.659	0.752	0.807	**0.899**
3 dB	0.735	0.791	0.830	**0.914**

First, the PESQ and STOI scores of all SE models were higher than those of the original y_n. Second, the model combining visual and voice information had better speech quality and speech understanding than the AO-SE model. The proposed AV-3D-SE system exhibited the highest PESQ and STOI scores. The PESQ score improved from 1.687 to 3.084 at an SNR of 3 dB, and from 1.826 to 3.229 at an SNR of 0 dB. The STOI scores increased by approximately 0.24 and 0.196, respectively. Comparing the AV-3D-SE model with the AV-2D-SE model, whether at an SNR of 0 or 3 dB, the PESQ and STOI scores increased significantly. From the above results, we infer that point clouds and scene flows as model inputs can better represent face deformation when speaking than that of 2D images. Accordingly, this shows that the AV-3D-SE model improves speech quality and comprehension compared to the other models.

Fig. 3 (A) and (B) present the average HASQI scores of the proposed and baseline systems at SNRs of 0 and 3 dB, respectively. From these results (at two SNR levels), among the

seven types of hearing loss, these three systems can provide enhancement benefits compared with those of noisy conditions; the AV-3D-SE model has the best performance, followed by the AV-2D-SE and AO-SE models. The three models also showed significant differences ($p<0.001$). However, compared with AO-SE, the performance of the AV-2D-SE model did not significantly improve under the conditions of audiograms 1, 2, and 3. The performance of AV-3D remained the best. For steep-drop-type high-frequency hearing loss (audiograms 4 to 7), the benefit of the AO-SE system is much lower, whereas the two systems with visual assistance maintain good performance. Consequently, we believe that it can improve the ability of hearing-impaired people to concentrate and improve the quality of their hearing with 3D facial structures and deformations similar to those of the human brain processes mentioned earlier. Thus, 3D images and scene flows are helpful in improving the speech quality perceived by hearing-impaired people.

Tables 4 and 5 show the average HASPI scores for different types of hearing loss in 11 noise situations. After the three SE systems were processed at two different SNR levels, the HASPI scores of different hearing-impaired individuals improved to a very high level. Although the score of the AV-3D-SE system is higher than that of the other two baseline systems, the difference in listening comprehension between the three is not obvious, which indicates that the processed speech of the three systems can be understood by different hearing-impaired individuals when listening.

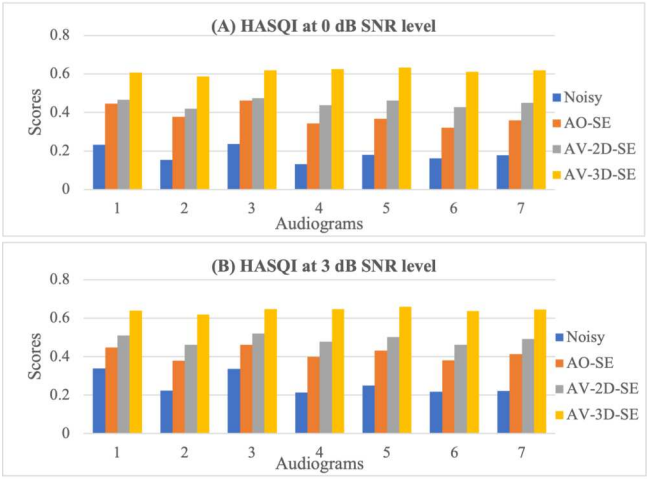


Fig. 3 Average HASQI scores of the proposed and baseline systems at (A) 0dB and (B) 3dB SNR levels.

IV. Conclusion

This study used the FlowNet3D model to extract the scene flow features of facial deformation around the lips when speaking. This information can be combined with audio information to create an innovative and progressive AV-3D-SE system for improving the noise reduction effect of hearing aids in noisy environments. The PESQ and STOI scores show that speech quality and comprehension are significantly improved compared with the other models. We used the HASPI and HASQI to compare the speech quality and comprehension of hearing-impaired individuals. The results show that all models perform intelligibly. However, in terms of speech quality, the AV-3D-SE model showed improved speech quality compared to the other algorithms. Finally, we showed that scene flow is a potential visual feature that enhances the benefits of the SE model and proved the effectiveness of the AV-3D-SE model proposed in this study.

Table 4. Average HASPI scores of the proposed and baseline systems at 0 dB SNR level.

Audiogram	Noisy	AO-SE	AV-2D-SE	AV-3D-SE
1	0.9640	0.9990	0.9988	**0.9998**
2	0.8316	0.9968	0.9961	**0.9997**
3	0.9609	0.9991	0.9987	**0.9998**
4	0.7669	0.9875	0.9945	**0.9998**
5	0.8305	0.9926	0.9966	**0.9998**
6	0.8040	0.9854	0.9934	**0.9997**
7	0.8655	0.9928	0.9966	**0.9998**

Table 5. Average HASPI scores of the proposed and baseline systems at 3 dB SNR level.

Audiogram	Noisy	AO-SE	AV-2D-SE	AV-3D-SE
1	0.9935	0.9990	0.9993	**0.9999**
2	0.9474	0.9968	0.9980	**0.9998**
3	0.9614	0.9992	0.9993	**0.9999**
4	0.9053	0.9961	0.9973	**0.9998**
5	0.9383	0.9977	0.9983	**0.9998**
6	0.9268	0.9953	0.9966	**0.9998**
7	0.9609	0.9977	0.9981	**0.9998**

V. Acknowledgment

This study was supported by the National Science and Technology Council of Taiwan (MOST 111-2221-E-A49-041-MY2).

References

[1] J. J. Eggermont, Hearing loss: Causes, prevention, and treatment. Academic Press, 2017.

[2] "WHO, Deafness and Hearing Loss," vol. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2021.

[3] J. L. Desjardins and K. A. Doherty, "The effect of hearing aid noise reduction on listening effort in hearing-impaired adults," in Hearing Review, vol. 19, no. 3, pp. 12-25, 2012.

[4] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[5] K. K. Paliwal and K. Yao, "Robust speech recognition under noisy ambient conditions," in *Human-centric interfaces for ambient intelligence: Elsevier*, pp. 135-162, 2010.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443-445, 1985.

[7] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, pp. 629-632, 1996.

[8] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, 2001.

[9] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, vol. 76, pp. 112-126, 2016.

[10] Y. H. Tu, J. Du and C. H. Lee, "Speech Enhancement Based on Teacher–Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080-2091, 2019.

[11] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092-7096, 2013.

[12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019.

[13] Y. H. Lai and W. Z. Zheng, "Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users," *Biomedical Signal Processing and Control*, vol. 48, pp. 35-45, 2019.

[14] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Evaluating the generalization of the hearing aid speech quality index (HASQI)," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 2, pp. 407-415, 2012.

[15] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75-93, 2014.

[16] X. Xu et al., "VSEGAN: Visual Speech Enhancement Generative Adversarial Network," *arXiv preprint arXiv:2102.02599*, 2021.

[17] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117-128, 2000.

[18] D. Michelsanti et al., "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421-2424, 2006.

[20] J. Besle, A. Fort, C. Delpuech, and M. H. Giard, "Bimodal speech: early suppressive visual effects in human auditory cortex," *Journal of Neuroscience of European*, vol. 20, no. 8, pp. 2225-2234, 2004.

[21] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”," *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417-1426, 2013.

[22] V. Van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proceedings of the National Academy of Sciences*, vol. 102, no. 4, pp. 1181-1186, 2005.

[23] C. E. Schroeder, P. Lakatos, Y. Kajikawa, S. Partan, and A. Puce, "Neuronal oscillations and visual amplification of speech," *Trends in cognitive sciences*, vol. 12, no. 3, pp. 106-113, 2008.

[24] L. H. Arnal, V. Wyart, and A. L. Giraud, "Transitions in neural oscillations reflect prediction errors generated in audiovisual speech," *Journal of Nature neuroscience*, vol. 14, no. 6, p. 797, 2011.

[25] J. C. Hou, S. S. Wang, Y. H. Lai, Y. Tsao, H. W. Chang, and H. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks,"in *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117-128, 2018.

[26] W. C. Yau, D. K. Kumar, and S. P. Arjunan, "Visual recognition of speech consonants using facial movement features," *Integrated computer-aided engineering*, vol. 14, no. 1, pp. 49-61, 2007.

[27] M. Breidt, H. H. Biilthoff and C. Curio, "Robust semantic analysis by synthesis of 3D facial motion," *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 713-719, 2011.

[28] M. Menze, C, Heipke, and A. Geiger, "Object scene flow," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 60-76, 2018.

[29] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 722-729, 1999.

[30] S. Vedula, P. Rander, R. Collins, T. Kanade, and m. intelligence, "Three-dimensional scene flow," in *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 475-480, 2005.

[31] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3061-3070, 2015.

[32] X. Liu, C. R. Qi and L. J. Guibas, "FlowNet3D: Learning Scene Flow in 3D Point Clouds," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 529-537, 2019.

[33] C. Vogel, K. Schindler, and S. Roth, "3d scene flow estimation with a piecewise rigid scene model" *International Journal of Computer Vision*, vol.115, pp. 1-28, 2015.

[34] Z. Aldeneh et al., "Self-supervised learning of visual speech features with audiovisual speech enhancement," *arXiv e-prints arXiv: 2004.12031*, 2020.

[35] Li Deng, J. Droppo and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," in *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 133-143, 2004.

[36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652-660, 2017.

[37] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," in *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1539-1548, 2017.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749-752, 2001.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE international conference on acoustics, speech and signal processing*, pp. 4214-4217, 2010.

[40] Y. H. Lai, P. C. Li, K. S. Tsai, W. C. Chu, and S. T. Young, "Measuring the long-term SNRs of static and adaptive compression amplification techniques for speech in noise," *Journal of the American Academy of Audiology*, vol. 24, no. 08, pp. 671-683, 2013.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering

Kaien Mo*, Xianrui Wang*†, Yichen Yang†, Tetsuya Ueda*,
Shoji Makino* and Jingdong Chen†
* Waseda University, Japan
† Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, China

***Abstract*—In order to improve both the separation performance and the convergence speed, several geometrically constrained independent vector analysis (GC-IVA) algorithms have been developed. Those algorithms are based on the multiplicative transfer function model, which assumes that the analysis window length is longer than the effective part of the room impulse responses. However, this assumption does often not hold in reverberant environments, particularly if the reverberation is strong, which makes the algorithms suffer from significant performance degradation. To circumvent this issue, an algorithm was developed, which jointly optimizes the weighted prediction error (WPE) dereverberation method and GC-IVA (GC-WPE-IVA). While it has demonstrated promising performance, this joint optimization method involves matrix inversion; so it is computationally very expensive. This work attempts to improve the efficiency and stability of GC-WPE-IVA. We develop an iterative source steering (ISS) updating algorithm in the framework of GC-WPE-IVA. The experimental results show that the developed method is computationally much more efficient yet it can achieve comparable separation performance in reverberation environments as compared to GC-WPE-IVA.**

I. Introduction

Speech enhancement is vital in speech communications systems as the source signals are always contaminated by interference, reverberation, and background noise [1]–[4]. When multiple sources are active simultaneously and no prior information is available, blind source separation (BSS) is usually used to separate source signals [1]. Independent component analysis (ICA) [5], one of the most adopted BSS methods, achieves source separation by exploiting the statistical independence among the sources. But it suffers from the problem of permutation. By considering the relationship among different frequency components, independent vector analysis (IVA) [6], [7], which is an extension of ICA, was developed to deal with the frequency permutation problem [8]. To accelerate the convergence and improve the robustness of IVA, the auxiliary function technique was developed [9] and the so-called iteration projection (IP) [9] and iterative source steering (ISS) [10] rules were proposed to optimize the auxiliary function. Although IVA has attracted great attention for its performance, the separated signals are in a random output order, which is known as the global permutation problem. In many applications, such as speech recognition, it is of great importance to match the separated sources and the corresponding speakers.

In real-world applications, the geometry of the array is fixed and the locations of the sources are often known or can be estimated [11]. To make use of such information, geometrically constrained independent vector analysis (GC-IVA) [12]–[16] algorithms were proposed. Recently, GC-IVA with the ISS-based updating rules (GC-IVA-ISS) was introduced, which demonstrated promising performance [17]; but its performance suffers from dramatic degradation if the reverberation is strong, which happens often in practice.

In order to mitigate the impact of reverberation on BSS, the weighted prediction error (WPE) [18], [19], one of the most widely used blind dereverberation methods, is usually implemented before separation [20]. Recently, convolutional beamformer (CBF) [21], which jointly updates the WPE and separation filters, was developed to improve the source extraction performance in strong reverberant environments. The source-wise factorization of the CBF [22] was further developed to reduce the computational complexity. ILRMA-T-ISS [23] is another work that deals with source separation in heavy reverberant environments with low computational cost. To cope with the problem of global permutation, the spatial information of the target source is used in the CBF framework (GC-WPE-IVA) [24], resulting in better performance than the conventional methods. Although GC-WPE-IVA is the state-of-the-art source separation method with dereverberation, it is still computationally costly due to the IP-based updating rules for the separation filters.

This work attempts to improve the efficiency and stability of the algorithm developed in [24]. We develop an ISS updating algorithm in the framework of GC-WPE-IVA, leading to a new algorithm called GC-WPE-IVA-ISS. Experiments are carried out and the results show that GC-WPE-IVA-ISS is computationally much more efficient yet it can achieve better or at least comparable separation performance in reverberation environments as compared to GC-WPE-IVA. Besides, since the spatial information of the sources is used, the global permutation problem is naturally solved with GC-WPE-IVA-ISS.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Signal model and problem formulation

Assume that there are N sources and M microphones. The source signals and observed signals can be expressed in a vector form as

$$\boldsymbol{s}_{f,t} = [s_{1,f,t} \quad s_{2,f,t} \quad \cdots \quad s_{N,f,t}]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}, \tag{1}$$

$$\boldsymbol{x}_{f,t} = [x_{1,f,t} \quad x_{2,f,t} \quad \cdots \quad x_{M,f,t}]^{\mathsf{T}} \in \mathbb{C}^{M\times 1}, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes transpose, $s_{n,f,t}$ and $x_{m,f,t}$ represent, respectively, the nth source signal and mth microphone observation signal in the short-time Fourier transformation (STFT) domain, $f = 1 \cdots F$ and $t = 1 \cdots T$ are frequency bin and time frame indexes with F and T being, respectively, the numbers of frequency bins and frames. $\boldsymbol{x}_{f,t}$ can be expressed as

$$\boldsymbol{x}_{f,t} = \sum_{\tau=0}^{L_A-1} \boldsymbol{A}_{f,\tau}\boldsymbol{s}_{f,t-\tau}, \tag{3}$$

where $\boldsymbol{A}_{f,\tau} \in \mathbb{C}^{M\times N}$ is the convolutional mixing matrix at time lag τ, and L_A is the order of time-lagged mixing filters. In this work, we only consider the determined case where $N = M$. Based on the multi-input multi-output (MIMO) CBF formulation [21], the beamformer's output can be expressed as

$$\boldsymbol{y}_{f,t} = \boldsymbol{W}_{f,0}\boldsymbol{x}_{f,t} + \sum_{\tau=D}^{L+D-1} \boldsymbol{W}_{f,\tau}\boldsymbol{x}_{f,t-\tau}, \tag{4}$$

where $\boldsymbol{W}_{f,0} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{W}_{f,\tau} \in \mathbb{C}^{N\times M}$ are the coefficient matrix of CBF, $\boldsymbol{y}_{f,t} = [y_{1,f,t} \quad y_{2,f,t} \quad \cdots \quad y_{N,f,t}]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}$ consists of the N separated signals, D and L are the time delay and the length of the CBF filters, respectively. According to source-wise factorization of CBF [22], the nth separated signal in $\boldsymbol{y}_{f,t}$ can be rewritten into two following form

$$\boldsymbol{z}_{n,f,t} = \boldsymbol{x}_{f,t} - \boldsymbol{G}_{n,f}^{\mathsf{H}}\overline{\boldsymbol{x}}_{f,t}, \tag{5}$$

$$y_{n,f,t} = \boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{z}_{n,f,t}, \tag{6}$$

where $(\cdot)^{\mathsf{H}}$ represents conjugate transpose, $\overline{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}_{f,t-D}^{\mathsf{T}} \quad \boldsymbol{x}_{f,t-D-1}^{\mathsf{T}} \quad \cdots \quad \boldsymbol{x}_{f,t-L-D+1}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML\times 1}$ contains past observed signals, $\boldsymbol{G}_{n,f} \in \mathbb{C}^{ML\times M}$ is the MIMO WPE filter, $\boldsymbol{z}_{n,f,t}$ is the dereverberated signal of nth source and $\boldsymbol{q}_{n,f}$ is the demixing filter. Note that (5) and (6) are strictly equal to (4) when $\boldsymbol{q}_{n,f}^{\mathsf{H}} = \boldsymbol{w}_{n,f,0}$ and $-\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{G}_{n,f}^{\mathsf{H}} = [\boldsymbol{w}_{n,f,D} \quad \boldsymbol{w}_{n,f,D+1} \quad \cdots \quad \boldsymbol{w}_{n,f,L+D-1}]$, where $\boldsymbol{w}_{n,f,\tau}$ is the nth row of $\boldsymbol{W}_{f,\tau}$.

To exploit the dependency between different frequency bins and thereby avoid the frequency permutation problem as IVA, CBF assumes that the sources follow a multivariate Gaussian distribution with time dependent variance $r_{n,t} = \sum_f |y_{n,f,t}|^2/F$. Therefore, the negative log-likelihood cost function can be derived as [22]

$$\mathcal{L} = -2\sum_f \log|\det \boldsymbol{Q}_f| + \frac{1}{T}\sum_{n,f,t}\left(\log r_{n,t} + \frac{|y_{n,f,t}|^2}{r_{n,t}}\right), \tag{7}$$

where $\boldsymbol{Q}_f = [\boldsymbol{q}_{1,f} \quad \boldsymbol{q}_{2,f} \quad \cdots \quad \boldsymbol{q}_{N,f}]^{\mathsf{H}}$ is the demixing matrix.

III. Proposed method

A. A Probabilistic model

In order to deal with the global permutation problem in strong reverberation conditions, we introduce the geometrical constraints [25] to the conventional CBF, which is expressed as

$$\mathcal{L}_{\mathrm{GC}} = \sum_{n,\varphi,f} \lambda_{n,\varphi}|\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2, \tag{8}$$

where $\lambda_{n,\varphi}$ is a non-negative weighting coefficient, $\boldsymbol{d}_{\varphi,f}$ is the steering vector along the direction φ and $c_{n,\varphi}$ is a non-negative-valued constraint. Note that if $c_{n,\varphi} = 1$, optimizing (8) gives the conventional delay-and-sum beamformer, which is steered to φ to extract the target signal whereas setting $c_{n,\varphi}$ to a small value will create a spatial null to suppress interference incident from the direction φ [14]. Combining (8) and (7) gives the cost function for the proposed method, i.e.,

$$\begin{aligned}\mathcal{L}(\Theta) = &-2\sum_f \log|\det \boldsymbol{Q}_f| + \frac{1}{T}\sum_{n,f,t}\left(\log r_{n,t} + \frac{|y_{n,f,t}|^2}{r_{n,t}}\right)\\ &+ \sum_{n,\varphi,f}\lambda_{n,\varphi}|\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2,\end{aligned} \tag{9}$$

where $\Theta = \{\Theta_{\boldsymbol{G}}, \Theta_{\boldsymbol{Q}}, \Theta_r\}$ is the parameter set to be estimated, $\Theta_{\boldsymbol{G}} = \{\boldsymbol{G}_{n,f}\}$, $\Theta_{\boldsymbol{Q}} = \{\boldsymbol{Q}_f\}$, and $\Theta_r = \{r_{n,t}\}$.

B. Optimization algorithm

According to the coordinate ascent method [20], each parameter in (9) can be updated iteratively by fixing the others until convergence.

1) Update of $\Theta_{\boldsymbol{G}}$: To update the WPE filter $\boldsymbol{G}_{n,f}$, the cost function (9) can be rewritten by fixing other parameters and ignoring the constant terms as

$$\mathcal{L}(\Theta_{\boldsymbol{G}}) = \frac{1}{T}\sum_{n,t,f}|\boldsymbol{q}_{n,f}^{\mathsf{H}}(\boldsymbol{x}_{f,t} - \boldsymbol{G}_{n,f}^{\mathsf{H}}\overline{\boldsymbol{x}}_{f,t})|^2/r_{n,t}. \tag{10}$$

Note that (10) is a quadratic function with respect to $\Theta_{\boldsymbol{G}}$. So, the parameter that minimizes (10) is [22]

$$\boldsymbol{G}_{n,f} \leftarrow \boldsymbol{R}_{n,f}^{-1}\boldsymbol{P}_{n,f}, \tag{11}$$

where

$$\boldsymbol{R}_{n,f} = \sum_t \frac{\overline{\boldsymbol{x}}_{f,t}\overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{r_{n,t}}, \tag{12}$$

$$\boldsymbol{P}_{n,f} = \sum_t \frac{\overline{\boldsymbol{x}}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}}{r_{n,t}}, \tag{13}$$

are spatio-temporal covariance matrices.

2) Update of $\Theta_{\boldsymbol{Q}}$: By fixing $\Theta_{\boldsymbol{G}}$ and Θ_r, the cost function (9) can be simplified to

$$\begin{aligned}\mathcal{L}(\Theta_{\boldsymbol{Q}}) = &-2\sum_f \log|\det \boldsymbol{Q}_f| + \sum_{n,f} \boldsymbol{q}_{n,f}^{\mathsf{H}} \boldsymbol{U}_{n,f} \boldsymbol{q}_{n,f} \\ &+ \sum_{n,\varphi,f} \lambda_{n,\varphi} |\boldsymbol{q}_{n,f}^{\mathsf{H}} \boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2,\end{aligned} \tag{14}$$

where

$$\boldsymbol{U}_{n,f} = \frac{1}{T}\sum_t \frac{\boldsymbol{z}_{n,f,t}\boldsymbol{z}_{n,f,t}^{\mathsf{H}}}{r_{n,t}} \tag{15}$$

is the weighted covariance matrix of $\boldsymbol{z}_{n,f,t}$. To accelerate the convergence of updating $\Theta_{\boldsymbol{Q}}$, instead of using the IP-based method that updates each row of $\boldsymbol{Q}_f$ iteratively [14] (which requires N times matrix inversion), the proposed method adopts the ISS updating rule [10] in which the whole filter is updated with a rank-1 matrix as

$$\boldsymbol{Q}_f \leftarrow \boldsymbol{Q}_f - \boldsymbol{v}_{k,f}\boldsymbol{q}_{k,f}^{\mathsf{H}}, \tag{16}$$

where $\boldsymbol{v}_{k,f} \in \mathbb{C}^{N\times 1}$ is a vector to be estimated. The update of $\boldsymbol{Q}_f$ will be repeated for $k = 1, ..., N$. Substituting (16) into the cost function (14) gives

$$\begin{aligned}\mathcal{L}_{\mathrm{ISS}}(\boldsymbol{v}_{k,f}) = &-2\sum_f \log|\det(\boldsymbol{Q}_f - \boldsymbol{v}_{k,f}\boldsymbol{q}_{k,f}^{\mathsf{H}})| \\ &+ \sum_{n,f} (\boldsymbol{q}_{n,f} - v_{n,k,f}^*\boldsymbol{q}_{k,f})^{\mathsf{H}} \boldsymbol{U}_{n,f} (\boldsymbol{q}_{n,f} - v_{n,k,f}^*\boldsymbol{q}_{k,f}) \\ &+ \sum_{n,\varphi,f} \lambda_{n,\varphi} |(\boldsymbol{q}_{n,f} - v_{n,k,f}^*\boldsymbol{q}_{k,f})^{\mathsf{H}}\boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2,\end{aligned} \tag{17}$$

where $v_{n,k,f}$ is the nth element of $\boldsymbol{v}_{k,f}$ and $(\cdot)^*$ denotes complex conjugate. Solving the equation $\partial\mathcal{L}_{\mathrm{ISS}}(\boldsymbol{v}_{k,f})/\partial v_{n,k,f}^* = 0$, which is inspired by the work in [17], we obtain the update rules for $v_{n,k,f}$, which is divided into the following two cases:

- $n \neq k$: In this case, the update rule is

$$v_{n,k,f} = \frac{\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{U}_{n,f}\boldsymbol{q}_{k,f} + \sum \lambda_{n,\varphi}(g_{n,\varphi,f} - c_{n,\varphi})g_{k,\varphi,f}^*}{\boldsymbol{q}_{k,f}^{\mathsf{H}}\boldsymbol{U}_{n,f}\boldsymbol{q}_{k,f} + \sum \lambda_{n,\varphi}|g_{k,\varphi,f}|^2}; \tag{18}$$

- $n = k$: In this case, $v_{k,k,f}$ is updated according to

$$v_{k,k,f} = \begin{cases} 1 - (\alpha_{k,f})^{-1/2} & \text{if } \beta_{k,f} = 0, \\ 1 - \beta_{k,f}^* \frac{|\beta_{k,f}| + \sqrt{|\beta_{k,f}^2| + 4\alpha_{k,f}}}{2\alpha_{k,f}|\beta_{k,f}|} & \text{otherwise,} \end{cases} \tag{19}$$

where

$$g_{n,\varphi,f} = \boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{d}_{\varphi,f}, \tag{20}$$

$$\alpha_{k,f} = \boldsymbol{q}_{k,f}^{\mathsf{H}}\boldsymbol{U}_{k,f}\boldsymbol{q}_{k,f} + \sum_{\varphi} \lambda_{k,\varphi}|g_{k,\varphi,f}|^2, \tag{21}$$

$$\beta_{k,f} = \sum_{\varphi} \lambda_{k,\varphi} c_{k,\varphi} g_{k,\varphi,f}. \tag{22}$$

Once $\boldsymbol{v}_{k,f}$ is obtained, $\boldsymbol{Q}_f$ can be updated through (16).

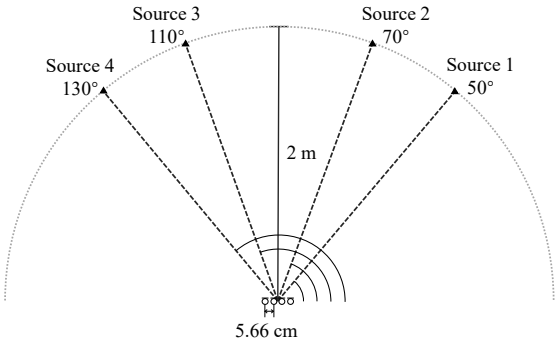


Fig. 1. Simulation layout.

3) Update of Θ_r: After updating the dereverberation ($\boldsymbol{G}_{n,f}$) and separating ($\boldsymbol{Q}_f$) filters, new separated signals $\boldsymbol{y}_{f,t}$ can be obtained through (5) and (6). Then, the time-varying variance $r_{n,t}$ is updated as

$$r_{n,t} \leftarrow \frac{1}{F}\sum_f |y_{n,f,t}|^2. \tag{23}$$

IV. Experiment

A. Experimental setup

The observation signals are generated by convolving the speech signals from the TIMIT database [26] with the room impulse responses (RIRs) from the RWCP dataset [27]. For every mixed signal, 4 speech segments, which are arbitrarily selected from different speakers, are concatenated to form 10 s long clean speech signals. The room reverberation time T_{60} is selected to 300 ms and 600 ms. The white Gaussian noise is added to control the signal-to-noise ratio (SNR) to 30 dB. The layout of sources and microphones is illustrated in Fig. 1, where a 4-element uniform linear microphone array (ULA) with an inter-element spacing of 5.66 cm is used. There are four sources in the sound field and the DOAs of the four sources are 50°, 70°, 110°, and 130°, respectively. The distance between the ULA center and the sources is 2 m. To measure the separation performance, twenty-five Monte Carlo simulations are carried out. All observed signals are sampled at 16 kHz and the STFT is conducted with the Hann window of 64 ms (1024 samples) and a window shift of 16 ms (256 samples).

We define two matrices, i.e., $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \ \ \boldsymbol{\lambda}_2 \ \cdots \ \boldsymbol{\lambda}_N]^{\mathsf{T}}$ and $\boldsymbol{C} = [\boldsymbol{c}_1 \ \ \boldsymbol{c}_2 \ \cdots \ \boldsymbol{c}_N]^{\mathsf{T}}$, with the parameters in (8) where $\boldsymbol{\lambda}_n = [\lambda_{n,\varphi_1} \ \ \lambda_{n,\varphi_2} \ \cdots \ \lambda_{n,\varphi_N}]^{\mathsf{T}} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{c}_n = [c_{n,\varphi_1} \ \ c_{n,\varphi_2} \ \cdots \ c_{n,\varphi_N}]^{\mathsf{T}} \in \mathbb{R}^{N\times 1}$ and $\Phi = \{\varphi_1, \varphi_2, \cdots, \varphi_N\}$ is the set of DOAs. According to the work in [17], we set $\boldsymbol{C} = \boldsymbol{E}$ so the output order is the same as the order of the input DOA and we use *NULL* constraints $\boldsymbol{\Lambda} = \Lambda(\boldsymbol{J} - \boldsymbol{E})$ where Λ is a non-negative number, $\boldsymbol{J}$ is an all-one matrix and $\boldsymbol{E}$ is the identity matrix. We initialize Λ as 8000 and decrease it over iterations according to [13], i.e.,

$$\Lambda_i = \Lambda_0\alpha^i, \tag{24}$$

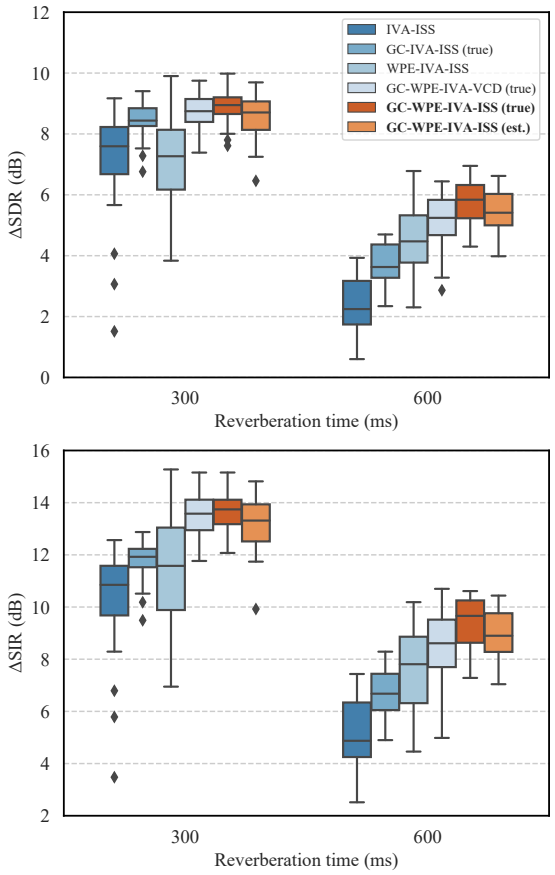


Fig. 2. Performance of the studied methods under different reverberation conditions.

where i represents the current number of iterations, Λ_0 is the initial value of Λ and the factor α is set to 0.8. For the WPE filter, the time delay D and filter length L are set to 2 and 10, respectively. The results for all the methods are obtained with 50 iterations and the WPE filter is updated every 10 iterations so both the convergence speed and computational cost are jointly considered. All the algorithms are implemented on a workstation powered by Intel Xeon E3-1505M. The improvement of signal-to-distortion ratio (ΔSDR) and signal-to-interference ratio (ΔSIR) [28] are used as the metrics to evaluate the separation performance.

The separation performance of the proposed algorithm (denoted as GC-WPE-IVA-ISS) is compared with IVA-ISS [10], GC-IVA-ISS [17], the joint optimization algorithm of WPE and IVA [22] with the ISS update method (WPE-IVA-ISS) and the geometrically constrained WPE-IVA with the vectorwise coordinate descent update method (GC-WPE-IVA-VCD) [24].

B. Experimental results

The SDR and SIR improvements of all the studied methods under different reverberation conditions are shown in Fig. 2. For the proposed GC-WPE-IVA-ISS, both the ground truth DOAs $\Phi_{\text{True}} = \{50°, 70°, 110°, 130°\}$ (denoted as true) and the DOAs with observation error $\Phi_{\text{True}} + \varepsilon$ (est.) are considered,

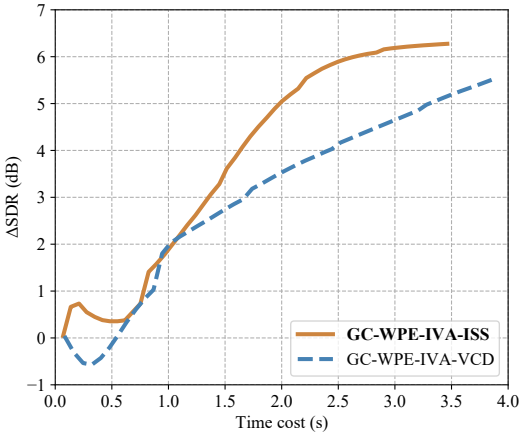


Fig. 3. Convergence speed of GC-WPE-IVA-ISS and GC-WPE-IVA-VCD.

where the error ε is computer generated random number with a uniform distribution in $[-15°, 15°]$ for every source. As seen from Fig 2, in the light reverberation condition with T_{60} = 300 ms, the ΔSDR of GC-IVA-ISS is close to that of the proposed method even without dereverberation. However, as the reverberation becomes stronger (with T_{60} = 600 ms), the separation performance of GC-IVA-ISS decreases significantly. In comparison, with the joint optimization of dereverberation filters and separation filters and the ISS-based updating rules, the proposed method generates a much better ΔSDR and ΔSIR than the compared conventional methods. In the presence of DOA errors, which is inevitable in practical applications with multi-sources and reverberation, the separation performance of the developed method decreases slightly. But its performance is still better than, or at least comparable to the compared conventional methods.

Figure 3 plots the convergence curves of GC-WPE-IVA-ISS and GC-WPE-IVA-VCD with ground true DOA. Note that the updating of the WPE filter is time-consuming. To better illustrate the acceleration of the convergence speed brought by the ISS-based optimization, the results in Fig. 3 show only the time cost for updating the separation filters. The results demonstrate that, due to the efficiency and stability of ISS, the developed method gets convergence with a much lower time cost than GC-WPE-IVA-VCD.

The accuracy of the output channel order of the proposed GC-WPE-IVA-ISS as well as GC-IVA-ISS, GC-WPE-IVA-VCD is listed in Table I. Note that the results of IVA-ISS and WPE-IVA-ISS are not presented here since their output orders are random. From Table I, it is seen that in the light reverberation condition, all the studied methods are able to produce accurate output order. But as the reverberation time becomes longer, the output accuracy of the conventional methods is affected while the proposed method can still guarantee the accuracy of the output order.

V. Conclusion

To make efficient use of the *a priori* spatial information of the sound sources, thereby improving the source separation

TABLE I
THE ACCURACY OF THE OUTPUT CHANNEL ORDER OF THE STUDIED METHODS UNDER DIFFERENT REVERBERATION CONDITIONS.

T_{60}	GC-IVA-ISS	GC-WPE-IVA-VCD	GC-WPE-IVA-ISS
300 ms	100%	100%	100%
600 ms	96%	96%	100%

performance in reverberant environments, we presented in this work a geometrically constrained algorithm, which jointly optimizes source separation and dereverberation filters with the ISS-based optimization. The experimental results show that the developed method is computationally much more efficient yet it can achieve better or at least comparable separation performance in reverberation environments as compared to the studied baseline algorithms. It is also robust to DOA estimation error. Another merit of the developed algorithm is that it is able to deal with the global permutation problem inherently.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] S. Makino, *Audio Source Separation*. Switzerland: Springer, 2018.
[2] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE Press., 2018.
[3] G. Huang, J. Benesty, and J. Chen, "Fundamental approaches to robust differential beamforming with high directivity factors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3074–3088, 2022.
[4] X. Wang, N. Pan, J. Benesty, and J. Chen, "On multiple input/binaural output antiphasic speaker signal extraction," in *Proc IEEE ICASSP*, 2023, pp. 1–5.
[5] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, 1st ed. Oxford, UK: Academic press/Elsevier, 2010.
[6] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.
[7] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.
[8] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Proc. IEEE ISCAS*, 2007, pp. 3247–3250.
[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, 2011, pp. 189–192.
[10] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.
[11] X. Wang, G. Huang, J. Benesty, J. Chen, and I. Cohen, "Time difference of arrival estimation based on a Kronecker product decomposition," *IEEE Signal Process. Lett.*, vol. 28, pp. 51–55, Dec. 2020.
[12] A. H. Khan, M. Taseska, and E. A. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.
[13] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. IEEE ICASSP*, 2018, pp. 746–750.
[14] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 846–850.
[15] Y. Yang, X. Wang, W. Zhang, and J. Chen, "Independent vector analysis assisted adaptive beamfomring for speech source separation with an acoustic vector sensor," in *Proc. IEEE IWAENC*, 2022, pp. 1–5.
[16] X. Wang, A. Brendel, G. Huang, Y. Yang, W. Kellermann, and J. Chen, "Spatially informed independent vector analysis for source extraction based on the convolutive transfer function model," in *Proc. IEEE ICASSP*, 2023, pp. 1–5, in press.
[17] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering," in *Proc. EUSIPCO*, 2022, pp. 757–761.
[18] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.
[19] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
[20] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2010.
[21] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2267–2282, Jul. 2020.
[22] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.
[23] T. Nakashima, R. Scheibler, M. Togami, and N. Ono, "Joint dereverberation and separation with iterative source steering," in *Proc. IEEE ICASSP*, 2021, pp. 216–220.
[24] Y. Yang, X. Wang, A. Brendel, W. Zhang, W. Kellermann, and J. Chen, "Geometrically constrained source extraction and dereverberation based on joint optimization," in *Proc. EUSIPCO*, 2023, in press.
[25] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 10, no. 6, pp. 352–362, Dec. 2002.
[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, Feb. 1993.
[27] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," pp. 965–968, 2000.
[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Study of Generative Adversarial Networks for Noisy Speech Simulation from Clean Speech

Leander Melroy Maben*, Zixun Guo†, Chen Chen†, Utkarsh Chudiwal‡ and Chng Eng Siong†
* Manipal Institute of Technology, Manipal, India
E-mail: leander.maben@gmail.com Tel/Fax: +91-8296366524
† Nanyang Technological University, Singapore
‡ Indian Institute of Technology, Indore, India

***Abstract*—The performance of speech processing models trained on clean speech drops significantly in noisy conditions. Training with noisy datasets alleviates the problem, but procuring such datasets is not always feasible. Noisy speech simulation models that generate noisy speech from clean speech help remedy this issue. In our work, we study the ability of Generative Adversarial Networks (GANs) to simulate a variety of noises. Noise from the Ultra-High-Frequency/Very-High-Frequency (UHF/VHF), additive stationary and non-stationary, and codec distortion categories are studied. We propose four GANs: the non-parallel translators, SpeechAttentionGAN, SimuGAN, and MaskCycleGAN-Augment, and the parallel translator, Speech2Speech-Augment. We achieved improvements of 55.8%, 28.9%, and 22.8% in terms of Multi-Scale Spectral Loss (MSSL) and 49.3%, 28.8%, and 18.2% in terms of Log Spectral Distance (LSD) as compared to the baseline for the RATS, TIMIT-Cabin, and TIMIT-Helicopter datasets, respectively, after training on small datasets of about 3 minutes.**

I. Introduction

Audio processing has wide-ranging applications in a multitude of tasks. Advancements in deep learning have further enabled the large-scale use of audio processing in tasks like Automatic Speech Recognition and Speech Enhancement [1]–[3]. The major challenge, however, is that the noisy environments that these models are exposed to in the real world severely impact their performance, making exposure to noise during training essential for good performance. Several works [4]–[6] use corpora of noises obtained from real-world settings. However, gathering noise datasets with sufficient variations and volume is not always economical or feasible.

Artificially generating noisy data is one viable option to remedy this issue. Several traditional models exist for audio augmentation, including parallel model matching [7], HMM adaptation using vector Taylor series [8], and stochastic matching [9]. There have also been works that directly add short static noise clips to clean audios for augmentation [10]. Obtaining significant durations of real-world noise, however, is often infeasible. Works like [11], and [12] use codecs to simulate distortion. Most of these traditional augmentation methods are based on using short clips of noise or pre-defined processing techniques. Using short noise clips entails the drawback of exposing the model to low levels of variation and hence lowers the model's generalizability to the broader variation levels present in the real world. Furthermore, if the augmentation techniques are predefined and not learned, they cannot adapt to different noise types.

Techniques involving deep learning or differentiable digital signal processing (DDSP) [13], [14] where model parameters are learned show potential in being more adaptable to different noise domains. Among deep learning techniques, Generative Adversarial Networks (GANs) show great promise in boosting model performance in noisy conditions due to their ability to simulate data realistically. GANs have the ability to perform both parallel and non-parallel translations from one domain to another. Parallel translation involves training the translation model with pairs of corresponding data points from both domains in a supervised fashion. On the other hand, non-parallel translation does not require training with input-target pairs, and hence, training is done in an unsupervised manner.

SimuGAN [15] has shown promise in simulating noise by learning noise statistics and has been shown to boost the performance of Automatic Speech Recognition (ASR). However, it has only been tested on a single type of noise and does not explore the usability of this GAN-based technique with other noise categories. DENT-DDSP [14] uses differentiable digital signal processing for parallel translation from the clean to the noisy domain in a very data-efficient manner. However, obtaining parallel data is difficult, and this may make the method unusable in applications where only non-parallel data is available.

There has been limited exploration of the ability of GANs to simulate noises of different types under conditions of low data volume constraints. In our work, we aim to explore the ability of four types of GANs to simulate four types of noises. We use SpeechAttentionGAN, MaskCycleGAN-Augment, and SimuGAN for non-parallel translation and Speech2Speech-Augment for parallel translation. SpeechAttentionGAN and MaskCycleGAN-Augment belong to the cycleGAN [16] family and use the 'Filling in Frames' technique during training. In addition, SpeechAttentionGAN uses attention masks for translation. SimuGAN uses the PatchNCE [17] loss for content preservation. Speech2Speech-Augment uses supervised learning with L1 loss to learn the translation. Furthermore, we use the RATS dataset, cabin and helicopter noise from the TIMIT dataset, and codec2 to represent the noise categories of Ultra-High-Frequency/Very-High-Frequency (UHF/VHF),

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

additive stationery, additive non-stationary, and codec-based distortion, respectively. We also compare the performance of noise simulation with GANs against a baseline.

Our work proves that GANs can be used to simulate noisy speech through experimental evidence. All four types of GANs show superior simulation quality in terms of LSD and MSSL when compared against the baseline. Unlike the traditional models, our proposed models learn to simulate the noise and are not based on a fixed processing algorithm or directly adding real-world noise clips. It thus provides great flexibility to be used for various kinds of noisy environments, as demonstrated by the experiments with multiple noise types. We also propose three GANs for non-parallel translation, which overcome the drawbacks of learning-based techniques like DENT-DDSP [14] that cannot be used without parallel data. Furthermore, our experiments were conducted using only about 3 minutes of clean and noisy clips each, demonstrating a way to effectively simulate noise even when the volume of data available is extremely low. Moreover, while traditional and learning-based methods involving DDSP cannot be used to simulate dynamic noise, we prove experimentally that GANs can successfully perform this task. Finally, the detailed experimental results provide researchers and developers of audio augmentation models with a comprehensive framework to use the correct type of GAN for the right kind of noise simulation task.

II. Methodology

A. Overview of the Audio Processing Pipeline

Figure 1 gives an overview of the prominent steps in the pipeline that convert a clean audio file to a noisy audio file. These steps are performed as follows. The loudness of the audio clip is first normalized to a fixed value. This is followed by Short Time Fourier Transform (STFT) to get a magnitude and phase spectrogram. Further processing is only done on the magnitude spectrogram. The phase spectrogram will be recycled and used in inverse STFT. The magnitude spectrogram goes through padding and componentization, which will be explained in detail in the following subsection. Each component then undergoes several processing steps, including conversion to decibels and scaling between 0 to 255. The magnitude spectrogram initially shows the lower frequencies on top and higher frequencies on the bottom. We flip it vertically to reverse this order. The components are then processed by the model to give us the generated components. The generated components are concatenated, and any padding added during componentization is removed. The magnitude spectrogram thus obtained is flipped, appropriately scaled, and converted to power from decibels. Inverse STFT is now performed using this generated magnitude spectrogram and the original phase spectrogram (from the clean audio) to get the target noisy audio clip.

B. Componentization

The lengths of spectrograms are variable and depend on the lengths of the audio clips. Longer audio clips would have longer spectrograms, while shorter audio clips would have

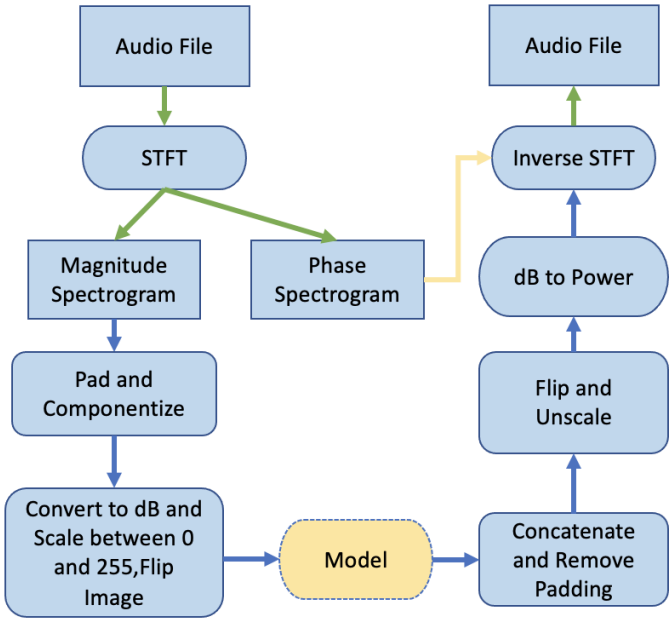


Fig. 1. Overview of the Audio Processing Pipeline. The trainable model is depicted by the yellow component. The flow of the magnitude spectrogram is depicted by blue arrows, the flow of the phase spectrogram is depicted by the yellow arrows, and the combined information flow is depicted by green arrows.

shorter ones along the horizontal axis. The height of the spectrogram depends on the window length (after padding) parameter of the STFT and would remain constant as this parameter is kept constant for all audio samples. During componentization, the spectrogram is broken into components of constant length, and the last component is padded with the required number of frames from the start of the spectrogram if it is not long enough. This ensures that the model inputs are kept at constant dimensions and helps a single audio clip serve as multiple training points if it is long enough.

C. GANS

1) SpeechAttentionGAN: SpeechAttentionGAN uses the base model architecture from AttentionGAN [18], [19], which was designed to be used for the non-parallel translation of images in a data-efficient manner. Information flow in SpeechAttentionGAN is depicted in figure 2.

In the proposed model, we introduce two novel modifications on top of the base model to help the model learn spectrogram structures better and to remedy the statistical averaging caused by the cycle consistency loss [16]. They are the addition of an input mask to implement the 'Filling in Frames' (FIF) approach in model training and a second adversarial loss, respectively.

To implement FIF, we generate a mask of zeros and ones by which the input spectrogram is then multiplied. The mask is generated in such a way that all frequencies for a random range of time frames of the spectrogram are removed during training. The masked spectrogram and the mask are given as inputs to

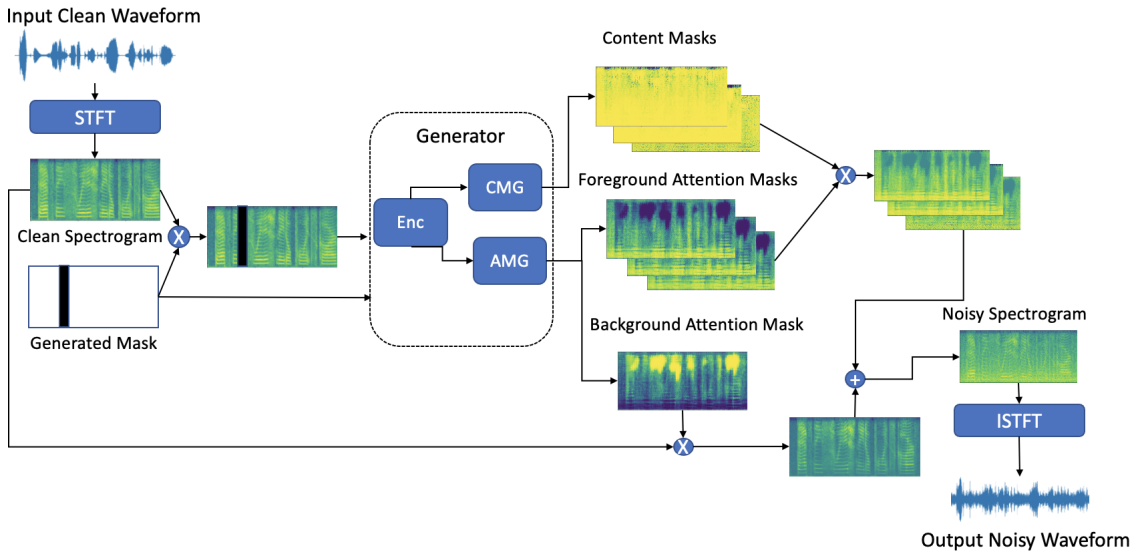


Fig. 2. SpeechAttentionGAN. The figure depicts the overall flow of data in one direction in the SpeechAttentionGAN pipeline with Filling in Frames. The generator consists of three parts - the Encoder (Enc), the Attention Mask Generator(AMG), and the Content Mask Generator(CMG). In this figure, spectrogram refers to magnitude spectrogram, STFT refers to Short-Time Fourier Transform, and ISTFT refers to Inverse Short-Time Fourier Transform.

the model. By training in this fashion, the model is forced to predict the missing content from the surrounding information and thus implicitly learn the spectrogram's structure. During testing, the entire mask is set to one, and no information loss takes place.

The second adversarial loss [20] is very similar in structure to the general adversarial loss [21] used in GANs. The difference lies in the fact that instead of using the real input and the output generated by one generator, the second adversarial loss uses real inputs and cycled outputs (outputs that have been cycled through both generators back to the original domain). The second adversarial loss for the translation from domain A to domain B and back to domain A (A→B→A) is described in (1). The second adversarial loss for the direction B→A→B is computed in a similar manner. In order to compute the second adversarial loss, we use two additional discriminators with identical architecture to the original two. We use one discriminator each to compute this loss in the two directions.

$$L_{adv2}^{A->B->A} = E_{a\sim A}[D'_A(a)] + E_{a\sim A}[D'_A(G_{B->A}(G_{A->B}(a)))] \quad (1)$$

In (1), $G_{A->B}$ represents the generator that translates from domain A to domain B, $G_{B->A}$ represents the generator that translates from domain B to domain A, D'_A represents the discriminator used to compute this loss in the direction A→B→A, and a represents a data point from domain A.

2) MaskCycleGAN-Augment: The MaskCycleGAN-Augment model uses the base architecture and framework followed by MaskCycleGAN-VC [20]. MaskCycleGAN-VC uses MelGAN [22] for the conversion between audio files and spectrograms. This approach, however, does not work well for our use case of converting clean speech to noisy speech. In particular, upon conversion of a noisy spectrogram to an audio clip, MelGAN performs poorly and outputs unintelligible clips. The reason for this behavior is that MelGAN is not explicitly designed to work with noisy spectrograms or noisy audio clips and, consequently, does not work well in our use case, which involves noisy audio clips and spectrograms. As a remedy to this issue, MaskCycleGAN-Augment uses the audio processing pipeline described in subsection II-A.

3) SimuGAN: SimuGAN [15] tackles the problem of non-parallel translation and does not use cycle consistency loss [16]. Instead, it promotes content preservation by encouraging the model to generate outputs such that the corresponding patches in the input and output spectrograms have more similar latent representations than the non-corresponding patches using a patchNCE loss [17].

4) Speech2Speech-Augment: Speech2Speech-Augment uses the base architecture of the Pix2Pix [23] model, which was designed for parallel image translation. It uses the adversarial loss [21] along with an L1 loss to encourage the model to produce outputs close to the expected targets.

III. Experiments and Results

A. Datasets

In order to study the potency and performance of GANs in simulating a wide range of noises, we select noises from 4 different categories. These include Ultra-High-Frequency/Very-High-Frequency (UHF/VHF) noise represented by the RATS dataset [24], additive stationary and non-stationary noises represented by the cabin and helicopter noises from the TIMIT dataset [25], respectively, and codec distortion generated using codec2 [26]. The RATS dataset consists of clean and noisy speech. The TIMIT dataset, however, consists of clean clips along with clips of pure noise. We use the method described in [27] to generate noisy speech clips using clean clips and

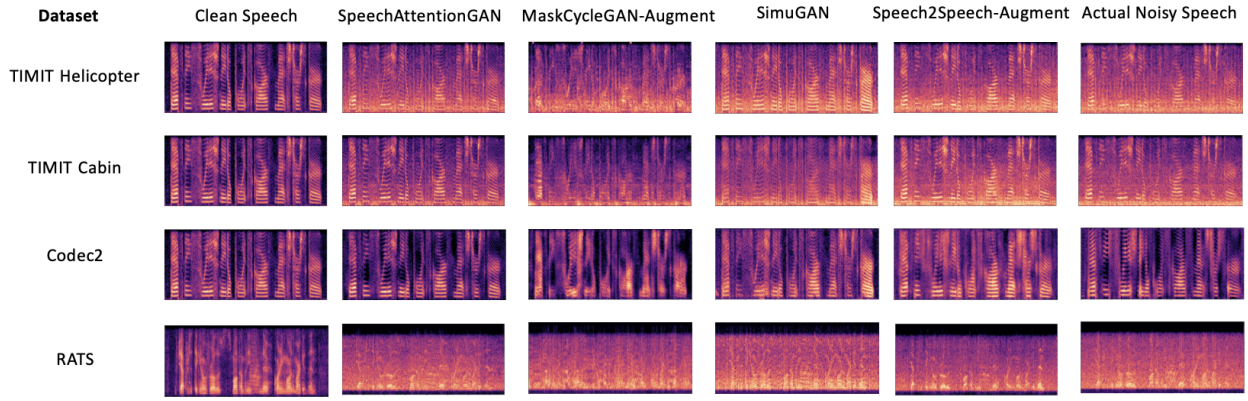


Fig. 3. Generated Samples. This figure shows the noisy spectrogram samples generated by the four models for each of the four noise types. It also shows the clean and actual noisy spectrogram samples from the four datasets.

noise clips. Similarly, we use the clean clips from the TIMIT dataset and pass them through the codec to get noisy speech clips with codec distortion. All datasets used are in the English language.

We prepare both parallel and non-parallel training sets for each type of noise. The training sets for all types of noises include approximately 3 minutes each of noisy and clean audio clips. The validation and test sets, respectively, contain approximately 100 and 200 seconds each of clean and noisy audio clips.

Furthermore, for the TIMIT stationary and non-stationary datasets, as well as the codec2 dataset, we ensured that there were no common speakers between any two of the three splits.

B. Experimental Setup

During training, model checkpoints are saved once every fixed number of epochs. In the validation phase, the validation set is used to evaluate model performance at different checkpoints. The model checkpoint with the best performance on the validation set is picked and used to generate simulations on the test set.

Furthermore, as a second mode of evaluation, we also calculate performance metrics across all checkpoints to compute the average performance of the model when we do not have a parallel validation dataset to pick the best checkpoint.

We use the augmentation technique of aggregating noise and adding it to the clean speech, then passing it through a codec (g726[28]) to introduce distortions as a baseline.

Generated samples from experiments can be heard at this website [1]. The codebase for training and testing the GANs can be found at this online repository[2].

C. Metrics Used

To measure model performance, we use the metrics Log Spectral Distance (LSD) [29] and Multi-Scale Spectral Loss (MSSL)[13]. Both of these metrics measure the distance or dissimilarity between the generated output and the target output, which implies that the lower the value of these metrics, the better the quality of the generated output. Prior to metric computation, the ground truth and generated audio clips go through normalization, energy equalization, and time alignment.

$$LSD = \frac{1}{L}\sum_{l=0}^{L-1}\sqrt{\frac{1}{N}\sum_{k=0}^{N-1}[10\log_{10}\frac{|\hat{\chi}(l,k)|^2}{|\chi(l,k)|^2}]^2} \quad (2)$$

The computation of LSD is described in (2), where l denotes the frame index, k denotes the index of the frequency bin, L is the total number of frames, and N is the total number of frequency bins. $\chi(l,k)$ and $\chi(\hat{l},k)$ denote the Discrete Fourier Transform (DFT) coefficients for the ground truth and generated audio clips, respectively, at the lth frame and kth frequency bin.

MSSL provides the benefit of averaging the distance between the spectrograms computed using multiple window sizes during STFT. A lower window size results in spectrograms with higher resolution in the time domain but lower resolution in the frequency domain. On the other hand, a higher window size results in spectrograms with higher resolution along the frequency domain but a lower resolution along the time domain. MSSL helps to average out this trade-off to an extent by taking into account multiple window sizes.

D. Result Analysis

The metrics computed for the test set using the GAN model checkpoints with the best validation MSSL are shown in Table I. The average of metrics for the validation set computed across all the checkpoints of the GANs is shown in Table II. Sample spectrograms from the experiments are depicted in figure 3

1) RATS Dataset (UHF/VHF): For the RATS dataset, we find that SimuGAN outperforms the other models when we pick the best checkpoint using a validation set. It outperforms the baseline by 49.3% and 55.8% with respect to LSD and MSSL, respectively. However, on computing the average of

[1]https://github.com/leandermaben/GANSpeechAugment
[2]https://leandermaben.github.io/GANSpeechAugment/

TABLE I
RESULTS USING THE VALIDATION SET TO PICK THE BEST CHECKPOINT. LSD STANDS FOR LOG SPECTRAL DISTANCE, AND MSSL STANDS FOR MULTI-SCALE SPECTRAL LOSS. CABIN AND HELICOPTER CORRESPOND TO NOISE TYPES FROM THE TIMIT DATASET.

Dataset	Model	Mean LSD	Mean MSSL	Mode
RATS	SpeechAttentionGAN	7.29	5.98	Non Parallel
	MaskCycleGAN-Augment	7.24	6.10	Non Parallel
	SimuGAN	**6.23**	**5.26**	Non Parallel
	Speech2Speech-Augment	7.02	5.83	Parallel
	Baseline	12.29	11.91	-
Cabin	SpeechAttentionGAN	6.35	4.92	Non Parallel
	MaskCycleGAN-Augment	7.56	6.12	Non Parallel
	SimuGAN	**5.48**	4.84	Non Parallel
	Speech2Speech-Augment	5.95	**4.69**	Parallel
	Baseline	7.70	6.60	-
Helicopter	SpeechAttentionGAN	7.16	**5.32**	Non Parallel
	MaskCycleGAN-Augment	8.50	6.48	Non Parallel
	SimuGAN	**6.98**	5.61	Non Parallel
	Speech2Speech-Augment	7.17	5.39	Parallel
	Baseline	8.53	6.90	-
Codec2	SpeechAttentionGAN	**6.18**	6.78	Non Parallel
	MaskCycleGAN-Augment	8.09	7.54	Non Parallel
	SimuGAN	6.28	7.04	Non Parallel
	Speech2Speech-Augment	6.55	**6.58**	Parallel

metrics across all checkpoints, we find that Speech2Speech-Augment has the best results with an improvement of 43.5% and 50.5% as compared to the baseline in terms of LSD and MSSL, respectively.

2) TIMIT Cabin Noise (Stationary Additive): In Table I, in terms of LSD, SimuGAN gives the best performance with an improvement of 28.8% as compared to the baseline. Speech2Speech-Augment performs the best in terms of MSSL, with an improvement of 28.9% with respect to the baseline. Using the second mode of evaluation as shown in Table II, we find that Speech2Speech-Augment gives the best performance in terms of both LSD and MSSL with improvements of 17.1% and 21.2% respectively, as compared to the baseline.

3) TIMIT Helicopter Noise (Non-Stationary Additive): Using the validation set, we find that SimuGAN outperforms other techniques in terms of LSD with an improvement of 18.1% over the baseline. SpeechAttentionGAN performs the best in terms of MSSL, with an improvement of 22.8% with respect to the baseline when we use the same evaluation metric. Using the average across checkpoints, we find that Speech2Speech-Augment outperforms the other models in terms of both LSD and MSSL with improvements of 11.7% and 16.9%, respectively, over the baseline.

4) Codec2: For Codec2 distortions, we find that SpeechAttentionGAN gives the most desirable LSD in both evaluation modes. Speech2Speech-Augment gives the most desirable MSSL in both modes.

E. Limitations

This work has two main limitations that can be explored and worked upon in future research. Firstly, we do not study how the performance of GANs varies across different languages, accent variations, age groups, and genders. Secondly, we only generate the magnitude spectrogram using GANs and recycle the phase spectrogram from the clean speech. Using GANs or other models to simulate the phase spectrogram would be a promising exploration avenue to improve simulation performance further.

IV. CONCLUSIONS

We perform a detailed study of the use of GANs in simulating different varieties of noises, including UHF/VHF, Additive Stationary and Non-Stationary, and Codec distortion. Through the study, we demonstrate that different types of GANs can be successfully used for noise simulation and that the performance of these GANs varies depending on the noise types. In addition, we show that GANs can be used for noise simulation even with small training datasets of the magnitude of 3 minutes. We also find that GANs give significant improvements over the baseline technique of using aggregated noise followed by codec (g726) distortion. We explore both parallel and non-parallel translation to evaluate the capacity of GANs in both modes of translation. These results would, in turn, help train more robust models for downstream processes. Finally, we document the performance of the various GANs on different datasets in terms of LSD and MSSL, which would help provide a framework to pick the right type of GAN for the desired task.

TABLE II
AVERAGE METRICS OVER ALL CHECKPOINTS. LSD STANDS FOR LOG SPECTRAL DISTANCE, AND MSSL STANDS FOR MULTI-SCALE SPECTRAL LOSS. CABIN AND HELICOPTER CORRESPOND TO NOISE TYPES FROM THE TIMIT DATASET.

Dataset	Model	Mean LSD	Mean MSSL	Mode
RATS	SpeechAttentionGAN	7.51	6.24	Non Parallel
	MaskCycleGAN-Augment	7.51	6.32	Non Parallel
	SimuGAN	7.77	7.82	Non Parallel
	Speech2Speech-Augment	**6.94**	**5.90**	Parallel
	Baseline	12.29	11.91	-
Cabin	SpeechAttentionGAN	7.04	5.71	Non Parallel
	MaskCycleGAN-Augment	7.91	6.61	Non Parallel
	SimuGAN	6.61	8.09	Non Parallel
	Speech2Speech-Augment	**6.38**	**5.20**	Parallel
	Baseline	7.70	6.60	-
Helicopter	SpeechAttentionGAN	7.86	5.89	Non Parallel
	MaskCycleGAN-Augment	9.01	7.15	Non Parallel
	SimuGAN	8.32	7.20	Non Parallel
	Speech2Speech-Augment	**7.53**	**5.73**	Parallel
	Baseline	8.53	6.90	-
Codec2	SpeechAttentionGAN	**6.05**	7.11	Non Parallel
	MaskCycleGAN-Augment	7.64	8.00	Non Parallel
	SimuGAN	6.96	9.14	Non Parallel
	Speech2Speech-Augment	6.55	**6.55**	Parallel

REFERENCES

[1] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 167–174. DOI: 10.1109/ASRU.2015.7404790.

[2] A. Hannun, C. Case, J. Casper, *et al.*, *Deep speech: Scaling up end-to-end speech recognition*, 2014. DOI: 10.48550/ARXIV.1412.5567. [Online]. Available: https://arxiv.org/abs/1412.5567.

[3] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by noisy2noisy signal mapping," *Applied Acoustics*, vol. 172, p. 107 631, 2021, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2020.107631. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X20307350.

[4] S. Badrinath and H. Balakrishnan, "Automatic speech recognition for air traffic control communications," *Transportation research record*, vol. 2676, no. 1, pp. 798–810, 2022.

[5] T. Pellegrini, J. Farinas, E. Delpech, and F. Lancelot, "The Airbus Air Traffic Control Speech Recognition 2018 Challenge: Towards ATC Automatic Transcription and Call Sign Detection," in *Proc. Interspeech 2019*, 2019, pp. 2993–2997. DOI: 10.21437/Interspeech.2019-1962.

[6] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, IEEE Computer Society, vol. 1, 1992, pp. 517–520.

[7] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Citeseer, 1995.

[8] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition.," in *INTERSPEECH*, 2000, pp. 869–872.

[9] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 1, 1995, pp. 121–124.

[10] D. Ma, G. Li, H. Xu, and E. S. Chng, "Improving code-switching speech recognition with data augmentation and system combination," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2019, pp. 1308–1312.

[11] M. Ferras, S. Madikeri, P. Motlicek, S. Dey, and H. Bourlard, "A large-scale open-source acoustic simulator for speaker recognition," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 527–531, 2016.

[12] T.-L. Vu, Z. Zeng, H. Xu, and E.-S. Chng, "Audio codec simulation based data augmentation for telephony speech recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 198–203. DOI: 10.1109/APSIPAASC47483.2019.9023257.

[13] J. Engel, L. (Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr.

[14] Z. Guo, C. Chen, and E. S. Chng, "DENT-DDSP: Data-efficient noisy speech generator using differentiable digital signal processors for explicit distortion modelling and noise-robust speech recognition," in *Proc. Interspeech 2022*, 2022, pp. 3799–3803. DOI: 10.21437/Interspeech.2022-763.

[15] C. Chen, N. Hou, Y. Hu, S. Shirol, and E. S. Chng, "Noise-robust speech recognition with 10 minutes unparalleled in-domain data," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4298–4302. DOI: 10.1109/ICASSP43922.2022.9747755.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

[17] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 319–345, ISBN: 978-3-030-58545-7.

[18] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

[19] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *International Joint Conference on Neural Networks (IJCNN)*, 2019.

[20] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5919–5923. DOI: 10.1109/ICASSP39728.2021.9414851.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[22] K. Kumar, R. Kumar, T. de Boissiere, *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632.

[24] *RATS Speech Activity Detection - Linguistic Data Consortium — catalog.ldc.upenn.edu*, https://catalog.ldc.upenn.edu/LDC2015S02, [Accessed 16-Aug-2022].

[25] *TIMIT Acoustic-Phonetic Continuous Speech Corpus - Linguistic Data Consortium — catalog.ldc.upenn.edu*, https://catalog.ldc.upenn.edu/LDC93s1, [Accessed 16-Aug-2022].

[26] *Codec 2; Rowetel — rowetel.com*, https://www.rowetel.com/wordpress/?page_id=452, [Accessed 16-Aug-2022].

[27] H.-Y. Lin, H.-H. Tseng, X. Lu, and Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 19 935–19 946. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/a5c7b30fb632c92feb59154517223dc9-Paper.pdf.

[28] G. R. Inc., *G.726 Speech Vocoder and Audio Codec (G726 Speech and Audio Compression and Processing Software) - GAO Research — gaoresearch.com*, http://www.gaoresearch.com/products/speechsoftware/other/g726.php, [Accessed 16-Aug-2022].

[29] S. Shirol, *Speech data augmentation in the frequency domain using deep learning methods*, https://shashankshirol.github.io/assets/pdfs/shan-bachelor-thesis.pdf, [Accessed 16-Aug-2022].

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Step Size Control of Shared-error Normalized Least Mean Square Algorithm for Acoustic Echo and Noise Canceller

Kenta Iwai* and Takanobu Nishiura*
* Ritsumeikan University, Japan
E-mail: {iwai18sp@fc}{nishiura@is}.ritsumei.ac.jp

***Abstract*—We propose a step size control method for an acoustic echo and noise canceller (AENC) based on a shared-error normalized least mean square (SENLMS) algorithm. The conventional AENC based on the SENLMS algorithm can optimize both the acoustic echo canceller (AEC) and the adaptive noise canceller (ANC). However, it is difficult to set the step size parameters to prevent the divergence of the conventional AENC because of the optimization of both AEC and ANC. To solve this problem, we investigate the constraint for the step size parameters of the AENC with the SENLMS algorithm and propose a step size control method. The constraint for the step size parameters is derived from the mean square value of the *a posteriori* error. Then, the step size parameter of the AEC is determined by the nonparametric variable step size (VSS) NLMS algorithm. Finally, the step size parameter of the ANC is automatically determined by the constraint and the variable step size parameter of the AEC. Simulation results show that the proposed AENC with the VSS-SENLMS algorithm can simultaneously prevent the divergence and maintain its the capability to reduce background noise and acoustic echo.**

I. Introduction

Hands-free telecommunication and video conference systems have developed significantly in recent years [1], [2]. Although the World Health Organization (WHO) had declared that Coronavirus disease (COVID-19) no longer represents a global health emergency [3], video conference systems continue to be used in many situations. The problems plaguing these systems are disturbance caused by acoustic echo and background noise. An acoustic echo is generated by acoustic coupling between a loudspeaker and a microphone. Telecommunication systems also suffer from the noise problem. To solve these problems, acoustic echo and noise cancellers (AENCs) have been developed [4]–[8]. An AENC [4]–[7] mainly consists of an acoustic echo canceller (AEC) and a noise suppressor.

We have proposed the AENC based on a shared-error normalized least mean square (SENLMS) algorithm [9]. This AENC consists of an AEC and Widrow's adaptive noise canceller (ANC) [10] with time-domain adaptive digital filters and can optimize both the AEC and ANC by the SENLMS algorithm. However, it is difficult to set the step size parameters to prevent the divergence of the conventional AENC because of the optimization of both AEC and ANC.

In this study, we investigate the constraint for the step size parameters of the AENC with the SENLMS algorithm and propose a step size control method. The constraint for the step size parameters is derived from the mean square value of the *a posteriori* error. Then, the control method for the step size parameters are derived on the basis of the constraint. The capability to reduce the acoustic echo and background noise is evaluated by computer simulations.

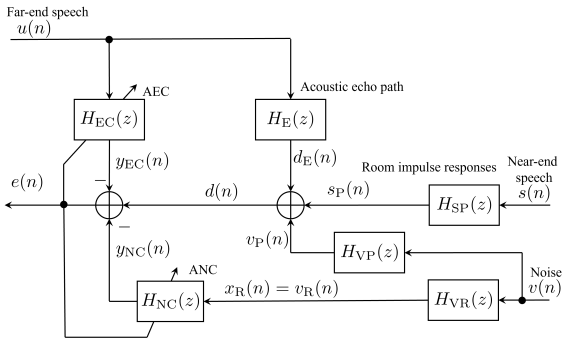


Fig. 1: Block diagram of AENC with shared-error NLMS (SENLMS) algorithm.

II. AENC with SENLMS algorithm [9]

This AENC utilizes a SENLMS algorithm that updates the coefficients of two adaptive filters for echo cancelling and noise reduction by using a shared error signal. The block diagram of the AENC utilizing the SENLMS algorithm is shown in Fig. 1. As shown in Fig. 1, the AENC with the SENLMS algorithm combines the AEC with the time-domain adaptive filter and Widrow's ANC [10]. With this structure, the AENC with the SENLMS algorithm can optimize the adaptive filters for the ANC and AEC simultaneously.

When the ANC and AEC are connected in parallel, their output signals interfere with each other; hence, their error signals become the same. Here, the error signals of the ANC and AEC are expressed as

$$e_{\mathrm{NC}}(n) = v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + d_{\mathrm{E}}(n) - y_{\mathrm{EC}}(n) + s_{\mathrm{P}}(n), \quad (1)$$

$$e_{\mathrm{EC}}(n) = d_{\mathrm{E}}(n) - y_{\mathrm{EC}}(n) + v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n), \quad (2)$$

where $v_{\mathrm{P}}(n)$, $d_{\mathrm{E}}(n)$, and $s_{\mathrm{P}}(n)$ are the background noise, acoustic echo, and near-end speech, and $y_{\mathrm{NC}}(n) = \mathbf{h}_{\mathrm{NC}}^{\mathrm{T}}(n)\mathbf{x}_{\mathrm{R}}(n)$ and $y_{\mathrm{EC}}(n) = \mathbf{h}_{\mathrm{EC}}^{\mathrm{T}}(n)\mathbf{u}(n)$ are the output signals of the ANC and AEC, respectively. Also, $\mathbf{h}_{\mathrm{NC}}(n)$ and $\mathbf{h}_{\mathrm{EC}}(n)$ are the filter coefficient vectors of the ANC and AEC,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

$\mathbf{x}_{\mathrm{R}}(n)$ and $\mathbf{u}(n)$ are the input signal vectors of the ANC and AEC, respectively. As can be seen from (1) and (2), the error signals $e_{\mathrm{NC}}(n)$ and $e_{\mathrm{EC}}(n)$ are the same. In Fig. 1, $d_{\mathrm{E}}(n)$ is disturbance for the ANC. In the case that the AEC converges to the optimal filter, the output of the AEC becomes

$$y_{\mathrm{EC}}(n) \approx d_{\mathrm{E}}(n). \tag{3}$$

Then, $d_{\mathrm{E}}(n)$ is cancelled by $y_{\mathrm{EC}}(n)$ and the error signal of the ANC becomes

$$e_{\mathrm{NC}}(n) \approx v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n). \tag{4}$$

Equation (4) is the ideal condition of the ANC. Similarly, $v_{\mathrm{P}}(n)$ is the disturbance for the AEC. In the case that the ANC converges to the optimal filter, the output of the ANC becomes

$$y_{\mathrm{NC}}(n) \approx v_{\mathrm{P}}(n). \tag{5}$$

Then, $v_{\mathrm{P}}(n)$ is cancelled by $y_{\mathrm{NC}}(n)$ and the error signal of the AEC becomes

$$e_{\mathrm{EC}}(n) \approx d_{\mathrm{E}}(n) - y_{\mathrm{EC}}(n) + s_{\mathrm{P}}(n). \tag{6}$$

Equation (6) is the ideal condition of the AEC. Thus, the AENC with the SENLMS algorithm realizes the optimal conditions of ANC and AEC simultaneously.

The adaptive filters of the ANC and AEC are updated by a normalized least mean square (NLMS) algorithm [11]–[13] combined with the shared error signal $e(n)$. The cost functions of the ANC and AEC are defined as

$$J_{\mathrm{NC}}(n) = e^2(n)/\|\mathbf{x}_{\mathrm{R}}(n)\|^2, \tag{7}$$
$$J_{\mathrm{EC}}(n) = e^2(n)/\|\mathbf{u}(n)\|^2, \tag{8}$$

where $e(n) = e_{\mathrm{NC}}(n) = e_{\mathrm{EC}}(n)$ is the shared error signal and $\|\cdot\|$ represents the norm of the vector. By partial differentiation of $J_{\mathrm{NC}}(n)$ and $J_{\mathrm{EC}}(n)$ with respect to $\mathbf{h}_{\mathrm{NC}}^{\mathrm{T}}(n)$ and $\mathbf{h}_{\mathrm{EC}}^{\mathrm{T}}(n)$, we can obtain the update rules of the ANC and AEC as

$$\mathbf{h}_{\mathrm{NC}}(n+1) = \mathbf{h}_{\mathrm{NC}}(n) + \alpha_{\mathrm{NC}} \frac{e(n)\mathbf{x}_{\mathrm{R}}(n)}{\|\mathbf{x}_{\mathrm{R}}(n)\|^2 + \beta_{\mathrm{NC}}}, \tag{9}$$
$$\mathbf{h}_{\mathrm{EC}}(n+1) = \mathbf{h}_{\mathrm{EC}}(n) + \alpha_{\mathrm{EC}} \frac{e(n)\mathbf{u}(n)}{\|\mathbf{u}(n)\|^2 + \beta_{\mathrm{EC}}}, \tag{10}$$

where α_{NC} and α_{EC} are the step size parameters of the ANC and AEC, and β_{NC} and β_{EC} are the regularization parameters of the ANC and AEC, respectively.

However, it is difficult to set the step size parameters to prevent the divergence of the conventional AENC because of the joint optimization of the ANC and AEC. Hence, we investigate the constraint for the step size parameters and propose the control method for the step size parameters.

III. Step size control of SENLMS algorithm

As mentioned in Sect. II, the AENC with the SENLMS algorithm has the problem of divergence depending on the fixed step size parameters. In this section, we investigate the constraint for the step size parameters and propose the control method for the step size parameters.

A. A posteriori error of SENLMS algorithm

In [14], a nonparametric VSS-NLMS (NP-VSS-NLMS) algorithm is derived via a mean square value of the *a posteriori* error signal of an adaptive digital filter. We investigate the constraint for the step size parameters on the basis of this idea. Firstly, the *a posteriori* error signal of the AENC with the SENLMS algorithm is represented by

$$\begin{aligned}\varepsilon(n) = {} & v_{\mathrm{P}}(n) - \mathbf{h}_{\mathrm{NC}}^{\mathrm{T}}(n+1)\mathbf{x}_{\mathrm{R}}(n) \\ & + d_{\mathrm{E}}(n) - \mathbf{h}_{\mathrm{EC}}^{\mathrm{T}}(n+1)\mathbf{u}(n) + s_{\mathrm{P}}(n).\end{aligned} \tag{11}$$

Here, we modify the update equations (9) and (10) as

$$\mathbf{h}_{\mathrm{NC}}(n+1) = \mathbf{h}_{\mathrm{NC}}(n) + \alpha_{\mathrm{NC}}(n)e(n)\mathbf{x}_{\mathrm{R}}(n), \tag{12}$$
$$\mathbf{h}_{\mathrm{EC}}(n+1) = \mathbf{h}_{\mathrm{EC}}(n) + \alpha_{\mathrm{EC}}(n)e(n)\mathbf{u}(n), \tag{13}$$

where $\alpha_{\mathrm{NC}}(n)$ and $\alpha_{\mathrm{EC}}(n)$ are the variable step size parameters of the ANC and AEC, respectively. From (11)–(13) the mean square value of (11) is represented by

$$\mathbb{E}\left[\varepsilon^2(n)\right] = \sigma_e^2(n)\left[1 - \|\mathbf{x}_{\mathrm{R}}(n)\|^2\alpha_{\mathrm{NC}}(n) - \|\mathbf{u}(n)\|^2\alpha_{\mathrm{EC}}(n)\right]^2, \tag{14}$$

where $\sigma_e^2(n) = \mathbb{E}\left[e^2(n)\right]$. Here, $\|\mathbf{x}_{\mathrm{R}}(n)\|^2$ and $\|\mathbf{u}(n)\|^2$ are treated as constants. Since the *a posteriori* error is obtained using $\mathbf{h}_{\mathrm{NC}}(n+1)$ and $\mathbf{h}_{\mathrm{EC}}(n+1)$, then $\mathbb{E}\left[\varepsilon^2(n)\right] = \mathbb{E}\left[s_{\mathrm{P}}^2(n)\right] = \sigma_s^2(n)$. Hence, the constraint of the variable step size parameters are derived as

$$\|\mathbf{x}_{\mathrm{R}}(n)\|^2\alpha_{\mathrm{NC}}(n) + \|\mathbf{u}(n)\|^2\alpha_{\mathrm{EC}}(n) = 1 - \sqrt{\frac{\sigma_s^2(n)}{\sigma_e^2(n)}}. \tag{15}$$

Here, the LHS of (15) is a function of $\alpha_{\mathrm{NC}}(n)$ and $\alpha_{\mathrm{EC}}(n)$. Therefore, (15) cannot be solved for $\alpha_{\mathrm{NC}}(n)$ and $\alpha_{\mathrm{EC}}(n)$. In this proposed method, we should firstly determine $\alpha_{\mathrm{NC}}(n)$ or $\alpha_{\mathrm{EC}}(n)$ before adopting (15).

B. Step size control of SENLMS algorithm

We utilize the NP-VSS-NLMS algorithm [14] to determine one of the variable step size parameters in the VSS-SENLMS algorithm. In this work, we apply the NP-VSS-NLMS algorithm to the update equation of the AEC (13). The update equation is modified as [1]

$$\alpha_{\mathrm{EC}}(n) = \frac{\alpha_{\mathrm{VSS,EC}}(n)}{\|\mathbf{u}(n)\|^2} = \frac{1}{\|\mathbf{u}(n)\|^2}\left[1 - \sqrt{\frac{\sigma_{v,\mathrm{EC}}^2(n)}{\sigma_e^2(n)}}\right], \tag{16}$$

where $\sigma_{v,\mathrm{EC}}^2(n) = \mathbb{E}\left[\{v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n)\}^2\right]$. From (15) and (16), the variable step size parameter of the ANC $\alpha_{\mathrm{NC}}(n)$ is automatically determined as

$$\alpha_{\mathrm{NC}}(n) = \frac{1}{\|\mathbf{x}_{\mathrm{R}}(n)\|^2}\left[1 - \sqrt{\frac{\sigma_s^2(n)}{\sigma_e^2(n)}} - \alpha_{\mathrm{VSS,EC}}(n)\right]. \tag{17}$$

[1] Considering only the AEC, $v_{\mathrm{P}}(n)$, $y_{\mathrm{NC}}(n)$, and $s_{\mathrm{P}}(n)$ are disturbances of the AEC.

As can be seen from (16) and (17), we should estimate the variances $\sigma_e^2(n)$, $\sigma_s^2(n)$, and $\sigma_{v,\mathrm{EC}}^2(n)$. $\sigma_e^2(n)$ is estimated in [14] as

$$\sigma_e^2(n) = \lambda\sigma_e^2(n-1) + (1-\lambda)\, e^2(n), \tag{18}$$

where λ is a small constant for averaging. Since $s_{\mathrm{P}}(n)$ and $v_{\mathrm{EC}}(n) = v_{\mathrm{P}}(n) - y_{\mathrm{NC}}(n) + s_{\mathrm{P}}(n)$ cannot be obtained directly, $\sigma_s^2(n)$ and $\sigma_{v,\mathrm{EC}}^2(n)$ are estimated in [15] as

$$\sigma_s^2(n) = \sigma_e^2(n) - \frac{\|\mathbf{r}_{eu}(n)\|^2}{\sigma_u^2(n)} - \frac{\|\mathbf{r}_{ex}(n)\|^2}{\sigma_x^2(n)}, \tag{19}$$

$$\sigma_{v,\mathrm{EC}}^2(n) = \sigma_e^2(n) - \frac{\|\mathbf{r}_{eu}(n)\|^2}{\sigma_u^2(n)}, \tag{20}$$

$$\sigma_x^2(n) = \lambda\sigma_x^2(n-1) + (1-\lambda)\, x_{\mathrm{R}}^2(n), \tag{21}$$

$$\sigma_u^2(n) = \lambda\sigma_u^2(n-1) + (1-\lambda)\, u^2(n), \tag{22}$$

$$\mathbf{r}_{eu}(n) = \lambda\mathbf{r}_{eu}(n-1) + (1-\lambda)\, e(n)\mathbf{u}(n), \tag{23}$$

$$\mathbf{r}_{ex}(n) = \lambda\mathbf{r}_{ex}(n-1) + (1-\lambda)\, e(n)\mathbf{x}_{\mathrm{R}}(n). \tag{24}$$

IV. Simulation results

We conducted computer simulations to evaluate the capability of the proposed AENC to utilize the variable step size parameters. The proposed AENC was compared with the conventional AENC [9]. In the simulation, the room impulse responses (RIRs) generated by the RIR simulator [16] were used. Here, the near-end speech was neglected because the objective of the simulation was to evaluate the capability of the proposed AENC to reduce acoustic echo and noise. This is equivalent to neglecting the double-talk condition. The RIRs were generated under the arrangement shown in Fig. 2. The generated RIRs are shown in Fig. 3. The ratio of the direct component to reverberant components, that is, the direct-to-reverberant ratio (DRR) [17], was about 26 dB in Fig. 3 (a).

The simulation conditions are shown in Table I. In the simulations, for the conventional AENC, the step size parameters for the ANC and AEC $(\alpha_{\mathrm{NC}}, \alpha_{\mathrm{EC}})$ were set to $(0.07, 0.03)$ and $(0.09, 0.09)$, which are hereafter referred to as "pattern 1" and "pattern 2," respectively. The averaging factor λ for the proposed AENC was set to $1 - 1/50N$, where N is the tap length of the adaptive digital filter. The capability to reduce acoustic echo and noise was respectively evaluated by echo return loss enhancement (ERLE) and reduction defined as

$$\mathrm{ERLE}(n) = 10\log_{10}\left[\frac{\sum_{i=0}^{N-1} d_{\mathrm{E}}^2(n-i)}{\sum_{i=0}^{N-1}\left\{d_{\mathrm{E}}(n-i) - y_{\mathrm{EC}}(n-i)\right\}^2}\right], \tag{25}$$

$$\mathrm{Re}(n) = 10\log_{10}\left[\frac{\sum_{i=0}^{N-1} v_{\mathrm{P}}^2(n-i)}{\sum_{i=0}^{N-1}\left\{v_{\mathrm{P}}(n-i) - y_{\mathrm{NC}}(n-i)\right\}^2}\right]. \tag{26}$$

In the simulation, (25) and (26) were calculated every $N = 1024$ samples. The time waveforms of the acoustic echo and background noise are shown in Fig. 4.

The time waveforms of the observed and error signals are shown in Fig. 5. As seen in Fig. 5, the proposed AENC prevents the divergence and maintains almost the same capability to reduce the background noise and acoustic echo. On the other hand, the conventional AENC with pattern 2 diverges after 25 s. This is because the step size parameters are too large to prevent the divergence.

TABLE I: Simulation conditions.

Far-end speech	Female speech (F1, Aesop)[18]
Background noise	Office noise [19]
Signal-to-noise ratio (SNR)	-5 dB
Sampling frequency	16 kHz
Tap length of each path	1024
Tap length of ADF of ANC	1024
Tap length of ADF of AEC	1024
Regularization parameter	1.0×10^{-6}

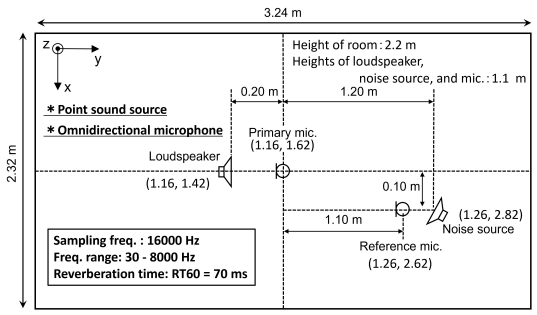
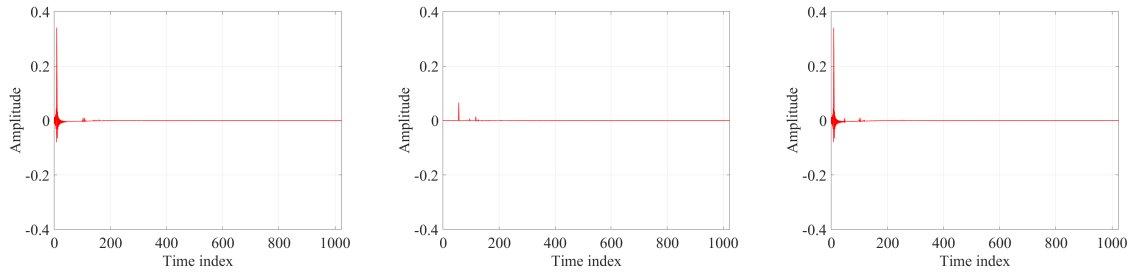


Fig. 2: Arrangement for RIR generation.

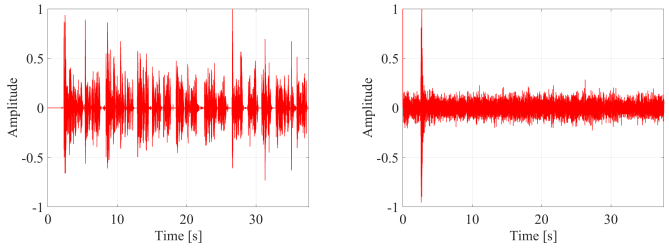
The amounts of noise reduction and ERLE are shown in Figs. 6 and 7. From Fig. 6, the proposed AENC is seen to reduce the background noise by about 25 dB when the conventional AENC with pattern 1 reduces the noise by about 28 dB. Moreover, from Fig. 7, the proposed AENC reduces the acoustic echo by about 13 dB when the conventional AENC with pattern 1 reduces the noise by about 20 dB. This is because the proposed AENC uses the estimates of each variance obtained using (18)–(24). These results can be explained by the dynamics of the variable step size parameters as shown in Fig. 8.

In Fig. 8(a), the variable step size parameter of ANC becomes a small value of about 5×10^{-3}, which leads to the prevention of the divergence of the AENC. Moreover, from Fig. 8(b), the variable step size parameter of AEC is a value of around 1×10^{-1}, which enables adaptation to the rapid fluctuations of acoustic echo. Thus, two variable step size parameters are adjusted to adequate values to a certain extent. On the other hand, the variable step sizes are smaller than those of the conventional AENC with pattern 1. This leads to small amounts of reduction of the background noise and acoustic echo. This is because the variances of each signal are estimated using (18)–(24) when the variable step size parameters are calculated. In this simulation, the office noise and speech signal are used, which are nonstationary signals. Hence, the variances obtained by the averaging process are not adequate for the variable step size parameters. Also, after 3 s, the variable step size for the ANC (17) is less than 1.0×10^{-2} while that for the AEC (16) is approximately 1.0×10^{-1}. This is because $\alpha_{\mathrm{VSS,NC}}(n) = 1 - \alpha_{\mathrm{VSS,EC}}(n)$ in a single-talk condition, and $\alpha_{\mathrm{VSS,NC}}(n) < \alpha_{\mathrm{VSS,EC}}(n)$ when



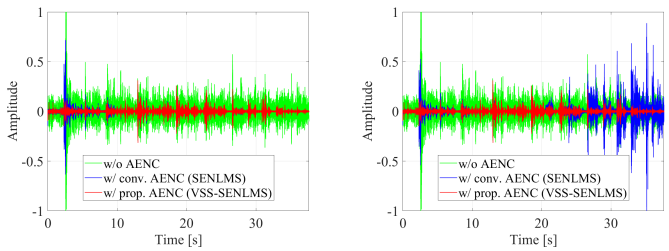
(a) Echo path (b) Noise path (primary microphone) (c) Noise path (reference microphone)

Fig. 3: Impulse responses of each path.



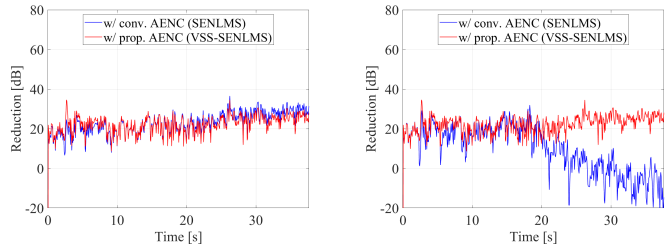
(a) Acoustic echo (b) Background noise

Fig. 4: Time waveforms of acoustic echo and background noise.



(a) Pattern 1 (b) Pattern 2

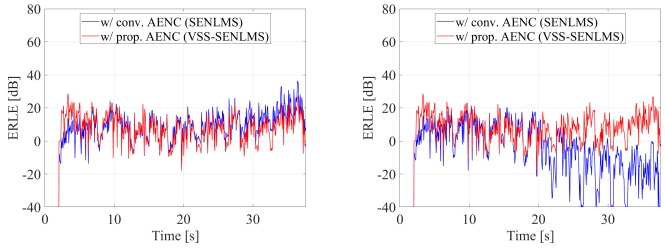
Fig. 5: Time waveforms of observed signals and outputs of AENCs .



(a) Pattern 1 (b) Pattern 2

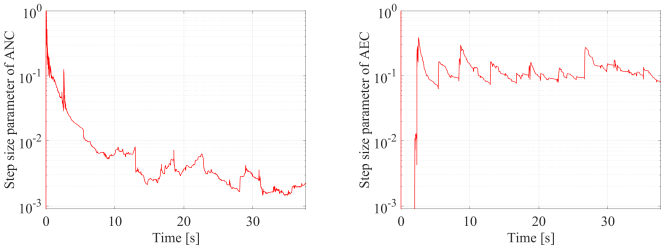
Fig. 6: Amount of noise reduction.

$\alpha_{\mathrm{VSS,NC}}(n)$ is small. However, the proposed AENC achieves almost the same amounts of reduction of the background noise and acoustic echo as the conventional AENC with pattern 1.



(a) Pattern 1 (b) Pattern 2

Fig. 7: Amount of ERLE.



(a) ANC (b) AEC

Fig. 8: Dynamics of variable step size parameters.

From the above results, it can be considered that the proposed AENC is effective in reducing both the acoustic echo and the background noise and preventing the divergence.

V. Conclusion

In this paper, we investigated the constraint for the step size parameters of the AENC with the SENLMS algorithm and proposed a step size control method. Simulation results showed that the proposed AENC with the VSS-SENLMS algorithm can simultaneously prevent the divergence and maintain its capability to reduce background noise and acoustic echo. In the future, we will investigate the step size control scheme for the double-talk condition and improve the accuracy of the estimation method for the variances for each signal.

Acknowledgements

This work was supported by JSPS KAKENHI (JP21K17773).

REFERENCES

[1] K. Sridhar, R. Cutler, A. Saabas, *et al.*, "ICASSP 2021 Acoustic Echo Cancellation Challenge: Datasets, Testing Framework, and Results," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 151–155.

[2] R. Cutler, A. Saabas, T. Parnamaa, *et al.*, "ICASSP 2022 Acoustic Echo Cancellation Challenge," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 9107–9111.

[3] S. Mundasad and P. Roxby, *Covid global health emergency is over, who says*, British Broadcasting Corporation, May 2023. [Online]. Available: https://www.bbc.com/news/health-65499929.

[4] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Process.*, vol. 64, no. 1, pp. 21–32, 1998.

[5] S. J. Park, C. G. Cho, C. Lee, and D. H. Youn, "On integrating acoustic echo and noise cancellation systems for hands-free telephony," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2, 2001, pp. 961–964 vol.2.

[6] M. Fukui, S. Shimauchi, Y. Hioka, A. Nakagawa, and Y. Haneda, "Acoustic echo and noise canceller for personal hands-free video IP phone," *IEEE Trans. Consum. Electron.*, vol. 62, no. 4, pp. 454–462, 2016.

[7] S. Ruiz, T. van Waterschoot, and M. Moonen, "Distributed combined acoustic echo cancellation and noise reduction in wireless acoustic sensor and actuator networks," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 30, pp. 534–547, 2022.

[8] K. Nathwani, "Joint acoustic echo and noise cancellation using spectral domain Kalman filtering in double-talk scenario," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC)*, 2018, pp. 326–330.

[9] K. Iwai and T. Nishiura, "Acoustic echo and noise canceller using shared-error normalized least mean square algorithm," in *Proc. 2022 Asia-Pacific Signal Inf. Process. Association Annual Summit Conf. (APSIPA ASC)*, 2022, pp. 281–285.

[10] B. Widrow, J. Glover, J. McCool, *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[11] S. O. Haykin, *Adaptive Filter Theory*, Fourth. Upper Saddle River, NJ: Prentice Hall, 2001.

[12] A. Zerguine, "Convergence behavior of the normalized least mean fourth algorithm," in *Conf. Record Thirty-Fourth Asilomar Conf. Signals Syst. Comput.*, vol. 1, 2000, pp. 275–278 vol.1.

[13] E. Eweda and A. Zerguine, "A normalized least mean fourth algorithm with improved stability," in *Proc. Conf. Record Forty Fourth Asilomar Conf. Signals Syst. Comput.*, 2010, pp. 1002–1005.

[14] J. Benesty, H. Rey, L. Vega, and S. Tressens, "A nonparametric VSS NLMS algorithm," *IEEE Signal Process. Lett.*, vol. 13, no. 10, pp. 581–584, 2006.

[15] M. A. Iqbal and S. L. Grant, "Novel variable step size NLMS algorithms for echo cancellation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2008, pp. 241–244.

[16] E. Habets, *Room impulse response generator*, the International Audio Laboratories Erlangen, 2010.

[17] J. Eaton, A. H. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-Reverberant Ratio estimation using a null-steered beamformer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 46–50.

[18] *University of tsukuba multilingual speech corpus (utml)*, National Institute of Informatics, 2006. DOI: https://doi.org/10.32130/src.UT-ML.

[19] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. 21st International Congress on Acoustics (ICA 2013)*, 2013.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhancing Spectrogram for Audio Classification Using Time-Frequency Enhancer

Haoran Xing* Shiqi Zhang* Daiki Takeuchi† Daisuke Niizumi† Noboru Harada† Shoji Makino*
* Waseda University, Japan
E-mail: haoranxing@suou.waseda.jp, tiosisai@gmail.com, s.makino@waseda.jp
† NTT Corporation, Japan
E-mail: noboru.harada.pv@hco.ntt.co.jp, daisuke.niizumi@ntt.com, daiki.takeuchi.ux@hco.ntt.co.jp

***Abstract*—It is challenging to deploy Transformer-based audio classification models on common terminal devices in real situations due to their high computational costs, increasing the importance of transferring knowledge from the larger Transformer-based model to the smaller convolutional neural networks (CNN)-based model via knowledge distillation (KD). Since an audio spectrogram can be regarded as an image, image-based models with CNN-based structures are used as the aforementioned smaller model for KD in several studies. However, the physical meanings of spectrograms differ from that of images in general. This fact possibly leads to the issue that the image-based model may not effectively extract features from a pure spectrogram. Thus, improving the spectrogram can help these models perform better on audio classification tasks. To implement our hypothesis, we propose a new Time-Frequency Enhancer (TFE), which is designed to learn how to enhance input spectrograms to make them effective for audio classification. In addition, we also propose TFE-ENV2, which extends EfficientNetV2 (ENV2), an image-based backbone model. To verify the effectiveness of the proposed method, we compare the performance between the original ENV2 and the proposed TFE-ENV2. In our experiments, the proposed TFE-ENV2 outperformed the original ENV2 on the ESC-50 and Speech Commands V2 datasets, demonstrating that the proposed TFE enhances spectrograms to assist image-based models in audio classification.**

I. Introduction

Audio classification task is aimed at classifying various audios and understanding the acoustic environment [1]. It has been used to assist speech enhancement [2], recognizing speech commands [3], and automatic speech recognition (ASR) [4], etc.

Convolutional Neural Network (CNN)-based models [5] have built-in inductive biases, which enable them to learn from limited audio data with small sizes [6, 7]. Recently, researchers have integrated image-based models with CNN architectures into audio classification tasks, leveraging their superior performance in the image domain to classify the spectrogram, which could be considered as the generalized image. At the same time, although some Transformer-based models [8–10] have shown better performance than CNN recently by incorporating techniques such as attention mechanisms [11], they typically have larger sizes and require higher computational costs, making them impractical for low-resource devices. For more usability on terminal devices in practical circumstances, previous studies have done research to transfer knowledge from larger Transformer-based models to smaller CNN-based models using knowledge distillation (KD) [12].

To enable both parameter/computationally efficiency and a high-performance model for audio classification tasks, this study focuses on training image-based models with CNN architectures via transferring knowledge from a large, high-performance Transformer-based model using KD. Both the Transformer-based model and the image-based model take the spectrogram as input.

Instead of feeding the original spectrogram to image-based models, we think that improving the spectrograms can help these models perform better on audio classification tasks since spectrograms and common images have different physical meanings. Values in spectrograms describe the energy distribution of the audio signal in the time-frequency domain, while images do not have this physical meaning.

To address the issue, we propose a new Time-Frequency Enhancer (TFE) to enhance the features of spectrograms in the time-frequency domain. The TFE consists of several learnable Enhancer Blocks, which are responsible for extracting features and generating a mask used for enhancing the spectrogram. The mask assigns importance weights to the features in the spectrogram, resulting in enhanced spectrograms. The TFE module is trained jointly with the image-based model using KD. While the TFE acts as a front end to provide enhanced spectrograms, the image-based model serves as the backbone. This joint optimization enables them to work together and achieve improved performance in audio classification. The TFE learns to enhance the spectrogram, thereby enhancing the audio classification performance.

Cross-Model Knowledge Distillation (CMKD) method [13] has been proven to be effective for transferring knowledge on audio classification. Besides, the Audio Spectrogram Transformer (AST) model has outstanding performance on audio classification [8] and EfficientNet (EN) models [14, 15] have been proven to perform well on audio classification in previous studies [9, 13]. Among all EN models, EfficientNetV2 (ENV2) [16] has been proven to outperform previous work on the ImageNet [17] dataset. The employment of the Fused-Mobile Convolution (Fused-MBConv) blocks [18] in the early stages of ENV2 can reduce training time and provide better efficiency for image classification tasks. Thus, to further improve the performance of audio classification, we transfer knowledge

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

from an AST model with good performance to the audio classification model with the backbone of ENV2 through CMKD.

The remainder of the paper is organized as follows: Section II provides an overview of the related work in this paper. In Section III, we introduce the proposed method and discuss its architecture. Experimental settings and results are presented in Section IV. Finally, in Section V, we conclude the paper.

II. Related Work

A. Audio Spectrogram Transformer

Audio Spectrogram Transformer (AST) [8] is a Transformer-based model for audio classification tasks that outperformed other CNN-based models at the time. Unlike CNN-based models for audio classification, AST works purely based on the Self-Attention mechanism [19] to capture long-range global information from spectrograms. However, AST is large and needs large-scale datasets for training, which leads to high training costs and high computational complexity. Thus, it is difficult to deploy AST on typical terminal devices in practical settings.

B. EfficientNet Model for Audio Classification

EfficientNet (EN) is a CNN-based model originally proposed for image classification tasks. EN works based on the structure of multiple Mobile Convolution (MBConv) blocks [20]. MBConv is a residual block with Squeeze-and-Excitation (SE) module [21], which is designed to capture channel-wise dependencies and generate importance weights for different channels. Several studies have used AudioSet [22] to pretrain EN to further improve audio classification performance. However, direct pretraining EN on AudioSet is time-consuming and leads to high training costs. Transferring knowledge from pretrained models to EN is important for this reason.

C. Cross-Model Knowledge Distillation

Cross-Model Knowledge Distillation (CMKD) [13] method is aimed at transferring knowledge between audio classification models with different architectures using KD, which is a common method allowing small student models to learn from large teacher models [23]. In the original paper of CMKD, KD is employed to transfer knowledge from a pretrained AST to an EN. It has been demonstrated that EN performs better after learning from pretrained AST. Therefore, CMKD is a practical method to obtain an audio classification model with good performance and a smaller size.

D. Learnable Frontend for Audio Classification

The related work Learnable Frontend (LEAF) [24] for audio classification is designed as a replacement for mel-filterbank. It learns audio feature extraction and converts the audio signal from the time domain into the time-frequency domain. However, due to KD being used for training, the input for the image-based CNN model is expected to be the same as that for the AST model, which is the Log Mel Spectrogram, a typical representation of audio signals in the time-frequency domain. Thus, enhancing the spectrogram in the time-frequency domain is required for image-based models rather than feature extraction from the time domain to the time-frequency domain.

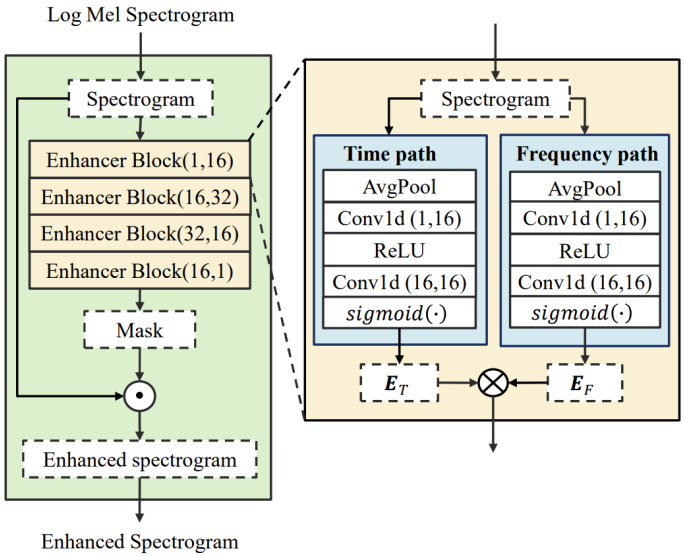


Fig. 1. Illustration for the architecture of the proposed TFE module

III. Proposed Method

We propose a new Time-Frequency Enhancer (TFE), which is intended to enhance the original spectrogram and provide the enhanced spectrogram as input for image-based models. In addition, we propose TFE-ENV2, in which we apply TFE as a front end to ENV2 for audio classification.

A. Time-Frequency Enhancer (TFE)

We propose the TFE module, inspired by Time-Frequency Attention (TFA) [25], which was designed for the speech enhancement task. TFA is a lightweight attention mechanism that provides the mask for the spectrogram in the time-frequency domain. It can be regarded as an Enhancer Block for the spectrogram, which employs two independent feature extractors with the same structures for the time frame path and the frequency bin path in the spectrogram, respectively. Both extractors consist of the average pooling layer (AvgPool), 1D convolution layer (Conv1d), ReLU activation function (ReLU), and Sigmoid activation function ($sigmoid(\cdot)$). In each path, at first, the 2D spectrogram is compressed into a 1D feature map through the average pooling operation. Then the 1D convolution layer is employed to extract features from the 1D feature map. Each extractor outputs an enhanced 1D feature vector. Two 1D vectors from the time path and frequency path would be used to generate a 2D mask for the input spectrogram by matrix multiplication, which can be written as

$$\mathbf{M} = \mathbf{E_T} \otimes \mathbf{E_F}, \tag{1}$$

where $\mathbf{M}$ represents the 2D mask, $\mathbf{E_T}$ represents the enhanced feature vector from the time path, $\mathbf{E_F}$ represents the enhanced

feature vector from the frequency path, and $\otimes$ represents matrix multiplication. The architecture of the Enhancer Block is shown in the yellow block (right) of Fig. 1.

The structure of the proposed TFE module, made up of multiple Enhancer Blocks, is shown in the green block (left) of Fig. 1. In the remaining paper, we use TFE(**n**) to represent the TFE module, which consists of **n** Enhancer Blocks. The TFE module in Fig. 1 includes 4 Enhancer Blocks, denoted as TFE(4). Compared with TFA, which uses only one Enhancer Block for spectrogram enhancement, the proposed TFE(**n**) module works based on **n** Enhancer Blocks to enhance the spectrogram in multiple channels. By deepening the network structure, the proposed TFE(**n**) should learn rich features that enable the model to capture more helpful information from the spectrogram. The TFE(**n**) module takes the original Log Mel Spectrogram as the input and generates the mask through **n** Enhancer Blocks. After that, the mask and input spectrogram are combined into the enhanced spectrogram using the element-wise product, which is expressed as

$$\widetilde{\mathbf{S}} = \mathbf{S} \odot \mathbf{M}, \tag{2}$$

where $\mathbf{S}$ represents the original spectrogram, $\mathbf{M}$ represents obtained mask, $\widetilde{\mathbf{S}}$ represents the enhanced spectrogram, and $\odot$ represents the element-wise product.

The proposed TFE module produces an attention mask to enhance the original spectrogram by multiplying two feature vectors, which are output from the time path and frequency path in the spectrogram, respectively. These two vectors emphasize the important points in these two paths based on the energy distribution of the spectrogram, respectively. As a result, the attention mask is intended to emphasize the important points in the whole spectrogram and provide significance weights for the spectrogram. Thus, as the combination of the attention mask and the original spectrogram, the enhanced spectrogram should provide more helpful features for audio classification.

B. EfficientNetV2 with TFE module (TFE-ENV2)

In this work, we propose a combined structure for audio classification that integrates EfficientNetV2 with the TFE module. This structure, named TFE-ENV2, incorporates the TFE module for spectrogram enhancement and the ENV2 module for classification.

Fig. 2(a) illustrates the architecture of the original ENV2 while Fig. 2(b) illustrates the architecture of the proposed TFE-ENV2. The original ENV2 consists of a 2D convolutional layer (with a kernel size of 3×3), Fused-MBConv blocks, MBConv blocks, and a classifier, which is the same as that employed for image domain [16].

The proposed TFE-ENV2 structure utilizes a Log Mel Spectrogram as input. The TFE module serves as the front end, while the ENV2 module functions as the backbone for audio classification. This combination is chosen based on specific advantages and expected improvements. By incorporating the TFE module, we aim to enhance the feature extraction process, leading to improved classification performance compared to the original ENV2.

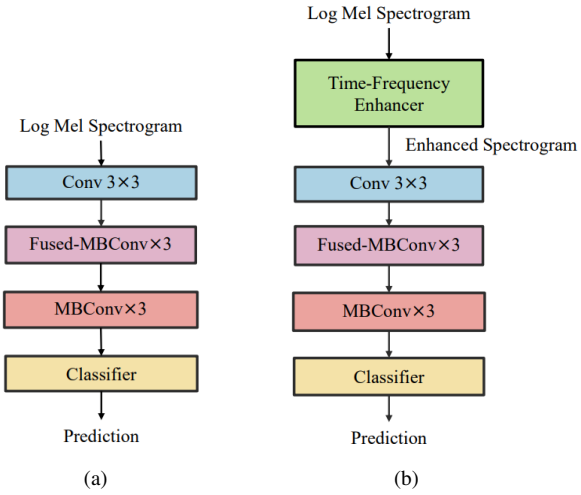


Fig. 2. Illustration for architectures of (a) the original ENV2 and (b) the proposed TFE-ENV2.

IV. Experiment

We validated the proposed method on ESC-50 [26] dataset and Speech Commands V2 [27] dataset.

To enhance audio classification performance, we employed the CMKD method to transfer knowledge from the AST model to the TFE-ENV2. And to ensure consistency in our approach, we reused the experimental settings of CMKD as described in the previous work [13] as shown in Table I.

A. Datasets

The experiments were conducted on two labeled datasets: ESC-50 and Speech Commands V2. The ESC-50 dataset is made up of 2,000 5-second environmental sounds organized into 50 classes. We followed the 5-fold cross-validation and recorded the mean accuracy. The speech Commands V2 dataset contains 105,829 different 1-second recordings divided into 35 classes. It is divided into a training set, a validation set, and an evaluation set, made up of 84,843, 9,981, and 11,005 samples, respectively. We followed the 35-class classification task and recorded the accuracy of the evaluation set.

B. Experimental Setup

To confirm the effectiveness of the proposed TFE, we conducted ablation experiments to compare the original ENV2 with the proposed TFE(**n**)-ENV2. We assessed three TFE models with varying numbers of Enhancer Blocks. **n** was adjusted to 1, 2, and 4 in our experiments. The objective was to evaluate the impact of different TFE configurations. TFE(1)-ENV2 is made up of Enhancer Block(1,1), TFE(2)-ENV2 is made up of Enhancer Block(1,32), and Enhancer Block(32,1). For a fair comparison, all student models were trained with the same training settings and learned from the same AST model. All experiments were repeated 5 times and the mean accuracy was reported.

Model settings: On the ESC-50 dataset, AST, the teacher model, was initialized by the weights pretrained on two

datasets (ImageNet and AudioSet) and then fine-tuned for 5 epochs. While on the Speech Commands V2 dataset, it was only pretrained on ImageNet and also fine-tuned for 5 epochs. Then, the fine-tuned models were used for knowledge distillation. The ENV2 was initialized with weights pretrained on ImageNet, serving as the backbone for audio classification. In contrast, the proposed TFE(**n**) module did not undergo any pretraining.

Data processing and data augmentation: For the ESC-50 dataset, to keep consistency, we followed the same data processing and data augmentation recipes as described in the CMKD method in previous work [13]. The teacher model was fine-tuned with SpecAugment [28] (frequency and time masking). After that, student models were trained using SpecAugment, random time-shift (rolling spectrograms in the time axis), random noise (generating distribution to spectrograms), and label smoothing (smoothing one hot label) [29]. For the Speech Commands V2 dataset, we followed the same data augmentation recipes as those used for the original AST in previous work [8]. The teacher model was trained with SpecAugment and mix-up (randomly mixing up two spectrograms and labels) [30]. After that, student models were trained using the same method. Input for the teacher model and student models were spectrograms with the same data augmentations during training, while there was no data augmentation during testing.

Training settings: All the details for implementation in experiments are shown in Table I. For ESC-50, to keep consistency, student models were trained with the same training settings used for the AST model in previous work [8], except for the initial learning rate, learning scheduler, and epochs. For ESC-50, the learning rate decreased by a factor of 0.85 every 5 epochs. For Speech Commands V2, the learning rate decreased by a factor of 0.85 for every epoch after the first 5 epochs. The Adam optimizer was used for both datasets.

TABLE I
TRAINING SETTINGS ON ESC-50 AND SPEECH COMMANDS V2

Settings	ESC-50	Speech Commands V2
Spectrogram size (Time frames/Frequency bins)	512/128	128/128
Initial learning rate	1e-4	5e-4
Batch size	48	128
Epochs	100	30
Loss for ground truth	Cross Entropy	Brainy Cross Entropy
SpecAugment (Time/Frequency)	(96/24)	(48/48)
Random noise on spectrogram	U(0,0.05)	U(0,0.1)
Random time-shift	±10 time frames	±10 time frames
Mix-up ratio	0	0.6
Label smoothing factor	0.1	0

Knowledge distillation: Distillation loss used in training can be written as

$$\mathcal{L} = \lambda \mathcal{L}_g(\delta(Z_S), y) + (1-\lambda)\mathcal{L}_d(\delta(Z_S), \delta(Z_T/\tau)), \quad (3)$$

where Z_S and Z_T present the logits from the student model and teacher model, respectively. $\mathcal{L}_g$ and $\mathcal{L}_d$ are losses for ground truth and distillation, respectively. y represents ground-truth labels. λ represents the balancing coefficient. δ is the activation function. τ is KD temperature. Cross-entropy (CE) loss and Kullback-Leibler divergence loss were used for $\mathcal{L}_g$ and $\mathcal{L}_d$, respectively. We fixed balancing coefficient $\lambda = 0.5$ and KD temperature $\tau = 5$. Soft-max was used as the activation function δ.

C. Results

TABLE II
RESULTS FOR ABLATION EXPERIMENTS ON THREE TFE-ENV2 MODELS

Model Name	Parameters	Accuracy	
		ESC-50	Speech Commands V2
TFE(1)-ENV2	8.75M	89.42% (±0.43)	97.78% (±0.04)
TFE(2)-ENV2	8.77M	89.23% (±0.27)	97.79% (±0.09)
TFE(4)-ENV2	8.82M	**89.67%** (±0.46)	**97.85%** (±0.04)

Table II shows the results of ablation experiments for TFE(**n**)-ENV2 models. According to results in Table II, for the Speech Commands V2 dataset, the accuracy increased as the number of Enhancer Blocks in TFE(**n**)-ENV2 increased. Compared with TFE(1)-ENV2, the accuracy of TFE(**n**)-ENV2 increased by 0.01% and 0.07% when **n** is set to 2 and 4, respectively. While for ESC-50, TFE(1)-ENV2 performed better than TFE(2)-ENV2. One possible reason might be that there was no pretraining for the TFE(**n**) module, so the increasing number of Enhancer Blocks caused the increase in unpretrained parameters, which need more data for training. As a result, TFE(**n**)-ENV2 models might overfit on the ESC-50 dataset, which has a smaller data size than Speech Commands V2. Thus, we believe that the trend of the accuracy of TFE(**n**)-ENV2 on the Speech Commands V2 is more deserving of reference.

TABLE III
COMPARISONS FOR THE PERFORMANCE BETWEEN AST, ENV2, AND THE BEST TFE-ENV2

Model Name	Parameters	Accuracy	
		ESC-50	Speech Commands V2
AST	85.70M	95.85%	97.46%
ENV2	8.74M	88.63% (±0.34)	97.74% (±0.07)
TFE(4)-ENV2	8.82M	**89.67%** (±0.46)	**97.85%** (±0.04)

Table III shows the comparison of the performance of AST, ENV2, and the best TFE-ENV2. According to the results, the accuracy of TFE(4)-ENV2, which is the best of all, increased by 1.04% and 0.11% compared with the original ENV2 on ESC-50 dataset and Speech Commands V2 dataset, respectively. Besides, all the proposed TFE(**n**)-ENV2 models outperformed the original ENV2 model. What's more, on Speech Commands V2, the accuracy of TFE(4)-ENV2 got almost 0.4% higher than AST with only 10.29% of its parameters. All these results demonstrated the effectiveness of the proposed TFE.

TABLE IV
PERFORMANCE OF ENV2 AND TFE(4)-ENV2 WHEN CMKD WAS SKIPPED FOR TRAINING

Model Name	Parameters	Accuracy	
		ESC-50	Speech Commands V2
ENV2	8.74M	86.74% (±0.29)	97.10% (±0.06)
TFE(4)-ENV2	8.82M	**87.13%** (±0.43)	**97.19%** (±0.09)

We also explored the impact of knowledge distillation on training models. In the experiment, by fixing the balancing coefficient $\lambda = 1$, the CMKD method was not performed to train the original ENV2 and the proposed TFE(4)-ENV2. Table IV shows the performance of ENV2 and TFE(4)-ENV2 when the CMKD was skipped. According to the results, the proposed TFE(4)-ENV2 still had better performance than the original ENV2 on both two datasets. However, compared to training with knowledge distillation, the accuracy of both ENV2 and TFE(4)-ENV2 dropped. For datasets ESC-50 and Speech Commands V2, the accuracy of the original ENV2 dropped from 88.63% to 86.74% and from 97.74% to 97.10%, respectively; the accuracy of the proposed TFE(4)-ENV2 dropped from 89.67% to 87.13% and from 97.85% to 97.19%, respectively. These findings demonstrate the importance of knowledge distillation for training the image-based model in the audio classification, and they also confirm that the proposed TFE module can still help improve the performance of the original ENV2 for speech classification even when knowledge distillation is not performed for training.

V. CONCLUSIONS

Unlike the general field of image processing, where the vertical and horizontal axes typically have similar physical meanings, the audio spectrogram exhibits distinct interpretations for these axes. This unique characteristic poses a challenge when utilizing spectrograms for image-based classifier models. To address this issue and obtain a more suitable spectrogram for the image-based classifier model, we proposed a new TFE module for spectrogram enhancement. The proposed model, TFE, consists of multiple Enhancer Blocks to extract time-frequency features in high dimensions, providing enhanced spectrograms to ENV2 for classification. In our experiments, compared with the original ENV2, the accuracy of the proposed TFE-ENV2 increased by 1.04% and 0.11% on ESC-50 and Speech Commands V2 respectively. All of these results demonstrated the effectiveness of the proposed TFE for ENV2 on audio classification tasks. Moreover, the evidence supports the idea that enhancing spectrograms through the extraction of high-dimensional time-frequency features based on energy distributions can be beneficial for audio classification. These findings not only contribute to the current understanding of audio classification but also provide a potential direction for future research in this area.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech & Audio Process*, vol. 10, no. 7, pp. 504–516, 2002.

[2] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[3] B. McMahan and D. Rao, "Listening to the world improves speech command recognition," in *arXiv preprint arXiv:1710.08377v1*, 2018.

[4] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, 2018.

[5] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," vol. 3361, no. 10, pp. 255–258, 1998.

[6] Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, 2017, pp. 131–135.

[7] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[8] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[9] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[10] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. ICASSP*, 2022, pp. 646–650.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] G. Aguilar, Y. Ling, Y. Zhang, B. Yao, X. Fan, and C. Guo, "Knowledge distillation from internal representations," in *Proc. AAAI*, 2020, pp. 7350–7357.

[13] Y. Gong, S. Khurana, A. Rouditchenko, and J. Glass, "CMKD: CNN/Transformer-based cross-model knowledge distillation for audio classification," *arXiv preprint arXiv:2203.06760*, 2022.

[14] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[16] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. ICML*, 2021, pp. 10 096–10 106.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image

database," in *Proc. CVPR*, 2009, pp. 248–255.
[18] S. Gupta and M. Tan, "EfficientNet-EdgeTPU: Creating accelerator-optimized neural networks with automl," *Google AI Blog*, vol. 2, no. 1, 2019.
[19] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. ICML*, 2019, pp. 7354–7363.
[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. CVRP*, 2018, pp. 4510–4520.
[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVRP*, 2018, pp. 7132–7141.
[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.
[23] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.
[24] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.
[25] Q. Zhang, Q. Song, Z. Ni, A. Nicolson, and H. Li, "Time-frequency attention for monaural speech enhancement," in *Proc. ICASSP*, 2022, pp. 7852–7856.
[26] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM*, 2015, pp. 1015–1018.
[27] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.
[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
[29] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *arXiv preprint arXiv:1906.02629v3*, 2019.
[30] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion

Wen-Chin Huang and Tomoki Toda
Nagoya University, Japan
E-mail: wen.chinhuang@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Foreign accent conversion (FAC) is a special application of voice conversion (VC) which aims to convert the accented speech of a non-native speaker to a native-sounding speech with the same speaker identity. FAC is difficult since the native speech from the desired non-native speaker to be used as the training target is impossible to collect. In this work, we evaluate three recently proposed methods for ground-truth-free FAC, where all of them aim to harness the power of sequence-to-sequence (seq2seq) and non-parallel VC models to properly convert the accent and control the speaker identity. Our experimental evaluation results show that no single method was significantly better than the others in all evaluation axes, which is in contrast to conclusions drawn in previous studies. We also explain the effectiveness of these methods with the training input and output of the seq2seq model and examine the design choice of the non-parallel VC model, and show that intelligibility measures such as word error rates do not correlate well with subjective accentedness. Finally, our implementation is open-sourced to promote reproducible research and help future researchers improve upon the compared systems.**

I. Introduction

Voice conversion (VC) is the task of converting between two types of speech without changing the linguistic content [1], [2]. While speaker conversion has been the main focus of previous VC research, in this paper we aim to tackle a special application: foreign accent conversion (FAC)[1] [3]. As illustrated in the right side of Figure 1, given an accented speech utterance spoken by a non-native source speaker, FAC aims to generate a native-sounding version with the same speaker identity as the source speaker. Applications of FAC include computer-aided language learning [3]–[5] and entertainment such as movie dubbing [6].

At first sight, one might try to apply existing state-of-the-art deep learning-based VC methods, such as sequence-to-sequence (seq2seq) modeling [7], [8] or non-parallel recognition-synthesis-based models [9], to FAC. However, such direct application is considered infeasible because of the absence of the ground-truth training target: it is impossible to collect native speech from a non-native speaker. In fact, there are many other VC applications that face the same problem, such as dysarthric speech conversion [10], [11], which aims to convert from disordered speech to healthy speech while preserving the speaker identity.

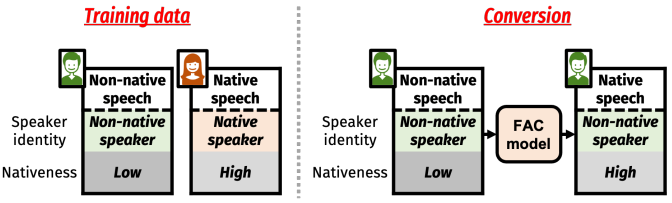


Fig. 1: Left: the training data, which is a parallel corpus between the source non-native speaker and a reference native speaker. Right: the goal in the conversion phase of FAC. The nativeness is expected to be increased while maintaining the speaker identity.

Researchers have proposed several methods for such a type of ground-truth-free[2] VC task. They all share a main idea, which is to first collect a training corpus from the source non-native speaker and then collect the native counterpart from a native reference speaker with the same prompt set, as depicted in the left side of Figure 1. Then, state-of-the-art VC methods for disentangling the speaker and content are designed to achieve FAC. However, these works were developed in parallel, and comparisons were often conducted on a system-to-system basis, leading to a lack of comparison and understanding of each method.

In this work, we aim to systematically evaluate three methods [10], [12], [13] for FAC. Experiments were conducted in a unified setting using a shared database, model architecture, and vocoder. We conducted a subjective evaluation test assessing three different aspects (naturalness, speaker similarity, accentedness) of the synthesized samples, allowing us to compare the methods. As we will show in our experimental evaluation section, we found that **no single method was significantly better than the others in all evaluation axes**, which is in contrast to conclusions drawn in previous studies [13]. We also present results of an objective intelligibility measure which was used in previous studies [12], and show that it might not correlate well to subjective accentedness. Finally, to promote reproducible FAC research, we open-source our implementation to help future researchers improve upon our system[3].

[1]Readers should note that the term "accent conversion" can be referred to many different tasks in the literature. While many have used this term to refer to the conversion between different accents or from native to accented speech, in this work we focus on the task of "de-accenting".

[2][12] used the term 'reference-free', but we found it confusing since a reference speaker is needed for training. Therefore we use "ground-truth-free" in the rest of the paper, where "ground-truth" refers to the ground-truth used as the training target.

[3]https://github.com/unilight/seq2seq-vc

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Related works and background knowledge

A. Non-native to native foreign accent conversion

In the FAC literature, most works tried to utilize accent-independent features to decompose accent from voice identity. For instance, early attempts made use of articulatory trajectories (e.g., lips and tongue movements) [14]–[16] and vocal tract length normalization [17]. More recently, more simplified features such as phonetic posteriorgrams (PPGs) [18], [19] and text [20] are combined with advanced deep neural network architectures, especially seq2seq VC models, whose ability to model segmental and prosody features simultaneously play a crucial role in FAC. However, only very few works have tried to address ground-truth-free FAC [12], [13]. We believe this difficult yet practical setting is worth further investigation.

B. Seq2seq VC modeling

Seq2seq modeling learns the alignment between the source and target data sequences in a data-driven manner, equipping the model with the ability to generate outputs of various lengths and capture long-term dependencies. This is considered particularly important in FAC, as duration and supra-segmental characteristics like f_0 play an important role in prosody conversion, which strongly affects the perception of foreign accents. In addition, in this work, we apply the text-to-speech (TTS) pre-training technique proposed in [7], [8] to increase the performance. However, a major drawback of seq2seq models is the requirement of parallel data, which is impractical to collect in FAC. Therefore, we resort to the help of another family of VC models.

C. Non-parallel frame-based VC modeling

Non-parallel VC relaxes the requirement of a parallel corpus, and one of the mainstream methods is autoencoder-style training. An encoder first extracts latent features to filter out certain attributes (e.g. speaker identity in speaker conversion), and then a decoder tries to generate the input speech by reconstructing the missing information. In contrast with seq2seq models, non-parallel models are frame-based, i.e. the duration and other supra-segmental attributes are not converted. It is therefore assumed that non-parallel frame-based models convert only global characteristics such as the speaker identity while maintaining local characteristics, such as pronunciation.

Although the extractor and the decoder can be jointly trained, as reported in [21], it is more effective to first train the extractor and then used the extracted features to train the decoder. The extractor is usually pre-trained using a large, multi-speaker dataset, and as a result an any-to-one (A2O) VC model is obtained as it generalizes well to unseen speakers. Among the many choices, the most widely used latent feature is the PPG [9] extracted from a supervisedly trained automatic speech recognition (ASR) model, as it preserves strong linguistic clues and serves as a strong speaker information bottleneck. Recently, self-supervised speech representations (S3Rs) are attractive in that they can be trained without labeled data and thus benefit from training on larger scales. Many have applied them to VC as an alternative [22], [23], and the best performing one is vq-wav2vec [24]. In this work, we evaluate the effectiveness of PPG and vq-wav2vec in the context of FAC.

III. Evaluated methods

In all three methods we evaluate, we assume two materials are prepared beforehand. First, as shown in Figure 1, access to a parallel dataset between the non-native speaker and a reference native speaker is assumed. Second, the dataset of the native speaker is used to train a non-parallel frame-based model, which will then be fixed in all three methods. This process is illustrated in the top of Figure 2. In the following subsections, the detailed procedures of all three methods are described. Figure 2 shows a complete illustration.

A. Method 1: cascade

The cascade method was originally proposed for dysarthric VC [10], but as mentioned in Section I, since dysarthric VC and FAC share a common problem of lacking ground-truth training target, here we examine whether it could be applied to FAC.

In the cascade method, a seq2seq model is trained to map from the source non-native speech to that of the reference native speaker. During conversion, the source speech is first sent into the seq2seq model to get the first stage converted speech. Although the nativeness is improved, the speaker identity is unwantedly changed into that of the reference speaker. Therefore, the non-parallel VC model is then used to change the identity back to that of the native speaker, while maintaining the pronunciation.

B. Method 2: synthetic target generation (STG)

In STG [12], the non-parallel VC model converts the training dataset of the native speaker such that the generated speech has the same nativeness of the input but with the speaker identity of the non-native speaker. We refer to this step as synthetic target generation. Then, the seq2seq model is trained using the non-native training set as the source and the synthetic native speech with the speaker identity of the same non-native speaker as the target. The conversion process is then as simple as using the seq2seq model to generate the de-accented speech with the identity of the non-native speaker.

C. Method 3: latent space conversion (LSC)

The LSC method [13] first uses the latent feature extractor module of the non-parallel VC model to transfer the training datasets of the source non-native and target native speakers from the speech space to the latent space. Then, the seq2seq model is trained to map the source latent features to the target latent features. During conversion, the latent features of the source non-native speech are first extracted and transformed to their native counterpart using the seq2seq model. Finally, the decoder of the non-parallel VC model is used to inject the identity of the non-native speaker into the converted latent features in order to generate the final converted speech.

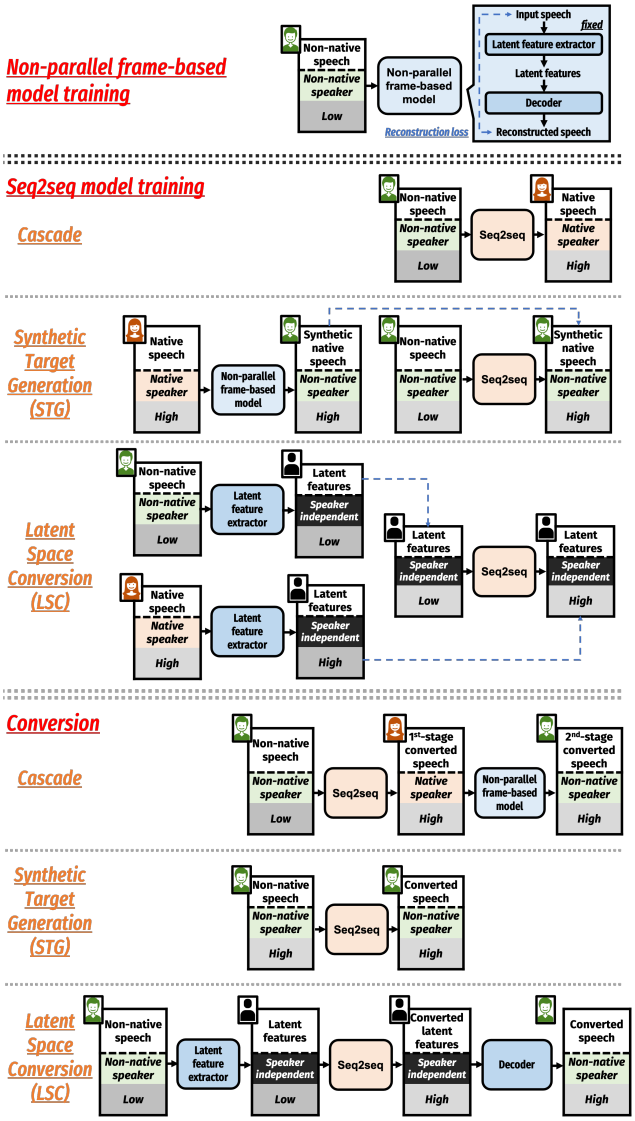


Fig. 2: Illustration of the training and conversion processes of the three evaluated methods (cascade, STG, LSC). Top: non-parallel frame-based VC model training using the data from the source non-native speaker. Middle: seq2seq model training. All the components from the non-parallel frame-based VC model (marked in blue), including the latent feature extractor and the decoder, are fixed in this phase. Bottom: the conversion phases.

D. What is the difference between these three methods?

As these methods might be seemingly complicated in their individual ways, the difference lies in the training input and output of the seq2seq model. This can be clearly observed in the right column of the middle part in Figure 2. First, in the cascade method, the seq2seq model needs to simultaneously convert the speaker identity and the nativeness, which is considered to be the most difficult. In contrast, STG first changes the speaker identity of the native training set, such that the seq2seq model only needs to model the pronunciation pattern. Finally, to further ease the job of the seq2seq model, LSC projects the training datasets onto the latent space, which is speaker-independent, and this is easier to model than the speech space. Based on this observation, we may hypothesize that the performance of each method should correlate with the difficulty of the mapping function that the seq2seq model needs to find. We will discuss this hypothesis in the results section.

These three methods also have their own weakness. For instance, during the conversion phase, not only cascade but also LSC does the input speech pass through a pipeline consisting of multiple modules, suffering from potential error propagation. STG, on the other hand, requires only the seq2seq module during conversion and thus does not suffer from error propagation, but the synthetic target data inevitably contains artifacts. The performance of the seq2seq model is then bounded by how imperfect the synthetic data is. With these unique limitations, we note to the readers that it is difficult to fairly compare these methods.

IV. Experimental Evaluation and Discussion

A. Experimental setting

The non-native and reference native speakers are THXC (Chinese male) and bdl (English male) from the L2-ARCTIC [25] and ARCTIC datasets, respectively. There are 1032/50/50 training/development/testing parallel utterances respectively, and the total duration of the training set is around one hour. All samples are in 16 kHz. The downsampled LJSpeech dataset [26] was used for the seq2seq model pretraining. The LibriSpeech (960 hours) [27] and LibriLight [28] (60k hours) datasets were used to train the PPG and vq-wav2vec models, respectively.

Mel-spectrograms were used as the acoustic features with a hop size of 256. ParallelWaveGAN [29] was used as the neural vocoder, trained with only the training set of the source non-native speaker. The seq2seq model is based on the Transformer [7], [30] and the decoder of the non-parallel frame-based VC model resembles Tacotron2 [31], following [22]. The implementation is open-sourced[4], so we refer readers to the source code for detailed hyperparameters due to space limits. An NVIDIA V100 GPU was used to train all the models, and the seq2seq model and non-parallel VC model in total took 6 hours to train.

B. Evaluation protocols

We conducted subjective evaluation tests on three axes, following previous works [12], [13]. In the naturalness and accentedness tests, listeners were asked to rate the samples on a 5-point and a 9-point scale, respectively. Samples of the source non-native and target native speech were also included. In the similarity test, a source non-native sample and a converted sample were presented, and listeners were asked to judge whether the two samples were uttered by the same speaker on a four-point scale, following the same convention in voice conversion challenges [21]. The naturalness and similarity tests were conducted using Amazon Mechanical Turk, and each of

[4]See footnote 3

TABLE I: Objective and subjective evaluation results with 95% confidence interval of samples from the evaluated methods, source and target.

Method	Extractor	CER/WER	Naturalness ↑ (1-5)	Similarity ↑ (0% -100%)	Accentedness ↓ (1-9)
Source (non-native)		5.3/12.3	4.18±0.19	–	6.06±0.38
Cascade	vq-wav2vec	29.1/52.5	3.17±0.23	28.7%±6.7%	5.41±0.32
	PPG	30.4/52.7	3.50±0.22	45.7%±7.3%	4.18±0.30
STG	vq-wav2vec	25.3/45.0	3.23±0.21	37.0%±7.0%	5.27±0.31
	PPG	17.7/40.9	3.66±0.20	57.3%±7.8%	4.36±0.32
LSC	vq-wav2vec	33.4/52.5	3.65±0.25	36.0%±7.0%	4.61±0.32
	PPG	9.8/19.5	3.64±0.22	43.8%±7.5%	3.95±0.31
Target (native)		1.3/4.3	4.42±0.18	–	1.49±0.21

the 65 workers we recruited rated 20 samples (or sample pairs). As for the accentedness, due to the difficulty of the task itself, we conducted an in-lab study by recruiting 19 listeners to each listen to 40 samples. Audio samples are available for readers[5].

Finally, we consider one objective measure for calculating speech intelligibility: character/word error rates (CER/WER) obtained by running an ASR model on the speech samples. We used a pretrained ASR model based on wav2vec 2.0[6] [32].

C. Design choice of the non-parallel frame-based model

In Secion II-C, we mentioned we experimented with two types of latent features in the non-parallel frame-based model: PPG and vq-wav2vec. As shown in Table I, in all three subjective evaluation axes (naturalness, similarity, and accentedness) and all three evaluated methods, using PPG was almost always significantly better than using vq-wav2vec. The only exception was that the naturalness scores were nearly identical when using vq-wav2vec and PPG in the LSC scenario.

Although previous works [22], [23] have already shown that PPG outperforms vq-wav2vec in terms of naturalness and similarity, the superiority of PPG in accentedness implies the importance of linguistic supervision in the training of the latent extractor. In the rest of the section, we focus on the results of the three methods using PPG.

D. Effectiveness of the three evaluated methods

In this subsection, we try to compare the performance tendency with the hypothesis described in Section III-D. We first look at the naturalness scores of the three methods in Table I. As the confidence intervals overlap, there is no statistically significant difference between the three methods. This suggests that naturalness is not affected by the difficulty of the seq2seq mapping.

Next, for similarity, STG is significantly better than cascade and LSC, which again violates the above-mentioned hypothesis. Nonetheless, this result leads to two implications. First, as the seq2seq model in STG operates in the speech domain, the model prioritizes the learning of speaker identity generation over accent removal. Second, the relatively low similarity of LSC implies that the assumption of speaker independence of the latent features may be invalid in the context of FAC.

Finally, in terms of accentedness, the only significant difference that can be observed is the superiority of LSC over STG. Although this does not match the hypothesis, we note that the accentedness score of LSC is significantly better than cascade and STG when using vq-wav2vec. This suggests that LSC is more robust to the choice of the latent feature.

E. Is character/word error rate a proper objective measure for FAC?

The feasibility of using objective measures to predict subjective results is a long-standing problem in VC research [33], [34]. Developing such a measure allows us to inspect the performance during system development without the expensive subjective evaluation process. Some previous works on FAC reported CER/WER as an indirect measure of accentedness, with the expectation that reducing accentedness can also reduce the error rates. Nonetheless, we try to investigate whether such a hypothesis is valid.

With the 8 data points in Table I, the linear correlation coefficients between accentedness and CER/WER are 0.413 and 0.442, respectively. It can be then inferred that there is a weak yet insignificant correlation between accentedness and CER/WER. We thus conclude that there are other factors than intelligibility when it comes to accentedness, thus using CER/WER solely as an objective measure for FAC is unreliable.

V. Conclusion and Future Works

In this work, we systematically compared three methods for ground-truth-free FAC. Experiments were carried out in a unified setting, and subjective tests were conducted in terms of naturalness, speaker similarity, and accentedness. In addition to the detailed discussion of each method and evaluation axis presented in Section IV, the most important message that the evaluation results show is that **no single method was significantly better than the other two in all evaluation axes**. While this may arise from the insufficiency of the evaluated methods, we also doubt if the evaluation protocols which we follow are proper. Below we list two directions that we wish to improve upon in the future.

[5]https://unilight.github.io/Publication-Demos/publications/fac-evaluate

[6]Performance and APIs can be found at https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

A. Adopting non-autoregressive (non-AR) seq2seq modeling

We observed that the seq2seq VC models suffer from intelligibility issues, evidenced by the high error rates in Table I. This mainly comes from the autoregressive (AR) design of the seq2seq models we adopted. With the advance of non-AR seq2seq modeling [35], [36] which is known for its robustness compared to AR models, it can be expected that adopting non-AR models in the FAC methods can further improve the performance.

B. Better subjective evaluation protocols

Although we adopted the absolute rating evaluation protocol following previous works [12], [13], the large confidence intervals observed in the naturalness and accentedness axes in Table I suggest that a more reliable protocol needs to be developed. Listener feedback suggests that a 9-point scale test as used in [12], [13], [37] is too fine-grained to give precise ratings. Also, while it is easy to tell whether a sample is native or not, rating the degree of accentedness is rather difficult. Comparative measurements such as preference tests might be more suitable, as advised in [38].

C. More accurate accentedness evaluation

Listeners mentioned that even as native English speakers, it is difficult to confidently rate accentedness. One way to improve this is to provide a training section containing utterances with different levels of accentedness, as the one provided in [39]. Another alternative is to directly recruit linguistics or educators, as someone with in-depth professional knowledge may make judgments more confidently.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Number 21J20920, JST CREST Grant Number JPMJCR19A3, and a project, JPNP20006, commissioned by NEDO, Japan.

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM TASLP*, vol. 29, pp. 132–157, 2021.

[3] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009.

[4] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors – in search of the golden speaker," *Speech Communication*, vol. 37, no. 3, pp. 161–173, 2002.

[5] S. Ding, C. Liberatore, S. Sonsaat, *et al.*, "Golden speaker builder – an interactive tool for pronunciation training," *Speech Communication*, vol. 115, pp. 51–66, 2019.

[6] O. Turk and L. M. Arslan, "Subband based voice conversion," in *Proc. 7th International Conference on Spoken Language Processing*, 2002, pp. 289–292.

[7] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[8] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining Techniques for Sequence-to-Sequence Voice Conversion," *IEEE TASLP*, vol. 29, pp. 745–755, 2021.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, 2016, pp. 1–6.

[10] W.-C. Huang, K. Kobayashi, Y.-H. Peng, *et al.*, "A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion," in *Proc. Interspeech*, 2021, pp. 1329–1333.

[11] W.-C. Huang, B. M. Halpern, L. Phillip Violeta, O. Scharenborg, and T. Toda, "Towards Identity Preserving Normal to Dysarthric Voice Conversion," in *Proc. ICASSP*, 2022, pp. 6672–6676.

[12] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM TASLP*, vol. 29, pp. 2367–2381, 2021.

[13] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Proc. Interspeech*, 2022, pp. 4920–4924.

[14] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 433–446, 2015.

[15] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign Accent Conversion Through Concatenative Synthesis in the Articulatory Domain," *IEEE TASLP*, vol. 20, no. 8, pp. 2301–2312, 2012.

[16] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech and Language*, vol. 36, pp. 260–273, 2016.

[17] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?" In *Proc. ICASSP*, 2014, pp. 7879–7883.

[18] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," in *Proc. ICASSP*, 2018, pp. 5314–5318.

[19] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams," in *Proc. Interspeech*, 2019, pp. 2843–2847.

[20] S. Liu, D. Wang, Y. Cao, *et al.*, "End-to-end accent conversion without using native utterances," in *Proc. ICASSP*, 2020, pp. 6289–6293.

[21] Y. Zhao, W.-C. Huang, X. Tian, *et al.*, "Voice Conversion Challenge 2020 - Intra-lingual semi-parallel and cross-lingual voice conversion -," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 80–98.

[22] W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, S. Watanabe, and T. Toda, "S3PRL-VC: Open-Source Voice Conversion Framework with Self-Supervised Speech Representations," in *Proc. ICASSP*, 2022, pp. 6552–6556.

[23] W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, "A Comparative Study of Self-Supervised Speech Representation Based Voice Conversion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1308–1318, 2022.

[24] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," in *Proc. ICLR*, 2020.

[25] G. Zhao, S. Sonsaat, A. Silpachai, *et al.*, "L2-ARCTIC: A Non-native English Speech Corpus," in *Proc. Interspeech*, 2018, pp. 2783–2787.

[26] K. Ito and L. Johnson, *The lj speech dataset*, https://keithito.com/LJ-Speech-Dataset/, 2017.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[28] J. Kahn, M. Rivière, W. Zheng, *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. ICASSP*, 2020, pp. 7669–7673.

[29] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[30] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is All you Need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[31] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[32] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeruIPS*, 2020.

[33] R. K. Das, T. Kinnunen, W.-C. Huang, *et al.*, "Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 99–120.

[34] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540. DOI: 10.21437/Interspeech.2022-970.

[35] Y. Ren, Y. Ruan, X. Tan, *et al.*, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Proc. NeruIPS*, vol. 32, 2019.

[36] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *Proc. ICASSP*, 2021, pp. 7068–7072.

[37] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[38] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright, "When is it better to compare than to score?" *arXiv preprint arXiv:1406.6618*, 2014.

[39] C. Cucchiarini, H. Van hamme, O. van Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," in *Proc. LREC*, May 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

DisC-VC: Disentangled and F_0-Controllable Neural Voice Conversion

Chihiro Watanabe* and Hirokazu Kameoka†
* NTT Communication Science Laboratories, Japan
E-mail: ch.watanabe@ntt.com Tel/Fax: +81-462404058
† NTT Communication Science Laboratories, Japan
E-mail: hirokazu.kameoka@ntt.com Tel/Fax: +81-462403645

***Abstract*—Voice conversion is a task to convert a non-linguistic feature of a given utterance. Since nuance of speech strongly depends on its pitch pattern, in some applications, it would be desirable to keep the original rise/fall pitch pattern while changing the speaker identity. Some of the existing methods address this problem by either using a source-filter model or developing a neural network that takes an F_0 pattern as input to the model. Although the latter approach can achieve relatively high sound quality compared to the former one, there is no consideration for discrepancy between the target and generated F_0 patterns in its training process. In this paper, we propose a new variational-autoencoder-based voice conversion model accompanied by an auxiliary network, which ensures that the conversion result correctly reflects the specified F_0/timbre information. We show the effectiveness of the proposed method by objective and subjective evaluations.**

I. Introduction

Voice conversion (VC) is a task to convert a non-linguistic feature of a given utterance and it can be used for various applications [1] including personalizing or changing emotion of text-to-speech systems [2]–[4]. In this paper, we focus on the conversion of F_0 pattern and timbre (i.e., speaker identity).

A basic VC framework consists of feature extraction from a speech signal, feature conversion, and speech synthesis based on a converted feature. With regard to feature conversion, some of the existing methods require parallel data set (e.g., Gaussian-mixture-model-based [5]–[7], deep-neural-network-based [8], [9], and sequence-to-sequence-based [10], [11] methods), while others do not (e.g., recognition-synthesis-based [12]–[14], phonetic-posteriorgram-based [15], variational-autoencoder (VAE) [16]-based [17]–[19], generative-adversarial-network-based [20]–[22], flow-based [23], [24], and score-based [25], [26] methods). As for feature extraction and speech synthesis, the existing methods can be classified roughly into two approaches. One is the parametric-vocoder-based (PV) approach [6], [8], [17], where we first apply timbre conversion based on mel-cepstral coefficients (MCCs) and F_0 conversion (e.g., linear transformation) separately and then synthesize a waveform with a parametric vocoder such as WORLD [27]. The other is the neural-vocoder-based (NV) approach [18], [28], [29], where we use a neural network for speech synthesis. While early neural vocoders have required spectral envelope- and excitation-related features as input [30], [31], the recent mainstream is to synthesize a waveform from a mel-spectrogram [32], [33].

In some applications such as online communication, it would be desirable to keep the original rise/fall pitch pattern, since it strongly affects the nuance of speech. Inappropriate modification of the relative pitch pattern would cause miscommunication in regard to the speaker's intent and/or emotion. One way to securely keep the relative pitch pattern of input speech would be to take the PV approach, which allows us to explicitly specify the F_0 pattern of the converted speech. However, one disadvantage is that it cannot generally achieve as high sound quality as the NV approach. Recently, some F_0-controllable NV methods have been proposed, with which we can specify both F_0 and timbre conditionings explicitly [34]–[36]. However, even with these models, a generated speech does not necessarily have the same F_0 pattern as the input conditioning. This is because the relationship between the specified F_0 pattern and the conversion result has not been considered in the training criteria, and thus the decoder may disregard the specified F_0 pattern and not reproduce it faithfully in the reconstruction process.

In this paper, we develop a disentangled and F_0-controllable neural VC (DisC-VC). DisC consists of the VAE-based generator and auxiliary network. The auxiliary network, which is composed of F_0 extractor and speaker classifier, ensures that the conversion result correctly reflects the specified F_0/timbre information. We show the effectiveness of DisC by both objective and subjective evaluations.

II. Proposed Method

A. Preprocessing of training samples

Fig. 1 shows the architecture of the DisC model. During the training phase, we use the standardized log mel-spectrogram $\boldsymbol{x} \in \mathbb{R}^{F \times T}$, the log F_0 pattern $\boldsymbol{\lambda} \in \mathbb{R}^{T}$, and the speaker index $s \in \{1, 2, \ldots, S\}$ as the input features, where T, F, and S indicate the numbers of time and frequency bins and speakers, respectively. To define $\boldsymbol{x}$, we first compute the log mel-spectrogram $\boldsymbol{x}^{(0)}$. Then, each entry of the column vectors of $\boldsymbol{x}^{(0)}$ is standardized according to the mean and standard deviation for all the time bins and training samples. To compute $\boldsymbol{\lambda}$, we extract the F_0 pattern from an input

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

speech signal with the WORLD F_0 estimator [27] and apply logarithmic transformation to its non-zero entries.

B. Generator

The generator consists of content encoder $\mathrm{ENC_C}$ and decoder DEC. To reflect the discrete nature of the phonemic units of speech, we adopt the encoder of Vector Quantised-Variational AutoEncoder or VQ-VAE [40] as $\mathrm{ENC_C}$. In the generator, we first extract the content feature $\boldsymbol{c} = \mathrm{ENC_C}(\boldsymbol{x}) \in \mathbb{R}^{N_\mathrm{C} \times T}$ from $\boldsymbol{x}$, where we set $N_\mathrm{C} = 16$ in the experiment. Then, the decoder reconstructs the mean and variance of the log mel-spectrogram $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}) = \mathrm{DEC}(\boldsymbol{c}, \boldsymbol{\lambda}, s)$, where $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}} \in \mathbb{R}^{F \times T}$.

C. Training objective

1) Reconstruction error: The first training objective is the reconstruction error of the generator. Let $q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})$ be a probability density function (PDF) of content feature $\boldsymbol{c}$ conditioned on log mel-spectrogram $\boldsymbol{x}$, based on which the content encoder with parameter $\boldsymbol{\phi}$ generates the content feature. Let $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}, s)$ be a PDF of $\boldsymbol{x}$ conditioned on $(\boldsymbol{c}, \boldsymbol{\lambda}, s)$, based on which the decoder with parameter $\boldsymbol{\theta}$ generates the output mel-spectrogram. We assume the prior distribution of $\boldsymbol{c}$ to be a discrete uniform distribution independent of the F_0/timbre features (i.e., $p(\boldsymbol{c}|\boldsymbol{\lambda}, s) = p(\boldsymbol{c})$). By using Jensen's inequality, the negative log marginal distribution can be upper bounded by

$$
\begin{aligned}
&-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\lambda}, s) = -\log \int q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}, s) p(\boldsymbol{c})}{q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})} \mathrm{d}\boldsymbol{c} \\
&\leq -\int q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x}) \log \left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}, s) p(\boldsymbol{c})}{q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})} \right) \mathrm{d}\boldsymbol{c} \\
&= D - \mathbb{E}_{\boldsymbol{c} \sim q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}, s)], \qquad (1)
\end{aligned}
$$

where $D \coloneqq D_\mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})||p(\boldsymbol{c})]$ and $D_\mathrm{KL}(q||p)$ indicates the KL divergence of q from p. It can be shown that the difference between the right- and left-hand sides of (1) is equal to $D_\mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{x}, \boldsymbol{\lambda}, s))$, which is minimized when $q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{x}, \boldsymbol{\lambda}, s)$ (i.e., the content feature $\boldsymbol{c}$ extracted by the encoder is independent of $(\boldsymbol{\lambda}, s)$, given $\boldsymbol{x}$). When $p(\boldsymbol{c})$ is a discrete uniform distribution, D becomes constant [40]. By assuming $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}, s)$ to be a Gaussian distribution with mean $\hat{\boldsymbol{\mu}}$ and variance $\hat{\boldsymbol{\sigma}}$, where all the entries are mutually independent, the second term in (1) can be interpreted as the reconstruction error of $\boldsymbol{x}$ [16]. Given training samples of $(\boldsymbol{x}, \boldsymbol{\lambda}, s)$, the actual training criterion $\mathcal{L}_\mathrm{like}$ will be the expectation of the right-hand side of (1) over $(\boldsymbol{x}, \boldsymbol{\lambda}, s)$.

2) Negative mutual information: As we later show experimentally in Section III, even if the conversion model is conditioned on specified F_0/timbre information $(\boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R})$, the converted speech does not always reflect them sufficiently with a model trained only based on the reconstruction loss $\mathcal{L}_\mathrm{like}$, since the decoder may disregard $(\boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R})$ and only use $\boldsymbol{c}$ to reconstruct the log mel-spectrogram. Therefore, as the second training objective, given $\boldsymbol{c} \sim q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x})$ and s_R, we consider the negative mutual information $\mathcal{I}_\mathrm{P}(\boldsymbol{\theta})$ between output log mel-spectrogram $\boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R})$ and log F_0 pattern $\boldsymbol{\lambda}_\mathrm{R}$, which is given by

$$
\begin{aligned}
\mathcal{I}_\mathrm{P}(\boldsymbol{\theta}) &= -\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R}), \boldsymbol{\lambda}'_\mathrm{R} \sim p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{\lambda}'_\mathrm{R}, \boldsymbol{x})}{p(\boldsymbol{\lambda}'_\mathrm{R}) p(\boldsymbol{x})} \right] \\
&= -\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R}), \boldsymbol{\lambda}'_\mathrm{R} \sim p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})} [\log p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})] + H_\mathrm{P}, \qquad (2)
\end{aligned}
$$

where $H_\mathrm{P} \coloneqq \mathbb{E}_{\boldsymbol{\lambda}_\mathrm{R} \sim p(\boldsymbol{\lambda}_\mathrm{R})} [\log p(\boldsymbol{\lambda}_\mathrm{R})]$ is a constant that does not depend on $\boldsymbol{\theta}$. By using the fact that a KL divergence is non-negative and the law of total expectation (i.e., Lemma 5.1 in [41]), the expectation of the first term in (2) over $\boldsymbol{\lambda}_\mathrm{R}$ can be upper bounded by

$$
\begin{aligned}
&-\mathbb{E}_{\boldsymbol{\lambda}_\mathrm{R} \sim p(\boldsymbol{\lambda}_\mathrm{R}), \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R}), \boldsymbol{\lambda}'_\mathrm{R} \sim p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})} [\log p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})] \\
&\leq -\mathbb{E}_{\boldsymbol{\lambda}_\mathrm{R} \sim p(\boldsymbol{\lambda}_\mathrm{R}), \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R}), \boldsymbol{\lambda}'_\mathrm{R} \sim p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})} [\log q(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})] \\
&= -\mathbb{E}_{\boldsymbol{\lambda}_\mathrm{R} \sim p(\boldsymbol{\lambda}_\mathrm{R}), \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R})} [\log q(\boldsymbol{\lambda}_\mathrm{R}|\boldsymbol{x})] =: R_\mathrm{P}, \qquad (3)
\end{aligned}
$$

where $q(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})$ is an arbitrary conditional PDF. The equality holds if $q(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x}) = p(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})$. Therefore, assuming $q(\boldsymbol{\lambda}'_\mathrm{R}|\boldsymbol{x})$ in some parametric form, the left-hand side of (3) can be decreased by alternately decreasing R_P with respect to $(\boldsymbol{\phi}, \boldsymbol{\theta})$ and q. Specifically, for all $i \in \{1, \ldots, T\}$, we represent $q((\lambda'_\mathrm{R})_i|\boldsymbol{x})$ as the Laplace distribution with location parameter $(\hat{\lambda}_\mathrm{R})_i$ and scale parameter of one, where $\hat{\boldsymbol{\lambda}}_\mathrm{R} = \mathrm{P_{ext}}(\hat{\boldsymbol{\mu}}_\mathrm{R})$ is an output of an auxiliary neural network (i.e., F_0 extractor) $\mathrm{P_{ext}} : \mathbb{R}^{F \times T} \mapsto \mathbb{R}^T$ and $(\hat{\boldsymbol{\mu}}_\mathrm{R}, \hat{\boldsymbol{\sigma}}_\mathrm{R}) = \mathrm{DEC}(C, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R})$, and consider training its parameters $\boldsymbol{\psi}_\mathrm{P}$ along with $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Hence, $\mathrm{P_{ext}}$ is trained to estimate the log F_0 pattern from an input log mel-spectrogram.

Given $\boldsymbol{c}$ and $\boldsymbol{\lambda}_\mathrm{R}$, we also consider using the negative mutual information $\mathcal{I}_\mathrm{T}(\boldsymbol{\theta})$ between $\boldsymbol{x}$ and s_R as a training criterion. As above, the expectation of $\mathcal{I}_\mathrm{T}(\boldsymbol{\theta})$ except for a constant term over s_R can be upper bounded by $R_\mathrm{T} \coloneqq -\mathbb{E}_{s_\mathrm{R} \sim p(s_\mathrm{R}), \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R}, s_\mathrm{R})} [\log r(s_\mathrm{R}|\boldsymbol{x})]$, where $r(s_\mathrm{R}|\boldsymbol{x})$ is an arbitrary categorical distribution and the exact bound is achieved when $r(s_\mathrm{R}|\boldsymbol{x}) = p(s_\mathrm{R}|\boldsymbol{x})$. Thus, it can be decreased by alternately decreasing R_T with respect to r and $(\boldsymbol{\phi}, \boldsymbol{\theta})$. Here, we assume $r(s_\mathrm{R}|\boldsymbol{x})$ to be a categorical distribution with each probability given as an output of an auxiliary neural network (i.e., speaker classifier) $\mathrm{CLS} : \mathbb{R}^{F \times T} \mapsto \mathbb{R}^{T_\mathrm{S}}$ [1], and consider training its parameters $\boldsymbol{\psi}_\mathrm{T}$ along with $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\psi}_\mathrm{P}$. Hence, CLS is trained to detect the speaker from an input log mel-spectrogram.

The criteria R_P and R_T are defined under specific $(\boldsymbol{c}, s_\mathrm{R})$ and $(\boldsymbol{c}, \boldsymbol{\lambda}_\mathrm{R})$, respectively. Since $\boldsymbol{c}$, s_R, and $\boldsymbol{\lambda}_\mathrm{R}$ are determined stochastically according to the training examples, the actual criteria used for training will be the expectations of R_P and R_T with respect to $\boldsymbol{c}$, s_R, and $\boldsymbol{\lambda}_\mathrm{R}$. Let $\mathcal{L}_\mathrm{P} \coloneqq \mathbb{E}_{\boldsymbol{x}' \sim p(\boldsymbol{x}'), \boldsymbol{c} \sim q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x}'), s_\mathrm{R} \sim p(s_\mathrm{R})} [R_\mathrm{P}]$ and $\mathcal{L}_\mathrm{T} \coloneqq \mathbb{E}_{\boldsymbol{x}' \sim p(\boldsymbol{x}'), \boldsymbol{c} \sim q_{\boldsymbol{\phi}}(\boldsymbol{c}|\boldsymbol{x}'), \boldsymbol{\lambda}_\mathrm{R} \sim p(\boldsymbol{\lambda}_\mathrm{R})} [R_\mathrm{T}]$, where $\boldsymbol{x}' \sim p(\boldsymbol{x}')$ indicates the log mel-spectrogram of a randomly drawn training sample. The process $\boldsymbol{\lambda}_\mathrm{R} \sim p(\boldsymbol{\lambda}_\mathrm{R})$ can be implemented as random resampling (RR), which converts an input log F_0 pattern $\boldsymbol{\lambda}$ into random one $\boldsymbol{\lambda}_\mathrm{R}$. In the experiment, we defined

[1] T_S is an arbitrary natural number. As shown in Fig. 1, CLS contains convolutional layers and thus T_S depends on T in the experiment.

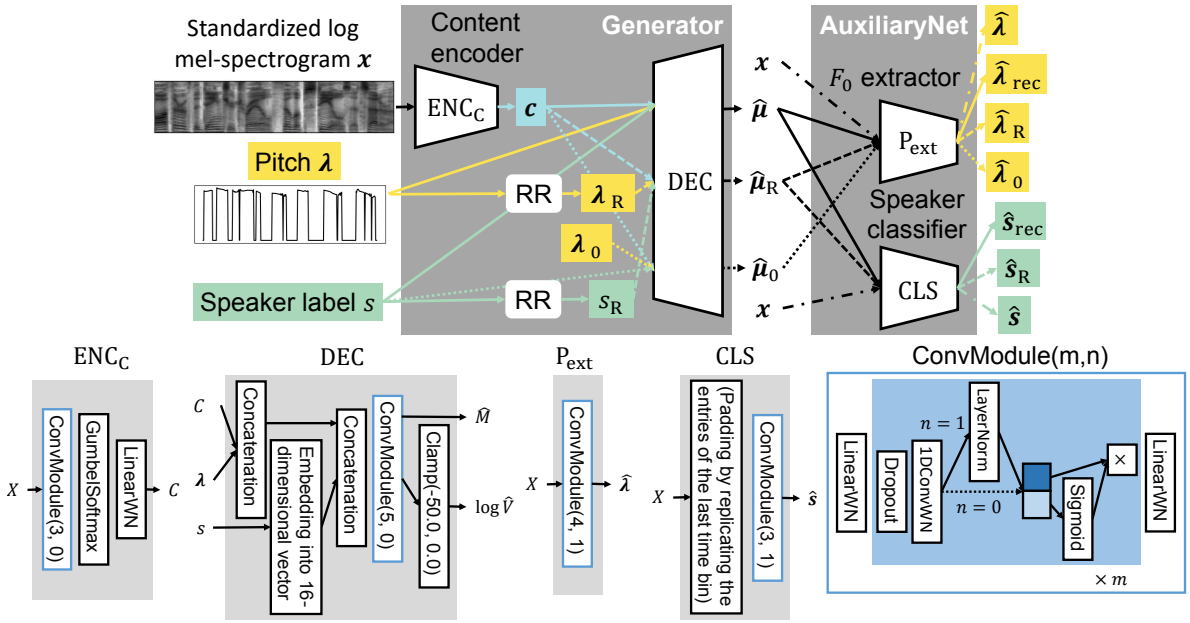


Fig. 1. The network architecture of the DisC-VC. "GumbelSoftmax" means sampling from the Gumbel-Softmax distribution [37]. "LinearWN" and "1DConvWN," respectively, are linear and 1D convolution layers with weight normalization [38] (i.e., the weight vectors are decoupled into their norms and directions). $\mathrm{Clamp}(a, b)$ is a function that clamps input into the range $[a, b]$ by setting the values less than a to a and by setting those more than b to b. "LayerNorm" means layer normalization [39] (i.e., normalization over the hidden units in a layer).

the RR of F_0 by $(\lambda_\mathrm{R})_i = \lambda_i + \beta$ with $\beta \sim \mathrm{Uniform}([0.3, 3))$, if $\lambda_i \neq 0$, and $(\lambda_\mathrm{R})_i = \lambda_i$, otherwise. We also defined the RR of speaker index by generating s_R from a discrete uniform distribution on $\{1, \ldots, S\}$. The idea of RR has also been adopted in SpeechFlow [19], however, there has been no consideration for the discrepancy between the specified F_0/timbre information and those of the conversion result.

The actual flow of computing $\mathcal{L}_\mathrm{P}$ and $\mathcal{L}_\mathrm{T}$ can be explained as follows. For each training sample $\boldsymbol{x}'$, the decoder generates $\boldsymbol{x}$ using $\boldsymbol{c} = \mathrm{ENC_C}(\boldsymbol{x}')$ along with $\boldsymbol{\lambda}_\mathrm{R}$ and s_R randomly drawn from predefined distributions $p(\boldsymbol{\lambda}_\mathrm{R})$ and $p(s_\mathrm{R})$, respectively. The auxiliary network then extracts $\hat{\boldsymbol{\lambda}}_\mathrm{R}$ and $\hat{\boldsymbol{s}}_\mathrm{R}$ from $\boldsymbol{x}$ and checks whether they are consistent with $\boldsymbol{\lambda}_\mathrm{R}$ and s_R. In addition, to better express unvoiced and voiced speech differently, we define $\mathcal{L}_{\mathrm{P}0}$ in a similar way as $\mathcal{L}_\mathrm{P}$ for a zero vector $\boldsymbol{\lambda}_0 \in \mathbb{R}^T$ and $\hat{\boldsymbol{\lambda}}_0 = \mathrm{P_{ext}}(\hat{\boldsymbol{\mu}}_0)$ as a training criterion, where $(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\sigma}}_0) = \mathrm{DEC}(C, \boldsymbol{\lambda}_0, s)$. We also define $\mathcal{L}_{\mathrm{P}1}$ for $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}} = \mathrm{P_{ext}}(\boldsymbol{x})$, $\mathcal{L}_{\mathrm{P}2}$ for $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}}_\mathrm{rec} = \mathrm{P_{ext}}(\hat{\boldsymbol{\mu}})$, $\mathcal{L}_{\mathrm{T}1}$ for s and $\hat{\boldsymbol{s}} = \mathrm{CLS}(\boldsymbol{x})$, and $\mathcal{L}_{\mathrm{T}2}$ for s and $\hat{\boldsymbol{s}}_\mathrm{rec} = \mathrm{CLS}(\hat{\boldsymbol{\mu}})$.

We train the entire network to minimize $\mathcal{L} := \eta\mathcal{L}_\mathrm{like} + \eta_\mathrm{P}(\mathcal{L}_\mathrm{P} + \mathcal{L}_{\mathrm{P}0} + \mathcal{L}_{\mathrm{P}1} + \mathcal{L}_{\mathrm{P}2}) + \eta_\mathrm{T}(\mathcal{L}_\mathrm{T} + \frac{1}{2}\mathcal{L}_{\mathrm{T}1} + \frac{1}{2}\mathcal{L}_{\mathrm{T}2})$ based on back-propagation. In the experiment, we set $\eta = 1/FT$, $\eta_\mathrm{P} = 1/4T$, and $\eta_\mathrm{T} = 1/2T_\mathrm{S}$, and replaced expectation over $\boldsymbol{x}'$ in computing each loss with a sample mean of a given mini-batch, which consists of 16 randomly selected training samples. A training sample was defined as $\{(\boldsymbol{x}_s, \boldsymbol{\lambda}_s, s) | s = 1, \ldots, S\}$, where $\boldsymbol{x}_s$ and $\boldsymbol{\lambda}_s$ were a log mel-spectrogram and a log F_0 pattern of the sth speaker, respectively. It must be noted that the criterion $\mathcal{L}$ can be used not only for training the generator considered in this study, but also for training other existing generator networks that allow F_0/timbre conditioning (e.g., SpeechFlow [19]).

D. Conversion Process

By using the trained DisC model, we can apply the following types of conversion to a given test sample $(\boldsymbol{x}, \boldsymbol{\lambda}, s)$: (1) change log F_0 pattern by a given value $\beta \in \mathbb{R}$, (2) change log F_0 pattern from speaker s to $\check{s}_\mathrm{P}$, and (3) change timbre from speaker s to $\check{s}_\mathrm{T}$. By defining $\beta = 0$, $\check{s}_\mathrm{P} = s$, or $\check{s}_\mathrm{T} = s$, we can also perform each type of conversion individually.

The target log F_0 pattern $\check{\boldsymbol{\lambda}} \in \mathbb{R}^T$ is defined as follows. Let $\mu_\mathrm{P}(s)$ and $\sigma_\mathrm{P}(s)$, respectively, be the sample mean and sample standard deviation of the log F_0 pattern (except for the zero entries) of the sth speaker in the training data set. For all $i \in \{1, \ldots, T\}$, we define $\check{\lambda}_i = [\sigma_\mathrm{P}(\check{s}_\mathrm{P})/\sigma_\mathrm{P}(s)][\lambda_i - \mu_\mathrm{P}(s)] + \mu_\mathrm{P}(\check{s}_\mathrm{P}) + \beta$, if $\lambda_i \neq 0$, and $\check{\lambda}_i = 0$, otherwise.

We obtain $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}) = \mathrm{DEC}(\mathrm{ENC_C}(\boldsymbol{x}), \check{\boldsymbol{\lambda}}, \check{s}_\mathrm{T})$ and reconstruct a waveform from $\hat{\boldsymbol{\mu}}$ with a neural vocoder. In the experiment, we used Parallel WaveGAN [33].

III. Experiment

A. Experimental settings

By using the CMU Arctic speech databases [42], we trained the DisC model with the $1,000$ training samples per speaker for $5,000$ epochs, applied F_0 and/or timbre conversion to the 132 test samples per speaker, and compared the results with those obtained with the following baseline methods. The number of speakers was four in both the training and test data sets.

DisC-NoAux is DisC with no auxiliary network, and it is trained by minimizing $\mathcal{L}_\mathrm{like}$ in Section II-C1. ACVAE-C [17] is

the original ACVAE that takes MCCs as input and reconstructs a waveform from converted MCCs and F_0 pattern by using a WORLD vocoder [27]. ACVAE-S [29] is the extended ACVAE that takes mel-spectrogram as input and reconstructs a waveform from a converted mel-spectrogram by using a neural vocoder. For reference, we also used WORLD vocoder [27] itself for F_0-only conversion by using the MCCs of an input signal and $\check{\boldsymbol{\lambda}}$ given in Section II-D. We also tried F_0 and timbre conversion with BNE-Seq2seqMoL (BNE) [35] as an example of F_0-controllable methods. It must be noted that we cannot compare its result with those of the other methods as it is, since we used the pretrained synthesis model provided by the authors of [35], which was trained with a different data set than the CMU Arctic one.

To compare these methods, we tried **task P** to change log F_0 pattern by $\beta \in \{\log(1.5), -\log(1.5)\}$, **task T** to change timbre to that of a target speaker, and **task PT** to change both F_0 pattern and timbre to those of a target speaker. Since only simultaneous conversion of F_0 pattern and timbre is possible with ACVAE-S, we use the same results in tasks T and PT for this method. Before evaluation, we trimmed the leading and trailing silent sections of all the target and converted samples. In tasks T and PT, we tried all 12 patterns of combinations of source and target speakers.

B. Objective evaluation

To evaluate F_0 and timbre conversion performance, we adopted the F_0 error ΔF_0 (i.e., the root mean squared error between two log F_0 patterns) and the mel-cepstral distortion (MCD), respectively. Since target and converted speech signals are not necessarily aligned in the time direction, we first applied endpoint-free Dynamic Time Warping to a given pair of log F_0 patterns or MCCs and then computed ΔF_0 and the MCD with the optimally aligned sequences. For all the tasks, we computed ΔF_0 between the target log F_0 pattern $\check{\boldsymbol{\lambda}}$ and that of the converted speech. We evaluated the MCD between the samples before and after conversion for task P and those between the target and converted samples for tasks T and PT.

Tables I, II, and III show the results of tasks P, T, and PT, respectively. Table I shows that the timbre of the source speech was maintained more successfully with DisC than with DisC-NoAux in F_0-only conversion. In the same task, we also see that the F_0 conversion performance of DisC was better than that of DisC-NoAux. Tables II and III show that timbre conversion was more successfully done by DisC than by ACVAE-C, whereas F_0 conversion was more successfully done by DisC than by ACVAE-S (especially in timbre-only conversion task). ACVAE-C and -S achieved the best performance in terms of ΔF_0 and the MCD, respectively, however, they could not keep these criteria small at the same time.

C. Subjective evaluation

To conduct subjective evaluation, we randomly chose three test samples for each of the quality and timbre evaluations and for each setting (i.e., a combination of β, a source speaker, and a conversion method in task T and a combination of a source and target speakers and a conversion method in tasks T and PT). For timbre evaluation, we concatenated target and converted speech samples with a 1 s silent interval. We defined the target speech sample as a randomly selected one of the target speaker (i.e., the text content is not necessarily the same as that of the converted one). By considering the fact that we use the same result for tasks T and PT with regard to ACVAE-S, each of 15 listeners was presented 324 samples in each of quality and timbre evaluations.

We conducted a mean opinion score (MOS) test for sound quality, where each listener scores the sound quality of each sample by choosing one out of the five options, "1: Bad," "2: Poor," "3: Fair," "4: Good," and "5: Excellent." Timbre evaluation was done based on the ABX test for speaker similarity, where he/she chooses one out of the four options, "1: Different (sure)," "2: Different (not sure)," "3: Same (not sure)," and "4: Same (sure)," for each concatenated sample.

Tables IV and V, respectively, show the results of the quality and timbre evaluations. In task P, F_0 conversion by WORLD vocoder achieved the best sound quality, however, DisC could maintain the timbre of the source speaker the best. When converting only timbre (i.e., task T), neither DisC nor DisC-NoAux could achieve as high sound quality and timbre

TABLE I
MCD FROM A SOURCE SPEAKER AND F_0 ERROR ΔF_0 IN TASK P. THE VALUES BEFORE AND AFTER $\pm$, RESPECTIVELY, INDICATE THE MEAN AND THE STANDARD DEVIATION FOR ALL THE SETTINGS AND TEST SAMPLES.

	DisC	DisC-NoAux	(WORLD)
MCD	6.14 ± 0.37	6.33 ± 0.47	4.23 ± 1.11
ΔF_0	0.22 ± 0.10	0.24 ± 0.11	0.14 ± 0.12

TABLE II
MCD FROM A TARGET SPEAKER AND ΔF_0 IN TASK T.

	DisC	DisC-NoAux	ACVAE-C	ACVAE-S
MCD	7.29 ± 0.74	7.33 ± 0.78	7.62 ± 0.73	6.65 ± 0.68
ΔF_0	0.24 ± 0.14	0.30 ± 0.20	0.12 ± 0.11	0.36 ± 0.18

TABLE III
MCD FROM A TARGET SPEAKER AND ΔF_0 IN TASK PT.

	DisC	DisC-NoAux	ACVAE-C	ACVAE-S	(BNE)
MCD	6.83 ± 0.63	6.80 ± 0.63	7.62 ± 0.74	6.65 ± 0.68	7.81 ± 0.67
ΔF_0	0.13 ± 0.11	0.14 ± 0.11	0.12 ± 0.10	0.14 ± 0.10	0.19 ± 0.13

TABLE IV
MOS FOR SOUND QUALITY. THE VALUES BEFORE AND AFTER $\pm$, RESPECTIVELY, INDICATE THE MEAN AND THE STANDARD DEVIATION FOR ALL THE LISTENERS, SETTINGS, AND SELECTED TEST SAMPLES.

Task	DisC	DisC-NoAux	WORLD	ACVAE-C	ACVAE-S
P	2.62 ± 1.05	2.42 ± 1.12	2.95 ± 1.17	–	–
T	2.57 ± 1.15	2.42 ± 1.19	–	2.24 ± 0.98	3.55 ± 0.94
PT	3.45 ± 0.91	3.77 ± 0.89	–	2.36 ± 1.04	3.55 ± 0.94

TABLE V
SCORE OF ABX TEST FOR TIMBRE SIMILARITY. EACH VALUE INDICATES SIMILARITY WITH SOURCE SPEAKERS FOR TASK P AND THAT WITH TARGET SPEAKERS FOR TASKS T AND PT.

Task	DisC	DisC-NoAux	WORLD	ACVAE-C	ACVAE-S
P	2.03 ± 0.83	1.85 ± 0.85	1.34 ± 0.59	–	–
T	2.06 ± 1.18	2.06 ± 1.10	–	1.59 ± 0.85	3.41 ± 0.69
PT	3.33 ± 0.80	3.37 ± 0.75	–	2.36 ± 0.88	3.41 ± 0.69

maintenance as ACVAE-S. However, by converting F_0 pattern and timbre simultaneously (i.e., task PT), they could achieve comparable performance to ACVAE-S. A possible reason for this would be that speaker identity strongly depends on the average fundamental frequency, as also pointed out in [43].

IV. Conclusion

We developed a new F_0-controllable VC method DisC that takes into account discrepancy between the target and generated F_0 patterns in the training process. In this paper, we assumed that the same set of speakers appears in training and test data sets. It would be important to extend DisC so that it can be applied to unseen speakers in the test phase. Another future direction would be to take into account other prosody information (e.g., rhythm and stress) as controllable features.

References

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] A. Kain and M. W. Macon, "Text-to-speech voice adaptation from sparse training data," in *ICSLP*, 1998, p. paper 0902.

[3] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[4] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of F0 for speech emotion conversion," in *ICASSP*, 2019, pp. 6830–6834.

[5] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, 1998, pp. 285–288.

[6] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *TASLP*, vol. 6, no. 2, pp. 131–142, 1998.

[7] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *TASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *ICASSP*, 2009, pp. 3893–3896.

[9] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *TASLP*, vol. 18, no. 5, pp. 954–964, 2010.

[10] M. V. Ramos, "Voice conversion with deep learning," Master's thesis, Instituto Superior Técnico, 2016.

[11] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *TASLP*, vol. 27, no. 3, pp. 631–644, 2019.

[12] S. Liu, Y. Cao, S. Kang, N. Hu, X. Liu, D. Su, D. Yu, and H. Meng, "Transferring source style in non-parallel voice conversion," in *Interspeech*, 2020, pp. 4721–4725.

[13] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *ICASSP*, 2021, pp. 5944–5948.

[14] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H. y. Lee, and L. s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP*, 2021, pp. 5939–5943.

[15] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*, 2016, pp. 1–6.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *TASLP*, vol. 27, no. 9, pp. 1432–1443, 2019.

[18] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *ICML*, vol. 97, 2019, pp. 5210–5219.

[19] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *ICML*, vol. 119, 2020, pp. 7836–7846.

[20] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *Interspeech*, 2017, pp. 3364–3368.

[21] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *EUSIPCO*, 2018, pp. 2100–2104.

[22] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion," in *Interspeech*, 2019, pp. 679–683.

[23] J. Serrà, S. Pascual, and C. S. Perales, "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion," in *NeurIPS*, vol. 32, 2019, pp. 6793–6803.

[24] T. Merritt, A. Ezzerg, P. Biliński, M. Proszewska, K. Pokora, R. Barra-Chicote, and D. Korzekwa, "Text-free non-parallel many-to-many voice conversion using normalising flow," in *ICASSP*, 2022, pp. 6782–6786.

[25] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and S. Seki, "VoiceGrad: Non-parallel any-to-many voice conversion with annealed Langevin dynamics," arXiv:2010.02977, 2020.

[26] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *ICLR*, 2022.

[27] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[28] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Interspeech*, 2019, pp. 4115–4119.

[29] H. Kameoka, T. Kaneko, K. Tanaka, and S. Seki, "ACVAE-VC: Auxiliary classifier variational autoencoder voice conversion," http://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/acvae-vc3/index.html.

[30] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *ASRU*, 2017, pp. 712–718.

[31] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Interspeech*, 2017, pp. 1118–1122.

[32] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[33] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020, pp. 6199–6203.

[34] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP*, 2020, pp. 6284–6288.

[35] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *TASLP*, vol. 29, pp. 1717–1728, 2021.

[36] K.-W. Kim, S.-W. Park, J. Lee, and M.-C. Joe, "ASSEM-VC: Realistic voice conversion by assembling modern speech synthesis techniques," in *ICASSP*, 2022, pp. 6997–7001.

[37] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *ICLR*, 2017.

[38] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *NeurIPS*, vol. 29, 2016, pp. 901–909.

[39] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *NeurIPS 2016 Deep Learning Symposium*, 2016.

[40] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NeurIPS*, vol. 30, 2017, pp. 6309–6318.

[41] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *NeurIPS*, vol. 29, 2016, pp. 2172–2180.

[42] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *SSW5-2004*, 2004, pp. 223–224.

[43] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Signal Processing*, vol. 23, no. 2, pp. 176–182, 1975.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Speech Synthesis Using Ambiguous Inputs From Wearable Keyboards

Matsuri Iwasaki*, Sunao Hara* and Masanobu Abe*
* Okayama University, Okayama, Japan
E-mail: px9h3vtm@s.okayama-u.ac.jp, {hara, m-abe}@okayama-u.ac.jp

***Abstract*—This paper proposes a new application in speech communication using text-to-speech (TTS), and the goal is to enable dysarthria, articulation disorder, or persons who have difficulty in speaking to communicate anywhere and anytime using speech to express their thoughts and feelings. To achieve this goal, an input method is required. Thus, we propose a new text-entry method based on three concepts. First, from an easy-to-carry perspective, we used a wearable keyboard that inputs digits from 0 to 9 in decimal notation according to 10-finger movements. Second, from a no-training perspective, users input sentences in a way of touch typing using the wearable keyboard. Following this method, we obtained a sequence of numbers corresponding to the sentence. Third, a neural machine translation (NMT) method is applied to estimate texts from the sequence of numbers. The NMT was trained using two datasets; one is a Japanese-English parallel corpus containing 2.8 million pairs of sentences, which were extracted from TV and movie subtitles, while the other is a Japanese text dataset containing 32 million sentences, which were extracted from a question-and-answer platform. Using the model, phonemes and accent symbols were estimated from a sequence of numbers. Thus, the result accuracy in symbol levels was 91.48% and 43.45% of all the sentences were completely estimated with no errors. To subjectively evaluate feasibility of the NMT model, a two-person word association game was conducted; one gave hints using synthesized speech that is generated from symbols estimated by NMT, while the other guessed answers. As a result, 67.95% of all the quizzes were correctly answered, and experiment results show that the proposed method has the potential for dysarthria to communicate with TTS using a wearable keyboard.**

I. Introduction

The intelligibility and naturalness of text-to-speech (TTS) have greatly improved in recent years by employing deep neural networks, allowing its widespread use in commercial products, such as smart speakers (e.g., Amazon Echo and Google Home) and voice assistance applications (e.g., Apple Siri). A promising application area for TTS is to support speech communications for persons who have difficulty in speaking, such as dysarthria, and articulation disorder [1], [2]. TTS is useful for people who have problems with articulators, and many patients enjoy the benefits of conveying their intentions through TTS. However, the range of use is limited because sentences were input using a keyboard. In general, keyboards have a certain size and weight, so users cannot always input sentences anywhere and anytime, especially while walking. Therefore, to extend the availability and create a new application in speech communication using TTS, we propose a new text-entry method based on three concepts: (1) using a wearable keyboard, (2) touch typing, and (3) employing neural machine translation (NMT) method.



Fig. 1. A wearable keyboard TAP

As the wearable keyboard, as shown in Figure 1, we used TAP [3], which consists of 10 finger rings attached to sensors and Bluetooth. When a user wears TAP and types, the sensors detect an occurrence of finger's movements, encode on-off patterns of the occurrence, and send the code using Bluetooth. To avoid training for usage and the burden of remembering the code, we used the concept of touch typing. The QWERTY keyboard is familiar to most of today's computer users, and users remember the QWERTY layout allowing them to "touch type." We specified numbers 1–5 to the little, ring, middle, index, and thumb fingers of the left hand, respectively. Meanwhile, numbers 6–9 and 0 were specified to the thumb, index, middle, ring, and little fingers of the right hand, respectively. For touch typing, mapping was created between fingers and letters. In general, the mapping is ambiguous. For example, number 1, representing the little finger of the left hand, covers "q," "a," "z," and "1." Therefore, numbers 1, 4, and 3 may represent the words "are" or "ate." Following this method, we obtained a sequence of numbers corresponding to a sentence via the wearable keyboard. Then, NMT is trained and has shown powerful performance in mapping variable-length sequences into continuous vector representations [4], [5], [6], [7], [8], [9], [10], [11]. This paper aims to use the NMT method to estimate a sentence from an ambiguous input sequence of numbers.

The Japanese writing system has several types of characters, including Kanji and Kana characters. Notably, the Kanji characters have multiple corresponding pronunciations, which are determined by different contexts, showing that a Kanji character has several corresponding sequences of letters.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Moreover, the number of Kanji characters is more than 50,000. Therefore, sentence estimation from a sequence of numbers might be complicated. However, conventional Japanese TTS systems usually have a preprocessing module that estimates a sequence of phonemes and accent symbols from a sentence represented by Kanji and Kana characters. Because of the drawbacks of multiple pronunciations in Kanji and a large number of Kanji characters, it might be another choice to estimate a sequence of phonemes and accent symbols from a sequence of numbers. This point was also examined in this paper.

This paper is organized as follows: Section 2 mentions related works, Section 3 describes the outline of the proposed method, Section 4 explains the setup for conducting the evaluation experiments of the proposed method, Section 5 shows the objective experiment methods and evaluation results, Section 6 illustrates the subjective experiment methods and evaluation results, and Section 7 discusses the conclusions and proposes avenues for future work.

II. Related works

Text entry is a fundamental function essential not only for desktops and personal computers but also for mobile devices. The QWERTY keyboard is currently common to most computer users. Furthermore, soft QWERTY keyboards have become a popular text-entry method for smartphones. However, for wearable devices, such as smartwatches, soft QWERTY keyboards may be too large to be optimally displayed on touchscreens. Therefore, a remarkable challenge is encountered while trying to enter text on small-sized devices. An approach to solve the challenge involves two-step key selection methods [13], [14], [15], [16], [17]. Here, the QWERTY keyboards are divided into multiple sections. First, after a user selects a section, the section is displayed larger. Second, the user selects a desired character from the enlarged section. The biggest drawback of this approach is that the user must train for at least 10 hours to fully understand the system.

In terms of text-entry methods for ubiquitous computing, an eyes-free feature that enables users to input texts without looking at the keyboard is sometimes essential. This method is applied because, in a mobile scenario, users must pay visual attention to other tasks during text entry. Speech input is a promising candidate, but it has not been appropriate for every circumstance because of noisy environments and privacy reasons. Therefore, touch-based and eyes-free text-entry methods have been majorly used instead [13], [14], [19], [20], [21]. Since the minimum efforts of learning to type are one of the most essential requirements, the QWERTY layout has been used in many cases. The biggest drawback of this method is that the eyes-free touch system is inaccurate on a touch panel [18]. Thus, errors in text entry occur in most cases.

Within the existing study, [22] is the most essential work to ours, which eliminates the keyboard and replaces it with gloves sensing only the pressure of a depressed finger and not location. Since missing location information creates ambiguity

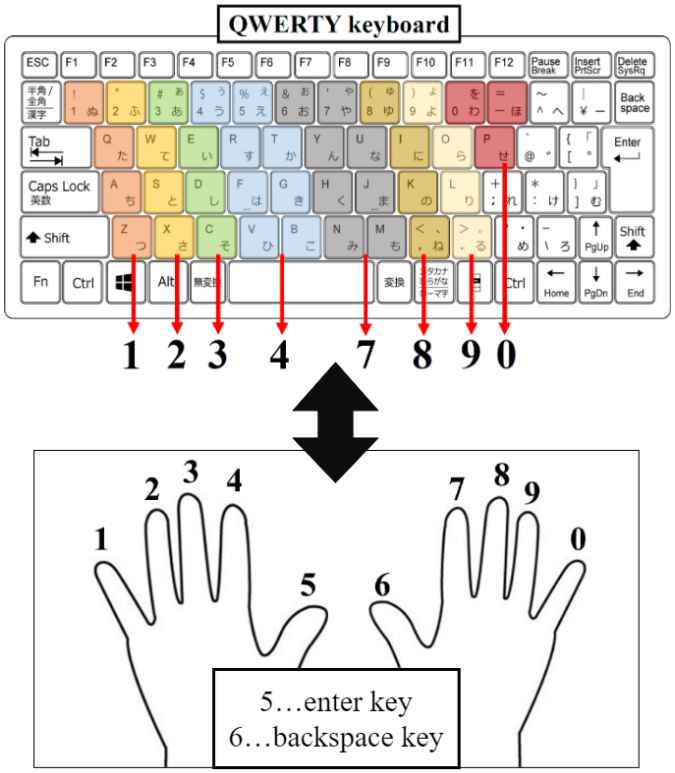


Fig. 2. The mapping between numbers and alphabets using a QWERTY keyboard

in the input stream, the researchers in the previous study attempted to estimate words based on lexical level matching and frequency-ranking estimation [22]. However, their vocabulary was small, and the execution of their study was limited. Thus, the proposed method includes the NMT. In terms of input devices, we will use a wearable QWERTY keyboard based on the "FingeRing" concept [23]. Examples of this method can be found in previous studies [24], [25], [3], [26].

III. Proposed Method

An NMT model is a neural network that directly predicts the conditional probability $p(Y|X)$ of translating a source sentence $X : x_1, \cdots, x_n$ into a target sentence $Y : y_1, \cdots, y_m$. Here, x_i or y_i represents a word. An NMT model consists of two components: (a) an encoder that computes a representation X for each source sentence, and (b) a decoder that generates a target word y_i at a single time. This process is formulated as

$$\log p(Y|X) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, X). \quad (1)$$

In the proposed model, x_i is a number corresponding to a letter using a QWERTY keyboard layout. Figure 2 shows the mapping between the fingers and letters. Here, as special mappings, numbers 5 and 6 represent the enter key and the backspace key, respectively. As shown in Figure 2, the mapping is ambiguous. For example, the number 1 may represent either "q," "a," "z," or "1." However, the word "the" is uniquely mapped to number

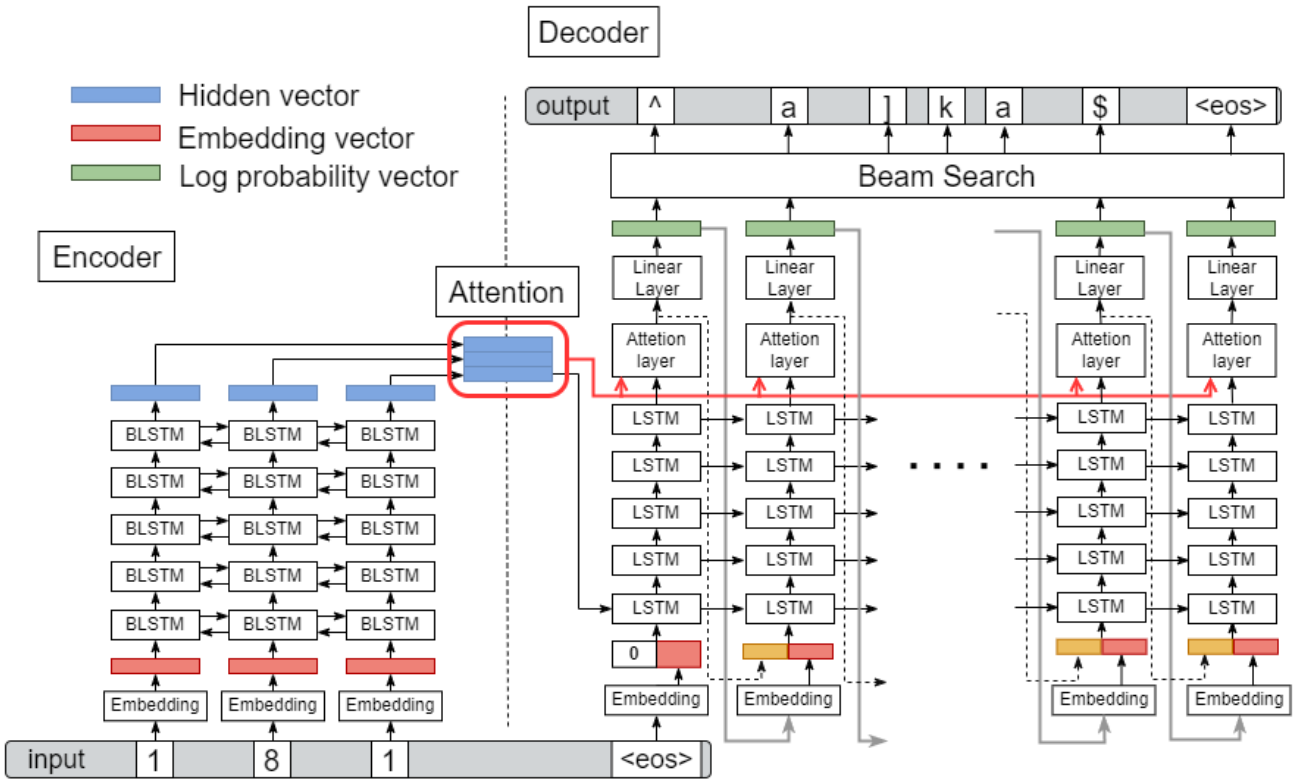


Fig. 3. An outline of the phoneme-based NMT model

473. To resolve such ambiguity, we used two types of NMT models. A character-based model estimates Japanese sentences including Kanji and Kana characters, while a phoneme-based model estimates a sequence of phonemes and accent symbols. Hence, y_i is either a Kanji and Kana character or a phoneme and accent symbol.

The text-entry problem is developed to estimate the sequence of output symbols $Y : y_1, \cdots, y_m$ from a sequence of numbers $X : x_1, \cdots, x_n$. The NMT model is trained sentence-by-sentence. For example, the number 21412871318418723833275 corresponds to the letter sequence of "watasihadaigakuseidesu ('I am a college student' in English)." The character- and phoneme-based models are expected to estimate "私は大学生です" and "ˆ w a [t a sh i w a # d a [i g a] k u s e e d e s u $," respectively. Here, ˆ, [,], #, $ shows the start of a sentence, F0-raise, F0-fall, accent phrase boundary, and end of the sentence, respectively.

Figure 3 shows an outline of the phoneme-based model. As shown in Figure 3, we used a stacking bidirectional long short-term memory (Bi-LSTM) for the encoder and a stacking LSTM for the decoder. In the encoder, each character is first represented by a 1-of-K representation, which is compressed by the embedding layer. The obtained hidden vector is inputted into the Bi-LSTM. The hidden vector becomes an input to the next layer's Bi-LSTM and is saved as an input to the attention mechanism [6]. Thus, it is essential to learn the context and alignment of characters by the attention mechanism because the proposed method has an ambiguity in the transformation. In the decoder, a character is estimated using both hidden vectors of the encoder and a character estimated in the previous time step.

IV. Experiment Data

To confirm the feasibility of the proposed method, experiments on two datasets were conducted. One is a Japanese-English parallel corpus (hereafter referred to as JESC) [28] containing 2.8 million pairs of sentences extracted from TV and movie subtitles. The other dataset is Yahoo! Chiebukuro data (version 3) [29] containing 32 million Japanese sentences. This corpus consists of sentences posted on a bulletin board system operated by Yahoo!

A. Dataset

1) Japanese-English Subtitle Corpus (JESC): JESC is a Japanese-English parallel corpus that consists of 2.8 million pairs of crawled TV and movie subtitles [28]. JESC consists mainly of short bursts of conversational dialog. The corpus contains a substantial number of colloquial expressions commonly found in spoken language. In this paper, we extracted and used the Japanese sentences from the paired dataset.

2) Yahoo! Chiebukuro data (version 3): Yahoo! Chiebukuro is a Japanese question-and-answer platform operated by Yahoo!. Users can post questions on various topics and receive

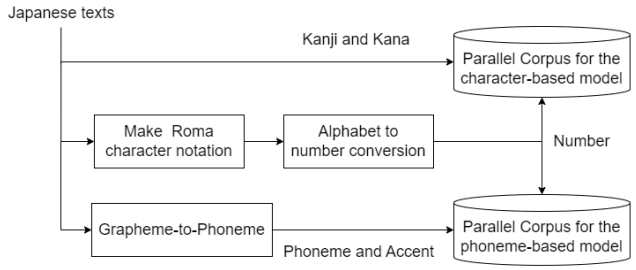


Fig. 4. Parallel corpus generation

responses from the community. The platform covers a wide range of topics, meaning that the corpus includes vocabulary on various topics, and the sentences in the corpus majorly consist of casual registers in written communication.

B. Parallel corpus construction for NMT training

A parallel corpus, which consists of pairs of characters sequences and corresponding numbers sequences, has been generated.

Figure 4 shows an outline to create a parallel corpus for sequences. First, morphological analysis is performed to assign Roman character notations to correspond to Kanji characters. Second, a number is generated by referring to a table that maps QWERTY keyboards to respective fingers corresponding to the keys at the time of touch typing. As a parallel corpus for the character-based model, the original Japanese sentences are paired with the generated number sequences. However, for the phoneme-based model, after generating a sequence of phonemes and accent symbols using the Grapheme-to-Phoneme function in ESPnet2 [30], the sequence is paired with the generated number sequences.

C. Experimental conditions

Table I shows the model configuration and training conditions. For the number of Kanji characters, all Kanji characters included in the training data were sorted in order of their appearance frequency. Then, the top 50,000 Kanji characters were selected. The 50,000-dimensional 1-of-K vectors representing the vocabularies are then embedded into 512 dimensions. These were later used as input to the Bi-LSTM.

V. Objective Evaluation

To evaluate the performance of the different kinds of output characters and the influence of the dataset size, we evaluated accuracy.

A. Configurations of Models

Four models were constructed for evaluation. Table II shows details of their training and evaluation data. Three models estimate the sequences of phonemes and accent symbols, and the other estimates sequences of Kanji and Kana characters. The models for the sequences of phonemes and accent symbols are distinguished by the training data. The datasets for Ph_JESC, Ph_Yahoo, and Ph_JESC+Yahoo are 2.7M, 32M, and 35M in size, respectively. Ja_JESC+Yahoo is trained on the same dataset as Ph_JESC+Yahoo.

TABLE I
Training conditions

Model	RNN-based encoder-decoder model
(Encoder)	-
RNN	Bi-LSTM
#layers	5 layers
#units	512 units in each layer
(Decoder)	-
RNN	LSTM
#layers	5 layers
#units of 1st layer	$1024 \rightarrow 512$
#units of 2nd–5th layers	$512 \rightarrow 512$
(Linear)	-
#units	$512 \rightarrow$ number of symbol types
activation function	Softmax
mini-batch	128 sentences
optimizer	Adam
loss function	Log-Softmax cross entropy

B. Evaluation Results

The models are evaluated by the accuracy of the estimated sentences. The evaluation indices are symbol and sentence accuracy. Symbol accuracy represents the percentage of estimated symbols that match symbols in reference sentences, while sentence accuracy indicates the percentage of estimated sentences that match the reference. For phonemes and accent symbols, the references are the sequences of phonemes and accent symbols generated by grapheme-to-phoneme. For Kanji and Kana characters, the references are original Japanese texts. Table III shows the experimental results.

In terms of output character types, phoneme and accent symbols are superior to Kanji characters in accuracy. The sentence accuracy of Ph_JESC+Yahoo is 12.15 points higher than that of the Ja_JESC+Yahoo. This shows that the sequences of phonemes and accent symbols are more suitable for estimation from ambiguous input than the sequences of Kanji and Kana characters.

Regarding the size of the training data, the symbol accuracy for Ph_JESC, Ph_Yahoo (32M), and Ph_JESC+Yahoo were 84.02%, 89.77%, and 91.48%, respectively. The symbol accuracy rate increases in proportion to the size of the training data. This shows that increasing the amount of data is effective in improving accuracy.

For phonemes and accent symbols, the accuracy for just phonemes symbols is higher than that for both phonemes and accent symbols in all models. The symbol and sentence accuracy for phonemes symbols are on average 1.07 and 5.02 points higher than for both phonemes and accent symbols, respectively. This shows that estimating accent symbols is more difficult than estimating phonemes.

TABLE II
TRAINING AND EVALUATION DATA FOR THE NMT MODELS

Model name	Type of output symbols	Dataset	Number of sentences		
			Training	Validation	Test
Ph_JESC	phoneme and accent	JESC	2.7M	2000	2000
Ph_Yahoo	phoneme and accent	Yahoo	32M	2000	2000
Ph_JESC+Yahoo	phoneme and accent	JESC and Yahoo	35M	2000	2000
Ja_JESC+Yahoo	Kanji and Kana	JESC and Yahoo	35M	2000	2000

TABLE III
AVERAGE ACCURACY OF NMT MODELS

Model name	accuracy per symbol		accuracy per sentence	
	all symbols	phoneme	all symbols	phoneme
Ph_JESC	84.02 %	85.27 %	26.85 %	31.55 %
Ph_Yahoo	89.77 %	90.62 %	43.65 %	48.10 %
Ph_JESC+Yahoo	91.48 %	92.55 %	43.45 %	49.35 %
Ja_JESC+Yahoo	84.02 %	-	31.10 %	-

TABLE IV
ESTIMATED SENTENCES FOR THE TEST DATA.
WORDS WRITTEN IN RED INDICATE ERRORS

answer 1	^ **b a [r e t e** # t o [r i k e s h i n i # n a] r u n a r a # ···
estimated 1	^ **t a [b e] t e** # t o [r i k e s h i n i # n a] r u n a r a # ···
answer 2	^ k o [o k o o s e e n o # j i [k y u **]** **u** s o o n i # ts u] i t e $
estimated 2	^ k o [o k o o s e e n o # j i [k y u **u** **]** s o o n i # ts u] i t e $
answer 3	··· # o [n o] o n o n o # y o] i # t e [N # w a [r u] i # t e [N g a # h i] cl t o # ···
estimated 3	··· # o [n o] o n o n o # y o] i # t e [N # w a [r u] i # t e [N **# w a [r u] i # t e [N # w a [r u] i # t e [N # w a [r u] i #** ···

C. Analysis of Estimated Errors

Table IV shows the estimated sentences for the test data. The sentences shown in the table are not complete, but rather part of them. Words written in red indicate errors. From the table, we categorize these errors into three types.

The first-type of error is due to word conflicts. As mentioned in the section III, a numbers sequence can correspond to more than one word. For example, in the estimated 1, "tabete (eating)" is estimated from 414347 although "barete (exposing)" is correct. Since the two words are completely different in meaning, this error could be resolved by considering the wider context of the conversation.

The second-type is an error in the estimation of accent symbols as shown in the estimated 2. The first cause is a selection of homonyms. Japanese words contain some homonyms with different accents. For example, we distinguish between bridges, chopsticks, and edges by differences in their accents. Context-aware NMT models could also improve the selection of homonyms. Another reason is that NMT models cause excess and deficiency of accent symbols. This error might be negligible in conversation since the intended sentences could be assumed even when they include incorrect accent symbols. The frequency of the second-type error can be observed in Table III. The category "all symbols" in Table 3 represents accuracy for both phonemes and accent symbols, whereas "phonemes" represents accuracy for phonemes only. Therefore, the differences between sentence accuracy for all symbols and that for phonemes indicate the percentage of sentences containing errors in phonemes. In the case of Ph_JESC+Yahoo, the difference is 5.9 points. According to detail analysis, the value means that there are 118 sentences with errors in phoneme estimation, which corresponds to 10.4% of the error sentences.

The third-type of error is the repetition of identical phrases as shown in the estimated 3. The phrase "waruiten" is generated repeatedly. This error occurs in a sentence in which similar phrases appear several times, attributing that NMT models focus on the same position repeatedly. Therefore, a new attention algorithm that prevents NMT from repeatedly focusing on the same position would be effective in reducing the third-type of error.

VI. SUBJECTIVE EVALUATION

To subjectively evaluate feasibility of the NMT model, we carried out conversational experiments. The participants performed a guessing game task in pairs using QWERTY keyboards instead of the wearable keyboards. In this experiment, we use the phoneme-based model for the translation from numbers to characters. The participants are four pairs of eight.

A. Experimental Procedures

Participants conducted the guessing task in pairs, one of them using the proposed system, while the other used his voice. The participant who used the proposed system is a speaker, and the other is a listener. The guessing task consisted of the following steps:

1) The speaker chooses an item from a set of items. Each set consists of three items that are familiar for all participants.
2) The speaker enters sentences that describe the characteristics of the item by touch typing.
3) The system synthesizes speech from the inferred sequence and plays it back in real-time.
4) The listener predicts the item based on the synthesized speech.
5) The speaker repeats steps 2-3 until the listener could speak out an answer.

Note that the listeners were allowed to respond only once to prevent random guesses. When the system made an estimation

TABLE V
AVERAGE ACCURACY OF NMT MODELS

	all symbols	phoneme only
accuracy per symbol	90.57 %	92.86 %
accuracy per sentence	50.00 %	53.85 %

error and synthesized speech different from that of the intended speaker, the speakers are permitted to enter sentences that state the same feature in different expressions.

Five trials were performed for each pair. In the experiment, the speakers touch-typed sentences using a QWERTY keyboard in Roman character notations, then the Roman character notations were converted to numbers sequences for the NMT. We recorded the sequences of the Roman character notations, the numbers sequences inputted to NMT, the estimated symbols, and the synthesized speech. In the evaluations, the Roman character notations were used as answers: what the speakers had tried to say.

B. Accuracy of estimated sentences

Table V shows the accuracy of the sequences estimated in the conversational experiment. The results showed the almost same tendencies with the results of Section V.

C. Evaluation by Comprehensibility

After each trial, we let the participants look back on the conversation. To indicate what the speaker had tried to say, we showed the participants the sentences in Roman character notations which the speaker input by keyboard. To evaluate comprehensibility, which is the extent to which a listener understands a speech produced by a speaker in the guessing tasks, we asked the listeners to recall the synthesized speech and evaluate its comprehensibility using three categories: understandable, not understandable but guessable, and unguessable. The evaluation was possible for listeners, because each trial took a short period of time; i.e., an average of 3.8 sentences were synthesized, and the evaluation was carried out just after each trial. For speakers, we showed both of the estimated symbols and input sentences in Roman character notations and asked them to evaluate their satisfaction with a corresponding estimated symbols using two categories: intended or unintended.

Figure 5-(a) shows the results for the listener and Figure 5-(b) for the speaker. For 5-(a), the total of understandable and guessable is 67.95%, meaning that the comprehensibility of the proposed method is 67.95%. By referring to Table V sentence accuracy is 17.95 points less than comprehension. This shows that an utterance can be understood in some cases even if it is not perfectly estimated. For example, three sentences contained only errors in accent symbols. The listeners evaluated all of these sentences as understandable. The speaker evaluations show that 60.26% of the synthesized speech is as intended. Moreover, 10.26% of the synthesized speech was evaluated as unintended because of typing errors.

Table VI shows the comprehensibility each pair. The comprehensibility of the best pair is 82.35%, while that of the

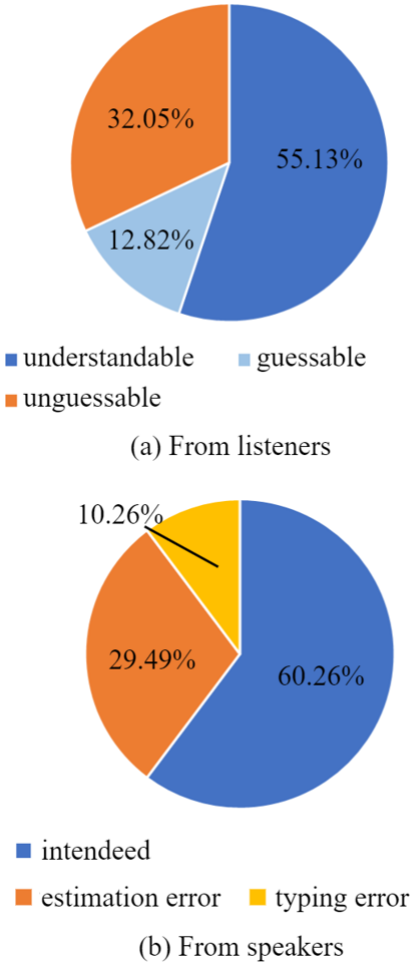


Fig. 5. The results for subjective evaluation

TABLE VI
COMPREHENSIBILITY FOR EACH GROUP

pair name	Comprehensibility
pair A	82.35%
pair B	50.00%
pair C	71.46%
pair D	68.18%

worst pair is 50.00%. This observation suggests that the comprehensibility depends on users. Among the speakers, the participant from pair A might be a quick learner. Actually, the accuracy of the estimation results depends on grammatical complexity and word difficulty. When an appropriate sentence is inputted into the NMT model, it has the potential to enhance the accuracy of the estimation. The speaker of the pair A might find the points during the experiments. These facts indicate that the comprehensibility is expected to increase as users get used to the proposed system.

VII. CONCLUSIONS

An NMT-text-entry method was applied in this paper using the QWERTY keyboard layout, and a digit from 0 to 9 in

decimal notation was inputted by fingers, resulting in an ambiguous input for alphabets. NMT, an attention-based encoder-decoder model, estimated a sentence from a sequence of numbers, and four NMT models were trained using two kinds of datasets. The experimental results revealed the following: (1) The most accurate model correctly estimates 91.48% of the characters in a sentence, (2) the performance of phoneme and accent symbols is better than the performance of the Kanji and Kana characters, and (3) the speech synthesized from estimated sequences conveys 67.95% of the sentences in a conversation. Thus, judging from the results, we can conclude that the proposed model has high potential as a text-entry method. Consequently, we plan to evaluate the proposed method using a wearable keyboard as an input device in the future. Furthermore, we need to investigate dealing with typing errors.

VIII. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP21K11963.

References

[1] Junichi Yamagishi, Christophe Veaux, Simon King, and Steve Renals. 2012. *Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction*, Acoustical Science and Technology 33.1: 1-5.

[2] Keisuke Matsubara, Takuma Okamoto, Ryoichi Takashima, Tetsuya Takiguchi, Tomoki Toda, and Yoshinori Shiga et al. 2021. *High-intelligibility speech synthesis for dysarthric speakers with LPCNet-based TTS and CycleVAE-based VC*, ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 7058-7062.

[3] Tap 2023. https://www.tapwithus.com/introducing-tap/

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. *Sequence to sequence learning with neural networks*, In Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS'14, pages

[5] Kyunghyun Cho, Bart van Merrienboer, C¸ aglar G¨ulc¸ehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning phrase representations using rnn encoder-decoder for statistical machine translation*, CoRR.

[6] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. *Effective approaches to attention-based neural machine translation*, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP):1412–1421.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. *Neural machine translation by jointly learning to align and translate*, In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

[8] Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. *Character-based neural machine translation*, arXiv preprint arXiv:1511.04586.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Neural machine translation of rare words with subword units*, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): 1715–1725, Association for Computational Linguistics.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*, In Advances in Neural Information Processing Systems 30: 6000–6010. Curran Associates, Inc.

[11] Makoto Morishita, Jun Suzuki and Masaaki Nagata. 2018. *Improving Neural Machine Translation by Incorporating Hierarchical Subword Features*, Proceedings of the 27th International Conference on Computational Linguistics: 618-629.

[12] Barry Brown, Moira McGregor and Donald McMillan. 2014. *100 days of iPhone use: understanding the details of mobile device use*, In Proceedings of the 16th international conference on Human-computer interaction with mobile devices and services (MobileHCI '14): 223-232.

[13] Sunggeun Ahn, Seongkook Heo and Geehyuk Lee. 2017. Alternation. *Typing on a Smartwatch for Smart Glasses*, In Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17): 201-209.

[14] Xiang 'Anthony' Chen, Tovi Grossman and George Fitzmaurice. 2014. *Swipeboard: a text entry technique for ultra-small interfaces that supports novice to expert transitions*, In Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14): 615-620.

[15] Hyeonjoong Cho, Miso Kim and Kyeongeun Seo. 2014. *A text entry technique for wrist-worn watches with tiny touchscreens*, Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology (UIST '14): 79-80.

[16] Luis A. Leiva, Alireza Sahami, Alejandro Catala, Niels Henze and Albrecht Schmidt. 2015. *Text Entry on Tiny QWERTY Soft Keyboards*, In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15): 669-678.

[17] Yuan-Fu Shao, Masatoshi Chang-Ogimoto, Reinhard Pointner, Yu-Chih Lin, Chen-Ting Wu and Mike Chen. 2016. *SwipeKey: a swipe-based keyboard design for smartwatches*, In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16): 60-71.

[18] Yuntao Wang, Chun Yu, Jie Liu, and Yuanchun Shi. 2013. *Understanding Performance of Eyes-free, Absolute Position Control on Touchable Mobile Phones*, In Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '13). ACM, New York, NY, USA: 79–88.

[19] Yiqin Lu, Chun Yu, Xin Yi, Yuanchun Shi, and Shengdong Zhao. 2017. *BlindType: Eyes-Free Text Entry on Handheld Touchpad by Leveraging Thumb's Muscle Memory*, Proc. ACM Interact. Mob. Wearable Ubiquitous

[20] Suwen Zhu, Tianyao Luo, Xiaojun Bi and Shumin Zhai. 2018. *Typing on an Invisible Keyboard*, In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18): 1-13.

[21] Tovi Grossman, Xiang Anthony Chen and George Fitzmaurice. 2015. *Typing on Glasses: Adapting Text Entry to Smart Eyewear*, In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services

[22] M. Goldstein, R. Book, G. Alsi¨o, and S. Tessa. 1999. *Non-keyboard QWERTY touch typing: a portable input interface for the mobile user*, In Human Factors in Computing Systems.

[23] M. Fukumoto and Y. Tonomura. 1997. *"Body Coupled FingeRing": Wireless Wearable Keyboard*, In Proceedings of CH1'97, ACM Press: 147-154.

[24] R. Rosenberg and M. Slater. 1999. *The chording glove: a glove-based text input device*, IEEE Trans. Syst., Man, Cybern 29, 2 : 186–191

[25] J. K. Cho and S. Lee. 2004. *Wearable Finger Motion Sensor for Sensing Finger Motion and Method of Sensing Finger Motion Using the Same*, Google Patents. US Patent App. 10/876,5. Retrieved from https://www.google.com/patents/US20040 3473.

[26] J. Lehikoinen and M. R¨oykkee. 2001. *N-fingers: A finger-based interaction technique for wearable computers*, Interacting with Computers 13, 5 (2001), 601–625

[27] Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn and Christof Monz. 2018. *Findings of the 2018 Conference on Machine Translation (WMT18)*, In Proceedings of the Third Conference on Machine Translation: Shared Task Papers: 272-303.

[28] Reid Pryzant, Youngjoo Chung, Dan Jurafsky and Denny Britz. 2018. *JESC: Japanese-English Subtitle Corpus*, In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[29] Yahoo Japan Corporation 2018. *Yahoo! Chiebukuro data (3rd edition)*. Informatics Research Data Repository, National Institute of Informatics. (dataset). https://doi.org/10.32130/idr.1.3

[30] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan. 2020. *Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to- Speech Toolkit*, In Proceedings of 2020 IEEE International Con- ference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7654–7658

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Accent-Preserving Voice Conversion between Native-Nonnative Speakers for Second Language Learning

Iago Lourenço Correa, Sei Ueno and Akinobu Lee
Nagoya Institute of Technology, Japan
E-mail: i.lourencocorrea.747@stn.nitech.ac.jp, sei.ueno@nitech.ac.jp, ri@nitech.ac.jp

***Abstract*—The use of generated corrected accent speech with a learner's own voice holds potential for second language learners to enhance their pronunciation through self-imitation exercises. However, synthesizing speech with a corrected accent poses significant challenges, particularly due to the unavailability of data from the same speaker exhibiting both native and non-native accents. To address this, voice conversion (VC) and foreign accent correction (FAC) approaches have been explored. VC aims to transform the voice of a native speaker source while preserving linguistic content, whereas FAC targets accent conversion of a non-native source while retaining the speaker's voice. This study investigates the application of the MaskCycleGAN-VC model as an approach for generating corrected accent speech from native sources. The model is evaluated as it was originally proposed and with the adoption of speaker embeddings, speaker native language ID, conditioned discriminator, and accent loss. Objective and subjective metrics are assessed on English and Japanese corpora, and including speaker embedding, speaker native language ID, and conditioned discriminator showed improvements.**

I. Introduction

In order to learn a second language, students need to devote time outside of class to study [1]. In addition, many teachers focus on reading and writing skills, while neglecting the importance of speech in communication [2], [3]. In this context, computer-assisted language learning (CALL) systems appear as a solution applying computers in the language learning environment. However, most research works focus on teaching vocabulary, ignoring other skills, especially speech. Thus, it is worth exploring an automated method for evaluating the speech of non-native (L2) learners.

Given an L2 language learner, one approach is to synthesize speech with a native-like accent (L1) but transplant the L2 speaker's voice onto it, creating a "golden speaker". Since the only difference between the golden speaker and the L2 learner's own voice is the accent, the mispronunciations become more salient [4], enabling practice by self-imitation, i.e., the learner would be able to imitate their own voice, a pedagogically effective practice [5].

To achieve this type of feedback, two different tasks can be explored, the first one is voice conversion (VC) [6], and the second one is foreign accent conversion (FAC) [4]. In the VC task, the goal is to convert the speaker's voice while preserving linguistic content and speech style from the source audio. To generate self-imitation feedback for language learners with this approach, a native source audio is required for the conversion. On the other hand, FAC seeks to transform the features of an utterance that contribute to accent while maintaining those that carry the identity of the speaker [7], being able thus to reduce the foreign accent in the non-native speaker's speech. Since it is not feasible to have paired data from the same speaker with and without a native accent, generating self-imitation feedback without a native source audio becomes a more challenging task than traditional VC. In this work, some different works on VC and FAC will be briefly explained.

In order to generate a native accented reference for self-imitation, in this work the MaskCycleGAN-VC [8] model was explored with English and Japanese native and non-native data. This model was originally proposed and evaluated for VC task with native data only, which is different from converting voice with two different accents involved, after all, the non-native speaker is not guaranteed to produce correct phonemes in their speech. However, the proposed approach is capable of generating corrected accent audio with non-native speakers' voices, such as language learners. To try to improve the model conversion, some alterations to the original MaskCycleGAN-VC were explored. First, the incorporation of a speaker embedding and a speaker native language ID as conditional inputs to the generator, giving information related to the desired speaker voice and accent for the conversion. Then, in addition to the speaker embedding and native language ID for the generator, the discriminator was also conditioned to the desired accent with the speaker native language ID. Finally, an accent loss was also explored to encourage the generator to output the desired accent.

This paper is organized as follows. Section II briefly discusses different VC and FAC works. Section III describes the proposed methodology with MaskCycleGAN-VC. Section IV presents the results obtained with two different datasets and their discussion. Finally, conclusions and future work are drawn in Section V.

II. Related Work

Corrected accent speech for self-imitation can be achieved with two possible approaches. The first one is VC, in which the primary goal is to change the speaker identity of a given speech into another speaker, i.e., convert from an utterance

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

source speaker's voice onto a target speaker while preserving the linguistic content and the speech style from the source. With this approach, corrected accent speech can be synthesized by converting a native source utterance onto the voice of a non-native speaker, hence achieving native-accented speech with the voice of a non-native speaker. Applying different approaches and architectures, recent VC models can be trained without parallel data or text reference.

Sequence-to-sequence (seq2seq) models are one common approach for achieving VC. In general, those models consist of encoder and decoder modules, in which the first one extracts latent space representations from a given input, and the second one should be able to decode that representation into the desired output sequence. In many cases when parallel data is available, seq2seq models can be considered a state-of-art approach for sequence conversion [9]. For the VC task, they aim to disentangle the input into the speaker and linguistic information so that VC can be achieved by changing only the speaker information [9]–[11].

In [11], two encoders and one decoder are used for the conversion, with an additional encoder during the training step. The two encoders used during the conversion extract linguistic and speaker representations, and the additional encoder used during the training also extracts linguistic representation from phoneme transcriptions. In sequence, the decoder recovers acoustic features from those disentangled linguistic and speaker representations combined. The work [10] extracts bottleneck features (BNF) as a linguistic representation with a phoneme recognizer, and then combines that information with some speaker representation so that a seq2seq synthesis model can reconstruct the spectral features. To achieve normal speech to dysarthric speech conversion, [9] trains a many-to-one seq2seq model to convert normal speech into dysarthric speech, and subsequently applies a non-parallel frame-wise model to convert the identity to the desired normal speaker's voice.

Another approach is to use generative adversarial networks (GANs), such as in [8], [12], [13]. In general terms, GANs are composed of generator and discriminator networks, in which the first one is trained to generate data that resembles real data from noise input, and the second should distinguish whether the generated data looks real or not. For VC task, the generator is trained to convert the input speech speaker identity while the discriminator assesses if the generated data belongs to the real distribution.

In [12] a cycle-consistent adversarial network-based non-parallel VC model is proposed for mel-cepstrum conversion. This model is later improved in [13] by adopting a two-step adversarial loss and changing the generator and discriminator architectures. Then, for achieving mel spectrogram conversion, [8] proposes MaskCycleGAN-VC, which improves the previous model [13] by training the model with an auxiliary filling-in-frames task. In this task, a mask is applied to the input spectrogram so that the model is encouraged to learn time-frequency structures in a self-supervised manner. For the VC task, other approaches can also be found such as the flow-based VC model by [14], or the diffusion-based approach by [15], [16].

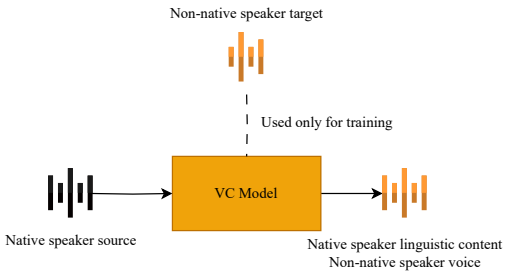


Fig. 1. Voice conversion scheme to generate feedback for self-imitation.

In the FAC approach, the desired input for the model is not a source native, but rather a non-native speaker. The goal is to convert the accent while preserving the voice. Since it seeks to not require a native reference for conversion, this task is more aligned with some practical implementations for feedback generation for learners, however, it is a more challenging task given that it is not feasible to have data from the same speaker with and without a native accent. Considering it is a harder task there are fewer works related to FAC.

Translator and synthesizer models are proposed in [17]. The first one translates BNF from an L2 speaker into L1, and the second converts the BNF into a mel spectrogram. On the other hand, [4] creates a "golden-speaker" utterance that is used for training a pronunciation correction model, so that in the inference it is capable of achieving FAC. Both of these works, incorporate seq2seq architecture models. Then, [18] propose a Tacotron-based text-to-speech model that is capable of synthesizing L2 speaker speech with corrected accent without parallel data in the training. In addition to those works, [19] trains a seq2seq conversion model that disentangles speaker identity and accent information, achieving a model capable of both VC and FAC between Chinese and Tianjin-accented Mandarin. In [20] VC model is used in order to synthesize a parallel dataset for FAC.

III. Methodology

This work proposes to apply the MaskCycleGAN-VC model for the conversion of a native source speaker's voice to a non-native target voice without altering the linguistic content, including pronunciation. This approach generates the corrected accent speech for self-imitation with the VC approach as in Fig. 1. A native source is used as input for the generation, and the non-native data is only needed for training the model. This approach differs from the FAC that takes as input non-native speech.

In addition to applying the MaskCycleGAN-VC model as it is, some alterations to the model were also explored and evaluated. Those alterations are as follows:

1) Applying a speaker embedding and speaker native language ID as conditional inputs to the generator.

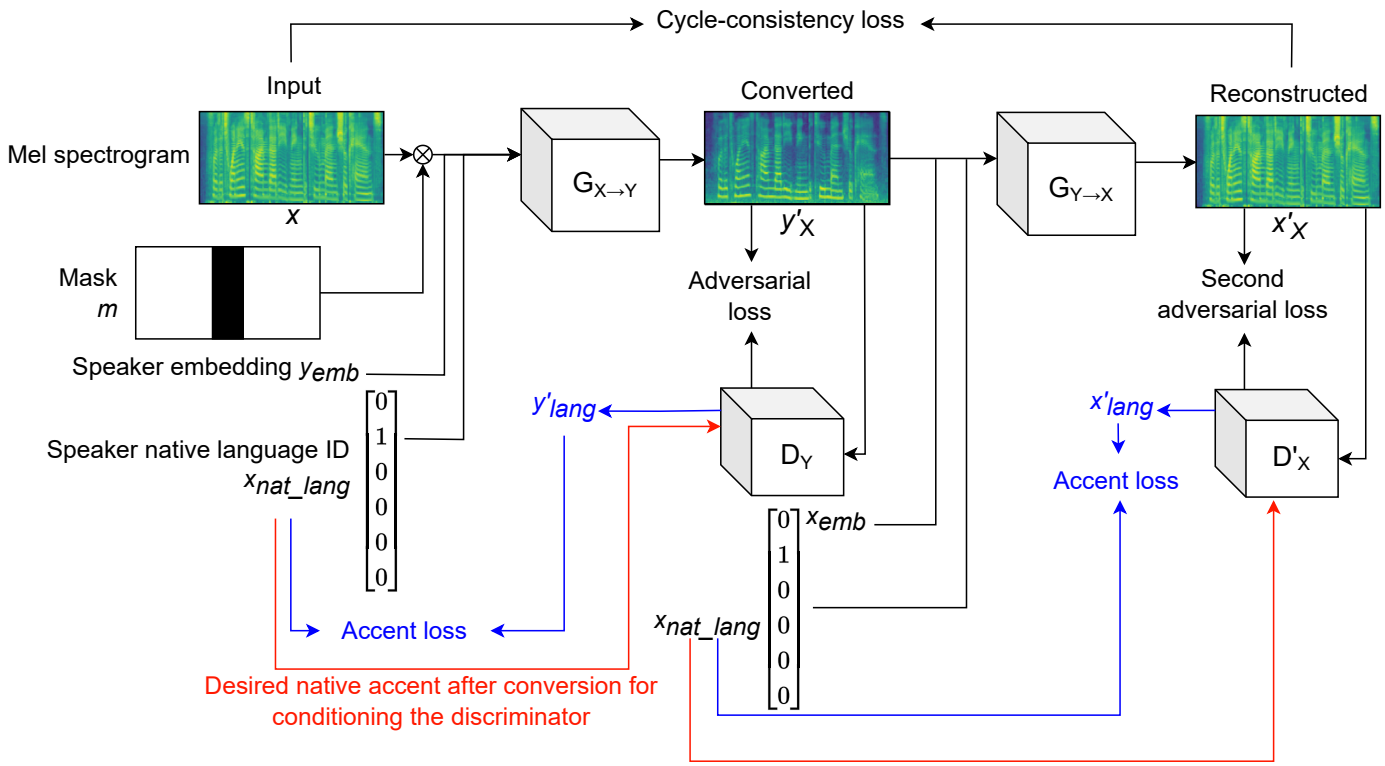


Fig. 2. Illustration of one training step from x to y (the step from y to x is ommited). Accent loss and discriminator are both depicted in the figure, but they are not applied simultaneously. For the accent loss, in addition to y' and x', the accent loss is also calculated in relation to the real input x in relation to both discriminators.

2) Conditioning the discriminator to the desired accent in addition to 1. The conditioning of the discriminator during training is illustrated in Fig. 2 in red.
3) Calculating an accent loss in relation to an additional classifier layer for the discriminator in addition to 1. The calculation of the accent loss during training is illustrated in Fig. 2 in blue.

A. MaskCycleGAN-VC

MaskCycleGAN-VC is a non-parallel cycle-consistent adversarial network-based voice conversion model [8]. It extends previously proposed cycle-consistent models works [12], [13] by adding an auxiliary filling-in-frames training task, outperforming them for pair-wise conversion of native speakers' mel spectrograms. In this auxiliary task, a mask is applied to the input mel spectrogram during training, therefore the generator should learn the time-frequency structures of a mel spectrogram in a self-supervised way. In comparison to the previous works, this addition enabled mel spectrogram conversion instead of mel cepstrum.

Different from the original work, in this work the model was applied to the voice conversion between a native speaker and a non-native speaker with the goal of generating corrected accent speech. Considering that non-native data presents different intonation and may not present all the correct phonemes from a native speaker, this task is expected to be harder than the native-to-native conversion. More details about the architecture of the MaskCycleGAN-VC model are presented in IV-B.

B. Speaker embedding and speaker native language ID

Speaker identity representations, such as speaker embeddings extracted from other models or simple speaker IDs, can be used as an additional input to the VC model in order to condition the speaker identity of the generated speech [4], [10], [11], [17], [19], [21]. In a similar way, adopting a class identification to condition the generator is also common for conditional GAN (CGAN) architectures [22].

In consideration of that, the adoption of a speaker embedding and a speaker native language ID as additional inputs to the generator was explored to condition the model generation to the desired speaker identity and accent, possibly increasing voice similarity and the output of native accented speech. Both are added to condition the input mel spectrogram after their dimensions are adjusted by a linear transformation.

C. Discriminator conditioning

For this alteration, the speaker embedding and speaker native language ID are still used as input to the generator. However, instead of conditioning only the generator, the native language ID of the desired accent is also used as an additional input to the discriminator. Similarly, the native language ID has its

dimensions adjusted by a linear transformation and is added to the discriminator input before its first layer.

By employing this, the discriminator is expected to judge if a given input belongs to the real distribution given the accent condition. This alteration is explored to try to achieve a better native-like accent after conversion since the discriminator could penalize generated data with the incorrect accent.

D. Accent loss based on speaker native language ID

For the MaskCycleGAN-VC [8] four different losses were originally calculated as presented below. For the total loss calculation, weight parameters $\lambda_{cyc} = 10$ and $\lambda_{id} = 5$ are applied to the cycle-consistency loss and identity-mapping loss consecutively.

1) Adversarial loss: ensures that the conversion results appear to be from the target speaker.
2) Cycle-consistency loss: checks if the results from the circular conversion are consistent, which assists the model training without parallel supervision.
3) Identity-mapping loss: encourages the input preservation.
4) Second adversarial loss: adds two additional discriminators and imposes the adversarial loss on the circular conversion results. In total, there are four discriminators in the model.

Given the incorporation of the speaker embedding and the speaker native language ID, for this alteration, an additional loss related to the native language of the accent is calculated. For this loss calculation, an additional fully-connected layer is added to the end of the discriminator as an additional classifier layer, so that the outputs from the discriminator are two, one judging whether the generated data resembles real data, and the other classifying its accent in relation to its native language, i.e., the speaker native language ID. For example, for a generated English sample, the discriminator could classify its native language as American, meaning it sounds native-like, or Chinese, meaning it does not sound native-like. For the sake of briefness, in this work this loss is abbreviated as accent loss.

In this work, the MaskCycleGAN-VC model is trained for converting from native speech to non-native speaker voice (X to Y), and also from non-native speech to native speaker voice (Y to X). Considering the first conversion, $X \rightarrow Y$, the generator $G_{X\rightarrow Y}$ converts from the native source acoustic features $\boldsymbol{x} \in X$ with native language $\boldsymbol{x_{nat_lang}}$ into the non-native target voice $\boldsymbol{y} \in Y$ maintaining the same $\boldsymbol{x_{nat_lang}}$ native accent. Then, D_Y predicts the native language accent for this $G_{X\rightarrow Y}(\boldsymbol{x}, \boldsymbol{x_{nat_lang}})$, and the difference between the predicted accent to the expected native accent $\boldsymbol{x_{nat_lang}}$ is calculated, as illustrated in (1). The $D_Y(*)[1]$ and $D_X(*)[1]$ in the next equations indicate that only one of the outputs of the discriminator is used for the calculation of this loss, the output related to classifying the accent.

$$\begin{gathered} \boldsymbol{y'_X} = G_{X\rightarrow Y}(\boldsymbol{x}, \boldsymbol{x_{nat_lang}}) \\ {\mathcal{L}_{acc}^{X\rightarrow Y}}_1 = \mathbb{E}_{x\sim P_X}\left[\left\|D_Y(\boldsymbol{y'_X})[1] - \boldsymbol{x_{acc}}\right\|_1\right] \end{gathered} \tag{1}$$

Considering the circular conversion, the previous output is used as input for the second generator, $G_{Y\rightarrow X}(\boldsymbol{y'_X}, \boldsymbol{x_{nat_lang}})$, and D'_X classifies its output native language accent for the second part of the accent loss, as in (2).

$$\begin{gathered} \boldsymbol{x'_X} = G_{Y\rightarrow X}(\boldsymbol{y'_X}, \boldsymbol{x_{nat_lang}}) \\ {\mathcal{L}_{acc}^{X\rightarrow Y}}_2 = \mathbb{E}_{x\sim P_X}\left[\left\|D_Y(\boldsymbol{x'_X})[1] - \boldsymbol{x_{nat_lang}}\right\|_1\right] \end{gathered} \tag{2}$$

Finally, the real data x is used as input for the discriminators D_X and D'_X to ensure both are being trained to identify correctly the native language accent. The outputs from both discriminators are compared to the real expected accent as in (3) and (4), and the final accent loss for the $X \rightarrow Y$ is defined as the sum of equations (1) to (4).

$${\mathcal{L}_{acc}^{X\rightarrow Y}}_3 = \mathbb{E}_{x\sim P_X}\left[\left\|D_X(\boldsymbol{x})[1] - \boldsymbol{x_{nat_lang}}\right\|_1\right] \tag{3}$$

$${\mathcal{L}_{acc}^{X\rightarrow Y}}_4 = \mathbb{E}_{x\sim P_X}\left[\left\|D'_X(\boldsymbol{x})[1] - \boldsymbol{x_{nat_lang}}\right\|_1\right] \tag{4}$$

The loss is also calculated for the conversion $Y \rightarrow X$, but instead of preserving the native accent $\boldsymbol{x_{nat_lang}}$, it seeks to preserve the accent from the native language of the speaker that is non-native $\boldsymbol{y_{nat_lang}}$. For this conversion, the equations are defined in (5) to (8):

$$\begin{gathered} \boldsymbol{x'_Y} = G_{Y\rightarrow X}(\boldsymbol{y}, \boldsymbol{y_{nat_lang}}) \\ {\mathcal{L}_{acc}^{Y\rightarrow X}}_1 = \mathbb{E}_{y\sim P_Y}\left[\left\|D_X(\boldsymbol{x'_Y})[1] - \boldsymbol{y_{nat_lang}}\right\|_1\right] \end{gathered} \tag{5}$$

$$\begin{gathered} \boldsymbol{y'_Y} = G_{X\rightarrow Y}(\boldsymbol{x'_Y}, \boldsymbol{y_{nat_lang}}) \\ {\mathcal{L}_{acc}^{Y\rightarrow X}}_2 = \mathbb{E}_{y\sim P_Y}\left[\left\|D_X(\boldsymbol{y'_Y})[1] - \boldsymbol{y_{nat_lang}}\right\|_1\right] \end{gathered} \tag{6}$$

$${\mathcal{L}_{acc}^{Y\rightarrow X}}_3 = \mathbb{E}_{y\sim P_Y}\left[\left\|D_Y(\boldsymbol{y})[1] - \boldsymbol{y_{nat_lang}}\right\|_1\right] \tag{7}$$

$${\mathcal{L}_{acc}^{Y\rightarrow X}}_4 = \mathbb{E}_{y\sim P_Y}\left[\left\|D'_Y(\boldsymbol{y})[1] - \boldsymbol{y_{nat_lang}}\right\|_1\right] \tag{8}$$

Finally, when the accent loss is calculated, the discriminator is not conditioned with the desired accent, as described previously in III-C, but the generator is conditioned to both speaker embedding and speaker native language ID.

IV. Experimental Results

A. Experimental Conditions

In this work, two experiments were conducted; English conversion and Japanese conversion. For native English conversion, we used the CMU-ARCTIC[1] [23] which consists of English native speakers (mostly Americans), and the L2-ARCTIC[2] [24] which consists of non-native English speakers with 6 different native languages. Both are composed of read speech from out-of-copyright texts from Project Gutenberg.

[1]http://festvox.org/cmu_arctic/

[2]https://psi.engr.tamu.edu/l2-arctic-corpus/

TABLE I
AMOUNT OF DATA PER SPEAKER FOR THE ENGLISH DATA.

Speaker	Gender	Native language	Train (min.)	Evaluate (min.)
clb	Female	US	51.39	12.76
rms	Male	US	52.75	13.35
lxc	Female	ZH	60	15.49
hqtv	Male	VN	55.05	13.93

TABLE II
AMOUNT OF DATA PER SPEAKER FOR THE JAPANESE DATA.

Speaker	Gender	Native language	Train (min.)	Evaluate (min.)
JJ_KYO_F01	Female	JP	6.85	0.73
JJ_KYO_M02	Male	JP	6.15	0.74
FJ_IWA_F03	Female	ZH	12.63	1.30
FJ_KYO_M03	Male	ZH	11.81	1.35

Since they have partially parallel data, those datasets are commonly used in conversion tasks with foreign-accented speech, such as in [4], [17], [18]. Considering that different speakers may read the same text, the data used for training is not totally non-parallel, but data used in the training for one speaker is selected so as to not be present in the evaluation. In this experiment 2 native and 2 non-native speakers were selected, and their amount of data used for training and evaluation is presented in Table I.

For native Japanese conversion, we used the UME-JRF[3] [25] dataset. This dataset is also composed of read speech, but the amount of data available per speaker is considerably shorter. In the same way as the datasets for English, the data used for training is not completely unparallel, but samples used for training are not present in the evaluation. Again, 4 speakers were selected, and their corresponding amount of data used for training and evaluation is present in Table II.

For both language experiments, models were trained for pair-wise conversion between the same genders and different genders, i.e., female-female, male-male, female-male, and male-female. The audio data were sampled at 16kHz, and 80-dimensional log mel spectrogram with a window length of 1024 and hop length of 256 samples were extracted according to the original MaskCycleGAN-VC [8]. For the English experiments, the model was trained for 50 epochs with batch size of 1, and for the Japanese experiments, the model was trained for 100 epochs with batch size of 1. In order to synthesize a waveform from the converted mel spectrograms, the HiFi-GAN [26] vocoder was fine-tuned for the selected speakers using the same data used in the conversion model training.

Furthermore, the speaker embeddings, for the altered MaskCycleGAN-VC, were extracted with the Resemblyzer[4]. The embeddings are 256-length vector representations for the speakers. As for the native language ID, it is a one-hot representation of an ID associated with the accent of the speaker. The speaker embedding of the target speaker and the native language ID of the source speaker (desired accent) are both added as input to the generator.

B. Model Configuration

The MaskCycleGAN-VC model is composed of two generators and four discriminators. The generator structure is composed of two 2D convolutions for downsampling, six residual blocks with 1D convolutions, and two 2D convolutions for upsampling, following the structure of the 2-1-2D CNN [13]. As for the discriminator, it follows the structure from PatchGAN [27], which consists of four 2D convolution layers for downsampling.

For the generator, both downsample and upsample layers have a factor of 2 and kernel sizes of 5x5, for the residual blocks the kernel size is 1x3 and 256 channels. For the discriminator, the first three downsample layers have a scale factor of 2 and kernel sizes of 3x3, but the last downsample has a scale factor of 1 and kernel size of 1x10. Finally, 1x1 convolution was applied before and after reshaping the feature map [13] for both the generator and discriminator.

C. Evaluation Metrics

For the evaluation of each model, 3 objective metrics and 3 subjective metrics were evaluated. The objective metrics were the mel-cepstral distortion (MCD), root of mean square errors of F_0 (F0 RMSE), word error rate (WER), and character error rate (CER). MCD goal is to assess the spectral conversion, F0 RMSE the fundamental frequency conversion, and the WER or CER tries to assess the intelligibility to some extent. For both WER and CER, two ESPnet[5] [28] pre-trained ASR models were utilized. WER is evaluated only for the English language with a transformer-based ASR model trained on LibriSpeech[6], meanwhile, since the division of a sentence into words is less clear in Japanese, CER is calculated for the Japanese language with a conformer-based ASR model trained on LaboroTVSpeech[7].

For the subjective evaluation, experiments were conducted with 14 English speakers and 10 Japanese speakers. For both languages, the calculated metrics were naturalness, native-likeness, and voice similarity. For both naturalness and native-likeness, the participants listened to converted audio and were asked to evaluate it from 1 to 5. In terms of naturalness 1 is equivalent to bad, and 5 to excellent. For native-likeness, 1 is equivalent to heavily foreign-accented and 5 to native-like pronunciation. In the voice similarity experiment, the participants listened to 3 audios: 1 converted, 1 from the source speaker, and 1 from the target speaker. In sequence, they were asked to choose which one the converted one had the voice more similar to. The results for voice similarity are expressed in the percentage of participants who chose that the converted audio is more similar to the target.

[3]https://research.nii.ac.jp/src/UME-JRF.html
[4]https://github.com/resemble-ai/Resemblyzer
[5]https://github.com/espnet/espnet
[6]https://www.openslr.org/12
[7]https://github.com/laboroai/LaboroTVSpeech

TABLE III
NON-ALTERED MODEL OBJECTIVE EVALUATION FOR ENGLISH CONVERSION PER SPEAKERS PAIRS.

Conversion pair	MCD (↓)	F0 RMSE (↓)	WER (↓)
Native Female to Non-native Female	8.79	32.35	6.2
Native Male to Non-native Male	10.12	73.34	15.1
Native Female to Non-native Male	11.20	34.77	10.3
Native Male to Non-native Female	9.34	88.35	14.9

TABLE IV
ALTERED MODEL OBJECTIVE EVALUATION FOR ENGLISH CONVERSION FOR ALL THE FOUR SPEAKERS PAIRS.

Model version	MCD (↓)	F0 RMSE (↓)	WER (↓)
No alterations	**9.86**	57.20	**11.6**
Speaker native language ID + Speaker embedding	9.88	57.11	12.3
Speaker native language ID + Speaker embedding + Conditioned discriminator	10.88	**56.78**	14.8
Speaker native language ID + Speaker embedding + Accent loss	11.14	60.49	34.2

TABLE V
NATURALNESS AND NATIVE-LIKENESS FOR ENGLISH CONVERSION FOR ALL THE FOUR SPEAKERS PAIRS WITH 95% CONFIDENCE INTERVALS.

Model version	Naturalness	Native-likeness
No alterations	**3.37** $\pm$ 0.22	2.67 $\pm$ 0.23
Speaker native language ID + Speaker embedding	3.21 $\pm$ 0.25	2.63 $\pm$ 0.25
Speaker native language ID + Speaker embedding + Conditioned discriminator	3.13 $\pm$ 0.23	**2.76** $\pm$ 0.24
Speaker native language ID + Speaker embedding + Accent loss	2.6 $\pm$ 0.27	2.29 $\pm$ 0.25

TABLE VI
VOICE SIMILARITY FOR ENGLISH CONVERSION FOR ALL THE FOUR SPEAKERS PAIRS WITH 95% CONFIDENCE INTERVALS.

Model version	Similarity to target
No alterations	0.56 $\pm$ 0.12
Speaker native language ID + Speaker embedding	0.54 $\pm$ 0.13
Speaker native language ID + Speaker embedding + Conditioned discriminator	0.56 $\pm$ 0.12
Speaker native language ID + Speaker embedding + Accent loss	**0.58** $\pm$ 0.12

D. English Conversion

1) Objective evaluation: The objective evaluation results obtained for the English data are summarized in Table III and Table IV. Table III presents the results for each conversion pair trained with the non-altered MaskCycleGAN-VC model, while Table IV shows the overall results for the different alterations tested with the MaskCycleGAN-VC model.

As it can be observed from Table III, the female-female conversion has the lowest values within all the pairs. Moreover, the conversions with native male source yield the highest F0 RMSE and WER values. Then, from Table IV, the adoption of speaker embedding, speaker native language ID, and discriminator conditioning, resulted in a slight decrease in the F0 RMSE, but most alterations did not improve the other two metrics.

2) Subjective evaluation: The subjective evaluation for English language data is presented in Table V and Table VI. From Table V it is possible to observe that the increase in the accent reduction (higher native-likeness), does not follow an increase in naturalness. In fact, the highest naturalness observed was with the non-altered MaskCycleGAN-VC model, while the highest native-likeness was observed when adopting speaker embedding, speaker native language ID, and conditioned discriminator. In terms of voice similarity, the proposed alterations did not present any improvement.

E. Japanese Conversion

1) Objective evaluation: In the same manner as for the English language results, for the Japanese data, the objective evaluation results are summarized in Table VII and Table VIII. As it can be observed from Table VII the female-female conversion has the lowest values for F0 RMSE and CER, but not for MCD. The conversions with native male sources yield the highest CER values. Then, from Table VIII, the adoption of speaker embedding, native language ID, and discriminator conditioning, resulted in a decrease for all three metrics, indicating its potential for improving the conversion between Japanese native and non-native speakers. The adoption of speaker embedding and native language ID also brings some improvement to MCD and WER.

TABLE VII
NON-ALTERED MODEL OBJECTIVE EVALUATION FOR JAPANESE CONVERSION PER SPEAKERS PAIRS.

Conversion pair	MCD (↓)	F0 RMSE (↓)	CER (↓)
Native Female to Non-native Female	9.82	35.84	32.8
Native Male to Non-native Male	8.44	45.53	43.2
Native Female to Non-native Male	8.21	111.33	34.4
Native Male to Non-native Female	11.7	89.78	56.9

TABLE VIII
ALTERED MODEL OBJECTIVE EVALUATION FOR JAPANESE CONVERSION FOR ALL THE FOUR SPEAKERS PAIRS.

Model version	MCD (↓)	F0 RMSE (↓)	CER (↓)
No alterations	9.54	70.62	41.8
Speaker native language ID + Speaker embedding	9.43	71.25	36.8
Speaker native language ID + Speaker embedding + Conditioned discriminator	**9.02**	**67.88**	**35.4**
Speaker native language ID + Speaker embedding + Accent loss	12.36	83.40	64.8

TABLE IX
NATURALNESS AND NATIVE-LIKENESS FOR JAPANESE CONVERSION FOR ALL THE FOUR SPEAKERS PAIRS WITH 95% CONFIDENCE INTERVALS.

Model version	Naturalness	Native-likeness
No alterations	2.44 ± 0.23	2.94 ± 0.27
Speaker native language ID + Speaker embedding	**2.7** ± 0.27	3.05 ± 0.28
Speaker native language ID + Speaker embedding + Conditioned discriminator	2.56 ± 0.23	**3.11** ± 0.25
Speaker native language ID + Speaker embedding + Accent loss	1.56 ± 0.19	1.94 ± 0.26

TABLE X
VOICE SIMILARITY FOR JAPANESE CONVERSION FOR ALL THE FOUR SPEAKERS PAIRS WITH 95% CONFIDENCE INTERVALS.

Model version	Similarity to target
No alterations	0.51 ± 0.15
Speaker native language ID + Speaker embedding	**0.64** ± 0.13
Speaker native language ID + Speaker embedding + Conditioned discriminator	0.35 ± 0.18
Speaker native language ID + Speaker embedding + Accent loss	0.36 ± 0.17

2) Subjective evaluation: The subjective evaluation for Japanese language data is presented in Table IX and Table X. In the same way, as for the English data, the highest native-likeness was observed when adopting speaker embedding, native language ID, and conditioned discriminator. For the Japanese data, the naturalness presents a slight increase in naturalness as well. In terms of voice similarity, only the adoption of speaker embedding and native language ID was able to bring some minor improvement, but the approach with the greater accent correction presented the worst voice similarity.

F. Discussion

Although the MaskCycleGAN-VC model was not originally proposed for conversion between native-nonnative speech, it is still capable of achieving the conversion, but not as well as between two native speakers. Some tentative changes were experimented with, and from the results, it is possible to observe that the adoption of speaker embedding, native language ID, and conditioned discriminator was the model alteration that rose the most improvement, although not for all cases.

Moreover, the adoption of an accent loss proved to worsen the non-altered model for most cases. Since the models were trained for pair-wise conversion, it is not surprising that the actual improvement was slim to non-existing, after all, in a pair-wise conversion the speaker and accent information are always the same, and there is not much gain of information from a speaker embedding or a speaker native language ID.

Overall, even though it was trained for more epochs, the results for Japanese data were worse, as was expected since the amount of data available was smaller. However, conducting experiments with another language besides English is important from a practical perspective.

V. CONCLUSIONS

This work presented the results of applying the MaskCycleGAN-VC model in the task of converting from a native source speaker voice to a non-native target speaker. This model was trained pair-wise for four different pairs, with two different languages, English and Japanese. In addition to the direct application of the models, some alterations to the model were also proposed. Those results can be divided into objective and subjective evaluation, and the metrics calculated were the MCD, F0 RMSE, CER/WER, naturalness, native-likeness, and voice similarity.

Although originally tested only on the native-to-native conversion, the model still is capable of performing VC, but with more limited performance. According to the experimental results, the inclusion of speaker embedding, native language ID, and conditioned discriminator led to improvements in some cases. However, the adoption of an accent loss did not yield positive results. It is supposed that since the models were trained for pair-wise conversion, the benefits of speaker embedding and native language ID were limited, as they provide little additional information when the speaker and accent remain constant.

For future works, exchanging the pair-wise frame of MaskCycleGAN-VC into a many-many conversion can be interesting. Furthermore, experiments with different types of models should also be conducted, especially focusing on experiments with different languages. Finally, trying to apply the VC approach to generate data for training a FAC model is also a work of interest as in [20].

REFERENCES

[1] Y. Tabata, C. Yin, H. Ogata, and Y. Yano, "An iPhone quiz system for learning foreign languages," in *CAR*, 2010.

[2] S. Junjie, Y. Nishihara, R. Yamanishi, and J. Fukumoto, "Analysis of Dialogues Difficulty in Anime Comparing with JLPT Listening Tests," *Procedia Computer Science*, vol. 112, 2017.

[3] S. Junjie, Y. Nishihara, and R. Yamanishi, "A System for Japanese Listening Training Support with Watching Japanese Anime Scenes," *Procedia Computer Science*, vol. 126, 2018.

[4] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting Foreign Accent Speech Without a Reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[5] E. Pellegrino and D. Vigliano, "Self imitation in prosody training: A study on Japanese learners of Italian," in *Annual Meeting of the The Association for Natural Language Processing*, 2015.

[6] S. Zhao, S. N. Koh, S. I. Yann, and K. K. Luke, "Feedback utterances for computer-adied language learning using accent reduction and voice conversion method," in *ICASSP*, 2013.

[7] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, 2009.

[8] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "MaskCycleGAN-VC: Learning Non-parallel Voice Conversion with Filling in Frames," in *ICASSP*, 2021.

[9] W.-C. Huang, B. Halpern, L. Violeta, O. Scharenborg, and T. Toda, "Towards Identity Preserving Normal to Dysarthric Voice Conversion," in *ICASSP*, 2022.

[10] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[11] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-Parallel Sequence-to-Sequence Voice Conversion with Disentangled Linguistic and Speaker Representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, 2020.

[12] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," in *EUSIPCO*, 2018.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," in *ICASSP*, 2019.

[14] M. Proszewska, G. Beringer, D. S. Trigueros, T. Merritt, A. Ezzerg, and R. Barra-Chicote, "GlowVC: Mel-spectrogram space disentangling model for language-independent text-free voice conversion," in *Interspeech*, 2022.

[15] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme," in *ICLR*, 2022.

[16] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and S. Seki, *VoiceGrad: Non-Parallel Any-to-Many Voice Conversion with Annealed Langevin Dynamics*, 2020. arXiv: 2010.02977.

[17] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Interspeech*, 2022.

[18] R. Chung and B. Mak, "Synthesizing Near Native-accented Speech for a Non-native Speaker by Imitating the Pronunciation and Prosody of a Native Speaker," in *Interspeech*, 2022.

[19] Z. Wang, W. Ge, X. Wang, *et al.*, "Accent and Speaker Disentanglement in Many-to-many Voice Conversion," in *ICSLP*, 2021.

[20] T. N. Nguyen, N.-Q. Pham, and A. Waibel, "Accent Conversion using Pre-trained Model and Synthesized Data from Voice Conversion," in *Interspeech*, 2022.

[21] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing Voice Conversion Network with Cycle Consistency Loss of Speaker Identity," in *SLT*, 2021.

[22] M. Mirza and S. Osindero, *Conditional Generative Adversarial Nets*, 2014. arXiv: 1411.1784.

[23] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *SSW*, 2004.

[24] G. Zhao, S. Sonsaat, A. Silpachai, *et al.*, "L2-ARCTIC: A Non-native English Speech Corpus," in *Interspeech*, 2018.

[25] S. Nakagawa, *UME Japanese Speech Database Read by Foreign Students (UME-JRF)*, Speech Resources Consortium, National Institute of Informatics, 2007.

[26] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NeurIPS*, 2020.

[27] C. Li and M. Wand, "Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks," in *ECCV*, 2016.

[28] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Interspeech*, 2018.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Increasing Speech Intelligibility by Mimicking Professional Announcers' Voices and Its Physical Correlates

Dung Kim Tran*, Masato Akagi*, and Masashi Unoki*
* School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
E-mail: {kimdungtran,akagi,unoki}@jaist.ac.jp

***Abstract*—Previous studies found that speech uttered by professional announcers is more intelligible than that by non-experts in noisy environments. On the basis of this finding, we developed a voice-conversion (VC) system to mimic professional announcers' voices by modifying the speaker embedding of non-expert speech. The results from our experiments to evaluate this system indicated that intelligibility increased significantly with this system. In this paper, to discuss what physical features correlate to the intelligibility, the following two issues are investigated by analyzing this system: (1) whether speech intelligibility can be changed gradually even by shifting one PCA (principal component analysis) component of the speaker embedding of the above VC system and (2) what physical features are changed when the PCA component is shifted, we retrained the VC system with a larger amount of training data. Comparing the speech intelligibility and candidate features that were changed with the shift of one axis of PCA, we found that spectral tilt, spectral plateau, and cepstral peak prominence are strongly correlated with intelligibility.**

Keywords: spectral tilt, spectral plateau, cepstral peak prominence, PCA, STOI, voice conversion

I. Introduction

Speech intelligibility is important in providing general information such as the arrival/departure time of trains and flights; driving instructions, nearby gas stations, and parking lots. It is especially important in announcing emergency situations such as fires, earthquakes, and landslides; as well as guiding people to safety. Making speech more intelligible is specifically challenging in noisy and reverberant environments such as airport/train stations, kitchens, and shower rooms. It is even more difficult when listeners are elderly with or without hearing impairments.

The easiest and most natural response to the challenges is turning up the volume of playback devices; however, increasing the volume is not equivalent to increasing the intelligibility of speech. A person can involuntarily increase the intelligibility of his/her speech in noisy places, known as the Lombard effect [1]. But increasing the volume of playback devices does not have such effect. Various approaches have been proposed to increase speech intelligibility in noisy environments without increasing the total power of speech. These include modification of spectral properties [2]–[4], dynamic range compression [5]–[7], modification of speech modulation spectrum [8], [9], and time-scale modification [10], [11].

Several studies have found that voice-related professions, e.g., professional announcers, voice actors, and singers, can produce speech that is clearer and easier to hear than non-professional people [12], [13]. Other studies have shown that the speech by professional announcers can maintain intelligibility better than that by non-experts in very noisy environments [14]. Inspired by this phenomenon, a previous study [15] developed an end-to-end deep neural network (DNN) based voice conversion (VC) system that mimicked the speaking style of professional announcers to increase speech intelligibility of non-expert people.

As reported in the previous study [15], by analyzing the principal component analysis (PCA) of speaker embedding in the proposed VC system, the first PCA component was found to be related to gender difference and the second PCA component of speaker embedding captured the difference between non-expert speakers and professional announcers. This is advantageous as a non-expert voice can be converted to have the voice style of a professional announcer only by changing the second PCA component of a speaker embedding. Since professional announcers' voices are more intelligible even in a noisy environment, it is expected that the second PCA component can be used to increase the intelligibility of non-expert voices. To clarify this point, the authors [15] proposed replacing the second PCA component of a non-expert speaker embeddings with the average value calculated from the second principal component of all male professional announcers. The obtained speaker embedding was then used to synthesize converted stimuli.

Experimental results [15] from objective measurements and subjective evaluation of the converted stimuli confirmed that adapting to an announcer's voice can increase the intelligibility of a non-expert speaker's voice. By modifying the second PCA component of speaker embedding, we can manually control the announcer speaking style, hence increase the intelligibility of speech in a noisy environment without completely changing speaker individuality. Statistical analysis also showed that modifying the second PCA component yields the highest performance. These results suggest that the second PCA com-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

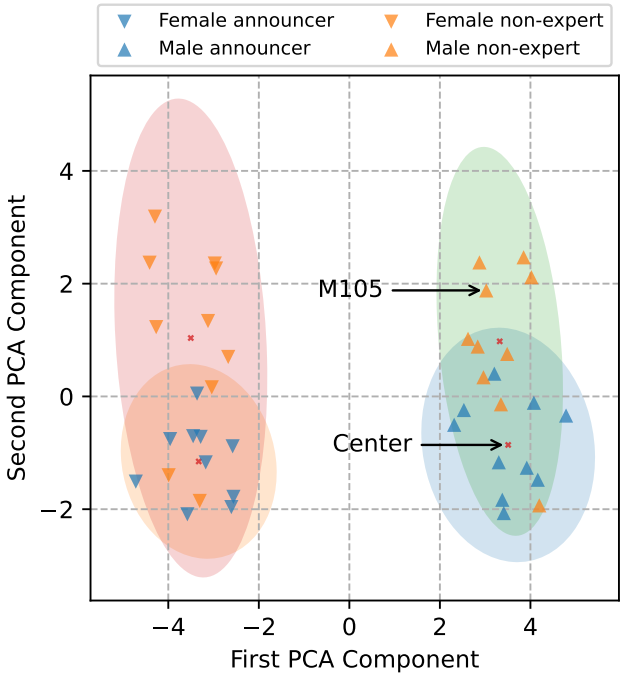


Fig. 1. Speaker embeddings in PCA domain

ponent is related to the intelligibility of speech.

However, the DNN-based VC system is a black box; it is unknown what happens that improves the intelligibility of speech when the second PCA component is swapped. It is also unclear whether speech intelligibility can be changed gradually by shifting only the second PCA component. This paper analyzes speaker embeddings in the system [15] that is thought to be a black box in a DNN-based VC for improving speech intelligibility and discusses physical cues for improving intelligibility. More specifically, we retrain the above VC system with a larger amount of training data and use it to control the second PCA component. We then study the subsequent effects on physical properties and intelligibility of speech and present two discussion points:

- Whether speech intelligibility can be changed gradually even by shifting one PCA component of the speaker embedding.
- What physical features are changed when the PCA component is shifted.

Possible candidates of physical correlates are spectral tilt and plateau and cepstral peak prominence (CPP).

II. Voice-conversion system

A. Training voice-conversion system

StarGANv2-VC [16] is one of the most effective VC systems. On the basis of this system, the authors [15] proposed a method for converting the voice of a non-expert speaker to that of a professional announcer to increase speech intelligibility. Their goal was to increase the effectiveness of StarGANv2-VC by adding automatic speech recognition to the discriminator module. They also used a Parallel WaveGAN [17] vocoder to generate converted speech utterances from the corresponding converted mel-spectrograms.

We retrained the above VC system with a larger amount of training data to discuss the properties of the PCA second component of the speaker embedding related to expert versus non-expert speakers. More specifically, the training data consist of utterances from 20 professional announcers from the ATR dataset A-set and 20 non-expert speakers from the ATR dataset C-set [18]. All the utterances are pre-processed by resampling to 24 kHz, removing leading and trailing silence; and they are combined to 5-second chunks. There are total 29,956 utterances, in which 500 utterances are used for validation. The 80-band log-mel spectrogram with band-limited frequency range (0 to 8 kHz) is extracted using short-time Fourier transform. The window length and frame shift are set to 1024 and 256, respectively.

We follow the training strategy described in the above studies [15], [16] with the same objective functions and hyper-parameters. We then used the style encoder of our retrained system to generate the PCA of the speaker embeddings for non-experts and professional announcers to verify that the second PCA component is related to expert and non-expert speakers.

Figure 1 shows the visualization of the first two PCA components of the speaker embedding. It can be seen that the first PCA component is related to gender with female and male speakers distributed on the left and right. The second PCA component is related to expert and non-expert speakers. The clusters of female/male professional announcers are compact and those of non-expert speakers are wide on the second PCA component. This phenomenon suggests that the professional announcers can control the configurations of their voice organs and high intelligible speech is the result of such effort, resulting that the variance of their voices are smaller than those of non-expert speakers.

B. Shifting second PCA component

We referred to the second PCA component of a non-expert speaker (M105 in the ATR dataset C-set) as the source and average value of the second PCA component of ten professional male announcers in the ATR dataset A-set as the Center. The PCA components of M105 and the Center are visualized in Fig. 1. Let d be the distance between the PCA values (the second PCA component) of the source and Center. We calculated three checkpoints, i.e., C1 $= 0.25d$, C2 $= 0.5d$ and C3 $= 0.75d$ from the source, respectively. For example, assuming that the second PCA components of the source and Center were vectors: $\mathbf{a}$ and $\mathbf{b}$; the checkpoints were calculated as: C1 $= 0.25d = (1-0.25)\times\mathbf{a}+0.25\times\mathbf{b}$, C2 $= 0.5d = (1-0.5)\times\mathbf{a}+0.5\times\mathbf{b}$, and C3 $= 0.75d = (1-0.75)\times\mathbf{a}+0.75\times\mathbf{b}$.

With our retrained voice-conversion system, we first calculated the PCA components for a non-expert speaker and all ten male professional announcers. We then calculated the Center by taking the average of the second PCA components corresponding to the professional announcers and calculated

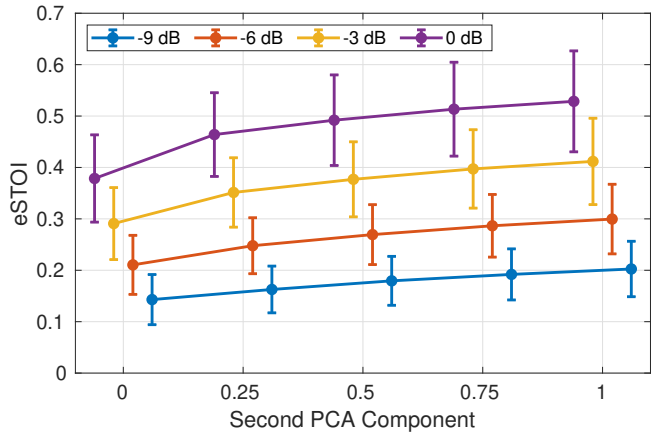


Fig. 2. eSTOI scores at four levels of noise (-9, -6, -3, 0 dB). On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

the distance between the second PCA components of the non-expert speaker and the Center as d. We then used d to calculate the PCA values for checkpoints C1, C2, and C3 accordingly. Finally, in turns, we replaced the second PCA component of the non-expert speaker embedding with the PCA values of C1, C2, C3, and Center and used our retrained voice-conversion system to produce the converted speech stimuli.

III. Second PCA Component and Intelligibility in Noise

The first discussion point is investigated in this section. We conducted an experiment to gain insights about the correlation between the second PCA component and speech intelligibility. The consistency of the correlation was also examined in various levels of noise.

We selected 520 Japanese words, each containing 1 to 4 morae, from the ATR dataset A-set and ATR dataset C-set as the clean stimuli for conversion-target and source speakers. All speech waveforms were pre-processed to a 16-kHz sampling rate with a single channel. We followed the instructions in [15] and used 16 kHz here instead of 24 kHz as described in Section II. There were 6 types of speech stimuli in the experiments: S_N is the natural speech of a non-professional speaker, collected from speaker M105 in the C-set; S_A is the natural speech of professional announcers, collected from 20 male and female speakers in the A-set; and S_1, S_2, S_3, and S_C are converted speech produced by shifting the second PCA components of speaker embedding of the source (M105) to checkpoint C1, C2, C3, and Center, respectively.

To create the noisy stimuli, we masked the clean stimuli with pink noise at four different signal-to-noise-ratio (SNR) levels: -9, -6, -3, and 0 dB. We calculated the root-mean-square of the speech signal only in the speech-presence region and scaled the noise signal to match the desired SNR level. The speech-presence region is derived from the text labels of the ATR dataset. To avoid the effect of different onset and offset timing between speech stimuli, the duration of each stimulus was adjusted to contain the same 200 ms of leading noise and 200 ms of trailing noise. Speech stimuli were gated with two raised cosine onset and offset windows of 40 ms to avoid overshoot distortion.

To objectively measure the intelligibility of speech in noise and determine whether speech intelligibility is increased gradually by shifting values along the second PCA component, we calculate the extended Short Term Objective Intelligibility (eSTOI) of the speech stimuli at 4 SNR levels: -9, -6, -3 and 0 dB at 5 checkpoints (Source, C1, C2, C3, and Center) using pySTOI python package [19]. The clean speech by speaker M105 were used as the reference signals at checkpoint 0 and the clean-converted speech were used as the reference signals at other checkpoints for eSTOI calculation.

Figure 2 shows the experimental results. The vertical lines at each checkpoint correspond to the standard deviations of eSTOI scores. It appears that intelligibility was increased gradually when the second PCA component was shifted from source to checkpoints C1, C2, C3, and Center. The eSTOI scores were also consistent with various levels of noise. The results provide an important insight in that the intelligibility could be tuned adaptively to various levels of noise on the basis of the voice of an arbitrary person by shifting the second PCA component using our retrained system.

IV. Intelligibility and its Physical Correlates

A. Spectral tilt, spectral plateau, and intelligibility

Studies on laryngeal anomalies have found that breathiness is a prominent feature used for determining various pathological conditions [20]. Breathiness is a phenomenon caused by the incomplete closure of vocal fold. As a result, breathy voices are more difficult to hear (less intelligible) in noisy environments. Compared with normally phonated signals, breathy glottal signals have steeper downward spectral slopes, less energy in the high-frequency region, and are less periodic [21]. Several acoustical studies have reported correlations between spectral tilt (downward spectral slope), spectral plateau (energy in the high-frequency region); and breathiness ratings [22]. Based on these knowledge, we hypothesize that a speech signal that has a flatter spectral tilt and larger plateau energy could have higher intelligibility and vice versa.

Figure 3 (a) shows an example of the spectral tilt we calculated for this study. It is the frequency spectrum of the vowel /a/ with the spectral tilt presented as the red line. We used the label of the ATR dataset to extract a segment of vowel /a/. We then calculated the Fourier spectrum of the segment. Finally, we calculated the regression line of the spectrum between 300 and 4,000 Hz. We used the slope of the regression line as the spectral tilt. Figure 3 (b) shows an example of the spectral plateau. On the same spectrum shown in Fig. 3 (a), the spectral plateau is presented as the red line between 2,000 and 4,000 Hz. We subtracted the entire spectrum from its peak magnitude. We then calculated the average magnitude of the spectrum in the 2,000-4,000-Hz band. We used this average magnitude to represent the spectral plateau and compared it with the plateaux of other spectra.

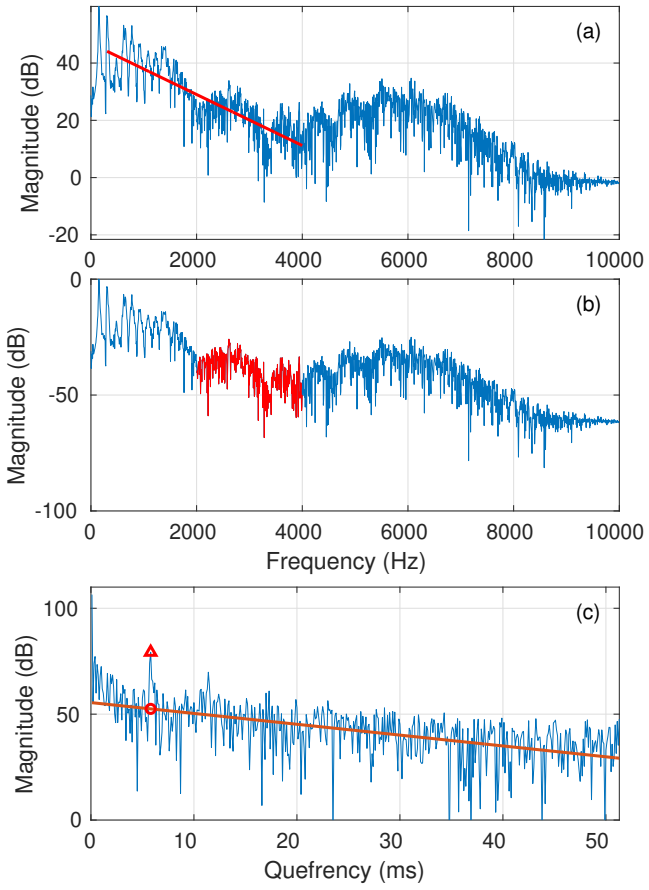


Fig. 3. Frequency spectrum and quefrency cepstrum of vowel /a/

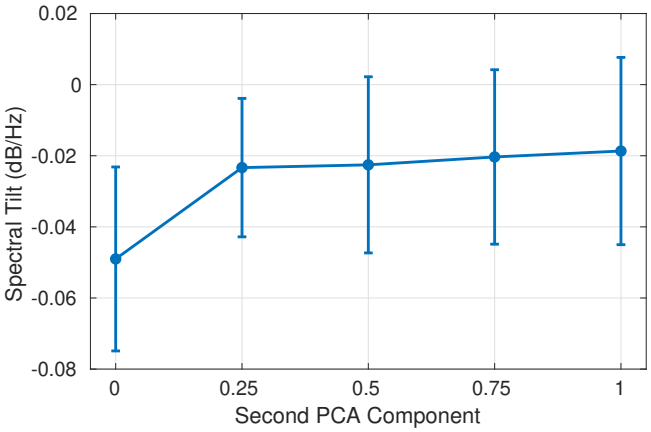


Fig. 4. Average spectral tilt and standard deviation of vowels /a/, /e/, /i/, /o/, /u/. On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

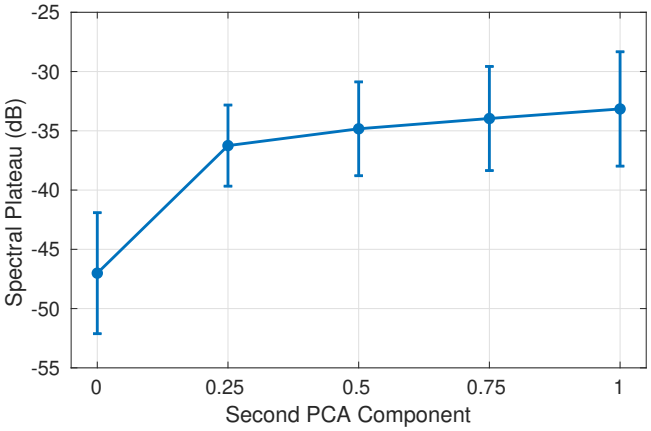


Fig. 5. Average spectral plateau and standard deviation of vowels /a/, /e/, /i/, /o/, /u/. On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

B. Periodicity, CPP, and intelligibility

A breathy voice is less periodic compared with a normal one and the quefrency domain is effective for analyzing signal periodicity. A crude explanation of the quefrency domain is that we treat the spectrum of a speech waveform as a signal and apply Fourier transform to it; the result is the cepstrum of the speech waveform in the quefrency domain. It can be observed that a highly periodic speech spectrum should result in clear harmonics in the quefrency domain; therefore, clearer cepstral peaks.

Several analyses have found a strong relation between the periodicity of a signal and its CPP [20], [23]. A CPP is the distance between the most prominent peak and the level of the cepstral background noise on the regression line immediately below the peak. Experimental results [20]–[23] indicated that the CPP of a breathy voice is significantly lower than that of a normal voice. In other words, a higher CPP level is an indicator of higher speech intelligibility. Figure 3 (c) shows an example of the CPP in the quefrency domain. It is the cepstrum of the vowel /a/ with its regression line and CPP shown as the red line and red triangle, respectively. We use the regression line of the cepstrum as the referenced background noise. The CPP is calculated as the difference in magnitude of the highest peak (the red triangle) and the red circle directly under it on the regression line.

C. Second PCA component, physical correlates, and intelligibility

Thus far, we have learned from the literature that a shallower spectral tilt, larger amount of energy of the spectral plateau, and higher CPP are associated with higher speech intelligibility. The next step is examining the correlation between the second PCA component and acoustical properties. We shifted the second PCA component, as described in Section II-B, and used our retrained voice conversion system to generate five types of stimuli. We then used the ATR dataset label to extract the vowel /a/, /e/, /i/, /o/, /u/ segments of five types of stimuli (S_N, S_1, S_2, S_3, and S_C). Finally, we used these segments to calculate spectral tilt, spectral plateau, and CPP.

The results of this experiment are shown in Figs. 4, 5, and 6. The vertical lines at each checkpoint correspond to the standard deviations of the averaged spectral tilts, spectral plateaux, and

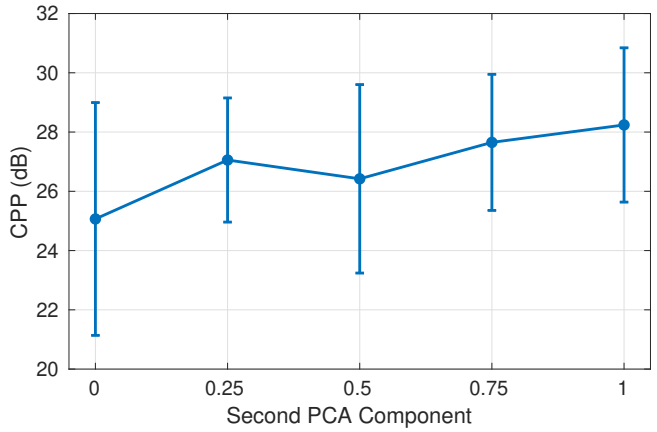


Fig. 6. Average cepstral peak prominence and standard deviation of /a/, /e/, /i/, /o/, /u/. On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

CPPs. It is known that a breathy voice (low intelligibility) has a steeper spectral slope (steeper tilt) and less energy in the high-frequency region (small or no plateau). It can be seen in Fig. 4 that the averaged slopes of the spectral tilts are reduced gradually from about -0.05 dB/Hz at checkpoint 0 to above -0.02 dB/Hz at the Center. Similarly, it appears in Fig. 5 that the average energy in the high-frequency regions (spectral plateaux) are rised from about -47 dB at checkpoint 0 to about -33 dB at the Center. Observing the experimental results visualized in Figs. 4, 5, and 2, there is a relation between the spectral tilt, spectral plateau, and intelligibility, i.e., the improvement of these physical properties is consistent with the improvement of speech intelligibility. More specifically, the non-expert speaker has the steepest spectral tilt and smallest spectral plateau, suggesting that the non-expert speaker has the lowest speech intelligibility. In contrast, the Center appears to have the shallowest spectral tilt and largest spectral plateau; therefore, it has the highest intelligibility. Between the non-expert speaker and Center, checkpoints C1, C2, and C3 have a small fluctuation; however, the overall trend is that the spectral tilts become shallower and spectral plateaux larger, indicating that the intelligibility becomes higher when the second PCA component is shifted from the non-expert speaker to checkpoints C1, C2, C3, and finally the Center.

It has been reported that the spectrum of a breathy voice (low intelligibility) is less periodic, and a less periodic spectrum results in a less prominent cepstral peak. It can be seen in Fig. 6 that the averaged cepstral peaks have become more prominent, rising from about 25 dB at checkpoint 0 to above 28 dB at the Center. Observing the experimental results visualized in Figs. 6 and 2, there is a relation between the CPP and intelligibility, i.e., the improvement of the CPP is consistent with the improvement of speech intelligibility. More specifically, the non-expert speaker has the lowest CPP; in other words, the speaker has the least speech intelligibility. The CPP values of other checkpoints are all larger than that of the non-expert. Although there is a small fluctuation, it appears that the CPP values becomes larger, indicating that the intelligibility becomes higher when the second PCA component is shifted from the non-expert speaker toward the Center.

The experimental results visualized in Fig. 2 show that speech intelligibility is increased gradually when the second PCA component is shifted gradually. The shift also produces enhancements of spectral tilt, spectral plateau, and CPP as shown in Figs. 4, 5, and 6, respectively. These results provide evidence that there is a correlation between the improvement of intelligibility and enhancement of physical properties of speech.

V. Conclusion

We presented two points of discussion for finding the correlations between the second PCA component, physical properties, and intelligibility of speech in order to learn why speech intelligibility can be increased by mimicking professional anouncers' voices using the end-to-end DNN-based VC system [15]. The first point is whether speech intelligibility can be changed gradually even by shifting one second PCA component of the speaker embedding and the second point is what physical features are changed when the second PCA component is shifted. To do so, we retrained the VC system [15] with a larger amount of training data. We then used our retrained system to shift the second PCA component of a non-expert speaker to those of checkpoints C1, C2, C3, and the Center. We used the eSTOI to evaluate the intelligibility of speech stimuli with four levels of noise (-9, -6, -3, and 0 dB) at five checkpoints. Experimental results confirm for the first point of discussion that speech intelligibility can be increased gradually by shifting the second principal component toward those of professional announcers.

We investigated why it is possible to increase speech intelligibility by using the second PCA component of professional announcers. We looked into the spectral tilt and spectral plateau in the frequency domain, as well as the CPP in the quefrency domain. The results confirm for the second point of discussion that when the second PCA component is shifted from a non-expert speaker through checkpoints C1, C2, C3, and Center, all the acoustical properties related to intelligibility appear to be enhanced proportionally.

VI. Acknowledgements

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (Grant number: 201605002) and Grant-in-Aid for Scientific Research (Grant number: 20H04207).

References

[1] H. Brumm and S. A. Zollinger, "The evolution of the Lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11/13, pp. 1173–1198, 2011.

[2] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, 2015. DOI: 10.1109/MSP.2014.2365594.

[3] C. H. Taal and J. Jensen, "SII-based speech preprocessing for intelligibility improvement in noise," in *Proc. Interspeech 2013*, 2013, pp. 3582–3586. DOI: 10.21437/Interspeech.2013-770.

[4] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech and Language*, vol. 28, no. 4, pp. 858–872, 2014, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2013.11.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230813001101.

[5] T.-C. Zorilă and Y. Stylianou, "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," in *Proc. Interspeech 2014*, 2014, pp. 2050–2054. DOI: 10.21437/Interspeech.2014-466.

[6] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech 2012*, 2012, pp. 635–638. DOI: 10.21437/Interspeech.2012-197.

[7] C. Chermaz and S. King, "A Sound Engineering Approach to Near End Listening Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 1356–1360. DOI: 10.21437/Interspeech.2020-2748.

[8] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45, no. 2, pp. 101–113, 2005, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2004.06.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639304000597.

[9] T. Ngo, R. Kubo, and M. Akagi, "Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function," *Speech Communication*, vol. 135, pp. 11–24, 2021, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2021.09.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763932100100X.

[10] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech 2011*, 2011, pp. 345–348. DOI: 10.21437/Interspeech.2011-126.

[11] V. Aubanel and M. Cooke, "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. Interspeech 2013*, 2013, pp. 3592–3596. DOI: 10.21437/Interspeech.2013-772.

[12] H. Noh and D.-H. Lee, "How does speaking clearly influence acoustic measures? a speech clarity study using long-term average speech spectra in korean language," *Clinical and Experimental Otorhinolaryngology*, vol. 5, pp. 68–73, 2012.

[13] C. Kashimada, K. Ogita, T. Ishikawa, H. Hasegawa, and M. Ayama, "Effects of voice training on subjective evaluation of voice quality," *The Journal of The Institute of Image Information and Television Engineers*, vol. 63, no. 12, pp. 1818–1823, 2009. DOI: 10.3169/itej.63.1818.

[14] M. Kobayashi and M. Akagi, "Intelligibility of announcer's speech in noisy environments," *IEICE Technical Report*, vol. 119, pp. 95–99, 2020.

[15] T. Vu Ho, M. Kobayashi, and M. Akagi, "Speak Like a Professional: Increasing Speech Intelligibility by Mimicking Professional Announcer Voice with Voice Conversion," in *Proc. Interspeech 2022*, 2022, pp. 171–175. DOI: 10.21437/Interspeech.2022-124.

[16] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1349–1353. DOI: 10.21437/Interspeech.2021-319.

[17] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203. DOI: 10.1109/ICASSP40776.2020.9053795.

[18] A. T. R. I. International, *Digital voice database*, http://www.atr-p.com/.

[19] P. Manuel, *Python implementation of STOI*, https://github.com/mpariente/pystoi.

[20] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994. DOI: 10.1044/jshr.3704.769.

[21] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 1064–1077, 1999. DOI: 10.1121/1.427116.

[22] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 6, pp. 1212–1223, 1995. DOI: 10.1044/jshr.3806.1212.

[23] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 2014, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2014.07.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809414000986.

Robust Networked Federated Learning for Localization

Reza Mirzaeifard*, Naveen K. D. Venkategowda§, Stefan Werner*
*Dept. of Electronic Systems, Norwegian University of Science and Technology-NTNU, Norway
§Department of Science and Technology, Linköping University, Sweden
E-mails: {reza.mirzaeifard, stefan.werner}@ntnu.no, naveen.venkategowda@liu.se

***Abstract*—This paper addresses the problem of localization, which is inherently non-convex and non-smooth in a federated setting where the data is distributed across a multitude of devices. Due to the decentralized nature of federated environments, distributed learning becomes essential for scalability and adaptability. Moreover, these environments are often plagued by outlier data, which presents substantial challenges to conventional methods, particularly in maintaining estimation accuracy and ensuring algorithm convergence. To mitigate these challenges, we propose a method that adopts an L_1-norm robust formulation within a distributed sub-gradient framework, explicitly designed to handle these obstacles. Our approach addresses the problem in its original form, without resorting to iterative simplifications or approximations, resulting in enhanced computational efficiency and improved estimation accuracy. We demonstrate that our method converges to a stationary point, highlighting its effectiveness and reliability. Through numerical simulations, we confirm the superior performance of our approach, notably in outlier-rich environments, which surpasses existing state-of-the-art localization methods.**

***Index Terms*—Federated learning, Robust learning, distributed learning, localization, non-convex and non-smooth optimization**

I. Introduction

The increasing prevalence of edge devices in the era of the internet-of-things (IoT) and cyber-physical systems has resulted in a significant upsurge in available data. These devices, functioning within a federated network, present unique challenges. In particular, traditional machine learning methodologies often require a central server for data processing, raising privacy concerns [1]. In response to these concerns, federated learning (FL) has emerged as an innovative solution [2]–[4] for edge devices to collectively train a global model using their locally stored data, eliminating the need for direct data sharing [2], [4]. However, federated settings are frequently besieged by outlier data, often exhibiting heavy-tailed distributions [3], [5]. These outliers can considerably distort the accuracy of learning outcomes. Nevertheless, developing robust and computationally efficient solutions for managing outlier data in FL remains a critical research question, which this paper aims to address in the context of localization.

Localization is the process of estimating the location of an object or an event within a given environment. This problem has given rise to a broad range of applications in vehicle localization [6], aviation [7], healthcare [8], environmental [9], and industrial fields [10]. Based on the type of measurements used to estimate position, the localization problem has different formulations which include time-of-arrival (ToA) [11], received-signal-strength (RSS) [12], time-difference-of-arrival (TDoA) [13], angle-of-arrival (AoA) [14], and frequency-difference-of-arrival (FDoA) [15]. In IoT applications, ToA and RSS are particularly effective as they provide relatively simple and cost-effective ways to estimate distances using the existing communication infrastructure. Therefore, in this paper, we focus on distance-based localization that inherently encompasses ToA and RSS techniques, as they can be interpreted as distance measurements [16]. This approach finds relevance across a spectrum of practical scenarios and continues to be an area of research interest.

Conventionally, various methods have been proposed for localization problems, including the parallel projection method [17], projection-onto-convex-sets method [18], nearest local minimum [19], boundary-of-convex-sets [20], recursive weighted least-squares algorithms [21], augmented Lagrangian-based method for localization [16], and iterative re-weighted least-squares [22]. However, these methods have limitations, such as being centralized or sequentially updating estimates, which hinders parallelization and flexibility. More recently, a parallel distributed alternating projection algorithm (DAPA) has been proposed for the distributed case, formulating the localization problem as a ring intersection problem [23]. Despite the advancements of DAPA, challenges remain, particularly in handling three-dimensional localization problems and targets lying outside the sensor's convex hull [24]. To address these limitations, an alternative approach called EL-ADMM has emerged. EL-ADMM directly solves the non-convex and non-smooth event localization problem using the Alternating Direction Method of Multipliers [24]. By bypassing the need for high-computation convex relaxation techniques, EL-ADMM shows promise in handling three-dimensional environments [24]. However, EL-ADMM still faces challenges in outlier handling and lacks comprehensive convergence proof. Given these constraints, there is a compelling research imperative to develop a robust, distributed algorithm capable of efficiently handling three-dimensional localization problems with outliers.

In this paper, we address a challenging class of robust localization problems that require tackling non-convex and non-smooth optimization within a federated setting. The decentralization, energy constraints, and the presence of outliers further

This work was supported by the Research Council of Norway.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

exacerbate the difficulty of accurately estimating locations in such settings. To overcome these challenges, we propose a novel distributed sub-gradient-based algorithm specifically designed to address this problem. Our contributions can be summarized as follows:

- **Efficient Optimization Algorithm**: We introduce a novel optimization algorithm that operates within a single loop framework. This algorithm directly solves the localization problem in its original form, utilizing simple updating steps. The use of this approach leads to increased accuracy and eliminates the need for iterative approximation processes.
- **Theoretical Analysis**: We provide comprehensive theoretical insights into the convergence behavior of our algorithm. Through rigorous mathematical analysis, we elucidate the conditions under which our method is guaranteed to converge, thereby offering deeper insights into its reliability and robustness.
- **Empirical Validation**: We substantiate our theoretical results through extensive numerical simulations. Notably, our algorithm exhibits exceptional resilience in the presence of heavy-tailed noise and outliers. Furthermore, a comparative study with the state-of-the-art EL-ADMM algorithm highlights the superior accuracy and comparable fast convergence rate of our approach.

Mathematical Notations: Scalars are denoted by lowercase letters, column vectors by bold lowercase letters, and matrices by bold uppercase letters. The transpose of a matrix is signified by $(\cdot)^{\mathrm{T}}$. The jth column of a matrix $\mathbf{A}$ is represented as $\mathbf{a}j$. The element in the ith row and jth column of $\mathbf{A}$ is represented as a_{ij}. The sub-gradient of a function $f(\cdot)$ at a given point u is signified by $\partial f(u)$.

II. Preliminaries

We consider the problem of localization with L sensors, where the location of the i-th sensor is denoted by $\mathbf{a}_i$, for $i = 1, 2, \ldots, L$, and $A = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_L\}$ represents the complete set of sensors' locations. The unknown source is located at $\mathbf{x} \in \mathbb{R}^n$, which must be estimated using the range information between the sensors and the source. Let $d_i > 0$, $i = 1, 2, \ldots, L$, represents a noisy observation representing the range between source location $\mathbf{x}$ and i-th sensor location $\mathbf{a}_i$. The range d_i can be expressed as

$$d_i = \|\mathbf{x} - \mathbf{a}_i\| + \epsilon_i, \quad i = 1, 2, \ldots, L, \tag{1}$$

where ϵ_i denotes the noise associated with each range observation.

Additionally, the network of L sensors is modeled as an unweighted undirected graph $\mathcal{G}$ consisting of vertices $\mathcal{V} = 1, \cdots, L$ and bidirectional communication links represented by the edge set $\mathcal{E}$. Each sensor $i \in \mathcal{V}$ can communicate with those in its neighborhood $\mathcal{N}_i$ of cardinality $|\mathcal{N}_i|$. We assume that graph $\mathcal{G}$ is strongly connected.

The maximum-likelihood estimate for the source location $\mathbf{x}$ with additive independent and identically distributed Gaussian noise affecting range measurements can be determined as the solution of the optimization problem [16]:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{L} \sum_{i=1}^{L} \left(\|\mathbf{x} - \mathbf{a}_i\| - d_i\right)^2. \tag{2}$$

However, this approach encounters challenges when faced with outlier measurements that do not conform to a Gaussian distribution. These outliers can drastically bias location estimate due to the magnified impact of their squares in each term. One could use conventional filtering techniques to reduce this effect. However, such filtering methods often presuppose knowledge about the noise characteristics, which is not always available or accurate. Additionally, filtering can induce a 'masking effect', in which crucial signal components, particularly those smaller ones sharing frequency characteristics with the noise, are suppressed or obscured [22]. Such an outcome is undesirable in our context as it may lead to the loss of vital information [22]. To counteract this, robust estimation theory offers a way of softly rejecting outliers by using the L_1 loss [25] as given by

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{L} \sum_{i=1}^{L} \left|\|\mathbf{x} - \mathbf{a}_i\| - d_i\right| = \arg\min_{\mathbf{x}} \sum_{i=1}^{L} f_i(\mathbf{x}), \tag{3}$$

where the local objective function $f_i(\mathbf{x})$ at sensor i is defined as $f_i(\mathbf{x}) = \frac{1}{L}\left|\|\mathbf{x} - \mathbf{a}_i\| - d_i\right|$. The L_1 objective function diminishes the impact of outliers by minimizing the average absolute discrepancy between the Euclidean distance from $\mathbf{x}$ to each sensor $\mathbf{a}_i$ and the corresponding observed range d_i. This approach is particularly beneficial as it ensures that smaller yet crucial variations in the data are preserved, and not overshadowed by outliers.

Even though the L_1 robust formulation can enhance accuracy and handle outliers more efficiently, it is inherently non-smooth and non-convex and not even a difference of convex functions, which makes (3) difficult to solve. Traditionally, convex relaxation approaches have been considered to tackle this problem [22], [25]. These methods, however, can compromise accuracy and are often time-consuming and computation-intensive iterative processes. In the distributed case, the problem exacerbates as one must ensure consensus among the nodes as well as convergence to a solution. To that end, this paper proposes a method for addressing the problem directly with the L_1 loss in a distributed manner.

III. Distributed Robust Localization

The proposed algorithm for localization in a distributed setting called distributed sub-gradient method for robust localization (DSRL), updates each node's estimate of the actual parameter $\mathbf{x}^*$ through a diffusion step and a sub-gradient update step. In the diffusion step, each node uses a diffusion variable, $\mathbf{v}_i$, to incorporate information from its neighboring nodes. At each time instant k, the diffusion step takes place as follows:

$$\mathbf{v}_i^{(k)} = \mathbf{x}_i^{(k)} + \alpha^{(k)} \sum_{j \in \mathcal{N}_i} \left(\mathbf{x}_j^{(k)} - \mathbf{x}_i^{(k)}\right). \tag{4}$$

Algorithm 1: Distributed Sub-gradient Method for Robust Localization (DSRL)

Initialize $\mathbf{x}_i^{(1)}$ for each node i, the weight parameter $\{\alpha_k\}_{k=1}^K$, the step-size $\{\beta_k\}_{k=1}^K$ and the number of iterations K;
for $k = 1, \cdots, K$ **do**
  for $i = 1, \cdots, L$ **do**
    Receive $\mathbf{x}_j$ from neighbors in $\mathcal{N}_i$;
    Update $\mathbf{v}_i^{(k)}$ by (4);
    Update $\mathbf{g}_i^{(k)}$ by (6);
    Update $\mathbf{x}_i^{(k+1)}$ by (5);
  end
end

In this formula, $\mathbf{x}_i^{(k)}$ is the current estimate at node i, while $\mathbf{x}_j^{(k)}$ are the estimates at each neighboring node j. The term $\alpha^{(k)}$ is a scalar weight parameter that adjusts how strongly the estimates of the neighbors influence the node's update. The diffusion step seeks to decrease the discrepancy between each node's estimate and the average estimate of its neighbors, working as a type of approximation to the average consensus algorithm. Instead of a simple average, we apply a weighted adjustment based on the difference between a node's estimate and its neighbors' estimates, thereby aiding the nodes to reach a consensus on the parameter estimate.

In the subsequent step, termed the sub-gradient step, each node updates its estimate of the true parameter $\mathbf{x}^*$ in the manner depicted below:

$$\mathbf{x}_i^{(k+1)} = \mathbf{v}_i^{(k)} - \beta^{(k)} \mathbf{g}_i^{(k)}, \tag{5}$$

where $\beta^{(k)}$ is the step-size at iteration k and $\mathbf{g}_i^{(k)} \in \partial f_i(\mathbf{x}_i^{(k)})$ is any element of the sub-differential set of $f_i(\cdot)$. Each function $f_i(\cdot)$ is neither convex nor smooth, so the distributed sub-gradient method proposed in [26] is adopted for iterative computation of the estimate of $\mathbf{x}$ at each node. The sub-gradient of the local robust localization function $f_i(\cdot)$ with respect to the coefficient $\mathbf{x}_i^{(k)}$ is computed as follows:

$$\partial f_i(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}-\mathbf{a}_i}{L\|\mathbf{x}-\mathbf{a}_i\|} \text{sign}(\|\mathbf{x}-\mathbf{a}_i\| - d_i), & \|\mathbf{x}-\mathbf{a}_i\| > 0 \\ \mathbf{0}. & \|\mathbf{x}-\mathbf{a}_i\| = 0 \end{cases} \tag{6}$$

Algorithm 1 summarizes the proposed method for solving distributed robust localization.

IV. Convergence Proof

In this section, a convergence analysis of the Algorithm 1 is presented. We construct our convergence proof based on the following assumptions.

Assumption 1. *Consider the function $f(\cdot) = \sum_{i=1}^{L} f_i(\cdot)$ and let $CP_f \subset \mathbb{R}^n$ represent the set of its critical points. Then, the complement of CP_f in $\mathbb{R}^n$, expressed as $\mathbb{R}^n \setminus CP_f$, is a dense set in $\mathbb{R}^n$.*

Assumption 2. *The weight parameter $\alpha^{(k)}$ and the step-size $\beta^{(k)}$ are derived from $\alpha^{(k)} = \frac{a}{k^{\tau_\alpha}}$ and $\beta^{(k)} = \frac{b}{k^{\tau_\beta}}$ respectively. Here, $0 < \tau_\alpha < \tau_\beta$, $\frac{1}{2} < \tau_\beta \leq 1$, and both a and b are positive constants.*

While Assumption 1 might appear to be technical, it is in fact a relatively mild condition that guides the positioning and quantity of sensors to ensure convergence to the critical points in our algorithm. As an illustration, this assumption is invariably satisfied when employing an odd number of sensors. Having a non-critical dense set ensures that, through communication and updates, the sensors have enough room to maneuver and collaborate towards converging to a consensus that aligns with the global objective. On the other hand, Assumption 2 guarantees that information can be effectively disseminated across the network.

In order to show the regularity of the objective function we provide the following Lemmas.

Lemma 1. *each $f_i(\cdot)$ is locally Lipschitz continuous.*

Proof. Given that the sub-gradient of each $f_i(\cdot)$ is bounded, it follows that these functions are locally Lipschitz continuous. □

Lemma 2. *There exists a radius $R > 0$ and constants $C_1, C_2 > 0$ such that:*

$$\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|}, \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\rangle \geq C_1 \quad \textit{and} \quad \|\mathbf{g}\| \leq C_2 \|\mathbf{x}\|,$$

for all $\mathbf{g} \in \partial f_n(\mathbf{x})$ and $\|\mathbf{x}\| \geq R$.

Proof. We select a radius $R = (3 \max_i \max_j |a_{ij}| + \max_i d_i) + 1$. Consequently, $\text{sign}(\|\mathbf{x} - \mathbf{a}_i\| - d_i) = 1$. We proceed to examine the first condition of the lemma. It is evident that

$$\begin{aligned} \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|}, \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\rangle &= \left\langle \frac{\mathbf{x}-\mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|}, \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\rangle \\ &\geq \frac{\|\mathbf{x}\|}{\|\mathbf{x}-\mathbf{a}_i\|} - \frac{\mathbf{x}^T \mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|\|\mathbf{x}\|} \\ &\geq 1 - \frac{\|\mathbf{a}_i\|}{\|\mathbf{x}-\mathbf{a}_i\|} \geq 1 - \frac{\|\mathbf{a}_i\|}{2\|\mathbf{a}_i\|} = \frac{1}{2}. \end{aligned}$$

These inequalities are derived from the conditions that $\|\mathbf{x}\|$ always surpasses $\|\mathbf{x}-\mathbf{a}_i\|$, and that $\|\mathbf{x}\|$ always exceeds $3\|\mathbf{a}_i\|$. These conditions are a direct result of our choice of R. This effectively verifies that $\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|}, \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\rangle \geq C_1$ for $C_1 = \frac{1}{2}$. For the second condition of the lemma, we are aware that $\|\mathbf{g}\| = 1$ when $\|\mathbf{x} - \mathbf{a}_i\| > d_i$. This is always true given our selection of R. Also, taking into account that $1 \leq \|\mathbf{x}\|$, the inequality $\|\mathbf{g}\| \leq C_2 \|\mathbf{x}\|$ holds for $C_2 = 1$. □

Theorem 1. *Given that the static graph $\mathcal{G}$ is strongly connected, the objective function fulfills Assumption 1, and both the step-size and weight parameter adhere to Assumption 2, the DSRL algorithm is ensured to converge to a critical point.*

Proof. The convergence of our proposed DSRL algorithm is substantiated by verifying that all the prerequisites set forth in

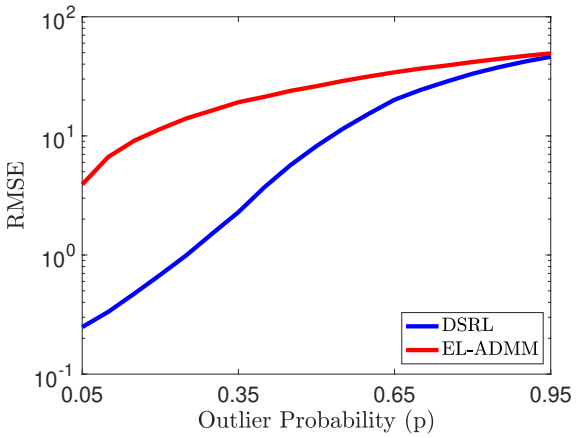


Fig. 1: Steady-state RMSE versus the outlier probability.

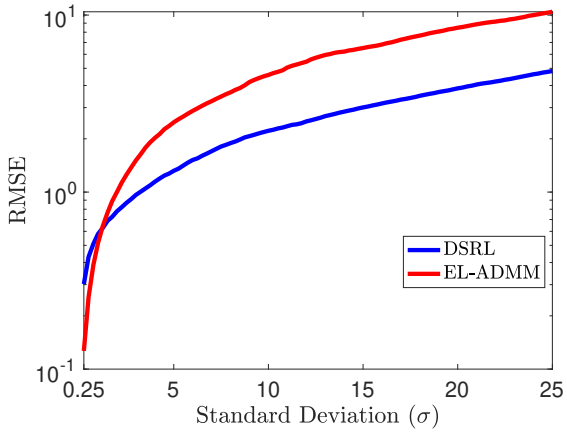


Fig. 2: Steady-state RMSE versus the standard deviation of mixture of Laplacian noise.

[26, Theorem 7], inclusive of the regularity conditions of our robust localization function, are met. The fulfillment of these conditions is demonstrated in Lemma 1 and Lemma 2. □

V. Simulation Results

In this section, we assess the performance of the proposed DSRL algorithm via simulations and draw comparisons with the distributed event localization via the Alternating Direction Method of Multipliers (EL-ADMM) algorithm, as outlined in [24]. The sensor network $\mathcal{G}$ under consideration comprises $L = 31$ nodes, which are uniformly and randomly distributed within a $[-30, 30]^3$ cubic region, and is fully connected. Connections are established between nodes if their distance is less than 25, ensuring a minimum of 2 and a maximum of 10 neighbors per node. In addition, the target is assumed to be uniformly and randomly distributed within the same cubic region. For each iteration, the weight parameter $\alpha^{(k)}$ and the step-size $\beta^{(k)}$ are set to $\frac{0.3}{k^{0.55}}$ and $\frac{3.5}{k^{0.5}}$, respectively. We employ the root-mean-square error (RMSE), denoted by $\sqrt{\frac{\sum_{i=1}^{L} ||\hat{\mathbf{w}}_i - \mathbf{x}||_2^2}{L}}$, as the performance metric. The results are obtained by averaging the outcomes of 1000 independent trials.

In the first scenario, we focus on comparing the accuracy of the algorithms when handling outliers in the measurements. We introduce outliers at varying probability, denoted as p, ranging from 0.05 to 0.95 with increments of 0.05. The outliers, represented as ϵ_i, are drawn from a uniform distribution over the interval $[0, 60\sqrt{3}]$. This is formalized as:

$$y_i = \begin{cases} \epsilon_i, & \text{if } u_i < p, \\ \|\mathbf{x} - \mathbf{a}_i\|, & \text{otherwise}, \end{cases} \tag{7}$$

where u_i is a realization of the uniform distribution on the interval [0,1]. As illustrated in Figure 1, the DSRL algorithm outperforms EL-ADMM consistently across the various frequency of outliers.

In the second scenario, we evaluate the performance of the algorithms under noisy conditions, where the noise follows

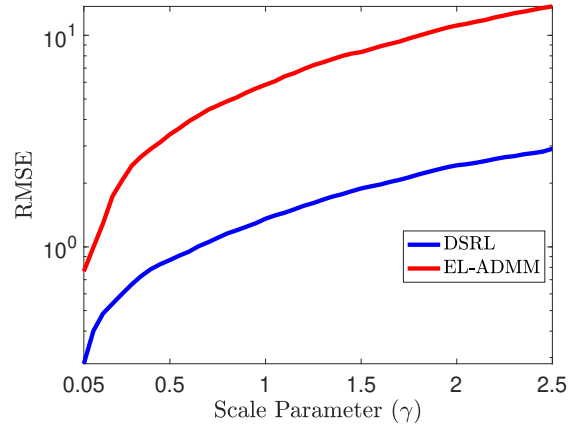


Fig. 3: Steady-state RMSE versus scale parameter of Cauchy noise.

a mixture of Laplacian distributions. The measurements are generated as follows:

$$y_i = \|\mathbf{x} - \mathbf{a}_i\| + \epsilon_i, \quad \forall i \in 1, \cdots, N, \tag{8}$$

where $\{\epsilon_i\}_{i=1}^{N}$ denotes the components of the noise, which are independently and identically sampled from a mixture of Laplacian distributions given by $\sum_{i=1}^{2} c_i \frac{\lambda_i}{2} e^{-\lambda_i |v|}$ Here, $c_1 = 0.9$, $c_2 = 0.1$, and $\lambda_2 = \frac{\lambda_1}{10}$. The parameter λ_1 is related to the standard deviation (σ) and can be calculated as $\lambda_1 = \sqrt{\frac{\sigma^2}{10.9}}$. The value of 10.9 in the denominator arises from the specific weights c_1 and c_2, and it represents the effective contribution of the two components of the Laplacian mixture in relation to the variance. In this setup, 10% of the measurements can be considered outliers. We obtain simulation results for varying standard deviation values, ranging from 0.25 to 25, with increments of 0.25. As illustrated in Figure 2, the DSRL algorithm exhibits superior performance in terms of RMSE, particularly for standard deviation values greater than 1.

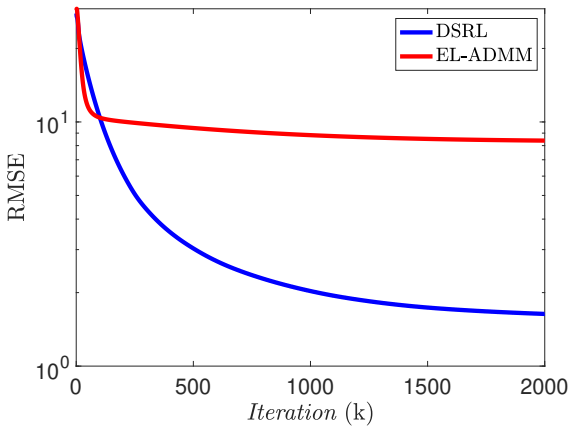


Fig. 4: RMSE versus iterations.

In the third scenario, we assess the accuracy of the algorithms in the presence of Cauchy noise. The measurements are formulated as in (8), where each component ϵ_i is independently and identically distributed according to the Cauchy distribution with the scale parameter γ, i.e., $\epsilon_i \sim \text{Cauchy}(0, \gamma)$. We conduct simulations for varying scale parameters, ranging from 0.05 to 2.5, in increments of 0.05. Figure 3 illustrates that the DSRL algorithm outperforms EL-ADMM in terms of RMSE across different scale parameters. Additionally, for the scale parameter set at 1, Figure 4 demonstrates that DSRL maintains a competitive convergence speed compared to EL-ADMM.

VI. Conclusion

This paper presents a sub-gradient algorithm devised for the distributed robust localization problem that directly tackles non-convex and non-smooth objective functions. This proposed algorithm has been rigorously proven to be convergent. Our simulation results highlight the superiority of our algorithm in terms of root-mean-squared error, compared to the existing state-of-the-art EL-ADMM algorithm. Particularly in outlier settings, the algorithm's effectiveness was consistently apparent.

References

[1] X. Yin, Y. Zhu, and J. Hu, "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–36, July 2021.

[2] X. Zhou, W. Liang, I. Kevin, K. Wang, Z. Yan, L. T. Yang, W. Wei, J. Ma, and Q. Jin, "Decentralized p2p federated learning for privacy-preserving and resilient mobile robotic systems," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 82–89, Apr. 2023.

[3] J. Zhao, X. Chang, Y. Feng, C. H. Liu, and N. Liu, "Participant selection for federated learning with heterogeneous data in intelligent transport system," *IEEE Transactions on Intelligent Transportation Systems*, Feb. 2022.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[5] V. Tsouvalas, A. Saeed, T. Ozcelebi, and N. Meratnia, "Federated learning with noisy labels," *arXiv preprint arXiv:2208.09378*, Aug. 2022.

[6] O. Kaiwartya, Y. Cao, J. Lloret, S. Kumar, N. Aslam, R. Kharel, A. H. Abdullah, and R. R. Shah, "Geometry-based localization for gps outage in vehicular cyber physical systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3800–3812, Jan. 2018.

[7] M. Strohmeier, I. Martinovic, and V. Lenders, "A k-nn-based localization approach for crowdsourced air traffic communication networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 3, pp. 1519–1529, Jan. 2018.

[8] S.-B. Han, J.-H. Kim, and H. Myung, "Landmark-based particle localization algorithm for mobile robots with a fish-eye vision system," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 6, pp. 1745–1756, Aug. 2012.

[9] Z. Zhou, Z. Peng, J.-H. Cui, Z. Shi, and A. Bagtzoglou, "Scalable localization with mobility prediction for underwater sensor networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 3, pp. 335–348, Sep. 2010.

[10] S. Petersen, P. Doyle, S. Vatland, C. S. Aasland, T. M. Andersen, and D. Sjong, "Requirements, drivers and analysis of wireless sensor network solutions for the oil & gas industry," in *2007 IEEE Conference on Emerging Technologies and Factory Automation*. IEEE, Sep. 2007, pp. 219–226.

[11] Y.-M. Pun and A. M.-C. So, "Local strong convexity of source localization and error bound for target tracking under time-of-arrival measurements," *IEEE Transactions on Signal Processing*, vol. 70, pp. 190–201, Dec. 2021.

[12] F. Yin, Y. Zhao, F. Gunnarsson, and F. Gustafsson, "Received-signal-strength threshold optimization using gaussian processes," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 2164–2177, Jan. 2017.

[13] N. Okello, F. Fletcher, D. Musicki, and B. Ristic, "Comparison of recursive algorithms for emitter localisation using tdoa measurements from a pair of uavs," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 3, pp. 1723–1732, July 2011.

[14] S. Xu and K. Doğançay, "Optimal sensor placement for 3-d angle-of-arrival target localization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 3, pp. 1196–1211, Feb. 2017.

[15] Y. Lin, Z. Zhang, and H. E. Najafabadi, "Underwater source localization using time difference of arrival and frequency difference of arrival measurements based on an improved invasive weed optimization algorithm," *IET Signal Processing*, vol. 16, no. 3, pp. 299–309, May 2022.

[16] D. R. Luke, S. Sabach, M. Teboulle, and K. Zatlawey, "A simple globally convergent algorithm for the nonsmooth nonconvex single source localization problem," *Journal of Global Optimization*, vol. 69, pp. 889–909, Dec. 2017.

[17] T. Jia and R. M. Buehrer, "A set-theoretic approach to collaborative position location for wireless networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 9, pp. 1264–1275, Dec. 2010.

[18] D. Blatt and A. O. Hero, "Energy-based sensor network source localization via projection onto convex sets," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3614–3619, Aug. 2006.

[19] Q. Shi and C. He, "Distributed source localization via projection onto the nearest local minimum," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Mar. 2008, pp. 2553–2556.

[20] J. Wang and P. A. Regalia, "Sensor network localization via boundary projections," in *2009 43rd Annual Conference on Information Sciences and Systems*. IEEE, Mar. 2009, pp. 224–229.

[21] C.-L. Wang, D.-S. Wu, S.-C. Chen, and K.-J. Yang, "A decentralized positioning scheme based on recursive weighted least squares optimization for wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4887–4893, Nov. 2014.

[22] A. Zaeemzadeh, M. Joneidi, B. Shahrasbi, and N. Rahnavard, "Robust target localization based on squared range iterative reweighted least squares," in *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, Oct. 2017, pp. 380–388.

[23] Y. Zhang, Y. Lou, Y. Hong, and L. Xie, "Distributed projection-based algorithms for source localization in wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3131–3142, Feb. 2015.

[24] C. Zhang and Y. Wang, "Sensor network event localization via non-convex nonsmooth ADMM and augmented lagrangian methods," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 4, pp. 1473–1485, Dec. 2019.

[25] P. Oguz-Ekim, J. P. Gomes, J. Xavier, and P. Oliveira, "Robust localization of nodes and time-recursive tracking in sensor networks using

noisy range measurements," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3930–3942, May 2011.

[26] B. Swenson, R. Murray, H. V. Poor, and S. Kar, "Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 14 751–14 812, Oct. 2022.

Continual Local Updates for Federated Learning with Enhanced Robustness to Link Noise

Ehsan Lari[1], Vinay Chakravarthi Gogineni[2], Reza Arablouei[3], Stefan Werner[1]
[1]Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway
[2]The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
[3]CSIRO's Data61, Pullenvale QLD 4069, Australia

***Abstract*—Communication errors caused by noisy links can negatively impact the accuracy of federated learning (FL) algorithms. To address this issue, we introduce an FL algorithm that is robust to communication errors while concurrently reducing the communication load on clients. To formulate the proposed algorithm, we consider a weighted least-squares regression problem as a motivating example. We recast this problem as a distributed optimization problem over a federated network, which employs random scheduling to enhance communication efficiency, and solve the reformulated problem via the alternating direction method of multipliers. Unlike conventional FL approaches employing random scheduling, the proposed algorithm grants the clients the ability to continually update their local model estimates even when they are not selected by the server to participate in FL. This allows for more frequent and ongoing client involvement, resulting in performance improvement and enhanced robustness to communication errors compared to when the local updates are only performed when the respective clients are selected by the server. We demonstrate the effectiveness and performance gains of the proposed algorithm through simulations.**

I. Introduction

Federated learning (FL) [1]–[6] is a distributed machine-learning approach that enables the training of models across multiple decentralized devices without transferring the data to any global server. FL allows devices to share their knowledge in terms of model estimates or gradients without revealing their raw data, thereby improving data privacy and security [7]. FL holds significant potential in a range of critical applications, including healthcare [8], finance [9], and industrial IoT [10], [11]. These data-sensitive domains demand utmost priority on data privacy, making FL an ideal approach to safeguard sensitive information while enabling effective machine learning.

The literature on FL encompasses a diverse array of methods that have been extensively investigated to address a multitude of aspects and challenges such as preserving privacy [9], [12]–[14], handling Byzantine attacks [15], [16], and improving communication efficiency [7], [17]. However, a significant number of these works operate under the assumption of ideal communication links, neglecting the presence of communication errors or noise [18]–[22]. In practical real-world applications, however, the communication links connecting clients and the server are susceptible to noise corruption, posing a potential threat to the performance of the model [23]. To tackle this challenge, several studies have introduced different techniques aimed at enhancing the accuracy of FL models when confronted with communication noise. Some works focus on the uplink noise and neglect the downlink noise [24], [25] while some other also consider the effects of the downlink noise [26].

Distributed algorithms that are based on the alternating direction method of multipliers (ADMM) can exhibit a certain degree of robustness to additive communication noise [27]. This is due to the inherent characteristics of ADMM that allow it to alleviate the impact of noise present in the communication channel. However, it requires complete collaboration among all clients, which stands in contrast to the objectives of FL, particularly those associated with system heterogeneity, such as varying computation and storage capacities across different clients. Hence, there exists a need for noise-robust ADMM-based algorithms that achieve convergence even when only a subset of clients participate in each iteration.

In this paper, we introduce a new FL algorithm that exhibits both communication efficiency and robustness to communication noise/errors. Additionally, our algorithm facilitates continual local updates at the clients, even when they are not selected by the server. This results in improved accuracy with no extra communication. We derive the proposed algorithm by using ADMM to solve a weighted least-squares (WLS) regression problem. We consider the communication links in both uplink and downlink to be noisy. To achieve noise robustness, we eliminate the dual variable update step at each client and transmit a linear combination of the last two global model updates. Our extensive simulation results corroborate the effectiveness of the proposed algorithm.

II. Background

Let us consider a federated network with K clients and a server. Each client k has an exclusive dataset denoted by $\mathcal{D}_k = \{\mathbf{H}_k, \mathbf{y}_k\}$ where $\mathbf{y}_k \in d_k$ is a column vector and $\mathbf{H}_k$ is a matrix of size $d_k \times L$. Each client independently trains a local model on its respective dataset using an FL algorithm. The process entails exchanging model updates with the global server, allowing for collaborative model training while preserving data privacy. For client k, a linear regression model relating $\mathbf{H}_k$ to $\mathbf{y}_k$ can be described as

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}^* + \boldsymbol{\nu}_k, \tag{1}$$

where $\mathbf{x}^* \in \mathbb{R}^L$ is the global regression parameter vector and $\boldsymbol{\nu}_k$ represents noise or perturbation.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

The goal of FL is to find an estimate of $\mathbf{x}^*$ that is the optimal solution to

$$\min_{\{\mathbf{x}_k\}} \sum_{k=1}^{K} \mathcal{J}_k(\mathbf{x}_k) \ \text{ s.t. } \mathbf{x}_k = \mathbf{x}, \ k = 1, 2, \cdots, K, \tag{2}$$

where $\mathcal{J}_k(\mathbf{x}) = \|\mathbf{y}_k - \mathbf{H}_k\mathbf{x}\|^2_{\mathbf{W}_k}$ is the local objective function at client k and $\mathbf{W}_k$ is the appropriate weight matrix of client k. In addition, $\mathbf{x}_k$ represents the local model estimate at client k, and $\mathbf{x}$ denotes the global model estimate.

We use an ADMM-based approach to solve (2). In this approach, primal and dual updates are computed by each client and transmitted to the global server via a noisy communication channel. Each client implements local iterations to update its local model and shares its updated local model estimate with the global server. The global server aggregates the received model estimates from the clients and updates the global model estimate. The server then sends the updated global model estimate to the clients over a noisy channel. The clients utilize the received global model estimate in updating their local estimates and the process continues until a convergence criterion is met.

To solve (2) using ADMM, we express the related augmented Lagrangian function as

$$\sum_{k=1}^{K} \mathcal{L}_k(\mathbf{x}, \mathbf{x}_k, \boldsymbol{\pi}_k) = \sum_{k=1}^{K} \mathcal{J}_k(\mathbf{x_k}) + \langle \mathbf{x}_k - \mathbf{x}, \boldsymbol{\pi}_k \rangle + \frac{\rho_k}{2} \|\mathbf{x}_k - \mathbf{x}\|_2^2, \tag{3}$$

where $\boldsymbol{\pi}_k \in \mathbb{R}^L$ and $\rho_k > 0$ are, respectively, the Lagrange multiplier vector and the penalty parameter associated with client k. Hence, the corresponding ADMM iterations at each client k are given as

$$\boldsymbol{\pi}_{k,n} = \boldsymbol{\pi}_{k,n-1} + \rho_k(\mathbf{x}_{k,n} - \mathbf{x}_n) \tag{4a}$$

$$\mathbf{x}_{k,n+1} = \hat{\mathbf{x}}_k - \mathbf{N}_k^{-1}(\boldsymbol{\pi}_{k,n} - \rho_k \mathbf{x}_n) \tag{4b}$$

and at the global server as

$$\mathbf{x}_{n+1} = \frac{1}{\sum_{k=1}^{K} \rho_k} \sum_{k=1}^{K} (\rho_k \mathbf{x}_{k,n+1} + \boldsymbol{\pi}_{k,n}) \tag{5}$$

where we define $\mathbf{N}_k = 2\mathbf{H}_k^\intercal \mathbf{W}_k \mathbf{H}_k + \rho_k \mathbf{I}$ and $\hat{\mathbf{x}}_k = 2\mathbf{N}_k^{-1}\mathbf{H}_k^\intercal \mathbf{W}_k \mathbf{y}_k$, and the index n denotes the iteration number.

In the above iterations, each client shares the local estimate $\rho_k \mathbf{x}_{k,n+1} + \boldsymbol{\pi}_{k,n}$ with the global server after computing (4a) and (4b). The global server then aggregates the received estimates to obtain the global estimate as in (5).

As it is evident from the recursions (4) and (5), both primal and dual model updates are sent to the server for it to estimate the global model update. However, by making a careful selection of the initial values, the information of the dual update can be incorporated into the primal update. Therefore, using the initial values $\mathbf{x}_{-1} = \mathbf{0}$, $\boldsymbol{\pi}_{k,-1} = \mathbf{0}$, and $\mathbf{x}_{k,0} = \hat{\mathbf{x}}_k$, we can eliminate the Lagrange multiplier vectors $\boldsymbol{\pi}_{k,n}$ and restate the recursions (4)-(5) as

$$\mathbf{x}_{k,n+1} = (\mathbf{I} - \rho_k \mathbf{N}_k^{-1})\mathbf{x}_{k,n} + \rho_k \mathbf{N}_k^{-1}(2\mathbf{x}_n - \mathbf{x}_{n-1}) \tag{6a}$$

$$\mathbf{x}_{n+1} = \frac{1}{\sum_{k=1}^{K} \rho_k} \sum_{k=1}^{K} \rho_k \mathbf{x}_{k,n+1}. \tag{6b}$$

By defining $\mathbf{s}_{k,n} = 2\mathbf{x}_{k,n} - \mathbf{x}_{k,n-1}$ and $\mathbf{s}_n = 2\mathbf{x}_n - \mathbf{x}_{n-1}$, we can further rewrite (6) as

$$\mathbf{x}_{k,n+1} = (\mathbf{I} - \rho_k \mathbf{N}_k^{-1})\mathbf{x}_{k,n} + \rho_k \mathbf{N}_k^{-1}\mathbf{s}_n \tag{7a}$$

$$\mathbf{s}_{k,n+1} = 2\mathbf{x}_{k,n+1} - \mathbf{x}_{k,n} \tag{7b}$$

$$\mathbf{s}_{n+1} = \frac{1}{\sum_{k=1}^{K} \rho_k} \sum_{k=1}^{K} \rho_k \mathbf{s}_{k,n+1}. \tag{7c}$$

In this algorithm, the clients share $\mathbf{s}_{k,n+1}$ with the server, and the server broadcasts $\mathbf{s}_{n+1}$ to the clients. When the exchange of model parameters occurs over noisy communication channels, the iterations (7) are more robust to communication noise compared to (6) as they utilize a single noisy global estimate received from the server to update the local estimate, i.e., $\mathbf{s}_n$, rather than two in (6), i.e., $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$.

In the downlink, noisy versions of the aggregated global model updates are received by the clients as $\tilde{\mathbf{s}}_n^k = \mathbf{s}_n + \boldsymbol{\zeta}_n^k$ where $\boldsymbol{\zeta}_n^k$ denotes the downlink noise of client k at iteration n. In the uplink, a noisy version of the local model update of each client is received by the server as $\tilde{\mathbf{s}}_{k,n} = \mathbf{s}_{k,n} + \boldsymbol{\eta}_{k,n}$ where $\boldsymbol{\eta}_{k,n}$ denotes the uplink noise of client k at iteration n.

In recursions (7), all clients are required to take part in each global model update. However, FL clients often have limited and diverse communication capabilities. Hence, participation of all clients in each global update round can be costly. Therefore, FL servers usually employ random scheduling of their clients by selecting only a subset of the clients, denoted by $\mathcal{S}_n$, to participate in model aggregation during each iteration n. The scheduling reduces the communications required at each iteration, leading to improved efficiency and better resource management. Utilizing random scheduling and considering noisy communications, we can rewrite the recursions (7) as mentioned in [28] as

$$\mathbf{x}_{k,n+1} = \begin{cases} \left(\mathbf{I} - \rho_k \mathbf{N}_k^{-1}\right) \mathbf{x}_{k,n} + \rho_k \mathbf{N}_k^{-1} \tilde{\mathbf{s}}_n^k, & k \in \mathcal{S}_n \\ \mathbf{x}_{k,n}, & \text{otherwise} \end{cases} \tag{8a}$$

$$\mathbf{s}_{k,n+1} = 2\mathbf{x}_{k,n+1} - \mathbf{x}_{k,n}, \quad k \in \mathcal{S}_n \tag{8b}$$

$$\mathbf{s}_{n+1} = \frac{1}{\sum_{k \in \mathcal{S}_n} \rho_k} \sum_{k \in \mathcal{S}_n} \rho_k \tilde{\mathbf{s}}_{k,n+1}. \tag{8c}$$

III. Proposed Algorithm

It is important to enhance the energy efficiency of FL by reducing its communication load, particularly considering that clients often have constraints on their energy resources in real-world scenarios. Lowering the communication burden on the clients can also improve the scalability and cost-effectiveness of

FL algorithms. The efficiency of communication during model training is related to the amount of data exchanged between the clients and the server. Thus, minimizing the exchanges of models, parameters, gradients, or other relevant information while upholding high accuracy levels presents a significant challenge in FL.

Random scheduling of clients is a practical method to enhance communication efficiency within FL, using which the overall communication load can be effectively reduced and resource utilization can be optimized more efficiently. It helps alleviate the potential bottlenecks that can arise from simultaneous communication by all clients. As a result, it improves the efficiency of data transmissions and plays a role in enhancing the overall performance of FL. However, in the conventional random scheduling approach as in (8), the clients that are not selected during each iteration do not carry out any local update and their most recent local model estimates are not incorporated into the global model aggregation process.

Here, we propose to allow all clients, including those not selected through random scheduling, to update their local model estimates during every iteration. As we will show later, this can improve the performance without introducing any additional communication overhead or imposing any significant increase in computations on the clients or the server. To realize the proposed algorithm, we let the clients store the most recent global model estimate received from the server and the server store the latest local model estimates received from the clients. Hence, the clients continually update their local models using their most recent global model estimate and the server updates the global model using the latest local updates from all clients. When a client is selected at iteration n, its latest local model estimate is made up-to-date at the server and the global model estimate received from the server replaces its older version at the client.

Therefore, the recursions of the proposed resource-efficient and noise-robust FL algorithm featuring continual local updates are given by

$$\mathbf{x}_{k,n+1} = (\mathbf{I} - \rho_k \mathbf{N}_k^{-1})\mathbf{x}_{k,n} + \rho_k \mathbf{N}_k^{-1}\left[a_{k,n}\tilde{\mathbf{s}}_n^k + (1-a_{k,n})\tilde{\mathbf{s}}_m^k\right] \tag{9a}$$

$$\mathbf{s}_{k,n+1} = 2\mathbf{x}_{k,n+1} - \mathbf{x}_{k,n}, \quad k \in \mathcal{S}_n \tag{9b}$$

$$\mathbf{s}_{n+1} = \frac{1}{\sum_{k=1}^{K} \rho_k} \sum_{k=1}^{K} \rho_k \left[a_{k,n}\tilde{\mathbf{s}}_{k,n+1} + (1-a_{k,n})\tilde{\mathbf{s}}_{k,m}\right], \tag{9c}$$

where $a_{k,n}$ represents random scheduling, i.e., $a_{k,n} = 1$ when $k \in \mathcal{S}_n$ and $a_{k,n} = 0$ otherwise. In addition, $\tilde{\mathbf{s}}_m^k$ represents the most recent global model estimate received from the server and stored in client k, which is utilized when the client is not chosen by the server. Moreover, $\tilde{\mathbf{s}}_{k,m}$ denotes the most recent local model estimate associated with client k, which is stored at the server and utilized during iterations when the client is not selected through random scheduling. We summarize the proposed algorithm in Algorithm 1.

Algorithm 1 The proposed communication-efficient and noise-robust FL algorithm with continual local updates.

Parameters: penalty parameters ρ_k, number of clients K, set of clients $\mathcal{S}$

Initialization: global model $\mathbf{x}_0 = \mathbf{x}_{-1} = \mathbf{0}$, local models $\mathbf{x}_{k,0} = \hat{\mathbf{x}}_k$

For $n = 1, \cdots, N$

The server randomly selects a subset $\mathcal{S}_n$ of its clients and sends the aggregated global model $\mathbf{s}_n$ to them.

Client Local Update:

If $k \in \mathcal{S}_n$

Receive $\tilde{\mathbf{s}}_n^k$, a noisy version of $\mathbf{s}_n$, from the server.
Store the latest global model $\tilde{\mathbf{s}}_m^k = \tilde{\mathbf{s}}_n^k$.
Update the local model as
$\mathbf{x}_{k,n+1} = (\mathbf{I} - \rho_k \mathbf{N}_k^{-1})\mathbf{x}_{k,n} + \rho_k \mathbf{N}_k^{-1}\tilde{\mathbf{s}}_n^k$.
Send $\mathbf{s}_{k,n+1} = 2\mathbf{x}_{k,n+1} - \mathbf{x}_{k,n}$ to the server.

Else

Update the local model as
$\mathbf{x}_{k,n+1} = (\mathbf{I} - \rho_k \mathbf{N}_k^{-1})\mathbf{x}_{k,n} + \rho_k \mathbf{N}_k^{-1}\tilde{\mathbf{s}}_m^k$.

EndIf

Aggregation at the Server:

The server receives $\tilde{\mathbf{s}}_{k,n+1}$, noisy versions of the locally updated models from the selected clients $k \in \mathcal{S}_n$ and aggregates them with $\tilde{\mathbf{s}}_{k,m}$, the stored local model estimates of the non-selected clients via

$$\mathbf{s}_{n+1} = \frac{1}{\sum_{k=1}^{K} \rho_k} \sum_{k=1}^{K} \rho_k \left[a_{k,n}\tilde{\mathbf{s}}_{k,n+1} + (1-a_{k,n})\tilde{\mathbf{s}}_{k,m}\right].$$

EndFor

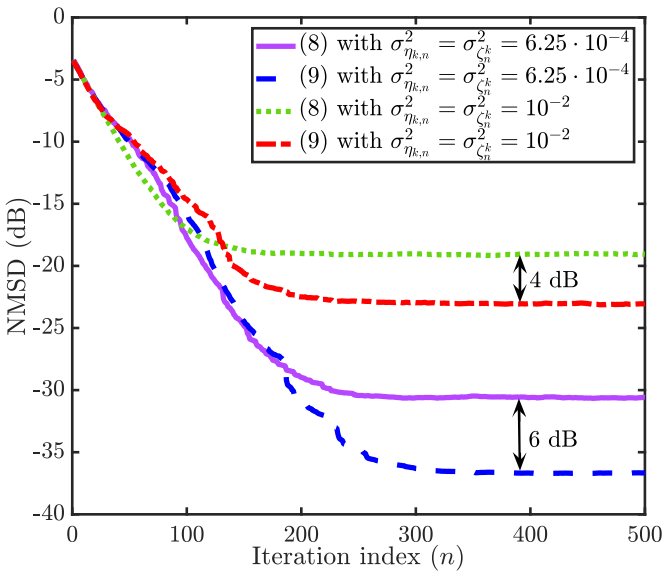


Fig. 1. NMSD of (8) and (9) versus iteration number for $|\mathcal{S}_n| = 4$ and different link noise variances.

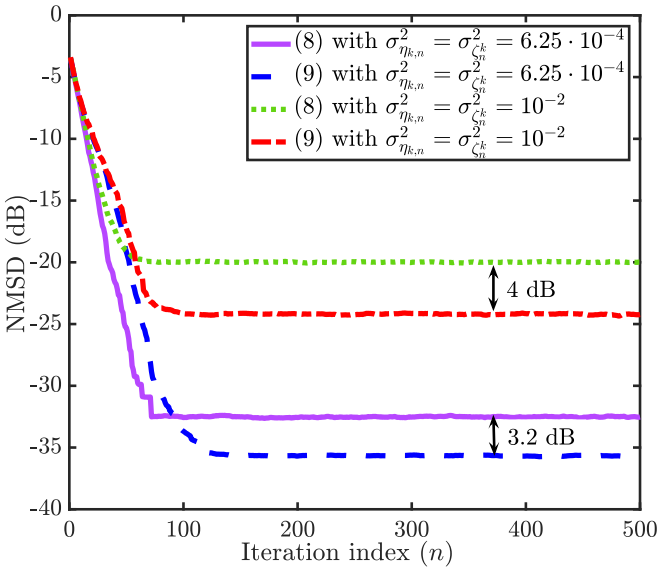


Fig. 2. NMSD of (8) and (9) versus iteration number for $|\mathcal{S}_n| = 10$ and different link noise variances.

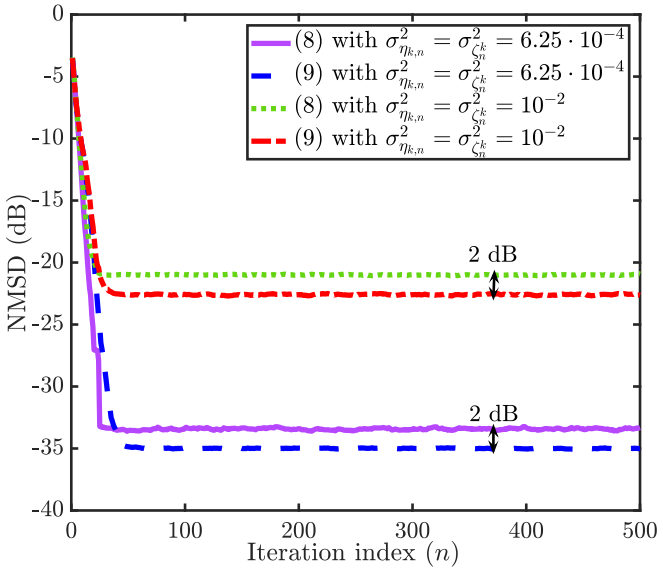


Fig. 3. NMSD of (8) and (9) versus iteration number for $|\mathcal{S}_n| = 25$ and different link noise variances.

IV. Simulation Results

In this section, we conduct a series of numerical experiments to examine the performance of the proposed algorithm. We consider a scenario with $K = 100$ clients directly connected to a global server. The goal of the federated network is to estimate a global model $\mathbf{x}^*$ of dimension $L = 128$. To induce data imbalance among the clients, we draw the size of each client's local dataset, d_k, from a uniform distribution, i.e., $d_k \in \mathcal{U}(50, 90)$. Each client k has synthetic and imbalanced non-IID data $\{\mathbf{H}_k, \mathbf{y}_k\}$ with the matrices $\mathbf{H}_k$ drawn from a multivariate normal distribution $\mathcal{N}(\mu_{\mathbf{H}_k}, \sigma^2_{\mathbf{H}_k})$ where $\mu_{\mathbf{H}_k} \in \mathcal{U}(-0.5, 0.5)$ and $\sigma^2_{\mathbf{H}_k} \in \mathcal{U}(0.5, 1.5)$. The weight matrices are set to the inverse covariance matrix of $\mathbf{y}_k$ at each client k, i.e., $\mathbf{W}_k = \Sigma^{-1}_{\mathbf{y}_k} = \mathbb{E}[(\mathbf{y}_k - \mathbb{E}[\mathbf{y}_k])(\mathbf{y}_k - \mathbb{E}[\mathbf{y}_k])^\intercal]^{-1}$. We set the global parameter vector $\mathbf{x}^*$ arbitrarily by drawing each entry from a standard normal distribution $\mathcal{N}(0, 1)$. The observation noise $\boldsymbol{\nu}_k$ at each client k is zero-mean IID Gaussian with variance 10^{-4}. The additive noise in both uplink and downlink are zero-mean IID white Gaussian. In all experiments, we set the penalty parameter to $\rho_k = 1$ for all clients. At each iteration n, the server selects a subset of the clients with equal probability of selection assigned to each client. We evaluate the performance using the network-wide average normalized mean-square deviation (NMSD) defined at each iteration n as

$$\text{NMSD}(n) = \frac{1}{K} \sum_{k=1}^{K} \frac{\|\mathbf{x}_{k,n} - \mathbf{x}^\star\|_2^2}{\|\mathbf{x}^\star\|_2^2}. \tag{10}$$

We obtain the learning curves (i.e., NMSD in dB vs. iteration number n) by averaging over 100 independent trials.

In Figs. 1-3, we present a performance comparison between the proposed algorithm, i.e., (9), and its closest contender, i.e. (8), which does not feature continual local updates. We obtain the results by including noise in the communication links and utilizing random scheduling for communication efficiency. We set the number of clients selected at each iteration to $|\mathcal{S}_n| \in \{4, 10, 25\}$ and use different link noise variances of $\sigma^2_{\boldsymbol{\eta}_{k,n}} = \sigma^2_{\boldsymbol{\zeta}_n^k} \in \{6.25 \cdot 10^{-4}, 10^{-2}\}$. We plot the corresponding learning curves in the figures. We observe that the proposed algorithm exhibits robustness against communication noise/error even when a small portion of the clients participate in every FL round. It also outperforms (8) significantly in terms of the steady-state NMSD in all considered cases. In summary, the results demonstrate that permitting the clients, which are not selected by random scheduling, to continually update their local models leads to a notable reduction in the steady-state NMSD without compromising the convergence rate. Moreover, as expected, when the variance of the noise in the communication links increases, the learning accuracy decreases.

V. Conclusions

We proposed a new FL algorithm that leverages random scheduling to enhance communication efficiency while being robust to additive noise in the communication links, in part, due to eliminating the dual parameters and minimizing the use of noisy estimates in update equations. The key novel feature of the proposed algorithm is that it allows the clients to update their model estimates locally even when they are not selected by the global server for participation in FL as per random scheduling. Our simulation results showed that the continual local updates lead to performance improvements in terms of both learning accuracy and robustness to link noise.

Acknowledgement

The Research Council of Norway supported this work.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. and Stat.*, Apr. 2017, pp. 1273–1282.

[2] V. Smith, C. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Proc. Adv. Neural Inf. Process. Sys.*, vol. 30, Jan. 2017.

[3] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.

[4] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: algorithms and theory," *IEEE Trans. Automat. Contr.*, vol. 51, no. 3, pp. 401–420, Mar. 2006.

[5] Y. Zhao, J. Zhao, L. Jiang, R. Tan, D. Niyato, Z. Li, L. Lyu, and Y. Liu, "Privacy-preserving blockchain-based federated learning for IoT devices," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1817–1829, Feb 2021.

[6] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 700–10 714, 2019.

[7] H. Zhao, W. Du, F. Li, P. Li, and G. Liu, "FedPrompt: Communication-efficient and privacy-preserving prompt tuning in federated learning," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2023, pp. 1–5.

[8] T. Xia, J. Han, A. Ghosh, and C. Mascolo, "Cross-device federated learning for mobile health diagnostics: A first study on COVID-19 detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2023, pp. 1–5.

[9] Z. Li, V. Sharma, and S. P. Mohanty, "Preserving data privacy via federated learning: Challenges and solutions," *IEEE Consumer Electronics Mag.*, vol. 9, no. 3, pp. 8–16, 2020.

[10] R. Xie, C. Li, X. Zhou, and Z. Dong, "Asynchronous federated learning for real-time multiple licence plate recognition through semantic communication," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2023, pp. 1–5.

[11] Z. Zheng, Y. Zhou, Y. Sun, Z. Wang, B. Liu, and K. Li, "Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges," *Connection Science*, vol. 34, no. 1, pp. 1–28, 2022.

[12] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454–3469, 2020.

[13] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 603–618.

[14] Y. Chang, K. Zhang, J. Gong, and H. Qian, "Privacy-preserving federated learning via functional encryption, revisited," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1855–1869, 2023.

[15] C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *arXiv preprint arXiv:1808.04866*, Aug. 2018.

[16] X. He, J. Zhang, and Q. Ling, "Byzantine-robust and communication-efficient personalized federated learning," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2023, pp. 1–5.

[17] V. C. Gogineni, S. Werner, Y. Huang, and A. Kuh, "Communication-efficient online federated learning strategies for kernel regression," *IEEE Internet Things J.*, pp. 1–1, Nov 2022.

[18] ——, "Communication-efficient online federated learning framework for nonlinear regression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2022, pp. 5228–5232.

[19] ——, "Decentralized graph federated multitask learning for streaming data," in *Proc. 56th Annu. Conf. Inf. Sci. Syst.*, Apr. 2022, pp. 101–106.

[20] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. on Selected Areas in Commun.*, vol. 37, no. 6, pp. 1205–1221, Mar. 2019.

[21] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated learning for ultra-reliable low-latency v2v communications," in *Proc. IEEE Global Commun. Conf.*, Feb. 2018, pp. 1–7.

[22] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. on Commun.*, Jul. 2019, pp. 1–7.

[23] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3452–3464, Mar. 2020.

[24] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, May 2020.

[25] H. Guo, A. Liu, and V. K. N. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 197–210, Jan. 2021.

[26] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1422–1437, Mar. 2022.

[27] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc wsns with noisy links—part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.

[28] E. Lari, V. C. Gogineni, R. Arablouei, and S. Werner, "Resource-efficient federated learning robust to communication errors," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, 2023, pp. 265–269.

Gaussian Process Learning for Location-Based Service Data

Ekin Ugurel[1], Shuai Huang[2] and Cynthia Chen[*1,2]
[1] Department of Civil and Environmental Engineering, University of Washington, Seattle
[2] Department of Industrial & Systems Engineering, University of Washington, Seattle
E-mail: ugurel@uw.edu Tel: 001-206-685-0326
E-mail: shuaih@uw.edu Tel: +001-206-685-2953
E-mail: qzchen@uw.edu Tel: +001-206-543-8974 (corresponding author[*])

***Abstract*—We develop a Gaussian process regression model to fill gaps in LBS (Location-Based-Service) data. We demonstrate our model's performance in two experiments with real app-based data, in which it outperforms alternative imputation methods.**

I. Introduction

The proliferation of Location-Based Services (LBS) in the past two decades have generated massive amount of time- and location-stamped data. Human mobility patterns can be inferred from such data and those patterns, in terms of individual users' visit patterns in time and space have found a wide variety of applications including, for example, quantifying urban vitality [1], understanding commuting patterns [2], and predicting pandemic spreading [3].

While LBS data offers sample sizes orders of magnitudes greater than traditional travel surveys, it is also sparse—for a substantial portion of users, there is a significant number of gaps in the data (extended periods where there is no data). When the LBS dataset has gaps, derived mobility patterns are vulnerable to bias, meaning that they may misrepresent individuals' actual movements [4]–[6]. Filling in gaps in LBS data is thus important to researchers and practitioners alike.

In this study, we develop a GP framework to fill in gaps in LBS data at the individual level. Our framework imputes a user's latitudes and longitudes simultaneously based on a set of predictor variables (primarily time). We train the framework's parameters on longitudinal trajectory data.

Our contributions include: 1) We introduce the use of smooth kernels (i.e., the rational quadratic kernel) and kernels with periodicities (i.e., the periodic kernel) in modeling human mobility data; and 2) We benchmark our model against other gap-filling imputation methods, demonstrating its effectiveness under a range of scenarios.

The rest of the paper is organized as follows. Section II briefly summarizes related work on gap-filling in LBS data. Section III outlines our methodology. Section IV describes the dataset we used in our experiments. Sections V and VI illustrate our results and provide a brief discussion.

II. Related Work

In general, missing data imputation methods for LBS data fall in one of three categories: (1) leverage time-series smoothing and interpolation methods, (2) integrate external data sources to facilitate learning or (3) employ machine learning-based methods. The first approach uses existing methods for smoothing and time-series interpolation to impute missing data points. These range from simpler methods like linear interpolation to more complex ones like seasonal auto-regressive integrated moving average (SARIMAX) [5], [7]–[9]. The drawback of these models is they are often not flexible enough to handle the non-linear nature of individual-level trajectory data.

The second approach leverages external data sources to make inferences about the original dataset. Common external datasets include real-time video feeds, transit agency-issued smart cards, and even aggregated credit card purchase histories [10], [11]. The drawback of such approaches is the difficulty of obtaining additional datasets—for many applications, external sources may not be readily available. Data acquisition and pre-processing for multiple data sources can also be time-consuming.

The third approach define kernels or activation functions to try predicting the similarity of two trajectories, then using the more complete trajectory to impute missing values in the more sparse trajectory [12]–[14]. The drawback of these approaches is that even if two trips have the exact same ground-truth trajectory, varying data sampling rates may distort the shape of the trajectory observed in the data.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

III. METHODOLOGY

Human mobility exhibits some distinct features. The first one is their spatio-temporal autocorrelations—observations at nearby time inputs tend to be spatially close to one another. This is termed as history dependency, meaning that where one is at time t depends upon their location at $t-1$. To model this phenomenon, we use the rational quadratic kernel

$$\boldsymbol{K}_{RQ}(x,x') = \Omega^2(1+\frac{(x-x')^2}{2\alpha l^2})^{-\alpha} \quad (1)$$

where x and x' are two scalar inputs (representing time), Ω is the output variance, determining the average distance of the function away from its mean, $\alpha > 0$ is the scale mixture parameter, determining the relative weighting of large-scale and small-scale variations in the data, and $\ell > 0$ is the lengthscale, specifying the smoothness of the function (i.e., the frequency of the gradient of the function to change sign). Note that as x and x' approach each other, the term with the exponent $-\alpha$ goes to 1, signifying the increasing likelihood of nearby points to covary.

The second feature of human mobility is that there exist anchor points such as home and work and visits to other activity locations such as stores and recreation sites often center around anchor points *[16]–[18]*. Additionally, those visits often exhibit return patterns of different lengths. In other words, there are rhythms of activity and travel patterns—people usually go to their workplaces and return home every day, inducing correlations between the days of the week. Other behavior like grocery shopping, going out with friends, or attending family gatherings happens less frequently, and hence has longer intervals between consecutive instances. These activities tend to be correlated between the weeks of the month and sometimes even the months of the year (i.e., Christmas, Thanksgiving, etc.). To reflect these patterns, we incorporate categorical variables like days and weeks by representing them as sets of binary variables, using a one-of-k encoding. For example, as the days of the week x_d can take one of seven values, $x_d \in \{M, Tu, W, Th, F, Sa, Su\}$, then a one-of-k encoding of x_d will correspond to seven binary inputs, and one-of-k $(Tu) = [0,1,0,0,0,0,0]$. One approach to embed multiple inputs in a GP framework is to multiply kernels defined on each individual input. This family of kernels is called Automatic Relevance Determination (or ARD), so named due to the existence of a different lengthscale parameter for each input dimension d [19]. In this case, the kernel function $\boldsymbol{K}_{RQ}$ in Equation 1 becomes:

$$\boldsymbol{K}_{RQ-ARD}(x,x') = \prod_{d=1}^{D} \Omega_d^2 \left(1+\frac{(x-x')^2}{2\alpha\ell_d^2}\right)^{-\alpha} \quad (2)$$

where ℓ_d is the lengthscale parameter for input dimension d. Thus, if the optimal lengthscale for a given categorical variable (i.e., day of the week) is small, the model has determined relatively low correlation between data in that category. On the other hand, categorical input dimensions with large lengthscales imply relatively little variation (in other words, high correlation) along those dimensions in the output variable (i.e., locations).

In addition to a one-of-k encoding strategy, we also use periodic kernels, which allow GPs to model functions that repeat themselves

$$\boldsymbol{K}_{PER}(x,x') = \Omega^2 \, exp\left(-\frac{2\sin^2(\pi|x-x'|/p)}{\ell^2}\right) \quad (3)$$

where the period length p determines the distance between repetitions of the function and the lengthscale ℓ works in the same way as in the RQ kernel.

IV. DATASET

We employ privacy-protected, passively-generated LBS data from Spectus, a U.S.-based data solution provider specializing in geospatial analytics. The dataset contains time-stamped location traces of 50 anonymous, opted-in individuals in the Greater Seattle Area between December 2019 and July 2020. More specifically, the data includes timestamps, unique device identifiers, latitude and longitude coordinates, and a measure of data precision (i.e., a location radius for which we have 95% confidence). The dataset does not include any demographic or socioeconomic information. Locational data is sent to Spectus servers in encrypted form, via an HTTPS protocol through three options: an Application Protocol Interface (API), through a publisher that has licensed Spectus' software development kit, or via direct server-to-server integration [20].

V. EXPERIMENTS AND RESULTS

We benchmark our model against alternative missing data imputation methods, including simple exponential smoothing, exponential smoothing [8], Holt-Winters [7], ARIMA and SARIMAX [9]. Simple exponential smoothing (SES), exponential smoothing (ES), and the Holt-Winters (Holt) method all work similarly—they weigh the average of past observations in producing forecasts, giving exponentially lower weights as the time gap increases. SES is most suitable for forecasting data with no clear trend or seasonal pattern. ES allows for forecasting with a trend and the Holt-Winters method uses three types of exponential smoothing to model the level (the typical value), the trend (the slope), and the

seasonality (repeating patterns) of data. ARIMA and SARIMAX are both linear regression models fit for forecasting univariate time series data. Each of these methods are regarded as reliable, well-studied alternatives commonly used to impute missing values in time-series, and therefore make a useful comparison. We "simulate" gaps in the selected 50 users by reserving a subset of their data for testing, and we choose this subset such that the temporal occupancy of their training data meets a lower target temporal occupancy (i.e., between 0 and the user's value) according to the temporal resolution being tested. We tested six different temporal resolutions (τ) from one week to fifteen minutes. Due to space limits, we show three intervals here: one week, one day and one hour.

For the benchmark algorithms, we used only the monotonically-increasing $\boldsymbol{t}_u$ as an input since they do not have a straightforward way to deal with categorical or binary variables. For each benchmark algorithm, we selected the best-fit parameters after cross-validating with multiple attempts.

Figure 1 shows the mean absolute error, the root mean square error, and the median absolute deviation of the six models we compared across for 1 week and 1 day. In all of the tests, our model outperformed the other models: The average error was always the lowest, and the 5th and 95th percentile bounds tended to be closer.

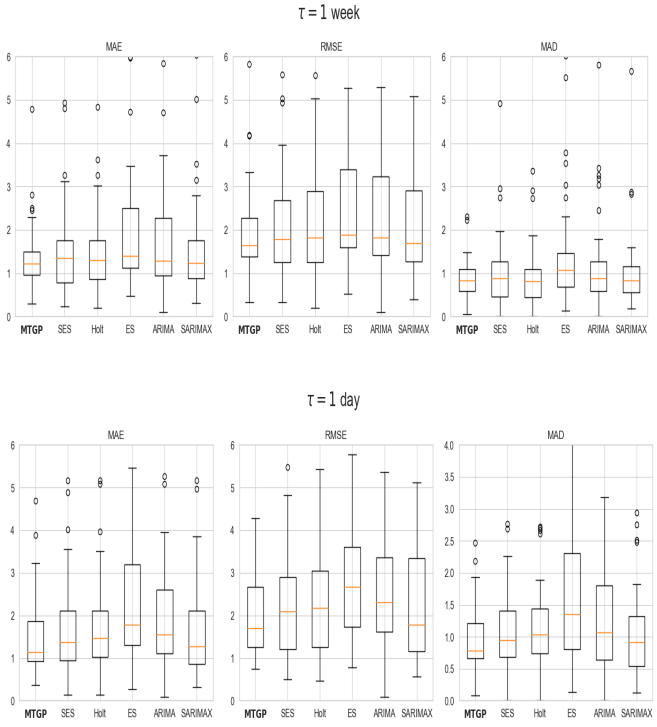


Figure 1. Median absolute error, root mean square error and median absolute deviation of predicted coordinates across different methods (MTGP is our method)

VI. Conclusions

To fill in gaps in the LBS data, we have developed a novel framework using Gaussian processes. We introduced and demonstrated the effectiveness of RQ and PER kernels in the context of LBS data for human mobility modeling, and highlighted the ARD extension of these kernels for fitting high-dimensional data (whether continuous or binary). In results, we show that our model outperformed five existing methods for filling gaps of different lengths in LBS data.

While the current study is motivated from the need to directly address the significant gap issue exhibited in passively-generated LBS data [5], the implications of this work can be significant in a number of aspects. The study opens us the opportunity of creating city-wide simulations of human mobility patterns using those passively-generated LBS data directly. This is in contrast with the current process where parametric models are first estimated using the household travel survey data (which often represents about or less than 0.1% of the population in an urban area) and then those models are extrapolated to a synthetic population.

Acknowledgment

The authors are grateful to the funding support from the National Science Foundation for the project titled as "A whole-community effort to understand biases and uncertainties in using emerging big data for mobility analysis" (award number 2114260). EK and CC are also supported by the Center for Teaching Old Models New Tricks (TOMNET), a University Transportation Center sponsored by the US Department of Transportation through Grant No. 69A3551747116.

References

[1] P. Sulis, E. Manley, C. Zhong, and M. Batty, "Using mobility data as proxy for measuring urban vitality," *J. Spat. Inf. Sci.*, vol. 2018, no. 16, pp. 137–162, 2018, doi: 10.5311/JOSIS.2018.16.384.

[2] V. Frias-Martinez, C. Soguero, and E. Frias-Martinez, "Estimation of urban commuting patterns using cellphone network data," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, in UrbComp '12. New York, NY, USA: Association for Computing Machinery, Aug. 2012, pp. 9–16. doi: 10.1145/2346496.2346499.

[3] Q. Hao, L. Chen, F. Xu, and Y. Li, "Understanding the Urban Pandemic Spreading of COVID-19 with Real World Mobility

Data," *26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. KDD 2020*, pp. 3485–3492, 2020, Accessed: Jun. 09, 2022. [Online]. Available: https://doi.org/10.1145/3394486.3412860

[4] X. (Jeff) Ban, C. Chen, F. Wang, J. Wang, Y. Zhang, and United States. Federal Highway Administration, "Promises of Data from Emerging Technologies for Transportation Applications: Puget Sound Region Case Study," FHWA-HEP-19-026, Oct. 2018. Accessed: May 17, 2022. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/55796

[5] D. McCool, P. Lugtig, and B. Schouten, "Maximum interpolable gap length in missing smartphone-based GPS mobility data," *Transportation*, Sep. 2022, doi: 10.1007/s11116-022-10328-2.

[6] N. H. Merrill, S. F. Atkinson, K. K. Mulvaney, M. J. Mazzotta, and J. Bousquin, "Using data derived from cellular phone locations to estimate visitation to natural areas: An application to water recreation in New England, USA," *PLOS ONE*, vol. 15, no. 4, p. e0231863, Apr. 2020, doi: 10.1371/journal.pone.0231863.

[7] T. Cipra, J. Trujillo, and A. Robio, "Holt-Winters Method with Missing Observations," *Manag. Sci.*, vol. 41, no. 1, pp. 174–178, Jan. 1995, doi: 10.1287/mnsc.41.1.174.

[8] J. Huo, C. D. Cox, W. L. Seaver, R. B. Robinson, and Y. Jiang, "Application of Two-Directional Time Series Models to Replace Missing Data," *J. Environ. Eng.*, vol. 136, no. 4, pp. 435–443, Apr. 2010, doi: 10.1061/(ASCE)EE.1943-7870.0000171.

[9] R. Kohn and C. F. Ansley, "Estimation, Prediction, and Interpolation for ARIMA Models with Missing Data," *J. Am. Stat. Assoc.*, Mar. 2012, Accessed: Feb. 02, 2023. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478332

[10] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Towards mobile intelligence: Learning from GPS history data for collaborative recommendation," *Artif. Intell.*, vol. 184–185, pp. 17–37, Jun. 2012, doi: 10.1016/j.artint.2012.02.002.

[11] X. Gong, Z. Huang, Y. Wang, L. Wu, and Y. Liu, "High-performance spatiotemporal trajectory matching across heterogeneous data sources," *Future Gener. Comput. Syst.*, vol. 105, pp. 148–161, Apr. 2020, doi: 10.1016/j.future.2019.11.027.

[12] G. Liu and J.-P. Onnela, "Bidirectional imputation of spatial GPS trajectories with missingness using sparse online Gaussian Process," *J. Am. Med. Inform. Assoc.*, vol. 28, no. 8, pp. 1777–1784, Jul. 2021, doi: 10.1093/jamia/ocab069.

[13] H. Ren *et al.*, "MTrajRec: Map-Constrained Trajectory Recovery via Seq2Seq Multi-task Learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Virtual Event Singapore: ACM, Aug. 2021, pp. 1410–1419. doi: 10.1145/3447548.3467238.

[14] Z. Wang, S. Zhang, and J. J. Q. Yu, "Reconstruction of Missing Trajectory Data: A Deep Learning Approach," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, Rhodes, Greece: IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ITSC45102.2020.9294402.

[15] R. Kitamura and M. Kermanshah, "Identifying Time and History Dependencies of Activity Choice," *Transp. Res. Rec.*, p. 9.

[16] H. Barbosa, F. B. de Lima-Neto, A. Evsukoff, and R. Menezes, "The effect of recency to human mobility," *EPJ Data Sci.*, vol. 4, no. 1, p. 21, Dec. 2015, doi: 10.1140/epjds/s13688-015-0059-8.

[17] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nat. Phys.*, vol. 6, no. 10, pp. 818–823, Oct. 2010, doi: 10.1038/nphys1760.

[18] D. do C. Teixeira, J. M. Almeida, and A. C. Viana, "On estimating the predictability of human mobility: the role of routine," *EPJ Data Sci.*, vol. 10, no. 1, p. 49, Dec. 2021, doi: 10.1140/epjds/s13688-021-00304-8.

[19] D. K. Duvenaud, "Automatic Model Construction with Gaussian Processes," University of Cambridge, 2014.

[20] "Privacy-First Approach | Policy - Spectus," Jan. 24, 2022. https://spectus.ai/privacy/privacy-policy/ (accessed Dec. 01, 2022).

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Distributed on-line anomaly detection using kernel methods

Anthony Kuh* and Tyler Baguio†
* University of Hawaii, Honolulu, HI 96822
E-mail: kuh@hawaii.edu
† University of Hawaii, Honolulu, HI 96822
E-mail: tylerb8@hawaii.edu

***Abstract*—In this paper we develop an online distributed learning algorithm for unlabeled data. Here a goal is to detect anomalies from streaming data. The assumption is that data is drawn from a probability distribution, but that at some unknown time data could be drawn from a different distribution. We assume the data comes from distributed sources and develop an online Federated Learning (FL) algorithm based on online kernel methods using Random Fourier Features (RFF). This work combines previous work on online unsupervised kernel algorithms with online supervised kernel algorithms using FL and RFF. There are many applications of this work including detecting bad power grid data.**

I. Introduction

In this day and age we are seeing an explosion of data coming from a variety of sources and locations at different rates. Much of this data come from distributed sources such as sensors, IoT devices, mobile devices, cameras, individuals, groups, or organizations. There is a need to process, learn, and make inferences from the data. Here we consider data gathered by edge devices or clients with information passed to a central server or the cloud. In the past, data was sent from the clients to the cloud where processing, learning and making decisions would occur as shown in Fig. 1a) and labeled as the centralized learning model. However, edge devices now have much more computational power with faster processors (GPUs) allowing more computations and learning to occur at the edge. In addition there are many other concerns with sending data from edge devices to the cloud. There are security and privacy concerns of sending data and information to the cloud. There is also concerns about an excess amount of communications between the edge devices and the cloud. Federated Learning (FL) was developed by [1] to address these issues where processing and learning takes place at the edge device and not passed to the cloud. This is shown in Fig. 1b) where information, but not data is passed to the cloud. The cloud assists in the learning process by sending collective information back to each edge device so that collaborative learning can occur. The FL model is also able to handle heterogeneous data and heterogenous systems.

There are many applications for FL where there are concerns about data privacy, communication costs, and handling heterogeneous data and systems. These include mobile communications, autonomous systems with communications between autonomous cars for real-time information and decision making,

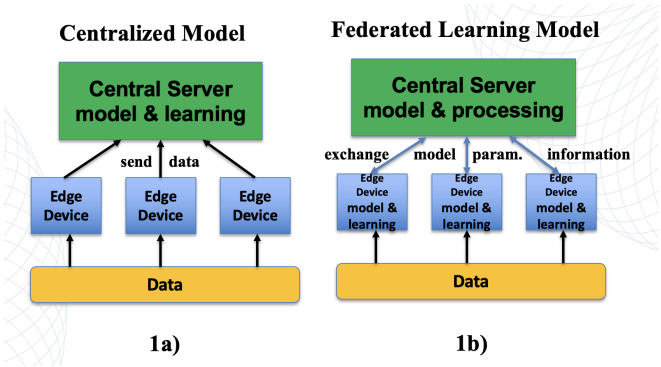


Fig. 1. Collaborative learning models

healthcare applications include sharing information between different hospitals, manufacturing for predictive maintenance, and smart grid applications for demand-side management, [2]. For demand-side management consider a group of homes in a community. Each home may be identical or similar in terms of the architecture, but energy usage in each home may be very different depending on the occupants, their energy preferences, and energy consuming devices in the home. This includes heating, air conditioning, hot water heaters, and appliances. Homes may also have rooftop solar panels, battery storage, and charging stations for Electric Vehicles (EV)s. This is a good application for FL as at each home gets data from many heterogeneous sources and there are privacy issues. This include environmental sensors monitoring local climate including solar irradiance, electrical grid data, data from sensors monitoring energy usage, and possibly data from sensors monitoring EV charging. From this data we can learn about household energy usage while preserving comfort, conserving energy and reducing costs.

There has been considerable research and interest in FL algorithms with the following papers giving a good discussion of FL research [3], [4]. Most of the learning algorithm for FL are supervised learning algorithms. In previous work we developed online kernel regression algorithms that used principles of FL, [5], [6]. The algorithms developed in [6] are based on edge devices (clients) learning data received using online kernel methods implemented using Random Fourier Features (RFF),

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

[7]. The clients send weight information to a server where the server coordinates learning by aggregating information from the clients and then sending this information back to the clients. Communication savings are achieved as we only sample a small number of clients at each update and the client only sends a small fraction of RFF coefficients when sampled. This is called Partial Sharing-Based Online Federated Learning (PSO-FED). PSO-FED is shown to converge and through simulations have good learning performance (even when a small fraction of RFF are transmitted). The algorithm is also shown to have some robustness against data poisoning (Byzantine attacks).

In this paper we extend our learning algorithms to unsupervised learning. Although the vast majority of machine learning deals with supervised learning algorithms where a target output is associated with each input data a considerable amount of data is unlabeled. Many learning algorithms have been developed for unsupervised learning ranging from clustering algorithms where similar input data is grouped together to Principal Component Analysis (PCA) which is a statistical method for reducing the dimensionality of a dataset by linear transforming the dataset to find the coordinates with the largest variance. Here we will focus on anomaly detection with a survey of machine learning methods presented in [8]. Here we will examine where data is drawn from a probability distribution in an online manner. At some point data is drawn from an anomalous distribution and the goal is to detect this change. We first formulate an online kernel algorithm following [9] and then use RFF to approximate the kernels to detect the change. We then adapt this algorithm based on distributed learning using principles of FL.

Let us briefly discuss two applications for anomaly detection. Consider a network of environmental sensors monitoring water quality. The sensors could be placed under water in a lake. Each sensor would periodically sample environmental conditions such as temperature, pH, turbidity, and perhaps other chemicals. PAD-OFL could learn the support of the distribution under normal conditions. Anomalies could occur due to natural conditions such as unusual algae growth or due to human factors such as the release of pathogens and/or pollutants into the lake. In a second application consider monitoring a power grid for bad data. Sensor could be distributed on the grid to monitor voltage, current, and other electrical measurements. Bad data could be the result of natural circumstances such as a downed power line due to fallen trees or a malicious man-made attack where bad data is injected. A more detailed discussion presenting simulation results is presented for detecting bad power grid data in the applications section.

The paper is organized as follows. In Section 2 we first discuss using a one class least squares kernel algorithm for modeling a probability distribution. We then approximate the kernels using RFF, and develop an online learning algorithm for anomaly detection. In Section 3 we assume the data comes from distributed sources (clients) and information is processed by a central server. We then develop a FL model for anomaly detection. Section 4 presents a simple example and an application detecting bad data for the electric power grid where algorithms of Section 3 are used. Section 5 summarizes the paper and discusses further directions.

II. Modeling Distributions and Anomaly Detection

A. Using Kernel Methods

In a key paper by Schölkopf, [10] a one-class support vector is used to model the support of a probability distribution. Here we will assume inputs $x(i), 1 \leq i \leq m$ are drawn from a sample of a random vector $X \in \mathcal{R}^n$. This can be represented by

$$\min \frac{1}{2}||w||^2 - \rho + \gamma \sum_i \xi_i \tag{1}$$

$$\text{subject to } w^T\phi(x(i)) \geq \rho - \xi_i. \text{ and } \xi_i \geq 0,$$

where $\phi(\cdot) : \mathcal{R}^n \rightarrow \mathcal{Z}$ is the mapping to a high dimensional feature space such that the dot product of ϕ is usually computed by evaluating a kernel $k(x, y) = \phi(x)^T\phi(y)$. The parameter $\gamma > 0$ is a regularization parameter and controls the fraction of possible outliers [10] and ξ_i are non-zero slack variables. The goal here in feature space is to have as many data points (in feature space) $\phi(x(i))$ as possible to lie on one side of a hyperplane (described by $f(x) = w^T\phi(x) - \rho = 0$) away from the origin such that the distance to the origin, ρ/w is as large as possible. This is know as the One-Class Support Vector Machine (SVM). The hyperplane solution can be found by solving the dual optimization problem [10]

$$\min \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j k(x(i), x(j)) \tag{2}$$

$$\text{subject to } 0 \leq \alpha_i \leq \gamma \text{ and } \sum_i \alpha_i = 1,$$

where α_i are the Lagrange multipliers. The solution for the weight vector w is given by

$$w = \sum_i \alpha_i \phi(x(i)).$$

We would like to consider online learning algorithms for this problem and can use a different squared error cost function which leads to a simpler online learning algorithms based on linear adaptive filters. This was implemented by Choi in [11] using a squared error loss function and is referred to as the Least Squares OC-SVM (LS-OC-SVM). For this optimization problem the goal is to perform linear regression in feature space with the hyperplane again as far away from the origin as possible. In the input space the manifold formed by the equation $f(x) = w^T\phi(x) - \rho = 0$ describes a surface where most of the support of the probability distribution lie near to f(x). The LS-OC-SVM optimization problem can be formulated as

$$\min J(w, \rho) = \min \frac{1}{2}||w||^2 - \rho + \frac{\gamma}{2}||\xi||^2 \tag{3}$$

$$\text{subject to } w = \mathbf{\Phi}\alpha \text{ and } \xi_i = \mathbf{1}\rho - \mathbf{\Phi}^T w.$$

where the feature input matrix $\mathbf{\Phi} = [\phi(x_1), \ldots, \phi(x_m)]$, $\alpha = [\alpha_1, \ldots, \alpha_m]^T$, and $\mathbf{1}$ represents a column vector of 1s. Note hear that for the LS-OC-SVM we have constraints that are equality constraints as opposed to the OC-SVM where we have constraints that are inequality constraints. Defining the kernel matrix $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$ we can then make substitutions in equation (3) to get the following

$$J(\alpha, \rho) = \frac{1}{2}\alpha^T\mathbf{K}\alpha + \frac{\gamma}{2}||\mathbf{1}\rho - \mathbf{K}\alpha||^2 \tag{4}$$

This is a quadratic function of the variables α and ρ with the solution found by solving a system of linear equations given by

$$\begin{bmatrix} \mathbf{1}^T\mathbf{1} & -\mathbf{1}^T\mathbf{K} \\ -\mathbf{1} & \mathbf{K} + \mathbf{I}/\gamma \end{bmatrix} \begin{bmatrix} \rho \\ \alpha \end{bmatrix} = \begin{bmatrix} 1/\gamma \\ \mathbf{0} \end{bmatrix} \tag{5}$$

where $\mathbf{0}$ is a column vector of zeros and $\mathbf{I}$ is an identity matrix.

Here all the input data are support vectors and the hyperplane in the dual space is given by $f(x) = \alpha^T\mathbf{k}(x) - \rho = 0$ where $\mathbf{k}(x) = \mathbf{\Phi}^T\phi(x)$. A problem with solving the set of linear equations, from equation (5) is that the matrix $\mathbf{K}$ grows as the number of input data grow, m as all input data are support vectors for the LS-OC-SVM. For m data the computational complexity for the solution grows as m^3 and alternative subspace methods are needed when m grows large. Subspace methods require that a dictionary of support vectors are chosen. For online learning the dictionary is grown online as data is processed. In [9] an online dictionary is grown using an approximate linear dependence (ALD) criterion, [12]. Alternatively, in [13] a dictionary is grown using a simpler coherence criterion. A problem with these methods is that they add complexity to the learning problem and we propose an alternate solution in the next subsection.

B. Approximating Kernels with Random Fourier Features

In a landmark paper by Rahimi and Recht, [7] they showed that if the kernel function can be expressed in terms of one variable, that is $k(u, v) = k(u - v)$, then we can have an alternative approach to construct a set of basis functions. Here assume that $k(u, u) = 1$. From Bochner's theorem we can then represent the kernel function in terms of its inverse n-dimentsional Fourier transform by

$$\begin{aligned} k(u-v) &= \left(\frac{1}{2\pi}\right)^n \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p_W(\omega) \exp(j\omega^T(u-v)) d\omega_1 \ldots d\omega_n \\ &= \mathbf{E}_W[\psi_\omega(u)\psi_\omega(v)^*] \end{aligned} \tag{6}$$

where $\psi_\omega(u) = \exp(j\omega^T u)$ and $*$ represents complex conjugate. From the righthand side of the equation we have $\psi_\omega(u)\psi_\omega(v)^*$ is an unbiased estimate of $k(u-v)$ as $p_W(\omega)$ is an n-dimensional probability density function. Note that

$$k(u - v) = \mathbf{E}_W \cos(\omega^T(u - v))$$

since $p_W(\omega)$ is real and even function of its n components (as $k(u, v)$ is real, even, and positive semi-definite). Let $c_\omega(u) = \sqrt{2}\cos(\omega^T u + \theta)$ where ω is drawn from $p_W(\omega)$ and θ is drawn from a uniform $[0, 2\pi]$ random variable. We then have

$$\begin{aligned} \mathbf{E}_{W,\theta}[c_\omega(u)c_\omega(v)] &= \mathbf{E}_{W,\theta}[\cos(\omega^T(u-v)) + \cos(\omega^T(u+v) + 2\theta)] \\ &= k(u - v) \end{aligned}$$

where the last equality holds as the expectation of the second term is 0.

Let $(\omega_j, \theta_j), 1 \leq j \leq D$ be D samples drawn from W and θ. We then have

$$k(u, v) = \langle \phi(u), \phi(v) \rangle \approx \frac{1}{D} \sum_{j=1}^{D} c_{\omega_j}(u) c_{\omega_j}(v) = c^T(u)c(v) \tag{7}$$

where $c(u) = \frac{1}{\sqrt{D}}[c_{\omega_1}(u), \ldots, c_{\omega_D}(u)]^T$. Here we approximate the kernel by drawing D samples and averaging. As an example of W let the kernel function be

$$k(u, v) = \exp(\sigma^2||u - v||^2/2)$$

then the kernel function is a Gaussian kernel and the inverse Fourier transform gives a Gaussian random vector with each component being iid with zero mean and variance σ^2. From the kernel representation result we then have the estimate of $f(\cdot)$ given by

$$\begin{aligned} f(x) &= \alpha^T k(x) - \rho = \sum_{i=1}^{m} \alpha_i k(x(i), x) - \rho \\ &\approx \sum_{i=1}^{m} \alpha_i c(x(i))^T c(x) - \rho. \end{aligned} \tag{8}$$

If we let $\beta = \sum_{i=1}^{m} \alpha_i c(x(i))$, then we have that

$$f(x) \approx \beta^T c(x) - \rho. \tag{9}$$

This reduces the problem to a linear regression problem in D-dimensional space generated by the RFF with the goal of finding the parameters β and b. Let $\mathbf{C} = [c(x(1)), \ldots, c(x(m))]$. Then with regularization we are trying to find β and ρ to minimize

$$J(\beta, \rho) = \frac{1}{2}||\beta||^2 + \frac{\gamma}{2}||e||^2 - \rho$$

where $e = \mathbf{C}^T\beta - \rho\mathbf{1}$. Solving this optimization problem is found by taking gradients and partial derivatives and setting to 0 resulting in

$$\nabla_\beta J(\beta, \rho) = (\mathbf{I} + \gamma\mathbf{C}\mathbf{C}^T)\beta - \gamma\mathbf{C}\mathbf{1}\rho = 0 \tag{10}$$

and

$$\frac{\partial J(\beta, \rho)}{\partial \rho} = \gamma m\rho - \gamma\mathbf{C}^T\beta - 1 = 0 \tag{11}$$

resulting in the following set of linear equations,

$$\begin{bmatrix} 1 & -{m_c}^T \\ -m_c & \mathbf{R}_c + \mathbf{I}/(m\gamma) \end{bmatrix} \begin{bmatrix} \rho \\ \beta \end{bmatrix} = \begin{bmatrix} 1/(m\gamma) \\ 0 \end{bmatrix} \tag{12}$$

where $m_c = \frac{1}{m}\mathbf{C}\mathbf{1}$ and $\mathbf{R}_c = \frac{1}{m}\mathbf{C}\mathbf{C}^T$. The solution is found by solving a system of $D + 1$ linear equations.

C. Anomaly Detection and Online Learning

Solving for β and ρ gives us a manifold described by $g(x) = \beta^T c(x) - \rho = 0$. Given we use enough training data m and enough RFF, D the support of the random variable X lies close to this manifold with the distance of a point x from the manifold given by

$$d(x) = \frac{|g(x)|}{\sqrt{\beta^T c(x)^T c(x)\beta}} \tag{13}$$

Following [9], if the distance of a data point is far from the manifold we classify this as an anomalous data point. Here we define a threshold value $\delta > 0$ such that if $d(x) > \delta$ then the data point is an anomaly. The setting of δ determines the fraction of points that are classified as anomaly data.

Here we assume that the first m data points are drawn from the random variable X. This can be considered the baseline data. The LS-OC-SVM can be trained to find the manifold using batch methods solving equation (12) or by online methods. Online methods include using the Recursive Least Square (RLS) algorithm or the Least Mean Square (LMS) algorithm described in [14]. The nice thing is that instead of working with kernel online learning algorithms we work with simpler linear adaptive filter algorithms.

After seeing the m data points we are presented with more data and need to determine if the data is anomalous data. Here we set the parameter δ, initialize $\beta(m)$ and $\rho(m)$, and let $z(k) = c(x(k))$ by solving equation (12), and then use the LMS learning algorithm to update the parameters. The learning algorithm is given below.

==========

Centralized Anomaly Detection with Online Learning using Kernels and RFF (CAD-OL)

1) **Initialization: Learn** $\beta(m)$ **and** $\rho(m)$**, set** $\delta > 0$**, set step sizes** μ_1, μ_2**, and** $k = m$**.**
2) **Get data point,** $x(k+1)$
3) **If** $x(k+1)$ **is anomaly data that is (**$d(x(k+1)) > \delta$**), label point** $A(k+1) = 1$ **and**

$$\beta(k+1) = \beta(k), \;\; \rho(k+1) = \rho(k)$$

4) **otherwise update parameters using LMS algorithm estimating gradients and derivatives from equations (10) and (11)**

$$e(k+1) = z(k+1)^T \beta(k) - \rho(k),$$

$$\beta(k+1) = (1 - \frac{\mu_1}{\gamma m})\beta(k) - \mu_1 e(k+1) z(k+1),$$

$$\rho(k+1) = \rho(k) + \mu_2(\frac{1}{\gamma m} + e(k+1)),$$

and $A(k+1) = 0$

5) $k \leftarrow k+1$**, go to 2)**

==========

Note that CAD-OL can detect anomalies if δ is set correctly with the vector of anomalies recorded in A. It also tracks slowly varying changes in the distribution as the manifold parameters slowly change as new data is presented.

III. Online Unsupervised Federated Learning using PSO-FED

As mentioned in the introduction many sources of data come from distributed sources and this Section develops the models and Federated Learning (FL) algorithms for this scenario. Here we assume we have L clients and also assume that clients receive data synchronously at each update, but this can be easily modified to asynchronous updates [5], [15]. At time k client l receives input data $x(k,l)$ with $z(k,l) = c(x(k,l))$. There is also a server that communicates with all the clients and helps to coordinate learning. One approach as mentioned earlier is a centralized learning approach where the clients send all data to the server. Then the server can just use CAD-OL discussed in the previous section. As mentioned earlier this results in high communication costs and loss of data privacy. This motivates the need to use FL algorithms.

Here we combine the unsupervised anomaly detection learning algorithm discussed in the previous section, CAD-OL with PAO-FED discussed in [6]. At each update the server chooses some random set of $P << L$ clients who each communicate with the server about their parameter information. When a server chooses a client, that client gets the server model of the parameters and adaptively updates their parameter based on its current input, $x(k,l)$ (CAD-OL step 4) unless that data point is an anomaly. The server then takes the sum of all the P clients updates as its new parameter estimate. The process then repeats as the server chooses a new set of P clients at the next update. In [6] this is referred to as Online-Fed learning.

To implement PAO-FED, [6] we start at each update by choosing $R << D$ parameters of the vector β for each client. Each client can have the same components chosen (coorinated) or choose different components (uncoordinated). Initially, these can be chosen randomly or sequentially. To denote which components of β are chosen we will define a selection matrix $S(k,l)$ at time k for client l which is a diagonal matrix of dimension D with 1s on the diagonal element where there will be an update on that component of β and 0s otherwise. We also let $\mathcal{P}(k)$ denote the set of clients chosen at time k. The algorithm for each client is similar to CAD-OL for edge devices with a server aggregating the selected clients and sending this information to each of the chosen clients. At each update the clients must also rotate which components of β are updated to iterate for the next update. We initially assume that we learn the support of the distribution.

==========

Partial Sharing Anomaly Detection with Online FED Learning using Kernels and RFF (PAD-OFL)
(learning support of distribution)

1) **Initialization: Set** $\beta(0)$**,** $\rho(0)$ **,** $\beta(0,l), \rho(0,l), 1 \leq l \leq L$**, set step sizes** μ_1, μ_2**, determine initial chosen components,** $S(0,l),\ 1 \leq l \leq L$**, and** $k = 0$**.**
2) **Get data points,** $x(k,l)$**, choose** P **clients,** $\mathcal{P}(k)$**, and communicate** $S(k,l)\beta(k)$ **to selected clients.**

3) **Client Update: If** $j \in \mathcal{P}(k)$

$$\beta'(k,j) = S(k,j)\beta(k) + (\mathbf{I}_D - S(k,j))\beta(k,j)$$

$$e(k+1,j) = z(k+1,j)^T\beta'(k.j) - \rho(k,j),$$

$$\beta(k+1,j) = (1-\frac{\mu_1}{\gamma m})\beta'(k,j) - \mu_1 e(k+1,j)z(k+1,j),$$

Else

$$e(k+1,j) = z(k+1,j)^T\beta(k.j) - \rho(k,j),$$

$$\beta(k+1,j) = (1-\frac{\mu_1}{\gamma m})\beta(k,j) - \mu_1 e(k+1,j)z(k+1,j),$$

Endif

$$\rho(k+1,j) = \rho(k,j) + \mu_2(\frac{1}{\gamma m} + e(k+1,j),$$

4) **Selection matrix update:** $S(k,j+1) = \mathbf{circshift}(\mathbf{diag}(S(k,j)),r)$**. Shift of** r**.**
5) **Server update: average client update parameters to get global update**

$$\beta(k+1) = \frac{1}{P}\sum_{j\in\mathcal{P}(k)} \beta(k+1,j)$$

$$= \frac{1}{P}\sum_{j\in S_k} S(k+1,j)\beta(k+1,j) + (\mathbf{I}_D - S(k+1,j))\beta(k)$$

$$\rho(k+1) = \frac{1}{P}\sum_{j\in\mathcal{P}(k)} \rho(k+1,j)$$

6) $k \leftarrow k+1$**, go to 2)**

==========

The second part of PAD-OFL is to detect anomalies. Here the algorithm is the same as for learning support of distribution except that we have a check if data point $x(k,l)$ is an anomaly by checking if $d(x(k,l)) >$ after step 2). The procedure is the same as CAD-OL. For PAD-OFL note that at each update the client l only communicates R components of $\beta(k,l)$ to the server and the server only communicates R components of $\beta(k)$ to each client.

IV. Application of Algorithm

Let us first examine a simple example that was presented in [9]. This is the synthetic square-noise dataset with normal points labeled in blue and noisy points labeled as red. This is shown in Fig. 2. If we assume the normal points are drawn from a probability distribution we can get the distribution support lies within two square boxes that contain the blue points. It is not difficult to simulate the LS-OC-SVM to construct the square boundary boxes approximately and then being able to detect the red point anomalies. The online CAD-OL has also been implemented and can get a good approximation of the two square boundary boxes and detect anomalies using Gaussian kernels and RFF. Similarly, PAD-OFL can be implemented with a number of clients to get the two square boundaries and detect the red point anomalies.

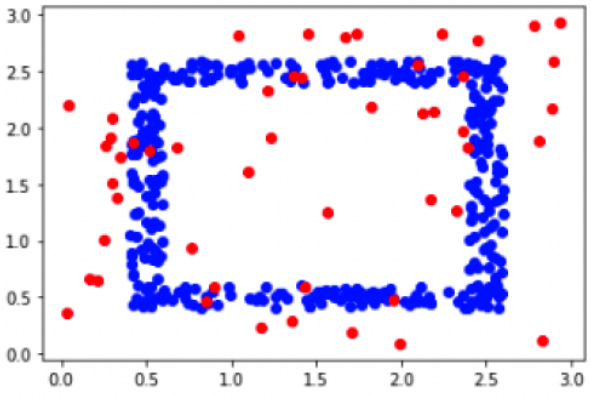


Fig. 2. Synthetic square-noise data set

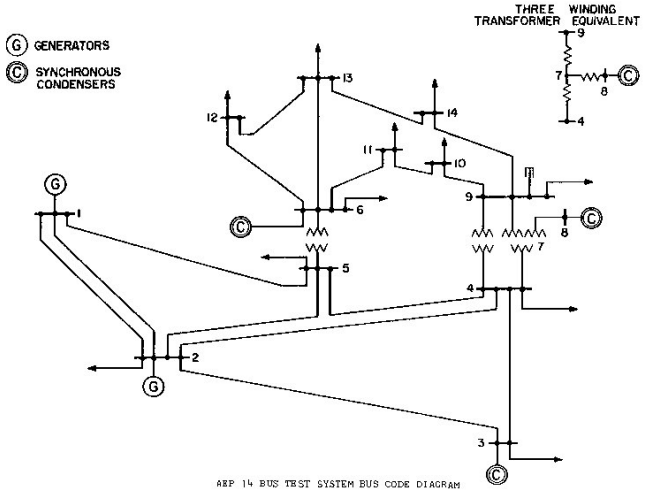


Fig. 3. IEEE 14-bus test system

For real applications, the probability distributions could be in high dimensions with complicated support. A goal is to implement the algorithms developed here on some of these applications. Here we briefly discuss one example of detecting bad electric power grid data.

Power injection	At buses 1, 2, 3, 8
Power flow	At branches 1-2, 1-5, 2-4, 2-5, 4-5, 4-7, 4-9, 5-6, 6-11, 6-14, 9-10, 9-14, 12-13
Voltage magnitude	At buses 3, 6, 8, 10, 14

Table 1. Measurement configuration for IEEE 14-bus system

In [9], detection of bad power grid data was considered. The IEEE 14-bus system [16] shown in Fig. 3 was used to test for bad grid data. A standard residual test for non-critical branch measurements was able to detect bad grid data, however branch P4-7 is a critical branch that is not observable and the standard residual test does not work. If we use an online LS-OC-SVM

using a kernel RLS algorithm we can detect bad grid data for branch P4-7. The experiment was done by considering a vector $x = [P_{4-7}, P_{4-2}, P_{4_5}, P_{4-9}]^T$ which contains the critical measurement along with measurements from neighboring buses. The true value for this flow is $P_{4-7} = 0.2285 p.u.$ We created a stream of 1000 samples by adding zero mean Gaussian noise with variance .0004 to the true value. The data was trained on a LS-OC-SVM to produce a manifold where the support of the distribution lies close to the manifold. Gaussian kernels were used and first trained on 700 training data. Then 300 test data were used with 5 of the data points injected with bad data. The LS0-OC-SVM was able to detect all five bad data points.

We then used CAD-OL using RFF using the LMS learning algorithm with suitably adjusted hyperparameters and small learning step sizes. This algorithm like the LSO-OC-SVM was also able to detect all five bad data points. Note that much power grid data is measured from distributed sources and we then applied the PAD-OFL to attempt to detect the bad grid data. We used 10 clients where data was equally and randomly distributed among these clients and then used the LMS algorithm where at each update we chose 3 of the clients to update. A total of 30 RFF were used and at each update we chose 5 of these RFF to update. The savings in communication cost over using all RFF is a factor of six. PAD-OFL using these parameters was also able to detect all five bad data points. PAD-OFL achieved communication savings and also reduced computations compared with centralized approaches and using LSO-OC-SVM.

We have discussed a power grid application where we could implement online FL algorithms such as PAD-OFL. There are many other applications where online anomaly detection for distributed sources using PAD-OFL could be used. These include personalized healthcare, networked autonomous vehicles, and precision agriculture with some of these applications are discussed in [2].

V. Summary and Further Direction

This paper discussed implementing online unsupervised learning algorithms for anomaly detection. We consider situations where data is coming from distributed sources and use principles of Federated Learning. We also use kernel methods and represent kernels using random Fourier Features (RFF). The paper combines implementation of LS-OC-SVM [9], with RFF [7], and using the PAO-FED algorithm [6]. The algorithm developed is called Partial Sharing Anomaly Detection with Online FED Learning using Kernels and RFF (PAD-OFL). We simulated results detecting bad power grid data showing the efficacy of both CAD-OL and PAD-OFL.

There is much further work to be considered including using PAD-OFL on numerous other applications and understanding the analytical behavior of PAD-OFL using tools from signal processing and information theory.

Acknowledgements

This work was supported in part by the United States National Science Foundation (NSF) grant ECCS-2142987.

References

[1] J. Konečný, B. McMahan, and D. Ramage. Federated optimization:distributed optimization beyond the datacenter, 2015.

[2] V. C. Gogineni, S. Werner, F. Gauthier, Y.-F. Huang, and A. Kuh. Personalized online federated learning for IOT/CPS: challenges and future directions. *IEEE Internet of Things Magazine*, 5:78–84, 2022.

[3] P. Kairouz and et. al. Advances and open problems in federated learning, 2021.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Proc. Magazine*, 37(3):50–60, May 2020.

[5] A. Kuh. Real time kernel learning for sensor networks using principles of federated learning. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 2089 – 2093, Tokyo, Japan, Dec. 2021.

[6] V. C. Gogineni, S. Werner, Y.-F. Huang, and A. Kuh. Communication-efficient online federated learning strategies for kernel regression. *IEEE Internet of Things Journal*, 10(5):4531–4544, Nov. 2022.

[7] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[8] L. Ruff, J. R. Kauffmann, R. A Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K. R. Müller. A unifying review of deep and shallow anomaly detection. *IEEE Proceedings*, 109(5):756–795, May 2021.

[9] M. Uddin and A. Kuh. Online least-squares one-class support vector machine for outlier detection in power grid data. In *2016 IEEE ICASSP*, Shanghai, China, Mar. 2016.

[10] B Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C.Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[11] Y.-S. Choi. Least squares one-class support vector machine. *Pattern Recognition Letters*, 30(13):1236–1240, 2009.

[12] Y. Engel, S. Mannor, and R. Meir. The kernel recursive least-squares algorithm. *IEEE Trans. on Signal Proc.*, 52(3):2275–2285, Aug. 2004.

[13] C. Richard, J.-C. Bermudez, and P. Honeine. Online prediction of time series data with kernels. *IEEE Trans. on Signal Proc.*, 57(3):1058–1067, Mar. 2009.

[14] S. Haykin. *Adaptive Filter Theory, 5th Ed.* Pearson, 2014.

[15] F. Gauthier, V. C. Gogineni, S. Werner, Y.-F. Huang, and A. Kuh. Asynchronous online federated learning with reduced communication requirements. Technical report, arXiv preprint arXiv:2303.15226, 2023.

[16] Power system test case archive. Technical report, Univ. of Washington, http://www.ee.washington.edu/research/pstca/.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Communication-Efficient Design of Learning System for Energy Demand Forecasting of Electrical Vehicles

Jiacong Xu*, Riley Kilfoyle†, Zixiang Xiong† and Ligang Lu‡
*Dept of CS, Johns Hopkins University, Baltimore, MD 21218
†Dept of ECE, Texas A&M University, College Station, TX 77843
‡Shell International Exploration and Production Inc., Houston, TX 77082
jxu155@jhu.edu; rskilfoyle@tamu.edu; zx@ece.tamu.edu; ligang.lu@shell.com

***Abstract*—Machine learning (ML) applications to time series energy utilization forecasting problems are a challenging assignment due to a variety of factors. Chief among these is the non-homogeneity of the energy utilization datasets and the geographical dispersion of energy consumers. Furthermore, these ML models require vast amounts of training data and communications overhead in order to develop an effective model. In this paper, we propose a communication-efficient time series forecasting model combining the most recent advancements in transformer architectures implemented across a geographically dispersed series of EV charging stations and an efficient variant of federated learning (FL) to enable distributed training. The time series prediction performance and communication overhead cost of our FL are compared against their counterpart models and shown to have parity in performance while consuming significantly lower data rates during training. Additionally, the comparison is made across EV charging as well as other time series datasets to demonstrate the flexibility of our proposed model in generalized time series prediction beyond energy demand. The source code for this work is available at https://github.com/XuJiacong/LoGTST_PSGF.**

I. Introduction

The Alternative Fuels Data Centre lists more than 54,000 EV charging stations (CSs) currently in operation in the United States. It is projected that the number of EVs on the road will rise by more than 4000% by 2030 [1]. As a result, there is a strong need to effectively and intelligently predict energy demands for EV CSs so as to mitigate their impact on the power grids without upgrading/expansion capability. Generally, the power grid supplies the energy for CSs once requests from EVs are received. Predicting the amount of energy needed at each CS across different periods of time will allow the power suppliers of CSs to purchase the desired amount of electricity at lower rates in order to save money and make charging at public stations more efficient – a key step towards smart charging [2]. In addition, the power grid can coordinate energy consumption via schedule management to reduce energy costs and waste of resources.

Energy utilization is a form of time series prediction algorithm that has been thoroughly studied in the past. Majidpour et al. [3] compared fast machine learning-based time-series prediction algorithms and found that the nearest neighbor algorithm showed improved accuracy. Ryu et al. [4] proposed deep learning (DL) load forecasting models and showed that DL methods exhibited better performance compared to other forecasting models. Paterakis et al. [5] compared a DL method with the eight most commonly used machine learning methods such as nearest neighbors, support vector machines, Gaussian processes, and regression trees, in time series energy prediction and showed that the DL method outperformed all eight other methods. However, conventional DL algorithms trained on individual CS nodes may not achieve high prediction accuracy due to insufficient data and failure to consider the influence of other CSs, particularly the ones in the neighborhood [6].

We introduce machine learning-based approaches that can not only improve the accuracy of time series prediction but also significantly reduce the communication overhead involved in training a machine learning algorithm involving a dispersed series of nodes. In particular, we first introduce a communication model using the power provider as a centralized node to gather all information from the individual CS nodes in a considered metropolitan area (e.g., Houston, or London, or Amsterdam, etc.). We then develop an ML algorithm to help the power provider more accurately predict energy demands for the CSs while simultaneously utilizing less data than previous algorithms. Our focus thus is two-fold. First to more accurately predict energy usage at CSs and second to lower the data utilization required to implement the prediction algorithm across a broad collection of CSs spread out geographically.

Specifically, we adopt a federated learning (FL) approach [7], [8] to energy demand prediction. As seen in Fig. 1, FL is a form of distributed learning wherein updates to model parameters are calculated locally on individual devices rather than transmitting massive volumes of data to centralized servers for global model development. These local model updates are then transmitted to a centralized server for aggregation into a global model. Once the global model has been updated according to various algorithms, the updated global model is transmitted out to the CSs for the next round of training. In this fashion the individual data never leaves the local CSs, some of the computation is offloaded to local CSs, and a globally

Work supported in part by NSF grants ECCS-1923803, CCF-2007527, CCF-2324397, and Shell.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

useful DL model is created. This reduces the increasing strain placed on communication networks while accomplishing the goal of training a useful global model. Thus, the advantage with FL is that the CSs only need to share their trained model parameters obtained from their individual local datasets instead of sharing their entire datasets, which generally are of large volume and often are protected by data privacy laws and security regulations.

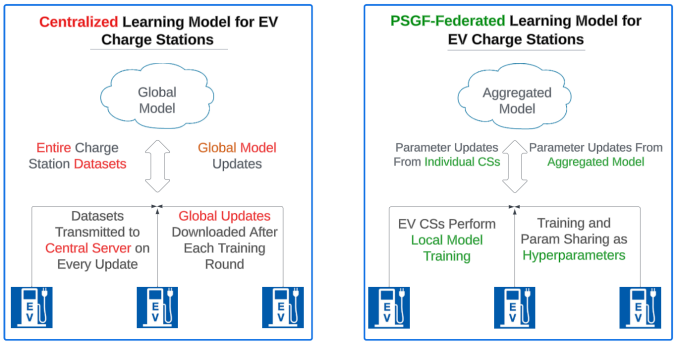


Fig. 1. FL architectures vs Centralized: In centralized learning (left), data is sent to the cloud, where the ML model is built. The model is accessed by a user through an API sending a request to access one of the available services. In FL [7,8] (right), each CS trains a model and sends only its updated parameters to a server for aggregation. Data is kept locally on devices and knowledge is shared selectively with peers through an aggregated model.

The EV charging prediction task for an individual client is a uni-variate time-series forecasting problem, which has been previously investigated. For example, using the UK EV dataset [9], Saputra et al. [10] showed that simple location-based node clustering improves the accuracy of energy demand prediction up to 24.63% and decreases communication overhead by 83.4% compared with other baseline machine learning algorithms. Different from previous works [6], [10] that treat the hourly energy consumption as the prediction output, we aggregate 24 hours of energy consumption as one data point upon which to perform daily prediction. This time aggregation is done to assist in dealing with inconsistencies present in various portions of the UK EV dataset [9] wherein certain chargers were offline for maintenance etc. at random times and therefore present discontinuities on an hourly level, whereas the daily data exhibits more clear trends of energy consumption variation and less meaningless randomness.

Inspired by the great success of transformer architectures for language and vision tasks, many advanced transformer architectures have been proposed for time series forecasting in the last two years. These transformer-based models have demonstrated better prediction performance compared with traditional RNNs and recent work PatchTST [11] outperforms its multi-layer perceptron (MLP) counterpart and shows state-of-the-art performance when tested on standard time-series forecasting benchmark datasets. In this paper, we incorporate the ideas from recent advances in efficient 2-D vision transformers and propose an improved lighter model called Local and then Global Time Series Transformer (LoGTST).

For generic FL, sharing model parameters among all clients and the server will involve heavy communications overhead costs which has the unwanted byproduct of slower convergence in real-time. To reduce the communication cost and convergence time, Online-Fed randomly down-selects a specific set of CS clients for model updating. Partial Sharing-based Online FL (PSO-Fed) [12] further alleviates this problem by only randomly sharing a partial subset of model parameters to those selected clients. However, in each global iteration, the clients in both of these two models can only access their own information to train on or stay idled not training during this particular iteration, which will hinder the convergence speed and limit that specific CS's generalization ability. Thus, we propose Partial Sharing Global Forwarding FL (PSGF-Fed) that is built upon PSO-Fed but enables the server to randomly share a small number of partial parameters to all the clients thus enabling all CSs to train during each iteration. In this way, PSGF-Fed is able to reduce the total communication overhead by accelerating the convergence speed of each local model so that fewer total iterations of training are required.

Through the generated results, we demonstrate that our proposed novel LoGTST architecture can achieve similar prediction performance with current baseline model: PatchTST [11] but involves around half of its number of trainable parameters, and our PSGF-Fed updating strategy offers better trade-off between prediction accuracy and communication cost compared with PSO-Fed [12] and Online-Fed. Also, we present additionally the flexibility via hyper-parameters between model performance and data overhead. This serves to demonstrate that our combination of LoGTST prediction model and PSGF-Fed distributed training strategy constitutes an efficient ML pipeline for energy demand forecasting of electrical vehicles.

II. Method

The learning system in this application consists of two major parts: the DL model and the FL strategy. For better prediction accuracy, the DL model should be able to capture the local and global trends of the input time series, and the FL policy should aggregate the information from all the clients and lead the model to a quick convergence with minimal bandwidth consumption. With the constraints of communication overhead driving larger parameter sets, the space complexity of the learning model should be as small as possible and the time and the amount of information sharing between server and clients should be minimized.

A. MetaFormer

Current research on the model architecture design for time series forecasting can be divided into two camps: Transformer and MLP. Different from previous transformer architectures, PatchTST [11] splits a time-series into patches and tokenizes these patches into vectors by patch embedding following Vision Transformer (ViT) [13]. This new transformer outperforms MLPs and other transformers such as DLinear [14] and FEDFormer [15] on many popular benchmarks.

The self-attention operation [16] in traditional transformers possesses the advantages of global dependency parsing and dynamic weight generation. However, self-attention brings a

large amount of space and time complexity to the overall model. The authors of MetaFormer [17] summarize recent advances on vision transformer and argue that it is the transformer architecture itself that contributes to the performance of ViT on vision tasks. They also replace the self-attention operation with a simple pooling operation and demonstrate that PoolFormer shows better efficiency than most recent variants of ViT considering the complexity and the accuracy.

Similar to PatchTST, we introduce the MetaFormer architecture into the task of 1-D time series forecast. As shown in Figure 2, we replace self-attention with Time-MLP and Identity operations as the token-mixer and propose two variants of MetaFormer: MLPFormer and IDFormer. In our experiments, we observed the surprising finding that the simplest identity operation performs as well as PatchTST. Since our goal is to reduce the model complexity while minimizing loss of accuracy, we favor most of the transformer blocks to be the simplest IDFormer in our final model.

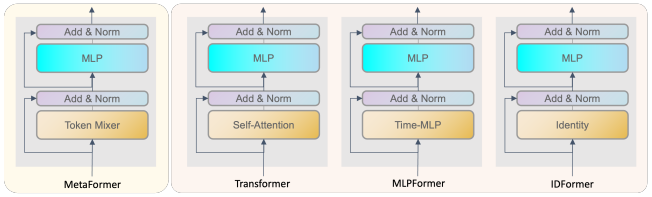


Fig. 2. Architectures of MetaFormer and its variants. Time-MLP refers to the MLP operation along the series of tokens; Identity means there is no operation.

B. LoGTST

Figure 3 depicts the basic architecture of our proposed LoGTST model.

RevIN. The Reversible Instance Normalization (RevIN) [18] module normalizes the input signal for each sample, which consists of one look-back window and one prediction horizon, and records the normalization factors for future denormalization of the output signal. In this way, RevIN is able to symmetrically remove and restore the statistical information of a time-series instance and improve the model's stability.

Tokenization and DeTokenization. The tokenization process is the same as Patch Embedding in PatchTST [11]. Here we call it Tokenization to follow the convention in computer vision and natural language processing. Given the input uni-variate time series as $\mathbf{x}$ and its length as L, the tokenization process can be simply accomplished by applying 1-D convolution on $\mathbf{x}$ with a predefined kernel of size P and stride S. The kernel size can also be seen as the patch length and the number of tokens is $N = [L/S]$. The DeTokenization process directly flattens feature vectors $\mathbf{x}_{hidden}^{(i)}$, where $i = 1, 2, ..., N$, concatenates the flattened vectors, and applies MLP for prediction, which can be represented as

$$\text{Pred} = \text{MLP}\{\text{Concat}[\text{Flat}(\mathbf{V}_0), \text{Flat}(\mathbf{V}_1), ...]\} \quad (1)$$

IDFormer and Transformer: In the transformer branch, we replace the early two transformer blocks by IDFormer blocks to reduce total model complexity and prevent early attention on naive features and keep the last transformer block to parse dependencies between hidden vectors. The patch embeddings $\mathbf{x}_p \in \mathbb{R}^{D\times N}$ should be merged with additive learnable positional encoding $\mathbf{W}_{pos}$ by $\mathbf{x}_d = \mathbf{x}_p + \mathbf{W}_{pos}$ in case of the attention mechanism treating all the feature vectors equally. For each head $h = 1, 2, ..., H$, we define three learnable matrices: $W_h^Q, W_h^K \in \mathbb{R}^{D\times d_k}$ and $W_h^V \in \mathbb{R}^{D\times D}$. Then, the calculation of multi-head self-attention can be written as:

$$O_h^T = \text{Attention}(Q_h, K_h, V_h) = \text{softmax}\{\frac{Q_h K_h^T}{\sqrt{d_k}}\}V_h \quad (2)$$

Loss: Following previous work for time series forecasting [11], [15], we also use MSE to measure the discrepancy between the prediction and the reality. Defining the number of input variables as M and the prediction horizon as T, then the total loss can be calculated by $\mathcal{L} = 1/M \sum_{i=1}^{M} \|\hat{\mathbf{x}}_{L+1:L+T}^{(i)} - \mathbf{x}_{L+1:L+T}^{(i)}\|^2$. Note that there is only one channel ($M = 1$) for the task EV charging forecasting.

C. PSGF-Fed

The difference between our proposed PSGF-Fed and previous works is illustrated in Figure 4.

Online-Fed. Instead of exchanging the model parameters with all the clients, the server of Online FL (Online-Fed) will randomly select a subset of clients in every iteration for communication efficiency. Denote the selected subset of client indices as S_n, where $C = |S_n|$ refers to the number of selected clients and n represents the current iteration. There is no need to perform local updates for unselected clients because the local model will be directly replaced by the server model when these clients are selected in future iterations. Define the local model parameters for client i in iteration n after local update as $\mathbf{w}_{n+1}^i$, then the global model can be updated by:

$$\mathbf{w}_{n+1} = \frac{1}{C}\sum_{i\in S_n} \mathbf{w}_{n+1}^i \quad (3)$$

PSO-Fed [12]. It reduces the granularity of communication to the parameter level for better efficiency and randomness. PSO-Fed randomly selects a subset of clients S_n and also randomly selects a subset of parameters $\mathbf{S}_n^i$ for client i for parameter exchange between server and clients. Different from Online-Fed, the unselected clients can still update their local models because when they are selected, not all of the local parameters will be replaced. The $\mathbf{S}_n^i$ can be a $D \times D$ diagonal matrix with M ones for selected diagonal elements and $D-M$ zeros. Then the selected local model can be updated by:

$$\mathbf{w}_{n+1}^i = \text{LocalUpdate}(\mathbf{S}_n^i \mathbf{w}_n + (\mathbf{I}_D - \mathbf{S}_n^i)\mathbf{w}_n^i) \quad (4)$$

The server generates the global model by aggregating the local parameters and this can be accomplished by:

$$\mathbf{w}_{n+1} = \frac{1}{C}\sum_{i\in S_n} \mathbf{S}_{n+1}^i \mathbf{w}_{n+1}^i + (\mathbf{I}_D - \mathbf{S}_{n+1}^i)\mathbf{w}_n \quad (5)$$

The communication overhead is then significantly reduced compared with Online-Fed while maintaining comparative convergence speed.



Fig. 3. Overview of the architecture of our proposed **LoGTST**. The input and output signals are processed by Reversible Instance Normalization (RevIN) module before and after the learning model, respectively. For fair comparison with PatchTST, we only change the first and second transformer blocks to IDFormer blocks so that the model can fully process the local features and keep the final transformer blocks for parsing of global dependency.

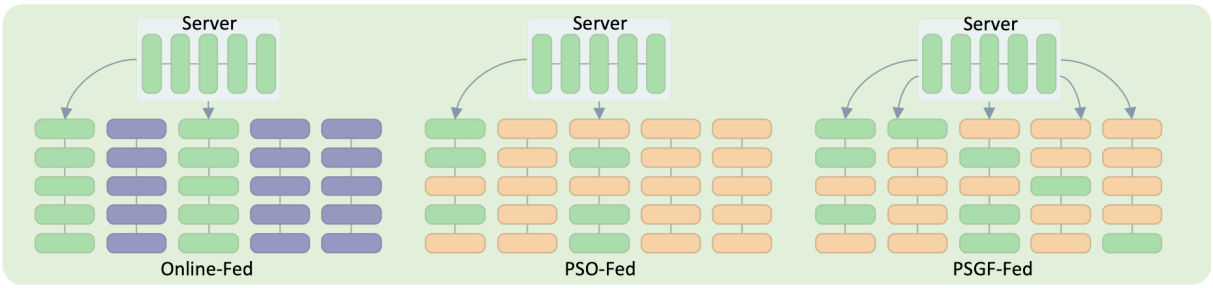


Fig. 4. Illustration of the model parameters' sharing from server to clients for Online-Fed, PSO-Fed [12], PSGF-Fed (ours). The green rectangles refer to shared parameters. The purple rectangles will remain idle during the following local update while the orange and green rectangles will be updated.

PSGF-Fed. PSO-Fed enables self-learning for each unselected client, which guarantees its fast convergence speed even though only partial parameters are shared. However, training a local model only on limited local data for many local iterations (or even several global iterations) without any global information has the risk of overfitting or weakening the generalization ability. To alleviate this issue, we propose a new online FL policy, namely PSGF-Fed, where the server will randomly select a small subset of model parameters and share them with each client. In this way, all the clients will receive some global information from the server, which will regularize the local training process. Define $\mathbf{F}_n^i$ as the parameter forwarding diagonal matrix of unselected client i with N ones for diagonal elements (selected forwarding parameters) and $D-N$ zeros, then the local training for this client can be represented by

$$\mathbf{w}_{n+1}^i = \text{LocalUpdate}(\mathbf{F}_n^i \mathbf{w}_n + (\mathbf{I}_D - \mathbf{F}_n^i)\mathbf{w}_n^i) \qquad (6)$$

The number of selected forwarding parameters (N) for each client can be adjusted to reach the best trade-off between convergence speed and communication overhead. All the selected local models will still be updated by (4) following PSO-Fed.

III. Experiments

A. Centralized Time Series Forecasting

1) Datasets: Following previous works, we evaluate our model performance on five public-domain datasets: Weather and four electricity transformer temperature (ETT) datasets (ETTh1, ETTh2, ETTm1, ETTm2) [19], which contain multivariate time series with more than 10k timesteps. These five datasets have been extensively utilized for benchmarking time series forecasting models and are introduced in detail by [20]. Since our ultimate goal is to deal with the prediction of univariate EV charging task, the forward process for each channel uni-variate series is independent of LoGTST and the model weights are shared between different channels.

2) Setting: For fair comparison, we build our model based on the codebase of PatchTST [11], FEDformer [15], and Autoformer [20] to guarantee the same experimental setup. Specifically, we train the model by stochastic gradient descent using Adam optimizer [21] and adjust the learning rate according to the cycle learning rate policy [22]. The total number of epochs is 100, but the training process will be early stopped with patience of 20 epochs to prevent overfitting. We also include previous representative models Informer [19] and Pyraformer [23] in our result comparison.

3) Results: Results in Table I show the accuracy of our LoGTST design when compared against other industry-leading architectures while functioning with less than half of the parameters available to most other models. Comparison is made across five different datasets using both the mean square error and mean absolute error. In both error measurements, across varying lengths of time prediction and across multiple time series datasets the LoGTST consistently improves upon or nearly matches the PatchTST/64 design which proved the closest competition. LoGTST does so while having only 45% the available parameters of PatchTST/64 and 58% of the parameters of PatchTST/42. When examined more closely, the toughest competition is in the area of shorter-term look-ahead where the metric is held to only 96. In this case, our LoGTST was marginally outperformed by the PatchTST networks but only by MSE of less than 0.01 and only in 3/5 cases. When the time length metric was increased to a longer look-ahead of 192, 336, or 720, the LoGTST is superior in 10/15 cases and is within 0.004 or less MSE across all other five.

B. Distributed Time Series Forecasting

1) Datasets: The EV dataset [9] was obtained from charging stations in Dundee city, the UK during 2017-2018, which has 65,601 transactions that include IDs from 58 CSs, transaction ID for each CS, EV charging date, EV charging time, and consumed energy (in kWh) for each transaction. There are

TABLE I

EXPERIMENTAL RESULTS OF LOGTST (OURS) AND OTHER RECENTLY PROPOSED MODELS FOR CENTRALIZED TIME SERIES FORECASTING TASK. WE CALCULATE MEAN SQUARE ERROR (MSE) AND MEAN ABSOLUTE ERROR (MAE) OF THE PREDICTED SERIES AND GROUND TRUTH FOR ALL THE MODELS TO QUANTIFY THEIR PERFORMANCE. FOLLOWING PREVIOUS WORKS, WE ALSO CHANGE THE PREDICTION LENGTH TO BE 96, 192, 336, 720 TO FULLY EXAMINE THE LONG-TIME SERIES FORECASTING PERFORMANCE. NOTE THAT PATCHTST/64 HAS A LONGER LOOK-BACK WINDOW THAN PATCHTST/42.

Models		LoGTST		PatchTST/64 [11]		PatchTST/42 [11]		FEDformer [15]		Autoformer [20]		Informer [19]		Pyraformer [23]	
#Parameters		**5.39E+05**		1.19E+06		9.21E+05		1.63E+07		1.05E+07		1.13E+07		1.01E+07	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.151	0.199	**0.149**	**0.198**	0.152	0.199	0.238	0.314	0.249	0.329	0.354	0.405	0.896	0.556
	192	0.195	**0.240**	**0.194**	0.241	0.197	0.243	0.275	0.329	0.325	0.370	0.419	0.434	0.622	0.624
	336	0.246	**0.280**	**0.245**	0.282	0.249	0.283	0.339	0.377	0.351	0.391	0.583	0.543	0.739	0.753
	720	0.318	**0.333**	**0.314**	0.334	0.320	0.335	0.389	0.409	0.415	0.426	0.916	0.705	1.004	0.934
ETTh1	96	0.379	0.404	**0.370**	0.400	0.375	**0.399**	0.376	0.415	0.435	0.446	0.941	0.769	0.664	0.612
	192	**0.413**	**0.421**	**0.413**	0.429	0.414	**0.421**	0.423	0.446	0.456	0.457	1.007	0.786	0.790	0.681
	336	0.426	**0.431**	**0.422**	0.440	0.431	0.436	0.444	0.462	0.486	0.487	1.038	0.784	0.891	0.738
	720	**0.447**	**0.463**	**0.447**	0.468	0.449	0.466	0.469	0.492	0.515	0.517	1.144	0.857	0.963	0.782
ETTh2	96	0.275	**0.336**	**0.274**	0.337	**0.274**	**0.336**	0.332	0.374	0.332	0.368	1.549	0.952	0.645	0.597
	192	**0.338**	**0.379**	0.341	0.382	0.339	0.379	0.426	0.446	0.426	0.434	3.792	1.542	0.788	0.683
	336	**0.327**	0.381	0.329	0.384	0.331	**0.380**	0.477	0.447	0.477	0.479	4.215	1.642	0.907	0.747
	720	**0.378**	**0.421**	0.379	0.422	0.379	0.422	0.453	0.469	0.453	0.490	3.656	1.619	0.963	0.783
ETTm1	96	**0.288**	**0.342**	0.293	0.346	0.290	**0.342**	0.326	0.390	0.510	0.492	0.626	0.560	0.543	0.510
	192	**0.331**	0.370	0.333	0.370	0.332	**0.369**	0.365	0.415	0.514	0.495	0.725	0.619	0.557	0.537
	336	**0.360**	**0.390**	0.369	0.392	0.366	0.392	0.392	0.425	0.510	0.492	1.005	0.741	0.754	0.655
	720	**0.416**	0.425	**0.416**	**0.420**	0.420	0.424	0.446	0.458	0.527	0.493	1.133	0.845	0.908	0.724
ETTm2	96	**0.163**	**0.253**	0.166	0.256	0.165	0.255	0.180	0.271	0.205	0.293	0.355	0.462	0.435	0.507
	192	0.221	0.293	0.223	0.296	**0.220**	**0.292**	0.252	0.318	0.278	0.336	0.595	0.586	0.730	0.673
	336	0.278	0.330	**0.274**	**0.329**	0.278	**0.329**	0.324	0.364	0.343	0.379	1.270	0.871	1.201	0.845
	720	0.366	**0.383**	**0.362**	0.385	0.367	0.385	0.410	0.420	0.414	0.419	3.001	1.267	3.625	1.451

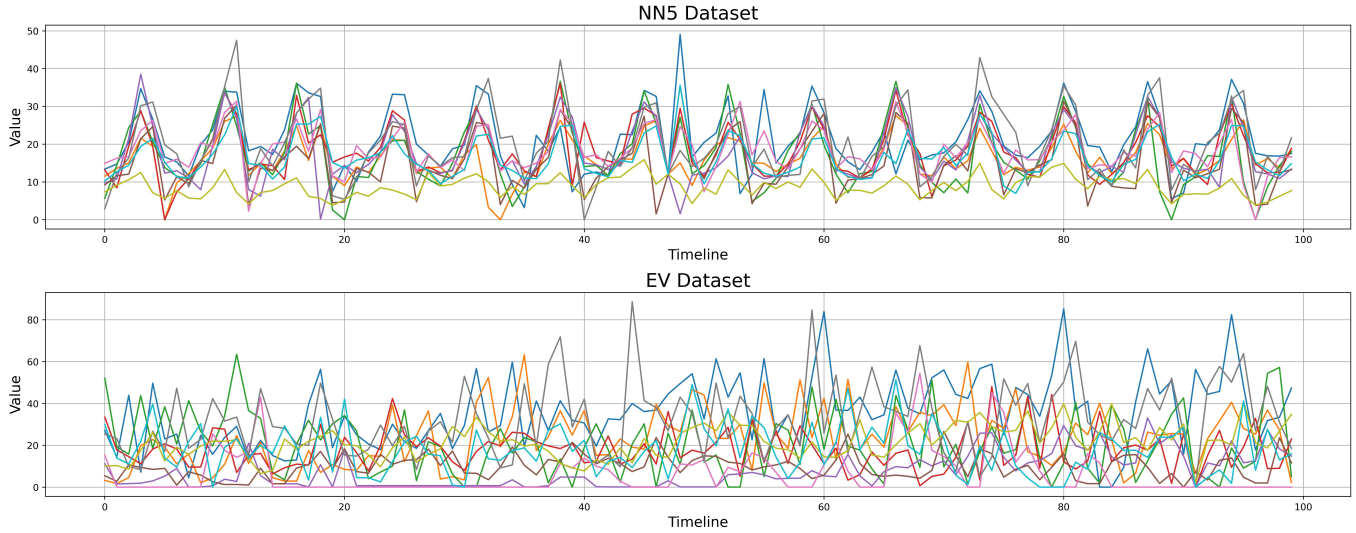


Fig. 5. Overview of the data for NN5 and EV charging datasets. We randomly select 10 clients from day 0 to day 99 of each dataset for data visualization. We can notice that the pattern of NN5 is much more obvious than EV charging dataset, where some of the data points are missing for some clients.

a large amount of missing data in the dataset, so we clean the entire dataset by removing the charging stations that stopped providing data during the period. Different from previous works [6], [10], we merge the 24-hour energy consumption into one day and focus on daily prediction since most of the hourly data is zero and involves much more randomness compared with daily data. As shown in Figure 5, even after the data cleaning, the data quality for EV dataset [9] is still quite low, which may not be able to fully benchmark the performance of current models. Thus, we also conduct experiments on NN5 dataset [24], which as shown in Figure 5 is of high quality and has a clear seasonal pattern to demonstrate the generalization ability of our method. The NN5 data consists of two years of daily cash money demand at various automatic teller machines (ATMs or cash machines) at different locations in England.

2) Setting: Following previous work [6], [10], all the clients are clustered using K-means clustering algorithm based on the distances measured by dynamic time warping (DTW) [25]. Then, the FL process are conducted independently between different clusters. The look-back window is set to be 128 steps and the prediction horizons are scheduled to be 4 and 2 for NN5 and EV datasets considering their dataset sizes. We still

use Adam optimizer with the initial learning rate to be 10^{-3}. The ratio of selected clients is set to be 50% for all three methods: Online-Fed, PSO-Fed, and PSGF-Fed. Both of the ratio of forward parameters of PSGF-Fed and the ratio of sharing parameters of PSO-Fed and PSGF-Fed are adjusted accordingly to find t heir o ptimal p erformance. T he training process will be stopped when the model reaches convergence (the training loss stops decreasing for 10 rounds). RMSE is used to quantify the prediction performance.

3) Results (NN5): As seen in Table II, when trained and tested against the NN5 time series dataset, the initial online FL, data-intensive training performs well with an MSE of 6.02 but at the cost of transferring approximately 1.5 billion parameters during training. The previous generation PSO FL architecture achieves slightly worse performance of 6.10 while only transferring 0.75 billion parameters. Our PSGF FL model achieves 6.08 MSE while only transferring 0.38 billion parameters under the 30/30 hyper-parameter implementation. This is a significant i mprovement i n c ommunication c ost overhead with a marginal performance improvement in addition when compared to the PSO FL model.

Figure 6 depicts the trade-off between communication cost and prediction loss. It is seen that With similar prediction accuracy, PSGF-Fed reduces the communication cost by at least 25% of PSO-Fed for the NN5 dataset.

TABLE II
EXPERIMENTAL RESULTS OF DIFFERENT FEDERATED LEARNING POLICIES FOR NN5 DATASET. THE SECOND COLUMN REFERS TO THE RATIO OF SHARED PARAMETERS BETWEEN SELECTED CLIENTS AND THE SERVER. THE RESULTS FOR THE LOWEST COMMUNICATION COST AND LOSS ARE FOLDED AND THE RESULTS FOR THE BEST TRADE-OFF ARE UNDERLINED.

Method		#Params (Comm.)	Loss (RMSE)
Online-Fed		1.53E+09	**6.02**
PSO-Fed [12]	50%	7.35E+08	6.10
	40%	4.03E+08	6.14
	30%	3.52E+08	6.15
	20%	**1.24E+08**	6.28
PSGF-Fed-20%	50%	5.75E+08	6.10
	40%	4.21E+08	6.10
	30%	3.45E+08	6.09
	20%	2.80E+08	6.14
PSGF-Fed-30%	50%	5.21E+08	6.11
	40%	5.25E+08	6.10
	30%	3.77E+08	6.08
	20%	3.18E+08	6.11

4) Results (EV): Similar to the NN5 results, the PSO federate learning is able to reduce the number of parameters passed during training by approximately 50 percent while maintaining performance within 2 percent of the original online federate learning model. As shown in Table III, our PSGF architecture is demonstrated to reduce passed parameter numbers by 77 percent while maintaining performance similar to that of the PSO model. Additionally, when given a matched communication budget as the PSO model, our PSGF architecture is able to outperform the PSO by 0.25 MSE. This shows the flexibility and robustness of our improved PSGF model in data overhead as well as performance.

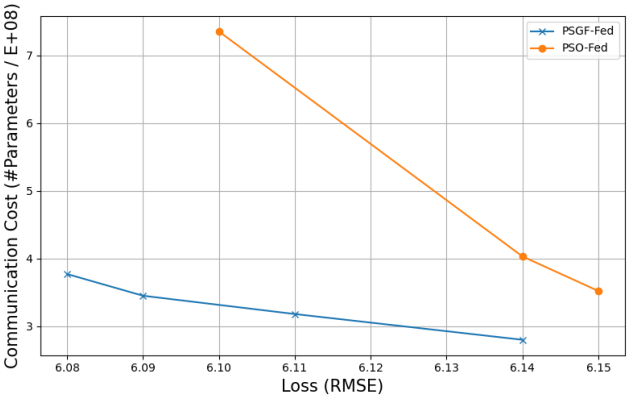


Fig. 6. The trade-off between communication cost and prediction loss (bottom left is better). With similar prediction accuracy, PSGF-Fed reduces the communication cost by at least 25% of PSO-Fed.

TABLE III
EXPERIMENTAL RESULTS OF DIFFERENT FEDERATED LEARNING POLICIES FOR UK EV DATASET. THE SECOND COLUMN REFERS TO THE RATIO OF SHARED PARAMETERS BETWEEN SELECTED CLIENTS AND THE SERVER. THE RESULTS FOR THE LOWEST COMMUNICATION COST AND LOSS ARE FOLDED AND THE RESULTS FOR THE BEST TRADE-OFF ARE UNDERLINED.

Method		#Params (Comm.)	Loss (RMSE)
Online-Fed		9.07E+06	10.46
PSO-Fed [12]	50%	4.84E+06	10.68
	40%	3.77E+06	10.89
	30%	2.75E+06	10.85
	20%	2.11E+06	11.14
PSGF-Fed-20%	50%	4.82E+06	**10.43**
	40%	4.28E+06	10.54
	30%	2.96E+06	10.67
	20%	2.11E+06	10.64
PSGF-Fed-30%	50%	4.46E+06	10.64
	40%	4.18E+06	10.63
	30%	3.28E+06	10.65
	20%	**2.08E+06**	10.60

IV. CONCLUSIONS

In this paper, we have proposed a communication-efficient time series forecasting model by combining LoGTST in transformer design and a PSGF-Fed strategy in FL. Simulations have demonstrated that our LoGTST trained with PSGF-Fed offers the best trade-off between communication overhead and performance for energy demand forecasting of a geographically dispersed series of EV charging stations. Tests of our proposed forecasting model on other time series datasets also showed superior performance over existing approaches, making it applicable to generalized time series prediction beyond energy demand for EV charging.

V. ACKNOWLEDGEMENTS

Z. Xiong would like to acknowledge fruitful discussions with Michail Gkagkos.

REFERENCES

[1] "The growth of electric vehicles," *https://www.cnbc.com/2018/05/30/electric-vehicles-will-grow-from-3-million-to-125-million-by-2030-iea.html*,

[2] M. Muratori, "Role of electric vehicles in the U.S. power sector transition," in *Keynote Session, 4-th E-Mobility Power System Integration Symposium*, November 3rd, 2020.

[3] M. Majidpour, C. Qiu, P. Chu, R. Gadh, and H. R. Pota, "Fast prediction for sparse time series: Demand forecast of ev charging stations for cell phone applications," *IEEE Trans. on Industrial Informatics*, vol. 11, no. 1, pp. 242–250, 2014.

[4] S. Ryu, J. Noh, and H. Kim, "Deep neural network based demand side short term load forecasting," *Energies*, vol. 10, no. 1, p. 3, 2017.

[5] N. G. Paterakis, E. Mocanu, M. Gibescu, B. Stappers, and W. van Alst, "Deep learning versus traditional machine learning methods for aggregated energy demand prediction," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pp. 1–6, 2017.

[6] D. Perry, N. Wang, and S. Ho, "Energy demand prediction with optimized clustering-based federated learning," in *2021 IEEE Globecom Conference*, pp. 1-6, 2021.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[8] S. Abdulrahman, H. Tout, O. Hanine, *et al.*, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 8, pp. 5476-5497, Oct. 2020.

[9] "Electric vehicle charging sessions dundee," *https://data.dundeecity.gov.uk/dataset/ev-charging-data*,

[10] Y. M. Saputra, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, M. D. Mueck, and S. Srikanteswara, "Energy demand prediction with federated learning for electric vehicle networks," in *2019 IEEE Globecom Conference*, pp. 1-6, 2019.

[11] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations*, 2023.

[12] V. C. Gogineni, S. Werner, Y.-F. Huang, and A. Kuh, "Communication-efficient online federated learning strategies for kernel regression," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4531–4544, 2022.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[14] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" In *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 11 121–11 128.

[15] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International Conference on Machine Learning*, PMLR, 2022, pp. 27 268–27 286.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] W. Yu, M. Luo, P. Zhou, *et al.*, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 819–10 829.

[18] T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *International Conference on Learning Representations*, 2021.

[19] H. Zhou, S. Zhang, J. Peng, *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 11 106–11 115.

[20] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 419–22 430, 2021.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, SPIE, vol. 11006, 2019, pp. 369–386.

[23] S. Liu, H. Yu, C. Liao, *et al.*, "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *International conference on learning representations*, 2021.

[24] "NN5 forecasting competition," *http://www.neural-forecasting-competition.com/NN5/index.htm*,

[25] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Radiated Sound Field Reproduction for Surrounding Loudspeaker Array Based on Higher-order Ambisonics

Shota Naiki*, Shumpei Miura*, Kenta Iwai*, Takanobu Nishiura* Yoshiharu Soeta†
* Ritsumeikan University, Shiga, Japan
E-mail: {is0514hi@ed, is0431vp@ed, iwai@fc, nishiura@is}.ritsumei.ac.jp
† National Institute of Advanced Industrial Science and Technology (AIST), Osaka, Japan
E-mail: y.soeta@aist.go.jp

***Abstract*—This paper proposes a radiated sound field reproduction method for a surrounding loudspeaker array. The conventional method controls the direction of the virtual sound image by a vector base amplitude panning (VBAP). The real sound source like human speech and musical instruments has a directivity pattern and should be considered to reproduce the sound field. However, the VBAP-based method cannot represent the directivity pattern because VBAP only controls the amplitude of an audio signal. In this paper, we propose the radiated sound field reproduction method for a surrounding loudspeaker array based on higher-order ambisonics (HOA). Based on HOA, the proposed method calculates the driving function of the virtual sound source by using the directivity pattern of the virtual sound source. Simulation results show that the proposed method achieves to represent the radiation sound with a normalized reproduction error of -20 dB in the target region.**

I. Introduction

Recently, the demand for highly realistic sound field reproduction systems has been increasing with the development of imaging technologies. 22.2 multichannel audio [2] is one of the sound field reproduction systems and consists of a surrounding loudspeaker array. In 22.2 multichannel audio, vector base amplitude panning (VBAP) [3] is used to control the direction of a three-dimensional (3-D) sound image. VBAP controls the direction of the sound image by selecting three loudspeakers from a surrounding loudspeaker array and weighing each speaker. However, VBAP controls only amplitude, and is difficult to reproduce a radiated sound image. To solve this problem, Iwai et al., proposed the 3-D sound image reproduction method based on spherical harmonic expansion [1]

The directivity pattern of a real sound source changes depending on the frequency and is important to percept the direction of sound [4]. Therefore, it is required to consider the frequency-dependent directivity of the virtual sound image for the reproduction of the radiation sound.

Radiating speaker arrays consisting of multiple loudspeaker elements mounted on a polyhedron or sphere enclosure are conventionally used to control the radiation directivity characteristics [5]. However, it is difficult to reproduce a radiated sound image using a surrounding loudspeaker array.

In this paper, we propose a system to construct a virtual sound source that has desired radiation characteristics in the 22.2 multichannel sound system. The proposed method calculates the driving function of the virtual sound source by using the directivity pattern of the virtual sound source based on higher-order ambisonics (HOA) [6]. We evaluate reproducibility by computer simulation.

II. Proposed method for radiated sound field reproduction for surrounding loudspeaker array based on higher-order ambisonics

A. Outline

In this paper, we propose radiated sound field reproduction for the surrounding loudspeaker array. A 3-D sound field can be represented by spherical harmonic expansion [7]. First, the proposed system calculates the coefficients of the spherical harmonic expansion which represent desired radiation characteristics. Then, the proposed system calculates sound pressure on each microphone when we locate a spherical microphone in the radiation sound in a computer simulation. Then, the proposed system calculates the driving signals of each loudspeaker which reproduces radiation sound based on HOA from sound pressures on each microphone. In this section, we state in detail.

B. Spherical harmonic expansion

The spherical harmonic function $Y_{nm}(\theta, \phi)$ is shown as the general solution of the 3-D wave equation and expressed by the following equation [7]:

$$Y_{nm}(\theta,\phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta) e^{jm\phi}, \quad (1)$$

here, $\theta(0 \leq \theta \leq \pi)$ is polar angle, $\phi(0 \leq \phi \leq 2\pi)$ is azimuth, $n(0 \leq n \leq \infty)$ is degree of spherical harmonic function, $P_{nm}(\cdot)$ is Lejandre's function, j is imaginary unit. Figure 1 shows the real part of the spherical harmonic function. The spherical harmonic function is an orthogonal function. Therefore, it is possible to express a sound field as expansion coefficients using the spherical harmonic function. In the case

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

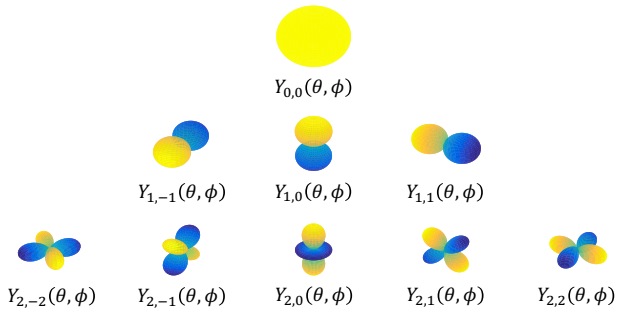


Fig. 1: Real part of spherical harmonics

where the sound source is outside the radius r, one of the solutions for the wave equation is represented with polar coordinates expressed by the following equation:

$$p(\boldsymbol{r},k)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}A_{nm}^{\text{int}}(k)y_n(kr)Y_{nm}(\theta,\phi), \tag{2}$$

here, $\boldsymbol{r}=(r,\theta,\phi)$, k is the wavenumber, $A_{nm}^{\text{int}}(k)$ is the expansion coefficient of the internal acoustic field, $y_n(kr)$ is the spherical Bessel function. When the sound pressure $p(r,k)$ is known at $r=a$, $A_{nm}{}^{\text{int}}(k)$ is expressed by the following equation:

$$A_{nm}^{\text{int}}(k)=\int_0^{\pi}\int_0^{2\pi}\frac{p(a,\theta,\phi,k)}{y_n(ka)}Y_{nm}(\theta,\phi)^*\sin\theta d\theta d\phi, \tag{3}$$

here, $(\cdot)^*$ is the complex conjugate. Then, the radiated sound field where the sound source is inside of the radius r is expressed by the following equation:

$$p(\boldsymbol{r},k)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}A_{nm}^{\text{ext}}(k)h_n(kr)Y_{nm}(\theta,\phi), \tag{4}$$

here, $A_{nm}^{\text{ext}}(k)$ is the expansion coefficient of radiated sound field. Assuming that sound pressure $p(\boldsymbol{r},k)$ is known at $r=b$, A_{nm}^{ext} is expressed by following equation:

$$A_{nm}^{\text{ext}}(k)=\int_0^{\pi}\int_0^{2\pi}\frac{p(b,\theta,\phi,k)}{h_n(kb)}Y_{nm}(\theta,\phi)^*\sin\theta d\theta d\phi. \tag{5}$$

C. Higher-order ambisonics (HOA)

In HOA, the desired sound field is reproduced in the reproduction field by HOA encoding and decoding based on spherical harmonic expansion.

The procedure of HOA encoding is as follows. First, we estimate $A_{nm}^{\text{int}}(k)$ by recording a sound field using a small spherical microphone array located in the desired sound field. As (2) calculates the infinite sum, it is required to the finite number of expansion order N_{int}. $(N_{\text{int}}+1)^2$ microphones are required to reproduce the desired sound field at least when limiting the order by N_{int}. When microphone array which radius r_q used, the signal $s(\boldsymbol{r}_q,k)$ recorded by a microphone located on (θ_q,ϕ_q) expressed by the following equation:

$$s(\boldsymbol{r}_q,k)=\sum_{n=0}^{N_{\text{int}}}\sum_{m=-n}^{n}A_{nm}^{\text{int}}(k)b_n(kr_q)Y_{nm}(\theta_q,\phi_q), \tag{6}$$

here, $b_n(kr_q)$ is the radial function decided by the shape of the microphone array. Table I shows $b_n(kr)$ for the shape of the microphone array and Fig. 2 shows the amplitude of $b_n(kr)$. The signal $s(r_q,k)$ expressed by (6) can be rewritten as the following equation:

$$\boldsymbol{s}(k)=\boldsymbol{Y}\cdot\text{diag}\left[b_n(kr_q)\right]\boldsymbol{A}(k), \tag{7}$$

here, diag$[\cdot]$ is diagonal matrix.

$$\boldsymbol{s}=\left[s(\boldsymbol{r}_1,k),\ldots,s(\boldsymbol{r}_q,k),\ldots,s(\boldsymbol{r}_Q,k)\right], \tag{8}$$

$$\boldsymbol{Y}=\begin{bmatrix}Y_{0,0}(\theta_1,\phi_1) & \cdots & Y_{N_{\text{int}},N_{\text{int}}}(\theta_1,\phi_1)\\ \vdots & \ddots & \vdots\\ Y_{0,0}(\theta_Q,\phi_Q) & \cdots & Y_{N_{\text{int}},N_{\text{int}}}(\theta_Q,\phi_Q)\end{bmatrix}, \tag{9}$$

$$\boldsymbol{A}=\left[A_{0,0}^{\text{int}}(k),\ldots,A_{N_{\text{int}},N_{\text{int}}}^{\text{int}}(k)\right]. \tag{10}$$

Then, $s(k)$ is the recording signal vector in the wave domain, Y is the matrix of the spherical harmonic function, and $A(k)$ is the expansion coefficient vector. From (7), $\boldsymbol{A}$ is expressed by the following equation:

$$\boldsymbol{A}(k)=\boldsymbol{Y}^{+}\cdot\text{diag}\left[\frac{1}{b_n(kr_q)}\right]\boldsymbol{s}(k), \tag{11}$$

here, $(\cdot)^{+}$is the Moore-Penrose inverse matrix. An expansion coefficient becomes unstable in HOA encoding when we use a hollow omnidirectional array in prohibited frequency at which $y_n(kr)=0$ [8]. Then, radial functions of a ridged microphone array or a hollow cardioid microphone array are proposed to avoid prohibited frequency [8].

Then, we calculate the driving signals of each loudspeaker based on the expansion coefficient by HOA decoding. The spherical harmonic expansion on location $\boldsymbol{r}_s$ expressed by the following equation [6]:

$$\frac{e^{jk|\boldsymbol{r}-\boldsymbol{r}_s|}}{4\pi|\boldsymbol{r}-\boldsymbol{r}_s|}=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}jkh_n(kr_s)y_n(kr)Y_{nm}(\theta,\phi)Y_{nm}(\theta_s,\phi_s)^*. \tag{12}$$

From (2) and (12), the expansion coefficients of a point source are expressed by the following equation:

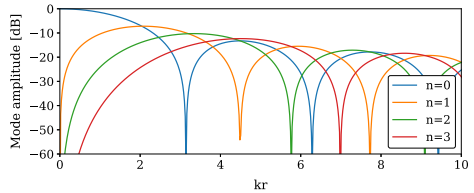
$$A_{nm}^{\text{int}}(k)=jkh_n(kr_s)Y_{nm}(\theta_s,\phi_s)^*. \tag{13}$$

In this paper, we decode HOA coefficients assuming a point source. Thus, the driving signal $d(\boldsymbol{r}_l,k)$ of loudspeaker $l(1\leq l\leq L)$ which located $\boldsymbol{r}_l=(r_l,\theta_l,\phi_l)$ and decoding matrix $\boldsymbol{C}$ are expressed by the following equations:

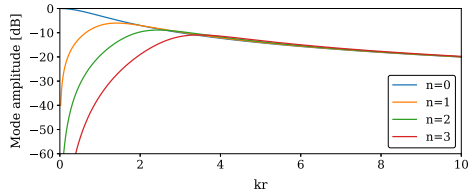
$$\boldsymbol{d}(k)=\boldsymbol{C}^{+}\cdot\text{diag}\left[\frac{j^n}{jkh_n(kr_l)}\right]\boldsymbol{A}(k), \tag{14}$$

TABLE I: $b_n(kr)$ for the various array configurations.

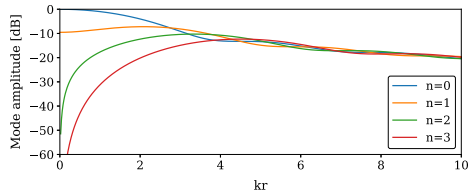
Array configuration	$b_n(kr)$
Open sphere	$y_n(kr)$
Rigid sphere	$y_n(kr) - \frac{y_n'(ka)}{h_n'(ka)} h_n(kr)$
Cardioid	$\alpha y_n(kr) - (1-\alpha) j y_n'(kr)$



(a) Open sphere



(b) Rigid sphere



(c) Cardioid

Fig. 2: Mode amplitude of $b_n(kr)$ for the various array configurations

$$\boldsymbol{d} = [d(\boldsymbol{r}_1, k), \ldots, d(\boldsymbol{r}_l, k), \ldots, d(\boldsymbol{r}_L, k)]\,, \tag{15}$$

$$\boldsymbol{C} = \begin{bmatrix} Y_{0,0}(\theta_1, \phi_1)^* & \ldots & Y_{N_{\mathrm{int}}, N_{\mathrm{int}}}(\theta_1, \phi_1)^* \\ \vdots & \ddots & \vdots \\ Y_{0,0}(\theta_L, \phi_L)^* & \ldots & Y_{N_{\mathrm{int}}, N_{\mathrm{int}}}(\theta_L, \phi_L)^* \end{bmatrix}, \tag{16}$$

here, the loudspeaker-drive-signal $\boldsymbol{r}$

D. Radiation sound reproduction

In this paper, we aim to construct a virtual sound source that has desired radiation characteristics by reproducing a radiated sound field based on HOA. The procedures of the proposed method are shown below.

[Step 1]Calculating radiation characteristics

The expansion coefficients of the radiated sound field with directivity patterns are calculated by the signals recorded by microphones located around the actual sound source. The discrete formula of (5) which sampled by microphones $q'(1 \leq q' \leq Q')$ located $\boldsymbol{r}_{q'} = (r_{q'}, \theta_{q'}, \phi_{q'})$ is expressed by the following equation:

$$A_{nm}^{\mathrm{ext}}(k) = \sum_{q'=1}^{Q'} \frac{p(\boldsymbol{r}_{q'}, k)}{h_n(kr_{q'})} Y_{nm}(\theta_{q'}, \phi_{q'})^*, \tag{17}$$

[Step 2]Calculating sound pressure on a microphone in a radiation sound

We assume a spherical microphone array with multiple microphones arranged on a sphere centered at position r_c in the radiated sound field. Sound pressure at the position of the q-th microphone $\boldsymbol{r}_q$ is calculated by (4). In this paper, we consider the directivity of microphones. Thus, recorded signals by each microphone are expressed by the following equation:

$$s(\boldsymbol{r}_q, k) = w_q p(\boldsymbol{r}_q, k), \tag{18}$$

$$w_q = \alpha + (1-\alpha) \frac{(\boldsymbol{r}_s - \boldsymbol{r}_c) \cdot (\boldsymbol{r}_s - \boldsymbol{r}_q)}{|\boldsymbol{r}_s - \boldsymbol{r}_c||\boldsymbol{r}_q - \boldsymbol{r}_q|}, \tag{19}$$

here, $\boldsymbol{r}_s$ is the position of the sound source. The directivity coefficient w_q corresponds to the recording of the open sphere when $\alpha = 1$ and cardioid when $\alpha = 0.5$, respectively.

[Step 3]Calculating loudspeaker driving signal based on HOA

The driving signals for each loudspeaker are obtained by encoding into expansion coefficients using (11) and decoding the expansion coefficients into driving signals using (14).

III. Objective Evaluation

We perform the computer simulation to evaluate the effectiveness of the proposed method.

A. Simulation conditions

We evaluate the reproducibility of the radiated sound field by comparing the desired sound field and the sound field reproduced by the proposed method in the computer simulation. Here, the desired sound field is generated by the radiation that has desired radiation characteristic and calculated by (4). We use radiation characteristics of the human voice shown in [4]. Figure 3 shows the polar pattern of the used radiation characteristic. We calculate the reproduced sound field assuming 22.2 multichannel audio as a surrounding loudspeaker array. In this simulation, a reproduced sound field is calculated by the open sphere ($\alpha = 1$) and the cardioid ($\alpha = 0.5$) in (11) and (19). The reproduced radiated sound field is represented by the following equations:

$$p(\boldsymbol{r}, k) = \sum_{l=1}^{L} d(\boldsymbol{r}_l, k) G(\boldsymbol{r}|\boldsymbol{r}_l), \tag{20}$$

$$G(\boldsymbol{r}|\boldsymbol{r}_l) = \frac{e^{jk|\boldsymbol{r}-\boldsymbol{r}_l|}}{4\pi|\boldsymbol{r}-\boldsymbol{r}_l|} \left(\beta - (1-\beta)\left(1 + \frac{j}{k|\boldsymbol{r}-\boldsymbol{r}_l|}\right)\cos\gamma_l\right), \tag{21}$$

here, $G(\cdot)$ is the transfer function of a directional sound source [9]. In this simulation, we assume that each loudspeaker is unidirectional ($\beta = 0.5$) and set γ_l to maximize the outputs of each loudspeaker on the center of the 22.2 multichannel audio. Table II shows simulationg conditions. Microphones are arranged as a Fibonacci spiral.

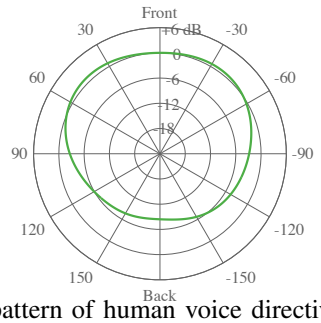


Fig. 3: Polar pattern of human voice directivity (500 Hz)

TABLE II: Experimental conditions

Frequency	500 Hz
Expansion order	$N_{\text{ext}} = 6, N_{\text{int}} = 4$
Number of microphones	$Q = 49$
Radius of microphones	$r_q = 0.05$ m
Number of loudspeakers	$L = 22$
Radius of loudspeakers	$r_l = 1.9$ m

B. Evaluation metric

In this simulation, a squared error between the desired sound field $p_{des}(r, k)$ and the reproduced sound field $p_{rep}(r, k)$ calculated by (20) is evaluated by normalized reproduction error (NRE). NRE is expressed by the following equation.

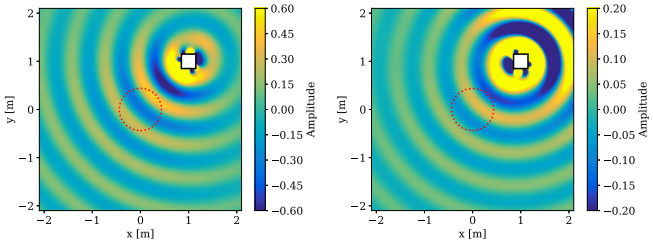
$$\text{NRE}(\boldsymbol{r}, k) = 10 \log_{10} \frac{|p_{\text{des}}(\boldsymbol{r}, k) - p_{\text{rep}}(\boldsymbol{r}, k)|^2}{|p_{\text{des}}(\boldsymbol{r}, k)|^2}, \qquad (22)$$

here, the lower the NRE value is, the better the reproduction ability is.

C. Simulation results

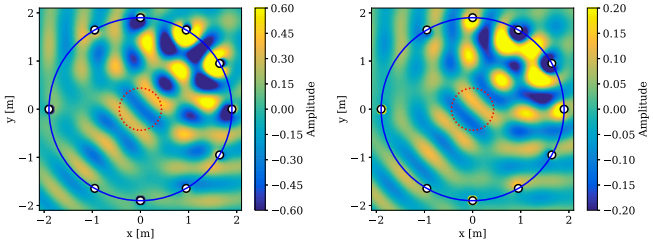
Figure 4 and 9 show the desired sound field. In Figs. 4 and 9, the white square on $(x, y) = (1, 1)$ shows the location of a sound source. Figures 5 and 10 show the reproduced sound field using the open sphere, Figs. 6 and 11 show NRE (open sphere), Figs. 7 and 12 show the reproduced sound field using a Cardioid, Fig. 8 shows NRE (cardioid). In Fig. 4-13, "Front" shows that the sound source radiates to the audience and "Back" shows that the sound source radiates to the opposite side of the audience. A red circle displayed by a dashed line shows the reproduction area r_0, a blue circle displayed by a solid line shows the radius of the loudspeaker array, white squares on the loudspeaker array show loudspeaker positions on the middle layer of the 22.2 multichannel audio. r_0 shows area that the reproduction error is about 4%, and $r_0 \approx N_{\text{int}}/2\pi f$ [10]. Here, f shows the frequency of the sound.

From Figs. 4, 5, 7, 9, 10 and 12 it can be seen that the amplitude of the reproduced area is almost similar to that in the desired sound field. Here, the frequency response of the sound source in each direction is reproduced with NRE = -20 dB. However, the reproduction accuracy of "Back" is less than "Front" according to Figs. 6, 8, 11 and 13. This is because "Back" includes many higher-order components of spherical harmonic function and has complex radiation characteristics. Thus, "Back" should be expanded with a higher expansion order than "Front".



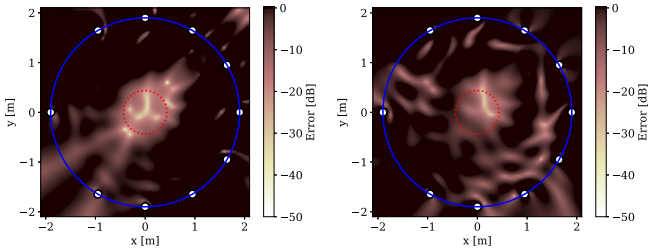
(a) Front (b) Back

Fig. 4: Desired sound field (500 Hz)



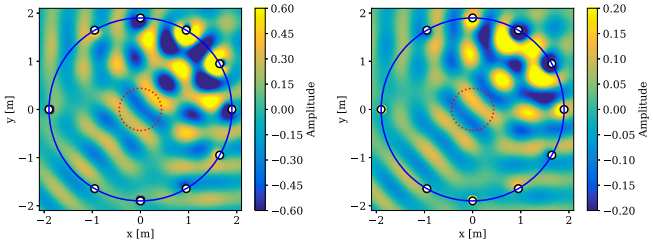
(a) Front (b) Back

Fig. 5: Reproduced sound field (open sphere) (500 Hz)



(a) Front (b) Back

Fig. 6: NRE (open sphere) (500 Hz)



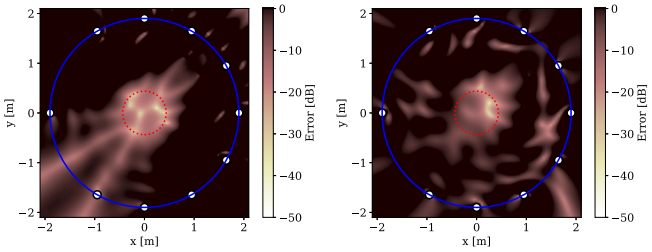
(a) Front (b) Back

Fig. 7: Reproduced sound field (cardioid) (500 Hz)

In Figs. 6, 8, 11 and 13, NRE for "Front" and "Back" are similar regardless of the directivity of microphones. Therefore, it is possible to avoid an inhibited frequency by using cardioid microphones when reproducing the radiation sound from the radiated directivity data by the proposed method. The reproduced area r_0 depends on reproduction frequency although NRE is the same regardless of reproduction frequency. From the above, the effectiveness of the proposed method is shown.

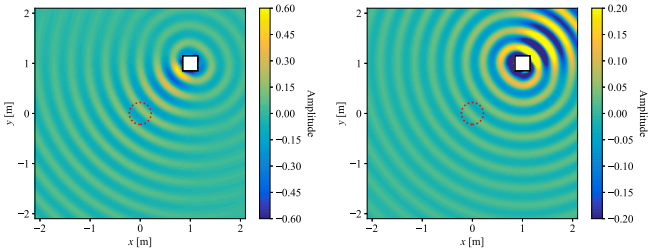
IV. Conclusions

In this paper, we propose a radiated sound field reproduction method for a surrounding loudspeaker array. Simulation results



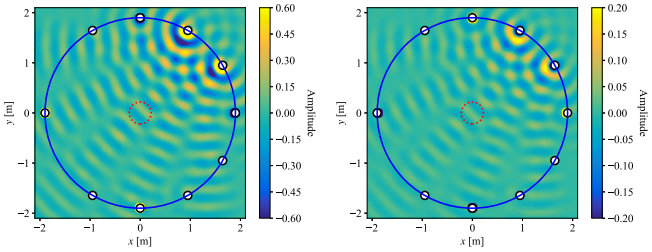
(a) Front (b) Back

Fig. 8: NRE (cardioid) (500 Hz)



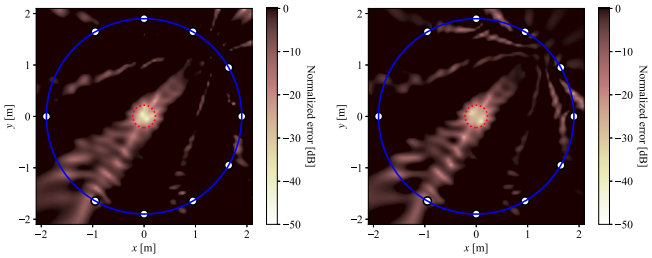
(a) Front (b) Back

Fig. 9: Desired sound field (1000 Hz)



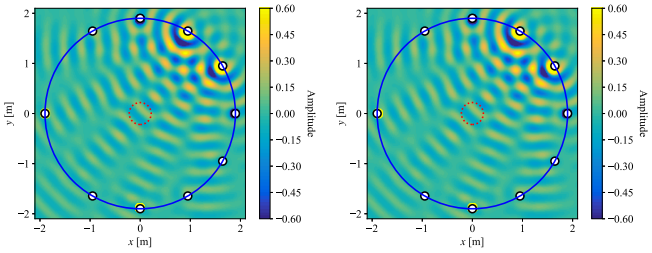
(a) Front (b) Back

Fig. 10: Reproduced sound field (open sphere) (1000 Hz)



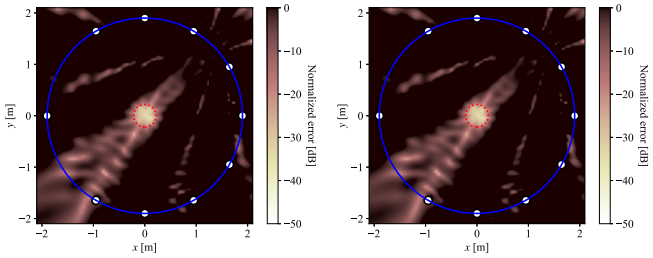
(a) Front (b) Back

Fig. 11: NRE (open sphere) (1000 Hz)



(a) Front (b) Back

Fig. 12: Reproduced sound field (cardioid) (1000 Hz)



(a) Front (b) Back

Fig. 13: NRE (cardioid) (1000 Hz)

showed that the proposed method reproduced desired sound field with NRE = −20 dB in cases of "Front" and "Back". Simulation results also showed that the proposed method can avoid prohibiting frequency by using cardioid microphones when reproducing a radiated sound field. In the future, we aim to reproduce the radiation sound with high accuracy regardless of the direction of radiation.

ACKNOWLEDGMENT

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP19H04142, JP21H03488, JP21H04427, and JP23H03425.

REFERENCES

[1] Kenta IWAI, Hiromu Suzuki, Takanobu Nishiura, "3-D Sound Image Reproduction Method Based on Spherical Harmonic Expansion for 22.2 Multichannel Audio," Applied Sciences, vol. 12, no. 4, 2022.

[2] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama, A. Ando, "A 22.2 multichannel sound system for ultrahigh-definition TV (UHDTV)," SMPTE Motion Imaging Journal, Vol. 117, Num. 3, pp. 40–49, 2008.

[3] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," Journal of the Audio Engineering Society, Vol. 45, Num. 6, pp. 456–466, 1997.

[4] C. Pörschmann, J. M. Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns," The Journal of the Acoustical Society of America, Vol. 149, Num. 6, pp. 4553–4564, 2021.

[5] K. Bando, Y. Haneda, "Interactive directivity control using dodecahedron loudspeaker array," Journal of Signal Processing, Vol. 20, Num. 4, pp. 209–212, 2016.

[6] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," Journal of the Audio Engineering Society, Vol. 53, Num. 11, pp.1004–1025, 2005.

[7] E. G. Williams, J. A. Mann iii "Fourier acoustics: sound radiation and nearfield acoustical holography," Academic Press, 1999.

[8] I. Balmages, B. Rafaely, "Open-sphere designs for spherical microphone arrays," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, Num. 2, pp. 727–732, 2007.

[9] M. A. Poletti, F. M. Fazi, P. A. Nelson, "Sound-field reproduction systems using fixed-directivity loudspeakers," The Journal of the Acoustical Society of America, Vol. 127, No. 6, pp. 3590–3601, 2010.

[10] D. B. Ward, T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," IEEE Transactions on speech and audio processing, Vol. 9, Num. 6, pp. 697–707, 2001.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multichannel learning-based spatially extended active noise control via model matching and sensor transfer function interpolation

Pei-Lin Zhong, You-Siang Chen, and Mingsian R. Bai
Department of Power Mechanical Engineering, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013.
Corresponding author: Mingsian R. Bai, Email: msbai@pme.nthu.edu.tw

Abstract—**This paper presents a learning-based multichannel active noise control (L-MANC) method that utilizes a model-matching framework along with a sensor interpolation approach. In a feedforward ANC scenario, L-MANC aims to cancel the primary noise field for an extended control region with a limited number of measurements. To achieve global control, the training data is generated by interpolating the acoustic transfer functions (ATFs) based on a small number of measurements via kernel ridge regression. Noise control filters for the secondary loudspeakers are learned using a convolutional recurrent neural network (CRNN) in an encoder-decoder architecture. In the training phase, the network parameters are obtained using the mean square error criterion in a model-matching paradigm. Simulations are performed for an ANC system consisting of a nine-loudspeaker linear array and eight measured control points. The proposed L-MANC approach is tested for various unseen scenarios with variations in reverberation time and location of the primary noise source. The proposed system has demonstrated superior ANC performance in terms of mean square noise reduction (MSNR) over several baselines.**

I. INTRODUCTION

Active Noise Control (ANC) is a noise control technique that aims to suppress unwanted noise by using an anti-noise signal generated by secondary loudspeakers [1] . In recent years, spatial ANC has received considerable attention [2] . ANC systems using multiple secondary loudspeakers have been suggested to achieve a broadened control region [3] . ANC systems have typically been implemented using adaptive filters such as the Filtered-x Least Mean Square (Fx-LMS) and Filtered-x Recursive Least Square (Fx-RLS) algorithms [4] [5] . These algorithms attempt to minimize the power of the residual noise at the error microphone locations [6] . The single-channel adaptive ANC was extended to adaptive MANC [7] for global ANC. However, the computational complexity of adaptive filters increases considerably with the number of loudspeakers and microphones [8] Moreover, the learning nature that involves convergence of error hampers the performance of noise reduction, especially for the transient primary noise. Adaptive algorithms could have convergence problems in the face of perturbation of environment or transient signals, resulting in poor noise reduction performance [9] .

Recently, the advent of deep neural networks (DNNs) has shed some light on ANC research [10] In this paper, a learning-based feedforward multichannel ANC approach is proposed to achieve the ANC performance within an extended region. In addition, a special technique is developed to interpolate the acoustic transfer functions (ATFs) based on a limited number of measurements. In light of the model matching principle [11] we design a DNN-based controller for multichannel ANC. Our approach adopts a CRNN implemented in an encoder-decoder structure to generate the loudspeaker signals for noise suppression. Furthermore, to achieve spatially extended control based only on a limited number of ATF measurements, we employ the kernel ridge regression method [12] to interpolate the ATFs for training data generation. This approach reduces the need for tedious room impulse response (RIR) measurements. By using a large number of interpolated ATFs, the residual noise signal can be minimized over an extended region.

The paper is organized as follows. Sec. II provides an overview of related work, with a particular focus on model-matching ANC approaches. Sec. III introduces the proposed L-MANC method and outlines the construction of a CRNN. Detailed information on the experimental setup, dataset

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

selection, and CRNN parameters are presented in Sec. IV. The results and discussions are presented in Sec. V, while the conclusion is given in Sec. VI.

II. RELATED WORK

In this section, the ANC problem via a model-matching principle is introduced. Multichannel ANC based on Tikhonov regularization (TIKR), linearly constrained minimum variance (LCMV), and the Fx-LMS methods are used as the baselines.

A. ANC as a Model Matching Problem

The architecture of a model-matching ANC is illustrated in Fig. 1. The single-channel input signal $x(f)$ at the frequency f is multiplied by the noise control filter vector $\mathbf{h}(f) \in \mathbb{C}^{J\times 1}$ to obtain the outputs. Each output is multiplied elementwise by the secondary-path ATF matrix $\mathbf{G}^s(f) \in \mathbb{C}^{K\times J}$ measured between the secondary loudspeaker inputs and the pressure responses $\mathbf{y}(f)$ at the control points. The objective of this architecture is to minimize the matching error between the predicted signal $\mathbf{y}(f)$ and the desired signal $\mathbf{d}(f)$. The vector $\mathbf{g}^p(f) \in \mathbb{C}^{K\times 1}$ represents the primary path ATFs from the primary noise source to all control points. The frequency index is omitted hereafter for brevity. Therefore, the goal of the model-matching principle is to generate the opposite phase of $\mathbf{g}^p$, as given by:

$$\underbrace{\begin{bmatrix} g_1^p \\ \vdots \\ g_K^p \end{bmatrix}}_{\mathbf{g}^p} = \underbrace{\begin{bmatrix} G_{11}^s & \cdots & G_{1K}^s \\ \vdots & \ddots & \vdots \\ G_{J1}^s & \cdots & G_{JK}^s \end{bmatrix}}_{\mathbf{G}^s} \underbrace{\begin{bmatrix} h_1 \\ \vdots \\ h_J \end{bmatrix}}_{\mathbf{h}}, \tag{1}$$

where G^s_{jk}, j = 1, …, J, and k = 1, …, K denotes the ATF between the jth loudspeaker and the kth measured point, h_j represents the noise control filter for the jth loudspeaker, and g^p_k is the ATF between the primary noise and the kth measured control point. In this model-matching ANC system, the learnable CRNN is employed to obtain the ANC controllers.

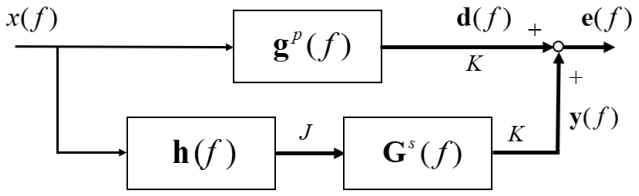


Fig. 1: The framework of the model-matching ANC.

For underdetermined systems ($J > K$), the solution of a regularized least squares approach, TIKR [13] can be acquired through employing the cost function as follows:

$$\min_{\mathbf{h}_{TIKR}} \|\mathbf{G}^s\mathbf{h}_{TIKR} + \mathbf{g}^p\|_2^2 + \beta^2 \|\mathbf{h}_{TIKR}\|_2^2, \tag{2}$$

where $\mathbf{h}_{TIKR}$ is the noise control filter, β is the regularization parameter tuned via the L-curve method [14] and $\|\cdot\|_2$ denotes the $l2$-norm. Thus, the solution of (2) can be obtained [15]

$$\mathbf{h}_{TIKR} = -(\mathbf{G}^{sH}\mathbf{G}^s + \beta^2\mathbf{I})^{-1}\mathbf{G}^{sH}\mathbf{g}^p. \tag{3}$$

Although TIKR has good ANC performance at the control points, the noise reduction performance degrades significantly at the adjacent region of the control points. To overcome this limitation, the LCMV [11] method can be adopted to globally enhance the ANC performance. The constrained optimization problem can be defined as:

$$\min_{\mathbf{h}_{LCMV}} (\mathbf{e}_i^H\mathbf{e}_i), \text{ s.t. } \mathbf{G}_m^s\mathbf{h}_{LCMV} = -\mathbf{g}_m^p \tag{4}$$

where $\mathbf{e}_i = \mathbf{g}_i^p + \mathbf{G}_i^s\mathbf{h}_{LCMV}$ represents the error vector at the interpolated points, $\mathbf{g}_i^p$ and $\mathbf{g}_m^p$ are the ATF vectors from the noise to the interpolated and measured points, respectively, $\mathbf{G}_i^s$ and $\mathbf{G}_m^s$ denote the ATF matrices from the secondary loudspeakers to the interpolated and measured points, respectively, and $\mathbf{h}_{LCMV}$ is the optimal noise control filter of the LCMV approach. By the method of Lagrange multipliers, the optimal filter can be solved as:

$$\begin{aligned}\mathbf{h}_{LCMV} = {} & (\mathbf{G}_i^{sH}\mathbf{G}_i^s + \varsigma\mathbf{I}_J)^{-1}\mathbf{G}^{sH}[\mathbf{G}^s(\mathbf{G}_i^{sH}\mathbf{G}_i^s + \varsigma\mathbf{I}_J)^{-1}\mathbf{G}^{sH} + \mu\mathbf{I}_K]^{-1}\cdot \\ & [\mathbf{G}^s(\mathbf{G}_i^{sH}\mathbf{G}_i^s + \varsigma\mathbf{I}_J)^{-1}\mathbf{G}_i^{sH}\mathbf{g}_i^p - \mathbf{g}^p] - (\mathbf{G}_i^{sH}\mathbf{G}_i^s + \varsigma\mathbf{I}_J)^{-1}\mathbf{G}_i^{sH}\mathbf{g}_i^p,\end{aligned} \tag{5}$$

where μ and ζ are the regularization parameters selected using particle swarm optimization (PSO) [16]

B. Filtered-x multichannel least mean square algorithm

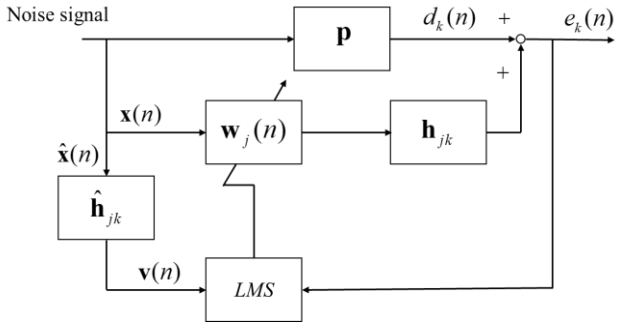


Fig.2: Block diagram of the ANC system using the Fx- LMS algorithm

The block diagram of the Fx- LMS algorithm is shown in Fig. 5. Consider the noise source input vector as follows:

$$\mathbf{x}(n) = [x(n)\ \cdots x(n-L+1)]^T \tag{6}$$

where L represents the filter length and n is the sample index. In the primary path, the desired signal $d_k(n)$ can be obtained by convolving the noise signal with the primary path $\mathbf{p} \in \mathbb{R}^K$. In

the intermediate path, the kth error signal $e_k(n)$ can be derived as follows:

$$e_k(n) = d_k(n) + \sum_{j=1}^{J} \mathbf{h}_{j,k}^T [\mathbf{w}_j^T(n)\mathbf{x}(n)] \quad (7)$$

where the adaptive filter coefficient is denoted as $\mathbf{w}_j(n) = [w_{j,0}(n) \ \dots \ w_{j,L-1}(n)]^T$, and the secondary path RIR $\mathbf{h}_{j,k}$ can be written as follows:

$$\mathbf{h}_{j,k} = [h_{j,k,1} \quad \cdots \quad h_{j,k,N}] \quad (8)$$

where N represents the FIR filter length. In the lower path, without loss of generality, we simply treat $\hat{\mathbf{h}}_{j,k} = \mathbf{h}_{j,k}$, then the filtered input $\mathbf{v}_{jk}(n) = \hat{\mathbf{x}}(n) * \hat{h}_{jk}(n)$, where $\hat{\mathbf{x}}(n) = [x(n) \ \dots \ x(n-N+1)]^T$. Lastly, the filter coefficient can be updated by

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) - \mu \sum_{k=0}^{K} \mathbf{v}_{jk}(n)^T e_k(n). \quad (9)$$

III. PROPOSED METHOD

This section describes the proposed L-MANC system within a model-matching framework. L-MANC adopts CRNN as a learnable function $CRNN(\cdot)$ for the ANC controllers. The kernel ridge regression method is employed for interpolated ATFs between measured points. In particular, this approach allows for more efficient training by using fewer measured points, resulting in improved control performance. During the training stage, the pre-generated RIRs at measured and interpolated points are utilized for the target signals captured by the error microphones. L-MANC aims to cancel the primary noise at an extended area.

A. Learning-based Model-Matching Approach

The L-MANC system is formulated in the time-frequency domain, as shown in Fig. 3. The ATF between the primary noise and the kth microphone is denoted as $g_k^p(t,f)$. $G_{jk}^s(t,f)$ represents the secondary path, where the ATF between the jth secondary loudspeaker and the kth error microphone is pre-measured or pre-interpolated. The indices t and f represent the time frame and the frequency bin, respectively. Therefore, the kth desired signal $d_k(t,f)$ can be obtained from the noise source $X(t,f)$ as

$$d_k(t,f) = X(t,f) g_k^p(t,f). \quad (10)$$

Next, the input signal $X(t,f)$ is decomposed into its real and imaginary parts, $X_r(t,f)$ and $X_i(t,f)$. These two parts are concatenated to construct the input matrix $\mathbf{X}(t,f) = [X^r(t,f) \ X^i(t,f)]^T \in \mathbb{R}^2$. Then, $\mathbf{X}(t,f)$ is fed into the neural network, $CRNN(\cdot)$ to obtain filtered signals, $\mathbf{Y}(t,f) \in \mathbb{R}^{2J}$:

$$\mathbf{Y}(t,f) = CRNN(\mathbf{X}(t,f)), \quad (11)$$

where $\mathbf{Y}(t,f)$ is formed by the real and imaginary parts of the J signals for the loudspeaker, $Y_j^r(t,f)$ and $Y_j^i(t,f)$, $j = 1, \dots, J$, as given by $\mathbf{Y}(t,f) = [Y_1^r(t,f) \ Y_1^i(t,f) \ \dots \ Y_J^r(t,f) \ Y_J^i(t,f)]^T$.

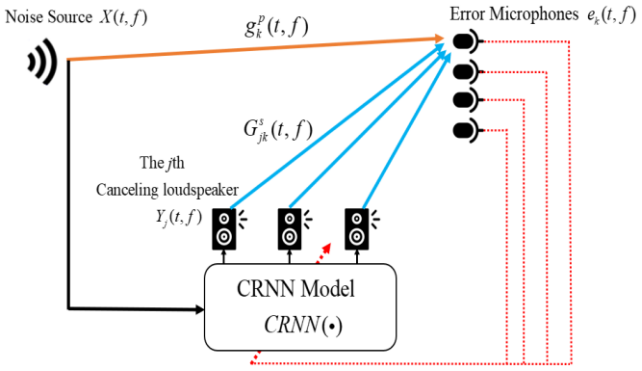


Fig. 3: The proposed L-MANC based on the model matching principle.

Then, the network output for each jth secondary loudspeaker is reformulated into the complex form of $Y_j(t,f)$ and multiplied by the ATF $G_{jk}^s(t,f)$ from the jth secondary loudspeakers to the kth error microphones. The error signal $e_k(t,f)$ between the desired and estimated signals can be expressed as

$$e_k(t,f) = d_k(t,f) + \sum_{j=1}^{J} G_{jk}^s(t,f) Y_j(t,f). \quad (12)$$

For global ANC, the kernel ridge regression method [12] is used to interpolate the ATFs between the measured points, which allows for accurate control for both measured and interpolated points.

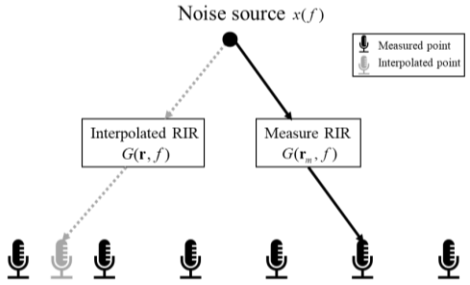


Fig. 4: Schematic diagram for defining the sensor interpolation problem.

B. ATF interpolation using kernel method

The schematic diagram of the sensor interpolation problem is shown in Fig. 4. In practical implementation, the training performance can be affected by the limited usage of the microphones. Hence, the goal of sensor interpolation is to increase the ATFs in between the control region and improve the performance in the CRNN training process. The objective

function of the ATF interpolation using kernel method is formulated as

$$\min_{\hat{G}} \mathcal{J} = \sum_{m=1}^{M} \left| \hat{G}(\mathbf{r}_m) - G(\mathbf{r}_m) \right|^2 + \beta \left\| \hat{G} \right\|_{\mathcal{H}}^2, \tag{13}$$

where $\hat{G}(\mathbf{r}_m)$ and $G(\mathbf{r}_m)$ denote the interpolated and measured ATFs at the position vector $\mathbf{r}_m$ of the mth microphone, respectively, M represents total number of measure points, β is the regularization parameter, and $\|\cdot\|_{\mathcal{H}}^2$ is the inner product within the Hilbert space $\mathcal{H}$. A closed-form solution of the interpolated $\hat{G}(r)$ can be obtained as [12]

$$\hat{G}(\mathbf{r}) = \boldsymbol{\kappa}^T(\mathbf{r}) \left(\mathbf{K} + \beta \mathbf{I}_M \right)^{-1} \mathbf{G}, \tag{14}$$

where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix, and $\mathbf{r}$ is position vector within the control region,

$$\mathbf{K} = \begin{bmatrix} j_0(k\|\mathbf{r}_1 - \mathbf{r}_1\|_2) & \cdots & j_0(k\|\mathbf{r}_1 - \mathbf{r}_M\|_2) \\ \vdots & \ddots & \vdots \\ j_0(k\|\mathbf{r}_M - \mathbf{r}_1\|_2) & \cdots & j_0(k\|\mathbf{r}_M - \mathbf{r}_M\|_2) \end{bmatrix}, \tag{15}$$

$$\boldsymbol{\kappa}^T(\mathbf{r}) = \begin{bmatrix} j_0(k\|\mathbf{r} - \mathbf{r}_1\|_2) & \cdots & j_0(k\|\mathbf{r} - \mathbf{r}_M\|_2) \end{bmatrix}, \tag{16}$$

is the Gram matrix and kernel function in which k is the wave number, and $j_0(\cdot) = \sin(\cdot)/(\cdot)$ denotes the zeroth-order spherical Bessel function, also known as the sinc function.

C. Network Architecture

The CRNN [17] in encoder-decoder architecture is employed in the proposed L-MANC, as illustrated in Fig. 5. The encoder consists of five convolutional layers to extract acoustic features from the input signal $\mathbf{X}(t, f)$. Batch normalization (BN) along with the exponential linear unit (ELU) activation function is applied after each convolutional layer, hyperbolic tangent (Tanh) activation function for last convolutional layer. The decoder part comprises five deconvolutional layers that reconstruct the signal. The skip

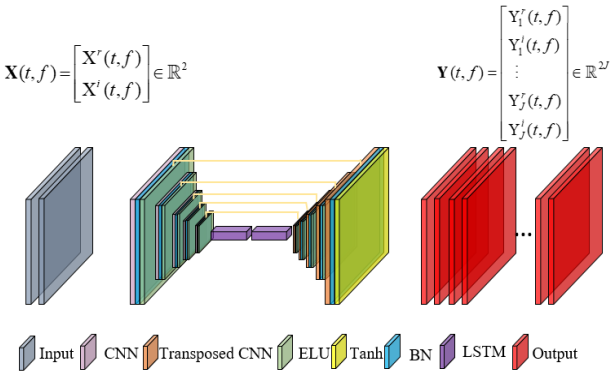


Fig. 5: The network architecture of CRNN.

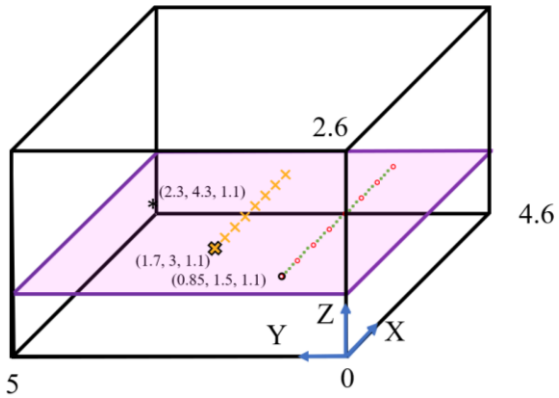


Fig .6: Environmental setups for the data generation. "*" represents the noise source, and "×" represents the secondary loudspeakers. "◦" represents the measured point, while "•" denotes the interpolated points.

connections concatenate each output layer in the encoder and the previous layer in the decoder. These skip connections allow the network to preserve salient features for the decoder. To effectively capture a broader context in the frequency domain, a stride of 2 is applied to all convolutional and deconvolutional layers along the frequency dimension. Zero padding is applied in each encoder-decoder layer to maintain the size consistency. Two long short-term memory (LSTM) units [18] between the encoder and decoder are included in the proposed architecture. The loss function of L-MANC is defined as the averaged error signal power

$$\mathcal{L} = \frac{1}{KFT} \sum_{k=1}^{K} \sum_{f=0}^{F} \sum_{t=0}^{T} \left| e_k(t,f) \right|^2 \tag{17}$$

where $|\cdot|$ denotes the absolute operation.

IV. SIMULATION SETTINGS

A. Dataset Generation and Parameter Settings

The simulation environment is illustrated in Fig. 6. A rectangular room with dimensions of 4.6 × 5 × 2.6 m was considered in the simulation. The ATFs at the control points were generated by the image source method (ISM) [19] In the L-MANC system, one noise source, nine secondary loudspeakers, and eight error microphones were used. The noise source was placed at (2.3, 4.3, 1.1) m. The nine secondary loudspeakers were placed from (1.7, 3, 1.1) m to (2.9, 3, 1.1) m with an equal spacing of 0.15 m between the loudspeakers. The eight error microphones were placed from (0.85, 1.5, 1.1) m to (3.65, 1.5, 1.1) m with an equal spacing of 0.4 m between the microphones. To extend the control region within the limited

control points, the kernel ridge regression method was used to interpolate the ATFs at the intervals between each pair of error microphones, with equal spacing of 0.1 m. A total of 21 additional interpolated points of ATFs were generated. The control frequency range is within the aliasing frequency of the deployment of the loudspeaker array for the control region, which is 1143 Hz to be specific. Within the control frequency range, the interpolated ATFs show good agreement with the ground-truth ATFs generated by ISM. In this room, the reverberation time (T60s) of 0.2 seconds was used for the training and testing stage. To improve the performance of the CRNN in the proposed L-MANC, a large variety of noise data were adopted in the training stage. For the generation of the training set, several 2-second audio clips were randomly selected from the NOISEX-92 database [20] to create 10000 training data. For the testing dataset, we used five types of unseen input noise, including four types of industrial machine sounds (fan, pump, valve, slider) [21] and speech data [22] 1200 non-speech and 300 speech unseen clips were generated for the validation of L-MANC. Each noise signal from the training and testing dataset was sampled at a rate of 16 kHz. The short-time Fourier transform (STFT) with a 20 ms Hann window and a 10 ms frame shift was applied to obtain the network input, $X(t, f)$. The optimizer used in the training process was AMSGrad [23] with a learning rate of 0.001. The batch size was set to be 15. The training process was stopped when the training loss was not decreased after 10 consecutive epochs. The network architecture and the hyperparameters are summarized in Table. 1. Hyperparameters are specified in the format of (kernel size, strides, output channel number), and T denotes the number of the time frames in an STFT spectrogram.

B. Evaluation Metrics

The validation of the proposed method and the baselines is performed on the time domain processed signals. Therefore, the evaluation metric, MSNR [24] is defined as follows:

$$\mathrm{MSNR} = 10\log_{10}\frac{\sum_{k=1}^{K}\sum_{n=0}^{T_s}\left|d_k(n)\right|^2}{\sum_{k=1}^{K}\sum_{n=0}^{T_s}\left|e_k(n)\right|^2}, \tag{18}$$

where $d_k(n)$ represents the nth sample point of the primary noise signal that needs to be canceled at the kth error microphone, $e_k(n)$ represents the nth sample point of the error signal at the kth error microphone, and T_s is the total number of samples. The larger MSNR value indicates better ANC performance.

Table. 1: Detail of the hyperparameters of the CRNN architecture

Layer name	Input size	hyperparameters	Output size
Conv2d_1	2×513×T	(3, 2), (2, 1), 20	20×256×T
Conv2d_2	20×256×T	(3, 2), (2, 1), 40	40×127×T
Conv2d_3	40×127×T	(3, 2), (2, 1), 80	80×63×T
Conv2d_4	80×63×T	(3, 2), (2, 1), 128	128×31×T
Conv2d_5	128×31×T	(3, 2), (2, 1), 256	256×15×T
Reshape_1	256×15×T	-	T×3840
LSTM_1	T×3840	3840	T×3840
LSTM_2	T×3840	3840	T×3840
Reshape_2	T×3840	-	256×15×T
Deconv2d_5	512×15×T	(3, 2), (2, 1), 256	128×31×T
Deconv2d_4	256×31×T	(3, 2), (2, 1), 128	80×63×T
Deconv2d_3	160×63×T	(3, 2), (2, 1), 80	40×127×T
Deconv2d_2	80×127×T	(3, 2), (2, 1), 40	20×256×T
Deconv2d_1	40×513×T	(3, 2), (2, 1), 20	2J×513×T

V. SIMULATION RESULTS

A. Comparison of L-MANC and the baselines

Table. 2 summarizes the MSNR values by using the proposed method and the baselines. The value of the measured point is determined as the average of the MSNR values of all the measurement points, and the same principle applies to the interpolated point. It can be seen that L-MANC demonstrates superior MSNR performance compared to other baseline approaches. Since TIKR attempts to minimize the error at the measured control points, the MSNR performance obtained at the measured points can be compared to that of L-MANC.

Table. 2: MSNR obtained by (a) L-MANC, (b) Fx-MCLMS, (c) TIKR, (d) LCMV

	Measured	Interpolated
L-MANC	21.04	19.38
Fx-MCLMS	6.62	2.77
TIKR	20.65	1.09
LCMV	13.48	7.93

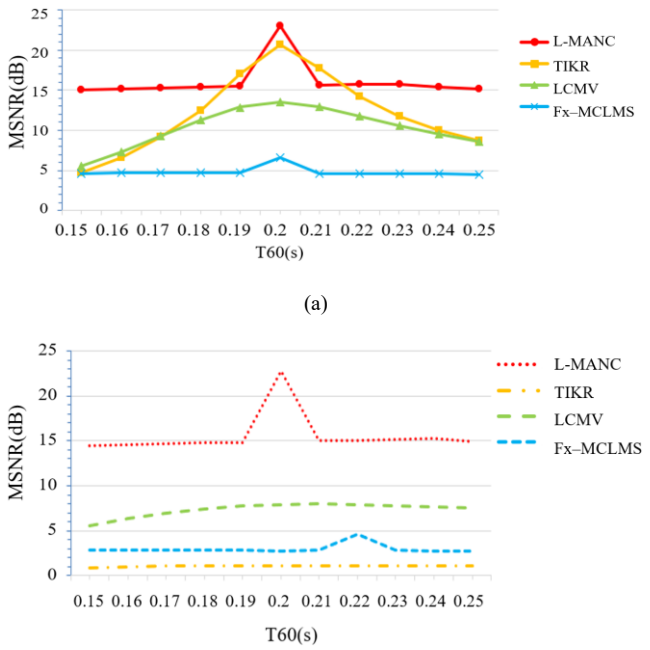


Fig. 7: MSNR values for four different ANC methods under different reverberation time (T60) from 0.15 s to 0.25 s. (a) MSNR value at the measured points, and (b) MSNR value at the interpolated points

However, its performance degrades significantly at the adjacent region of control points. In contrast, because LCMV addresses the error at the interpolated points while maintaining the performance at the control points, the MSNR obtained within the extended control points is not considerably degraded. However, L-MANC still shows superior performance at both measured and interpolated points.

Table. 3: MSNR obtained by L-MANC under different T60 environments

T60(s)	Measured	Interpolated
0.8	22.74	20.75
1.0	21.67	19.71
1.3	21.89	19.93

To ensure the versatility and generalizability of the proposed methodology, variations were introduced in the T60 values. As described in the section of data generation, the measured RIRs were generated with T60 = 0.8, 1.0, and 1.2s to simulate more realistic environment. Multiple networks were trained according to a specific T60 value based on the procedure of our L-MANC approach. The results in Table 3 present the summary of the ANC performance in terms of MSNR values across different T60 environments. The proposed method demonstrates consistent ANC performance across different environment and delivers outstanding results no matter in the measured or interpolated points.

B. Robustness of L-MANC

In this section, we investigate the robustness of the ANC methods in the face of variations on reverberation time and the primary noise positions.

1. Variations in reverberation time

The effectiveness of the ANC algorithms is evaluated by considering various reverberation times, covering a range of T60 from 0.15 s to 0.25 s. The results of the MSNR are shown in Fig. 7. In the study, the ATFs for the L-MANC and other algorithms were generated using T60 values of 0.2 s. All the approaches showed their highest performance at T60 = 0.2s at the measured points. Furthermore, our proposed approach consistently demonstrated the effectiveness across all T60 scenarios compared to the baselines. At the interpolated points, only the L-MANC method demonstrated performance comparable to that of the measured points. The other algorithms exhibited variation more than 5dB from the performance at the measured points, providing further confirmation of the superior results across different T60 values.

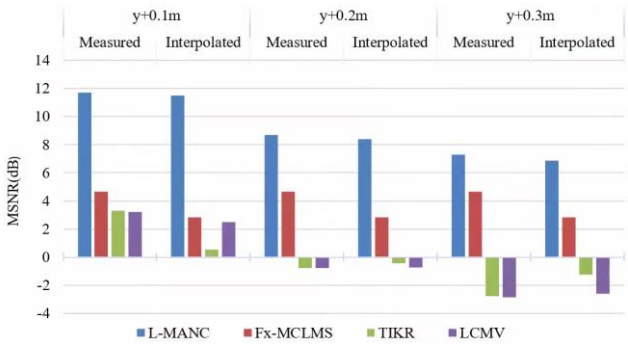


Fig. 8: MSNR performance for perturbed primary noise positions.

2. Variations in the primary noise positions

The proposed ANC is evaluated for perturbed noise source positions. The noise source that is initially located at (2.3, 4.3, 1.1) m is moved along the y-axes from 0.1 to 0.3 m with the increments of 0.1 m. The results are summarized in Fig. 8. It is clear that a slight movement of the noise source causes the MSNR to decrease for these ANC approaches. However, L-MANC shows superior performance compared to the baselines at all measured points and interpolated points. Furthermore, because the interpolated points were considered

during the training phase, the spatial variance in terms of MSNR in L-MANC between the measurement and interpolation points is small. These results indicate that the ANC performance of L-MANC can be achieved in a broadened region in the face of the primary noise position variations.

VI. CONCLUSIONS

In this paper, a learning-based feedforward ANC technique, L-MANC, has been formulated in the light of the model-matching principle. The CRNN network with an encoder-decoder structure is used to generate the loudspeaker signals and to reduce the error signal captured by the microphones. To ensure practicality in real-world scenarios, the unmeasured ATFs within a given region can be interpolated by the kernel ridge regression method. This allows the DNN to be trained with very limited measurements. The experimental results have demonstrated the superior ANC performance achieved by the proposed L-MANC compared to several baseline methods by using the objective metric, MSNR. The proposed method outperforms other baseline approaches in both the measured and interpolated points, indicating that comprehensive ANC control within a defined region can be achieved. The robustness of L-MANC is also validated for the variations in the reverberation time and noise positions compared to the baselines.

VII. REFERENCES

[1] S. M. Kuo and D. R. Morgan, "Active noise control: a tutorial review," Proc. of the IEEE, vol. 87, pp. 943–973, 1996.

[2] N. Kwon, M. Park, H.-S. Lee, J. Ahn, and M. Shin, "Construction noise management using active noise control techniques," J. Constr. Eng. Management, 142, 2016.

[3] H. Zhang and D. Wang, "Deep MCANC: A deep learning approach to multi-channel active noise control," Neural Netw., vol. 158, pp. 318–327, 2023.

[4] M. H. Costa, J. C. M. Bermudez, and N. J. Bershad, "Stochastic analysis of the filtered-X LMS algorithm in systems with nonlinear secondary paths," in IEEE Trans, vol. 50, no. 6, pp. 1327-1342, June 2002.

[5] X. Liang, J. Yao, L. Luo, W. Zhu, W. Zhang, Y. Wang, "A New Proportionate Filtered-x RLS Algorithm for Active Noise Control System," *Sensors* **2022**, *22*, 4566. https: //doi.org/ 10.3390/ s22124566.

[6] T. Murao, C. Shi, W.-S. Gan, and M. Nishimura, "Mixed-error approach for multi-channel active noise control of open windows," Appl. Acoust., vol 127, pp. 305-315, 2017.

[7] X. Sun, S. M. Kuo, and G. Meng, "Adaptive algorithm for active control of impulsive noise," J. Sound Vib., vol 291, issues 1–2, pp. 516-522, 2006.

[8] V. Patel and N. V. George, "Multi-channel spline adaptive filters for non-linear active noise control," Appl. Acoust., vol 161, pp. 107142, 2020.

[9] D. Shi, W.-S. Gan, B. Lam, and S. Wen, "Feedforward Selective Fixed-Filter Active Noise Control: Algorithm and Implementation," in IEEE/ACM Trans. Audio Speech Lang. Process, vol. 28, pp. 1479-1492, 2020.

[10] J. C. Young, A. Mostafavi, and S. S. Benipal, "DNoiseNet: Deep learning-based feedback active noise control in various noisy environments," Eng. App. of AI, vol 121, pp. 105971, 2023.

[11] S. C. Huang, C. H. Ma, Y. C. Hsu, and M. R. Bai, "Feedforward active noise global control using a linearly constrained beamforming approach," J. Sound and Vib. vol 537, pp. 117190, 2022.

[12] N. Ueno, S. Koyama, and H. Saruwatari, "Kernel ridge regression with constraint of Helmholtz equation for sound field interpolation," in Proc. Int. Workshop Acoust. Signal Enhancement, pp. 436-440, 2018.

[13] F. Ma, W. Zhang, and T. D. Abhayapala, Active control of outgoing broadband noise fields in rooms, in IEEE/ACM Trans Audio Speech Lang. Process, vol. 28, pp. 529-539, 2020.

[14] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," SIAM Journal on Sci. Comput., vol 14, no. 6, pp. 1487-1503, 1993.

[15] A. N. Tikhonov, "Solution of nonlinear integral equations of the first kind," Soviet Math. Dokl. vol. 5, pp. 835-838, 1964.

[16] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proc. of IEEE international conference on neural networks, vol.4, pp. 1942-1948, 1995.

[17] C. Sun, M. Zhang, and R. Wu, et al., "A convolutional recurrent neural network with attention framework for speech separation in monaural recordings," Scientific Reports, vol. 11, no. 1, pp. 1434, 2021.

[18] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol., vol. 1 (Long Papers), pp. 799–808, 2018.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. vol. 65, no. 4, pp. 943-950, 1979.

[20] A. Varga and H. J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., vol. 12, no. 3, pp. 247-251, 1993.

[21] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: sound dataset for malfunctioning industrial machine investigation and inspection," arXiv preprint arXiv:1909.09347, 2019.

[22] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., pp. 5206-5210, 2015.

[23] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," arXiv preprint arXiv:1904.09237, 2019.

[24] D. P. Das and G. Panda, "Active mitigation of nonlinear noise Processes using a novel filtered-s LMS algorithm," in IEEE Transactions on Speech and Audio Process, vol. 12, no. 3, pp. 313-322, May 2004.

A Study of the Microphone Protection of Active Noise Control for Axial Fan

Yi-Tsung Shen and Cheng-Yuan Chang
Chung Yuan Christian University, Taiwan
E-mail: im362512@gmail.com Tel: +886-979025977
E-mail: ccy@cycu.edu.tw Tel: +886-3-2654838

***Abstract* — In this paper, we propose an active noise control (ANC) system for axial fan and conduct a real time experiment. The experimental results compare three different experimental setups to verify the impact of microphone protection on system performance. The experimental results show that the final experimental setup can achieve satisfactory performance.**

I. INTRODUCTION

In the operation of electrical and electronic products, heat dissipation occurs between power consumption and energy conversion. In order to prevent high temperature from causing system malfunction, shutdown or even loss, we must rely on cooling devices to reduce the ambient temperature to stabilize the system, and fans play an important role here. Usually, traditional passive noise control is used to prevent fan noise, such as using sound-absorbing materials to isolate noise, or changing the material, number, and angle of fan blades, and using institutional design to make the noise generated by the fan less harsh.

Since the main function of the fan is to generate air pressure and air flow, at least the air flow out of the entrance must be retained, the passive noise elimination material to isolate the noise has its imperfections. The ANC system uses the principle of superposition [1] to offset the noise energy by using the anti-noise generated by the speakers. Compare to the passive noise control method, it can handle the noise better without affecting the function of the fan itself.

This paper introduces an axial fan system using ANC, which aims to reduce fan noise and avoid affecting the airflow in and out of the fan, and finds that wind interference directly affect the system performance by comparing the performance of three different settings and two different algorithms.

II. PROPOSE SYSTEM

In this section, we introduce the fan fundamental frequency equation and analyze the noise spectrum of the axial fan, and then propose the fan ANC algorithms. We set up the experiment on a steel frame mold to perform noise reduction for the location of the axial fan outlet, and compare the performance difference between them in three different settings.

A. Noise analysis of axial fan

Fan noise consists of two components, narrowband noise and broadband noise. Broadband noise is generated by the turbulent structure of the fan and the collision of airflow. The narrowband noise of a fan is generated when the fan blade rotates and radiates outward from the gap between the end of the fan blade and the fan casing. The Blade Pass Frequency (BPF) of this noise is proportional to the fan speed s (rpm) and the number of fan blades n [2], The BPF is generated as:

$$BPF = n * s / 60 \tag{1}$$

The fan noise is composed of this frequency and its multiplier frequency.

The fan used in this experiment is an axial fan with 7 blades, an operating voltage of 12 volts, and a rotational speed of 7000 rpm. According to the fundamental frequency calculation formula (1), the noise fundamental frequency can be obtained to be about 817Hz. We use a microphone to measure the noise spectrum of the fan at 30 cm in front of the air outlet, as shown in Fig. 1. Since this experiment focuses on eliminating noise within 3000Hz, the fundamental frequency of 817Hz, its double frequency of 1634Hz and its triple frequency of 2451Hz are the most important narrowband noises. From the spectrum, it could be seen that there is a high-energy noise in the 0~500Hz part. Since we are recording from the air outlet of the fan, it is the sound caused by the wind hitting the microphone. In this experiment, the 1x1 feedback FXLMS [3] algorithm and the 1x1 parallel form narrowband FXLMS

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

algorithm would be used to suppress the noise at the air outlet of the fan for narrowband noise.

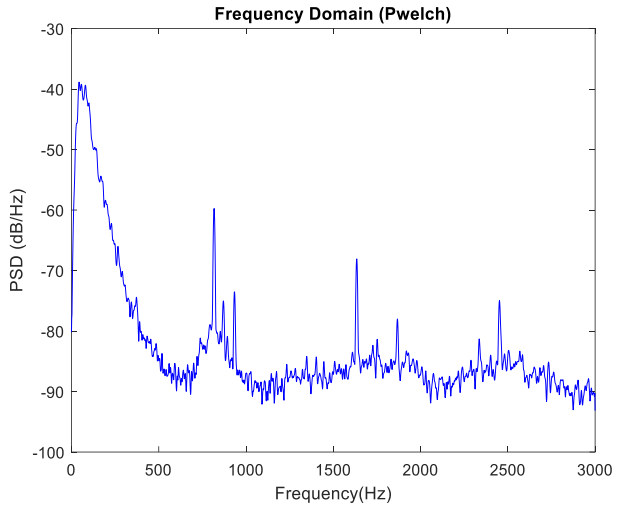


Fig. 1 The noise spectrum of the fan at 30 cm in front of the air outlet.

B. *Feedback FXLMS algorithm*

The architecture of the single channel feedback ANC system is shown in Fig. 2. In the feedback FXLMS algorithm [4], the reference signal $x(n)$ is obtained by combining the error signal $e(n)$ and the output signal $y(n)$, and the calculation formula is as follows:

$$x(n) \equiv d'(n) = e(n) + \sum_{m=0}^{M-1} \hat{s}_m y(n-m) \qquad (2)$$

where $\hat{s}_m$ is the coefficient of the M-order estimated secondary path filter $\hat{S}(z)$. The system obtains the output signal $y(n)$ from the reference signal $x(n)$ through the filter $W(z)$:

$$y(n) = \sum_{l=0}^{L-1} w_l(n) x(n-l) \qquad (3)$$

The signal $x(n)$ becomes $x'(n)$ after passing through the estimated secondary path:

$$x'(n) = \sum_{m=0}^{M-1} \hat{s}_m x(n-m) \qquad (4)$$

And update the filter coefficients with $x'(n)$, μ and $e(n)$. Where μ is the step size:

$$w_l(n+1) = w_l(n) + \mu x'(n-l) e(n), l = 0,1,\dots,L-1 \qquad (5)$$

Update the coefficient $w_l(n)$ obtained by the finite impulse response (FIR) of the adaptive filter until the system converges to achieve the effect of denoising.

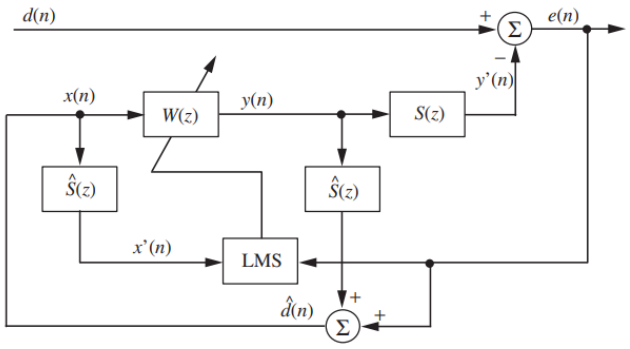


Fig. 2 The block diagram of adaptive feedback ANC system.

C. *Parallel form narrowband FXLMS algorithm*

Parallel form narrowband ANC [5] is based on the FXLMS algorithm, which splits the signal obtained by the non-acoustic sensor into multiple reference signals at a single frequency. Its block diagram is shown in Fig. 3.

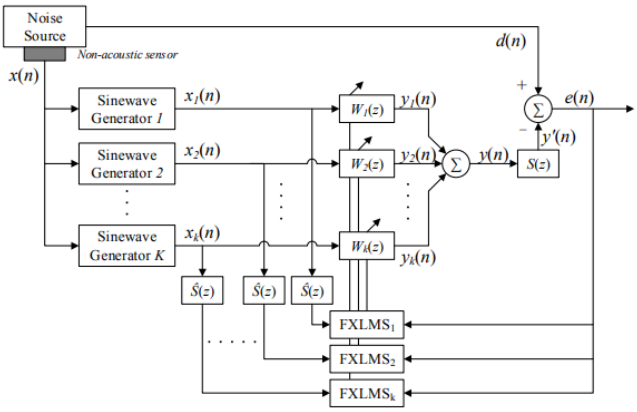


Fig. 3 The block diagram of parallel form narrowband ANC system.

The reference signal $x(n)$ can be split into multiple sine wave signals $x_1(n)$, $x_2(n)$ to $x_K(n)$:

$$x_k(n) = A_k \cos(\omega_k n), k = 1,2,\dots,K \qquad (6)$$

where K is the total number of sine wave signals. And the output signal $y(n)$ is accumulated after the filter of the adaptive filter $W_k(z)$:

$$y(n) = \sum_{k=l}^{K} \sum_{l=0}^{L-1} w_{kl}(n) x_k(n-l) \qquad (7)$$

The reference signal $x_k(n)$ becomes $x'_k(n)$ after passing through the estimated secondary path:

$$x'_k(n) = \sum_{m=0}^{M-l} \hat{s}_m x_k(n-m), k = 1, \dots, K \quad (8)$$

where $\hat{s}_m$ is the coefficient of the M-order estimated secondary path filter $\hat{S}(z)$.

$$w_{kl}(n+1) = w_{kl}(n) + \mu_k x'_k(n-l)e(n) \quad (9)$$

where $l = 0,1, \dots, L; k = 1,2, \dots, K$ and μ_k is the step size. Constantly update and adjust the filter coefficients in (9) until the system converges.

D. Experimental Setup

We would use a 10cm acrylic duct to stick to the air outlet of the fan and seal it. The purpose is to make the noise easy to control, and use the acrylic duct to limit the direction of sound travel. The installation configuration is shown in Fig. 4. The speaker was installed in the center of the fan outlet where there

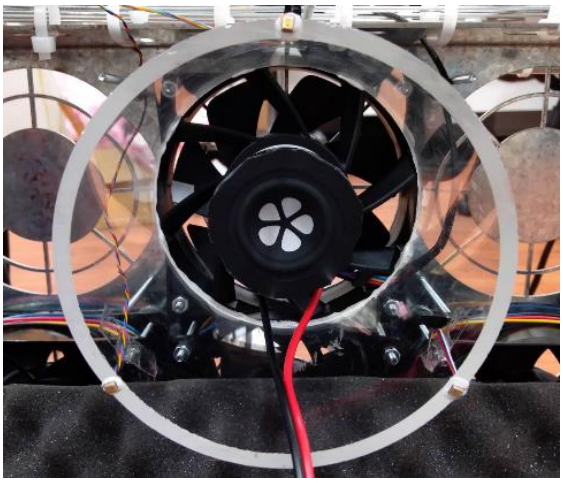


Fig. 4 Experimental setup of ANC fan system.

was no wind. Three error microphones were arranged on the nozzle. The angle between the microphones was 120°, and the received signals were added and averaged.

In order to measure the global noise cancellation effect, 25 microphone measurement points were placed 20 cm in front of the fan outlet, as shown in Fig. 5. And according to equation (10), the noise energy distribution is calculated. P is the energy, N is the sampling number, $y(n)$ is the amplitude of the sampling point.

$$P = 10 * log_{10}[\ 1/N * \sum_{n=1}^{N-1} y(n)^2\ ] \quad (10)$$

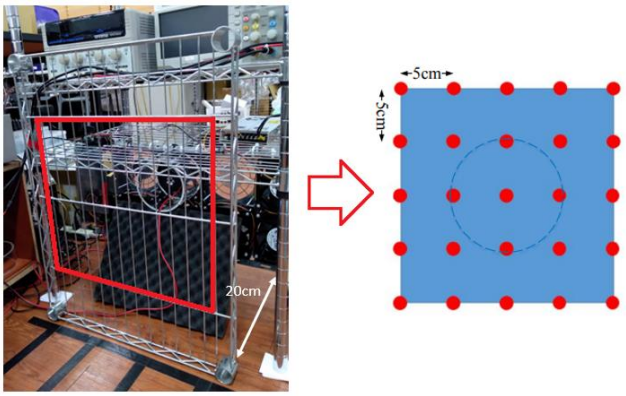


Fig. 5 Axial fan noise distribution measurement points.

III. Real-Time Experiments Results And Analysis

This section introduces the developed axial fan ANC real-time experiments. We improved the initial settings to verify the impact of microphone protection on system performance. After analyzing and comparing, we got the best setting results and showed the noise reduction results of the experiment.

A. Microphones without any protection

Since the microphones were not protected in the settings of the Fig. 4, its frequency spectrum was largely masked by broadband, resulting in only the fundamental frequency and its triple frequency being exposed. We used the feedback algorithm for noise reduction, as shown in Fig. 6. Since the feedback algorithm replaces the reference microphone by

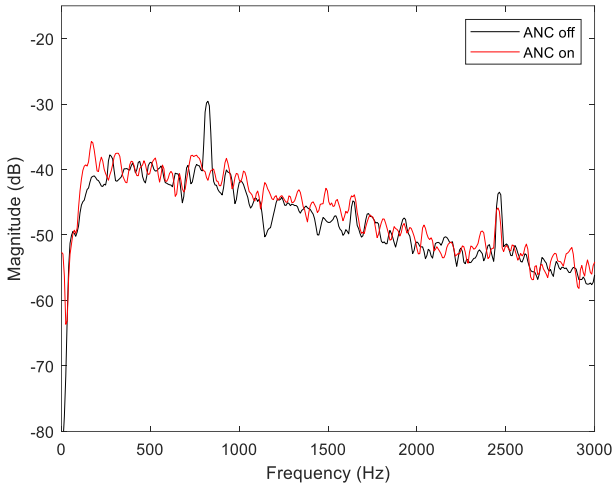


Fig. 6 The spectrum measured at the error microphone in the feedback system.

combining the past error signal with the output signal, the noise must be predictable [6]. However, because the microphones were not protected, the error microphones would feedback a broadband noise generated by the wind, that is, unpredictable noise, to the system for update, which would affect the system performance.

Next, we used the parallel form narrowband algorithm to perform ANC, and set 817Hz and 2451Hz as reference signals. The frequency spectrum is shown in the Fig. 7. It could be found that when ANC was on, although the fundamental frequency dropped, a broadband noise was generated, which exceeded the energy of ANC off. We speculate that this is also the result of the microphone not being protected.

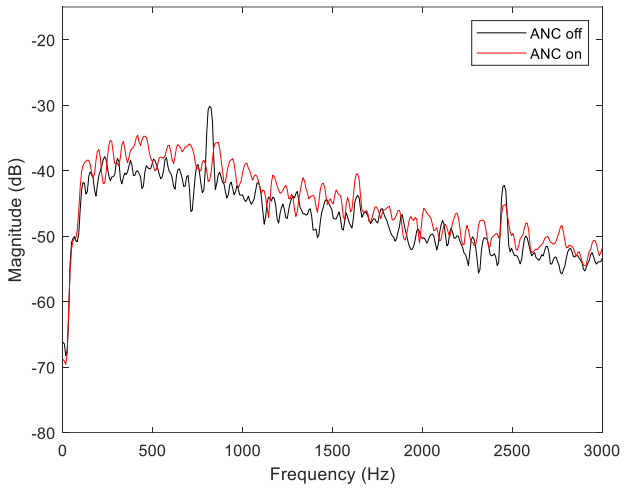


Fig. 7 The spectrum measured at the error microphone in the narrowband system.

B. Microphones with tape protection

In order to solve the above problem, we would use the settings in Fig. 8 and use the tape to protect the microphones. First, we performed the feedback algorithm, as shown in Fig. 9, it could be seen that when the ANC was on, there was no additional broadband noise. The effect of ANC was much better than the settings of unprotected microphones and the fundamental frequency could be reduced by 15dB. In addition, we used parallel form narrowband to perform ANC, as shown in Fig. 10, which was also better than the previous settings, the double frequency of 1634Hz could also be reduced by 2 or 3dB.

C. Microphones with acoustic sponge protection

We would use the settings in Fig. 11 and use the sound-absorbing sponges to protect the microphones. First, we

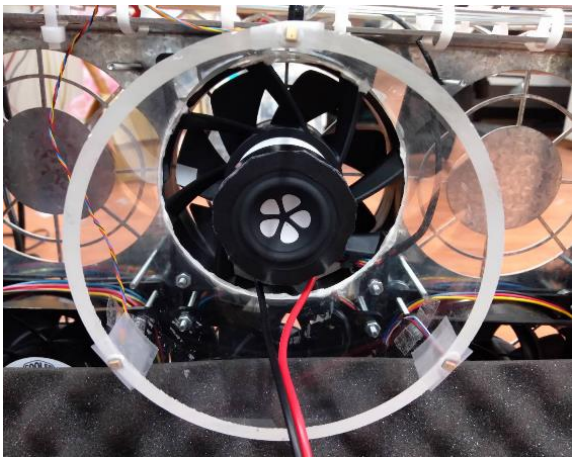


Fig. 8 We used the tape to protect the microphones.

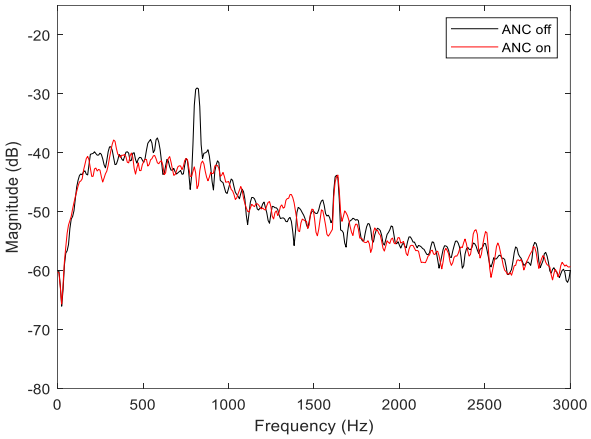


Fig. 9 The spectrum measured at the error microphone in the feedback system.

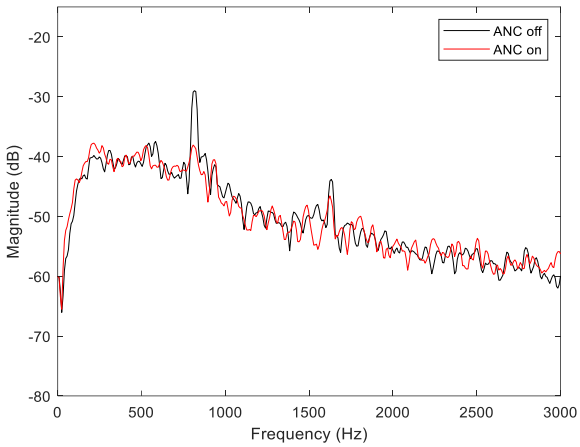


Fig. 10 The spectrum measured at the error microphone in the narrowband system.

performed the feedback algorithm, as shown in Fig. 12, it could be seen that when the ANC was on, there was no additional broadband noise. Applying the feedback algorithm to this setting could reduce the fundamental frequency by about 18dB.

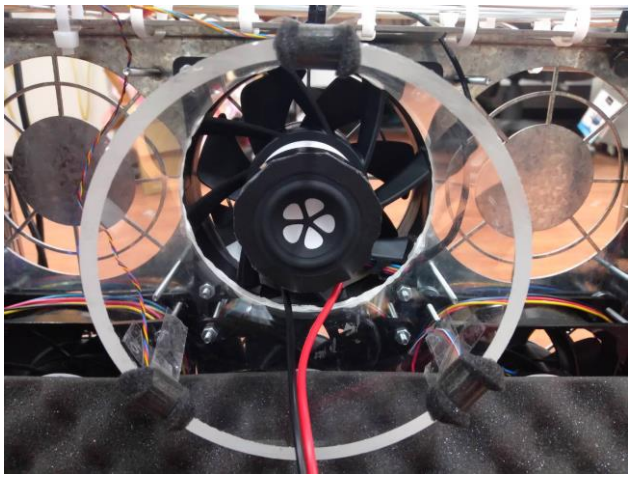


Fig. 11 We used the sound-absorbing sponge to protect the microphone.

From this spectrum, it could be seen that the energy after the reduction of the fundamental frequency was still higher than the energy of the double and triple frequencies, meaning that the noise energy of 1634Hz and 2451Hz was too low, so that the system only processed the noise with stronger energy. That's why we can see that only the fundamental frequency has effect.

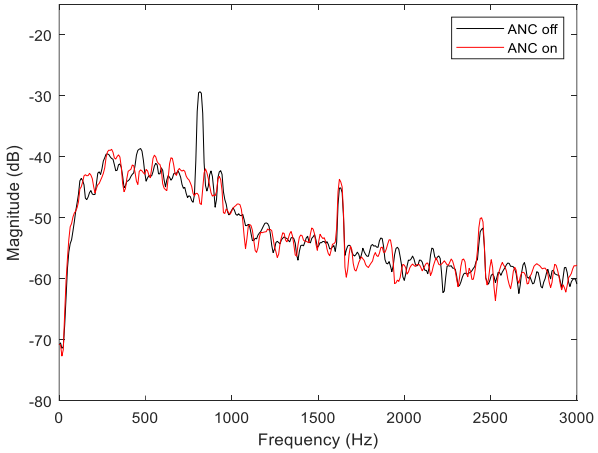


Fig. 12 The spectrum measured at the error microphone in the feedback system.

We used parallel form narrowband to perform ANC, and set 817Hz, 1634Hz, and 2451Hz as reference signals. The frequency spectrum is shown in the Fig. 13. We can see from

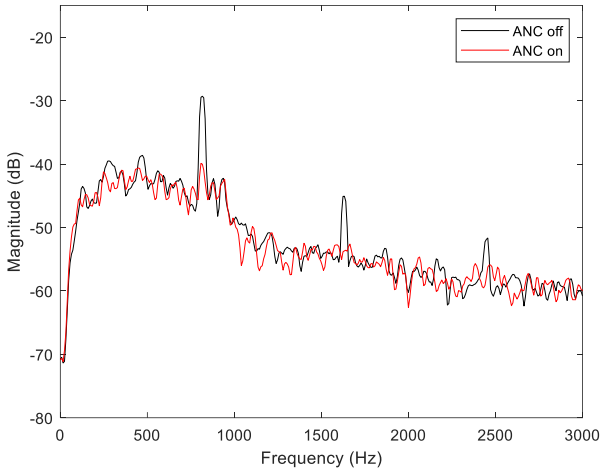


Fig. 13 The spectrum measured at the error microphone in the narrowband system.

the spectrum that there was no broadband noise, and the single frequencies were eliminated at the frequencies of the reference signals we set. This proves that the poor performance of the previous setup was due to the influence of the wind, and that if we protected the microphones well, we could make the system perform better.

We also used the method in Fig. 5 to measure the noise energy distribution diagram of the parallel form narrowband system with acoustic sponge protection. We subtracted the energy of ANC off from the energy of ANC on, and plotted the distribution diagram of the noise reduction energy, as shown in Fig. 14. It could be seen that the main noise reduction was distributed at the overlap of the microphones and the upper right corner, which showed that this system was effective and could achieve global noise reduction.

IV. CONCLUSIONS

We optimized the ANC performance of the axial fan system by changing the setting. From the experiment, we can know that the important factor will affect the system performance, it is the wind. And the microphones need to be protected to avoid the noise caused by the wind. It has been proved by experiments that if the microphone is well protected, the performance of ANC system can be improved.

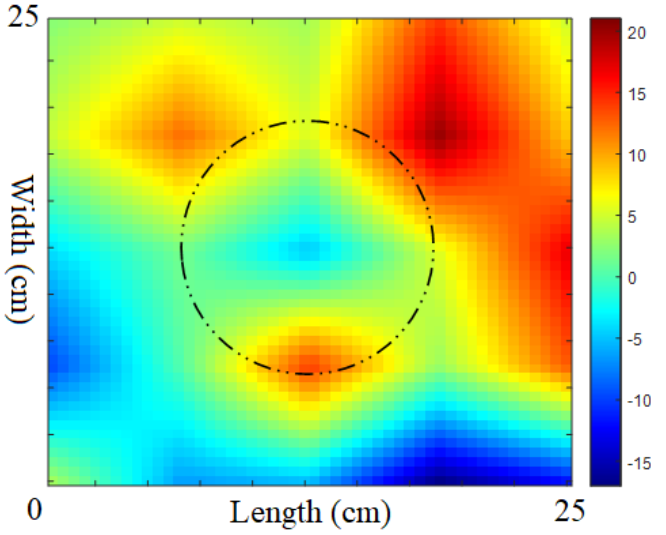


Fig. 14 Axial fan noise energy distribution diagram in narrowband system.

REFERENCES

[1] P. Lueg Process of silencing sound oscillations. U.S. patent, 1936, No. 2,043,416.

[2] J. J. Esplin, Active noise control of a centrifugal fan mounted in a mock laptop enclosure, M.S. Thesis, Dept. of Physics and Astronomy. Brigham Young University, Aug, 2012.

[3] J. C. Burgess, "Active Adaptive Sound Control in a Duct: A Computer Simulation," Journal of Acoustical Society of America, vol. 70, no. 3, pp. 715-726, 1981.

[4] S. M. Kuo, D. R. Morgan, "Active Noise Control: a tutorial review," Proceedings of the IEEE, vol. 87, no. 6, pp.943-975, 1999.

[5] Y. Kajikawa, W. S. Gan, and S. M. Kuo, "Recent Advances on Active Noise Control: Open Issues and Innovative Applications" APSIPA Transactions on Signal and Information Processing, vol. 1, pp. 1-21, August 2012.

[6] S. M. Kuo and D. R. Morgan, Active Noise Control Systems Algorithms and DSP Implementations, NY: Wiley, 1996.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

SFANC with Compensation Filter Based on MEFxDCTLMS Algorithm

Kenya Doi* and Yoshinobu Kajikawa†
Kansai University, Osaka, Japan
* E-mail: k333286@kansai-u.ac.jp
† E-mail: kaji@kansai-u.ac.jp

***Abstract*—This paper proposes a hybrid system of the selective fixed-filters active noise control (SFANC) and the modified error filtered-x DCT-LMS (MEFxDCTLMS) algorithm. Conventional SFANC systems have a problem that the noise reduction performance deteriorates due to the incorrect selection of fixed filters. In addition, the conventional SFANC system uses the spectrogram by short-time Fourier transform (STFT) of the reference signal detected by the reference microphone as an input in a two-dimensional convolutional neural network (2D-CNN). However, the processing delay for calculating the spectrogram delays the fixed filter selection in SFANC, resulting in degraded noise reduction performance. To compensate for the fixed filter, we propose a system in which an adaptive filter based on MEFxDCTLMS is vertically connected to SFANC. This improves the degradation of the noise reduction performance due to the mis-selection of the fixed filter and improves the robustness against changes in the target noise. Furthermore, by using Sliding DCT instead of STFT, the fixed filter selection can be performed at the appropriate timing while maintaining the selection accuracy of 2D-CNN, thereby improving the noise reduction performance.**

I. Introduction

In recent years, the development of transportation and industrial equipments has caused the noise in a variety of situations. Active noise control (ANC) systems [1]-[3] are used to reduce such noise. It is an effective method to reduce the noise level at the noise control point using anti-noise. Anti-noise with the same amplitude and opposite phase for the unwanted noise is generated and emitted to reduce the noise at the noise control point based on the principle of superposition. Conventional ANC systems typically use noise control filters designed by adaptive algorithms [4] such as filtered-x normalized least mean square (FxNLMS) algorithm to generate anti-noise.

Among the ANC methods that have been proposed, the selective fixed-filters ANC (SFANC) [5]-[9] is expected to be a computationally efficient noise reduction method because it is based on pre-designed fixed filters. However, there are some problems in SFANC. The SFANC system uses the convolutional neural network (CNN) [10]-[13] as the noise classifier and switches multiple pre-designed noise control filters according to the noise. This allows the system to respond quickly to changes in the noise. In conventional 2D-CNN based SFANC, acoustic signals are transformed to a frequency-domain representation using the short-time Fourier transform (STFT) [14]. The STFT calculation is performed by applying a window function to the signal and summing the extracted samples. Therefore, to increase the frequency resolution, it is necessary to use a window function with a long window length. However, a longer window length increases the computational complexity and decreases the response speed to noise changes, making it unsuitable for implementation in ANC systems. In addition to the STFT, the sliding discrete cosine transform (SDCT) is another method of transforming a signal into the frequency domain [15]. SDCT can instantly convert a signal from the time domain to the frequency domain using a recursive formula and requires far fewer operations than STFT. Sliding Discrete Fourier Transform (STFT), a method similar to SDCT, requires an additional operation to calculate absolute values to convert from the real time domain to the complex frequency domain. In this paper, SDCT is used as input for 2D-CNN in SFANC.

Another problem in SFANC is that immediately after noise variation, the filter selected by the classifier may not match the correct filter suitable for the noise. To solve this problem, a hybrid algorithm of SFANC and FxNLMS has been proposed in applied research of SFANC [8]. The hybrid system achieved high tracking performance for noise fluctuations. Moreover, it is stated that future developments in this research are expected to combine SFANC with other adaptive algorithms. Therefore, we propose to combine SFANC with modified error FxDCT-LMS (MEFxDCTLMS) algorithm. In this paper, we show the effectiveness of the hybrid SFANC-MEFxDCTLMS system.

II. Hybrid SFANC-MEFxDCTLMS

In the Hybrid SFANC-MEFxDCTLMS, 2D-CNN using SDCT is used as the classifier and the MEFxDCTLMS based adaptive filter fine-tunes the secondary noise.

A. Sliding Discrete Cosine Transform

DCT is a method of converting a time domain signal into the frequency domain. This method is more efficient than DFT because it uses a cosine function and outputs only the real part. The DCT coefficients of the signal sequence $u(n)$ at time n are defined as

$$C_m(n) = \frac{1}{2} k_m (-1)^m W_{2M}^{m/2} A_m(n), \quad (m = 0, 1, \cdots, M-1), \quad (1)$$

where

$$k_m = \begin{cases} \sqrt{2} & (m = 0) \\ 1 & (\text{otherwise}) \end{cases}, \quad (2)$$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

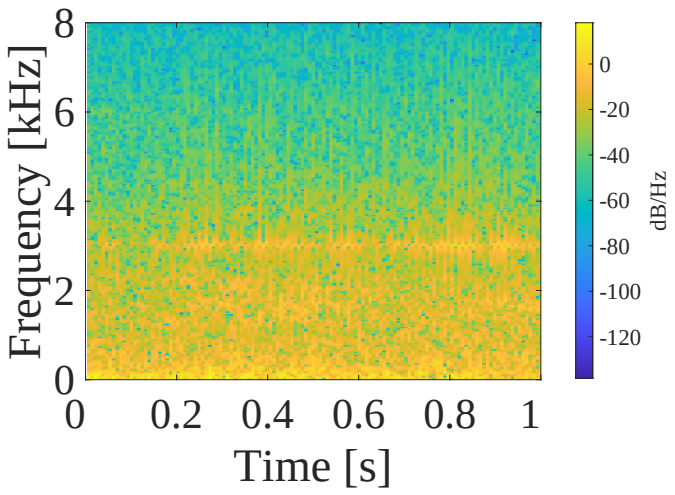


Fig. 1. DCT coefficient color map of airplane noise.

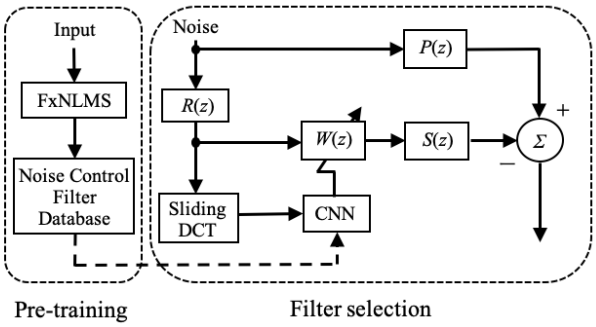


Fig. 2. Block diagram of SFANC.

$$W_{2M} = \exp\left(-\frac{j2\pi}{2M}\right). \tag{3}$$

M is the number of taps, $A_m(n)$ is the sum of two complementary DFT presented by

$$A_m(n) = A_m^{(1)} + A_m^{(2)}. \tag{4}$$

where

$$A_m^{(1)}(n) = W_{2M}^{m} A_m^{(1)}(n) + u(n) - (-1)^m u(n-M), \tag{5}$$

$$A_m^{(2)}(n) = W_{2M}^{-m} A_m^{(2)}(n) + W_{2M}^{-m}\left(u(n) - (-1)^m u(n-M)\right). \tag{6}$$

The SDCT is obtained from the first-order difference equation, so the computational complexity is negligible. Fig. 1 shows an example of the SDCT coefficients obtained from equation (1) arranged in a time series. Fig. 1 plots the SDCT coefficients for 1 second of aircraft noise on a color map.

B. Selective Fixed-filter ANC with SDCT input CNN

Fig. 2 shows a block diagram of the SFANC system, where $P(z)$ is the primary path, $S(z)$ is the secondary path, and $W(z)$ is the noise control filter selected from the filter database by CNN. The SFANC system is divided into two main stages. The first is the fixed filter pre-design stage. Fixed filters are pre-designed and stored in a database using adaptive algorithms for noise with different frequency spectra. The second is the noise control stage. The noise control filter that provides the best reduction for the noise acquired with the reference microphone is selected from the database and applied to the ANC system. In SFANC systems, CNNs are commonly used as noise classifiers because they are well suited to handle the frequency representation of noise signals.

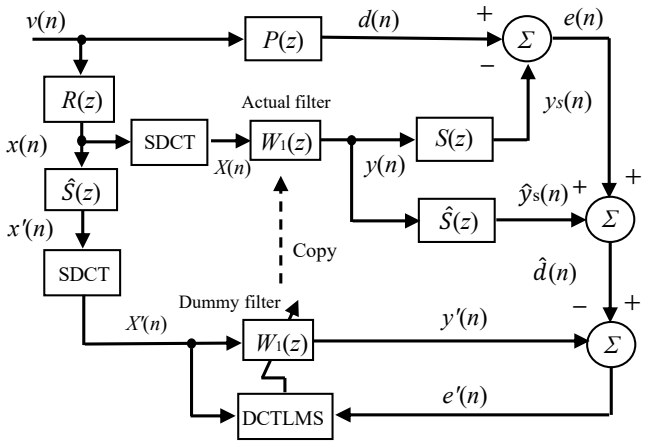


Fig. 3. Block diagram of MEFxDCTLMS based ANC system.

C. MEFxDCTLMS Algorithm

MEFxDCTLMS is an application of MEFxNLMS with a different adaptive algorithm. Fig. 3 shows a block diagram of the MEFxDCTLMS based ANC system, where $P(z)$ is the primary path, $S(z)$ is the secondary path, $\hat{S}(z)$ is the model of the $S(z)$, and $W_1(z)$ is the noise control filter based on MEFxDCTLMS.

In Fig. 3, the noise control filter $W_1(z)$ is updated by

$$w_m(n+1) = w_m(n) + \frac{\mu}{\hat{\lambda_m}(n)} X'_m(n) e'(n). \tag{7}$$

where μ is the step size parameter for the noise control filter, $X'_m(n)$ is the filtered SDCT coefficients, $e'(n)$ is the error signal given as $e'(n) = \hat{d}(n) - y'(n)$, and $\hat{\lambda}$ is the eigenvalue expressed by the recursive equation

$$\hat{\lambda_m}(n) = \gamma \hat{\lambda_m}(n-1) + \frac{1}{n}(X'^2_m(n) - \gamma \hat{\lambda_m}(n-1)). \tag{8}$$

The forgetting factor γ is used in the range of 0 to1. After updating, the filter weight coefficients are copied to noise control filter $W_1(z)$. This system uses a secondary path model to generate the corrected error signal, so the delay caused by the secondary path can be disregarded. Therefore, the step size parameter can be larger than when using the error signal, which speeds up the convergence of the algorithm.

D. Implementation of MEFxDCTLMS in SFANC

Fig. 4 shows a block diagram of the hybrid MEFxDCTLMS-SFANC system. In the proposed method, an adaptive filter

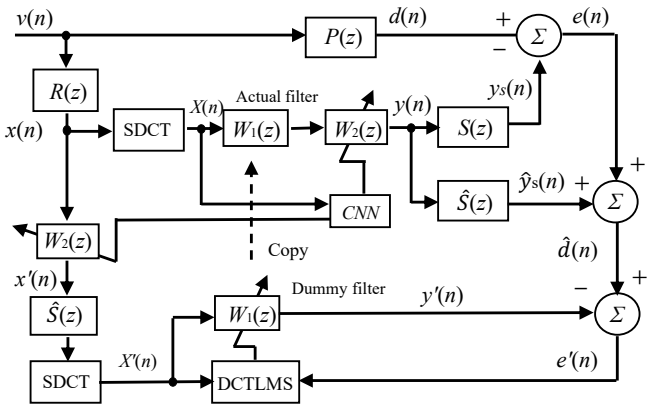


Fig. 4. Block diagram of the hybrid MEFxDCTLMS and SFANC system.

for fine tuning is connected in series with a fixed filter. This allows the adaptive filter to play a supplementary role to the fixed filter, and it is thought that it can achieve higher noise reduction than only the fixed-filter ANC. MEFxDCTLMS performs background optimization with dummy filters. Therefore, if adaptive algorithms are not needed, computational efficiency can be improved by running only the SFANC system. The system determines that the type of noise has changed by comparing the filter selection results with those of the past. Only fixed filters can be introduced at the moment the noise changes to allow time for the adaptive filter to converge.

III. Simulation Results

In this section, a 2D-CNN to be used as a noise classifier is trained by using audio datasets and implemented in SFANC. Further, this section verifies effectiveness of the proposed method for non-stationary noise through noise reduction simulations.

A. 2D-CNN Learning

Fig. 5 shows the CNN model architecture used for SFANC system. The middle layer of the CNN uses two convolution layers, batch normalization, ReLU activation function, and a Maxpooling layer. Of the 96,000 acoustic datasets, including synthetic and environmental sounds, e.g. household appliances and factory noise, car and airplane engine noise, 10,000 each are used for testing and validation, with the remainder used for training.The acoustic datasets are converted to the frequency domain by SDCT and the training labels are given in 15 classes based on the frequency characteristics of the acoustic signal. The number of training epochs is 50 and the Adam algorithm is used for optimization. The classification accuracy of the 2D-CNN on the test data was 97.06%.

B. Pre-designed Noise Control Filters

The pre-design of noise control filters by optimization using the broadband noise is shown in Table I. Design conditions for noise control filters are shown in Table II. Noise control

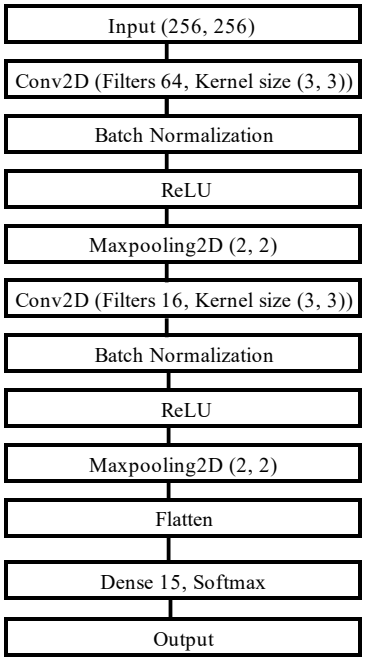


Fig. 5. The 2D-CNN architecture in SFANC system.

TABLE I
Labels for training datasets and frequency bands of broadband noise for filter design.

Trainning data	SDCT signal
Number of data	96000
Input size	(256,256)
Number of labels	15
Label 1-8	20-1000 Hz, 1000-2000 Hz, ⋯, 7000-8000 Hz
Label 9-12	20-2000Hz, 2000-4000 Hz, ⋯, 6000-8000 Hz
Label 13,14	20-4000 Hz, 4000-8000 Hz
Label 15	20-8000 Hz

filters for SFANC are pre-designed using the MEFxDCTLMS algorithm. Each filters are effective for signals with a frequency response similar to that of the signal used in the pre-design. These filters are stored in a database and the appropriate filters are selected from the CNN classification results and used in the SFANC system. Each of the filters created in this study can reduce the target noise by 15 to 20 dB. In this study, the labeling of datasets is done using these 15 filters. Each noise control filter is applied to the signal data and a label is given corresponding to the filter with the highest reduction.

C. Comparison of Noise Reduction Performances

The noise reduction performance of the proposed and conventional methods is compared by computer simulations. In the proposed method, the compensation filter $W_1(z)$ based on MEFxDCTLMS is initialized at system startup, constantly updated in the background, and copied to the system. Fig. 6 shows spectrogram of input noise. The input signal for verification is switched every 0.5 seconds between six different types of noise with different frequency characteristics, such

TABLE II
DESIGN CONDITIONS FOR NOISE CONTROL FILTERS.

Sampling rate	16000 Hz
Iteration	20000
Tap length of noise control filters M	256
Tap lengths of primary, secondary, and reference paths	256
Adaptive algorithm	MEFxDCTLMS
Step size parameter α	$1/2M$
Recursive factor β	0.99
Forgetting factor γ	0.99

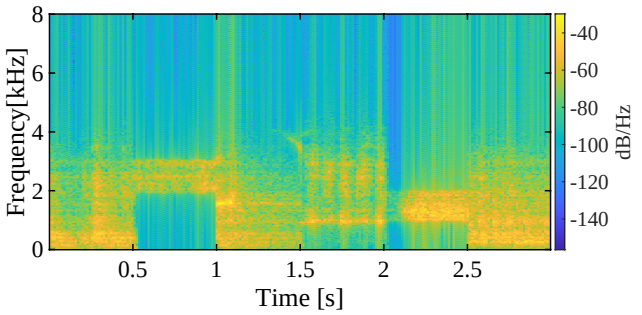
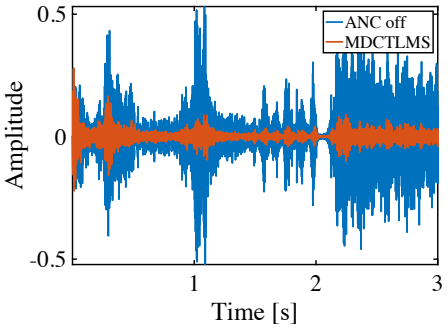
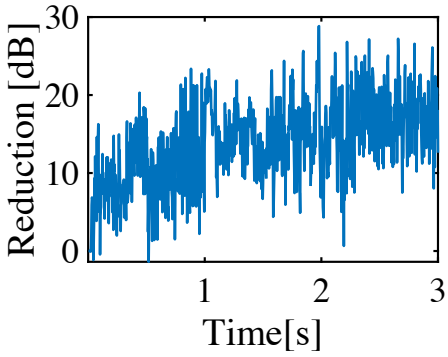


Fig. 6. Spectrogram of input noise.

as train, washing machine, and factory noise. Figs. 7-9 show noise reduction performance of conventional and proposed ANC systems, and Fig. 10 compares them. Fig. 11 shows the transition for the selection of noise control filter. From Fig. 10, the conventional SFANC method degrades the reduction performance due to the incorrect filters, whereas the proposed method improves it. This is because the MEFxDCTLMS-based filter works to compensate for the fixed filter output. In another aspect, ANC based on MEFxDCTLMS has degradation in reduction performance from the time of system startup until the adaptive filter converges. On the other hand, the proposed method immediately selects a fixed filter when the noise changes, and thus has better tracking performance than ANC based on MEFxDCTLMS. However, the noise reduction effect after adaptive algorithm convergence is the same, the proposed method has instantaneous performance degradation immediately after noise change. This is caused by a delay between the occurrence of a noise change and its detection. To avoid this problem, delay compensation could be added when switching filters.

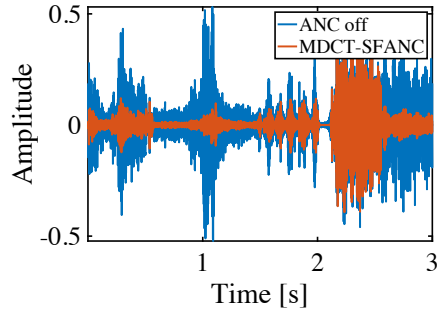


(a) Time wave form

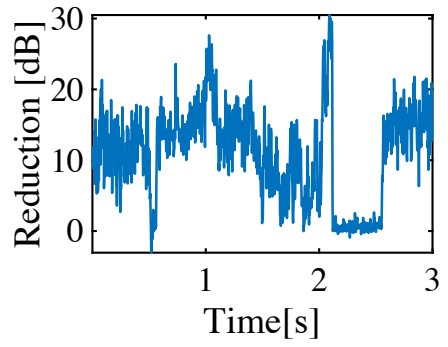


(b) Noise attenuation

Fig. 7. Noise reduction perfomance for MEFxDCTLMS based ANC.

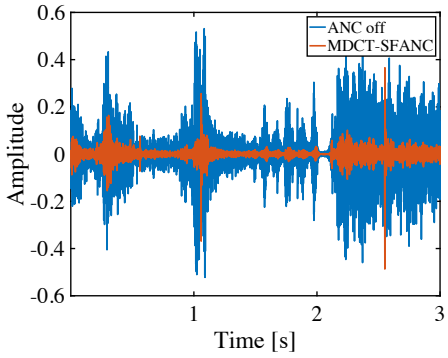


(a) Time wave form

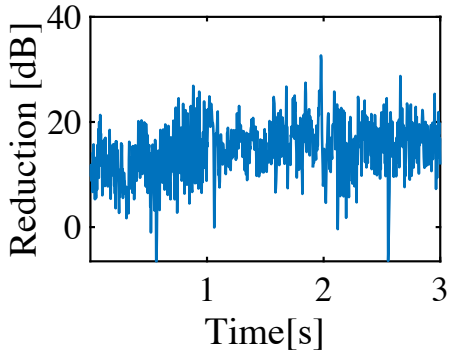


(b) Noise attenuation

Fig. 8. Noise reduction perfomance for SFANC.



(a) Time wave form



(b) Noise attenuation

Fig. 9. Noise reduction perfomance for hybrid MEFxDCTLMS-SFANC.

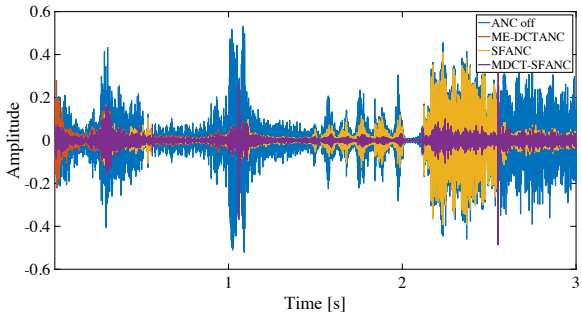


Fig. 10. Comparison of noise reduction perfomance.

IV. Conclusions

In this paper, we proposed the hybrid MEFxDCTLMS-SFANC system with SDCT input 2D-CNN. In this system, an adaptive compensation filter is added to SFANC to improve robustness. Computer simulation results show that the proposed method is robust against noise variations and filter misselection. However, it was also found that the reduction performance drops momentarily immediately after the noise changes. To solve this problem, we plan to add delay compensation for filter selection. We will also study systems that are robust to transmission path variations in noise.

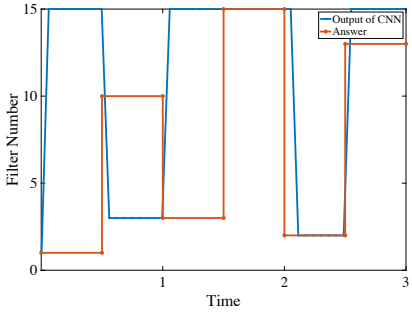


Fig. 11. Transition for the selection of noise control filter.

References

[1] S. M. Kuo and D. R. Morgan, "Active Noise Control: A Tutorial Review," Proceeding of the IEEE, vol. 87, no. 6, pp. 943–973, Jun. 1999.
[2] S. J. Elliott, Signal Processing for Active Control, Academic Press, San Diego, CA, 2001.
[3] Y. Kajikawa, W.-S. Gan and S. M. Kuo, "Recent Advances on Active Noise Control: Open Issues and Innovative Applications," APSIPA Transactions on Signal and Information Processing, vol. 1, Aug. 2012.
[4] S. Haykin, Adaptive Filter Theory, 5th Edition, Pearson, 2013.
[5] D. Shi, W. -S. Gan, B. Lam, S. Wen, "Feedforward Selective Fixed-Filter Active Noise Control: Algorithm and Implementation," IEEE /ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1479–1492, May 2020.
[6] D. Shi, B. Lam, K. Ooi, X. Shen, W. -S. Gan, "Selective Fixed-filter Active Noise Control Based on Convolutional Neural Network," Signal Processing, vol. 90, 108317, 2022.
[7] Z. Luo, D. Shi, W. -S. Gan, L. Zhang, Q. Huang, "Performance Evaluation of Selective Fixed-filter Active Noise Control based on Different Convolutional Neural Networks," arXiv:2208.08082, Aug. 2022.
[8] Z. Luo, D. Shi and W. -S. Gan, "A Hybrid SFANC-FxNLMS Algorithm for Active Noise Control Based on Deep Learning," IEEE Signal Processing Letters, vol. 29, pp. 1102–1106, 2022.
[9] D. Shi, W. -S. Gan, Z. Luo, B. Lam, X. Shen, "Transferable Latent of CNN-based Selective Fixed-filter Active Noise Control," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 14, no. 8, pp. 1–12, Mar. 2023.
[10] G. Huang, Z. Liu, L. van der Maaten, "Densely Connected Convolutional Networks," CVPR 2017, Jan. 2017.
[11] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, Apr. 2015.
[12] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," arXiv:1512.03385, Dec. 2015.
[13] M. Sandler, A. Howard, M. Zhu A. Zhmoginov L. -C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," arXiv:1801.04381, Jan. 2018.
[14] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. Acoust. Speech Signal Process., vol. 25, no. 3, 235-238, 1977.
[15] V. Kober, "Fast Algorithms for the Computation of Sliding Discrete Sinusoidal Transforms," IEEE Trans. on Signal Processing, vol. 52, no. 6, 1704–1710, Jun. 2004.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Practical Active Noise Control: Restriction of Maximum Output Power

Woon-Seng Gan*, Dongyuan Shi* and Xiaoyi Shen*
* Digital Signal Processing Lab, School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore
E-mail: ewsgan@ntu.edu.sg

***Abstract*—This paper presents some recent algorithms developed by the authors for real-time adaptive active noise (AANC) control systems. These algorithms address some of the common challenges faced by AANC systems, such as speaker saturation, system divergence, and disturbance rejection. Speaker saturation can introduce nonlinearity into the adaptive system and degrade the noise reduction performance. System divergence can occur when the secondary speaker units are over-amplified or when there is a disturbance other than the noise to be controlled. Disturbance rejection is important to prevent the adaptive system from adapting to unwanted signals. The paper provides guidelines for implementing and operating real-time AANC systems based on these algorithms.**

Keywords: adaptive active noise control algorithm, output saturation, 2GD-FxLMS, leaky FxLMS

I. Introduction

Active noise control (ANC) system produces anti-noise to cancel out unwanted noise [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. However, adaptive ANC (AANC) in real-time [9], [11], [12], [13], [14] may fail if the audio amplifier is overdriven to saturation, causing the adaptive algorithm to diverge. Saturation distortion affects both the amplitude and phase of the secondary path in ANC [15], [16]. Some nonlinear algorithms based on neural networks or Volterra filters [17] can handle saturation distortion, but they are costly and complex to implement, and the vanishing gradient issue is a problem for real-world scenarios. Moreover, they do not solve the problem of insufficient power when the control signal exceeds the amplifier threshold. Thus, these nonlinear solutions are not practical for real applications and may only be used for partially overdriven actuators. Furthermore, the nonlinear adaptive algorithm without constraints is not a desired solution. A better approach is to limit the amplifier's output power and keep it within a certain power budget. Some of the common algorithms for this approach are (i) clipping or rescaling algorithms [18], which either truncate the output signal above the threshold or adjust the weight of the control filter; (ii) leaky-type filtered-reference Least Mean Square (FxLMS) algorithm [19], which adds a leak term or penalty factor to stabilize the algorithm, but requires trial and error to choose the optimal factor; (iii) A two-gradient FxLMS (2GD-FxLMS) [20] algorithm, which imposes a specific output constraint with no extra computation compared to the conventional FxLMS algorithm. Recent research has revealed new insights into constraint-FxLMS algorithms. This paper highlights the key differences and evaluates their merits and drawbacks. A typical block diagram of the ANC with saturation effect can be viewed in Fig. 1 and will be used throughout the discussion of different constraint FxLMS.

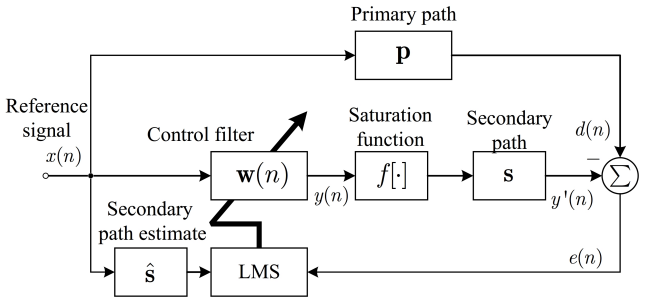


Fig. 1. Block diagram of the ANC with the saturation effect.

Before we move into the definition, explanation, and implementation of the various types of constraint FxLMS algorithms, we have defined a table of variables and their descriptions for ease of reading.

$\mathbf{w}(n)$: weight vector of the control filter	$e(n)$: error signal
$\mathbf{x}(n)$: reference signal vector	$\mathbf{x}'(n)$: filtered reference signal vector
μ: step size of the adaptive algorithm	γ: leaky factor
ρ^2: output power constraints	$\mathbf{s}$: impulse response of secondary path
L: length of the control filter	$d(n)$: disturbance
λ: Lagrangian factor	σ^2:variance of disturbance
y_{o}: optimal output	σ_{x}: variance of reference
$\mathbf{w}_{\mathrm{o}}$: optimal weight vector	η: degree of nonlinearity
$\mathbf{R}_{\mathrm{x}'}$: autocorrelation matrix of $\mathrm{x}'(n)$	$\mathbf{R}_{\mathrm{y}}$: autocorrelation matrix of $y(n)$
I: identity matrix	$\mathbf{P}_{\mathrm{dx}'}$: cross-correlation vector of $d(n)$ and $x'(n)$
λ_{o}: optimal Lagrangian factor	γ_{o}: optimal leaky factor
G_{s}: power gain of the secondary path	$S(e^{j\omega})$: transfer function of the secondary path

II. Clipping FxLMS

The clipping algorithm mainly truncates the parts of the output signal that exceeds a certain voltage level. However,

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP50122-0018).

this output amplitude constraint approach has a drawback of stability, distorts the attenuated noise, and may converge slowly. Therefore, in this paper, we will not consider this approach. A similar type of clipping algorithm is the nonlinear FxLMS algorithm [21], which consists of an exponential term that clips off the output signal when the control signal exceeds the output constraints.

Even though the clipping FxLMS algorithms can deal with an amplitude that exceeds the constraint, these simple approaches may not be ideal as these algorithms distort the attenuated noise, require an accurate model of the output-saturation model, and have significant residual distortion.

III. Output Constraint: Leaky-type FxLMS Algorithms

The leaky FxLMS (LFxLMS) algorithm [22], [23] introduces a leak term (or penalty coefficient) that constrains the control effort and stabilizes the algorithm. The general LFxLMS algorithm is given as

$$\mathbf{w}(n+1) = (1-\mu\gamma)\mathbf{w}(n) + \mu e(n)\mathbf{x}'(n), \tag{1}$$

$$e(n) = d(n) - f\left[\sum_{l=0}^{L-1} s_l \mathbf{w}^{\mathrm{T}}(n-l)\mathbf{x}(n-l)\right]. \tag{2}$$

In the LFxLMS algorithm, the selection of the leak factor plays a crucial role in balancing noise control and constraint satisfaction. However, due to the ad-hoc selection of the leak factor, the output signal may not always adhere to the imposed constraint. The time-domain version of the LFxLMS algorithm results in noise reduction errors and slower convergence. To address this issue, recent versions of the algorithm use frequency-domain constraints that only penalize frequency bins exceeding the constraint. Despite these improvements, the performance of the algorithm is still greatly influenced by the empirical choice of the leak factor.

Recently, an optimal leak factor selection method for output-constrained LFxLMS algorithm [23], [24] has been proposed. This latter technique ensures that the LFxLMS reaches optimal control within its specific output power constraint. A leak factor can be derived on-the-fly based on the application's measured primary and secondary acoustic paths. The cost function with a specific output constraint [25]

$$\begin{aligned} &\min_{\mathbf{w}} \mathrm{J}(\mathbf{w}) = \mathbb{E}\left[\left|d(n) - \sum_{l=0}^{L-1} s_l \mathbf{w}^{\mathrm{T}}(n-l)\mathbf{x}(n-l)\right|^2\right] \\ &\text{s.t. } g(\mathbf{w}) = \mathbb{E}\left[\left|\mathbf{w}^{\mathrm{T}}(n)\mathbf{x}(n)\right|^2\right] \le \rho^2, \end{aligned} \tag{3}$$

where $\mathbb{E}(\cdot)$ denotes the expectation of the argument. The Lagrangian function [26] based on (3) is defined as

$$\mathbb{L}(w,\lambda,\theta) = J(w) + \lambda\left[g(w) + \theta^2 - \rho^2\right]. \tag{4}$$

The Lagrangian factor λ is used in the above cost function to introduce constraints into the optimization problem, and the optimal Lagrangian factor, λ_{o} is derived as:

$$\lambda_{\mathrm{o}} = \frac{\mathbb{E}\left\{y'_{\mathrm{o}}(n)d(n)\right\} - \mathbb{E}\left\{y'_{\mathrm{o}}(n)^2\right\}}{\rho^2}. \tag{5}$$

The optimal Lagrangian factor, which includes the variance of the output signal from the control filter and the output filtered that passes through the secondary path (includes DAC, amplifier, secondary speaker, and the acoustic transfer path from speaker to error microphone), is shown below:

$$\lambda_{\mathrm{o}} = \frac{\sigma^2_{y_o}\sum_{l=0}^{L-1} s_l^2}{\rho^2}\left(\sqrt{\frac{\sigma_d^2}{\sum_{l=0}^{L-1} s_l^2\sigma^2_{y_o}}} - 1\right). \tag{6}$$

To further simplify the above expression, we can introduce a new term, known as the degree of nonlinearity of the system, η^2:

$$\eta^2 = \max\left(\frac{\sigma_d^2}{\sum_{l=0}^{L-1} s_l^2\rho^2}, 1\right), \tag{7}$$

which leads to

$$\lambda_{\mathrm{o}} = \sum_{l=0}^{L-1} s_l^2(\eta - 1). \tag{8}$$

The optimal control filter with output constraint is given as:

$$\mathbf{w}_{\mathrm{o}} = \left(\lambda_{\mathrm{o}}\sigma_{\mathrm{x}}^2\mathbf{I} + \mathbf{R}_{\mathrm{x}'}\right)^{-1}\mathbf{P}_{\mathrm{dx}'}. \tag{9}$$

The optimal control filter using LFxLMS is given as:

$$\mathbf{w}_{\mathrm{o}} = \left(\gamma\mathbf{I} + \mathbf{R}_{\mathrm{x}'}\right)^{-1}\mathbf{P}_{\mathrm{dx}'}. \tag{10}$$

Therefore, comparing these two optimal control filters, we have a sufficient condition (under white noise):

$$\gamma = \lambda_{\mathrm{o}}\sigma_{\mathrm{x}}^2. \tag{11}$$

In practical cases, we have to replace the sufficient condition as:

$$\gamma = \lambda_{\mathrm{o}}\mathbf{R}_{\mathrm{x}}. \tag{12}$$

Substituting the optimal Lagrangian factor into the above yields the optimal leak factor as shown:

$$\gamma_{\mathrm{o}} = G_{\mathrm{s}}(\eta - 1)\mathbf{R}_{\mathrm{x}}, \tag{13}$$

where $G_s = \sigma^2_{y'_{\mathrm{o}}}/\rho^2$ is the power gain of the secondary path and can be written as the energy of the secondary path if the control signal is narrowband over $[\omega_1, \omega_2]$ and the degree of nonlinearity of the system are:

$$G_{\mathrm{s}} = \frac{1}{2\pi}\int_{\omega_1}^{\omega_2}\left|S\left(e^{j\omega}\right)\right|^2 d\omega, \tag{14}$$

$$\eta^2 = \max\left(\frac{\sigma_{\mathrm{d}}^2}{G_{\mathrm{s}}\rho^2}, 1\right). \tag{15}$$

All the above statistic terms can be measured from the ANC system before its normal control operation. It is important to note that the LFxLMS algorithm with optimal leak factor is useful in ensuring that the solution does not violate the constraint, and at the same time, obtains a satisfying noise reduction performance. Since the optimal LFxLMS does not

allow the control filter to operate beyond the constraint, there is no need to include any system saturation model in the derivation. Furthermore, unlike other algorithms, no primary path is required in the implementation of the optimal LFxLMS algorithm.

As mentioned, there is still a drawback of the above optimal LFxLMS algorithm as several parameters can only be obtained from the control signal's statistical observation before its normal control operation. As the autocorrelation matrix of reference input, secondary path, disturbance power, and the secondary path may change during real-time control, the pre-derived leaky term may not be suitable during its real-time control process.

A more practical solution, which is based on inverse adaptive modeling [27], [28], can estimate the optimal LFxLMS for different types of noise distribution in practice. This latter method also does not require the assumption of the type of noise signal. The power gain of the secondary path, G_s can be approximated as the ratio of the variance of disturbance and its corresponding control signal.

$$G_s = \frac{\sigma_{\mathrm{d}}^2}{\sigma_{\hat{\mathrm{y}}_{\mathrm{d}}}^2}. \tag{16}$$

But to predict the control signal from the disturbance, we need to estimate the inverse model $c(n)$ of the secondary path, as shown in Fig. 2, and the predicted control signal (or anti-noise) is $\hat{y}_d = \mathbf{c}_{\mathrm{o}}^{\mathrm{T}}\mathbf{d}(n)$. From these equations, we can estimate the power gain across a frame of K samples

$$G_{\mathrm{s}} = \frac{\sum_{k=0}^{K} d^2(n-k)}{\sum_{k=0}^{K} \hat{y}_{\mathrm{d}}^2(n-k)}, \tag{17}$$

and the optimal leak term is given as

$$\gamma_{\mathrm{o}} = \frac{\sum_{k=0}^{K} d^2(n-k)}{\sum_{k=0}^{K} \hat{y}_{\mathrm{d}}^2(n-k)}(\eta - 1)\mathbf{R}_{\mathrm{x}}. \tag{18}$$

Our proposed algorithm OLFxLMS [27] can be applied to different ANC applications and handling different types of noise, with only a slight increase in computational complexity compared to other algorithms.

A recent paper [29] introduces the modified FxLMS algorithm to continuously estimate the disturbance power and the power of the secondary path, and thus, a time variable penalty factor can be derived iteratively. This latter approach expands the ability of the constraint-based leaky FxLMS algorithm to handle dynamic noise sources in real-time, and still be able to constraint within its maximum power output.

IV. Output Constraint: gradient project based FxLMS Algorithms

In this class of output constraint FxLMS algorithm, modifying the direction of the error gradient, or gradient projection, does not alter the cost function. The search is usually taken along or below the constraint contour. For example, the rescaling FxLMS algorithm [30], which rescales the weight and output signal of the control filter, is a form of gradient

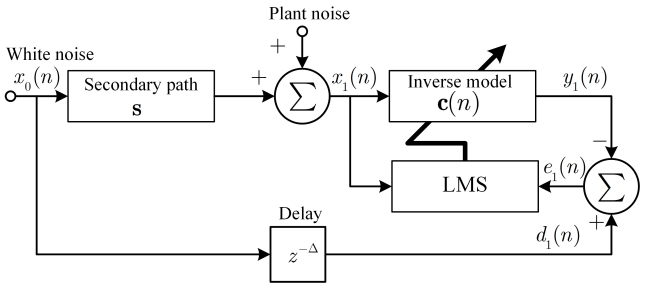


Fig. 2. Block diagram of adaptive inverse modeling for the secondary path.

TABLE I
Pseudocode of 2GD-FxLMS Algorithm

Two gradient FxLMS algorithm(2GD-FxLMS)
Input: Reference signal $x(n)$ and error signal $e(n)$ **Output**: Clipped output signal $y_{\text{out}}(n)$
Step 1: $y(n) = \mathbf{w}^{\mathrm{T}}(n)\mathbf{x}(n)$; $\mathbf{x}'(n) = \hat{s}^{\mathrm{T}}(n)\mathbf{x}(n)$. **Step 2**: If $\lvert y(n)\rvert \le \rho$ $\mathbf{w}(n+1) = \mathbf{w}(n) + \mu_1 e(n)\mathbf{x}'(n)$; else $\mathbf{w}(n+1) = \mathbf{w}(n) - \mu_2 y(n)\mathbf{x}(n)$. **Step 3**: If $y(n) > \rho$, $y_{\text{out}} = \rho$; else if $y(n) < -\rho$, $y_{\text{out}} = -\rho$; else $y_{\text{out}} = y(n)$.

projection algorithm. However, this algorithm increases computational load and can lead to a reduction in sampling frequency for practical implementation. To combine the modification of search direction and imposing a constraint, we can derive an optimal solution, which is referred to as the output constraint two-gradient descent (2GD) FxLMS algorithm [20], [31]. Several variants of the 2GD FxLMS take into account faster convergence, minimum mean square error, and extension to multiple channels. A pseudocode of the baseline 2GD-FxLMS algorithm is listed in Table I.

When the average output power of the control filter is within the constraint ($\mathbb{E}\left[y^2(n)\right] \le \rho^2$), the weight update is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \frac{1}{2}\mu\frac{\nabla J(\mathbf{w})}{\|\nabla J(\mathbf{w})\|}. \tag{19}$$

In contrast, when the average output power exceeds the constraint ($g(\mathbf{w}) > \rho^2$), the weight update changes to

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \frac{1}{2}\mu\frac{\nabla g(\mathbf{w})}{\|\nabla g(\mathbf{w})\|}. \tag{20}$$

By replacing the normalized gradient and average power constraint with the instantaneous gradient and amplitude constraint, the update equation can now be expressed as

$$\begin{cases} \mathbf{w}(n+1) = \mathbf{w}(n) + \mu_1 e(n)\mathbf{x}'(n), & \mathbb{E}\left\{y(n)^2\right\} \le \rho^2 \\ \mathbf{w}(n+1) = \mathbf{w}(n) - \mu_2 y(n)\mathbf{x}(n), & \mathbb{E}\left\{y(n)^2\right\} > \rho^2 \end{cases} \tag{21}$$

The main advantage of the 2GD-FxLMS algorithm is its ability to continuously update even when the output of the control filter exceeds the amplifier threshold; therefore, it possesses

TABLE II
COMPARATIVE EVALUATION OF THE DIFFERENT TYPES OF CONSTRAINT-BASED FXLMS ALGORITHMS

Algorithms:	**2GD-FxLMS**	**Optimal Leaky FxLMS**
Advantages	· Light computational load · A simple operator to switch between modes · Smaller residual error · Perform better in broadband noise · Enhance system stability and quality of anti-noise. · Able to implement a light-weight multi-channel 2GD-FxLMS.	· An optimal leaky FxLMS can avoid output saturation distortion and provides good noise reduction. · Online estimation of optimal leak factor that can handle different types of noise.
Disadvantages	· Noise reduction performance is usually inferior to other constraint-based algorithms. · May result in slower convergence, but can be solved using momentum and variable step size.	· Higher computational cost may not be desired for MCANC. Can use MOV-FxLMS [29] to reduce complexity. · More steps in determining the optimal leak term, but can be done online.
Usages	In practical real-time implementation of MCANC for open aperture and spatial ANC for controlling high-amplitude acoustic noise sources.	

the switching mechanism, and its computational complexity is the same as the iterative FxLMS algorithm. To put it simply, the 2GD-FxLMS forces the amplifier and speaker to operate in their linear region. Several real-time implementations of the 2GD FxLMS algorithms have been tested in air ducts.

There are several recent improvements in the 2GD FxLMS algorithm, namely:

(i) **Momentum 2GD-FxLMS** algorithm [32] with variable step size. We introduce a momentum factor to improve the convergence speed of the baseline 2GD-FxLMS algorithm and added a computation-effective variable step size approach to further reduce the steady-state error due to the switching nature of the 2GD-FxLMS.

(ii) **Multi-channel 2GD-FxLMS (or 2GD-MCFxLMS)** [33] is an extension of the single-channel 2GD-FxLMS algorithm. However, with the increase in the number of multiple output amplifiers, the output saturation nonlinearity also increases in the multiple-channel compared to the single-channel 2GD-FxLMS algorithm. Additionally, having multiple channel ANC (MCANC) applications means more effort is required to maintain the system components. Constantly subjecting the system to high-amplitude anti-noise reproduction can shorten the lifespan of the amplifiers/speakers and increase maintenance costs.

V. CONCLUSIONS

Table II provides a comprehensive comparison of various constraint-based FxLMS algorithms. These algorithms effectively address the issues of actuator overdriving and nonlinearity caused by saturation. Both the 2GD-FxLMS and optimal Leaky FxLMS algorithms employ a similar constraint-based approach to derive the adaptive FxLMS algorithm.

Typically, imposing a strong constraint on the FxLMS algorithm directs more effort towards achieving the desired output power. However, this approach compromises the level of noise reduction. In other words, constraint-based FxLMS algorithms are tailored to specific requirements and struggle to perform optimally in meeting both objectives simultaneously. Therefore, this paper proposes deriving an optimal leak factor that strikes a balance between noise reduction and satisfying power constraints. We provide a summary of two existing approaches, namely the 2GD-FxLMS and optimal leaky FxLMS algorithms, highlighting their unique properties, advantages, disadvantages, and practical applications.

The focus on real-time implementation in this study is of great value to practitioners, as it enables them to incorporate the optimal leak factor and 2GD into their adaptive ANC (FxLMS) systems. This applied research approach effectively bridges the gap between academic studies and practical implementations, facilitating more efficient noise control in real-world applications.

REFERENCES

[1] Sen M Kuo and Dennis R Morgan, *Active noise control systems*, vol. 4, Wiley, New York, 1996.

[2] Stephen J Elliott and Philip Arthur Nelson, "Active noise control," *IEEE signal processing magazine*, vol. 10, no. 4, pp. 12–35, 1993.

[3] Yoshinobu Kajikawa, Woon-Seng Gan, and Sen M Kuo, "Recent advances on active noise control: open issues and innovative applications," *APSIPA Transactions on Signal and Information Processing*, vol. 1, 2012.

[4] Colin N Hansen, *Understanding active noise cancellation*, CRC Press, 1999.

[5] Feiran Yang, Yin Cao, Ming Wu, Felix Albu, and Jun Yang, "Frequency-domain filtered-x lms algorithms for active noise control: a review and new insights," *Applied Sciences*, vol. 8, no. 11, pp. 2313, 2018.

[6] Jihui Zhang, Huiyuan Sun, Prasanga N Samarasinghe, and Thushara D Abhayapala, "Active noise control over multiple regions: Performance analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8409–8413.

[7] Cheng-Yuan Chang, Xiu-Wei Liu, Sen M Kuo, et al., "Active noise control for centrifugal and axial fans," *Noise Control Engineering Journal*, vol. 68, no. 6, pp. 490–500, 2020.

[8] Xiaoyi Shen, Woon-Seng Gan, and Dongyuan Shi, "Multi-channel wireless hybrid active noise control with fixed-adaptive control selection," *Journal of Sound and Vibration*, vol. 541, pp. 117300, 2022.

[9] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, Rina Hasegawa, and Yoshinobu Kajikawa, "Feedforward multichannel virtual-sensing active control of noise through an aperture: Analysis on causality and sensor-actuator constraints," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 32–48, 2020.

[10] Bhan Lam, Dongyuan Shi, Woon-Seng Gan, Stephen J Elliott, and Masaharu Nishimura, "Active control of broadband sound through the open aperture of a full-sized domestic window," *Scientific reports*, vol. 10, no. 1, pp. 1–7, 2020.

[11] Jordan Cheer and Stephen J. Elliott, "Multichannel control systems for the attenuation of interior road noise in vehicles," *Mechanical Systems and Signal Processing*, vol. 60-61, pp. 753–769, 2015.

[12] Yoshinobu Kajikawa, Woon-Seng Gan, and Sen Maw Kuo, "Recent applications and challenges on active noise control," in *2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2013, pp. 661–666.

[13] Xiaoyi Shen, Dongyuan Shi, Woon-Seng Gan, and Santi Peksi, "Adaptive-gain algorithm on the fixed filters applied for active noise control headphone," *Mechanical Systems and Signal Processing*, vol. 169, pp. 108641, 2022.

[14] Junwei Ji, Dongyuan Shi, Zhengding Luo, Xiaoyi Shen, and Woon-Seng Gan, "A practical distributed active noise control algorithm overcoming communication restrictions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] Sen M Kuo, Hsien-Tsai Wu, Fu-Kun Chen, and Madhu R Gunnala, "Saturation effects in active noise control systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 6, pp. 1163–1171, 2004.

[16] Dongyuan Shi, Chuang Shi, and Woon-Seng Gan, "Effect of the audio amplifier's distortion on feedforward active noise control," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 469–473.

[17] Li-Zhe Tan and Jean Jiang, "Filtered-x second-order volterra adaptive algorithms," *Electronics letters*, vol. 33, no. 8, pp. 671–672, 1997.

[18] Xiaojun Qiu and Colin H Hansen, "A study of time-domain fxlms algorithms with control output constraint," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2815–2823, 2001.

[19] Orlando José Tobias and Rui Seara, "Leaky-fxlms algorithm: Stochastic analysis for gaussian data and secondary path modeling error," *IEEE Transactions on speech and audio processing*, vol. 13, no. 6, pp. 1217–1230, 2005.

[20] DongYuan Shi, Woon-Seng Gan, Bhan Lam, and Chuang Shi, "Two-gradient direction fxlms: An adaptive active noise control algorithm with output constraint," *Mechanical Systems and Signal Processing*, vol. 116, pp. 651–667, 2019.

[21] Sen M Kuo and Hsien-Tsai Wu, "Nonlinear adaptive bilinear filters for active noise control systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 3, pp. 617–624, 2005.

[22] Lifu Wu, Xiaojun Qiu, and Yecai Guo, "A generalized leaky fxlms algorithm for tuning the waterbed effect of feedback active noise control systems," *Mechanical Systems and Signal Processing*, vol. 106, pp. 13–23, 2018.

[23] Dongyuan Shi, Bhan Lam, Woon-Seng Gan, and Shulin Wen, "Optimal leak factor selection for the output-constrained leaky filtered-input least mean square algorithm," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 670–674, 2019.

[24] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, and Xiaoyi Shen, "Comb-partitioned frequency-domain constraint adaptive algorithm for active noise control," *Signal Processing*, vol. 188, pp. 108222, 2021.

[25] Farzin Taringoo, Javad Poshtan, and Mohammad Hossein Kahaei, "Analysis of effort constraint algorithm in active noise control systems," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–9, 2006.

[26] Brian Beavis and Ian Dobbs, *Optimisation and stability theory for economic analysis*, Cambridge university press, 1990.

[27] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, and Xiaoyi Shen, "Optimal penalty factor for the mov-fxlms algorithm in active noise control system," *IEEE Signal Processing Letters*, vol. 29, pp. 85–89, 2021.

[28] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, and Xiaoyi Shen, "A frequency-domain output-constrained active noise control algorithm based on an intuitive circulant convolutional penalty factor," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1318–1332, 2023.

[29] Chung Kwan Lai, Dongyuan Shi, Bhan Lam, and Woon-Seng Gan, "Mov-modified-fxlms algorithm with variable penalty factor in a practical power output constrained active control system," *IEEE Signal Processing Letters*, 2023.

[30] Hui Lan, Ming Zhang, and Wee Ser, "A weight-constrained fxlms algorithm for feedforward active noise control systems," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 1–4, 2002.

[31] Dongyuan Shi, Woon-Seng Gan, Bhan Lam, and Shulin Wen, "Practical consideration and implementation for avoiding saturation of large amplitude active noise control," *Proc. 23rd Int. Congr. Acoust*, pp. 6905–6912, 2019.

[32] Xiaoyi Shen, Dongyuan Shi, Zhengding Luo, Junwei Ji, and Woon-Seng Gan, "A momentum two-gradient direction algorithm with variable step size applied to solve practical output constraint issue for active noise control," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[33] Dongyuan Shi, Bhan Lam, Xiaoyi Shen, and Woon-Seng Gan, "Multi-channel two-gradient direction filtered reference least mean square algorithm for output-constrained multichannel active noise control," *Signal Processing*, p. 108938, 2023.

A QoS Throughput Performance Measurement Comparison between UGS and BE Services of a Real-time FPGA Based OFDM Multi-user System Design Implementation

Trio Adiono
Department of Electrical Engineering
Bandung Institute of Technology
Bandung, Indonesia
tadiono@stei.itb.ac.id

Michael Jonathan
Department of Electrical Engineering
Bandung Institute of Technology
Bandung, Indonesia
m.jonathan9999@gmail.com

Erwin Setiawan
Department of Electrical Engineering
Bandung Institute of Technology
Bandung, Indonesia
erwin.setiawan789@gmail.com

Nana Sutisna
Department of Electrical Engineering
Bandung Institute of Technology
Bandung, Indonesia
nana.sutisna@gmail.com

Rahmat Mulyawan
Department of Electrical Engineering
Bandung Institute of Technology
Bandung, Indonesia
rahmat.mulyawan@gmail.com

Infall Syafalni
Department of Electrical Engineering
Bandung Institute of Technology
Bandung, Indonesia
infall@staff.stei.itb.ac.id

***Abstract*— It is defined in IEEE 802.16e that there are five different Quality of Service (QoS) classes, which include: UGS, rtPS, nrtPS, ertPS, and BE. This paper compares the throughput performance measurement results between the UGS and BE scheduling services for every modulation and coding scheme (MCS), which have been implemented in our FPGA-based real-time OFDM multi-user system. There are two FPGA-based RF transceiver boards involved in the one-slot experiment and three FPGA-based RF transceiver boards involved in the two-slot experiment. The downlink transmission is done in FDD, using 241 MHz as the center frequency, and the uplink transmission is done in TDD, using 261 MHz as the center frequency. The transmission scenario used in this experiment is point-to-multipoint (P2MP). The throughput measurement is taken under LOS conditions using cables and quarter-wave antenna connections. We observed that the BE service has higher downlink and uplink throughput compared to the UGS service by utilizing one uplink slot. However, there are almost 50% drops in the throughput results for every MCS in the BE implementation when two uplink slots are utilized. There are some limitations to this experiment, involving the number of subscriber stations, a connection timeout issue, and the experimental condition. Therefore, it is advised to increase the number of subscriber stations and test the system under several wireless conditions for future development.**

Keywords*— *QoS, UGS, BE, OFDM, Multi-user, P2MP, Throughput

I. Introduction

The IEEE 802.16 standard, commonly known as Worldwide Interoperability for Microwave Access (WiMAX), is a recent wireless broadband standard that has promised high bandwidth over long-range transmission [1]. It is defined in IEEE 802.16e that in order to meet different Quality of Service (QoS) requirements, such as throughput, packet error rate, jitter, system availability, and delay, there are five different scheduling mechanism categories to allocate the downlink and uplink transmission opportunities for different PDUs [2]. Those categories are: unsolicited grant service (UGS), real-time polling service (rtPS), non-real-time polling service (nrtPS), extended real-time polling service (ErtPS), and best effort service (BE).

The summary of the applications for each category can be observed in Table I.

TABLE I
THE SUMMARY OF QOS CATEGORIES APPLICATIONS [1]

QoS Category	Applications	QoS Specifications
UGS Unsolicited grant service	VoIP	Maximum sustained rate, maximum latency tolerance, and jitter tolerance.
rtPS Real-time polling service	Streaming audio or video	Maximum reserved rate, maximum sustained rate, maximum latency tolerance, and traffic priority.
ErtPS Extended real-time polling service	Voice with activity detection (VoIP)	Minimum reserved rate, maximum sustained rate, maximum latency tolerance, jitter tolerance, and traffic priority.
nrtPS Non-real-time polling service	File Transfer Protocol (FTP)	Minimum reserved rate, maximum sustained rate, and traffic priority.
BE Best Effort service	Data transfer, Web browsing, etc.	Maximum sustained rate and traffic priority.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

These five service classes can be divided into two main categories: non-real-time and real-time. UGS, rtPS, and ErtPS are categorized in the real-time category. The nrtPS and BE are categorized in the non-real-time category. The real-time services have strict constraints on delay and delay jitter. This makes scheduling more difficult to meet the delay constraint and tolerate the delay jitter with optimal throughput [3]. On the other hand, the non-real-time category has no strict constraints on these parameters. Therefore, common scheduling schemes can be applied, such as weighted fair queue (WFQ) and a variation of the round-robin (RR) algorithm [3].

In this paper, we present throughput measurement results of the UGS and BE scheduling services of a real-time orthogonal frequency division multiplexing (OFDM) system that is developed from our previous work [4] and implemented in the AD9361 SDR-FPGA platform [5]. The throughput results are used to obtain a performance measurement comparison between the implementations of each scheduler for every MCS in our system.

This paper is organized as follows. Section II presents an overview of the designed OFDM system. Section III describes the experimental setup, including hardware configuration and specifications. Section IV discusses measurement results and the achieved performance of each implemented service. The conclusions of this work are summarized in Section V.

II. System Design

A. The Design of The OFDM Multi-user System

The OFDM multi-user system is composed of the hardware, kernel space, and user space layers. The hardware layer consists of the Ethernet (PS) and the PHY transceiver (FPGA). The kernel space layer includes drivers (the Ethernet driver and the OFDM driver), the MAC layer (the OFDM MAC), and the net device. TCP/IP, and system calls. The user space is responsible for IP table forwarding. The overall system design is depicted in Fig. 1.

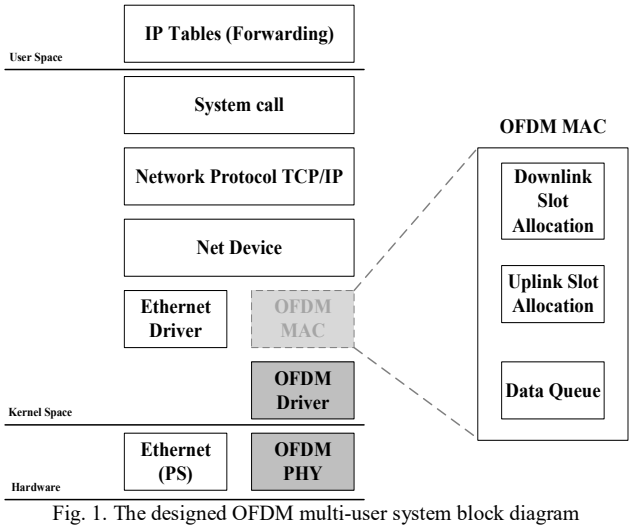


Fig. 1. The designed OFDM multi-user system block diagram

B. MAC Header Frame Structure

The QoS scheduler is implemented in the OFDM MAC, which is located in the kernel space layer. The OFDM MAC is also responsible for slot allocation, TDMA synchronization, and the data queue. Fig. 2 describes the frame structure of the MAC header. The MAC header is arranged by the frame control, the source address, and the destination address. The total length of the MAC header is 13 bytes.

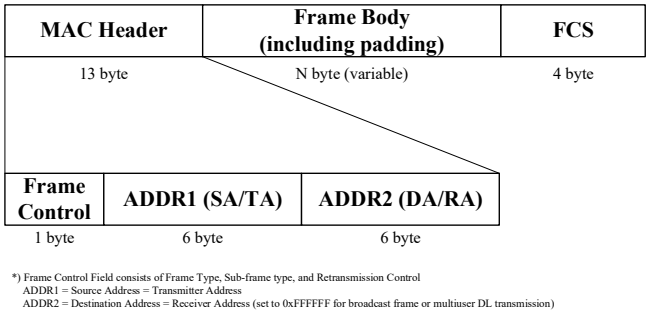


Fig. 2. MAC header frame structure

Furthermore, Table II presents the detailed specification of the frame control of the MAC header. The total size of the frame control is 1 byte and consists of the frame type (from bit-0 up to bit-1), the sub-frame type (from bit-2 up to bit-5), and the retransmission control (bit-6).

TABLE II
THE MAC HEADER FRAME CONTROL

Frame Type		Sub Type		Retry	
Bit-1 … Bit-0		Bit-5 … Bit-2		Bit-6	
00	Management	0001	Beacon	0	Normal
		0010	Association request	1	Retry
		0011	Association response		
		0100	Authentication request		
		0101	Authentication response		
		0110	RTS (Request-to-Send)		
		0111	CTS (Clear-to-Send)		
		1000	Slot reservation		
		1001	Slot confirmation		
		1010	Slot ranging		
01	Control Frame	0001	ACK (Acknowledge)		
10	Data Frame	0001	Multi-user Data Frame		
		0010	Broadcast Data Frame		

C. TDMA of The UGS Scheduler

The TDMA of the UGS scheduler is shown in Fig. 3. Each uplink slot has a fixed delay value in order to transmit the uplink frames.

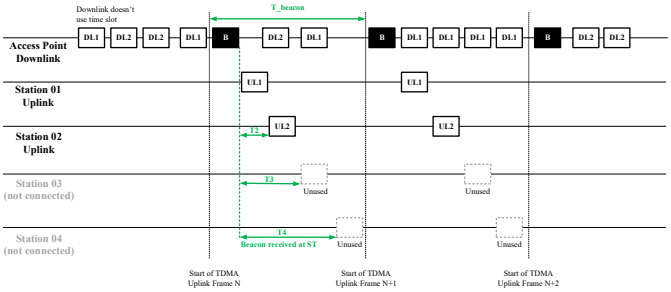


Fig. 3. The TDMA of the UGS scheduler

The delay value for each uplink timeslot is determined by the uplink slot number and the frame duration of each MCS. The delay value for each slot is calculated by (1), where T(n) is the

delay of the uplink slot, n is the uplink slot number (starts from zero), and T_{frame} is the frame duration for each MCS.

$$T(n) = n * T_{frame} \quad (1)$$

The frame duration of each MCS is obtained from the total of the number of symbols, the signal, the preamble, and the guard band, which can be calculated using (2), times the sum of symbol and cyclic prefix duration, which is calculated by (3).

$$N_{frame} = N_{sym} + N_{signal} + N_{preamble} + N_{guard} \quad (2)$$

$$T_{frame} = N_{frame} * (T_{symbol} + T_{CP}) \quad (3)$$

It is known that for 3.5 MHz bandwidth, the sum of the symbol and cyclic prefix durations is equal to 80 µs. A 0.5 ms guard duration is added to the final calculation of the frame duration to ensure every frame is fully transmitted. Therefore, the frame duration for each MCS in the 3.5 MHz bandwidth can be observed in Table III.

TABLE III
THE FRAME DURATION FOR EVERY MCS IN 3.5MHZ BANDWIDTH

BW [MHz]	MCS	N_frame	T_symbol + T_CP [µs]	Frame Duration [ms]
3.5	0	151	80	12.58
	1	75	80	6.50
	2	40	80	3.70
	3	51	80	4.58
	4	28	80	2.74

The uplink timeslot delay value for every MCS using the 3.5 MHz bandwidth is presented in Table IV.

TABLE IV
THE UPLINK TIMESLOT DELAY VALUE FOR EVERY MCS IN 3.5MHZ BANDWIDTH

BW [MHz]	MCS	Frame Duration [ms]	Delay T0 [ms]	Delay T1 [ms]	Delay T2 [ms]	Delay T3 [ms]
3.5	0	12.58	0	12.58	25.16	37.74
	1	6.50	0	6.50	13.00	19.50
	2	3.70	0	3.70	7.40	11.10
	3	4.58	0	4.58	9.16	13.74
	4	2.74	0	2.74	5.48	8.22

D. TDMA of The BE Scheduler

The TDMA of the BE scheduler is shown in Fig. 4. It can be observed that, if there is no other station connected, the only station connected is able to utilize more uplink slots for the uplink frame transmission.

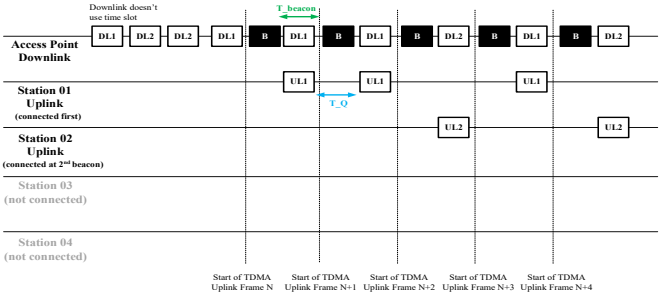


Fig. 4. The TDMA of the BE scheduler

However, if there is a new station connected, the transmission of the uplink frame of each connected station will be scheduled by the round-robin algorithm. The round-robin algorithm scheme can be observed in Fig. 5.

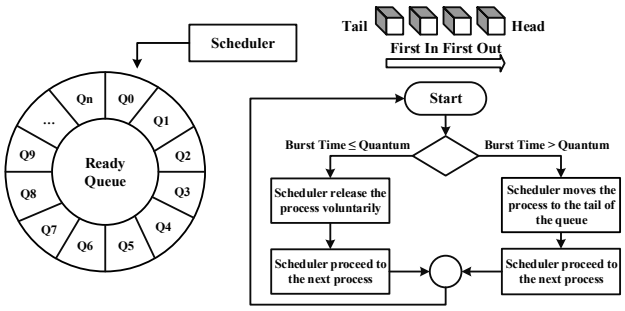


Fig. 5. The round-robin algorithm scheme

It is known that the round-robin algorithm is generally used for a timesharing system. This algorithm is similar to the first come, first served (FCFS) algorithm; however, there is a preemption in order to do the switching of the processes. A time unit, called a time quantum or time slice, is defined to determine the interval between processes. Every process will be inserted into a ready queue, which can be visualized as a circular queue.

Let S_0, S_1 … S_n be n-stations having A_i, B_i, and r_i where i=0,1,2 … n, as their arrival time, burst time, and number of rounds required for completion, respectively. Let Q_0, Q_1 … Q_n be n-station slots in the ready queue. Therefore, we use (4) to arrange every station into the station slots in an FCFS manner.

$$Q_{A_n} = S_n \quad (4)$$

Furthermore, let q denote a fixed-time-quantum in time units. The number of rounds required for completion, r_i, can be calculated using (5) [6].

$$r_i = \begin{cases} B_i \text{ div } q, if\ B_i \bmod q = 0 \\ (B_i \text{ div } q) + 1, otherwise. \end{cases} \quad (5)$$

The symbols 'div' and 'mod' denote modulus and division operators, respectively. Both B and q are assumed to be integers. Every process will be executed for one complete time quantum in each round except for the last round, where it may be less than or equal to one time quantum. The time requirement for a station slot Q_i in its last round, denoted by q_i, is the remainder of B_i /q.

The remaining burst time is updated after every round. The remaining burst time and the executed portion are denoted respectively by $B_{i,j}$ and $q_{i,j}$ for the i^{th} slot in j^{th} round. Therefore, the remaining time requirement of a station slot Q_i in r^{th} round can be given by (6) and the q_{r-1} can be calculated by (7) [6].

$$B_{i,r} = B_{i,r-1} - q_{i,r-1} \quad (6)$$

$$q_{r-1} = \begin{cases} q_i, if\ B_{i,r-1} < q \\ q, if\ B_{i,r-1} \geq q \end{cases} \quad (7)$$

The output of the round-robin algorithm is a Gantt chart consisting of the scheduled station slots. The quantum time, which is used in the scheduler implementation, is 1 time unit.

III. Performance Measurement Setup

In order to measure the throughput performance of the implemented schedulers in our real-time OFDM multi-user system, several variables, such as the AD9361 parameter configuration and the experiment setup, need to be configured and defined.

A. AD9361 Configuration

The AD9361 configuration includes the bandwidth, sampling frequency, downlink center frequency, uplink center frequency, and gain control configuration. We employ *iperf* software [7] in the AD9361 software driver to calculate the throughput. The AD9361 configuration is provided in Table V.

TABLE V
THE AD9361 PARAMETERS CONFIGURATION

Parameter	Value
Bandwidth [MHz]	3.5
Sampling Frequency [Msps]	20
Downlink Center Frequency [MHz]	241
Uplink Center Frequency [MHz]	261
TX Gain [dB]	-10
RX Gain Control	Fast Attack

B. Experiment Setup for One-slot Performance Measurement

The one-slot performance measurement is taken in ideal conditions using two U.FL-to-U.FL cables. It involves two transceiver boards: one behaves as an access point (AP), and the other behaves as a subscriber station (ST-01). The maximum number of slots connected is set to 4, interconnected with a point-to-multipoint (P2MP) topology. The one-slot performance measurement diagram and experiment setup can be observed in Figs. 6 and 7, respectively.

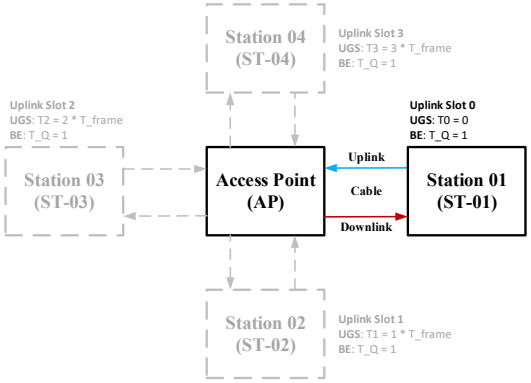


Fig. 6. The one-slot performance measurement setup diagram

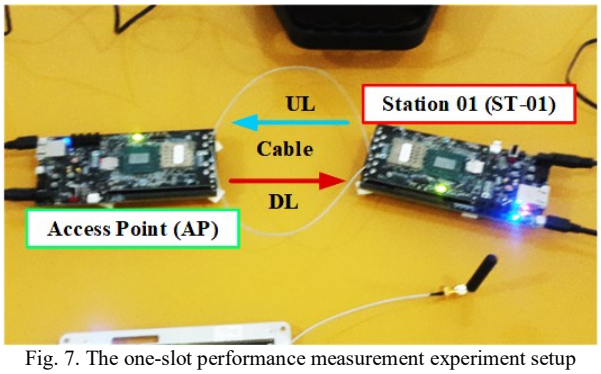


Fig. 7. The one-slot performance measurement experiment setup

C. Experiment Setup for Two-slot Performance Measurement

The two-slot performance measurement is taken in an indoor line-of-sight (LOS) condition using six quarter-wave antennas. It involves three transceiver boards: one behaves as an access point (AP), and the others behave as subscriber stations (ST-01 and ST-02). The maximum number of slots connected is also set to 4, interconnected with point-to-multipoint (P2MP) topology. The two-slot performance measurement diagram and experiment setup can be observed in Figs. 8 and 9, respectively.

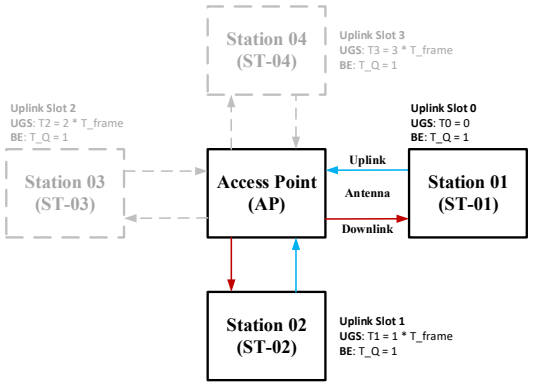


Fig. 8. The two-slot performance measurement setup diagram

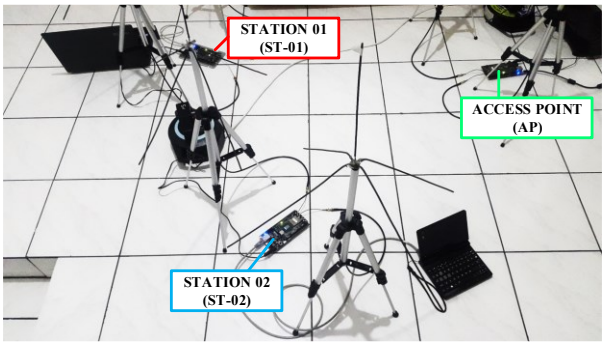


Fig. 9. The two-slot performance measurement experiment setup

IV. Performance Measurement Result

A. One-slot UGS Throughput Measurement Result

The throughput measurement results for the UGS implementation involving only one subscriber station connected to the access point and utilizing only one uplink slot are presented in Table VI.

TABLE VI
ONE-SLOT UGS THROUGHPUT MEASUREMENT RESULT

MCS		Throughput [Mbps]	
		Downlink	Uplink
0	BPSK FEC 1/2	Timeout	Timeout
1	QPSK FEC 1/2	0.83	0.48
2	16-QAM FEC 1/2	1.52	0.86
3	QPSK FEC 3/4	1.27	0.67
4	16-QAM FEC 3/4	2.22	1.17

There is a connection timeout issue for the MCS 0 (BPSK FEC ½) which causes failure in the throughput measurement process. It can be observed that the highest achievable downlink and uplink throughput by implementing the UGS in the one-slot experiment, respectively, is 2.22 Mbps and 1.17 Mbps by employing MCS 4: 16-QAM FEC ¾.

B. *One-slot BE Throughput Measurement Result*

The throughput measurement result for the BE implementation by involving only one subscriber station connected to the access point and utilizing only one uplink slot is presented in Table VII.

TABLE VII
ONE-SLOT BE THROUGHPUT MEASUREMENT RESULT

MCS		Throughput [Mbps]	
		Downlink	Uplink
0	BPSK FEC 1/2	0.64	0.53
1	QPSK FEC 1/2	1.48	1.16
2	16-QAM FEC 1/2	2.01	1.45
3	QPSK FEC 3/4	1.87	1.32
4	16-QAM FEC 3/4	3.04	1.87

Unlike the UGS service, there is no connection timeout issue in any MCS in the BE service implementation. It can be observed that the highest throughput achieved by the implementation of the BE service in the one-slot experiment is respectively 3.04 Mbps and 1.87 Mbps by employing MCS 4: 16-QAM FEC ¾.

C. *One-slot Throughput Comparison between UGS and BE*

The throughput comparison between the UGS and the BE for the one-slot experiment for every MCS can be visualized in Fig. 10.

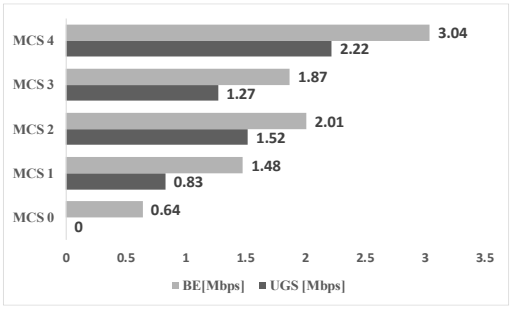


Fig. 10. The downlink throughput results comparison between the UGS and BE for every MCS from the one-slot experiment

As shown in Fig. 10, for every MCS, the throughput result of the BE implementation surpasses the throughput result of the UGS implementation. The range of differences in the throughput results between the UGS and the BE service for each MCS varies from 0.48 Mbps up to 0.82 Mbps.

The uplink throughput comparison between the UGS and the BE for the one-slot experiment for every MCS can be visualized in Fig. 11.

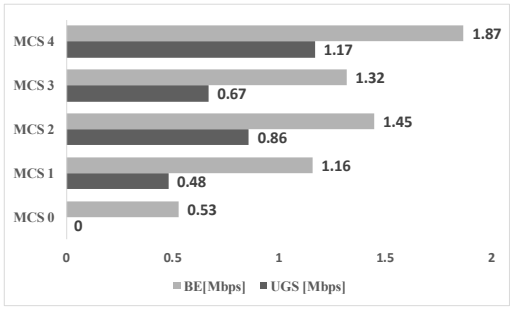


Fig. 11. The uplink throughput results comparison between the UGS and BE for every MCS from the one-slot experiment

As shown in Fig. 11, for every MCS, the throughput result of the BE implementation surpasses the throughput result of the UGS implementation. The range of differences in the uplink throughput results between the UGS and the BE service for each MCS varies from 0.53 Mbps up to 0.70 Mbps.

D. *Two-slot UGS Throughput Measurement Result*

The throughput measurement results for the UGS service implementation by involving two subscriber stations connected to the access point, utilizing two uplink slots, is presented in Table VIII.

TABLE VIII
TWO-SLOT UGS THROUGHPUT MEASUREMENT RESULT

MCS		Throughput [Mbps]					
		ST-01		ST-02		Average	
		DL	UL	DL	UL	DL	UL
0	BPSK FEC 1/2	Timeout					
1	QPSK FEC 1/2	0.81	0.51	0.69	0.48	0.75	0.50
2	16-QAM FEC 1/2	1.52	0.86	1.29	0.85	1.41	0.86
3	QPSK FEC 3/4	1.15	0.67	1.11	0.56	1.13	0.62
4	16-QAM FEC 3/4	1.83	1.17	1.75	1.18	1.79	1.18

Based on Table VIII, the highest average downlink and uplink throughput achieved by each station by employing the UGS service is 1.79 Mbps and 1.18 Mbps, respectively, using MCS 4: 16-QAM FEC ¾. The timeout issue still persists and causes failure to the throughput performance measurement of the MCS 0: BPSK FEC ½.

E. *Two-slot BE Throughput Measurement Result*

Table IX presents the throughput measurement results for the BE service implementation by involving two subscriber stations connected to the access point, utilizing two uplink slots.

TABLE IX
TWO-SLOT BE THROUGHPUT MEASUREMENT RESULT

MCS		Throughput [Mbps]					
		ST-01		ST-02		Average	
		DL	UL	DL	UL	DL	UL
0	BPSK FEC 1/2	0.27	0.27	0.27	0.27	0.27	0.27
1	QPSK FEC 1/2	0.88	0.39	0.74	0.51	0.81	0.45
2	16-QAM FEC 1/2	1.58	0.97	1.39	0.97	1.49	0.97
3	QPSK FEC 3/4	1.08	0.79	1.07	0.79	1.08	0.79
4	16-QAM FEC 3/4	1.89	1.02	1.48	1.25	1.69	1.14

By observing Table IX, the highest average downlink and uplink throughput achieved by each station by employing the BE service is 1.69 Mbps and 1.14 Mbps, respectively, using MCS 4: 16-QAM FEC ¾.

F. *Two-slot Throuhgput Comparison between UGS and BE*

For every MCS in the two-slot experiment, the average throughput result achieved by each subscriber station by implementing the UGS service is almost the same as the average throughput result achieved by each subscriber station by implementing the BE service. There are slight differences in the average throughput results between the UGS and the BE service for each MCS, which ranged from 0.05 Mbps up to 0.27 Mbps for the downlink and from 0.04 Mbps up to 0.27 Mbps for the uplink. Fig. 12 shows the average downlink throughput comparison between the UGS and the BE implementation for every MCS in the two-slot experiment.

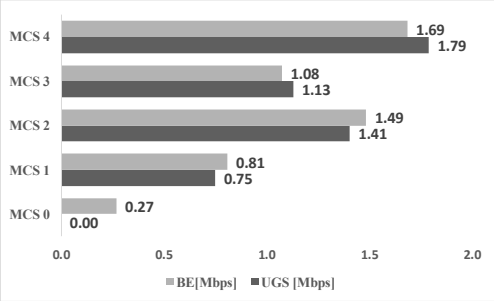


Fig. 12. The average downlink throughput results comparison from each station between the UGS and BE for every MCS from the two-slot experiment

Fig. 13 shows the average uplink throughput comparison between the UGS and the BE implementation for every MCS in the two-slot experiment.

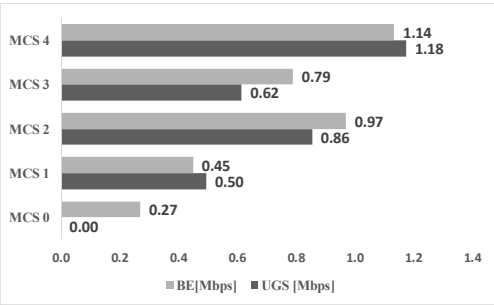


Fig. 13. The average uplink throughput results comparison from each station between the UGS and BE for every MCS from the two-slot experiment

G. *One-slot and Two-slot Throughput Comparison*

Fig. 14 visualizes the comparison of the downlink throughput results achieved by each station from both one-slot and two-slot experiments for every MCS by implementing UGS and BE services.

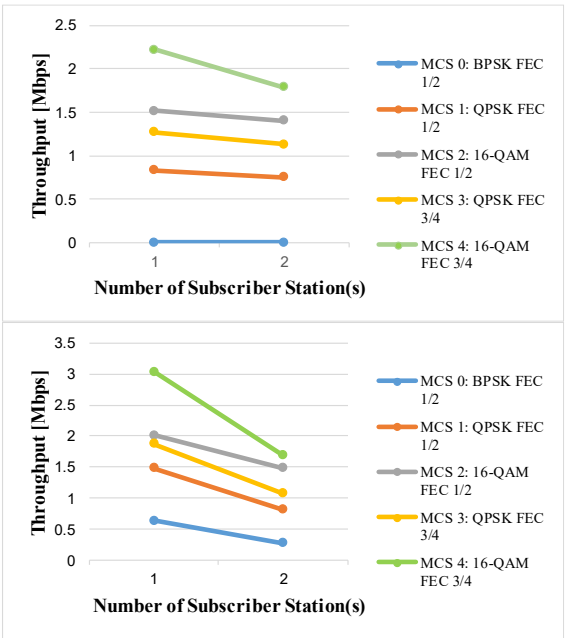


Fig. 14. The downlink throughput results achieved by each station from both one-slot and two-slot experiments for every MCS by implementing UGS (above) and BE (below) services

Fig. 15 visualizes the comparison of the uplink throughput results achieved by each station from both one-slot and two-slot experiments for every MCS by implementing UGS and BE services.

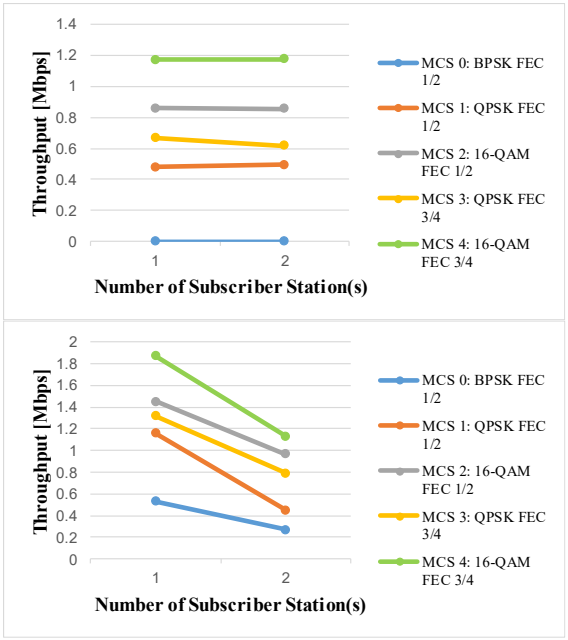


Fig. 15. The uplink throughput results achieved by each station from both one-slot and two-slot experiments for every MCS by implementing UGS (above) and BE (below) services

V. Conclusion

In this paper, we have presented a throughput performance measurement between two QoS implementations (UGS and BE) of a real-time OFDM multi-user system. The performance measurement is aimed at comparing the throughput result for every MCS with the implementation of the UGS and BE schedulers, utilizing one uplink slot and two uplink slots. Based on the real-time throughput evaluation, the throughput of the BE implementation is higher than the throughput of the UGS implementation for every MCS when only one uplink slot is utilized. However, when two uplink slots are utilized, there are almost 50% drops in both downlink and uplink throughput of the BE implementation. On the other hand, the throughput of the UGS remains relatively stable when two uplink slots are utilized.

There are some limitations to this experiment. One of them is the number of subscriber stations involved in this experiment, which is less than enough to show the reliability of each scheduler. We also found that in the UGS implementation, there is a connection timeout issue for the MCS 0: BPSK FEC ½. This issue is not found in the BE implementation. Therefore, it is advised to do further investigation in order to solve the timeout issue of MCS 0 in the UGS implementation. The other limitation is that the condition of the experiment is still under ideal conditions. Therefore, for future development, it is advised to increase the number of subscriber stations in order to gain insight into the schedulers' reliability and to test the system under several wireless conditions in order to show how environmental non-idealities can affect the transmission.

REFERENCES

[1] B. Li, Y. Qin, C. P. Low, C. L. Gwee, "A Survey on Mobile WiMAX," *IEEE Commun. Mag.,* pp. 70-75, December 2007.

[2] A. K. M. Abdelhadi, H. E. Idris, A. A. Mohamed, "A QoS Analysis of UGS and ERTPS Scheduling Flows in Video Confrence over Worldwide Interoperability for Microwave Access (Wi-MAX)," *IJEAT*, pp. 145-150, April 2017.

[3] C. So-In, R. Jain, A. Al-Tamimi, "A Scheduler for Unsolicited Grant Service (UGS) in IEEE 802.16e Mobile WiMAX Networks," *IEEE Systems Journal*, Vol. 4, No. 4, pp. 487-494, December 2010.

[4] T. Adiono, et al. "Performance Measurement of Real-time FPGA based OFDM System Implementation," presented at the 8th ICEEI Conference, Malaysia, 2021.

[5] Analog Devices, *RF Agile Transceiver: AD9361 Datasheet*. Norwood, USA: One Technology Way, 2013, pp. 1.

[6] D. Pandey, Vandana, "Improved Round Robin Policy A Mathematical Apporach," *IJCSE*, Vol. 2, No. 4, pp. 948-954, 2010.

[7] Jon Dugan, Seth Elliott, Bruce A. Mah, Jeff Poskanzer, and Kaustubh Prabhu. "iPerf," [Online]. Available: https://iperf.fr.

Algorithm Development for Stepwise Valve Deflation Method in Blood Pressure Measurement

Trio Adiono*, Reina Puteri Ramadhani*, Clarance Amadeus*, Sindy Novaria Cicilya Sinaga◦
*School of Electrical Engineering and Informatics, Bandung Institute of Technology
◦PT. Xirka Dama Persada, Indonesia.

***Abstract*— This paper aims the development a Pump-based Leakless NIBP (Non-Invasive Blood Pressure) system using the oscillometric principle and a stepwise valve deflation method. The stepwise deflation method involves a controlled and gradual reduction of pressure inside the arm cuff, which enables the utilization of a single controlled valve. Evaluation results show error rates between 2.05 and 5.6, accuracy ranging from 86.0% to 99.1%, and low standard deviations (1.57 to 3.91) for error, and (0.84% to 9.78%) for accuracy. Most presets meet the ±5 error threshold, but further optimization is needed for preset 5 to align with Ministry of Health regulations, eliminating the need for multiple valves and air leakage.**

Keywords*—NIBP, Blood pressure measurement, Stepwise deflation, oscillometric algorithm, valve controlled.*

1. Introduction

Creating a portable, accurate, and practical blood pressure monitor is very important because blood pressure is one of the most vital metrics of healthiness [1]. An accurate and practical blood pressure monitor empowers individuals to measure their blood pressure independently, without relying on the medical staff, facilitating easier and faster monitoring.

Currently, numerous portable blood pressure monitors are available in medical stores, known for their affordability, user-friendliness, and compliance with medical standards [1][12]. Although, the key to their effective performance of the right algorithm. This algorithm includes how the device reads the raw pressure data and processes it into systolic and diastolic blood pressure values. Despite their significance, many of these algorithms are not publicly disclosed and remain an open research problem.

In this paper, we aim to introduce an algorithm for blood pressure measurement, utilizing the stepwise deflation method in conjunction with an oscillometric-based approach. Stepwise deflation is a method of acquiring raw pressure data [3]. Oscillometric based algorithm is a well-known algorithm to process raw pressure data. Nevertheless, there is no universally optimal set of parameters for this algorithm, as its performance can vary across different conditions. Therefore, the objective of this paper is to provide a comprehensive explanation of all the algorithm's parameters employed for blood pressure measurement.

2. Previous Work

In this section, some of the previous works will be explained. This section will be divided into 2 parts, each to address the previous works in the cuff deflation algorithm and the previous works in the oscillometric algorithm.

2.1. Cuff Deflation Algorithm

Basically, there are 2 different algorithms for cuff deflation, which are linear deflation and stepwise deflation. In the linear deflation algorithm, the cuff is pumped to a pressure higher than the systolic pressure of the patient and deflated at a constant rate. The pressure of the cuff is monitored until the pressure in the cuff is fully deflated. The resulting raw data is a linear declining cuff pressure waveform with an oscillometric waveform superimposed in it. In the stepwise deflation algorithm, the cuff is pumped in the same way as the linear deflation algorithm but is held and deflated for a certain period. The pressure of the cuff is monitored every time the cuff pressure is held. The resulting raw data is a stepwise declining cuff pressure waveform with an oscillometric waveform superimposed on each level. Stepwise deflation method needs an electronically controlled valve to control the deflation of the cuff. The deflation process usually takes 10-12 steps [1]. The pressure drop in each step can differ, but often is about 4-6 mmHg [2]. In each step, one or more pulses are detected before the cuff is deflated into the next step.

Typically, the commercially available blood pressure monitor device uses stepwise deflation [3]. The stepwise deflation is preferred because it is less susceptible to noise. However, this method increases the measurement time, and this method requires strict timing control to control the deflation time. To reduce the measurement time, an adaptive timing control algorithm has been proposed. In [4], the pressure difference in each deflation step is changed corresponding to the amplitude of the pulse found in each pressure level. In [5], the valve open time is altered in each deflation step corresponding to the amplitude of the pulse found in each pressure level.

2.2. Signal Processing Algorithm

The oscillometric algorithm is a widely used algorithm in blood measurement devices. It is done by separating the Oscillometric Waveform (OMW) from the cuff pressure waveform, then processing it to obtain Mean Arterial Pressure (MAP), Systolic Blood Pressure (SBP), and Diastolic Blood

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Pressure (DBP). To process the raw data and calculate blood pressure, many methods have been proposed. These methods can be statistical-based, DSP-based, Neural Network Based, or empirical.

MAP, SBP, and DBP can be calculated directly from the oscillometric pulse shape [6]. This method can be an excellent one to study the oscillometric pulse, but it needs a good understanding of Digital Signal Processing (DSP) and is noise prone. The more common way to extract MAP, SBP, and DBP is by using a ratio. SBP and DBP can be found as fixed ratios from the MAP. To find these ratios, statistical methods such as Bayes [7] [8], AI methods such as Neural Networks [9] and Fuzzy [10], and empirical methods [11] [12] [13] have been proposed. These methods usually have trade-offs between accuracy, robustness, and complexity.

3. Method

3.1. Pump-based Leakless NIBP Measurement Hardware

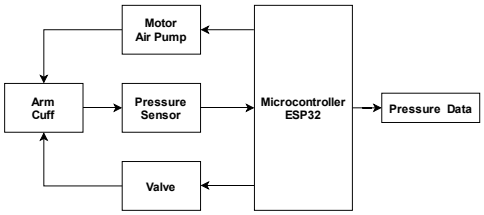


Fig. 1. Block Diagram of Hardware System

Figure 1 depicts the block diagram of the pump-based leakless NIBP system that we have implemented. In this method, we utilized an internal cuff in NIBP Simulator for the arm cuff. The Microcontroller ESP32 controls the motor air valve for cuff inflation and the valve for deflation at the desired time intervals. The cuff pressure is measured by the air pressure sensor. The pump-based leakless NIBP measurement eliminates the need for a precision dump valve for controlling the deflated airflow. Instead, this system employs a single controlled valve to manage the deflation of the airflow. Fig.2 illustrates the direction of airflow during inflation (indicated by the blue arrow) and deflation (indicated by the red arrow).

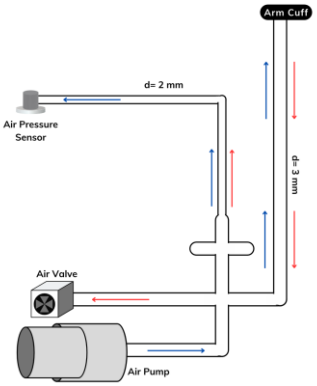


Fig. 2. Illustration of Airflow when inflated (blue) and deflated (red)

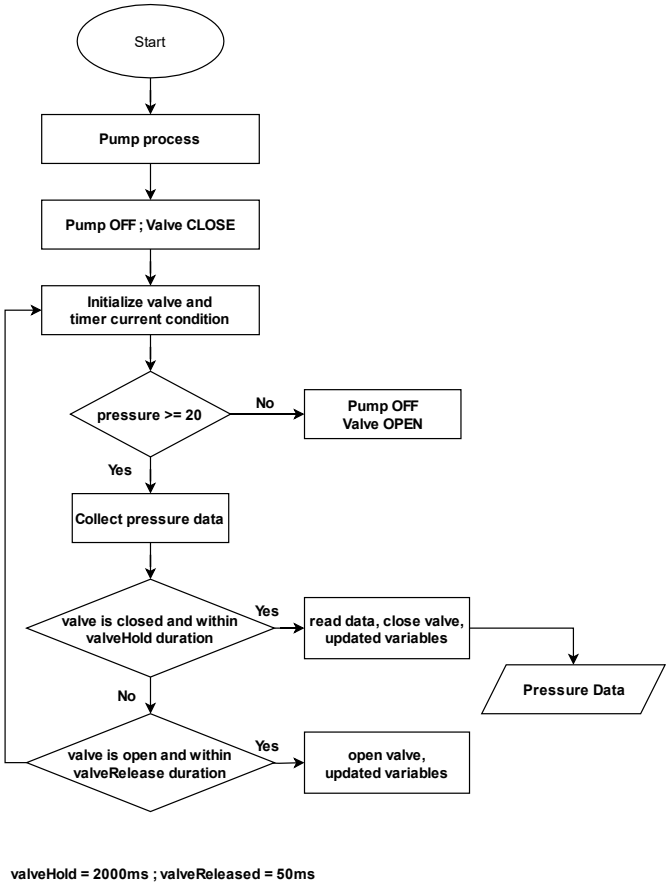


Fig. 3. Flowchart for stepwise deflation (*valve controlled*)

The pressure data acquisition process is performed in the digital domain within the microcontroller, where data is collected with valve control. After the pumping process is completed, the valve is controlled at pressures above 20 mmHg. During deflation, the valve remains closed for the valveHold duration and then opens for the valveRelease duration. Throughout the open-close cycle, pressure data will be captured only when the close cycle.

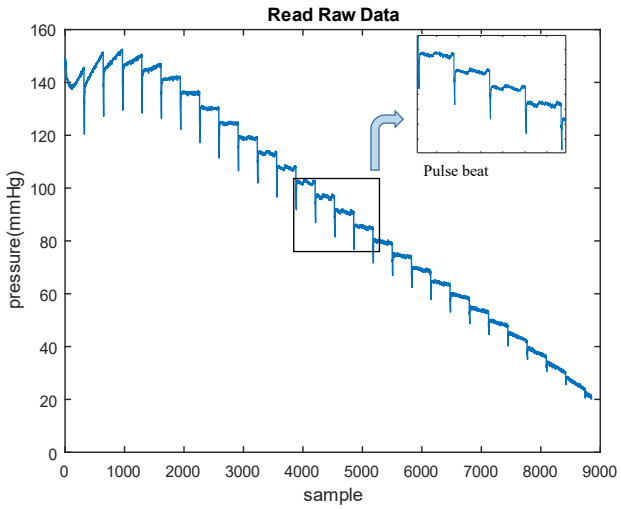


Fig. 4. Raw Pressure Data of step-wise deflation

In this step-wise deflation method, the control of the valve is very important. To ensure seamless recording of the pulse signal without interruptions, we have set the valve hold duration to 2000ms. Furthermore, we're considering establishing the lowest measurable heart rate at 30 bpm (0.5Hz or 2 seconds). The overarching goal is to detect one pulse beat for each step when the lowest bpm value is attained. For the valve release duration, we have chosen 50 ms, resulting in pressure exiting within the range of 4-6 mmHg. After the pressure falls below 20 mmHg, indicating the absence of a pulse beat, the valve will fully open until reaching 0 mmHg. The pressure data values will then be forwarded to the signal processing stage to estimate the values of MAP, SBP, and DBP.

3.2. Dataset Creation

We created and evaluated a dataset to test the accuracy of our algorithm. The dataset is distributed in accordance with the NIBP testing standards for bedside monitor evaluation, as specified by the Ministry of Health of the Republic of Indonesia in 2018, as outlined in Table 1. Data acquisition was conducted using the CONTEC MS200 NIBP Simulator to generate cuff pressure waveforms for specific BPM, SBP, and DBP values. We introduced variations by using all standard presets, each with BPM variations of 30, 60, 120, and 180, and this process was repeated 5 times for each setting, resulting in a total dataset of 140 data points.

The pressure threshold for data acquisition in presets 1, 5, 6, and 7 was established at 150 mmHg. As for presets 2, 3, and 4, the pressure threshold is determined by adding 30 to the systole score for each respective preset. The data acquisition results were stored in .txt format.

TABLE I. DATASET DISTRIBUTION

Setting Standard Variation	Blood Pressure (mmHg)		
	Systole	*Mean*	*Diastole*
Preset 1	120	93	80
Preset 2	150	116	100
Preset 3	200	166	150
Preset 4	255	215	195
Preset 5	60	40	30
Preset 6	80	60	50
Preset 7	100	76	65

3.3. Algorithm and Parameter Selection

The signal processing algorithm is implemented using MATLAB R2023a, benefiting from its versatile functions for signal graph plotting, facilitating signal analysis, and designing filters. The utilization of MATLAB's visualization and signal processing capabilities offers a robust foundation for the effective exploration and interpretation of NIBP data. With its powerful capabilities and various built-in tools for signal processing, MATLAB allows for efficient development and analysis of NIBP data. Fig. 5 illustrates the block diagram of the implemented algorithm.

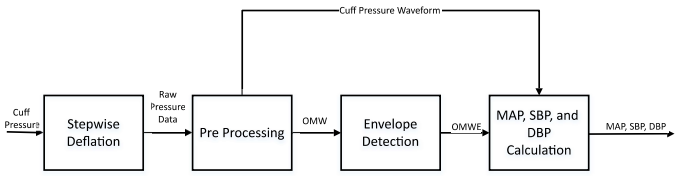


Fig. 5. Block Diagram of the Algorithm

- ***Pre-processing***

The primary goal of the algorithm is to detect the oscillometric signal to identify the values of MAP, SBP, and DBP. To achieve this, we utilize Low Pass Filter, High Pass Filter, and Moving Average filter. The LPF filter allows low-frequency components below its cut-off frequency to pass through, as shown in Fig. 6 (top), effectively eliminating environmental noise and producing a smoother cuff pressure signal. On the other hand, the HPF filter passes signals with frequencies higher than its cut-off frequency, as depicted in Fig. 6 (middle), facilitating the extraction of the oscillometric waveform. To further improve the signal quality, we apply a moving average filter, resulting in the smoothed signal displayed in Fig. 6 (bottom). For precise control over the filtering process, detailed parameters are utilized and listed in Table II.

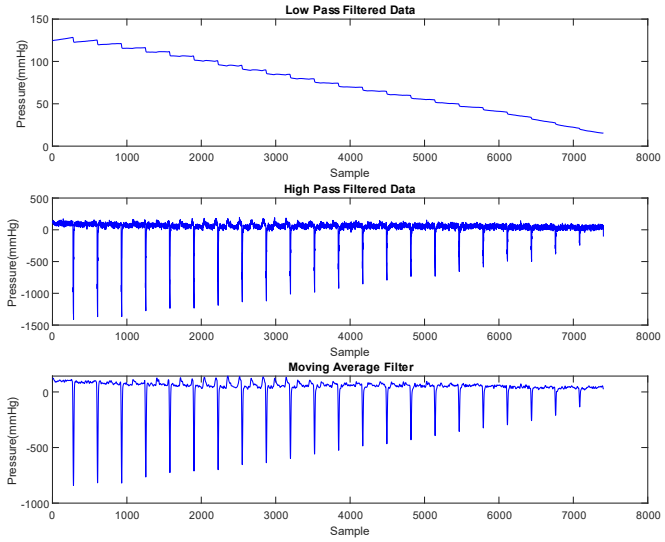


Fig. 6. Low Pass Filtered (top) , High Pass Filtered (middle), Moving Average Filter (bottom)

In the preprocessing stage, alongside filtering, we implement the process of signal amplification and offsetting data is utilized to increase the magnitude of the oscillometric waveform. Oscillometric waveforms may occasionally display small fluctuations or low amplitudes, especially in clinical scenarios involving weaker pulses. Moreover, this process enables the distinction between peak signals and noise,

contributing to consistent data processing and facilitating waveform comparisons across different signals. The application of signal amplification and offsetting assists in detecting and analyzing oscillometric waveforms, thereby simplifying the examination process.

- ***Envelope Detection***

After obtaining the oscillometric waveform, an upper envelope of the signal will be generated. This envelope is commonly referred to as the Oscillometric Waveform Envelope (OMWE). To obtain OMWE, an envelope detector algorithm is used. The algorithm takes the maximum amplitude of the oscillometric waveform in a certain frame length and adjusts the signal envelope to reach that maximum value. If the current envelope amplitude is lower than the maximum value, the envelope amplitude is incremented proportional to the error between the current envelope amplitude and maximum amplitude, and if the current envelope amplitude is higher than the maximum value, the envelope amplitude is decremented proportional to the error between current envelope amplitude and maximum amplitude. The presence of lower spikes in the HPF result is inconsequential as we focus on the upper envelope.

- ***MAP, SBP, and DBP Calculation***

The MAP, SBP, and DBP values are calculated based on the envelope values. The first step is to determine the MAP value from the maximum envelope value. In searching for the maximum MAP value, we limit it to ensure that the value does not reside at the beginning or at the end of the signal. Once the MAP value is obtained, the SBP and DBP values are determined by considering the ratios and control points in relation to data segments on both sides of the MAP value. Fig. 7 demonstrates the synchronization of the resulting indices on the envelope signal with the cuff pressure signal data obtained from the LPF.

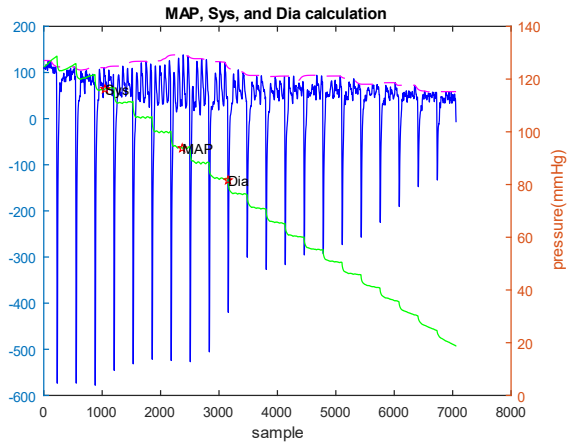


Fig. 7. MAP, Sys, and Dia Estimation

Several parameters have been selected for implementation in the algorithm. Detailed information regarding the values of each parameter used can be found in Table II.

TABLE II. ALGORITHM PARAMETER

Parameters	Value
Cut-off LPF	15 Hz
Cut-off HPF	0.3 Hz
Frame Moving Average	15 ; 10
Frame Length of Envelope	300
Ratio Sys	0.7
Ratio Dia	0.8

4. Result and Discussion

In the cuff pressure data acquisition process using the stepwise deflation method, we control the valve according to the flowchart shown in Fig.3, where the valve is closed during the valve hold period (2000ms) and opened during the valve release period (50ms). Performance graphs for each hold and release step are presented in Fig.10 and Fig.11, illustrating the time (milliseconds) taken for each operation. Table III provides valuable insights into the valve-controlled performance, showcasing the mean and standard deviation of errors and accuracy for each valve condition.

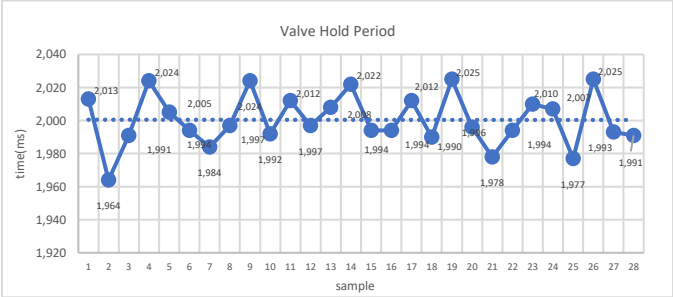


Fig. 8. Graph of Valve Hold Period in Measurement

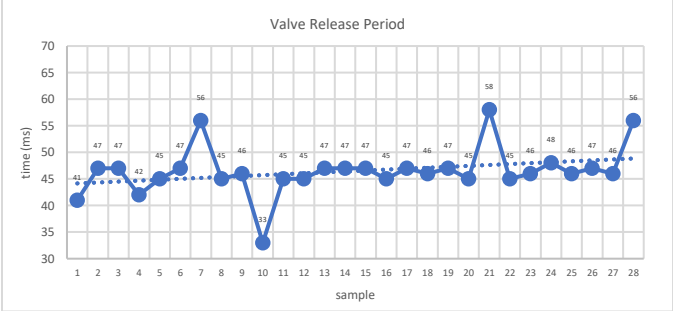


Fig. 9. Graph of Valve Release Period in Measurement

TABLE III. VALVE CONTROLLED PERFORMANCE

Valve Condition	Error (s)		Accuracy	
	Average	*St.Deviation*	*Average*	*St.Deviation*
Hold (Closed)	0.012893	0.008702	99.36%	0.44%
Release (Opened)	0.00493	0.002918	90.14%	5.84%

Notably, there is a significant difference in standard deviation values between the valve hold and valve release periods. The larger standard deviation during the valve release period suggests increased data variability, likely influenced by the instrument's ability to synchronize timing in very short intervals during the 50ms valve release.

The algorithm's performance in measuring MAP, SBP, and DBP values through cuff pressure data acquisition using the stepwise deflation method was assessed. We examined the results from the bpm-based approach (Table IV and Table V) and the preset-based approach (Table VI and Table VII) as listed in Table I. We observed the mean and standard deviation of errors and accuracy for each MAP, SBP, and DBP value.

TABLE IV. AVERAGE OF BPM-BASED PERFORMANCE

BPM	Error (mmHg)			Accuracy		
	Sys	Dia	MAP	Sys	Dia	MAP
30	3.91	3.20	3.11	96.83%	95.65%	96.03%
60	3.176	3.353	4.176	97.07%	95.06%	95.03%
120	3.94	4.29	4.83	95.75%	93.56%	94.10%
180	3.74	4.51	4.89	96.26%	92.97%	93.80%

TABLE V. STANDARD DEVIATION OF BPM-BASED PERFORMANCE

BPM	Error (mmHg)			Accuracy		
	Sys	Dia	MAP	Sys	Dia	MAP
30	3.27	2.31	2.22	2.31%	4.12%	4.02%
60	2.61	2.22	3.41	3.48%	4.69%	6.36%
120	2.97	2.52	2.99	4.63%	5.53%	5.32%
180	2.92	2.72	3.01	4.01%	7.57%	5.35%

TABLE VI. AVERAGE OF PRESETSS-BASED PERFORMANCE

Standard Setting Variation	Error (mmHg)			Accuracy		
	Sys	Dia	MAP	Sys	Dia	MAP
1	3.55	4.40	4.10	97.04%	94.50%	95.59%
2	4.80	3.70	4.25	96.80%	96.30%	96.34%
3	2.80	4.05	3.55	98.60%	97.30%	97.86%
4	2.30	3.95	4.85	99.10%	97.97%	97.74%
5	5.10	4.10	5.60	91.50%	86.33%	86.00%
6	4.25	4.50	4.40	94.69%	91.00%	92.67%
7	3.90	2.05	5.10	96.10%	96.85%	93.29%

TABLE VII. STANDARD DEVIATION OF PRESET-BASED PERFORMANCE

Standard Setting Variation	Error (mmHg)			Accuracy		
	Sys	Dia	MAP	Sys	Dia	MAP
1	2.98	2.95	3.74	2.48%	3.68%	4.02%
2	3.05	2.00	2.81	2.04%	2.00%	2.42%
3	2.67	2.61	3.50	1.33%	1.74%	2.11%
4	2.15	2.70	3.07	0.84%	1.39%	1.43%
5	3.65	2.57	3.91	6.09%	8.58%	9.78%
6	3.19	2.28	3.27	3.99%	4.57%	5.45%
7	2.47	1.57	2.77	2.47%	2.42%	3.64%

From the obtained measurements, we calculated the average values of errors and accuracy for both the bpm-based and preset-based approaches. The maximum error value was found to be 5.6, while the minimum was 2.05. As for accuracy, the maximum value reached 99.1%, and the minimum value was 86.0%. Additionally, we computed the standard deviation values for both error and accuracy in both approaches. The maximum standard deviation for error was 3.91, and the minimum was 1.57. For accuracy, the maximum standard deviation was 9.78%, and the minimum was 0.84%.

According to the NIBP testing standards for bedside monitor evaluation, as defined by the Ministry of Health of the Republic of Indonesia in 2018, a desirable error value is ±5 to avoid false diagnoses for individuals. Referring to Table VI, presets 1, 2, 3, 4, 6, and 7 exhibit an average error <5 for SBP, DBP, and MAP. Although, for preset 5, the SBP and MAP values have an error >5. As indicated in Table I, which presents the dataset distribution, preset 5 possesses the lowest values. This algorithm has not yet yielded optimal outcomes for presets with lower values.

The standard deviation values provide insights into the consistency of the algorithm's performance. The relatively low values of standard deviation for error (ranging from 1.57 to 3.91) and accuracy (ranging from 0.84% to 9.78%) indicate that the algorithm's estimations are relatively stable across multiple runs and scenarios. The lower the standard deviation, the more consistent the algorithm's performance.

5. Conclusion and Recommendations

This paper focuses on the cuff pressure data acquisition process using the stepwise deflation method. The selection of a 2000ms duration for the valve to remain closed is appropriate as it allows for accurate recording of pulse beats. Similarly, the 50ms duration for the valve to open is also suitable as it effectively reduces the pressure by 4-6 mmHg, enabling the recording of data without prolonging the process.

The signal processing algorithm has provided optimal results (±5 mmHg error) for presets 1, 2, 3, 4, 6, and 7 at each BPM. The relatively low values of standard deviation for error and accuracy indicated a consistent performance of the algorithm across multiple runs and scenarios. These findings emphasize the stability and reliability of the algorithm's estimations.

In the future, it is expected that there will be further development and optimization of the algorithm to yield favorable outcomes within a wider range of measurement values, particularly for low blood pressure data.

References

[1] Thomas and D. Rees, “Monitoring arterial blood pressure,” Anaesthesia amp;amp; Intensive Care Medicine, vol. 19, no. 4, p. 194a197, ˆ 2018J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] B. S. Alpert, D. Quinn, and D. Gallick, “Oscillometric blood pressure: a review for clinicians,” Journal of the American Society of Hypertension, vol. 8, no. 12, pp. 930–938, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1933171114007487R. Nicole.

[3] W. Van Moer and K. BarbA©, “Influence of the cuff deflation mode ˜ on oscillometric blood pressure measurements,” in 2011 IEEE International Symposium on Medical Measurements and Applications, 2011, pp. 652–656. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[4] H. Oh, Y. Chee, J. Lee, I. Kim, S. Kim, and Y. Kim, “The modified step-wise deflation method in blood pressure measurement,” in 2008 Computers in Cardiology, 2008, pp. 169–172.

[5] J. Booth and R. Medero, “Oscillometric blood pressure monitor employing deflation periods of alternating durations,” May 1996.

[6] M. Mafi, S. Rajan, M. Bolic, V. Z. Groza, and H. R. Dajani, “Blood pressure estimation using maximum slope of oscillometric pulses,” in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012, pp. 3239–3242.

[7] S. Lee, G. Jeon, and G. Lee, “On using maximum a posteriori probability based on a bayesian model for oscillometric blood pressure estimation,” Sensors (Basel, Switzerland), vol. 13, pp. 13 609–23, 10 2013.

[8] S. Lee, S. Rajan, H. R. Dajani, V. Z. Groza, and M. Bolic, “Determination of blood pressure using bayesian approach,” in 2011 IEEE International Instrumentation and Measurement Technology Conference, 2011, pp. 1–5.

[9] A. Aolak and C. Isik, “ Blood pressure estimation using neural networks,” 08 2004, pp.21-25.

[10] M. Forouzanfar, H. R. Dajani, V. Z. Groza, M. Bolic, and S. Rajan, “Adaptive neuro-fuzzy inference system for oscillometric blood pressure estimation,” in 2010 IEEE International Worksho p on Medical Measurements and Applications, 2010, pp. 125–129.

[11] R. Jaafar, H. M. Desa, Z. Mahmoodin, M. R. Abdullah, and Z. Zaharudin, “Noninvasive blood pressure (nibp) measurement by oscillometric principle,” in 2011 2nd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering, 2011, pp. 265–269.

[12] S. Kumar, S. Yadav, and A. Kumar, “Oscillometric waveform evaluation for blood pressure devices,” Biomedical Engineering Advances, vol. 4, p. 100046, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667099222000226

[13] P. Huang, J. Fu, and Y. Zhao, “A matlab demo software for nibp measurement based on hybrid oscillometric method,” in 2011 4th International Congress on Image and Signal Processing, vol. 5, 2011, pp. 2536–2539

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

SUMO Based Hardware/Software Co-simulation for Two-Intersection Adaptive and Collaborative Traffic Signal Controller

Kendrik Emkel Ginting*, Nana Sutisna†, Infall Syafalni‡, and Trio Adiono§
* Institut Teknologi Bandung, Indonesia
E-mail: 13218077@std.stei.itb.ac.id Tel/Fax: +62-22-2506291
† Institut Teknologi Bandung, Indonesia
E-mail: nana.sutisna@staff.itb.ac.id Tel/Fax: +62-22-2506291
‡ Institut Teknologi Bandung, Indonesia
E-mail: infall@ieee.org Tel/Fax: +62-22-2506291
§ Institut Teknologi Bandung, Indonesia
E-mail: tadiono@stei.itb.ac.id Tel/Fax: +62-22-2506291

***Abstract*—Due to safety reasons, the development of Reinforcement Learning (RL) based Traffic Signal Controller (TSC) can not be perform in real-world environment. A training and hardware/software co-simulation platform is needed to verify that the RL-based TSC works well before being deployed in real-world environment. In this paper we discuss the design and development of such platform using SUMO and python for 2-agent Q-Learning Algorithm. The platform receives Q-Matrix from an external device and then process the Q-Matrix to control the environment. Overall, the design works well. The platform could train the RL Agent and could affect the environment. Improvement could be made in minimizing the execution time of the simulation.**

I. Introduction

Traffic signals are designed to regulate traffic at an intersection. In Indonesia, many of the traffic signals are pre-timed [1]. The Department of Transportation first do a survey of traffic condition on a certain area and then create traffic signal plans based on the traffic. The plans are designed to be dynamically switched based on the time of the day. Some of the traffic signals are also designed to be coordinated with its adjacent intersections. The commonly used way to perform coordination is by setting green time offset between adjacent intersections which is called as "green wave" [2]. However, this method is not optimised for a highly fluctuating traffic and prolonged congested intersections [1], [3].

A more advanced way in traffic signal control that is deployed in big cities are Adaptive Traffic Control System (ATCS) which is a self-optimising control based on real-time vehicle data [1]. However, the ATCS are unable to handle oversaturated traffic well [3]. In addition, the abundance of traffic data and the very complex nature of traffic results in the need of a model-free traffic signal controller (TSC) [4], [5].

Reinforcement Learning (RL) is one of the methods that is being developed to solve the problems. RL does not require a model for complex traffic behavior, thus has a good outlook for future TSC [4]–[6]. RL-based TSC is, however, infeasible to be directly deployed in real-world due to safety reasons. Therefore, a platform to train RL-based TSC is needed. The platform must perform hardware/software cosimulation, that is verifying that hardware and software function correctly together [7].

In this paper, we discussed the development of a training platform for a Q-learning TSC. The platform provides Q-learning training environment that represents real-world condition of Dago-Cikapayang and Dago-Sulanjana intersections in Bandung. The traffic volume data is based on traffic counting at those intersection during peak hour. The TSC is also designed to comply with regulations of minimum duration of a traffic signal phase. Several simplification is, however, made, that is, the geometric of the intersection to become two adjacent 4-way intersection and the vehicle that is simulated is only car.

II. Related Works

A. Q-Learning

Reinforcement Learning is a branch of machine learning where the learning *agent* picks an action given a state with a goal to maximize its long-term reward. The agent at time t observes the current state of the environment, s_t, and takes an action, a_t. This in turn changes the environment to its next state, s_{t+1}, and gives the agent a reward, r_t. One of the learning and exploring methods for RL agent is Q-Learning Algorithm. The Bellman Equation, used for calculating the maximum future reward, is given in the equation below.

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(a_t, s_t) + \alpha\delta_t(s_t, a_t) \tag{1}$$

B. Simulator of Urban MObility (SUMO)

Environement simulation is needed for the training of RL agents. There are three modelling levels in traffic simulator: macroscopic, mesoscopic, and microscopic. Macroscopic traffic simulator simulates the aggregate behaviour of the traffic flow, while microscopic traffic simulator simulates the individual behaviour of the vehicle. On the other hand, mesoscopic

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

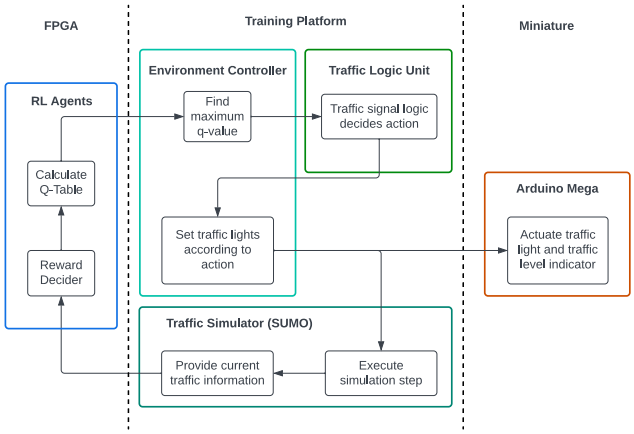


Fig. 1. Process Diagram for the RL Training Platform

model of traffic simulation is in the middle between microscopic and macroscopic [8]. The complexity of the simulation increased as the detail of the simulation increased.

In this paper, we use a microscopic traffic simulator to simulate the environment due to the required level of details. Some examples of microscopic traffic simulator are Vissim and SUMO. SUMO is chosen due to the open-source nature of the program, which requires no license, and due to the availability of its interface to Python called TraCI (Traffic Control Interface). TraCI provides us convenience in getting and providing data from and to the simulation.

C. Serial Communication

The training platform is designed to send the decided action to a two adjacent 4-way intersection miniature. The decided action is taken shown by the traffic light LEDs and the queue length is shown by LED strips on each approach of the intersection. The miniature is run by Arduino Mega. Communication between the training platform and Arduino Mega is required. Communication is established by serial communication due to its practicality both on Arduino and on a PC.

III. Design

The training platform consists of four main parts: Simulator, Environment Controller, Traffic Logic Controller, and Communication Module. The interaction between those parts is shown in the figure 1. RL agents, which are planted on an FPGA, produce q-tables. The q-tables are in the form of 3D array. Environment controller find the maximum q-value from the q-tables. The j and k index of the maximum q-values corresponds to the wished action and target level. However, this action and target level do not directly be actuated in SUMO or miniature, since the traffic signals needs to comply to other rules. The wished action and target level is then processed in a Traffic Logic Controller.

Traffic Logic Controller decides which action to take based on several conditions. The conditions are made such that the resulting action complies with regulations of traffic signals. The decided action, then, is sent back to Environment controller to set the traffic signals. The traffic signal states is then transferred both to Arduino Mega and SUMO Traffic Simulator. After displaying the current traffic signal state, SUMO will take the next step in the simulation and updates the current traffic conditions. The information then transmitted to FPGA to be considered for the reward of the action.

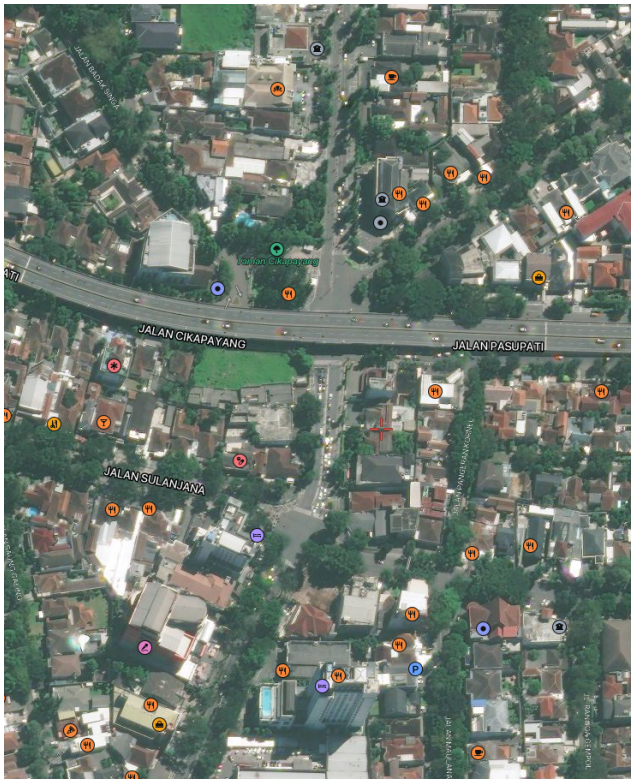


Fig. 2. Satellite imagery of Dago-Cikapayang and Dago-Sulanjana intersections.

A. Simulator Design

The simulator design involves intersection geometric design and traffic condition reproduction. Our aim is to design a simulation that reproduce peak hour traffic condition at Dago-Sulanjana and Dago-Cikapayang intersection.

1) Intersections Geometric Design: Figure 2 shows the satellite imagery of Dago-Cikapayang and Dago-Sulanjana Intersections which is the focus of this paper. The simplified version of the Intersections needs to be produced. Dago-Sulanjana intersection is kept as it is. It has 4 approaches. The western and eastern approaches of Dago-Sulanjana is a two-lane two-way undivided roads (2/2 UD). The northern and southern approach of the intersection is a four-lane two-way divided roads (4/2 D).

At the Dago-Cikapayang intersection, originally there are five approaches. The eastern, northern, western, and the southern approaches of Dago-Cikapayang intersection are available for both entering and exiting the intersection. The north-eastern approach is an exit-only approach to Jalan Dipatiukur. The

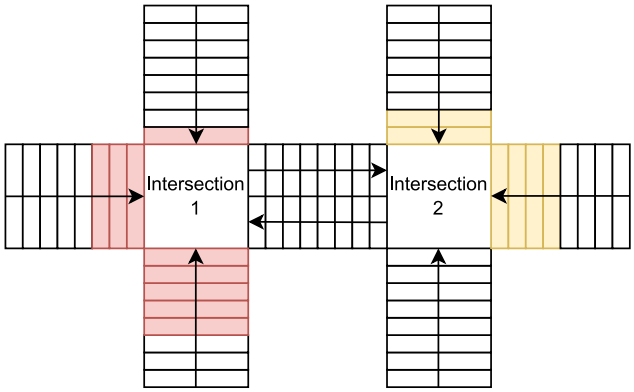


Fig. 3. Dago-Cikapayang and Dago-Sulanjana Intersection Model

TABLE I
NUMBER OF VEHICLES CROSSING DAGO-SULANJANA AND DAGO-CIKAPAYANG INTERSECTIONS (PASSENGER CAR EQUIVALENT)

Direction of Origin	Dago-Cikapayang	Dago-Sulanjana
Eastern Approach	784	648
Northern Approach	1548	1204
Western Approach	1116	420
Southern Approach	1780	1264

simulation will focus on the intersection and the link between Dago-Cikapayang and Dago-Sulanjana. Thus, vehicles exiting the approaches — except for the link — is not relevant for the simulation. On the other hand, running Multi-agent Q-Learning algorithm for two 4-way intersections is more feasible memory wise. Therefore, the Dago-Cikapayang intersection is simplified to a 4-way intersection. Each approaches of the intersection is a 4/2 D roads.

As Q-Learning algorithm requires discrete states, the queue length of each approach at both intersection must be divided into discrete levels. Considering the memory limitation of the Xilinx PYNQ FPGA that is used for RL Algorithm computation, the queue length on each approach is divided into 4 levels. The design of the intersection geometric is shown by figure 3.

2) Traffic Flow Design: To produce a more realistic traffic data we did a traffic counting on Dago-Cikapayang and Dago-Sulanjana intersection at peak hour. During the peak hour, bottleneck that causes queue in the middle of the intersection often happens. The simulation should be able to reproduce this event. The obtained data is shown in table I, table II, and table III. Table I shows the traffic counting result. The number of vehicles in the table is a result of Passenger Car Equivalent (PCE) conversion. The numbers of motorcycle, light vehicles, and heavy vehicles are all converted to PCE. The turning ratio in Dago-Cikapayang intersection and Dago-Sulanjana intersection are shown in table II and table III correspondingly. The data in the tables is recorded in an hour, thus in the simulation the data should be spread in an hour too.

TABLE II
TURNING RATIO OF DAGO-CIKAPAYANG INTERSECTION

Direction of Origin	Left-turn ratio	Right-turn ratio
Eastern Approach	0.133	0.623
Northern Approach	0.067	0.299
Western Approach	0.073	0.656
Southern Approach	0.162	0.309

TABLE III
TURNING RATION OF DAGO-SULANJANA INTERSECTION

Direction of Origin	Left-turn ratio	Right-turn ratio
Eastern Approach	0.32	0.336
Northern Approach	0.149	0.066
Western Approach	0.202	0.202
Southern Approach	0.123	0.164

B. Traffic Logic Unit Design

Traffic Logic Unit receives maximum q-value of the current state with its j and k index from the q-matrix 3D array. The j index corresponds to which action to take, while the k index corresponds to the target level of the lane to which an action is given. Traffic Logic Unit decides which action to take by considering the current state, thus q-value, and other considerations.

One of the considerations that should be taken into account by Traffic Logic Unit is regulations of the local government. In Indonesia, the government regulates the minimum duration of yellow signal of 3 seconds [9]. In addition to that, other knowledge or wisdoms could be added too. During our observation, we noticed that the shortest duration of green signal in Dago-Cikapayang and Dago-Sulanjana intersection is 20 seconds. We decided to retain the minimum green time of 20 seconds to prevent quick change of only several seconds of green. A quick change of traffic signal could poses danger to crossing pedestrians due to the size of the intersection. The Traffic Logic Unit's workflow is illustrated by the Finite State Machine (FSM) on the Figure 4 below.

C. Environment Controller Design

Environment controller has the job to influence the environment simulated by the training platform. Control of the environment is done by giving action based on the Q-Learning Algorithm. The action is giving green signal to a certain lane based on q-value and other rules. As figure 1 describes, Q-table is received by Environment Controller from FPGA. The first task that the environment controller must do is to find the maximum q-value at current traffic condition. After the maximum q-value is obtained, Environment Controller send the data to Traffic Logic Unit to be processed further.

The Environment Controller then received the final action that should be applied to the environment from the Traffic Logic Unit. The second task of the Environment Controller is to codify the actions to be applied to the environment in accordance to the agreed communication mechanism.

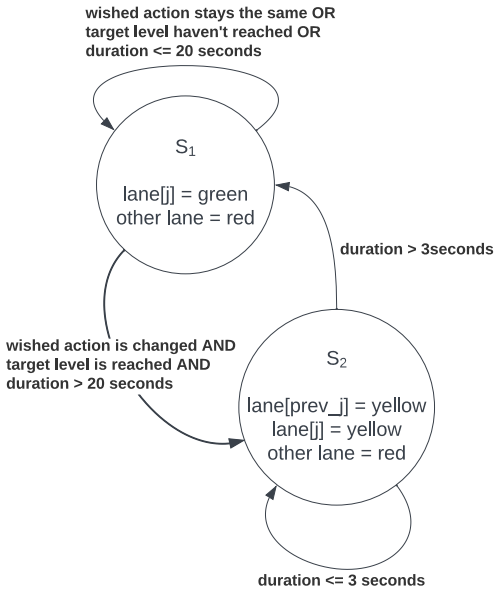


Fig. 4. Finite State Machine of RL-mode Traffic Logic Unit

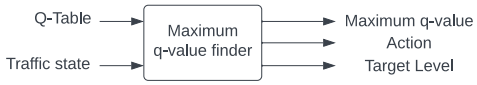


Fig. 5. Diagram of Maximum Q-Value Finder

1) Maximum Q-Value Finder: The Maximum Q-Value Finder should accomplishes the first task of the Environment Controller. It receives Q-Table from FPGA and current traffic state of the intersection as an input and produce maximum Q-value, action, and target level of the current state. It works by converting received Q-Table into 3D array. The function then find the row of the current state and find the maximum q-value on that row. It returns the maximum q-value itself, the index j of the q-value which corresponds to the action that should be done, and the index k of the q-value which corresponds to the target level that should be achieved by lane-j. The maximum Q-Value Finder function is illustrated in Figure 5.

2) Communication Module: A communication module is required for communication between FPGA, the training platform, and the miniature as illustrated in figure 1. In addition, a communication between Python and SUMO is also needed to be established. Miniature, which is run by Arduino Mega, and Training Platform, which is based on Python, communicates through serial connection. The library *pyserial* is used in Python. As mentioned in section II-B, the communication with SUMO is done through TraCI.

As the communication is established, the form of communicated message is needed to be agreed. TraCI has its own message format for traffic signals phase actuation. The Environment Controller needs to produce a message that comply to this format. For the serial communication with Arduino, a message format is also needed.

a) SUMO Phase State Encoder: SUMO's TraCI receives traffic light phase state in a string as a tuple of light definitions from rugGyYuoO, for red, red-yellow, green, yellow, off. In our Dago-Cikapayang and Dago-Sulanjana network, the phase state definition consists of 16 characters. The first four characters is controls the traffic signals of the eastern approach, the second four characters for the northern approach, the second last four characters for the western approach, and the last four character for the southern approach.

SUMO Phase State Encoder encode the actions that needs to be actuated to the TraCI string format. Based on the FSM of the Traffic Logic Unit in Figure 4, the state of the FSM, the action — index j of maximum q-value, which corresponds to the lane number that should be given green — and the previous action should be codified into TraCI string format. The state of the FSM is relevant to decide which action should be taken: giving green light to lane-j or giving yellow light to lane-j and lane-$previous_j$. The final string is later sent to SUMO through TraCI.

b) Arduino Phase State Encoder: Since a traffic intersection miniature is connected to the training platform, a communication mechanism between the training platform to Arduino Mega is needed. Serial communication is chosen because of its availability in Arduino Mega and its practicality. A message format is also needed in order to communicate well between devices.

One of the design consideration of the message format is the length of message. Serial communications send data bytes-by-bytes. Therefore, a message that is too long is undesirable since it could have a big delay. It also consumes more memory and costs more processing power. The size of a char that is transmitted with serial communication is 1 byte. The number of bytes will later affect the chosen baud rate of the serial communication. Due to the fact that a simulation of 1 step in SUMO corresponds to 1 second in the real-world, length of the transmitted message do not matter much. However, to make it more time efficient in testing, we would like to run the simulation with delay of each step of 0.1 second. With a typical baud rate of $9,600$ bits per second, the size of transmitted data needs to be at most 12 bytes, thus 12 characters.

According the FSM in Figure 4, there are 3 data that are needed to be transferred for a Traffic Light in an intersection: state of the FSM, current action, and the previous action. There are 2 states of the FSM, 4 possibilities of current action, and 3 possibilities of previous action. In summary, there are 24 states for a Traffic Light in an intersection. Since we have 2 intersection, we have 576 combined states.

We decided to convert the FSM's state, current action, and previous action of both intersection into integer. We then represent the integer in 3-digit hexadecimal and send the hexadecimal to Arduino Mega using serial communication. To convert those 6 elements to integer, the formula (2) is used. DC corresponds to Dago-Cikapayang and DS corresponds to Dago-Sulanjana.

$$msg_{int} = DC_{state} + 2 \cdot DS_{state} + 4 \cdot DC_{action} + 16 \cdot DS_{action} + 64 \cdot DC_{prev_action} + 256 \cdot DS_{prev_action} \quad (2)$$

Other than traffic light information, the transfer of current traffic level information of both Dago-Cikapayang and Dago-Sulanjana intersection is needed. Each approach of the intersection has 4 levels. This results in 256 states per intersection. We converted each intersections state into 2-digit hex and then append them to the 3-digit hex of traffic light states. In the end, we will have a 7-digit hex with the first three digit represent the traffic light states, the fourth and fifth digit represent Dago-Cikapayang traffic levels, and the last two digits represent Dago-Sulanjana traffic levels.

D. GUI Design

The training platform needs a GUI to show the simulation data in a more precise manner. The GUI will show the number of vehicles on each lanes and its level. Other than that, the GUI will also have the button to start the simulation.

IV. Design Implementation

In this part, the design stated in the previous chapter is implemented. The results are also shown to proof that the design is implemented correctly.

A. SUMO Implementation & Results

The microscopic traffic is simulated with SUMO. The design is being implemented by importing Dago-Cikapayang and Dago-Sulanjana intersections from OpenStreetMaps. The road network is then edited to conform to the design. Discrete levels on each of the approaches is implemented by placing a detector with the length of 100 meters. Each levels then equivalent of 25 meters queue. The traffic flow is implemented by creating random trips of vehicles and then sample the random trips based with turn counts data. The sampling results in $6,974$ trip routes with total counting of $8,760$. The number is very close to the sum of traffic counting data in Table I which is $8,764$. Figure 6 shows the final intersection geometric implemented in SUMO. When the simulator is run, the traffic congestion and bottleneck occurs. It shows that the simulation has been correctly implemented.

B. Traffic Logic Unit Implementation

Traffic Logic Unit is implemented with Python as smaller functions that is illustrated in Figure 7. Each of the functions is tested individually. The individual testing runs well. The testing of Traffic Logic Unit will be combined with the Environment Controller in the next section since the two is closely related.

C. Environment Controller Implementation & Results

The Environment Controller is implemented with Python too. The first function that is implemented is the maximum q-value finder. The maximum q-value finder is implemented as *q_matrix_2d_to_matrix* function. The testing is done by loading a CSV with 256 rows 16 columns. The columns represents 4 elements of array of 4. The Q-Table from the CSV is successfully loaded to 3D matrix. The result that is printed in the console is in accordance to the source CSV file.

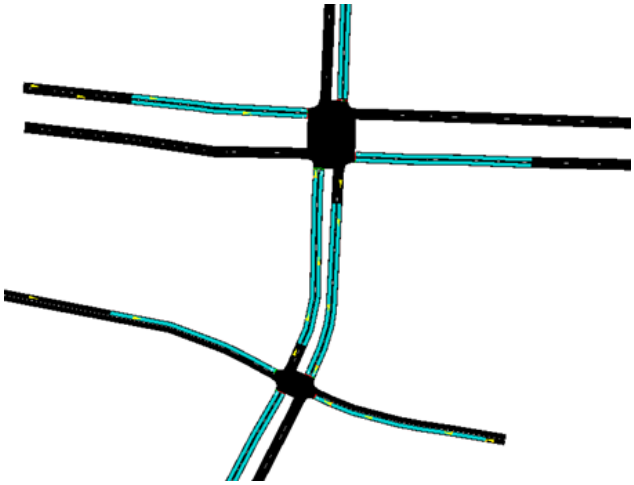


Fig. 6. Intersection geometry in SUMO

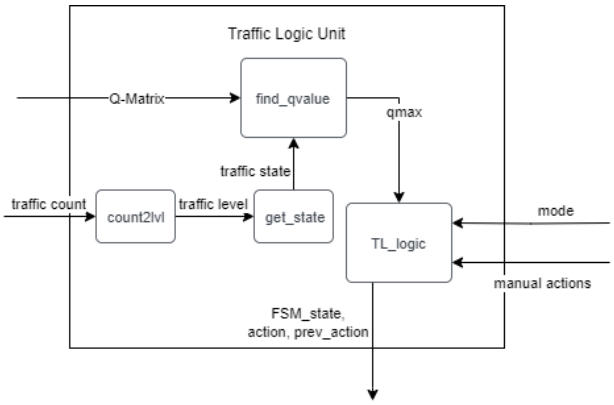


Fig. 7. Smaller function in Traffic Logic Unit Implementation

The second part of the Environment Controller that is implemented is the communication module. In accordance to the design, we implemented the SUMO Phase State Encoder and Arduino Phase State Encoder. The testing of Arduino Phase State Encoder is done by giving entering input of FSM state of Dago-Cikapayang (in the Table IV indicated by number 1), FSM state of Dago-Sulanjana (2), action of Dago-Cikapayang (3), action of Dago-Sulanjana (4), previous action of Dago-Cikapayang (5), previous action of Dago Sulanjana (6), traffic state of Dago-Cikapayang (7), and traffic state of Dago-Sulanjana (8). We send the message to Arduino and programmed the Arduino to resend the received message. The correct response would be, our console receives the message back. The result of this testing is shown in Table IV. The results in the table shows that the module has been implemented correctly since our console receive the exact message that we sent.

D. GUI Implementation & Results

The GUI is implemented using tkinter library on Python. The GUI shows global variables of vehicle counts and queue length in the form of discrete levels. Besides, there is also

TABLE IV
TESTING OF ARDUINO COMMUNICATION MODULE

Input								Output	
1	2	3	4	5	6	7	8	From Arduino	Ref
1	1	0	2	3	1	228	27	b'1e0e41b\r \n'	1e0e41b
2	2	3	0	1	1	217	82	b'14fd952\r \n'	14fd952
1	1	3	0	1	1	10	111	b'14c0a6f\r \n'	14c0a6f
2	2	1	2	3	0	106	84	b'0e76a54\r \n'	0e76a54
1	1	1	1	0	0	0	0	b'0140000\r \n'	0140000
2	2	3	3	2	2	255	255	b'2bffffff\r \n'	2bffffff

a button to start the simulation. As the simulation run, the Python script calls TraCI to updates the number of vehicles in each step. The number of vehicles then processed that results in levels. At the same time, GUI needs to show the data. Therefore, threading is used to allows part of programs to run concurrently.

The GUI is shown by Figure 9. The number of vehicles that is shown in the GUI is consistent with the number of vehicle that is in the simulation. The simulation is shown in Figure 10. From Figure 10, we could manually count that on the eastern approach of Dago-Cikapayang (the northern junction), there are 0 cars, 13 cars on the southern approach, and 28 cars in the detector area on the western approach. We can conclude that the GUI shows the data appropriately and has been implemented correctly.

V. ANALYSIS & DISCUSSION

In this paper, the execution time of the simulation is analysed and the match between SUMO and miniature traffic light phase and traffic state is verified.

A. Execution Time

Execution time of SUMO is measured between simulation without miniature and simulation with miniature. When the PC that runs simulation is connected to the miniature, communication between PC and miniature (Arduino Mega) is done. Table V shows the result of execution time in 10 different measurements with miniature. On the other hand, Table VI shows the result of execution time in 10 measurements without miniature In each measurement the simulation is set to have 3601 steps.

Based on the measurement, the average execution time is 574.987 seconds for total simulation which means an average duration of 159.674 milliseconds per step. Without miniature, the average execution time for 3601 steps is 460.323 seconds and 127.82 milliseconds per step. It is shown that there is a speedup of 0.8 when miniature is connected to the PC. With 95% confidence interval, the difference between simulation with- and without miniature is in the range of 58.663 seconds and 170.663 for 3601 simulations or between 25.84 milliseconds and 37.84 milliseconds for each step.

One of the predicted causes of this delay is the communication delay between PC and Arduino. Theoretically, for the transmission of 7 characters using serial, 7 bytes is needed.

TABLE V
MEASUREMENT OF EXECUTION TIME WITHOUT MINIATURE

Measurement Number	Duration (s)	Duration per step (s)	Step per second
1	552.436	153.412	6.518
2	575.949	159.941	6.252
3	577.042	160.244	6.240
4	578.972	160.781	6.219
5	576.653	160.137	6.244
6	575.841	159.911	6.253
7	579.044	160.801	6.219
8	577.409	160.347	6.236
9	578.351	160.609	6.226
10	578.172	160.559	6.228

TABLE VI
MEASUREMENT OF EXECUTION TIME WITH MINIATURE

Measurement Number	Duration (s)	Duration per step (s)	Step per second
1	480.863	133.536	7.489
2	458.438	127.309	7.855
3	456.613	126.802	7.886
4	457.938	127.170	7.863
5	455.050	126.367	7.913
6	460.809	127.967	7.815
7	457.657	127.092	7.868
8	460.323	127.832	7.822
9	463.118	128.608	7.776
10	452.417	125.636	7.959

With baud rate of 9600 bits per second, 5.6 milliseconds is needed. However, as the simulation is done in discrete steps, the duration of the simulation also play an important role in determining the communication duration.

Figure 8 shows the execution time for each step in simulation. It is shown that the execution time increases as the step increases and forms a logarithmic curve. This causes by the increase of simulation complexity as the simulation goes on. In the early stage of the simulation, the number of cars simulated is low, therefore the computational power needed to process the simulation is low. This results in faster execution time. As the step increases, more cars spawn in the simulation. Thus, more computation is needed. The simulation is a microscopic simulation, where each car is simulated uniquely in detail. Therefore the need of computational power is high.

B. SUMO and Miniature Traffic Light Phase and Traffic State Verification

As the miniature and the simulator GUI together display traffic light conditions, a verification should be done. Verification is done by comparing the traffic light phase in simulator and in Arduino. The simulation is set to have 3601 steps and is controlled by RL TSC. The verification results in 97.501% match between the traffic light phase and traffic state that is shown in the miniature and in SUMO.

The error is caused by delay between subsystems. Every change in traffic light phase or traffic state has a delay of 2 steps, e.g., if a change occurs at step 38 in SUMO, the change will be presented in miniature in step 40. However,

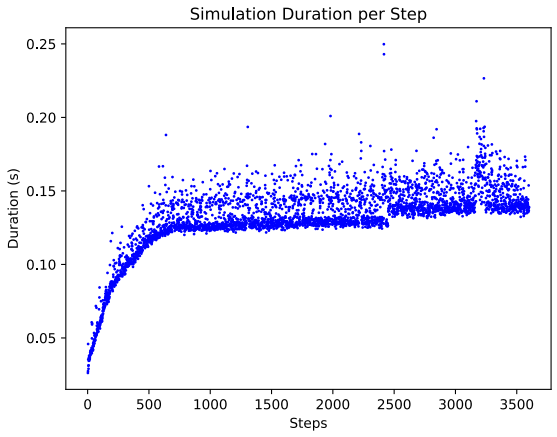


Fig. 8. Execution time per step

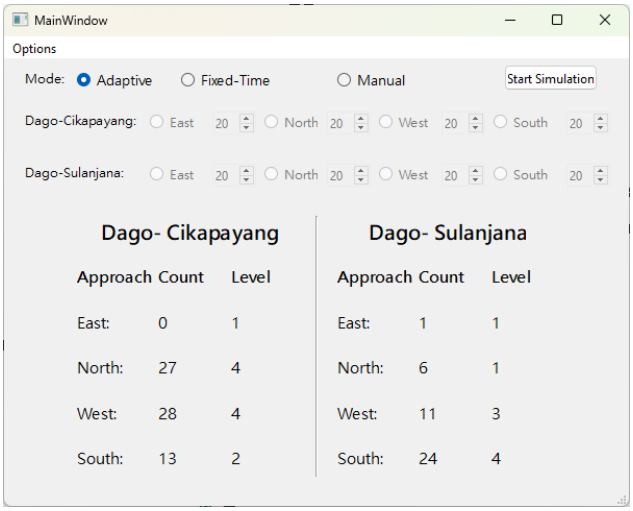


Fig. 9. GUI of the training platform

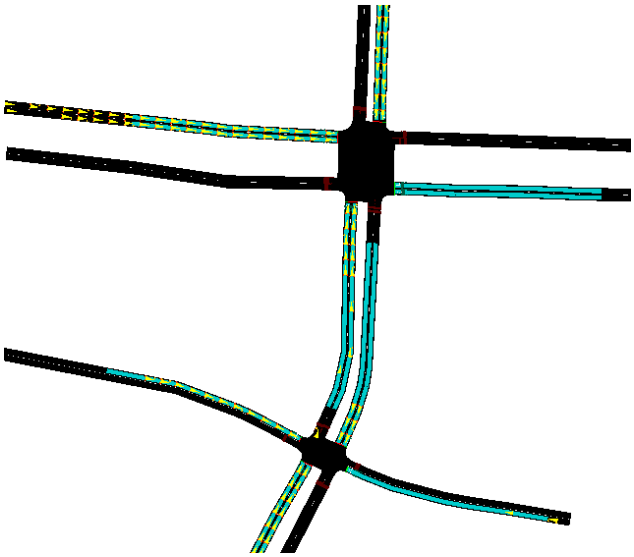


Fig. 10. SUMO simulation for GUI testing

this is recorded when the duration of each simulation step is set to 10 milliseconds. Referring to Figure 8, the duration of delay is varied according to the step of the simulation. An adjustment to the simulation step duration is needed so that the unsynchronization is reduced.

VI. Conclusion

Overall, the system has been implemented correctly and works appropriately. This is shown during the testing of the GUI in Figure 9 and Figure 10. The simulation is running well. It means that the underlying functions is implemented correctly. The behavior of the traffic light comply with regulations and the intended design. However, the system currently take Q-Matrix from CSV. The intended design was, the system is connected to FPGA and communicates directly with the FPGA. Further development in the communications with FPGA is needed. The mechanism of the communications should be designed.

Improvement could also be made as an effort in making the simulation faster. A device dedicated to simulate the environment could provide a good computational power for simulation, especially when the complexity of the simulation is high. In addition, an adjustment to the simulation step time is needed in order to reduce the unsynchronization between simulator and miniature.

References

[1] B. Yulianto, "Detector technology for demand responsive traffic signal control under mixed traffic conditions," in *AIP Conference Proceedings*, Author(s), 2018. DOI: 10.1063/1.5042991. [Online]. Available: https://doi.org/10.1063/1.5042991.

[2] R. P. Roess, E. S. Prassas, and W. R. McShane, *Traffic Engineering*, en, 3rd ed. Upper Saddle River, NJ: Pearson, May 2004.

[3] S. Shepherd, "Traffic control in over-saturated conditions," *Transport Reviews*, vol. 14, no. 1, pp. 13–43, Jan. 1994. DOI: 10.1080/01441649408716864. [Online]. Available: https://doi.org/10.1080/01441649408716864.

[4] Y. Wang, X. Yang, H. Liang, and Y. Liu, "A review of the self-adaptive traffic signal control system based on future traffic environment," *Journal of Advanced Transportation*, vol. 2018, pp. 1–12, Jun. 2018. DOI: 10.1155/2018/1096123. [Online]. Available: https://doi.org/10.1155/2018/1096123.

[5] M. Gregurić, M. Vujić, C. Alexopoulos, and M. Miletić, "Application of deep reinforcement learning in traffic signal control: An overview and impact of open traffic data," *Applied Sciences*, vol. 10, no. 11, p. 4011, Jun. 2020. DOI: 10.3390/app10114011. [Online]. Available: https://doi.org/10.3390/app10114011.

[6] A. Müller, V. Rangras, T. Ferfers, *et al.*, "Towards real-world deployment of reinforcement learning for traffic signal control," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2021, pp. 507–514. DOI: 10.1109/ICMLA52953.2021.00085.

[7] J. Rowson, "Hardware/software co-simulation," in *31st Design Automation Conference*, Jun. 1994, pp. 439–440. DOI: 10.1109/DAC.1994.204143.

[8] Z. H. Khan and T. A. Gulliver, "A macroscopic traffic model for traffic flow harmonization," *European Transport Research Review*, vol. 10, no. 2, Jun. 2018. DOI: 10.1186/s12544-018-0291-y. [Online]. Available: https://doi.org/10.1186/s12544-018-0291-y.

[9] *Indonesian highway capacity manual (ihcm)*, Directorate General Bina Marga, Feb. 1997.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Sparsity Exploration for Structured and Unstructured Weight Formations in CNN Architecture

Devi Noor Endrawati[§],Infall Syafalni[‡§],
Nana Sutisna[‡] and Trio Adiono[‡§]
[‡]School of Electrical Engineering and Informatics
[§]University Center of Excellence on Microelectronics
Bandung Institute of Technology

***Abstract*—The Convolutional Neural Network (CNN) is a frequently used algorithm used technique for processing pixel data and visual images. Nevertheless, the utilization of extensive network sizes poses significant obstacles to the underlying technology's throughput and energy efficiency. Various methods have been devised to enhance the efficiency and computing speed of Convolutional Neural Network (CNN) acceleration, improving the accuracy of the CNN process. Compression techniques have been employed in the algorithmic aspect to minimize the number of parameters in each layer, enhancing the detection process's efficiency. Pruning is a compression technique that involves the removal of neurons and synapses that do not play a substantial role in the detection process, hence minimizing their impact on the Neural Network layer. The compression method generates sparse data, wherein a significant amount of data has a value of zero. Various pruning techniques provide distinct types of sparse data, encompassing organized structured sparse and unstructured sparse data.**

The consequent form of sparse data structure has many advantages and downsides. Based on an analysis of various extant sparse data structures, an innovative accelerator has been developed to perform convolution operations using various types of sparse data obtained via distinct pruning procedures. The design architecture outlined in this study involves the removal of all zero-valued data weights through a pruning technique following the training process. The accelerator will process all non-zero weights by performing the convolution process promptly, utilizing the "weight stream" data. The findings from the conducted design testing indicate that the utilization of trimmed sparse data leads to improved performance in terms of faster output during the convolution process. The researchers conducted experiments utilizing diverse types of sparse weighted data, encompassing both structured and unstructured forms. These experiments provided evidence that the hardware accelerator is capable of performing convolutional computing operations with different variations of sparse data. The convolution procedure utilizing the developed accelerator is projected to yield a reduction of 56% in the computation results of all layers in YoloV3-Tiny. In the context of data calculations, including a sparsity percentage of up to 30%, it can be observed that the convolution process exhibits a reduction of 25% in relation to the overall convolution process.

I. Introduction

CNN is widely recognized as a prominent approach for constructing a robust architecture to facilitate model detection. Several detection model algorithms have been devised in image processing to comprehend images, discern objects, and effectively categorize them. The utilization of Convolutional Neural Networks (CNN) in image processing has been observed in a multitude of applications, including but not limited to autonomous cars, video surveillance, facial recognition, and other activities related to image processing. The structure of CNN exhibits similarities to the cognitive processes involved in visual information processing in the human brain. The layers of the CNN algorithm progressively extract more intricate information as the data traverses the network. This method enables the model to comprehend abstract and detailed elements inside images.

The CNN architecture incorporates a mechanism for parameter sharing. In the convolution layer, every neuron employs uniform weights to identify characteristics throughout the entirety of the image. This approach effectively decreases the number of parameters the model has to learn, enabling CNN to be trained using extensive datasets without requiring excessive computer resources. The utilization of CNN, which has the ability to learn and identify common characteristics seen in extensive image datasets, renders it suitable as an initial framework for transfer learning. This conditions implies that a model trained on a substantial dataset can be applied to several tasks, or a smaller dataset can be utilized with slight modifications. Additionally, this practice aids in conserving time and computational resources while training novel models. The advent of CNN has led to significant progress in image processing and computer comprehension. The model's performance and efficiency in image processing applications have been further improved by the development of more complex CNN architectures like ResNet, VGG, and Yolo.

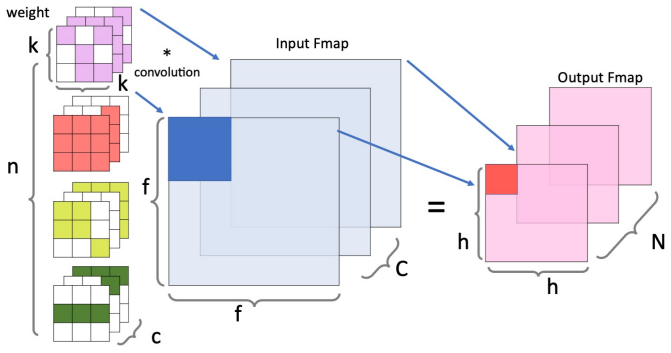


Fig. 1. Convolution Process

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

CNN is a deep learning method specifically developed to analyze and manipulate data presented in a grid-like structure, such as a two-dimensional image. There has been a proliferation of studies involving CNN, leading to widespread implementations of AI in fields as disparate as computer vision, robotics, autonomous vehicles, and medical diagnostics. CNN, short for Convolutional Neural Network, is a computational model that draws inspiration from the mechanisms behind visual perception in mammals. Compression is a viable technique for expediting the acceleration process, achieved by enhancing the speed of the detection process through the utilization of the Convolutional Neural Network (CNN) algorithm. The CNN process acceleration improvement procedure encompasses three main stages: the Algorithm, Structural, and Hardware level [1]. Compression techniques, such as pruning, occupy a position inside the redundancy process at the structural level. At this stage of the structure, various compression techniques are implemented, including pruning and quantization, in order to compress the weights of data or filters used in the detection phase. The acceleration process is implemented at the algorithmic level through the incorporation of additional algorithms into the architecture of the detection model. Similar to the convolution procedure, the efficacy of the convolution process can be enhanced by including convolution algorithms such as Winograd, Fast Fourier Transform (FFT), and General Matrix Multiply (GEMM). At the hardware level, the acceleration of the CNN process is achieved through the development of specialized devices capable of enhancing the speed and efficiency of the fundamental computations involved in the CNN algorithm. Furthermore, a pertinent concern that arises in the realm of hardware pertains to the methods of parallelization and data movement reduction. In the context of detection, particularly real-time detection, a substantial volume of data will be necessary for the detection process.

Commonly operated hardware components for the detecting process encompass Field-Programmable Gate Arrays (FPGAs), Central Processing Units (CPUs), Graphics Processing Units (GPUs), and Application-Specific Integrated Circuits (ASICs). The CPU stands out as a widely utilized and essential component among the array of available hardware components. However, the CPU exhibits limitations in terms of efficiency due to its nature as a hardware component designed for general-purpose computing operations. The utilization of hardware accelerators, such as GPU, FPGA, and ASIC, in conjunction with the CPU, is a common practice in computing. A hardware accelerator is a dedicated device designed to execute specific activities with enhanced efficiency.

Hence, this study aims to develop an innovative design architecture for conducting convolution on the Tiny Yolo V3 image detection or inference object process. The proposed architecture will leverage sparse weight data to effectively decrease the computing burden while maintaining the accuracy of the system's outcomes. The convolution process will be reduced due to convolving just non-zero weight data since it excludes the processing of zero weight values and the subsequent input feature map. However, the architecture being constructed must maintain accuracy. This design architecture is achieved by ensuring that the architectural design does not alter the weight value and accurately maps the output to return the convolution result data to the appropriate output index.

II. Prelinimary

A. Previous Work

Research has been conducted in the field of hardware acceleration for the purpose of manufacturing specialized hardware for sparse data. Notably, there exists hardware that employs the YOLO (You Only Look Once) detection algorithm, which boasts a streamlined architecture and enables real-time detection with enhanced speed. Hardware improvements have been made to the convolution architecture of Yolo by incorporating Winograd into the convolution process [2]. In addition, the hardware described in [3] employs an alternative convolution method that utilizes GEMM to enhance the speed of the convolution process. In the realm of hardware optimization for convolution processes, a notable approach has been the integration of the FFT method with the Winograd algorithm. This combination, as discussed in [4], has enhanced convolution operations efficiency.

The third category of accelerator employs sophisticated convolution algorithms, such as the frequency domain convolution scheme [5], in its hardware design to minimize the computational burden, precisely the number of multiply-accumulate (MAC) operations, in contrast to direct spatial convolution algorithms. Consequently, this approach yields a significant performance improvement for the implemented accelerator. The accelerator architecture described in reference [2] employed the Winograd convolution technique, resulting in a notable 47% reduction in workload for the YOLOv1 algorithm. This improvement was achieved with a little trade-off, as the detection accuracy only experienced a slight 1% loss. The ultimate implementation attains a detection rate of 15.3 frames per second (fps) and utilizes 1024 digital signal processing (DSP) blocks on a Xilinx FPGA. The author of this study implemented the YOLO algorithm by effectively utilizing pipeline architecture, loop unrolling, and data reordering techniques on a PYNQ FPGA platform. The hardware accelerator sequentially processes each of these hidden layers. Additionally, the initial and final layers operate through software implementation, resulting in diminished frame rates. In the study conducted by the authors in [6], they introduced a variant of YOLO-v2 that incorporates binary networks and a regression support vector machine (SVM). The author devised the co-stream binary convolutional circuit, wherein each layer is processed sequentially. While previous iterations of designs achieved improved efficiency by decreasing algorithmic complexity, they failed to consider the corresponding decrease in external memory access.

B. Sparse Structured

The pruning algorithm is employed to selectively remove network connections to preserve the detection results accuracy

concerning the original network configuration. Observing the actuality of the training step for the detection model, a procedure is implemented to eliminate all connections that possess weights below a specified threshold. The pruning process involves converting a densely linked neural network layer into a sparser layer or network. The initial stage of this procedure involves an examination of the network topology, followed by an identification of critical connections and subsequent removal of non-essential connections. Subsequently, the network undergoes retraining to enable the surviving connections to compensate for removing the previously eliminated connections. The process of pruning and retraining can be iteratively employed to decrease the complexity of a network, resulting in the generation of sparse data. [7].

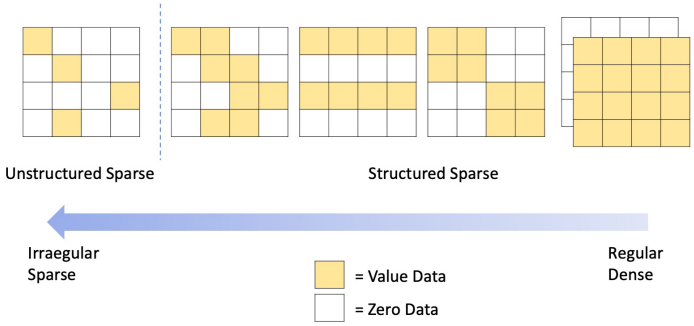


Fig. 2. Structure of data sparse

Sparsity refers to conditions in which a significant portion of a given dataset consists of zero values. Sparsity data can manifest in various forms, such as arrays, vectors, matrices, and tensors. In contrast, dense data refers to data that has a higher proportion of non-zero values. The sparsity value of the data weight is found to be significant during the pruning process. Hence, this study aims to develop a specific architecture that can effectively execute convolution for image detection objects in Tiny Yolo model inference. This architecture will leverage sparse weight data to minimize the computing burden while ensuring that the accuracy of the system's outcomes remains unaffected. The convolution process can be optimized by performing computation only on non-zero data. This optimization reduces the computational load by excluding the processing of weight values that are 0, as well as the input feature map that follows them. Nevertheless, the architectural integrity will be preserved as the design ensures that the weight value remains unchanged and the output is accurately mapped to the appropriate output index.

III. Proposed Method

A. Bitmap Indexing

Indexing or creating an index in the context of sparse data involves identifying and numbering elements inside a data structure. This index serves to provide information about the position of non-zero elements or elements that possess significant values within the sparse data structure. Given that a significant portion of the values inside the sparse data exhibit null characteristics, it becomes necessary to only retain and retrieve the non-zero elements for further computational procedures. Indexes facilitate the association of these items with their respective positions inside a data structure optimized for efficiency. Numerous indexing techniques have been used to leverage sparse data, including CSR (Compressed Sparse Row), CSC (Compressed Sparse Column),COO(Coordinate Format) Bitmap, etc.

Figure 3, part a, depicts the data structure known as CSR, which represents sparse matrices. The matrix rows are organized in an array format in this representation, requiring three distinct data types. Data weight is an Array that consists of non-zero values extracted from the matrix sequentially based on the rows. Furthermore, the Array contains the column index of every non-zero element within the matrix. The third component is called the Indice Pointer, an array that stores the starting index of each row within the data and indices array. CSC format is a sparse matrix representation with the same structure as the CSR format. The matrix is partitioned into vertical columns organized in a two-dimensional array. In the context of Indice Pointer data, the Array comprising the index consists of each column of the matrix, commencing with the data array and corresponding indices.

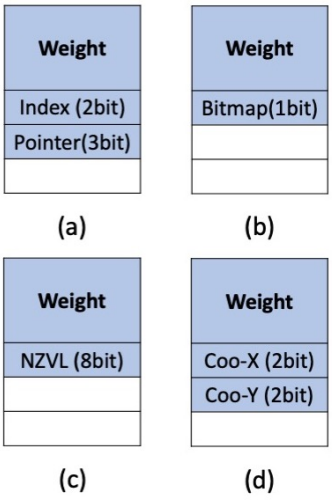


Fig. 3. Indexing method to exploit sparse position, (a)CSC&CSR, (b)Bitmap, (c)NZVL[8], (d)COO

COO refers to the non-zero elements of a matrix, which are identifiable by their corresponding coordinate pairs (row, column). A bitmap is a form of binary representation that represents a collection of elements. In the Bitmap, there exists an entry for every element, with a value of 1 denoting the presence of the element (non-zero) and a value of 0 indicating the absence of the element (zero). Bitmaps are commonly employed in situations where a significant proportion of elements possess a value of zero, whereas only a small subset of elements exhibit non-zero values. Indexing plays a crucial role in all of these representations as it enables the identification of

the precise location of non-zero items. This, in turn, facilitates rapid execution of mathematical operations or other forms of data manipulation. The hardware accelerator effectively utilizes indexing to expedite the retrieval of certain elements, hence avoiding the need to traverse the whole data structure. Additionally, it is advantageous in terms of conserving storage capacity and reducing execution time, which are crucial factors when handling big and intricate datasets, particularly in the context of real-time detection procedures.

The analysis of Figure 3 reveals that the utilization of distinct indexing strategies has a discernible impact on the amount of memory capacity that is needed. The process of indexing utilizing CSR and CSC representations requires the inclusion of two supplementary data elements in order to determine the location of non-zero values. The utilization of 2-bit and 3-bit data widths in this context bears a resemblance to the indexing technique employed in COO, wherein two additional 2-bit data elements are utilized. The study by the authors [8] employs an alternative indexing technique, namely a binary map and NZVL (Not Zero Value List), with the NZVL data width set at 8 bits. Hence, the hardware accelerator architecture in this study incorporates a bitmap indexing technique, which necessitates a data width of only 1 bit. The utilization of bitmap indexing also enables the efficient representation of diverse types of sparse data inside the dataflow process through "weight streams".

B. *Weight Stream*

The convolution procedure involves utilizing the input data, which consists of the pruning weight input and the Fmap input. Fmap, short for Feature Map or Activation Map, refers to the outcome of the convolution process. The proposed accelerator will prioritize the convolution process in the layer with the highest sparsity value, specifically layer 6. This implies that the input picture utilized will be the convolution output from the preceding layer.

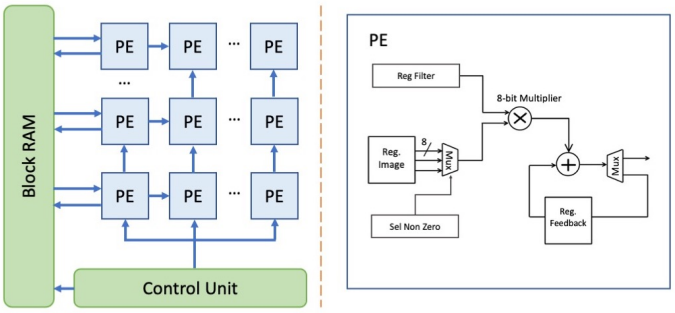


Fig. 4. Architecture Hardware Accelerator

As depicted in Figure 1, the convolution process involves an input weight size and an input feature map (fmap). The variables K, C, and N are utilized to denote the magnitude of the input weights. The kernel, a fundamental component in model detection designs, can be categorized into two sizes: 3×3 and 1×1 in terms of weight. In the process of convolution, the dimensions of the output channels (C) must match the dimensions of both the input weights and the input feature map (fmap). The variable N represents the number of filters present in a single layer. The size of the fmap input, denoted as $F \times F \times C$, refers to the dimensions of the input matrix. Consequently, the implementation of a heavy data flow architecture is deemed necessary for this accelerator. Subsequently, employing a single fmap operation, the input received via channel 0 will undergo convolution with distinct weights from channel 0 for a total of N iterations. The planned accelerator will utilize a Fmap input of dimensions 3×3. The utilization of a structured format, such as the one depicted in Figure 2, offers the benefit of facilitating a systematic approach to reading and selecting non-zero data. Nevertheless, a lower sparsity number has an impact on the computational load required by the hardware during the inference procedure. In the context of data analysis, unstructured data can give rise to sparse data sets characterized by a substantial number of non-zero values. This abundance of non-zero values can have a considerable impact on computational processes, leading to a significant reduction in efficiency. However, the presence of unstructured data poses challenges in terms of data readability once the final mapping procedure is conducted to generate suitable output data that can be further processed in subsequent stages. The present research introduces an architectural framework capable of effectively handling two distinct types of sparse data representations, namely structured and unstructured. This is achieved through the utilization of a data flow architecture that incorporates a weight stream.

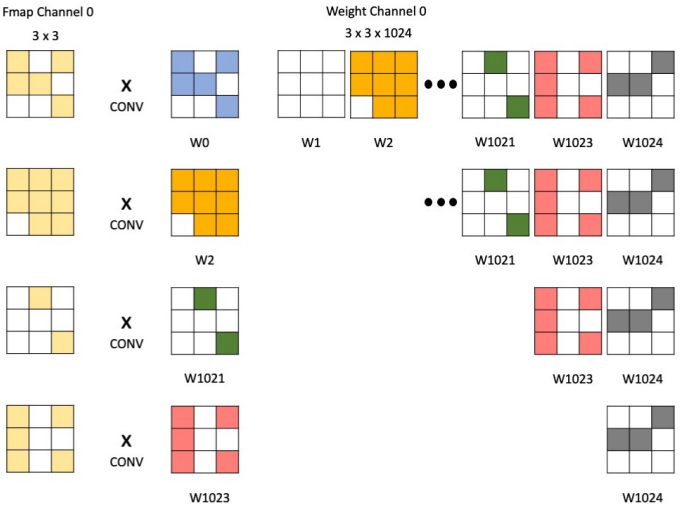


Fig. 5. Weight Stream Data Flow Architecture

The weights that have undergone compression during the software process result in the retention of solely non-zero or value-based data inside the weight data. When all data zeros are present on the same channel, it enables a continuous flow of weight through the accelerator, generating a convolution product that aligns with the input bitmap utilized for selecting the feature map input. Hence, invoking the fmap input several

times is unnecessary to execute the convolution procedure N times. However, invoking the function once for a given channel is sufficient. Utilizing a convolution accelerator with a data flow of this nature facilitates the efficient reuse of data, hence obviating the need for more power during the iterative process of utilizing the same data.

Figure 4 depicts the intended architectural architecture, whereby the data selection process that will be fed into the convolution process is present in each processing node. Additionally, each processing element is interconnected inside a parallel network. Figure 5 depicts the process known as the ”weight stream.” The data from current weights will be sequentially processed without any processing delay for different weights. Simultaneously, the selection of fmap data for the convolution process will also be conducted without any delay in the overall process.

IV. Experimental Results

In the hardware design process or circuit design, it is essential to conduct testing to achieve the required output before its implementation into the Field-Programmable Gate Array (FPGA). During the accelerator design phase for the convolution process, a simulation methodology is employed to verify the alignment of the design outcomes with the system design and the intended objectives. Furthermore, within the context of this simulation process, it will provide an assessment of the performance of the designed system. The simulation method is conducted using various input conditions to evaluate the efficacy of the design.

The decision was made to utilize a timing diagram to modify this accelerator's architecture. A timing diagram is a graphical representation that depicts the temporal relationship between one or more elements by displaying their respective values at specific points in time. Utilizing a timing diagram in the simulation aims to offer valuable insights during the design phase, facilitating the attainment of the intended output. The timing diagram can also depict the outcomes of the convolution accelerator architecture, wherein the input weight is provided to the hardware consecutively. This enables the identification of any potential calculation errors, facilitating accurate convolution results.

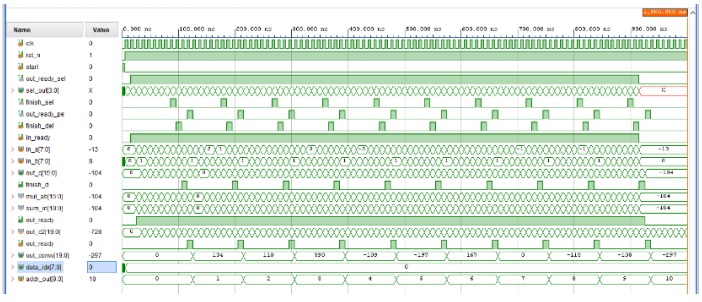


Fig. 6. Output Simulation using dense data

The timing diagram results depicted in Figure 6 demonstrate that each weight necessitates an equivalent amount of time throughout the convolution output acquisition operation, specifically 12 clock cycles for each datum. The time diagram illustrates the convolution process of a processing unit characterized by sequential multiplication and propagation of data flow. Each multiplication between one kernel weight requires a total of nine clock cycles. Performing a 3×3 matrix multiplication with an input feature map of size 3×3.

The output value of the data selected is determined by the non-zero data position of the input data bitmap and is activated when the 'in_ready' signal is active. Additionally, the input weight process is initiated. However, it is essential to note that the input fmap and input weight cannot coexist simultaneously. This is due to the fact that the input fmap experiences a delay of one clock cycle as a result of the initial 'sel_out' selection process. Consequently, the data input weight is also delayed by one clock cycle before continuously flowing into the module processing element. As depicted in Figure 6, the input weight 'in_a' and the input feature map data 'in_b' are concurrently inputted. The requirement of 92 clock cycles is due to the dense nature of the input data, which ensures that no information is omitted from the 10 data weights. In the meanwhile, it is seen that each weight necessitates a total of 11 clock cycles. Given the presence of 9 input weights and the utilization of a sequential PE architecture, the multiplication operation involving 9 data inputs necessitates nine clock cycles. However, an additional two clock cycles are incurred due to the requirement of a preliminary select index process within the architecture for data entry. The initial clock cycle is allocated for the select index operation. In contrast, an additional clock cycle is required for the final data storage step, which corresponds to the number of convolution outputs.

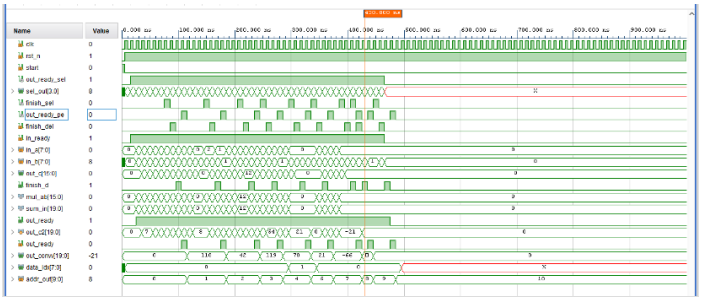


Fig. 7. Output Simulation using sparse data

In this experiment involving sparse data, we will utilize the same dataset as the previous one. However, specific data points will be modified to have a value of 0. The objective is to examine the impact of employing dense data in the results of the timing diagram from the previous simulation. It is important to note that all bitmap index data and relative index have been generated prior to conducting the simulation. Nevertheless, in the context of simulating sparse data, using the same input data weight in conjunction with bitmap data allows for incorporating sparse data and input index bitmap. This facilitates the multiplication process within the convolution, resulting in the propagation of zero data. It is anticipated that

the computational time required to perform the convolution process ten times will be shorter than the dense data procedure.

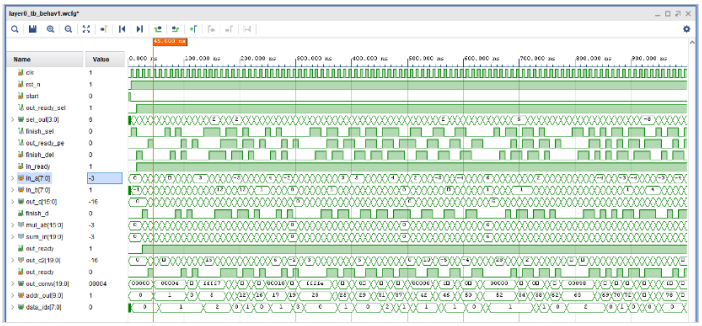


Fig. 8. Output Simulation using all kind of sparse and dense data

The convolution procedure for the inference process at one layer is shown in Figure 8. The results presented in this study demonstrate a range of weight data in different forms and quantities, spanning from sparse to dense distributed weight data. It is seen that both structured and unstructured sparse data can yield satisfactory outputs without inducing processing delays, even in cases where there are variations in the data structure across different weight data sets. Established on the test results, which were conducted in adherence to the prescribed architecture, specifically by employing the techniques of skipping zero data and utilizing compressed bitmap data, it is evident that there exists a disparity in the number of clock cycles required to accomplish the convolution of the data weights. Due to the compression of the bitmap data, the accelerator will exclusively handle nine data weights, while the remaining data will be contained within each weight. According to the data presented in Figure 7, it can be observed that the 5th data weight possesses a cumulative data weight of zero. Additionally, there is an absence of non-zero data inside this weight. Consequently, the bitmap data included within this weight will be eliminated, excluding the 5th weight from the accelerator process.

At the outset, there were nine data weights assigned to each weight. Consequently, for a total of ten data weights, the convolution process will involve a cumulative dataset of 90 data points. The sparse data simulation procedure involves skipping zero data points and not utilising compressed bitmap data. In this particular scenario, only 45 data points are obtained from a set of 10 data weights. The process of convolving the entire data weight, starting from the initial data bitmap entry, requires a total of 47 clock cycles. Compared to the convolution outcomes depicted in Figure 7, the technique as mentioned earlier yields identical results consisting of nine convolutions.

Upon conducting the simulation, it was determined that the number of convolution processes required to handle sparse data can be calculated. The results of this calculation indicate that the convolution process utilizing the designed accelerator is projected to yield a reduction of 56% in the calculation results of all YoloV3-Tiny layers. In the context of data calculations

TABLE I
COMPARISON OF SPARSE HARDWARE DESIGN

No	*Architecture*	*Sparse*	*Indexing*	*Dynamic*	*Pruning Rate*
1	Sparse-Yolo[9]	Unstructured	CSR	Dynamic	89%
2	RDL[8]	Unstructured	NZVL	Dynamic	59.5%
3	This Work	Unstructured Structured	Bitmap	Static	70% 30%

involving sparsity percentages of up to 30%, it can be observed that the convolution process experiences a reduction of 25% concerning the overall convolution process. Determining the number of multiplication operations involved in a convolution operation is contingent upon the resultant dimensions of the convolution output. The utilization of an equation can achieve the determination of the output size of the convolution process:

$$h = \frac{f - k + 2P}{S} + 1, \tag{1}$$

where h represents the output size, which can refer to the length or width dimension depending on the context in which the size is calculated. The variable f represents the input size, which can refer to length or width, depending on the specific dimensions being calculated. The variable k denotes the size of the kernel, which refers to the length or width of the kernel. P represents the quantity of padding applied to the input. Without any additional padding, the value of P equals zero. When padding is employed, the variable P encompasses the total quantity of padding on both sides, which is typically executed symmetrically. And the variable S represents the step or stride length employed during the execution of the convolution operation. Stride refers to the displacement distance by which the kernel is displaced during each iteration of the convolution process. Without a stride, the value of S is equal to 1.

The formula above will quantitatively assess the production within a single dimension. The output data for the convolution process will be specified by the number of input channels included in the weight data. The number of multiplication operations performed in a convolution process can be determined by utilizing the equation subsequent to acquiring the size of the output resulting from the convolution process using the equation :

$$Num_{mult} = h \times h \times k \times k \times c \times N, \tag{2}$$

where Num_{mult} denotes the quantity of multiplications, which is determined by the product of the number of weight data and output data. The reduction in data weight resulting from the pruning procedure and the acquisition of numerous zero data points can effectively decrease the computational burden associated with multiplication operations. Quantifying the multiplications performed in each layer makes it possible to determine the percentage of convolutional operations before and after implementing the architectural design proposed in this research study.

Furthermore, compared to the hardware architectures presented in Table 1, the architectural design proposed in this paper offers the benefit of employing several types of sparse data structures, encompassing organized and unstructured formats. able 1 presents a comparison between the implementation of the YOLO version employed in this study and the YOLO version utilized in the paper referenced as [9]. This study employs an alternative iteration of the convolution-based design while maintaining the fundamental structure. The article titled "Sparse-Yolo" [9] achieved a sparsity value of 89% through the implementation of the pruning process approach. This implies that a mere 11% of the data necessitates processing during the convolution procedure. In addition, the study employs an alternative detection model, specifically the general Convolutional Neural Network (CNN) model, with a structured sparse block form, resulting in a sparsity level of 50% [10]. However, in the case of other works, such as [11], the hardware accelerator is constructed using either random sparse or non-random data. The structure of the data exhibits a sparsity of 62.5%, while the data index width is relatively high, measuring 10 bits. Hence, the architectural design of this paper confers benefits in the processing of many types of sparse data. The present study explores the implementation of the hardware architectural design for conducting computational processes, including both structured and unstructured data. The findings indicate that the accuracy achieved by this approach is comparable to that of alternative hardware-based computing methods.

The determination of the reduction percentage value for the convolution procedure is derived by evaluating the number of convolution computations that need to be performed utilizing the available sparse data. There is a discernible distinction in computational outcomes between utilizing a data sparsity of 30% exclusively with non-zero data and employing the same sparsity value of 30% while still processing all zero values without skipping them. Hence, employing different sparse data structured leads to a higher percentage value of computational speed while effectively handling zero data by skipping it.

V. Conclusions

Once the procedure has determined the output size in each layer, it becomes possible to calculate the total number of clock cycles for each layer. The number of clock cycles required for processing sparse data in a processing element (PE) is equivalent to the number of multiplications executed in the PE. This is due to the sequential nature of the multiplication process in the PE. The subject in question. The accelerator system requires an additional clock cycle at its initiation to process bitmap data specifically for the purpose of picking fmap input data. In addition, it is necessary to allocate an additional clock cycle at the conclusion of the convolution operation in order to save the resultant value. By using this architecture and "weight stream" dataflow, implementing zero data skipping from any structured sparse data (structured / untructured) in the convolution process can potentially enhance hardware's computational efficiency. The accelerator's efficiency is enhanced due to the decreased memory-to-weight ratio after trimming, leading to a reduction in the number of chip design regions. This reduction is attributed to the ability to decrease the required processing parts for computation.

Acknowledgment

This work was supported by Indonesia Endowment Fund for Education (LPDP) grant /TRP/D/9/lpdp2022.

References

[1] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49908466.

[2] C. Bao, T. Xie, W. Feng, L. Chang, and C. Yu, "A power-efficient optimizing framework fpga accelerator based on winograd for yolo," *IEEE Access*, vol. 8, pp. 94 307–94 317, 2020. DOI: 10.1109/ACCESS.2020.2995330.

[3] T. Adiono, A. Putra, N. Sutisna, I. Syafalni, and R. Mulyawan, "Low latency yolov3-tiny accelerator for low-cost fpga using general matrix multiplication principle," *IEEE Access*, vol. 9, pp. 141 890–141 913, 2021. DOI: 10.1109/ACCESS.2021.3120629.

[4] Y. Liang, L. Lu, Q. Xiao, and S. Yan, "Evaluating fast algorithms for convolutional neural networks on fpgas," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 4, pp. 857–870, 2020. DOI: 10.1109/TCAD.2019.2897701.

[5] H. Nakahara, M. Shimoda, and S. Sato, "A tri-state weight convolutional neural network for an fpga: Applied to yolov2 object detector," in *2018 International Conference on Field-Programmable Technology (FPT)*, 2018, pp. 298–301. DOI: 10.1109/FPT.2018.00058.

[6] H. Nakahara, H. Yonekawa, T. Fujii, and S. Sato, "A lightweight yolov2: A binarized cnn with a parallel support vector regression for an fpga," in *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '18, Monterey, CALIFORNIA, USA: Association for Computing Machinery, 2018, pp. 31–40, ISBN: 9781450356145. DOI: 10.1145/3174243.3174266. [Online]. Available: https://doi.org/10.1145/3174243.3174266.

[7] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1510.00149.

[8] C.-B. Wu, C.-H. Chen, and C.-P. Kuan, "Reconfigurable deep learning accelerator hardware architecture design for sparse cnn," in *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2021, pp. 1–2. DOI: 10.1109/ICCE-TW52618.2021.9602959.

[9] Z. Wang, K. Xu, S. Wu, L. Liu, L. Liu, and D. Wang, "Sparse-yolo: Hardware/software co-design of an fpga accelerator for yolov2," *IEEE Access*, vol. 8, pp. 116 569–116 585, 2020. DOI: 10.1109/ACCESS.2020.3004198.

[10] Z.-g. Liu, P. N. Whatmough, and M. Mattina, "Systolic tensor array: An efficient structured-sparse gemm accelerator for mobile cnn inference," *IEEE Computer Architecture Letters*, vol. 19, pp. 34–37, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:215854824.

[11] A. Parashar, M. Rhu, A. Mukkara, *et al.*, "Scnn: An accelerator for compressed-sparse convolutional neural networks," *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 27–40, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:1775161.

[12] S. Zhang, Z. Du, L. Zhang, *et al.*, "Cambricon-x: An accelerator for sparse neural networks," Oct. 2016, pp. 1–12. DOI: 10.1109/MICRO.2016.7783723.

[13] S. Han, J. Kang, H. Mao, *et al.*, "Ese: Efficient speech recognition engine with sparse lstm on fpga," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '17, Monterey, California, USA: Association for Computing Machinery, 2017, pp. 75–84, ISBN: 9781450343541. DOI: 10.1145/3020078.3021745. [Online]. Available: https://doi.org/10.1145/3020078.3021745.

[14] H.-J. Kang, "Accelerator-aware pruning for convolutional neural networks," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 30, no. 7, pp. 2093–2103, Jul. 2020, ISSN: 1051-8215. DOI: 10.1109/TCSVT.2019.2911674. [Online]. Available: https://doi.org/10.1109/TCSVT.2019.2911674.

[15] S. Han, X. Liu, H. Mao, *et al.*, "Eie: Efficient inference engine on compressed deep neural network," in *Proceedings of the 43rd International Symposium on Computer Architecture*, ser. ISCA '16, Seoul, Republic of Korea: IEEE Press, 2016, pp. 243–254, ISBN: 9781467389471. DOI: 10.1109/ISCA.2016.30. [Online]. Available: https://doi.org/10.1109/ISCA.2016.30.

[16] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," vol. 44, Jun. 2016. DOI: 10.1145/3007787.3001177.

[17] D. Kim, J. Ahn, and S. Yoo, "Zena: Zero-aware neural network accelerator," *IEEE Design and Test*, vol. 35, no. 1, pp. 39–46, 2018. DOI: 10.1109/MDAT.2017.2741463.

[18] J. Albericio, P. Judd, T. Hetherington, T. Aamodt, N. E. Jerger, and A. Moshovos, "Cnvlutin: Ineffectual-neuron-free deep neural network computing," in *Proceedings of the 43rd International Symposium on Computer Architecture*, ser. ISCA '16, Seoul, Republic of Korea: IEEE Press, 2016, pp. 1–13, ISBN: 9781467389471. DOI: 10.1109/ISCA.2016.11. [Online]. Available: https://doi.org/10.1109/ISCA.2016.11.

[19] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15, Montreal, Canada: MIT Press, 2015, pp. 1135–1143.

[20] S. Zhai, C. Qiu, Y. Yang, J. Li, and Y. Cui, "Design of convolutional neural network based on fpga," *Journal of Physics: Conference Series*, vol. 1168, p. 062 016, Feb. 2019. DOI: 10.1088/1742-6596/1168/6/062016.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, 2015.

[22] X. Zhang and S. Ding, "Research on traffic target detection method based on improved yolov3," May 2021, pp. 39–43. DOI: 10.1109/ICoIAS53694.2021.00015.

[23] D. T. Nguyen, T. N. Nguyen, H. Kim, and H.-J. Lee, "A high-throughput and power-efficient fpga implementation of yolo cnn for object detection," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 8, pp. 1861–1873, 2019. DOI: 10.1109/TVLSI.2019.2905242.

[24] S. Li, C. Yu, T. Xie, and W. Feng, "A power-efficient optimizing framework fpga accelerator for yolo," in *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2022, pp. 1–6. DOI: 10.1109/CISP-BMEI56279.2022.9980321.

[25] G. Zhang, K. Zhao, B. Wu, Y. Sun, L. Sun, and F. Liang, "A risc-v based hardware accelerator designed for yolo object detection system," in *2019 IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE)*, 2019, pp. 9–11. DOI: 10.1109/ICIASE45644.2019.9074051.

1M parameters are enough? A lightweight CNN-based model for medical image segmentation

Binh-Duong Dinh
Department of Automation Engineering
School of Electrical and Electronic Engineering
Hanoi University of Science and Technology
duong.db190043@sis.hust.edu.vn

Thanh-Thu Nguyen
Department of Automation Engineering
School of Electrical and Electronic Engineering
Hanoi University of Science and Technology
thu.nt192096@sis.hust.edu.vn

Thi-Thao Tran*
Department of Automation Engineering
School of Electrical and Electronic Engineering
Hanoi University of Science and Technology
thao.tranthi@hust.edu.vn

Van-Truong Pham
Department of Automation Engineering
School of Electrical and Electronic Engineering
Hanoi University of Science and Technology
truong.phamvan@hust.edu.vn

***Abstract*—Convolutional neural networks (CNNs) and Transformer-based models are being widely applied in medical image segmentation thanks to their ability to extract high-level features and capture important aspects of the image. However, there is often a trade-off between the need for high accuracy and the desire for low computational cost. A model with higher parameters can theoretically achieve better performance but also result in more computational complexity and higher memory usage, and thus is not practical to implement. In this paper, we look for a lightweight U-Net-based model which can remain the same or even achieve better performance, namely U-Lite. We designed U-Lite based on the principle of Depthwise Separable Convolution so that the model can both leverage the strength of CNNs and reduce a remarkable number of computing parameters. Specifically, we propose Axial Depthwise Convolutions with kernels 7×7 in both the encoder and decoder to enlarge the model's receptive field. To further improve the performance, we use several Axial Dilated Depthwise Convolutions with filters 3×3 for the bottleneck as one of our branches. Overall, U-Lite contains only 878K parameters, 35 times less than the traditional U-Net, and much more times less than other modern Transformer-based models. The proposed model cuts down a large amount of computational complexity while attaining an impressive performance on medical segmentation tasks compared to other state-of-the-art architectures. The code will be available at: https://github.com/duong-db/U-Lite.**

***Index Terms*—Depthwise Separable Convolution, Medical Image Segmentation, Lightweight Model, Convolutional Neural Networks, U-Lite.**

I. Introduction

Artificial Intelligence (AI) has been recently deployed in many hospitals for practical use in developed countries. According to Health Equity Magazine (2018), AI helps store and access a large amount of information effectively. Currently, the amount of medical information has doubled every three years. It is estimated that if a doctor wants to keep up with all medical news, he must read 29 hours a day, which is impossible. On the other hand, a deep learning model can also assist in diagnosing medical images. The IBM Watson healthcare system (USA) shows that 90% of the recommendations made by their AI system are consistent with the suggestions of medical experts but taking only 40 seconds to complete all processes. Realizing the importance of AI in the medical field, many study efforts have been made to improve the performance of deep learning models.

Computer Vision is one of the most prominent applications of AI in the medical field. Besides diagnosing diseases based on clinical symptoms, modern medicine also diagnoses diseases based on subclinical symptoms from images obtained from medical devices. Therefore, deep learning models are developed for segmenting tumors and cells or abnormal areas, thereby supporting doctors in disease identification and diagnosis and the severity of the disease. In 2015, Ronneberger et al. introduced U-Net [1] as a highly effective model for medical image segmentation. After the success of U-Net, many follow-ups work did different research to optimize U-Net and offered numerous variants with higher performance such as Unet++ [2], ResUnet++ [3], Double Unet [4], Attention Unet [5], to name just a few. They all are deep learning models developed based on CNNs. There is no denying that the appearance of CNNs has created a great revolution in computer vision fields.

Recently, Vision Transformer [6] and MLP-like architectures (MLPs) have been widely used and have become a new de facto standard in Computer Vision. Vision Transformer considers each patch of an image as a token and feeds them through the Multi-head Self-attention mechanism as they successfully did with sentences in Natural Language Processing (NLP). In the medical segmentation task, TransUnet [7] can be considered one of the high-performing models in accuracy and efficiency. Following the success of TransUnet, Transformer-based models continued to be developed. Pyramid Vision Transformer (PVT) [8] is utilized as the backbone of many high-performance models such as MSMA-Net [9], Polyp-PVT [10]. Meanwhile, MLP-like architectures are also very research-focused. MLPs leverage

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the advantage of conventional MLP to encode the features along each of their dimensions. AxialAtt-MLP-Mixer [11] gives exceptionally reliable performance on many medical image datasets by applying axial attention to replace the token mixing in MLP-Mixer [12]. Different from CNNs, the models based on Transformers or MLPs concentrate on the global receptive field of the image, thus costing much in computational complexity and being overly heavy for the training process.

To successfully implement in practice, a machine learning model first needs to achieve high accuracy and secondly, it should be fast and compact enough to integrate into mobile medical devices. Nevertheless, there is often a trade-off between the need for high accuracy and the desire for low computational cost. These above studies can theoretically achieve impressive performance, but many may give heavy operation and slow calculation speed due to the massive number of parameters. To solve this problem, some attempts for a lightweight architecture can be mentioned as Mobile-Unet [13], DSCA-Net [14], and MedT [15]. In this paper, we rethink an efficient lightweight architecture for the medical segmentation task to further explore a high-performing model which can effectively address this issue. In short, the main contributions of this paper are three-fold:

1) Propose the usage of Axial Depthwise Convolution module based on the concept of Depthwise Separable Convolution. This module helps the model solve every complex architecture problem: enlarging the model's receptive field while reducing the heavy computation burden.
2) Propose U-Lite, a lightweight and simple CNN-based architecture. As far as we know, U-Lite is one of the few models that surpasses a recent highly efficient compact network UNeXt [16] in terms of both accuracy and the number of parameters.
3) We have successfully implemented the model on medical segmentation datasets and achieved considerable results.

II. Related Work

U-Net. Introduced in 2015 [1], U-Net is an effective deep-learning CNNs-based model for biomedical image segmentation. U-Net follows the U-shape architecture with a typical encoder-decoder structure. The encoder uses convolution, activation, and normalization layers to extract high-level features of the input image. Meanwhile, max-pooling layers are applied to capture multi-scale features, allowing the model to handle different shapes and sizes of objects. The decoder utilizes upsampling layers to rescale the size of encoded features back to the original image's size. To preserve spatial information, U-Net adds skip connections between the encoder and decoder, which improves the model's performance and solves the vanishing gradient problem.

Depthwise Separable Convolution. Depthwise Separable Convolution splits the convolution into two steps: depthwise and pointwise. Depthwise convolution encodes the features per channel, while pointwise convolution encodes these features along their depth dimension. Depthwise Separable Convolution delivers almost the same performance as traditional convolution while using fewer parameters, thereby reducing computational complexity, and making the model more compact. Therefore, this technique has been widely used in many modern deep learning models, for example, InceptionNet [17] and MobileNets [18].

III. Proposed Model

An overview of our proposed model U-Lite is illustrated in Fig. 1. We follow the symmetric encoder-decoder architecture of U-Net and design U-Lite in an efficient way so that the model can leverage the strength of CNNs while maintaining the number of computational parameters as small as possible. To this end, an Axial Depthwise Convolution module is thoughtfully proposed, shown in Fig. 2. Describing the operation of U-Lite, an input image at shape $(3,H,W)$ is fed to the network through 3 stages: encoder stage, bottleneck stage, and decoder stage. U-Lite follows the hierarchical architecture, where the encoder extracts six different level features in shape $\left(C_i, \frac{H}{2^i}, \frac{W}{2^i}\right)$, where $i \in \{0,1,2,..,5\}$. The bottleneck and decoder take part in processing these features as well as upscaling them to the original shape to obtain the segmented mask. We also use skip connections between the encoder and decoder. It is worth noting that although the design of U-Lite is simple, the model still performs well on the segmentation task thanks to the contribution of Axial Depthwise Convolution module.

A. Axial Depthwise Convolution module

The success of Vision Transformer [6] promotes various works on researching and improving this special structure. Swin Transformer [19] reduces the computational complexity of Transformer by limiting self-attention computation to non-overlapping local windows of size 7×7. ConvNext [20] realizes this modification and adopts convolutions with kernel size 7×7 to CNN architecture, bringing ResNet up to 86.4% top-1 accuracy on ImageNet. Meanwhile, a recent new paradigm, Vision Permutator [21] makes use of linear projections to separately encode the feature representations along the height and width dimensions. This variant of MLP-like architecture is supposed to readily attain promising results in Computer Vision. Our exploration is motivated by a natural question: *What would happen if we replaced the cruciform receptive field of Vision Permutator with a local receptive field version, in the comparable manner as Swin Transformer did with Vision Transformer?*

To give a simple answer, we proposed Axial Depthwise Convolution module, as a combination of Vision Permutator's and ConvNext's designs (Fig. 2). The mathematical formulation of this operator is expressed as follows:

$$x' = x + \mathrm{DW}_{1\times n}(x) + \mathrm{DW}_{n\times 1}(x) \tag{1}$$

$$y = \mathrm{GELU}\Big(\mathrm{PW}_{C_1\to C_2}\big(\mathrm{BN}\left(x'\right)\big)\Big) \tag{2}$$

where, x is input feature, y is output feature; DW, PW and BN stand for Depthwise Convolution, Pointwise Convolution, and Batch Normalization respectively; $1 \times n$ and $n \times 1$ are the kernel sizes of the convolutions; C_1 and C_2 represent the number of input and output channels of the feature map. In our experiment, $n = 7$. To achieve a minimal and flexible design, we use a unique Pointwise Convolution

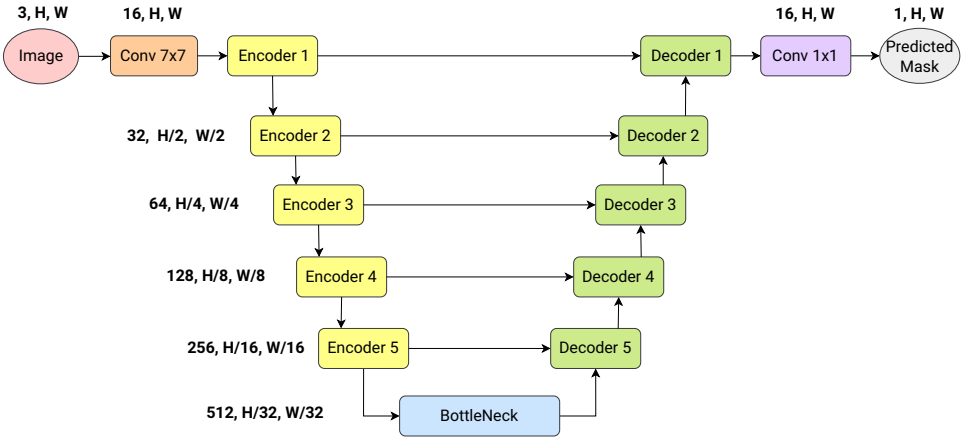


Fig. 1: The proposed U-Lite architecture.

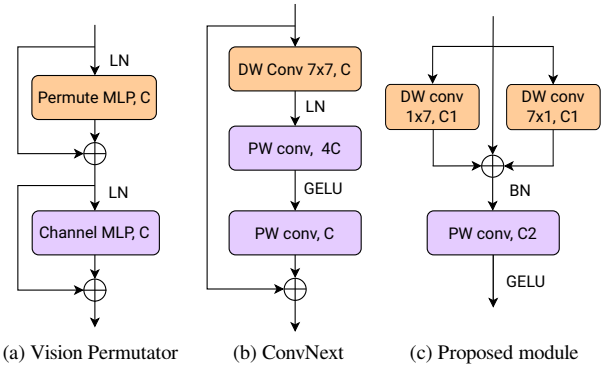


Fig. 2: Architectures of (a) Vision Permutator, (b) ConvNext, and (c) Proposed Axial DW Convolution module. The proposed module is inspired by Vision Permutator's and ConvNext's designs.

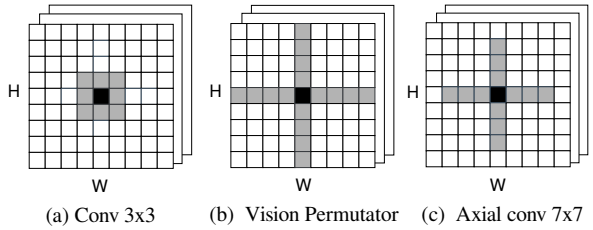


Fig. 3: The receptive field comparison between Convolution 3×3, Vision Permutator, and Axial convolution 7×7. Axial convolution 7×7 offers a large receptive field compared with Convolution 3×3 while using fewer computational parameters than Vision Permutator.

without adding residual connection, allowing us to adjust the number of input channels adaptively.

B. Encoder Block and Decoder Block

The design principles of the encoder and decoder blocks are established as follows:

- Utilize the Depthwise Separable Convolution architecture. This is an important key to successfully building a lightweight model from scratch. Depthwise Separable Convolution delivers almost the same performance as traditional convolution while using fewer parameters, therefore reducing computational complexity, and making the model more compact.

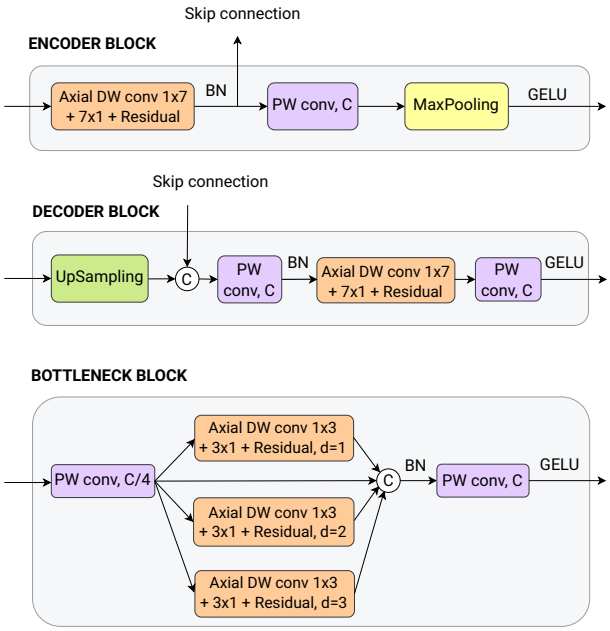


Fig. 4: Encoder, decoder, and bottleneck blocks. Designed based on Depthwise Separable Convolution concept. Each block adopts one Batch Normalization layer and ends with a GELU activation function.

- Limit the use of unnecessary operators. Simply using normal MaxPooling and UpSampling layers. There is no need for a high-parameter-consuming operator such as Transposed Convolution. A Pointwise Convolution operator can play two roles simultaneously: encoding features along the depth of the feature map while flexibly changing the number of input channels.
- Each encoder and decoder block adopts one Batch Normalization layer and ends with a GELU activation function. We have made a performance comparison between Batch Normalization and Layer Normalization but there is not much difference. GELU is applied since it is proven to give an improvement in accuracy when using GELU compared to ReLU, and ELU.

The successful encoder and decoder structures of U-Lite are depicted in Fig. 4.

C. Bottleneck Block

To further boost the performance of U-Lite, we apply Axial Dilated Depthwise Convolutions with kernel size $n = 3$ to the bottleneck block (Fig. 4). The applied dilation rates are $d = 1, 2, 3$. We use axial dilated convolutions with a kernel of size 3 for two reasons: 1) a kernel of size 3 is more suitable with the spatial shape of bottom-layer features, where the height and width of such features are reduced multiple times, and 2) it gives a better performance when using dilated convolutions with different rates to capture multi-spatial representations of high-level features at the bottom stage. To further reduce the number of learnable parameters, a pointwise convolution layer is adopted at the beginning of the bottleneck block. This helps scale down the channel dimension of the last feature layer before feeding them to the Axial Dilated Depthwise Convolution mechanism.

IV. Experiments

A. Implementation Detail

Dataset. To evaluate the performance in the segmentation task, we trained our proposed model on three datasets. The first dataset, ISIC 2018, contained 2594 images of skin lesions. We used 2334 images for training and 260 images for testing. The second dataset, DSB 2018, included 617 images of cell nuclei. We split this dataset into 80% for training and 20% for testing. The third dataset, GlaS, consisted of 165 images of colorectal adenocarcinoma, of which 85 were used for training and 80 for testing. In all the experiments, the images were resized to 256×256.

Training strategy. We conducted U-Lite on the PyTorch framework running on NVIDIA Tesla T4 GPU with 15GB of memory. During the training phase with 200 epochs, training data were augmented by several techniques including random rotation, horizontal flip, and vertical flip. We employed Adam optimization [25] algorithm with the initial learning rate $1e-3$ and used the Dice Loss function, which is defined as:

$$\mathscr{L}_{\text{Dice}}(G,P) = 1 - \frac{2\sum_{i=1}^{N} G_i P_i}{\sum_{i=1}^{N} (G_i + P_i)} \tag{3}$$

where N is the total number of pixels of the image, $G_i \in \{0,1\}$ and $P_i \in (0,1)$ present the value of the i-th pixel of the ground truth and the predicted mask, respectively.

Evaluation metric. We utilized Dice and IoU, two typical metrics for semantic segmentation, to quantitatively measure the model performance. These metrics evaluate the overlap between the ground truth and the predicted mask of an image by the following formulas:

$$\text{Dice} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

where TP, FP and FN stand for True Positives, False Positives and False Negatives, respectively.

B. Representative Results

We compare the segmentation quality of U-Lite and UNeXt [16] on three medical datasets: ISIC 2018, DSB 2018, and GlaS. Fig. 5, Fig. 6, and Fig. 7 show some representative sesults of testing images and their predicted masks by both models, along with the ground truths. U-Lite provides more accurate results than UNeXt, as highlighted by the red circles in the last two columns of each figure. These circles indicate the areas where UNeXt failed to segment correctly.

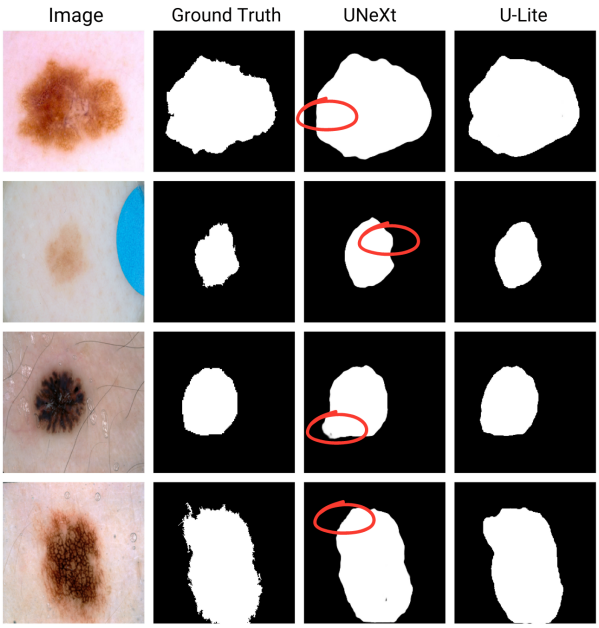


Fig. 5: Comparisons between UNeXt and U-Lite on ISIC 2018 dataset.

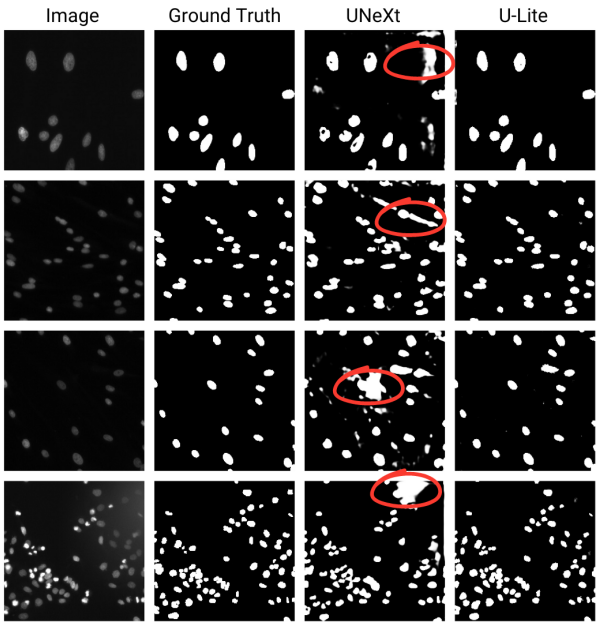


Fig. 6: Comparisons between UNeXt and U-Lite on DSB 2018 dataset.

C. Comparative Results

To prove the outstanding performance of the proposed model, we compare the performance indicators of U-Lite with other previously proposed models in the medical image segmentation task.

As shown in Table I, for the ISIC 2018 dataset, U-Lite achieves 90.39% Dice and 83.63% IoU, with the Dice

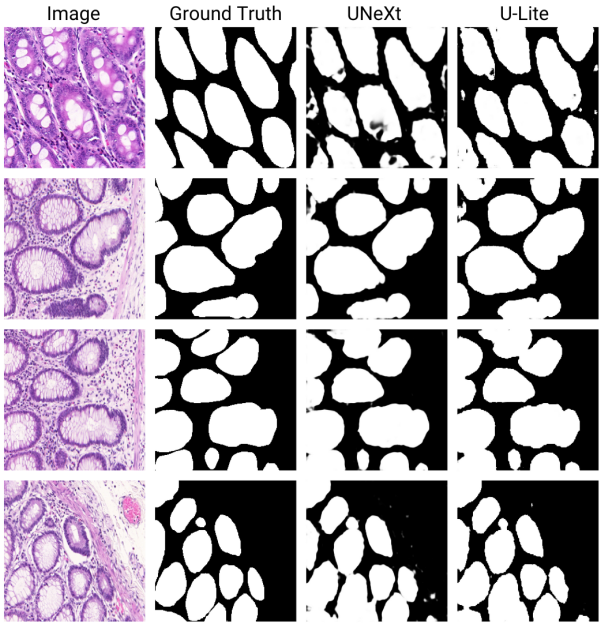


Fig. 7: Comparisons between UNeXt and U-Lite on GlaS dataset.

performing 2.02% higher than U-Net (a traditional CNNs model) and 0.53% higher than ConvUNeXt (a recent high-performing CNN-based model). The total parameters of the U-Lite model are small, only 878K parameters, 35 times less than U-Net. The number of parameters of U-Lite is also smaller than that of some recently proposed lightweight models: ConvUNeXt with 3.5M parameters, Unet++ with 9.2M parameters, and UNeXt with 1.47M parameters.

Table II and Table III evaluate the performance of U-Lite on DSB 2018 and GlaS datasets, respectively, compared to previously proposed models. On DSB 2018, our proposed model achieves 91.37% Dice and 83.63% IoU. Meanwhile, on GlaS, these indices are 86.65% Dice and 77.84% IoU. These results are all higher than the other models mentioned in the tables.

As shown in tables I, II, and III, our proposed model occupies only 3.5MB of memory for storage, much smaller than the state-of-the-art models. Thus, it can be suitable for real-time applications on mobile devices.

In the experiment, we also compared testing times per image on the CPU between U-Lite and previously proposed models. It is observed that UNeXt has the fastest processing speed, followed by U-Lite. Meanwhile, other models all give longer testing times than our model. Although U-Lite is not as fast as UNeXt, it provides more acceptable results when we prioritize the accuracy and compactness of the model, especially for medical image segmentation tasks.

V. Discussion

A. Analysis on the Axial Depthwise Convolution operator

To evaluate the performance of the proposed Axial Depthwise Convolution module, we replace the Axial Depthwise Convolution operator, in Fig. 2, with the conventional Depthwise Convolution operator and experiment with many sizes of kernel to compare their performances. The results are shown in Table IV.

Surprisingly, Axial Depthwise Convolution still performs better and again shows simplicity in parameter configuration

TABLE I: QUANTITATIVE EVALUATION RESULTS ON ISIC 2018 DATASET COMPARED TO PREVIOUSLY PROPOSED MODELS

Model	Params	Memory size	Testing time	Dice	IoU
U-Net [1]	31.2M	124.1MB	927ms	88.60	81.58
Attention Unet [5]	31.4M	139.5MB	1482ms	88.66	81.68
Unet++ [2]	9.2M	36.7MB	898ms	87.78	80.72
ConvUNeXt [22]	3.5M	14.0MB	543ms	89.81	83.19
UNeXt [16]	1.47M	5.9MB	**61ms**	89.56	82.97
U-Lite (Ours)	**878K**	**3.5MB**	140ms	**90.39**	**83.63**

TABLE II: QUANTITATIVE EVALUATION RESULTS ON DATA SCIENCE BOWL 2018 DATASET COMPARED TO PREVIOUSLY PROPOSED MODELS

Model	Params	Memory size	Testing time	Dice	IoU
U-Net [1]	7.9M	31.4MB	366ms	88.93	82.57
Unet++ [2]	9.2M	36.7MB	898ms	88.96	82.77
ConvUNeXt [22]	3.5M	14.0MB	543ms	89.09	83.64
UNeXt [16]	1.47M	5.9MB	**61ms**	88.13	82.15
U-Lite (Ours)	**878K**	**3.5MB**	140ms	**91.37**	**84.86**

TABLE III: QUANTITATIVE EVALUATION RESULTS ON GLAS DATASET COMPARED TO PREVIOUSLY PROPOSED MODELS

Model	Params	Memory size	Testing time	Dice	IoU
U-Net [1]	7.9M	31.4MB	366ms	77.78	65.34
Unet++ [2]	9.2M	36.7MB	898ms	78.03	65.55
UNeXt [16]	1.47M	5.9MB	**61ms**	86.49	77.77
U-Lite (Ours)	**878K**	**3.5MB**	140ms	**86.85**	**77.84**

TABLE IV: PERFORMANCE COMPARISON BETWEEN DEPTHWISE CONVOLUTION AND AXIAL DEPTHWISE CONVOLUTION WITH DIFFERENT SIZES OF KERNEL

Dataset	Operator	Metric	n = 3	n = 5	n = 7
ISIC 2018	DW Conv	Dice	89.71	89.91	89.89
		IoU	82.92	83.19	83.48
	Axial DW Conv	Dice	90.04	90.31	**90.39**
		IoU	83.35	**83.68**	83.63
DSB 2018	DW Conv	Dice	90.95	91.01	91.27
		IoU	84.30	84.36	84.74
	Axial DW Conv	Dice	91.21	91.34	**91.37**
		IoU	84.60	84.77	**84.86**
GlaS	DW Conv	Dice	85.88	85.87	86.14
		IoU	76.45	76.53	76.86
	Axial DW Conv	Dice	86.00	85.80	**86.85**
		IoU	77.20	76.32	**77.84**

and computational complexity. This result can be theoretically explained by the following reasons: 1) By using a horizontal filter along with a vertical filter instead of a unique Depthwise Convolution operator, the model is more focused on which interested spatial aspect to aggregate information. 2) More flexible. We give an example, for the Sobel edge detection problem, the filter $n \times 1$ can act as a vertical edge detector while the $1 \times n$ can act as a horizontal edge detector. 3) Without losing the effect on the receptive field. On the spatial area, if the pixel (a,b) wants to aggregate information from pixel (x,y), it can contact the others through the middle point (a,y) or (x,b), based on the axial mechanism.

To our knowledge, the main factors that make our Axial Depthwise Convolution module differ from the previously

proposed architecture of InceptionNet [17] are that, firstly, we use the Depthwise Separable Convolution concept to reduce the number of parameters, secondly, we take the sum of two convolutions with kernels $1 \times n$ and $n \times 1$ which leads to the cruciform receptive field (while they are connected in series in InceptionNet), and lastly, a residual connection is added through the depthwise convolution layer to retain the important information of feature maps.

B. Effect of Axial Dilated Depthwise Convolutions to the bottleneck

Table V compares the performance of U-Lite before and after integrating the model with the Axial Dilated Depthwise Convolution (ADDC) mechanism. In the case of not using ADDC, we adopt the Axial Depthwise Convolution module with the kernel size $n = 7$ only to the bottleneck of U-Lite as an alternative. Overall, this model contains 817K parameters. It is observed that ADDC boosts the performance of U-Lite in both Dice and IoU coefficients on all three datasets while not overly increasing the number of learnable parameters.

TABLE V: PERFORMANCES WHEN USING (w) AND WITHOUT USING (w/o) THE AXIAL DILATED DEPTHWISE CONVOLUTIONS (ADDC) IN THE BOTTLENECK

Dataset	ADDC	Dice	IoU
ISIC 2018	w/o	89.96	83.21
	w	**90.39**	**83.63**
DSB 2018	w/o	91.21	84.60
	w	**91.37**	**84.86**
GlaS	w/o	84.54	74.82
	w	**86.85**	**77.84**

VI. Conclusion

In this paper, we proposed a lightweight CNN-based architecture for the medical segmentation task. The suggested model U-Lite overcomes the limitation of many complex models that need a high number of parameters and thus results in much memory size and low inference speed. To this end, we proposed the Axial Depthwise Convolution module as a combination of Vision Permutator and ConvNext blocks. We further boost the performance of U-Lite by integrating an Axial Dilated Depthwise Convolution mechanism to the bottleneck of the model. Our experiments on various medical datasets show that U-Lite has a smaller number of parameters and costs less in computational complexity while also performing promising results.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2021.34.

References

[1] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *International Conference On Medical Image Computing And Computer-assisted Intervention*. pp. 234-241 (2015)

[2] Zhou, Z., Siddiquee, M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning In Medical Image Analysis And Multimodal Learning For Clinical Decision Support*. pp. 3-11 (2018)

[3] Jha, Debesh, et al. "Resunet++: An advanced architecture for medical image segmentation." *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019.

[4] Jha, D., Riegler, M., Johansen, D., Halvorsen, P. & Johansen, H. Doubleu-net: A deep convolutional neural network for medical image segmentation. *2020 IEEE 33rd International Symposium On Computer-based Medical Systems (CBMS)*. pp. 558-564 (2020)

[5] Oktay, Ozan, et al. "Attention u-net: Learning where to look for the pancreas." *arXiv preprint arXiv:1804.03999* (2018).

[6] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[7] Chen, Jieneng, et al. "Transunet: Transformers make strong encoders for medical image segmentation." *arXiv preprint arXiv:2102.04306* (2021).

[8] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 568-578).

[9] Trung-Kien Le, Hoang-Minh-Quang Le, Thi-Thao Tran, and Van-Truong Pham. "MSMA-Net: A Multi-scale Multidirectional Adaptation Network for Polyp Segmentation." *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2022.

[10] Dong, Bo, et al. "Polyp-pvt: Polyp segmentation with pyramid vision transformers." *arXiv preprint arXiv:2108.06932* (2021).

[11] Hong-Phuc Lai, Thi-Thao Tran, and Van-Truong Pham. "Axial attention mlp-mixer: A new architecture for image segmentation." *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*. IEEE, 2022.

[12] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J. & Others MLP-Mixer: An all-mlp architecture for vision. *Advances In Neural Information Processing Systems*. 34 (2021)

[13] Jing, Junfeng, et al. "Mobile-Unet: An efficient convolutional neural network for fabric defect detection." *Textile Research Journal 92.1-2* (2022): 30-42.

[14] Shan, Tong, et al. "DSCA-Net: A depthwise separable convolutional neural network with attention mechanism for medical image segmentation." *Mathematical Biosciences and Engineering 20.1* (2023): 365-382.

[15] Valanarasu, J., Oza, P., Hacihaliloglu, I. & Patel, V. Medical transformer: Gated axial-attention for medical image segmentation. *International Conference On Medical Image Computing And Computer-Assisted Intervention*. pp. 36-46 (2021)

[16] Valanarasu, Jeya Maria Jose, and Vishal M. Patel. "Unext: Mlp-based rapid medical image segmentation network." *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. Cham: Springer Nature Switzerland, 2022.*

[17] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[18] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

[19] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[20] Liu, Zhuang, et al. "A convnet for the 2020s." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[21] Hou, Qibin, et al. "Vision permutator: A permutable mlp-like architecture for visual recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence 45.1* (2022): 1328-1334.

[22] Han, Z., Jian, M., & Wang, G. G. (2022). ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowledge-Based Systems, 253, 109512.*

[23] Gu, Ran, et al. "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation." *IEEE transactions on medical imaging 40.2 (2020)*: 699-711.

[24] Trockman, Asher, and J. Zico Kolter. "Patches are all you need?." *arXiv preprint arXiv:2201.09792* (2022).

[25] Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*. (2014)

Imaging Ultrasound Scattering Targets using Density-Enhanced Chaotic Compressive Sampling

Luong Thi Theu
Hoa Binh University, Hanoi, Vietnam
lttheu@daihochoabinh.edu.vn

Tran Quang Huy*
Faculty of Physics Hanoi Pedagogical University 2, Hanoi, Vietnam
tranquanghuy@hpu2.edu.vn

Tran Thi Thuy Quynh
Faculty of Electronics & Telecommunications, VNU University of Engineering & Technology, Hanoi, Vietnam
quynhttt@vnu.edu.vn

Duc-Tan Tran*
Faculty of Electrical and Electronic Engineering, Phenikaa University, Hanoi 12116, Vietnam
tan.tranduc@phenikaa-uni.edu.vn

***Abstract*—The ultrasound imaging method is a highly promising technique used in various fields. The utilization of ultrasound technique's backscattering could also allow humans to observe objects as small as the size of the incident wavelength. However, the employing backscattering approach requires high computational calculations and poses a significant obstacle to the production of equipment using the ultrasonic method. In this work, we utilized the pseudo-random compression sampling method in the extended DBIM model, which takes tissue density factor. This is necessary because tissue density provides quantitative information about soft tissue and serves as an image contrast source. This method combines pseudo-random measurement and compressed sensing data recovery technique. The simulation results have proved the effectiveness of the proposed CCS-DBIM method. In terms of imaging quality, at compression ratios as low as 0.5, the large difference in imaging quality is evident. After 5 iterations, the normalization error of the DBIM and CCS-DBIM methods was 0.8828 and 0.5281 respectively, i.e. a 40% reduction in normalization errors when the number of measurements was only half of the variables. The disadvantage of the proposed CCS-DBIM method is that the imaging time is significantly greater than that of the DBIM method. However, with today's parallel computing technology, large amounts of data and computational time will no longer be a big problem.**

Keywords—Ultrasonic imaging, distorted Born iterative method, pseudo-random, scattered data, Born approximation.

I. INTRODUCTION

To construct quantitative ultrasound images, the simplest method is to extract information directly from the B-mode image. The image texture analysis was performed using a number of techniques, such as first order and second parameters (mean value, standard deviation, entropy, run-length parameters) [1], contour-based statistical parameters [2], and wavelet analysis [3]. However, these quantitative parameters are not completely independent of imaging systems, and therefore, these solutions have certain limitations and are not applied in current biomedical devices. For quantitative ultrasound imaging, two of the most interesting techniques are ultrasound tomography and backscatter coefficient analysis. Ultrasound tomography is based on backscattering, where a significant portion of energy scatters in multiple directions when the acoustic wave encounters an inhomogeneous field. A variety of measurements in the scattering medium can be conducted using the wave equation. The ultrasonic tomography problem consists of the following tasks: measuring and determining the change of sound parameters in the scattering medium such as variations in sound velocity, density, attenuation, intensity and other parameters. Nevertheless, the backscattering technique used has certain limitations, which is reason tomography ultrasonic devices have not been as successful as other biomedical imaging modalities such as X-ray imaging, nuclear-based imaging, or harmonic imaging. One limitation is that backscattering encounters some problems when restoring high-contrast objects due to convergence effects. This limitation has restricted the use of backscatter imaging to breast tissue imaging, as breast tissue consists mainly of soft tissue. Some independent studies have explored imaging of bone [4], which may expand the application range of the backscatter technique to some extent. Another requirement is that backscattered data must be acquired from multiple angles ranging from 0° to 360° to achieve the best image quality. This is another reason why most backscatter research focuses on breast tissue imaging. Some approaches are proposed in order to improve imaging quality in ultrasound tomography [5], [6], [7], [8], [9].

In 2006, the compressed sensing (CS) technique was developed by works [10], [11], [12], [13], which analyzes and restores signals at a much lower sampling rate than the standard Nyquist sampling rate, but ensures accuracy. That proves that, if compression sampling is used, only a much lower amount of data is required than conventional sampling techniques, but ensures accurate signal recovery. Therefore, the amount of data required for collection and analysis will be significantly reduced, leading to a high increase in processing speed and minimizing system complexity. However, with compression sampling techniques, deploying a hardware random number generator (HRNG) or pseudorandom number generator (PRNG) to generate a random sequence will be very difficult in practice. Therefore, we propose a pseudo-random compression sampling solution to recover images in the DBIM model [14] taking into account the density parameters of the imaging object. The hardware of the pseudo-random

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

transmitter is the basic electronic circuit of the nonlinear system and is much simpler than the random transmitter [15].

II. Pseudo-random compression sampling

In traditional compression sampling techniques, measuring matrix $\bar{\Phi}$ is random. However, it can be noticed that the method of deployment on the hardware of pseudorandom systems is usually simpler than that of random systems. Therefore in this study, the matrix $\bar{\Phi}$ built using pseudo-random methods. We consider a dynamic pseudo-stochastic system of pseudo-like nature. Since the operational properties of the system are very sensitive to initial conditions, the output of the system is pseudo-random. This is presented in the study [16]. This technique ensures accuracy in recovery similar to random compression sampling, as noted in [17], [18]. In this work, we look at a logistic map based on a dynamic structure that generates a very simple pseudorandom sequence that behaves like a Gauss chain [16]. To build $\bar{\Phi}$, we use the following logical chain:

$$Hq(n+1) = \rho q(n)\big(1 - q(n)\big), \tag{1}$$

Then, convert the string using the logit transform to get the Gauss chain as follows:

$$q_G(n) = \ln\left[\frac{q(n)}{1 - q(n)}\right]. \tag{2}$$

When control parameters $\rho = 4$, the dynamicity described by Eq. (1) is pseudo-random; this means that the dynamics of Eq. (1) are very sensitive to initial conditions $q(0)$. A slight variation of $q(0)$ will cause big change $q(n)$ in a relatively short period of time. Details about the construction $\bar{\Phi}$ from $q_G(n)$ can be referenced from the work [16]. Finally, we restore the image by using a sparse approximation algorithm.

III. Proposed Distorted Born Iterative method combined with pseudo-random compression sampling technique

Suppose we have $N_t \times N_r$ probes, where N_t is the transmitter number and N_r is the receiver number (therefore, we have $N_t \times N_r$ measurements). We need to restore the ideal target function, it is divided into N pixels vertically and horizontally, i.e. N^2 is variables.

If $N_t \times N_r$ is less than N^2, i.e. the number of equations is less than the number of variables, this is an "under-determinded problem". In principle, when the number of equations is less than the number of variables, we cannot get the correct solution; we need to use approximation methods. When $N_t \times N_r$ is much smaller than N^2, we call this a case of collecting little information. We define that the ratio between $N_t \times N_r$ and N^2 is the compression ratio r:

$$r = \frac{N_t N_r}{N^2}. \tag{3}$$

It is clear that in the problem of collecting small amounts of scattered information, we have $r < 1$. When r is small, the measurement system will be simple and the data collected will be small; Therefore, system setup and calculations will be less complicated. Therefore, we focus on the case of the smallest value of r while ensuring acceptable image recovery quality.

With the traditional DBIM method [14], where the ideal target function is the input and the recovery target function is the output. This method uses measurement parameters with evenly arranged receiver and transmitter positions, and applies the Tikhonov approach to restore the ideal target function. With the CCS-DBIM proposed method, the measurement configuration is applied with uniformly distributed transmitter positions and pseudo-randomly distributed receiver positions. Then, the l_1 optimization method is used. The l_1-LSP method solves the least squares problems based on the l_1 norm using the truncated Newton interior-point method and is expressed as follows:

$$min\|\bar{A}\bar{w} - \bar{u}\|^2 + \zeta\|\bar{w}\|_1, \tag{4}$$

in which $\bar{A}$ is a data matrix in dense or sparse form with n columns and m lines; $\bar{w}$ is a vector of length n; $\bar{u}$ is a vector of length m. In the CCS-DBIM method, $\bar{A}$, $\bar{w}$, $\bar{u}$ are $\overline{M_t}$, $\Delta\bar{O}$, and $\Delta\bar{p}^{sc}{}_t$, respectively. Note that, the size of $\overline{M_t}$ depends on the number of transmitters (N_t), receiver number (N_r) and the number of variables (N^2). Therefore, it will change in different scenarios. In the CCS-DBIM method, the l_1 norm is represented as follows:

$$\Delta\bar{O} = \arg\min_{\Delta\bar{T}} \left\|\Delta\bar{p}^{sc}{}_t - \overline{M_t}\Delta\bar{O}\right\|_2^2 + \zeta\|\Delta\bar{O}\|_1 , \tag{5}$$

where ζ is the regularization parameter in the problem l_1, the object function $\bar{O}$ is defined as follows:

$$\bar{O}(\mathrm{r}) = \begin{cases} \omega^2\left(\frac{1}{v^2} - \frac{1}{v_0^2}\right) - \rho^{1/2}(r)\nabla^2\rho^{-1/2}(r) & if\ r \le R \\ 0 & if\ r > R \end{cases} \tag{6}$$

where R is the radius of the target, $\rho(r)$ is the target density, ω is the frequency of the incident signal ($\omega = 2\pi f$), v_0 is the sound speed difference of the background medium, and v is the target medium. In this paper, the pseudo-random compression sampling method is utilized in the extended DBIM model with taking account of tissue density parameter.

The normalized error between the obtained image and the initial image at every iteration is determined by using this formula:

$$\varepsilon = \frac{1}{V \times W}\sum_{i=1}^{V}\sum_{j=1}^{W}\frac{\left|m_{ij} - \hat{m}_{ij}\right|}{\left|m_{ij}\right|} \tag{7}$$

where m is a $V \times W$ initial image (i.e. hypothetical target function) and $\hat{m}$ is the obtained image.

IV. Numerical simulation

Frequency $f = 0.64$ MHz; total loops $N_{sum} = 5$; N = 10 (i.e., the number of variables equal to $N^2 = 10 \times 10 = 100$); diameter of the object = 7.3 mm; the variation in the sound contrast is 7%; signal-to-noise ratio (SNR) = 20 dB; the distance from the transmitter and receiver to the center of the object is 100 mm. The number of measurements (N_m) is defined as the product of

the number of transmitters (N_t) and receivers (N_r), i.e. $N_m = N_t \times N_r$. The number of measurements in the selected simulation scenarios is correlated with the number of variables as shown in Eq. (3).

Fig. 1 shows the measurement configuration of the traditional DBIM method with a measurement number of N_m = 100 (transmitter number of 20 is indicated by a red star, receiver number of 5 is indicated by a blue rectangle). The positions of the transmitter and receiver are evenly spaced on the measuring system. Similarly, Fig. 2 shows the measurement configuration of the proposed CCS-DBIM method with the same number of measurements, however it differs only in that the positions of the receivers are arranged pseudo-randomly on the measurement system. This is done to ensure that recovery is possible with the sparse and random signal recovery algorithm l_1. Fig. 3 shows the ideal object that needs to be reconstructed.

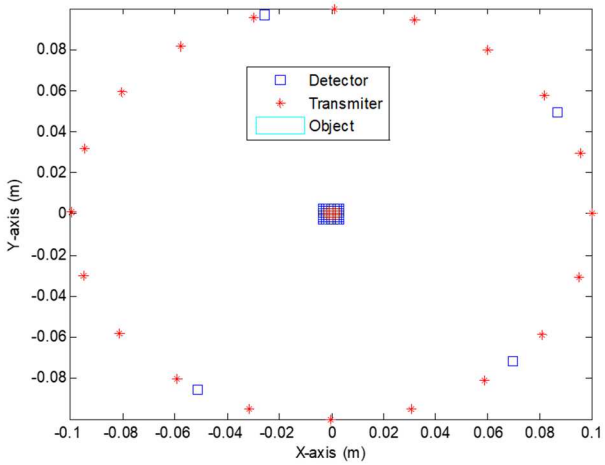


Figure 1. Measurement configuration of traditional DBIM method (N_m = 100)

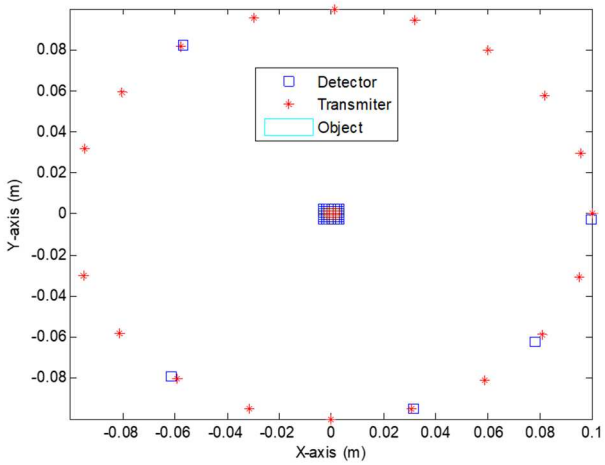


Figure 2. Measurement configuration of the proposed CCS-DBIM method (N_m = 100)

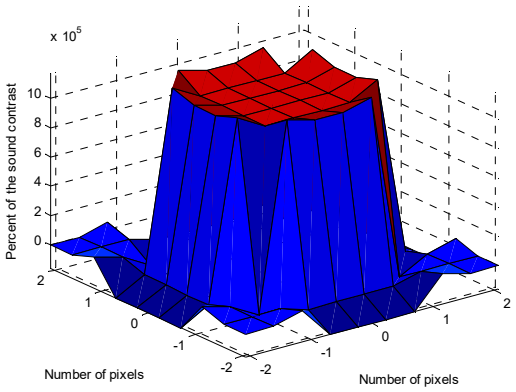


Figure 3. Ideal image that needs to be reconstructed

The image restored using the DBIM and CCS-DBIM methods in case the compression ratios r = 1, r = 0.8 and r = 0.5 in loop 1 are shown in Figure 4, 6 and 8, respectively, we see that, when the compression ratio is large (r = 1), i.e. the number of measurements equals to the number of variables, the recovery results in the first loop with the DBIM method are also quite good. However, when the number of measurements is much reduced, i.e. when the compression ratio is small (r = 0.5), the recovery results in the first loop with the DBIM method are quite poor. In particular, with the CCS-DBIM method in the first loop as well as in all other loops, the recovery results are always significantly better than the DBIM method, especially right at the first loop. In addition, the noise handling of the CCS-DBIM method is done quite well, we see that background noise in the CCS-DBIM method is significantly reduced compared to the DBIM method at the same compression ratio, and at different compression ratios, we also observe a sharp reduction in background noise.

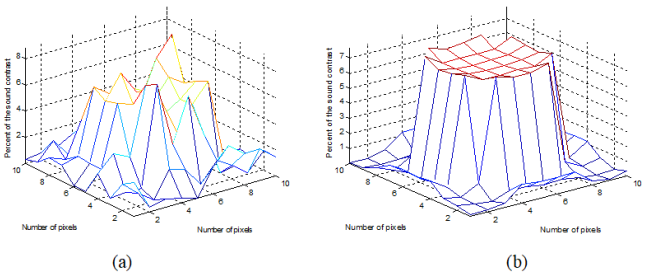


Figure 4. Recovery image results in loop 1 of DBIM using Tikhonov regularization (a) and CS-DBIM using l_1 norm (b) methods in the case of N_m = 100

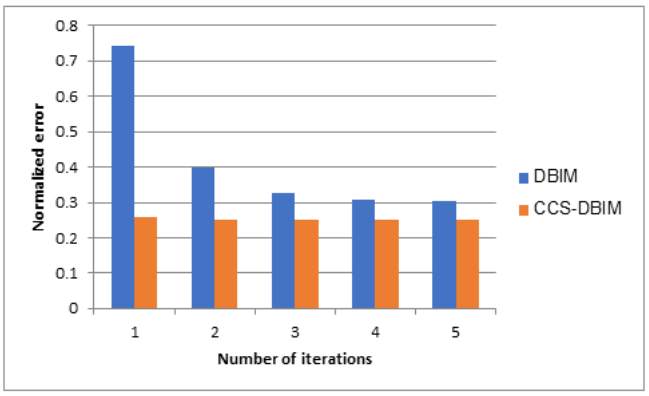
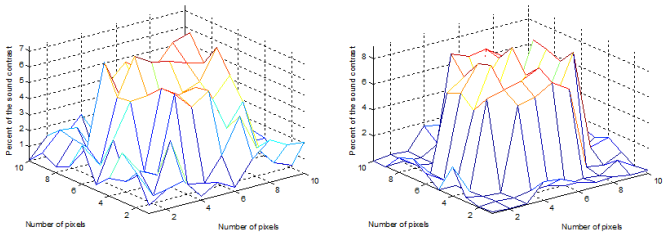


Figure 5. The normalization error result corresponds to the loop number of DBIM and CCS-DBIM methods in the case of $r = 1$



(a) (b)

Figure 6. Recovery image results of DBIM (a) and CCS-DBIM (b) methods in the case of $r = 0.8$

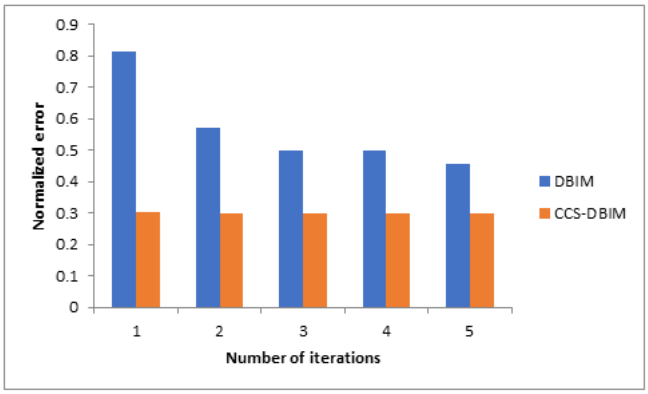
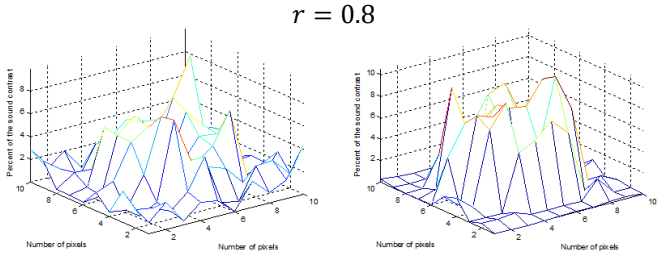


Figure 7. The normalization error result corresponds to the loop number of DBIM and CCS-DBIM methods in the case of $r = 0.8$



(a) (b)

Figure 8. Recovery image results of DBIM (a) and CCS-DBIM (b) methods in the case of $r = 0.5$

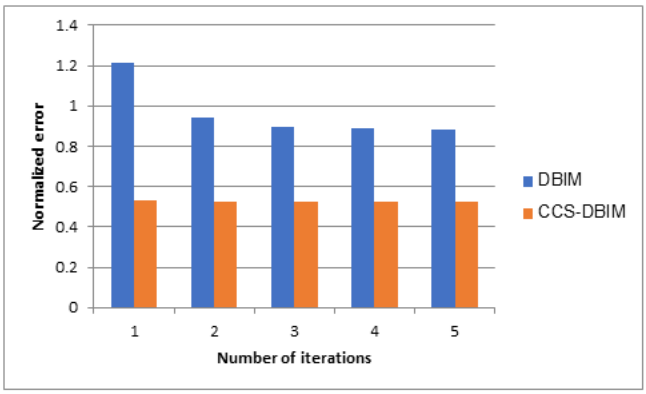


Figure 9. The normalization error result corresponds to the loop number of DBIM and CCS-DBIM methods in the case of $r = 0.5$

Figure 7 shows the normalization error result corresponding to the loop number of DBIM and CCS-DBIM methods in the case of $r = 0.8$. Although the normalized error seems not to be saturated after 5 iterations, we want to observe the convergence rate of the two methods in iterations. The CCS-DBIM method reaches saturation very quickly after the first 1-2 iterations. However, the DBIM method after 5 iterations still has not reached saturation. This shows that the convergence speed of the CCS-DBIM is much better than that of the DBIM.

Figure 10 shows the normalization error results corresponding to the different compression ratios of DBIM and CCS-DBIM methods. It can be seen that, when the compression ratio decreases, normalization errors tend to increase, this is completely normal, because the compression ratio decreases in proportion to the number of measurements decreases, so the collected data will decrease and consequently the image recovery will be reduced. However, we see that, at smaller compression ratios, especially at a compression ratio of 0.5, the greater the difference in error intervals in the 2 methods, specifically DBIM and CCS-DBIM methods have normalization errors after 5 loops of 0.8828 and 0.5281, i.e. the normalized method for the proposed method decreases by 40% when the number of measurements is half of the variables. In the case of a compression ratio of 1, the normalization error of DBIM and CCS-DBIM methods after 5 loops is 0.3024 and 0.2517, i.e. the normalization error of the proposed method decreases by 16.7% when the number of measurements is equal to the number of variables. This makes sense in practice, since when we need to create a high-resolution image of an object, i.e. a large number of variables, meanwhile, our hardware configuration is limited, so restoring the object at small compression ratios gives as accurate a result as possible.

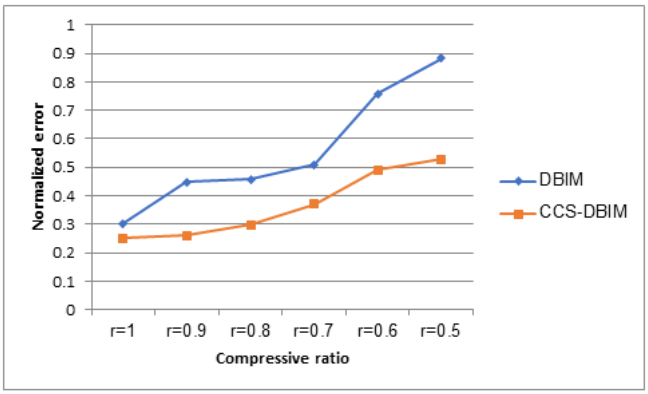


Figure 10. Normalization error results corresponding to different compression ratios of DBIM and CCS-DBIM methods

Figure 11 shows the imaging time results corresponding to the different compression ratios of DBIM and CCS-DBIM methods. We see that, the smaller the compression ratio, the smaller the imaging time, which can be explained by the fact that, when the compression ratio is small, i.e. the number of measurements is small, the amount of information to be processed and calculated will be less, therefore, the time to create the image will be less. We realize that the disadvantage of the CCS-DBIM method is that the imaging time is significantly larger than the traditional DBIM method. However, in Figures from 4 to 10, we see that the CCS-DBIM method recovers the object very well from the first loop, and the normalization error decreases after the first few loops, and then it quickly reaches a threshold and does not decrease anymore. That means, in actual image creation, with the CCS-DBIM method we only need to restore the image in the first few loops (i.e. do not have to create images in all 5 loops as with the DBIM method), if so, then the imaging time will be comparable to the DBIM method. Moreover, with today's parallel computing technology, large amounts of data and computational time will no longer be a big problem.

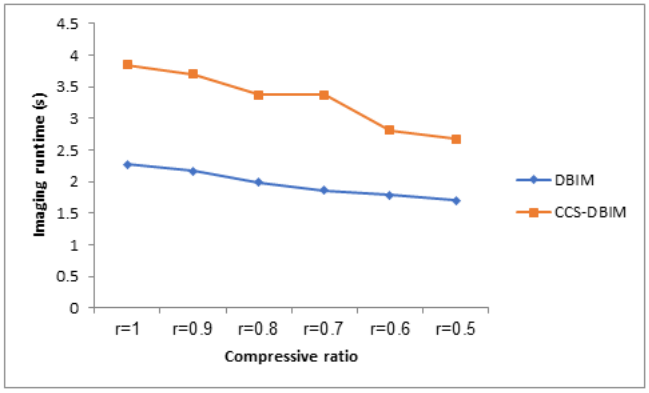


Figure 11. Imaging time results corresponding to different compression ratios of DBIM and CCS-DBIM methods

The complexity of the reconstruction system depends on the number of pixels $N \times N$, the number of iterations N_{iter}, the number of transmitters N_t and receivers N_r. With the proposed CCS-DBIM, the number of measurements and number of iterations required to reconstruct an image is greatly reduced. Therefore, this method provides much reduced computational complexity.

V. Conclusions

This paper applied pseudo-random compression sampling in an extended DBIM model that takes into account the density factor for imaging. The proposed CCS-DBIM method is easier to implement and more effective than previous methods. To prove the effectiveness of this method, many hypothetical simulation scenarios were performed. The results showed that a clear difference in imaging quality was expressed in compression ratios as low as 0.5. After 5 iterations, the CCS-DBIM method reduced normalization errors by 40% compared to the DBIM method, and with only half of the measurements. However, one of the disadvantages of the CCS-DBIM method is that the imaging time is significantly greater than that of the DBIM method. However, with parallel computing technology, the problem of large amounts of data and computational time can be solved.

References

[1] V.Murmis, J.Gisvold, T.Kinter, and J.Greenleaf, "Texture analysis of ultrasound B-scans to aid diagnosis of cancerous lesions in the breast," in Proceedings of the IEEE Ultrasonics Symposium, pp. 839–842, 1988.

[2] Dutt and J.Greenleaf, "Ultrasound echo envelope analysis using a homodyned K distribution signal model," Ultrasonic Imaging, vol. 16, no. 4, pp. 265–287, October 1994.

[3] A.Mojsilovic, MN. Popovic, S.Markovic, and M.Krstic, "Characterization of visually similar diffuse diseases from B-scan liver images using nonseparable wavelet transform," IEEE Transactions on Medical Imaging, vol. 17, no. 4, pp. 541–549, August 1998.

[4] O.Lasaygues, F.Franceschini, R.Guillermin, J.-P.Lefebvre, N.Salaud, and P.Petit, "Two-dimensional ultrasonic computed tomography of growing bones," in IEEE Ultrasonics Symposium, vol. 1, 2007, pp. 1816–1819.

[5] P.Lu, & P.Kosmas, "Three-dimensional microwave head imaging with gpu-based fdtd and the dbim method," Sensors, vol. 22, no. 7, pp. 2691, 2022.

[6] L. T.Theu, T.Quang-Huy, T.Duc-Nghia, V. K.Solanki, T.Duc-Tan, & J. M. R.Tavares, "High Sound-Contrast Inverse Scattering by MR-MF-DBIM Scheme," Electronics, vol. 11, no. 19, pp. 3203, 2022.

[7] T. Quang-Huy, K. T.Nguyen, P. T.Doan, & D. T.Tran, "Interpolated hybrid DBIM approach for enhanced imaging in ultrasound tomography," Research on Biomedical Engineering, pp. 1-12, 2021.

[8] O.Karadima, M.Rahman, I.Sotiriou, N.Ghavami, P.Lu, S.Ahsan, & P.Kosmas, "Experimental validation of microwave tomography with the DBIM-TwIST algorithm for brain stroke detection and classification," Sensors, vol. 20, no. 3, pp. 840, 2020.

[9] P.Lu, J.Corcoles, & P.Kosmas, "Enhanced FEM-based DBIM approach for two-dimensional microwave imaging," IEEE Transactions on Antennas and Propagation, vol. 69, no. 8, pp. 5187-5192, 2020.

[10] Candès, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 489-509, 2006.

[11] E.Candes and J.Romberg, "Sparsity and incoherence in compressive sampling", Inverse Problem, vol. 23, pp. 969, 2007.

[12] E.J.Candes, "The restricted isometry property and its implication for compress sensing", Compte Rendus de l'Academie des Sciences, vol. 346, pp. 589-592, 2008.

[13] E.J.Candes and M.Wakin, "An Introduction To Compressive Sampling", IEEE Signal Processing Magazine, vol. 25, pp. 21-30, 2008.

[14] R.Lavarello, M.Oelze, "Tomographic Reconstruction of Three-Dimensional Volumes Using the Distorted Born Iterative Method",

IEEE Transactions on Medical Imaging, vol. 28, no. 1, pp. 1643-1653, 2009.

[15] Mohammed Bakiri, “Hardware implementation of a pseudo random number generator based on chaotic iteration”, Cryptography and Security [cs.CR]. Université Bourgogne Franche-Comté, 2018. English. ffNNT : 2018UBFCD014ff. fftel-02080150.

[16] J.C. Sprott, “Chaos and time-series analysis”, Oxford University Press, 2003.

[17] N.Linh-Trung, D.Van Phong, Z. M.Hussain, H. T.Huynh, V. L.Morgan, & J. C.Gore, “Compressed sensing using chaos filters”, In 2008 Australasian Telecommunication Networks and Applications Conference, pp. 219-223, 2008.

[18] L. T.Nguyen, M. C.Truong, D. T.Tran, V. H.Le, & N. M.Do, “Chaotic Compressed Sensing and its Application in Magnetic Resonance Imaging”, REV Journal on Electronics and Communications, vol. 3, no. 3-4, pp. 84-92, 2013.

Segmentation and observation of hand rehabilitation exercises by supporting of acceleration signals

Sinh-Huy Nguyen
MITI, AMST, Vietnam
huyns76@gmail.com

Thi-Thu-Hong Le
MITI, AMST, Vietnam
lethithuhong1302@gmail.com

Hoang-Bach Nguyen
MITI, AMST, Vietnam
nhbach2203@gmail.com

Ngoc-Bach Duong
SEEE, HUST, Vietnam
ngocbach99ltv@gmail.com

Hung-Cuong Nguyen
SEEE, HUST, Vietnam
nhcuongit@gmail.com

Chi-Thanh Nguyen
MITI, AMST, Vietnam
thanhnc80@gmail.com

Van-Loi Nguyen
SEEE, HUST, Vietnam
loi.nv182646@sis.hust.edu.vn

Hai Vu
SEEE, HUST, Vietnam
hai.vu@hust.edu.vn

***Abstract*—In recent years, there has been growing interest in using sensors worn on patients' bodies to assist doctors in monitoring and evaluating rehabilitation therapy. In this study, we toward an assisted tool that aids in the rehabilitation assessments of patients with impaired hands. The proposed method automatically spots the patient's exercises using utilizing accelerometers and displays the results on the collected video. This tool helps doctors to monitor and evaluate the patient's progress more concentrated and easily. The proposed method utilizes wavelet signal transformations, and signal segmentation by Sliding Window and Short-time Energy techniques. These techniques are not only simple calculations but also offer promising results. In the experimental results, the exercises, which are set up in real clinical situations, have been segmented with 87% of accuracy and 80% of sensitivity rate. These results confirm a potential solution to improve the efficiency and effectiveness of the rehabilitation process using the automatic assisted application.**

***Index Terms*—wearable sensor, accelerometer, rehabilitation, wavelet transform, sliding window, short-time energy**

I. Introduction

Modern medicine not only aims to treat life-threatening diseases but also to facilitate the recovery of patients' motor functions and their reintegration into society. Thus, rehabilitation is a crucial process for patients following injuries or strokes. For hand rehabilitation, patients frequently perform exercises as per the physician's prescribed regimen. Currently, wearable sensor technologies offer promising prospects for the development of innovative tools to facilitate rehabilitation assessment. These technologies allow for the capture of data from sensors such as cameras and accelerometers, which can then be analyzed and applied to create applications that assist healthcare professionals in tracking and evaluating patients' rehabilitation progress.

Automatic segmentation of the rehabilitation exercises or sessions performed by patients is the first step of an assisted tool. As shown in Fig. 1, within a rehabilitation session, the patients repeatedly practice exercises. Log activities during a session could be weak/normal hands in actions, or idle times where the patients are relaxed or in the exercises under

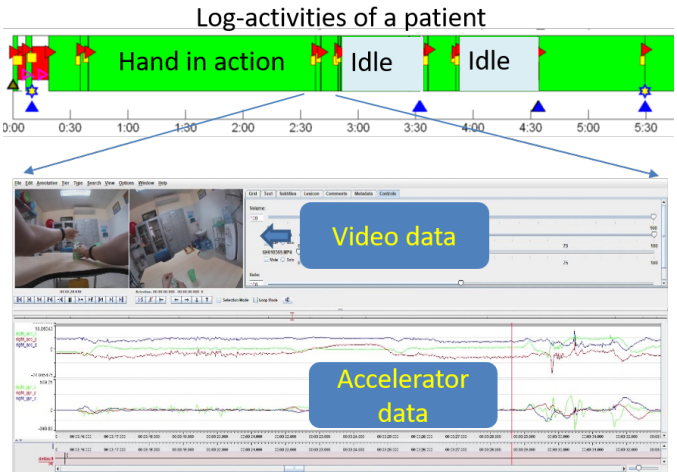


Fig. 1. An illustration of the study's purpose. Upper-panel: Log activities during a rehabilitation session. Lower-panel: The corresponding data collected by wearable devices. A spotting technique helps to segment the period when the patient's hand is in action. The idle times, therefore, can be eliminated.

construction or preparations. A signal processing technique could be applied for spotting these log activities. Based on the log activities during a session, the technical doctors are more convenient to jump and observe a certain exercise. This approach is especially critical in healthcare settings with a high patient load but limited staffing resources.

To address this issue, wearable sensors including accelerometers and image sensors are set up in rehabilitation exercises at a healthcare center. Data collected by wearable sensors are synchronized. To automatically spot each exercise, the acceleration data is analyzed by a series of signal processing techniques. Assessment of the patient's rehabilitation ability is based on the video data of each segmented period. Therefore, the proposed method handles three key issues: determining the start and end times of hand movements during exercise, identifying the hand involved in the exercise, and integrating the detected outcomes into corresponding exercise videos. The evaluation results ensure that the proposed method is able to

978-1-6654-5245-8/23/$31.00 ©2023 IEEE

healthcare professionals to rapidly assess a patient's exercise progress, measure training duration, and compare the strength and dexterity of the patient's weaker and unaffected hands.

II. RELATED WORKS

Applying wearable-sensor techniques for rehabilitation monitoring has received attention in recent years. For instance, the roles of the hand in egocentric videos such as hand detection, segmentation, localization as well as activity recognition, and hand-object manipulations in activities of daily living (ADL) have been widely explored in [1]. Researchers in [2], [3] focus on the video data collected by wearable camera only to measure meaningful information of a patient's hands such as the number and speed of hand-to-object interaction. Combining wearable sensors such as accelerator and video collected simultaneously has not been deeply attempted.

This study demonstrates the importance of accelerometer signals for applications in rehabilitation monitoring. Accelerometer signal processing has been an active research area in recent years due to its widespread use in wearable devices and healthcare applications. One commonly used technique for accelerometer signal processing is wavelet transform, which has been applied in many studies to decompose the signal into different frequency bands and extract time-frequency features. In [4] and [5], the authors used wavelet transform to decompose the accelerometer signal into frequency subbands and extract time-domain features for activity recognition. Accelerometer signal also helps divide the signal into segments based on some criteria such as amplitude, frequency, or time intervals. Segmentation is often used to detect specific activities or events in the signal. For example, Zhou, Bangwen, et al. [6] used segmentation with a multi-probability threshold on accelerometer signals for human activity recognition. Wavelet transforms and segmentation are two commonly used techniques for processing accelerometer signals and extracting meaningful features.

III. PROPOSED METHOD

A. Data acquisition setting

Patients undergoing rehabilitation are outfitted with a wearable chest camera and smartwatch featuring accelerometers on both hands. The clinical setting is shown in Fig. 2 During the exercises, they utilize both hands (the right hand and the left hand) to facilitate the comparison between the normal hand and the weak hand. We used a GoPro Hero4 camera from San Mateo, California, USA, mounted on the chest and head to collect image data (Fig. 2(a)-(c)), while the accelerometer data were obtained by a Gear S3 Frontier, a Samsung watch worn on the patient's arms. During recording, the image data and accelerometer were manually synchronized. The smartwatch Samsung Gear 3 (SG3) employs different sensors such as an accelerometer, a gyroscope, a heart rate sensor, a thermal and a light sensor. In this study, sensing signals from the accelerometer will be used as inputs of the system. The sampling frequency of the SG3 is set at 50Hz. Fig. 2(d) depicts

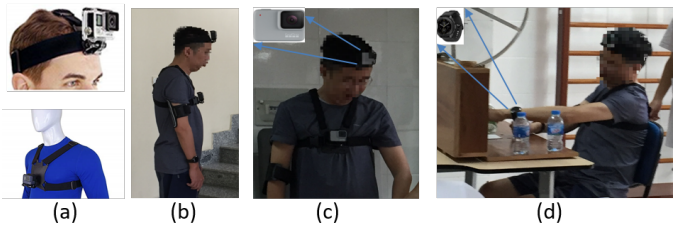


Fig. 2. Set up the wearable devices. a) Mounting wearable cameras on chest and head; b) Wearable devices see from side and frontal-view; c) a Patient wear cameras on chest and head; d) A patient doing a physical exercise wearing a smart-watch G3.

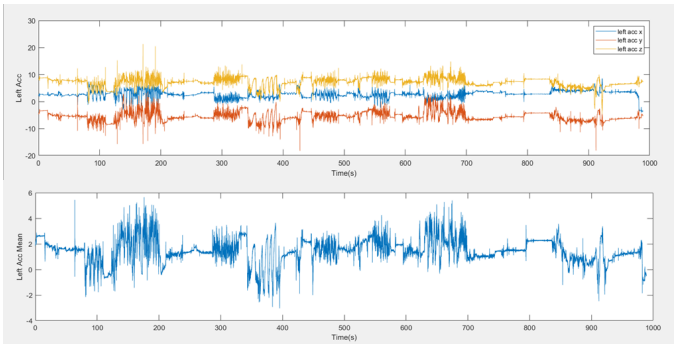


Fig. 3. Calculate the average value of the acceleration signal

actual photos of a patient wearing the sensors during a physical exercise.

B. Extract acceleration data

Acceleration data obtained from the sensor with 3 dimensions x, y, and z is averaged according to the following formula:

$$mean_{acc} = (acc_x + acc_y + acc_z)/3 \quad (1)$$

The original value and the average value of the acceleration are shown on the graph as shown in Fig.3:

C. Wavelet transform

Wavelet transforms have emerged as a widely used and effective technique for signal and image processing, particularly in the areas of noise removal and field separation, as described in references [7] and [8]. This technique, also known as multi-resolution analysis, involves decomposing a given signal into two key components: the approximation component A (also referred to as the Approximation), which corresponds to the low-frequency component, and the detail component D, which corresponds to the high-frequency component. The decomposition is achieved using two filters: a low-pass filter that is based on the scaling function $(\Psi(x))$ and a high-pass filter that utilizes the wavelet function $(\Phi(x))$. The scaling function captures the global features of the signal, while the wavelet function captures local features. By performing a wavelet transform on a given signal, it can be broken down

into its different frequency components, each of which can then be analyzed and processed independently. This makes the wavelet transform a powerful tool for processing complex signals and images, enabling more effective noise removal and signal separation. The relationship between these two functions is given by the following formulas:

$$\Phi(x) = \sum_{k=0}^{N-1} (c_k.\Phi(2x-k)) \tag{2}$$

$$\Psi(x) = \sum_{k=0}^{N-1} ((-1)^k.c_k.\Phi(2x+k-N+1)) \tag{3}$$

In order to minimize computational load, filters are applied at multiple levels. At each level, the signal is down-sampled to a factor of 2 after being passed through the filter. This results in each layer of the signal having a different resolution. The discrete wavelet transform, therefore, is known as a multi-resolution analysis (MRA).
At each filter stage, the filter expression is given by the formulas:

$$y_{\text{hight}}(n) = \sum_{k=0}^{N-1} (S(n).g(2k-n)) \tag{4}$$

$$y_{\text{low}}(n) = \sum_{k=0}^{N-1} (S(n).h(2k-n)) \tag{5}$$

whereas S(n) represents the signal, h(n) represents the impulse response of the low-pass filters corresponding to the scaling function $(\Psi(x))$, and g(n) represents the impulse response of the high-pass filters corresponding to the wavelet function $(\Phi(x))$. These two filters are related by the formula:

$$h(N-1-n) = (-1)^n g(n) \tag{6}$$

where N is the number of samples in the signal. In this work, we used a 5-level discrete wavelet transform to reduce the noise amplitude and align the signal's approximation with the horizontal axis.

D. Signal segmentation

In signal processing, there are various methods available for segmenting signals, including the Sliding Window algorithm, Top-Down algorithm, Bottom-Up algorithm, and SWAB algorithm (Sliding Window and Bottom-up). For this project, we prefer to use the Sliding Window algorithm to segment the accelerometer signal, due to its simplicity, low complexity, and ease of implementation in programming. In this study, the steps of the sliding window algorithm are depicted in Fig.4. The Sliding Window algorithm utilizes windows of varying size (initially set to 1) that move along the time axis of the signal. The algorithm begins by establishing the left boundary of the prediction segment, which also serves as the starting point for the sliding window on the right. While moving along the data points, the window size gradually increases until the mean squared error approximation of all data points within the window surpasses the user-defined maximum error threshold

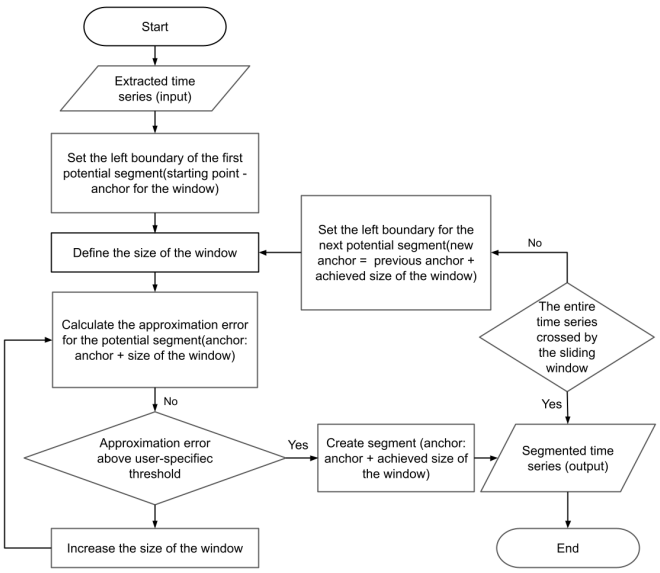


Fig. 4. Sliding Window algorithm diagram

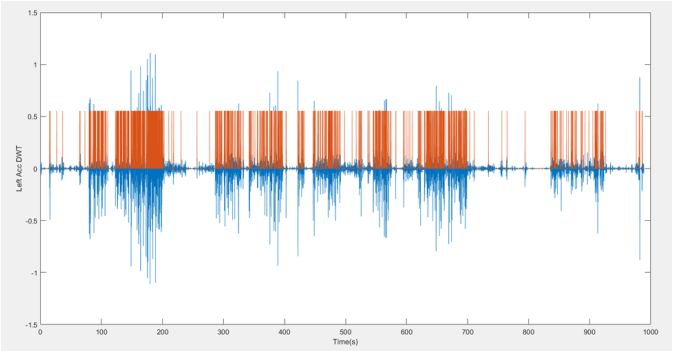


Fig. 5. Signal segmentation using Sliding Window (orange signal)

for the signal approximation. At this point, the right boundary of the window becomes the endpoint for a segment. The breakpoint of the previous segment becomes the starting point of the next segment, and the window size returns to the initial level, beginning the increment process again. This process continues until all data points are segmented. The Sliding Window method helps to segment the signal into segments with strong signal amplitude as shown in Fig.5

E. Detect time frame when hand engages activity

The energy E of a discrete signal x(n) is determined by the following expression:

$$E = \sum_{n=-\infty}^{\infty} \mathrm{x}^2(\mathrm{n}) \tag{7}$$

While energy calculation provides valuable information about the signal's amplitude, it does not provide any insight into the signal's time dependence. As we are aware, the amplitude of the acceleration signal varies over time. To account for this time-varying aspect, we calculate the short-time signal energy,

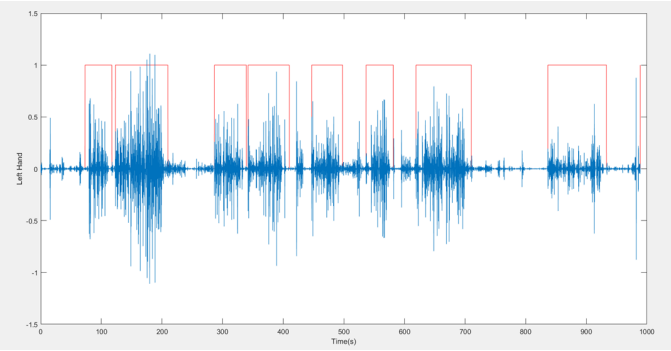


Fig. 6. Segment signal after calculating STE

commonly referred to as Short-time Energy (STE), to identify periods with high energy concentration in the signal. The STE is computed as follows:

$$\mathrm{E_m} = \sum_{n=-\infty}^{\infty} [x(\mathrm{n})\mathrm{w}(\mathrm{m}-\mathrm{n})]^2 \tag{8}$$

Once the E_m signal has been calculated, the energy threshold is applied and appropriate short time frames are eliminated to attain greater accuracy in comparison to the actual signal. Figure 6 depicts the outcome of signal segmentation during hand activity (red area).

IV. EXPERIMENTAL RESULTS

A. Dataset and Metric

The proposed method is evaluated by real situations of patient rehabilitation in which the data is collected at a Rehabilitation healthcare center. Each dataset comprises video captured by a chest-mounted camera and accelerometer data for both left and right-hand. To prepare ground-truth, the video data are utilized to denote exercise and non-exercise segments.

The performance of the proposed method was assessed using the Intersection over Union (IoU), which measures the overlap between the actual active interval and the detection interval. Furthermore, we employ two metrics including specificity and sensitivity indicators with an IoU threshold above 0.5. Two metrics are to comprehensively evaluate the performance and gain a better understanding of its capabilities and limitations.

B. Evaluation results

The results of the evaluation on four data sets with measurements for both hands are presented in Table I. In this context, the mean Intersection over Union (IoU) represents the degree of overlap between the predicted and actual time intervals. To accurately identify the rehabilitation interval, the start and end times of hand movements are recorded using accelerometers and overlaid onto corresponding exercise videos as shown in Fig. 7. The rehabilitation interval can be precisely determined by indicators (e.g., red and blue arrows for the right and left hand, respectively). The video results are made available at https://bit.ly/3laTBWg

TABLE I
EVALUATION OF THE PROPOSED MODEL'S ACTIVITY DETECTION

Data	Hand	IoU mean	Spec (%)	Sens (%)
Video 1	Left	0.79	1	1
	Right	0.78	1	1
Video 2	Left	0.7	0.92	0.92
	Right	0.43	0.55	0.4
Video 3	Left	0.53	0.92	0.69
	Right	0.61	0.875	0.875
Video 4	Left	0.59	0.875	0.7
	Right	0.7	0.82	0.75
Average	Both hands	0.64	0.87	0.80

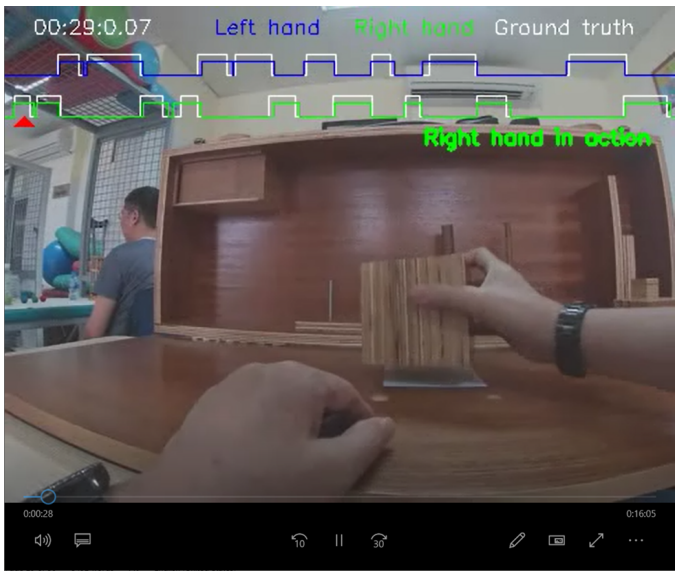


Fig. 7. Hand activity segment results on video

V. CONCLUSION

After completing the processing of the accelerometer signal received from the wearable device to obtain the results above and comparing them with the actual video of the exercises, we can see that the achieved results for the start and end times of one hand's participation in the exercise are quite good, but the accuracy for the starting and ending times of the activity is still not precise. There is still the phenomenon of wrong-hand detection in some activities. This occurs because the noise filtering process is not effective, and the error threshold at the processing steps is still done manually and cannot be automated, resulting in lower accuracy. In addition, computer vision studies should be combined to identify which exercises patients do and how often they should exercise to assess hand rehabilitation. These tasks will be our research direction in the future.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under

grant 102.01-2017.315.

REFERENCES

[1] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] J. Likitlersuang and J. Zariffa, "Interaction detection in egocentric video: Toward a novel outcome measure for upper extremity function," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 561–569, 2016.

[3] J. Likitlersuang, E. R. Sumitro, T. Cao, R. J. Visée, S. Kalsi-Ryan, and J. Zariffa, "Egocentric video: a new tool for capturing hand use of individuals with spinal cord injury at home," *Journal of neuroengineering and rehabilitation*, vol. 16, no. 1, pp. 1–11, 2019.

[4] D. Vishwakarma, P. Rawat, and R. Kapoor, "Human activity recognition using gabor wavelet transform and ridgelet transform," *Procedia Computer Science*, vol. 57, pp. 630–636, 2015.

[5] A. Nedorubova, A. Kadyrova, and A. Khlyupin, "Human activity recognition using continuous wavelet transform and convolutional neural networks," *arXiv preprint arXiv:2106.12666*, 2021.

[6] B. Zhou, C. Wang, Z. Huan, Z. Li, Y. Chen, G. Gao, H. Li, C. Dong, and J. Liang, "A novel segmentation scheme with multi-probability threshold for human activity recognition using wearable sensors," *Sensors*, vol. 22, no. 19, p. 7446, 2022.

[7] I. Daubechies, "Wavelets: Algorithms and applications," *Science*, vol. 262, no. 5139, pp. 1589–1592, 1993.

[8] M. Yves, "Wavelets and operators," 1992.

Investigating the Role of Human Action Detector in Visual-guide Audio Source Separation System

Thanh Thi-Hien Duong*, Trung-Hieu Nguyen†, The Thanh-Dat Le†, Thi-Lich Nghiem‡§, Duc-Huy Pham†, Thi-Lan Le§¶
*Dept. Information Technology, Hanoi University of Mining and Geology, Vietnam
†School of Information Technology, Hanoi University of Science and Technology, Vietnam
‡Thuongmai University, Vietnam
§Computer vision department, MICA, Hanoi University of Science and Technology, Vietnam
¶School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Vietnam
E-mail: duongthihienthanh@humg.edu.vn

***Abstract*—Visual-guided Audio Source Separation (VASS) is the task that deals with using available visual information to guide the audio separation from the mixture signal consisting of the sounds of many simultaneous sound sources. Visual information can be images of sound sources (e.g., musical instruments) or human gestures and activities (e.g., musicians). The visual features, audio features, and the correlation between audio and visual can be used to estimate the audio mask in the audio source separation model to improve the separation performance. In this study, we introduce a new multi-modal audio separation framework to jointly train the processing blocks helping the network to find the features of the target sounds more optimally. The proposed framework is introduced with three main blocks: visual extractor, action extractor, and audio separator. They combine visual, action, and spectral features to estimate the separated sounds' spectral masks. Extensive experiments have been conducted on the MUSIC dataset containing musical performance videos to evaluate the quality of human joint estimation and gesture representation and their role in audio separation. Experimental results obtained from different settings have confirmed the effectiveness of combining an action extractor with the audio-visual separation model.**

I. Introduction

As a topic in signal processing and machine learning area, audio source separation focuses on techniques to extract individual sound sources from mixed audio signals. The goal is to separate and isolate different sources, such as vocals, instruments, or background noise, from a recorded mixture [16], [24]. Audio source separation has received the interest of the research community for many decades due to its importance in many application fields, for example, telecommunication, healthcare, robotics, entertainment, music information retrieval, etc.

Early studies used digital signal processing techniques and probabilistic statistical models, to calculate and estimate the desired audio source [14], [19]. In current years, the strong development of machine learning and deep learning models has allowed researchers to develop universal deep learning models for audio source separation in particular as well as audio processing in general. These approaches use additional information about either the sources themselves or the environmental condition in order to guide the separation process. In case the input data is video clips, including audio and visual, the exploitation of visual information to help the audio separation process in the universal models is the ideal proposed by many researchers [8], [10]. Such universal models were shown to be effective in guiding the source separation algorithm, giving a promising performance.

Inspired by this approach, some recent studies including ours have used the semantic relationship between sound information and visual information about the objects existing in the video, to support the source separation. We have introduced an empirical study on isolating the sounds of individual instruments in the video [5]. Our other one has investigated an across-modal retrieval framework for simultaneous training of the audio component and visual component, which optimizes the correlation of sound and image of the object that is the sound source [12].

In this study, we present an extension of our previous work [12], to investigate the role of human action in the visual-guide audio source separation by combining the human action extractor with our previous audio-visual model. The proposed multi-modal framework includes three main blocks: (1) a visual extractor to detect and extract the visual feature of objects in the video; (2) an action extractor, to detect and extract the action feature of humans in the video; and (3) audio separator combines visual feature, action feature, and spectral feature for estimating the spectral masks of the separated sounds.

The rest of this paper is structured as follows. We first briefly present the state of the art on audio source separation and visual-audio source separation in Section II. Section III introduces the proposed VASS framework, integrating new processing blocks for extracting the human action feature. We describe experimental settings, results, and comments in Section IV. Finally, some conclusion is shown in Section V.

II. Related works

Audio source separation is an actively researched topic with various approaches and scenarios that have been published in

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

recent years. Some of the main approaches can be mentioned as follows:

Audio-only-based approaches have usually depended on prior knowledge or assumptions about the statistical characteristics of the target sounds (such as independence, sparsity, and so on) [4]. While traditional approaches used NMF [21], [3], [4] to separate the sound, type of supervised learning can generate synthetic samples frequently including combined known sounds in training data, then training the model to extract the desired sounds from the combination [11]. In settings where separated target sources are not available, weak sound-event activity labels [13] have been proposed.

In the audio-visual-based approaches, the challenge of figuring out the relationship between the audio signal and its matching pattern in the field of visuals is considered by Audio-Visual Source Separation. Some approaches have been used for tasks including voice isolation [6], [1], acoustic instrument audio isolation [23], [8], and separation of on-screen sounds of generic objects [18]. Integrating visual information into object and action recognition systems enhances their performance and versatility. By combining visual and auditory cues, these systems can better handle challenging scenarios and improve recognition accuracy. Deep learning techniques have been instrumental in achieving this integration. Arandjelović *et al.* demonstrated the joint learning of audio and visual representations for object recognition and retrieval [2]. Recently, authors have attempted to incorporate data on movement into the visual representation using these approaches, either in the form of pixel trajectories [23] or human posture [8]. Zhao *et al.* illustrated a PixelPlayer to learn from the unlabelled videos to separate input sound in the form of pixel trajectories [23]. Gan *et al.* proposed a novel approach that utilizes music-related visual cues, specifically hand gestures, to guide the sound separation process. The authors argued that certain hand gestures performed by musicians during instrument playing can provide cues about the sound source associated with each gesture. By analyzing video data of musicians playing different instruments, they aim to infer the sound source corresponding to each gesture and use it to separate the audio sources in a music mixture [8]. In 2020, Afouras *et al.* also exploited the temporal correspondence between audio and visual modalities in videos to learn meaningful audio-visual associations without explicit supervision. They proposed a self-supervised learning approach to learn aligning audio and visual streams from different video segments. The learned audio and visual features are used to cluster video segments based on the objects present in them [1]. Nevertheless, these strategies employ a video-level "mix-and-separate" training model.

Besides exploring the correlation between objects and sound sources, it is indeed important to investigate the relationship between actions. Shlizerman *et al.* proposed a novel approach that can predict human body actions and poses based solely on the audio input. They used a neural network model on a large dataset of synchronized audio and motion capture data, where the audio signals are used as input to predict the corresponding body poses. The model is designed to learn the complex mapping between audio and body dynamics, capturing the underlying patterns and relationships [15]. By analyzing the connection between action and sound, researchers can develop models that capture the multi-modal nature of audio-visual data more comprehensively [9].

These advancements in sound source separation and the integration of visual information into object and action recognition systems have paved the way for improved audio-visual analysis and understanding. By leveraging the complementary nature of audio and visual cues, researchers have made strides in enhancing recognition accuracy, robustness, and applicability in various domains, such as surveillance, robotics, and human-computer interaction [8]. Continued research in these areas holds great potential for further advancements and innovation in audio and visual signal processing.

III. Proposed VASS framework

A. General workflow

Aiming to separate different audio sources from video, we propose a VASS system composed of three key processing components: object detection and visual feature extractor, action detection and action feature extractor, and audio-visual-pose separator, as depicted in Figure 1. These blocks work together to process input data consisting of video and single-channel audio signals of such video. The object detector's role is to identify objects in the video frames that are likely to be audio sources. Once detected, the visual feature extraction block will extract features for the identified object. To extract the action features of the musician in the video, the action feature extractor is utilized. In the audio-visual-pose separation block, we employ a convolutional neural network, which combines audio features with visual features and action features through convolutional layers to estimate the spectral masks of the audio that need to be separated.

Below, we will present more details about the processing blocks in the proposed workflow:

1) Object detection and visual feature extraction: The object detector's role is to identify objects in the video frames that are likely to be audio sources. Once identified, the visual features of such objects are extracted by the visual feature extractor. In this module, we use the same way as in our previous study [12], in which the YOLOv5 model is used to detect instruments in the video (i.e., image frames) first, then the ResNet-18 is used to exploit the visual features of the identified instruments.

2) Human joints estimation: Besides objects, the actions of musicians, while they are playing various musical instruments, may bring useful information for sound source separation. To incorporate player action in the sound source separation framework, we extract human joints from RGB images and compute action embedding features via a Graph Convolutional Network (GCN). Specifically, AlphaPose [7], a deep learning method, to harness the power of CNN for accurate and reliable estimation of human joints is used for this task. By combining the capabilities of object detection and neural networks,

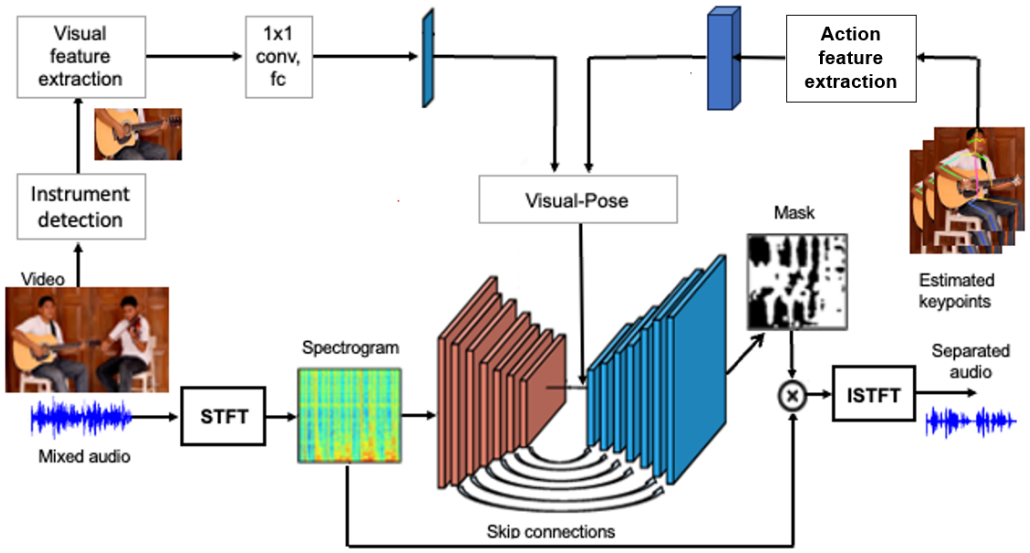


Fig. 1. The general workflow of proposed Visual-guide Audio Source Separation system.

AlphaPose provides precise estimations of the positions and actions of key joints within the human body. To ensure the system's effectiveness, AlphaPose has undergone extensive training using a large dataset. This training allows the system to learn and recognize the distinctive features and characteristics of various joints, including the shoulders, elbows, wrists, fingers, knees, and ankles. By learning from diverse examples, AlphaPose becomes capable of generalizing its knowledge and accurately estimating joints even in challenging scenarios such as occlusions and varying lighting conditions.

Figure 2 illustrates the index of 68 joints (26 joints for the body and 21 joints for each hand) of AlphaPose. We extract a sequence of skeletons with 68 joints from each 10-second clip. Each joint is represented by 2D coordinates (x, y) and its corresponding confidence score. This information will be fed to the GCN model to generate the embedding feature of the musician action.

3) Human action feature extraction and recognition: We use a Spatial-Temporal Graph Convolutional Network (ST-GCN) [22] to extract action features of musicians because compared to other approaches, this model has been proven robust in capturing human action's spatial and temporal information. The human action extraction module consists of two steps. First, an ST-GCN $G = (V, E)$ is built, where each node $v_i \in V$ corresponds to body joints of the body skeleton and each edge $e_i \in E$ represents the natural connection between those joints. We use 2D coordinates of detected body joints along with the confidence score extracted by AlphaPose as presented in the previous section as input for each node and construct a space-time graph by connecting body joints in a single frame and the joints in a sequence of frames. Then, we train the model on the training set of the MUSIC dataset to recognize the musical instrument playing gesture of the musician in the video. The main aim of the second step is to use the trained model as an action feature extractor. In this way, the framework can extract relevant features and representations specific to each instrument-playing activity. These features provide a distinctive representation of the actions performed while playing each instrument and will be used to help the model to separate sound sources better.

4) Audio separation: For audio separation, we use a U-NET convolutional network due to its proven effectiveness in audio prediction in this block. U-NET first extracts audio features of the spectrogram of the input mixed signal by applying down-sampling convolutional layers. These audio features are then conjugated with visual features and action features before being sent through up-sampling convolutional layers to estimate the spectral masks of each sound separately.

The audio-visual-pose separator is trained with a new loss function that we call the *audio-visual-pose loss function*. This loss function aims at exploiting not only audio spectral features, but also consistency between the separated spectrograms with the visual object categories, and the action categories.

B. Audio-visual-pose loss function

The proposed VASS system is implemented to separate the sound of each individual instrument from the music video. The original input data are videos and the single channel mixtures of audio of such videos. To train the audio-visual-pose separator aimed at achieving accurate audio source separation, the loss function is designed to enforce both spectral features and the consistency between the isolated spectrograms with the visual object categories, and the action categories.

Let's consider a pair of videos before mixing audio O_{v1} and O_{v2}, assuming the audio signal of each video is a mixture of individual sources as follows [5]:

$$x_1(t) = \sum_{i=1}^{N_1} s_i^1(t) \quad \text{and} \quad x_2(t) = \sum_{i=1}^{N_2} s_i^2(t), \tag{1}$$

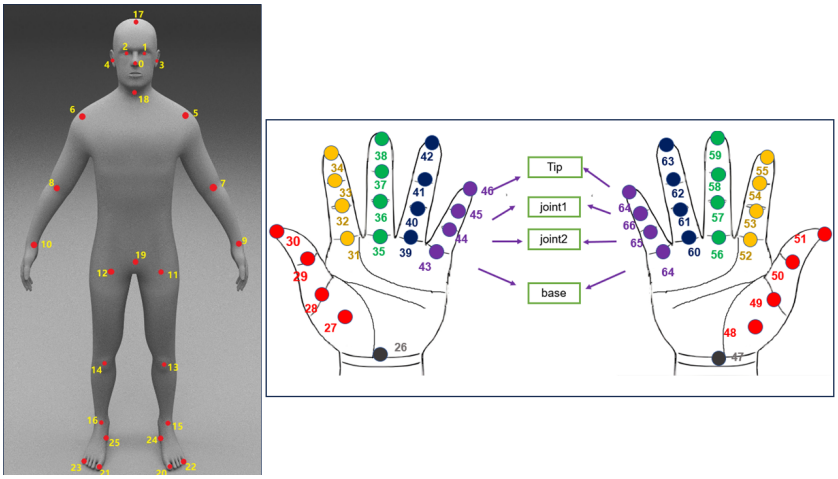


Fig. 2. Skeleton with join indexes extracted by Alphapose including 26 body joints and 21 joints for each hand.

where $s_i^1(t)$ and $s_i^2(t)$ represent the individual sources in the first and second videos, respectively.

The audio signals are transformed to the frequency domain by computing their magnitude spectrograms [5]:

$$\mathbf{X}_1 \approx \sum_{i=1}^{N_1} \mathrm{S}_i^1 \text{ and } \mathbf{X}_2 \approx \sum_{i=1}^{N_2} \mathrm{S}_i^2, \tag{2}$$

where $\mathbf{S}_i^1$ and $\mathbf{S}_i^2$ denote the magnitude spectrograms of the estimated sources $\hat{s}_i^1(t)$ and $\hat{s}_i^2(t)$ in the first and second videos, respectively. Similarly, $\mathbf{X}_1$ and $\mathbf{X}_2$ represent the magnitude spectrograms of the mixed signals $x_1(t)$ and $x_2(t)$, respectively.

The audio separation model is trained to aim for minimizing the audio separation loss component, which measures the discrepancy between the sum of separated spectrograms and the corresponding magnitude spectrograms of the mixed audio signals. Alternatively, The audio separation loss can be formulated in terms of the predicted ratio masks of the separated sound sources. The ratio mask represents the proportion of each source at each time-frequency bin. In this case, the separation loss $L_{sep-mask}$ can be approximated as [5]:

$$\begin{aligned} L_{sep-mask} = &|| \sum_{i=1}^{N_1} \mathcal{M}_i^1 - \mathcal{M}^{O_{v1}} ||_1 \\ &+ || \sum_{i=1}^{N_2} \mathcal{M}_i^2 - \mathcal{M}^{O_{v2}} ||_1, \end{aligned} \tag{3}$$

where $\mathcal{M}i^1$ and $\mathcal{M}i^2$ are the predicted spectrogram ratio masks for the sources in the first and second video, respectively. $\mathcal{M}^{Ov1}$ and $\mathcal{M}^{Ov2}$ represent the ground-truth spectrogram ratio masks for the sources in two videos, written as:

$$\mathcal{M}^{O_{v1}} = \frac{\mathbf{X}_1}{\mathbf{X}_1 + \mathbf{X}2} \quad \text{and} \quad \mathcal{M}^{Ov2} = \frac{\mathbf{X}_2}{\mathbf{X}_1 + \mathbf{X}_2}. \tag{4}$$

Realizing that the action of musicians when playing different instruments is also different, and assuming that the number of such action classes is equal to the number of instrument object classes, we introduce an object-pose-consistency loss for each predicted audio spectrogram. This loss ensures the separated spectrograms align with the visual object and the musician's action, and is calculated via cross-entropy loss as:

$$L_{obj-pose} = \frac{1}{N_1 + N_2} \sum_{i=1}^{N_1+N_2} \sum_{c=1}^{C} -y_{i,c} \log(p_{i,c}), \tag{5}$$

where N_1 and N_2 are the numbers of objects in the first and second video, respectively, and C represents the total number of classes. The binary index $y_{i,c}$ determines if class c is the correct class of the predicted spectrogram $\mathbb{S}i$, while $p_{i,c}$ denotes the prediction probability corresponding to the class c.

Finally, our overall loss function is defined as:

$$L = L_{\text{sep-mask}} + \lambda L_{\text{obj-pose}}, \tag{6}$$

where λ is the weight assigned to the object-pose-consistency loss. This weight allows us to balance the importance of separation accuracy with object consistency and action pose consistency in the overall training objective.

By optimizing this combined loss function, we aim to achieve a trade-off between accurately separating the audio sources using the predicted ratio masks and maintaining consistency between the separated spectrograms and the visual object and action categories. The parameter λ allows us to control the relative importance of the object and action poses consistency in relation to separation accuracy.

IV. Experimental Results

A. Dataset

The MUSIC dataset comprises YouTube videos collected through keyword queries. It consists of 685 video clips featuring musical duets and solos, with 536 solos and 149 duets. We used a subset that includes 536 solo videos to implement the proposed method, called *solo MUSIC*. This dataset covers a category of 11 instruments, including accordion, violin, guitar,

clarinet, tuba, acoustic cello, xylophone, flute, erhu, trumpet, and saxophone.

The original solo videos were cut into 10-second clips with various overlap ratios ranging from 50% to 70% to normalize and ensure and balanced distribution of solo clips of instrument classes. As a result, a total of 9760 10-second clips were obtained. They were allocated to training, evaluation, and testing datasets of 9260 clips, 250 clips, and 250 clips, respectively.

B. Evaluation metrics

Metrics of the human gesture recognition: We use four metrics: *Precision, Recall, F1-Score* [17] to evaluate the quality of the trained ST-GCN model for human gesture recognition.

Metrics of sound source separation: Three metrics *SDR (Signal-to-Distortion-Ratio)*, *SIR (Signal-to-Interference Ratio)*, and *SAR (Signal-to-Artifact Ratio)* are used to evaluate the quality of sound source separation. These criteria have been commonly used in the source separation community, measured in dB where the higher the better [20].

C. Implementation details

Our model is implemented using PyTorch, a popular deep-learning framework. The entire process involves several steps, including video segmentation, audio preprocessing, object detection, image feature extraction, pose estimation, pose feature extraction, and feature fusion.

1) Video segmentation: To begin with, we divide the input videos into clips of 10 seconds each. This ensures that the model can process manageable segments of the videos, facilitating better analysis and feature extraction.

2) Audio preprocessing: For the audio component, we use the STFT for converting the waveform audio signals into spectrograms and analyzing the frequency content of audio signals over time. The STFT is set up with a window size of 1022 and a hop length of 256, which enables capturing fine-grained details of the audio signals.

3) Object detection and image feature extraction: For object detection, we utilize the YOLOv5, a popular object detection algorithm. This model is trained on a comprehensive dataset consisting of 11 object categories derived from the Music dataset for fine-tuning and customizing for our specific task. Then, the ResNet18 is used to extract high-level features from the instrument images, which are crucial for further analysis and classification. The output of ResNet18 is passed through the 1x1 convolution fully connected layer to make the visual feature vector output with the size of (1, 256).

4) Pose estimation and human action feature extraction: In addition to object detection and image feature extraction, we utilize the Alpha Pose algorithm to extract skeletal joint positions from each frame of the video clips. By analyzing the action and configuration of these skeletal joints, we can infer valuable information about the musicians' poses and gestures. For each 10-second input clip, we divide it into 80 frames. These skeletal joint positions are then processed by the ST-GCN model, which has been pre-trained on our dataset. ST-GCN effectively captures temporal dynamics and encodes motion-related features from the skeletal joint sequences. It should be noted that both Alpha Pose and ST-GCN are pre-trained on our dataset. The output of this process is a vector of size (20, 68, 256), where 20 represents the 20 specific frames, 68 represents the number of skeletal joints, and 256 represents the number of corresponding features for each joint in each frame.

5) Feature fusion with Visual-Pose module: After obtaining the visual and pose features, we pass both of these features through the Visual-Pose Module to combine them. Note that the sizes of the visual features vector and pose features vector are different, so we first repeat the visual features to match the pose features' size before concatenating them to form a vector of size (20, 68, 512). Next, this vector is fed through two convolutional layers, which we refer to as spatial and temporal convolutions. The output of this process is a vector with dimensions (1, 512) used to fuse with the audio features generated by the Encoder in the U-NET. This fusion occurs at the bottleneck of the U-NET network.

6) Training setup: The model is trained for 2 epochs, with a batch size of 1. We use Adam optimizer to implement. The learning rate for the overall model was initialized as 1×10^{-4}, while for the image feature extraction component, we use a learning rate of 1×10^{-5}. This differential learning rate helps fine-tune the image feature extractor more effectively. In the loss function, the weight of object-pose-consistency loss is set $\lambda = 0.05$.

D. Results and discussion

1) Evaluation of the quality of human pose estimation: The main objective of this evaluation is to quantify the quality of human pose estimation. Given an RGB image, AlphaPose will result in 68 joints as explained in the previous section. Figure 3 shows an example of estimated joints for playing guitar, trumpet, and flute activities from RGB images while Fig. 4 illustrates a sequence of a skeleton of playing guitar.

We conducted an empirical study to examine the quality of joint estimation, focusing on three distinct activities in the dataset: guitar, trumpet, and flute playing. The dataset encompassed a wide range of guitarists and trumpeters with diverse playing styles, techniques, and skill levels.

To evaluate the confidence of the participation estimation, we calculate a threshold by analyzing the range of confidence values. This threshold serves as a reference to determine whether a joint is detected with high confidence or not. Joints with higher confidence values than the threshold are considered successfully detected, while those with lower confidence values may be deemed undetected. When setting the confidence threshold at 0.3, we observed that the pose estimation system captured more poses. However, this lower threshold also resulted in an increased likelihood of estimation errors. For example, some key points may be falsely detected or placed in incorrect positions, leading to less accurate pose representation. However, when setting the confidence threshold higher at 0.5, we observed more accurate estimations with

fewer errors. The system focused on capturing poses with higher confidence, resulting in higher accuracy and confidence levels. However, this also meant that some valid poses with lower confidence may have been excluded from the estimation.

Therefore, we have come up with a solution, by computing the average confidence values from the generated JSON result files, which provide detailed data on the estimated main joints including their corresponding confidence values, we calculate the average confidence value for each video. From the computed average confidence values, we establish a threshold that allows us to evaluate the confidence of the main joint estimations. This threshold enables us to accurately assess the performance of the pose estimation system on music-playing videos, providing reliable evaluations of joint positions and actions with higher confidence.

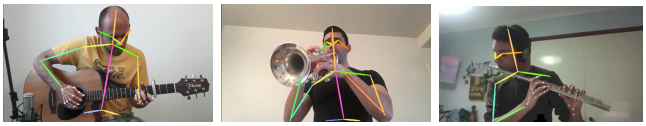


Fig. 3. The estimated joints of three musical playing activities.

Through the analysis of joint detection results during the performance of guitar, trumpet, and flute, we discovered guitar and flute playing activity exhibited consistently higher confidence scores, indicating robust and reliable joint detection. On the other hand, when trumpet playing activity, lower confidence scores were observed. This can be attributed to the occlusion of hand joints, meaning that certain parts of the hands were obstructed or hidden from view during the performance. The occlusion of hand joints in trumpet playing can occur due to the nature of holding the instrument, which may partially or fully cover some hand joints.

By aggregating the results from various videos, we generated a comprehensive report highlighting the joints that frequently go missing. This report provides valuable insights into the specific joints or body parts that present challenges for accurate estimation. Figure 5 shows the analysis results of three activities. We can observe that in guitar and flute playing activity, the bone joints were consistently and clearly detected, enabling accurate estimation of their positions and actions. Specifically, in the guitar playing activity, keypoints with ID numbers 50 (Right Thumb joint 2), 57 (Right Middle Finger joint 1), 49 (Right Thumb joint 1), 51 (Right Thumb tip), and 61 (Right Ring Finger joint 1) demonstrate the greatest occurrence of disappearances. These keypoints predominantly belong to the right hand, indicating the inherent challenges in detecting them due to the various playing techniques and hand positions involved. Furthermore, the disappearance rates of these keypoints consistently range from 8.85% to 9.69% out of a total of 58,603 frames. This implies that, on average, approximately 5186 to 5836 frames out of 58,603 frames experience the disappearance of these keypoints. This relatively low disappearance rate is advantageous in effectively monitoring the keypoint positions throughout the guitar-playing activity. These findings indicate the effective performance of the pose estimation system in monitoring the keypoints during guitar playing, enabling easier analysis and identification of guitar playing activities.

Similarly, in the flute playing activity, we achieve promising results with relatively low disappearance rates for the top 5 keypoints that experience the most frequent disappearances. The disappearance rates of these keypoints range from 6.40% - 7.78% out of a total of 64,363 frames, approximately 3750 to 4559 frames out of 64,363 frames experience the disappearance of these keypoints. The keypoints with the highest disappearance rates in flute playing are identified as ID 43 (Left Little finger base), 36 (Left Middle Finger joint 1), 9 (Left Wrist), 39 (Left Ring Finger base), and 35 (Left Middle Finger base). These keypoints primarily pertain to the left hand and predominantly represent the end joints of each finger, which is a characteristic feature observed during flute playing. Moreover, the disappearance rates of these keypoints in flute playing are the lowest among the three activities examined, indicating the exceptional performance of the pose estimation system in accurately monitoring these keypoints during flute playing.

The precise monitoring of keypoints during guitar-playing activities facilitates effective analysis and detection of guitar and flute-related actions. This capability significantly contributes to tasks such as instrument classification, enhancing our ability to identify and understand specific musical activities related to guitar and playing.

For the trumpet-playing activity, we observed significant differences compared to the previous two activities. The disappearance rates of keypoints suddenly increased, ranging from 7.98% to 15.59% out of a total of 83,698 frames, approximately 6679 to 13,048 frames out of 83,698 frames experienced the disappearance of these keypoints. The keypoints that experience the most frequent disappearances include the main keypoints with ID numbers 9 (Left Wrist), 7 (Left Elbow), 27 (Left Thumbbase), 10 (Right Wrist), and 26 (Left Palm). The disappearance rates of these keypoints exhibited a substantial deviation of up to 7.61%, whereas the discrepancy for guitar and flute playing activities was only around 0.84% and 1.48%, respectively. These keypoints encompass both sides of the hands and include larger joints in the arms, such as the Elbow. This indicates that playing the flute presents greater challenges in detecting keypoint positions compared to the previous activities. The playing techniques used in flute playing often obscure the keypoint locations, making them harder to discern. In particular, keypoints such as the elbow and wrist are frequently obstructed, making it extremely difficult to accurately determine these keypoints. The difficulty in identifying these keypoint locations significantly impedes the accuracy of the pose estimation system for flute playing. As a result, the precision of instrument recognition and performance analysis may be compromised, leading to a decrease in the quality of the results.

In summary, different playing postures in musical instrument

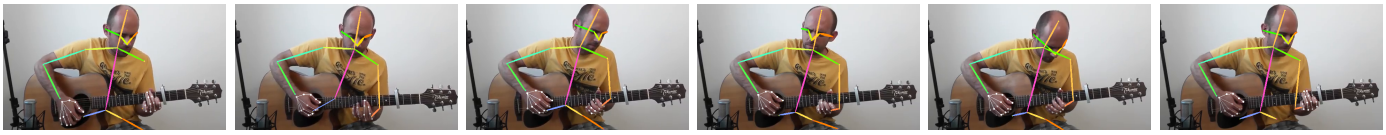


Fig. 4. Skeleton action during guitar playing

performance have a significant impact on the detection of keypoints and the determination of instrument-specific actions. This has a profound effect on the analysis and evaluation of musical instruments. The variations in hand positions, playing techniques, and instrument-specific actions make it challenging to accurately detect and track keypoints, leading to potential limitations in the analysis and assessment of musical instruments.

2) Evaluation of the human gesture recognition: Table I shows the recognition results of different activities while Figure 6 illustrates the confusion matrix. We can observe that ST-GCN is quite robust in recognizing these instrument activities. The F1-Score value for most activities is quite high, ranging from 0.86 to 0.99. The lowest F1-score (0.89) is obtained for "playing erhu" activity while the highest ones are achieved for "playing guitar" and "playing accordion". This shows that the model is capable of properly recognizing various activities based on human skeletons. Therefore, in this work, we employ the model as an extractor to compute the embedding feature for human action.

TABLE I
RECOGNITION RESULTS OF ACTIVITY RECOGNITION.

ID	Activity	precision	recall	f1-score
0	playing accordion	0.99	0.99	0.99
1	playing acoustic-guitar	0.98	1.00	0.99
2	playing cello	0.95	0.96	0.96
3	playing clarinet	0.93	0.85	0.89
4	playing erhu	0.82	0.98	0.89
5	playing flute	1.00	0.98	0.99
6	playing saxophone	0.95	0.79	0.86
7	playing trumpet	0.95	0.95	0.95
8	playing tuba	0.94	0.82	0.88
9	playing violin	0.98	0.97	0.98
10	playing xylophone	0.96	0.96	0.96

3) Evaluation of sound source separation: To investigate the effectiveness of the VASS model when exploiting more human action features, we compare the audio separation results obtained by our VASS model with three different setting cases on the MUSIC dataset:

- **Audio-Visual**: We do not perform the pose estimation, pose features extraction, and the Visual-Pose module. The output of the visual feature extraction process will have a size of 256, it was directly combined with the audio features generated by the Encoder at the bottleneck of the U-NET [5].
- **Audio-Pose**: For this case, object detection, image feature extraction, and the Visual-Pose module were not used. The output of the pose feature extractor will have a size of 512 instead of 256, it was directly combined with the audio features generated by the Encoder at the bottleneck of the U-NET.
- **Audio-Visual-Pose**: All the described activities in the proposed VASS model are performed sequentially.

The sound source separation results of our VASS model with three setting cases over the solo MUSIC dataset are shown in Table II.

TABLE II
AVERAGE AUDIO SOURCE SEPARATION RESULTS OF THE PROPOSED VASS MODEL OVER THE SOLO MUSIC DATASET.

Test case	SDR	SIR	SAR
Audio-Visual	4.92	6.54	**16.91**
Audio-Pose	3.88	6.11	15.72
Audio-Visual-Pose	**5.42**	**7.07**	16.78

As you can see, the results of the Visual-Object model are higher than those of the Visual-Pose model in both SDR and SIR, especially the most important metric SDR. This points out that using a combination of audio features and visual features of the instrument object for guiding the audio separation process gives better results than the field using a combination of audio features and musician action features.

The proposed Audio-Visual-Pose method offers the best instrument sound separation performance in both SDR and SIR when compared to the two others. Specifically with SDR, the Audio-Visual-Pose model provides 1.54 and 0.50 higher than the Audio-Pose model and Audio-Visual models, respectively. This confirmed that additional use of human action features along with audio features and object features can increase the performance of audio source separation from video.

V. CONCLUSIONS

Focus on the visual-guide audio source separation problem, this study investigates a specific task of separating different instrument sounds in the music performance video. Aiming to exploit the connection between action and sound, we proposed a novel deep learning model to jointly train the visual feature, action feature, and audio features of a multi-modal framework from the input video. Our VASS model was tested with three different setting cases: exploit audio features and visual features, exploit audio features and action features, and exploit all together audio, visual, and action features simultaneously.

Due to the limited server configuration for experimental, the audio separation results on the solo-solo MUSIC dataset were not highly competitive compared with some current studies. However, our study has confirmed the role of human action extraction in visual-guide audio source separation tasks, combining the human action extractor with the audio-visual model can increase the quality of separated sounds.

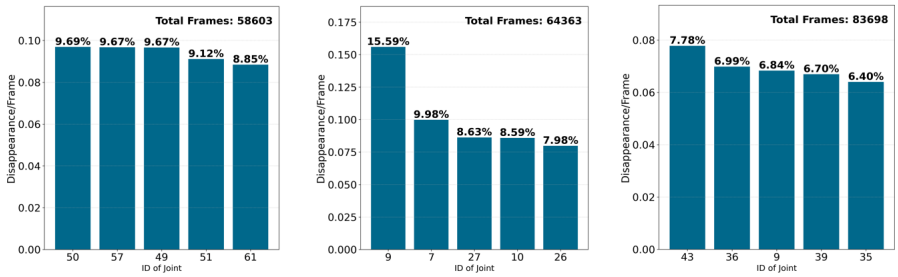


Fig. 5. Top 5 joints with the highest rate of low confidence in playing guitar (left figure), playing trumpet (center), and playing flute activity (right figure).

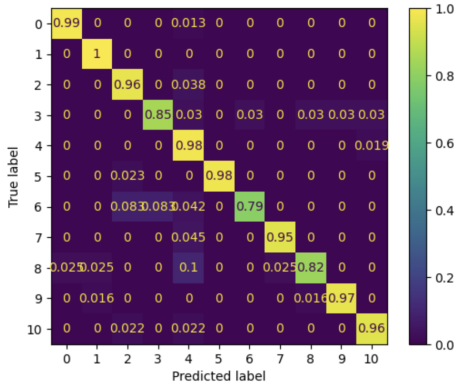


Fig. 6. Confusion matrix of action recognition.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education and Training of Vietnam under grant number B2023-MDC-02.

REFERENCES

[1] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision – ECCV 2020*. Springer International Publishing, 2020.

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

[3] H.-T. T. Duong, Q.-C. Nguyen, C.-P. Nguyen, T.-H. Tran, and N. Q. K. Duong. Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint. In *Proceedings of the Sixth International Symposium on Information and Communication Technology*, SoICT 2015, page 247–251, 2015.

[4] T. T. H. Duong, N. Q. K. Duong, P. C. Nguyen, and C. Q. Nguyen. Gaussian modeling-based multichannel audio source separation exploiting generic source spectral model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):32–43, 2019.

[5] T. T.-H. Duong, M. N. Huu, H. N. Thi, T.-L. Le, P.-L. Nguyen, and Q.-C. Nguyen. Visual-guided audio source separation: an empirical study. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2021.

[6] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4), July 2018.

[7] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[8] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[9] R. Gao and K. Grauman. Co-separating sounds of visual objects, 2019.

[10] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500, 2021.

[11] Y. Luo and N. Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, aug 2019.

[12] D.-H. Pham, Q.-A. Do, T. T.-H. Duong, T.-L. Le, and P.-L. Nguyen. End-to-end Visual-guided Audio Source Separation with Enhanced Losses. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 2022–2028, Chiang Mai, Thailand, Nov. 2022. IEEE.

[13] F. Pishdadian, G. Wichern, and J. L. Roux. Finding strength in weakness: Learning to separate sounds with weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1, 2020.

[14] H. Sawada, S. Araki, and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:516–527, 2011.

[15] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to body dynamics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[16] O. Slizovskaia, G. Haro, and E. Gomez. Conditioned Source Separation for Musical Instrument Performances. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2083–2095, 2021.

[17] I. Tomek. Two modifications of cnn. 1976.

[18] E. Tzinis, S. Wisdom, A. Jansen, S. Hershey, T. Remez, D. P. Ellis, and J. R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020.

[19] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. Duong. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, 2012. Latent Variable Analysis and Signal Separation.

[20] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[21] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[22] B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[23] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 587–604, Cham, 2018. Springer International Publishing.

[24] L. Zhu and E. Rahtu. V-SlowFast Network for Efficient Visual Sound Separation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2182–2192, Waikoloa, HI, USA, Jan. 2022. IEEE.

A combination of time and frequency synchronization with Doppler compensation for coded OFDM-based UWA systems

Hoai Linh Nguyen Thi, Quoc Khuong Nguyen, Van Duc Nguyen*
Wireless Communication Laboratory, School of Electrical and Electronic Engineering
Hanoi University of Science and Technology, Hanoi, Vietnam
* Corresponding author: duc.nguyenvan1@hust.edu.vn

***Abstract*—The paper proposes a combination of time and frequency synchronization with Doppler compensation for coded OFDM-based UWA systems. In addition, we use a soft-decision for M-QAM schemes and Viterbi decoding to enhance the system's performance. A novel pilot arrangement scheme is proposed to mitigate inter-carrier and intersymbol interference. The simulation results show that the system performance obtained by the proposed concept significantly outperforms the state-of-the-art methods.**

***Index Terms*—OFDM, underwater acoustic (UWA), Forward Error Correction (FEC), Doppler compensation, and non-uniform FFT.**

I. Introduction

Over the last decade, researchers and industry have been interested in investigating underwater acoustic (UWA) communications using the orthogonal frequency division multiplexing (OFDM) technology [1]–[3]. The deployment of the OFDM in underwater communication offers a variety of applications for both military and commercial purposes. To communicate in the UWA channel, researchers have faced a lot of challenges. The main drawbacks of communication in underwater channels are the high attenuation, high bit error rates, large and erratic propagation delays, high Doppler shift influence, and narrow bandwidth of acoustic channels [4]. In this research, we propose some techniques to overcome these disadvantages to enhance OFDM-based UWA performance.

The main contributions and novelties are given as follows:

- To solve these problems, we combine time and frequency synchronization in coded OFDM-based UWA systems. The time and frequency synchronization has been described in [5], [6].
- We provide soft-decision in comparison to the standard approach for coded OFDM-based UWA systems with M-QAM. Using this technique, the obtained system performance has a significantly lower BER than that obtained by the hard-decision technique.

The rest of our paper is organized as follows. The architecture of the OFDM system is described in Section II. Sec. III presents the suggested soft decision for the demodulation algorithms for 4-QAM, 16-QAM, and 64-QAM. The system performance simulation results are discussed in Sec. IV. Finally, Sec. V draws conclusions.

II. The architecture of the OFDM-based UWA system

A. Transmitter achitechture

Fig. 1(a) depicts the OFDM-based UWA transmitter structure. Firstly, the bits stream will be encoded by the convolutional coder. The encoded bits will be punctured to have a desired code rate. Afterwards, these bits are modulated by using the M-QAM modulation. The pilot symbols are inserted for the purpose of channel estimation. Virtual sub-carriers are inserted before sending to the IFFT to form the OFDM signal. The guard interval (GI) is inserted in the OFDM signal to protect the received signal from multipath propagation. The modulated signal is then sent to the receiver over the UWA channel [7].

B. Receiver achitechture

The receiver performs the inverse functions of the transmitter as shown in Fig. 1(b). In order to remove out-of-band noise, the received signal is initially fed to a bandpass filter (BPF). Later, the coarse Doppler frequency shift is estimated based on a pilot carrier frequency (PCF) signal and compensated by a resampling block. The next two blocks are OFDM frame synchronization and OFDM symbol separation. The fine Doppler frequency shift is performed to compensate for the remainder of the Doppler shift. The Least Square (LS) method is used for channel estimation. The Zero Forcing (ZF) method is deployed to perform the data equalization. The M-QAM demodulation block demodulates the equalized signal. Finally, a forward error correction (FEC) decoder corrects the erroneous caused by the fading channels. To enhance the system performance, we use soft-decision in the FEC decoders.

In this paper, we use a systematic convolutional encoder structure. The encoder has a code rate of 1/2. Its constraint length is 5.

C. Description of the proposal of the transmitted OFDM signal structure

As illustrated in Fig. 2, we suggest a pilot arrangement for both channel estimation and Doppler shift estimation, where Doppler shift is estimated based on the PCF signal. Additionally, this PCF signal transmits a larger power compared to other

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

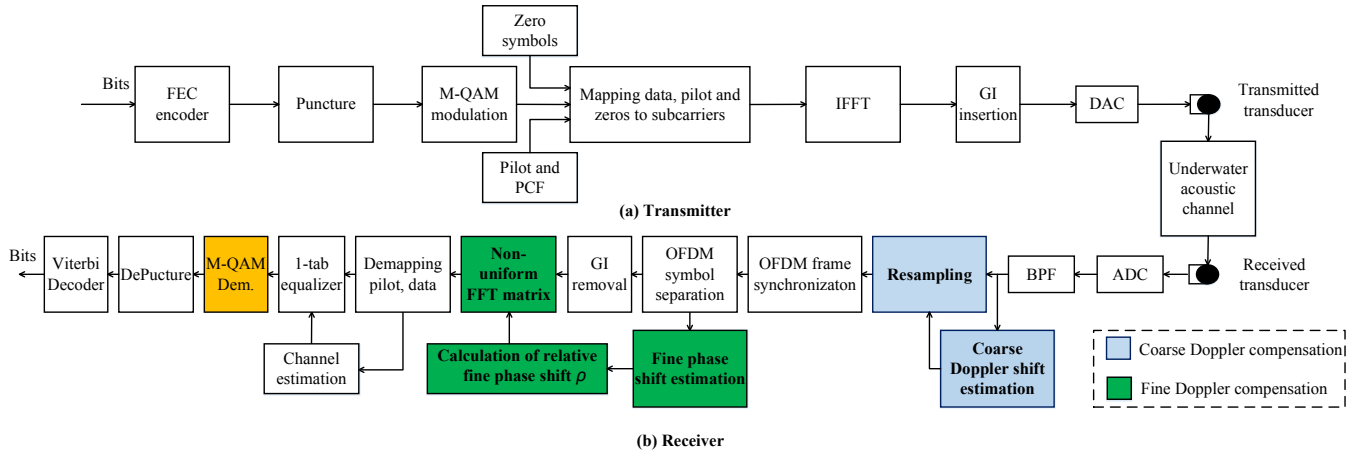


Fig. 1: The architecture of the proposed OFDM-based UWA system.

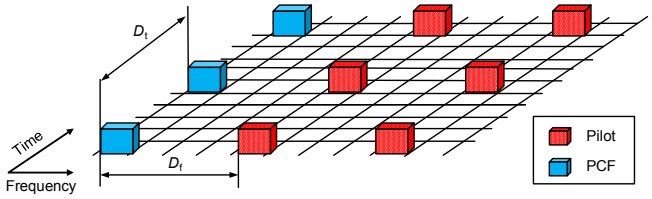


Fig. 2: Our proposal for a pilot strategy for OFDM-based UWA systems.

sub-carriers as shown in Fig. 3(a). D_t and D_f represent the pilot distance in the time and frequency domains, respectively. Moreover, we apply the Zadoff-chu sequence to create pilot value.

Let us consider the symbol $\vec{d}_{i,k}$ in the i^{th} OFDM symbol, on the k^{th} sub-carrier, then $\vec{d}_{i,k}$ can be written as

$$\vec{d}_{i,k} = [p_{i,2K}, d_{i,2K+1}[1], p_{i,2K+2}, \ldots, d_{i,3K-1}[K/2]]^T, \quad (1)$$

where $p_{i,2K}$ and $d_{i,2K+1}[n]$ represent a pilot in the i^{th} OFDM symbol, on the $2K^{th}$ sub-carrier and a data in the i^{th} OFDM symbol, on the $(2K+1)^{th}$ sub-carrier and at the index n, respectively. The index n runs from 1 to $K/2$.

Let us describe the modulated M-QAM symbols vector $\boldsymbol{s}_i$ after zeros insertion follow as:

$$\boldsymbol{s}_i = \begin{bmatrix} [0_{i,0}, \ldots 0_{i,2K-2}]^T \\ \mathrm{PCF}_{i,2K-1} \\ \vec{d}_{i,2K \to 3K-1} \\ [0_{i,3K}, \ldots, 0_{i,5K}]^T \\ \vec{d}^{*}_{i,5K+1 \to 6K} \\ \mathrm{PCF}^{*}_{i,6K+1} \\ [0_{i,6K+2}, \ldots, 0_{i,8K-1}]^T \end{bmatrix} \quad (2)$$

where $\mathrm{PCF}_{i,2K-1}$ represents the PCF signal in the i^{th} OFDM symbol on the $(2K-1)^{th}$ sub-carrier. Supposed that the FFT length is $N_{\mathrm{FFT}} = 8K$, the total number of data and pilot sub-carriers is $2K+2$, and the remaining $6K-2$ sub-carriers are zero.

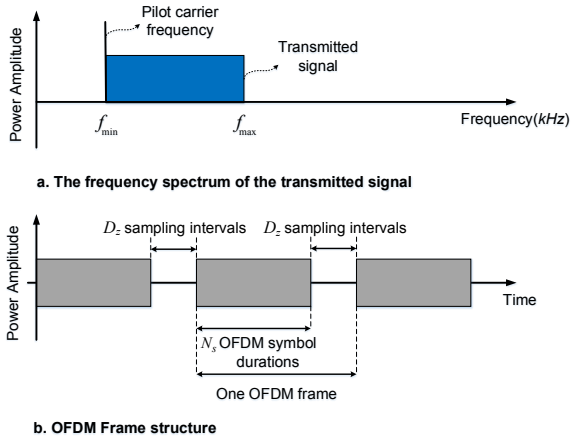


Fig. 3: (a) the frequency spectrum of the transmitted signal and (b) OFDM frame structure.

In Fig. 3(b), we can observe the structure of an OFDM-based UWA frame. Each OFDM frame is composed of a total of N_s OFDM symbols. Here, we use T_g to represent the duration of GI and T_{sub} to represent the duration of a single OFDM symbol. A pause period is created between two OFDM frames by the preamble D_z. Consequently, the duration of an entire OFDM frame, denoted as T_s, can be expressed as:

$$T_s = N_s(T_{sub} + T_g) + D_z \quad (3)$$

D. Doppler compensation by using the PCF signal

We make the assumption that the time-domain received signal, denoted as y, can be represented as $y = h*x+w$, where x signifies the transmitted signal, h characterizes the channel coefficients specific to the UWA channel, and w corresponds to the additive white Gaussian noise (AWGN).

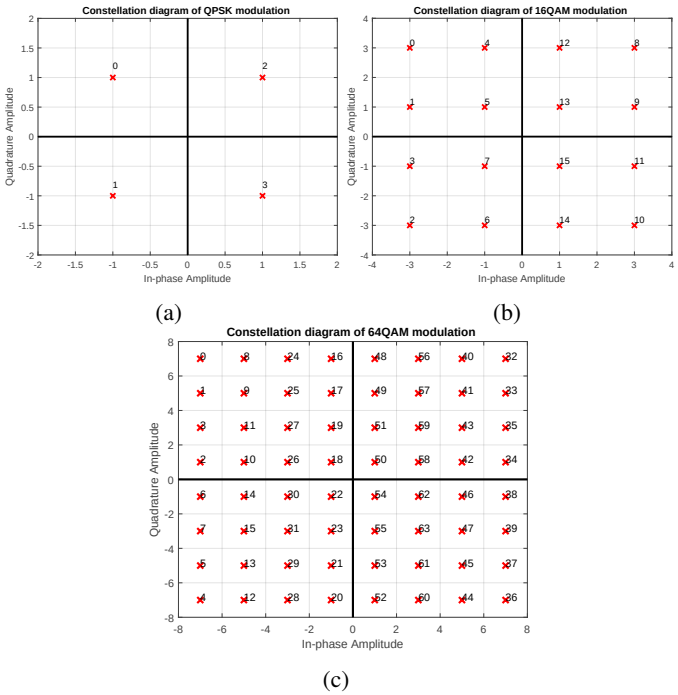
First, to eliminate the out-of-band noise impact on the received signal, y is fed into the BPF block. Then, the coarse Doppler shift estimation and resampling will be done on the PCF signal. Based on the guard interval, the beginning of each OFDM frame is detected by using a time synchronization method described in detail in [5]. Afterward, the fine Doppler compensation for the received signal y is performed after removing the guard interval, which is clearly presented in [6]. The fine Doppler compensation is to minimize the remaining frequency offset by using a non-uniform FFT matrix, namely, $[\mathbf{G}_{\mathbb{RS}}]$, as follows:

$$\hat{\mathbf{Y}} = [\mathbf{G}_{\mathbb{RS}}]y \tag{4}$$

After the Doppler shift is compensated, we can estimate channel coefficients at the positions of the reference signals based on the pilot signal. Following the acquisition of all the channel coefficients at these points, we have to estimate channel coefficients at the positions of the data signals. The most conventional method is the interpolation technique. In terms of equalization, the signal may be easily implemented by applying the formula $\tilde{X} = \frac{\hat{Y}}{H}$ in the frequency domain. Before decoding FEC, the signal is then fed to M-QAM demodulation which is discussed in detail in Sec. III.

III. Proposal of a Soft-Decision Algorithm for M-QAM Demodulation

To improve the performance of our OFDM system, we propose a soft-decision detection instead of using hard-decision for M-QAM demodulation. Fig. 4 depicts the M-QAM (namely, 4-QAM, 16-QAM, and 64-QAM) constellation.



(a) (b)

(c)

Fig. 4: (a) 4-QAM constellation, (b) 16-QAM constellation, and (c) 64-QAM constellation.

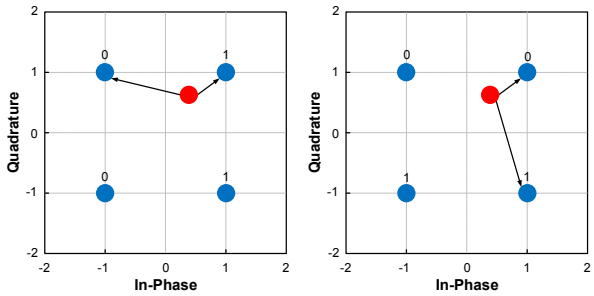
Let S_{I}^{+} and S_{I}^{-} represent a set of the symbols whose I^{th} bit is 1 and 0, respectively. To make the decision on I^{th} bit according to soft output, we define the log-likelihood ratio (LLR) value calculated as follows

$$\begin{aligned}\mathrm{LLR}(\mathrm{b_I}) &= \ln\frac{\sum\limits_{\mathrm{X}\in\mathrm{S_I^-}} \mathrm{f}(\tilde{\mathrm{X}}|\mathrm{X})}{\sum\limits_{\mathrm{X}\in\mathrm{S_I^+}} \mathrm{f}(\tilde{\mathrm{X}}|\mathrm{X})} \\ &\approx \frac{1}{2\sigma^2}\left(|\tilde{X} - X_{\mathrm{I,opt}}^{+}|^2 - (|\tilde{X} - X_{\mathrm{I,opt}}^{-}|^2\right)\end{aligned} \tag{5}$$

where $\mathrm{f}(\tilde{\mathrm{X}}|\mathrm{X})$ stands for the probability density function that a symbol can be received given its transmission, $X_{\mathrm{I,opt}}^{+} = \arg\min_{X\in\mathrm{S_I^+}}|\tilde{X} - X|^2$, and $X_{\mathrm{I,opt}}^{-} = \arg\min_{X\in\mathrm{S_I^-}}|\tilde{X} - X|^2$. Based on Eq. (5), we calculate the LLR values for different demodulation schemes.

A. Soft-decision for 4-QAM demodulation

Let us $b_0 b_1$ denotes two bits of 4-QAM demodulation. For each bit of the 4-QAM constellation, The calculations of LLR value are depicted in Fig. 5.



(a) 1^{st} bit (b) 2^{nd} bit

Fig. 5: The 4-QAM constellation's LLR computations for each bit: examples.

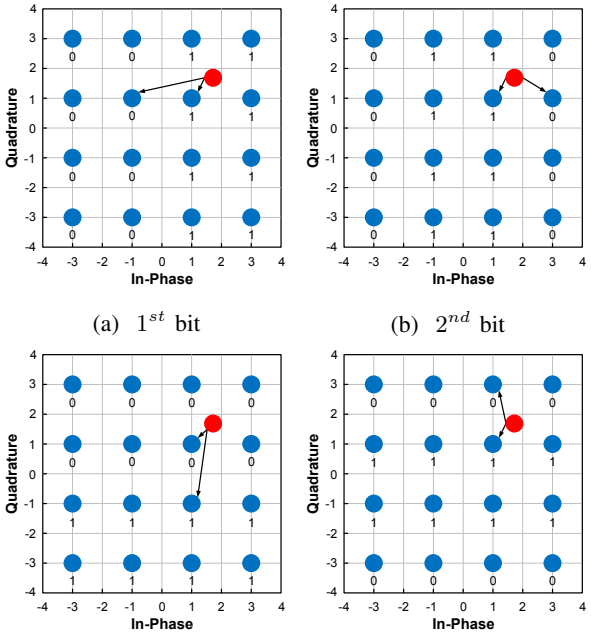
It follows from Fig. 5 that the LLR values of 1^{st} bit $\mathrm{LLR}\,(b_0)$ depend on the real part $\tilde{X}_{\mathrm{I}}$ whereas the LLR values of 2^{nd} bit $\mathrm{LLR}\,(b_1)$ rely on the imaginary part $\tilde{X}_{\mathrm{I}}$. Suppose that the red point is represented by the symbol $\tilde{X}_{\mathrm{I}}$. In terms of LLR calculates of 1^{st} bit, the real part of the 4-QAM constellation takes the value 1 when $b_0 = 1$ and -1 when $b_0 = 0$. Similar to the LLR calculation of 2^{nd} bit, the imaginary part of the 4-QAM constellation takes the value -1 when $b_0 = 1$ and 1 when $b_0 = 0$. Using the Eq.(5), two LLR values of each bit in 4-QAM demodulation are given as:

$$\mathrm{LLR}\,(b_0) = -\frac{2}{\sigma^2}\Re(\tilde{X}_{\mathrm{I}}), \quad \forall\Re(\tilde{X}_{\mathrm{I}}) \tag{6}$$

$$\mathrm{LLR}\,(b_1) = \frac{2}{\sigma^2}\Im(\tilde{X}_{\mathrm{I}}), \quad \forall\Im(\tilde{X}_{\mathrm{I}}) \tag{7}$$

B. Soft-decision 16-QAM demodulation algorithm

Let us $b_0b_1b_2b_3$ represent four bits of 16-QAM demodulation. From equations (8), (9), (10) and (11), it can be seen that the LLR values of the 1^{st} bit LLR (b_0) and 2^{nd} bit LLR (b_1) depend on the real part $\tilde{X}_{\mathrm{I}}$, while the values of the 3^{rd} bit LLR (b_2) and 4^{th} bit LLR (b_3) rely on the imaginary part $\tilde{X}_{\mathrm{I}}$. Fig. 6(a) shows LLR computation for 1^{st} bit. In this constellation, the real part has values of 1 and 3 when b_0 is 1, and -1 and -3 when b_0 is 0. According to the Hamming distance, the numbers -1 and 1 correspond to $b_0 = 0$ and $b_0 = 1$, respectively, which may be disregarded. Similar to the remaining bits, the LLR values can be described as follows:



(a) 1^{st} bit (b) 2^{nd} bit

(c) 3^{rd} bit (d) 4^{th} bit

Fig. 6: Examples of The 16-QAM constellation's LLR computations for each bit.

$$\mathrm{LLR}\,(b_0) = \begin{cases} -\frac{4}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) - 1], & 2 \le \Re(\tilde{X}_{\mathrm{I}}) \\ -\frac{2}{\sigma^2}\Re(\tilde{X}_{\mathrm{I}}), & -2 \le \Re(\tilde{X}_{\mathrm{I}}) < 2 \\ -\frac{4}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) + 1)], & \Re(\tilde{X}_{\mathrm{I}}) < -2 \end{cases} \tag{8}$$

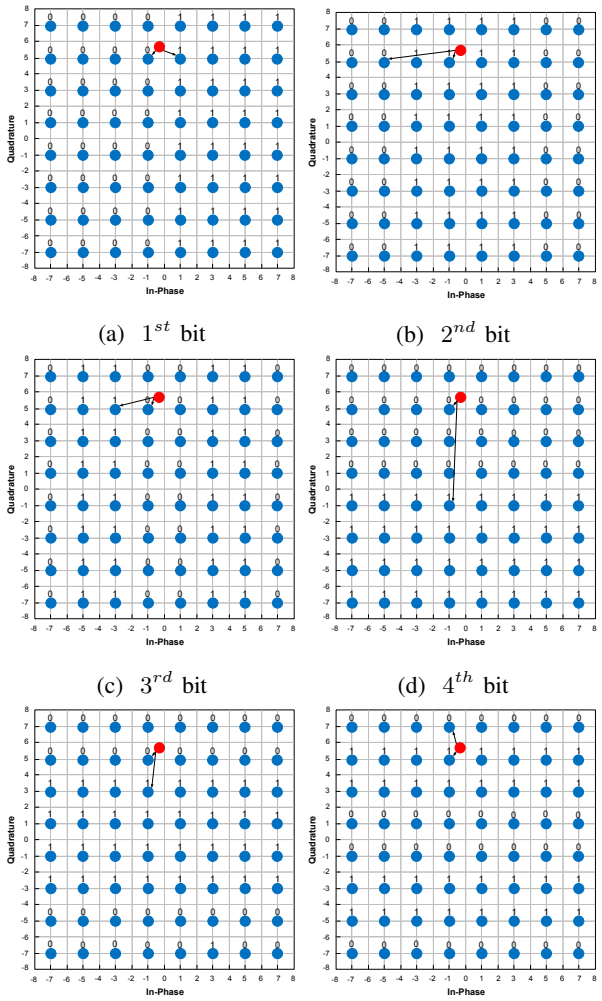
$$\mathrm{LLR}\,(b_1) = \frac{2}{\sigma^2}[|\Re(\tilde{X}_{\mathrm{I}})| - 2], \quad \forall \Re(\tilde{X}_{\mathrm{I}}) \tag{9}$$

$$\mathrm{LLR}\,(b_2) = \begin{cases} \frac{4}{\sigma^2}[\Im(\tilde{X}_{\mathrm{I}}) - 1)], & 2 \le \Im(\tilde{X}_{\mathrm{I}}) \\ \frac{2}{\sigma^2}\Im(\tilde{X}_{\mathrm{I}}), & -2 \le \Im(\tilde{X}_{\mathrm{I}}) < 2 \\ \frac{4}{\sigma^2}[\Im(\tilde{X}_{\mathrm{I}}) + 1], & \Im(\tilde{X}_{\mathrm{I}}) < -2 \end{cases} \tag{10}$$

$$\mathrm{LLR}\,(b_3) = \frac{2}{\sigma^2}[|\Im(\tilde{X}_{\mathrm{I}})| - 2], \quad \forall \Im(\tilde{X}_{\mathrm{I}}) \tag{11}$$

C. Soft-decision 64-QAM demodulation algorithm

Let us $b_0b_1b_2b_3b_4b_5$ represent six bits of 64-QAM demodulation. Similar to Sec. B, the values of the first three bits (LLR (b_0), LLR (b_1) and, LLR (b_2)) depend on the real component $\tilde{X}_{\mathrm{I}}$, while the values of the last three bits (LLR (b_3), LLR (b_4) and, LLR (b_5)) depend on the imaginary part $\tilde{X}_{\mathrm{I}}$ as shown in Fig. 7. These LLR values can be obtained as follows:



(a) 1^{st} bit (b) 2^{nd} bit

(c) 3^{rd} bit (d) 4^{th} bit

(e) 5^{rd} bit (f) 6^{th} bit

Fig. 7: Examples of The 64-QAM constellation's LLR computations for each bit.

$$\mathrm{LLR}\,(b_0) = \begin{cases} -\frac{8}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) - 3], & 6 \le \Re(\tilde{X}_{\mathrm{I}}) \\ -\frac{6}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) - 2], & 4 \le \Re(\tilde{X}_{\mathrm{I}}) < 6 \\ -\frac{4}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) - 1], & 2 \le \Re(\tilde{X}_{\mathrm{I}}) < 4 \\ -\frac{2}{\sigma^2}\Re(\tilde{X}_{\mathrm{I}}), & -2 \le \Re(\tilde{X}_{\mathrm{I}}) < 2 \\ -\frac{4}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) + 1], & -4 \le \Re(\tilde{X}_{\mathrm{I}}) < -2 \\ -\frac{6}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) + 2], & -6 \le \Re(\tilde{X}_{\mathrm{I}}) < -4 \\ -\frac{8}{\sigma^2}[\Re(\tilde{X}_{\mathrm{I}}) + 3], & \Re(\tilde{X}_{\mathrm{I}}) < -6 \end{cases} \tag{12}$$

$$\mathrm{LLR}\,(b_1) = \begin{cases} \frac{4}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) - 5], & 6 \le \Re(\tilde{X}_\mathrm{I}) \\ \frac{2}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) - 4], & 2 \le \Re(\tilde{X}_\mathrm{I}) < 6 \\ \frac{4}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) - 3], & 0 \le \Re(\tilde{X}_\mathrm{I}) < 2 \\ -\frac{4}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) + 3], & -2 \le \Re(\tilde{X}_\mathrm{I}) < 0 \\ -\frac{2}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) + 4], & -6 \le \Re(\tilde{X}_\mathrm{I}) < -2 \\ -\frac{4}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) + 5], & \Re(\tilde{X}_\mathrm{I}) < -6 \end{cases} \tag{13}$$

$$\mathrm{LLR}\,(b_2) = \begin{cases} \frac{2}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) - 6], & 4 \le \Re(\tilde{X}_\mathrm{I}) \\ -\frac{2}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) - 2], & 0 \le \Re(\tilde{X}_\mathrm{I}) < 4 \\ \frac{2}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) + 2], & -4 \le \Re(\tilde{X}_\mathrm{I}) < 0 \\ -\frac{2}{\sigma^2}[\Re(\tilde{X}_\mathrm{I}) + 6], & \Re(\tilde{X}_\mathrm{I}) < -4 \end{cases} \tag{14}$$

$$\mathrm{LLR}\,(b_3) = \begin{cases} \frac{8}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 3], & 6 \le \Im(\tilde{X}_\mathrm{I}) \\ \frac{6}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 2], & 4 \le \Im(\tilde{X}_\mathrm{I}) < 6 \\ \frac{4}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 1], & 2 \le \Im(\tilde{X}_\mathrm{I}) < 4 \\ \frac{2}{\sigma^2}\Im(\tilde{X}_\mathrm{I}), & -2 \le \Im(\tilde{X}_\mathrm{I}) < 2 \\ \frac{4}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 1], & -4 \le \Im(\tilde{X}_\mathrm{I}) < -2 \\ \frac{6}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 2], & -6 \le \Im(\tilde{X}_\mathrm{I}) < -4 \\ \frac{8}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 3], & \Im(\tilde{X}_\mathrm{I}) < -6 \end{cases} \tag{15}$$

$$\mathrm{LLR}\,(b_4) = \begin{cases} \frac{4}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 5], & 6 \le \Im(\tilde{X}_\mathrm{I}) \\ \frac{2}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 4], & 2 \le \Im(\tilde{X}_\mathrm{I}) < 6 \\ \frac{4}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 3], & 0 \le \Im(\tilde{X}_\mathrm{I}) < 2 \\ -\frac{4}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 3], & -2 \le \Im(\tilde{X}_\mathrm{I}) < 0 \\ -\frac{2}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 4], & -6 \le \Im(\tilde{X}_\mathrm{I}) < -2 \\ -\frac{4}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 5], & -6 \le \Im(\tilde{X}_\mathrm{I}) < -4 \end{cases} \tag{16}$$

$$\mathrm{LLR}\,(b_5) = \begin{cases} \frac{2}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 6], & 4 \le \Im(\tilde{X}_\mathrm{I}) \\ -\frac{2}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) - 2], & 0 \le \Im(\tilde{X}_\mathrm{I}) < 4 \\ \frac{2}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 2], & -4 \le \Im(\tilde{X}_\mathrm{I}) < 0 \\ -\frac{2}{\sigma^2}[\Im(\tilde{X}_\mathrm{I}) + 6], & \Im(\tilde{X}_\mathrm{I}) < -4 \end{cases} \tag{17}$$

IV. Results and Discussions

In the following, the parameters of the experimental OFDM-based UWA system are described, where a single input and single output scenario are considered. The sampling frequency f_s is 96 kHz. The carrier frequency ranges from 12 kHz to 15 kHz. The guard interval length and the total number of sub-carriers are $GI = 1024$ and $N_{\mathrm{FFT}} = 2048$, respectively. The modulation techniques, namely, 4-QAM, 16-QAM, and 64-QAM, are applied. Regarding the simulation results, we will discuss the following issues: frame synchronization results, Doppler compensation performance, the effectiveness of the pilot strategy, and the performance of the proposed soft-decision method.

First of all, we discuss the synchronous signal. Fig. 8(a) represents the received signal after adding AWGN at SNR = 5 dB, respectively. Next, Fig. 8(b) is the synchronous signal, **SYN**, which is obtained on the basis of the designed guard interval.

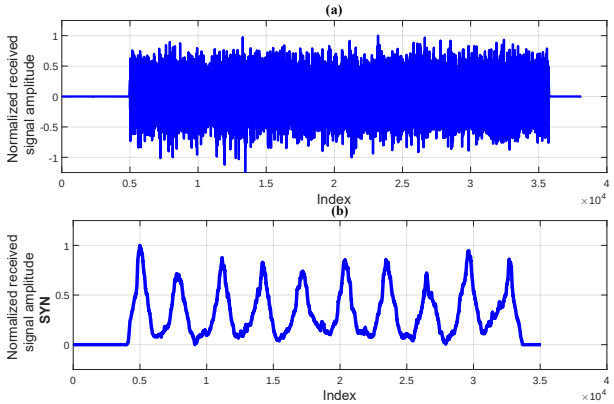


Fig. 8: (a) The received signal (**y**), (b) the synchronous signal (**SYN**)

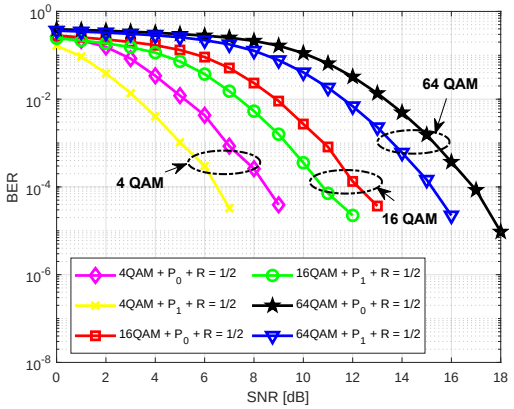


Fig. 9: BER of different pilot schemes.

Figure 9 shows the performance of the pilot strategy in our system using the convolution code $R = 1/2$. Instead of using the pilot sequence P_0 that is used in [6] for channel estimation, we employ the Zadoff-chu sequence P_1 in this article. As can be seen, our approach outperforms the method of [6] three distinct modulation schemes. Specifically, this method offers a system performance in terms of SNR gain of 2dB in the case of 4-QAM, 1dB in the case of 16-QAM, and 2dB in the remaining case.

In the following part, we discuss the performance of the Doppler compensation method. To evaluate the performance of the method in [6] when the system applies convolution codes $R = 1/2$ and soft-decision of M-QAM demodulation by adjusting the subcarrier frequency to shift it by 20 Hz. As shown in Fig. 10, the performance of the system is measured by BER. The case of without Doppler compensation is a reference. The performance of the OFDM-based UWA is clearly improved by using the proposed technique.

Finally, we discuss the performance of the proposed method if the demodulation scheme is combined with convolution code. The suggested technique performs better in both circumstances than the hard decision, as illustrated in Fig. 11(a), 11(b) as well as in 11(c). In particular, when using 4-QAM modulation, our technique outperforms other methods by 1 dB

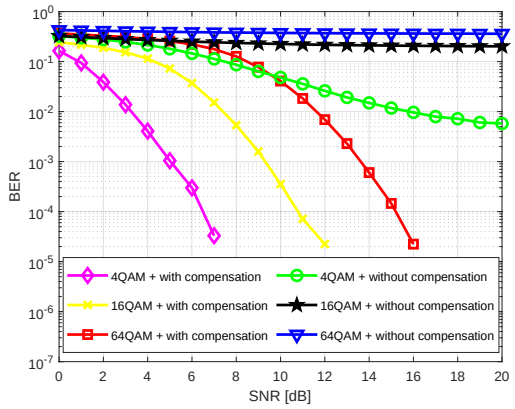


Fig. 10: BER of Doppler compensation with convolution code $R = 1/2$.

for code rates of $R = 1/2$ and 2 dB for rates of $R = 2/3$ as depicted in Fig. 11(a). According to the results shown in Fig 11(b) and 11(c), our method is better than 6 dB in the case of code rate $R = 1/2$ and 5dB in the case of $R = 2/3$ for both 16-QAM and 64-QAM.

V. Conclusion

In this paper, we propose a combination of time and frequency synchronization with Doppler compensation for coded OFDM-based UWA systems. For a given transmission scenario, the numerical simulations help us to optimize the system parameters to obtain the best performance. Numerical simulation results show also that the obtained system performance by using our proposed techniques outperforms the state of the art in terms of BER or in terms of SNR.

Acknowledgment

This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2023-PC-034.

References

[1] E. Panayirci, M. T. Altabbaa, M. Uysal, and H. V. Poor, "Sparse channel estimation for ofdm-based underwater acoustic systems in rician fading with a new omp-map algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 6, pp. 1550–1565, 2019.

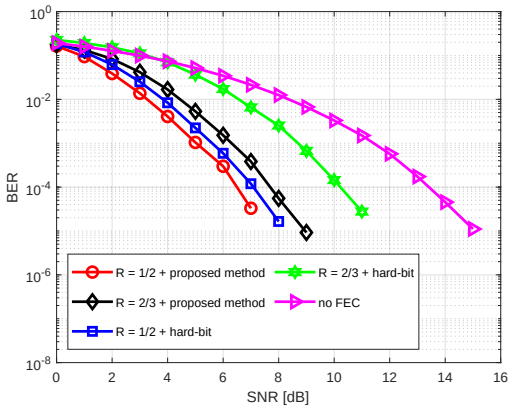
[2] M. R. Khan, B. Das, and B. B. Pati, "Channel estimation strategies for underwater acoustic (uwa) communication: An overview," *Journal of the Franklin Institute*, vol. 357, no. 11, pp. 7229–7265, 2020.

[3] X. Zhu, C.-X. Wang, and R. Ma, "A 2d non-stationary channel model for underwater acoustic communication systems," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. IEEE, 2021, pp. 1–6.

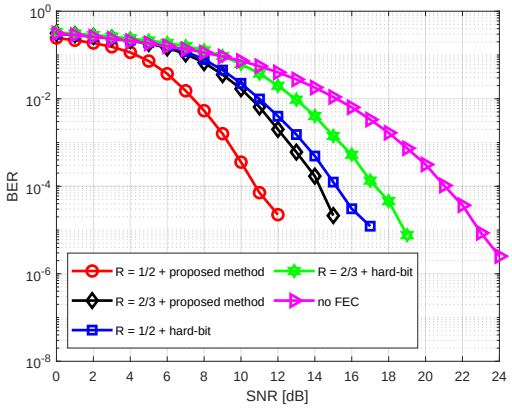
[4] L. Liu, Y. Wang, L. Li, X. Zhang, and J. Wang, "Design and implementation of channel coding for underwater acoustic system," in *2009 IEEE 8th International Conference on ASIC*. IEEE, 2009, pp. 497–500.

[5] D. H. Do, Q. K. Nguyen, N. Van Duc *et al.*, "A time synchronization method for ofdm-based underwater acoustic communication systems," in *2016 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2016, pp. 131–134.

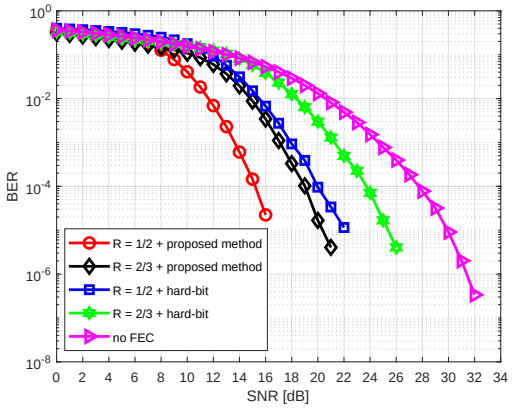
[6] H. L. N. Thi, Q. K. Nguyen, T. H. Nguyen *et al.*, "Low Complexity Non-Uniform FFT for Doppler Compensation in OFDM-Based Underwater Acoustic Communication Systems," *IEEE Access*, vol. 10, pp. 82 788–82 798, 2022.



(a)



(b)



(c)

Fig. 11: (a) BER of 4-QAM modulation, (b) BER of 16-QAM modulation, and (c) BER of 64-QAM modulation in combination with different code rate R.

[7] H. Trung, N. Van Duct *et al.*, "Symbol duration and adaptive guard length for underwater acoustic ofdm systems," in *2019 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2019, pp. 169–173.

Remarks on an Optimal Predictive Control Using a Quaternion Neural Network and a Metaheuristics Optimisation Method

Kazuhiko Takahashi, Remi Aoki, Eri Kagamiishi, Masafumi Hashimoto
Doshisha University, Japan
E-mail: {katakaha@mail, cguf1003@mail4, cguf1018@mail4, mhashimo@mail}.doshisha.ac.jp

***Abstract*—The present study explored the possibility of using a quaternion neural network (QNN) as a prediction model to achieve optimal predictive control of non–linear systems. Computational experiments for a tracking task of a non–linear plant confirmed the promising role of QNN for control applications.**

I. Introduction

Although deep neural networks (DNNs) have been successfully demonstrated to solve intractable problems in many fields [1], high–dimensional NNs using hyper–complex numbers have been explored as an alternative for handling such problems. The present study investigated a control systems application using a quaternion NN (QNN) [2]. We focused on a model predictive control (MPC) approach [3] to achieve optimal predictive control of non–linear systems.

II. Optimal Predictive Control

An optimal control input to the objective plant is determined by minimising the objective function J at each time step k: $J = \sum_{j=1}^{N_p} \|\hat{\epsilon}_{k+j|k}\|^2 + \sum_{j=1}^{N_u} \lambda_j \|\Delta u_{k+j-1}\|^2$, where N_p is the prediction horizon for the plant output $y_k \in \mathbb{H}$, $\hat{\epsilon}_{k+j|k} \in \mathbb{H}$ is the predicted control error defined by the difference between the predicted plant output $\hat{y}_{k+j|k} \in \mathbb{H}$ and reference output $r_{k+j} \in \mathbb{H}$, N_u is the control horizon for the control input $u_k \in \mathbb{H}$, $\Delta u_{k+j-1} \in \mathbb{H}$ is the control input increment, and $\lambda_j \in \mathbb{R}$ is the weight coefficient. A multi–layer QNN was used as the model of the objective plant to estimate the predictions of the plant output for the whole prediction horizon at each time k. The external input to the network consists of the tapped delay lines of the plant's input–output. The QNN was trained offline to minimise a real–valued cost function defined by the norm of the identification error defined by the difference between the plant and network outputs using a stochastic gradient descent method extended to quaternion systems on the input–output set of the plant. We introduced a metaheuristics optimisation method, such as simulated annealing (SA), genetic algorithm (GA), and particle swarm optimisation (PSO), which allowed us to minimise the objective function J without solving the non–linear programme.

III. Computational Experiments

To investigate the capability of using the QNN in the optimal predictive control, computational experiments were conducted using the following objective plant: $y_k = 1.3y_{k-1} - 0.3y_{k-2} + 0.03y_{k-3} + u_{k-1} + 0.7u_{k-2} + 0.2y_{k-1}^2 + \zeta_{k-1} - 0.3\zeta_{k-2}$, where $\zeta_k \in \mathbb{H}$ is the random noise. The control input had an inequality constraint of $|u_{k_i}| \leq 0.5$. The QNN had a 4–16–1 network topology. The total number of training datasets was 10^4. In the objective function J, N_p and N_u were set to one and λ_1 was set to 0.1. The control task of ensuring that the plant output followed the reference output could be achieved with the acceptable accuracy (Fig. 1). The mean squares of the control error with SA, GA and PSO were 0.0245 ± 0.0015, 0.0296 ± 0.0024 and 0.0240 ± 0.0011, respectively. The Welch one–way ANOVA test with a significance level of 5% showed statistical differences between GA and SA/PSO.

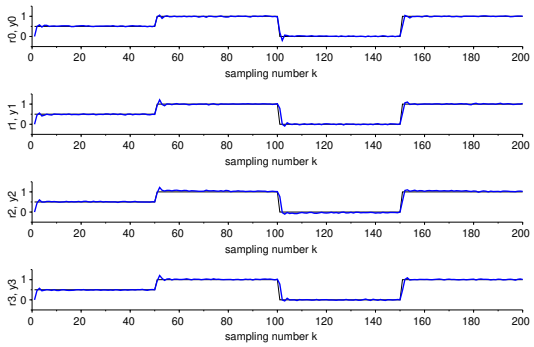


Fig. 1. Example of system responses using SA, where the reference changes 0.5, 1, 0 and 1 every 50 samples. The reference r_k and the plant output y_k are denoted using thin black and blue lines, respectively

IV. Conclusions

To achieve an optimal predictive control for non–linear systems, we applied a QNN to a prediction model in the MPC approach and an optimal control input was defined by minimising the quadratic objective function composed of the predicted control error and control input increment using a metaheuristics optimisation. Our findings confirmed that the optimal predictive control using the QNN could ensure the control task with the required performance.

References

[1] S. Pouyanfar, *et al.*, "A survey on deep learning: Algorithms, techniques, and applications", *ACM Comput Surv*, 51(5), 92:1–92:36, 2018.
[2] E. Bayro-Corrochano, "A survey on quaternion algebra and geometric algebra applications in engineering and computer science 1995–2020", *IEEE Access*, 9, 104326–104355, 2021.
[3] M. Schwenzer, *et al.*, "Review on model predictive control: an engineering perspective", *The Int. J. Adv. Manuf. Technol.*, 117, 1327–1349, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Classification of Normal vs. Pathological Infant Cries Using Morse Wavelets

Priyanka Gupta, Aastha Kachhi, and Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India.
E-mail: {priyanka_gupta, aastha_k, hemant_patil}@daiict.ac.in

***Abstract*—This study proposes Morse wavelet-based scalogram features for classification of normal *vs.* pathological infant cry. To the best of the authors' knowledge and belief, this study is the first of its kind to exploit analytic wavelet-based features for this task. For comprehensive evaluation of the proposed feature set, two datasets have been considered in this work, namely, Baby Chillanto, and DA-IICT corpora, and a third dataset is created by the combining these corpora. Furthermore, three data augmentation techniques (namely, tempo, volume, and speed perturbation) have also been applied on these datasets, and the performance of the proposed feature set is evaluated. Our study finds that the proposed Morse wavelet-based features shows the best performance of 98.19% (94.51%) as overall accuracy without augmentation, with an absolute improvement of 6.03% (6.06%) and 0.88% (5.10%) in accuracy compared to Mel-STFT and CQT baselines, respectively, on Baby-Chillanto (DA-IICT) datasets. Further, the performance of wavelet-based features outperforms the Mel-STFT and CQT on Baby Chillanto dataset, whereas it gives comparable results on DA-IICT and combined datasets.**

Index Terms: Infant cry classification, Morse wavelet, scalogram, data augmentation.

I. Introduction

Crying is the known mode of communication for infants, which is an essential evolutionary signal that enables infants to convey discomfort. During the first three months after birth, infant cries are known to carry neurological and health status of the infant [1]. About 2.5 million infants succumb to various vaccine-preventable and other ailments within initial months of birth. Several diagnosis methods, such as magnetic resonance imaging, computed tomography scans, and head ultrasound are used to identify various pathologies. However, pathology diagnosis is costly, and also time taking, which impacts the health of the baby. For example, asphyxia is identified by pale and bluish limbs. However, by the time these visual symptoms are visible, notable neurological damage to neonate would have already been occurred [2]. Hence, the need to develop a robust assistive pathology detection algorithm to aid paediatricians in identifying such pathologies is increasing [3]. Some of the prominent causes of infant's death are asthma, Sudden Infant Death Syndrome (SIDS), and asphyxia.

The infant cry classification and analysis involves neurological, physiological, psychology, engineering, developmental linguist, and paediatrics making this field highly interdisciplinary in nature. Early, in 1960s, the preliminary analysis of infant cries for pathological cry detection was pioneered using spectrograms by *Xie et. al.* [4]. The characteristics of normal cry sounds, generally defined as "cry modes" or "cry phonemes," including vibration, dysphonation, inhalation, and hyperphonation, were investigated for both the time and frequency-domains [4]. This study was extended to the pathological cries in [5], where some of these cry modes were found to be correlated with pathology. In [6], infants are found to have melodic pattern in their cry. To that effect, a recent Constant Q Transform Cepstral Coefficients (CQCC)-based study was also explored in [7] where the *form-invariance* property of Constant Q Transform (CQT) was explored for infant cry classification. Given CWT is a type of CQT, in this study, we explore the effectiveness of Continuous Wavelet Transform (CWT) (i.e., scalogram-based features) for infant cry classification task. Furthermore, similar to Mel-Short Time Fourier Transform (Mel-STFT), and CQT, the CWT-based scalogram features are also known to have better resolution in lower frequency regions [8], [9]. In addition to the existing baselines reported in [10], [11], [12] that use spectrograms (with or without mel scale) as input to neural network-based classifiers, we use CQT-based baseline as well, and we propose Morse wavelet-based features for infant cry classification. Moreover, the spectrogram has fixed time-frequency resolution in the entire time-frequency plane and hence, it fails to preserve *form-invariance* property of Fourier transform, i.e., $F\{x(\alpha t)\} = \frac{1}{|\alpha|} X\left(\frac{\omega}{\alpha}\right)$, because the analysis window used in STFT is a function of *only* time parameter.

It should also be noted that there is existing literature which uses wavelets for infant cry classification [13], [14], [15]. However, the existing works propose the use of variants of Daubechies wavelets, which differ from our proposed work in the following 3 aspects:

- The existing approaches use **real** wavelets, whereas, our work proposes the use of **analytic** wavelets which are known to capture instantaneous frequency. Moreover, Morse wavelets show *strict analyticity* which enables **no spectral leakage** (as discussed later in this paper) and superior time localization.
- The existing approaches are based on DWT-based features, where our work is based on the analytic CWT. The analytic CWT is robust to noise and stable to perturbations and higher-order modulations, which makes it more suitable for feature extraction as compared to DWT.
- Our work additionally addresses the issue of selection of an appropriate wavelet by using *Generalized* Morse

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

wavelets as a superfamily of analytic wavelets. To the best of the authors' knowledge, this issue of selection of appropriate wavelet has not been addressed in the literature, especially for infant cry classification. Therefore, our work exhibits novelty in these three aspects.

In particular, we present the following contributions in this paper:

- We propose Generalized Morse Wavelet (GMW)-based features for pathology detection from infant cries.
- Given that Baby Chillanto, and DA-IICT datasets have utterances recorded in different acoustic conditions, performance is also evaluated on a third dataset - combined dataset.
- Three data augmentation techniques (speed, volume, and tempo perturbation) on all the three datasets.
- For all these datasets, the proposed features are evaluated for two cases (i.e., with and without augmentation).

II. Proposed GMW-based Features

Analytic wavelets are *causal* in the frequency-domain, and are of different types, including complex Shannon, log-normal, Gaussian derivative, Cauchy, and Morlet wavelets [16], [17], [18], [19]. However, as there are numerous wavelet types, selecting the most suitable one for a given task is a significant consideration. GMWs, on the other hand, are regarded as a super-family of analytic wavelets, which addresses the problem of choosing a specific type of analytic wavelet [20], [21], [22]. Moreover, infant cry is an oscillatory signal, and its zero-crossings capture important acoustic information pertaining to possible pathology. Thus, we employ Morse wavelet that is analytic and hence, it can capture the instantaneous frequency content in the cry segment.

GMWs are a set of eigenvectors w.r.t. to a time-frequency localization operator, whose details can be studied from [23]. In this work, out of these set of eigenvectors, only the 0^{th} GMW is considered because it corresponds to the *largest* eigenvalue, and is known for its *optimal energy concentration* properties [23], [24]. To that effect, we represent the 0^{th} GMW as the Morse wavelet, which is mathematically expressed in the frequency-domain as [20], [21], [24]:

$$\Psi_{\beta,\gamma}(\omega) = \int_{-\infty}^{+\infty} \psi_{\beta,\gamma}(t) e^{-j\omega t} dt = U(\omega) a_{\beta,\gamma} \omega^{\beta} e^{-\omega^{\gamma}}, \quad (1)$$

where β and γ control the *shape* and *size* of the wavelet, as shown in Fig. 1. Furthermore, $a_{\beta,\gamma}$ is a real-valued normalizing constant, and $U(\omega)$ is the unit-step function in the frequency-domain due to the causality of $\Psi_{\beta,\gamma}(\omega)$. The Morse wavelet, however, is better analyzed using its dimensionless derivatives [22]. To that effect, its 2^{nd} derivative is called the *wavelet duration* or *inverse bandwidth* and is denoted as $P^2_{\beta,\gamma}$. The significance of 2^{nd}-order derivative is discussed in detail in the Appendix of this paper. The quantity $P^2_{\beta,\gamma}/2\pi$ sets the number of oscillations in the wavelet, where $P^2_{\beta,\gamma}$ is given as [22]:

$$P^2_{\beta,\gamma} \equiv -\frac{\omega^2_{\beta,\gamma} \Psi''_{\beta,\gamma}(\omega_{\beta,\gamma})}{\Psi_{\beta,\gamma}(\omega_{\beta,\gamma})} = \beta\gamma. \quad (2)$$

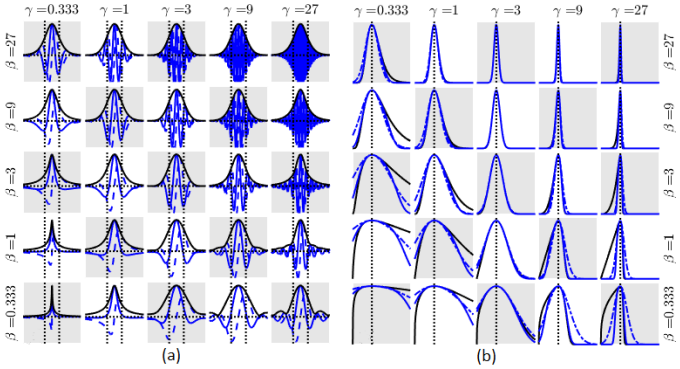


Fig. 1. Morse wavelets for varying values of β and γ in (a) time-domain, and (b) frequency-domain. After [21].

When the parameter $P^2_{\beta,\gamma}$ is kept fixed, varying the parameter γ, shifts the wavelet transform between different families of wavelets of fixed durations. Therefore, the parameter γ is termed as *family*. In this work, the Morse wavelet with parameter γ = 3 (also known as *'Airy family'*) is chosen because we get the most optimum Heisenberg area $A_{\beta,\gamma}$ for γ = 3 even for a small wavelet duration (as shown in Fig. 2) [21]. The minimum value of $A_{\beta,\gamma}$ is governed by $\sigma_t^2 \cdot \sigma_\omega^2 \geq 1/4$, where σ_t^2 and σ_ω^2 are the variances in time and frequency domains, respectively [17]. Further details of the Heisenberg's area $A_{\beta,\gamma}$ can be studied from [8], [24]. Additionally, at $\gamma = 3$,

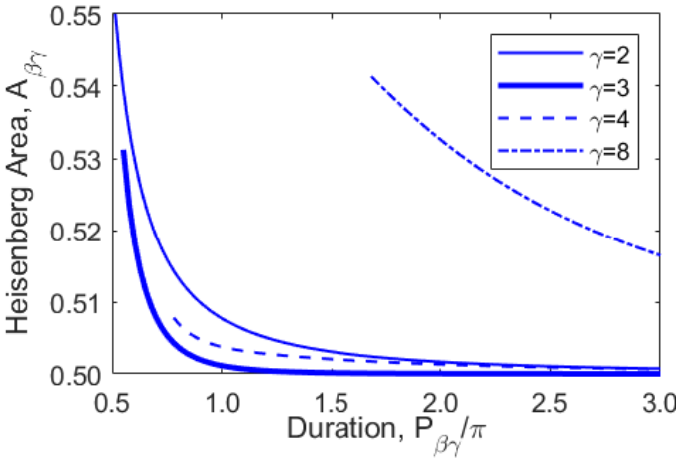


Fig. 2. Effect of γ parameter on the time-frequency Heisenberg area $A_{\beta,\gamma}$ w.r.t. wavelet duration $P_{\beta,\gamma}/\pi$. After [21].

the Morse wavelet is said to be similar to the conventional Morlet wavelet. However, the conventional Morlet wavelet defies analyticity property under special cases, but the Morlet-like wavelet achieved from by taking $\gamma = 3$ in Morse wavelet, shows *strictly* analytic behaviour. In other words, as observed in Fig. 3, a *spectral leakage* can be observed in the case of the conventional Morlet wavelet, whereas the Morse wavelet remains to be strictly analytic, i.e., its spectrum *does not leak* to the negative frequency region, i.e., it is zero in the negative frequencies as seen in Fig. 3 (d).

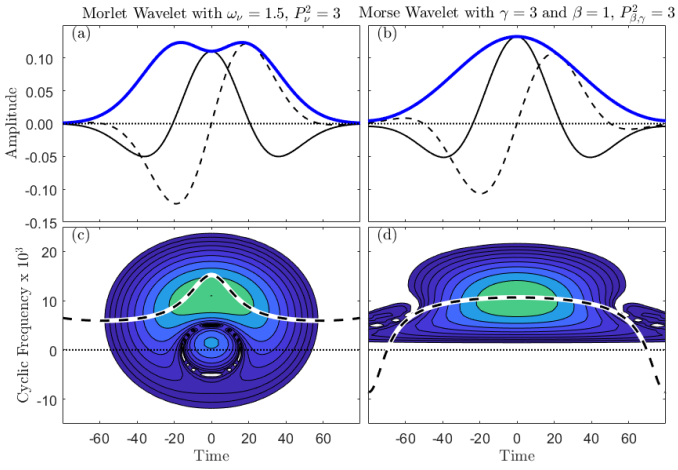


Fig. 3. Analysis of spectral leakage (a) Morlet wavelet, *vs.* (b) Morse wavelet and (c), (d) their respective Weigner-Ville distributions.

A. Morse Wavelet-Based Feature Extraction

The CWT of a signal $x(t)$ is estimated as

$$W_a x(u,s) =< x(t), \psi_{u,s}(t) >, \\ = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t)\psi^* \left(\frac{t-u}{s}\right) dt, \quad (3)$$

where $< \cdot, \cdot >$ indicates inner product operation to compute wavelet coefficients, and $*$ denotes the complex conjugate operator. The dilation (scaling) coefficient is denoted by s, and translational (positional) coefficients are denoted by u. From Fig. 4 (a), it can be observed that the scalogram corresponding to healthy (normal) cry exhibits a *continuous* band of energy in the high frequency region (particularly near 1 kHz, as shown by the red box). However, a clear distinction can be observed in Fig. 4 (b), which shows scalogram corresponding to pathological (asphyxia) infant cry. In particular, it can be observed that there is *absence* of continuous band of energy in this case. Moreover, a *short burst of energy* (as indicated by the red box) is observed in the lower frequency region, indicating the possible presence of pathology. In this work, we extract CWT-based representations (i.e., scalograms), as described in Algorithm 1. These scalogram-based features are then fed as input to the GMM classifier.

Algorithm 1: MATLAB pseudocode for the proposed Morse wavelet-based scalogram extraction

Input: Speech signal $x(t)$
Output: scalo_image
w_name='morse' // Taking Morse wavelet
[cwt_coeffs, F] ← cwt(x(t), w_name)
/* Mapping CWT coefficients to scalogram RGB image of 512×512. */
tpoints ← time vector of $x(t)$
/* Converting to Pseudocolor image */
pcolor(tpoints, cwt_coeffs)
/* Converting scalorgram as RGB color image */
scimg=ind2rgb(rescale(cwt_coeffs), jet(320));
scalo_image← imresize(scimg, 512, 512);

III. Experimental Setup

A. Datasets Used

In this study, we use three datasets, namely, Baby Chillanto database, the DA-IICT corpus, and the combined (Baby Chillanto + DA-IICT) dataset. Baby Chillanto dataset was designed using recordings made by doctors and is the property of Mexico's NIAOE-CONACYT and its statistics can be studied from [25], [26]. The second dataset known as the DA-IICT corpus was collected by [27], [28]. The third dataset called as combined dataset consists of all utterances of Baby Chillanto dataset along with DA-IICT corpus, with 1842 utterances in the healthy class, and 1616 utterances in the pathology class. For fair experimentation, we have resampled the utterances of all the datasets to a uniform sampling frequency of 16 kHz. Experiments in this study are performed using *10*-fold cross-validation (CV) method. Table I shows the data partitioning of each of the *10*-folds.

TABLE I
Number of Utterances in Data Partitions in Each Fold

Dataset →	Baby Chillanto	DA-IICT	Combined
Train	2041	1071	3112
Test	227	119	346
Total	2268	1190	3458

B. Data Augmentation

To increase the robustness towards *realistic* scenarios, data augmentation has also been performed on all the datasets. Data augmentation is particularly useful for infant cry classification problem because of the limited publicly available datasets. The data augmentation in this work has been performed using "SOX" tool for three augmentation techniques, namely, speed, tempo, and volume perturbation. For speed augmentation, the audio is perturbed *w.r.t.* time to get the speed augmented utterance, expressed as $y(t)$=$s(b*t)$, where $s(t)$ is the input cry utterance, and b is scaling factor [29]. Tempo augmentation is a STFT-based perturbation technique where speech segment is re-positioned based on speech synthesis instants as in [30]. For volume augmentation, the utterance is scaled as $y(t)$=$b * s(t)$. The value of b takes values as 0.9 and 1.1 for speed and volume perturbation each.

C. Classifier Used

The classifier used in this work is Gaussian Mixture Model (GMM). For the case of Baby Chillanto dataset, the number of mixtures taken is 512, while it is 128 for the DA-IICT corpus. The difference in the number of mixtures for both the datasets is due to the difference in the size (i.e., the number of cry utterances) of both the datasets.

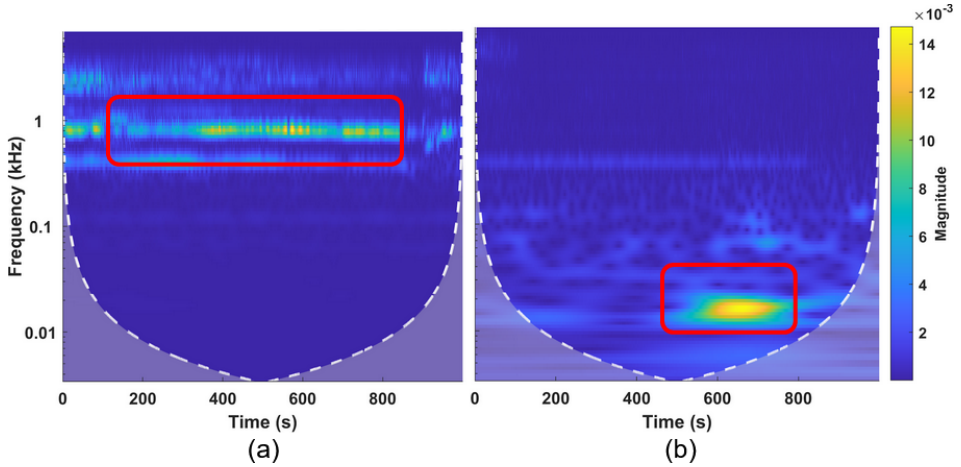


Fig. 4. Proposed Morse scalogram-based analysis for (a) healthy, and (b) pathology (asphyxia) infant cry.

Moreover, it should be noted that due to the relatively smaller test size of the dataset, neural network-based classifiers, such as CNN could not be used as it results in *overfitting*. However, we also performed experiments using SVM as the classifier to observe the effect of the classifier.

D. Baselines

Given that the proposed feature set is *non-cepstral*, in this work, for fair comparison, two non-cepstral baselines are considered, namely, Mel STFT, and CQT. Mel STFT-based features are evaluated by keeping window size of 30 ms and window shift of 15 ms, whereas the CQT-based baseline is using 96 bins and F_{min} = 100 Hz. It should be noted that both the baselines and the proposed feature set have *high* resolution in the lower frequency regions.

IV. Experimental Results

A. Effect of Classifier Structure

Given the small size of the test data, neural network based classifiers such as CNNs are not considered, to observe the effect of the classifier. Table II shows the performance of Mel STFT, CQT, and the proposed Morse wavelet-based scalogram features using SVM and GMM classifiers on the standard Baby Chillanto dataset. It can be observed that the proposed features outperform Mel STFT and CQT features for both SVM and GMM. However, since GMM gives absolute increment of 7% on all feature sets as compared to SVM, the remaining experiments in this paper are performed by considering GMM-based classifier.

TABLE II
% Accuracies using SVM and GMM

Classifier	Mel STFT	CQT	Morse Wavelet-based Scalogram
SVM	88.55	87.84	**90.75**
GMM	92.16	97.31	**98.19**

B. Effect of Wavelet Duration ($P^2_{\beta,\gamma}$)

In this sub-Section, we present the results obtained by varying the wavelet duration parameter $P^2_{\beta,\gamma}$ of the Morse wavelet. Fig. 5 shows the effect of $P^2_{\beta,\gamma}$, keeping γ fixed as 3, on the two datasets. It can be observed from Fig. 5 that the results are degraded for relatively lower values of $P^2_{\beta,\gamma}$, more so, for the DA-IICT corpus. From eq. (2), it can be deduced that for a fixed value of γ (i.e., $\gamma = 3$), a low value of $P^2_{\beta,\gamma}$ corresponds to a low value of β. Therefore, in Fig. 1, as we keep $\gamma = 3$ fixed, and move towards lower values of β (i.e., lower values of wavelet duration $P^2_{\beta,\gamma}$), poor quality subband filters (with low quality factor, i.e., with increased bandwidth) are produced. Therefore, the performance is degraded for lower values of wavelet duration for both the datasets. Furthermore,

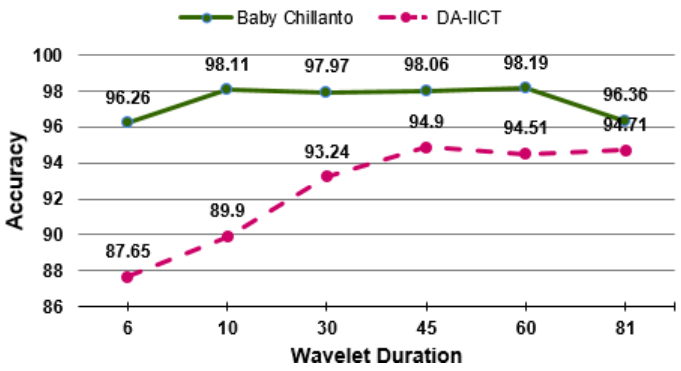


Fig. 5. Effect of the Morse wavelet parameter $P^2_{\beta,\gamma}$ on Baby Chillanto, and DA-IICT corpora.

the optimal value of $P^2_{\beta,\gamma}$ obtained is 60 for the case of Baby Chillanto dataset, and 45 for the case of DA-IICT corpus. Given that Baby Chillanto is statistically significant standard dataset for infant cry classification problem, we choose to keep $P^2_{\beta,\gamma}$ as 60, for the remaining set of experiments.

C. Overall Performance

As discussed in sub-Section IV-B, keeping the parameter $P^2_{\beta,\gamma}$ as 60, we now present the overall performance in terms

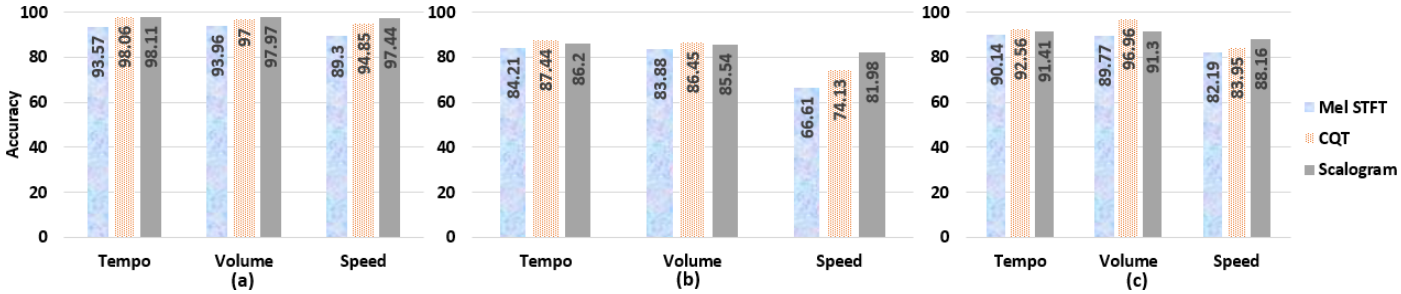


Fig. 6. Effect of three data augmentation techniques (i.e., tempo, volume, and speed perturbations), on (a) Baby Chillanto, (b) DAIICT, and (c) combined corpora. Best viewed in color.

of % accuracy, obtained on the three datasets used in this work. To that effect, Table III shows the performance of Mel STFT, CQT, and the proposed Morse wavelet-based scalogram features. It can be observed that the proposed features

TABLE III
OVERALL PERFORMANCE (IN % ACCURACY) OF BASELINES AND THE PROPOSED FEATURES ON THE THREE DATASETS

Datasets →	**Baby Chillanto**	**DA-IICT**	**Combined**
Mel-STFT	92.16	87.45	89.57
CQT	97.31	89.41	**92.05**
Morse wavelet -based scalogram	**98.19**	**94.51**	91.70

outperform Mel STFT and CQT features on Baby Chillanto and DA-IICT corpora. On the Baby Chillanto dataset, the absolute improvement in % accuracy is 6.03 and 0.88, for Mel STFT and CQT, respectively. On the DA-IICT dataset, the absolute improvement in % accuracy is 7.06 and 5.1, for Mel STFT and CQT, respectively. Furthermore, for the combined dataset, CQT still remains to be performing the best. However, the absolute difference of % accuracy of CQT and Morse wavelet-based scalogram features is only 0.35. This marginally improved performance of CQT can be due to the data partitioning effect on the combined data.

D. Effect of Data Augmentation

In this sub-Section, the results (in % accuracy) pertaining to the three data augmentation techniques are presented, namely, tempo, volume, and speed perturbation. Fig. 6 (a) shows the results pertaining to Baby Chillanto dataset, wherein the performance on three types of data augmentation techniques (tempo, volume, and speed perturbation) is shown. It can be observed that Morse wavelet-based scalogram approach outperforms the existing approaches for *all* the three augmentation techniques in the case of Baby Chillanto dataset. Furthermore, Fig. 6 (b) shows the results pertaining to DA-IICT dataset. It can be observed that CQT-based method is observed to perform the best for tempo and volume perturbation (with accuracies of 87.44% and 86.45%, respectively), whereas for speed perturbation, Morse wavelet-based features show the best performance of 81.98% accuracy. A similar trend can be observed in Fig. 6 (c), which shows the results pertaining to the combined dataset. Furthermore, volume perturbation should ideally not affect classification, however, it does have a detrimental effect based on the results shown in Fig. 6, thereby interfering with pathological markers.

It is worth noting that *all* the datasets, the performance of speed augmentation is relatively lower than tempo and volume perturbation for each of the Mel STFT and CQT-based features. However, the performance degradation is negligible for the proposed Morse wavelet-based scalogram feature set.

V. SUMMARY AND CONCLUSIONS

This study proposed the use of Morse wavelet-based scalograms features for the classification of normal *vs.* pathological infant cries. Given the form-invariance property of CQT, the proposed features are CWT-based, which is also a type of CQT. However, while dealing with CWT, choosing an appropriate wavelet function is an issue. Therefore, we proposed to exploit GMWs, which encompass most of the families of analytic wavelets. To that effect, we showed detailed analyses pertaining to the choice of Morse wavelet parameters, which are followed by experimental fine-tuning. Moreover, we evaluated the performance of the proposed system using three datasets, namely, Baby Chillanto, DA-IICT, and combined dataset. Lastly, we also show the effect of three data augmentation techniques (tempo, volume, and speed perturbation). Our results, particularly on volume perturbation, indicate that it is likely that the augmentation has interfered with pathological markers, and hence, an acceptable method of volume augmentation which does not interfere with pathological markers is a future research direction.

VI. ACKNOWLEDGEMENTS

The authors sincerely thank MeitY, Govt. of India, for project 'Speech Technologies in Indian Languages' BHASHINI', (Grant ID: 11(1)2022-HCC (TDIL)).

APPENDIX

The 2^{nd} order derivative $P_{\beta,\gamma}$ is the second-order term in the time-domain Taylor expansion of the wavelet demodulated by its own peak frequency. The following discussion derives the expression and assesses the significance of the 2^{nd} order derivative.

For a complex random variable X, the n^{th} moment is $E[X^n]$ which is estimated as [31]

$$E[X^n] = \int_{-\infty}^{+\infty} x^n f_X(x) dx, \tag{4}$$

where $f_X(x)$ is the probability density function of X. The characteristic function $C_x(t)$ is defined as

$$C_x(t) = E[e^{-itx}] = \int_{-\infty}^{+\infty} e^{-itx} f_X(x) dx, \tag{5}$$

where $f_X(x)$ is the probability density function of X. Interestingly, *the characteristic function of X is nothing but the Fourier transform of $f_X(x)$*. The moments ($E[X]$, $E[X^2]$, $\dots E[X^n]$) can be generated by differentiating the characteristic function $C_x(t)$ or by differentiating the Fourier transform of $f_X(x)$.

Now, consider $f_X(x)$ as the Morse wavelet demodulated by its peak frequency ω_ψ, and denoting it by $f(t)$ i,e.,

$$f(t) = \psi(t) e^{-i\omega_\psi t}. \tag{6}$$

Then the time-domain n^{th} order moment of the demodulated wavelet can be written as

$$m_{n;\psi} \equiv \int_{-\infty}^{+\infty} t^n f(t) dt = \int_{-\infty}^{+\infty} t^n \psi(t) e^{-i\omega_\psi t} dt. \tag{7}$$

Given the relationship between the time domain moments and the frequency domain derivatives

$$\frac{m_{n;\psi}}{m_{0;\psi}} = i^n \frac{\psi^{(n)}(\omega_\psi)}{\tilde{\Psi}(\omega_\psi)}, \tag{8}$$

where $\tilde{\Psi}(\omega_\psi)$ is the wavelet's *dimensionless derivatives* at $\omega = \omega_\psi$, and defined as [22], [24]:

$$\tilde{\Psi}_n(\omega) \equiv \omega^n \frac{\Psi^{(n)}(\omega)}{\Psi(\omega)}, \tag{9}$$

where n as superscript denotes the n^{th} derivative and $\tilde{\Psi}(\omega_\psi)$ is nothing but $\tilde{\Psi}_n(\omega)$ evaluated at $\omega = \omega_\psi$. Therefore,

$$\frac{m_{n;\psi}}{m_{0;\psi}} = i^n \frac{\psi^{(n)}(\omega_\psi)}{\tilde{\Psi}(\omega_\psi)} = i^n \frac{\tilde{\Psi}_n(\omega_\psi)}{\omega_\psi^n}. \tag{10}$$

Since the wavelet satisfies the admissibility criteria, its mean is zero; i.e., $m_{1;\psi} = 0$. Therefore, the higher-order demodulate moments can be normalized by the 2^{nd} moment, and hence, the dimensionless measure using the 2^{nd} moment is defined as

$$P_{\beta,\gamma} = \omega_\psi \sqrt{\frac{m_{2;\psi}}{m_{0;\psi}}} = \sqrt{|\tilde{\Psi}_2(\omega_\psi)|}, \tag{11}$$

where $\tilde{\Psi}_2(\omega_\psi) = -\beta\gamma$ (as shown in Appendix D of [22]).

REFERENCES

[1] S. A. Reijneveld, E. Brugman, and R. A. Hirasing, "Excessive infant crying: the impact of varying definitions," *Pediatrics*, vol. 108, no. 4, pp. 893–897, 2001.

[2] C. C. Onu, J. Lebensold, W. L. Hamilton, and D. Precup, "Neural transfer learning for cry-based diagnosis of perinatal asphyxia," *ICLR Workshop, Graz, Austria*, 2019.

[3] K. Manickam and H. Li, "Complexity analysis of normal and deaf infant cry acoustic waves," in *Fourth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, october 29-31, 2005, Firenze, Italy, pp. 105–108.

[4] Q. Xie, R. K. Ward, and C. A. Laszlo, "Determining normal infants' level-of-distress from cry sounds," in *Proceedings of Canadian Conference on Electrical and Computer Engineering, Vancouver, BC, Canada*, 1993, pp. 1094–1096.

[5] H. A. Patil, "Cry baby: Using spectrographic analysis to assess neonatal health status from an infant's cry," in *Amy Neustein (Ed.), Advances in Speech Recognition*. Springer, 2010, pp. 323–348,.

[6] L. Armbrüster, W. Mende, G. Gelbrich, P. Wermke, R. Götz, and K. Wermke, "Musical intervals in infants' spontaneous crying over the first 4 months of life," *Folia Phoniatrica et Logopaedica*, vol. 73, no. 5, pp. 401–412, 2021.

[7] H. A. Patil, A. T. Patil, and A. Kachhi, "Constant Q Cepstral Coefficients for classification of normal *vs*. pathological infant cry," in *ICASSP, Singapore*, 2022, pp. 7392–7396.

[8] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[9] P. Gupta, P. K. Chodingala, and H. A. Patil, "Morse wavelet features for pop noise detection," in *SPCOM, IISc Bangalore, India*, 2022, pp. 1–5.

[10] J. Chunyan, M. Chen, L. Bin, and Y. Pan, "Infant cry classification with graph convolutional networks," in *ICCCS*. IEEE, 2021, pp. 322–327.

[11] C. Ji, Y. Jiao, M. Chen, and Y. Pan, "Infant cry classification based-on feature fusion and mel-spectrogram decomposition with cnns," in *International Conference on AI and Mobile Services,*. Springer, 2022, pp. 126–134.

[12] H.-N. Ting, Y.-M. Choo, and A. A. Kamar, "Classification of asphyxia infant cry using hybrid speech features and deep learning models," *Expert Systems with Applications*, vol. 208, p. 118064, 2022.

[13] M. Hariharan, S. Yaacob, and S. A. Awang, "Pathological infant cry analysis using wavelet packet transform and probabilistic neural network," *Expert systems with applications*, vol. 38, no. 12, pp. 15 377–15 382, 2011.

[14] J. Saraswathy, M. Hariharan, V. Vijean, S. Yaacob, and W. Khairunizam, "Performance comparison of daubechies wavelet family in infant cry classification," in *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, 2012, pp. 451–455.

[15] M. Hariharan, R. Sindhu, V. Vijean, H. Yazid, T. Nadarajaw, S. Yaacob, and K. Polat, "Improved binary dragonfly optimization algorithm and wavelet packet based non-linear features for infant cry classification," *Computer Methods and Programs in Biomedicine*, vol. 155, pp. 39–51, 2018.

[16] M. Holschneider, "Wavelets : An Analysis Tool," *Oxford Science Publications*, 1995.

[17] S. G. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 2^{nd} Edition, 1999.

[18] H. Knutsson, C.-F. Westin, and G. Granlund, "Local multiscale frequency and bandwidth estimation," in *Proceedings of* 1^{st} *International Conference on Image Processing*, vol. 1. ICIP, 13-16 Nov. 1994, pp. 36–40.

[19] X. Zhu and J. Kim, "Application of analytic wavelet transform to analysis of highly impulsive noises," *Journal of Sound and Vibration*, vol. 294, no. 4-5, pp. 841–855, 2006.

[20] S. C. Olhede and A. T. Walden, "Generalized morse wavelets," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2661–2670, 2002.

[21] J. M. Lilly and S. C. Olhede, "Generalized morse wavelets as a superfamily of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6036–6041, 2012.

[22] Lilly, Jonathan M. and Olhede, Sofia C., "Higher-order properties of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 146–160, 2008.

[23] I. Daubechies and T. Paul, "Time-frequency localization operators-a geometric phase space approach: II. The use of dilations," *Inverse problems*, vol. 4, no. 3, p. 661, 1988.

[24] J. M. Lilly and S. C. Olhede, "On the analytic wavelet transform," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 4135–4156, 2010.
[25] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi, "Classifying infant cry patterns by the genetic selection of a fuzzy model," *Biomedical Signal Processing and Control*, vol. 17, pp. 38–46, 2015.
[26] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, "Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach," in *2008 Mexican International Conference on Computer Science*. IEEE, 2008, pp. 261–267.
[27] N. Buddha and H. A. Patil, "Corpora for analysis of infant cry," O-COCOSDA, Vietnam 2007.
[28] A. Chittora and H. A. Patil, "Data collection of infant cries for research and analysis," *Journal of Voice*, vol. 31, no. 2, pp. 252–e15, 2017.
[29] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH, Dresden, Germany*, 2015.
[30] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *ICASSP*, vol. 2, 1993, pp. 554–557.
[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Compressive Sensing Based Algorithms for Limited-View PAT Image Reconstruction

Mary Josy John and Imad Barhumi
College of Engineering, United Arab Emirates University, Al Ain, UAE.
201990198@uaeu.ac.ae, imad.barhumi@uaeu.ac.ae

***Abstract*—Limited-view sensor arrangement is a major concern in medical imaging as it limits the data that the sensor could acquire. However, this limitation, signal sparsity, can be exploited using compressive sensing (CS) techniques to reconstruct high-resolution images. The objective of this research paper is to develop CS-based algorithms for reconstructing images in limited-view photoacoustic tomography. Various CS reconstruction algorithms and sensor arrangements were assessed to identify the optimal approach for reconstructing images from limited-view sensor data. The results show that the split Bregman total variation (SBTV)-l_1 CS algorithm is the most efficient for all sensor arrangements. The study also reveals that the convex sensor array yields the best results among all sensor arrangements. Additionally, the implementation of SBTV-l_1 using Cholesky factorization requires less computation time and is 10 to 15 times faster than the direct implementation.**

***Index Terms*—Photoacoustic Tomography, Compressive Sensing, Split Bregman, Limited view.**

I. Introduction

Medical imaging is a technology that is crucial for the analysis of drug distribution, disease diagnosis, and therapy. Radiographic imaging systems allow to save and easily access images or videos enabling accurate disease diagnosis. Without these developments, it would be impossible to identify these diseases before they progress to a far more serious level that endangered human life. There are several imaging techniques currently in use, including X-rays, optical coherence tomography (OCT), ultrasound scanning (USG), magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) [1]. MRI takes longer and costs more than an X-ray. In comparison to a CT scan, it exposes people to less radiation. However, X-rays are primarily employed to scan bones, and as they only provide a single image from a single angle, they cannot image tissues or give a detailed perspective of the body [2]. Photoacoustic tomography (PAT) is a novel medical imaging technique that addresses the limitations of existing methods while incorporating the desired benefits. PAT employs the principle of light absorption, which causes a localized temperature increase and generates sound waves or pressure differences. These sound waves are detected by a sensor array and used to produce images of specific tissues or regions of interest [3].

Given that an imaging system is modeled as a linear system, the least squares (LS) solution resolves it as precisely as possible [4]. However, the LS solution can suffer from being ill-posed, computationally complex, non-sparse, sensitive to noise, and non-unique. Regularization is a technique that can encourage the solution to exhibit sparsity, smoothness, or other desired structures that align with the properties of the signal by incorporating a penalty term. A common regularization technique in compressive sensing (CS) is the use of Lasso regularization, which involves adding an l_1-norm penalty to the objective function. This penalty discourages large coefficients in the solution and promotes small, sparse solutions, ultimately encouraging sparsity in the reconstructed signal [5]. Sparsity is necessary for reconstruction because it enables the reconstruction of the original signal from a small number of non-zero coefficients, resulting in less processing, less computing, and faster signal acquisition. Thus, by combining PAT and CS, the necessary signal can be reconstructed with very high contrast, good resolution, and faster computing speeds.

This paper is structured as follows: Section II presents an overview of the theory behind compressive sensing algorithms. Section III details the proposed work and the algorithms used. Section IV discusses the results obtained from the study. Finally, Section V presents the conclusions drawn from the study.

II. Theory

A. CS-PAT algorithms

In this section, a brief overview of the existing CS-PAT reconstruction algorithms is provided. In photoacoustic tomography (PAT), a laser is utilized to stimulate a specific region within a compressible medium, ensuring thermal confinement and minimal viscosity. This leads to the generation of sound waves or pressure variations [6]. The acoustic pressure can be described by the linear approximation of the wave equation, and its solution, representing the acoustic pressure at a specific location and time, is obtained through the application of the Green's function approach [7]. The Green's function corresponds to a solution of the wave equation that incorporates a localized source term at a specific point in space and time. By integrating the Green's function with the source term, the solution to the wave equation can be obtained for any arbitrary source distribution. In practice, the pressure is computed on grid points, with grid spacing adhering to the Nyquist criterion [8] to prevent aliasing. The grids are typically divided into an imaging grid and a k-space computation grid. The k-space representation of the solution involves expressing the wave equation in terms of the wavenumber (or frequency) components of the solution. By taking the Fourier transform of the equation,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the partial differential equation is transformed into a set of ordinary differential equations in terms of the wavenumber components. These equations can then be solved numerically using techniques such as the Fast Fourier Transform (FFT). The k-space computation is particularly advantageous when dealing with problems that exhibit periodicity or translation symmetry, as the Fourier transform naturally captures these properties. Additionally, the use of Fourier-based methods allows for high accuracy and spectral convergence, similar to the pseudospectral method.

The PAT imaging system is designed as a linear system as

$$\mathbf{y} = \mathbf{M}\mathbf{x}. \tag{1}$$

Here $\mathbf{y} \in \mathbb{R}^{N_s N_t \times 1}$ is the measurement vector, $\mathbf{M} \in \mathbb{R}^{N_s N_t \times N_{in}^2}$ is the system matrix and $\mathbf{x} \in \mathbb{R}^{N_{in}^2 \times 1}$ is the image to be reconstructed (in vectorized form); N_s, N_t and N_{in}^2 are the number of sensors, number of time samples and the number of imaging grid points respectively. The system matrix arises when discretizing the partial differential equations (PDE) problem, which involves approximating the continuous solution and differential operators on a discrete grid or mesh. The discretization divides the spatial and temporal domains into a set of discrete points or elements, and the system matrix relates the values of the solution at these discrete points. Once the system matrix is formed, it can be combined with the vector representing the right-hand side of the discretized equations, which contains the source terms and boundary conditions, to form a linear system of equations. This linear system can then be solved using various numerical techniques, such as direct methods (e.g., LU decomposition) or iterative methods (e.g., conjugate gradient method), to obtain the numerical solution of the PDE problem. In recent years, CS has emerged as a relatively new field that has garnered significant attention. It concerns the retrieval of a sparse or compressible signal from a small number of linear measurements, often substantially less than the signal's dimension. However, standard least squares regression may not be the most suitable approach in this context since it necessitates more measurements than the signal's dimension. To tackle this challenge, multiple regularized least squares regression techniques have been proposed, which exploit the sparsity or compressibility of the signal to enhance the recovery performance. These techniques play a crucial role in compressive sensing since they enable us to recover the signal accurately from the limited measurements. The first and the best way to enforce sparsity is to do the l_0 norm minimization. It is formulated as

$$\mathbf{x}^* = \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2 + \lambda \|\mathbf{x}\|_0 . \tag{2}$$

It is NP-hard and one approach to solving such a problem is to use a smoothed differentiable approximation of the l_0 norm. l_1-norm minimization-based optimization problems such as Basis Pursuit (BP) [9], and Least Absolute Shrinkage and Selection Operator (LASSO) [10] are also commonly used. The BP optimization problem is formulated as

$$\mathbf{x}^* = \min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{s.t}\ \mathbf{M}\mathbf{x} = \mathbf{y}, \tag{3}$$

and the LASSO problem is formulated as

$$\mathbf{x}^* = \min_{\mathbf{x}} \|\mathbf{M}\mathbf{x} - \mathbf{y}\|_2^2 \ \text{s.t}\ \|\mathbf{x}\|_1 , \tag{4}$$

The formulation (4) is utilized for scenarios when the measurement data is noisy. Two of the most common approaches for solving l_1-minimization is by using the CVX tool [11] and l_1-magic [12]. Both of these methods are time-consuming and computationally expensive [13], [14].

An l_2-norm when applied as a regularizer to a least-square (LS) problem causes diagonal loading [15] and is formulation as

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 , \tag{5}$$

where $\lambda > 0$. Alternatively, it is known as Tikhonov regularization [16]. Since it results in uniform distribution of error and a subpar reconstructed image, this kind of regularization is not preferred. To further exploit sparsity, $l_1 - l_2$ regularization which is non-convex and Lipschitz continuous is used for image reconstruction and is formulated as

$$\mathbf{x}^* = \min_{\mathbf{x}} \|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \ \text{s.t}\ \mathbf{M}\mathbf{x} = \mathbf{y}. \tag{6}$$

The two most popular techniques for implementing such a regularization are forward-backward splitting (FBS) [17], and difference of convex algorithms (DCA) [18].

The method of total variation regularization is a frequently employed technique in compressive sensing and is commonly used to preserve edges, denoising, and image enhancement. The method of total variation regularization entails augmenting the objective function of the least squares regression by introducing a penalty term. This penalty term minimizes the sum of absolute differences between neighboring pixels or samples in the signal. The introduction of this penalty term promotes sparsity in the gradient domain of the solution, resulting in improved recovery performance. It is formulated as

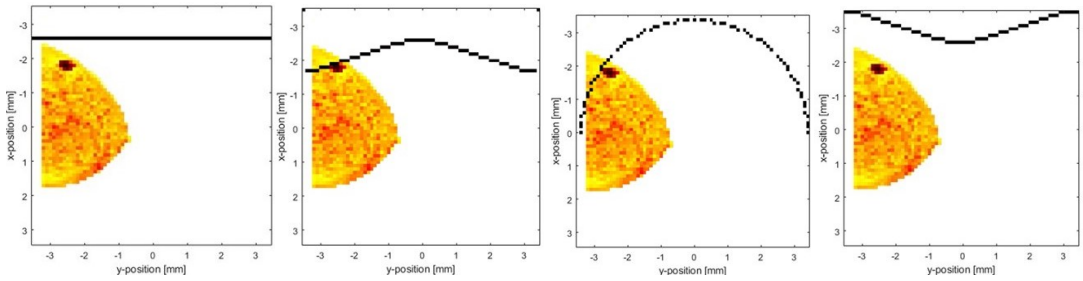
$$\mathbf{x}^* = \min_{\mathbf{x}} \|\mathbf{D}\mathbf{x}\|_1 \ \text{s.t.}\ \mathbf{M}\mathbf{x} = \mathbf{y}, \tag{7}$$

where $\mathbf{D}$ is a difference matrix defined as:

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \cdots & \cdots \\ & & & -1 \end{bmatrix}.$$

III. Proposed work

In this section two CS algorithms are proposed for the reconstruction of images from limited-view sensor arrangements. The first one is relaxed BP and the second one is l_1 regularized split Bregman (SB) TV (SBTV-l_1). The algorithms and their corresponding implementations are also discussed.



(a) (b) (c) (d)

Fig. 1: Sensor arrangement: (a) line sensor array, (b) concave sensor array (c) semicircular sensor array, and (d) convex sensor array

A. rBP-ADMM

Initially, BP-ADMM problem is formulated by introducing an auxiliary variable $\mathbf{z}$ and the augmented Lagrangian can be formulated as

$$L(\mathbf{x},\mathbf{z}) = \gamma \left\|\mathbf{z}\right\|_1 + \mathbf{u}^T(\mathbf{x}-\mathbf{z}) + \frac{\rho}{2}\left\|\mathbf{x}-\mathbf{z}\right\|_2^2 + \frac{\alpha}{2}\left\|\mathbf{M}\mathbf{x}-\mathbf{y}\right\|_2^2, \tag{8}$$

where $\mathbf{u}$ is the dual variable, ρ is the augmented Lagrangian parameter and $\alpha, \gamma > 0$. The rBP-ADMM includes a relaxation parameter r, applied to the $\mathbf{u}$-update in the $\mathbf{x}$-subproblem, which is proved to be much more efficient than its conventional BP-ADMM implementation. rBP-ADMM implementation is shown in Algorithm 1. Relaxation is necessary as the BP-ADMM is highly sensitive to modeling errors.

Algorithm 1 rBP-ADMM

Require: ρ, α, γ, r, maxiter, tol.
Initialize: $k = 0$, $\mathbf{z}^{(0)} = \mathbf{u}^{(0)} = 0$, NMSE^0 = ∞
while $k <$ maxiter $\wedge$ NMSE$^k >$ tol **do**
$\mathbf{x}^{k+1} \leftarrow \left(\rho\mathbf{I} + \alpha\mathbf{M}^T\mathbf{M}\right)^{-1}\left(\rho\mathbf{z}^k + r\mathbf{u}^k + \alpha\mathbf{M}^T\mathbf{y}\right)$
$\mathbf{z}^{k+1} \leftarrow \text{Shrink}\left(\frac{\mathbf{u}^k}{\rho} + \mathbf{x}^{k+1}, \frac{1}{\gamma}\right)$
$\mathbf{u}^{k+1} \leftarrow \mathbf{u}^k + \rho\left(\mathbf{x}^{k+1} - \mathbf{z}^{k+1}\right)$
NMSE$^{k+1} \leftarrow \frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2}{\|\mathbf{x}^{k+1}\|_2^2}$
$k \leftarrow k+1$
end while

B. SBTV-l_1

An unconstrained TV problem after the introduction of the auxiliary variable $\mathbf{d}$ is written as

$$\mathbf{x}_{TV} = \min_{\mathbf{x},\mathbf{d}} \left\|\mathbf{d}\right\|_1 + G(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}\|_1 + \frac{\gamma}{2}\left\|\mathbf{d}-\mathbf{D}\mathbf{x}\right\|_2^2, \tag{9}$$

where $G(\mathbf{x}) = \frac{\beta}{2}\|\mathbf{M}\mathbf{x}-\mathbf{y}\|_2^2$, and $\alpha, \beta, \gamma > 0$. Instead of solving the optimization problem in (9) in a single step, an SB formulation is employed which allows splitting the full optimization problem in (9) into its equivalent subproblems. The advantage of this method is that it allows for exact minimization, fast convergence, and less computation time. The subproblems are solved iteratively as depicted in Algorithm 2, where $\mathbf{b}$ is the Bregman parameter.

Algorithm 2 SBTV-l_1

Require: $\alpha, \beta, \gamma > 0$, tol, maxit
Initialize: $k = 0$, NMSE^0 = ∞, $\mathbf{d}^0 = \mathbf{b}^0 = \mathbf{0}$
while $k <$ maxiter $\wedge$ NMSE$^k >$ tol **do**
$\mathbf{x}^{k+1} \leftarrow$ Shrink$((\beta\mathbf{M}^T\mathbf{M} + \gamma\mathbf{D}^T\mathbf{D})^{-1}(\beta\mathbf{M}^T\mathbf{y} + \gamma\mathbf{D}^T(\mathbf{d}^k - \mathbf{b}^k)), \frac{1}{\alpha})$
$\mathbf{d}^{k+1} \leftarrow$ Shrink$(\mathbf{D}\mathbf{x}^{k+1} + \mathbf{b}^k, \frac{2}{\gamma})$
$\mathbf{b}^{k+1} \leftarrow \mathbf{b}^k + \mathbf{d}^{k+1} - \mathbf{D}\mathbf{x}^{k+1}$
NMSE$^{k+1} \leftarrow \frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2}{\|\mathbf{x}^{k+1}\|_2^2}$
$k \leftarrow k+1$
end while

IV. Results

This section presents a comparison of the performance of the proposed compressive sensing algorithms with existing methods. The performance evaluation is based on the following metrics:

1) Structural similarity index (SSIM): It is used to assess image quality and has a value range of -1 to 1. When the reconstructed image closely resembles the original image, its value is close to 1.
2) Peak Signal to Noise Ratio (PSNR): It is the ratio between the maximum signal power to the noise power. It is measured in decibels.
3) Number of iterations: The number of iterations is also taken into account as a metric for assessing how well iterative algorithms work.
4) Normalized mean-squared error (RMSE): It is the normalized mean-squared error between the reconstructed image and the original image defined as:
$$\text{NMSE} = \frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2^2}{\|\mathbf{x}\|_2^2} \tag{10}$$
where $\hat{\mathbf{x}}$ is the estimated image (or the reconstructed image from measurements).
5) Gini Index (GI): It reflects how sparse the signal is. The range is from 0 to 1, where 1 means highly sparse.
6) Computation time: It is the amount of time taken by an algorithm to complete the iteration and it depends on the

TABLE I: Limited view - concave sensor array test results

Algorithm used	SSIM	*PSNR (dB)*	iterations	*NMSE*	*GI*	*time (sec)*
SBTV	0.908	34.253	14	0.003	0.766	6.55
c-SBTV	0.908	34.253	14	0.003	0.766	0.81
SBTV-l_1	0.962	33.586	10	0.005	0.769	5.08
c-SBTV-l_1	0.962	33.586	10	0.005	0.769	0.56
c-rBP	0.894	27.805	53	0.034	0.762	8.57
DCA	0.574	26.033	56	0.124	0.760	4.95
FBS	0.560	17.603	101	0.637	0.819	53.78
BP-l_1-magic	0.999	52.529	8	1.0×10^{-4}	0.864	1382
LASSO-l_1-magic	0.999	52.692	18	9.8×10^{-5}	0.864	74.59
RLS	0.424	13.530	NA	0.995	0.728	49.62
LS	-1.0×10^{-10}	-106.94	NA	2.7×10^{11}	0.651	34.32

TABLE II: Limited view - line sensor array test results

Algorithm used	SSIM	*PSNR (dB)*	iterations	*NMSE*	*GI*	*time (sec)*
SBTV	0.871	31.418	17	0.006	0.765	0.80
c-SBTV	0.871	31.418	17	0.006	0.765	9.72
SBTV-l_1	0.938	30.986	12	0.010	0.770	5.08
c-SBTV-l_1	0.938	30.983	12	0.010	0.770	0.72
c-rBP	0.905	28.841	65	0.024	0.763	17.67
DCA	0.692	24.100	200	0.178	0.785	17.58
FBS	0.484	17.180	101	0.698	0.842	53.90
BP-l_1-magic	0.966	33.800	8	0.010	0.864	808.6
LASSO-l_1-magic	0.970	34.223	15	0.010	0.864	65.53
RLS	0.424	13.529	NA	0.995	0.731	49.13
LS	-1.3×10^{-13}	-143.14	NA	1.2×10^{15}	0.655	44.36

!

TABLE III: Limited view - semicircular sensor array test results

Algorithm used	SSIM	*PSNR (dB)*	iterations	*NMSE*	*GI*	*time (sec)*
SBTV	0.938	33.659	10	0.003	0.763	3.95
c-SBTV	0.938	33.659	10	0.003	0.763	0.47
SBTV-l_1	0.952	33.196	6	0.005	0.767	4.06
c-SBTV-l_1	0.952	33.196	6	0.005	0.767	0.27
c-rBP	0.932	31.773	32	0.012	0.776	5.33
DCA	0.144	13.518	200	0.540	0.769	17.67
FBS	0.617	18.001	101	0.580	0.810	54.82
BP-l_1-magic	1	90.583	8	1.8×10^{-8}	0.846	1488
LASSO-l_1-magic	1	81.85	28	2.5×10^{-7}	0.864	94.27
RLS	0.424	13.531	NA	0.995	0.636	48.94
LS	2.2×10^{-14}	-165.374	NA	3.0×10^{17}	0.740	34.43

stopping criteria. It is used to evaluate the performance of various algorithms because it is the crucial metric for determining an algorithm's effectiveness and efficiency.

The processor used in the simulation is Intel®Core™ i7-8565U 1.8GHz with MATLAB®.

The purpose of this study is to determine the optimal sensor arrangement and compressive sensing algorithm for reconstructing images from limited-view sensor arrangements. Cholesky factorization (c) is used to implement the inverse in the iterations which can decrease the computation time. The study evaluated the performance of four different limited view sensor arrangements (linear, convex, concave, and semicircular), as shown in Figure 2, and several compressive sensing algorithms (including SBTV, SBTV-l_1, Cholesky factorized-rBP (c-rBP), $l_1 - l_2$ using DCA and FBS, BP, LASSO, regularized LS (RLS), and LS) to compare their effectiveness.

For the line sensor array, the maximum SSIM obtained is 0.938 when c-SBTV-l_1 algorithm is used. Additionally, it is demonstrated to have a maximum PSNR of 30.983 dB and a reconstruction time of under a second, which falls within the acceptable limit. The c-BP also has an SSIM of 0.905, but when comparing the computation time, SBTV-l_1 algorithm is preferable.

For the concave sensor arrangement, the SSIM for the SBTV implementation is better and the highest SSIM and PSNR of 0.962 and 33.586 dB is obtained when SBTV-l_1 is used. It is observed that the c-SBTV-l_1 has the least computation time, which makes it the more suitable algorithm for reconstruction. For the semicircular sensor arrangement, the performance of both c-rBP and c-SBTV-l_1 is comparable with both having an SSIM of 0.952 and 0.932 respectively. But the reconstruction time of c-SBTV-l_1 is the lesser and thus more preferable. When comparing the SSIM of all sensor arrangements the convex sensor array exhibits the best SSIM value of 0.97 for SBTV-

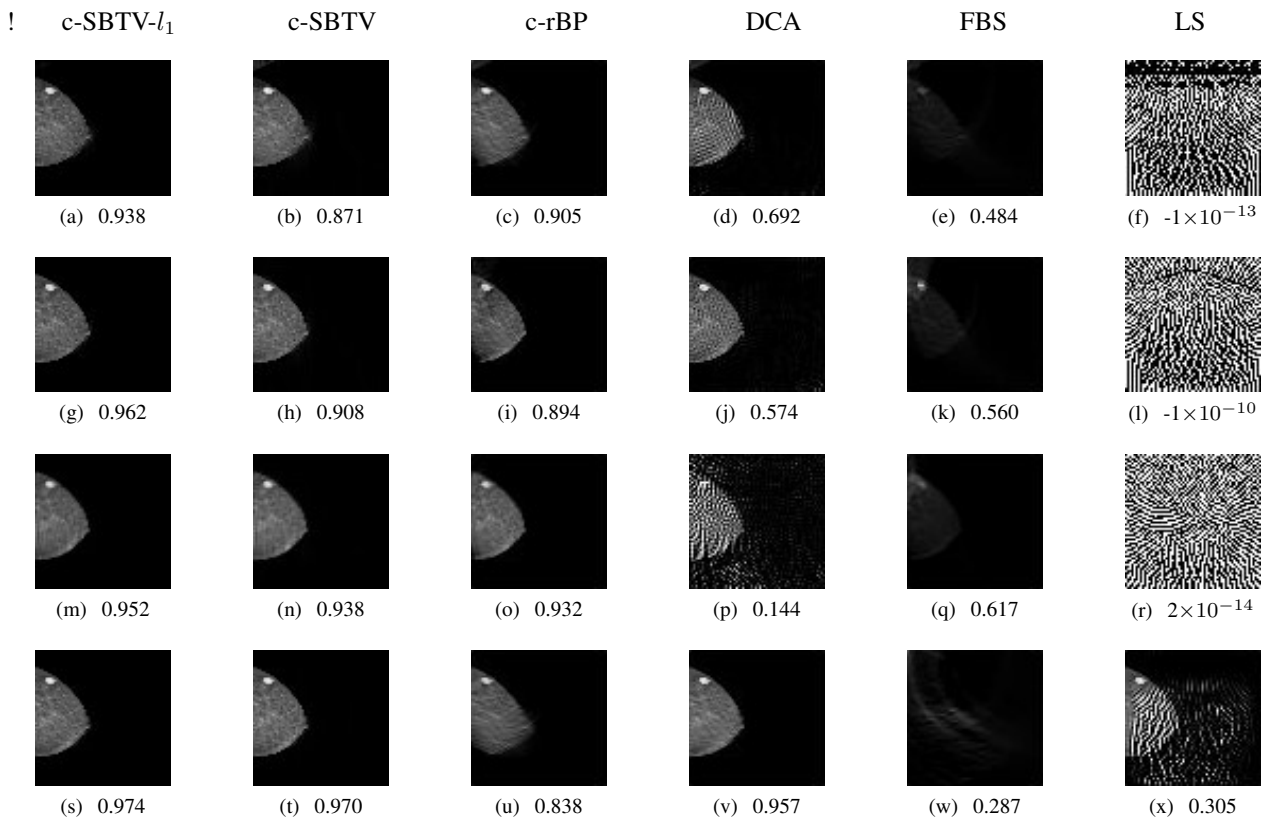


Fig. 2: Reconstructed images for different sensor arrangements: Line sensor array (a)-(f); Concave sensor array (g)-(l); Semicircle sensor array (m)-(r); Convex sensor array (s)-(x) with the corresponding SSIM's.

TABLE IV: Limited view - convex sensor array test results

Algorithm used	**SSIM**	***PSNR (dB)***	**iterations**	***NMSE***	***GI***	***time (sec)***
SBTV	0.971	35.851	15	0.003	0.770	5.35
c-SBTV	0.970	36.150	15	0.003	0.770	0.86
SBTV-l_1	0.975	36.347	10	0.021	0.770	5.37
c-SBTV-l_1	0.974	36.304	9	0.005	0.767	0.76
c-rBP	0.838	25.017	54	.065	0.756	8.68
DCA	0.957	33.299	200	0.483	0.768	17.33
FBS	0.287	16.610	101	0.743	0.707	53.96
BP-l_1-magic	0.986	38.715	8	0.002	0.845	1011.33
LASSO-l_1-magic	0.990	39.921	15	0.001	0.865	67.53
RLS	0.424	13.528	NA	0.995	0.764	48.79
LS	0.305	10.937	NA	0.503	0.756	35.57

l_1 with a reconstruction time of 0.76 seconds. Also, it can be observed that the implementation of Cholesky factorization has made the SBTV-l_1 algorithm 7 times faster for line and convex arrays and 9 and 15 times faster for concave and semicircular arrays respectively.

It is to be noted that even though the BP and LASSO implementation using l_1-magic gives good SSIM and PSNR values, the reconstruction time is very high and unacceptable for a fast image reconstruction. The non-iterative algorithms, LS and RLS have the worst results as it is highly dependent on the data availability, and it consumes more time for computation.

The results showed that convex sensor arrangement provided the best outcomes for all the CS algorithms, with the c-SBTV-l_1 algorithm being the most efficient. The SSIM and PSNR values were highest for the convex sensor arrangement in comparison to the other arrangements. The results of the different sensor arrangements for different CS algorithms are presented in Tables II - IV, and the corresponding reconstructed images are shown in Fig. 2.

V. CONCLUSION

This work aimed to identify the most efficient sensor arrangement and compressive sensing algorithm for reconstructing PAT images from limited-view sensor data. Four different sensor arrangements and several compressive sensing algorithms were evaluated. According to the findings, the SBTV-l_1 algorithm was successful in reconstructing the PAT image

for all sensor arrangements, demonstrating its superiority over other CS algorithms. Moreover, the study revealed that the use of a convex sensor array resulted in higher image quality compared to other sensor configurations. Additionally, implementing Cholesky factorization in the inverse computation stage of the SBTV-l_1 algorithm accelerated the reconstruction process by 15 times compared to the implementation without Cholesky. The results presented in the study can provide useful insights into selecting the optimal sensor arrangement and algorithm for PAT image reconstruction from limited-view sensor data in various applications.

REFERENCES

[1] H. Kasban, M. El-Bendary, and D. Salama, "A comparative study of medical imaging techniques," *International Journal of Information Science and Intelligent System*, vol. 4, no. 2, pp. 37–58, 2015.

[2] X. Ou, X. Chen, X. Xu, L. Xie, X. Chen, Z. Hong, H. Bai, X. Liu, Q. Chen, L. Li *et al.*, "Recent development in x-ray imaging technology: Future and challenges," *Research*, vol. 2021, 2021.

[3] J. Shah, S. Park, S. R. Aglyamov, T. Larson, L. Ma, K. V. Sokolov, K. P. Johnston, T. E. Milner, and S. Y. Emelianov, "Photoacoustic imaging and temperature measurement for photothermal cancer therapy," *Journal of biomedical optics*, vol. 13, no. 3, p. 034024, 2008.

[4] E. Deeba and A. Gunawardena, *Interactive linear algebra with Maple V*. Springer Science & Business Media, 1998.

[5] N. Vaswani and W. Lu, "Modified-cs: Modifying compressive sensing for problems with partially known support," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4595–4607, 2010.

[6] P. Kuchment and L. Kunyansky, "Mathematics of thermoacoustic tomography," *European Journal of Applied Mathematics*, vol. 19, no. 2, pp. 191–224, 2008.

[7] L. V. Wang and X. Yang, "Boundary conditions in photoacoustic tomography and image reconstruction," *Journal of biomedical optics*, vol. 12, no. 1, p. 014027, 2007.

[8] D. F. Resende, M. Khosravy, H. L. Monteiro, N. Gupta, N. Patel, and C. A. Duque, "Neural signal compressive sensing," *Compressive Sensing in Healthcare*, pp. 201–221, 2020.

[9] S. Chen and D. Donoho, "Basis pursuit," in *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, vol. 1, 1994, pp. 41–44 vol.1.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[12] E. Candes and J. Romberg, "l1-magic: Recovery of sparse signals via convex programming," *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf*, vol. 4, p. 14, 2005.

[13] M. John and I. Barhumi, "Fast and efficient pat image reconstruction algorithms: A comparative performance analysis," *Signal Processing*, vol. 201, p. 108691, 2022.

[14] M. J. John and I. Barhumi, "Total variation algorithms for pat image reconstruction," in *Proceedings of 2022 APSIPA Annual Summit and Conference*, 2022, pp. 1165–1169.

[15] I. Selesnick, "Least squares with examples in signal processing," *Connexions*, vol. 4, 2013.

[16] M. T. Heath, "Numerical methods for large sparse linear least squares problems," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 497–513, 1984.

[17] Y. Lou and M. Yan, "Fast l1–l2 minimization via a proximal operator," *Journal of Scientific Computing*, vol. 74, no. 2, pp. 767–785, 2018.

[18] Y. Lou, P. Yin, Q. He, and J. Xin, "Computing sparse representation in a highly coherent dictionary based on difference of l1 and l2," *Journal of Scientific Computing*, vol. 64, no. 1, pp. 178–196, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Towards AST-LLDs for the Analysis of Depression in Speech Signals

Sidharrth Nagappan, Chern Hong Lim and Anuja Thimali Dharmaratne
School of Information Technology, Monash University Malaysia
E-mail: {sidharrth.nagappan, lim.chernhong, anuja.dharmaratne}@monash.edu

***Abstract*—Recent advancements in deep learning allowed the deployment of large, multimodal models for depression analysis, but the increasing complexity of such models resulted in slow deployment times. This work proposes multi-stream audio-only models for depression analysis, that use transformer weights attended to by low-level descriptors (LLD) through an attention-weighted sum. It operates on the hypothesis that handcrafted feature sets will ameliorate extensive transformer pre-training. Extensive experimentation on the DAIC-WOZ test dataset shows that a combination of an audio spectrogram transformer (AST) and a Mel-frequency cepstral coefficient (MFCC) based convolutional neural network (AST-MFCC) produces the highest accuracy in our suite of models, but reports marginally lower macro F1 scores than both a naive AST and pure LLD-based models, suggesting that the injection of extra feature streams adds a sensibility element to models and limits false positives. However, the naive transformer-based and LLD-based models are surprisingly more effective at flagging depressed patients, although at the cost of an acceptable number of false positives. Our work suggests in totality that the addition of extra feature streams adds a distinct and controllable discriminating power to existing models and is able to assist lightweight models in low-data, audio-only settings.**

I. Introduction

Automatic Depression Detection (ADD) is a principal research area, stemming from the need for comprehensive screening processes in clinical psychotherapy. In a preliminary stage, current medical teams require patients to self-diagnose using the Patient Health Questionnaire (PHQ), where the patient answers a series of ordinal questions that quantify the severity of depression. In later stages, trained physicians interview patients and make an official diagnosis based on the replies and behavioural patterns of the patient. The preliminary assessment attempts to quantify a phenomenon that is subjective, while the physician interview is time-consuming and may not be required in most cases [1].

To address the above-mentioned problems, ADD systems were introduced [2]. However, the growing computational complexity of multimodal depression analysis engines makes us question whether the same task can be achieved through the analysis of only speech signals. This allows lightweight model deployments, but the real question is whether the performance trade-off as compared to a full-blown multimodal video-audio-text-based engine is significant.

The state-of-the-art in audio-based analysis have either been (i) pure deep learning models, (ii) low-level feature extraction with convolutional neural networks or (iii) context-aware question-answer attention networks [1]. Each method outperformed its predecessor, but to the best of our knowledge, little research has studied the individual contributions of each component. In this work, we test different combinations of state-of-the-art audio models and low-level feature sets, to deliver a robust model that leverages the best of each feature stream and operates in production environments with limited data.

The principal contributions of our work are: (i) testing the audio spectrogram transformer for deep depression analysis in a pioneering multi-stream approach, (ii) proposing inspired attention-based multi-stream fusion for LLD-assisted audio analysis and (iii) multi-stage cooling-based training processes for multimodal transformers. To ensure that our results are reproducible, we make our code publicly available [1].

II. Related Work

In clinical environments, a variety of depression scales are used, including physician-operated instruments such as the Diagnostic and Statistical Manual of Mental Disorders (DMS), Hamilton Depresion Rating Scale (HDRS), Montgomery-Åsberg Depression Rating Scale (MADRS) and self-diagnosing scales such as the Patient Health Questionnaire (PHQ-8) and Beck's Depression Inventory [3]. The principal difference between physician-operated surveys and self-diagnosing surveys is the human element that is able to make nuanced decisions about behavioural and psychopathological tendencies observed during conversations with the patients [4]. During conversations, simple voice modulations, pauses, vocabulary choice and other fine-grained observations play a role in analysing the severity of depression. To automate screening technology, the majority of production systems have focused on distributing self-diagnosing questionnaires. However, numerically quantifying highly subjective illnesses is widely disputed as an accurate assessment [5], and the only other alternative (live clinical interviews) is blockaded by a limited supply of medical professionals.

Due to rapid developments in the macroscopic fields of speech transcription and speech emotion recognition (SER) [6] and a proven relationship between prosodic and formant voice features and the severity of depression, recent research has applied machine learning to automate the screening process [7].

[1] https://github.com/sidharrth2002/speech-depression

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Initial work fundamentally relied on low-level descriptors (LLDs) such as spectral, cepstral, formant, and energy coefficients as the only representatives of an audio signal. Such features would then be treated as input features to the standard linear discriminant machine learning classifiers, such as Gaussian Mixture Models, Hidden Markov Models, k Nearest Neighbour (kNN) and Support Vector Machines [8]. The advent of deep learning mimicked the same structure, instead using multilayer perceptrons to build decision structures [9]. Later work used the same LLDs but in combination with convolutional neural networks, that would effectively slide across feature tensors and build an understanding of the relationships between sub-groups of audio features.

A more evolved understanding of audio signals was inspired by mel-spectrograms, allowing one to convert audio signals to visual representations. As analysing the 2-dimensional spectrograms became a computer vision task, a multitude of deep image models were applied. These included Long Short Term Memory (LSTM) models, Convolutional Neural Networks (CNNs), LSTM-CNNs, question-answer attention networks and LLD-based graph CNNs [10]. However, the increased dependence on deep learning comes with a large data barrier. For example, the Distress Analysis Interview Corpus/Wizard-of-Oz (DAIC-WOZ), the most widely used depression corpus, only contains 189 patient interviews.

More recent research understood the contributing effects of both LLD-inspired learning and deep audio modelling, eventually proposing inspired methods to fuse their contributions. Newer state-of-the-art transformer-based audio models such as the audio spectrogram transformer (AST) have been released, allowing systems to comprehend global context with minimal memory overhead, even when audio streams are lengthy [11]. The AST achieves state-of-the-art performance on AudioSet, ESC-50 and Speech Commands V2. The AST's ability to process lengthy streams and understand context seems particularly useful in depression analysis, as interviews are often extensive and a range of voice modulations are observed; a medical professional takes a collective view of the entire interview before reaching a diagnosis, as should a model. In an increasingly multimodal era of deep learning, our research attempts to strike a balance by proposing lightweight, audio-based methods to understand the extent to which audio signals alone are effective at depression detection.

III. Methods

Single-stream methods employ only one model type, while multi-stream networks fuse the neural outputs of different independent models through early or late fusion. In this work, we propose a two-stream approach, consisting of a Deep Audio Model (DAM) that directly manipulates the audio signal and a tensor of low-level descriptors (LLD) working in tandem. Both feature streams are combined through an attention-weighted sum, where outputs from the DAM and LLD streams are summed before the final tensor is queried by DAM outputs. Our initial hypothesis is that the low-level nature of LLDs will alleviate the black box of deep audio models.

A. Audio Spectrogram Transformer (AST)

An Audio Spectrogram Transformer (AST) is employed as the primary deep audio-based model [12]. It outperforms existing CNN-based architectures in a variety of generic audio classification tasks, using self-attention to learn the temporal and spectral dependencies of audio signals.

The audio spectrogram is first computed through strategies such as short-time fourier transforms (STFT) and embedded into 2-dimensional patches. Each embedding records positional, spatial and temporal relationships within a spectrogram. Subsequently, embeddings are passed into a transformer encoder, constituting self-attention modules and multi-headed attention blocks. After each head attends to different input patches, the final output is concatenated and passed through fully connected layers and a classification module. The architecture is illustrated in Figure 1.

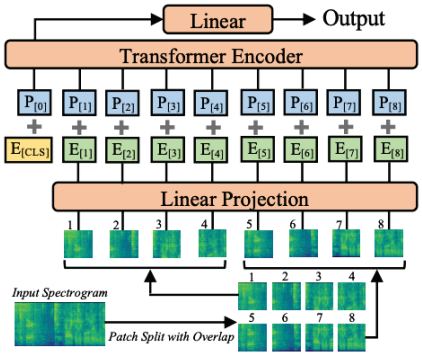


Fig. 1. AST architecture [12]

This research is the first to use the AST in the context of depression detection and also the first to use the AST in a multi-stream setting.

B. Low Level Descriptor (LLD) Convolutional Neural Network

Low-level descriptors are still stalwarts in audio processing, allowing the computationally efficient extraction of very low-level features. While they do not encompass an understanding that is as extensive as deep audio models, we hypothesise that their inclusion through attention querying will ameliorate the performance of its deep counterpart. The processing of LLDs is done through both 1-dimensional and 2-dimensional convolutional neural networks (CNNs). CNNs traverse through feature tensors, subsequently passing through pooling and batch normalization layers before returning a *batch size* $\times$ *512* tensor to be combined with other feature streams. The individual success of CNNs for LLD processing is apparent in [3]. CNNs were so predominant in the literature that several models went so far as to employ dual-stream CNNs [13], where one 2D stream would process a mel-spectogram while the other 1D stream would iterate through Mel-frequency cepstral coefficients (MFCCs).

We employ three commonly used feature sets as candidates for LLDS: IS09, eGemaps and MFCCs.

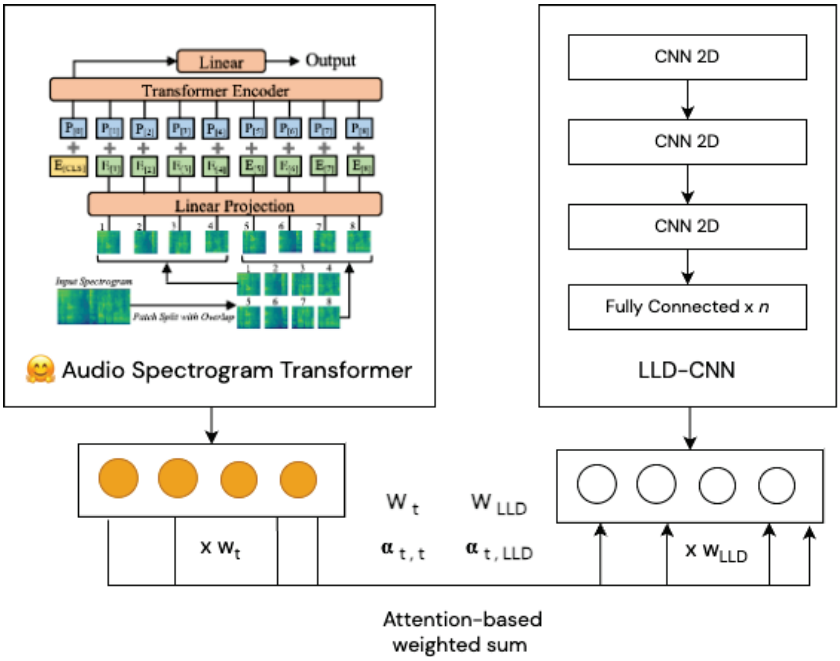


Fig. 2. Attentive AST-LLDs

1) MFCCs - 2-dimensional feature tensors that capture spectral characteristics of audio signals by converting a frequency-domain representation to a cepstral-domain representation through a combination of framing, windowing, fast fourier transforms, mel filterbanks, logarithms, discrete cosine transforms (DCT) and cepstral lifting.

2) IS09 - 348 functional units are extracted from each audio sample, encompassing fundamental frequency, frame energy, the zero-crossing rate and 12 selected MFCCs and their corresponding derivatives.

3) Extended Geneva Minimalistic Acoustic Parameter Set (eGeMaps) [14] - 88 acoustic features comprising F0-related, energy, spectral, harmonics-to-noise ratio, voice quality and prosodic features augmented with additional formant features.

C. Attention-based Tabular Combination

A distinct contribution of our research is examining unique combinations of multiple feature streams, going a step beyond standard concatenation. Early and late fusion functions ensure that stream features play a contributing role.

A weighted sum is first computed between the modalities, with the weight of each component becoming a trainable parameter. Subsequently, a $LeakyReLU$ based attention function uses transformer outputs as keys and the output of the other modalities as values. In these equations, α is the attentional matrix, F is the final feature vector, W is the weight matrix, x is the AST's audio features, while lld represents the LLD features.

$$F = \alpha_{x,x} W_x x + \alpha_{x,c} W_{lld} lld \tag{1}$$

$$\alpha_{i,j} = \frac{exp(LeakyReLU(a^T [W_i x_i || W_j x_j]))}{\sum_{k \in \{x,c,n,w\}} exp(LeakyReLU(a^T [W_i x_i || W_k x_k]))} \tag{2}$$

The dimensions of the fully connected layers in both the DAM and LLD blocks are determined by halving the number of neurons in each subsequent layer; the outputs of the final layer in each block are then passed to the fusion module. By understanding the relative weights given to each feature stream and complementing individual transformer features with their corresponding LLDs, we hypothesise that these key-value pairs will result in performance improvements [15]. The final architecture is illustrated in 2.

IV. Experiment

A. Preprocessing on the dataset

The DAIC-WOZ corpus is a set of compiled clinical interviews and ground truth depression diagnoses. It is scraped in it's entirety from the University of South Carolina's portal [16]. A total of 180 participants are available, with pre-defined train, test, and validation splits. Each participant is given a PHQ-8 score and a binary identifier of depression; often, both values are linked where a PHQ-8 score above 10 signifies depressive symptoms. Interviews are conducted by Elle, a computer agent, who asks succinct questions that are both pre-defined and conceived in real-time to logically sustain the conversation.

We use interviewer-interviewee timestamps to remove the interviewer's voice, and subsequently concatenate each participant's answers into a single audio file. Significant pauses

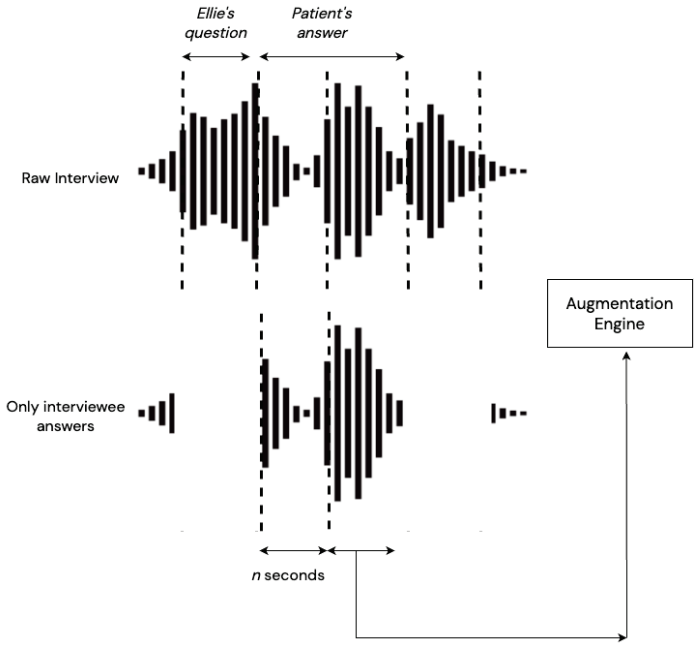


Fig. 3. Preprocessing and sampling interview data, consisting of removing the interviewer's voice, splitting the audio into n second samples and augmenting each sample through the proposed augmentation pipeline

of above 3 seconds are also removed to build more compact spectrograms and LLD feature sets, in line with the work of [1], [17]. Based on the pioneering sampling strategy of [17], we segment each participant's audio file into 5-second segments. This presents a balance between having sufficient data in an audio sample and controlling the sparsity of sample-based spectograms. The preprocessing and sampling strategy is illustrated in Figure 3.

Given the vast disparity between depressed and non-depressed samples and the possibility of overfitting, we use a random selection of augmentation strategies including Vocal Tract Length Perturbation (VTLP) [18], α-controlled white noise injection (noise factor of 0.05), semitone-based pitch adjustment (pitch factor of 1.5) and audio shifting (maximum shift of 0.2s) to augment non-depressed samples [19], culminating in a balanced training set. Manual intervention is done to ensure that the augmentation of the audio clip does not make the voice indiscernible.

Recent work that specifically studied the effects of audio augmentation strategies argued that the best performance is obtained through different combinations of augmentation methods [20], [21]. However, the literature is ambiguous about whether the augmentation strategies above are applied sequentially to signals, or chosen at random. We, therefore, test our baseline model with different combinations of augmentation strategies on the training set, while keeping our validation and test datasets untouched. All augmentation is done through a function that chooses a strategy through a temperature-based roulette wheel. We employ audio wrangling functions from *librosa* in tandem with *numpy*.

TABLE I
ABLATION STUDY BETWEEN AUGMENTATION STRATEGIES ON OUR BEST MODEL (AST + MFCC)

Strategy	Accuracy	F1 Score
No augmentation	0.7210	0.4304
VTLP [18]	0.7037	**0.4317**
Random augmentation	**0.7253**	0.4310

Our worst performance is obtained when all four augmentation strategies are applied sequentially, while our best is a random selection between noise injection, pitch adjustment and audio shifting (as shown in Table I).

B. Feature Extraction

A total of 3 different feature sets are extracted from each sample: (i) IS09, (ii) eGemaps and (iii) Mel-frequency cepstral coefficients (MFCCs). Librosa is used for (i) and (ii), while we employ the OpenSmile toolkit's audio processing binary for (iii). Hamming windows, Fast Fourier Transforms (FFT), Mel filter banks, logarithmic transformations and Discrete Cosine Transforms (DCT) are sequentially applied to each sample. All computation happens in a cached dataloader in a Linux-based high-performance cluster. Feature tensors are directly reshaped and sent to the GPU prior to modelling.

C. Cooled Training

Our architectures constitute blocks of different maturity levels; transformer weights are extensively pretrained and are capable of sensible computation out of the box, while randomly initialised convolutional neural networks and fully connected layers require significant training epochs to start producing discernible results. We, therefore, train all weights in the transformer block for 5 epochs, before freezing transformer weights, increasing the learning rate, and training the entire network for up to 30 epochs. We name this method "cooled training".

V. RESULTS AND DISCUSSIONS

All models are implemented using Pytorch, with base transformer architectures being imported from Huggingface.

The first set of experiments (as shown in Table II attempts to identify the best-performing LLD feature set for this task, where each feature set is passed in its entirety to a 1D convolutional model. The first model in this series is tuned to identify an optimal number of layers (4 CNN-BatchNorm blocks) and this configuration is carried forward. We can observe that both eGeMaps and MFCC features deliver similar performance. The efficiency of MFCC features falls in line with the findings of [3], [21], [22], although the minor performance boost obtained by eGeMaps is surprising. When testing MFCC features in tandem with a 2D CNN, we find that performance worsens further, even after hyperparameter tuning.

To understand if multi-stream additions are in-fact superior to naive transformer models, we experiment with different combinations of the AST with the top two LLDs (MFCC and

TABLE II
MODEL STREAM RESULTS WITH ACCURACY, PRECISION, RECALL, MACRO F1 SCORE AND F1 SCORES FOR POSITIVE AND NEGATIVE CLASSES

Model	Accuracy	Precision	Recall	F1	$F1_P$	$F1_N$
eGemaps-Conv1D	**0.6566**	**0.7394**	0.8115	**0.5265**	**0.2793**	**0.7737**
is09-Conv1D	0.6224	0.7136	0.8011	0.4665	0.1783	0.7548
MFCC-Conv1D	0.6417	0.7160	**0.8389**	0.4638	0.155	0.7727
MFCC-Conv2D	0.6161	0.7257	0.7573	0.4993	0.257	0.741

TABLE III
SUITE OF MULTI-STREAM MODELS AND COMPARISON WITH AN LSTM BENCHMARK

Model	Accuracy	F1
MFCC-LSTM [20]	0.5862	0.3263
AST (Baseline)	0.5945	**0.4735**
AST + eGemaps	0.6244	0.4305
AST + MFCC	**0.7037**	0.4325

eGemaps) in Table III, combined using an attention-weighted sum.

Here, we find that a naive AST model performs worse than *eGemaps-Conv1D*, which could be attributed to the transformer model's inability to completely generalise to the new problem domain of depression detection. Perhaps lower learning rates and more extensive fine-tuning can tackle this issue.

In terms of multi-stream models, we find that *AST + eGemaps* performs worse than *eGemaps* alone (by up to 4.91%). The *AST + MFCC* combination all its counterparts with a 16.83% accuracy improvement from the baseline AST model. However, it is important to note that the macro F1 score drops by 9%, indicating that additional keyed components increase the model's power to identify non-depressed samples, but hinder its ability to flag depressed samples.

We compare our multi-stream model to an LSTM-based reproduction of [20]. Their final model was an LSTM-AU model that considered MFCC features using an LSTM model in tandem with handcrafted action units from the video modality. Our reproduction only uses the LSTM component of their model, proving that the multi-stream model outperforms its LSTM-based counterpart in an audio-only context.

VI. CONCLUSION

In this work, we explored different LLD-based convolutional neural networks in tandem with state-of-the-art audio transformers, and analysed their performance limitations with regards to fully-fledged video+text+audio modality models. We find that the most complex model is not necessarily the best performant and basic LLD-based models perform unusually well in low-data, audio-only contexts. Our multi-stream model attains the highest accuracy in our suite of models, but has a diminished ability to identify depressed samples. Nevertheless, our results suggest that LLD-assisted transformers are a promising research direction since they produce comparable results in low-data settings and have discriminating powers that can be more closely controlled as compared to black-box deep learning models.

ACKNOWLEDGEMENTS

This work is supported by the Monash Malaysia NEED Grant Scheme under reference code IT/NEED/11-2020/001.

REFERENCES

[1] M. Niu, K. Chen, Q. Chen, and L. Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4235–4239. DOI: 10.1109/ICASSP39728.2021.9413486.

[2] M. Morales, S. Scherer, and R. Levitan, "A cross-modal review of indicators for depression detection systems," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Vancouver, BC: Association for Computational Linguistics, Aug. 2017, pp. 1–12. DOI: 10.18653/v1/W17-3101. [Online]. Available: https://aclanthology.org/W17-3101.

[3] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, "Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis," *Machine Learning with Applications*, vol. 2, p. 100 005, 2020, ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2020.100005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827020300050.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021. DOI: 10.1109/ACCESS.2021.3068045.

[5] S. Eack, C. Greeno, and B.-J. Lee, "Limitations of the patient health questionnaire in identifying anxiety and depression: Many cases are undetected," *Research on social work practice*, vol. 16, pp. 625–631, Nov. 2006. DOI: 10.1177/1049731506291582.

[6] M. Khan and S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *International Journal of Intelligent Systems*, vol. 36, p. 20, May 2021. DOI: 10.1002/int.22505.

[7] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bioacoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, p. 105 020, 2023, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.105020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423004536.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019. DOI: 10.1109/ACCESS.2019.2936124.

[9] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041249. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1249.

[10] M. Ishimaru, Y. Okada, R. Uchiyama, R. Horiguchi, and I. Toyoshima, "A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network," *Diagnostics*, vol. 13, no. 4, 2023, ISSN: 2075-4418. [Online]. Available: https://www.mdpi.com/2075-4418/13/4/727.

[11] Y. Zhang, B. Li, H. Fang, and Q. Meng, "Spectrogram transformers for audio classification," in *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2022, pp. 1–6. DOI: 10.1109/IST55454.2022.9827729.

[12] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," *CoRR*, vol. abs/2104.01778, 2021. arXiv: 2104.01778. [Online]. Available: https://arxiv.org/abs/2104.01778.

[13] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.05.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S153204641830090X.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016. DOI: 10.1109/TAFFC.2015.2457417.

[15] K. Gu and A. Budhkar, "A package for learning on tabular and text data with transformers," in *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, Mexico City, Mexico: Association for Computational Linguistics, Jun. 2021, pp. 69–73. DOI: 10.18653/v1/2021.maiworkshop-1.10. [Online]. Available: https://www.aclweb.org/anthology/2021.maiworkshop-1.10.

[16] J. Gratch, R. Artstein, G. Lucas, *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf.

[17] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16, Amsterdam, The Netherlands: Association for Computing Machinery, 2016, pp. 35–42, ISBN: 9781450345163. DOI: 10.1145/2988257.2988267. [Online]. Available: https://doi.org/10.1145/2988257.2988267.

[18] N. Jaitly and E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," 2013.

[19] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103 107, 2022, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103107. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421007047.

[20] M. Muzammel, H. Salam, and A. Othmani - Hiring Postdocs, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106 433, Sep. 2021. DOI: 10.1016/j.cmpb.2021.106433.

[21] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103 107, 2022, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103107. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421007047.

[22] H. Wang, Y. Liu, X. Zhen, and X. Tu, "Depression speech recognition with a three-dimensional convolutional network," *Frontiers in Human Neuroscience*, vol. 15, 2021, ISSN: 1662-5161. DOI: 10.3389/fnhum.2021.713823. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2021.713823.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

ecVoice: Audio Text Extraction Optimization of Video Based on Idioms Similarity Replacement

Jinwei Lin*
* Monash University, Malaysia
E-mail: Jinwei.Lin@monash.edu ORCID:0000-0003-0558-6699

***Abstract*—The Text Extraction of the Audio from the Video plays an important role in multimedia editing and processing. As a popular open source toolkit, Whisper performs fast in human voice recognition. However, the recognition performance is dependent on the computing resource, which makes the low computing memory running Whisper become difficult. Our paper presents an available solution to extract the human voice from the video and gain the high quality text generation from the voice. The generated voice can be used in video language translation and translated voice simulation. To improve the extraction and transform quality of human voice, we present ecVoice, a method using the idioms similarity computation and analysis to improve the quality of audio text extraction. Relative experiments are held to verify that the ecVoice can improve the idiom grammar correction rate to 90% on average. The method is simple but fast which means this method will cause less bad influence of consuming computing resources when improving the voice recognition rate. Our method and solution can significantly enhance the Whisper recognition with low computing memory.**

I. Introduction

Extracting and gaining a high quality audio text from the video is a significant application and implementation in multimedia. The text extracted from the video can be used in multiple fields of multimedia applications or research implementations, such as translated language of video playing, audio editing and internationalisation of game video design. With the technologies of super Artificial Intelligence (AI), AI assistants have more and more important roles in multimedia and computer games design. One important process or research problem of these applications is transforming the voice to text, which is also one of the most important points of our research.

However, there are still some difficulties or challenges in transforming the voice to text, for example, improving the recognition accuracy and processing speed.Usually, methods like Deep Neural Network (DNN) are used to address this issue. But the recognition accuracy may of course have a large fluctuation influence which is related to the parameter size of the neural network model. For some machines or hardware with limited gpu performance or memory, the network parameter size of the model is one of the main influence factors. For the limitation of computing hardwares, one feasible method to improve the recognition performance is using the software solutions, like improving the algorithms design and architecture. In this paper, these software methods can be summarised into three categories, the first is pre-processing methods, the second is the during-processing, the third is post-processing methods. We will focus more on the post-processing method in research of the paper.

In the paper, we will present an available solution to split the audio from the video and extract the accurate text from the audio with an efficient and prolongable algorithm to improve the performance and accuracy of audio text extraction based on idioms similarity replacement. The relative experiment code is open source in GitHub:https://github.com/JYLinOK/ecVoice.

II. Literature Review

A. Extract Text from Voice

Extracting the text from voice files is related to the research field of voice recognition. The technologies of voice recognition have been used in various fields in AI, such as using speech recognition to recognize the emotion of the speaker [1], the end-to-end automatic speech recognition [2], as well as using different methods to improve the speech recognition accuracy in normal speech recognition filed [3]. The efficiency of speech recognition is closely related to the volume of the given information from the voice and how the recognition performs a high quality recognition in limited conditions [4]. The accuracy of speech recognition can be improved by improving the process algorithms [5], or using special processes on voice data [6]. Some researchers try to improve the recognition accuracy by focusing on the text processing after the post-processing of speech recognition [7].

Our research is focused on extracting the text from the voice in the wild that includes human speech, and improving the speech text recognition by post-process of the recognized text.

B. Voice in Multimedia Editing

In this research, we focus on improving the efficiency of the voice processing in multimedia editing. Voice processing plays an important role as visual information processing in modern multimedia editing [8]. Voice processing is also an important part in video editing, and plays an important role in video text translation [9] and game design [10]. The audio from the video and then translate the speech voice to text for further multimedia processing [11], which is considered as a significant application in voice recognition. And the research of the paper is focused on how to improve the accuracy of speech recognition from the audio that was extracted from the video, by presenting an algorithm on the post-processing of voice recognition. This research work will be applied and accessed

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

in our further research on automatic intelligent multimedia editing and NeRF AI game engine design in the future.

C. Text Grammar Correction

After translating the voice to text from the audio, two factors will influence the currency of the speech recognition. One is the recognition success rate of the speech recognition model, which can only be improved by updating and optimising the design or training strategy of the recognition model [12]. The other is the pre-processing [13] or post-processing and optimization of the translated text from the recognition model. For the post-processing of the translated text, the text syntax correction process in the field of natural language [14] is to directly improve the textual expression quality of the translated text to improve the accuracy of the voice recognition.

One of the most important keys of our method to improve the quality of the audio text extracted from the video is to improve the efficiency of text grammar correction by correcting the wrong expression of the idioms with idioms similarity replacement. Our method is simple but fast, which makes it can be used in the optimising and correcting translated text extracted from the speech audio with a high speed requirement.

III. Methodology Analysis

A. Running Architecture

Optimising textual expression of the extracted text that is translated from the audio of the video is the main process and algorithm analysis of this research. Beside this, splitting the audio data from a specific video, and making the pre-processing of the voice data, are also an important part of the running architecture of the research. As shown in Figure 1, the running architecture of this research includes 7 key components: splitting audio from video, splitting speech from audio, split the main speech, speech text recognition, creating the pinyin analysis function, generating the pinyin dictionary of idioms data-set, finding the closest idioms algorithm. Following are the detailed analysis of the components mentioned above.

As shown in Figure 1, first, process splitting audio from video, is related to the item $videoAudioSplit$. The target of the process is extracting the audio files from the specific video file. Second, process splitting speech from audio, is related to the item $UltimateVocalRemover$. Because the toolkit Ultimate Vocal Remover will be used to gain the pure human speech data with a high quality from the audio files. Third, the process of splitting the main speech is related to the item $segmentation$, which means making the segmentation for the speech file to get the main useful segment data. Fourthly, the processing of speech text recognition is related to the item $useWhisper$, which means using the robust toolkit whisper to generate the accurate translated text of the speech file. Fifthly, the process of analysing pinyin and generating the pinyin dictionary is related to the item $getPinYin + genIdomPinyin$, which means creating function to get the basic information of the segments of words, and generating a pinyin dictionary with a special data format. Last, the process of finding the closest idioms algorithm is related to the item $findClosest$, which means analysing the expression of a segment of words to check if it is a normal idiom with the correct writing format or not, and get the most closest right idiom expression if the words segment is wrong.

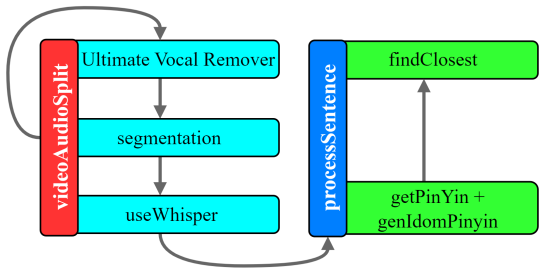


Fig. 1. Running and design architecture of this whole research.

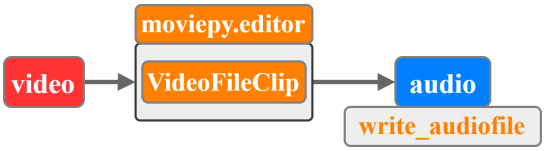


Fig. 2. Using Moviepy to splitting the audio data from video.

B. Splitting Audio From Video

If we want to implement the optimization for the text expression that is extracted from the audio file, the primary process is getting the audio files. Due to focusing on the target of multimedia processing, the audio files used in this research are all split from the video file, which is corresponding to the item $videoAudioSplit$ in Running Architecture. As shown in Figure 2, we utilise the Python library $moviepy$ to design a processing algorithm to get the split audio file from one specific video file. Using the function $mp.VideoFileClip$ to generate a video clip of an input video with the mp4 format, followed by using the function $write_audiofile$ to resave the audio data as a wav format file. Thus getting the audio file of one specific video file. The format of the extracted audio file can be custom defined in the processing code.

C. Audio Segmentation

After gaining the extracted audio file from the video file, further processing related to the voice optimization can be implemented. In our research, to ensure the high quality of the audio data before it is processed in future procedures, like splitting the human speech voice and the background audio made by other objects but not human voice, we make the process of cutting off the two ends audio segments of the previous audio that is extracted from the video. This process and design is corresponding to the item $segmentation$ in Running Architecture. As shown in Figure 3, we use the $pydub$ library to make the segmentation of the audio. Detailedly, Using the function $AudioSegment.from_wav$ to transform the audio file, for example in wav format, subsequently, using the index of list to make the segmentation of the audio file.

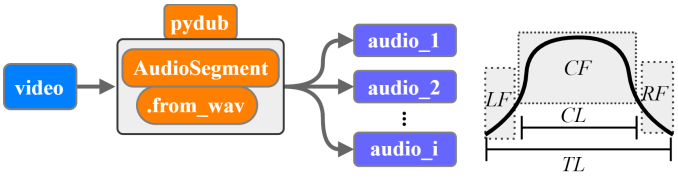


Fig. 3. Using pydub to make the segmentation of the extracted audio.

During this processing, the audio file is presented as one kind of editable data with the list format. After the segmentation edit of the audio, calling the function $export$ of the split audio. Eventually, the previous audio will be split into a specific number of segments.

As shown in Figure 3, assume the total length of audio is represented as STLS, when the audio is generated or recorded, usually the two ends of the audio are sick in recording the information, which will cause some mistakes in information representation. Therefore in this research, processing algorithm will delete the two ends of the audio to reduce the related interference, and use the middle segment of the audio as the processed audio. As shown in Figure 3, the LF represents the left end of audio length, the RL represents the right end of audio length, the CL represents the central of audio length.

D. Audio to Text

After the segmentation of audio, the next process is translating the speech audio to text, which means the process needs the speech recognition to get the text of the human speech in the audio. To ensure the quality of this process of speech recognition, we use the toolkit $UltimateVocalRemover$ to split the human speech from the audio. Except for human speech audio, the background music or voice of the previous audio will be saved for future works, like video or audio speech language translation and multimedia application editing.

Subsequently, we used whisper to implement the speech recognition. We design a custom selection function with five model parameters: t, b, s, m, l to select the five types of model of whisper by using the $load_model$ function. Each model from whisper needs a different size of Visual RAM (VRAM) to run. As shown in Figure 4. After a series of experiments, we found that the minimum practical required VRAM is bigger than the requirement provided by whisper, so we provide the recommended size parameters for this, as shown in Figure 4. The output result of this process is the text recognized from the speech audio file, which will be used in the next procedure.

E. Redundancy Removal

After getting the text that is recognized and translated from the speech audio, the next procedure is optimising the translated text. In our research, we prepared several methods to make the optimization, of which the first one is, as shown in Figure 5, the redundancy removal of the translated text. The key point of this method is that, comparing the start and end parts of two closed sentences to find the redundant words in some combined context. Because, before the speech audio

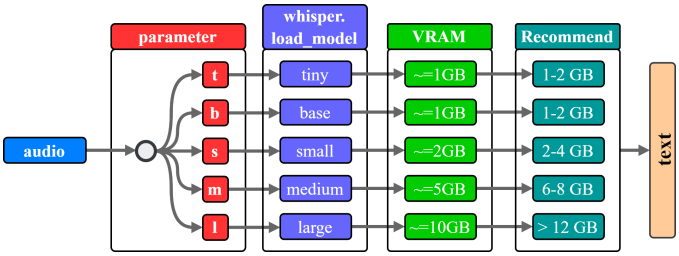


Fig. 4. Using the Whisper to get translate the speech of audio to text.

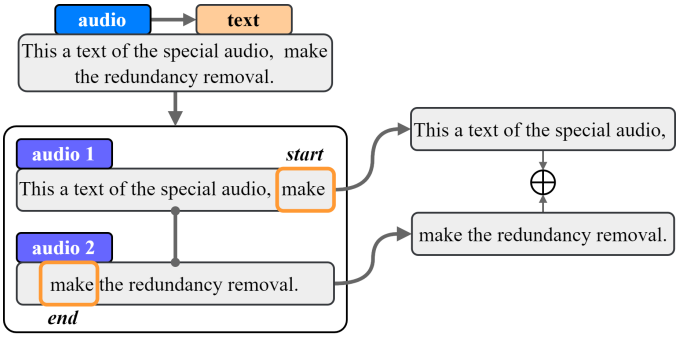


Fig. 5. Comparing two closed sentences to make redundancy removal.

is sent to make the speech recognition to get the translated text, the speech audio is gained by segment or splitting, which means some repetitive words may be represented in the start and end parts of two closed sentences. Therefore this process will be meaningful to get more accurate translated results in the further speech text processing.

As shown in Figure 5, for example, suppose there is an audio with a speech text "This is a text of the special audio, make the redundancy removal.", when this audio is recorded or split from the previous process, the audio may be split or recorded into two sentences which have the same repetitive words in the start and end parts of two closed sentences, like $audio1$ and $audio2$. Our method is to remain one side of repetitive words in one sentence, and remove the repetitive words in the other sentence. Subsequently, making the further text processing for them, before combining them to get the terminal output.

F. Data Format Conversion

After implementing the redundancy removal of the translated text of the speech audio, the next process for the audio text is building a dataset that follows the special data format with the existing idiom data files. As shown in Figure 6, the previous data format of the idioms json files is combined with:$derivation, example, explanation, pinyin, word$ and $abbreviation$. Obviously some of these items are non-useful for us to build the idioms for syntax correction. Therefore, we just selected the item$word$ as one component of our new designed idioms dataset. Beside this, the tone of the idiom word is the other required component, which is gained from the program coded by the library $pypinyin$ of Python.

As shown in Figure 6, we design a data format of key-value to store the information of the idioms and their corresponding

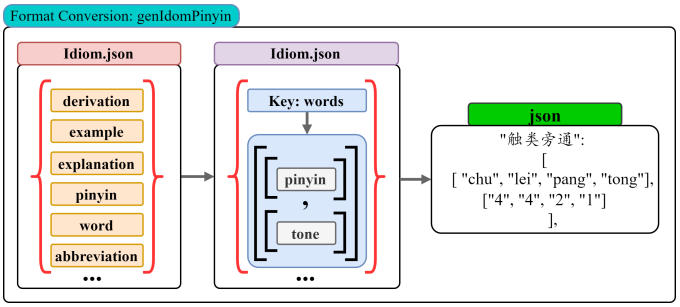


Fig. 6. Data format conversion of the idioms to make the dataset.

tones. One unit of this dataset is combined with a Chinese idiom as the kay and a list of tone information of this idiom as the value. The data format can be described as follows:

$$word: \left\{ \begin{matrix} [[pinyin_1, pinyin_2, ..., pinyin_n] \\ [tone_1, tone_2, ..., tone_n]]] \end{matrix} \right\}, \quad (1)$$

The Equation 1 can be used in the data format conversion of other languages that are similar with Chinese, and has phonetic symbols and corresponding tones. With this kind of data format, we can quickly and individually deal with the calculation and comparison between different idiom words, which will promote the further development and research.

G. Similarity Computation

After finishing the building of the dataset of idioms with the special defined format, the next process is analysing and estimating whether the idiom in the text has a correct expression or not. To address this problem, we design a comparison calculation algorithm to judge and analyse the expression of the idiom. If the expression of the idiom in the text is wrong, the algorithm will find the correct expression from the idiom dataset and replace it in the text with the right expression.

To approach this problem, there are four procedures to do. First, get the wrong idiom expression in the audio text. This process is difficult with uncertain interfering factors, for those kinds of languages like Chinese, the words are independent characters, which means it will be difficult to distinguish whether a word character is relative to the word before it or the word behind it. Therefore, a word character in a Chinese sentence, maybe can become an idiom when combining with the words before and after. For this issue, we design an algorithm to analyse the word in a Chinese sentence one by one. As shown in Figure 7, for general Chinese sentences, the character length of a classical Chinese idiom is 4, so with gradual bit positional processing, selecting the sentence segments $seg_1, seg_2, seg_3, \ldots, seg_i$ to make the positional processing $p_1, p_2, p_3, \ldots, p_i$ correspondingly. If the current process index of the sentence is i, then capture a segment of sentence with a length of 4, which will select the indexes:$i, i+1, i+2, i+3$ to combine a selected sequence $[c_i, c_{i+1}, c_{i+2}, c_{i+3}]$. The length of the idiom is set to be 4, and can also be other numbers like $n, n \in \mathbb{N}^+$.

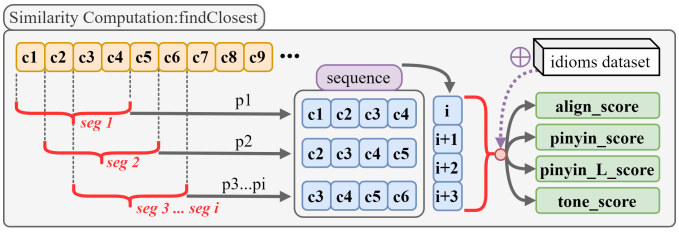


Fig. 7. Similarity computation and replacement processing for idioms.

As shown in the following code, the $word_1$ represents the sequence extracted from the sentence and $word_2$ represents the idiom selected from idioms dataset. max_L1 means the length of the sequence is bigger than the idiom. The judge criteria is relative to length of $word_2$ and parameter $align$.

```
if len(words1) >= len(words2):
    max_L1 = True
else:
    max_L1 = False

if max_L1:
    for word_i in range(len(words2)):
        if words1[word_i] == words2[word_i]:
            align_score += a

        if align_score < int(len(words2) / align):
            return_ok = False
else:
    for word_i in range(len(words1)):
        if words1[word_i] == words2[word_i]:
            align_score += a

    if align_score < int(len(words1) / align):
        return_ok = False
```

When an idiom expression with a length of $n (n \in \mathbb{N}^+)$ is gained, the next process is calculating the similarity score S_s of the idiom expression with the idiom dataset, to test whether the idiom expression is correct or wrong. We design four kinds of sub-similarity-scores to calculate the total similarity score S_a. As shown in Figure 7, the four scores are: $align_score, pinyin_socre, pinyin_L_score$ and the $tone_score$.

$$S_a = \sum_i^n \left(\begin{matrix} if\ c_{si} = c_{di} : a \\ if\ c_{si} \neq c_{di} : 0 \end{matrix} \right) \quad (2)$$

In Equation 2, there is the following relationship, the parameter $a = 1, (a \in \mathbb{N}^+)$ is the proportional control value, $c_{si} \in [c_{s1}, c_{s2}, ..., c_{sn}], c_{di} \in [c_{d1}, c_{d2}, ..., c_{dn}$, the parameter c_{si} represents the character on the index i in the $sequence$ selected from the processing sentence. The parameter c_{di} represents the character on the index i in the $idiom$ selected from the idiom dataset. This operation likes bit and operation, if parameter c_{si} equals parameter c_{di}, then the total $align_score$ which represented as S_s, will add $a = 1$, else will add 0, that means no more transformation.

Second, for each sequence selected from the sentence, we need to calculate the second evaluation score, that is $pinyin_score$, which will follow the operation idea as the $align_score$ and calculate the similarity of the pinyin character

by bit. The calculation operation is to compare and calculate the similarity between the pinyin expression of the sequence selected from the sentence and the idiom selected from the idioms dataset. Calculating the similarity between two item that all have the same expression format method in pinyin will be useful for the comparison and similarity calculation of the text correct expression, because the pinyin, as the main articulatory representation item of these languages like Chinese, is most relative expression about the pronunciation of human speech recognition. Therefore, the parameter $pinyin_score$ is useful for our similarity replacement algorithm.

$$S_p = \sum_i^n (\begin{Bmatrix} if\ cp_{si} = cp_{di} : b \\ if\ cp_{si} \neq cp_{di} : 0 \end{Bmatrix}) \tag{3}$$

In Equation 3, there is the following relationship, the parameter $b = 1, (b \in \mathbb{N}^+)$ is the proportional control value, $cp_{si} \in [cp_{s1}, cp_{s2}, ..., cp_{sn}], cp_{di} \in [cp_{d1}, cp_{d2}, ..., cp_{dn}$, the parameter cp_{si} represents the pinyin character on the index i in the pinyin expression of the $sequence$ selected from the processing sentence. The parameter cp_{di} represents the pinyin character on the index i in the pinyin expression of $idiom$ selected from the idiom dataset. This operation likes bit and operation, if parameter cp_{si} equals parameter cp_{di}, then the total $pinyin_score$ which represented as S_p, will add $b = 1$, else will add 0, that means no more transformation.

Third, besides the small-grained comparison and calculation mentioned above, we design another calculation algorithm for the large comparison and for the pinyin expression of each sequence and idiom. For each sequence selected from the sentence, we need to calculate the third evaluation score, that is $pinyin_L_score$, which will calculate the similarity of the pinyin character in the unit level of an individual word of the expression of sequence or idiom . The calculation operation is to compare and calculate the similarity between the word pinyin expression of the sequence selected from the sentence and the word pinyin expression of idiom selected from the idioms dataset. Calculating the similarity between two word items to test whether these two expressions are equal will be useful for the comparison and similarity calculation of the text correct expression, because the comparison for the word pinyin expressions between these two items will quickly get the similarity level in the big level of word. The L in $pinyin_L_score$ means the $largelevel$, this is the word level.

$$S_pl = \sum_i^n (\begin{Bmatrix} if\ cpl_{si} = cpl_{di} : c \\ if\ cpl_{si} \neq cpl_{di} : -c \end{Bmatrix}) \tag{4}$$

In Equation 4, there is the following relationship, the parameter $c = 2, (c \in \mathbb{N}^+)$ is the proportional control value, $cpl_{si} \in [cpl_{s1}, cpl_{s2}, ..., cpl_{sn}], cpl_{di} \in [cpl_{d1}, cpl_{d2}, ..., cpl_{dn}$, the parameter clp_{si} represents, in word level, the pinyin unit expression on the index i in the expression of the $sequence$ selected from the processing sentence. The parameter cpl_{di} represents, in word level, the pinyin unit expression on the index i in the pinyin expression of $idiom$ selected from the idiom dataset. If parameter cpl_{si} equals parameter cpl_{di}, then the total $pinyin_L_score$ which represented as S_pl, will add $c = 2$, else will add $-c = -2$, that means reverse transformation.

Last, after the calculation of the part of pinyin, for the other important part of the tone of idiom, we also design an algorithm to calculate the similarity between tone expressions of each sequence and idiom. For each sequence selected from the sentence, we need to calculate the fourth evaluation score, that is $tone_score$, which will calculate the similarity of the pinyin tone level number of the expressions of sequence or idiom. The calculation operation is to compare and calculate the similarity distance between the word pinyin tone level number of the sequence selected from the sentence and the pinyin tone level number of idiom selected from the idioms dataset. Calculating the similarity distance between two word items to test whether these two pinyin tone level number of items are equal or their distance will be useful for the comparison and similarity calculation of the correct pronunciation of the idiom item. A correct pronunciation is an important issue to ensure the right expression of the idioms.

$$S_t = p \cdot \frac{\sum_i^n (int(tone_{si}) - int(tone_{di}))^2)}{n} \tag{5}$$

In Equation 5, there is the following relationship, the parameter $p = 1, (p \in \mathbb{N}^+)$ is the proportional control value, $c_{si} \in [c_{s1}, c_{s2}, ..., c_{sn}], c_{di} \in [c_{d1}, c_{d2}, ..., cl_{dn}$, the parameter c_{si} represents, the pinyin tone level number of the word in the index$i$ of expression of the $sequence$ extracted from the processing sentence. The parameter c_{di} represents, in word level, the pinyin tone level number of the word in the index$i$ of expression of $idiom$ selected from the idiom dataset. We use the calculation method of square deviation to calculate the offset distance of similarity of pinyin tone of the sequence and idiom. The parameter p is used to fine-tune the effort of the calculator of the similarity distance of pinyin tone.

During the design and implementation process of similarity calculation, double for loop is usually used to make the double ergodic calculation. To ensure the correct calculation, the iteration number should be the smaller level compared with the second for loop. So $min(iteration\ number)$ should be dimension number of first for loop, the other $bigger(iteration\ number)$ should be dimension number of second for loop.

After all calculations of these four scores, the final evaluation similarity score S can be gained as follow as shown in Equation 6:

$$\begin{aligned} S = pinyin_score \times pp - tone_score \times pt \\ + pinyin_L_score \times ppl \end{aligned} \tag{6}$$

Using the S parameter of Equation 6 to evaluate the similarity degree between the $sequence$ extracted from the processing sentence and the $idiom$ selected the idioms data set. Each total calculator score S of each similarity calculation all will be saved in a list. Finally, make the operation $max(scores_list)$ to get the maximum of the total score. Subsequently, make a

judgement of the parameter S to decide whether to replace the sequence with the right idiom expression. If the total similarity score is too small, then remain the previous expression of the extracted sequence, otherwise, replace the sequence with the right expression of the idiom selected from idioms data set.

IV. Experiment and Evaluation

To evaluate the performance of our process algorithm, we design the following evaluation method to test the performance of $ecVoice$ from different dimensions.

A. Idiom Correction by Word Expression

First we make the detection experiment for the single idiom correction by using almost some pinyin tone and pronunciation but different expressions in word writing. As shown in Table I, the parameter $word$ represents the writing of the sequence extracted from the sentence. Due to the sequence word maybe having different expressions in typing or characters writing, we use the pinyin characters and the _ symbol to connect different pinyin of different words. The parameter $n, n \in \mathbb{N}^+$. represents $group_n$, which means the number of the items of a test group of sequence words. All of the items in the group of sequence words will be tested with our algorithm and count the number of correct test samples, of which number is recorded with the parameter $c, c \in \mathbb{N}^+$. The parameter $time, time \in \mathbb{N}$ represents the time in seconds of the whole implementation of the group of sequence words. The parameter $s1, s1 \in \mathbb{N}$ represents the similarity score between the correct sequence and the correct idiom, which means the score $s1$ should be the largest one among the similarity scores of the sequence group. Because the first item in the sequence group has the same typing or writing expression of the idiom from the idioms dataset. The parameters $s2, s3 (s2, s3 \in \mathbb{N})$ represent the similarity scores between two incorrect sequences and the correct idiom from the idioms dataset. Therefore parameters $s2$ and $s3$ should be never larger than $s1$. The parameter $sa, sa \in \mathbb{N}$ represents the average of the parameters $s2$ and $s3$, which represents the average of the normal similarity score of this sequence test group.

TABLE I
WORD ERROR INDEX = 1.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.207	19.0	19.0	14.0	16.5
chu_mu_cheng_song	3	3	1.18	22.0	16.0	22.0	19.0
fang_ge_zong_jiu	3	3	1.156	21.0	21.0	18.75	19.9
jing_bing_qiang_jiang	3	3	1.203	26.0	23.75	23.75	23.8
man_shan_bian_ye	3	3	1.151	21.0	21.0	21.0	21.0
po_bu_ji_dai	3	3	1.123	17.0	17.0	17.0	17.0
qing_chu_yu_lan	3	3	1.121	20.0	20.0	20.0	20.0
ren_shou_nian_feng	3	3	1.135	23.0	23.0	23.0	23.0
shi_wai_tao_yuan	3	3	1.127	21.0	21.0	21.0	21.0
wu_ti_tou_di	3	3	1.126	17.0	17.0	17.0	17.0

As shown in Table I, the word error index is 1, that means for all sequences with the incorrect expression of the test sequence group, the first character is wrong for the idiom expression. Here for convenience to represent the word site of the sequence, the number of the index starts from 1. The result shows all of the sequences with the wrong idiom expression have been detected. For different sequences, the similarity score used to judge is different. The approximate similarity score is in the range of from 17.0 to 26.0. The implementation time for each sequence group is about 1.2 seconds.

Subsequently, we implement the experiment of sequence groups that have the wrong textual expression sequences in the index of 2, which means the second characters of those sequences are in the wrong expression. As shown in Table II, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 17.0 to 26.0. The implementation time for each sequence group is about 1.15 seconds.

TABLE II
WORD ERROR INDEX = 2.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.149	19.0	19.0	19.0	19.0
chu_mu_cheng_song	3	3	1.166	22.0	22.0	22.0	22.0
fang_ge_zong_jiu	3	3	1.255	21.0	18.75	18.75	18.8
jing_bing_qiang_jiang	3	3	1.27	26.0	23.75	23.75	23.8
man_shan_bian_ye	3	3	1.209	21.0	21.0	21.0	21.0
po_bu_ji_dai	3	3	1.192	17.0	17.0	17.0	17.0
qing_chu_yu_lan	3	3	1.147	20.0	20.0	20.0	20.0
ren_shou_nian_feng	3	3	1.156	23.0	23.0	23.0	23.0
shi_wai_tao_yuan	3	3	1.137	21.0	21.0	21.0	21.0
wu_ti_tou_di	3	3	1.089	17.0	17.0	17.0	17.0

Next, we implement the experiment of sequence groups that have the wrong textual expression sequences in the index of 3, which means the third characters of those sequences are in the wrong expression. As shown in Table III, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 17.0 to 26.0. The implementation time for each sequence group is about 1.15 seconds.

TABLE III
WORD ERROR INDEX = 3.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.149	19.0	19.0	19.0	19.0
chu_mu_cheng_song	3	3	1.166	22.0	22.0	22.0	22.0
fang_ge_zong_jiu	3	3	1.255	21.0	18.75	18.75	18.8
jing_bing_qiang_jiang	3	3	1.27	26.0	23.75	23.75	23.8
man_shan_bian_ye	3	3	1.209	21.0	21.0	21.0	21.0
po_bu_ji_dai	3	3	1.192	17.0	17.0	17.0	17.0
qing_chu_yu_lan	3	3	1.147	20.0	20.0	20.0	20.0
ren_shou_nian_feng	3	3	1.156	23.0	23.0	23.0	23.0
shi_wai_tao_yuan	3	3	1.137	21.0	21.0	21.0	21.0
wu_ti_tou_di	3	3	1.089	17.0	17.0	17.0	17.0

Subsequently, we implement the experiment of sequence groups that have the wrong textual expression sequences in the index of 4, which means the fourth characters of those sequences are in the wrong expression. As shown in Table IV, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 17.0 to 26.0. The implementation time for each sequence group is about 1.14 seconds.

TABLE IV
WORD ERROR INDEX = 4.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.159	19.0	19.0	19.0	19.0
chu_mu_cheng_song	3	3	1.149	22.0	22.0	22.0	22.0
fang_ge_zong_jiu	3	3	1.155	21.0	21.0	21.0	21.0
jing_bing_qiang_jiang	3	3	1.259	26.0	23.75	23.75	23.8
man_shan_bian_ye	3	3	1.152	21.0	21.0	21.0	21.0
po_bu_ji_dai	3	3	1.137	17.0	17.0	17.0	17.0
qing_chu_yu_lan	3	3	1.107	20.0	20.0	20.0	20.0
ren_shou_nian_feng	3	3	1.146	23.0	23.0	23.0	23.0
shi_wai_tao_yuan	3	3	1.152	21.0	21.0	21.0	21.0
wu_ti_tou_di	3	3	1.096	17.0	17.0	17.0	17.0

The operations mentioned above is all about controlling one of the characters of the sequence to be wrong in expression, followed by detecting and calculating the similarity between the sequence with wrong expression and the idiom with correct expression. To better test the performance of our algorithm, we design the following experiments, which set two characters of the sequence to be wrong and test the similarity score. As shown in Table V, setting the first and second characters of the sequences to be wrong and then testing the similarity score, meanwhile recording the implementation time for each sequence group. As shown in Table V, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 17.0 to 26.0. The implementation time for each sequence group is about 1.14 seconds.

TABLE V
WORD ERROR INDEX = 1,2.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.161	19.0	19.0	14.0	16.5
chu_mu_cheng_song	3	3	1.188	22.0	16.0	22.0	19.0
fang_ge_zong_jiu	3	3	1.142	21.0	21.0	18.75	19.9
jing_bing_qiang_jiang	3	3	1.155	26.0	26.0	21.0	23.5
man_shan_bian_ye	3	3	1.168	21.0	21.0	21.0	21.0
po_bu_ji_dai	3	3	1.133	17.0	17.0	17.0	17.0
qing_chu_yu_lan	3	3	1.126	20.0	20.0	19.75	19.9
ren_shou_nian_feng	3	3	1.13	23.0	23.0	23.0	23.0
shi_wai_tao_yuan	3	3	1.143	21.0	21.0	21.0	21.0
wu_ti_tou_di	3	3	1.086	17.0	17.0	17.0	17.0

After testing the influence of the two characters in the start of the sequence to the similarity score calculation, we have also tested the influence of the two characters in the end of the sequence to the similarity score calculation. As shown in Table VI, setting the third and fourth characters of the sequences to be wrong and then testing the similarity score, meanwhile recording the implementation time for each sequence group. As shown in Table VI, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 17.0 to 26.0. The implementation time for each sequence group is about 1.25 seconds.

B. Idiom Correction by Tone Pronunciation

Due the pinyin tone of the sequence is another important factor for the similarity calculation for our similarity analysis and replacement algorithm, we implemented the following two kinds of experiments to test the performance of our algorithm under the influence of wrong pinyin tone. Because the tone of pinyin can make the word character, so in this condition, the textual expression of the sequence usually are different.

TABLE VI
WORD ERROR INDEX = 3,4.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.163	19.0	19.0	19.0	19.0
chu_mu_cheng_song	3	3	1.154	22.0	22.0	22.0	22.0
fang_ge_zong_jiu	3	3	1.222	21.0	18.75	18.75	18.8
jing_bing_qiang_jiang	3	3	1.24	26.0	23.75	23.75	23.8
man_shan_bian_ye	3	3	1.646	21.0	21.0	21.0	21.0
po_bu_ji_dai	3	3	1.14	17.0	17.0	17.0	17.0
qing_chu_yu_lan	3	3	1.111	20.0	20.0	20.0	20.0
ren_shou_nian_feng	3	3	1.141	23.0	23.0	23.0	23.0
shi_wai_tao_yuan	3	3	1.185	21.0	21.0	21.0	21.0
wu_ti_tou_di	3	3	1.148	17.0	17.0	17.0	17.0

As shown in Table VII, setting the first and second tone numbers of the sequence pinyin to be wrong and then testing the similarity score, meanwhile recording the implementation time for each sequence group. As shown in Table VII, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 8.75 to 21.75. The implementation time for each sequence group is about 1.30 seconds.

TABLE VII
TONE ERROR INDEX = 1,2.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.169	19.0	12.75	11.5	12.1
chu_mu_cheng_song	3	3	1.293	22.0	8.75	10.75	9.8
fang_ge_zong_jiu	3	3	1.233	21.0	11.5	15.75	13.6
jing_bing_qiang_jiang	3	3	1.18	26.0	20.0	17.75	18.9
man_shan_bian_ye	3	1	1.154	21.0	9.75	11.75	10.8
po_bu_ji_dai	3	2	1.139	17.0	16.75	5.75	11.2
qing_chu_yu_lan	3	3	1.152	20.0	14.5	15.5	15.0
ren_shou_nian_feng	3	3	1.144	23.0	21.75	15.0	18.4
shi_wai_tao_yuan	3	3	1.114	21.0	7.75	12.75	10.2
wu_ti_tou_di	3	3	1.101	17.0	16.5	16.5	16.5

As shown in Table VIII, setting the third and fourth tone numbers of the sequence pinyin to be wrong and then testing the similarity score, meanwhile recording the implementation time for each sequence group. As shown in Table VIII, the result shows all of the sequences with the wrong idiom expression have been detected. The approximate similarity score is in the range of from 8.5 to 25.5. The implementation time for each sequence group is about 1.35 seconds.

C. Optimization for Audio Text Recognition

To test the optimization and assistance of our whole algorithm to the audio text recognition, we implemented the following comparison experiments. We use some video selected by random and use our algorithms to make the audio text recognition, comparison the output of the methods and the original output from the small size of the whisper model.

As shown in Table IX, HanLP-v1.8.4 [15] is very famous Chinese syntax correction and NLP process framework, but

TABLE VIII
TONE ERROR INDEX = 3,4.

word	n	c	time	s1	s2	s3	sa
ri_chu_san_gan	3	3	1.146	19.0	18.0	8.5	13.2
chu_mu_cheng_song	3	3	1.153	22.0	14.75	21.75	18.2
fang_ge_zong_jiu	3	3	1.162	21.0	10.75	18.5	14.6
jing_bing_qiang_jiang	3	3	1.2	26.0	25.5	23.5	24.5
man_shan_bian_ye	3	3	1.182	21.0	18.75	20.5	19.6
po_bu_ji_dai	3	3	1.129	17.0	16.5	14.5	15.5
qing_chu_yu_lan	3	3	1.149	20.0	18.75	19.0	18.9
ren_shou_nian_feng	3	3	1.141	23.0	15.75	22.75	19.2
shi_wai_tao_yuan	3	3	1.147	21.0	19.75	20.5	20.1
wu_ti_tou_di	3	3	1.102	17.0	15.75	16.5	16.1

TABLE IX
MODEL OR METHOD PERFORMANCES COMPARISON

methods	performance	useful
HanLP	60-85%	less useful
original whisper	81%	normal
our ecVoice	81-95%	useful

in special application like idiom syntax problem analysis or detection is needed more update. Our algorithm performances well in this issue, which performance good in small size model.

V. Conclusion

By the experiments analysis mentioned above, we can get this conclusion: the similarity calculator and judgement part of the whole similarity replacement algorithm has a high efficiency. The running time for each sequence group is about 1.5 seconds, which can be optimised by upgrading the search algorithm in the further research. Compared with tone wrong expression detection, the word wrong expression detection needs less implementation time and has a higher efficiency. For application situations that without non-real-time requirements, this still can provide a strong and fast process performance.

References

[1] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional lstm," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 680–688, 2019.

[2] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[3] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust self-supervised audio-visual speech recognition," *arXiv preprint arXiv:2201.01763*, 2022.

[4] J. L. K. E. Fendji, D. C. Tala, B. O. Yenke, and M. Atemkeng, "Automatic speech recognition using limited vocabulary: A survey," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2 095 039, 2022.

[5] T. Kumar, M. Mahrishi, and G. Meena, "A comprehensive review of recent automatic speech summarization and keyword identification techniques," *Artificial Intelligence in Industrial Applications: Approaches to Solve the Intrinsic Industrial Optimization Problems*, pp. 111–126, 2022.

[6] T. Green, G. Hilkhuysen, M. Huckvale, *et al.*, "Speech recognition with a hearing-aid processing scheme combining beamforming with mask-informed speech enhancement," *Trends in Hearing*, vol. 26, p. 23 312 165 211 068 629, 2022.

[7] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Communication*, 2022.

[8] B. Çeken and N. Taşkın, "Multimedia learning principles in different learning environments: A systematic review," *Smart Learning Environments*, vol. 9, no. 1, pp. 1–22, 2022.

[9] J. Dıéaz-Cintas and S. Massidda, "Technological advances in audiovisual translation," 2019.

[10] L. A. Sparrow, M. Gibbs, and M. Arnold, "The ethics of multiplayer game design and community management: Industry perspectives and challenges," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[11] A. Rajeev and P. Raviraj, "An insightful analysis of digital forensics effects on networks and multimedia applications," *SN Computer Science*, vol. 4, no. 2, p. 186, 2023.

[12] Y. He, T. N. Sainath, R. Prabhavalkar, *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6381–6385.

[13] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, 2019.

[14] A. Solyman, W. Zhenyu, T. Qian, A. A. M. Elhag, M. Toseef, and Z. Aleibeid, "Synthetic data with neural machine translation for automatic correction in arabic grammar," *Egyptian Informatics Journal*, vol. 22, no. 3, pp. 303–315, 2021.

[15] H. He and J. D. Choi, "The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5555–5577. [Online]. Available: https : / / aclanthology.org/2021.emnlp-main.451.

Heart Rate Acquisition and Processing Techniques for a Miniature Wearable Microphone Sensor

Yi Yang Ang*, Kirish Boodhoo*, Wee Ser*, and Rex Xiao Tan*
*R&D Group, Aevice Health Pte Ltd, Singapore

***Abstract* — This paper presents an adaptive windowing approach to advanced heart rate vital estimation from cardiovascular acoustic captured using the recently proposed miniature wearable microphone sensor. The sensor is sufficiently light to securely attach to the user's chest with a medical silicone adhesive and can now estimate vitals beyond the upper and lower limits previously identified. The proposed approach consists of acquiring an initial vital estimate from its spectrogram and then deriving an optimal window length based on the frequency resolution and physical limits of vitals to refine the estimation. Finally, we proposed an estimation correction step if the current estimates differ from past estimates' trends using a 1-step autoregressive model. The proposed approach was verified experimentally in a practical context of overnight vital sign monitoring with healthy volunteers against a medical gold-standard device commonly used in healthcare settings. The proposed approach achieved an accuracy of ± 2.8 beats per minute (bpm) for heart rate measurement against a heart rate measurement reference of ± 3 bpm accuracy.**

Keywords- wearable acoustic sensor, IoT system, heart rate estimation, FDA-cleared

I. INTRODUCTION

World Health Organization (WHO) and Centers for Disease Control and Prevention published a fact sheet identifying cardiovascular diseases as one of the leading causes of death worldwide [1], [2]. Accordingly, in recent years the number of wearable devices has grown tremendously to address this market demand [3]–[5] as early detection of cardiovascular diseases through continuous monitoring of heart rate (HR), followed by early interventions, can help alleviate mortalities.

When compared to the wearable single-lead ECG monitoring device [6], wearable acoustic sensors have demonstrated their potential in continuous monitoring of HR [7]–[12], respiration rate (RR) [11], [13], and the detection of anomalies in the heart [14], [15] and lung sounds [16], [17]. Still, a wearable device that measures multiple features or serves multiple functions tends to face hardware limitations of memory and battery capacities, further complicated or worsened by computational complexity and cost.

One cardiac cycle of heart acoustics consists of two sounds (commonly annotated as lub-dub corresponding to the first, S1, and second, S2). The first and second heart sounds are due to the closure of atrioventricular values at the beginning of systole and semilunar valves during diastole, respectively [18]. Recently, our group developed a miniature wearable acoustic sensor for continuously monitoring vitals [19]. Then, we further showed that the HR estimates perform well in most of the HR ranges [20]; however, the performances were challenged in the extreme ends of HR (i.e., HR<60 bpm and 120 bpm<HR).

One of the primary shortcomings of the previous HR estimation algorithm was using a fixed window size in the peak detection of the autocorrelation function, which poses the risk of under- and over-estimation of HR in the two extreme HR ranges.

In this work, we advanced the prior art by proposing an adaptive peak detection approach to improve the HR algorithm performance in the lower and higher ends of HR. Further, we proposed including an estimation correction step using a 1-step autoregressive (AR) model to filter away any predicted miscalculation due to poor signal-to-noise ratio (SNR). Finally, we analyzed the performance of the proposed HR estimation against a non-acoustic-based United States Food and Drug Administration *(*FDA) cleared medical heart rate monitoring device MightySat® Rx (Masimo, United States).

II. METHODS

The proposed algorithms were implemented in a wearable acoustic microphone sensor, AeviceMD (Aevice Health Pte Ltd, Singapore). The wearable weighs approximately 10 g and securely attaches to a user's chest with a medical silicone adhesive. Since the paper's contribution is the proposed approach to enhance HR estimation robustness, we direct interested readers to the device's design and other algorithms

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

besides the HR estimation algorithms to our other works [19]–[22].

A. Overview of HR Estimation Algorithm

AeviceMD has a sample rate of 4000 Hz to enable capturing of adventitious auscultation, $\mathbf{x} \in \mathbb{R}^L$, where L is the total number of samples in 5.12 s [21], [22]. However, since heart sound typically spans a much lower frequency [23], the acoustics were down-sampled to 250 Hz before estimation. Down-sampling serves as a low pass filter [24] and reduces the memory allocation and processing needed in the wearable to perform the task of HR estimation. Next, the time-domain acoustic is peak-enhanced to emphasize the S1-S2 before acquiring the principal frequency of its autocorrelation for peak-interval detection. Fig. 1 illustrates the proposed approach for HR estimation for our miniature wearable microphone sensor.

B. Adaptive Windowing for Peak Detection

Our previous work found that HR estimation based on peak detection from a fixed window interval was not optimal for robustly estimating HR of extreme ends. This is because the earlier fixed window interval prevents the algorithm from identifying S1-S2 peaks with larger or smaller systolic periods corresponding to the lower and upper HR limits, respectively. As part of the effort to advance the signal processing, we proposed to utilize an adaptive window interval to segment the acoustic prior to peak detection at the two extreme ends, as illustrated in Fig. 1. The proposed adaptive window intervals are designed based on the reported systolic period range of 0.2 to 0.4 s [23]–[26].

Concretely, each auscultation, $\mathbf{x} \in \mathbb{R}^L$, S1 and S2 are first enhanced from the down-sampled, $\mathbf{x}_d \in \mathbb{R}^{L_d}$, acoustic by peak thresholding to arrive at $\mathbf{x}_t \in \mathbb{R}^{L_d}$ before estimating the frequency of the smoothed autocorrelation, $\boldsymbol{\gamma}_t \in \mathbb{R}^{L_d}$, from its spectrogram. Here, we proposed to use $N = 1024$ to achieve a frequency resolution of $r_f = 0.244$ Hz/bin, corresponding to approximately $r_b = 14$ beats per min (bpm). Notice because L_d corresponds to 5 s of auscultation, an r_b of 14 bpm can theoretically resolve at least one pair of S1-S2 in each 5 s acoustic.

It is clear from the above that HR estimates that are based solely on frequency estimation of autocorrelation, $\boldsymbol{\gamma}_t$, cannot differentiate S1-S2 and resolve the ambiguity of ± 14 bpm; hence, peak detection is an essential step to arrive at an accurate estimate [12].

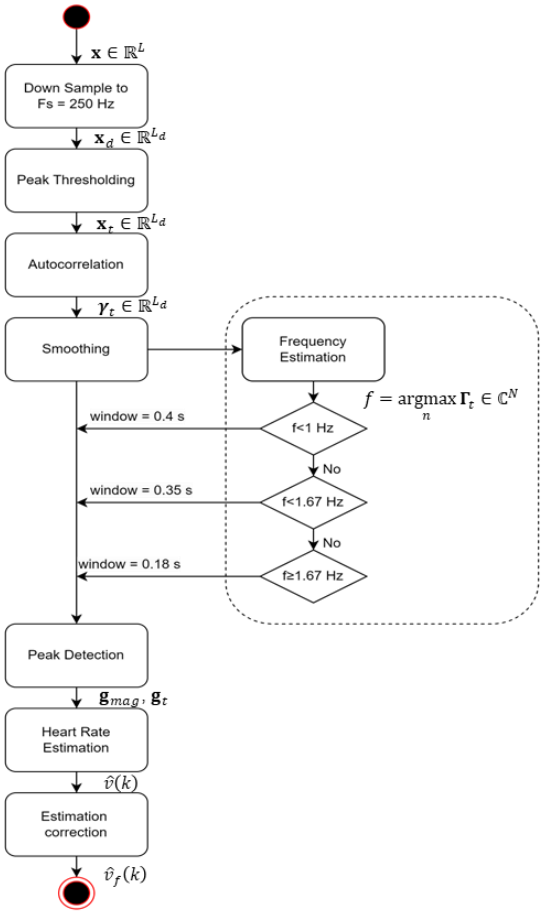


Fig. 1: The proposed HR estimation algorithm based on acoustics from the miniature wearable microphone sensor

The main objective of peak detection is to identify the occurrence of S1 or S2 events. Accordingly, from the frequency estimation, $\boldsymbol{\gamma}_t \in \mathbb{R}^{L_d}$ is segmented into $\boldsymbol{\gamma}_{t,w} \in \mathbb{R}^{\lfloor L_d/W \rfloor}$, where $w = [1, \cdots, W]$ and W is the total number of window segments. Note that $\lfloor . \rfloor$ The local peak magnitude, $\mathbf{p}_{mag} \in \mathbb{R}^W$, and their relative time position, $\mathbf{p}_t \in \mathbb{R}^W$, within each window segment are identified. Lastly, the global peak, $\mathbf{g}_{mag}$, and the global peak's relative time position, $\mathbf{g}_t$, corresponding to the S1 events, can be localized by finding all the stationary points in $\mathbf{p}_{mag}$, i.e., localizing the sign changes from the 2nd-order difference of $\mathbf{p}_{mag}$.

With the global peak magnitude and relative position ($\mathbf{g}_{mag}$ and $\mathbf{g}_t$) corresponding to all the S1 events identified, the kth instance HR, $\hat{v}(k)$, is estimated from the averaged interval difference of $\mathbf{g}_t$.

.

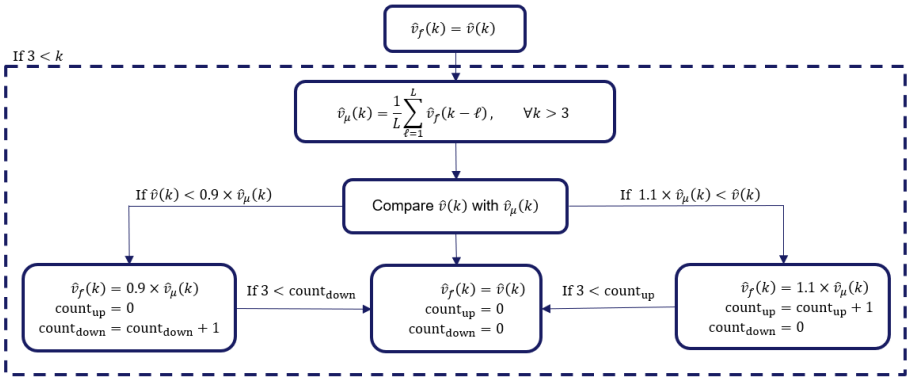


Fig. 2: The proposed process for estimation correction with 1-step AR model when $k < 3$.

C. *Estimation Correction with 1-step AR model*

The estimation correction approach aims to predict the upward or downward trend of the HR estimation algorithm after estimating $3 < k$ HR estimates. By comparing the current estimate, $\hat{v}(k)$, with its previous average HR estimation, $\hat{v}_\mu(k)$, this reduced the occurrence of displaying an instantaneous spike or dip in HR value due to the poor SNR to arrive at the final HR estimation, $\hat{v}_f(k)$. A summary of the approach is presented in Fig. 2. To address the shortcoming of updating an actual high and low HR, we proposed the following conditions if there are three consecutive downtrends (e.g., $3 < \text{count}_{\text{down}}$) or uptrends (e.g., $3 < \text{count}_{\text{up}}$) detected, the current HR will be replaced by the latest computed HR.

D. *Algorithm Performance Evaluation*

The algorithm was evaluated in two stages to demonstrate the improvements attributed to the adaptive peak detection and estimation correction approaches.

In the first stage, we analyzed the HR estimation results of AeviceMD with and without the proposed approaches. In the second stage, we analyzed the results from a main group of 26 healthy subjects and a secondary group of 15 healthy subjects. The performance of the HR estimation from the AeviceMD was compared against MightySat® Rx (Masimo, United States), a gold standard medical heart rate monitoring device.

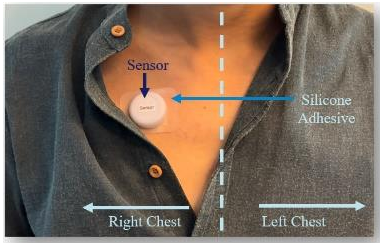


Fig. 3: Placement of the sensor on the right side of the chest wall

The main group has an age distribution between 21 to 80 (40% of the main group is male) with a Cohort body mass index (BMI) of $23.8 \pm$ kg/m.

The participant age distributions are as follows:

- 21 - 29 years old: six subjects
- 30 - 39 years old: five subjects
- 40 - 49 years old: five subjects
- 50 - 59 years old: five subjects
- 60 - 69 years old: three subjects
- 70+ years old: two subjects

On the other hand, the secondary group age distribution is between 20 to 35, where the subjects' HR was analyzed after intense exercise to elevate their heart rate.

All participants were seated on a reclining chair throughout the validation study with AeviceMD attached to their right upper chest using a silicone adhesive, as depicted in Fig. 3. The reference device was fitted to the left index finger of the subjects.

The HR of each participant was first recorded for ten minutes at rest, followed by another five minutes post-movement. At the onset of the tenth minute, the subjects were requested to simulate a tossing scenario that may happen if the device was to be worn throughout the sleep cycle of the subject. The tossing scenario involves turning the body side way (left-to-right and right-to-left). The subjects could adjust their body position for comfort during the recording session. For the secondary group recruited, the participants were asked to perform 2 minutes of jumping jacks to accelerate their HRs. The HR output of AeviceMD is updated every five seconds, whereas the HR reference device's HR is updated every second. Therefore, the data points of the HR reference device were averaged to produce five-second-average data in the validation study.

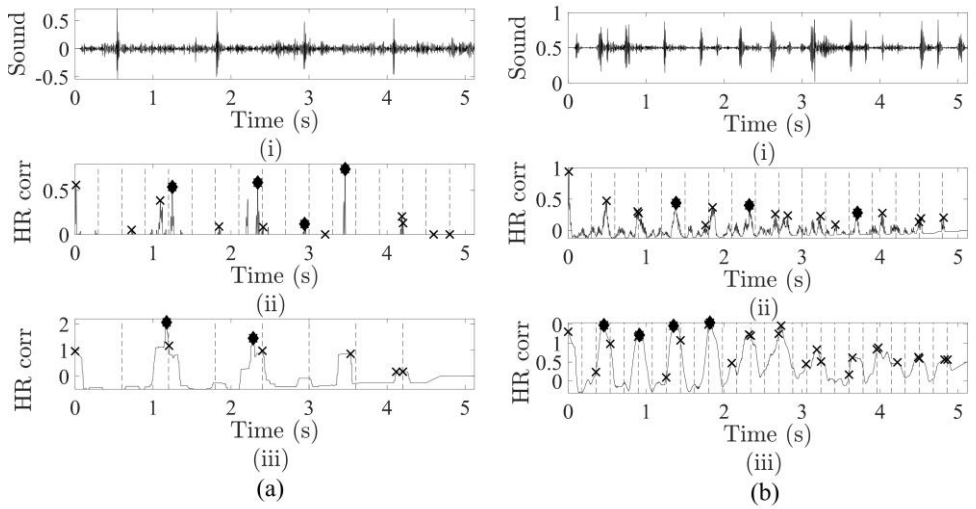


Fig. 4. Comparison in the results of HR estimation algorithm with the (ii) non-adaptive and (iii) adaptive windowing approach employed in the peak detection of HR algorithm. Two scenarios were evaluated: (a) low HR and (b) high HR. HR: Heart rate, HR corr: HR autocorrelation output, marker 'x': local peaks, marker '♦': global peaks.

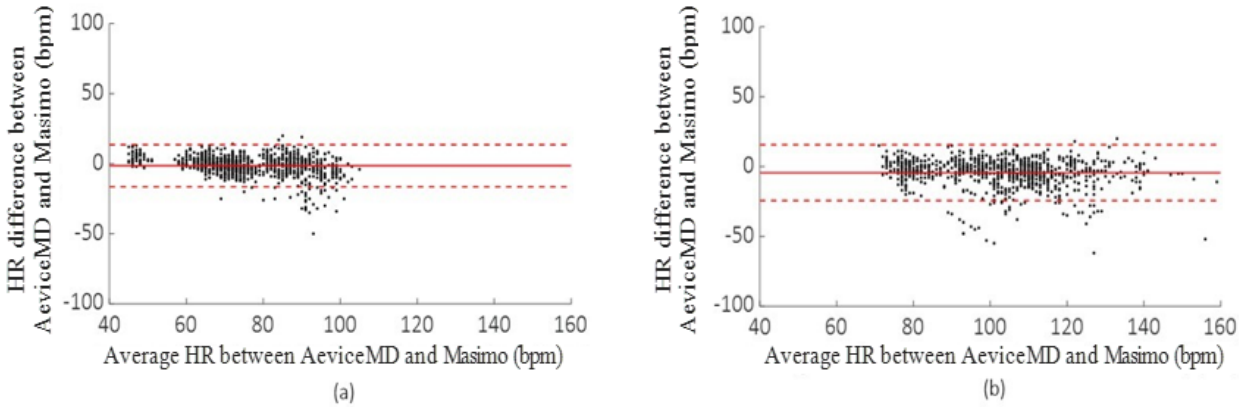


Fig. 5. The Bland-Altman heart rate distribution at (a) rest, main group, and (b) post-exercise, secondary group.

Table 1: Mean Bias and Standard Deviation of the AeviceMD HR Measurement

	Main Group		Secondary Group	
	Rest	Post Movement	Rest	Post Exercise
Mean Bias (bpm)	2.8	3.0	-1.4	-4.4
Standard Deviation (bpm)	8.3	8.1	7.5	10.0

III. Results

Because the heart rate measurement reference has an accuracy of ± 3 bpm, the mean bias of AeviceMD should also be within ± 3 bpm of the reference device, i.e., the maximum acceptable bias is ± 6 bpm of the reference. Overall, we found that the AeviceMD HR estimates' mean bias is within the acceptable range.

Specifically, the result from the first stage HR estimation analysis is depicted in Fig. 4. In the case of lower reference HR (i.e., 47 bpm in Fig. 4(a)), the estimated HR without and with the adaptive peak detection was 90 bpm and 53 bpm, respectively. The adaptive peak detection has reduced the error from 43 bpm to 6 bpm. While in the high reference HR (i.e., 129 bpm in Fig. 4(b)), the estimated HR without and with the adaptive peak detection was 54 bpm and 133 bpm, respectively. Likewise, the adaptive peak detection significantly reduced the error from 75 bpm to 4 bpm.

Table 1 tabulates the results from the second stage study. Here, we report a mean bpm bias of 2.8 and 3.0 for subjects at rest and post-movement, respectively, for the main group. While the second group reports a mean bpm bias of -1.4 and -4.4 at rest and post-exercise, respectively. The bland-Altman plots of the second stage HR estimation analysis result for at-rest and post-exercise are depicted in Fig. 5a and Fig. 5b, respectively.

IV. Conclusion

An adaptive window interval and estimation correction approach for a miniature wearable microphone sensor has been proposed and implemented on the measured heart acoustic signal for HR estimation. The proposed approach results were benchmarked against an FDA-cleared HR monitoring device, MightySat® Rx, in two extreme HR ends. In the first stage, the adaptive window interval for peak detection and estimation correction approach reduced the error from 75 bpm and 43 bpm to 4 bpm and 6 bpm, respectively. In the second stage, the maximum reported mean absolute bias was 2.8 bpm, 3.0 bpm, and 4.4 bpm for at-rest, post-movement and post-exercise, respectively. Even though the post-exercise scenario was not the intended use case of the AeviceMD, the maximum mean was still within ±6 bpm of the reference device. The purpose of including the post-exercise test was to validate the higher HR range, which was unavailable in the resting HR of healthy subjects. Overall, this work advances the HR estimation algorithm for a miniature wearable acoustic microphone sensor to estimate the HR of the upper and lower extreme ends.

Acknowledgment

The authors would like to thank Olivia Lee, Shanelle Ng, Wilson Quek, and the management of Aevice Health Pte Ltd for their help and permission to publish this paper.

References

[1] World Health Organization, "The Top 10 Causes of Death," *Fact Sheets*, 2020.

[2] Centers for Disease Control and Prevention, "Underlying cause of death 1999-2020," *CDC Wonder Online Database*.

[3] S. M. A. Iqbal, I. Mahgoub, E. Du, M. A. Leavitt, and W. Asghar, "Advances in healthcare wearable devices," *NPJ Flex. Electron.*, vol. 5, no. 1, p. 9, 2021.

[4] S. M. Lee and D. Lee, "Healthcare wearable devices: an analysis of key factors for continuous use intention," *Serv. Bus.*, vol. 14, no. 4, pp. 503–531, 2020.

[5] M. Smuck, C. A. Odonkor, J. K. Wilt, N. Schmidt, and M. A. Swiernik, "The emerging clinical role of wearables: factors for successful implementation in healthcare," *NPJ Digit. Med.*, vol. 4, no. 1, p. 45, 2021.

[6] D. D. Mehta, N. T. Nazir, R. G. Trohman, and A. S. Volgman, "Single-lead portable ECG devices: Perceptions and clinical accuracy compared to conventional cardiac monitoring," *J. Electrocardiol.*, vol. 48, no. 4, pp. 710–716, 2015.

[7] P. Sharma, S. A. Imtiaz, and E. Rodriguez-Villegas, "An algorithm for heart rate extraction from acoustic recordings at the neck," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 246–256, 2018.

[8] P. Corbishley and E. Rodriguez-Villegas, "Breathing detection: towards a miniaturized, wearable, battery-operated monitoring system," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 196–204, 2007.

[9] G. Chen, S. A. Imtiaz, E. Aguilar–Pelaez, and E. Rodriguez–Villegas, "Algorithm for heart rate extraction in a novel wearable acoustic sensor," *Healthc. Technol. Lett.*, vol. 2, no. 1, pp. 28–33, 2015.

[10] S. Verma, M. Dhanak, and J. Frankenfield, "Visualizing the effectiveness of face masks in obstructing respiratory jets," *Phys. Fluids*, vol. 32, no. 6, p. 61708, 2020.

[11] R. Abbasi-Kesbi, A. Valipour, and K. Imani, "Cardiorespiratory system monitoring using a developed acoustic sensor," *Healthc. Technol. Lett.*, vol. 5, no. 1, pp. 7–12, 2018.

[12] T. T. Zhang *et al.*, "Sound based heart rate monitoring for wearable systems," in *2010 International Conference on Body Sensor Networks*, 2010, pp. 139–143.

[13] J. Zhang, W. Ser, and D. Y. T. Goh, "A novel respiratory rate estimation method for sound-based wearable monitoring systems," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 3213–3216.

[14] T. Wang, M. Gong, X. Yu, G. Lan, and Y. Shi, "Acoustic-pressure sensor array system for cardiac-sound acquisition," *Biomed. Signal Process. Control*, vol. 69, p. 102836, 2021.

[15] H. Ren, H. Jin, C. Chen, H. Ghayvat, and W. Chen, "A novel cardiac auscultation monitoring system based on wireless sensing for healthcare," *IEEE J. Transl. Eng. Heal. Med.*, vol. 6, pp. 1–12, 2018.

[16] S.-H. Li, B.-S. Lin, C.-H. Tsai, C.-T. Yang, and B.-S. Lin, "Design

of wearable breathing sound monitoring system for real-time wheeze detection," *Sensors*, vol. 17, no. 1, p. 171, 2017.

[17] D. Oletic, B. Arsenali, and V. Bilas, "Low-power wearable respiratory sound sensing," *Sensors*, vol. 14, no. 4, pp. 6535–6566, 2014.

[18] M. Malik *et al.*, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, 1996.

[19] A. Ibrahim, V. C. A. Koh, S. H. Abdul Hamid, X. R. Tan, Y. Y. Ang, and S. Wee, "System and Method For Monitoring A Pluraity of Bio-Signals, And A Gateway Device Operable Therein.," WO2021236008A1, 2021

[20] V. C. A. Koh, R. X. Tan, and Y. Y. Ang, "A Miniature Wearable Microphone Sensor for Cardiopulmonary Monitoring," in *2020 IEEE SENSORS*, 2020, pp. 1–4.

[21] V. C. A. Koh, Y. Y. Ang, W. Ser, and R. X. Tan, "Remote monitoring of patient respiration with mask attachment—A pragmatic solution for medical facilities," *Inventions*, vol. 6, no. 4, 2021, doi: 10.3390/inventions6040081.

[22] B. Xue *et al.*, "Distance-Based Detection of Cough, Wheeze, and Breath Sounds on Wearable Devices," *Sensors*, vol. 22, no. 6, 2022, doi: 10.3390/s22062167.

[23] P. C. Adithya, R. Sankar, W. A. Moreno, and S. Hart, "Trends in fetal monitoring through phonocardiography: Challenges and future directions," *Biomed. Signal Process. Control*, vol. 33, pp. 289–305, 2017.

[24] K. Barnova, R. Kahankova, R. Jaros, M. Litschmannova, and R. Martinek, "A comparative study of single-channel signal processing methods in fetal phonocardiography," *PLoS One*, vol. 17, no. 8, p. e0269884, 2022.

[25] H. Boudoulas, P. Geleris, R. P. Lewis, and S. E. Rittgers, "Linear relationship between electrical systole, mechanical systole, and heart rate," *chest*, vol. 80, no. 5, pp. 613–617, 1981.

[26] R. Sarnari, R. Y. Kamal, M. K. Friedberg, and N. H. Silverman, "Doppler assessment of the ratio of the systolic to diastolic duration in normal children: relation to heart rate, age and body surface area," *J. Am. Soc. Echocardiogr.*, vol. 22, no. 8, pp. 928–932, 2009.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Study on Counter-Movement Detection on Dumbbell Curl Using Convolutional Autoencoder

Atsuya Emoto*, Koh Inoue†, and Ryo Matsuoka*‡
*The University of Kitakyushu, Fukuoka, Japan
†Kagawa University, Kagawa, Japan
†National Institute of Advanced Industrial Science and Technology (AIST), Kagawa, Japan
E-mail: ‡r-matsuoka@kitakyu-u.ac.jp

***Abstract*—This paper proposes a counter-movement detection method for dumbbell curl using a convolutional autoencoder in the frequency domain. Specifically, using an inertial motion capture system, we make an acceleration dataset of dumbbell curl without counter-movement and with counter-movement. Then, we convert the counter-movement-free dataset to 3D spectrograms and train a convolutional autoencoder model using them as inputs. Anomaly score statistics are calculated using the counter-movement-free motion test data, and based on them, a threshold value is set to detect counter-movement. Experiments confirmed that the proposed method can detect counter-movement with the highest accuracy compared to the comparison methods.**

***Index Terms*—Weight training, counter-movement detection, spectrogram, convolutional autoencoder.**

I. Introduction

Free weight training with dumbbells is helpful to compensate for the lack of physical activity because it is easy to do at home. However, weight training with counter-movement forms can increase the risk of muscle, tendon, and joint injury. This problem is severe for beginners not supervised by a trainer with expertise or experience [1]–[4].

Recently, "activity recognition" using machine learning approaches has attracted much attention. Activity recognition research analyzes motion data, such as coordinates and acceleration, to recognize what kind of activity a person is engaged in. Research interest is also shifting from "activity recognition" to "activity quality analysis," which is expected to be used for digital healthcare and improving the performance of athletes.

Velloso et al. proposed a model that estimates wrist, elbow, and shoulder joint angles from data captured by Microsoft Kinect and measures the range of motion of joint angles and repetitive movements [3]. A rule-based algorithm can classify some weight-lifting exercises using the motion features extracted by their model. Moreover, they provided a user interface system that provides real-time activity quality and error feedback. Kowsar et al. measure unilateral biceps curl exercises using an accelerometer attached to glove, upper arm, dumbbell, and belt, mathematically model and extract features of normal motion from these measured data, and threshold them to propose an algorithm for detecting anomalies in which the motion is partially different from the correct motion [4]. These methods predict and detect incorrect form movements in weight-lifting exercises. Therefore, they did not explicitly consider the detection of counter-movement during the use of heavy weights with high loads on the human body.

In this paper, we propose a counter-movement detection method for a dumbbell curl based on a convolutional autoencoder (CAE). The acceleration data of dumbbell arm curl movements with and without counter-movement are measured by inertial motion capture to make a dataset. A CAE model is trained using only the counter-movement-free dataset. Then, we input test acceleration data with and without counter-movement to the trained CAE and detect counter-movement by thresholding the reconstruction error.

II. Proposed Method

A standard dumbbell curl is a single-joint exercise primarily involving elbow joint movement. When using an appropriate weight, the dumbbells can be raised and lowered slowly without needing additional movement. This is considered the correct form. However, it becomes challenging to maintain a slow and controlled movement when attempting to perform the exercise with heavy dumbbells. In such cases, a slight counter-movement may be necessary to generate enough force to lift the weight. This is considered the incorrect form of this exercise because it deviates from the original purpose of stimulating only the targeted muscles. We aim to detect this counter-movement based on the acceleration data measured by the inertial motion capture system. The procedure of the proposed method is briefly described below.

A. Dataset

Inertial sensors were attached to four locations: the back of the hand, the proximal end of the forearm, the proximal end of the upper arm, and the middle trapezius of the subjects. We measured the acceleration data when the subjects performed dumbbell curls for 10 repetitions at regular intervals according to a metronome. The motion data without counter-movement was obtained by measuring the arm curl with a 2kg dumbbell. On the other hand, the following three types of motion data with counter-movement were measured based on the base weight that each subject could barely lift for 10 repetitions without counter-movement.

1) 10 repetitions with a dumbbell of the base weight +2kg.
2) the first five of 10 repetitions with the base weight +4kg.
3) the last five of 10 repetitions with the base weight +4kg.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I: Counter-movement detection results: (I)-(III) the comparison methods and (IV) our method.

Method	Input	Model	Recall (%)
(I)	1D time-series signal	AE	16
(II)	1D power spectrum	AE	46
(III)	1D spectrogram (3D $\rightarrow$ 1D)	AE	78
(IV)	3D spectrogram	CAE	88

The lifting and lowering motion in the dumbbell curl was defined as one repetition with the signal length T, and a total of 6899 samples of data without counter-movement and 80 samples of data with counter-movement were collected from five subjects, respectively.

B. Training the CAE Model

The data without counter-movement were considered normal data, and the data with counter-movement were considered anomalies data, and only the normal data were used to train the proposed CAE model consisting of encoders and decoders in the 6 layer. Note that 5899 of the data without counter-movement was used as training data and 1000 as test data. Each acceleration data was converted into a 3D spectrogram and input into the CAE using the following procedure. Each time series signal from the four measurement locations along the X-axis was transformed by STFT with 64 blocks and 80% overlap, generating a 2D spectrogram. Then, four 2D spectrograms were horizontally concatenated. We performed the same process on the Y- and Z- axis acceleration data, concatenated the spectrograms in the three-dimensional direction, and normalized them to their maximum value.

We used MSE (mean squared error) as the loss function of the CAE.

C. Counter-movement detection

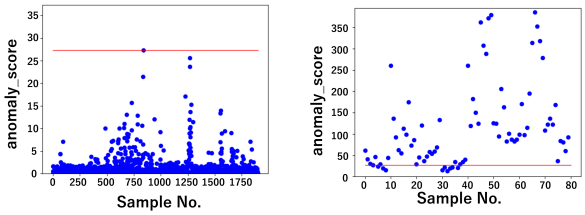
We assume that the distribution of the MSE (mean squared error) of the correct dataset not used for training the CAE can be approximated by a normal distribution with mean μ and variance σ, the anomaly score α is calculated by

$$\alpha = \frac{(\mathrm{MSE}(\mathbf{x}, \mathbf{y}) - \mu)^2}{\sigma^2}, \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the N-dimensional vectors of the input and output of the CAE, respectively. In the experiments, we calculated the anomaly score of the K normal data not used for training and set the maximum value of them as the threshold value $\alpha_{\mathrm{th}} = \max\{\alpha_1, \ldots, \alpha_K\}$. When the acceleration data with or without counter-movement is input into the trained CAE, if the reconstruction error is less than the threshold α_{th}, it is classified as counter-movement-free. In contrast, if the error is higher, it is classified as counter-movement.

III. Experiments

To verify the effectiveness of the proposed method, we compared our method with three methods based on autoencoder (AE) (for details, see Method (I)-(III) in Table I). We trained the comparison methods using the same dataset as the proposed method and followed the same procedure to set the threshold value for counter-movement detection (see Section II-C).



(a) (b)

Fig. 1: The anomaly scores of the proposed method calculated from the test data (a) without and (b) with counter-movement, where the red line indicates the threshold value.

Figures 1 (a) and (b) show the calculated anomaly scores for the test data without and with counter-movement, respectively, and the red line indicates the threshold value (where $\alpha_{\mathrm{th}} = 27.3$). A recall ratio was calculated from these anomaly scores and the threshold value, indicating the proportion of those labeled counter-movement predicted to be an anomaly. The same procedure was used to calculate the recall ratio of the comparison methods.

The fourth column in Table I shows the recall ratio of all the methods. The proposed method has the highest counter-movement detection accuracy of 88% in all methods, which is about five times more accurate than that of the AE method with 1D time-series acceleration signals as input.

IV. Conclusion

This paper proposed a counter-movement detection method for dumbbell curl based on a CAE model. We applied STFT to the acceleration data and obtained a 3D spectrogram. Then, we trained the CAE using the counter-movement-free dataset. Anomaly score statistics were calculated using the counter-movement-free test data, and based on them, a threshold value was set to detect counter-movement. We confirmed that experiments showed the superiority of the proposed method compared with the comparative method based on AE.

Acknowledgment

This work was partially supported by MEXT Promotion of Distinctive Joint Research Center Program Grant Number# JPMXP 0621467946. Acceleration datasets were measured with Xsens MTw Awinda, and analysed using Deeplearning-BOX/Alpha Workstation at The University of Kitakyushu.

References

[1] David A Winter, *Biomechanics and motor control of human movement*, John Wiley & Sons, 2009.

[2] Vladimir M Zatsiorsky, William J Kraemer, and Andrew C Fry, *Science and practice of strength training*, Human Kinetics, 2020.

[3] Eduardo Velloso, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks, "Qualitative activity recognition of weight lifting exercises," in *Proc. of Augmented Human Int. Conf.*, 2013, pp. 116–123.

[4] Yousef Kowsar, Masud Moshtaghi, Eduardo Velloso, Lars Kulik, and Christopher Leckie, "Detecting unseen anomalies in weight training exercises," in *Proc. of Australian Conf. Comput.-Human Interact.*, 2016, pp. 517–526.

Causal Analysis of Hepatic Fat Accumulation by Linear Mixed Causal Model

Takashi Nojima*, Yoshito Itoh†, Kota Yano†, Kanji Yamaguchi†, Yuya Seko†, Satoaki Matoba†,
Kenichi Nishioji‡, Koichi Fujiwara§, Shota Kato*, Manabu Kano*
* Kyoto University, Japan
E-mail: nojima.takashi.44e@st.kyoto-u.ac.jp, manabu@human.sys.i.kyoto-u.ac.jp
† Kyoto Prefectural University of Medicine, Japan
‡ Japanese Red Cross Kyoto Daini Hospital, Japan
§ Nagoya University, Japan

***Abstract*—Non-alcoholic fatty liver disease (NAFLD) is a fatty liver disease without a history of excessive alcohol intake. We aim to identify the causal factors contributing to the onset of NAFLD by analyzing medical data with a causal discovery approach. The linear mixed causal model (LiM) is a causal discovery method applicable to mixed data containing both continuous and discrete variables and thus suitable for medical data analysis. However, the computational time of LiM increases substantially when the number of variables reaches dozens. To solve this problem, we improved the optimization procedure used in LiM in this study. Furthermore, we applied this improved model to a causal analysis of medical data involving physical examination and blood test results. As a result, alanine aminotransferase (ALT), triglyceride (TG), gamma-glutamyl transpeptidase (γ-GTP), uric acid (UA), high sensitive C-reactive protein (hs-CRP), body fat mass (BFM), body mass index (BMI), and fecal hemoglobin (F-Hb) were estimated to be factors for the accumulation of fat in the liver.**

I. Introduction

Non-alcoholic fatty liver disease (NAFLD) is a type of fatty liver disease that develops without a clear history of alcohol consumption, and the number of patients with this disease has been increasing [1]. Previous studies have reported correlations between many medical examination items and the onset of NAFLD [2], [3]. However, few studies have focused on analyzing the causal relationship between these indicators and the NAFLD development [4]. We aim to identify causal factors of NAFLD by analyzing medical data.

The method of discovering causal relationships from data is called causal discovery. One of the causal discovery methods that can be applied to data containing both continuous and discrete variables is the linear mixed causal model (LiM) [5]. LiM has the problem of being computationally expensive when applied to data with dozens of variables. In this study, we extend LiM so that it can be applied to data with dozens of variables. We then estimate the causes of NAFLD by applying our proposed method to medical data.

II. Method

The LiM model estimates the adjacency matrix $\boldsymbol{B}$ representing causal relationships between variables by minimizing the negative log-likelihood. Zeng [5] employed a two-stage optimization process involving `GlobalSearch` and `LocalSearch` to determine $\boldsymbol{B}$. In the first stage, `GlobalSearch` estimates the causal skeleton, and in the second step, `LocalSearch` searches the space of causal relationships having the estimated skeleton to obtain the one with the best log-likelihood score. `LocalSearch` was introduced to avoid falling into local optima, but requires a computational complexity of $\mathcal{O}(2^{P^2})$, where P is the number of variables. The conventional LiM model therefore increases the computational time significantly when the number of variables reaches dozens.

To reduce the computational time, we use `GlobalSearch` with Monotonoic Basin Hopping (MBH) instead of the two-stage optimization process. MBH is a stochastic optimization algorithm that involves iteratively performing a sequence of actions, which include randomly perturbing the current local optimum and conducting optimization with the point after perturbation as the initial value. We adopt the L-BFGS-B method [6], which is an optimization method used in the original LiM. During each iteration k of L-BFGS-B, the computational complexity for computing the Hessian matrix $\boldsymbol{H}_k$ and updating the current solution $\boldsymbol{B}_k$ is $\mathcal{O}(MNP^2)$. Here, N denotes the number of samples, and M represents the parameter within L-BFGS-B that determines the number of saved BFGS corrections, typically ranging between 3 and 20 [7]. Consequently, this method can be applied to data consisting of dozens of variables, where causal discovery is not feasible within a realistic time using LiM.

III. Analysis of Medical Data

The proposed method is applied to medical data provided by Kyoto Prefectural University of Medicine, involving physical examination and blood test results from 323 males aged from 65 to 101 years and 523 females aged from 65 to 99 years. The presence of fatty liver was determined based on computed tomography (CT) scan data, considering liver-to-spleen CT attenuation values ratio (L/S ratio) below 1.0 as indicative of fatty liver [8]. The number of test items is 42, with only the presence or absence of fatty liver being a binary variable,

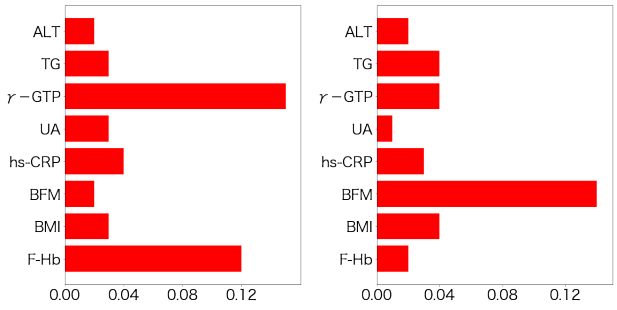
TABLE I: The number of persons with fatty liver.

Sex \ Fatty liver	Negative	Positive	Total
Men	328	27	355
Women	485	38	523
Men (oversampled)	328	328	656
Women (oversampled)	485	485	970

while the others are continuous variables. To address missing values within the medical data, collaborative filtering [9] was employed for imputation. Notably, the number of persons with fatty liver is significantly lower than those without it, necessitating measures to rectify this imbalance. To tackle this issue, the data of persons with fatty liver were oversampled using SMOTE [10]. Table I presents the counts of persons with and without fatty liver before and after oversampling in male and female datasets. Data collection and analysis received approval from the Research Ethics Committee of the Hospital.

IV. Results and Discussion

Figure 1 illustrates the onset factors of fatty liver estimated by our method for each gender and their causal effects on the development of fatty liver. In both genders, alanine aminotransferase (ALT), triglyceride (TG), gamma-glutamyl transpeptidase (γ-GTP), uric acid (UA), high sensitive C-reactive protein (hs-CRP), body fat mass (BFM), body mass index (BMI), and fecal hemoglobin (F-Hb) were estimated to have a positive effect on the accumulation of hepatic fat. This result shows the contributory role of indicators associated with metabolic syndrome, such as TG, BMI, and BFM, in fostering fatty liver onset. Previous studies have reported the link between metabolic syndrome and NAFLD development [11]. Therefore, the outcome of our estimations aligns with existing research, suggesting the involvement of TG, BMI, and BFM in liver fat accumulation. A previous study, which followed 4,138 healthy men for seven years, reported that relatively high levels of hs-CRP increase the risk of developing NAFLD [12]. The results of this analysis corroborate the claims of this preceding study. Furthermore, a preceding study that tracked 4,954 participants for five years reported that high uric acid levels increase the risk of developing NAFLD [13], and the findings of this analysis support this assertion. Correspondingly, prior research has also documented a correlation between elevated levels of ALT and the NAFLD presence [14]. Based on these results, the factors indicated in Fig. 1 were estimated to be the causes of hepatic fat accumulation.



(a) Male **(b)** Female

Fig. 1: (a) Factors contributing to the onset of fatty liver in males as estimated by our method. (b) Factors contributing to the onset of fatty liver in females as estimated by our method.

References

[1] S. L. Friedman, B. A. Neuschwander-Tetri, R. Mary, and A. J. Sanyal, "Mechanisms of nafld development and therapeutic strategies," *Nature Medicine*, vol. 24, pp. 908–922, 2018.

[2] H. Imaizumi, A. Takahashi, N. Tanji, *et al.*, "The association between sleep duration and non-alcoholic fatty liver disease among japanese men and women," *Obesity Facts*, vol. 8, pp. 234–242, 2015.

[3] G. Tsuruta, N. Tanaka, M. Hongo, *et al.*, "Nonalcoholic fatty liver disease in japanese junior high school students: Its prevalence and relationship to lifestyle habits," *Journal of Gastroenterology*, vol. 45, pp. 666–672, 2010.

[4] T. Uchida, K. Fujiwara, K. Nishioji, *et al.*, "Medical checkup data analysis method based on lingam and its application to nonalcoholic fatty liver disease," *Artificial Intelligence in Medicine*, vol. 128, no. 102310, 2022.

[5] Y. Zeng, S. Shimizu, H. Matsui, and F. Sun, "Causal discovery for linear mixed data," in *Proceedings of Machine Learning Research*, vol. 177, 2022, pp. 994–1009.

[6] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, pp. 1190–1208, 1995.

[7] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software*, vol. 23, pp. 550–560, 1997.

[8] M. Iwasaki, Y. Takada, M. Hayashi, *et al.*, "Noninvasive evaluation of graft steatosis in living donor liver transplantation," *Transplantation*, vol. 78, pp. 1501–1505, 2004.

[9] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, pp. 61–70, 1992.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[11] M. Hamaguchi, T. Kojima, N. Takeda, *et al.*, "The metabolic syndrome as a predictor of nonalcoholic fatty liver disease," *Annals of Internal Medicine*, vol. 143, pp. 722–728, 2005.

[12] J. Lee, K. Yoon, S. Ryu, Y. Chang, and H. R. Kim, "High-normal levels of hs-crp predict the development of non-alcoholic fatty liver in healthy men," *Plos One*, vol. 12, no. e0172666, 2017.

[13] J. W. Lee, Y. K. Cho, M. Ryan, *et al.*, "Serum uric acid as a predictor for the development of nonalcoholic fatty liver disease in apparently healthy subjects: A 5-year retrospective cohort study," *Gut Liver*, vol. 4, pp. 378–383, 2010.

[14] S. Jimba, T. Nakagami, M. Takahashi, *et al.*, "Prevalence of non-alcoholic fatty liver disease and its association with impaired glucose metabolism in japanese adults," *Diabetic Medicine*, vol. 22, pp. 1141–1145, 2005.

Investigation of the neuromodulation effect of tVNS on electrical brain activity

W.-J. Kim[1], Y.-S. Lee[1,2], H. Choi[3], J.-J. Song[4], K.-H. Hong[5], H.-J. Hwang[1,2*]

[1]Department of Electronics and Information Engineering, Korea University, Sejong 30019, Republic of Korea

E-mail: kwj970612@korea.ac.kr, ysung947@korea.ac.kr, *hwanghj@korea.ac.kr, Tel: +82-44-860-1762

[2]Interdisciplinary Graduate Program for Artificial Intelligence Smart Convergence Technology,

Korea University, Sejong 30019, Republic of Korea

[3]Department of Medical Sciences, Graduate School of Medicine, Korea University, Seoul 028411, Republic of Korea

E-mail: hyuk76@korea.ac.kr Tel: +82-2-2626-3302

[4]Department of Otorhinolaryngology-Head and Neck Surgery, Korea University Guro Hospital, Seoul 08308, Republic of Korea

E-mail: jjsong23@gmail.com Tel: +82-2-2626-1280

[5]Neurive Co., Ltd., Gimhae, 50969, Republic of Korea

E-mail: kihwan.hong@neurive.com Tel: +82-70-4044-5255

***Abstract* — We investigated the neuromodulation effect of transcutaneous vagus nerve stimulation (tVNS) on brain activity using electroencephalography (EEG). We measured resting-state EEG from ten participants before, during, and after tVNS. The power spectral density (PSD) was estimated in alpha band (8 – 13 Hz) and it was confirmed that significantly increased in frontal and right temporal areas during and after tVNS than before tVNS application. These results demonstrated that the tVNS could modulate the electrical brain activities during and this brain modulation effect was maintained even after tVNS application.**

I. INTRODUCTION

The vagus nerve was known as a parasympathetic branch of the autonomic nervous system. Recently, some previous studies have been conducted to induce symptom relief of neurophysiological diseases such as epilepsy by stimulating the autonomic nervous system using a vagus nerve stimulation (VNS) [1].

VNS was one of the invasive brain stimulation methods to modulate neural activity by applying a weak electric current to the vagus nerve. VNS was known to be effective in improving various neurological disorders such as epilepsy, headaches, and Alzheimer's disease. In addition, VNS could minimize drug-induced side effects and can be an effective treatment option for patients who are resistant to drugs [2]. However, VNS has a limitation in that used only in cases of high severity because stimulation electrodes and generator must be directly implanted into the human body through a surgical operation [3].

To overcome this limitation, recently non-invasive vagus nerve stimulation (transcutaneous VNS: tVNS) has been proposed, and some previous studies conducted to confirm the neuromodulatory effect of tVNS [4]. Some Previous studies reported the clinical effects of tVNS and other studies mainly used the functional magnetic resonance imaging (fMRI) to confirm the neural mechanism of tVNS. However, the different effects of tVNS have been reported depending on the experimental group and stimulation parameters, and previous studies on the neural mechanism of tVNS were still lacking. In particular, studies about the effects of tVNS on neural activity based on electroencephalography (EEG) are lacking compared to studies based on fMRI.

Therefore, we tried to investigate the effect of tVNS on neural activity based on EEG. To this end, resting-state EEG data were measured before, during, and after tVNS application, and we confirmed changes in EEG patterns during each stimulus condition.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. METHOD

A. Data and data acquisition

Ten participants took part in the experiment, and resting state EEG was measured before, during, and after tVNS application. EEG data were recorded at a sampling rate of 1000 Hz through 31 scalp electrodes based on the international 10-20 system (Figure 1).

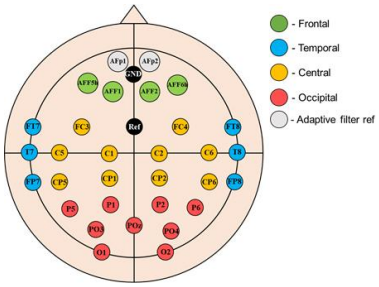


Fig. 1 The montage of EEG electrodes

tVNS was applied to the antihelix of both ears to stimulate the auricular branch of the vagus nerve (Figure 2-a, b). tVNS was applied for 1 minute and then applied through an on/off method in which stimulation was not applied for 1 minute. The corresponding process was repeated 20 times, and stimulation was applied for a total of 40 minutes.

tVNS was applied at 30 Hz with a pulse width of 200 μs. Moreover, tVNS intensity was individually determined from level 1 to 10, depending on the presence of the pain for each participant. In addition, the participants who presence the pain even at the minimum intensity of level 1 were excluded from the experiment (see Table I).

Table I. Intensity of tVNS

Level	Maximum output voltage (Vpp)	Output current ch1 (mA)	Output current ch2 (mA)
1	11.0	0.464	0.558
2	19.6	0.850	0.976
3	22.4	1.014	1.153
4	25.7	1.154	1.293
5	28.2	1.262	1.415
6	30.2	1.370	1.522
7	32.4	1.453	1.619
8	34.0	1.540	1.695
9	36.2	1.161	1.776
10	37.4	1.136	1.827

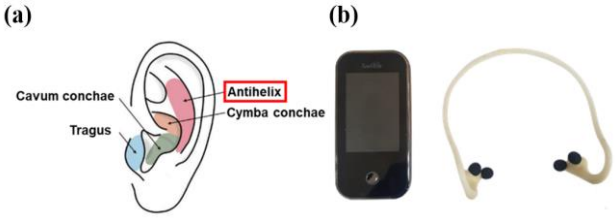


Fig. 2 (a) Location of the tVNS application and (b) tVNS device (Neurive Co., Ltd.)

B. Paradigm

All participants maintained their eyes open (EO) and eye closed (EC) state for 1 minute before and after tVNS application, respectively. In addition, they only maintained the EO state at during tVNS. To compare the EEG changes in before, during, and after stimulation, only EO data was used for further analysis. In particular, only tVNS off data was used to further analysis because of the tVNS on data was contaminated by electrical noise induced from the tVNS (Figure 3).

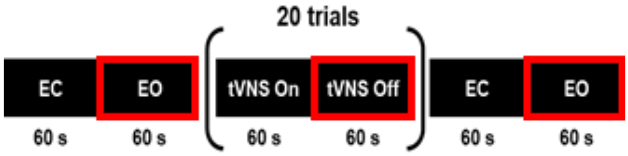


Fig. 3 The experimental paradigm

C. Preprocessing

The recorded EEG raw signals were downsampled to 200 Hz to reduce computational load. The common average reference (CAR) and a band-pass filter (1 – 50 Hz) were performed to remove the noise components. In addition, an adaptive filter applied using the AFp1 and AFp2 electrodes to remove the electrooculography (EOG) components. EEG data in before, during and after tVNS were segmented at 1-second intervals. Moreover, segmented data containing values exceeding ± 75 (μV) excluded from the analysis. Alpha band (8 – 13 Hz) power spectral density (PSD) was estimated using the fast fourier transform (FFT) to confirm the neuromodulation effects on resting state EEG. In addition, to investigate the effects of tVNS according to brain areas, the EEG data was divided into five regions of interests (ROIs). Each alpha band power of before, during and after tVNS application were statistically evaluated using the Wilcoxon signed-rank test with the false discovery rate (FDR) correction.

III. RESULT

Figure 4 shows the average alpha band frequency power topography in the before, during and after tVNS for all participants. It shows the alpha band PSD was increased in the whole brain areas at during tVNS. Moreover, the alpha band PSD was increased in the central and occipital areas at after tVNS.

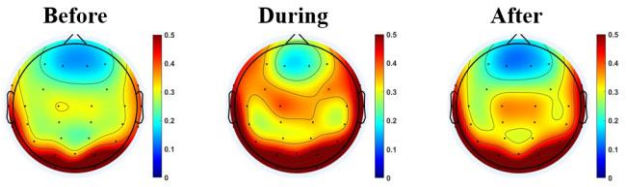


Fig. 4 Topography of the alpha band before, during and after tVNS

Further analysis was performed to precisely compare the differences between before tVNS and during and after tVNS (Figure 5). Figure 5-a shows the differences in average of alpha band PSD during and after tVNS compared to before tVNS for whole participants. Figure 5-b shows the topographic map of -values for the differences in alpha band PSD between during and after tVNS compared to before tVNS, for each channel. Channels exhibiting statistically significant differences ($p < 0.05$) were indicated by red circles in Figure 5-b. As a results, alpha band PSD significantly increased in the frontal and right temporal areas at during tVNS than before tVNS. Especially, significantly increased of alpha band PSD in right temporal area was maintained even after tVNS.

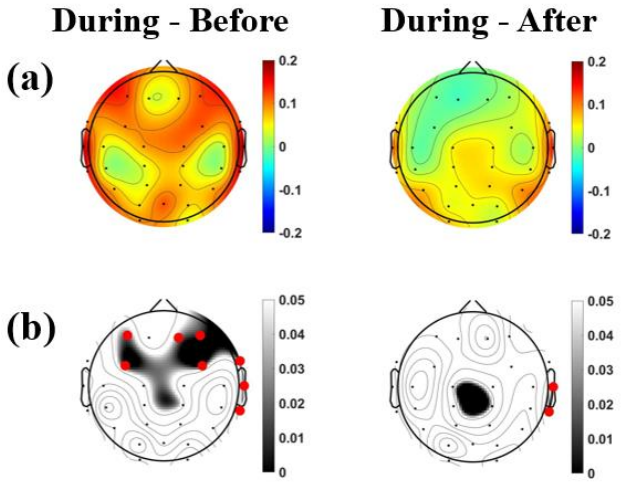


Fig. 5 Differences in alpha band PSD during and after tVNS compared to before tVNS.

Figure 6 shows the alpha band PSD for the ROIs across individual participants. Alpha band PSD was generally increased during tVNS than before tVNS in the frontal and right temporal areas. However, alpha band PSD was decreased during and after tVNS than before tVNS in few participants. These results show that tVNS may not cause the same neuromodulation effects in whole participants.

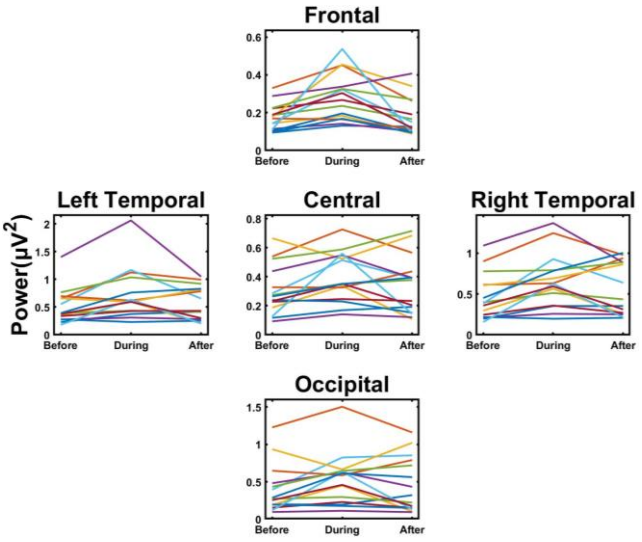


Fig. 6 Changes in individual alpha band PSD according to tVNS condition for each ROIs

Figure 7 shows the average alpha band PSD at the each ROIs for whole participants. As a results, alpha band PSD significantly increased during tVNS than before tVNS in frontal and right temporal areas. In addition, significantly increased of alpha band PSD in right temporal area was maintained even after tVNS than before tVNS (* $p < 0.05$, ** $p < 0.01$).

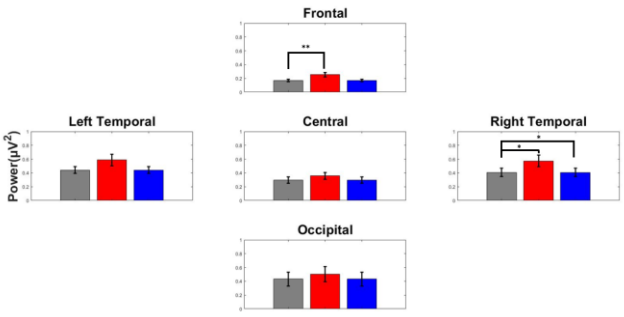


Fig. 7 Changes in mean alpha band PSD according to tVNS condition for each ROIs

IV. Conclusions

In this study, we investigated the neuromodulation effects of tVNS using the resting state EEG data measured before, during, and after tVNS. We could confirm that the alpha band PSD was significantly increased across frontal and right temporal areas during tVNS than before tVNS. Especially, these effects were found to maintained even after tVNS in right temporal area. However, we confirmed the neuromodulation effects of tVNS only on the alpha band. Therefore, in further study, will investigate the neuromodulation effects of tVNS on various frequency bands.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2019R1I1A3A01060732).

References

[1] Ben-Menachem E, Manon-Espaillat R, Ristanovic R et al. (1994) Vagus nerve stimulation for treatment of partial seizures: 1. A controlled study of effect on seizures, Epilepsia 35: 616– 26.

[2] Andreas Straube, Ozan Eren (2021) Vagus Nerve Stimulation for Pediatric and Adult Patients with Pharmaco-resistant Epilepsy, Autonomic Neuroscience - Basic & Clinical 236 102875

[3] Ventureyra ECG (2000) Transcutaneous vagus nerve stimulation for partial onset seizure therapy (2000), Child's Nerv. Syst., 16, 101–102.

[4] Ricci L, Croce P, Lanzone J, Boscarino M, Zappasodi F, Tombini M, Lazzaro VD, and Assenza G (2020) Transcutaneous vagus nerve stimulation modulates EEG microstates and delta activity in healthy subjects, Brain Sci., 10:668.

[5] Andreas Straube, Ozan Eren (2021) tVNS in the management of headache and pain, Autonomic Neuroscience., 236 102875

Gender-Specific EEG Feature-Based Diagnostic Tool for Male Depression Based on Machine-Learning Techniques

Eun-Gyoung Yi[1-3], Miseon Shim[1], Seung-Hwan Lee[3-5], and Han-Jeong Hwang[1, 2*]
[1] Department of Electronics and Information Engineering, Korea University, Sejong
E-mail: yeg0311@korea.ac.kr, shim.miseon@gmail.com, hwanghj@korea.ac.kr Tel: +44-860-1762
[2] Artificial Intelligence Smart Convergence Technology, Korea University, Sejong
E-mail: yeg0311@korea.ac.kr, hwanghj@korea.ac.kr Tel: +44-860-1762
[3] Clinical Emotion and Cognition Research Laboratory, Department of Psychiatry, Inje University, Goyang
E-mail: yeg0311@korea.ac.kr, lshpss@hanmail.net Tel: +44-860-1762, Tel: +31-811-8430
[4] Psychiatry Department, Ilsan Paik Hospital, Inje University, Goyang
E-mail: lshpss@hanmail.net Tel: +31-811-8430
[5] Bwave Inc, Goyang
E-mail: lshpss@hanmail.net Tel: +31-811-8430

***Abstract* — Machine-learning-based computer-aided diagnostic (CAD) systems using neurophysiological data could complement conventional interview-based diagnostic method for major depressive disorder (MDD). In particular, gender-specific and medication-controlled EEG-features can improve the diagnostic performance of CAD system because electroencephalography (EEG) features of MDD patients are influenced from both factors. In this study, we developed a machine-learning-based CAD system using gender-specific and medication-controlled EEG features for MDD patients to improve diagnostic performances. To this end, we used eyes-closed resting-state EEG data recorded from 30 male MDD patients and 30 male healthy controls (HCs). Power spectral density (PSD) was computed to use as EEG-based features and a support vector machine (SVM) classifier with a leave-one-out cross-validation method was used to differentiate two groups. The best diagnostic accuracy of 93.33 % was obtained when 43 PSD features.**

I. INTRODUCTION

After COVID-19 pandemic, mental problems, such as depression, anxiety, and stress, have been increased due to social isolation. Among various mental disorders, the patients with major depressive disorder (MDD) have been increased by approximately 30.00 %, compared to before COVID-19 [1]. Since MDD is the most common mental disorder and is considered less severe disorder compared to other mental disorders like schizophrenia, some MDD patients do not actively treat their symptoms. However, MDD is also a severe mental disorder because that can lead to suicide if not treated in a timely manner.

Fortunately, a recovery rate of MDD is over 54 % higher than other mental disorder. However, if not treated promptly, the recovery rate is decreased until 16 %. Therefore, it is important that MDD patients should be diagnosed and treated in timely manner. Conventionally, mental disorders, including MDD, are diagnosed by interview with clinical experts. Even though it conducted under the well-organized diagnostic criteria, the conventional interview-based diagnosis might be contaminated by subjective interpretation of experts or deception of patients [2].

To assist accurate diagnosis of mental disorders, researchers have been striving to develop the machine-learning-based computer-aided diagnosis (CAD) systems using objective physiological data, like electroencephalography (EEG). When developing EEG-based CAD system for MDD patients, both gender and medication should be controlled because the EEG-features of MDD patients are influenced from those of two factors [3, 4, 5]. In terms of gender, male MDD patients showed the increased long-distance connectivity between the left parieto-occipital and frontal areas regardless of hemispheres, while female MDD patients showed the short-distance connectivity between the left posterior and central areas [6]. Moreover, treated MDD patients under the medications, such as fluoxetine or venlafaxine, showed reduced prefrontal cortex activity, as compared to unmediated MDD patients [7]. Therefore, it is necessary to develop a gender-specific and

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

medication-controlled EEG features-based CAD system for MDD patients to ensure an accurate and precise diagnosis.

In this study, we developed a machine-learning-based CAD system using gender-specific and medication-controlled EEG-based neurophysiological features. Furthermore, we investigated spatial distributions of selected features when achieving the best classification performance to improve the understanding of what neurophysiological traits affect on the diagnostic performance of male MDD patients.

II. METHODS

A. Data acquisition

In the present study, 30 drug-naïve male MDD patients and 30 male healthy controls (HCs) were recruited. To evaluate the symptoms of depression and anxiety in MDD patients, the Hamilton Depression Rating Scale (HAM-D) and Hamilton Anxiety Rating Scale (HAM-A) were used, respectively. Demographic data are summarized in Table 1. This study was approved by the Institutional Review Board of Inje University Ilsan Paik Hospital [2015-04-316/2016-08-017] and conducted in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki), and all participants submitted written informed consent before the experiment.

Table 1. Demographic data of drug naïve male major depressive disorder (MDD) patients and healthy controls (HCs). All participants are male. Abbreviations: HAM-D – Hamilton Depression Rating Scale; HAM-A – Hamilton Anxiety Rating Scale

	MDD	HC
Cases (N)	30	30
Age (years)	36.73±12.37	30.07±7.82
Education	14.97±1.10	15.13±0.97
HAM-D	22.93±5.81	
HAM-A	27.83±6.47	

B. Data analysis

Eyes-closed resting-state EEG data were recorded for 5 minutes using a NeuroScan SynAmps2 (Compumedics USA El Paso, TX, USA) from 64 Ag/AgCl scalp electrodes evenly mounted on a QuikCap according to the extended 10-20 international system. Since two channels were used for references (M1 and M2), only 62 electrodes were used for further analysis. To eliminate artifacts of recorded resting-EEG data, band-pass filtering was conducted based on frequency ranges from 1 Hz to 55 Hz. Then, independent component analysis (ICA) was performed to remove various external artifacts, such as eye blinks and eye movements.

Artifact-free EEG data were down-sampled from the original sampling rate of 1,000 Hz to 200 Hz to reduce the computational cost of a deep-learning model. Then, EEG data was segmented into the length of 35,945-time points (about 3 minutes).

C. Feature extraction

Power spectral density (PSD) was used for neurophysiological features to differentiate MDD patients and HCs. To this end, a Short-time Fourier transform (STFT) method, which can compute spectral power across both time and frequency, was used. Computed spectral powers were averaged across the all-time and across the specific frequency bands. In this study, PSD were quantified for seven different frequency bands [delta: 1-4 Hz; theta: 4-8 Hz; low-alpha: 8-10 Hz; high-alpha: 10-12 Hz; low-beta: 12-22 Hz; high-beta: 22-30 Hz; gamma: 30-55 Hz]. Therefore, the size of extracted PSD-based features was [60 (subjects) x 434 (62-channels x 7 frequency bands)].

D. Classification

Since too many features can hide accurate classification performances occurring bias when using machine-learning algorithms, feature selection method was used to select important features in this study. Among various feature selection methods, an independent *t*-test was used, and features were selected based on *t*-values. We evaluated classification performances sequentially using the selected features from 1 to 50. To differentiate MDD patients from HCs, a support vector machine (SVM) classifier with leave-one-out cross-validation was used.

Based on the classification performances, we investigated the spatial distribution of selected features. To this end, we represented the selected channels and frequency bands to topographical map when achieving the best classification performance.

III. RESULTS

A. Classification performances

Table 2 represents the classification performances between MDD patients and HCs (accuracy, sensitivity, and precision). The best classification accuracy of 93.33 % was achieved when using 43 features.

Table 2. Classification performances (units: %)

Accuracy	Sensitivity	Specificity
93.33	93.33	93.33

B. *Identification of features used for diagnosis*

To evaluate what frequency band plays an important role when differentiating MDD patients and HCs, we investigated the characteristics of 43 features selected when achieved the best classification. As a results, four frequency bands (delta, low-beta, high-beta, and gamma) were used, and the largest number of features was selected from gamma frequency band (the number of features: delta – 11, low-beta – 1, high-beta – 7, and gamma – 24). Fig. 1 represents the characteristic of selected features in terms of number.

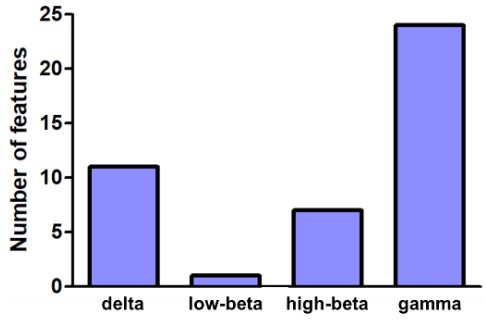


Fig. 1 Features used for classification for each frequency band (delta band: 11, low-beta band: 1, high-beta band: 7, gamma band: 24)

We also investigated the spatial distribution of selected 43 features where is the important brain locations when differentiating MDD patients from HCs. Most of selected channels were located in central area regardless of frequency bands. Fig. 2 represents the location of selected features, red color indicates the selected channels and blue color represents the not-selected channels.

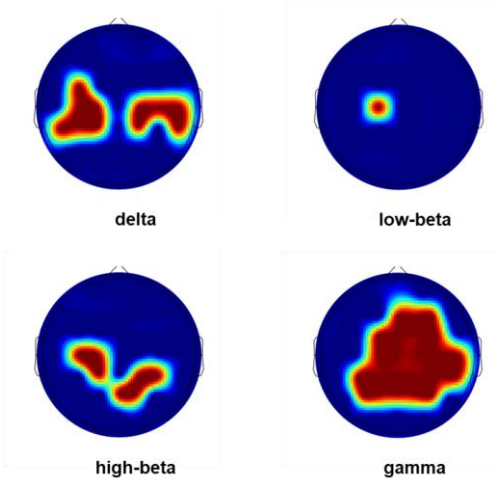


Fig. 2 Topographical map to represent selected 43 features when achieving the best classification performance.

IV. Conclusions

In the present study, we developed a machine-learning-based diagnostic tool utilizing the EEG-based PSD features to assist accurate diagnosis for drug-naïve male MDD patients. The best classification accuracy of 93.33 % was obtained when using 43 PSD features, and the more than half of selected features were allocated in high-frequency bands, such as low-beta, high-beta, and gamma. Moreover, the selected features were predominantly located in the central area of the brain.

Even though the developed EEG-based CAD system showed the high diagnostic performances, there were limitations. First, in this study, only male MDD patients were recruited. To improve the feasibility, the generalized EEG-based CAD system, which can be used regardless of gender, by recruiting female MDD patients. Moreover, in this study, EEG-based features were extracted by hand-craft engineering, which cause low computation efficacy and human-related diagnostic error. Therefore, to improve efficacy and minimize human-related error, the semi-automatic diagnostic tools should be developed by using deep-learning techniques.

In conclusion, we developed the machine-learning-based CAD system using gender-specific EEG-based features to assist accurate and objective diagnosis of drug-naïve male MDD patients and achieved the high diagnostic accuracy. Our developed machine-learning-based CAD system can be a promising diagnostic tool to timely diagnose the male MDD patients.

Acknowledgment

This work was supported by 2022 Startup Growth Technology Development Project (Tips Program, S3197996) supported by the Ministry of SMEs and Startups.

References

[1] A. Saeedi, M. Saeedi, A. Maghsoudi, and A. Shalbaf, “Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach,” *Cognitive Neurodynamics*, vol. 15, no.2, pp. 239-252, JUL 2020.

[2] W. H. Ming, A. Bahar, K. R. Bahar, and M. M. Addi, “Early Detection of Depression using Screening Tools and Electroencephalogram (EEG) Measurements,” *International Journal of Integrated Engineering*, vol. 12, no. 6, pp. 216-228, 2020.

[3] M. Arns, G. Bruder, U. Hegerl, C. Spooner, D. M. Palmer, A. Ethkin, K. Fallahpour, J. M. Gatt, L. Hirshberg, and E. Gordon, “EEG alpha asymmetry as a gender-specific of outcome to

acute treatment with different antidepressant medications in the randomized iSPOT-D study," *Clinical Neurophysiology*, vol. 127, no. 1, pp. 509-519, JAN 2016.

[4] J. L. Stewart, A. W. Bismark, D. N. Towers, J. A. Coan, and J. J. B. Allen, "Resting Frontal EEG Asymmetry as an Endophenotype for Depression Risk: Sex-Specific Patterns of Frontal Brain Asymmetry," *Journal of Abnormal Psychology*, vol. 119, no. 3, pp. 502-512, AUG 2010.

[5] F. S. D. Neto, and J. L. G. Rosa, "Depression biomarkers using non-invasive EEG: A review," *Neuroscience and Biobehavioral Reviews*, vol. 105, pp. 83-93, OCT 2019.

[6] S. Tement, A. Pahor and N. Jaušovec, "EEG alpha frequency correlates of burnout and depression: The role of gender," *Biological Psychology*, vol. 114, pp. 1-12, FEB 2016.

[7] I. A. Cook, A. F. Leuchter, M. Morgan, and E. Witte, "Early changes in prefrontal activity characterize clinical responders to antidepressants," *Neuropsychopharmacology*, vol. 27, no. 1, pp. 120-131, JUL 2002.

Optimal Input Data Search of Deep Learning Model for Cognitive Task Classification Based on NIRS

Ahyun Kim[1,2], Hyung-Tak Lee[1,2] and Han-Jeong Hwang[1,2*]
[1] Department of Electronics and Information Engineering, Korea University, Sejong, Korea
2 Interdisciplinary Graduate Program for Artificial Intelligence Smart Convergence Technology, Korea University, Sejong, Korea
E-mail: kahkah05@korea.ac.kr, htlee@korea.ac.kr Tel: +44 860-1422
E-mail: hwanghj@korea.ac.kr Tel: +44 860-1762

***Abstract* — The near-infrared spectroscopy (NIRS) measures changes in brain signals which quantify the concentration changes of oxygenated hemoglobin and deoxygenated hemoglobin. Previous studies mainly focused on developing models that can learn the characteristics of NIRS signals to improve the performance of classification models. In this study, we investigate to optimize the input data shape in order to extract the relevant features from NIRS signals using Shallow ConvNet, a deep learning model in a brain-computer interface research.**

I. Introduction

A brain-computer interface (BCI) is a technology that reads neural signals from the brain using invasive and non-invasive devices, processes the signals, and interprets them using computers. In BCI research, two commonly used brain neural signals are electroencephalography (EEG) and near-infrared spectroscopy (NIRS) signals. EEG is measured the electrical activity in the brain, while NIRS is measured changes in the concentrations of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) in the cerebral region.

Reference [1] showed that a mental arithmetic task can be interpreted using BCI. During mental calculations of simple math problems, the subjects' brain activity was measured. Based on the brain signals, it was determined whether or not the participants had successfully performed the task. Reference [1] reported classification accuracy for NIRS data using a machine learning model called shrinkage linear discriminant analysis (sLDA). Recent research has been actively conducted not just on machine learning models, but also on deep learning models. Because deep learning models have the advantage of being able to automatically learn features from data and effectively capture complex patterns or abstract features with their hierarchical structures. In particular, researchers have focused on developing deep learning models to enhance classification performance.

However, the purpose of this study is to find optimal input data for effective learning rather than developing complex models. To be specific, we identify the optimal input data that enables effective learning of the signal's characteristics in a deep learning model, Shallow ConvNet, using NIRS signals.

We also test the model with two different input data used in NIRS signals to investigate its applicability on NIRS data.

II. Method

A. Used Deep Learning Model

Shallow ConvNet is a deep learning model inspired by the filter bank common spatial patterns (FBCSP) pipeline [2]. It contains temporal and spatial convolutional layers, each of which learns information about time and channels, respectively (see Fig. 1). Therefore, for the above reasons, this model has been widely used for EEG data analysis.

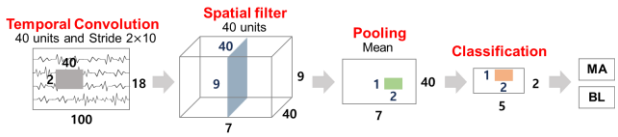


Fig. 1 Shallow ConvNet architecture with hyperparameters adjusted through a cross-validation.

We adjusted the hyper-parameters specifically for NIRS since the data size is small using the inner loop of nested cross-validation. Additionally, we modified the kernel size and stride in the pooling layer to address the issue of the kernel size being larger than the data size. Several other hyper-parameters were also fine-tuned, such as dropout rate, learning rate, and batch size.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I

HYPER-PARAMETER SETTINGS

Hyper-parameter	Hyper-parameter setting
Kernel size in temporal	40
Stride in temporal	10
Kernel size in pooling	2
Stride in pooling	1
Kernel size in final	2
Drop out	0.5
Learning rate	0.1
Batch size	2

Temporal: Temporal layer. Pooling: Pooling layer. Final: Final layer.

B. *Experimental procedure*

An open dataset for an EEG and NIRS hybrid BCI was used in this study [1]. We only used NIRS data of twenty-nine subjects which were measured while the subjects performed mental arithmetic (MA) and resting-state (RS) tasks. The experiment consisted of three sessions, and each of which consisted of 10 trials for the MA and RS tasks, respectively; a single trial consisted of an instruction period of 2 s, a task period of 10 s, and a rest period of between 15 to 17 s (see Fig. 2).

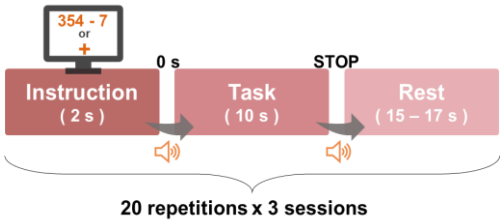


Fig. 2 Schematic sequence diagram of the experimental paradigm.

C. *Data collection and processing*

The NIRS signal was collected from thirty-six channels and and one channel consists of a pair of source-detectors. However, we used nine NIRS channels in the prefrontal cortex, which are well known to be associated with cognitive tasks (see Fig. 3).

The data was collected at 12.5 Hz sampling rate. The collected data were downsampled at 10 Hz, meaning that 10

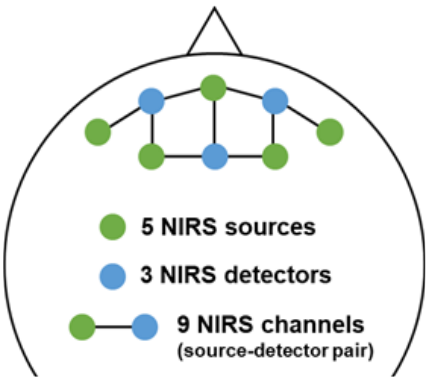


Fig. 3 Placement of NIRS sources and detectors. Gray solid lines are NIRS channels.

data samples exist per second. Concentration changes of the HbO and HbR were estimated using the modified Beer-Lambert law [3]. High-frequency instrument noise, as well as low-frequency components such as heartbeats and respiration, were removed from the HbO and HbR signals using bandpass filtering with cutoff frequencies of 0.01 and 0.1 Hz. The filtered data were epoched from -5 to 10 s based on the onset time, and baseline correction was performed using average values of -5 to -2 s for each epoch.

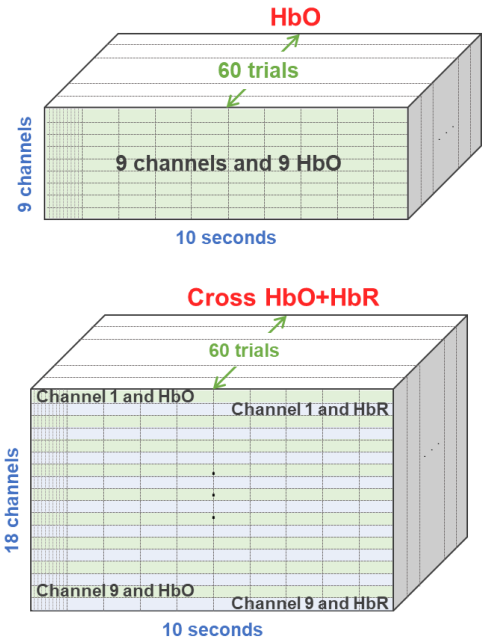


Fig. 4 Input data shape of Shallow ConvNet

After performing baseline correction, we generated two types of input data: HbO data and Cross HbO+HbR data (see Fig. 4). The latter involves concatenating the HbO and HbR

data for each channel, column by column. Each type of input data was used as input for the Shallow ConvNet.

In order to find hyper-parameters for NIRS data, cross-validation was performed using the input data. Nested cross-validation involves an outer loop and an inner loop. In the outer loop, a 5-fold cross-validation was performed to evaluate the performance of the model. In the inner loop, a 4-fold cross-validation method was used to determine the optimal hyper-parameters.

III. Results

We examine the performance of the Shallow ConvNet model to determine which data type, HbO or Cross HbO+HbR, better captures the information present in the NIRS data.We present the classification accuracy for each subject when HbO and Cross HbO+HbR are used. When using HbO, and Cross HbO+HbR data, the classification accuracies were 74.94 ± 10.43% and 76.26 ± 11.68%, respectively (See TABLE II).

TABLE II

Classification Accuracy per Subject in The Shallow ConvNet

Input data	HbO					Cross HbO+HbR				
Classification accuracy (accuracy ± standard deviation %)	Sub. 1	Sub. 2	Sub. 3	Sub. 4	Sub. 5	Sub. 1	Sub. 2	Sub. 3	Sub. 4	Sub. 5
	86.67 ± 8.50	78.33 ± 8.50	78.33 ± 12.47	68.33 ± 13.33	68.33 ± 13.33	85.00 ± 13.33	86.67 ± 8.50	86.67 ± 8.50	88.33 ± 8.50	68.33 ± 11.06
	Sub. 6	Sub. 7	Sub. 8	Sub. 9	Sub. 10	Sub. 6	Sub. 7	Sub. 8	Sub. 9	Sub. 10
	71.67 ± 17.16	71.67 ± 4.08	81.67 ± 9.72	90.00 ± 8.16	90.00 ± 6.24	68.33 ± 13.33	75.00 ± 9.13	90.00 ± 9.72	86.67 ± 11.30	93.33 ± 6.24
	Sub. 11	Sub. 12	Sub. 13	Sub. 14	Sub. 15	Sub. 11	Sub. 12	Sub. 13	Sub. 14	Sub. 15
	66.67 ± 14.34	66.67 ± 4.08	51.67 ± 12.25	60.00 ± 5.27	80.00 ± 4.08	70.00 ± 11.30	70.00 ± 13.54	55.00 ± 10.00	68.33 ± 15.28	85.00 ± 3.33
	Sub. 16	Sub. 17	Sub. 18	Sub. 19	Sub. 20	Sub. 16	Sub. 17	Sub. 18	Sub. 19	Sub. 20
	55.00 ± 20.82	81.67 ± 13.33	91.67 ± 5.27	75.00 ± 5.27	63.33 ± 11.30	56.67 ± 11.06	83.33 ± 9.13	91.67 ± 9.13	71.67 ± 11.30	50.00 ± 11.79
	Sub. 21	Sub. 22	Sub. 23	Sub. 24	Sub. 25	Sub. 21	Sub. 22	Sub. 23	Sub. 24	Sub. 25
	81.67 ± 14.34	61.67 ± 4.08	68.33 ± 12.25	83.33 ± 5.27	80.00 ± 4.08	85.00 ± 9.72	60.00 ± 6.24	61.67 ± 15.46	75.00 ± 17.48	76.67 ± 6.24
	Sub. 26	Sub. 27	Sub. 28	Sub. 29	Avg.	Sub. 26	Sub. 27	Sub. 28	Sub. 29	Avg.
	75.00 ± 9.13	85.00 ± 11.06	86.67 ± 8.50	75.00 ± 9.13	74.94 ± 10.43	71.67 ± 15.46	85.00 ± 6.24	83.33 ± 10.54	83.33 ± 5.27	76.26 ± 11.68

Sub: Subject, Avg: Average.

The classification accuracy was relatively higher when using both HbO and HbR as input data compared to using them independently. Furthermore, we confirmed that the fNIRS-PreT, a newly proposed deep learning model in previous studies, achieved a maximum classification accuracy of 78.53% using the same dataset as our study (see Figure 5).

Therefore, when comparing it with the results of previous research, we confirmed that optimizing the input data of existing deep learning models alone can achieve similar classification performance to developing complex algorithms.

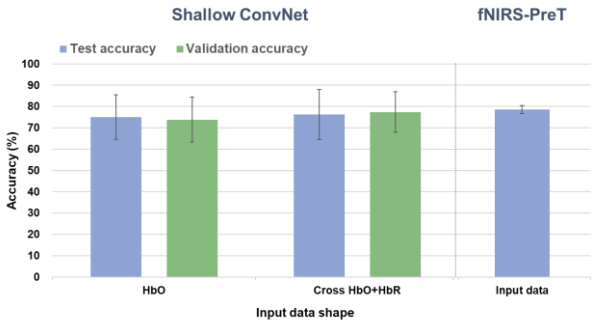


Fig. 5 The average classification accuracies in Shallow ConvNet,and fNIRS-PreT

IV. Conclusions

EEG signals contain temporal and spatial information, while NIRS signals provide temporal, spatial, and hemodynamic response information. Many researchers have conducted numerous studies, including proposing new models, to improve the classification performance of deep learning models using these signals, and this research has continued until recently. Among them, Shallow ConvNet is a widely used model for classification using EEG data, as it is considered effective in extracting temporal and spatial features. However, if NIRS signals are used as input data for this model, it is possible that the effective learning of hemodynamic response information in the signals cannot be guaranteed.

Therefore, it is necessary to develop a new model that can effectively capture the characteristics of these signals or find an input data shape that enables existing models, such as a shallow ConvNet, to extract the features of NIRS signal, in order to enhance the classification performance of the deep learning model using NIRS signals. However, we have confirmed in previous studies that the structure of the newly developed model is complex. Therefore, in this study, our aim was to find the optimal input data shape that enables the existing model to effectively learn the hemodynamic information contained in NIRS signals, without the necessity of developing complex models. We used NIRS data collected during MA and RS tasks, which underwent pre-processing to generate two different shapes of input data. These data shapes were subsequently fed into the Shallow ConvNet, and their performances were compared to determine the optimal input data shape.

As a result, we confirmed that optimizing the input data of the deep learning model, we can achieve classification performance comparable to developing complex algorithms.

In future works, we aim to enhance the classification performance by utilizing a more diverse combination of HbO and HbR, as well as optimizing the deep learning model.

ACKNOWLEDGMENT

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2021M3E5D2A01019547).

REFERENCES

[1] J. Shin, A. V. Lühmann, B. Blankertz, D. W. Kim, J. Jeong, H. J. Hwang, and K. R. Müller, Open Access Dataset for EEG+NIRS Single-Trial Classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering 25 (2017), 1735-1744.

[2] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, and F. Hutter, Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization, Human Brain Mapping 38 (2017), 5391-5420.

[3] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, Transformer Model for Functional Near-Infrared Spectroscopy Classification, IEEE Journal of Biomedical and Health Informatics 26 (2022), 2559-2569.

[4] Z. Wang, J. Zhang, Y. Xia, P. Chen, B. Wang, A General and Scalable Vision Framework for Functional Near-Infrared Spectroscopy Classification, Human Brain Mapping 30 (2022), 1982-1991.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Detection and Correction of Defective Relative Humidity Data Collected from the Greenhouse Environment Using Nested Kalman Filters with Standard Deviation Analysis

Kraithep Sirisanwannakul*, Nutchanon Siripool*, Kenji Suzuki†, Waree Kongprawechnon*, and Jessada Karnjana‡
* Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
E-mail: kraithep.sr@gmail.com Tel: +66-94-139-4420
† Tokyo Institute of Technology, Ookayama Campus
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan
E-mail: suzuki.k.di@m.titech.ac.jp
‡ NECTEC, National Science and Technology Development Agency, Pathum Thani, Thailand
E-mail: jessada.karnjana@nectec.or.th

***Abstract*—This study proposes a method for detecting and correcting defective data read from drifted humidity sensors used in the greenhouse, leveraging a nested Kalman filter approach with two components: an inner Kalman filter and an outer Kalman filter. The inner Kalman filter is built using data from neighboring sensors. It serves as the prediction for the outer Kalman filter, where its measurement is the data from the sensor of interest (in a defective data case) or from interpolation of neighbor sensors (in a severely defective data case). Analysis based on the standard deviation value of sensor data is used to classify the data into three cases: normal, defect, and severe defect. In the first case, the sensor data are considered trustable. In the last two cases, the data are corrected accordingly using the nested Kalman filters. Experimental results demonstrated significant improvement in accuracy. The mean absolute deviation (MAD) and the root-mean-square error (RMSE) between the corrected data and ground truth reduced when applying the proposed method. The experiments were conducted in two scenarios. The first scenario involves a sensor of interest that exhibits all three data cases: normal, defective, and severely defective data. The second scenario focuses on a sensor that mainly experiences two cases: normal and defective. The MAD decreased by approximately 75.71% and 50.93% in the former and the latter, respectively. Likewise, the RMSE dropped by about 66.41% and 41.80%, respectively. Therefore, the proposed method provides an effective and reliable technique for defective data detection and correction in humidity sensors where accurate and stable humidity measurement is critical.**

I. Introduction

Intelligent agriculture has emerged as a vital factor in improving the quality and quantity of farm products [1]. The main objective of intelligent farming systems is to supervise and manage the environmental conditions essential for the well-being and growth of plants and animals [2][3]. Temperature, humidity, and light intensity are among the critical environmental factors that play a crucial role in various applications, including agriculture, horticulture, and environmental monitoring [4]. In this study, the primary focus is on maintaining optimal conditions in orchid greenhouses, where the quality of the orchids heavily depends on precise humidity control. Humidity is a crucial parameter that significantly influences the health and growth of orchids, and any deviation from the required levels could adversely affect their well-being. Despite its importance, monitoring humidity poses a significant challenge for automated systems. Humidity sensors, particularly when exposed to high humidity environments like orchid greenhouses for prolonged periods, often degrade over time [5]. This degradation can lead to incorrect values over time, adding another layer of complexity to the precise maintenance of the desired humidity levels. Consequently, data from humidity sensors can become defective, leading to inefficient control and decision-making in intelligent farming systems. In order to address the challenges associated with defective data, farmers must regularly inspect the status of sensors installed in greenhouses and the quality of the collected data. Intelligent farming systems can support farmers in identifying defective data, potentially caused by drifting sensors or other factors, and automatically correcting the data to enhance the overall performance and accuracy of the system. Therefore, the scope of this work is focused on developing methods to detect and correct defective humidity data in high-stakes environments, such as orchid greenhouses, where the maintenance of humidity sensors is of paramount importance. Consequently, data from humidity sensors can become defective, leading to inefficient control and decision-making in intelligent farming systems. In order to address the challenges associated with defective data, farmers must regularly inspect the status of sensors installed in greenhouses and the quality of the collected data. Intelligent farming systems can support farmers in identifying defective data, potentially caused by drifting sensors or other factors, and automatically correcting the data to enhance the overall performance and accuracy of the system.

Various techniques for detecting and correcting sensor drift

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

have been proposed [6][7][8]. For instance, Dheeraj *et al.* proposed a spatial kriging and Kalman filtering-based automatic drift detection and correction approach [6]. Their method proved resilient to drift and surpassed traditional average interpolation techniques [6]. However, kriging interpolation faces accuracy limitations when dealing with a few data points. Zhan *et al.* adopted a clustering method and a Kalman filter to address the same challenge [7]. Their strategy did not rely on heavy assumptions or prior data model knowledge, and the results demonstrated improved data reliability and robustness in noisy situations [7]. However, the method's singular value decomposition (SVD) led to considerable computational time complexity. Therefore, it is not suitable for real-time applications. Zenisek *et al.* used a machine learning approach for drift detection to forecast maintenance needs for industrial equipment [8]. They applied three machine learning algorithms, i.e., linear regression, random forest regression, and symbolic regression, to assess the machine's state and generate alerts if data began to drift into unknown territory [8]. However, while such machine learning techniques can offer valuable solutions, they also present certain challenges. Primarily, the performance of these algorithms heavily relies on the quality of the training dataset. Furthermore, ensuring that the employed dataset genuinely represents the complete range of potential data variations is another challenge. If the dataset lacks representativeness, the effectiveness and reliability of the machine learning model could be compromised.

In our preliminary research, we focused on enhancing the accuracy of humidity sensors using nested Kalman filters. This approach significantly improved the precision of the sensor readings [9]. However, that approach lacked a detection mechanism. It showcased the potential of nested Kalman filters for addressing drift in humidity sensors, indicating that further improvement could be achieved by incorporating a detection component. In another previous work, we proposed a method that utilized a Kalman filter with an artificial neural network to detect sensor drift and correct humidity data obtained from drifting sensors [10]. The method demonstrated a significant reduction in accumulated mean absolute deviation. However, the detection process required daily data samples to classify the sensor's status, and the neural network relied solely on time as input. This led to the Kalman filter use a single prediction value for each sample every day, resulting in an unreliable outcome due to its inability to fully capture the complexities of sensor drift. Additionally, a parameter relating to the variance of the process noise of the Kalman filter was fixed daily, which is an unrealistic assumption concerning the dynamic nature of sensor drift. Therefore, in this study, we propose a new method that focuses on analyzing and detecting data status every 4 hours, providing a more adaptive approach to correcting defective and severely defective data in drifting humidity sensors. This method aims to address the limitations of previous approaches by leveraging the strengths of the nested Kalman filter. Our current work incorporates two main components for correcting defective and severely defective data: the inner Kalman filter, which is built using data from neighboring sensors and serves as the prediction state for the outer Kalman filter, and the outer Kalman filter, which uses the inner Kalman filter's prediction state and measurements from the sensor of interest or bicubic interpolation of neighbor sensors to correct the inaccurate humidity value. By focusing on fixing defective data and refining the nested Kalman filter approach, our method aims to accurately identify and rectify the distorted humidity readings, restoring the data to its appropriate values and maintaining the sensor's accuracy and reliability.

The rest of this paper is structured as follows: Section 2 offers a concise overview of the essential background information and concepts relevant to our proposed framework and an outline of the monitoring system and the one-dimensional Kalman filtering algorithm employed in our experiments. In Section 3, we delve into the specifics of the proposed method. Section 4 presents the experimental results and performance evaluation of the proposed method. Section 5 discusses our findings and interprets the outcomes, highlighting the strengths and potential limitations of the method and its implications for future research and practical applications. Finally, Section 6 summarizes the work presented in this paper.

II. Background

This section provides an overview of our monitoring system implemented to gather the data utilized in this study. Additionally, a brief introduction to the one-dimensional Kalman filter is also included in this section.

A. Overview of a monitoring systems

Our monitoring system is implemented in greenhouses at NECTEC, NSTDA, Thailand, spanning an area of 6×20 square meters. This system is comprised of 21 humidity sensors with two data loggers (logger (a) and logger (b)). Each humidity sensor is responsible for monitoring a 2×2.5 square meter area, as illustrated in Fig. 1. Data from the humidity sensors is transmitted to a data logger, which reformats and forwards the information to the gateway. The data from all the loggers is periodically transmitted to a database via a GPRS network using the gateway. This allows users to access the data through a web browser. Digital SHT31 humidity sensors from Sensirion are used in this greenhouse, which operate by measuring capacitance. Over time, the collected humidity data may become defective due to factors such as sensor aging, environmental changes, or manufacturing defects. Inaccurate measurements lead to the generation of defective data and can adversely affect the efficiency and effectiveness of greenhouse operations.

B. One-dimensional Kalman filtering

The Kalman filtering algorithm encompasses two primary stages: prediction and measurement update [11]. The fundamental assumption of this algorithm is that a linear system can

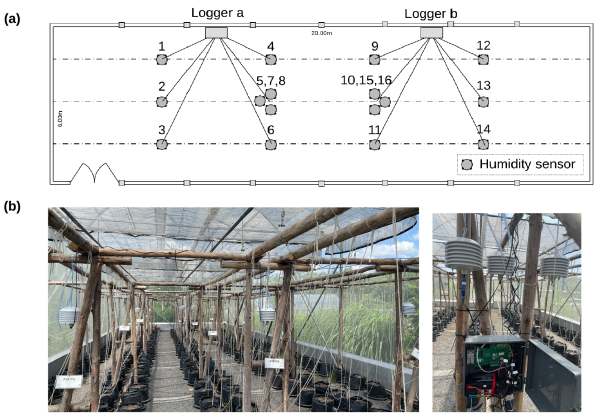


Fig. 1. Sensor location inside the greenhouse in NECTEC, NSTDA, Thailand: (a) diagram of the locations in association with the dataloggers and (b) image of the inside environment.

be delineated using one state equation and one measurement equation. These equations are as follows:

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}, \tag{1}$$

and

$$y_k = Cx_k + v_k, \tag{2}$$

where x_k denotes the state at time k, u_{k-1} is an input at time $k-1$, w_{k-1} is process noise with a mean of zero and a variance of Q at time $k-1$, y_k represents a measured output at time k, v_k is measurement white noise with a variance of R at time k, and A, B, and C are factors associated with the state transition, control input, and the correlation between the measurement and the state, respectively.

The prediction stage involves two steps. Initially, the algorithm utilizes the previous state estimate $\hat{x}_{k-1}$ to predict the *a priori* state estimate $\hat{x}_k^-$ through the equation:

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1}. \tag{3}$$

Note that for $k=1$, $\hat{x}_0$ is an the initial state.

Subsequently, the *a priori* error variance P_k^- is calculated using:

$$P_k^- = AP_{k-1}A + Q, \tag{4}$$

Here, P_{k-1} represents the *a posteriori* error variance, which is defined as the expectation of a posteriori estimate error, $x_k - \hat{x}_k$, times its transpose. For $k=1$, P_0 is an initial error variance.

The measurement update stage is composed of three steps. Initially, the algorithm computes the Kalman gain K:

$$K = \frac{P_k^- C}{CP_k^- C + R}. \tag{5}$$

Then, it estimates the predicted state estimate $\hat{x}_k$:

$$\hat{x}_k = \hat{x}_k^- + K(y_k - C\hat{x}_k^-). \tag{6}$$

Finally, the state error variance is updated:

$$P_k = (1 - KC)P_k^-. \tag{7}$$

III. Proposed Method

The proposed method, as illustrated in Fig. 2, is divided into two main stages: detection and correction. In the detection stage, data from the sensor of interest are compared with the standard deviation (SD) values of humidity readings from neighboring sensors over a four-hour period. This allows us to identify if the data significantly diverges from its neighbors. The thresholds of mean - $2\cdot$SD and mean + $2\cdot$SD are selected based on our observation and the fact that approximately 95% of the data in a normally distributed dataset falls within these boundaries. Here, the mean refers to the average of the neighboring humidity sensor. If the data point is outside these boundaries, the data is considered defective; if it's within, the data is deemed normal. If the data point lies beyond mean - $4\cdot$SD or mean + $4\cdot$SD, it is classified as severely defective. The $4\cdot$SD thresholds are chosen based on our observation and the understanding that the probability of observing a data point beyond these boundaries in a normally distributed dataset is less than 0.00006, indicating a highly unlikely deviation. The primary task of the first stage is to categorize the data into three groups: normal, defective, and severely defective.

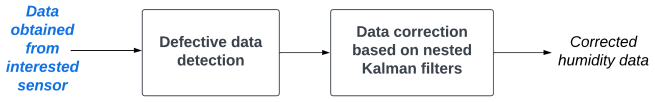


Fig. 2. Framework of proposed method.

The second stage, correction, depends on the classification of the data obtained in the first stage. If the data is normal, it remains unchanged and is used as is. If the data is defective, a nested Kalman filter is employed, structured as shown in Fig. 3(a). For severely defective data, the nested Kalman filter is also used with the structure depicted in Fig. 3(b). The difference between Fig. 3(a) and Fig. 3(b) is the use of bicubic interpolation in the latter. Bicubic interpolation estimates missing data using values and derivatives of humidity values from neighboring sensors. Performed on a 3×3 grid of data points, it considers the 8 nearest neighbor sensors, providing a smooth and continuous estimation of severely defective humidity data. Furthermore, this approach ensures the accurate estimation of severely defective values even in challenging scenarios, as it takes advantage of both the prediction and measurement states. In defective data and severely defective data cases, nested Kalman filters are employed to correct the humidity values. The outer Kalman filter uses the following state and measurement equations:

$$x_k = x_{k-1} + w_{k-1}, \tag{8}$$

and

$$y_k = x_k + v_k, \tag{9}$$

where x_k is the estimated humidity value at the location of interest at time k, w_{k-1} is the process noise at time $k-1$, y_k is the data read from the sensor at the location of interest at time

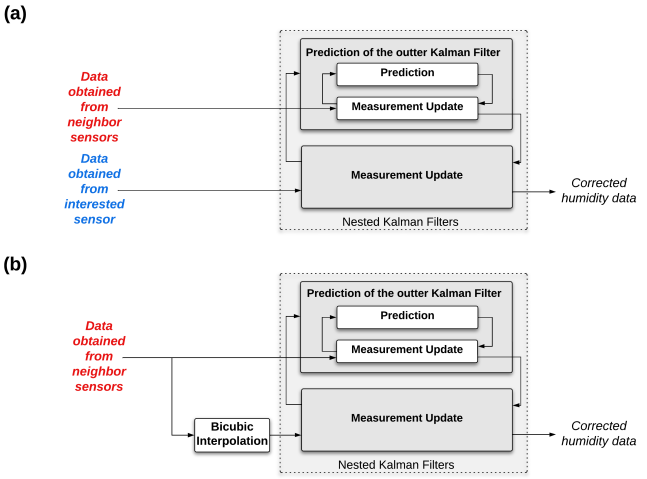


Fig. 3. Block diagrams of proposed method: (a) defective case and (b) severely defective case.

k in the *defective data* case, and y_k is the data obtained from the bicubic interpolation at time k in the *severely defective data* case. Lastly, v_k represents the measurement noise at time k.

The inner Kalman filter serves as the predictive model for the outer Kalman filter. Let N be a set of indices representing the neighbor sensors to understand this relationship. The proposed method considers a sequence y_l^p for $l = 1$ to $n(N)$, where $n(N)$ is the cardinality of N. Each y_i in the sequence corresponds to the data read from the neighbor sensor with index i for $i \in N$. Furthermore, the initial state estimates x_0^p are assumed to be the average value of the humidity values of y_i for all is. With these assumptions, the inner Kalman filter employs the following state and measurement equations:

$$x_l^p = x_{l-1}^p + w_{l-1}^p, \tag{10}$$

and

$$y_l^p = x_l^p + v_l^p, \tag{11}$$

where x_l^p is the estimated humidity value at the neighboring sensor at time l, w_{l-1}^p is the process noise at time $l-1$, y_l^p is the data read from the neighboring sensor at time l, and v_l^p is the measurement noise at time l.

The Q values for the inner and outer Kalman filters are set to 0.01 and 1, respectively, obtained through trial and error. These values represent the process noise covariance for each filter, reflecting the uncertainty in the model and the correction process.

IV. Experiment and Results

This section presents the experimental setup and results of the proposed method for detecting and correcting defective relative humidity data in greenhouse environments. First, the experimental setup is described, including the data collection process, sensor arrangement, and parameters used in the study. Next, the experimental results are discussed, highlighting the method's performance in detecting and correcting defective data. Finally, the analysis of the results demonstrates the effectiveness of the proposed method.

A. Experimental Setup

In this study, humidity data are collected from sensors placed in a greenhouse and transmitted to the database every 5 minutes, resulting in 288 data points per day. The experiment is divided into two scenarios to demonstrate the effectiveness of the proposed method in handling varying sensor conditions. Each scenario utilizes humidity data gathered from distinct sensor locations within the greenhouse, as illustrated in Fig. 4.

The first scenario involves a sensor of interest that exhibits all three data cases: normal, defective, and severely defective data. The second scenario focuses on a sensor that mainly experiences two cases: normal and defective. This distinction between scenarios allows for a comprehensive evaluation of the proposed method's ability to detect and correct the defective data issues read from drifted humidity sensor, improving the accuracy and reliability of the sensor readings across different situations.

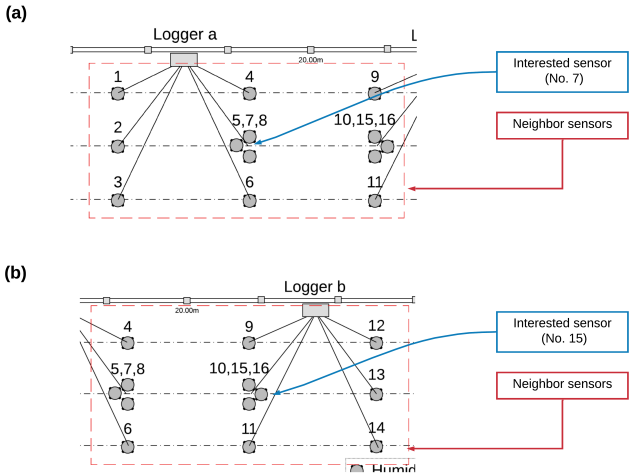


Fig. 4. Positions of sensors used in this experiment: (a) first scenario and (b) second scenario.

Since N is the set of indices of the neighbor sensors in the correction part, in this experiment,
$N = \{1, 2, 3, 4, 6, 9, 10, 11\}$, for the first scenario, and
$N = \{4, 5, 6, 9, 11, 12, 13, 14\}$, for the second scenario.

The humidity data in the NECTEC greenhouse were used. The duration for data analysis is about 61 days, i.e., from 1 June 2021 to 31 August 2021. This experiment aims to apply the proposed method to estimate the actual humidity value read by sensor No. 7 for the first scenario or sensor No. 15 for the second scenario, which is the location of interest. In other words, sensors No. 7 and No. 15 are somehow defective or drifted.

To evaluate the proposed framework, we use accumulated mean absolute deviation (MAD) and the root mean squared error (RMSE) to evaluate this experiment. In the first scenario,

sensor No. 5 serves as the reference for evaluating MAD and RMSE, as it provides readings close to sensor No. 8, which is well-maintained and installed to read values only during specific periods of certain days (to minimize sensor drift). Thus, sensor No. 5 is the reference sensor for this scenario. Similarly, in the second scenario, sensor No. 10 serves as the reference for MAD and RMSE, as it is positioned closely and provides readings close to sensor No. 16, which is also installed and maintained in the same manner as sensor No. 8. Therefore, in both scenarios, the reference sensors (No. 5 for the first scenario and No. 10 for the second scenario) are used to compare the outputs of the proposed method and calculate MAD and RMSE values.

The following equation defines the MAD:

$$\text{MAD} = \frac{1}{n}\sum_{k=1}^{n} |x_k - z_k| \,. \tag{12}$$

The following equation defines the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - z_k)^2}, \tag{13}$$

where n is the number of data points, and z_k is humidity value from reference sensor No. 5 for the first scenario, or sensor No. 10 for the second scenario at time k.

B. Experimental Result

Upon applying the proposed method for detecting and correcting defective and severely defective data read from drifted humidity sensor, a substantial reduction in accumulated mean absolute deviation with reference to the sensor was observed. For the first scenario, the decrease was approximately 75.71%, while for the second scenario, it was approximately 50.93%. Similarly, the RMSE for the first and second scenarios dropped by approximately 66.41% and 41.80%, respectively, as shown in Table I. In comparison, the results from the previously proposed method, which utilized two nested Kalman filters without the detection part [9] (mentioned as *previous* in Table I), showed lower reductions in RMSE and MAD. This significant improvement in the accuracy of humidity measurements in both scenarios highlights the effectiveness of the proposed method in detecting and correcting issues related to defective and severely defective data in humidity measurements.

TABLE I
COMPARISON OF RMSE AND MAD REDUCTION BETWEEN THE PREVIOUS AND PROPOSED METHODS FOR BOTH SCENARIOS.

Sc.	Method	RMSE reduction	MAD reduction
1	Previous [9]	42.87%	43.18%
	Proposed	66.41%	75.71%
2	Previous [9]	30.24%	31.78%
	Proposed	41.80%	50.93%

The simulations were conducted on an ASUS ROG STRIX G17 laptop equipped with an AMD Ryzen 7 6800H processor, 16 GB DDR4 RAM, and GEFORCE RTX3050. The computational time for simulating the proposed method for a single day is approximately 5-7 seconds.

Preliminary experimentation with adjusting thresholds for detecting defective data has revealed a pattern when broadening or narrowing the thresholds. It was found that excessively widening the detection range leads to an inability to distinguish severely defective data in certain instances. This implies that the detection process predominantly identifies only moderately defective data, consequently impacting the overall performance of the subsequent defective data correction steps. On the other hand, setting the detection threshold too narrowly results in the identification of mainly severely defective data. This implies that each instance of correcting defective data necessitates extensive interpolation, increasing computational time, i.e., an increase of up to around 30% was observed when setting the threshold too narrowly to the point where all data were detected as severely defective.

V. DISCUSSION

In both scenarios, the proposed method demonstrated significant improvements in the accuracy of humidity sensor readings, effectively detecting and correcting defective and severely defective data issues. Fig. 5 shows the example results in one day. The enhanced reliability and consistency of the sensors' measurements are crucial for various applications in agriculture, meteorology, and environmental monitoring. Notably, the proposed method showed a considerable improvement compared to the past work when the sensor was in a very bad state, having both defective and severely defective data. Furthermore, the method's superior performance in the more challenging first scenario, where the sensor experienced all three statuses (normal, defective, and severely defective), indicates its effectiveness in handling complex and varying sensor conditions.

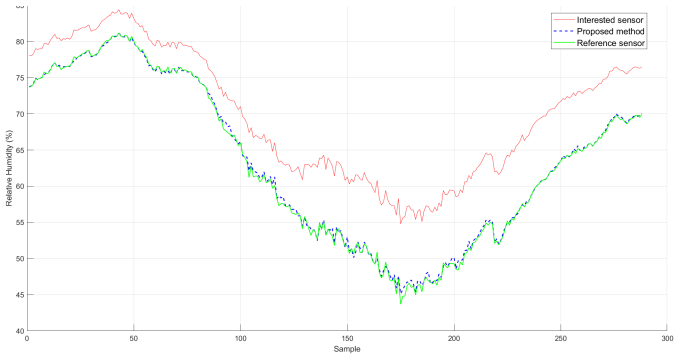


Fig. 5. Applying the proposed method to correct the defective data in the interested sensor.

The second scenario also demonstrated the method's adaptability and robustness in dealing with less severe drift issues, where the sensor exhibited mainly normal and defective statuses. In this scenario, the interested sensor was not very bad, leading to a lower percentage improvement. Nevertheless, this versatility suggests the potential applicability of the proposed

method to other types of sensors and situations where similar data issues may arise, making it a valuable tool in sensor data correction with possible applications beyond humidity sensors in various domains.

The chosen positions for the interested sensor placement influence the performance of the proposed method. A higher density of neighboring or surrounding sensors with normal reading capabilities enhances the performance of the proposed method. Acknowledging that accurate corrections cannot be achieved if most neighboring sensors cannot provide normal readings is crucial.

In the current study, the inner and outer Kalman filter's Q values were fixed based on trial and error, similar to our previous work. Although this approach yielded significant improvements in the accuracy of humidity sensor readings, there is still room for further investigation in optimizing the Q values. Adapting the Q values dynamically or exploring alternative optimization techniques could enhance the performance of the proposed method, leading to even better results in detecting and correcting defective and severely defective data. Exploring these possibilities in future research could contribute to developing more robust and efficient solutions for fixing drift and malfunction issues in humidity sensors, ultimately benefiting various agriculture, meteorology, and environmental monitoring applications.

VI. Conclusions

In conclusion, this research has presented and evaluated a novel method for detecting and correcting defective and severely defective data in humidity measurements using nested Kalman filters and bicubic interpolation. The proposed method combines standard deviation-based detection nested Kalman filters. The experimental results have demonstrated significant improvements in the accuracy of humidity measurements across two different scenarios, with varying degrees of defective and severely defective data.

Acknowledgment

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) under the TAIST Tokyo Tech Program. The authors would also like to express their gratitude to the National Research Council of Thailand (NRCT) for providing scholarships to support this study.

References

[1] McBratney, A., Whelan, B., Ancev, T., & Bouma, J. (2005). Future directions of precision agriculture. Precision agriculture, 6(1), 7-23.

[2] Du, X., Wang, J., Ji, P., & Gan, K. (2011, July). Design and implement of wireless measure and control system for greenhouse. In Proceedings of the 30th Chinese Control Conference (pp. 4572-4575). IEEE.

[3] Xing, X., Song, J., Lin, L., Tian, M., & Lei, Z. (2017, July). Development of intelligent information monitoring system in greenhouse based on wireless sensor network. In 2017 4th International Conference on Information Science and Control Engineering (ICISCE) (pp. 970-974). IEEE.

[4] Stuckey, I. H. (1967). Environmental factors and the growth of native orchids. American Journal of Botany, 54(2), 232-241.

[5] Griesel, S., Theel, M., Niemand, H., & Lanzinger, E. (2012). Acceptance test procedure for capacitive humidity sensors in saturated conditions. WMO CIMO TECO-2012, Brussels, Belgium, 1-7.

[6] Kumar, D., Rajasegarar, S., & Palaniswami, M. (2013, May). Automatic sensor drift detection and correction using spatial kriging and kalman filtering. In 2013 IEEE International Conference on Distributed Computing in Sensor Systems (pp. 183-190). IEEE.

[7] Li, Z., Wang, Y., Yang, A., & Yang, H. (2015, October). Drift detection and calibration of sensor networks. In 2015 International Conference on Wireless Communications & Signal Processing (WCSP) (pp. 1-6). IEEE.

[8] Zenisek, Jan, Florian Holzinger, and Michael Affenzeller. "Machine learning based concept drift detection for predictive maintenance." Computers & Industrial Engineering 137 (2019): 106031.

[9] Dangsakul, P., Siripool, N., Sirisanwannakul, K., Keinprasit, R., Rungprateeptavorn, K., Keerativittayanun, S., & Karnjana, J. (2020). Humidity sensor accuracy improvement based on two nested Kalman filters for commercial cultivation of tropical orchids. In Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27 (pp. 104-112). Springer International Publishing.

[10] Sirisanwannakul, Kraithep, et al. "Humidity Sensor Drift Detection and Correction Based on a Kalman Filter with an Artificial Neural Network for Commercial Cultivation of Tropical Orchids." International Conference on Computational Intelligence in Information System. Springer, Cham, 2021.

[11] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Pedestrian Crossing Intention Prediction with Multi-Modal Transformer-Based Model

Ting Wei Wang* and Shang-Hong Lai*†
*Institute of Information Systemas and Applications
†Dept. of Computer Science
National Tsing Hua University, Taiwan
E-mail: *gs109065509@gapp.nthu.edu.tw, †lai@cs.nthu.edu.tw

***Abstract*—The popularity of autonomous driving and advanced driver assistance systems can potentially reduce thousands of car accidents and casualties. In particular, pedestrian prediction and protection is an urgent development priority for such systems. Prediction of pedestrians' intentions of crossing the road or their actions can help such systems to assess the risk of pedestrians in front of vehicles in advance. In this paper, we propose a multi-modal pedestrian crossing intention prediction framework based on the transformer model to provide a better solution. Our method takes advantage of the excellent sequential modeling capability of the Transformer, enabling the model to perform stably in this task. We also propose to represent traffic environment information in a novel way, allowing such information can be efficiently exploited. Moreover, we include the lifted 3D human pose and 3D head orientation information estimated from pedestrian image into the model prediction, allowing the model to understand pedestrian posture better. Finally, our experimental results show the proposed system provides state-of-the-art accuracy on benchmarking datasets.**

I. Introduction

Automotive technologies are developing toward autonomous driving. Autonomous driving or ADAS (Advanced Driver Assistance Systems) can significantly improve safety by reducing accidents caused by human mistakes or careless driving behaviors. Pedestrian Crossing Intention Prediction is a crucial task in this type of system. It tries to predict the pedestrian's intention of crossing the road by analyzing pedestrian images, postures, behaviors, and information about the surroundings of pedestrians.

In recent years, research in this field has considerably increased. The introduction of JAAD [1] and PIE [2] datasets has enabled researchers to develop pedestrian trajectory, behavior, and crossing intention prediction tasks more effectively. A model using CNN and RNN architectures for pedestrian crossing intention and trajectory prediction was proposed in the paper of PIE dataset, which demonstrates the RNN model's ability to process sequence information in this task. Other methods [3] [4] [5] [6] [7] have started to introduce a series of information related to pedestrians, such as human pose and trajectory. Besides, surrounding environment data has also been introduced, such as ego-vehicles speed, surrounding environment images, and traffic status.

However, there are still more additional information that has not be exploited into this task. Both datasets provide quite rich road environment information that has not been fully utilized. In addition, with the introduction of the Transformer [8], it becomes a more suitable model architecture for predicting pedestrian crossing intention because of the advantage of being parallelizable for real-time computation and the ability to model the temporal information of sequential data.

A benchmark for evaluating pedestrian action prediction [9] has also been released. This evaluation standard solves the problem of inconsistent standards for comparing previous methods. However, we have observed that the testing environment in several previous works may not necessarily follow the benchmark settings.

In this paper, We focus on exploring different data and model structures to find a solution that could handle more complex scenarios and provide reliable performance while following the Benchmark [9] settings. The main contributions of this paper are summarized as follows:

- We propose a multi-modal method based on Transformer architecture and use nine different types of input data for predicting pedestrian crossing intention, and it achieves state-of-the-art performance.
- We are the first to combine Traffic light, Crosswalk, and Road Sign data into new Traffic Perception Data. This strategy has been demonstrated to improve the accuracy of our method through experiments.
- Our method uses lifted 3D human pose and 3D head orientation information to provide 3D pedestrian information for the model, which allows it be adapted to a wide variety of scenarios.

II. Related Work

Pedestrian crossing intention prediction is an important part of ADAS systems and self-driving cars, so there have been many related studies in the past. Initially, some methods directly analyze the possibility of pedestrians crossing through a single image [1] [10] [9]. However, because only a single image was used for analysis, there was insufficient contextual information to infer pedestrian behavior accurately.

A. Sequential Modeling

Some methods that use sequential data as input have also been proposed [2] [3]. These methods consider the changes

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

in various features of pedestrians over a period of time. The model can analyze pedestrians' motion which is related to its possible future location. RNN models were introduced to process these sequential data. They can handle the temporal information at different time steps in contextual data. Compared with those based on single-time information, this approach brings breakthroughs and accuracy improvements to this field.

B. Exploration of Novel Inputs

Since then, many representative trajectory-based methods have been developed [11] [12]. These methods determine the likelihood of a pedestrian crossing by analyzing the pedestrian's trajectory, thus enhancing the reliability of the model in predicting the crossing intention through explicit future trajectories of pedestrians. Previous works have also tried to utilize more information about pedestrians from different angles of pedestrian images. For example, in [13] [11], they attempted to determine pedestrians' trajectories, appearance, or social states from an eagle-view perspective, which can prevent pedestrians from being blocked by other objects on roads.

However, pedestrian trajectories, i.e., pedestrian motion, are extremely diverse and unpredictable. For example, when a pedestrian is stationary or has minimal movement, there is almost no "trajectory" for the pedestrian to predict their future movement. Furthermore, eagle view data is difficult to obtain in real-world scenarios.

Some works enhance the prediction accuracy through the visual aspects. First, [3] [2] [7] [9] used the image of the pedestrian itself and also considered the surrounding images of the pedestrian. These data show the interaction between pedestrians and surrounding objects and supplement information about the environment in which the target pedestrian is located. Some methods introduced semantic segmentation information [7] [4] so that the model can analyze the road appearance and traffic conditions where the target pedestrian is located. However, relying only on visual information is easily affected by the diversity of training data and the quality of images. In addition, semantic segmentation information may bring adverse results when computing in real-time scenarios because of its large amount of computation.

Therefore, some literature turned their attention to pedestrians themselves and introduced human pose information due to the posture and behavior of pedestrians are closely related to whether they will cross the road. For example, pedestrians often use gestures to communicate with drivers to express their desire to cross. However, we believe there is room for improvement in these methods, as most of the methods lack 3D information on pedestrian posture. This 3D information would provide information on pedestrian orientation, hand movements, and potential crossing behavior in such tasks.

Pedestrians also will swing their heads to look both ways to confirm whether cars are coming and interact with the vehicle driver with eye contact and nods. Thus, the pedestrian head pose is valuable for reference. Head orientation is less mentioned in this field than the human pose. In [6] [4], they segment the pedestrian's head or body orientation into eight discrete directions as input. However, they only focus on a single dimension of the horizontal rotation of the pedestrian. In [14], they introduced 3D head orientation including the dimensions pitch, roll, and yaw, and used the head rotation angle as a scale. But this method relies mainly on the head direction to determine pedestrian intention and does not consider other pedestrian features, which is still unreliable.

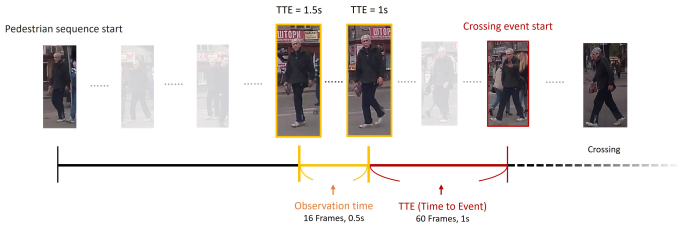


Fig. 1. Definition of TTE and observation time. The yellow part indicates the range where the model can observe and make predictions. This range must be prior to the time "TTE" before the pedestrian starts to cross the road.

In addition, the environment around pedestrians and traffic-related information are also very critical. In [2], they included ego-vehicle speed as an input to their prediction model and mentioned that vehicle speed critically influences whether pedestrians cross the road. Since then, vehicle speed information has become common in most literature. Another paper [5] used the presence of traffic lights or crosswalks and the distance between the pedestrian and vehicle to calculate a value as input information. Although the PIE and JAAD datasets provide a wealth of information about the traffic environment above, much of the previous methods did not consider this, and this information would be worth exploring.

III. Proposed Method

A. Problem Statement

According to the definition of the benchmark, pedestrian crossing intention prediction is defined as a binary classification problem. Various information related to pedestrians is used as model input data. The model will give a prediction result of "crossing" or "non-crossing," which corresponds to the "1" or "0" of the output.

A pedestrian crossing in front of the ego-vehicle is defined as crossing, while any pedestrian that does not cross the path of the ego-vehicle is defined as non-crossing, e.g., a pedestrian walking in the same direction as the ego-vehicle.

The model needs to complete the prediction in advance before the pedestrians start to cross within the time that model can observe the input sequence. This advance is called TTE (Time-To-Event), which refers to the interval between the last frame of model observation time and the start of the pedestrian crossing. The concept of TTE and observation time can refer to Figure 1.

B. Module Architecture

The overall architecture of the proposed pedestrian crossing intention prediction system is illustrated in Figure 2. The

proposed system consists of two parts – the Feature Pre-processing Module and the Prediction Module. More details on these two modules will be described in the following sections.

*1) **Feature Pre-processing Module**:* To predict the pedestrian crossing intention, we rely on several different types of data as a basis for model prediction. Each data has a different relationship with the target pedestrian and can be used to determine whether the pedestrian wants to cross the road under different circumstances. The Feature Pre-processing Module is responsible for converting or extracting the original input data so that various types of data can be used to help the model train in the most effective way.

Traffic Perception Data: In our method, the Traffic Perception Data can help the model understand the surrounding traffic conditions in which the pedestrian is located. It consists of three different types of information: The traffic light state for the lane where the ego-vehicle is located, the presence of specific signs, and the presence of crosswalks in the current frame. These three data types are represented in a unique way to provide road environment traffic information.

*a) **Traffic Light**:* Based on the observations of failure cases, we found that many samples can preliminarily infer the possibility of a pedestrian crossing through the status of traffic lights for the lane in which the ego vehicle is traveling. In this method, we directly use the traffic light status provided by the dataset and transform the representation into a 3-dimensional array as training data. Where [1,0,0] represents a red light, [0,1,0] represents a yellow light, and [0,0,1] represents a green light, [0,0,0] represents no traffic light in the current frame. The values of each dimension have two states - 1 or 0, representing whether the corresponding light is on or off.

*b) **Road Sign**:* Road signs provide information on the right-of-way and the order of passage among road users. Pedestrians or passing vehicles will refer to the road signs to take corresponding and appropriate actions. This information can also be found in the dataset, which provides detailed information on the type of road signs in the current frame. However, the training results using detailed road sign types as training data were unsatisfactory due to redundant and divergent information.

Instead, we select the signs directly related to the pedestrian crossing and use the existing state of these signs in the current frame as input data. For example, When a stop sign appears, vehicles on the road will tend to slow down or stop. Or, if a pedestrian sign is present, vehicles will tend to yield to pedestrians. Therefore, we represent the information of specific road signs as a 1-dimensional array. When the array is [1], it represents a stop sign or pedestrian sign in the current frame; when the array is [0], it is the opposite.

*c) **Crosswalk**:* : From the perspective of pedestrians, they are more willing to cross the road in places with crosswalks, which is also a factor to consider as needed. The 1-dimension array [1] represents when the crosswalk is present in the current frame and [0] when it is not.

Traffic Perception Data Fusion: After experimenting with the above three data types, we found that performing an early fusion on these data is more effective for training than feeding the data into the model separately because the model can learn the correlation between different data. Therefore, we fuse these three data by concatenating them into a 5-dimension array.

Pedestrian Behavior Data:

*d) **3D Head Orientation**:* From the discussion in the previous chapter, we can found that knowing whether the target pedestrian is looking at the driver or noticing the vehicle can help the model to analyze pedestrians' behavior. Since head orientation is not provided in the dataset, we need to apply a head orientation detector to obtain this information.

We choose "6DRepNet" [15] as our method to generate the 3D head orientation. It can provide the orientation in three dimensions, i.e., pitch, roll, yaw of the subject's head rotation. Furthermore, the detection angle of pitch and roll covers the range [-90°, +90°], and yaw covers the range of [-180° , +180°]; these ranges cover the full range of human head rotatable angles. The final data format is a 3-dimensional array, e.g., [-10, 23, 128], which represents the current head pitch, roll, and yaw angles, as if a person were tilting their head a bit to the side and lowering their head slightly.

*e) **3D Pedestrian Human Pose**:* The 2D human pose data generated by OpenPose [16] is provided in the dataset. However, as 3D pose information can provide richer information for the model, we use the method proposed by [17] to lift 2D pose data to 3D. The original 2D pose sequence must be converted to the Human3.6M representation the lifting model uses before processing.

Therefore, we interpolate the coordinates to calculate the missing keypoints, such as head, spine, and middle of the hip in Openpose, then convert the pose representation to the format of human3.6M. The lifted format will become 17 joints with three dimensions x, y, z. The final pose data will be represented by a 51-dimensional array like [x1, y1, z1 … x51, y51, z51].

*f) **Pedestrian Bounding Box**:* The pedestrian bounding box coordinates allow us to know the location of pedestrians and the direction of pedestrian movement. The bounding box in the dataset is a 4-dimensional array of [x1, y1, x2, y2], where x1 and y1 correspond to the coordinate points at the top left-hand corner of the bounding box, and x2 and y2 correspond to those at the bottom right-hand corner.

Driving Behavior Data:

*g) **Ego-Vehicle Speed**:* Pedestrians often judge whether it is appropriate to cross the road based on the vehicle's speed. For example, fast-approaching vehicles may cause pedestrians to decide against crossing. The ego-vehicle speed data are available in the datasets and represent as a 1-dimension array. The data in PIE dataset is the GPS speed value of the vehicle driving. In JAAD, it is represented by text labels of the vehicle's motion state, such as [0] represents the stationary state, [1] is moving slow, [2] is moving fast, [3] is decelerating, [4] is accelerating.

Pedestrian Appearance Data:

*h) **Pedestrian Bounding Box Image**:* Pedestrians' appearance information can provide many details and complement other input data, such as pedestrians' behavior and

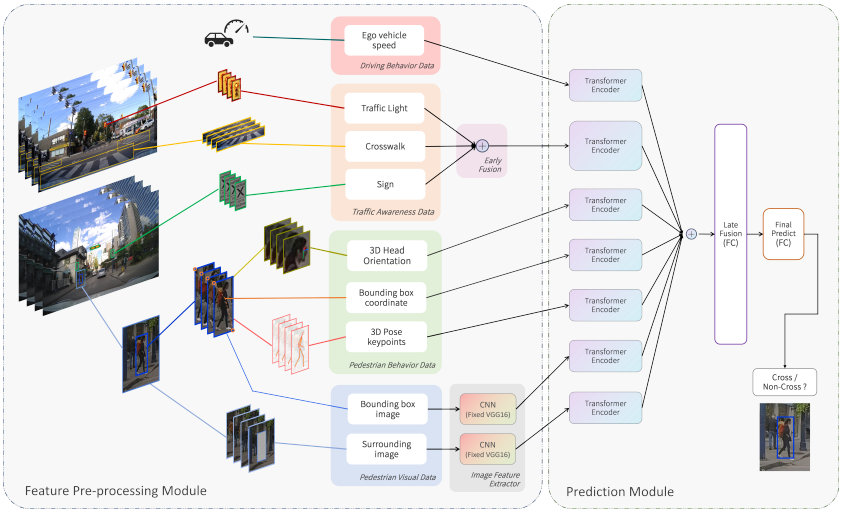


Fig. 2. The overall architecture of proposed pedestrian crossing intention prediction system.

orientation. The pedestrian bounding box image is generated by cropping the pedestrian bounding box coordinates provided by the dataset from the complete image where the pedestrian is located. The size of the image will change according to the pedestrians' location and distance in the image.

i) ***Pedestrian Surround Image****:* Besides the pedestrian appearance information, the surrounding environment around pedestrians is also vital information, which can provide a lot of additional information, such as the location where pedestrians stand, the interaction between pedestrians and surrounding objects, or whether specific objects are around them. The way to generate this information follows the setting of the benchmark dataset. It is cropped and obtained by expanding 1.5 times according to the size of the pedestrian bounding box and filling the original pedestrian bounding box area with gray pixels to present a rectangle with a hollow center. This makes the model pay more attention to visual information in the surrounding environment and avoid being disturbed by pedestrians' appearance.

The above two types of visual information will be pre-processed by a pre-trained VGG16 convolutional neural network for feature extraction same, as in the datasets. The image feature extractor does not participate in the training process. It is only used to generate and extract Image Features for subsequent model training and does not include the top layer part used as a classification. The pedestrian bounding box image and the surround image are fed to the feature extractor to generate a 512-dimensional image feature as training data.

2) ***Prediction Module****:* The Prediction Module is responsible for the part of performing pedestrian intention prediction.

a) Prediction Model Flow: Here, we introduce the model flow for the Prediction Module. The process diagram can refer to the Prediction Module part on the right side of Fig. 2.

After the Data Pre-processing Module generates all the data we need, it will first enter different branches of the Transformer Encoder to encode temporal information and extract features in each sequence. So that it can observe pedestrians' possible crossing intentions within different time steps. The output of the encoder blocks then enters the Late Fusion stage. It will merge data from different branches, comprehensively referencing changes in different data under different conditions and further extracting these features for the final prediction. The fusion stage has two main steps: integrating different features through concatenation, then reducing dimensionality and extracting through a FC(Fully Connected) layer. In the Final Prediction layer, a fully connected layer and a Sigmoid activation function will help to give the final prediction result.

b) Transformer Encoder: The Transformer Encoder block is mainly used in the proposed model. The transformer is released by [8]. One of its major features is that it can analyze and model the information from different time steps in a sequence through a novel architecture and the power of self-attention. Moreover, Transformer has a computational advantage that can be parallelized for processing. Such characteristics and advantages are particularly suitable for use in the pedestrian crossing intention prediction task.

Note that we use only the encoder part of the original Transformer and modify the input embedding layer to a learnable fully connected layer. For the weighted output of the Transformer encoder, we perform a 1D convolutional neural network to integrate the output from different time steps.

C. Loss Function

We use the Focal loss [18] as our loss function instead of Cross Entropy which is commonly used in this field. Focal loss can deal with the fact that models tend to bias towards easy samples in training data and ignore hard samples. This will cause the model to make mistakes when more complex or uncommon data appear.

Focal loss [18] reduces the weight of easy samples and forces the model to pay more attention to hard samples. To be

more specific, the focal loss is given in the following equation.

$$FocalLoss(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

$$\alpha_t = \begin{cases} \alpha & ,y=1 \\ 1-\alpha & ,y=0 \end{cases} \qquad p_t = \begin{cases} p & ,y=1 \\ 1-p & ,y=0 \end{cases} \tag{2}$$

where $\gamma \geq 0$. The parameter α can reduce the effect of imbalanced positive and negative samples. We adjust the unbalanced data using the same strategy in [9] with class weights, so the α_t here is kept at 1, which means no adjustment is applied. The parameter γ controls the importance of easy and hard samples in training. With the larger value of γ, the more the influence of the loss of the positive samples, which are the easy samples here, will be weakened. This achieves the purpose of controlling easy samples and hard samples' influence on the overall training goal. Here we set 2 for the parameter γ after the experiment.

IV. Experiments

A. Experimental Setting

In accordance with the benchmark for predicting pedestrian crossing intentions, two datasets are included in the evaluation of the model.

1) Joint Attention Autonomous Driving (JAAD) Dataset: JAAD [1] is a pedestrian crossing and behavior annotation dataset launched by A. Rasouli et al. earlier in 2017. It contains 391K of the pedestrian sample and was collected from urban scenes in North America and Europe under several weather conditions. The 80% of the sample video resolution is 1280x720 and 20% is 1920x1080. It has two subsets JAAD Beh and JAAD All. The former focused on pedestrians' behavior while they cross the road and annotated with pedestrian behavior. The majority of this subset of samples are crossing samples, so the imbalanced and rare samples along with the lower video quality, cause most methods to perform poorly. The latter adds 2100 noncrossing samples with similar characteristics to the former. Therefore, this subset is highly skewed towards the noncrossing samples, resulting in most of the methods being able to improve their performance over a large number of non-crossing samples.

2) Pedestrian Intention Estimation (PIE) Dataset: In 2019, the author of JAAD launched the PIE dataset [2]. There are 740K pedestrian samples with bounding box annotations in this dataset, all at 1920x1080 resolution and recorded in downtown Canada in clear weather. Although it is still imbalanced and the number of noncrossing samples is larger than the crossing samples. The more diverse pedestrian samples, higher resolution, and more uniform weather conditions, most methods perform better on PIE. The model performance characteristics described above can be observed in the subsequent evaluation experiments.

3) Metrics: Our method follows the five metrics used by the benchmark as our evaluation criteria. These include Accuracy, AUC (Area Under Curve), F1-Score, Precision and Recall.

B. Implementation Details

The adjustment of hyperparameters can significantly affect the training results. In our implementation, there are two types of parameter settings to be aware of - Data sampling parameters for the data interface and the model training hyperparameters based on the Transformer encoder.

1) Data Sampling Parameters of Pedestrian Crossing Prediction: The model is sensitive to parameters like "observation time" and "TTE", the impact of these two parameters on model training has been investigated in previous papers [9] [3]. These two constraints must be strictly defined and set for a fair comparison. According to the benchmark, TTE here is set to 30 to 60 frames in this paper, i.e., 1 to 2 seconds, and the observation time is 16 frames, about 0.5 seconds for 30 frames per second videos.

2) Training Details and Hyperparameters: We are primarily concerned with the following four hyperparameters in Transformer Encoder. The one controls the dimension of input embedding and output of multi-head self-attention, the dimension in the feed-forward network, the number of layers of the transformer encoder, and the number of heads of the multi-head self-attention. The hyperparameter settings we apply to each data branch are listed below in the order mentioned above. Bounding box and surrounding image: 128, 128, 1, 4. 3D pose: 128, 64, 2, 4. 3D head orientation: 128, 64, 1, 4. Bounding box: 128, 128, 1, 4. Ego-vehicle speed: 128, 128, 1, 4. Traffic perception data: 128, 128, 1, 4.

The hyperparameters that are relevant to the training process are listed below. The batch size is set to 128, and the learning Rate is set to 0.001 in our experiments. The output dimensions of the two FC layers for fusion and final prediction are 128 and 1, respectively. Furthermore, we use Adam optimizer and apply the same class weights as in the benchmark method[9] to mitigate the problem of imbalance between crossing and non-crossing samples.

C. Evaluation of the Proposed Model

The comparison of the evaluation results of our proposed method with other state-of-the-art baselines on the benchmark[9] is presented in Table I. Our proposed method achieves the best accuracy results in all evaluation datasets. Although not all of the remaining evaluation metrics are best achieved by our method, it can be claimed that our method achieves better balance across the three datasets.

The best results for different metrics are generally achieved by different methods. Some methods have achieved a significant lead in specific metrics but at a relative cost. For example, in the PIE comparison, I3D achieves a recall of 0.9 but a precision of 0.6, compared to 0.85 and 0.84 from our method which has a better balance. This is reflected in the AUC and F1.

For JAAD beh, our method outperforms all the methods that provide JAAD beh scores. Only the recall did not reach the 0.91 obtained by TrouSPI-Net, but its precision was reduced to 0.65 compared to ours. Our method maintains better accuracy,

TABLE I

COMPARISON WITH OTHER METHODS. WHERE "-" INDICATES THAT THE METHOD DOES NOT PROVIDE THESE METRICS.

Method	PIE					JAAD all					JAAD beh				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
SFGRU[3]	0.82	0.79	0.69	0.67	0.7	0.84	0.84	0.65	0.54	**0.84**	0.51	0.45	0.63	0.61	0.64
I3D[9]	0.81	0.83	0.72	0.60	**0.9**	0.84	0.8	0.63	0.55	0.73	0.62	0.51	0.75	0.65	0.88
TrouSPI-Net[19]	0.88	0.87	0.80	0.77	0.84	0.82	0.77	0.58	0.49	0.70	0.64	0.55	0.76	0.65	**0.91**
MultiRNN[20]	0.83	0.8	0.71	0.69	0.73	0.79	0.79	0.58	0.45	0.79	0.61	0.5	0.74	0.64	0.86
D. Yang .et al.[7]	-	-	-	-	-	0.83	0.82	0.63	0.51	0.81	0.62	0.54	0.74	0.65	0.85
PCPA[9]	0.87	0.86	0.77	-	-	0.85	**0.86**	0.68	-	-	0.58	0.5	0.71	-	-
IntFormer[21]	0.89	**0.92**	0.81	-	-	0.86	0.78	0.62	-	-	0.59	0.54	0.69	-	-
Yu Yao .et al.[22]	0.84	0.90	**0.88**	**0.96**	-	0.87	0.70	**0.92**	0.66	-	-	-	-	-	-
BiPed[23]	**0.91**	0.90	0.85	0.82	-	0.83	0.79	0.60	0.52	-	-	-	-	-	-
Ours	**0.91**	0.89	0.84	0.84	0.85	**0.89**	0.78	0.66	**0.72**	0.61	**0.68**	**0.63**	**0.76**	**0.71**	0.81

TABLE II

WE CONDUCTED EXPERIMENTS ON DIFFERENT COMBINATIONS OF INPUT DATA TO VERIFY THE CONTRIBUTION OF EACH INPUT FEATURE. "B" IS BOUNDING BOX COORDINATES, "TP" IS TRAFFIC PERCEPTION DATA, "LB" IS LOCAL BOUNDING BOX IMAGE, "LS" IS LOCAL SURROUNDING IMAGE, "S" IS EGO-VEHICLE SPEED, "H" IS 3D HEAD ORIENTATION, "P" IS 3D HUMAN POSE.

Features	PIE					JAAD all					JAAD beh				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
LB, LS	0.71	0.61	0.43	0.49	0.38	0.86	0.75	0.60	0.62	0.58	0.63	0.53	0.75	0.64	0.90
B, S	0.90	**0.89**	0.82	0.79	**0.85**	0.84	0.77	0.59	0.55	0.65	0.61	0.57	0.70	0.68	0.72
B, S, TP	**0.92**	**0.89**	**0.85**	**0.86**	0.84	0.83	0.78	0.59	0.51	0.70	0.62	0.57	0.71	0.68	0.75
B, S, TP, H	0.91	**0.89**	0.84	0.84	**0.85**	0.83	0.78	0.59	0.51	**0.71**	0.63	0.62	0.68	**0.71**	0.65
LB, LS, B, S	0.89	0.86	0.80	0.81	0.80	0.88	0.78	0.65	0.64	0.65	0.65	0.55	0.77	0.65	**0.93**
LB, LS, B, S, TP	0.89	0.86	0.81	0.83	0.79	0.88	0.75	0.63	0.72	0.55	0.66	0.57	**0.76**	0.66	**0.93**
LB, LS, B, S, H, P	0.88	0.85	0.79	0.81	0.77	0.88	0.74	0.60	**0.74**	0.51	0.64	0.56	**0.76**	0.66	0.88
LB, LS, B, S, TP, H, P	0.91	**0.89**	0.84	0.84	**0.85**	**0.89**	0.78	**0.66**	0.72	0.61	**0.68**	**0.63**	**0.76**	**0.71**	0.81

while the model is more precise in discriminating different situations, and the difference between precision and recall is also smaller. In JAAD All, our method leads in accuracy and precision but does not perform as well on other metrics as on other datasets. It might be the problem of significant gaps between crossing and non-crossing samples and can be observed in the decreasing Recall of our method. Reducing the impact of imbalanced samples will be an area that our method needs to consider carefully in the future.

Despite this, our approach strikes a balance performance between the three datasets. This is valuable when all three datasets have their own significant bias and unbalanced data and can be observed in the ablation study.

V. ABLATION STUDY

A. Importance of Input Data

Table II shows how the different input data combinations impact the training results of our method. Due to space limitations, we have selected several representatives or more significantly varied test results to present in Table II.

Many previous studies [24] [21] [9] have pointed out that the PIE and JAAD datasets are more sensitive to specific input data. This means we can achieve good baseline results with only certain input data. For example, as shown in rows B, S and B, S, TP in Table II. The performance on PIE improved significantly after using bounding boxes and speed. Even after fusing our proposed traffic perception data, the score surpassed our proposed method and was ahead of other methods.

However, as mentioned previously, we incorporated more information to support the stability and balance of the model in a variable scenario. This ensures that the scores on the PIE are comparable to other methods and also cover the performance in other datasets.

Then, as shown in the lower third of the table, a significant performance drop occurs after removing any of the TP, H, or P data. We conclude that these three types of data complement each other, allowing the model to adapt to more scenarios and proving the validity of these data.

TABLE III

COMPARISON OF DIFFERENT FUSION METHODS. WHERE THE ABBREVIATION "CH ATT" IS CHANNEL ATTENTION.

Fusion Method	PIE					JAAD all					JAAD beh				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
Hierarchical	**0.91**	**0.89**	**0.84**	**0.84**	0.83	0.88	0.75	0.62	**0.72**	0.54	0.67	0.58	**0.77**	0.67	0.91
Eearly Fusion	0.86	0.81	0.73	0.79	0.68	0.87	0.78	0.63	0.62	**0.64**	0.64	0.54	0.76	0.64	**0.93**
Late Fusion + CH Att	0.90	0.88	0.83	**0.84**	0.82	0.88	**0.79**	**0.66**	0.68	0.63	0.64	0.56	0.75	0.66	0.87
Late Fusion + FC	**0.91**	**0.89**	**0.84**	**0.84**	**0.85**	**0.89**	0.78	**0.66**	**0.72**	0.61	**0.68**	**0.63**	0.76	**0.71**	0.81

B. Comparison of Different Fusion Methods

We performed an experiment to verify the effectiveness of our model architecture and fusion methods. In addition to the Late Fusion method used in our method, we compare three different solutions. The results are shown in Table III.

The first is a hierarchical fusion method inspired by [3], which integrates the input information by progressively fusing different inputs in every layer. In the second method, Early fusion is used to fuse data before it enters the Transformer encoder block. Finally, based on the proposed method, we replace the Fully connected layer of late fusion with channel attention mechanism, hoping that the channel attention mechanism can automatically adjust the weights of different input information to optimize the training results.

Among these four fusion schemes, the scheme of combining Late fusion with the fully connected layer gives the most satisfactory results. In the hierarchical and channel attention scheme, these two more complex models tend to overfit as they rapidly converge to the training set during training. Early fusion, on the other hand, gives poor results because the input data is fused before the feature extraction. It is also not able to learn complex scenes well due to the lack of model complexity, which can be observed in the reduced accuracy on PIE dataset.

TABLE IV

COMPARISON OF DIFFERENT TRAFFIC PERCEPTION DATA FUSION STRATEGIES. "T" IS TRAFFIC LIGHT, "S" IS SIGN, "C" IS CROSSWALK. WHERE "NO FUSION & NO TRANSFORM" MEANS NO FUSION AND FORMAT TRANSFORM IS APPLIED. "T+SC" MEANS T IS A SINGLE BRANCH WITHOUT FUSION, WHILE S AND C ARE FUSED. THE OTHER TWO FOLLOW THIS PATTERN. "TCS EARLY FUSION" IS THE PROPOSED METHOD WITH THREE DATA FUSIONS.

TP Fusion Strategy	PIE					JAAD all					JAAD beh				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
No Transform & Fusion	0.89	0.85	0.81	0.84	0.78	0.87	0.76	0.61	0.65	0.58	0.65	0.56	0.77	0.66	0.92
No Fusion	0.90	0.87	0.83	0.85	0.80	0.88	0.77	0.64	0.70	0.59	0.66	0.56	0.77	0.66	**0.94**
T + SC	0.90	0.87	0.82	0.84	0.80	0.88	0.73	0.60	0.72	0.51	0.69	**0.67**	0.76	**0.75**	0.76
TS + C	0.90	0.87	0.82	0.85	0.78	**0.89**	0.72	0.59	**0.79**	0.47	**0.70**	0.66	**0.78**	0.73	0.83
TC + S	0.90	0.86	0.82	**0.87**	0.77	0.88	0.77	0.64	0.69	0.60	**0.70**	**0.67**	0.76	**0.75**	0.78
TCS Early Fusion	**0.91**	**0.89**	**0.84**	0.84	**0.85**	**0.89**	**0.78**	**0.66**	0.72	**0.61**	0.68	0.63	0.76	0.71	0.81



Fig. 3. The input features introduced in our method will help the model to predict more accurately in specific situations and improve the cases that are easily misclassified in the figure. Green bounding boxes indicate pedestrian samples that cross, while Red ones indicate those that do not.

C. Comparison of Traffic Perception Data Fusions

We represent the traffic lights, signs and crosswalks in a special format and integrate them using concatenant operations, so that this information brings a visible improvement to the model training. Here, we test six different fusion strategies. Without the format transformation and fusion, such data's contribution almost negligibly to the training results, i.e. the "No Transform & No Fusion" row in Table IV.

Compare to other fusion strategies, our final solution TSC Eealy Fusion, demonstrates the best results, maintaining a decent level of performance in all three datasets, but was not clearly biased towards certain datasets. For example, the TS+C and TC+S fusion strategies show excellent results in JAAD beh, but they did not provide improvement in PIE.

D. Qualitative Justification

Figure 3 shows the visualization prediction results of this method. The model can correctly predict complex samples. The addition of traffic perception data, 3D pedestrian head orientation, and 3D human pose helps the model improve the misclassification in many cases before.

E. Discussion of Failure Case

Among the failed cases, we found a condition that confuses the model and repeatedly occurs in incorrect cases. As shown in Figure 4, these cases occur when the vehicles are about to turn and the directions of the vehicles are changed after the observation. For example, the original direction of the vehicle is parallel to the pedestrian, and it does not change in the observation time. When a vehicle turns after observation and

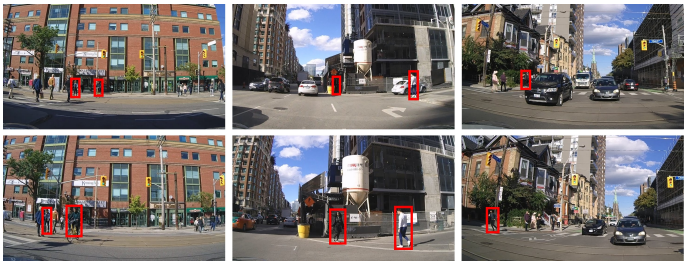


Fig. 4. These cases illustrate the impact of ego-vehicle turning after the model observation. Where Green bounding boxes indicate pedestrian samples that cross, while Red ones indicate those that do not.

a pedestrian starts to cross the road at the same time, the route of the vehicle turns to intersect with the pedestrian.

This case is confusing to the model because the model naturally defines such pedestrians as non-crossing since there is no overlap between pedestrian and vehicle trajectories during the model observation time. Such a condition would contradict with the sample of vehicles going straight because there is no significant difference between the two cases in the model observation condition, but the ground truth shows two opposite results. Therefore, we suggest that future datasets could include this kind of information, which would be beneficial for related research.

For other failures such as extreme situations where the pedestrian appearance is not visible or when there is no motion information available due to minimal pedestrian and vehicle movement. We believe that image processing or restoration models can be used to try to cope with the first case, while the second case can be aided by obtaining more accurate and rich information about pedestrian posture to predict the scene.

VI. CONCLUSIONS

In this paper, we proposed a multi-modal Transformer-based approach to predict the intention of a pedestrian crossing the road. In this work, we explore how people determine the crossing intention of pedestrians from the perspective of human drivers and pedestrians themselves and integrate the pieces of critical information to help the model make accurate predictions.

To the best of our knowledge, we are the first to represent traffic lights, road signs, crosswalks in a novel way and consider the lifted 3D human pose and 3D head orientation information to help the model better understand the traffic environment and the behavior of pedestrians. Our experimental results demonstrated that the proposed model achieves state-of-the-art performance across the three benchmark datasets. In addition, we performed several experiments to verify the effectiveness of various components and different input data of our method.

References

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," pp. 206–213, 2017.

[2] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6261–6270. DOI: 10.1109/ICCV.2019.00636.

[3] A. Rasouli, I. Kotseruba, and J. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," in *Proceedings of the British Machine Vision Conference (BMVC)*, K. Sidorov and Y. Hicks, Eds., BMVA Press, Sep. 2019, pp. 49.1–49.13. DOI: 10.5244/C.33.49.

[4] Z. Sui, Y. Zhou, X. Zhao, A. Chen, and Y. Ni, "Joint intention and trajectory prediction based on transformer," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 7082–7088.

[5] B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, "Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5338–5349, 2021.

[6] A. T. Schulz and R. Stiefelhagen, "Pedestrian intention recognition using latent-dynamic conditional random fields," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 622–627. DOI: 10.1109/IVS.2015.7225754.

[7] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 221–230, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2017. arXiv: 1706.03762 [cs.CL].

[9] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for Evaluating Pedestrian Action Prediction," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1258–1268.

[10] D. Varytimidis, F. Alonso-Fernandez, B. Duran, and C. Englund, "Action and intention recognition of pedestrians in urban traffic," in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2018, pp. 676–682. DOI: 10.1109/SITIS.2018.00109.

[11] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971. DOI: 10.1109/CVPR.2016.110.

[12] S. A. Bouhsain, S. Saadatnejad, and A. Alahi, "Pedestrian intention prediction: A multi-task perspective," *arXiv preprint arXiv:2010.10270*, 2020.

[13] R. Rozenberg, J. Gesnouin, and F. Moutarde, "Asymmetrical bi-rnn for pedestrian trajectory encoding," *arXiv preprint arXiv:2106.04419*, 2021.

[14] M. I. Perdana, W. Anggraeni, H. A. Sidharta, E. M. Yuniarno, and M. H. Purnomo, "Early warning pedestrian crossing intention from its head gesture using head pose estimation," in *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2021, pp. 402–407. DOI: 10.1109/ISITIA52817.2021.9502231.

[15] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6d rotation representation for unconstrained head pose estimation," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2496–2500. DOI: 10.1109/ICIP46576.2022.9897219.

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, *Openpose: Realtime multi-person 2d pose estimation using part affinity fields*, 2019. arXiv: 1812.08008 [cs.CV].

[17] Z. Chen, A. Sugimoto, and S.-H. Lai, "Learning monocular 3d human pose estimation with skeletal interpolation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4218–4222.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[19] J. Gesnouin, S. Pechberti, B. Stanciulcscu, and F. Moutarde, "Trouspi-net: Spatio-temporal attention on parallel atrous convolutions and u-grus for skeletal pedestrian crossing prediction," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, IEEE, 2021, pp. 01–07.

[20] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4194–4202.

[21] J. Lorenzo, I. Parra, and M. Sotelo, "Intformer: Predicting pedestrian intention with the aid of the transformer architecture," *arXiv preprint arXiv:2105.08647*, 2021.

[22] Y. Yao, E. Atkins, M. J. Roberson, R. Vasudevan, and X. Du, "Coupling intent and action for pedestrian crossing behavior prediction," *arXiv preprint arXiv:2105.04133*, 2021.

[23] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 600–15 610.

[24] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is attention to bounding boxes all you need for pedestrian action prediction?" In *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 895–902. DOI: 10.1109/IV51971.2022.9827084.

Revolutionizing Formative Assessment in STEM Fields: Leveraging AI and NLP Techniques

Chi Wee Tan* and Khai Yin Lim
* Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, 53300 Setapak, Kuala Lumpur, Malaysia E-mail: chiwee@tarc.edu.my
Department of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Penang, Malaysia E-mail: limky@tarc.edu.my

***Abstract* —Artificial intelligence (AI) has been extensively studied in science, technology, engineering, and mathematics (STEM), but there is a disparity between AI-generated and human-written scientific content. To bridge this gap, a prototype utilizing Natural Language Processing (NLP) techniques and a large language model (LLM) generates assessment questions and evaluates student answers. This formative assessment system offers a user-friendly and scalable solution for higher education educators. It tailors' assessments to individual students, accommodates varying capabilities, and facilitates performance analysis. Through rigorous evaluation and benchmarking, the prototype ensures alignment with High-Level Performance (HLP) standards. This AI-assisted formative assessment system enhances efficiency and efficacy by providing accurate and timely feedback. It has the potential to significantly improve STEM education through scalable and personalized formative assessment experiences. AI and NLP enable educators to access tailored assessment options, enhancing learning outcomes and the overall educational experience.**

I. INTRODUCTION

Formative assessment is a fundamental process in the provision of feedback to students regarding their learning progress. It constitutes a vital component of effective teaching and learning practices in the fields of science, technology, engineering, and mathematics (STEM). This approach enables students to identify areas requiring additional support and make necessary adjustments to their learning strategies. Historically, formative assessment has been undertaken by educators utilizing diverse methods such as quizzes, tests, and classroom discussions. Nonetheless, these approaches can be time-consuming, labor-intensive, and challenging to implement on a larger scale. The advent of artificial intelligence (AI) presents an opportunity to revolutionize formative assessment in STEM disciplines by providing a more efficient and scalable feedback mechanism to students.

AI-powered formative assessment systems have the potential to automatically evaluate student work, pinpoint areas necessitating further support, and deliver personalized feedback aimed at enhancing learning outcomes. In this paper, we introduce a prototype AI-assisted formative assessment system leveraging natural language processing (NLP) techniques. Our system is specifically designed for implementation in STEM courses and has the capability to assess short-answer questions, essays, and other types of student assignments. We conducted an evaluation of our system involving a group of STEM educators. The results demonstrated that our system was capable of delivering accurate and timely feedback to students, leading to improvements in their learning outcomes. We firmly believe that our system possesses the capacity to enhance the quality of STEM education by providing a more efficient and scalable approach to formative assessment. Our ongoing efforts involve further development of the system, with the aim of making it accessible to a wider audience in the future.

II. LITERATURE REVIEW

A. Digital Formative Assessment

The inclusion of digital formative assessment has become an essential component of contemporary education, enabling educators to collect immediate feedback and effectively monitor student progress. This literature review aims to examine the concept of digital formative assessment and emphasize the range of tools and technologies available to facilitate its integration into educational environments. Formative assessment involves gathering data on students' learning advancement to offer prompt feedback and adjust teaching strategies accordingly. Digital formative assessment leverages technology to enhance this process, creating

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

opportunities for interactive, personalized, and instantaneous feedback. Research indicates numerous advantages of implementing digital formative assessment in classrooms. Firstly, it fosters active student engagement by incorporating interactive elements like quizzes, polls, and discussions. Secondly, it enables teachers to efficiently track individual student progress and identify areas requiring improvement. Thirdly, digital formative assessment supports differentiated instruction, enabling educators to tailor their teaching to meet each student's specific needs. Lastly, it delivers timely and constructive feedback to students, cultivating a growth mindset and enriching their learning experience. In the pursuit of enhancing support for the development of adaptive learning environments, the interdisciplinary field of learning science has attracted the attention of computer scientists for over a decade. Researchers from the education and computer science domains have been working together to optimize the complex and time-consuming design of adaptive learning environments through the utilization of advanced techniques such as data mining, machine learning, and deep learning. Their objective is to enhance the capabilities of Intelligent Tutoring Systems (ITS) in order to provide more effective assistance to students. Numerous endeavours have been witnessed in the development of ITS for various subjects such as mathematics [1], [2], programming [3]–[6], languages [7]–[10] and more. Efficiently and effectively guiding students in their learning process remains a recurring topic in educational research of the 21st century.

B. Tools for Digital Formative Assessment

Learning Management Systems (LMS): LMS platforms like Moodle, Canvas, and Schoology offer features for creating online assessments, tracking student performance, and providing feedback. These systems often include built-in grading tools, discussion boards, and analytics, enabling teachers to monitor student progress comprehensively [11], [12].

Quiz and Polling Tools: Tools like Kahoot, Quizizz, and Mentimeter facilitate the creation of interactive quizzes and polls. These tools engage students in a gamified learning experience, encouraging participation and enabling instant feedback [13].

Online Survey Tools: Survey tools such as Google Forms and Microsoft Forms provide educators with the means to collect data from students, gauge their understanding, and assess their learning progress. These tools offer a variety of question types and data analysis features [14].

Digital Whiteboarding Tools: Whiteboarding tools like Jamboard, Miro, and Explain Everything enable collaborative and interactive formative assessment. Teachers can create visual presentations, interactive exercises, and virtual whiteboards to facilitate student engagement and feedback [15].

Video Assessment Tools: Video assessment tools, such as Flipgrid and Edpuzzle, allow students to record their responses, presentations, or explanations. Teachers can review and provide feedback on these videos, promoting self-reflection and communication skills [16]–[18].

C. Artificial Intelligence as a tool for formative assessment

There has been a growing interest in the use of AI for formative assessment in recent years. Several research studies have been conducted to evaluate the effectiveness of AI-assisted formative assessment systems [19], [20]. The findings from these studies have yielded varied results, but on the whole, they indicate that AI-supported formative assessment systems have the potential to enhance student learning outcomes. Incorporating AI into formative assessment poses certain challenges, including the difficulty of developing accurate and reliable systems. Additionally, the development and implementation of AI-assisted formative assessment systems can be costly. Despite these obstacles, the potential benefits of AI in formative assessment are substantial. AI-supported formative assessment systems have the capacity to enhance the efficiency and scalability of assessment practices, while providing personalized feedback to students, thereby aiding their learning progress. Although the use of AI in formative assessment is still in its early stages, it holds the potential to revolutionize the evaluation methods employed for students. As AI-assisted formative assessment systems continue to advance, they can significantly improve the efficiency, scalability, and personalization of assessment processes, consequently leading to notable enhancements in student learning outcomes. Below are a few instances showcasing the application of AI in formative assessment:

- **Grading student work:** AI-assisted formative assessment systems can automatically grade student work, such as essays and short-answer questions. This can free up teachers' time so they can focus on providing feedback to students.
- **Identifying areas of need:** AI-assisted formative assessment systems can identify areas where students need additional support. This information can be used to provide targeted interventions to help students improve their learning.

- **Providing personalized feedback:** AI-assisted formative assessment systems can provide personalized feedback to students. This feedback can help students to identify their strengths and weaknesses, and to develop strategies for improving their learning.

D. Large language models (LLMs)

Large language models (LLMs) are large neural networks trained on enormous amounts of text data, sometimes at the petabyte scale. They could generate text that's far more fluent and coherent than previous language models, and they can also be used as a strong foundation for other NLP tasks. LLMs represent a major advancement in AI, with the promise of transforming domains through learned knowledge [21]. A bibliometric review of LLMs research has shown that LLMs are a class of language models that have demonstrated outstanding performance across a range of natural language processing (NLP) tasks and have become a highly sought-after research area, because of their ability to generate human-like language and their potential to revolutionize science [22]. Large Language Models (LLMs) are sophisticated computer programs that introduce new possibilities for text comprehension and generation within software systems like Google Search and Bing. These models undergo extensive training on vast amounts of text, enabling them to learn intricate relationships between terms. Once trained, LLMs have the capability to produce text that closely resembles human-written content. Numerous large language models have been developed, including OpenAI's GPT-3 and GPT-4, Meta's LLaMA, and Google's PaLM2. These models possess language understanding and text generation capabilities. For instance, GPT-3, with its 175 billion parameters, can comprehend and generate text across more than 95 natural languages and 12 programming languages. Google's BERT, with 340 million parameters, can understand and generate text in 104 languages through its multilingual model. BigScience's BLOOM, with 176 billion parameters, can handle text comprehension and generation in 46 natural languages and 13 programming languages. NVIDIA's NeMo LLM, with 530 billion parameters, specializes in English text understanding and generation exclusively. LLMs have been used for various applications such as chatbots, machine translation, question answering systems, summarization systems, sentiment analysis systems, speech recognition systems, image captioning systems, among others. In this context, is believe that LLMs can be used in assisting educators in formative assessment by providing feedback to students and modifying their teaching based on the information derived from assessments [23].

Table. 1 Comparison of Different Large Language Models.

Model (Year Founded)	**Type of license**	**Supported features**	**Limitations**	**Advantages**
MarianMT (2017)	Open Source	Language translation	Limited support for other NLP tasks	Free and open source
BERT (2018)	Open Source	Language understanding, text classification	May require large amounts of labelled training data	Strong performance in various NLP tasks
Megatron LM (2019)	Open Source	Text generation, language translation, Q&A, chatbots	Requires significant computational resources	Optimized for training large-scale models
RoBERTa (2019)	Open Source	Language understanding, text classification	Limited support for fine-grained text generation	Pretrained on large corpus for improved accuracy
CTRL (2019)	Open Source	Text generation, language translation, Q&A, chatbots	Less optimized for large-scale language tasks	Fine-tuned for content creation and chatbots
T5 (2020)	Open Source	Text generation, language translation, Q&A, chatbots	Less emphasis on model customization and fine-tuning	Unified architecture for diverse language tasks
GPT-2 (2019)	Proprietary	Text generation, language translation, Q&A, chatbots	Limited fine-grained control over model behaviour	Strong text generation capabilities
Turing-NLG (2021)	Open Source	Text generation, language translation, Q&A, chatbots	Limited availability and support	Innovative model architecture
GPT-Neo (2021)	Open Source	Text generation, language translation, Q&A, chatbots	Less optimized for commercial	Open-source alternative to GPT models

			deployment	
GPT-3 (2020)	Proprietary	Text generation, language translation, Q&A, chatbots	Limited fine-grained control over model behaviour	State-of-the-art performance in various tasks
GPT-4 (2022)	Proprietary	Text generation, language translation, Q&A, chatbots	Limited fine-grained control over model behaviour	Improved performance and capabilities
OPT (OpenAI) (2022)	Proprietary	Text generation, language translation, Q&A, chatbots	Limited fine-grained control over model behaviour	Latest model with enhanced features and quality

III. METHODOLOGY

In this section, we will provide a detailed explanation of the methodology employed in developing the AI-assisted prototype, which utilizes NLP techniques to enhance educators' capabilities in the context of formative assessment (as illustrated in Figure 1). The proposed solution incorporates the utilization of the OpenAI API key. The OpenAI API is an Application Programming Interface (API) that provides access to advanced AI models developed by OpenAI. It empowers our proposed system to comprehend natural language and generate responses that closely resemble human-like outputs (Odunayo, 2023).

OpenAI provides a diverse array of models, each catering to specific application requirements and offering varying degrees of capabilities. The primary function of OpenAI API are:

Understanding Educator Input. The OpenAI GPT API takes the input provided by the educator as a starting point. This input can include prompts, questions, or specific instructions.

Contextual Understanding. GPT (Generative Pre-trained Transformer) models, such as GPT-3, are designed to understand and generate text in a contextual manner. They analyse the input provided and use the context to generate meaningful responses.

Question Generation. If the educator's input contains prompts or instructions to generate questions, the OpenAI GPT API can generate a series of questions based on the given

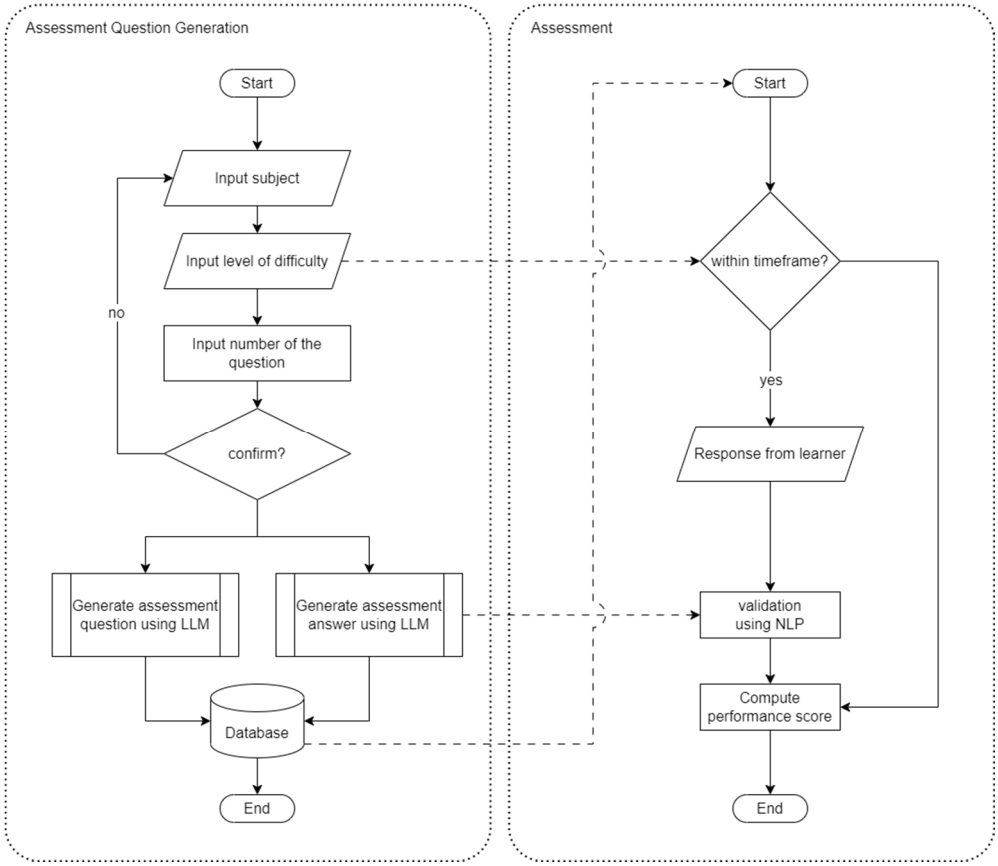


Fig. 1 Proposed solution architecture.

context. These questions are designed to be coherent and contextually relevant.

To facilitate effective preprocessing and enhance the overall functionality of our solution, we have integrated the Natural Language Toolkit (NLTK), a robust Python library designed explicitly for efficient and proficient natural language processing tasks. With NLTK, our system can effectively process human language, ensuring optimal performance:

Text Tokenization. NLTK is used for breaking down the text into smaller units such as words or sentences. Tokenization is an important step in understanding and processing text data.

Lowercasing. NLTK can be used to convert all text to lowercase. This standardizes the text and ensures that the processing is not case-sensitive.

Removing Stop Words. Stop words are common words (e.g., "the," "and," "is") that do not carry significant meaning. NLTK provides tools for removing these stop words from the text, which can improve the accuracy of text analysis.

Stemming and Lemmatization. NLTK offers stemming and lemmatization techniques that reduce words to their base or root forms. This helps in standardizing variations of words (e.g., "running" becomes "run") and improves text analysis accuracy.

Additionally, we leverage the Twinword Text Similarity API to facilitate a comparative analysis of learner responses against correct answers. This API serves as a semantic keyword research tool, employing a ranking mechanism that assesses keyword relevance to user-defined topics. This functionality empowers our system to compute the degree of similarity between learner-provided responses and the corrected answers generated by OpenAI GPT-3. Consequently, our solution streamlines the automated grading of open-ended questions, as it can objectively assess a learner's response by comparing it to the correct answer and assigning a score based on their similarity. This streamlined process saves valuable time for educators, provides learners with immediate feedback, and delivers a more objective evaluation of their performance.

IV. Result And Discussion

Table 3 shows the sample assessment question generated by the proposed framework with its sample marking scheme.

Table. 2 Sample Output.

	Assessment Question	**Sample Marking Scheme**
1	What is an algebraic expression?	An algebraic expression is a combination of numbers, variables and operations that are written out in a mathematical statement.
2	What is the difference between an equation and expression?	An equation is a statement that two expressions are qual, while an expression is composed of numbers and symbols that represents a mathematical statement.
3	How do you solve an algebraic equation?	To solve an algebraic equation, isolate the variable on one side of the equation then use inverse operations to get the variable alone on the on the other side of the equation. Once the variable is isolated, substitute the value for the variable into the equation.

Evaluation and benchmarking of the assessment system can be conducted by comparing its automated grading system, which utilizes AI and Natural Language Processing (NLP) with the Twinword semantic similarity feature, against the standard of Human Level Performance (HLP). HLP (as shown in Table 3) serves as a reference point, representing the level of grading accuracy and understanding achieved by human assessors. is employed to validate the questions generated by the proposed system, ensuring their validity, accuracy, and relevance. Additionally, it is utilized to verify that the sample answers generated align correctly with the corresponding questions. Through this evaluation, the system's performance can be assessed by comparing its grading results with those of human assessors.

Table. 3 Sample Data with Benchmarking with HLP.

Test Plan				**Performance**	
Subject	Level	Question Generated	Suggested Answer	HLP on Question Generated (Validity)	HLP on Answer Generated (Accuracy)
Mathematics	Easy	Q1. What is Algebra?	A1. Algebra is a branch of mathematics that uses symbols and equations to represent numbers, relationships, and objects.	Yes	Yes
		Q2. What are variables in Algebra?	A2. Variables are letters or symbols that stand for a particular number in an equation.	Yes	Yes
		Q3. What is an equation?	A3. An equation is a mathematical statement expressing the equality of two expressions. It is the statement of equality between two expressions that have the same value.	Yes	Yes
	Moderate	Q1. What is a Linear Equation?	A1. A linear equation is an equation of a straight line, written in the form ax + b = 0;	Yes	Yes

		where 'a' and 'b' are constants and 'x' is a variable.		
	Q2. How do you solve a System of Linear Equations?	A2. A system of linear equations can be solved using various methods such as Substitution Method, Elimination Method, and Gaussian Elimination.	Yes	Yes
	Q3. What is the Quadratic Formula?	A3. The Quadratic Formula is a method used to solve a quadratic equation, written in the form ax2 + bx + c = 0; where 'a', 'b', and 'c' are constants and 'x' is a variable. The formula is x = [-b ± √(b2 - 4ac)] / 2a.	Yes	Yes
Hard	Q1. Solve for x in the equation: 3x - 6 = 12	A1. x = 8	Yes	Yes
	Q2. Divide the polynomial 4x^3 + 27x^2 - 73x + 60 by 2x + 5	A2. 4x^2 + 17x - 43	Yes	Yes
	Q3. Find the product of polynomials (2x^2 + 3x)(5x + 4)(x - 1)	A3. 10x^3 + 31x^2 - 23x - 4	Yes	Yes
Extreme Hard	Q1. Prove that the product of two odd numbers is always an odd number.	A1. Let a and b be two odd numbers. This means a = 2n + 1 and b = 2m + 1, where n and m are integers. Thus, their product a*b = (2n+1)*(2m+1) = 4nm + 2(n+m) + 1. Since n and m are both integers, the whole expression is an integer plus one - thus it is always an odd number.	Yes	Yes
	Q2. Prove that (2x^2 -5x + 3) is a prime number for any integer value of x.	A2. Let x be any integer. Then, 2x^2 - 5x + 3 can be written as (2x - 3)(x + 1). Since at least one of the factors is always a non-zero integer, the expression 2x^2 - 5x + 3 is not prime for any integer value of x.	Yes	Yes
	Q3. Find the value of x in the equation x^3 + 3x -2 = 0.	A3. Using the cubic formula, x = (∛(-2 + √48) + ∛ (-2 - √48))/3. Therefore, the value of x is approximately -1.9021.	Yes	Yes

The AI and NLP algorithms employed in our system employ advanced techniques to analyze and evaluate student responses. Utilizing the Twinword semantic similarity feature, the system measures the degree of similarity between student answers and the correct answers. By benchmarking the system's performance against Human-Level Performance (HLP), we can assess the system's accuracy and proficiency in automated grading. This evaluation approach holds immense value in determining the effectiveness and reliability of the system. It offers valuable insights into the system's capacity to comprehend and accurately assess student answers, comparable to human assessors' expertise. Furthermore, it facilitates the identification of areas where the system excels or requires further improvement.

By subjecting the assessment system to evaluation against HLP using AI and NLP techniques, educators and developers gain crucial insights into its performance. This process enables them to refine and enhance the system's automated grading capabilities, aligning it with human-level grading standards while harnessing the power of AI and NLP technologies. It is worth highlighting that the proposed solution successfully satisfies all the criteria and test scenarios for the HLP evaluation. The assessment system demonstrates its ability to achieve grading accuracy and understanding at a level comparable to human assessors. This significant achievement attests to the system's effectiveness and reliability in automated grading, instilling confidence in its capacity to deliver high-quality assessments.

V. CONCLUSIONS

In conclusion, formative assessment holds a pivotal role in effective teaching and learning, particularly within STEM

fields, by offering students invaluable feedback on their progress. However, conventional assessment methods like quizzes and tests are often plagued by time constraints, resource demands, and scalability challenges. The integration of artificial intelligence (AI) has emerged as a transformative solution, revolutionizing formative assessment in STEM education by presenting a more efficient and scalable approach to delivering feedback. This paper introduces a prototype AI-assisted formative assessment system enriched with natural language processing (NLP) techniques, underscoring its potential to elevate formative assessment practices.

Through rigorous evaluation involving STEM educators, this system has demonstrated its prowess in providing accurate and prompt feedback, culminating in tangible enhancements in learning outcomes. By automating the grading process, identifying areas necessitating additional support, and delivering personalized feedback, the system optimizes the efficiency and efficacy of formative assessment, adeptly catering to individual student requirements. A cornerstone of this system's success lies in the incorporation of large language models (LLMs) and NLP techniques, which amplify its proficiency in comprehending natural language and generating coherent responses. Furthermore, the integration of supplementary technologies like the Twinword Text Similarity API bolsters the system's capacity to objectively assess student responses, juxtapose them against correct answers, and assign precise scores. This empowerment equips educators with time-saving features, grants students' immediate feedback, and institutes a more objective performance evaluation process.

The utilization of LLMs, such as GPT-4 and OPT (OpenAI), signifies a strategic advantage, harnessing their language understanding and text generation capabilities, as showcased in Table 1. These models have evolved as powerful tools with varying strengths and weaknesses. For instance, models like BERT excel in language understanding and text classification, while GPT-3 and GPT-4 exhibit strong text generation capabilities. MegatronLM, on the other hand, is optimized for training large-scale models, making it suitable for extensive natural language processing tasks. Each model offers a unique set of features, limitations, and advantages, enabling educators to tailor their formative assessment strategies to specific needs. While challenges persist, encompassing aspects like development and deployment costs, the potential merits of AI-assisted formative assessment systems are substantial. Their scalability and personalization capabilities hold the promise of revolutionizing education, affording educators comprehensive insights into student progress, and facilitating tailored interventions. Ongoing development endeavours aim to hone and broaden the system's capabilities, with the overarching objective of rendering it accessible to a wider audience and magnifying its impact on STEM education. By harnessing AI and NLP technologies, educators and developers can incessantly refine the system's accuracy, reliability, and alignment with human-level grading standards. This iterative approach guarantees that the system remains at the forefront of advancements in formative assessment practices, ultimately fostering enhanced student learning outcomes within STEM disciplines.

In summary, the fusion of AI and NLP techniques in formative assessment signifies a paradigm shift in STEM education. The presented AI-assisted formative assessment system underscores its capacity to furnish precise and timely feedback, thereby heightening the efficiency and effectiveness of formative assessment practices, while strategically leveraging advanced large language models like GPT-4 and OPT (OpenAI) and considering the unique features of various LLMs, as highlighted in Table 1. As further strides are made, this system possesses the potential to profoundly ameliorate the quality of STEM education by delivering scalable and personalized formative assessment experiences.

REFERENCES

[1] L. Razzaq *et al.*, "The Assistment project: Blending assessment and assisting," *Proceedings of the 12th Annual Conference on Artificial Intelligence in Education*, no. 0231773, 2005.

[2] K. R. Koedinger and J. R. Anderson, "Illustrating Principled Design: The Early Evolution of a Cognitive Tutor for Algebra Symbolization," *Interactive Learning Environments*, vol. 5, no. 1, pp. 161–179, Mar. 1998, doi: 10.1080/1049482980050111.

[3] J. Figueiredo and F. J. García-Peñalvo, "Intelligent Tutoring Systems approach to Introductory Programming Courses," in *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, New York, NY, USA: ACM, Oct. 2020, pp. 34–39. doi: 10.1145/3434780.3436614.

[4] A. Mitrovic, "An intelligent SQL tutor on the web," *Int J Artif Intell Educ*, vol. 13, no. 2–4, 2003.

[5] E. R. Sykes and F. Franek, "A Prototype for an Intelligent Tutoring System for Students Learning to Program in Java™," *Advanced Technology for Learning*, vol. 1, no. 1, 2004, doi: 10.2316/Journal.208.2004.1.202-1454.

[6] D. Weragama and J. Reye, "The PHP intelligent tutoring system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7926 LNAI, pp. 583–586, 2013, doi: 10.1007/978-3-642-39112-5_64/COVER.

[7] M. L. Swartz and M. Yazdani, Eds., *Intelligent Tutoring Systems for Foreign Language Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992. doi: 10.1007/978-3-642-77202-3.

[8] V. Slavuj, B. Kovacic, and I. Jugo, "Intelligent tutoring systems for language learning," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, May 2015, pp. 814–819. doi: 10.1109/MIPRO.2015.7160383.

[9] A. Ferreira and J. Atkinson, "Designing a feedback component of an intelligent tutoring system for foreign language," *Knowl Based Syst*, vol. 22, no. 7, pp. 496–501, Oct. 2009, doi: 10.1016/j.knosys.2008.10.012.

[10] M. W. Evens *et al.*, "CIRCSIM -Tutor : An Intelligent Tutoring System Using Natural Language Dialogue," *Cogn Sci*, 2001.

[11] S. Salah and M. Thabet, "E-Learning Management Systems- A Feature-based Comparative Analysis," *Journal of Information Systems and Technology Management*, vol. 18, p. e202118003, May 2021, doi: 10.4301/S1807-1775202118003.

[12] Y. Zhang, A. Ghandour, and V. Shestak, "Using Learning Analytics to Predict Students Performance in Moodle LMS," 2020, Accessed: Jun. 20, 2023. [Online]. Available: https://digitallibrary.aau.ac.ae/handle/123456789/434

[13] R. Raju, S. Bhat, S. Bhat, R. D'Souza, and A. B. Singh, "Effective Usage of Gamification Techniques to Boost Student Engagement," *Journal of Engineering Education Transformations*, vol. 34, no. 0, p. 713, Jan. 2021, doi: 10.16920/jeet/2021/v34i0/157171.

[14] A. Z. Mansor, "Managing Student's Grades and Attendance Records using Google Forms and Google Spreadsheets," *Procedia Soc Behav Sci*, vol. 59, pp. 420–428, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.296.

[15] C. H. Ng, "MICROTEACHING VIA ZOOM IN PRESERVICE ENGLISH LANGUAGE TEACHER TRAINING: TECHNOLOGY, PEDAGOGY, AND CONTENT KNOWLEDGE," in *EDULEARN22 Proceedings*, IATED, Jul. 2022, pp. 2078–2084. doi: 10.21125/edulearn.2022.0541.

[16] V. R. Manipatruni and Dr. N. S. Kumar, "The Efficacy of Edtech Tools In Higher Education Online for The Provision of Quality Education in India During Covid-19 Pandemic," *EPRA International Journal of Research and Development (IJRD)*, vol. 7, no. 3, pp. 91–95, Mar. 2022, doi: 10.36713/epra2016.

[17] O. F. Ipek and Z. Ustunbas, "Applications Used in Distance Education for English Language Classrooms," *JET (Journal of English Teaching)*, vol. 7, no. 3, pp. 260–272, Oct. 2021, doi: 10.33541/jet.v7i3.3027.

[18] O. F. Ipek and Z. Ustunbas, "Applications Used in Distance Education for English Language Classrooms," *JET (Journal of English Teaching)*, vol. 7, no. 3, pp. 260–272, Oct. 2021, doi: 10.33541/jet.v7i3.3027.

[19] S. Lloyd, M. Beckman, D. Pearl, R. Passonneau, Z. Li, and Z. Wang, "Foundations for AI-Assisted Formative Assessment Feedback for Short-Answer Tasks in Large-Enrollment Classes," in *Bridging the Gap: Empowering and Educating Today's Learners in Statistics. Proceedings of the Eleventh International Conference on Teaching Statistics*, International Association for Statistical Education, Dec. 2022. doi: 10.52041/iase.icots11.T3C3.

[20] Sein Minn, "AI-assisted knowledge assessment techniques for adaptive learning environments," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100050, Jan. 2022, doi: 10.1016/j.caeai.2022.100050.

[21] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[22] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A Bibliometric Review of Large Language Models Research from 2017 to 2023," Apr. 2023, Accessed: Jun. 17, 2023. [Online]. Available: http://arxiv.org/abs/2304.02020

[23] K. Schildkamp, F. M. van der Kleij, M. C. Heitink, W. B. Kippers, and B. P. Veldkamp, "Formative assessment: A systematic review of critical teacher prerequisites for classroom practice," *Int J Educ Res*, vol. 103, p. 101602, Jan. 2020, doi: 10.1016/j.ijer.2020.101602.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Biased Mixed-Precision Convolution Engine for Hardware-Efficient Computational Imaging CNN

Hao-Jiun Tu*, Yu-Feng Ou, Yong-Tai Chen, and Chao-Tsung Huang
Department of Electrical Engineering
National Tsing Hua University, Hsinchu, Taiwan
* E-mail: tuj0419@gapp.nthu.edu.tw

***Abstract*—Mixed-precision networks have demonstrated significant advantages in CNN acceleration for edge inference. They can achieve high throughput and energy efficiency while maintaining quality in resource-constrained environments. However, existing high-performance accelerators for bit-flexible computation suffer from significant hardware overheads due to the extra sign extension in bit-level decomposition, severely confining the performance enhancement in mixed-precision networks. In this paper, we aim to facilitate efficient inference for mixed-precision networks in computational imaging applications. First, we propose a biased multiplier to optimize hardware performance for bit-flexible architectures. Unlike the conventional bit-level decomposition, the proposed biased representation (BR) avoids sign extension to enhance hardware efficiency. On top of that, we build 4/8-bit biased mixed-precision networks and convolution engines with highly-parallel eCNN accelerator for evaluation. We implement convolution engines with three types of multipliers, uniform-precision, sign-extended, and biased multipliers, on TSMC 40 nm technology to support advanced super-resolution and denoising at 4K UHD 30fps, FHD 60fps and FHD 30fps. Compared to the sign-extended counterpart, synthesis results show that our biased mixed-precision convolution engine can provide area overhead and energy reduction by 50%. At the same level of image quality, it can also deliver 33% energy reduction on average over the uniform-precision method.**

I. Introduction

Convolutional neural networks (CNNs) have demonstrated exceptional capability in computational imaging applications, such as super-resolution (SR) [13], [15], [17] or denoising (Dn) [5], [25], [26]. The CNN-based methods aim to generate latent images from given low-resolution or noisy images by well-learned filters, and can deliver superior quality than traditional algorithms. However, high computational demands and energy burden for CNNs make it hard to deploy on edge devices. Therefore, reducing model complexity while maintaining quality becomes a critical issue for edge CNN inference.

Depth reduction and quantization are commonly adopted to reduce complexity for CNNs. Recent evidence [7], [9], [23] indicates that the model depth has a great impact on the performance of both recognition and image restoration tasks. As a result, sufficient depth for a compact model is essential to maintain quality on edge devices. As for quantization, early works [2], [11], [12], [27] employ uniform-precision quantization for all parameters/activations across CNN layers to further reduce complexity. However, computational imaging CNNs aim to render fine-textured and high-precision feature maps. In this case, the precision of MAC operations could be at least 8-bit for representing a sufficient dynamic range. For example, hardware accelerators optimized for computational imaging use 8-bit precision to maintain image quality for hardware implementation, such as eCNN [10], RingCNN [9] and DepFiN [14]. Other frameworks like Diffy [19] and IDEAL [18] employ higher precision configurations. Thus, uniform-precision quantization restricts the potential for complexity reduction in computational imaging applications.

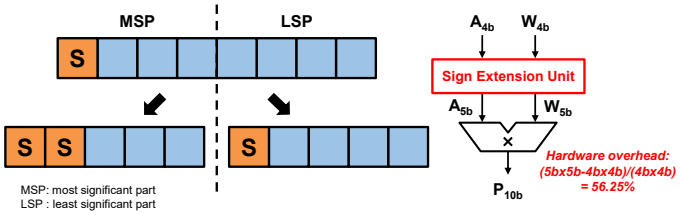


Fig. 1. Conventional bit-level decomposition with sign extension for two's complement 8-bit data (left). Hardware overheads of a sign-extended multiplier in 4-bit granularity (right).

An alternative for addressing this issue is to utilize mixed-precision quantization. Previous studies [4], [20] have shown that parameters/activations exhibit distinct properties and distributions across layers. Each layer may vary in quantization level and accordingly require different precision allocation. By exploiting the different quantization sensitivity [3], [16], [24] of each layer, the precision can be further reduced while mitigating quality degradation for non-sensitive layers within the network. This approach enables better tradeoffs between complexity and quality for mixed-precision networks.

Recently, advanced bit-flexible architectures have been developed to support various bitwidth configurations for energy-efficient inference. However, most bit-flexible accelerators, such as Bit Fusion [22], Bitblade [21] and HNPU [6], suffer from significant hardware overheads due to the conventional bit-level decomposition method as shown in Fig. 1. Since two's complement data after the decomposition is either signed or unsigned, a sign extension unit is required for signed MAC operations. The increased bitwidth on the most significant bit (MSB) of the decomposed data could produce significant hardware overheads on multipliers, as well as the subsequent accumulators and pipeline registers. These overheads defy the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

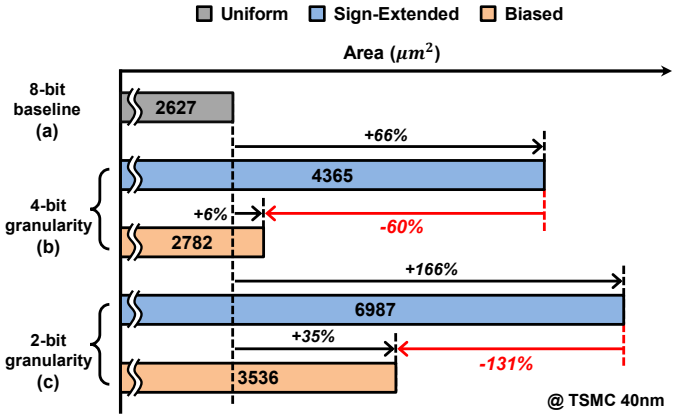


Fig. 2. Area comparison for three types of multipliers: uniform-precision, sign-extended, and the proposed biased multiplier. The computing capacity of above multipliers is the same as an 8-bit multiplication but with different granularity. (a) Baseline 8-bit. (b) Coarse-grained 4-bit. (c) Fine-grained 2-bit.

advantage of improved computational efficiency for mixed-precision networks. Therefore, an efficient hardware structure and decomposition method is crucial to facilitate optimizations of hardware performance.

In this work, we propose a biased multiplier to optimize hardware performance for bit-flexible architectures. Specifically, the proposed bit-level decomposition method, called biased representation (BR), converts the unsigned decomposed data into the signed one to fully utilize the dynamic range of two's complement input data. It can thus avoid sign extension to eliminate bitwidth-increasing overheads. Then, we build coarse-grained 4/8-bit biased mixed-precision networks and convolution engines for evaluation. We adopt eCNN [10], the state-of-the-art accelerator for computational imaging, as our engine backbone. With TSMC 40 nm technology, we implement convolution engines with three types of multipliers, uniform-precision, sign-extended, and biased multipliers, to show effectiveness on SR and Dn tasks at 4K UHD 30fps (UHD30), FHD 60fps (HD60), and FHD 30fps (HD30). Compared to the sign-extended counterpart, our biased mixed-precision convolution engine can provide area overhead and energy reduction by 50%. Moreover, it can also deliver 33% energy reduction on average over the uniform-precision method at the same level of image quality.

II. Preliminaries

A. Hardware Overhead

To accurately examine hardware overheads physically, we synthesized three types of multipliers, uniform-precision, sign-extended, and the proposed biased multiplier, on TSMC 40 nm technology as shown in Fig. 2. In comparison with an 8-bit uniform-precision multiplier, the conventional sign-extended multiplier in 4-bit granularity incurs 66% area overhead. However, our proposed biased multiplier only has 6% area overhead, providing a reduction of the area overhead by 60%, as illustrated in Fig. 2 (b). In addition to fine-grained 2-bit in Fig. 2 (c), the biased multiplier can further reduce the area overhead by 131%. This demonstrates that the biased multiplier can provide a more hardware-efficient computation for bit-flexible architectures. The key idea and details of the biased multiplier will be elaborated in Section III.

B. Bit-Level Complexity

Since the custom precision is allocated for parameters/activations across layers in mixed-precision networks, the number of MAC operations fails to exactly reflect the model complexity. Therefore, UNIQ [1] proposed the BOPs metric quantifying the number of bit operations for MAC units. Given that multiplications substantially dominate the hardware resources in MAC units, we define the metric $bitOps$ that quantifies the bit-level complexity focusing on multiplications:

$$bitOps = \sum_{l} (b_W^{(l)} h_W^{(l)} w_W^{(l)} ch_W^{(l)}) \times (b_A^{(l)} h_A^{(l)} w_A^{(l)} ch_A^{(l)}), \quad (1)$$

where l denotes the index of the layer. b_W and b_A denote the bitwidth of weight and input activation. h_W, w_W and ch_W denote the height, width and channel of weight. h_A, w_A and ch_A denote the height, width and channel of output activation.

Section I describes that the conventional bit-level decomposition induces significant hardware overheads, which can be evaluated by the metric $bitOps$, as illustrated in Fig. 1. For a 4-bit multiplier, the bit-level complexity is the product of the bitwidths of two operands ($b_{A_{4b}} \cdot b_{W_{4b}}$). The same calculation is applied to a 5-bit multiplier ($b_{A_{5b}} \cdot b_{W_{5b}}$). Therefore, a sign-extended multiplier in 4-bit granularity can exhibit 56.25% complexity overhead. In addition to 2-bit granularity, the complexity overhead increases to 125%. In the following sections, we will use the metric to evaluate the model complexity.

III. Biased Multiplier

In this section, we elaborate on the hardware-software co-design strategy for the biased multiplier. We first demonstrate how the proposed biased representation (BR) decomposes the data into two bit-slices, and how it is mapped to the hardware for reconfigurable computation. Then, we devise a BR-aware training process to optimize the model to fit the biased multiplier.

A. Biased Representation

The conventional bit-level decomposition method splits the data into the signed most significant part (MSP) and the unsigned least significant part (LSP) as shown in Fig. 1. To support reconfigurable computation, the sign extension is typically applied on the MSP and LSP of data for signed MAC operations. In contrast, the proposed BR decomposes the data without the extra sign extension as shown in Fig. 3. For example, the two's complement 8-bit data 10001100_2 (-116_{10}) can be divided into the MSP 1000_2 (-128_{10}) and LSP 1100_2 (12_{10}). Since the original unsigned bit $(+8)$ in the LSP is replaced by the sign bit (-8), it is required to compensate the offset $(+16)$ in the MSP when the sign bit of the LSP is 1_2. The MSP of data is then converted from 1000_2

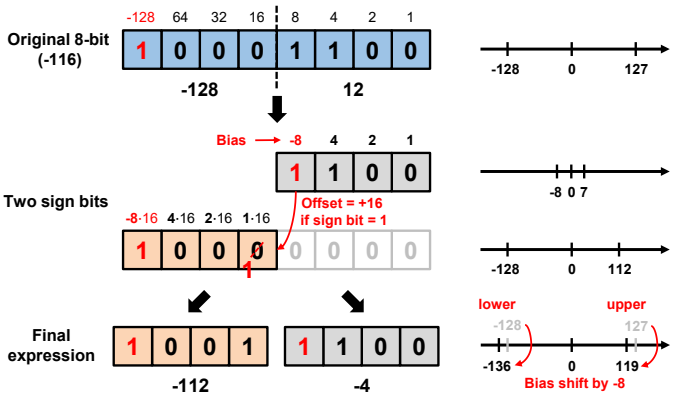


Fig. 3. Biased representation (BR) for decomposing two's complement 8-bit data into two 4-bit bit-slices (left) and the corresponding dynamic range (right).

to 1001_2 by adding 1_2. Thus, the 8-bit data 10001100_2 can be decomposed into the MSP 1001_2 (-112_{10}) and LSP 1100_2 (-4_{10}). By utilizing the BR, 8-bit data can be fitted into 4-bit signed multipliers because both decomposed MSP and LSP are signed.

However, the BR introduces a bias shift by -8 in signed 8-bit dynamic range, as illustrated in Fig. 3 on the right. With the sign bit replaced in the LSP of data, the dynamic range shifts from the original range of $[-128, 127]$ to $[-136, 119]$. As a result, the overflow issue occurs within the original range of $[120, 127]$, making it fail to represent the data in the BR. To address the issue, we propose the BR-aware training strategy, which will be discussed in Section III-C.

In general, the BR can apply to any two's complement data. The two's complement representation of an N-bit integer A can be defined as follows:

$$A = -a_{N-1}2^{N-1} + \sum_{i=0}^{N-2} a_i 2^i, \tag{2}$$

where $N \mid 2 \in \mathbb{N}$, $a_i \in \{0, 1\}$ for $i = N-1, N-2, \ldots, 0$, and a_{N-1} denotes the sign bit. To decompose A into two $N/2$-bit integers,

$$\begin{aligned} A = &- a_{N-1}2^{N-1} + a_{N-2}2^{N-2} + \cdots + a_{N'}2^{N'} \\ &+ a_{N'-1}2^{N'-1} + a_{N'-2}2^{N'-2} + \cdots + a_0 2^0, \end{aligned} \tag{3}$$

where $N' = N/2$. Equation (3) is extended by adding and subtracting $2a_{N'-1}2^{N'-1}$:

$$\begin{aligned} A = &\underbrace{-a_{N-1}2^{N-1} + a_{N-2}2^{N-2} + \cdots + a_{N'}2^{N'}}_{Most\ Significant\ Part} + \underbrace{a_{N'-1}2^{N'}}_{Offset} \\ &\underbrace{-a_{N'-1}2^{N'-1} + a_{N'-2}2^{N'-2} + \cdots + a_0 2^0}_{Least\ Significant\ Part}. \end{aligned} \tag{4}$$

Therefore, the BR can decompose an N-bit signed integer into two $N/2$-bit signed integers with the additional offset in the MSP integer.

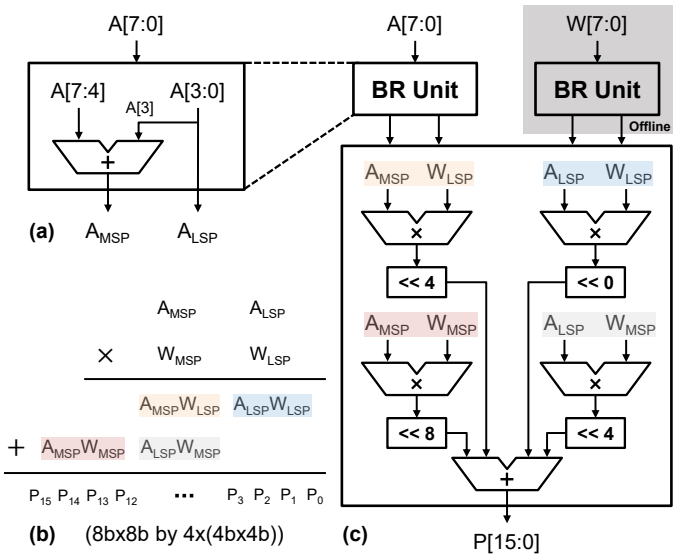


Fig. 4. Concept and structure for reconfigurable biased multiplication. (a) Schematic of a BR unit. (b) Column method for an 8-bit multiplication by using four 4-bit multiplications. (c) Overall structure for an 8-bit multiplication through biased multipliers.

B. Reconfigurable Computation for Biased Multiplication

To illustrate how biased multipliers compose the multiplication operation with 4/8-bit mixed-precision, the following discussion uses an 8-bit multiplication as an example. Fig. 4 presents the concept and structure for reconfigurable biased multiplication between two operands weight (W) and activation (A). First, a BR unit is responsible to convert the representation in bit-level decomposition, and it only consists of an adder for the 4-bit $A[7:4]$ and 1-bit $A[3]$ operands to generate the MSP of data A_{MSP}. In particular, the representation conversion for weight can be calibrated offline to reduce the hardware cost. In addition, Fig. 4 (b) also illustrates the mathematical method for an 8-bit multiplication, which can be decomposed into four 4-bit multiplications. These decomposed multiplications are then mapped to biased multipliers to generate partial products. The final result is obtained by summing all partial products through shift-and-add operations. Similarly, we can also perform 4-bit $\times$ 8-bit and 8-bit $\times$ 4-bit multiplications by using 4-bit multiplications through similar approaches as described above. Therefore, we can construct the reconfigurable biased mixed-precision convolution engine.

C. BR-Aware Training

Section III-A mentions that the BR introduces a bias shift by -8 in signed 8-bit dynamic range, leading to the overflow issue within the original range of $[120, 127]$. To avoid degrading the dynamic range through direct clipping, we propose the BR-aware training strategy to maintain quality. Fig. 5 illustrates an example of a 4/8-bit biased mixed-precision model. The dynamic range alignment is performed to preserve the full dynamic range. First, the signed 8-bit model input is shifted by -8 to align with the dynamic range of $[-136, 119]$ for the

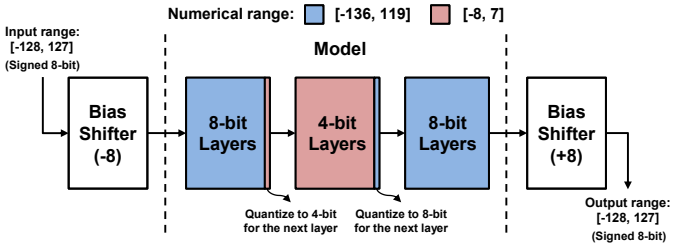


Fig. 5. An example of BR-aware training strategy for a 4/8-bit biased mixed-precision model. Assume that the activation precision is 4-bit in the middle of the model, while the rest of the model remains at 8-bit.

TABLE I
HARDWARE CONFIGURATIONS.

System		Real-Time Computation	
Technology	TSMC 40nm	UHD30	4K UHD 30fps (2.11 MbitOps/pixel)
Frequency	200 MHz	HD60	Full HD 60fps (4.21 MbitOps/pixel)
Performance	524.3 TbitOps (pure)	HD30	Full HD 30fps (8.43 MbitOps/pixel)
Computing Capacity	1. Uniform-Precision : 40960 MAC_{8x8} 2. Sign-Extended : 163840 MAC_{5x5} 3. Biased : 163840 MAC_{4x4}		

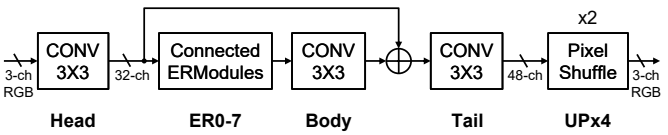


Fig. 6. SR4ERNet-B8R3N6 for four-times SR and the corresponding notation for each block.

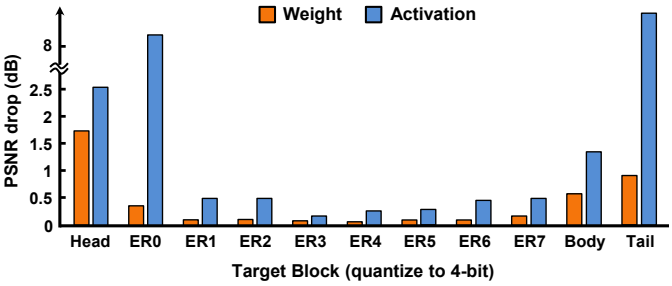


Fig. 7. PSNR drop of quantizing SR4ERNet-B8R3N6 weight and activation individually of each target block. The target block is quantized to 4-bit, while the others remain at full-precision.

BR. In addition, the numerical range for dynamic fixed-point quantization must be guaranteed within $[-8, 7]$ for 4-bit and $[-136, 119]$ for 8-bit in each layer during quantization-aware training. Note that the numerical range in Fig. 5 only focuses on activation layers, while the numerical range for parameters must also be considered. In the end, the model output data is shifted back by $+8$ to restore the original dynamic range. With the BR-aware training, a biased mixed-precision model can be optimized to fit the biased multiplier.

IV. Evaluation

The hardware configurations of our implementation are listed in Table I. The computation for three real-time specifications, UHD30, HD60, and HD30, aims to constrain the model complexity of uniform/mixed-precision networks. In the following, we will present our approaches for building mixed-precision networks and determining layerwise precision with a heuristic-based method. Then, we will show extensive evaluations for image quality and hardware performance on 4/8-bit mixed-precision networks with our proposed biased multipliers.

A. Experiment Setup

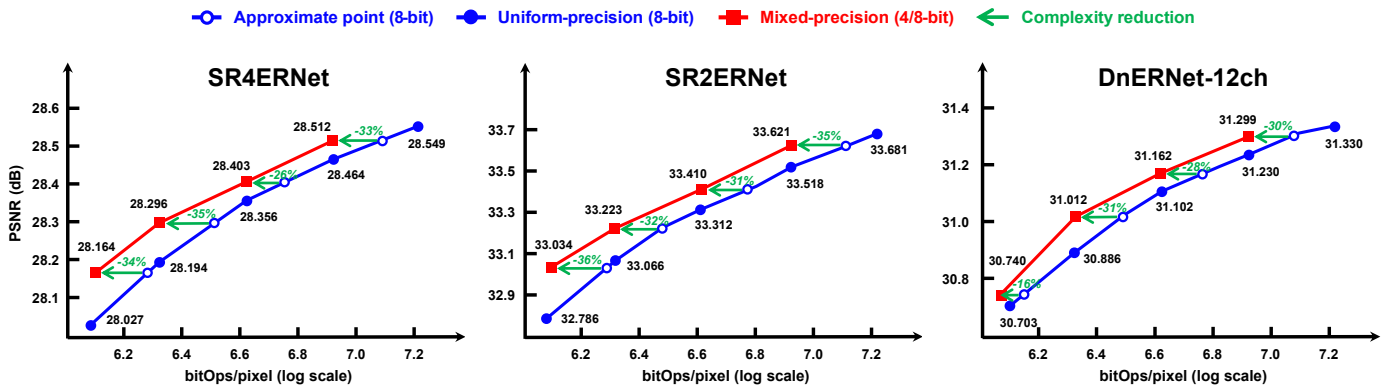
Model structure. For high-quality computational imaging, we adopt the advanced ERNets for eCNN [10] as backbone models. They can provide similar image quality compared to the well-known SRResNet [15] for SR and FFDNet [26] for Dn, and possess additional advantages on hardware implementation [10]. Fig. 6 shows the ERNet model SR4ERNet for SRx4 and its corresponding notation for each block. There are three hyperparameters for model construction in ERNet: ERModule number B, base pumping ratio R, and additional pumping layer number N. The basic building block, ERModule, expands the model width by R times by a CONV3x3 layer and reduces the width back by a following CONV1x1 layer. We use the similar model structure to build SR2ERNet and DnERNet-12ch [10] for SRx2 and Dn, respectively.

Sensitivity-Aware Layerwise Precision. In this work, we provide a heuristic-based method to determine layerwise precision for coarse-grained 4/8-bit mixed-precision networks. We employ post-training quantization for the pretrained floating-point model to sequentially quantize the weight and activation of the target block into 4-bit, while the others remain at full-precision as shown in Fig. 7. Note that the model input and output activation remain at 8-bit to approximate. Empirically, the performance drop of Head, ER0, Body and Tail in ERNets are higher than the others. Therefore, we define these four blocks as quantization sensitive blocks and maintain the weight and activation at 8-bit in the following experiments. For clarity, the uniform-precision 8-bit for weight and activation is denoted by W8A8, while the mixed-precision 4/8-bit is denoted by W4A8, W8A4, or W4A4.

Hardware Implementation. For hardware performance comparisons, we implement three types of multipliers, uniform-precision, sign-extended, and the proposed biased multiplier, on eCNN [10] accelerators. We implemented eCNN convolution engines in Verilog HDL and performed bit- and cycle-accurate simulation for functional verification. The verified design was then synthesized with Synopsys Design Compiler on TSMC 40 nm technology. Finally, we conducted power estimation on the synthesized design using Synopsys PrimeTime PX with signal switching activity from RTL simulation.

B. Quantitative Comparisons

Table II lists the hyperparameters used in our experiments. We train ERNet models on two datasets DIV2K and Waterloo for SR and Dn tasks. We consider three throughput targets for hardware acceleration: UHD30, HD60, and HD30. For each throughput target and application scenario, we adopt



(a) SRx4 (b) SRx2 (c) Dn

Fig. 8. Quality-complexity comparisons with uniform/mixed-precision ERNet models for SR and Dn tasks. The complexity reduction is calculated by the interpolation on the uniform-precision curve based on the same image quality of mixed-precision models.

TABLE II
ERNET TRAINING SETTINGS.

Model	Dataset	Patch	Batch	Batch #	Learning Rate
SR	DIV2K	96x96	16 Patches	600K	$4 \cdot 10^{-4} \cdot 2^{-k}$, k=0–3
Dn	Waterloo			1200K	$1 \cdot 10^{-3} \cdot 10^{-k}$, k=0–2

TABLE III
PSNR PERFORMANCE FOR UNIFORM/MIXED-PRECISION ERNET MODELS.

Fixed-Point Model	MbitOps/pixel		PSNR (dB)				
SR4ERNet	Complexity		Set5	Set14	BSD100	Urban100	Avg
B8R3N2-W8A8	UHD30	2.09	31.68	28.29	27.37	25.44	28.19
B11R3N6-W4A8		2.10	31.84	28.37	27.42	25.56	**28.30**
B13R3N3-W8A8	HD60	4.20	31.88	28.41	27.46	25.67	28.36
B17R3N14-W4A8		4.20	31.91	28.45	27.49	25.76	**28.40**
B18R3N16-W8A8	HD30	8.40	31.98	28.48	27.53	25.83	28.46
B25R4N0-W4A8		8.30	32.03	28.53	27.54	25.94	**28.51**
SR2ERNet	Complexity		Set5	Set14	BSD100	Urban100	Avg
B2R2N1-W8A8	≈UHD60	1.19	37.17	32.75	31.58	29.65	32.79
B3R2N1-W4A8		1.23	37.32	32.86	31.66	29.90	32.94
B6R2N0-W4A4		1.24	37.37	32.92	31.74	30.10	**33.03**
B4R2N1-W8A8	UHD30	2.07	37.39	32.96	31.76	30.15	33.07
B7R2N0-W4A8		2.03	37.53	33.08	31.86	30.42	**33.22**
B8R2N1-W8A8	HD60	4.12	37.59	33.13	31.92	30.61	33.31
B11R2N5-W4A8		4.11	37.65	33.23	31.97	30.78	**33.41**
B11R3N0-W8A8	HD30	8.33	37.75	33.33	32.02	30.98	33.52
B14R4N0-W4A8		8.37	37.83	31.16	32.07	33.43	**33.62**
DnERNet-12ch	Complexity		Set5	Set14	CBSD68	Avg	
B2R2N1-W8A8	≈UHD60	1.24	31.51	29.90	30.70	30.70	
B2R3N0-W4A8		1.15	31.55	29.93	30.72	30.73	
B5R2N0-W4A4		1.16	31.60	29.88	30.74	**30.74**	
B3R2N3-W8A8	UHD30	2.05	31.07	30.07	30.87	30.89	
B7R2N0-W4A8		2.10	31.84	30.21	30.99	**31.01**	
B8R2N1-W8A8	HD60	4.19	31.93	30.32	31.06	31.10	
B12R2N2-W4A8		4.15	31.99	30.38	31.12	**31.16**	
B12R2N7-W8A8	HD30	8.34	32.06	30.46	31.17	31.23	
B15R3N7-W4A8		8.33	32.12	30.55	31.23	**31.30**	(**σ = 25**)

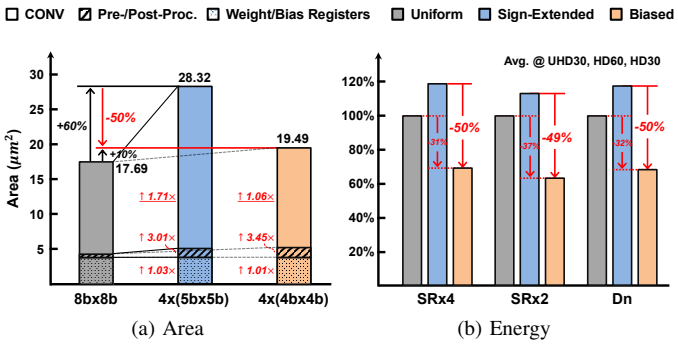
TABLE IV
SYNTHESIS PERFORMANCE FOR THREE ECNN CONVOLUTION ENGINES.

	Uniform-Precision	Sign-Extended	**Biased Multiplier**
Technology	TSMC 40nm		
Frequency	200 MHz		
Precision (W/A)	8/8	8/8, 4/8, 8/4, 4/4	
Performance	524.3 TbitOps (pure)		
	16.4 TOPS	16.4 – 65.5 TOPS	
Computing Capacity	40960 MAC_{8x8}	163840 MAC_{5x5}	163840 **MAC_{4x4}**
Area (mm²)	17.69	28.32	19.49
Power (W) (SRx4/SRx2/Dn)	2.80/2.81/2.88	4.84/4.67/4.83	2.81/**2.63/2.79**

the compact ERNet configuration for model evaluation as shown in Table III. Note that the quality difference between biased and "unbiased" mixed-precision models is minimal at the same level of model complexity, so we only show the latter in Table III. The mixed-precision models demonstrate significant gains over the uniform-precision models. Specifically, SR4ERNet-W4A8 and DnERNet-12ch-W4A8 models outperform the W8A8 models by 0.11 and 0.12 dB at UHD30, respectively. In general, the W4A8 models can show clear advantages over the W4A4 models because the activation are more sensitive to quantization than the weight as shown in Fig. 7. But for extremely low-complexity scenarios, the W4A4 models yield superior quality than the W4A8 models for SRx2 and Dn at near UHD60 due to the deeper layers. In addition, we compare quality-complexity tradeoff curves in Fig. 8. Based on the same image quality of the uniform/mixed-precision models, the mixed-precision models can provide the complexity reduction by 31%, 33%, and 30% on average for SRx4, SRx2, and Dn, respectively.

C. Hardware Performance

We evaluate the hardware performance of three types of multipliers on eCNN accelerators for comparison. Table IV summarizes the design configurations and synthesis perfor-



(a) Area (b) Energy

Fig. 9. Area and energy comparison for three eCNN convolution engines corresponded to Table IV. The eCNN convolution engines consist of convolution, pre-/post-processing units and local register files for parameters. The energy of both sign-extended and biased mixed-precision engines is based on the power consumption listed in Table IV and the complexity reduction with respect to uniform-precision counterpart in Fig. 8. It is then averaged over real-time inferences and normalized with the energy of the uniform-precision engine.

mance for three eCNN convolution engines. The eCNN convolution engines can run at 200 MHz and achieve up to 524.3 TbitOps of inference performance. Fig. 9 shows the area and energy comparison for the design. Compared to the uniform-precision counterpart, the sign-extended and biased mixed-precision convolution engines bring 60% and 10% area overhead, respectively. The proposed biased multipliers can provide the area overhead reduction by 50%. And the remaining overhead mainly arises from the datapath control and post-processing units to enhance the design throughput. On the other hand, we perform energy comparison among three engines, as shown in Fig. 9 (b), at the same level of image quality. The experiments are conducted on SR and Dn tasks in three real-time specifications: UHD30, HD60, and HD30. Based on the results, our biased mixed-precision convolution engine demonstrates higher energy efficiency than the other two engines. More specifically, our biased mixed-precision convolution engine can deliver 33% and 50% energy reduction on average over the uniform-precision and the conventional sign-extended methods.

V. Discussion

A. Mixed-Precision Quantization

There have been extensive explorations on mixed-precision quantization to achieve additional optimizations for model quality. HAWQ [3] introduces a sensitivity metric to systematically allocate quantization precision across CNN layers during quantization-aware training. HAQ [24] leverages a learning-based method that can automatically search the optimal quantization policy under different constraints. These approaches can further enhance model quality but require more computation and time. In this paper, we provide a heuristic-based method to determine layerwise precision for mixed-precision networks. It employs the post-training quantization to determine the precision according to PSNR drop of each target block, and we fix the precision during quantization-aware training to save the computation and time. Despite the non-optimal performance, the method is fast and effective to evaluate the model quality for mixed-precision networks.

B. Performance on Recognition Tasks

Unlike computational imaging CNNs, recognition CNNs demonstrate exceptional tolerance for low-precision conditions. Recent work [2] shows that 2-bit ResNet50 can even achieve less than 3% accuracy degradation on ImageNet dataset. Several bit-flexible accelerators, such as Bit Fusion [22] and Bitblade [21], are then dedicatedly optimized for recognition CNNs. Most of them employ the conventional sign-extended multipliers to construct the convolution engine. Due to the significant complexity reduction for recognition tasks, they can achieve great energy efficiency in spite of the hardware overheads. In this paper, the proposed biased multiplier can be applied on these bit-flexible accelerators to further enhance the hardware performance. Particularly in 2-bit granularity, our biased multiplier has only a 35% area overhead, whereas the sign-extended multiplier results in 166% area overhead, as illustrated in Fig. 2.

VI. Conclusion

In this paper, we present a bit-flexible convolution engine to achieve near-maximum hardware efficiency for dense CNN acceleration. The proposed biased multiplier contributes to hardware efficiency while maintaining quality for two key reasons. First, unlike the conventional sign-extended method, the BR avoids sign extension through the replaced sign bit in the unsigned decomposed data to significantly reduce hardware overheads in bit-level decomposition. Second, the BR-aware training strategy preserves the dynamic range of input data and enables model optimizations to fit the biased multiplier. Finally, we implement three eCNN convolution engines for verifying practical effectiveness. Synthesis results show that our biased mixed-precision convolution can further reduce the area overhead by 50%. Moreover, it can also deliver 33% and 50% energy reduction over the uniform-precision and sign-extended methods, respectively, for high-quality computational imaging at real-time applications. Based on these results, we believe that the biased multiplier demonstrates exceptional capability in hardware-efficient CNN inference for bit-flexible architectures.

References

[1] C. Baskin, N. Liss, E. Schwartz, E. Zheltonozhskii, R. Giryes, A. M. Bronstein, and A. Mendelson, "UNIQ: Uniform Noise njection for non-uniform quantization of neural networks," *ACM Transactions on Computer Systems (TOCS)*, vol. 37, 2021.

[2] J. Choi, S. Venkataramani, V. V. Srinivasan, K. Gopalakrishnan, Z. Wang, and P. Chuang, "Accurate and Efficient 2-bit Quantized Neural Networks," *Proceedings of Machine Learning and Systems*, vol. 1, 2019.

[3] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "HAWQ: Hessian Aware Quantization of Neural Networks with Mixed-Precision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[4] A. Elthakeb, P. Pilligundla, F. Mireshghallah, A. Yazdanbakhsh, S. Gao, and H. Esmaeilzadeh, "ReLeQ: An Automatic Reinforcement Learning Approach for Deep Quantization of Neural Networks," in *NeurIPS ML for Systems workshop*, 2019.

[5] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics (ToG)*, vol. 35, 2016.

[6] D. Han, D. Im, G. Park, Y. Kim, S. Song, J. Lee, and H.-J. Yoo, "HNPU: An Adaptive DNN Training Processor Utilizing Stochastic Dynamic Fixed-Point and Active Bit-Precision Searching," *IEEE Journal of Solid-State Circuits*, vol. 56, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] C.-T. Huang, "ERNet Family: Hardware-Oriented CNN Models for Computational Imaging Using Block-based Inference," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] C.-T. Huang, "RingCNN: Exploiting Algebraically-Sparse Ring Tensors for Energy-Efficient CNN-Based Computational Imaging," in *Proceedings of the 48th Annual ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2021.

[10] C.-T. Huang, Y.-C. Ding, H.-C. Wang, C.-W. Weng, K.-P. Lin, L.-W. Wang, and L.-D. Chen, "eCNN: A Block-Based and Highly-Parallel CNN Accelerator for Edge Inference," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2019.

[11] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations," *The Journal of Machine Learning Research*, vol. 18, 2017.

[12] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] G. Koen and V. Marian, "DepFiN: A 12nm, 3.8 TOPs depth-first CNN processor for high res. image processing," in *2021 IEEE Symposium on VLSI Circuits*, 2021.

[15] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] D. Lee, M. Cho, S. Lee, J. Song, and C. Choi, "Data-Free Mixed-Precision Quantization Using Novel Sensitivity Metric," *arXiv preprint arXiv:2103.10051*, 2021.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[18] M. Mostafa, Z. Bojian, L. A. Delmás, H. Felix, A. Jonathan, B. Paul, O. Emmanuel, and M. Andreas, "IDEAL: Image DEnoising AcceLerator," in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2017.

[19] M. Mostafa, S. Kevin, and M. Andreas, "Diffy: A Déjà vu-free Differential Deep Neural Network Accelerator," in *Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2018.

[20] E. Park, S. Yoo, and P. Vajda, "Value-Aware Quantization for Training and Inference of Neural Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[21] S. Ryu, H. Kim, W. Yi, and J.-J. Kim, "Bitblade: Area and Energy-Efficient Precision-Scalable Neural Network Accelerator with Bitwise Summation," in *Proceedings of the 56th Annual Design Automation Conference (DAC)*, 2019.

[22] H. Sharma, J. Park, N. Suda, L. Lai, B. Chau, V. Chandra, and H. Esmaeilzadeh, "Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Network," in *Proceedings of the 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "HAQ: Hardware-Aware Automated Quantization with Mixed Precision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, 2017.

[26] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a Fast and Flexible Solution for CNN-Based Image Denoising," *IEEE Transactions on Image Processing*, vol. 27, 2018.

[27] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients," *arXiv preprint arXiv:1606.06160*, 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Lightweight Speaker Verification Model For Edge Device

Ting-Wei Chen*, Chia-Ping Chen*,
Chung-Li Lu†, Bo-Cheng Chan†, Yu-Han Cheng†, Hsiang-Feng Chuang†, Wei-Yu Chen†
* National Sun Yat-sen University, Taiwan
E-mail: m103040017@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw
† Advanced Technology Laboratory, Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan
{chungli, cbc, henacheng, gotop, weiweichen}@cht.com.tw

***Abstract*—In this paper, we investigate lightweight models for automatic speaker verification (ASV) that can be feasibly implemented on edge devices. The proposed model architecture includes an improved and re-parameterized VGG-based front-end stem (RepVGG) and an ECAPA-TDNNLite-like backbone. We refer the proposed model by IM ECAPA Rep-TDNNLite. We further apply the decoupled knowledge distillation approach to improve the performance of the lightweight model. On the open VoxCeleb1-O test set, the proposed IM ECAPA Rep-TDNNLite achieves 1.76% equal error rate (EER) and 0.13 minimum decision cost function (minDCF) with 1.31M model parameters and 1.45G FLOPs. The performance can be improved to 1.66% EER and 0.12 minDCF by a different RepVGG front-end with 1.46M parameters and 2.67G FLOPs. Thus, the proposed lightweight model is capable of providing good performance on edge devices with limited computation resources.**

I. Introduction

Automatic speaker verification (ASV) is an identity authentication technology that has been receiving increasing attention in recent years. Modern ASV systems often convert the acoustic features of speech into the corresponding speaker embeddings for reliable verification. It can protect personal data from being forged and stolen by comparing the pre-enrollment speaker embedding with an unknown one.

Effective recognition models from the image processing field can be applied in the field of spoken language technology. In particular, such models have made very strong impact on ASV recently. With 1-D convolution (TDNN) as the backbone, ECAPA-TDNN [1], which introduced the multi-branch structure [2] and attention mechanism [3], has occupied top rankings in many competitions since its launch [4]–[6]. With 2-D convolution, ResNet [7] and RepVGG [8], [9] also work well in several ASV evaluations [9]–[11]. As both model architectures are capable of performing well in competitions, models combining the two architectures have been developed. ECAPA CNN-TDNN [12] connects the ResNet structure to ECAPA-TDNN as a front-end module, further improving system performance. CS-Rep [13] incorporates the re-parameterization method [8] and adjusts the fusion operation, and lets VGG-like TDNN structures achieve state-of-the-art performance.

Since the above models were introduced to ASV competitions, many variants have been introduced by adding or modifying different structures. Most of these models were, however, enhanced through the use of more complex convolution operations or deeper neural networks. In practice, these models are too complex to deploy on edge devices such as mobile phones. As a result of the hardware limitations on edge devices, it is important to design models with low complexity while maintaining sound performance level.

In this paper, we begin with IM ECAPA CNN-TDNN [14], which is an improved version of ECAPA CNN-TDNN, as baseline. We design a lightweight model by combining the baseline with dimensionality reduction and separable convolution [15]. Variants of RepVGG-Blocks [9] are applied to replace the original convolution layers inside the front-end module. We call this model IM ECAPA Rep-TDNNLite. Furthermore, we experiment knowledge distillation approaches to further improve the performance of the lightweight model. After training with knowledge distillation, the performance of the model on each dataset is further improved. Note that knowledge distillation is a key component as the final system outperforms other recent works without the assistance of teacher models.

The remainder of this paper is organized as follows. Section II introduces our complex and lightweight models in detail. Section III describes the knowledge distillation employed in our experiments. Section IV specifies the experimental setup and evaluation protocol. Section V presents the experimental results and associated discussion. Concluding remarks are summarized in Section VI.

II. Models

A. IM ECAPA CNN-TDNN

IM ECAPA CNN-TDNN is an improved version of ECAPA CNN-TDNN that we developed in previous study [14]. The model architecture is shown in Figure 1. The refinements are summarized as follows.

- **Adjust the neural networks of residual blocks in 2D-CNN stem to the pre-activation version.** According to the study in [16], any operation on the shortcut connections of residual blocks would lead to diminished performance of neural networks; on the other hand, switching the neural networks' sequence of residual blocks from

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

post-activation to pre-activation could increase the generalization ability of models and improve its training speed.

- **Replace Res2Block in ECAPA-TDNN with SC-Block.** SC-Block is the convolution block of Self-Calibration Network (SCNet) [17], which integrates the concepts of Res2Net and SKNet [18]. With its own self-calibration mechanism and segmentation convolution similar to Res2Block, SC-Block obtains a large receptive field and captures spatial information. This enables SC-Block to extract more detailed feature information than Res2Block.
- **Add an aggregation layer after every SC-Block.** These aggregation layers concatenate the previous output of different scales to perform feature reconstruction. They are combined with the multi-layer aggregation method in the original ECAPA-TDNN to create a hierarchical aggregation structure. At the network head (top), more features of different scales would be fused to extract speaker embeddings with more detailed information.
- **Change the attention mechanism from SE to CBAM.** SE is an attention mechanism module from Squeeze-and-Excitation Network (SENet) [3]. It only focuses on the channel dimension, while the spatial dimension still has significant information to exploit. Therefore, we introduce CBAM (convolution block attention module) [19] as a replacement for SE. Comparatively to SE, CBAM also takes spatial attention into account. By integrating channel attention and spatial attention, the 2D CNN stem could reconstruct the audio features in a more detailed manner.

IM ECAPA CNN-TDNN is a complex model with high model capacity. A trained IM ECAPA CNN-TDNN model is used as a teacher model in knowledge distillation for the proposed lightweight model, which is introduced next.

B. IM ECAPA Rep-TDNNLite

The proposed lightweight IM ECAPA Rep-TDNNLite model is based on the complex model aforementioned in Section II-A. The complete network topology is shown in Fig. 2. The modified parts are described as follows.

- **Reduce the number of convolution channels in the backbone structure.** Since this is a lightweight model, it is very intuitive to reduce the number of parameters and floating point operations (FLOPs) by reducing the number of convolution channels. Furthermore, we set the stride of the first convolution layer in the backbone structure to 2 according to ECAPA-TDNNLite [15]. This setting reduces computation by 50% while decreasing performance only slightly.
- **Replace SC-Block to separable Res2Block.** Based on ECAPA-TDNNLite, we replace the convolution blocks SC-Block inside the backbone structure with a separable Res2Block to further reduce the number of parameters.
- **Replace the 2D-CNN stem to a RepVGG-based front-end module.** As the front-end module of the complex model in Section II-A uses only the basic residual block structure, we replace it with RepVGG-Block [8]. With RepVGG, the model takes advantage of multi-branch convolution during training. In the inference stage, the model could switch to one-way convolution through re-parameterization, reducing parameters and FLOPs effectively while maintaining performance. The structure of RepVGG-Block has been experimented on speaker verification tasks [9]. Two of the RepVGG-Block variants designed in [9], RSBA and RSBB (Rep Speaker Block A & B), are implemented and evaluated in our experiments. These two structure are shown in the III part of Fig. 2.

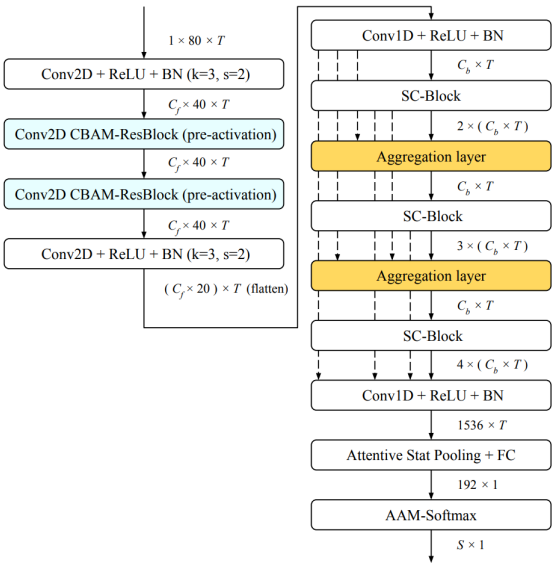


Fig. 1. Network topology of IM ECAPA CNN-TDNN. The left half is front-end module, and the right half is backbone structure. k represents kernel size, s represents stride. C_f and C_b denote the number of channels in the front-end module and the backbone structure respectively. T and S are the number of frames and speakers. Lines with dashes ignore the middle layer of the path and send the data directly to the specified layer.

In our experiments, we compare training alone and training with knowledge distillation on this lightweight model to investigate the impact of distillation on model performance.

III. Knowledge Distillation

Knowledge distillation is a well-known method for model compression. With the assistance of dark knowledge extracted from pre-trained complex model, the output of lightweight model would be as close as possible to the output of complex model. There are two mainstream knowledge distillation methods in the speaker verification task. One is the embedding-based distillation, another is the logit-based distillation. The simplified knowledge distillation diagram in speaker verification task is illustrated in Fig. 3.

In our experiments, we use the logit-based distillation. During the training phase of speaker verification systems, the output of the classification head consists of the posterior probability for the specified set of speakers. We let the output posterior of the student model imitate the output posterior of the teacher model. Our implementation of knowledge distillation includes the traditional Kullback-Leibler divergence loss

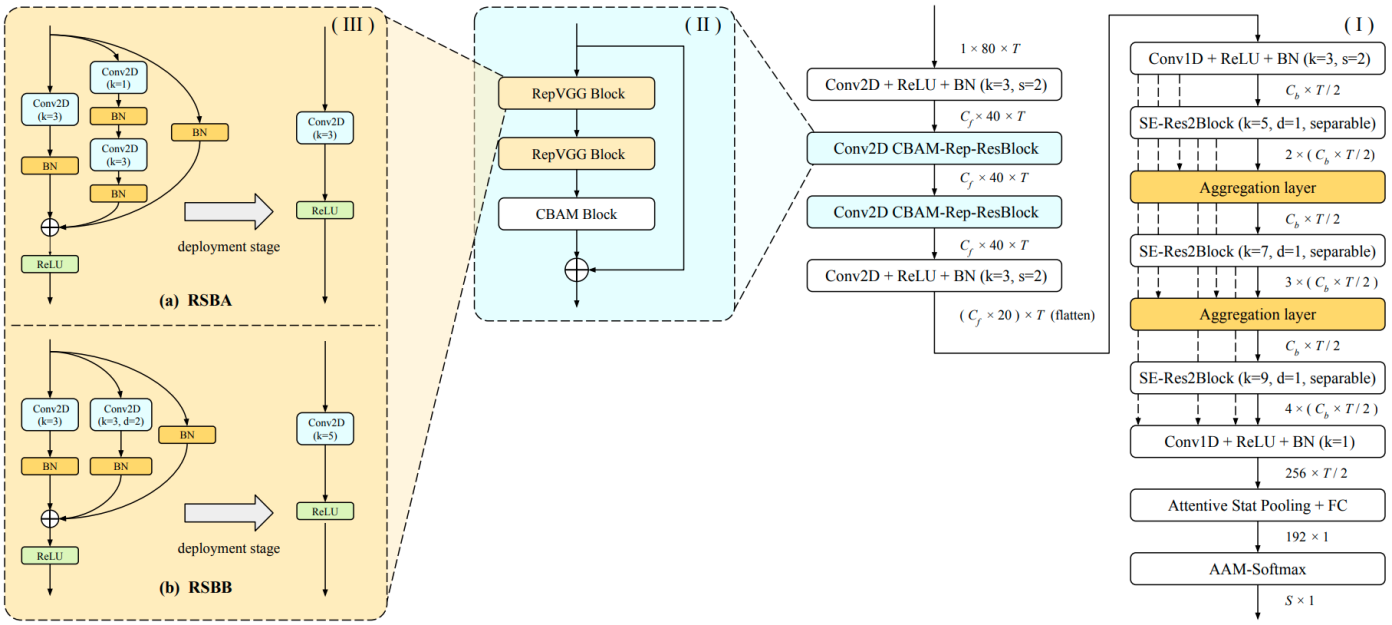


Fig. 2. Network topology of IM ECAPA Rep-TDNNLite. We expand the structural details in three levels: (I) Architecture of IM ECAPA Rep-TDNNLite, (II) Details of Conv2D CBAM-Rep-ResBlock, and (III) Details of two variant structures of RepVGG-Block. The "separable" in SE-Res2Block of (I) means we use depthwise separable convolution layers in SE-Res2Block. C_f and C_b denote the number of channels in the front-end module and the backbone structure respectively. k, s, d denote kernel size, stride, and dilation spacing. T and S correspond to the number of frames and the number of speakers. $\oplus$ denotes as element-wise addition. RSB denotes the RepSpk block from [9], and we choose two variants A & B to implement in our experiments.

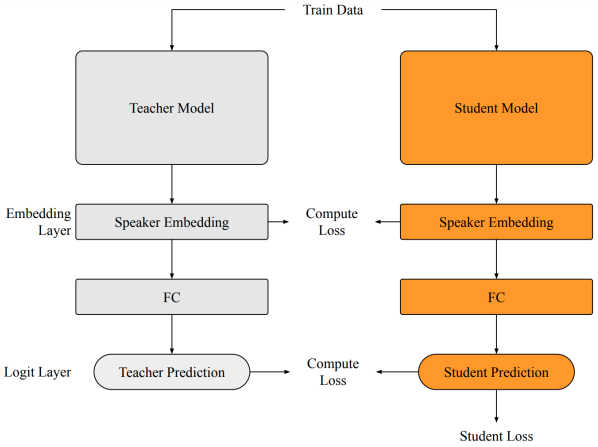


Fig. 3. Diagram of knowledge distillation for speaker verification systems. Embedding-based distillation means using the embedding layer output to compute consistency loss, while logit-based distillation uses the logit layer.

and the recently proposed decoupled knowledge distillation loss. In training stage, the distillation loss is a soft loss combined with the original student model loss (hard loss)

$$L = L_{\text{hard}} + \alpha L_{\text{soft}} \tag{1}$$

where α is a hyper-parameter for the weight of the soft loss.

A. Kullback-Leibler Divergence Loss

Kullback-Leibler divergence (KLD) measures the difference between two probability distributions. In the well-known study of knowledge distillation [20], KLD is used to calculate the distillation loss. Specifically

$$L_{\text{KLD}} = \sum_{i=1}^{N} D_{\text{KL}}\left(\tilde{\mathbf{p}}^i || \mathbf{p}^i\right) = \sum_{i=1}^{N}\sum_{j=1}^{C} \tilde{p}_j^i \log\left(\frac{\tilde{p}_j^i}{p_j^i}\right) \tag{2}$$

where N is the number of samples, C is the number of classes, $\tilde{\mathbf{p}}^i, \mathbf{p}^i$ are probability vectors, and $\tilde{p}_j^i, p_j^i$ are the probability of class j for sample i.

B. Decoupled Knowledge Distillation Loss

We also investigate decoupled knowledge distillation [21]. The loss L_{KLD} can be decomposed to two terms. One term is a binary logit loss for target classification knowledge distillation (TCKD). The other is a multi-category logit loss for non-target classification knowledge distillation (NCKD). Specifically

$$\begin{aligned} L_{\text{KLD}} &= \sum_{i=1}^{N} \text{TCKD}_i + (1 - p_t^T)\text{NCKD}_i \\ &= \sum_{i=1}^{N} D_{\text{KL}}(\mathbf{b}_t^i || \mathbf{b}_s^i) + (1 - p_t^T) D_{\text{KL}}(\hat{\mathbf{p}}_t^i || \hat{\mathbf{p}}_s^i) \end{aligned} \tag{3}$$

where s and t indicate student or teacher models, $\mathbf{b}^i$ is the binary probabilities for target classification, $\hat{\mathbf{p}}^i$ is the multi-category probabilities for non-target classification, and p_t^T is the probability of the target class predicted by teacher model.

The results of ablation experiments in [21] indicated that NCKD is the major cause affecting the distillation effect. However, the reformulated KLD in Eq. (3) shows that NCKD is constraint by weight $(1-p_t^T)$. As the teacher model predicts more accurately, i.e. the larger the p_t^T, the lower the impact of

NCKD. As a result, it is difficult to distill knowledge from the high performance model using the traditional KLD method. In order to resolve this issue, a simplified and effective method called decoupled knowledge distillation (DKD) is proposed

$$L_{\text{DKD}} = \sum_{i=1}^{N} \text{TCKD}_i + \gamma \text{NCKD}_i. \quad (4)$$

That is, $(1 - p_t^T)$ in Eq. (3) is replaced by hyper-parameter γ to simplify the balance between TCKD and NCKD.

IV. Experiments

A. Datasets

We use the development part of the VoxCeleb2 [22] dataset as our training data, which contains 1,092,009 utterances from 5994 speakers. Musan [23] corpus (additive speech, music, and noise) and RIR [24] corpus (reverb) are employed for online data augmentation. Each training step would randomly select one method for data augmentation.

B. Training Details

The input features are 80-dimensional log Mel-filterbank energies with a window length of 25 ms and a 10 ms frame-shift. In addition, the extracted features are mean-normalized without the application of voice activity detection (VAD). The features are cropped into 2-second segments and 256 segments form a mini batch. The RSBA structure can only be trained using 200 segments in a mini batch because of hardware limitations. The channel C_f is set to 64 in all front-end modules, and the channel C_b is set to 512 and 144, separately, in the backbone of complex models and lightweight models. All models are trained with an initial learning rate of 0.001, which is reduced by 25% every 10 epochs with Adam optimizer. AAM-softmax is chosen as the loss function of hard loss with a margin of $m = 0.2$ and a scale of $s = 32.0$. A weight decay of 2e-05 was applied during training to prevent the model from overfitting. The hyper-parameter α, β and γ are set to 1, 1 and 8. The temperature parameter in KLD-based loss is set to 4. The training phase of lightweight models terminates after 100 epochs.

C. Evaluation Protocol

In the experiment, the clean trials of VoxCeleb1 [25] are used as test sets. The features of test audios are cropped into 4-second segments. Equal error rate (EER) and minimum detection cost function (minDCF) are used as our evaluation metrics for measuring the performance of the systems. The computation of minDCF uses the same setting as VoxSRC 2022, with parameters $C_{\text{miss}} = 1, C_{\text{false}} = 1$, and $P_{\text{target}} = 0.05$. We did not implement any score normalization methods.

V. Results

We first present the ablation experiment results in our previous paper [14] to show the performance of IM ECAPA CNN-TDNN. After that we compare our proposed models with other recent works on lightweight models. Finally, we discuss the impact of different methods of knowledge distillation on model performance.

TABLE I
SIMPLIFIED EXPERIMENT RESULTS FROM [14].

Architecture	VoxCeleb1-O		VoxSRC2022 val	
	EER(%)	minDCF	EER(%)	minDCF
ECAPA-TDNN (Re-implemented)	1.3770	0.0931	3.6735	0.2479
IM ECAPA-TDNN	1.2600	0.0849	3.6824	0.2462
IM ECAPA CNN-TDNN				
SE module	1.0424	0.0739	34471	0.2198
GC module	1.2552	0.0992	3.4990	0.2282
CBAM module	1.1484	0.0817	3.1160	0.2053

TABLE II
COMPARISON OUR MODELS AND RECENT WORKS PERFORMANCE ON EER OF VOXCELEB1-O, THE NUMBER OF PARAMETERS AND FLOPS. THE SYMBOL '-' IN THE TABLE MEANS THAT THE REFERENCE PAPER DOES NOT PROVIDE THIS INFORMATION.

Architecture	EER(%)	#Params	FLOPs
Thin ResNet-34 [26]	2.36	1.40M	0.93G
Fast ResNet-34 [26]	2.37	1.40M	0.45G
DenseNet + global block [27]	1.94	1.20M	-
ECAPA-TDNNLite (symmetric) [15]	3.07	318K	11.6M
ECAPA-TDNNLite (asymmetric) [15]	2.31	318K	11.6M
IM ECAPA CNN-TDNNLite	1.99	1.31M	1.46G
IM ECAPA RepA-TDNNLite	1.88	1.47M	2.81G
⇒ ***deployment stage***	1.88	1.31M	1.45G
IM ECAPA RepB-TDNNLite	1.77	1.46M	2.67G
⇒ ***deployment stage***	1.77	1.57M	3.57G

A. Performance of IM ECAPA CNN-TDNN

Our previous paper compared the performance of IM ECAPA CNN-TDNN composed of different attention mechanisms. The simplified experimental results are shown in Table I. We only present the evaluation results of the simplest and most complex test sets. Note that some training details and evaluation protocol in the previous paper are different from this paper, this leads to differences in performance results of models. According to the experimental results, the IM ECAPA CNN-TDNN with CBAM module performed better on the complex test set. This shows that the model has better generalization ability, which is why we use this model as our baseline model and teacher model.

B. Performance of IM ECAPA Rep-TDNNLite

Table II compares the performance of our lightweight models and other recent lightweight models for speaker verification. The EER values were taken from the reported values in the reference papers. We first experiment with IM ECAPA CNN-TDNNLite, the lightweight version of our baseline model. It has the same backbone structure as IM ECAPA Rep-TDNNLite.The results show that this model has outperformed most other recent models. On the other hand, due to the multi-branch convolution design of the RepVGG structure, the performances of two IM ECAPA Rep-TDNNLite are even better than IM ECAPA CNN-TDNNLite. This shows that

TABLE III
PERFORMANCE OF OUR PROPOSED LIGHTWEIGHT MODELS WITH DIFFERENT KNOWLEDGE DISTILLATION METHODS. REPA (REPB) DENOTES RSBA (RSBB) STRUCTURE FOR REPVGG-BLOCKS IN FRONT-END MODULES.

Architecture	#Params	FLOPs	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
IM ECAPA CNN-TDNN (teacher)	10.04M	3.16G	1.0318	0.0786	1.2494	0.0816	2.3402	0.1447
IM ECAPA RepA-TDNNLite	1.31M	1.45G	1.8775	**0.1198**	1.9952	0.1258	3.2833	**0.1988**
+ KLD Loss	-	-	1.7864	0.1200	1.9731	0.1300	3.3519	0.2046
+ DKD Loss	-	-	**1.7551**	0.1303	**1.7682**	**0.1246**	**3.2755**	0.2134
IM ECAPA RepB-TDNNLite	1.46M	2.67G	1.7658	0.1215	1.9700	0.1242	3.2419	0.1979
+ KLD Loss	-	-	1.7658	**0.1124**	1.9024	0.1244	3.2092	**0.1918**
+ DKD Loss	-	-	**1.6647**	0.1194	**1.7175**	**0.1232**	**3.1784**	0.2112

using RepVGG-Block to replace the basic convolution layer in the front-end module can improve performance. However, the number of parameters and FLOPs also increases due to the use of multi-branch convolutions. This is where the reparameterization operation of RepVGG comes into play. By reparameterizing RepVGG-Blocks in the models, the model with the front-end module composed of RSBA structures managed to reduce 10.8% of the parameters and 48% of the FLOPs while maintaining the same performance, making the resource consumption close to other recent lightweight models. Note that it achieves better EER than the lightweight versions of ResNet-34 with a smaller model size. Using RSBB version of the front-end module will improve EER performance by over 6% in comparison to RSBA version. Note that RSBB will increase the parameters and FLOPs after reparameterization because it transforms into a 5×5 convolution layer. Therefore, RSBB should preserve its original structure and require more computing resources than RSBA.

C. Comparison between Knowledge Distillation Methods

The results of proposed systems under different knowledge distillation methods (KLD, DKD) are summarized in Table III. IM ECAPA CNN-TDNN is the teacher model, while IM ECAPA RepA/RepB-TDNNLite without knowledge distillation are two student model baselines. Comparing to baselines, both distillation methods we applied could boost the performance of the models to a certain extent. However, even though the traditional KLD method made a slight improvement on several test sets, the RepA model performed poorly than the baseline on VoxCeleb1-H. The results show that the KLD method fails to improve model generalization consistently. On the other hand, the DKD method brings significant enhancements to lightweight models. Particularly in terms of EER, the systems trained with the DKD method showed a clear reduction of EER from the simplest VoxCeleb1-O to the most complex VoxCeleb1-H. Based on our hyper-parameter settings during training, the result shows that the inference in [21] that NCKD is the major factor affecting the distillation effect is correct. Meanwhile, this also means that DKD can also be applied in the audio signal processing field in addition to being used in the image processing field.

VI. CONCLUSION

We propose IM ECAPA Rep-TDNNLite model, consisting of lightweight backbone of IM ECAPA CNN-TDNN and RepVGG-based front-end module, for automatic speaker verification. It achieves an EER of 1.77% on VoxCeleb1-O with only 1.46M parameters, outperforming other recently proposed lightweight models. The performance can be further improved by 7% on average across multiple test sets with knowledge distillation from complex models. Thus, the proposed model can be implemented within the computational resources of edge devices while maintaining sound performance levels.

REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[2] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[4] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxceleb speaker recognition challenge 2020 system description," *arXiv preprint arXiv:2010.12468*, 2020.

[5] L. Zhang, H. Zhao, Q. Meng, Y. Chen, M. Liu, and L. Xie, "Beijing zkj-npu speaker verification system for voxceleb speaker recognition challenge 2021," *arXiv preprint arXiv:2109.03568*, 2021.

[6] A. Avdeeva, A. Gusev, I. Korsunov, *et al.*, "Stc speaker recognition systems for the nist sre 2021," *arXiv preprint arXiv:2111.02298*, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 733–13 742.

[9] Y. Ma, M. Zhao, Y. Ding, Y. Zheng, M. Liu, and M. Xu, "Rep works in speaker verification," *arXiv preprint arXiv:2110.09720*, 2021.

[10] Y. Zheng, J. Peng, M. Zhao, *et al.*, "Thuee system description for nist 2020 sre cts challenge," *arXiv preprint arXiv:2210.06111*, 2022.

[11] R. Makarov, N. Torgashov, A. Alenin, I. Yakovlev, and A. Okhotnikov, "Id r&d system description to voxceleb speaker recognition challenge 2022," 2022.

[12] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," *arXiv preprint arXiv:2104.02370*, 2021.

[13] R. Zhang, J. Wei, W. Lu, *et al.*, "Cs-rep: Making speaker verification networks embracing re-parameterization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7082–7086.

[14] T.-W. Chen, W.-T. Lin, C.-P. Chen, *et al.*, "Investigation of feature processing modules and attention mechanisms in speaker verification system," in *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, 2022, pp. 84–91.

[15] Q. Li, L. Yang, X. Wang, X. Qin, J. Wang, and M. Li, "Towards lightweight applications: Asymmetric enroll-verify structure for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7067–7071.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 630–645.

[17] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 096–10 105.

[18] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[26] J. S. Chung, J. Huh, S. Mun, *et al.*, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[27] M. Kim, Z. Piao, S. Um, *et al.*, "Light-weight speaker verification with global context information," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022, 2022, pp. 5105–5109.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Efficient Incremental Text-to-Speech on GPUs

Muyang Du*, Chuan Liu*, Jiaxing Qi*, Junjie Lai*
* NVIDIA, Beijing, China
E-mail: {myrond, riverl, jqi, julienl}@nvidia.com

***Abstract*—Incremental text-to-speech, also known as streaming TTS, has been increasingly applied to online speech applications that require ultra-low response latency to provide an optimal user experience. However, most of the existing speech synthesis pipelines deployed on GPU are still non-incremental, which uncovers limitations in high-concurrency scenarios, especially when the pipeline is built with end-to-end neural network models. To address this issue, we present a highly efficient approach to perform real-time incremental TTS on GPUs with Instant Request Pooling and Module-wise Dynamic Batching. Experimental results demonstrate that the proposed method is capable of producing high-quality speech with a first-chunk latency lower than 80ms under 100 QPS on a single NVIDIA A10 GPU and significantly outperforms the non-incremental twin in both concurrency and latency. Our work reveals the effectiveness of high-performance incremental TTS on GPUs.**

I. Introduction

With the recent blossoming in deep learning, speech synthesis methods have switched from traditional concatenation-based[1] and HMM-based[2], [3] statistical parametric methods to neural network based methods and have been widely used in various application scenarios. Compared with traditional methods, neural networks can produce more natural and high-fidelity speech at the cost of more computing power and larger latency. Therefore, reducing the latency of the speech synthesis is vital to improve the user experience of applications that require instant response, such as the virtual agents of call centers, especially in the case of highly concurrent requests during peak periods. With such needs, incremental synthesis shows its benefits. Instead of synthesizing the entire speech audio before responding, incremental synthesis can produce speech chunk-by-chunk to provide a lower response latency. Once the first audio chunk is generated, the synthesis time of the subsequent audio chunks is hidden within the playback time of the preceding chunks.

Speech synthesis pipelines deployed in production environments typically comprises three major components: a frontend for extracting linguistic features from the text, an acoustic model for synthesizing acoustic features such as the Mel spectrogram, and a vocoder for converting acoustic features into waveform samples. The frontend typically performs text normalization, grapheme-to-phoneme[4] conversion, prosodic structure prediction, and named entity recognition. BERT[5], a transformer-based pre-training language model for natural language processing, has been applied in several unified TTS frontend research efforts[6] and has been proven capable of effectively extracting a wide range of linguistic features for realistic speech synthesis. Acoustic models can be grouped into two major types: autoregressive and parallel. Among them, the autoregressive model Tacotron2[7], [8] and its variants have dominated the acoustic models used in industry for many years due to their outstanding and stable synthesis quality. In recent years, parallel models, such as FastSpeech[9], [10] and FastPitch[11], have dramatically improved the controllability of speech, and are widely in personalized synthesis and singing synthesis. Parallel speech synthesis models often have architecture based on Transformer[12] and Convolution. They have higher throughput compared with autoregressive models for non-incremental synthesis. However, the synthetic speech quality of parallel models is highly dependent on the visibility of the entire feature sequence. In the case of incremental synthesis, the perceptual field is usually limited to the size of the chunk resulting in quality degradation of parallel models. In comparison, the hidden state of autoregressive models passed over time carrying contextual information ensures speech quality unharmed. Meanwhile, Samsung's study[13] also pointed out that incremental synthesis using an autoregressive model has a more stable latency when the text length increases, which is evident as the latency is positively correlated with the computational overhead.

Similar to acoustic models, there are autoregressive and parallel vocoders. Autoregressive vocoder models such as WaveNet[14], WaveRNN[15], and LPCNet[16] synthesize audio sample-by-sample based on the acoustic features and previously generated samples. Parallel vocoders such as Flow-based WaveGlow[17], GAN-based MelGAN[18], and HiFi-GAN[19] directly map the entire acoustic feature sequence to the waveform. Unlike acoustic models, autoregressive vocoders needs to perform an enormous number of regressions to generate a chunk of audio. Although the number of regressions can be reduced using the multi-band mechanism[20], further model compression may still be required to obtain real-time. In contrast, parallel vocoders can perform chunk-by-chunk generation for incremental TTS, which is friendly to GPUs.

With more applications relying on low latency speech synthesis in recent years, several studies[21]–[23] have been proposed to optimize the quality and naturalness of incremental TTS. However, previous studies have disregarded the performance of incremental TTS on GPUs, which should be worthy of attention, especially under high concurrency scenarios. To the best of our knowledge, many incremental TTS pipelines deployed in production are CPU-based, which typically requires reducing the computational cost by compressing[24], sparsi-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

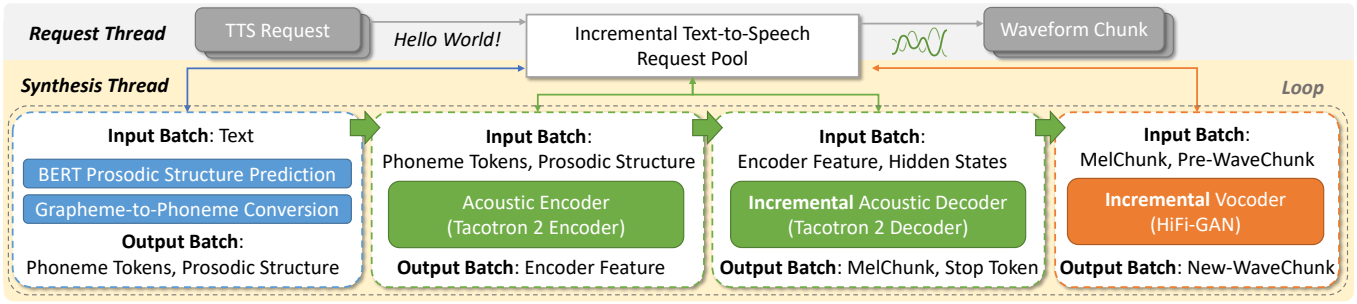


Fig. 1. Architecture of the proposed highly efficient incremental text-to-speech pipeline running on GPUs.

fying[25], and distilling[26] the model to obtain real-time. During inference, each TTS request is processed by one or more CPU cores, and batching is not involved. The additional optimization efforts for real-time on the CPU are usually time-consuming and may negatively impact the synthetic speech quality. Furthermore, the concurrency that each CPU server can handle is strictly limited by the number of CPU cores. In comparison, GPUs are better suited to computationally-intensive tasks and easier to scale. Despite the fact that GPUs have been widely used in TTS training and non-incremental synthesis as they are naturally friendly to batching, more needs to be investigated about efficient incremental TTS on GPUs.

To fill the research void, we propose an efficient approach for incremental text-to-speech synthesis on GPUs. The proposal contains a BERT-based frontend, a Tacotron 2-based acoustic model, and a HiFi-GAN-based vocoder. Furthermore, we propose the use of Instant Request Pooling and Module-wise Dynamic Batching strategies, which are vital for efficient incremental synthesis on GPUs. In the experiments, we make pressure and listening tests on the pipeline running on NVIDIA A10 GPU. Experiments prove that our approach can produce high-fidelity speech with ultra-low latency under high QPS.

II. Methods

A. Modules

*1) **Frontend**:* The frontend consists of a grapheme-to-phoneme (G2P) conversion unit and a BERT-based prosodic structure prediction model. G2P first converts the text into a Chinese pinyin sequence using the forward maximum matching algorithm based on a pronunciation dict of Chinese phrases, then converts the pinyin of each Chinese character into its phoneme tokens based on a mapping table to get the final phoneme sequence pho. During the G2P conversion, the number of phoneme tokens for each Chinese character $char_i$ is recorded as $count_i$. The prosodic structure of text can be divided into three levels: pw (prosodic word), pph (phonological phrase), and iph (intonational phrase). In order to predict the prosodic structure token sequence of each level, the text is first passed through a shared BERT backbone to extract hidden prosodic features h_i of each character using multi-head attention blocks. Then three separate linear layers are used to predict the prosody token (beginning **B** or inside **I**) of each level for $char_i$ based on h_i. Finally, the prosodic structure token sequences are regulated to have the same length as the phoneme token sequence using $count_i$.

*2) **Acoustic Encoder**:* The acoustic encoder is based on the Tacotron 2 encoder, which contains a token embedding layer, three stacked convolutional layers with batch regularization and ReLU activation, and a bi-directional LSTM layer. We refer to it as the CBRL network. In addition, three embedding layers are introduced to encode three prosodic structure sequences. The prosodic structure embedding sequence E_{pw}, E_{pph}, E_{iph} are first summed with the phoneme token embedding sequence E_{pho} then passed to the CBRL network to generate the encoded feature sequence F_{enc}.

$$\begin{aligned} E_{all} &= E_{pho} + E_{pw} + E_{pph} + E_{iph} \\ F_{enc} &= CBRL(E_{all}) \end{aligned} \tag{1}$$

*3) **Acoustic Decoder**:* The acoustic decoder is based on the auto-regressive Tacotron 2 decoder. It contains an information bottleneck PreNet, a location-sensitive attention module LSA, two stacked unidirectional LSTM layers, two linear layers for predicting Mel spectrogram and stop token, and a PostNet for improving the Mel reconstruction. Unlike the decoder used in non-incremental TTS that generates the entire Mel before vocoding. The decoder used in incremental synthesis auto-regressively generates only a chunk of Mel frames at each time, then the Mel chunk is passed directly to the vocoder for waveform chunk generation. In order to maintain the transfer of contextual information across chunks, the final decoder LSTM hidden states of the last chunk need to be explicitly cached to be used as the initial decoder states of the current chunk. These states are the last Mel frame M_{last}, the attention context A, the attention weight of the last Mel frame W_{last}, the accumulated attention weights W_{acc}, the hidden states H_{att}, H_{dec} and the cell states C_{att}, C_{dec} of the two LSTM layers.

$$\begin{aligned} H_{att}, C_{att} &= LSTM_{att}([Pre(M_{last}), A], H_{att}, C_{att}) \\ A, W_{last} &= LSA(H_{att}, F_{enc}, W_{acc}) \\ H_{dec}, C_{dec} &= LSTM_{dec}([H_{att}, A], H_{dec}, C_{dec}) \\ W_{acc} &= W_{acc} + W_{last} \\ M_{last} &= Linear(H_{dec}, A) \end{aligned} \tag{2}$$

*4) **Vocoder Model**:* The vocoder is a HiFi-GAN-based generator G that contains multiple transposed convolution and multi-receptive field fusion (MRF) blocks. In incremental TTS,

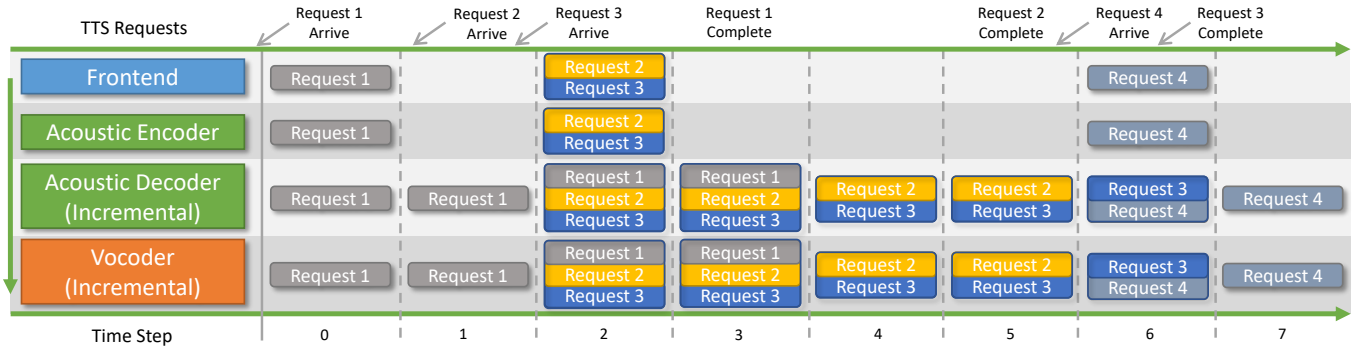


Fig. 2. Incremental synthesis timeline with the proposed instant request pooling and module-wise dynamic batching strategies.

the vocoder synthesizes one waveform chunk at a time based on a chunk of Mel frames. In order to provide a smooth articulation between two adjacent waveform chunks, we splice multiple frames from the tail of the previous Mel spectrogram chunk M_{pre} to the head of the current Mel spectrogram chunk M_{cur} and then pass it through the vocoder. Finally, the overlap areas between the current waveform chunk S_{cur} and the previous waveform chunk S_{pre} are fused by applying fade-in and fade-out coefficients α, β from equal power cross-fade, as illustrated in Equation 3.

$$\begin{aligned} S_{cur} &= G([tail(M_{pre}), M_{cur}]) \\ head(S_{cur}) &= \alpha \odot head(S_{cur}) + \beta \odot tail(S_{pre}) \end{aligned} \tag{3}$$

B. Incremental Synthesis

1) ***Incremental Synthesis of a Single Text****:* In order to achieve ultra-low latency, the proposed approach generates one short audio chunk at a time. In practice, each request contains a single text (sentence) corresponding to multiple responses. Each response carries an audio chunk of several hundred milliseconds in duration, depending on the configurable size of the chunk. The synthesis of each request involves four modules, the frontend (F), the acoustic encoder (E), the acoustic decoder (D), and the vocoder (V). The frontend and acoustic encoder modules are non-incremental and run only once per text to generated the encoded feature. On the other hand, the acoustic decoder and vocoder modules are incremental, running multiple times for each text. Each iteration produces a Mel chunk and a waveform chunk, respectively. The stop token output by the acoustic decoder exceeding threshold 0.5 flags the end of the synthesis.

2) ***Incremental Synthesis under High Concurrency****:* Incremental synthesis under high concurrency faces two challenges. Firstly, scheduling numerous requests while ensuring that each request is processed as soon as it reaches the server. Secondly, synthesizing multiple requests simultaneously while keeping low latency. As shown in Figure 1, we introduce two strategies, Instant Request Pooling and Module-wise Dynamic Batching, to guarantee new requests are synthesized instantly, together with all the other incomplete requests in the server.

■ **Instant Request Pooling**. Conventional approach to handle the incoming TTS requests while the server is processing the ongoing requests is to keep the new requests waiting inside a queue to be processed when the ongoing requests are complete. However, this approach can hardly meet the requirement of low-latency incremental TTS because the waiting time can be long under high concurrency. Therefore, we introduce an additional request pool to hold all the ongoing and new requests. We use an infinite loop to handle the incremental synthesis. Each loop iteration consistently generates one audio chunk for each ongoing request in the pool. Once a new request arrives, it is immediately added to the pool as a new pool item and available for synthesis. New requests received during the current iteration only need to wait for the next iteration to be processed, regardless of the number of ongoing requests in the pool. For each request, its pool item caches the following variables, including the model states and a module indicator:

1. Model states, including the phoneme and prosody token sequence extracted by the frontend, encoder output F_{enc}, decoder states M_{last}, W_{last}, A, W_{acc}, H_{att}, H_{dec}, C_{att}, C_{dec}, and vocoder states M_{pre}, M_{cur}, S_{pre}, S_{cur}.
2. Module indicator, indicating which module should process this item in the current loop iteration. Initialized to 0 for F, E, then changed to 1 for D, V when E completes.

■ **Module-wise Dynamic Batching**. Batching is an essential strategy to fully utilize GPU's parallel computing capability for both training and inference. For incremental TTS, we propose to use a dynamic batch size for each module. In each loop iteration, all the modules are executed sequentially, and the batch size of each module is determined by the number of items in the pool that each module should process. Specifically, each module first retrieves the pool items it should process based on the module indicator saved in each pool item. Then, it constructs an input batch for each model state from the retrieved items and performs a batched inference. Finally, it splits the output batch of each state and saves the updated state back to its corresponding pool item. Take the vocoder V as an example (F, E, D modules follows the same process), suppose N pool items are retrieved for the vocoder in the current iteration. Their states $[M^1_{pre}, M^2_{pre}, ..., M^N_{pre}]$, $[M^1_{cur}, M^2_{cur}, ..., M^N_{cur}]$, and $[S^1_{pre}, S^2_{pre}, ..., S^N_{pre}]$ are constructed to input batch M^B_{pre}, M^B_{cur}, and S^B_{pre}. After inference, the output batch S^B_{cur} is split

into $[S^1_{cur}, S^2_{cur}, ..., S^N_{cur}]$ and we update each retrieved item i:

$$M^i_{pre} = M^i_{cur},\ S^i_{pre} = S^i_{cur} \tag{4}$$

For each pool item, its F, E states keep unchanged after its first inference iteration, while its D, V states are constantly updated in the following iterations. Finally, if a request completes, its item is removed from the pool immediately.

*3) **Timeline and Latency Analysis**:* Figure 2 elaborates on the timeline of incremental synthesis with the proposed pooling and batching strategies. Each time step corresponds to one incremental synthesis iteration. In short, a new request can be processed in step t if it arrives before the start of the frontend module in step t. Request 1 arrives right before step 0 and is processed in step 0. Requests 2, 3 arrive during the execution of step 1 and are processed in step 2 in dynamic batches along with request 1. Request 1 is completed and removed immediately from the pool in step 3. Requests 2, 3 are processed together in dynamic batches in steps 4, 5. Request 2 is then completed and removed in step 5. Request 4 arrives at the end of step 5 and is processed in dynamic batches in step 6 with request 3, which is completed and removed in step 6.

Define the execution time of time step t as T_t. If a request arrives at step t, the maximum first audio chunk latency is $T_t + T_{t+1}$ (arrives right after the start of the frontend in step t), the minimum latency is $min(T_t, T_{t+1})$ (arrives right before the start of the frontend in step t, or right before the end of step t), and the expectation of the latency is $\mathbb{E} = T_t/2 + T_{t+1}$.

III. Experiments

A. Settings

All the models in our experiments are trained with the Chinese Standard Mandarin Speech Corpus (CSMSC)[27] open-sourced by Databaker. The corpus has 10,000 audio clips and a total duration of about 12 hours, each audio clip contains a short sentence of 4.27 seconds on average. We kept 100 clips for validation, 100 clips for test, and the rest for training. The Bert used in the frontend contains 12 Transformer layers with each layer 12 heads and 768 hidden units. The Tacotron 2 has the same model parameters as the original proposal with three additional 512-dim prosody embedding table. The HiFi-GAN vocoder with 128 upsample initial convolution channels and upsample factors of [8,8,2,2] is used to synthesize 22.05kHz 16bit PCM audio. All the models are trained on NVIDIA Tesla V100 GPU, and the inference pipelines are tested on NVIDIA Ampere A10 GPU. The proposed incremental pipeline is implemented in C++ as a NVIDIA Triton custom backend and the models are accelerated with TensorRT.

Since there is no publicly available performance benchmark for incremental TTS on GPU under high concurrency, we compare the proposed approach with a self-developed non-incremental pipeline using the exact same models. For the non-incremental twin, the TTS requests received over a short period are constructed into a batch for inference. The size of the batch keeps unchanged for all the models within each inference round. New requests received during the current round need to wait in the queue for the next round to be processed. Moreover, TensorRT and Triton are also applied for a fair comparison.

In the experiments, a fixed chunk size of 32 (∼372ms) Mel frames is used with an overlap (OL) of 4 and 8. We test both the incremental and non-incremental pipelines with three lengths of text (TL) and their mixture. The average audio duration (DUR) for each text type are Short (2.64s), Medium (4.08s), Long (7.00s), and Mixed (4.54s). For incremental synthesis (INCR), we measure the first-chunk latency (FCL) (between the request being sent and the first chunk being received), the last-chunk latency (LCL) (between the request being sent and the last chunk being received), and real-time factor RTF (LCL/DUR). For non-incremental synthesis (Non-INCR), we measure latency and RTF. Finally, we measure the mean opinion score (MOS) of the INCR, compared with the Non-INCR and the ground truth recordings. The MOS test is carried out by native Mandarin speakers hired from Amazon Mechanical Turk. A total of 50 fixed samples are selected from the test set for evaluating each experimental configuration. Each sample was listened by 10 evaluators and each evaluator was asked to give an opinion score from 1 to 5.

B. Discussion

TABLE I
Latency and RTF Comparison at 70 QPS

TL	OL	DUR(s)	FCL(ms)	LCL(ms)	RTF
Short	4	2.64	32.04	168.60	0.064
Medium		4.08	38.16	295.36	0.072
Long		7.00	77.05	1002.56	0.143
Mixed		4.54	44.47	400.32	0.087
Short	8	2.64	32.26	170.41	0.064
Medium		4.08	39.38	305.88	0.075
Long		7.00	91.42	1190.83	0.170
Mixed		4.54	46.29	418.37	0.091
Short	-	2.64	227.66		0.086
Medium		4.08	451.67		0.109
Long		7.00	1405.51		0.202
Mixed		4.54	1253.25		0.313

*1) **Response Latency and Real-Time Factor**:* We pressure-test the proposed approach using QPS (Queries-per-Second) from 10 to 100. Each test lasts 200 seconds. Requests sent to the server are evenly distributed within a second. Full performance metrics and samples can be found on our Github page[1]. Figure 3 shows the latency of mixed-length texts at different QPS and overlaps, which is more comparable to the real-world scenario. In the figure, we omit the data with a latency of more than 1s. For INCR, the first-chunk latency can be kept within 40 ms, and the last-chunk latency within 400 ms if the QPS does not exceed 60. In contrast, Non-INCR has a latency of more than 200 ms at 10 QPS and more than 800 ms at 60 QPS. The response latency of INCR has a 95.4%, 91.2% and 89.3% deduction compare with Non-INCR at 60, 30, and 10 QPS, respectively. At 100 QPS, INCR can synthesize in

[1]https://muyangdu.github.io/Efficient-Incremental-TTS-on-GPUs

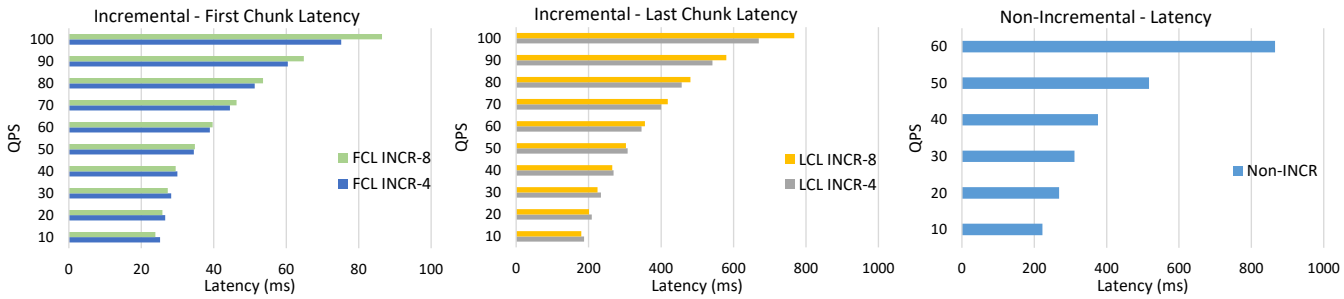


Fig. 3. Response latency of the incremental and non-incremental synthesis pipline under different QPS using mixed-length input.

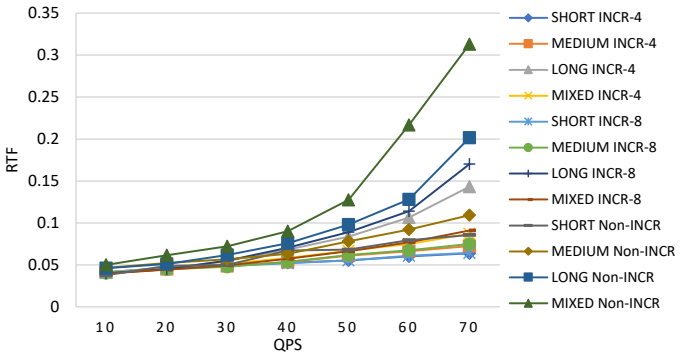


Fig. 4. Real time factor under different QPS.

real-time and provide a response latency of less than 100ms, while Non-INCR is not real-time and its latency already far exceeds 1s. In regard to the overlap, using 8 frames for overlap has slightly higher latency than 4 under high QPS.

Figure 4 shows the variation of RTF with QPS up to 70, below which all the configurations can synthesize in real-time. INCR of short, medium, and mixed text maintains a smooth RTF below 0.1 within 70 QPS. In comparison, the RTF of Non-INCR on the mixed text has a dramatic increase after 50 QPS. This proves the effectiveness and stability of the proposed strategies. Table 1 shows the performance details at 70 QPS. We observe that the mixed LCL is comparable to the medium LCL for INCR. In contrast, the mixed latency is only comparable to the long latency for Non-INCR. Similar observation is also shown in the RTF figure, in which the mixed RTF is even greater than the long RTF for Non-INCR. We attribute this observation to the redundant padding introduced for batched inference of all the modules in Non-INCR. Although padding to the maximum sequence length within a batch is still required for the F and E in INCR, the redundant padding of the D and V has been greatly reduced, leading to more efficient synthesis on GPUs.

*2) **Evaluation of Speech Quality**:* Table 2 indicates the MOS of INCR compared with Non-INCR and ground truth. The proposed approach can synthesize comparable speech quality with the non-incremental synthesis[1]. Moreover, a weak trade-off can be observed between the incremental speech quality and overlap length. Combined with the performance analysis that larger overlap only introduces slightly higher latency under high QPS on GPU, we believe it's worthwhile to use a larger overlap to achieve a better speech quality.

TABLE II
MOS WITH 95% CONFIDENCE INTERVALS

Type	Overlap	MOS (CI)
INCR	4	4.213 ± 0.120
INCR	8	4.237 ± 0.127
Non-INCR	-	4.362 ± 0.143
	Recording	4.625 ± 0.053

IV. LIMITATIONS AND FUTURE WORKS

One limitation we have identified is the use of an auto-regressive decoder, which requires multiple regressions to generate a chunk of Mel frames. Replacing it with a transformer-like parallel decoder could further speedup chunk generation. However, transformer lacks time-domain hidden states for transferring contextual information if it is used for chunk-by-chunk generation. Simply limiting the input of parallel decoder to an encoded feature chunk can result in discontinuity artifact between adjacent Mel chunks. Therefore, a potential future direction is exploring methods to use a parallel decoder for Mel chunk generation while maintaining Mel quality unharmed. Another limitation we have identified is the quality degradation caused by the overlap areas of the vocoder. While employing a larger overlap can alleviate the impact, it also reduces the ratio of valid payload of each chunk, leading to a marginally increased latency. Therefore, a potential future direction is to investigate a better streaming strategy for the parallel vocoder.

V. CONCLUSIONS

In this paper, we introduce a highly efficient approach for performing incremental text-to-speech synthesis on GPUs. Our approach allows for ultra-low latency TTS even in application scenarios with high concurrency by utilizing two important strategies: Instant Request Pooling and Module-wise Dynamic Batching. We show the effectiveness of our proposal through extensive experiments and timeline analysis. Furthermore, we offer a comprehensive performance benchmark of a complete incremental TTS system on GPUs. This benchmark can serve as a valuable reference and starting point for researchers and developers in this field. Future work will be focused on further enhancing efficiency while upholding synthetic quality.

REFERENCES

[1] L.-S. Lee, C.-Y. Tseng, and C.-J. Hsieh, "Improved tone concatenation rules in a formant-based chinese text-to-speech system," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, pp. 287–294, 1993.

[2] J. Yamagishi, T. Nose, H. Zen, *et al.*, "Robust speaker-adaptive hmm-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[3] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "Hmm training strategy for incremental speech synthesis," in *Interspeech 2015-16th Annual Conference of the International Speech Communication Association*, 2015, pp. 1201–1205.

[4] S. Hunnicutt, "Grapheme-to-phoneme rules: A review," *Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR*, no. 2-3, pp. 38–60, 1980.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

[6] Z. Bai and B. Hu, "A universal bert-based front-end model for mandarin text-to-speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6074–6078.

[7] Y. Wang, R. J. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed., ISCA, pp. 4006–4010.

[8] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, IEEE, pp. 4779–4783.

[9] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., pp. 3165–3174.

[10] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net.

[11] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6588–6592.

[12] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, *et al.*, Eds., pp. 5998–6008.

[13] N. Ellinas, G. Vamvoukakis, K. Markopoulos, *et al.*, "High quality streaming speech synthesis with low, sentence-length-independent latency," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds., ISCA, pp. 2022–2026.

[14] A. van den Oord, S. Dieleman, H. Zen, *et al.*, *Wavenet: A generative model for raw audio, corr, vol. abs/1609.03499*, 2017.

[15] N. Kalchbrenner, E. Elsen, K. Simonyan, *et al.*, "Efficient neural audio synthesis," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, J. G. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 2415–2424.

[16] J. Valin and J. Skoglund, "LPCNET: improving neural speech synthesis through linear prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, IEEE, 2019, pp. 5891–5895.

[17] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, IEEE, 2019, pp. 3617–3621.

[18] K. Kumar, R. Kumar, T. de Boissiere, *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 14 881–14 892.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-*

12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[20] T. Q. Nguyen, "Near-perfect-reconstruction pseudo-qmf banks," *IEEE Transactions on signal processing*, vol. 42, no. 1, pp. 65–76, 1994.

[21] B. Stephenson, T. Hueber, L. Girin, and L. Besacier, "Alternate endings: Improving prosody for incremental neural TTS with predicted future text input," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, ISCA, pp. 3865–3869.

[22] D. S. R. Mohan, R. Lenain, L. Foglianti, *et al.*, "Incremental text to speech for neural sequence-to-sequence models using reinforcement learning," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds., ISCA, pp. 3186–3190.

[23] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "What the future brings: Investigating the impact of lookahead for incremental neural TTS," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds., ISCA, pp. 215–219.

[24] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126–136, 2018.

[25] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, "Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks," *J. Mach. Learn. Res.*, vol. 22, 241:1–241:124, 2021.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *stat*, vol. 1050, p. 9, 2015.

[27] D. Baker, *Chinese Standard Mandarin Speech Corpus*, https://www.data-baker.com/open_source.html, [Online; accessed 7-March-2008], 2018.

Dual-Path Residual Attention Convolution Networks for Color-Embedded-Grayscale Image

Heri Prasetyo
Departemnt of Informatics
Universitas Sebelas Maret (UNS)
Surakarta, Indonesia
heri.prasetyo@staff.uns.ac.id

Abid Ammar Mahdy
Deptartment of Informatics
Universitas Sebelas Maret (UNS)
Surakarta, Indonesia
abidammar1611@student.uns.ac.id

Abrar Dwi Fairuz Nadhif
Department of Informatics
Universitas Sebelas Maret (UNS)
Surakarta, Indonesia
abrarnadip123@student.uns.ac.id

Taufiqurrakhman Nur Hidayat
Dept. Informatics Engineering
Universitas Sebelas Maret (UNS)
Surakarta, Indonesia
taufiqurrakhman.nh@staff.uns.ac.id

Rudi Hartono
Dept. Informatics Engineering
Universitas Sebelas Maret (UNS)
Surakarta, Indonesia
rudi.hartono@staff.uns.ac.id

***Abstract*—The reversible grayscale to color recovery is a challenging task due to its ill-posed inverse nature. The recovered color image produced by the color-embedded-greyscale process commonly does not have mesmerizing quality because of color distortion and checkerboard artifacts. To tackle this problem, we propose a Dual-Path Residual Attention Convolution Network (DPRACNet) to improve the quality of recovered color image. This method implements dual-path convolution network and residual attention as the building block. As reported in the Experiment Section, the DPRACNet provides best performance among the previous methods with Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) is 31.62 and 0.934, respectively.**

Keywords—color-embedded-grayscale, dual-path convolution network, image reconstruction, residual attention

I. INTRODUCTION

A color image with three channels (RGB) can be constructed into greyscale image, cited as color-embedded-greyscale image (de Queiroz & Braun, 2006). The term color-embedded means the greyscale image contains the color information (de Queiroz & Braun, 2006). Therefore, it is achievable to revert the greyscale image back to its color version. However, the recovered color image is not as good as the initial color image. Numerous methods have been proposed to tackle this issue such as swarm intelligence (Prasetyo & Suryani., 2019) and pseudo inverse (Bjerhammar, 1951). The development of Convolutional Neural Network (CNN) enable to overcome this problem using CNN based model. CNN model achieve outstanding results compared to traditional model. For instance, CorNet (Bagaskara et al., 2020) apply CNN model to fix the information distortion of recovered color image. RecovNet (Bagaskara, 2020) introduce the use of CNN end-to-end model for color-embedded-greyscale problem. CVT (Jannah, 2022) combine the transformer and CNN model. Residual Attention Network (Deann, 2023) utilize attention mechanism.

Deeper convolutional layer grant the network to learn more advance feature so it increase the ability to learn (Wu et al., 2023). However, deeper network means more computation complexity and slower to train. The performance of the network will not make a huge impact and even decline if the network too deep (He et al., 2016). To overcome this problem, He et al. (2016) proposed residual learning concept and succeed to increase the performance of the deep network. On the other side, wider network tend to train faster, but wider network has more parameters (Zagoruyko & Komodakis, 2016). Based on these problems, many researcher propose dual convolution to expand the network width but keep the parameters low. The dual-path convolution widely used in image denoising problem such as DualCNN (Pan et al., 2018), DudeNet (Tian, 2021), and DRANet (Wu et al., 2023). These method is reported to yield good performance.

As the CNN based model grow rapidly, various architecture has been proposed to tackle different computer vision problems. U-Net (Ronneberger et al., 2015) is one of the architecture that commonly used in segmentation problems. It contains encoder and decoder. U-Net achieves state-of-the-art performance and currently adopted in other computer vision problem such as image restoration (Liu et al., 2018) and image denoising (Gurrola-Ramos, 2021). More evidence of CNN growth is attention-based models. The attention framework has widely used in numerous problems, such as image classification (Hu et al., 2018), image segmentation (Oktay et al., 2018), image denoising (Wu et al., 2023), and many others. It has been verified that attention framework achieve remarkable performance.

Motivated from several concepts as mentioned above, in this paper, we present Dual-Path Residual Attention Convolution Network (DPRACNet) for color-embedded-greyscale. DRACNet divided into two sub networks, which is upper-network and lower-network. The upper-network has a u-shaped network, which use downsampling and upsampling. The lower-network consists dilated convolution to widen the receptive field of convolution. Each sub-networks is facilitated with skip connection to enhance the network training. The feature of upper-network and lower-network is concatenated and 1×1 convolution is applied to produce the recovery color image.

The remaining section of this paper are structured in the following manner. Section II provide the relevant studies about color-embedded-greyscale image and all proposed method. Proposed method is explained in Section III. Section IV report the experiment result. Finally, the experiment concluded in Section V.

II. RELATED WORKS

A. Color embedding process

Let I be a color image of size $H \times W$ in Red-Green-Blue color space. Image I can be defined as $I \in \{Ir, Ig, Ib\}$ or $I \in \mathbb{R}^{h \times w \times 3}$. The method in (de Queiroz & Braun, 2006) begins with transforming image I into Chrominance-Luminance space or YCbCr. The transformation is formulated as:

$$\{Y, C_b, C_r\} = C\{I_r, I_G, I_b\}, \quad (1)$$

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

where $C\{\cdot\}$ represents the transformation operator from RGB to YCbCr color space. The notation Y is luminent element, while C_b and C_r is chrominance element. The process continues with decomposing luminance element Y use Discrete Wavelet Transform (DWT) as follows:

$$\{Y_\theta|(\theta = LL, LH, HL, HH)\} = \Im\{Y\}, \quad (2)$$

where Y_θ state the DWT subbands of Y element. $\Im\{Y\}$ is DWT operation. Four image sub-bands of size $\frac{H}{2} \times \frac{W}{2}$ for each image are acquired from this process. The chrominance element, $\{Cb, Cr\}$ are downsampled with factor 0.5 as depicted bellow:

$$\{c_b, c_r\} = \phi(C_b, C_r), \quad (3)$$

$\{cb, cr\}$ is the downsampled product from downsampling operator $\phi\{\cdot\}$. Each size of cb and cr become $\frac{H}{2} \times \frac{W}{2}$.

The subsequent process insert the color information by replacing the luminance elements $\{LH, HL\}$ with chrominance elements $\{c_b, c_r\}$. The other DWT elements remain the same. The embedding procedure performed by following formula:

$$\tilde{I} = \Im^{-1}\{Y_{LL}, c_b, c_r, Y_{HH}\}, \quad (4)$$

$\tilde{I}$ is color-embedded-greyscale image of size HxW. The image obtained by inverse of DWT operator $\Im^{-1}\{\cdot\}$.

Due to possible pixel values of image I may beyond the range [0, 255], we perform the clipping and rounding process. The process can be formulated as follow:

$$\tilde{I}^* = Q(\tilde{I}), \quad (5)$$

where $\tilde{I}^*$ is corrupted color-embedded-greyscale image by clipping-arounding opeartor $Q(\cdot)$.

B. Residual learning

The evolution of CNN allows it to compose a very deep network. A deeper network should capture more detailed features as the layer increases. However, the more layers stacked to the network, the more performance exhibits a progressively deteriorating trend (He et al., 2016). This issue is referred to as a degradation problem (He et al., 2016). In order to overcome this problem, a residual learning concept is proposed (He et al., 2016). The main idea of residual learning is to solve learn residual mapping which is more manageable than learning original mapping (He et al., 2016). This concept enables to build a very deep network without worrying about the degradation problem. The residual learning concept has been used in many computer vision problems, such as ResNet (He et al., 2016) for image recognition, ResU-Net (Zhang et al., 2018) for image segmentation, Residual Attention Network (Deann, 2023) for color-embedded-greyscale, DRANet (Wu et al., 2023) for image denoising and many others.

C. Dilated Convolution

Widen the receptive field of convolution can lead the network to use more context to extract global feature information (Peng et al., 2019). This can be done by increasing the filter size of convolution. However, increasing the filter size will increase the network parameters and computational complexity. To overcome this problem, dilated convolution is utilized. Dilated convolution expands the receptive field without additional parameters (Peng et al., 2019). In dilated convolution, there is dilated rate, which means the spacing between the kernel. Dilated convolution-based model has been used in various computer vision problems. Some of them are image segmentation (Chen et al., 2017), object detection (Qiao et al., 2021), and image denoising (Wu et al., 2023).

D. Squeeze-and-excitation network

One of the crucial aspects of image processing is capturing and filtering proper features. The behavior of how the model captures the features affects the output produced. However, there is a possibility some feature maps produce less useful information (Hu et al., 2018). Thus, the squeeze-and-excitation network (Hu et al., 2018) was proposed to focus on informative features and suppress the less useful ones. The gating function and sigmoid function take a big part to determine the importance of the feature map (Hu et al., 2018). Because the Squeeze-and-excitation investigates the relationship of the feature map channel, it is also called channel attention. Squeeze-and-excitation network has been widely used in image classification (Hu et al., 2018), image denoising (Wu et al., 2023), image segmentation, (Mou et al, 2019), and many other fields.

E. Dual path network

A wide network is preferable to a deep network due to vanishing or exploding gradient descent as the network is deeper (Zagoruyko et al., 2016). A wide network has a downside in the aspect of network parameters. To tackle those issues, a dual path network is used. This strategy uses two parallel sub-networks to optimize the learning ability of the model (Wu et al., 2023). These sub-networks usually has different structure so it captures different features. Dual path network has been used in various problems, such as image denoising (Wu et al., 2023), object detection (Zhou et al., 2022), image segmentation (Lai et al., 2023), and image completion (Shamsolmoali et al., 2023).

III. Proposed Method

In this section, we present the overview of DPRACNet. Subsequently, we elaborate on the detailed structure of sub-networks.

A. Network architecture

The overview of DPRACNet architecture can is displayed in Fig. 1. The model is divided into two sub-networks: upper-network and lower-network. The upper network is inspired by U-Net, which consists of downsampling and upsampling. The Max Pooling is utilized to downsample the feature maps and for the upsampling, bilinear interpolation is used. The building block of the upper network is the residual block. The U-Net structure can capture multi-scale features. The lower network contains stacked dilated residual blocks. The dilation rate is shown by the number at the dilation residual block. Skip connections are employed in both sub-networks to improve the model performance. To fuse the extracted features from each sub-networks, a concatenation operation is used. The fused features passed to 1×1 convolutions to produce the recovery-colored image. The filters of all convolutional layer are set to $3 \times 3 \times 32$.

B. Upper Network

The upper network consists stacked residual block (ResBlock) with a U-shaped structure shown in Fig. 1. The structure of ResBlock is presented in Fig. 2. Each ResBlock contains two convolutions. The first convolution is followed by non-linearities. After the last convolution, we assign

channel attention to find the inter-channel correlation of extracted features. The shortcut connection is located after channel attention and followed by non-linearity. The sign '⊕' represent element-wise summation. The use of non-linearity will be determined by experimenting with several popular non-linearity, such as ReLU, Leaky ReLU, SiLU, and GeLU. The upper network is facilitated with a symmetric skip connection. A symmetric skip connection will help convolution to reconstruct the upsampled features after the spatial information loss because of downsampling.

C. Lower-Network

The lower network is composed of several ResBlock with dilated convolution. The dilated ResBlock is displayed in Fig. 3. The lower network is also facilitated by a symmetric skip connection like the upper network. Different from the upper-network, the lower network does not have a U-shaped structure. The structure of dilated ResBlock is similar to ResBlock except it uses dilated convolution. The dilation rate span from 2 to 4. The dilated convolution will widen the receptive fields so the captured features give more context and global information. The sign '⊕' denotes element-wise summation. Channel attention is also assigned in the dilated ResBlock. The non-linearity will follow the upper network and determained via experiment.

D. Channel attention

The channel attention design is inspired by the squeeze-and-excitation block. The main purpose of using channel attention is to find the correlation of the feature maps channel. Channel attention can boost the model performance and help the model to converge when training. The composition of channel attention can be viewed in Fig. 4. To get the relationship of inter-channel, first, we squeeze the dimension of feature maps by global pooling. Squeeze-and-excitation block uses Global Average Pooling, but CBAM (Woo et al., 2018), also uses Global Max Pooling. To find out the best global pooling, we will determine it by experiment. The squeezed features are then passed to two convolutions with the first convolution followed by non-linearity. The convolution will learn the global information from squeezed features. Subsequently, the learned features are sent to the sigmoid function to generate channel weight. Finally, the reweighted feature maps are obtained by element-wise product between the input features and channel weight. This produced feature map will highlight more useful features and diminish less useful ones.

E. Loss function

Mean squared error (MSE) is implemented as the loss function for DPRACNet. MSE is widely used for the image reconstruction problem. We try to reconstruct the color-embedded-greyscale image back to color image, so MSE will be suitable for the problem. The loss function can be expressed as:

$$MSE = \frac{1}{2N}\sum_{i=1}^{N}\left\| f_{\mathrm{DPRACNet}}\left(I_G^i\right) - I_C^i \right\|^2, \quad (6)$$

where N denotes the total data, $f_{\mathrm{DPRACNet}}(\cdot)$ is the prediction produced by feed-forward of DPRACNet with a color-embedded-greyscale image I_G^i as the input. I_C^i is the ground truth of the color image.

IV. Experimental Results

This section provides the full experiment environment and result. The evaluations used to measure the network performance are Peak Signal Noise Ratio (PSNR) and Structural SIMilarity (SSIM) index.

A. Dataset

The Div2K (Agustsson & Timofte, 2017) was used to train DPRACNet. We use 800 high-resolution images for training data and 100 high-resolution images for validation data. To uniform the size of all images, we resize them to a size of 512x512 pixels. Each image is patched into non-overlapping patch images size of 128x128 pixels. The train data is reproduced by image augmentation techniques, such as horizontal flip, vertical flip, and random rotation. The total number of images in train data after image augmentation is 12800 images. The model performance is evaluated using test data, containing 24 images from the kodak image dataset (Franzen, 1999).

B. Experimental setup

For this experiment, we use TensorFlow v2.11.0 framework. The model is trained in Google Colab which facilitated with GPU T4, Intel(R) Xeon(R) CPU @ 2.20GHz, 12GB RAM. We use adam optimizer with initial learning rate is 1e-3. We use default value for adam optimizer beside the learning rate. The batch size is set to 8. For network architecture tuning, the network is trained for 40 epochs. After obtaining the final architecture, the network is trained for 70 epochs with learning rate scheduler. For 1st-30th epochs, the learning rate is 1e-3. The learning rate of 31th-60th epochs is 1e-4. For the last ten epochs, the learning rate is set to 1e-5.

C. Network architecture tuning

In this experiment, there are two parts of network to tune. These are global pooling for channel attention and non linearity for the whole network. First, we find the best global pooling for channel attention. To conduct the first experiment, we try global average pooling, global max pooling, and combination of two of them. For the combination pooling, the convolution share the used weigth, so there is no addition parameters. We determine the result based on the loss in validation data. The result of global pooling experiments is reported by fig 5. This figure shows that using Global Max Pooling achieve lowest loss compare to other global pooling. Based on the reported result, we will use Global Max Pooling in channel attention.

Using Global Max Pooling, we held the second experiment. The second experiment is finding suited non-linearity for the whole network. We use some popular non-linearities, such as ReLU, Leaky ReLU, SiLU, and GeLU. The various loss decline on different non-linearities depicts in Fig. 6. As illustrated in this figure, the most suited non-linearities to implement for the whole network is ReLU. This can be considered based on the lowest loss value when the network uses ReLU as non-linearity.

Using Global Max Pooling as the global pooling in channel attention and ReLU as non-linearity, we train the final architecture. The final architecture is trained for 70 epochs with a learning rate varying from 1e-3 to 1e-5 using a learning rate scheduler. The training progress can be seen in Fig 7.

D. Performance comparisons with previous method

To discover how good DPRACNet is compared to the previous method, we evaluate the PSNR score and SSIM. The summarization of the comparison with the previous method is represented in Table 1. As shown in this table, DPRACNet has the highest PSNR and SSIM score among previous methods. It validates that DPRACNet is better than other methods by producing the recovery color image closer to its color version.

DPRACNet also has a relatively lower number of parameters compared to other deep learning methods. Although having a relatively lower number of parameters, DPRACNet has significant performance improvement compared to other previous methods.

E. Visual Investigation

The visual investigation of DPRACNet is shown in Fig 8. This figure shows how well the performance of the purposed method. As observed in this figure, we can conclude that DPRACNet has the best recovered color image compared to the previous method based on the PSNR.

V. Conclusion

In this paper, we present Dual-Path Residual Attention Convolution Network for color-embedded-greyscale images. The network consists of two sub-networks arranged in parallel. The upper network has U-shaped and ResBlock as the building block. Meanwhile, the lower network is composed of dilated ResBlock. Each sub-networks is facilitated by a symmetric skip connection to boost the performance and network training. Using dual-path convolution, the proposed method can capture more various features and achieve promising performance. Based on the experimental result, the proposed method yields the highest performance compared to the previous method based on PSNR and SSIM 31.62 and 0.9341, respectively. The proposed method also has a relatively low number of parameters so it has less complex computation.

References

[1] Agustsson, E., & Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 126-135).

[2] Bagaskara, A. R. (2020). Peningkatan Kualitas Citra Color-Embedded-Grayscale Menggunakan Deep Convolutional Networks. Surakarta: Universitas Sebelas Maret

[3] Bagaskara, A. R., Prasetyo, H., & Alim Wicaksono, H. P. (2020). Color-Embedded-Grayscale Quality Enhancement Using Deep Convolutional Networks. 2020 27th International Conference on Telecommunications (ICT). doi:10.1109/ict49546.2020.9239457

[4] Bjerhammar, A. (1951). Application of calculus of matrices to method of least squares: with special reference to geodetic calculations (pp. 1-86). Göteborg: Elander.

[5] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[6] Clerc, M., & Kennedy, J. (2002). The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE transactions on Evolutionary Computation, 6(1), 58-73.

[7] Das, S., Abraham, A., & Konar, A. (2008). Particle swarm optimization and differential evolution algorithms: technical analysis, applications and hybridization perspectives. Advances of computational intelligence in industrial systems, 1-38.

[8] Deann, K. A. (2023). Color-Embedded-Grayscale Image Menggunakan Residual Attention Network. Surakarta: Universitas Sebelas Maret

[9] de Queiroz, R. L., & Braun, K. M. (2006). Color to gray and back: color embedding into textured gray images. IEEE Transactions on Image Processing, 15(6), 1464-1470.

[10] Franzen, R. (1999) Kodak lossless true color image suite, http://r0k.us/graphics/kodak/

[11] Gurrola-Ramos, J., Dalmau, O., & Alarcón, T. E. (2021). A residual dense u-net neural network for image denoising. IEEE Access, 9, 31742-31754.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[13] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[14] Jannah, M. (2022). Rekonstruksi Citra Berderajat Keabuan-Tersisipi-Warna Menggunakan Convolutional Vision Transformer.

[15] Kennedy, J. (2006). Swarm intelligence (pp. 187-219). Springer US.

[16] Krohling, R. A., & dos Santos Coelho, L. (2006). Coevolutionary particle swarm optimization using Gaussian distribution for solving constrained optimization problems. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36(6), 1407-1416.

[17] Lai, H., Wang, T., & Zhou, S. (2023). DLUNet: Semi-supervised Learning based Dual-Light UNet for Multi-organ Segmentation. In Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation: MICCAI 2022 Challenge, FLARE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings (pp. 64-73). Cham: Springer Nature Switzerland.

[18] Liu, P., Zhang, H., Zhang, K., Lin, L., & Zuo, W. (2018). Multi-level wavelet-CNN for image restoration. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 773-782).

[19] Mou, L., Zhao, Y., Chen, L., Cheng, J., Gu, Z., Hao, H., ... & Liu, J. (2019). CS-Net: channel and spatial attention network for curvilinear structure segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22 (pp. 721-730). Springer International Publishing.

[20] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

[21] Pan, J., Liu, S., Sun, D., Zhang, J., Liu, Y., Ren, J., ... & Yang, M. H. (2018). Learning dual convolutional neural networks for low-level vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3070-3079).

[22] Peng, Y., Zhang, L., Liu, S., Wu, X., Zhang, Y., & Wang, X. (2019). Dilated Residual Networks with Symmetric Skip Connection for Image Denoising. Neurocomputing. doi:10.1016/j.neucom.2018.12.075

[23] Prasetyo, H., & Suryani, E. (2019). Swarm Intelligence on Color-Embedded-Grayscale Image. 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC). doi:10.1109/icsecc.2019.8907159

[24] Qiao, S., Chen, L. C., & Yuille, A. (2021). Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10213-10224).

[25] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

[26] Shamsolmoali, P., Zareapoor, M., & Granger, E. (2023, June). Image Completion Via Dual-Path Cooperative Filtering. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[27] Tian, C., Xu, Y., Zuo, W., Du, B., Lin, C. W., & Zhang, D. (2021). Designing and training of a dual CNN for image denoising. Knowledge-Based Systems, 226, 106949.

[28] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[29] Wu, W., Liu, S., Zhou, Y., Zhang, Y., & Xiang, Y. (2023). Dual Residual Attention Network for Image Denoising. arXiv preprint arXiv:2305.04269.

[30] Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv:1605.07146.

[31] Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters, 15(5), 749-753.

[32] Zhou, Q., Shi, H., Xiang, W., Kang, B., Wu, X., & Latecki, L. J. (2022). DPNet: Dual-Path Network for Real-time Object Detection with Lightweight Attention. arXiv preprint arXiv:2209.13933.

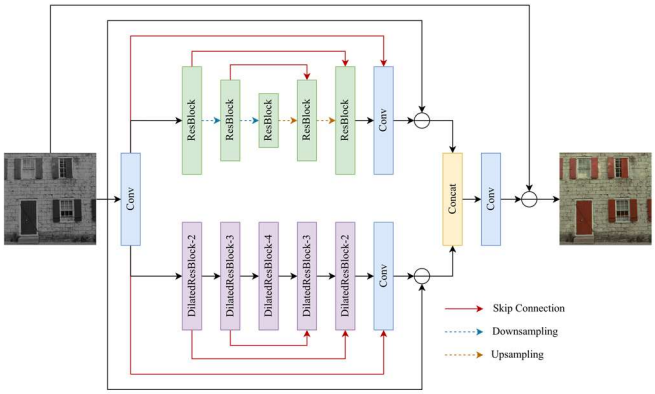


Fig. 1. The architecture of proposed method.

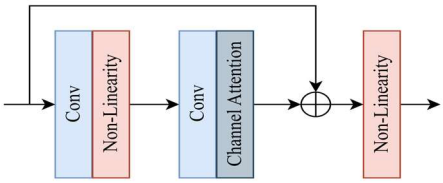


Fig. 2. The structure of ResBlock.

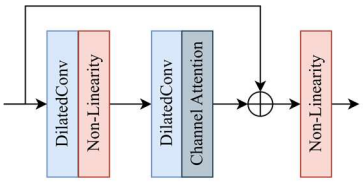


Fig. 3. Block arrangement in dilated ResBlock

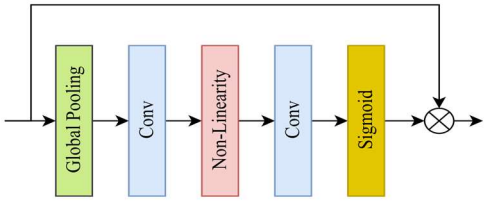


Fig. 4. The composition of channel attention

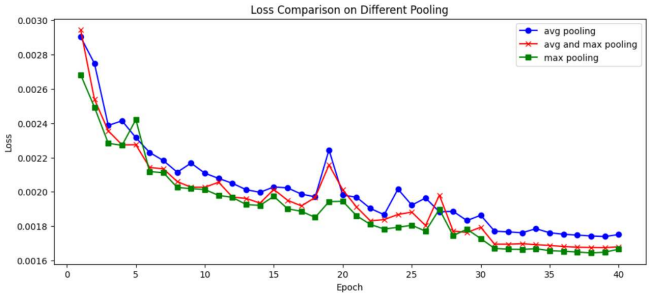


Fig. 5. Loss comparison on different pooling.

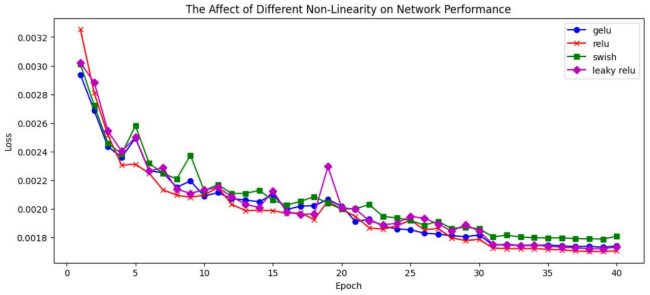
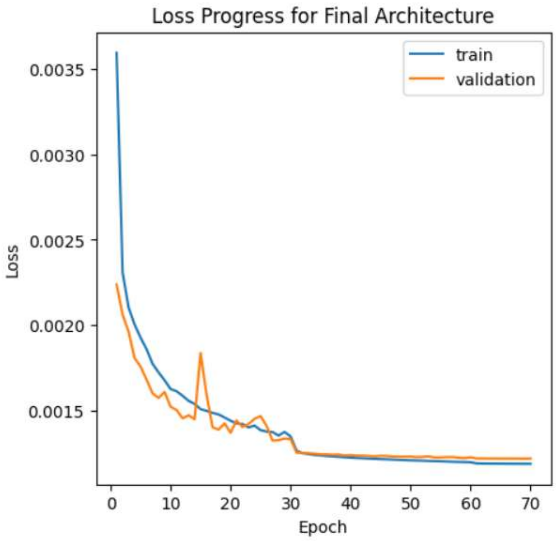
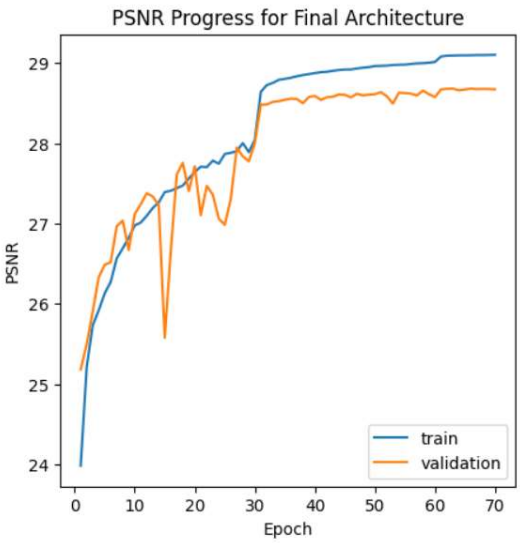


Fig. 6. The affect of different non-linearity on network performance.





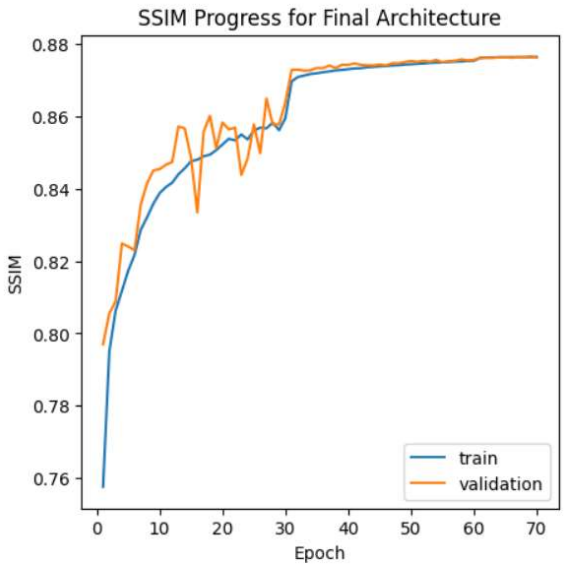


Fig. 7. Experiment summarization of final architecture.



Fig. 8. Comparison of recovered color image on different methods.

TABLE I. PERFORMANCE COMPARISON WITH PREVIOUS METHODS.

Methods	PSNR	SSIM
Former Scheme (de Queiroz & Braun, 2006)	16.70	0.5132
Classical PSO (Kennedy, 2006)	17.23	0.3863
LDWPSO (Das et al., 2006)	20.35	0.6151
CoPSO (Clerc & Kennedy, 2006)	20.34	0.6246
AGPSO (Krohling et al., 2006)	19.98	0.6246
Pseudo Inverse (Bjerhammar, 1951)	20.31	0.5296
CorNet (Bagaskara et al, 2020)	29.61	0.9320
RecovNet (Bagaskara et al., 2021)	29.57	0.9331
CVT (Jannah, 2022)	30.43	0.8373
Residual Attention Network (Deann, 2023)	31.11	0.9122
DPRACNet (proposed method)	**31.62**	**0.9341**

TABLE II. NUMBER OF PARAMETERS COMPARISON WITH PREVIOUS METHODS

Methods	Number of Parameters
CorNet (Bagaskara et al., 2020)	85,832
CVT (Jannah, 2022)	944, 803
Residual Attention Network (Deann, 2023)	210,051
DPRACNet (proposed method)	214,691

DOC: A Novel DOuble-Contour-Based Macro Placement Framework for Mixed-Size Designs

Yin-Rong Zhuo, Hui-Lin Chen, Yu-Guang Chen
Electrical Engineering of National Central University, Taiwan
E-mail: {108501017, 108501538, andyygchen}@cc.ncu.edu.tw
Tel: +886-3-4227151#34457

***Abstract* — Due to the rapid technology scaling and complexity increasing in modern IC designs, a chip often contains thousands of macros and millions of standard cells. In the back-end physical design, the placement becomes a significant challenge in the presence of large number of macros nowadays. To swiftly obtain low-power, high-efficiency, and low-cost chips, design factors such as total wirelength is usually considered as an important matric during placement stage. Conventionally, the position of the macro in initial placement is roughly determined by minimizing the wirelength at prototyping stage. After that, at macro placement stage, the location of each macro is adjusted to form a legal placement solution without increasing total wirelength. Finally, standard cells are placed based on the macro placement result. Among all three stages, the macro placement stage is crucial since the macro placement result will seriously impact the quality of standard cell placement as well as total wirelength. Therefore, in this paper, we propose a novel macro placement framework with the concept of "double contour" to find legal placement result of macros with minimum total wirelength. Specifically, our framework can be divided into two stages, the legal placement stage and optimization stage. In legal placement stage, we apply two contours which macro can be align with so the solution space is increased and the better legal placement result could be obtained. In optimization stage, we propose a simulated annealing based (SA-based) approach to further optimize the result by flipping and fine-grained moving macros. Experimental results show that our method can reduce total wirelength up to 12.5% compared with conventional single-contour-based method.**

1. INTRODUCTION

In modern IC design procedure, Intellectual Property (IP) modules and embedded memory [1] have been widely used in integrated circuit manufacturing processes to reduce design complexity and shorten time to market. Since such circuits usually contain millions of macros and standard cells, finding the appropriate position of each component is extremely difficult. The process of placing all the macros and standard cells into the chip is called mixed-size circuit placement design [2]. Currently, the most common adopted mixed-size circuit placement process can be divided into three stages: mixed-size prototyping, macro placement, and standard cells placement, as shown in Fig. 1. In mixed-size prototyping stage, we consider wirelength as the major factor to determine the initial spread of macros and standard cells on the chip, which is called the initial placement. The better initial placement should contain minimum total wirelength so that the circuit performance can be improved. In this stage, overlapping of components is allowed in order to find better solutions. In Macro placement stage, we only focus on legalization and optimization of macros based on the initial placement. In the standard-cell placement stage, macros will not be moved, and standard cells are placed in the remaining space.

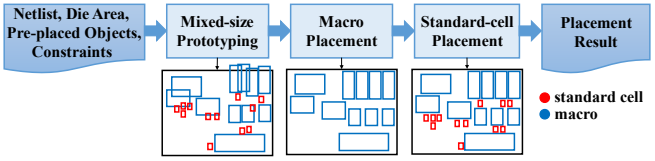


Fig. 1 Mixed-size circuit placement process.

Unlike standard cells have the same height, the size of each macro may be variant. Besides, macros are much larger than standard cells. If we place macros inappropriately, it may not only lead to longer total wirelength but also make the standard cells placement stage difficult to be implemented. Therefore, macro placement is especially important, and developing a comprehensive framework for automatically placing macros to obtain optimal macro placement efficiently is in demand. The primary goal of macro placement is to obtain a legal solution for macro placement. The solution should reserve enough space for standard cells and minimize total wirelength. Since the initial placement of macro at prototyping stage has given the wire length as the factor, it implies that in macro placement stage we should avoid increasing wirelength by minimizing the displacement of macro.

There are several papers concentrated on macro placement problems, and various data structure and algorithm such as MP-tree [3], CP-tree [4], B*-tree [5], Chang *et al.* [6], XDP [7], and ePlace-MS [8] are proposed. Among them, tree-based methods are most widely used since they are easily applied to represent the relative position of each macro in contrast of other methods. Moreover, MP tree and CP tree are inclined to reserving the center space of the chip by placing macros to the four boundaries of the chip. However, they have restriction in the case of non-boundary fixed macros. The circular-contour-based approach [9] can solve such case. Also, it usually leaves more room for standard cells than tree-based method. Nevertheless, the circular-contour-based approach will make the displacement of some macros from its original position increase, since it tends to pull the macro closed to the center to near the boundary of the chip. This may lead to longer total wirelength, and cause routing congestion at the boundary which limits the placement of standard cells.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In this paper, we propose the concept of double-contours method to take this issue into account. We use an example to illustrate our idea. Fig. 2(a) shows the initial illegal macro placement. The macros (m1~m7) inside the gray dashed rectangle are placed closed to the center of the chip. Comparing their final position after macro placement by circular-contour-based and our proposed double contour method respectively, as shown in Fig. 2(b) and (c), it can be observed that the circular-contour-based method only use single contour to decide the position where macros to be placed. Since it is limited to only a single contour, the next macro to be placed can only choose from the current contour to place. Thus, its placeable position is relatively limit and extremely close to the boundary of the chip. On the other hand, the novel concept of the double contour is shown in Fig. 2(c). Instead of single contour, we construct a second-layer contour according to the distribution positions of m1~m7 from Fig. 2(a). The second-layer is used to provide inner space for macros. In this way, the macro whose initial position is near to the outside can be placed based on the outer contour. Furthermore, the macros whose initial position is closer to the inside can be placed based on the inner contour, which will no longer be confined to the outer contour. It can be seen that our method can lead to wider solution space. Therefore, it can effectively reduce the displacement, and thus further avoid longer wirelength.

The major contributions of this paper are as follows:

- We propose a novel double contour data structure to provide a second layer of contour for macro placement Each macro can deserve more choice of placement, so that better legal placement can be found.
- We improve simulated annealing (SA) [10] which is combined with greedy algorithm to obtain optimal placement solution efficiently. Consequently, it can focus on solution quality based on HPWL by flipping and moving macros.
- Experimental results show that our algorithm is effective and robust. The proposed algorithm can achieve the better solutions from the view of routed wire length on average, compared with single contour.

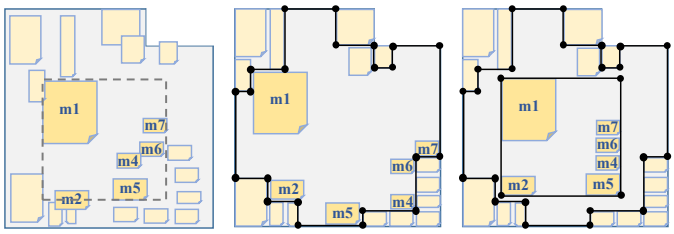


Fig. 2 Macro placement method by single contour and double contours. (a) prototype (b) single contour (c) double contour

2. PREVIOUS WORKS

Since our method refers to Slicing tree [11], contour, and SA [10], here we first introduce them below.

2.1. *Slicing tree-based approaches*

Its purpose is to divide the chip into some blocks. The method is to divide a rectangular region into two rectangles with vertical or horizontal lines each round. That is, the Slicing tree is a binary tree composed of internal nodes representing vertical or horizontal lines and leaf nodes representing rectangular domains. Its advantage is to avoid macros in a certain area to be put far from the original region. Therefore, the total displacement of the macros can be effectively reduced, and the complexity of routing is also reduced.

2.2. *Contour-based placement approaches*

There are some previous algorithms that use the concept of the contour, such as MP-tree [3], CP-tree [4], B*-tree [5], MDP-tree [12], and so on. These packing trees use only a contour in one direction (ex: horizontal contour or vertical contour). Their contour is defined as the surface that can be irradiated by imagining parallel light in the specified direction. It will cause great restrictions in the process of placing the macros. Furthermore, there will be some wasted area. Another paper proposed the concept of circular contour [9] which improves the disadvantages of the single-direction contour just mentioned above. However, the macro is limited to be put along the frame. Hence, we propose a double-contours structure, so that the macro can also be placed in the middle area of the frame to solve the above problems.

2.3. *Simulated Annealing (SA)*

SA [10] will find a new solution based on the current solution. Also, it has a chance to accept a poorer solution instead of being limited to the current best solution and stop. Hence, it may find a better solution. The probability of acceptance will decrease along with the gradually reduced temperature, so eventually it converges to a better solution. Therefore, in this paper, after we find the initial legal solution, we will use the SA to further optimize our placement.

3. PROBLEM FORMULATION

This paper targets the macro placement problem in the three-stage mixed-size placement, as shown in Fig. 3. Specifically, we assume the initial placement result is given, and our mission is to adjust the macros by moving or flipping. For the purpose that the placement result can meet all the constraints and have minimal total wirelength. There are five constrains mentioned above. First, all the macros should be placed inside the chip. Second, the fixed macros must not be moved. Third, the orientation is limited to four directions, namely N, FN (for Y-axis flip), S (for X-axis and Y-axis flip), and FS (for X-axis axis flip). Fourth, the minimum distance that must be preserved the channels between two adjacent macros. The last one is that macros comply with "buffer component area reservation restrictions." Namely, the region between the frame of each macro and the boundary expanded a fixed distance should be overlapped with the space where standard cells could be put in.

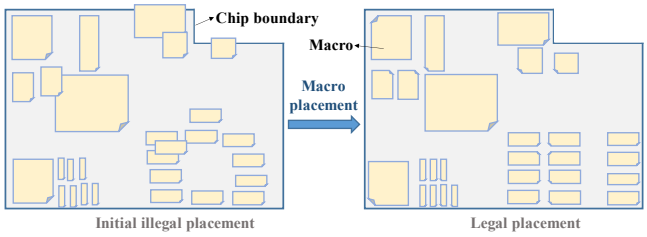


Fig. 3 Macro placement.

Eventually, after macro placement, we would use the NTUplace3 [13] to perform standard cell placement and then estimate total wirelength by half-perimeter wirelength model (HPWL).

4. ALGORITHM

Fig. 4 shows the flow of the entire algorithm. The input data are the initial illegal macro placement. First, deal with the irregular frame and the minimum channel distance limit. When generating the initial legal placement, we would avoid the huge difference between input data and our legal placement which caused to the longer total wirelength. Hence, slice the chip according to the connections between macros. Meanwhile, allocate macros to each domain. Then, decide the position of the macros by our proposed structure, double contours. The final step is to go through our modified SA algorithm in order to optimize the position and the orientation of the macros.

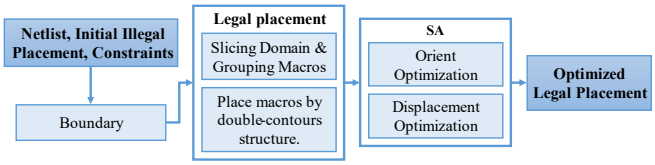


Fig. 4 Flow chart of our algorithm.

4.1. Boundary

A. Deal with the irregular frame and extend the macros

Use the fake fixed macro to fill the irregular frame into a rectangular frame, as shown in Fig. 5(a). In addition, we expand each macro by half of the minimum channel distance, as shown in Fig. 5(b). And the macro must be expanded toward four directions, from, bottom, left and right. Therefore, there is at least a certain distance between components.

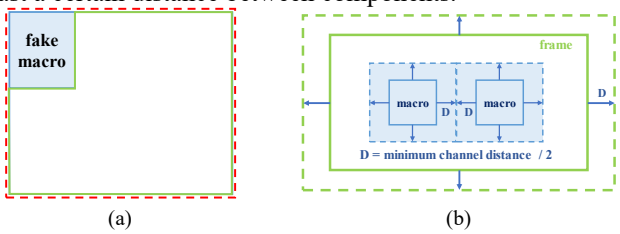


Fig. 5 (a) Deal with the irregular frame. (b) Extend the macros.

B. Integer multiples of the cell height and width

According to the distance between the macro and the four sides of the frames respectively, we will expand toward the direction with a larger distance, as shown in Fig. 6. The value of the expansion is an integer multiple of the least common multiple of the height and width of the standard cells. This purpose is to reserve the valid space to place the standard cells.

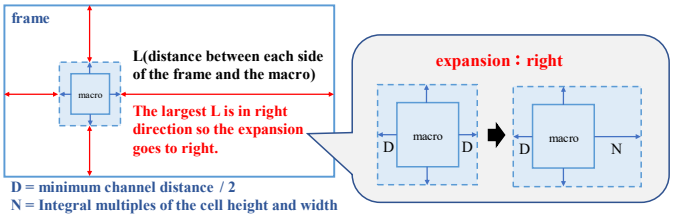


Fig. 6 An example of the macro expansion.

4.2. Slicing Domain and Grouping Macros

In order to minimize the total wirelength, we need to divide the chip into multiple domains at this stage according to the position of the macros on the chip and their connections. And each macro could only put within the block that it belongs to so as to avoid excessive displacement of the macro. Fig. 7 is the flow chart of slicing domain and grouping macros. We will store each domain in array, and then put the macros in the domain into the stack. The exact cut position is based on the net connections between the macros in the stack. Determine the cutline that if the domains on both sides of the cutline can hold macros. The information of the two domains will be added to the array, and then the above actions will be repeated until the domains in the array can no longer be divided. The following will tell more details about how to determine the cutline and how to construct domain information.

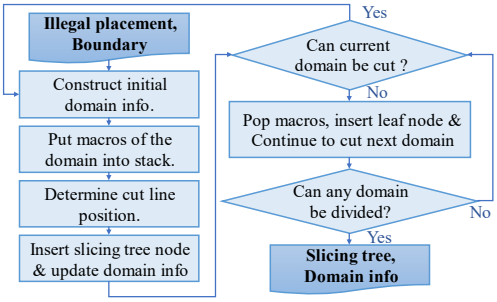


Fig. 7 Flow chart of "Slicing Domain and Grouping Macros."

A. Decide cutline

First, the cutline would be the edge of one of the macros in this domain. Besides, in order to avoid too much difference in the number of macros on both sides after cutting domain, we set the parameter r as the ratio of the number of macros on both sides, as shown in Fig. 8. And limit the cutline to fall in the range of $r >= R_{limit}$, where R_{limit} would decrease. If R_{limit} is closer to 1, indicate that the quantity on both sides is more balanced. Furthermore, if a group of the macros whose net connections between each one is more than others is divided into different domain, it would cause more displacement after placing macros. Therefore, in order to place densely connected macros in the same domain, among the cutlines within the range mentioned above, find the one that the smallest numbers of the nets cross it. In Fig. 8, the number of nets cut is 2 (n4 and n1).

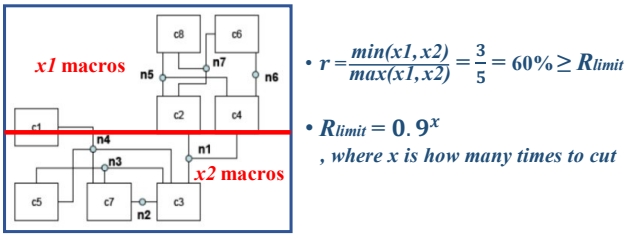


Fig. 8 An example selecting one of the cutline.

B. Construct domain info

Fig. 9 is a schematic diagram of constructing domain info. The upper part is the process of division, and the lower part is to store information according to the domain and macro in the

upper part. First, it will be divided according to the method described in the previous paragraph, and then it will be judged whether to continue to divide from the domain on the left/below of the cutline. Otherwise, it will be divided from the domain on the right/above of the cutline. Domain info is an one-dimensional array which contains the coordinates of the lower left corner of the domain, the length and width of the domain, and the macros in the domain. The macros are recorded in the form of linked-list.



Fig. 9 The process for constructing the domain info.

4.3. *Double contours*

In the previous step, the chip has been divided into multiple domains. In this step, the macros in each domain must be placed legally. Therefore, we determine the position of the macro by the double-contours structure. The advantage is that the macro in the middle will not be moved to the corner of the domain. As the name implies, the double-contours structure has two layers of contours. We assume that the macros could only place on the inner side of the contour, as shown in Fig. 10.

The algorithm we proposed is as shown in Fig. 11. It is mainly divided into two stages. First, try to place all macros. If there are some macros can't find the legal position to put, these macros will be placed in adjacent domains until all macros have a legal location. In the above first stage, each domain will first judge whether to add the second layer of contours. Then, initialize the contours according to this result. Since the fixed macros cannot change their positions, and the placed macros are the opposite, they need to be handled separately. Update the contour according to the location of the fixed macro first, and use the contour to obtain the legal placement of the placed macros afterwards. The following uses a domain as an example to illustrate.

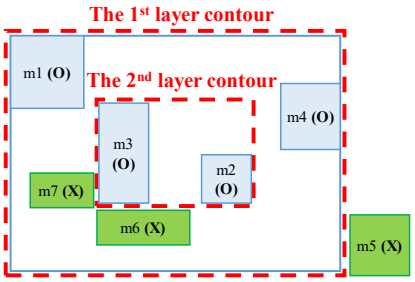


Fig. 10 The proper and the improper position where to place the macros.

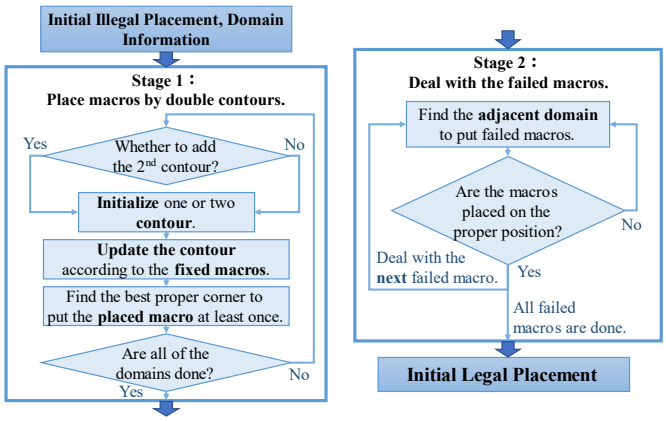


Fig. 11 Flow chart of the double-contours algorithm.

A. *Decide whether to join the second circular contour*

First, we set the second initial contour as a rectangle with the same aspect ratio as the domain block to which it belongs, as shown in Fig. 12. That is, the contour length and width are the domain length and width multiplied by a parameter (*shrink_ratio*) less than 1. If the ratio of the number of macros in the range of the second initial contour to the total number of macros in the domain exceeds *macros_percent* which is a self-defined parameter, the second layer of contour will be added. We set *macros_percent* as 40% in our experiment.

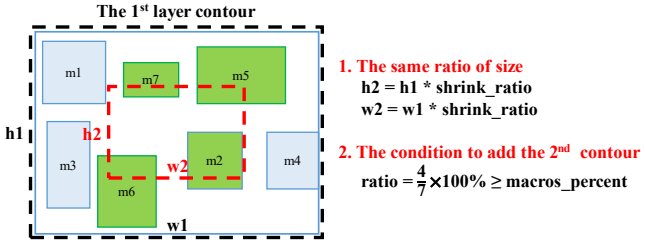


Fig. 12 Confirm whether to add the second layer of the contour.

B. *Deal with fixed macros*

Update the contour according to the location of the fixed macros. There are three possible situations, inside, outside, and on the edge of the domain. In the above three situations, you must first find the side closest to the domain frame, and update the contour so that it bypasses the macro, as shown in Fig. 13.

C. *Deal with placed macros*

The contour has been updated in the previous step. This step will try to place all the macros. First, calculate the distance between each macro and all the proper corners in the contour. Then, try to put the macro on the corner with the smallest distance and confirm whether it would overlap with other macros that have been placed and whether it would exceed the boundary of the domain. If there is a violation, try the next pair of macros and corners with the smallest distance, and so on, as shown in Fig. 14. If none of the macros are successfully placed in the first round, the failed macros will be placed first in the next round until all the macros are successfully placed. In the worst case, the macros cannot be placed successfully, and it would go to the stage 2, as shown in Fig. 11.

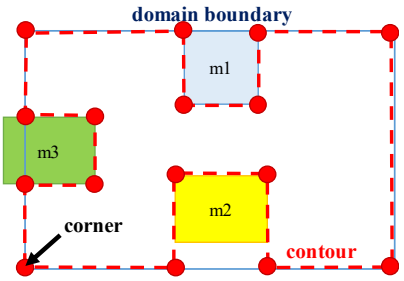


Fig. 13 The relationship between the fixed macro and the domain boundary.

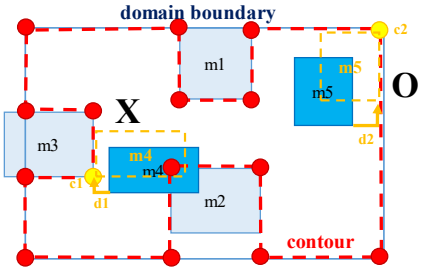


Fig. 14 The picture shows how to handle placed macros. d1 and d2 are the first and second smallest among all the calculated distances between macro and corner. However, when m4 is placed on c1, it will overlap with m2, so this round will choose to place m5 on c2. Eventually, the moving distance is small and the position is legal.

4.4. Modified-SA (simulated annealing)

After the double-contours stage is completed, a legal placement solution will be obtained. In order to optimize the initial placement solution, we adopt the simulated annealing (SA) [10] method in this step and add some additional action. The algorithm flow chart is our modified SA, as shown in Fig. 15.

The disadvantage of SA is that the final output solution may not be optimal due to the possibility that it will update to a poorer solution. Therefore, our modified algorithm is different from the traditional SA. We store the current solution and the better solution accepted previous time in addition in order to avoid the last updated placement solution after SA accepts a poor solution. We will also keep the best solution obtained in the process. Hence, at the end of SA, even if the last updated solution is worse, we still output the best solution. In addition, if SA doesn't accept the poor solution, it would operate the heuristic algorithm [14] trying to improve the current solution.

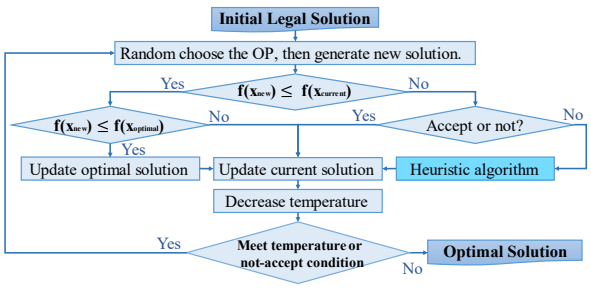


Fig. 15 Flow chart of the SA algorithm.

A. OP introduction

In order to get the optimal solution, how to adjust the direction and position of the macro is very important. The following are some operations (OP) we designed.

- OP1: Move a macro to another domain.
- OP2: Switch two macros in the different domain.
- OP3: Combine two adjacent domains.
- OP4: Slice one domain into two sub-domains.
- OP5: Add one fake macro.
- OP6: Flip one macro.
- OP7: Move a macro in random direction and by random distance.

B. Heuristic algorithm

In order to find the best solution more efficiently and effectively, we also add a Heuristic algorithm into the process. This algorithm will try to move or flip all the macros every time. If each macro is adjusted, it will have its own gain value which is the difference of the total wirelength. Thus, adjust the macro haven't been adjusted with the smallest gain each time. After all the macros are done, a pass is completed. After each pass is completed, it will be compared with the solution of the previous pass. If it improves, continue to the next pass. Otherwise, stop the loop. The advantage of the Heuristic algorithm is that it can efficiently find a better solution, even if there is no guarantee that the solution will be optimal. Therefore, when the OP selected by SA is unable to obtain an acceptable solution, this algorithm will be used to quickly find a better solution.

5. EXPERIMENT

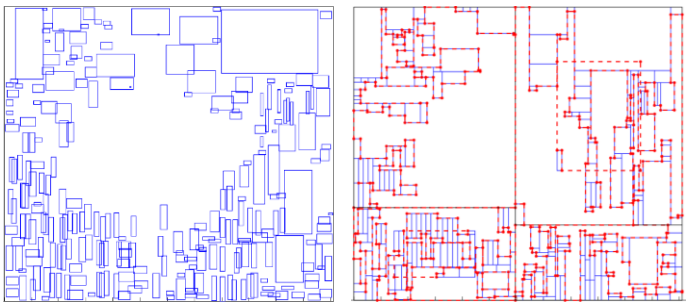
We implemented our macro placement framework in the C++ programming language. All experiments were performed on an Intel Xeon Gold 5320 2.2GHz Linux workstation with 131GB memory. We used the benchmark provided by CAD Contest 2022 [15] for testing. As shown in the second and third column of Table 1, "macro" and "component" represent the number of macros and all components respectively.

For the purpose of analysis, we use Matlab to generate visual images. Fig. 16(a) represents the layouts of initial illegal placement of the input file. Also, Fig. 16(b) represents the legal placement after using our double contour method. The blue rectangle represents the macro. The red dashed lines are contours that may be the first or second layer of the contour. It can be clearly seen that in the initial placement stage, there are many macros that overlap or outside the chip. Moreover, macros are scattered in a messy manner. Through our method to place macros, there are not any macros overlapping or outside the chip. Also, they are placed relatively neatly in the chip.

Here we recall that our framework could be divided into two main stages, "legal placement" and "SA". Therefore, we compare the HPWL result of three different approaches, as shown in Table 1. In Table 1, "single contour" indicates the HPWL estimated by the NTUplace3 [13] after only using the single contour to get the legal placement. The next two columns represent the HPWL after using double contours to get the legal placement and its improvement rate. The last two columns represent the HPWL after using double contours and SA to get the optimized legal placement and its improvement rate.

Table 1 improvement of HPWL of double contours compared to single contour

benchmarks	benchmark info.		single contour	double contour		double contour + SA	
	macro	component		HPWL	improvement	HPWL	improvement
Case 1	175	1,209,716	235,668,699,752	226,441,129,211	3.92%	225,994,886,353	4.10%
Case 2	84	1,876,103	366,912,034,519	363,289,921,227	0.99%	345,702,012,698	5.78%
Case 3	84	1,876,103	401,384,714,466	398,328,082,658	0.76%	353,207,445,548	12.00%
Case 4	175	1,209,716	202,787,906,675	197,502,553,240	2.61%	194,540,514,420	4.07%
Case 5	175	1,209,716	204,944,999,466	179,741,577,927	12.30%	179,440,258,447	12.44%



(a) (b)

Fig. 16 (a) Initial placement. (b) Macro placement using double-contours method.

6. CONCLUSIONS

In this paper, we propose a novel double-contours concept for macro placement in mixed-size circuits. By appending a second layer of contour, the macro will not be limited to be placed in the outer frame of the chip, so that there are more possibilities for placement. Moreover, through our improved SA, several operations we design will constantly flip and move macros to get a more optimal placement. The experimental results show that our method in most of the cases can achieve the effect of reducing the wirelength, and in case3 and case5, even improves the HPWL by more than 12%. We believe that after more detailed improvement, such a tool will bring great benefits to mixed-size circuit design.

REFERENCES

[1] E. Wein and J. Benkoski. “Hard macros will revolutionize SoC design.” EE Times Online, August 2004.

[2] T.-C. Chen, P.-Y. Lee, and T.-C. Chen, “Automatic Floorplanning for AI SoCs,” Proceedings of The 1st IEEE-TSA VLSI Design Automation and Test Conference (VLSI-DAT), 2020.

[3] T.-C. Chen, P.-H. Yuh, Y.-W. Chang, F.-J. Huang and T.-Y. Liu, “MP-trees: A packing-based macro placement algorithm for modern mixed-size designs”, IEEE Trans. Computer.-Aided Design Integr. Circuits Syst., vol. 27, no. 9, pp. 1621-1634, Sep. 2008.

[4] Y.-F. Chen, C.-C. Huang, C.-H. Chiou, Y.-W. Chang and C.-J. Wang, “Routability-driven blockage-aware macro placement”, Proc. 51st Annu. Design Autom. Conf. Design Autom. Conf. (DAC), pp. 1-6, 2014.

[5] Y.-C. Chang, Y.-W. Chang, G.-M. Wu, and S.-W. Wu. “B*-trees: A new representation for non-slicing floorplans.” In Proc. of ACM/IEEE DAC, pages 458–463, Los Angeles, CA, June 2000.

[6] C.-H. Chang, Y.-W. Chang and T.-C. Chen, “A novel damped-wave framework for macro placement”, Proc. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD), pp. 504-511, Nov. 2017.

[7] J. Cong and M. Xie. “A robust mixed-size legalization and detailed placement algorithm.” IEEE Trans. Computer-Aided Design, 27(8):1349–1362, August 2008.

[8] J. Lu et al., “EPlace-MS: Electrostatics-based placement for mixed-size circuits”, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 34, no. 5, pp. 685-698, May 2015.

[9] C.-H. Chiou, C.-H. Chang, S.-T. Chen and Y.-W. Chang, “Circular-contour-based obstacle-aware macro placement”, Proc. 21st Asia South Pacific Design Autom. Conf. (ASP-DAC), pp. 172-177, Jan. 2016.

[10] D. Wong, H. W. Leong, and C. L. Liu. “Simulated annealing for VLSI design.” Kluwer Academic Publishers, 1988.

[11] F. Y. Young and D. Wong. “How good are slicing floorplans?” Integration, the VLSI journal, 23(1):61–73, 1997.

[12] Y.-C. Liu, T.-C. Chen, Y.-W. Chang and S.-Y. Kuo,“MDP-trees: multi-domain macro placement for ultra large-scale mixed-size designs”, Proc. ASP-DAC, 2019.

[13] T.-C. Chen, Z.-W. Jiang, T.-C. Hsu, H.-C. Chen, and Y.-W. Chang. “NTUplace3: An analytical placer for large-scale mixed-size designs with preplaced blocks and density constraints.” IEEE TCAD, 27(7):1228–1240, July 2008.

[14] C. M. Fiduccia and R. M. Mattheyses. “A linear-time heuristic for improving network partitions.” In DAC ’82: Proceedings of the 19th conference on Design automation, pages 175–181, Piscataway, NJ, USA, 1982. IEEE Press.

[15] ICCAD 2022 CAD Contest Problem D: Wirelength-Driven Detailed Macro Placement.

[16] ICCAD 2021 CAD Contest Problem D: Macro Legalization.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Hindering Adversarial Attacks with Multiple Encrypted Patch Embeddings

AprilPyone MaungMaung[1], Isao Echizen[1,2], and Hitoshi Kiya[3]
[1]National Institute of Informatics, Tokyo, Japan [2]University of Tokyo, Tokyo, Japan [3]Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—In this paper, we propose a new key-based defense focusing on both efficiency and robustness. Although the previous key-based defense seems effective in defending against adversarial examples, carefully designed adaptive attacks can bypass the previous defense, and it is difficult to train the previous defense on large datasets like ImageNet. We build upon the previous defense with two major improvements: (1) efficient training and (2) optional randomization. The proposed defense utilizes one or more secret patch embeddings and classifier heads with a pre-trained isotropic network. When more than one secret embeddings are used, the proposed defense enables randomization on inference. Experiments were carried out on the ImageNet dataset, and the proposed defense was evaluated against an arsenal of state-of-the-art attacks, including adaptive ones. The results show that the proposed defense achieves a high robust accuracy and a comparable clean accuracy compared to the previous key-based defense.**

I. Introduction

Deep learning has brought major breakthroughs in many applications [1]. Some notable examples are visual recognition [2], natural language processing [3], and speech recognition [4]. Despite the remarkable performance, machine learning (ML) in general, including deep learning, is vulnerable to a wide range of attacks. Particularly, many ML algorithms, including deep neural networks, are sensitive to carefully perturbed data points known as adversarial examples that are imperceptible to humans, but they cause ML models to make erroneous predictions with high confidence [5]–[7]. Not only in the digital domain, adversarial examples can also appear in the physical domain. One work shows that an adversarial example can be photographed with a smartphone, and the taken picture can still fool a neural network [8]. Similarly, another work demonstrated that it is possible to construct 3D adversarial objects [9]. Such physical adversarial attacks may cause serious and dangerous problems by applying paintings or stickers on physical objects [10], [11]. For example, self-driving cars may misclassify a "Stop" sign as a "Speed Limit" [11], and face recognition systems may authenticate unauthorized users as authorized ones [12]–[15].

As adversarial examples are an obvious threat, numerous methods for generating adversarial examples (attack methods) and defenses against them have been proposed in the literature [16]. Adversarial attacks and defenses have entered into an arms race, and attack methods are ahead of defense ones. Once defense mechanisms are known to the attacker, adaptive attacks can be carried out [17], [18]. Most of the defense methods either reduce the classification or are completely broken. Therefore, defending against adversarial examples is still challenging and remains an open problem.

Inspired by cryptography, a new line of research on adversarial defense has focused the use of secret keys so that defenders have some information advantage over attackers [19]–[23]. Key-based defenses follow Kerckhoffs's second cryptographic principle, which states that a system should not require secrecy even if it is exposed to attackers, but the key should be secret [24]. The main idea of the key-based defense is to embed a secret key into the model structure with minimal impact on model performance. Assuming the key stays secret, an attacker will not obtain any useful information on the model, which will render adversarial attacks ineffective. The idea of making adversarial attacks expensive or ideally intractable is further supported on a theoretical basis that adversarially robust machine learning could leverage computational hardness, as in cryptography [25].

However, a recent study shows that applying cryptographic principles alone in algorithm design is insufficient to defend against efficient adversarial attacks [26]. To further harden key-based defenses, researchers have also proposed to use implicit neural representation [26] and ensembles of key-based defenses [27], [28]. Nevertheless, although these defenses seem promising, they have yet to tested on large datasets like ImageNet. Therefore, in this paper, we build upon the idea of key-based defense and propose a novel defense that is not only efficient to hinder adversarial attacks but also scalable to large datasets. The proposed defense utilizes multiple secret patch embeddings and classifier heads that are controlled by secret keys and a pre-trained isotropic network (with the same depth and resolution across different layers in the network) backbone such as a vision transformer (ViT) [29] and ConvMixer [30]. Our contributioins in this paper are as follows.

- We propose a novel defense idea that uses multiple secret patch embeddings.
- We show that the proposed defense is computationally feasible even on ImageNet and report empirical results on the strongest attacks.

In experiments, the proposed defense is confirmed not only to be efficient against attacks, but also to maintain a comparable clean accuracy.

II. Related Work

A. Adversarial Robustness

There are two distinct strategies in designing adversarial defenses: (1) classifiers are designed in such a way that they are

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

robust against all adversarial examples in a specific adversarial space either empirically (*i.e.*, adversarial training) or in a certified way (*i.e.*, certified defenses), and (2) input data to classifiers are pre-processed in such a way that adversarial examples are ineffective (*i.e.*, input transformation defenses, key-based defenses).

Adversarial Training. Current empirically robust classifiers utilize adversarial training, which includes adversarial examples in a training set. Madry *et al*. approach adversarial training as a robust optimization problem and utilizes projected gradient descent (PGD) adversary under ℓ_∞-norm to approximate the worst inputs possible (*i.e.*, adversarial examples) [31]. As PGD is iterative, the cost of computing PGD-based adversarial examples is expensive. Much progress has been made to reduce the computation cost of adversarial training as in free adversarial training [32], fast adversarial training [33], and single-step adversarial training [34]. However, adversarially trained models (with ℓ_∞ norm-bounded perturbation) can still be attacked by ℓ_1 norm-bounded adversarial examples [35].

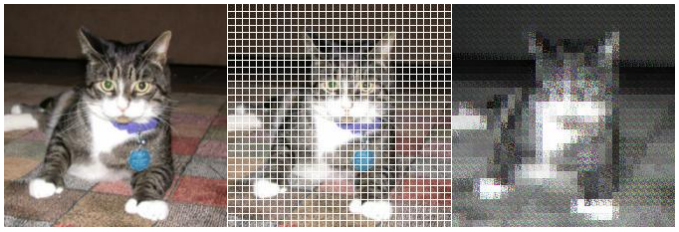
Certified Defenses. Robust classifiers in this direction use formal verification methods in such a way that no adversarial examples exist within some bounds [36]–[39]. Ideally, these defenses are preferred for achieving certain guarantees.. Although certified defenses are attractive, they can be bypassed by generative perturbation [40] or parametric perturbation (outside of pixel norm ball) [41].

Input Transformation. Adversarial defenses in this strategy aim to find a defensive transform to reduce the impact of adversarial noise or make adversarial attacks ineffective (*i.e.*, computing adversarial noise is either expensive or intractable). The works in this direction use various transformation methods, such as thermometer encoding [42], diverse image processing techniques [43], [44], denoising strategies [45], [46], GAN-based transformation [47], and so on. Although these input transformation-based defenses provided high accuracy at first, they can be attacked by adaptive attacks such as [17], [18].

Key-Based Defenses. The defenses with a secret key are similar to the input transformation-based defenses. However, the major difference is that defenders have an information advantage over attackers (*i.e.*, a secret key). The success of adaptive attacks is based on the ability to approximate a defended model's outputs on arbitrary inputs. On this insight, key-based defenses hide a model's decision from attackers by means of training the model with encrypted images. Such key-based defenses include [19], [20]. However, a recent study pointed out that such key-based defenses are still vulnerable to adaptive attacks [26]. Instead, they proposed a new key-based defense by using neural implicit representation [26].

B. Patch Embedding

Inspired by the success of the vision transformer (ViT) [29], a series of works such as [30], [48]–[52] utilizes an isotropic architecture that uses patch embeddings for the first layer. Especially a recent network architecture, ConvMixer [30] showed some evidence that patch embeddings play an important role in achieving high performance accuracy. With hindsight, we fine-tune the patch embedding layer by using block-wise encrypted images to obtain encrypted patch embeddings in this paper.



(a) (b) (c)

Fig. 1. Block-wise image transformation with block size $M = 7$. (a) Plain. (b) Block division. (c) Block-wise shuffling.

C. Block-Wise Image Transformation

The previous key-based defense proposed three block-wise transformation with a secret key for an adversarial defense: pixel shuffling, bit flipping, and FFX-based transformation [20]. In this paper, we adopt block-wise pixel shuffling from [20] for the proposed defense, and the detailed procedure for block-wise pixel shuffling is described as follows.

1) Divide a three-channel (RGB) color image, x with $w \times h$ into non-overlapping blocks each with $M \times M$ such that $\{B_1, \ldots, B_i, \ldots, B_{(\frac{h}{M} \times \frac{w}{M})}\}$.
2) Generate a random permutation vector, v with key K, such that $(v_1, \ldots, v_k, \ldots, v_{k'}, \ldots, v_{3M^2})$, where $v_k \neq v_{k'}$ if $k \neq k'$.
3) For each block B_i,
   flatten three-channel block of pixels into a vector, b_i such that $b_i = (b_i(1), \ldots, b_i(3M^2))$,
   permute pixels in b_i with v such that
   $$b'_i(k) = b_i(v_k), k \in \{1, \ldots, 3M^2\}, \quad (1)$$
   and reshape the permuted vector b'_i back into the three-channel block B'_i.
4) Integrate all permuted blocks, B'_1 to $B'_{(\frac{h}{M} \times \frac{w}{M})}$ to form a three-channel pixel shuffled image, x'.

As an example, Fig. 1 shows a plain image with visualization of blok division and block-wise shuffling.

III. Proposed Defense with Encrypted Patch Embeddings

A. Overview

We consider an image classification scenario where a provider deploys a key-based defense and provides application programming interface (API) to users (potentially adversarial) for inference as shown in Fig. 2.

Formally, given a dataset $\mathcal{D}$ with pairs of examples (images and corresponding labels), $\{(x, y) \mid x \in \mathcal{X},\ y \in \mathcal{Y}\}$, a key-based defense in [20] first maps the input space $\mathcal{X}$ to an encrypted space $\mathcal{H}$ by using a blok-wise transformation with

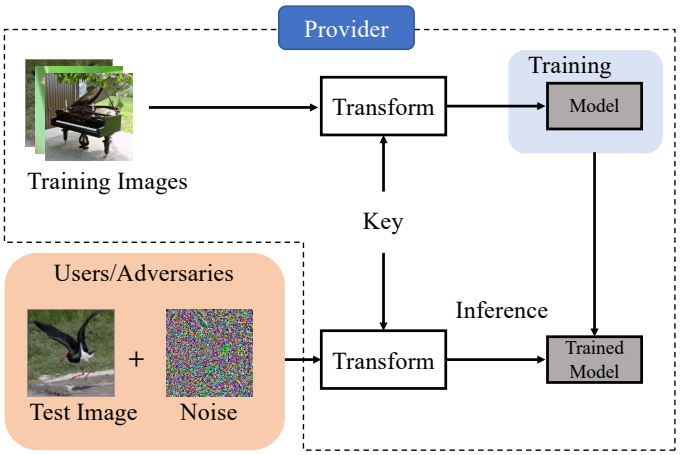
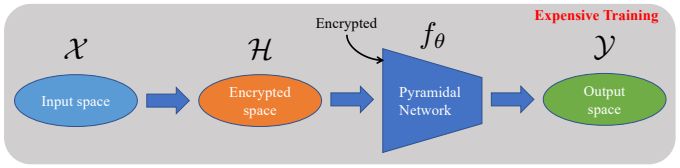
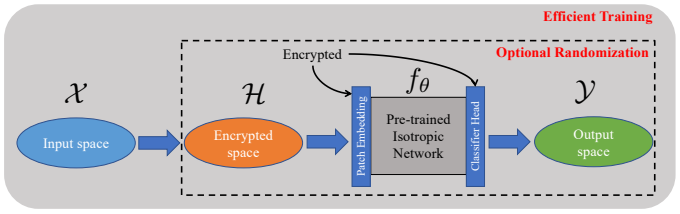


Fig. 2. Overview of image classification scenario with key-based defense.



(a) Previous [20]



(b) Proposed

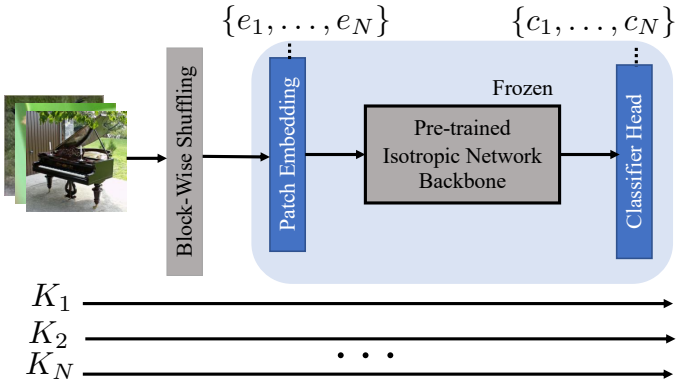
Fig. 3. Difference between previous and proposed key-based defense.

a secret key (*i.e.*, $x \mapsto t(x, K)$). Then, a robust classifier f_θ parameterized by θ ($f_\theta : \mathcal{X} \to \mathcal{Y}$) is trained by using encrypted images $t(x, K) \in \mathcal{H}$.

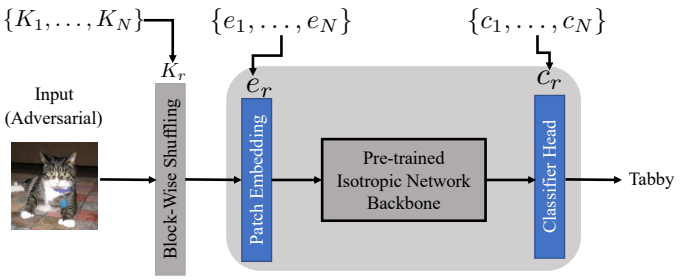
We build upon the robust classifier f_θ from [20] with two major improvements: (1) efficient training and (2) optional randomization. Figure 3 highlights the difference between the previous key-based defense and the proposed one. The proposed defense leverages pre-trained isotropic networks (such as ViT, ConvMixer, etc.), and fine-tunes only patch embeddings and classifier heads (optionally multiple times with different secret keys). When more than one secret patch embedding is used, the proposed defense becomes a radomized defense. The details of training and inference pipelines for the proposed defense are described in the following subsections.

B. Training

The proposed defense utilizes a patch-based pre-trained isotropic network (such as ViT, ConvMixer, etc.). The training pipeline of the proposed defense is depicted in Fig. 4a. To obtain encrypted patch embeddings, $\{e_1, \ldots, e_N\}$ and encrypted classifier heads, $\{c_1, \ldots, c_N\}$, only the patch embedding layer



(a) Training



(b) Inference

Fig. 4. Training and inference pipelines for proposed defense. During training, pre-trained isotropic network backbone is frozen.

and classifier head in a pre-trained isotropic network is fine-tuned for N times by using encrypted images that are encrypted by a block-wise pixel shuffling t with N secret keys $\{K_1, \ldots, K_N\}$. The detailed procedure of t was previously described in Section II-C.

C. Inference

Figure 4b shows the inference process of the proposed defense. We first sample a random key K_r from a pool of keys, $\{K_1, \ldots, K_N\}$, and select the corresponding patch embedding, $e_r \in \{e_1, \ldots, e_N\}$, and classifier head, $\{c_1, \ldots, c_N\}$. During inference, a test image (potentially adversarial) is transformed with K_r. Then, the corresponding patch embedding e_r and classifier head c_r are selected and attached to the pre-trained isotropic network backbone to carry out the forward pass. Note that when N is greater than 1, the proposed defense provides randomization ability for each inference.

D. Threat Model

We focus on hindering adversarial attacks on a classifier f_θ, which takes block-wise transformed input, $t(x, K)$ with a secret key K. To evaluate the proposed defense, we assume a gray-box threat model, where the attacker has full knowledge of the defense mechanism, pre-trained isotropic network backbone weights, and training data, except a secret key, encrypted patch embeddings weights, and classifier heads. The attacker performs state-of-the-art both targeted and untargeted attacks under ℓ_∞ and ℓ_2 norms.

IV. Experiments and Discussion

A. Setup

We carried out experiments for the proposed defense on ImageNet-1k classification. The ImageNet-1k dataset (with 1000 classes) consists of 1.28 million color images for training and 50,000 color images for validation [53]. To implement the proposed defense, we utilized a publicly available pre-trained isotropic network, ConvMixer-768/32[1] model, which uses a patch size value of 7 on 224×224 images. To inline with the patch size of the patch embedding in the pre-trained model, we also set the same value of 7 for the block size value in the key-based transformation (*i.e.*, $M = 7$). We fine-tuned only the patch embedding layer and classifier head with the same learning rate value of 0.01 as in the original training settings without default augmentation for only one epoch. The fine-tuning process was repeated N times to produce N secret patch embeddings and classifier heads with N secret keys. In our experiments, we used $N = 5$.

For evaluation, we deployed AutoAttack (AA) strategy, which is an ensemble of strong, diverse attacks [54]. We used AA's "standard" version with a perturbation budget value of $4/225$ under ℓ_∞ norm and a value of 0.5 under ℓ_2 norm for all attacks. To account for adaptive attacks, we also utilized a "random" version, which implements Expectation over Transformation (EoT) as described in [17].

B. Results

We performed ImageNet-1k classification of both clean images and adversarial examples on the proposed defended classifier.

Transfer Attacks. We generated adversarial examples for two scenarios.

- **Scenario 1.** Since the proposed defense is built on top of the pre-trained model, it is natural to generate adversarial examples directly on the pre-trained model (which includes pre-trained patch embeddings, isotropic network backbone, and a classifier head).
- **Scenario 2.** We assume the attacker has knowledge of the proposed defense mechanism. The attacker may fine-tune the pre-trained model with a guessed key. Thus, we also simulated this scenario to evaluate the proposed defense.

We carried out attacks on both scenarios, and the results are shown in Table I, where (–) denotes "there is no reported results" or "not applicable". Both clean and robust accuracies were calculated on the validation set of ImageNet (50,000) images. Although the un-defended model (plain) achieved the highest accuracy, it was most vulnerable to all attacks. The results suggest that adversarial examples generated on the pre-trained model were not effective on the proposed defense method with secret patch embeddings and a classifier head. Under attacks, the proposed method achieved comparable accuracy for Scenario 1 and higher accuracy for Scenario 2. However, the proposed method reduced clean accuracy, and we shall improve it in our future work.

Adaptive Attacks. When using $N > 1$, the proposed utilizes random secret patch embeddings and a classifier head from N options. Therefore, the proposed defense produces a randomized prediction on each inference. To counter the proposed randomized defense, we assume that the attacker prepares multiple secret patch embeddings as in the proposed defense. We carried out EoT attacks on the attacker's version of the proposed defense with $N \geq 1$ and evaluated the proposed defense. Table II presents the results of such adaptive attacks. For this experiment, attack success rate (ASR) was calculated for randomly selected 1000 images that were correctly classified by the proposed defense model with $N = 5$. Although the EoT attack was improved when using $N = 5$, the ASR was still low.

C. Discussion

Comparison with state-of-the-art methods. For experiments, we implemented the proposed method in a relatively small isotropic model, ConvMixer-768/32, which contains about 21.1 million parameters (*i.e.*, even smaller than ResNet50). We compare the proposed defense with the previous key-based defense and the top 3 adversarially trained (AT) models [57] in terms of clean and robust accuracy. Note that the key-based defense, including the proposed one, has an information advantage (secret key) over attackers. In contrast, AT models are trained to be robust against all adversarial examples under a specific adversarial space. Overall, the proposed defense improved the robust accuracy but reduced the clean accuracy.

Advantages. The main advantage of the proposed method is its feasibility. Much of the adversarial robustness has been done on low-resolution datasets like CIFAR-10 [58]. In contrast, much less is known for ImageNet due to high computing requirements. The proposed method overcomes this barrier by effectively re-using pre-trained weights and only fine-tuning a patch embedding layer and a classifier head.

Limitations. The proposed method shows the potential of making publicly available pre-trained models to be robust against adversarial examples. However, further evaluation and investigation of the proposed method against adaptive attacks is required. In addition, the proposed method in its current form reduces classification accuracy on clean images, although robust accuracy is improved. We shall address these limitations in our future work.

V. Conclusions

In this paper, we proposed a novel key-based defense by using one or more secret patch embeddings and classifier heads with a pre-trained isotropic network. The proposed method is not only efficient in training even for ImageNet but also effective in hindering strong adversarial attacks. Experiment results showed that the proposed method achieved high robust accuracy against a suit of powerful attacks under different

[1] https://github.com/locuslab/convmixer

TABLE I
CLEAN AND ROBUST ACCURACY (%) OF PROPOSED DEFENSE AND STATE-OF-THE-ART CLASSIFIERS

			ℓ_∞		ℓ_2		
	Model	Clean Accuracy	Robust (Scaneraio 1) Accuracy	Robust (Scenario 2) Accuracy	Robust (Scenario 1) Accuracy	Robust (Scenario 2) Accuracy	Defense
Proposed	ConvMixer-768/32 ($N = 1$)	71.98	64.74	**70.65**	**70.16**	**71.47**	Key
	ConvMixer-768/32 ($N = 3$)	71.43	64.75	70.05	69.64	70.86	
	ConvMixer-768/32 ($N = 5$)	71.55	64.74	70.05	69.72	70.92	
	ConvMixer-768/32	**80.15**	0.00	–	0.32	–	Plain
	†ResNet50 ([20])	75.69	**66.95**	10.06	–	–	Key
	Swin-L ([55])	78.92	59.56	–	–	–	AT
	ConvNeXt-L ([55])	78.02	58.48	–	–	–	
	ConvNeXt-L + ConvStem ([56])	77.00	57.70	–	–	–	

† The model was tested according to its stated threat model.

TABLE II
ATTACK SUCCESS RATE (%) OF ADAPTIVE ATTACKS ON PROPOSED DEFENSE

EoT	ASR (AA — ℓ_∞)	ASR (AA — ℓ_2)
EoT ($N = 1$)	8.46	7.13
EoT ($N = 3$)	10.93	7.62
EoT ($N = 5$)	14.77	8.41

norm metrics compared to the previous key-based defense. In addition, we also carried out adaptive attacks to further evaluate the effectiveness of the proposed method. The results also confirmed a low attack success rate suggesting the proposed method was resistant to such adaptive attacks.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant JP21H04907, JP21H01327, and 23K19983, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, Japan.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645–6649.

[5] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[6] B. Biggio, I. Corona, D. Maiorca, *et al.*, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, Springer, 2013, pp. 387–402.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[8] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, Chapman and Hall/CRC, 2018, pp. 99–112.

[9] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, PMLR, 2018, pp. 284–293.

[10] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[11] K. Eykholt, I. Evtimov, E. Fernandes, *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[12] K. Xu, G. Zhang, S. Liu, *et al.*, "Adversarial t-shirt! evading person detectors in a physical world," in *European Conference on Computer Vision*, Springer, 2020, pp. 665–681.

[13] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 819–826.

[14] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the*

2016 acm sigsac conference on computer and communications security, 2016, pp. 1528–1540.

[15] N. Guetta, A. Shabtai, I. Singh, S. Momiyama, and Y. Elovici, "Dodging attack using carefully crafted natural makeup," *arXiv preprint arXiv:2109.06467*, 2021.

[16] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[17] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*, PMLR, 2018, pp. 274–283.

[18] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *Advances in neural information processing systems*, vol. 33, pp. 1633–1645, 2020.

[19] O. Taran, S. Rezaeifar, and S. Voloshynovskiy, "Bridging machine learning and cryptography in defence against adversarial attacks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[20] A. MaungMaung and H. Kiya, "Block-wise image transformation with secret key for adversarially robust defense," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2709–2723, 2021.

[21] H. Kiya, A. MaungMaung, Y. Kinoshita, S. Imaizumi, and S. Shiota, "An overview of compressible and learnable image transformation with secret key and its applications," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, e11, 2022.

[22] A. MaungMaung and H. Kiya, "Encryption inspired adversarial defense for visual classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 1681–1685.

[23] R. Iijima, M. Tanaka, S. Shiota, and H. Kiya, "Enhanced security against adversarial examples using a random ensemble of encrypted vision transformer models," *arXiv preprint arXiv:2307.13985*, 2023.

[24] A. Kerckhoffs, "La cryptographie militaire," *Journal des sciences militaires*, pp. 5–38, 1883.

[25] S. Garg, S. Jha, S. Mahloujifar, and M. Mohammad, "Adversarially robust learning could leverage computational hardness.," in *Algorithmic Learning Theory*, PMLR, 2020, pp. 364–385.

[26] A. A. Rusu, D. A. Calian, S. Gowal, and R. Hadsell, "Hindering adversarial attacks with implicit neural representations," in *International Conference on Machine Learning*, PMLR, 2022, pp. 18 910–18 934.

[27] O. Taran, S. Rezaeifar, T. Holotyak, and S. Voloshynovskiy, "Machine learning through cryptographic glasses: Combating adversarial attacks by key-based diversified aggregation," *EURASIP journal on information security*, vol. 2020, no. 1, pp. 1–18, 2020.

[28] A. MaungMaung and H. Kiya, "Ensemble of key-based models: Defense against black-box adversarial attacks," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, IEEE, 2021, pp. 95–98.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[30] A. Trockman and J. Z. Kolter, "Patches are all you need?" *arXiv preprint arXiv:2201.09792*, 2022.

[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[32] A. Shafahi, M. Najibi, M. A. Ghiasi, *et al.*, "Adversarial training for free!" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[33] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *International Conference on Learning Representations*, 2020.

[34] P. de Jorge Aranda, A. Bibi, R. Volpi, *et al.*, "Make some noise: Reliable and efficient single-step adversarial training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 881–12 893, 2022.

[35] Y. Sharma and P.-Y. Chen, "Attacking the madry defense model with l_1-based adversarial examples," *arXiv preprint arXiv:1710.10733*, 2017.

[36] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International conference on machine learning*, PMLR, 2018, pp. 5286–5295.

[37] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *International Conference on Learning Representations*, 2018.

[38] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *international conference on machine learning*, PMLR, 2019, pp. 1310–1320.

[39] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," *Advances in neural information processing systems*, vol. 30, 2017.

[40] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4422–4431.

[41] H.-T. D. Liu, M. Tao, C.-L. Li, D. Nowrouzezahrai, and A. Jacobson, "Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer," in *International Conference on Learning Representations*, 2019.

[42] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: One hot way to resist adversar-

ial examples," in *International conference on learning representations*, 2018.

[43] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations*, 2018.

[44] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *International Conference on Learning Representations*, 2018.

[45] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1778–1787.

[46] Z. Niu, Z. Chen, L. Li, Y. Yang, B. Li, and J. Yi, "On the limitations of denoising strategies as adversarial defenses," *arXiv preprint arXiv:2012.09384*, 2020.

[47] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in *International Conference on Learning Representations*, 2018.

[48] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.

[49] H. Touvron, P. Bojanowski, M. Caron, *et al.*, "Resmlp: Feedforward networks for image classification with data-efficient training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[50] S. Chen, E. Xie, C. Ge, R. Chen, D. Liang, and P. Luo, "Cyclemlp: A mlp-like architecture for dense prediction," *arXiv preprint arXiv:2107.10224*, 2021.

[51] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9204–9215, 2021.

[52] Q. Hou, Z. Jiang, L. Yuan, M.-M. Cheng, S. Yan, and J. Feng, "Vision permutator: A permutable mlp-like architecture for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1328–1334, 2022.

[53] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[54] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*, PMLR, 2020, pp. 2206–2216.

[55] C. Liu, Y. Dong, W. Xiang, *et al.*, "A comprehensive study on robustness of image classification models: Benchmarking and rethinking," *arXiv preprint arXiv:2302.14301*, 2023.

[56] N. D. Singh, F. Croce, and M. Hein, "Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models," *arXiv preprint arXiv:2303.01870*, 2023.

[57] F. Croce, M. Andriushchenko, V. Sehwag, *et al.*, "Robustbench: A standardized adversarial robustness benchmark," *arXiv preprint arXiv:2010.09670*, 2020.

[58] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Implementation of PLIM on 429MHz LoRa/FSK with improved conversion table

Keita Takeda*, Ryuji Miyamoto*, and Osamu Takyu*
* Shinshu University, Japan
E-mail: takyu@shinshu-u.ac.jp Tel: +81-26-269-5255

***Abstract*—LPWA (Low Power Wide Area) is a communication method for the IoT that enables long-distance communication with low power consumption, and PLIM (Packet Level Index Modulation), an extension of LPWA, enables several bits of additional transmission. However, there is a problem of data loss due to packet collision when transmitting simultaneously and on the same frequency channel. The authors have implemented PLIM on a 429 MHz LoRa/FSK and improved the collision rate by randomly creating and assigning mappings in the implementation. In this paper, the authors conducted experiments on a real device using a method of designing optimal mappings using mathematical optimization, showed the effectiveness of the method and how it differs from other mappings, and evaluated the method in a real device environment.**

I. Introduction

In recent years, demand for IoT has been increasing in terms of social issues such as the COVID-19 and population decline, as well as the spread of smartphones and improved convenience of products. One of these is LPWA (Low Power Wide Area), which is a communication method that enables long-distance communication and low-power consumption, and it is attracting attention for the infrastructure of supporting IoT[1]. In LPWA, there is a problem that throughput cannot be increased due to transmission time constraints. PLIM (Packet Level Index Modulation), which switches frequency channels and transmission timing to transmit signals, is expected to be a solution to this problem[2]. PLIM can transmit several bits of data in addition to the existing LPWA, but a problem unique to PLIM is that when multiple sensors transmit data at the same time and on the same frequency channel, packet collisions occur and thus the sent data is lost. To solve this problem, research has been conducted to reduce the collision rate by improving the mapping between the index and the information to be sent[3].

The authors have implemented PLIM on a 429 MHz LoRa/FSK and have shown through experiments that the packet delivery rate (PDR) rate can be improved by using randomly generated mappings for each sensor, as opposed to the conventional method of using one mapping for all sensors. If the optimal mapping constructed by Ref. [3] is applied the implemented PLIM, the additional improvement of PDR is expected but it has not been clarified, yet. In the actual implementation of LPWA, the unsuspicious incidents which is not considered for constructing the mapping table occur, these may cause the degradation of the PDR. Therefore, the performance evaluation of the PLIM with the optimal mapping table in the actual equipment of LPWA is important.

In this paper, we implement the mapping created using a method in the literature[3] that uses mathematical optimization to create optimal mappings on an actual device and measure the PDR over a certain period of time to evaluate whether the improvement is effective in the actual device environment as well.

II. Description of PLIM

PLIM is a communication method in which information to be sent and received corresponds to an index consisting of a frequency channel and a time slot. The fig.1 below shows an image of PLIM. The vertical and horizontal axes are labeled as frequency channel and time slot, respectively. A mapping table is created in advance that maps the indexes to the information to be sent. When sending information, the appropriate index is selected and sent according to this pre-created mapping table. In the fig.1, the sensor sends a packet with the index 1, which corresponds to the information of the number 1, because it wants to send the information of the number 1. In this case, the communication is synchronized between all sensor nodes and the aggregation station, and all sensors transmit at regular intervals.

In PLIM, when multiple sensors send packets with the same index, there is a problem of packet collision and data cannot be demodulated. In particular, if there is spatial correlation in the data, the information sent between nearby sensors will be the same, and if they communicate using the same mapping, collisions are especially likely to occur. In addition, if the data is also spatially correlated and the same mapping is used, collisions are likely to occur between specific sensors, and there is also the problem of not being able to receive data from some sensors. To address this problem, methods to avoid collisions, such as improved mapping, are needed to aggregate data in a stable manner.

There are some frequency bands for using the LPWA standards. The actual device used in this paper uses the 429 MHz band; 429 MHz offers better radio diffraction than the larger frequency band such as 920 MHz and better communication connectivity even in mountains with many obstacles. In addition, the lack of transmission time constraints allows for real-time and flexible mapping changes between each sensor.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

III. Description of Optimal Mapping

To avoid collisions that occur in PLIM, this experiment uses the optimal mapping design using mathematical optimization found in Ref.[3]. In the optimal mapping, the mapping with the lowest two-packet collision probability is created using the pre-measured data as training data. The information used for mapping in this experiment is the received signal strength indicator (RSSI) value.

The objective function and constraint equations used in the optimal design should be in accordance with the literature[3], and the experimental parameters should be set appropriately. The following is a list of constraint equations and objective functions for actual machine experiments taken from Ref.[3].

$$\min \sum_{k=1}^{K}\sum_{i_1=1}^{I}\sum_{i_2=i_1+1}^{I}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J} P_{i_1,j_1}P_{i_2,j_2}x_{i_1,j_1,k}x_{i_2,j_2,k} \quad (1)$$

$$\text{subject to} \quad x_{i,j,k} \in \{0,1\} \qquad \forall i \in I, \forall j \in J, \forall k \in K$$

$$\sum_{k=1}^{K} x_{i,j,k} = 1 \qquad \forall i \in I, \forall j \in J \quad (2)$$

$$\sum_{j=1}^{J} x_{i,j,k} \leq 1 \qquad \forall i \in I, \forall k \in K \quad (3)$$

$$1 \leq \sum_{j=1}^{J}\sum_{i=1}^{I} x_{i,j,k} \leq I \qquad \forall k \in K \quad (4)$$

Where the i, j, and k denote the sensor number, RSSI number, and index number, respectively. The $P_{i,j}$ is the establishment of the i-numbered sensor observing the jth RSSI value, and $x_{i,j,k}$ means that the jth RSSI value of the ith sensor is assigned to the kth index when it is 1, and 0 means it is not assigned.The equation (2) indicates that all RSSI values are assigned to one of the indices for all sensors, the equation (3) indicates that at most one RSSI value is assigned to one index for all sensors, and the equation (4) indicates that at least one RSSI value is assigned to one index for all indexes, and at most one RSSI value can be assigned to a maximum number of sensors. In accordance with these constraints, the number of RSSI values should be set to the number of indexes in the actual environment as well.

Equation (1) is the objective function of mathematical optimization, which means minimizing the two-packet collision probability. It minimizes the sum of the collision probabilities between any two sensors, which is different from the original collision probability because some collisions are actually more than three packets. However, since the 2-packet collision probability includes the collision probability of 3 or more packets, reducing the 2-packet collision probability also reduces the probability of 3 or more sensors sending packets at the same time.

In addition, in optimal design, as the number of sensors increases, the number of possible combinations of mapping patterns increases, and the calculation time until the optimal solution is obtained increases. Therefore, in order to measure the environment as close as possible to the prior information in the actual experiment, it is necessary to adjust the number of sensors to minimize the computation time as much as possible and to minimize the number of collisions.

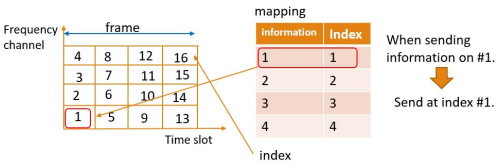


Fig. 1: Image of PLIM

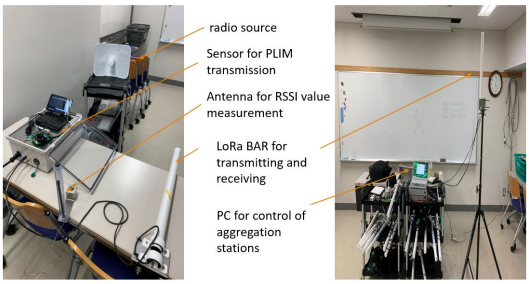


Fig. 2: Construction diagram of sensor (left) and aggregation station (right)

IV. Description of Actual Equipment

This experiment is performed using sensor terminals and aggregation stations implemented with PLIM on 429 MHz LoRa/FSK. The sensor terminals and aggregation stations constructed in Fig.2 are shown.

After receiving a synchronization packet from the aggregation station, the sensor terminal measures the temperature, humidity, and RSSI value, and transmits data at the appropriate timing and frequency channel according to the mapping stored in advance in accordance with the measured data. Since we are dealing with RSSI values, only the antenna for measuring RSSI values is attached to the sensor terminal in Fig.2. Up to 95 mapping patterns can be stored in advance. The information that determines the mapping can be switched by changing a switch on the board of the sensor terminal. The packet contains information such as sensor number, transmission time, index information (selected frequency channel and time slot, mapping number), and measured RSSI value. Up to eight frequency channels and up to 30 time slots can be used, and the mappings to be stored can be changed as needed.

The aggregation station consists of a PC for control, a LoRaBAR(Made by Circuit Design, Inc.) for transmission and reception, and a microcontroller for aggregation and 429 MHz RSSI measurement. The aggregation station can transmit/receive and measure RSSI values using up to eight frequency channels, but since this experiment deals only with transmission/reception using one frequency channel, one LoRaBAR is set up with a tripod. Fig.3 shows a capture of

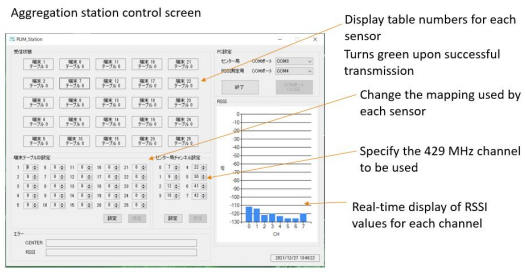


Fig. 3: Control PC screen

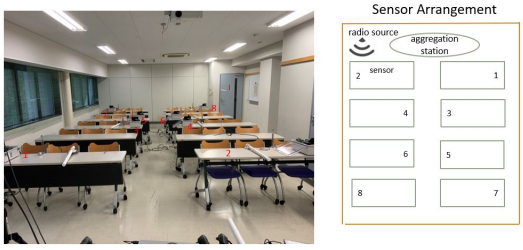


Fig. 4: Sensor arrangement diagram

the screen of the control PC. At the aggregation station, the control PC can change the mapping of each sensor, change the actual frequency channel to be used, and check in real time which sensors have been successfully demodulated and the RSSI values of each channel measured. The aggregated data is saved in a CSV file format on the control PC.

V. Experimental Methods

The experiment was conducted indoors. A diagram of the sensor arrangement is shown in Fig.4. In this experiment, eight sensor terminals are used in a crossed arrangement as shown in the figure. The information to be transmitted by PLIM in this experiment is the RSSI value, which is the received electric field strength of a 5 GHz Wi-Fi radio wave. The location of the radio source is the same for both the preliminary information acquisition and the experimental measurement.The construction of the sensor terminal is shown in the fig.2. Normally, it would be better to treat the antenna for transmission and reception in the upright position, but in this case, it is laid down for convenience and short-distance communication indoors.

The first step is to measure the prior information used for optimal mapping. The locations of the radio sources are shown in the fig.4, and RSSI values are measured at each sensor for 10 minutes to obtain preliminary information. The information obtained here is used to create the optimal mapping. The mapping parameters used to create the mapping are shown in the tableI.When using the optimal design, the number of indices and the number of information to be assigned to be handled for mapping creation must be appropriately determined. In this case, the appropriate quantization width of RSSI values is determined based on the maximum and minimum RSSI values of the prior information obtained; RSSI values are divided into a total of 10 values in 4 dB increments from -97 dBm with a margin between the maximum and minimum measured values. The number of indices is similarly prepared in accordance with the constraint formula, with 10 indices. For the number of indices and RSSI values, the number of indices is set to a small number in consideration of the time required for optimization and the environment in which collisions occur. In addition, since the information to be observed this time is the RSSI values of 5GWi-Fi, the RSSI values to be mapped become large if the quantization width of the RSSI values was small because the observed values are constantly changing. With this in mind, the quantization width of the RSSI value was set to a larger value of 4 dB. For the mappings to be compared, the same parameters as in the optimal design were used, and the RSSI values and indices were randomly assigned to prepare the necessary amount of mappings.

Next, measurements are performed using optimal mappings. For comparison, we also measure random mappings that randomly distribute the information to be sent to the index for each sensor, and fixed mappings that use a common mapping for each sensor. The measurement time is 10 minutes for each, and the number of transmissions is 120 times per sensor for each. For the prepared mappings, the indexes to be assigned are one frequency channel and 10 time slots for a total of 10 indexes this time. This is to ignore the effect of radio wave leakage into other channels when multiple frequency channels are used. As for the radio sources, the power is not turned off until the measurement in all mappings is completed. The position of the radio source is fixed at the one point shown in the fig.4. This is to avoid that when the radio source is moved, the frequency of RSSI values measured by each sensor is scattered, resulting in unbalanced frequency values with few characteristics, and thus reducing the merit of using the optimal design.

TABLE I: Mapping Parameters

Parameters	Value
Range of RSSI	4[dB]
Number of indexes	10
Number of RSSI values	10
Minimum RSSI value	-94[dBm]
Maximum RSSI value	-71[dBm]

TABLE II: experimental results

Mapping	PDR[%]
Optimal Mapping	93.75
Random Mapping	59.58
Fixed mapping	51.98

VI. Experimental Results

The tableII shows the measurement results. The packet delivery rate (PDR) is the percentage of successfully demodulated packets in all transmissions. Comparing the PDR, the optimal mapping improves the PDR by more than 30 percent compared to the random mapping. Random mapping also improves the PDR by several percent compared to fixed mapping, but the optimal mapping results in a much larger improvement. This result indicates that the optimal design rearranged the RSSI values to avoid collisions based on the frequency values of the RSSI values that serve as prior information. In addition, since this experiment is conducted in an indoor environment with a single radio source and a fixed radio source, the possible RSSI values are fixed within a certain range, so it is thought that collisions are more likely to occur with fixed mapping because the RSSI values are similar between sensors that are far from the radio source. When using random mapping, packets from sensors that are far from the radio source can be saved to some extent, but since they are reordered randomly, RSSI values that are frequently transmitted among other sensors are assigned to the same index, which is thought to make it difficult to improve the success rate of transmissions.

Fig.5 shows the indexes selected during transmission when optimal mapping is used, and Fig.6 shows a graph of the indexes selected during transmission when fixed mapping is used. This graph shows which index is selected by each sensor when transmitting. It can be seen that in the fixed case, the indexes used by each sensor for transmission are concentrated on the same index, whereas in the case of using the optimal design, the indexes used by each sensor for transmission are scattered. This also indicates that the optimal design allows each sensor to assign different RSSI values to different indices, which are frequently measured with prior information, resulting in fewer sensors selecting the same index and fewer packet collisions.

The fig.7 shows a graph summarizing the PDR for each sensor when each mapping is used. It can be seen that some sensors have particularly low PDR when using fixed or random mapping, while all sensors have improved PDR when using optimal mapping. As mentioned earlier, since the target of this measurement is the RSSI value of a stationary radio source, it is easy for each sensor to detect a certain RSSI value, and furthermore, since radio waves have spatial correlation, sensors closer to the radio source are more likely to detect a high RSSI value and sensors further away are more likely to detect a low RSSI value, resulting in many sensors detecting the same data. This means that sensors that detect the same data are more likely to be detected. Therefore, in fixed and random mapping, packet collisions frequently occur among certain sensors, resulting in a low PDR for some sensors. In addition, in fixed mapping, sensors 7 and 8, which are far from the radio source, have particularly low PDR, whereas in random mapping, sensors 7 and 8 have improved their PDR and sensor 4 has decreased its PDR. On the other hand, in

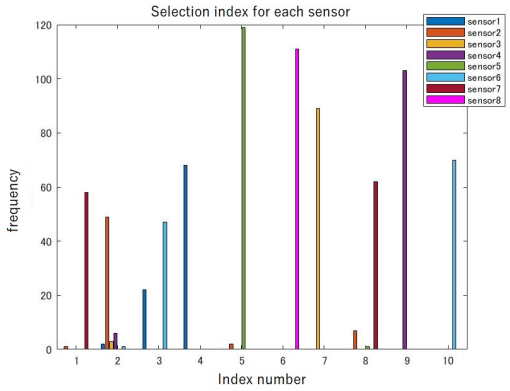


Fig. 5: Selection index per sensor (optimal mapping)

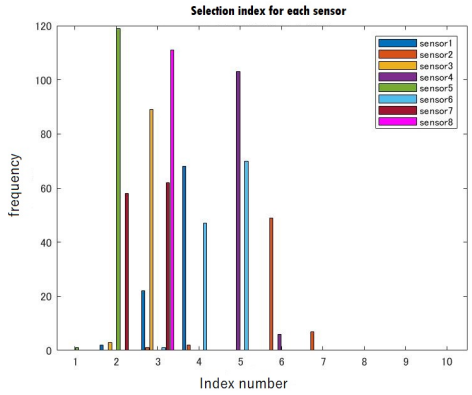


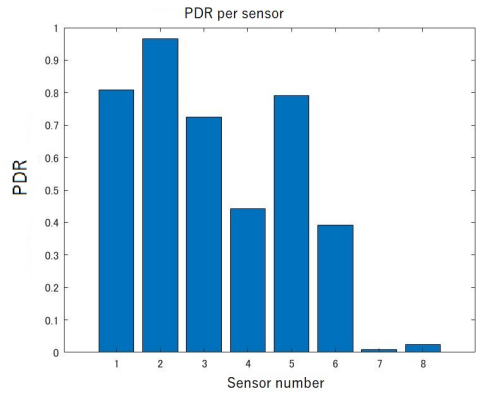
Fig. 6: Selection index per sensor (fixed mapping)

the optimal mapping, the frequently measured RSSI values are assigned to different indices for each sensor, as described earlier, resulting in a lower packet collision rate and a higher PDR for all sensors. This is one significance to the random mapping of the optimal design, indicating that it is possible to collect information at all sensors without discarding any one sensor.

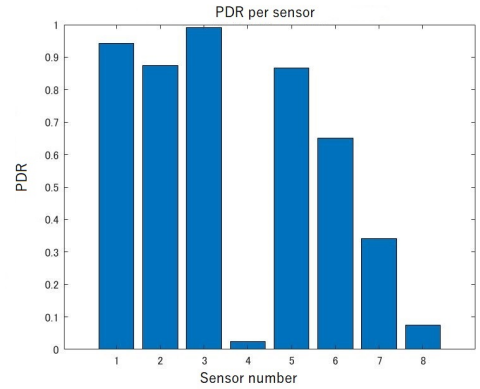
VII. Conclusions

In this paper, we implemented optimal mapping using mathematical optimization on an actual PLIM device and evaluated the comparison between fixed and random mapping in terms of PDR. Significant improvement in PDR was observed when using optimal mapping compared to fixed or random mapping, demonstrating the effectiveness of optimal mapping. This indicated that a lower collision rate also occurs in the actual environment under the same environment as the prior information, because optimal mapping is designed to avoid assigning frequently measured data to the same index based on the trend of data distribution using prior learning. We also showed that the PDR is improved for all sensors, indicating that the optimal design is clearly superior to the method of randomly creating mappings.

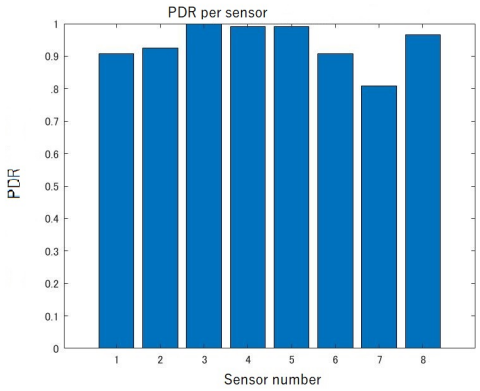
In this environment, the number of indexes and sensors was limited in order to ignore the time required for optimization



(a) PDR per sensor (fixed mapping)



(b) PDR per sensor (random mapping)



(c) PDR per sensor (optimal mapping)

Fig. 7: PDR per sensor

and the effects of using multiple frequency channels. Attempts to increase the number of indices and sensors will be the subject of future research.

ACKNOWLEDGEMENT

This research and development was supported by SCOPE (JP205004001, JP235004002) of the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] B. A. Homssi, A. Al-Hourani, K. Magowe, J. Delaney, N. Tom, J. Ying, H. Wolf, S. Maselli, S. Kandeepan, K. Wang, and K. M. Gomez, "A framework for the design and deployment of large-scale lpwan networks for smart cities applications," *IEEE Internet of Things Magazine*, vol. 4, no. 1, pp. 53–59, 2021.

[2] K. Adachi, K. Tsurumi, A. Kaburaki, O. Takyu, M. Ohta, and T. Fujii, "Packet-level index modulation for lorawan," *IEEE Access*, vol. 9, pp. 12 601–12 610, 2021.

[3] R. Miyamoto, O. Takyu, H. Fujiwara, K. Adachi, M. Ohta, and T. Fujii, "Performance evaluation of highly efficient information collection methods by trend analysis of sensor information using pre-learning," pp. 55–60, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Numerical Performance Evaluation of $\ell_1 - \ell_2$ Sparse Reconstruction Using Optical Analog Circuit

Soma Furusawa* , Kazunori Hayashi*, Kaito Kameda†, and Ryo Hayakawa†
* Graduate School of Informatics, Kyoto University, Kyoto, Japan, Tel/Fax: +81-75-753-9690/+81-75-753-9691
E-mail: furusawa.soma.82h@st.kyoto-u.ac.jp
E-mail: hayashi.kazunori.4w@kyoto-u.ac.jp
† Graduate School of Engineering Science, Osaka University, Osaka, Japan, Tel/Fax: +81-6-6850-6580
E-mail: kameda@sip.sys.es.osaka-u.ac.jp
E-mail: hayakawa.ryo.es@osaka-u.ac.jp

***Abstract*—Iterative shrinkage thresholding algorithm (ISTA), fast ISTA (FISTA) and alternating direction method of multipliers (ADMM) based algorithm are well known convex optimization algorithms based on proximal splitting for $\ell_1 - \ell_2$ optimization problem in compressed sensing. In our previous study, aiming at the implementation with optical analog devices, we have proposed constant inertial FISTA (CIFISTA) and an approximated ADMM based algorithm, which do not require the operation of division. In this paper, we consider to realize those algorithms with optical analog devices and show concrete configurations of analog circuits using optical devices. We have conducted computer experiments to evaluate the performance of each algorithm, taking into account the effects of the additive noise introduced by optical amplifiers assuming the implementation using optical devices for common optical fiber communications systems. From the numerical results, it has been confirmed that both the proposed CIFISTA and the approximated ADMM based algorithm can achieve faster convergence than ISTA, while the degradation of the steady state mean squared error (MSE) due to the circuit noise is much smaller than that of the conventional ADMM algorithm.**

I. Introduction

In order to achieve high-speed signal processing with low power consumption, the use of analog signal processing technology instead of conventional digital signal processing with electric devices has been attracting much attention [1]. Especially, optical analog device can compute the matrix-vector product by just passing optical signals through the device, and thus the implementation of machine learning techniques, such as deep learning, with the optical analog device has been investigated recently [2], [3], [4].

Compressed sensing is a method of reconstructing unknown sparse vector from underdetermined linear measurement which may include observation noise [5], [6], and various reconstruction algorithms have been proposed so far. One of well-known signal reconstruction approaches is the convex optimization method using proximal splitting [7]. Specifically, fast iterative shrinkage thresholding algorithm (FISTA) [8], which is the fast version of iterative shrinkage thresholding algorithm (ISTA) [9], is one of most famous algorithms of the approach. FISTA has a convergence rate of $O(k^{-2})$, which is known to be optimum for the first order method [10], [11], where k is the number of iterations in the algorithm. It is also known that alternating direction method of multipliers (ADMM) algorithm [12] can achieve comparable convergence rate to FISTA for $\ell_1 - \ell_2$ optimization problem.

In our previous study, we have considered to realize FISTA and the ADMM algorithm with optical analog devices. Since FISTA requires the operation of division by a dynamic value in the update equation of the inertial parameter, while it is difficult to calculate the division with optical analog device, we have proposed constant inertial FISTA (CIFISTA) [13] by simply setting the inertial parameter of FISTA to be a constant value. Moreover, we have developed an approximated ADMM based algorithm [14] by using different formulation from the existing ADMM algorithm for $\ell_1 - \ell_2$ optimization problem. Although both the conventional ADMM and the proposed approximated ADMM do not require any division operation in the update equations, and thus both algorithms could be implemented with optical devices, but the proposed algorithm has an advantage over the conventional ADMM that it is free from matrix inversion. In the previous studies [13], [14], we have numerically confirmed that both proposed algorithms can achieve comparable convergence performance to FISTA and the existing ADMM algorithm. However, the optical analog signal processing specific aspects, such as signal attenuation by optical devices or additive noise introduced by optical amplifiers, have not been taken into consideration.

In this paper, we evaluate the performance of the proposed CIFISTA and the proposed approximated ADMM algorithm to see the applicability of the proposed algorithms to the implementation with optical analog devices. In order to evaluate the power of the additive noise added at the optical amplifier, we present concrete optical circuit configuration of each algorithm and clarify the required amplifier power gain in the optical analog circuit. Since additive noise power at the amplifier depends on the amplifier power gain as well as the signal bandwidth, we can determine the noise power for each algorithm from the derived circuit configuration and the assumption of the signal bandwidth of 10 GHz, which corresponds to common optical fiber communications systems. From the numerical results, it can be observed that the CIFISTA, the approximated ADMM,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

and the conventional ADMM achieve comparable convergence rates, which is much faster than the rate of ISTA, even in the presence of the additive noise at the amplifier (circuit noise). On the other hand, when it comes to the degradation due to the circuit noise, ISTA achieves the lowest steady state mean squared error (MSE). However, the proposed CIFISTA and the proposed approximated ADMM still achieve acceptable steady state MSE, while the conventional ADMM suffers from larger degradation due to the circuit noise. After all, it can be concluded that both the proposed CIFISTA and the proposed approximated ADMM are suited for the implementation with optical analog circuits.

II. $\ell_1 - \ell_2$ Optimization Problem

In this paper, we consider the problem of estimating an unknown sparse vector $\boldsymbol{x} \in \mathbb{C}^N$ from the underdetermined linear measurement $\boldsymbol{y} \in \mathbb{C}^M$ obtained with a known sensing matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n} \in \mathbb{C}^M \quad (M < N), \tag{1}$$

where $\boldsymbol{n} \in \mathbb{C}^M$ is a zero mean white observation noise vector. Such a problem is called compressed sensing [15], and in particular, the $\ell_1 - \ell_2$ optimization problem given by

$$\hat{\boldsymbol{x}}_{\ell_1-\ell_2} = \arg \min_{\boldsymbol{x} \in \mathbb{C}^N} \left(\lambda \|\boldsymbol{x}\|_1 + \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right), \tag{2}$$

is often used to reconstruct the sparse signal. Here, $\lambda > 0$ is a regularization parameter that balances between ℓ_1 sparse regularization term of $\|\boldsymbol{x}\|_1$ and data fidelity term of $\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

Several algorithms have been proposed for solving the $\ell_1 - \ell_2$ optimization problem, but, in this paper, we focus on the convex optimization based algorithms using proximal splitting because they can achieve good reconstruction performance with low computational complexity in general.

III. Convex Optimization based Algorithms for $\ell_1 - \ell_2$ Optimization Problem

One of the simplest and most popular reconstruction algorithms will be ISTA shown in Algorithm 1. In the algorithm, $S_\delta(\cdot)$ is the proximity operator of ℓ_1 norm called as soft thresholding function and is defined by

$$S_\delta(x) := \begin{cases} (|x| - \delta)\dfrac{x}{|x|} & |x| \geq \delta \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

for $^\forall x \in \mathbb{C}$, $^\forall \delta > 0$ [16]. When the argument of $S_\delta(\cdot)$ is a vector, $S_\delta(\cdot)$ becomes a vector obtained by applying (3) for each element of the input vector. Since all operations involved in ISTA except for soft thresholding function* can be achieved with optical devices, ISTA could be implemented with optical analog devices, however, it suffers from slow rate of convergence.

*In this paper, we assume soft thresholding function is calculated in electrical domain by using efficient OEO (optical signal - electrical signal - optical signal) converter devices [17].

Algorithm 1 ISTA

Input: $\boldsymbol{x}[0]$, $\gamma > 0$.
Output: Estimate of $\boldsymbol{x}$
  for $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{x}[k+1] = S_{\gamma\lambda}(\boldsymbol{x}[k] - \gamma \boldsymbol{A}^{\mathrm{H}}(\boldsymbol{A}\boldsymbol{x}[k] - \boldsymbol{y}))$
  end for
  return $\boldsymbol{x}[k+1]$

Algorithm 2 FISTA

Input: $\boldsymbol{x}[0]$, $\boldsymbol{z}[0]$, $t[0] > 0$, $\gamma > 0$.
Output: Estimate of $\boldsymbol{x}$
  for $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{x}[k+1] = S_{\gamma\lambda}(\boldsymbol{z}[k] - \gamma \boldsymbol{A}^{\mathrm{H}}(\boldsymbol{A}\boldsymbol{z}[k] - \boldsymbol{y}))$
    $t[k+1] = \dfrac{1 + \sqrt{4t[k]^2 + 1}}{2}$
    $\boldsymbol{z}[k+1] = \boldsymbol{x}[k+1] + \dfrac{t[k] - 1}{t[k+1]}(\boldsymbol{x}[k+1] - \boldsymbol{x}[k])$
  end for
  return $\boldsymbol{x}[k+1]$

FISTA is a faster version of ISTA, where the acceleration is achieved by the employment of the update equation using not only the estimate in the previous iteration but also the estimate of one more previous iteration. Detailed steps of FISTA are shown in Algorithm 2. It is known that FISTA can achieve the optimum convergence rate among the first order methods [10], [11]. However, it is difficult to implement FISTA with optical analog devices, because it includes the operation of division with a dynamic value in the update equation of $\boldsymbol{z}[k+1]$.

In order to make FISTA realizable with optical analog devices, in our previous study [13], we have proposed CIFISTA given in Algorithm 3, in which $\frac{t[k]-1}{t[k+1]}$ is replaced by a constant inertial parameter c. In our preliminary numerical experiments, we have confirmed that CIFISTA can achieve comparable convergence performance to FISTA with a proper choice of the inertial parameter c.

ADMM is also a popular convex optimization algorithm using proximal splitting, and it is possible to solve the $\ell_1 - \ell_2$ optimization problem with ADMM as shown in Algorithm 4. ADMM could be implemented with optical analog devices because it does not require the division operation. However, one of major drawbacks of the conventional ADMM will be the necessity of the matrix inversion, which may restrict the

Algorithm 3 CIFISTA (proposed)

Input: $\boldsymbol{x}[0]$, $\boldsymbol{z}[0]$, $\gamma > 0$, $0 \leq c \leq 1$.
Output: Estimate of $\boldsymbol{x}$
  for $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{x}[k+1] = S_{\gamma\lambda}(\boldsymbol{z}[k] - \gamma \boldsymbol{A}^{\mathrm{H}}(\boldsymbol{A}\boldsymbol{z}[k] - \boldsymbol{y}))$
    $\boldsymbol{z}[k+1] = \boldsymbol{x}[k+1] + c(\boldsymbol{x}[k+1] - \boldsymbol{x}[k])$
  end for
  return $\boldsymbol{x}[k+1]$

Algorithm 4 ADMM

Input: $\boldsymbol{z}[0]$, $\boldsymbol{v}[0]$, $\eta > 0$.
Output: Estimate of $\boldsymbol{x}$
  for $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{x}[k+1] = \left(\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A} + \frac{1}{\eta}\boldsymbol{I}\right)^{-1} \cdot \left(\boldsymbol{A}^{\mathrm{H}}\boldsymbol{y} + \frac{1}{\eta}(\boldsymbol{z}[k] - \boldsymbol{v}[k])\right)$
    $\boldsymbol{z}[k+1] = S_{\eta\lambda}(\boldsymbol{x}[k+1] + \boldsymbol{v}[k])$
    $\boldsymbol{v}[k+1] = \boldsymbol{v}[k] + \boldsymbol{x}[k+1] - \boldsymbol{z}[k+1]$
  end for
  return $\boldsymbol{x}[k+1]$

Algorithm 5 Approximated ADMM (proposed)

Input: $\boldsymbol{x}[0]$, $\boldsymbol{z}[0]$, $\boldsymbol{v}[0]$, $\gamma > 0$, $\rho > 0$.
Output: Estimate of $\boldsymbol{x}$
  for $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{x}[k+1] = S_{\gamma\lambda/\rho}\Big(\boldsymbol{x}[k] - \gamma\boldsymbol{A}^{\mathrm{H}}\Big(\boldsymbol{A}\boldsymbol{x}[k] - \boldsymbol{z}[k] + \frac{1}{\rho}\boldsymbol{v}[k]\Big)\Big)$
    $\boldsymbol{z}[k+1] = \frac{1}{1+\rho}\left(\boldsymbol{y} + \rho\left(\boldsymbol{A}\boldsymbol{x}[k+1] + \frac{1}{\rho}\boldsymbol{v}[k]\right)\right)$
    $\boldsymbol{v}[k+1] = \boldsymbol{v}[k] + \rho(\boldsymbol{A}\boldsymbol{x}[k+1] - \boldsymbol{z}[k+1])$
  end for
  return $\boldsymbol{x}[k+1]$

application to large-scale problems†.

In our previous study [14], we have derived a novel ADMM based algorithm obtained by using a different formulation from the existing ADMM. Update equations of the proposed algorithm are shown in Algorithm 5. With the new formulation, the $\ell_1-\ell_2$ optimization problem appears in one of update equations of ADMM, therefore, we have approximated the update equation with a single iteration of ISTA in order to avoid the double loop algorithm. This is the reason why the algorithm is called "approximated" ADMM. Note that the proposed approximated ADMM does not include matrix inversion, and also that the approximated ADMM can achieve comparable convergence performance to the conventional ADMM in our preliminary numerical experiments.

IV. Proposed Optical Analog Circuit Configurations of $\ell_1 - \ell_2$ Optimization Algorithms

Since all reconstruction algorithms discussed in the previous section require iterative signal processing, it would be a natural approach to implement each algorithm with optical analog circuit, where one iteration of the algorithm corresponds to the optical signal propagation of one round in the circuit. As for the optical devices to implement the algorithms, we assume following optical devices are available: signal splitter (SS), adder, subtractor, matrix-vector multiplier, attenuator, amplifier, and delay element. Among the devices, we focus on the impact of SS, adder, and subtractor, because they are

†Note that we do not assume that the matrix inversion is calculated with optical analog devices.

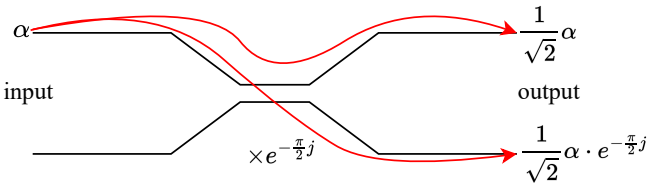


Fig. 1. Beam splitter (BS)

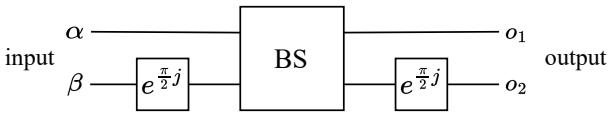


Fig. 2. Optical adder / subtractor

dominant devices in the circuit and are realized with a common optical device, namely the beam splitter (BS), which introduces large signal attenuation.

BS has two input ports and two output ports as depicted in Fig. 1. If we input signal $\alpha \in \mathbb{C}$ from one of the ports, then the output signal from the output port of the same side as the input is $\frac{1}{\sqrt{2}}\alpha$, while the output signal from the other side port is $\frac{1}{\sqrt{2}}\alpha \cdot e^{-\frac{\pi}{2}j}$, where j is an imaginary unit. Thus, if we give the phase shift of $e^{\frac{\pi}{2}j}$ to the lower port in Fig. 1, then BS can be used to realize SS, while the signal power of each output becomes half of the input signal power.

Optical adder and subtractor also can be realized by using BS and phase shifter. Fig. 2 shows the configuration of the adder and subtractor. Based on the property of BS, if we input signals $^{\forall}\alpha, \beta \in \mathbb{C}$ from the two input ports, then output signals from the two output ports will be

$$o_1 = \frac{1}{\sqrt{2}}\alpha + \frac{1}{\sqrt{2}}\beta \cdot e^{\frac{\pi}{2}j} \cdot e^{-\frac{\pi}{2}j} = \frac{1}{\sqrt{2}}(\alpha + \beta), \tag{4}$$

$$o_2 = \frac{1}{\sqrt{2}}\alpha \cdot e^{-\frac{\pi}{2}j} \cdot e^{\frac{\pi}{2}j} + \frac{1}{\sqrt{2}}\beta \cdot e^{\frac{\pi}{2}j} \cdot e^{\frac{\pi}{2}j} = \frac{1}{\sqrt{2}}(\alpha - \beta). \tag{5}$$

Thus, the addition and the subtraction can be realized with the optical analog devices in Fig. 2, while the power of the output signal becomes half of the desired signal as in the case of SS.

Optical analog circuits for the signal reconstruction have several SSs, adders and subtractors, therefore, the optical signal in the circuit suffers from signal attenuation during the propagation. However, at the input of soft thresholding function, which is required in all algorithms considered in this paper, the signal power must be an appropriate value because the soft thresholding function is a nonlinear function. Therefore, we insert optical amplifiers in the circuits before the soft thresholding function in order to compensate the signal attenuation due to BSs. Note that the number of required BSs in the optical circuit depends on the algorithms, and thus the required power gain of the amplifier also depends on the algorithms. Moreover, some optical attenuators must be

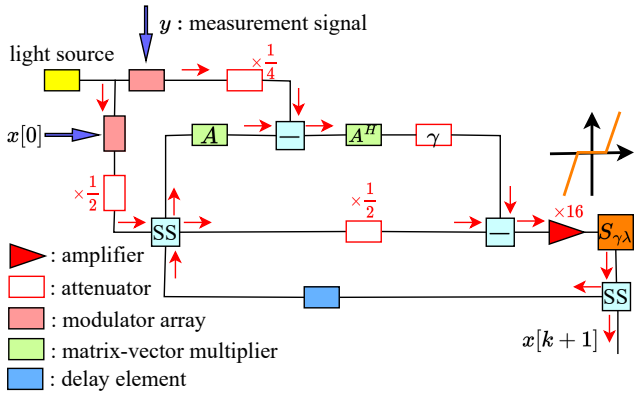


Fig. 3. Optical analog circuit configuration of ISTA

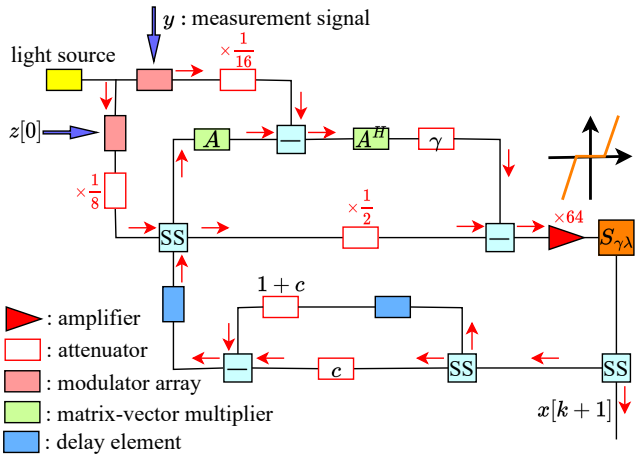


Fig. 4. Optical analog circuit configuration of CIFISTA

inserted at appropriate locations in the circuit to balance the signal power of each term before the addition or subtraction operations. Based on these considerations, we have derived the optical analog circuit configurations of ISTA, CIFISTA, conventional ADMM, and proposed approximated ADMM in Figs. 3 to 6, respectively.

V. Estimate of Additive Noise Power at Optical Amplifier

In the previous section, we have shown that optical amplifiers are necessary in the optical analog circuits of the $\ell_1 - \ell_2$ optimization algorithms. Generally speaking, additive noise is inevitable in the amplification of the signals. Therefore, the additive noise at the amplifier could be one of major performance limiting factors of the proposed approach. In this section, we estimate the power of the additive noise at the amplifier based on the noise model of optical amplifier.

According to [18], the power spectral density of amplified spontaneous emission (ASE) noise for a single erbium doped fiber amplifier (EDFA), which is one of most popular optical amplifiers, is given by

$$G_{\mathrm{ASE}} = F(G-1)h\mu, \tag{6}$$

where F is the amplifier noise figure (NF), G is the amplifier power gain, h is Planck's constant, and μ is the frequency. Thus, the noise power is given by the product of the power spectral density G_{ASE} and the signal bandwidth.

In order to estimate the noise power, we have assumed to use optical devices for common optical fiber communications systems. Specifically, if we assume the wavelength of $1550\,\mathrm{nm}$ corresponding to the standard single-mode fiber, then the frequency will be

$$\mu = \frac{3.0 \times 10^8}{1550 \times 10^{-9}} \approx 1.94 \times 10^{14}. \tag{7}$$

Moreover, by using signal bandwidth of $10\,\mathrm{GHz}$, which is typical for optical fiber communications, the ASE noise power added the optical amplifier with the power gain of G is given by

$$G_{\mathrm{ASE}} \times 10\,\mathrm{GHz} = (G-1) \cdot 2.56 \times 10^{-9}, \tag{8}$$

where the ideal value of $F = 2$ is employed for NF. Some specific values of the additive noise power at the optical amplifier with different power gain G are summarized in Table I.

TABLE I
Estimation of Additive Noise Power for Different Amplifier Power Gain

amplifier power gain (G)	additive noise power
8	1.79×10^{-8}
16	3.84×10^{-8}
32	7.94×10^{-8}
64	1.61×10^{-7}
128	3.25×10^{-7}
256	6.53×10^{-7}

VI. Numerical Results

Here, we evaluate the performance of ISTA, CIFISTA, the conventional ADMM, and the proposed approximated ADMM in the presence of additive noise at the optical amplifier (circuit noise), which is assumed to be white Gaussian noise, via computer experiments.

In the experiments, we set the dimension of unknown sparse vector $\boldsymbol{x}$ to be $N = 500$ and that of linear measurement vector be $M = 200$. We set the number of non-zero elements of $\boldsymbol{x}$ as $K = 25$, and the non-zero elements are randomly generated with independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian (CSCG) distribution with zero-mean and variance $\sigma_{\boldsymbol{x}}^2 = 1$. Note that typical signal power in the optical fiber transmission is around 0.001, thus we have used 1,000 times larger values for the additive noise power at the optical amplifier than the values in Table I. The measurement matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ is generated by randomly choosing M rows of N-point unitary discrete Fourier transform (DFT) matrix. The measurement noise $\boldsymbol{n} \in \mathbb{C}^M$

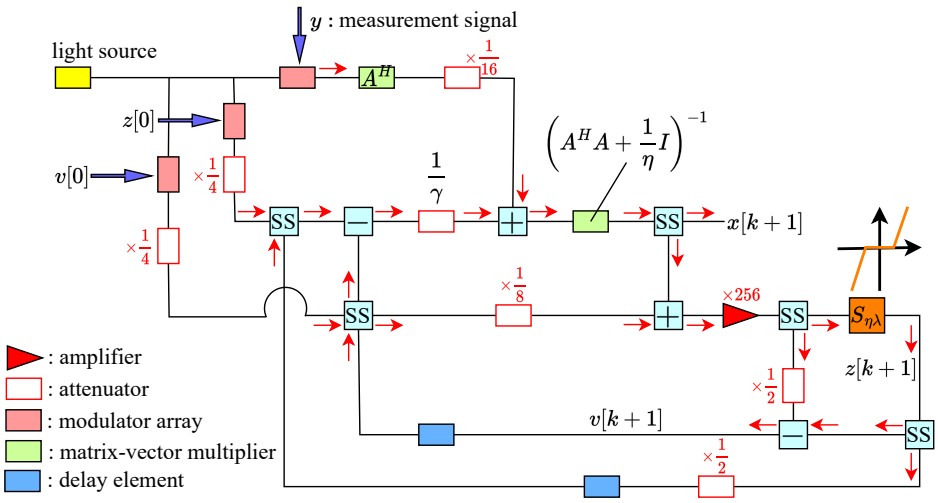


Fig. 5. Optical analog circuit configuration of conventional ADMM

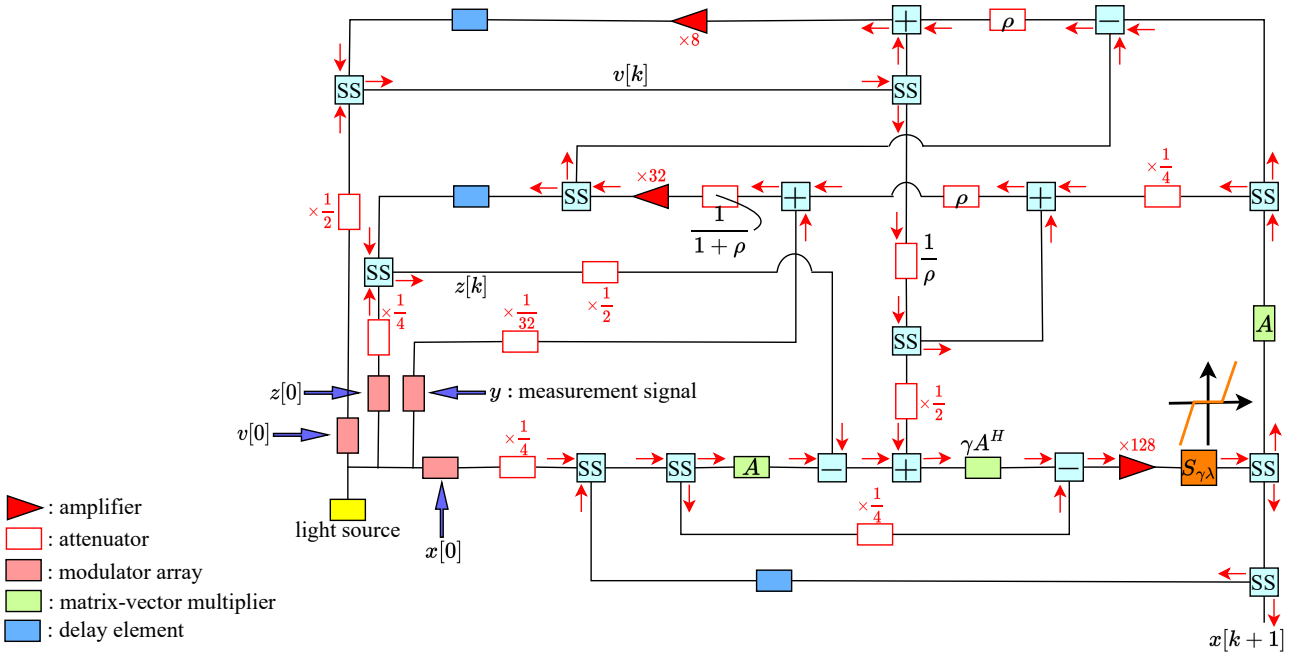


Fig. 6. Optical analog circuit configuration of approximated ADMM

is generated from zero-mean i.i.d. CSCG distribution with the variance corresponding to the measurement signal-to-noise ratio (SNR) of 30dB. We set the regularization parameter for the $\ell_1-\ell_2$ optimization problem to be $\lambda = 0.02$, the constant inertial parameter of CIFISTA $c = 0.65$, and other parameters $\gamma = 0.99$, $\eta = 4.5\times0.99$, and $\rho = 0.20$. The initial values of the vectors in the algorithms are all set to $\boldsymbol{x}[0] = \boldsymbol{z}[0] = \boldsymbol{v}[0] = \boldsymbol{0}$. Under these settings, the MSE performance of each algorithm is evaluated by averaging the results obtained with 100 trials.

Fig. 7 shows the MSE performance of ISTA, CIFISTA, the conventional ADMM, and the proposed approximated ADMM in the absence of the circuit noise. In this case, the performance of FISTA is also shown in the same figure for comparison purpose. From the figure, we can see that all algorithms achieve almost the same steady state MSE, while the convergence rate of ISTA is much slower than other four algorithms. CIFISTA achieves similar convergence rate to FISTA in spite of the simplified algorithm with a constant inertial parameter. Also, the proposed approximated ADMM can achieve comparable convergence rate to the conventional ADMM without using matrix inversion.

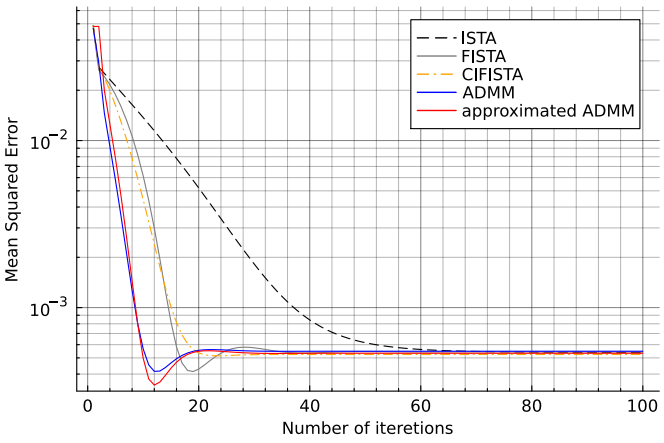


Fig. 7. Comparison of convergence characteristics (no circuit noise)

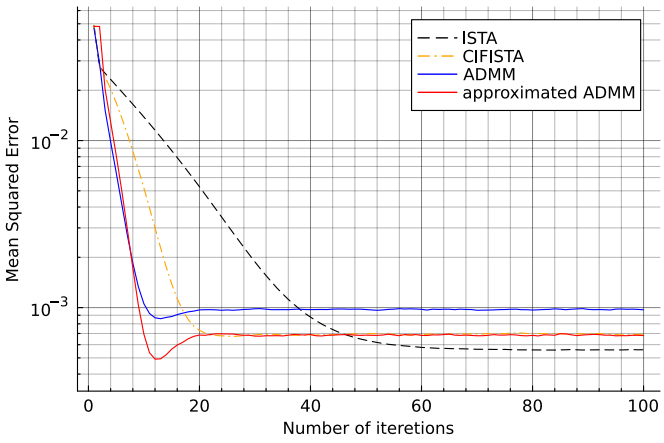


Fig. 8. Comparison of convergence characteristics (with circuit noise)

Fig. 8 shows the MSE performance of ISTA, CIFISTA, the conventional ADMM, and the proposed approximated ADMM in the presence of the circuit noise. As you can see from the figures, the degradation of the steady state MSE of ISTA is almost negligible. On the other hand, a certain degradation of the steady state MSE can be observed for CIFISTA and the approximated ADMM, while the conventional ADMM suffers from larger degradation of the steady state MSE. This could be mainly because the power gain of the optical amplifier is the smallest for the case with ISTA and is the largest for the case with the conventional ADMM. Moreover, the soft thresholding function also could have a certain impact on the steady state MSE performance, since it has a function to eliminate small noises, and only in the case of the conventional ADMM, half of the optical amplifier output with the largest power gain does not pass through the soft thresholding function.

Table II summarizes the comparison results of the convergence performance obtained by numerical experiments and the complexity of the optical analog circuits for the four algorithms. Judging from the balance between the steady state MSE and the convergence rate, both CIFISTA and the proposed approximated ADMM could be good choices for the implementation with optical analog circuits.

VII. Conclusions

In this paper, we have evaluated the performance of convex optimization based algorithms using proximal splitting for the $\ell_1 - \ell_2$ optimization problem assuming the implementation with optical analog devices. We have presented concrete configurations of optical analog circuits to realize ISTA, CIFISTA, the conventional ADMM, and the approximated ADMM. Moreover, we have estimated the additive noise power at the optical amplifier for each optimization algorithm assuming the implementation using optical devices for common optical fiber communications systems. From the numerical results obtained in the presence of the circuit noise, it can be concluded that both the proposed CIFISTA and the proposed approximated ADMM are more suited for the implementation with optical analog circuits than ISTA or the conventional ADMM.

Future work includes the investigation of the attenuation caused by other optical devices, and the implementation of soft thresholding function without using OEO converter.

Acknowledgment

This work has been supported in part by JST CREST JPMJCR21C3.

References

[1] N. Shlezinger, G. C. Alexandropoulos, M. F. Imani, Y. C. Eldar and D. R. Smith, "Dynamic Metasurface Antennas for 6G Extreme Massive MIMO Communications," in *IEEE Wireless Communications*, vol. 28, no. 2, pp. 106-113, April 2021.

[2] Wetzstein, G., et al, "Inference in artificial intelligence with deep optics and photonics," *Nature* 588, 39-47 2020.

[3] C. Huang, et al, "Prospects and applications of photonic neural networks," *Advances in Physics: X*, 7:1, 2022.

[4] Zhou, H., et al, "Photonic matrix multiplication lights up photonic accelerator and beyond." *Light Sci Appl* 11, 30 2022.

[5] M. Elad "Sparse and Redundant Representations," *Springer*, 2010.

[6] I. Rish and G. Grabarnik, "Sparse Modeling: Theory, Algorithms, and Applications," *CRC Press*, 2014.

[7] M. Nagahara, "Sparsity methods for systems and control," *now Publishers*, 2020.

[8] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 188-202, 2009.

[9] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413-1457, 2004.

[10] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Springer, 2004.

[11] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, Vol. 140, No. 1, pp. 125-161, 2013.

[12] Eckstein, J., Bertsekas, D.P., "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming* 55, 293-318 1992.

TABLE II
CHARACTERISTIC COMPARISON OF ALGORITHMS AS AN OPTICAL ANALOG CIRCUIT

	ISTA	CIFISTA	ADMM	approximated ADMM
Convergence Rate	slow	fast	fast	fast
Degradation due to Circuit Noise	very small	small	large	small
Circuit Complexity	simple	simple	complex	complex

[13] K. Kameda, R. Hayakawa, K. Hayashi and Y. Iiguni, "Performance Evaluation of FISTA with Constant Inertial Parameter," *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Chiang Mai, Thailand, 2022, pp. 1713-1719.

[14] R. Lin and K. Hayashi, "An Approximated ADMM based Algorithm for $\ell_1 - \ell_2$ Optimization Problem," *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Chiang Mai, Thailand, 2022, pp. 1720-1724.

[15] D. L. Donoho, "Compressed sensing," in *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306, April 2006.

[16] M. V. W. Zibetti, E. S. Helou, R. R. Regatte, and G. T. Herman, "Monotone FISTA With Variable Acceleration for Compressed Sensing Magnetic Resonance Imaging," in *IEEE Transactions on Computational Imaging*, vol. 5, no. 1, pp. 109-119, March 2019.

[17] K. Nozaki, S. Matsuo, T. Fujii, et al. , "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," *Nat. Photonics* Vol. 13, pp. 454-459, 2019.

[18] P. Poggiolini, G. Bosco, A. Carena, V. Curri, Y. Jiang, and F. Forghieri, "A Detailed Analytical Derivation of the GN Model of Non-Linear Interference in Coherent Optical Transmission Systems," *arXiv*: 1209.0394, 2014.

Assessing the Effects of Filtering Processing on Pulse Wave Transit Time Measured by Photoplethysmography from Earlobe

Shangdi Liao*, Haipeng Liu†, Dingchang Zheng† and Fei Chen*
* Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
E-mail: fchen@sustech.edu.cn Tel: +86-755-88018554
† Research Centre for Intelligent Healthcare, Coventry University, Coventry, United Kingdom
E-mail: haipeng.liu@coventry.ac.uk Tel: +44 (0)78 4642 4479

Abstract — **The waveform of a photoplethysmography (PPG) signal depends on individual physiological conditions. Pulse transition time (PTT), derived from electrocardiogram (ECG) and PPG waveform features, is an important physiological parameter for blood pressure estimation. Filtering can change PPG signal waveform and the timing of PPG feature points. This work aimed to quantitatively investigate the filtering-induced PTT changes at earlobe in healthy subjects of different ages and with different PTT definitions. The ECG and earlobe PPG signals were recorded simultaneously at a sample rate of 2500 Hz for 120 s from 100 healthy adults in 5 age groups (20-29, 30-39, 40-49, 50-59, and over 60 years). The PPG signals were preprocessed using band-pass filtering (pass and stop bands: >0.5 Hz and <0.2 Hz for high-pass filter, <20 Hz and >30 Hz for low-pass filter) and then low-pass filtered (pass and stop bands: <3 Hz and >5 Hz). The PTT difference between the preprocessed and filtered PPG signals was calculated and analyzed. Four pulse feature points were defined and extracted (i.e., maximum peak (MP), valley point (VP), maximal first derivative (M1D), and maximal second derivative (M2D)). The results showed significant effect of PTT definition on filtering-induced PTT relative difference (p<0.001). It was found that the age group (50-59) and the MP definition had the minimum filtering-induce PTT changes. This work suggested that the physiological factors including PTT definition and age should be considered in the PTT-based blood pressure estimation methods.**

I. Introduction

Photoplethysmography (PPG) signal reflects the volumetric changes in microcirculation. PPG signals can be easily measured at different body sites using wearable sensors. Recent years witnessed a rapid development of wearable PPG sensors and relevant applications in daily monitoring of cardiovascular parameters, e.g., heart rate, heart rate variability, blood oxygen saturation, and blood pressure [1, 2]. Pulse transit time (PTT), which is approximated as the interval between the R wave of electrocardiogram (ECG) and a PPG waveform feature (often the end-of-diastolic valley) in the same cardiac cycle [3], is negatively related to blood pressure, providing the possibility of cuffless blood pressure monitoring using wearable PPG sensors [4].

Many physiological factors, including age, body posture, and body site of measurement, can influence the waveform of a PPG signal [5-7]. Compared to those measured from wrist and fingers, the PPG signal from earlobe is less influenced by tissue composition and has less motion artifacts, and using sensors designed specifically for acquisition from ear might bring comparable or even better results than the PPG from fingertip [8]. Aging can influence the biomechanical properties of arteries, veins, and microcirculation, leading to the changes in PPG waveform [3].

In many wearable applications, the original PPG signals are filtered before feature extraction. Filtering can change the waveform of PPG signals and the timing of feature points [9]. Especially, the finite impulse response (FIR) and infinite impulse response (IIR) [10,11] filters are widely applied in PPG signal processing [12]. The nonlinear phase response of the IIR filter may deform the PPG waveform, leading to an inaccuracy of PTT estimation. Currently, the PTT-based blood pressure estimation methods have limited accuracy, where filtering may be a neglected factor. There is a lack of quantitative evaluation of the effect of physiological factors on IIR filtering-induced PTT changes. To address the research gap, in this pilot study, we aimed to quantitatively investigate the effects of age and PTT definition on the IIR filtering-induced PTT.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

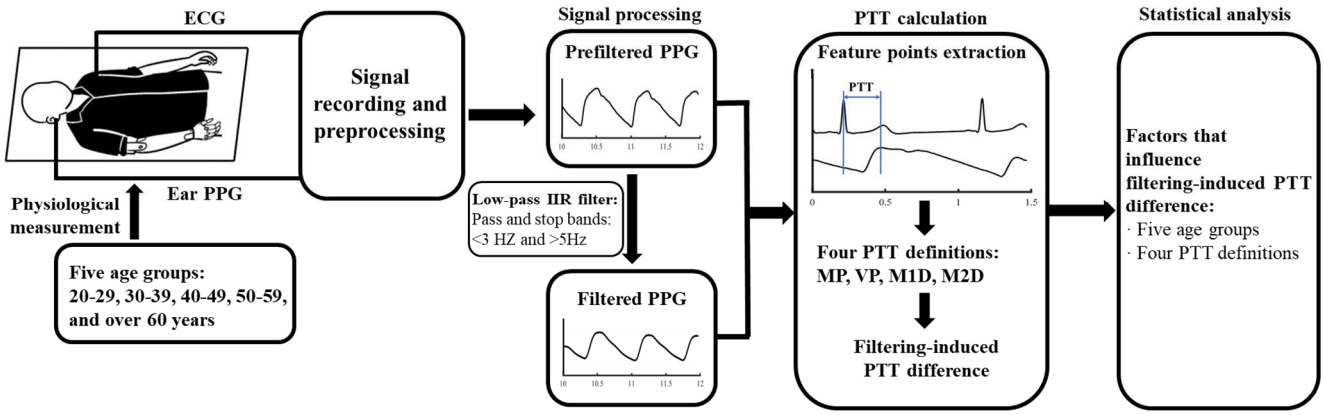


Fig. 1 Flowchart of PPG signal processing and analysis.

II. METHODS

A. Subject and Data Collection

One hundred participants (age: 44±14 (range: 20 to 71) years; 48 males and 52 females) were recruited from staff, students and their relatives in Newcastle Hospitals and Newcastle University with written informed consent. The subjects were equally distributed in five different ranges of age: 20-29, 30-39, 40-49, 50-59, and over 60 years old, i.e., 20 subjects in each age group. No participants suffered cardiovascular disease before. The experimental procedures involving human subjects described in this work complied with the principles in the Declaration of Helsinki, and received ethical permission from the Faculty Research Ethics Panel at Anglia Ruskin University.

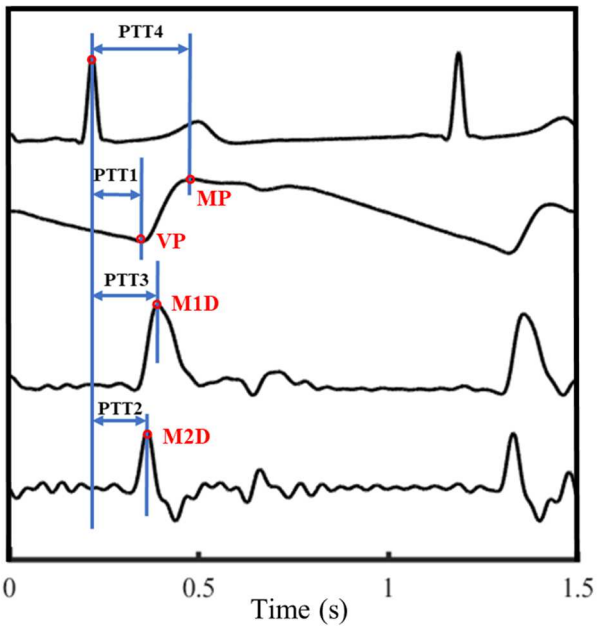


Fig. 2 Four definitions of pulse transit time. From top to bottom, the signals are ECG, PPG, the first derivative of PPG, and the second derivative of PPG.

The experiment began in a thermostatic room maintaining the temperature at 23±1 ℃. In order to stabilize the cardiovascular system, each subject rested in a supine posture on a couch for 5 minutes before signal recording. ECG and PPG signals from the right earlobe were recorded simultaneously for 120 seconds at a sampling rate of 2500 Hz when all signals were clear and stable. During the whole experiment, the subjects kept normal breathing without body movement.

B. Signal Preprocessing and Filtering

The recorded data were imported to MATLAB (R2021b; The MathWorks Inc. Natick, USA) for signal processing. A bandpass filter (0.1-35Hz) was used to preprocess the ECG signal. As shown in Fig. 1, the recorded raw PPG signals were preprocessed with a high-pass IIR filter (1 zero and 10 poles, passband: >0.5 Hz, stopband: <0.2 Hz) to remove the low-frequency and DC components (e.g., 'baseline wander'). A low-pass IIR filter (1 zero and 16 poles, passband: <20 Hz, stopband: >30 Hz) was then used to remove the high-frequency noises including the 50 Hz power line and electrophysiological noises. The preprocessed PPG signals were further filtered with a low-pass IIR filter (1 zero and 13 poles, passband: <3 Hz, stopband: >5 Hz) which was applied in our early study where the details of the filters could be found [11].

C. Definition of PTT

PTT is usually defined as the time interval between the R-peak of the ECG and a reference point on systolic PPG signal segment. The reference point can be derived from different PPG features (e.g., end-of-diastolic valley, systolic peak, etc.), which leads to different PTT values [7]. This work calculated the PTTs defined by four different PPG feature points: end-of-

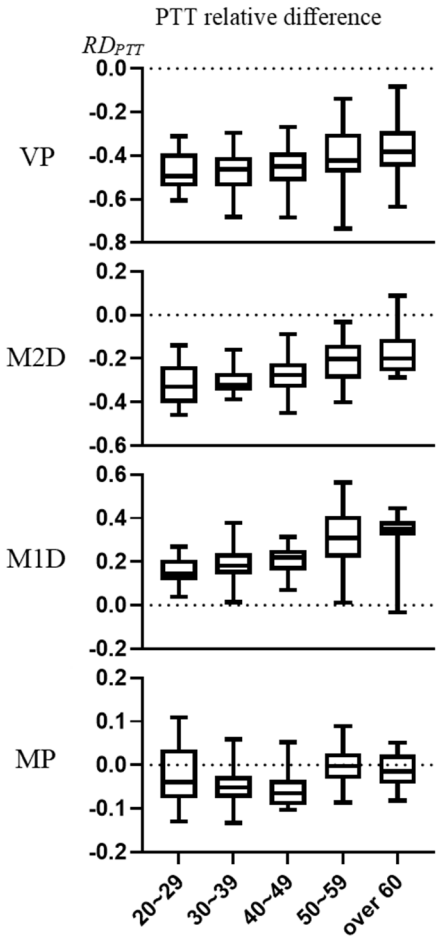


Fig. 3 Box and whisker plots of the PTT relative difference (i.e., RDPTT) of each age group at different PTT definition. The medians are represented by the horizontal lines within the boxes while the first and third quartiles are represented by the box boundaries, whereas the maximum and minimum values are represented by the whiskers.

diastolic valley (VP), systolic peak (MP), the peak of the first derivative of PPG (M1D) and the peak of the second derivative of PPG (M2D), as illustrated with PTT1, PTT4, PTT2 and PTT3, respectively, in Fig. 2.

D. Extraction of Feature Points

On ECG signals, the R wave peak was detected as the maximal value in a cardiac cycle using the Pan Tompkins method [13]. The four PPG feature points (peak, valley, maximal first derivative, maximal second derivative) were extracted from the prefiltered and filtered PPG signals and their derivatives (Fig. 2). The valley and peak detection algorithm were based on the calculation of differentials. Considering the high sampling frequency (2500 Hz), the error caused by the approximation is very limited ($<4 \times 10^{-4}$ s for the timing of any feature point). The details can be found in our published work [11]. Considering the noisy PPG signals in some cardiac cycles, only the valleys within 50-500 ms after the ECG R-peak

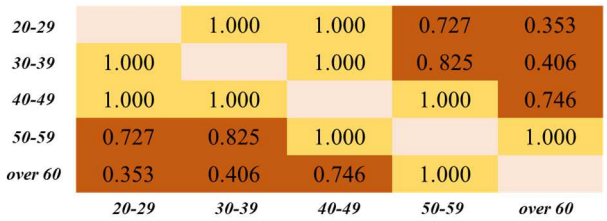


Fig. 4a Results of Dunn's Kruskal–Wallis multiple comparisons between the five age groups regarding the PTT relative difference. The red color indicates the significant effect, and the yellow color indicates a less significant effect.

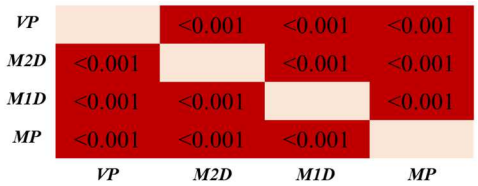


Fig. 4b Results of Dunn's Kruskal–Wallis multiple comparisons between the three types of PTT definition regarding the PTT relative difference. The red color indicates the more significant effect, and the yellow color indicates a less significant effect.

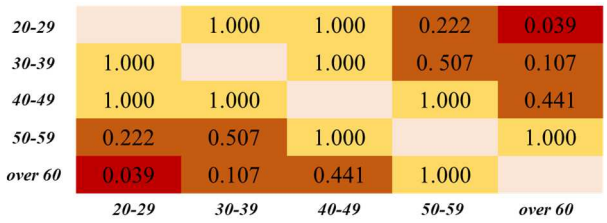


Fig. 4c Results of Dunn's Kruskal–Wallis multiple comparisons between the five age groups regarding the PTT relative difference (except MP). The red color indicates the significant effect, and the yellow color indicates a less significant effect.

(i.e., 50 ms≤ PTT ≤500 ms) were selected for analysis. For each PPG signal, the PTT was calculated as the mean value of PTTs of the selected cardiac cycles.

E. Statistical Analysis

For each subject, the filtering-induced PTT difference was calculated between the PTT values derived from filtered and preprocessed PPG signals. The PTT relative difference was calculated as:

$$RD_{PTT}=(PTT_{filtered}-PTT_{preprocessed})/PTT_{preprocessed} \quad . \qquad (1)$$

Statistical analysis was performed using SPSS (Version 24.0, IBM Corp) and R programming language (R Core Team, 2021). Considering the data size, Shapiro–Wilk test was performed to investigate the normality of data distribution. For normally distributed data (defined as $p>0.05$ in Shapiro–Wilk test). To investigate if there was any significant effect of age, PTT's definition, and their interaction on the filtering-induced PTT relative difference, the analysis of variance (ANOVA) or

Scheirer–Ray–Hare test was performed. ANOVA was performed on normally distributed data where the homogeneity of variance was satisfied (defined as p>0.05 the Levene's test), otherwise the Scheirer–Ray–Hare test was performed as a substitute. To further investigate the difference between age groups, or between PTT's definitions, and to identify the age group and the PTT definition with the highest reliability (i.e., with the least filtering-induced PTT relative difference), the post hoc analysis was performed, i.e., least significant difference and Dunn's Kruskal–Wallis multiple comparisons for ANOVA and the Scheirer–Ray–Hare test, respectively. In the Dunn's Kruskal–Wallis multiple comparisons, the p-value was adjusted via the Benjamini–Hochberg method [14].

III. Results

A. Effect of Age and PTT Definition on the Filtering-induced Differences

Fig. 3 shows the distribution of IIR filtering-induced PTT relative difference (i.e., RD_{PTT}) in five different age groups and four PTT definitions. After the filtering processing, PTTs were shortened except for M1D, with a consecutive trend in different age groups. The relative difference distributions did not satisfy the homogeneity of variance (p=0.16). The relative difference distributions did not satisfy the homogeneity of variance (p<0.05 in Levene's test). Therefore, the Scheirer-Ray-Hare test was performed. There was a significant effect of the PTT definition on the filtering-induced PTT relative difference (p<0.001), whereas the effect of age and their interaction are not significant (p=0.086 and p=0.999, respectively).

B. Comparison between Age Groups and PTT Definitions

The Fig. 4a shows the results of Dunn's Kruskal–Wallis multiple comparisons between the five age groups regarding the PTT relative difference. There are no significant difference between any paired age groups. In Fig. 4b, significant differences are observed between any two PTT definitions (p<0.001 for all). Overall, the age group (50-59) was least influenced by the filtering processing, indicating the highest robustness against filtering-induced PTT difference.

C. Comparison between Age Groups: MP Excluded

It was observed that all type's PTT relative difference showed age dependence except MP (Fig. 3). To investigate the correlation between age and filtering-induced PTT relative difference in the remaining three PTT definitions, the Scheirer-Ray-Hare test was performed without MP results. There was a significant effect of PTT definition on the filtering-induced PTT relative difference (p=0.014). Fig. 4c shows the results of Dunn's Kruskal–Wallis multiple comparisons between the five age groups except MP regarding the PTT relative difference. There is a significant difference between the youngest and oldest age groups (i.e., 20-29 and over 60) in PTT relative difference (p<0.05).

IV. Discussion and Conclusion

The results in this work showed that: 1) filtering processing could considerably change the PTT values. We observed that PTTs were shortened except for M1D after filtering; 2) filtering-induced PTT changes depended on age and the definition of PTT, where the effect of age was consecutive (i.e., higher PTT changes in younger subjects) except MP. Among different PTT definitions, MP showed the minimum filtering-induce PTT changes.

Vascular stiffness increases with age, which has a significant effect on PPG signal waveform. However, no quantitative observation has been disclosed. The results in this work fill this gap and highlight the importance of considering filtering-induced inaccuracy in PTT-based blood pressure estimation, especially when IIR filtering is applied on the earlobe PPG signals where the PTT is shorter compared with that from the fingertip or wrist.

At present, the filter parameters of many commercial wearable PPG sensors are unrevealed. In published works, the filtering parameters are not uniform and narrow frequency bands were widely used (e.g., 0.5-4 Hz [15], 0.5-4 Hz [16], and 0.8-4 Hz [17]). 0.5-10 Hz was suggested for estimation of heart rate from PPG signals [4]. And 0.5-10 Hz with 60 Hz notch filter was used to estimate blood pressure using the PPG signal obtaining from earlobe [18]. Our previous study demonstrated that age and measurement site had significant impacts on the time shift in feature points between prefiltered and filtered fingertip PPG signals [19]. Our results here highlighted the importance of considering the age and PTT definition in PTT measurement, providing the reference for future works on more accurate PTT measurement via a pathway of age- and definition-adjustment towards individualization.

In conclusion, the results in this work showed that the filtering-induced PTT difference was significantly different among different PTT types or among different age groups. The physiological factors including PTT definition and age should be considered in PTT-based blood pressure estimation using wearable sensors.

Acknowledgment

This work was supported by Shenzhen Key Technology Program Funding (JSGG20220831103803006).

REFERENCES

[1] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *International Journal of Biosensors and Bioelectronics*, vol. 4, no. 4, pp. 195–202, 2018.

[2] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, p. R1, 2007.

[3] M.H. van Velzen, A.J. Loeve, S.P. Niehof, and E.J. Mik, "Increasing accuracy of pulse transit time measurements by automated elimination of distorted photoplethysmography waves," *Medical and Biological Engineering and Computing*, vol. 55, no. 11, pp. 1989–2000, 2017.

[4] R. Mukkamala, J.O. Hahn, O.T. Inan, L.K. Mestha, C.S. Kim, H. Töreyin, and S. Kyal, "Toward ubiquitous blood pressure monitoring via pulse transit time: theory and practice," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 8, pp. 1879-1901, 2015.

[5] J. Allen, and A. Murray, "Age-related changes in the characteristics of the photoplethysmographic pulse shape at various body sites," *Physiological Measurement*, vol. 24, no. 2, p. 297, 2003.

[6] Y. Maeda, M. Sekine, T. Tamura, T. Suzuki, and K.I. Kameyama, "Position dependency in photoplethysmographic sensor-comparison of adjoining PPG signals in light sources and measurement sites," *Journal of Life Support Engineering*, vol. 23, no. 3, pp. 124–129, 2011.

[7] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, and D. Zheng, "Quantitative comparison of photoplethysmographic waveform characteristics: effect of measurement site," *Frontiers in Physiology*, vol. 10, p. 198, 2019.

[8] Špulák, D., Čmejla, R. and Fabián, V., "Experiments with blood pressure monitoring using ECG and PPG," in *Proc. Conf. Technical Computing Bratislava*, pp. 1-5.

[9] J. Allen and A. Murray, "Effects of filtering on multisite photoplethysmography pulse waveform characteristics," *Computers in Cardiology*, pp. 485-488, 2004.

[10] U. Yadav, S.N. Abbas, and D. Hatzinakos, "Evaluation of PPG biometrics for authentication in different states," in *2018 International Conference on Biometrics (ICB)*, pp. 277-282, 2018.

[11] H. Liu, J. Allen, S.G. Khalid, F. Chen, and D. Zheng, "Filtering-induced time shifts in photoplethysmography pulse features measured at different body sites: the importance of filter definition and standardization," *Physiological Measurement*, vol. 42, 074001, 2021.

[12] C. Park, H. Shin, and B. Lee, "Blockwise PPG enhancement based on time-variant zero-phase harmonic notch filtering," *Sensors*, vol. 17, no. 4, p. 860, 2017.

[13] L. Sathyapriya, L. Murali and T. Manigandan, "Analysis and detection R-peak detection using modified Pan-Tompkins algorithm," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, pp. 483–487, 2014.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289-300, 1995.

[15] L. Wang, B. P. Lo and G. Z. Yang, "Multichannel reflective PPG earpiece sensor with passive motion cancellation," in *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, no. 4, pp. 235-241, 2007.

[16] S. Vogel, M. HÜlsbusch, T. Hennig, V. Blazek and S. Leonhardt, "In-ear vital signs monitoring using a novel microoptic reflective sensor," in *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 882-889, 2009.

[17] M. Z. Poh, K. Kim, A. Goessling, N. Swenson and R. Picard, "Cardiovascular monitoring using earphones and a mobile device," in *IEEE Pervasive Computing*, vol. 11, no. 4, pp. 18-26, 2012.

[18] J. Gu-Young, K. H. Yu, and K. Nam-Gyun, "Continuous blood pressure monitoring using pulse wave transit time," *In Proc. ICCAS 2005 Int. Conf. on Control.* pp. 834-837, 2005.

[19] S. Liao, F. Chen, H. Liu., and D. Zheng, "Effect of filtering on pulse wave transit time Measured by photoplethysmography," in *Computers in Cardiology,* 2022.

Efficient Dictionary and Grid-Based Framework for Answering Durable k-Nearest Neighbor Queries on Time Series Data

Bagus Jati Santoso, Dwi Prasetya Armunanta, Baskoro Adi Pratomo, Hudan Studiawan
Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
E-mail: bagus@if.its.ac.id, armunanta.17051@mhs.its.ac.id, baskoro@if.its.ac.id, hudan@its.ac.id

***Abstract* — With the massive amount of data generated, a data engineering method is needed to process the data to obtain new information, which can later be used for further understanding and business decisions. One of the data processing methods is the k-nearest neighbor (*k*-NN) query. The *k*-NN queries aim to identify *k*-objects with the shortest distance based on object references over the network. However, a critical limitation of the *k*-NN query lies in its inability to process a query when an object has proximity durability to a reference object at a particular time interval. This limitation arises since the *k*-NN query does not consider the durability and time interval variables in its calculation algorithm.**

This article discusses the problem of time interval-based durable *k*-NN queries on time series data by introducing the algorithmic framework and employing dictionary and grid data structure to process these queries. The proposed algorithm, which is a dictionary and grid-based durable *k*-NN queries method, incorporates parallel computing techniques to efficiently search for a collection of *k* objects that exhibit durability in their proximity to a reference object within a predefined time interval. The effectiveness and efficiency of the devised algorithms were assessed through extensive testing using both real-world and synthetic datasets.

Keywords— ***continuous, k-NN queries, multidimensional, time series, grid, data engineering, data banks***

I. INTRODUCTION

One of the data processing methods is the *k*-nearest neighbor (*k*-NN) queries [10][11]. The k-NN query intends to search for a set of *k*-objects that exhibit the shortest distance in relation to the reference objects. However, this method is inadequate when confronted with a scenario wherein an object has proximity durability to a reference object within a specific time interval, as the conventional *k*-NN query is unable to handle such instances.

As an illustration, consider an individual denoted as person A as an asymptomatic carrier and has been infected by a respiratory virus. Person A engaged in conversation with three companions, namely persons B, C, and D. Each of them has a different incoming and leaving time from the meeting room. The three of them were found to be in breach of the health protocols, as they failed to maintain a safe physical distance and did not wear masks. The distance, incoming, and leaving time are shown in Table 1.

TABLE I. INCOMING TIME, LEAVING TIME, AND DISTANCE TO PERSON A

Persons	Incoming Time	Leaving Time	Distances from A
A	2:00 PM	3.00 PM	0 cm
B	1.45 PM	2.45 PM	90 cm
C	2:00 PM	2.55 PM	20 cm
D	2.55 PM	3.15 PM	50 cm

Let us assume the hypothetical scenario where a respiratory virus exhibits a significantly high transmission probability, which is happened when individuals engage in no-health protocol interactions lasting a minimum duration of 15 minutes within a distance of the asymptomatic carrier less than 1 meter. Now suppose a healthcare professional aims to identify individuals who are potentially susceptible to contracting a virus. To pinpoint those most at risk of infection, the healthcare professional intends to search for the two most suspicious infected persons, employing a threshold duration of 15 minutes.

To tackle this problem, we utilize a search strategy based on the *k*-NN query approach, with query point A, $k = 2$, and the durability threshold of 25% (equivalent to 15 minutes from the query point, or A's duration, which is 1 hour). Please note that the search duration is based on the active time of the query point, Person A, from 2 to 3 PM (the incoming and leaving time of A). In addition, to simplify the discussion, it is assumed that the motion of the object remains stationary throughout the duration of this particular example. Later, we use time series geolocation data to illustrate the dynamic situation.

From Table 1, we include B and C as the candidates for the query answers, since they interact with A in 45 and 55 minutes, respectively. Further, we can exclude D since he only interacts from 2.55 to 3.00 PM (5 minutes) and only fulfills 8 % of the search duration to be part of k-*NN* w.r.t. query point person A. Hence, B and C are the answer to the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

durable k-*NN* query with query point A, *k=2*, and durability threshold = 25%.

In this work, we formulate several issues as follows. First, we propose the design and implementation of data structures and algorithms to address the query of durable *k*-NN based on time intervals on time series data. Second, we conduct the performance evaluation of the proposed data structures and algorithms in answering the query of durable-*k*NN query based on time intervals on time-series data. Moreover, in this query, an indexing data structure, namely grid-based indexing, is highly utilized to perform indexing on the data, thereby expediting and facilitating the query process.

We organize our work presentation as follows. Chapter I delves into the introduction and contributions of this work. Chapter II discusses the related works. We address the problem environment of this work in Chapter III. Chapter IV discusses the solutions to the identified issues. Subsequently, Chapter V provides a comprehensive discussion of the performance evaluation results. Finally, Chapter VI concludes the findings and contributions of this study.

II. RELATED WORKS

The work by Ding, Lu, and Liu [5] focuses on developing algorithms for *k*NN queries and *k*NN-Joins in relational databases. Their objective is to design solutions that can be implemented using existing SQL operators without modifying the underlying database engine. Li, Liu, and Wang [6] propose a grid-based index structure suitable for parallel implementations, providing an efficient way to index and search large-scale and high-dimensional data.

In the realm of spatial databases, Chunming, Xiangfeng, and Liqiang [7] introduce the R-Grid algorithm, which enhances the performance of *k*NN retrieval through the utilization of a grid-based spatial index. This approach significantly speeds up the search process in location-based nearest-neighbor queries. Zhang, Yang, and Qin [8] present the GDCF algorithm, which employs bitmap indexing and union-find techniques to support efficient neighbor grid queries. Their approach reduces redundancies and enhances the efficiency of grid-based clustering algorithms.

In the broader context of machine learning, the article by Onel [9] provides an overview of the *kNN* algorithm as a simple and easy-to-implement supervised learning technique for classification and regression tasks.

Several studies have been conducted on the topic of durable queries. In the study by Wang et al. [2], the focus is on identifying objects with sustained quality over time in historical time series databases. The authors propose a framework for processing durable top-k queries, which efficiently retrieves objects that exhibit durable quality over time. Gao et al. [12] also present efficient algorithms for processing durable top-k queries specifically on temporal data.

Another study by Semertzidis [13] introduces a hybrid approach that effectively filters potential matches based on both structure and time in temporal graphs. This approach aims to enhance the identification of durable top-k objects within instant-stamped temporal records. Additionally, Gao et al. [14] further investigate the durable top-k framework and its application to instant-stamped temporal records.

III. PROBLEM BACKGROUND

A. Time Series with Geolocation Data

Time series with geolocation data is a collection of geolocation data (latitude and longitude) acquired through repetitive time measurements. The ubiquity of time series data is undeniable, as time constitutes an inherent facet of all observable phenomena. Table 2 shows time series data pertaining to the positional attributes of the observed object.

TABLE II. THE EXAMPLE OF TIME SERIES GEOLOCATION DATA OF PERSON A

Latitude	Longitude	Timestamp (hh:mm:ss)
-7.53824172425915	112.23824890861432	10:00:01
-7.53824172421915	112.23824890857432	10:00:30
-7.53824172426415	112.23824890861932	10:01:00
-7.53824172425415	112.23824890860932	10:01:30

B. Grid-Based Indexing

The grid index [1] serves as a data structure employed for indexing the data based on their location. Figure 1 presents a visual representation of the two-dimensional grid index.

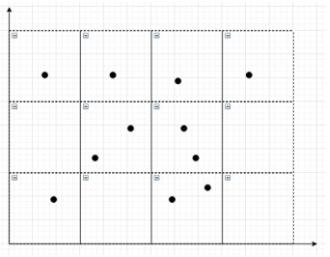


Fig.1 Grid Index

The data is indexed and mapped according to its grid cell. Grid cell *(i,j)* on column *i* and row *j* (started from left bottom corner) contain all objects with a range of coordinate *x* equals [i . θ, (i+1). θ] and coordinate *y* with values [*j. θ*, *(*j+1). *θ*] where *θ* is the length of the grid cell. Moreover, an object with coordinates (*x,y*) is located on grid cell *ci,j* where $i = \lfloor x/\theta \rfloor$ dan $j = \lfloor y/\theta \rfloor$ [1].

C. Durable Query

The concept of a durable query refers to a query that searches for objects with durable quality consistency within a specific time interval in a given dataset, using a time unit [2][12][13][14]. Durable quality can be represented by scores, weights, or other numerical attributes. In a durable query, a threshold is utilized as a boundary value for objects that

exhibit a certain level of resilience in their occurrence (durability) within a time interval.

Drawing from the illustration of the respiratory virus previously, the durable quality is represented by the distance between the individual and the query point (person A).

D. K-NN Query and Continuous Monitoring

The *k*-Nearest Neighbor query is one of the query processing algorithms widely utilized due to its ease of use, effectiveness, and efficiency. This algorithm has been implemented in various fields, including pattern recognition, text categorization, and moving object recognition, among others [3]. The k-NN query aims to find the k-nearest objects based on a reference object.

The search for the *k*-nearest objects can employ various distance functions such as Minkowski Distance, Manhattan Distance, Euclidean Distance, Cosine Distance, Jaccard Distance, and Hamming Distance. In this article, the author utilizes the Euclidean Distance function as it measures the straight-line distance between two points in Euclidean Space.

Moreover, continuous *k*-Nearest Neighbor Monitoring is a method employed to continuously monitor changes in the results of *k*-NN within specific time intervals. If an object identified by the *k*-NN moves away from the *k*-NN region or if there is a new object entering the k-NN region, the *k*-NN computation will be performed to obtain the latest results. Figure 2 provides an illustrative example of Continuous *k*-Nearest Neighbor Monitoring.

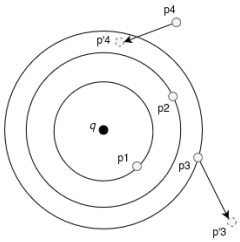


Fig. 2. Example of 3-NN Monitoring

In the illustrative example depicted in Figure 2, the object q has 3-nearest neighbor objects, namely *p1*, *p2*, and *p3*. At a certain point in time, object *p3* moves away from q. Simultaneously, object *p4* approaches *q* and has a closer distance to *q* compared to *p3* before it moves away. Due to these changes, a new search for the 3-nearest neighbors must be performed.

Initially, the system will search for a set of objects that qualify as candidates by comparing the distances between these objects and the reference object with the radius of the reference object [4].

Once the number of candidates meets the value of k, the computation of *k*-NN can be performed. When there are changes in *the k*-NN result objects (objects that exit the *k*-NN result region or new objects entering the *k*-NN result region), a new search will be conducted to identify fresh candidates that satisfy the value of k.

E. Durable k-NN Query on Time-Series Data

The Durable *k*-Nearest Neighbor Query on Time Series Data (DkNNT) is designed to retrieve a set of *k*-nearest objects that exhibit a durable level of proximity (durability) to a reference object within a specific time interval.

The significance of the DkNNT query, when applied to the context of tracing over respiratory virus disease lies in its ability to identify individuals who have the closest distance and a resilient level of proximity to individuals exposed to the virus within a defined time interval. This enables the tracing of individuals who are at risk of exposure and infection by the virus.

The brief example presented in Chapter I explains the concept of basic computations of DkNNT under the assumption of static movement. This example serves to illustrate the fundamental principles involved in the context of the given scenario. Clearly, persons *B* and *C* are the answer to the DkNNT query w.r.t. the query point A, *k*=2, and durability threshold = 25%.

It is important to note that there is a possibility of having more than *k* records that meet the criteria to be included as candidates for answering the Durable K-Nearest Neighbor in Time (DkNNT) query. However, due to the dynamic nature of the environment, we also take into consideration the number of time slices in which a record occurs as a factor in determining the query results on every periodic checking, denoted as *Δs*. The more frequently a record appears in each check, the higher it is regarded as a component of the DkNNT query. By incorporating this parameter, we can ensure obtaining a bounded number of DkNNT query results.

IV. Proposed Method

This section presents the approach employed to address the Durable *k*-Nearest Neighbor query based on time intervals, including the data structures and processing algorithms utilized. The utilization of symbolic notations throughout the article and its descriptions can be referred to in Table III.

TABLE III. List Of Symbols

Symbol	Description
r	Durability thresholds
tb	Start time of query interval
tc	End time of query interval
k	Number of query results
Δmin	Minimum time for the object to be part of the query result
Δs	Number of occurrences for an object to be part of DkNNT query results
sref	Reference object
candidates	DkNNT candidates

future candidates	Possible DkNNT candidate for the next timestamp
knn_list	*k*NN object list
radius	Minimum distances to find the candidates from the reference object
CS	Candidate of query results
RS	Query results

A. Data Structure

There exist three main data structures that are utilized to support the computation of the DkNNT query, they are Dictionary, Grid Index, along with Object Data Timestamp and Grid Position.

1. Dictionary
   The dictionary data structure is employed to construct the Grid Index, store data objects in the Grid Index, store the timestamp and grid position of data objects in the Object Data Timestamp and Grid Position, store the candidate of *k*-NN objects, and store the *k*-NN objects themselves. The key-value structure of each dictionary can be observed in Fig. 4 and Fig. 5.

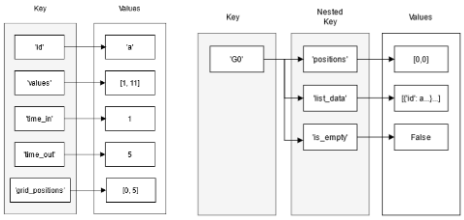


Figure 1. Structure of Dictionary for Data Object and Grid Index

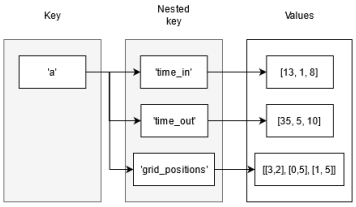


Figure 2. Structure of Dictionary for Object Data Timestamp and Grid Position

2. Grid Index
   The Grid Index is a data structure designed for indexing multidimensional data by grouping data according to their respective grid cells. It serves as an efficient mechanism for organizing and accessing data in a grid-based manner.

3. Object Data Timestamp and Grid Position
   We also utilize the Object Data Timestamp and Grid Position to store the list of arrival and departure times, as well as the list of Grid Cell positions for each object. This information proves valuable in tracking the movement of the reference object (*sref*) within the Grid Index. By maintaining these records, we can effectively monitor the trajectory and changes in the position of the reference object.

B. The Algorithms

As mentioned previously, the proposed algorithm to process DkNNT query consists of two-step processing, they are data precomputing and query processing tasks.

1. Data Precomputing
   Data precomputing is a process that enhances the performance of query processing algorithms by efficiently indexing the dataset using a Grid Index and storing the timestamp and grid position lists of each data object in the Object Data Timestamp and Grid Position structure.
   The process begins with the construction of the Grid Index data structure, followed by populating each cell in the Grid Index with the corresponding data objects. Once the Grid Index has accommodated all the data objects, the Object Data Timestamp and Grid Position are formed to store the lists of arrival and departure times, as well as the grid cell positions for each object. Upon completion, the Grid Index along with Object Data Timestamp and Grid Position are serialized and stored in disk files for future utilization during query processing. The overall workflow of the data precomputing process is implemented in Algorithm 1 and Algorithm 2.

Algorithm 1 Data Precomputing

Input: dataset
Output: Grid Index *GRID*, Object Data Timestamp and Grid Position *ODTGP*

```
list_object_data ← dataset
GRID ← init Grid Index
GRID.getListData(list_object_data)
for all id ∈ list_object_data do
    ODTGP ← GRID.getDataTimeIn(id)
    ODTGP ← GRID.getDataTimeOut(id)
    ODTGP ← GRID.getGridPosition(id)
Export GRID, ODTGP
```

Algorithm 2 *Grid* Class

Input: Size of Grid *grid_size*, dimension of Grid *d*, max value of Grid *max_value,* list of data *list_data,* object ID *id,* object position *position*
Output: Grid *G*

```
constructor Grid(grid_size, d, max_value)
    size ← grid_size
    dimensions ← d
    grid_length ← max_value / grid_size
    amount_data ← ∅
    grid_cell ← ∅
```

```
7:      for all segment in grid_size ^ d do
8:          Generate grid_cell of each segment
9:      procedure getAmountData
10:         return amount_data
11:     procedure addDataObject(list_data)
12:         for all data ∈ list_data do
13:             Generate data_location
14:             for all cell ∈ grid_cell do
15:                 if data_location = grid_cell['positions'] then
16:                     grid_cell['list_data'] ← data
17: procedure getGridPosition(id)
18:     position_list ← ∅
19:     for all cell ∈ grid_cell do
20:         if id in cell['list_data'] then
21:             position_list ← cell['list_data'][id]['grid_positions']
22:     return position_list
23: procedure getDataTimeIn(id)
24:     timestamp_list ← ∅
25:     for all cell ∈ grid_cell do
26:         if id in cell['list_data'] then
27:             timestamp_list ← cell['list_data'][id]['time_in']
28:     return timestamp_list
29: procedure getDataTimeOut(id)
30:     timestamp_list ← ∅
31:     for all cell ∈ grid_cell do
32:         if id in cell['list_data'] then
33:             timestamp_list ← cell['list_data'][id]['time_out']
34:     return timestamp_list
35: procedure getListData(position)
36:     if position = cell['positions'] then
37:         return cell['list_data']
38: procedure isValidNeighbor(position)
39:     counter ← 0
40:     for index in dimensions do
41:         if position[index] ≥ 0 and position[index] < size then
42:             counter++
43:     if counter = dimensions then
44:         return True
45:     return False
46: procedure getNeighborList(grid_position, level, previous_neighbor)
47:     valid_neighbor_list ← ∅
48:     iterators ← ∅
49:     lower_bound ← level * -1
50:     upper_bound ← level + 1
51:     for index in range(lower_bound, upper_bound) do
52:         iterators ← index
53;     Generate all possible neighbors_direction from iterators
54:     neighbor_candidates ← neighbor_direction
55:     neighbor_candidates ← neighbor_candidates – previous_neighbor
56:     for neighbor ∈ neighbor_candidates do
57:         if isValidNeighbor(neighbor) then
58:             valid_neighbor_list ← neighbor
59:     return valid_neighbor_list
```

2. Query Processing

Query processing is the second stage of the Durable k Nearest Neighbor query algorithm, aimed at processing user input queries using the Grid Index and Object Data Timestamp and Grid Position obtained from the data precomputing phase. The process begins by validating the time interval of the user's reference object. Next, an iteration is performed to search for k-nearest neighbors (k-NN) by identifying candidates and upcoming candidates from the reference object (sref). Once the candidate search process is completed, k-NN objects are obtained as the k-NN objects of *sref*. Subsequently, the monitoring process for changes in the k-NN is conducted until the durable query result objects are obtained, which are then returned and displayed to the user. The overall workflow of the query processing is implemented in Algorithm 3 and 4.

Algorithm 3 *Query Processing*

```
        Input: Reference object sref, number of nearest object k, initial interval
        timestamp tb, final interval timestamp tc, durability threshold r
        Output: list of durable k-NN objects RS
1:      grid ← GRID
2:      odtgp ← ODTGP
3:      data_length ← grid.getAmountData()
4:      isValidTimestamp, valid_sref ← checkObjectValidity(sref, tb, tc)
5:      if k > data_length -1 then
6:          return error
7:      if tb ≥ tc then
8:          return error
9:      if isValidTimestamp then
10:         Δmin ← ⌈r * (tb - tc)⌉
11:         radius ← sqrt( k / π * data_length)
12:         CS ← ∅
13:         RS ← ∅
14:         knn_list ← ∅
15:         knn_candidates ← ∅
16:         future_candidates ← ∅
17:         sref_position_list ← ∅
18:         sref_current_timestamp ← ∅
19:         t ← tb
20:         while t < tc do
21:             t_next ← t + 1
22:             sref_current_timestamp ←
                getCurrentTimestamp(sref_object_timestamp, t)
23:             if sref_current_timestamp is not null then
24:                 knn_candidates, future_candidates ← getKNNCandidates(sref,
                    valid_sref, k, t, radius)
25:             else
26:                 if t_next < tc then
27:                     t = t_next
28:                     continue
29:                 else
30:                     break
31:             knn_list ← getKNNList(k, knn_candidates)
32:             CS ← knn_list
33:         for time in range(t_next, tc) do
                if time in range(sref_current_timestamp['time_in'],
                sref_current_timestamp['time_out'])
34:                 if isKNNObjectNotGone(knn_list, time, k) and not
                    isCloserFutureCandidateComing(knn_list[x],
                    future_candidates, time) then
35:                     Increment Δs in CS corresponding with knn_list
36:                 elif not isKNNObjectNotGone(knn_list, time, k) and not
                    isCloserFutureCandidateComing(knn_list[x],
                    future_candidates, time) then
37:                     t = time
38:                     break
39:                 else
40:                     new_candidates ← ∅
41:                     new_candidates ← knn_candidates, future_candidates that
                        still exist at time
42:                     knn_list ← getKNNList(k, new_candidates)
43:                     Move new object of knn_list which is not yet in CS
```

```
                    Increment Δs in CS corresponding with knn_list
                        Move object in CS to RS where Δs ≥ Δmin
                    if time > tc – Δmin then
                    Remove object in CS that satisfy Δs < Δmin – tc + time + 1
                    if CS is null then
                        break
                else
                    t = time
                    break
            if CS is null then
                break
        return RS
    else
        return error
```

Algorithm 4 Procedure getKNNCandidates

Input: Object reference *sref*, Valid object reference *valid_sref*, number of object searched *k*, timestamp *t*, object reference radius *radius*
Output: k-NN candidate *candidates*, future k-NN candidate *future_candidates*

```
candidates ← ∅
future_candidates ← ∅
object_sref ← ∅
not_eligible_object ← ∅
sref_position ← getObjectPosition(valid_sref, t)
list_data ← grid.getListData(sref_position)
if sref in list_data then
    object_sref ← list_data[sref]
Collect candidates, future_candidates using thread
if candidates.length ≥ k then
    return candidates, future_candidates
else
    previous_neighbor ← ∅
    previous_neighbor ← sref_position
    level ← 1
    while level < grid.grid_size do
        neighbor_list ← grid.getNeighborList(sref_position, level,
        previous_neighbor)
        if neighbor_list is null then
            break
        previous_neighbor ← neighbor_list
        new_candidates ← ∅
        new_not_eligible ← ∅
        new_future_candidates ← ∅
        thread_list ← ∅
    for all position ∈ neighbor_list do
        Collect candidates, future_candidates in grid cell using thread
    if candidates.length < k then
        if not_eligible_object is not null then
            candidates ← not_eligible_object
        else
            candidates ← new_not_eligible
        not_eligible_object.clear
        not_eligible_object ← new_not_eligible
        if candidates.length ≥ k then
            not_eligible_object.clear
            break
        level++
    return candidates, future_candidates
```

V. PERFORMANCE EVALUATION

The experiments were conducted to assess the performance in terms of computational time and memory usage for each operation. The experimental environment is outlined in Table 4, providing insights into the specific conditions under which the tests were carried out.

TABLE IV. TESTING ENVIRONMENT

Environment	Specification
Hardware	Prosesor: Intel Core i5 7200U @ 2.50GHz (4 CPU) RAM: 8192 MB
Software	Operating System: Windows 10 Enterprise LTSC 64-bit

A. *Datasets*

There exist three datasets that are utilized to represent the location distribution of the record for the performance testing, they are Independent (IND), anti-correlated (ANT), and real-life data, Forest Cover (FC)

B. *Evaluation Scenario*

By default, we define the value of the performance evaluation setup as follows:

a. number of spatial dimensions: two dimensions
b. the data size: 50,000
c. the number of grids (in a dimension): 10
d. time interval: 100
e. k= 50; and
f. durability threshold: 70%;

Using the above default value, we vary the value of some testing parameters in each scenario of evaluation as follows:

1. effect of the number of spatial dimensions: two- and three-dimensions;
2. effect of the size of data: 10K, 30K, 50K, 100K, 200K, and 300K;
3. effect of the value of k: 25, 50, 75, and 100;
4. effect of the time interval: 10, 50, 100, 500, 1000;
5. effect of the durability threshold: 30%, 50%, and 70%.

C. *Evaluation Results of Query Processing*

1. Effect of the Number of Spatial Dimensions

In the case of independent data (IND), the variation in attribute dimensions in Table V has a highly significant influence on the execution time required by the query processing to handle user input queries. This is attributed to the random distribution of data values in each attribute, leading to a scattered placement of data points on the grid. As a result, the search process for the k-nearest neighbor (k-NN) candidate list and the search process for the k-NN object list become more time-consuming.

On the other hand, in the case of anti-correlated data (ANT) and forest cover data (FC), there is a decrease in execution time as the attribute dimensions increase. This is because in higher dimensions, the characteristics of the anti-correlated (ANT) and forest cover (FC) data cluster together to form an area or

region, reducing the need for extensive candidate verification operations.

TABLE V. EFFECT OF DIMENSIONAL SIZE IN QUERY PROCESSING

Size of Dimension	Running Time (s)		
	IND	ANT	FC
2	0,35	2.1	2,41
3	43.2	0.39	0,59

2. Effect of the Size of The Data

The variation in the amount of data in Fig.6 has a significant impact on the execution time of query processing for all three types of datasets. As the amount of data increases, the number of data points within each grid cell also increases. Consequently, the verification process for the k-nearest neighbor (k-NN) candidate list becomes more time-consuming.

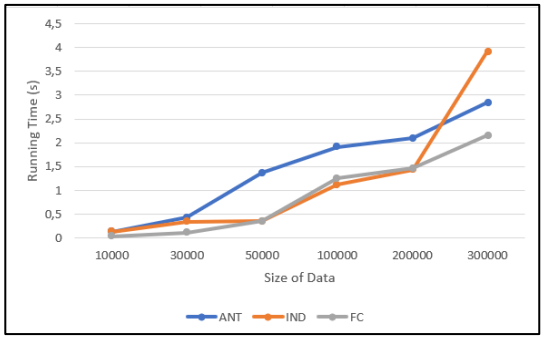


Figure 6. Size of Data vs Running Time

3. Effect of the *k*

Similar to the changes in the data quantity, the alteration of the value of *k* in Fig.7 has a significant impact on the execution time of query processing in all three types of datasets. This is due to the fact that as the value of *k* increases, the algorithm needs to expand the search area, resulting in a larger radius for the reference object (*sref*). Consequently, the process of gathering the list of *k*-NN candidate objects becomes more extensive as it requires visiting grid cells located within the radius region of *sref*.

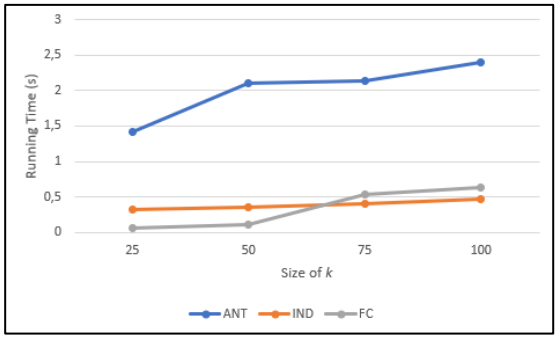


Figure 7. Size of *k* vs Running Time

4. Effect of the Time Interval

The variation in time interval magnitude in Figure 8 has a significant influence on the execution time of query processing in all three types of datasets. In the ANT dataset, the running time starts to increase from the search at a time interval of 100. Conversely, in the IND and FC datasets, the impact of the interval increase becomes significant only at a time interval of 1000. This is due to the fact that as the time interval expands under these conditions, the process of monitoring the list of *k*-NN query result objects becomes more time-consuming. This is because there are more objects that need to be checked for their appearance and departure within the search radius.

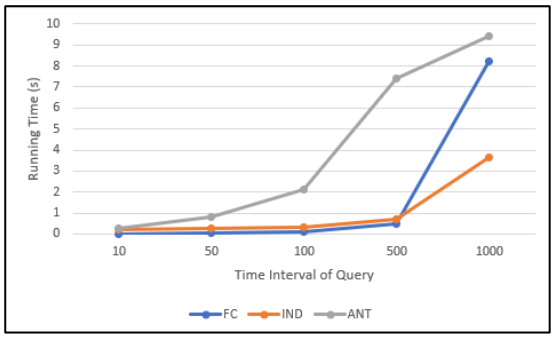


Figure 8. Time Interval vs Running Time

5. Effect of the Durability Threshold

Upon observing the overall graph in Fig.9, the movements of both indicator lines are not significantly pronounced across all three types of datasets. This indicates that the durability threshold does not have a substantial impact on the execution time of query processing, as the durability threshold primarily affects the size of a small search scope among the final candidates, and it is highly dependent on the size of the time interval.

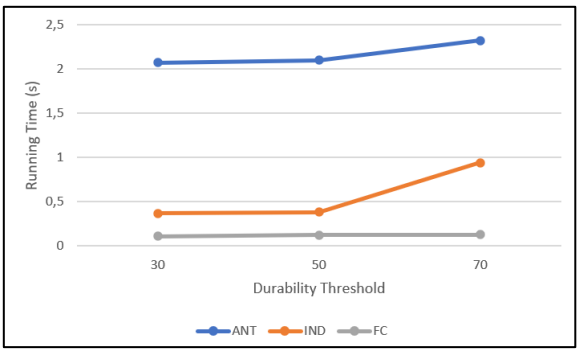


Figure 9. Durability Threshold vs Running Time

VI. CONCLUSIONS

In this study, we discussed the data structure and query algorithm for Durable k-Nearest Neighbor using Grid Indexing, which consists of two processes: data precomputing and query processing. A series of experiments have been conducted, comparing various parameter variations and utilizing both synthetic and real-life datasets, to evaluate the effectiveness of the proposed solution. From the experimental results, it can be concluded that the proposed solution is highly sensitive to the number of dimensions, data size, and time interval.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under the project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2023.

REFERENCES

[1] K. Mouratidis, M. Hadjieleftheriou and D. Papadias, "Conceptual Partitioning: An Efficient Method for Continuous Nearest Neighbor Monitoring," 2005.

[2] H. Wang, Y. Cai, Y. Yang, S. Zhang and N. Mamoulis, "Durable Queries over Historical Time Series," vol. 26, p. 595, 2014.

[3] S. Dhanabal and D. S. Chandramathi, "A Review of various k-Nearest Neighbor Query Processing Techniques," International Journal of Computer Applications (0975 – 8887), vol. 31, p. 14, 2011.

[4] K. Mouratidis, D. Papadias, S. Bakiras, and Y. Tao, "A Threshold-Based Algorithm for Continuous Monitoring of k Nearest Neighbor," IEEE Transactions on , vol. 17, pp. 1453, 1454, 2005.

[5] H. Ding, G. Lu, and L. Liu, "K Nearest Neighbor Queries and KNN-Joins in Large Relational Databases (Almost) for Free," 2010 IEEE 26th International Conference on Data Engineering, 2010.

[6] F. Li, L. Liu, and S. Wang, "Grid-Based Indexing and Search Algorithms for Large-Scale and High-Dimensional Data," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 5, pp. 887-900, 2018.

[7] D. Chunming, H. Xiangfeng, and W. Liqiang, "Grid-based Spatial Index Method for Location-based Nearest Neighbour Search," 2019 International Conference on Advances in Computational Science and Engineering, 2019.

[8] X. Zhang, Q. Yang, and Z. Qin, "Grid-based DBSCAN: Indexing and inference," Pattern Recognition, vol. 88, pp. 39-51, 2019.

[9] Onel Harisson. Towards Data Science, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[10] Kramer, Oliver, and Oliver Kramer. "K-nearest neighbors." Dimensionality reduction with unsupervised nearest neighbors (2013): 13-23.

[11] Dhanabal, S., and S. J. I. J. C. A. Chandramathi. "A review of various k-nearest neighbor query processing techniques." International Journal of Computer Applications 31.7 (2011): 14-22.

[12] Gao, Junyang, Pankaj K. Agarwal, and Jun Yang. "Durable top-k queries on temporal data." Proceedings of the VLDB Endowment 11.13 (2018): 2223-2235.

[13] Semertzidis, Konstantinos, and Evaggelia Pitoura. "Top-*k* Durable Graph Pattern Queries on Temporal Graphs." IEEE Transactions on Knowledge and Data Engineering 31.1 (2018): 181-194.

[14] Gao, J., Sintos, S., Agarwal, P. K., & Yang, J. (2021, April). Durable top-k instant-stamped temporal records with user-specified scoring functions. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 720-731). IEEE.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Retinex-based Low-Light Image Enhancement

Rui Luo*, Yan Feng†, Mingxin He* and Yuliang Zhang*
* Northwestern Polytechnical University, Xi'an
E-mail: rui_l @mail.nwpu.edu.cn, hemingxin@mail.nwpu.edu.cn, zyuliang@mail.nwpu.edu.cn
† Northwestern Polytechnical University, Xi'an
E-mail: sycfy@nwpu.edu.cn

***Abstract*—Image enhancement in low-light conditions is not solely a matter of adjusting the illumination. Due to the limited capabilities of capturing devices, there are also issues of hidden noise bursts and color distortion in dark environments. The Retinex model, which decouples observed image space into different subspaces for image enhancement, has been extensively validated as effective. Inspired by this model, this paper proposes a low-light image enhancement method based on Retinex model decomposition, guided by deep neural networks.The proposed approach decomposes the image into the illumination map responsible for light adjustment and the reflectance map responsible for removing degradation. A light enhancement network is trained using image pairs captured with different exposures to achieve user-friendly adjustable light intensity enhancement. To alleviate the issue of color distortion, a reflectance restoration network incorporates illumination guidance information. Additionally, a diffusion model is employed as a noise detector to eliminate unnecessary noise and other distortions in the brightened low-light images by fitting the conditional distribution between different images. Experimental results demonstrate that the proposed method outperforms in terms of color preservation in low-light images with varying light intensities. It effectively removes noise while preserving image details.**

I. Introduction

Images captured in low-light environments often have visually displeasing effects for both humans and machines. Simply increasing the ISO of the capturing device or adding additional background lighting often introduces additional challenges. While higher ISO enhances the sensitivity of the image sensor, it also amplifies image noise. Background lighting can partially illuminate the environment, but point light sources can cause uneven lighting and introduce unexpected highlights. Moreover, the high cost of hardware development has made the development of a practical low-light image enhancer a hot research topic.

In accordance with the Retinex theory [1], the Single-scale Retinex (SSR) [2] technique utilizes a Gaussian blurred version of the input as the illumination map. By subtracting this illumination map from the input, the final result is obtained. To improve upon this approach, the Multi-scale Retinex (MSR)[3] extends SSR by combining the results of multiple Gaussian blur functions with different variances. Additionally, LIME [4] proposes a method to refine the initial illumination through the Max-RGB assumption and the structure preserving constraint. However, these methods primarily focus on enhancing the brightness of low-light images and often overlook hidden degradation such as noise and color distortion.

The aforementioned approaches are traditional algorithms for enhancing low-light images. However, in recent years, deep learning methods have dominated the field due to their strong generalization capabilities and high performance. For instance, MSR-Net[5] integrates the MSR mechanism into a neural network and utilizes BM3D[6] for denoising.Although there have been advancements in visual quality, the approach still inherits the limitations from traditional MSR methods. In an effort to overcome these limitations, [8] proposes the integration of a semantic segmentation technique with a Retinex-based model. The objective is to leverage semantic clues for improved illumination adjustment. However, it is important to note that these methods, while yielding satisfactory results, still struggle to effectively tackle issues related to noise and color distortion.The Progressive Retinex method [7] addresses light and noise issues by iteratively estimating illumination maps and noise levels. However, it does not effectively tackle color distortion. On the other hand, [9] decouples images into low-frequency and high-frequency layers to reduce noise and enhance details respectively. TBEFN [10] combines enhanced and denoised results from two branches. However, most of these methods fail to consider the relationship between illumination changes and degradation, leading to issues such as residual noise and over-smoothing. EnlightenGAN [11] leverages larger-scale unpaired training data using the GAN mechanism. While it alleviates the need for paired data, its focus is mainly on addressing lighting issues, leaving other defects insufficiently addressed.

To effectively enhance the brightness of low-light images while preserving color accuracy and minimizing noise, achieving a satisfactory visual experience, our contributions can be summarized as follows:

- Cleverly partitioning the image into different subspaces and applying corresponding operations to address noise and color distortion issues separately.
- Enabling interactive adjustment of the illumination intensity, allowing users to have control over the brightness enhancement according to their preferences.
- Taking into consideration the relationship between color distortion and illumination adjustment, we propose an illumination-guided color restoration network. This network performs well in restoring colors for images with varying levels of brightness.
- By incorporating diffusion models from the image gener-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

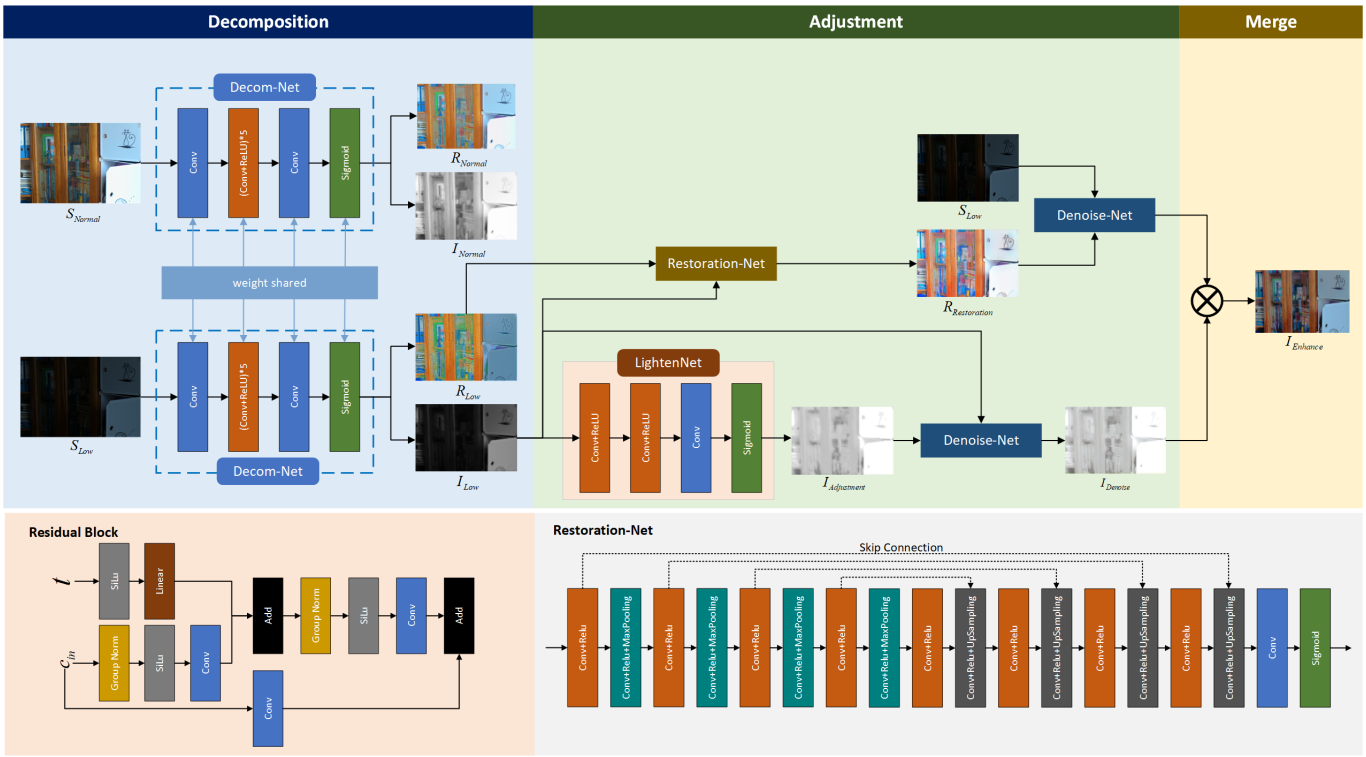


Fig. 1. the architecture of our proposed network.the architecture can be divided into three main blocks and four subnetworks with distinct functionalities. The bottom left corner depicts the residual blocks used to construct the diffusion model, while the detailed structure of the Restoration-Net is shown in the bottom right corner.



(a) Input (b) Rlow (c) BM3D (d) DMdenoise (e) Rrestoration (f) Ground Truth

Fig. 2. Decomposing low-light images of varying degrees to obtain reflectance maps and enhancing them,(a) Original low-light image, (b) Reflectance map, (c) Reflectance map processed with BM3D denoising, (d) Reflectance map processed with diffusion model denoising, (e) Reflectance map processed with Restoration method in this paper, (f) Ground truth image corresponding to the low-light image.

ation domain into the denoising domain, we can preserve more detailed image information in the denoised results.

This article is organized as follows: Section 2 provides areview of related works. Section 3 presents our methodology and the detials of our network. Section 4 reports on the experimental results and analysis. The conclusion is presented in Section 5.

II. Related Work

A. Adjust Reflection With Illumiation

Retinex theory decomposes an image I into its reflectance map R and illumination map L, represented in the subsequent formulation.

$$I = R \circ L \tag{1}$$

where $\circ$ designates the Hadamard product operator.Actually, the model may encounter challenges due to the limitations of the capturing sensor, leading to images suffering from complex degradation. As a result, according to [14] the expected rela-

tionship between the reflectance component and illumination is modeled as:

$$I = R \circ L + E = \tilde{R} \circ L = (R + \tilde{E}) \circ L \tag{2}$$

where E designates the pollution component, $\tilde{R}$ stands for the polluted reflectance, $\tilde{E}$ is defined as E/L with element-wise division. From this we can know that the degradation caused by R is related to changes in the illumination component.Moreover, when capturing images in low-light conditions, sensors (such as CCD or CMOS) exhibit sensitivity and non-linearity towards insufficient photons across different light spectrums. This sensitivity often leads to color distortion, even when proper measures have been taken to address illumination and noise issues. Consequently, to obtain high-quality results from degraded inputs, an effective algorithm must address the complex interplay of dim light, noise, and color distortion, which are closely intertwined.

B. Diffusion Model

Diffusion Models (DMs) [15] is a type of generative model proposed in 2020. Unlike other generative networks , the diffusion model applies noise progressively to the image during the forward diffusion process, gradually deteriorating the image into pure Gaussian noise. Then, during the reverse sampling process, it learns to reconstruct the real image from the Gaussian noise. This approach allows for connecting a complex data distribution with a simple and analyzable distribution using a Markov chain composed of a finite number of time steps.

To avoid the computationally expensive generative reverse process of DMs for sampling normally-exposed images, we focus on harnessing their capability to capture complex conditional data distributions. Specifically, [17] utilize DMs to capture the relationship between under-exposed and normally-exposed images. Previous research has demonstrated that DMs can also serve as backbone feature extractors, accurately predicting features based on noisy inputs[16].

III. Our Proposed Method

A. General Framework

As shown in Fig. 1,our network framework consists of three main components: the decomposition part, the adjustment part, and the fusion part. The decomposition part is accomplished by the Decom-Net, while the adjustment part includes the Lighten-Net for illumination adjustment, the Restoration-Net for color restoration, and the Denoise-Net for noise removal. Finally, the image fusion is performed using a simple fusion network. In the following sections, we will provide detailed descriptions of each network component.

B. Demon-Net

According to the Retinex theory, the reflectance of an image should remain constant under different lighting conditions. Therefore, in the decomposition network, we refer to [12] and utilize the linear combination of three different losses $\mathcal{L}^{Dn}$ to constrain the paired images during the decomposition process.

$$\mathcal{L}^{Dn} = \mathcal{L}_{re} + \beta_{ir}\mathcal{L}_{ir} + \beta_{is}\mathcal{L}_{is} \tag{3}$$

From (3), β_{ir} and β_{is} are constants. When the image pairs in the dataset are decomposed, the normal-exposed images I_{Normal} are split into $[L_{Normal}, R_{Normal}]$, while the low-light images I_{Low} are decomposed into $[L_{Low}, R_{Low}]$.The three losses correspond to the following (4), (5), (6).

$$\mathcal{L}_{re} = \sum_{i=Low,Normal} \sum_{j=Low,Normal} \lambda_{ij} \| R_i{}^{\circ} I_j - S_j \|_1 \tag{4}$$

$$\mathcal{L}_{is} = \sum_{i=Low,Normal} \| \nabla I_i{}^{\circ} \exp\left(-\beta_x \nabla R_i\right) \|_1 \tag{5}$$

$$\mathcal{L}_{ir} = \| R_{Low} - R_{Normal} \|_1 \tag{6}$$

where λ_{ij} and β_x are both constant coefficients. The symbol ∇ represents the gradient, which includes both the horizontal component and the vertical component.

The Decom-Net consists of five convolutional layers, with a ReLU activation applied between the first two convolutional layers. The remaining three convolutional layers do not have a ReLU activation function. The specific structure is shown in Fig. 1.

C. Restoration-Net

The reflectance maps extracted from low-light images, as depicted in Fig. 2. As the light becomes dimmer, the images are accompanied by increasingly severe color distortion. In order to find an appropriate restoration function, the objective can be simplified as follows[14]:

$$\begin{aligned} \mathcal{L}^{Rn} = & \left\| \widehat{R} - R_{Normal} \right\|_2^2 + \left\| \nabla \widehat{R} - \nabla R_{Normal} \right\|_2^2 \\ & - \mathrm{SSIM}\left(\widehat{R}, R_{_{Normal}} \right) \end{aligned} \tag{7}$$

where $SSIM(,)$ is the structural similarity measurement.The term $\widehat{R}$ refers to the restored reflectance map. In other words, it represents the estimate or reconstructed version of the reflectance map in the image. The loss focuses on capturing and preserving the texture details of the images during the restoration process. The specific configuration of the reflectance map restoration network can be found in Fig. 1, which provides a detailed breakdown of the network architecture, including the number of layers, types of layers. This information serves as a reference for understanding the internal structure and design choices of the network used for reflectance map restoration.

D. Lighten-Net

There is no standard for the intensity of light enhancement. To be able to adjust the intensity of light enhancement flexibly and meet different requirements, we will utilize existing pairs of$[L_{Normal}, L_{Low}]$ images. By calculating the intensity ratio between the two, we obtain the illumination adjustment



(a) Input (b) BIMEF [8] (c) EnlightGAN [9] (d) LIME [10] (e) RetinexNet [11] (f) Zero_DCE [1] (g) Ours (h) Ground Truth

Fig. 3. Visualized images of different low-light image enhancement algorithms on the LOL test dataset.

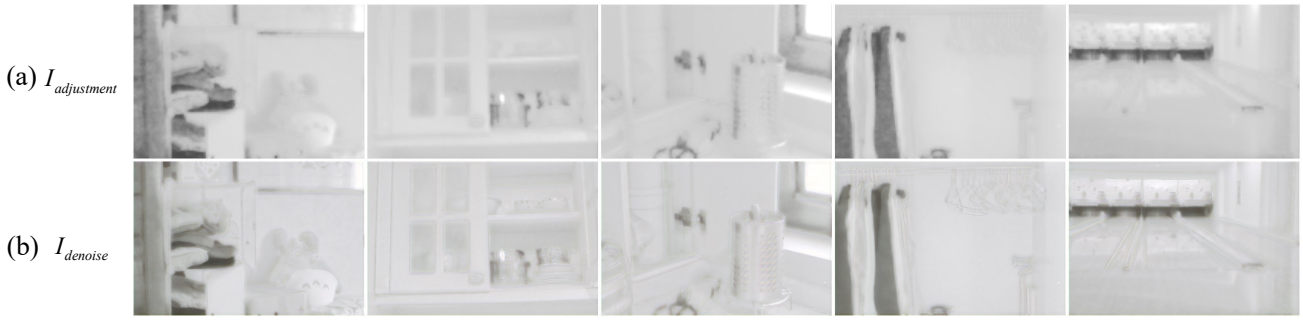


Fig. 4. Comparison images of the illumination component before and after denoising with the diffusion model.

coefficient. This coefficient is then expanded to a feature map and combined with the low-light illumination map as input for training and learning in the network. The following losses are used as constraints:

$$\mathcal{L}^{Ln} = \left\| \widehat{L} - L_t \right\|_2^2 + \left\| |\nabla \widehat{L}| - |\nabla L_t| \right\|_2^2 \tag{8}$$

The network architecture is designed to be lightweight, consisting of three convolutional layers. Specifically, it includes two convolutional layers followed by a ReLU activation function, and one additional convolutional layer. The output of the network is then passed through a Sigmoid activation function. This streamlined architecture aims to achieve computational efficiency while still delivering effective results.

E. Denoise-Net

Each network designed to address specific problems may introduce different forms of defects in its resulting outputs. Inspired by [17], we use the diffusion model as noise detector to alleviate this problem. We aim to capture the underlying data distribution between the under-exposed images b and the corresponding normally-exposed images I_0.

During the forward diffusion process, Gaussian noise is incrementally added to the data according to a predefined schedule of variances $\alpha_1, \ldots, \alpha_t$, Gaussian noise is defined as $z_1, z_2, z_3 \ldots \sim N(0, I)$. Mathematically, this can be represented as shown in (9).

$$I_t = \sqrt{\bar{\alpha}_t} I_0 + \sqrt{1 - \bar{\alpha}_t} z_t \tag{9}$$

TABLE I

RESULTS ON THE LOL TEST SET FOR DIFFERENT METHOD

Methods	SSIM	PSNR	MAE	LPIPS
LIME[4]	0.484	17.181	0.124	0.372
BIMEF[20]	0.595	13.875	0.206	0.326
RetinexNet[12]	0.425	16.774	**0.126**	0.474
EnlightenGAN[11]	0.652	17.483	0.135	0.322
ZeroDCE[21]	0.562	14.861	0.185	0.335
Ours	**0.761**	**17.656**	**0.126**	**0.233**

I_t represent the image after adding t steps of noise, $\bar{\alpha}_t{=}\alpha_t\alpha_{t-1}\alpha_{t-2}\cdots\alpha_1$ is a cumulative multiplication operation. The conditional diffusion model aims to maximize the likelihood of $p(\,I_0|\,b)$ The reverse diffusion process is characterized by parameterized Gaussian transitions[17]. It guides the diffusion model to predict the potential noise involved in the sampled I_t. Given I_t and the noise prediction from the diffusion model, we can compute an estimation of I_0:

$$I_0 \approx \widehat{I}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} I_t - \left(\sqrt{\frac{1}{\bar{\alpha}_t} - 1}\right) Z_\theta\left(I_t, t, b\right). \tag{10}$$

where Z_θ presents diffusion model. It utilizes an improved U-Net model for modeling, incorporating channel and time step embeddings through residual blocks. The structure of the residual block is illustrated in the bottom left corner of Fig. 1.

IV. Experimental Analysis

A. Implementation Details

To accommodate the diverse and sensor-specific forms of image degradation, we opted to train our model using a dataset called LOL. This dataset comprises 485 pairs of real-exposure images. The original images are in a portable network image format with a size of 400x600 pixels. For the Decom-Net, we set the batch size to 16 and the patch size to 48x48 pixels. For the Restoration-Net and Lighten-Net, the batch size is set to 4 and the patch size is set to 384x384 pixels. Regarding the diffusion model used for denoising, we fix the diffusion steps (T) to be 1000. We employ the AdamW [18] technique for optimization, with a learning rate of $1{\times}10^{-6}$.

B. Performance Evaluation

We visualize our algorithm and some state-of-the-art low-light image enhancement method on the LOL test set, as shown in Fig.4.Our method outperforms the competing method on the LOL test set. The results obtained by our method exhibit significantly better performance, with effectively reduced noise and fewer artifacts. In contrast, the outputs generated by the other competitors suffer from issues such as irregular illumination, residual noise, and texture/color distortions. Especially, comparing the images in the third and fourth rows, it is evident that our method successfully corrects the greenish hue, resulting in a more realistic tone in the enhanced images.

In this study, we utilize several full-reference metrics to evaluate the performance of the proposed method. These metrics include the structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), mean absolute error (MAE), and learned perceptual image patch similarity (LPIPS) [19]. By employing these metrics, we can assess various aspects of image quality, such as structural similarity, noise levels, pixel-wise differences, and perceptual similarity.A higher value for PSNR and SSIM indicates superior image or video quality. On the other hand, for metrics like MAE and LPIPS , a lower value is desirable as it signifies better quality. As shown in Table 1, it is evident that our method consistently outperforms the other methods.

C. Ablation

To validate the effectiveness of different subnetworks, we selected images with varying lighting conditions from the LOL test dataset for testing. Firstly, we utilized the trained decomposition network to obtain the reflectance map and illumination map of the low-light images.By comparing the original reflectance map obtained from the decomposition with the reconstructed reflectance map obtained after passing through the Restoration-Net in Fig.2, we can clearly observe that the green noise artifacts have been effectively removed, resulting in an overall improvement in image quality. Additionally, considering the denoising operation applied to the reflectance map during the network refinement stage, we compared the popular BM3D algorithm with the diffusion-based denoising method used in this study. We found that while traditional denoising methods like BM3D offer some benefits for smoothing images, they are unable to handle large areas of colored noise effectively. This finding further validates the effectiveness of our proposed Restoration-Net in recovering the color information in the reflectance map.

In Fig. 4, we selected the illumination component maps adjusted by the Lighten-Net and applied a diffusion denoising model(Denoise-Net). By comparing the first and second rows of the image, we can observe that the diffusion model not only helps in noise removal but also effectively restores image details. The illumination component map, which contains more lighting details, enables the recovery of additional detailed features in the enhanced images.

V. Conclusion

This paper reveals the challenges of noise and color shift (green tint) in low-light image enhancement. To address these specific issues, the approach involves decomposing the image into different subspaces. The Restoration-Net is proposed to tackle color distortion in the reflectance map, while the Lighten-Net addresses the problem of illumination adjustment and enhancement. The Denoise-Net is employed to remove noise while preserving image details. These three networks work on individual image components and, in the end, the improved outcomes are created by merging the altered components. Experimental results demonstrate our method demonstrates superior performance by effectively suppressing noise, reducing artifacts, addressing illumination irregularities, and rectifying color distortions, ultimately producing visually

pleasing and realistic results on the low-light images of the LOL test set.

REFERENCES

[1] E. H. Land, "The Retinex Theory of Color Vision," Scientific American, pp. 108–128, Dec. 1977.
G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London,* vol. A247, pp. 529-551, April 1955.

[2] D. J. Jobson, Z. Rahman and G. A. Woodell, "Properties and performance of a center/surround retinex," in IEEE Transactions on Image Processing, vol. 6, no. 3, pp. 451-462, 1997.

[3] D. J. Jobson, Z. Rahman and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," in IEEE Transactions on Image Processing, vol. 6, no. 7, pp. 965-976, 1997.

[4] X. Guo, Y. Li and H. Ling, "LIME: Low-Light Image Enhancement via Illumination Map Estimation," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 982-993, 2017.

[5] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-net: Low-light Image Enhancement Using Deep Convolutional Network.," Cornell University - arXiv, 2017.

[6] K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering," in IEEE Transactions on Image Processing, vol. 16, no. 8, pp. 2080-2095, 2007.

[7] Y. Wang et al., "Progressive Retinex: Mutually Reinforced Illumination-Noise Perception Network for Low Light Image Enhancement," Cornell University - arXiv, 2019.

[8] M. Fan, W. Wang, W. Yang, and J. Liu, "Integrating Semantic Segmentation and Retinex Model for Low-Light Image Enhancement," in Proceedings of the 28th ACM International Conference on Multimedia, Seattle WA USA, 2020.

[9] K. Xu, X. Yang, B. Yin and R. W. H. Lau, "Learning to Restore Low-Light Images via Decomposition-and-Enhancement," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 2278-2287, 2020.

[10] K. Lu and L. Zhang, "TBEFN: A Two-Branch Exposure-Fusion Network for Low-Light Image Enhancement," in IEEE Transactions on Multimedia, vol. 23, pp. 4093-4105, 2021.

[11] Y. Jiang et al., "EnlightenGAN: Deep Light Enhancement Without Paired Supervision," in IEEE Transactions on Image Processing, vol. 30, pp. 2340-2349, 2021.

[12] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," arXiv: Computer Vision and Pattern Recognition, 2018.

[13] X. Guo and Q. Hu, "Low-light Image Enhancement via Breaking Down the Darkness," International Journal of Computer Vision, pp. 48–66, 2023.

[14] Y. Zhang, J. Zhang, and X. Guo, "Kindling the Darkness: A Practical Low-light Image Enhancer," arXiv: Computer Vision and Pattern Recognition, 2019.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[16] B. Kim, Y. Oh, and J. Ye, "Diffusion Adversarial Representation Learning for Self-supervised Vessel Segmentation," 2022.

[17] S. Panagiotou and AnnaS. Bosman, "Denoising Diffusion Post-Processing for Low-Light Image Enhancement," 2023.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Cornell University - arXiv, 2017.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[20] Z. Ying, G. Li, and W. Gao, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," arXiv preprint arXiv:1711.00591, 2017.

[21] C. Guo, C. Li, J. Guo, C. C. Loy,J. Hou, S. Kwong, and R. Cong, "Zeroreference deep curve estimation for low-light image enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Fine-grained Face Anti-Spoofing based on Recursive Self-Attention and Multi-Scale Fusion

Shichuang Xie*, Jiasheng Wu*, Yanli Chen*, Meng Han†, Ting Wu‡, and Tong Qiao*§¶
* School of Cyberspace, Hangzhou Dianzi University, China
† Zhejiang University, Binjiang Institute of Zhejiang University, China
‡ Hangzhou Innovation Institute, Beihang University, China
§ State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, China
¶ Corresponding author, E-mail: tong.qiao@hdu.edu.cn

***Abstract*—Face anti-spoofing technique is a critical preset security measure for face recognition system. However, most existing detection techniques can only detect face spoofing and cannot provide fine-grained information on the type of spoofing attack. Additionally, DNN-based detection models lack the ability to extract diverse attack features distributed across different network depths and have a high number of parameters, making them challenging to train. To address these issues, we propose a face anti-spoofing method based on recursive self-attention and multi-scale fusion. Firstly, we introduce the design of recursive neural networks into self-attention mechanisms, combining multiple shared-weight self-attention blocks in a recursive form to extract features in the deep layer of the model while reducing the number of parameters. Furthermore, we combine atrous convolution with self-attention mechanisms, using atrous convolution with different atrous rates to provide multi-scale fusion capabilities for self-attention mechanisms. Extensive experiments on benchmark datasets demonstrate that our proposed method achieves superior performance while remarkably reducing the number of model parameters.**

I. Introduction

As facial recognition is increasingly used in various scenarios, malicious attackers are attempting to disrupt facial recognition systems by using techniques such as synthesizing realistic faces [1]–[4] or face spoofing [5]–[9]. According to research, there are currently five main types of face spoofing attacks, including print attacks [10], replay attacks [11], partial and full mask [12]–[16], and even make-up attacks [17]. We classify mainstream anti-spoofing systems into two categories: one based on external physical auxiliary equipment or user actions, such as using near-infrared cameras for auxiliary shooting judgment, or requiring user assistance for head shaking, mouth opening, and other actions. The other type is based on DNN (Deep Neural Networks) to develop automatic face anti-spoofing systems [18], such as [17], [19]–[26]. The former obviously increases the deployment cost of facial recognition systems and reduces their usability. As Fig. 1 shows, the latter is limited to binary classification, i.e., it can only detect whether a face is real or spoofed, and cannot distinguish specific types of facial spoofing attacks, which is disadvantageous for some demanding evidence collection scenarios.

Therefore, to provide stronger support for evidence collection work, [27] proposes a novel fine-grained classification detector that explores the potential of the Transformer structure



Fig. 1. The vanilla face anti-spoofing systems can only accomplish a binary classification task (left), while our proposed method can further distinguish among different types of spoofing attacks (right).

in face anti-spoofing forensics. They combine the task characteristics of face anti-spoofing forensics to introduce convolutional operations into self-attention mechanisms, achieving good results. Inspired by [27], we also consider that different types of facial spoofing attacks have different features, which are distributed in convolutional neural networks at different depths and are difficult to extract uniformly through a single CNN (Convolutional Neural Network) structure. Therefore, before making classification judgments, combining these features distributed at different convolutional depths plays a very important role in fine-grained classification. However, [27] fully explores the artifacts hidden in the shallow layers of DNNs in facial spoofing attacks, but there is still an issue with insufficient feature extraction in the abstract features of deeper networks. These abstract features play a crucial role in detecting more challenging attacks such as silicone masks. Therefore, different from [27], we attempt to deepen the network structure while maintaining the overall parameter reduction, thereby further improving the model's detection performance and training speed. Then the contributions of this paper are as follows:

- The design of RNN (Recurrent Neural Networks) is introduced into self-attention mechanisms, remarkably increasing the model's depth while maintaining a basic parameter number. This further enhances the model's ability to extract deep abstract features.
- An atrous self-attention mechanism with multi-scale feature fusion capability is achieved through the introduction of atrous convolution in the recurrent self-attention mechanism.
- Experimental results demonstrate that this method

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

achieves high accuracy and fine-grained classification ability on large-scale datasets.

II. Method

A. Preliminary

The detection model structure proposed in this paper is shown in the Fig. 2. The stride convolution self-attention module is derived from [27]. This module provides two advantages: Firstly, it reduces the overall computational cost as the feature map resolution decreases. Secondly, by varying the feature map resolution, it enables each stage of convolution to have a different receptive field, thereby improving the utilization of the convolutional spatial pyramid property and facilitating feature extraction at multiple levels. Then, new feature maps are obtained through stride self-convolution. To further explore the evidential clues, we use three recurrent self-attention modules with consistent structures to extract image features. Finally, we introduce a self-attention mechanism in the classifier and increase its depth.

B. Recursive Self-Attention

The depth of neural networks significantly affects the accuracy of model description, including Transformer-style models [28]. However, the self-attention mechanism requires an increase in weight matrices for each additional layer, resulting in a substantial increase in the overall number of model parameters. Therefore, most self-attention-based models use a limited number of layers. This issue also exists in the previous methods, where increasing model depth leads to a significant increase in parameters and computational power required for training. To address this, We propose to concatenate multiple self-attention modules with shared parameters into a whole recursive self-attention module.

In Fig. 2, we present a detailed construction of the first recursive self-attention module, each recursive self-attention module undergoes three rounds of multi-scale fusion self-attention processing, and the three multi-scale fusion self-attentions inside the module share the same parameters. The design of parameter sharing ensures that the total number of parameters in each recursive self-attention module is the same as that by using the multi-scale fusion self-attention mechanism only once, which controls the overall parameter amount while significantly improving the feature extraction ability of the model. Specifically, the output of each multi-scale fusion self-attention is processed by a layer of SiLU (Sigmoid Linear Unit) [29] activation function, which is shown below:

$$SiLU(x) = x \cdot Sigmoid(x), \tag{1}$$

where x represents the input feature map, and $Sigmoid(\cdot)$ represents the Sigmoid function, which is used to limit the feature values to the range [0, 1]. The reason for using this activation function is twofold: on the one hand, it plays a role similar to the gate controller in LSTM (Long Short-Term Memory) [30], filtering the data during the process of stacking residuals; on the other hand, it has been proven in the work of [29], that the self-calibration property of the SiLU activation function can help the model fit the data more accurately during the training phase. In addition, the recursive self-attention module uses residual operations similar to ResNet, which adds the input to the final output. This approach has the advantage of enhancing the flow of neural network gradients and mitigating the degradation of neural network gradients.

Furthermore, [31] shows that if a model is entirely composed of self-attention mechanisms, the model's weight matrix rank will quickly degrade to 1 due to the rapid convergence speed. Therefore, to mitigate the model's degradation, we add a MLP (Multilayer Perceptron) consisting of two linear mappings at the end of each recursive self-attention module. Finally, the output feature map y is normalized using LN (Layer Normalization) [32]:

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta \tag{2}$$

where x represents the input feature vector, $\mathrm{Var}[\cdot]$ represents the standard deviation, $\mathrm{E}[\cdot]$ represents the mean, ϵ is an infinitesimal quantity to solve the problem of division by zero, and γ and β are two trainable affine transformation parameters. Compared with the commonly used BN (Batch Normalization) [33] in computer vision, which normalizes across different channels of the feature map, LN normalizes each sample, [34] and [35] have conducted extensive comparisons and analyses of the two normalization methods. In short, LN can make the model converge more easily without introducing additional normalization, as there is no normalization operation in the feedforward network of the Transformer-style structure. Additionally, when the model settings cannot use a large batch size due to memory limitations, LN has better performance.

C. Multi-scale Fusion Optimization Based on Atrous Self-Attention

In order to further enrich the semantic features contained in the features, we introduce three atrous convolutions after linear mapping. As Fig. 3 illustrates, these three atrous convolutions share the same weights, with the only difference being their atrous rate. Since the convolutional stride in this part is 1, the resolution of the output feature map remains consistent with that of the input feature map. The final query value Q can be obtained by simply adding the outputs of the three atrous convolutions at corresponding positions. Additionally, since the three atrous convolutions share the same weights, they do not relevantly increase the number of model parameters. By using atrous convolutions with different atrous rate, our method achieves multi-scale feature fusion.

Our method combines features from multiple scales to reduce the risk of model misjudgment in face anti-spoofing. For instance, employing a single-scale feature extraction method with a small receptive field for extracting the specular reflection features of the pupil region exhibits certain similarities with the specular reflection features present in partial replay attacks and mask attacks. However, utilizing a multi-scale feature fusion approach, where the receptive field for feature

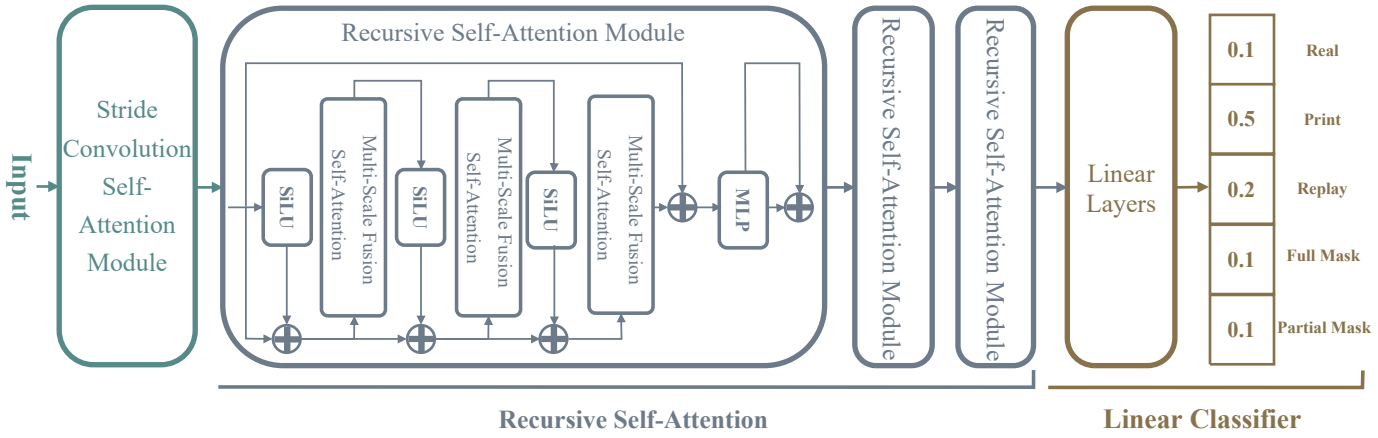


Fig. 2. Overview of our proposed fine-grained detection network.

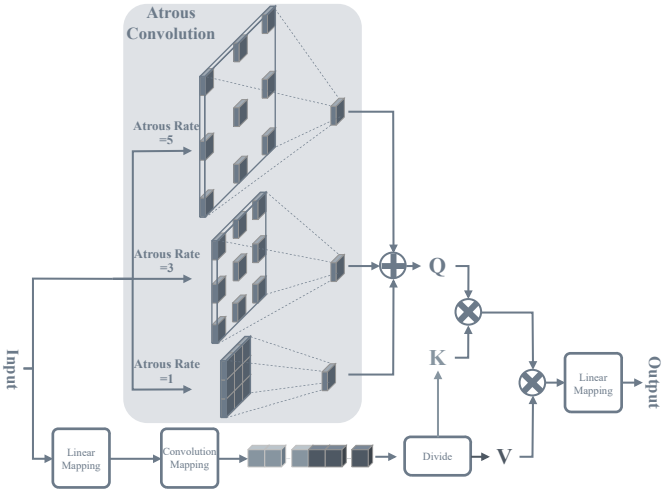


Fig. 3. Multi-scale self-attention structure.

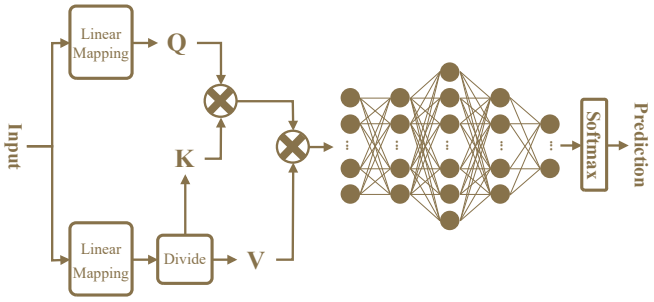


Fig. 4. Linear classifier structure.

extraction encompasses the entire eye reflection region, results in distinct differences between the extracted specular reflection features and the specular artifacts present in other types of attacks, thereby facilitating fine-grained classification. Therefore, we use multi-scale fusion to obtain Q, which reflects the correlation between feature points in the self-attention mechanism. This optimizes the weight matrix based on the judgment of the correlation weights between each feature point and other feature points at different scales.

In the process of computing K and V, since the key vector K and the value vector V respectively represent the key semantics of the feature point and the feature itself, they are semantically close in the self-attention mechanism. Therefore, the processing of K and V is combined here. First, a linear mapping layer is used to map the input feature matrix to another high-dimensional space by doubling the last dimension of the input matrix. Then, convolutional mapping is introduced to reduce the resolution of the output feature map. After rearranging the obtained feature map y_{output}, the front and back parts respectively represent K and V. The calculation method between Q, K, and V is as follows:

$$y_{output} = V \cdot (F_{dropout}(F_{softmax}(Q \cdot K))), \tag{3}$$

where $F_{softmax}(\cdot)$ represents the softmax operation, and $F_{dropout}(\cdot)$ represents the dropout operation.

D. Linear Classifier

As shown in Fig. 4, we add a self-attention module and slightly increas the number of linear layers in the classifier. The vanilla self-attention mechanism is directly used in the classifier. Since the linear classifier is at the final stage of the entire model, the resolution of the feature map at this point has become relatively low. That can effectively avoid the problem of parameter explosion when the input image is too large. At the same time, since it is located in the deep layer of the neural network, the deep model pays more attention to the abstract features in the feature map that are directly related to face anti-spoofing classification. Therefore, we directly use the linear mapping-based self-attention in the linear classifier. Additionally, since the self-attention mechanism itself is nonlinear, no additional activation function is added here. Subsequently, the features processed through the linear mapping-based self-attention are fed into the classification network, which consists of an input layer, three hidden layers composed of fully

connected layers, and an output layer with the same number of nodes as the fine-grained classification categories. Finally, we use cross-entropy with label smoothing as the loss function, and the overall loss function ℓ_{cls} is formulated as:

$$\ell_{cls}(x, y) = \sum_{i}^{n_{cls}} -y_i \times LogSoftmax(x_i), \tag{4}$$

$$LogSoftmax(x_i) = log\left(\frac{\exp(x_i)}{\sum_j \exp(x_j)}\right), \tag{5}$$

where x represents the predicted value, y represents the true sample label, and n_{cls} represents the number of categories in the fine-grained classification. Overall, this loss function is similar to the standard cross-entropy, but the difference is that when calculating the loss function, the cross-entropy with label smoothing first smooths the predicted value based on Eq. (5), thereby improving the model's generalization and accuracy.

III. Experiments

A. Datasets and Experiment Setup

In this section, we conduct extensive experiments to verify the effectiveness of our proposed model. We suggests classifying face spoofing attacks into four types, namely replay, print, partial mask, and full mask. To validate our approach, the model is tested on multiple datasets, including CeFA [36], WMCA [37], which encompass all attack modes, and HifiMask [38], which is currently the largest 3D mask attack dataset.

For all images in the dataset, facial detection is performed using 3DDFA-V2 [39], and the resulting facial images are cropped and resized to a uniform size of 128×128. Any image in the dataset that is not in the RGB format, such as near-infrared images, are discarded. The experimental results presented in this paper are obtained using only RGB images. The model is trained with the following parameters: an initial learning rate of 0.0002, a batch size of 64, a decay rate of 0.001, and optimization using the AdamW [40].

Experiments are primarily conducted using the standards specified in ISO/IEC 30107-3 [41]. Spoofing images are defined as positive samples, while genuine images are defined as negative samples. TP (True Positive) represents the number of spoofing attacks that are correctly identified, FP (False Positive) represents the number of spoofing attacks that are incorrectly identified as genuine faces, TN (True Negative) represents the number of genuine faces that are correctly identified, and FN (False Negative) represents the number of genuine faces that are incorrectly identified as spoofing attacks. The following sections will provide detailed explanations of these evaluation metric.

APCER (Attack Presentation Classification Error Rate) represents the probability of incorrectly identifying a spoofing attack as a genuine face. The calculation for APCER is as follows:

$$APCER = \frac{FP}{FP + TN}, \tag{6}$$

BPCER (Bona Fide Presentation Classification Error Rate) represents the probability of incorrectly identifying a genuine face as a spoofing attack. The calculation for BPCER is as follows:

$$BPCER = \frac{FN}{TP + FN}, \tag{7}$$

ACER (Average Classification Error Rate) represents the overall classification error probability. The calculation for ACER is as follows:

$$ACER = \frac{APCER + BPCER}{2}, \tag{8}$$

In addition, we use ACC (Accuracy) as an evaluation metric, measuring the ratio of correctly classified samples to the total number of samples across all categories.

B. Results on the Homogenous Datasets

The experiments in this section are conducted on CeFA [36], HifiMask [38] and WMCA [37] datasets, respectively. Specifically, CeFA provides four evaluation protocols, one of which is primarily used to measure the generalization ability of models in the face of multi-modal scenarios, while our method only focuses on RGB images as a single modality. Therefore, in the experiment, we use three of CeFA's evaluation protocols. Protocol 1 comprehensively evaluates the impact of different skin colors and races on face anti-spoofing model; Protocol 2 evaluates the model's generalization ability when facing unknown attack methods; Protocol 3 comprehensively considers the above two situations. HifiMask provides two evaluation protocols. Protocol 1 comprehensively evaluates the impact of different subjects on the model and is used to test the model's generalization ability when facing unknown faces. Protocol 2 evaluates the impact of different mask materials on the model, including transparent plastic, gypsum, and resin. As the WMCA dataset does not provide relevant protocol standards, we simply divide the training, validation, and testing sets in a 6:2:2 ratio.

Firstly, we conduct experiments on the benchmark dataset WMCA [37] to validate the performance of our proposed method against various types of face spoofing attacks. As shown in Table I, our proposed method outperforms the other methods on the WMCA [37] dataset. The multi-scale fusion feature extraction ability proposed in this paper enables effective detection of almost all attack types. However, full-face masks might be misjudged due to their similarity to real faces. Furthermore, due to the fact that the full-face masks attack does not compromise the digital integrity of the image but rather replaces the identity information at a physical level, it becomes challenging to learn the forgery characteristics of this type at the image level. As a result, the detection performance is degraded.

Next, we continue our experiments on the benchmark dataset CeFA [36]. As shown in Table II, the proposed method outperforms the baseline method remarkably in some scenarios of the CeFA [36] dataset. In Protocol 1 and Protocol 2, the performance of the proposed method is slightly lower than FGDNet, but the overall gap is not remarkable. This is because our proposed method incorporates certain design aspects from

TABLE I
PERFORMANCE COMPARISON ON THE WMCA DATASET.

Evaluation Metric(%)	Method	Partial Mask		Full Mask			Print	Replay	Avg.
		Glasses	Hard Mask	Fake Head	Silicone Mask	Paper Mask			
APCER	ResNet [42]	95.9	29.4	4.9	66.2	34.7	7.1	31.7	38.6
	MA-Net [43]	73.4	19.7	4.4	50.7	1.7	0.6	6.3	22.4
	FGDNet [27]	0	0	25.0	14.3	12.3	10.5	11.1	10.4
	Ours	0	0	9.0	6.1	5.2	2.2	2.6	3.6
BPCER	ResNet [42]	0.6	1.4	0	0.3	1.8	0	0	0.6
	MA-Net [43]	0	0	0	0	0	0	0	0
	FGDNet [27]	0	0	0	0	12.6	0	0	1.8
	Ours	0	0	0	0	0	0	0	0
ACER	ResNet [42]	48.3	15.4	2.5	33.3	18.2	3.6	15.9	19.6
	MA-Net [43]	36.7	9.8	2.1	25.3	0.9	0.3	3.2	11.2
	FGDNet [27]	0	0	12.5	7.1	12.4	5.3	5.6	6.1
	Ours	0	0	4.5	3.1	2.6	1.1	1.3	1.8

FGDNet, specifically in terms of fine-grained feature learning, which enables the method to retain the ability to learn subtle discriminative features. In Protocol 3, which is the most challenging, due to the incorporation of the recursive self-attention mechanism, we are able to highly correlate the long distance within the hidden layers, thereby facilitating the exploration of finer discrimitative features, the proposed method outperforms [27] remarkably, demonstrating the robustness of the proposed method.

TABLE II
PERFORMANCE COMPARISON ON THE CEFA DATASET.

Protocol	Method	APCER(%)	BPCER(%)	ACER(%)
CeFA_Protocol 1	ResNet [42]	28.1	6.4	17.3
	PSMM-Net [36]	17.4	13.8	15.6
	MA-Net [43]	16.7	12.4	14.6
	FGDNet [27]	0.4	0.3	0.4
	Ours	2.4	0	1.2
CeFA_Protocol 2	ResNet [42]	40.7	0.6	20.7
	PSMM-Net [36]	25.3	2.2	13.8
	MA-Net [43]	20.9	1.2	11.1
	FGDNet [27]	0.1	16.5	8.3
	Ours	0.1	17.9	9.0
CeFA_Protocol 3	ResNet [42]	49.4	5.3	27.4
	PSMM-Net [36]	33.1	17.2	25.2
	MA-Net [43]	30.4	18.1	24.3
	FGDNet [27]	1.1	19.3	16.2
	Ours	1.4	1.2	1.3

To further verify the model's ability to detect 3D mask attacks with a richer datadset, we conduct experiments on the HifiMask [38] dataset. As shown in Table III, our proposed method exhibits strong performance in detecting 3D mask attacks. Compared to prior detection methods, our proposed method, during the training stage, mitigates the excessive reliance on the dataset and to some extent prevents model overfitting by introducing the operation of stride convolution mapping. Therefore, the proposed method remarkably outperforms the baseline methods on the more challenging Protocol II, demonstrating the generalization ability of our approach.

TABLE III
PERFORMANCE COMPARISON ON THE HIFIMASK DATASET.

Protocol	Method	APCER(%)	BPCER(%)	ACER(%)
HifiMask_Protocol 1	ResNet [42]	3.7	5.7	4.7
	Auxiliary [19]	4.9	1.8	3.4
	CDCN [44]	3.3	3.9	3.6
	Ours	4.1	2.4	3.2
HifiMask_Protocol 2	ResNet [42]	22.4	14.6	18.5
	Auxiliary [19]	12.1	11.2	11.4
	CDCN [44]	12.6	16.8	14.7
	Ours	8.4	4.0	6.2

Finally, to demonstrate the performance of our proposed method in fine-grained classification, we compare it with another method FGDNet [27] for fine-grained face anti-spoofing. This part of the experiment is conducted on three datasets, CeFA, WMCA, and HifiMask, as shown in Table IV. The detection performance of our proposed method is superior

to FGDNet. This attributes to our improvement in the self-attention module, which enables more efficient extraction of long-distance features. Meanwhile, due to preserving the local features of the image, we have deepened the structure of the feature extraction network, resulting in a more robust feature extraction performance.

TABLE IV
PERFORMANCE COMPARISON OF FINE-GRAINED FACE ANTI-SPOOFING METHODS.

Method	CeFA	WMCA	HifiMask
FGDNet [27]	91.2	89.6	90.1
Ours	94.3	96.1	92.7

C. Ablation Study

In this section, we conduct a set of ablation experiments on the HifiMask dataset to demonstrate the effectiveness of the components in our proposed model.

TABLE V
RESULTS OF ABLATION EXPERIMENTS ON MULTI-SCALE SELF-ATTENTION.

Method	APCER	BPCER	ACER	ACC
Without Multi-Scale Self-Attention	15.2	7.8	11.5	77.9
With Multi-Scale Self-Attention	11.5	3.0	7.2	90.2

Firstly, let us validate the importance of the multi-scale feature fusion improvement. In the ablation experiment, we compare the three dilated convolutions carrying different dilation rates (with multi-scale self-attention), with three dilated convolutions carrying the same dilation rate of 1 (without multi-scale self-attention). As shown in Table V, the experimental results show that multi-scale feature fusion remarkably improves the performance of the model.

TABLE VI
RESULTS OF ABLATION EXPERIMENTS ON THE SELECTION OF ACTIVATION FUNCTION.

Activation Function	APCER	BPCER	ACER	ACC
RELU	15.7	6.1	10.9	77.5
GELU	10.7	4.0	7.4	86.0
SiLU	11.5	3.0	7.2	90.2

Secondly, in order to illustrate the impact of different activation functions on the performance of the detection model, we conduct ablation experiments on the selection of activation function. We replace the SiLU function in the recursive self-attention module with the commonly used ReLU function in computer vision and the Gaussian Error Linear Units (GELU) function. Table VI results show that the SiLU function remarkably enhances the performance of the model.

TABLE VII
RESULTS OF ABLATION EXPERIMENTS ON THE RECURSIVE SELF-ATTENTION.

Method	OULU	SiW	WMCA	CeFA
Without Recursive Self-attention	13.5	9.5	11.5	80.2
With Recursive Self-attention	11.5	3.0	7.2	90.2

Thirdly, in order to demonstrate the significant role played by recursive self-attention in model, we conduct ablation experiments on the recursive self-attention module. As shown in Table VII, "Without Recursive Self-attention" refers to the self-attention mechanism with only one multi-scale feature fusion operation, while "With Recursive Self-attention" refers to a recursive structure formed by using three multi-scale feature fusion operations in the self-attention mechanism, which is the structure employed in our proposed method. By observation, it can be seen that the use of recursive self-attention mechanism effectively improving the model's detection performance. By employing a recursive operation combined with three-level feature fusion, the depth of the feature extraction network is increased, thereby enhancing the model's capability to extract abstract features at deeper levels. Consequently, this leads to an improvement in the model's detection performance.

D. Analysis of computational complexity

TABLE VIII
COMPARISON OF POWER CONSUMPTION.

Method	Parameter Volume ($\times 10^6$)	FLOPs ($\times 10^9$)
STDN [17]	1.3	79.8
ViT-FAS [45]	85.8	33.7
FGDNet [27]	20.1	6.8
Ours	10.1	6.2

Finally, we validate the computational parameter and power consumption of the proposed model through comparative experiments. As shown in Table VIII, it can be seen that compared to the STDN model based on GAN [46] network and the ViT model based on the original Transformer network, the proposed method has significant advantages in computational complexity. Moreover, due to the design of shared weights in the recursive self-attention with atrous convolution kernel, the parameter volume of our proposed method is significantly lower than that of FGDNet, and the power consumption is slightly lower than that of FGDNet.

IV. CONCLUSIONS

In this work, we propose a novel face anti-spoofing method that leverages recursive self-attention and multi-scale feature fusion. To increase model depth without significantly increasing parameter volume, we introduce RNN into the self-attention mechanism. To extract local features, we employ a

self-attention module based on stride convolution and a recursive self-attention module with three consecutive parameter-sharing mechanisms and residual connections. Additionally, we introduce multiple weight-sharing dilated convolutions to extract features at different scales and improve the loss and activation functions. Our proposed method basically better than the state-of-the-art performance on the benchmark CeFA, WMCA, and Hifimask datasets, and the ablation experiments demonstrate the importance of key components.

V. Acknowledgements

This work was supported by Zhejiang Provincial Natural Science Foundation of China (No. LZ23F020006), the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK219909299001-007), the Open Projects Program of National Laboratory of Pattern Recognition, the Key Research Project of Zhejiang Province (No. 2023C01025), the National Natural Science Foundation of China (No.62376246) .

References

[1] Z. Xia, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Towards deepfake video forensics based on facial textural disparities in multi-color channels," *Information Sciences*, vol. 607, pp. 654–669, 2022.

[2] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised learning-based framework for deepfake video detection," *IEEE Transactions on Multimedia*, 2022.

[3] T. Qiao, Y. Chen, X. Zhou, *et al.*, "Csc-net: Cross-color spatial co-occurrence matrix network for detecting synthesized fake images," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[4] S. R. Ahmed, E. Sonuç, M. R. Ahmed, and A. D. Duru, "Analysis survey on deepfake detection and recognition with convolutional neural networks," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, 2022, pp. 1–7.

[5] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2016.

[6] K. Ito, T. Okano, and T. Aoki, "Recent advances in biometrie security: A case study of liveness detection in face recognition," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2017, pp. 220–227.

[7] X. Yang, W. Luo, L. Bao, *et al.*, "Face anti-spoofing: Model matters, so does data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3507–3516.

[8] Y. Du, T. Qiao, M. Xu, and N. Zheng, "Towards face presentation attack detection based on residual color texture representation," *Security and Communication Networks*, vol. 2021, 2021.

[9] K. Watanabe, K. Ito, and T. Aoki, "Spoofing attack detection in face recognition system using vision transformer with patch-wise data augmentation," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 1561–1565.

[10] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[11] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*, IEEE, 2012, pp. 26–31.

[12] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3d mask face anti-spoofing database with real world variations," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 100–106.

[13] P. C. Yuen, S. Liu, S. Zhang, and G. Zhao, *3d mask face anti-spoofing with remote photoplethysmography*, US Patent 10,380,444, Aug. 2019.

[14] A. Liu, C. Zhao, Z. Yu, *et al.*, "3d high-fidelity mask face presentation attack detection challenge," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 814–823.

[15] D. Deb and A. K. Jain, "Look locally infer globally: A generalizable face anti-spoofing approach," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1143–1157, 2020.

[16] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[17] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *European Conference on Computer Vision*, Springer, 2020, pp. 406–422.

[18] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[19] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[20] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 11 974–11 981.

[21] X. Tu, Z. Ma, J. Zhao, G. Du, M. Xie, and J. Feng, "Learning generalizable and identity-discriminative representations for face anti-spoofing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–19, 2020.

[22] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4680–4689.

[23] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "Nas-fas: Static-dynamic central difference network search for face anti-spoofing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[25] Z. Wang, Z. Wang, Z. Yu, *et al.*, "Domain generalization via shuffled style assembly for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4123–4133.

[26] Q. Zhou, K.-Y. Zhang, T. Yao, *et al.*, "Generative domain adaptation for face anti-spoofing," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, Springer, 2022, pp. 335–356.

[27] T. Qiao, J. Wu, N. Zheng, M. Xu, and X. Luo, "Fgdnet: Fine-grained detection network towards face anti-spoofing," *IEEE Transactions on Multimedia*, 2022.

[28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[29] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning.," *Neural Networks*, 2018.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[31] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," *arXiv: Learning*, 2021.

[32] J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv: Machine Learning*, 2016.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv: Learning*, 2015.

[34] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," *Cornell University - arXiv*, 2019.

[35] Z. Yao, Y. Cao, Y. Lin, Z. Liu, Z. Zhang, and H. Hu, "Leveraging batch normalization for vision transformers," *International Conference on Computer Vision*, 2021.

[36] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1179–1187.

[37] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, 2019.

[38] A. Liu, C. Zhao, Z. Yu, J. Wan, and D. Zhang, "Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection," 2021.

[39] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*, Springer, 2020, pp. 152–168.

[40] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2017.

[41] *Iso/iec jtc 1/sc 37 biometrics. information technology biometric presentation attack detection part 1: Framework*, https://www.iso.org/obp/ui/iso, international organization for standardization, 2016.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[43] A. Liu, Z. Tan, J. Wan, *et al.*, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2759–2772, 2021.

[44] Z. Yu, C. Zhao, Z. Wang, *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[45] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2021, pp. 1–8.

[46] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 2014.

StyleStegan: Leak-free Style Transfer Based on Feature Steganography

Xiujian Liang 1, Bingshan Liu 2, Qichao Ying 3, Zhenxing Qian 4*, Hsunfang Cho 5, and Xinpeng Zhang 6
1 Fudan University, China
E-mail: liangxj23@m.fudan.edu.cn
2 Engineering Research Center of Digital Forensics, China
E-mail: 202183290228@nuist.edu.cn
3 Fudan University, China
E-mail: 20110240050@fudan.edu.cn
4* Fudan University, China
E-mail: zxqian@fudan.edu.cn
5 Shanghai Data Exchange Corporation, China
E-mail: alex-zhuo@chinadep.com
6 Fudan University, China
E-mail: zhangxinpeng@fudan.edu.cn

***Abstract* — In modern social networks, existing style transfer methods suffer from a serious content leakage issue, which hampers the ability to achieve serial and reversible stylization, thereby hindering the further propagation of stylized images in social networks. To address this problem, we propose a leak-free style transfer method based on feature steganography. Our method consists of two main components: a style transfer method that accomplishes artistic stylization on the original image and an image steganography method that embeds content feature secrets on the stylized image. The main contributions of our work are as follows: 1) We identify and explain the phenomenon of content leakage and its underlying causes, which arise from content inconsistencies between the original image and its subsequent stylized image. 2) We design a neural flow model for achieving loss-free and biased-free style transfer. 3) We introduce steganography to hide content feature information on the stylized image and control the subsequent usage rights. 4) We conduct comprehensive experimental validation using publicly available datasets MS-COCO and Wikiart. The results demonstrate that StyleStegan successfully mitigates the content leakage issue in serial and reversible style transfer tasks. The SSIM performance metrics for these tasks are 14.98% and 7.28% higher, respectively, compared to a suboptimal baseline model.**

I. INTRODUCTION

With the rapid development of Internet social networks and AI image generation technology, multimedia communication through the use of digital images has become a prevalent method for information exchange. Style transfer techniques [1] have gained significant popularity in image processing software, allowing users to transform ordinary digital images into captivating artistic styles. However, these existing style transfer methods inadvertently introduce modifications to the subtle textures and content information within the images. Consequently, when performing serial style transfers with a stylized image, severe corruption and artifacts occur within the content features, giving rise to the content leakage phenomenon [2], as illustrated in Figure 1.

The phenomenon of content leakage pertains to the profound transformation undergone by a stylized image, rendering it virtually indiscernible to the unaided human vision when refer to its original content, subsequent to a series of stylistic transmutations, whether proceeding sequentially or reversely. Furthermore, current de-styling methods face challenges in accurately extracting the original content information directly from the stylized image, often resulting in circular inconsistencies during the image reconstruction process [3].

The presence of content leakage hinders the achievement of high-quality serial style transfer and reverse style transfer, thereby impeding the widespread dissemination of stylized images on social networks. Furthermore, the increasing attention from social users to privacy security and copyright protection issues arising from the dissemination of stylized images in social networks highlights the need for an effective

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



Fig. 1 Content leakage phenomenon in style transfer.

solution to control subsequent stylized permissions. This study aims to address two key problems:

1) How to achieve leak-free style transfer in social networks?

2) How to control the subsequent usage of original content?

To address the first problem, we have designed a reversible neural flow network as a fundamental solution for achieving leak-free style transfer. This network eliminates content reconstruction loss during the encoding-decoding process and mitigates feature offset errors in the style transformation process. To tackle the second problem, we have introduced image steganography as a means to embed content features into stylized images. This approach enables precise control over subsequent stylized content usage rights. Only social users with specific extraction methods can extract the embedded content features, allowing them to perform high-quality stylization while preserving the original content. By integrating the aforementioned methods, we have successfully resolved the issues of content leakage and subsequent control of stylized images in existing social networks. This achievement not only allows for the more seamless and controlled dissemination of stylized images within social networks, ensuring the preservation of their artistic effects and maximizing their impact during dissemination, but also hold significant practical implications for the diverse transmission of world art.

II. Related Work

The successful application of deep learning in computer vision tasks has opened up new possibilities for image style transfer, as initially proposed by [4] using the concept of image iteration. This approach utilized pre-trained convolutional neural networks to extract content and style features and encode them into semantic representations. Building upon this foundation, [1] further advanced the previous method [4] by incorporating additional control over spatial location, color information, and cross-space scales, resulting in significant improvements in the quality of stylized images. However, it is important to note that image iterative methods suffer from slow stylization due to the iterative pixel-by-pixel optimization required for matching style features between the output and input images.

To accelerate the style transfer process, [5] introduced a feedforward model approach aimed at achieving real-time style transfer by training feedforward networks to approximate the iterative optimization process. However, this approach prioritized transfer speed at the expense of image quality degradation. In order to address the trade-off between transfer speed and image quality, [6] made significant improvements to the network proposed in [5] by replacing the batch normalization layer with an instance normalization layer. Additionally, [7] introduced conditional instance normalization to enable real-time style transfer across multiple styles learned during training. Despite these advancements, both image iteration-based and feedforward model-based style transfer methods faced a common challenge in practical applications: they were limited to specific styles and struggled to generalize to arbitrary stylizations. To overcome these limitations, the Adaptive Instance Normalization (AdaIN) method [8] was developed. [8] investigated the impact of statistical properties of stylized image features on the stylization process and proposed an affine transformation based on the mean and standard deviation of stylized images. AdaIN had a significant influence on subsequent research in style transfer, leading to the proposal of various style transfer methods [9-16]. As a result, style transfer entered the era of arbitrary style transfer.

In practical applications on social networks, stylized images have the potential to alter the original content information. This leads to a significant issue of content leakage when social users perform serial and reverse stylization on these stylized images. While [17] focused on de-styling, their work only addressed face de-styling. To tackle this problem, [18] introduced STNet, a steganographic network based on style transfer, which embeds secret information within the style features of the stylized image. Building upon this, [19] proposed PSTNet, which enhances embedding capacity and steganalysis resistance during the stylization process. Motivated by advancements in steganography, [20] attempted to embed content image features on stylized images, while [21] explored hiding content image features among layered stylized features. However, the content features still experience changes in texture details and spatial locations. Existing image steganography-based style transfer methods only address the loss of content from the stylized images generated after transfer, failing to provide a fundamental solution to the content leakage problem that occurs during the transfer process.



Fig. 2 Test results of iterative inference with the AdaIN

III. PRE-ANALYSIS

A. Reasons

To address the content leakage problem, we first investigate and summarize the causes, which can be categorized as follows:

1. Lossy image reconstruction

The model architecture commonly adopted by existing methods for style transfer consists of encoder-decoder pairs. This architecture is based on the intuitive explanation that the decoders of existing style transfer methods cannot achieve lossless image reconstruction of the input content images. To illustrate this point, three methods [22-24] use VGG19 as an encoder and train a decoder to invert the features of VGG19 into the image space. However, despite using image reconstruction loss or content loss functions to train the decoder, it was found in [23] that the feature inversion is imperfect due to the loss of spatial information during the pooling operation in the encoder. Consequently, the accumulated errors in image reconstruction can gradually interfere with the content details and result in content leakage.

2. Biased decoding training

In addition to the image reconstruction loss, which can explain part of the content leakage phenomenon, the biased decoder training scheme is another major contributing factor. In this case, we will use the training scheme of AdaIN as an example to explain how its loss function is configured in a way that leads to content leakage. Specifically, AdaIN trains the decoder using a weighted combination of content loss and style loss, where

$$L_{content} = \left\| E(D(t)) - t \right\|_2 \tag{1}$$

$$L_{style} = \sum_{i=1}^{L} \left\| \mu(\varphi_i(D(t))) - \mu(\varphi_i(s)) \right\|_2 + \sum_{i=1}^{L} \left\| \sigma(\varphi_i(D(t))) - \sigma(\varphi_i(s)) \right\|_2 \tag{2}$$

In Eqs. (1) and (2), t denotes the output of AdaIN, E, D represent the encoder and decoder respectively, φ_i denotes the layer used to compute the style loss in VGG19, μ, σ represents the mean and standard deviation of the feature map.

Figure 2 illustrates the process of style transfer using the AdaIN self-encoder. Without any style image for transfer guidance, a content image is given as input. Each round involves using the output of the previous round's decoder as the input for the next round's encoder. As the number of inference rounds increases, the generated results start exhibiting strange artistic patterns. This phenomenon suggests that during training, AdaIN's encoder is biased towards style rendering.

3. Biased transfer module

In addition to the image reconstruction loss and decoding training bias, the transfer principle of some style transfer modules itself introduces a bias for the content features. Let's take the style decorator module in Avatar-Net as an example. The style decorator, inspired by deep image analogy [25], consists of two main steps. In the first step, the algorithm identifies a corresponding feature block in based on the content similarity between two feature blocks from . Then, in the second step, it finds a corresponding feature block in by replacing the corresponding feature block in . This replacement operation creates a feature block that cannot be recovered from , resulting in a bias towards stylistic features and content leakage.

B. Remedies

Based on the above analysis, realizing loss-free and bias-free style transfer is a necessary prerequisite to solve the content leakage problem. To achieve loss-free feature extraction, we propose a feature extraction network based on the neural flow model. This network maps the content features and style features of an image into the latent space in a loss-free manner, which is then inputted to the style transfer module. The specific details of the network will be introduced in Section IV. In addition, to achieve bias-free style transfer, we introduce the AdaIN-based style transfer module. This module ensures separable and bias-free image content and style features. The reasons for introducing this module will be described below.

The mechanism of the popular generic style transfer methods at this stage can be seen as a natural evolution of the bilinear model [26]. The model separates image features into content factors and style factors, and then performs style transfer by replacing the style factors in the content image with the style factors in the target image.

Definition 1 *Suppose there is a bilinear style transfer module* $f_{cs} = C(f_c)S(f_s)$*, where* C, S *denotes the content factor and style factor in the bilinear model, respectively. If* $C(f_{cs}) = C(f_c)$ *and* $S(f_{cs}) = S(f_s)$*, then* f_{cs} *is an unbiased style transfer module.*

Theorem 1 *The adaptive instance normalization in AdaIN is an unbiased style transfer module.*

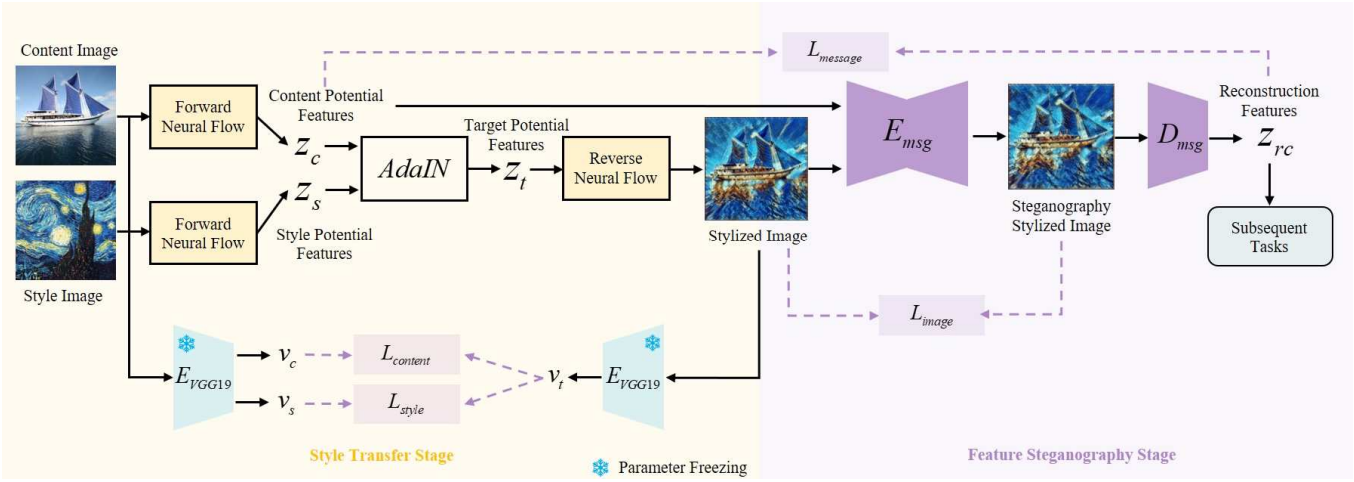


Fig. 3 The overview of proposed StyleStegan

Proof 1 *Assuming that both* f_c *and* f_s *satisfy the generality of the data distribution, it follows that:*

$$f_{cs} = C(f_c)S(f_s) = \frac{f_c}{\sigma(f_c)}\sigma(f_s) \tag{3}$$

where,

$$C(f) = \frac{f}{\sigma(f)}, \quad S(f) = \sigma(f) \tag{4}$$

since,

$$\sigma(f_{cs}) = \sigma\left(\frac{f_c}{\sigma(f_c)}\right)\sigma(f_s) = 1 \times \sigma(f_s) = \sigma(f_s) \tag{5}$$

we have,

$$C(f_{cs}) = \frac{f_{cs}}{\sigma(f_{cs})} = \frac{f_c}{\sigma(f_c)} = C(f_c) \tag{6}$$

$$S(f_{cs}) = \sigma(f_{cs}) = \sigma(f_s) = S(f_s) \tag{7}$$

Therefore, the adaptive instance normalization in AdaIN is an unbiased style transfer module.

IV. Proposed Method

A. Overview of StyleStegan

To address the two issues mentioned, namely achieving leak-free style transfer in social networks and controlling the subsequent usage of stylized original content, we propose a feature steganography-based approach. Our method involves a two-stage model that combines style transfer and feature steganography, as depicted in Fig. 3.

During the style transfer phase, we employ a reversible neural stream-based feature network to map the features of the style image I_s and the content image I_c to the latent space without loss, resulting in z_s and z_c. We then utilize the AdaIN-based unbiased style transfer module to perform stylization on the latent features. Finally, we project the target latent features back into the image using the reversible neural stream, yielding the stylized image I_t.

In the feature steganography phase, the steganographic network is trained to learn a feature steganography encoder. This encoder hides the content features in the latent space into the stylized image completed in the previous phase, resulting in a steganographic stylized image. Additionally, a feature extraction decoder is learned in conjunction with the feature steganography encoder. The purpose of this decoder is to extract content feature information from the steganography stylized image. The steganography stylized image is pre-embedded by the social user who owns the content copyright. Within this stage, the method offers flexibility in parameter *–stage* settings. It can be configured as a two-stage end-to-end model to complete the subsequent stylization of the social user's original content usage rights. The choice depends on the social user's preferences.

B. Style Transfer Stage

In the style transfer stage, we borrowed the idea of the Glow model [27] to construct a reversible neural flow-based feature network. This network serves as both a deep feature extractor and an image synthesizer within the overall framework. Figure

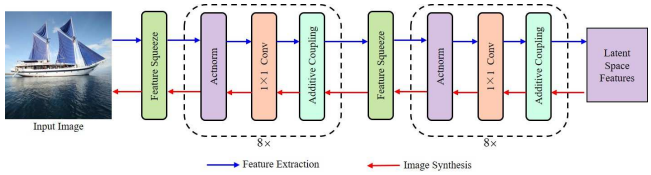


Fig. 4 Reversible neural flow based feature network

4 illustrates the network, which comprises three reversible transformations with learning parameters: additive coupling, reversible 1 × 1 convolution, and activation normalization. It is worth noting that all components of this network are reversible, ensuring loss-free information propagation during both forward and backward propagation. The specific description of the network is provided in the following section.

1. Additive coupling layer

[28,29] proposed an expressive and reversible layer called affine coupling. This network, however, utilizes a specific variant of affine coupling known as additive coupling. In the additive coupling layer, the parameters undergo positive calculations, as demonstrated in Equations (8) to (10):

$$x_a,\ x_b = split(x) \tag{8}$$

$$y_b = NN(x_a) + x_b \tag{9}$$

$$y = concat(x_a,\ y_b) \tag{10}$$

where function $split()$ divides a tensor x precisely along the channel dimension, while $NN()$ being an inherently symmetrical neural network boasting inputs and outputs of identical dimensions. Function $concat()$ adeptly interlaces two tensors along the channel dimension.

2. Reversible 1×1 convolutional layer

To ensure that all dimensions of the feature maps are equally affected, the channel dimensions of the feature maps in additive coupling processes should be swapped. For this purpose, a learnable reversible 1 × 1 convolutional layer is employed in the second transformation of this network, allowing for flexible channel permutation. The forward calculation of this layer is shown in Equation (11):

$$y_{i,j} = \mathrm{M}x_{i,j} \tag{11}$$

where M represents a weight matrix of dimensions $c \times c$, wherein c denotes the channel dimension of tensors x and y. The function responsible for the inverse computation of this weight matrix is denoted as $x_{i,j} = \mathrm{M}^{-1} y_{i,j}$.

3. Activating Normalization layer

The final transformation in this network replaces the conventional batch normalization method with activation normalization. Activation normalization involves applying an affine transformation to each channel of the tensor, as shown in Equation (12):

$$y_{i,j} = w \square\ x_{i,j} + b \tag{12}$$

where i, j symbolize the spatial coordinates within the tensor, w, b represent the erudite weights and bias parameters governing affine transformation. The inverse computational function is elegantly denoted as $x_{i,j} = (y_{i,j} - b) / w$

In this stage, the above three layers of reversible transforms combine to form a stream, and every eight streams create a block. Alongside the three transformations mentioned earlier, the compression operation divides the features into smaller feature blocks based on the spatial dimension, and then merges these blocks along the channel dimension. This compression operation is inserted between the input and block-to-block of the network, aiming to decrease the spatial size of the two-dimensional feature map. The loss function in this stage is determined by weighting the content loss and style features, which are calculated using the same equations as Equation (1) and (2), and will not be elaborated here.

C. Feature Steganography Stage

In the feature steganography stage, the network contains a steganography encoder and a corresponding steganographic decoder. The steganography encoder hides the content features z_c into the stylized image I_t and produces the encoded image I_e, which is the output of the two-stage model. And the steganography decoder tries to decode z_c from I_e.

As a steganographic scheme, the encoded image I_e and the stylized image I_t should exhibit a visually imperceptible difference. To achieve this, the steganography encoder is trained to minimize the disparity between the two images. This optimization is carried out using a loss function defined as:

$$L_{image} = \left\| I_e - I_t \right\|_2 \tag{13}$$

The Steganography encoder consists of 1 input layer, 3 convolutional layers, and 1 output layer. It takes the stylized image as input and performs splicing embedding of steganographic features in the channel dimension. The encoder continuously optimizes the embedding of secret information and image visual quality throughout the remaining convolutional layers. Finally, it produces the steganographic stylized image as output.

The steganographic decoder is trained to optimize the extraction of content features from the steganographic stylized image. This is achieved by minimizing the loss function, as defined in Equation (14):

$$L_{message} = \left\| D_{msg}(I_e) - z_c \right\|_2 \tag{14}$$

The steganography decoder is composed of 1 input layer, 6 convolutional layers, and 1 output layer. It takes the steganographic stylized image as input and continuously optimizes the extraction accuracy of steganographic features in the 6 convolutional layers.

Finally, it outputs the extracted steganographic features. The objectives of the feature steganography phase can be summarized as Equation (15):

$$L_{steganography} = \lambda_{img} L_{image} + \lambda_{msg} L_{message} \tag{15}$$

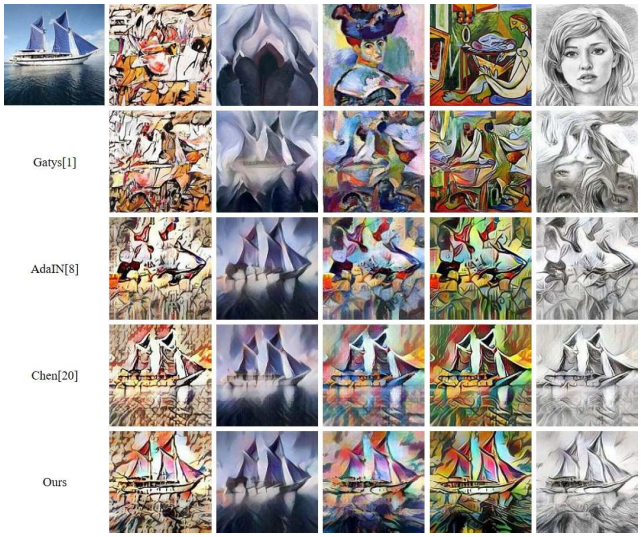


Fig. 5 Comparison of the effectiveness in serial style transfer tasks

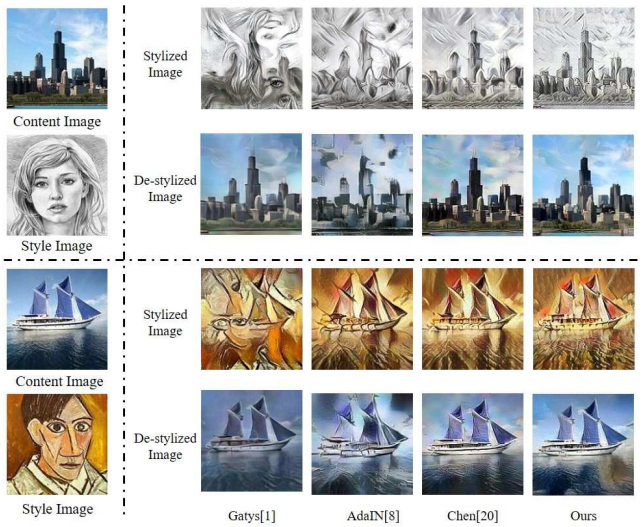


Fig. 6 Comparison of the effectiveness in reverse style transfer tasks

V. EXPERIMENTS

A. Dataset and Environment

This paper follows a similar dataset setup as described in [8] to build the training set for the model. The training dataset is constructed by randomly selecting 20,000 content images from the training set of MS-COCO [30] and 20,000 style images from the training set of WikiArt [31]. To ensure consistency, the selected images were first resized to a minimum size of 512 while maintaining the aspect ratio. Subsequently, random cropping was applied to obtain images of size 256 × 256. For the test dataset, 1000 content images and 1000 style images were randomly chosen from the test sets of MS-COCO and WikiArt respectively, without any crossover with the training dataset. During stylization, each content image was paired with one of the 10 randomly selected style images from the test set. Importantly, the test images used for experimentation were never included in the training process to prevent bias during learning and inference.

For the experimental environment setup, the Google online computing platform Colaboratory was used for training and inference with a Tesla T4 GPU with a video memory size of 15 GB, a system RAM size of 16 GB, a disk memory size of 128 GB, a CUDA version number of 12.0, a Python version number of 3.8, and the deep learning framework used was Pytorch, whose version is 2.0.0+cu118.

B. Visual Comparison

In the experimental design, we compare the proposed model in this paper with the baseline model of [1], the adaptive instance normalized baseline model of [8], and the end-to-end model of [30]. These comparisons are made to validate the qualitative results through serial and reverse style transfer tasks.

1. Serial style transfer

Serial style transfer aims to apply a new style to a stylized image while minimizing the impact of previous stylizations. The desired outcome of serial style transfer is to achieve an artistic rendering effect similar to directly and separately stylizing the original content with the new style. In the experiments conducted for this task, we compare the results of our method with existing methods, as depicted in Figure 7.

The obtained results demonstrate that our method effectively reduces the corruption of content features caused by intermediate style transfers. This is achieved by selectively and discreetly embedding content features into each stylized image, catering to the specific requirements of users for subsequent style transfers. By doing so, the original content features can be extracted and utilized in each style transfer operation.

2. Reverse style transfer

The goal of reverse style transfer is to de-stylize a stylized image while preserving its original content as much as possible. Ideally, the outcome of reverse style transfer should closely resemble the input content image. However, the baseline model of Gatys et al [1] and AdaIN [8] do not provide a direct approach for reverse style transfer using the original content image as the reference for the target style and applying the stylization to the given stylized image. This method requires access to the original content image. In contrast, the model developed by Chen et al [20] and the present method can extract the hidden features necessary for de-styling using only the given stylized image. To evaluate the effectiveness of the

Tab. 1 Quantitative experimental Results of serial and reverse style transfer

		Gatys[1]	AdaIN[8]	Chen[20]	Ours
Serial	L2↓	7.5239	0.0213	0.0148	**0.0104**
	SSIM↑	0.0473	0.5471	0.7143	**0.8213**
	LPIPS↓	0.4316	0.4614	0.2437	**0.1483**
Reverse	L2↓	4.4331	0.0368	**0.0187**	0.0193
	SSIM↑	0.2041	0.3826	0.4796	**0.5145**
	LPIPS↓	0.3694	0.4628	**0.3323**	0.3802

present method in reverse style transfer, it is compared with existing methods, and the experimental results are presented in Figure 6. The results demonstrate that the proposed method achieves better content image reconstruction in the reverse style transfer task.

C. Quantitative Comparison

Quantitative experiments entail the application of different models for conducting serial style transfer and reverse style transfer tasks. These models generate output images, which are subsequently compared to their respective desired images. The aim of this comparison is to assess the performance of the models in both serial style transfer and reverse style transfer. The findings from these quantitative experiments are presented in Table 1.

VI. CONCLUSIONS

We address the issue of content leakage that commonly arises in serial and reverse style transfer and propose a corresponding solution. Our experimental results demonstrate the superiority of our method in both serial and reverse style transfer tasks. Specifically, our method achieves a 14.98% and 7.28% higher SSIM compared to the suboptimal performing baseline model for the two tasks, respectively. The specific contributions of our work are as follows:

1. We analyze the phenomenon of content leakage in typical style transfer methods and identify image reconstruction loss, decoding training bias, and transfer module bias as potential causes.

2. To ensure content integrity and style consistency in stylized images, we design loss-free and bias-free reversible neural flow networks. These networks eliminate content reconstruction loss during encoding and decoding, as well as feature bias error during style transfer.

3. We develop a model for content feature steganography and extraction, which allows the embedding of content features into stylized images. This enables subsequent high-quality stylization while ensuring that only authorized users with specific extraction methods can retrieve the original content, providing control over the stylized images.

4. We conduct experimental validation on the MSCOCO and Wikiart public datasets to showcase the visual effectiveness and outstanding performance of our method in generating stylized images. Our results provide solid evidence for the effectiveness of our approach in addressing the content leakage problem in serial and reverse style transfer tasks.

ACKNOWLEDGMENT

This work was supported by Shanghai Data Exchange Corporation. Xiujian Liang and Bingshan Liu acknowledge the support from the National Student Innovation and Entrepreneurship Training Program Support Project of China (202210300028Z). Thanks to the Multimedia Artificial Intelligent Security Laboratory of Fudan University for the research direction guidance and experimental support.

REFERENCES

[1] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[2] Wei, Hua-Peng, et al. "A Comparative Study of CNN-and Transformer-Based Visual Style Transfer." *Journal of Computer Science and Technology* 37.3 (2022): 601-614.

[3] Wang, Hao, et al. "Cycle-consistent inverse GAN for text-to-image synthesis." *Proceedings of the 29th ACM International Conference on Multimedia*. 2021.

[4] Gatys, Leon, Alexander S. Ecker, and Matthias Bethge. "Texture synthesis using convolutional neural networks." *Advances in neural information processing systems* 28 (2015).

[5] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer International Publishing, 2016.

[6] Ulyanov, Dmitry, et al. "Texture networks: Feed-forward synthesis of textures and stylized images." *arXiv preprint arXiv:1603.03417* (2016).

[7] Dumoulin, Vincent, Jonathon Shlens, and Manjunath Kudlur. "A learned representation for artistic style." *arXiv preprint arXiv:1610.07629* (2016).

[8] Huang, Xun, and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization." *Proceedings of the IEEE international conference on computer vision*. 2017.

[9] Huang, Siyu, et al. "Parameter-Free Style Projection for Arbitrary Image Style Transfer." *ICASSP 2022-2022 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[10] Wang, Zhizhong, et al. "AesUST: towards aesthetic-enhanced universal style transfer." *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.

[11] Liu, Zhi-Song, et al. "Name your style: An arbitrary artist-aware image style transfer." *arXiv preprint arXiv:2202.13562* (2022).

[12] Wang, Quan, et al. "Multi-granularity brushstrokes network for universal style transfer." *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)* 18.4 (2022): 1-17.

[13] Kwon, Gihyun, and Jong Chul Ye. "Clipstyler: Image style transfer with a single text condition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[14] Chen, Zhe, et al. "Towards ultra-resolution neural style transfer via thumbnail instance normalization." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 1. 2022.

[15] Zhang, Yuxin, et al. "Domain enhanced arbitrary image style transfer via contrastive learning." *ACM SIGGRAPH 2022 Conference Proceedings*. 2022.

[16] Zhang, Yabin, et al. "Exact feature distribution matching for arbitrary style transfer and domain generalization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[17] Shiri, Fatemeh, et al. "Face destylization." *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2017.

[18] Wang, Zihan, et al. "STNet: A style transformation network for deep image steganography." *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part II 26*. Springer International Publishing, 2019.

[19] Xue, Yuliang, et al. "PSTNet: Protectable Style Transfer Network Based on Steganography." *2022 International Conference on Culture-Oriented Science and Technology (CoST)*. IEEE, 2022.

[20] Chen, Hung-Yu, et al. "Self-contained stylization via steganography for reverse and serial style transfer." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.

[21] Ubhi, Jagpal Singh, and Ashwani Kumar Aggarwal. "Neural style transfer for image within images and conditional GANs for destylization." *Journal of Visual Communication and Image Representation* 85 (2022): 103483.

[22] Gu, Shuyang, et al. "Arbitrary style transfer with deep feature reshuffle." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[23] Li, Yijun, et al. "Universal style transfer via feature transforms." *Advances in neural information processing systems* 30 (2017).

[24] Sheng, Lu, et al. "Avatar-net: Multi-scale zero-shot style transfer by feature decoration." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[25] Liao, Jing, et al. "Visual attribute transfer through deep image analogy." *arXiv preprint arXiv:1705.01088* (2017).

[26] Tenenbaum, Joshua B., and William T. Freeman. "Separating style and content with bilinear models." *Neural computation* 12.6 (2000): 1247-1283.

[27] Kingma, Durk P., and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions." *Advances in neural information processing systems* 31 (2018).

[28] Dinh, Laurent, David Krueger, and Yoshua Bengio. "Nice: Non-linear independent components estimation." *arXiv preprint arXiv:1410.8516* (2014).

[29] Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp." *arXiv preprint arXiv:1605.08803* (2016).

[30] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014.

[31] K. Nichol. Painter by numbers, wikiart. https://www. kaggle.com/c/painter-by-numbers, 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Robust Watermark Imaging via Graph-signal Optimization

Ruiguo Yang*, Xinhui Han*, Wenfa Qi*, Wei Hu*
* Wangxuan Institute of Computer Technology, Peking University
E-mail: {yangruiguo, hanxinhui, qiwenfa, forhuwei}@pku.edu.cn

***Abstract*—The topic of robust watermarking has long been a subject of interest in the field of information hiding, and in these days, the robustness against cross-media attacks, such as screen-capture and printer-capture attacks, have become an increasingly pressing concern. Previous work in the field, such as StegaStamp, one of the most popular open-source screen-shoot resilient watermark scheme, has achieved promising results in terms of decoding accuracy. However, the visual quality of some watermarked images by this scheme is still not inadequate for practical purposes. In this paper, we propose an optimization that could significantly enhances the visual quality of watermarked images. We use a graph-based method to build graph from image blocks, compute a loss designed with inspiration from GLR, and add it to the existing loss function of the neural network during training. As demonstrated by the experiment results, we have achieved substantial improvements in the quality of watermarked images, in both visual quality and metrics, while still maintaining strong robustness and decoding capability.**

I. Introduction

In the era of big data, there is an increased demand from individuals and organizations for the privacy, security, and authenticity of their information. Watermarking technology has emerged as an important solution to meet these needs. By embedding a unique identifier within sensitive documents or images, the authenticity of these files can be confirmed. Additionally, it can be used to trace the source of sensitive information leaks in the confidential departments of companies. To ensure that the watermark remains intact and recognizable during multiple transmissions, the robustness of the watermark is a crucial factor to consider.

Previous studies on robust image watermarking[1], [2] have focused on resisting attacks on the digital channel, such as noise, rotation, and JPEG compression, among others. These types of attacks are common in digital transmissions. However, with the proliferation of multimedia devices such as projectors, cameras, printers, and scanners, the way information is transmitted has changed dramatically. Smartphone performance enhancements allow eavesdroppers to more stealthily capture high-fidelity and less traceable documents, which has resulted in substantial business losses. Therefore, cross-media distortions must also be taken into account in the robustness of the watermark. For instance, in tracking the origin of sensitive information leaks from a screen-captured image by a mobile phone, or verifying the authenticity of a photo on the passport, new attacks such as Moiré patterns and perspective warps (as shown in Figure 1) can significantly affect the accuracy of watermark recognition. Compared to distortions on the digital channel, which has been studied for a long time and can be accurately simulated, the cross-media distortion is more complex and random in nature, making quantitative analysis challenging. This type of attack imposes higher demands on the robustness of watermarks.



Fig. 1. The serious distortion imported by screen-capture

Several methods have been proposed to address these challenges, such as the intensity-based SIFT algorithm and DCT-based method proposed by Han et al.[3]. While this work provides important new insights, its accuracy in terms of keypoint location and message decoding is not adequate. Another method, StegaStamp[4], which has almost became the most popular open source robust watermark scheme since its publication, uses a learned steganographic algorithm to achieve high bit accuracy in a variety of scenarios and seems to perform well in terms of watermarked image quality, also referred to as perceptual invisibility of watermark. However, the paper only shows the performance of watermarking on images with complex textures. Through our attempts, we found that it performs poorly on images with relatively simple textures. Let's take the most famous Lenna image as an example(shown in Figure 2). After applying StegaStamp watermarking, many areas in the face and shoulders of the image appear to be blurry or *dirty*.

To improve the perceptual invisibility of robust watermark, we consider graph-based methods, which have been successfully used in fields such as image denoising, image deblurring, and 3D point cloud denoising[5]–[7]. All of these tasks seem to be related to our task of reducing *dirty* watermarks from images. Among these methods, we found that GLR (Graph Laplacian Regularizer) is often used to establish similarity

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



(a) Original image (b) StegaStamp (c) Ours

Fig. 2. Result of original StegaStamp and our optimized version.

relationships among image pixels and enhance the model's understanding of the spatial structure of the image. This improves the performance of image processing tasks, and gives us a new insight to improve the quality of watermarked images. In this paper, inspired by the GLR, we present an optimization for improving the image quality of deep learning-based watermarking schemes. Our algorithm builds graph from the images and add a new loss function in the training process to reduce the *dirty* appearance of *flat* areas in watermarked images. The algorithm maintains high decoding accuracy against many real-world attacks in the experiment, while significantly improving the perceptual invisibility of watermark. To sum up, the contributions of this paper can be listed as follows:

- We present a novel optimization graph-based method to improve the visual quality of robust watermarked images.
- To demonstrate the effectiveness of our approach, we deployed this optimization on one of the previously most popular robust watermarking schemes, StegaStamp, and showed how our method can make watermarked images more practical and usable in the real world.
- A comprehensive set of experiments were conducted to show the performance of our optimization. In the experimental results, we demonstrated a visible improvement both in visual quality that can be perceived by the human eyes, and in metrics including PSNR and SSIM, and equivalent decoding accuracy compared to the original method under various real-world attacks. Our optimization enables the new scheme to produce better image quality even when embedding a message of three times the length.

II. Related Work

In this section, we overview previous works on robust watermark and graph-based digital image processing.

A. Robust Watermark

The field of robust watermarking has been the subject of extensive research, with many mature methods developed over time. Most of these methods embed watermarks in transformed domains, such as the DFT (Discrete Fourier Transform) or DWT (Discrete Wavelet Transform), to avoid spatial feature degradation that can occur as a result of image processing attacks like noise and compression. For instance, Lagzian et al. [2] presented a robust watermarking scheme that combined DWT and SVD, which demonstrated improved performance against common attacks compared to previous methods. Khedr [1] also proposed a robust watermarking scheme for multiple images that utilized both DWT and DCT for embedding.

However, the emergence of new real-world cross-media attacks has prompted researchers to develop new robust watermarking schemes. Han Fang et al. [3] presented a screen-shooting resilient watermarking scheme that aimed to address distortion in the screen-shooting process. The method used an intensity-based SIFT (scale-invariant feature transform) algorithm for keypoint detection, and proposed an embedding algorithm based on modification of DCT coefficients. However, after some production, we found that locating keypoints accurately in real-world complex environments was still challenging. Additionally, the method was prone to incorrect keypoint detection in simple images and also resulted in noticeable texturing in the plain areas of images, affecting image quality.

In addition to traditional methods, deep learning and neural network algorithms have also been applied in the field of robust watermarking. JiRen[8] proposed the first end-to-end method of watermark with neural network. Kangli [9] proposed a GAN-based robust watermarking scheme, where the generator and decoder played against a discriminator in an encoder-decoder setup. Huynh-The [10] presented a scheme using a convolutional encoder-decoder neural network that was resistant to multiple attacks.

For screen-shoot resilient watermarking using deep learning, the key issue to consider is how to address the non-differentiablility of the cross-media distortion. LFM[11] try to solve this problem by training an additional CDTF(camera display transfer function) function seperately to model the display-camera process, with a dataset of screen-captured images. This function is located between the embed layers and the recover layers, and is trained before the whole end-to-end system. Although it performs not bad, collecting such a dataset is very time-consuming and labor-intensive.

Not long after that, StegaStamp[4] is proposed and outperformed LFM significantly. It choose to use a UNet-based encoder-decoder network to embed the watermark. The key novelty of the strong robustness of this work is that, it analyze and model the distortion process manually, and apply a set of differentiable image perturbations to simulate the screen-camera distortion. This work performs well in terms of bit accuracy and was highly resistant to screen-camera and print-camera attacks compared to LFM. However, when used to embed watermarks in relatively simple images, the watermarked images often appeared *dirty*, particularly in important areas such as human faces and plain areas such as backgrounds. This significant drawback limits its application in critical use cases where images are not always as complex as those presented in the paper.

B. Graph-based Image Processing

Graph-based methods[12] have been widely employed in the domain of image processing for various tasks such as denoising, deblurring, and super-resolution. Normally, GSP(Graph

Signal Processing) is applied in data that can be represented by graph structure naturally, like social network, to deal with relationship among nodes in the network. Although image are typically represented in 2D form, the desire of representing the relationships between pixels in an image has led us also to apply graph signal processing here. The fundamental idea behind these methods is to represent pixels in an image as nodes in a graph and model their relationships as edges with assigned weights. In this way, notions like transform and smoothness can also be derived.

In [13], Pang et al. utilized graph Laplacian regularization for the purpose of image denoising, formulating the problem as an optimization problem. They derive the optimal edge weights for local graph-based filtering using gradient estimates, and analyze the effect of the gradient estimates both theoratically and experimentally. Bai et al. [6] also proposed a graph-based approach for blind image deblurring. They first proved that there is a skeleton image can be used to estimate the blur kernel, converting the blind blurring into non-blind blurring; and then they design a RGTV(reweighted graph total variation) prior for this problem. However, since the original RGTV is an non-convex non-differentiable optimization problem, they leverage the new graph spectral interpretation for RGTV and design a new algorithm that solves for the blur kernel. These studies demonstrate the effectiveness of graph-based methods in preserving the structure and visual content of images.

III. Methodology

In this section, we first review some basic concepts of graph and graph signal processing, and then propose our optimization loss.

A. Preliminary

Graph is a common data structure used to represent relationship among entities. An undirected edge-weighted graph can be represented by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ denotes the node set and $\mathcal{E}$ denote the edge set.

To deal with information on the graph, we need to introduce graph signal processing. Graph signal denotes a function $f : \mathcal{V} \Rightarrow R^n$, which means some n-dimentional signal on each node. A key component in graph signal processing is the graph Laplacian matrix L, which represents the relationships between nodes in the graph and encodes prior knowledge about the graph structure. The graph Laplacian matrix is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{1}$$

where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ is the adjacant matrix of the graph. This matrix works as the operator to compute divergence of the gradient at each node.

B. Loss Design

To first build graph from images, we split a image into 16*16 blocks with step length 8. Each node in the graph denotes a pixel in the block, and the edge weight is defined as

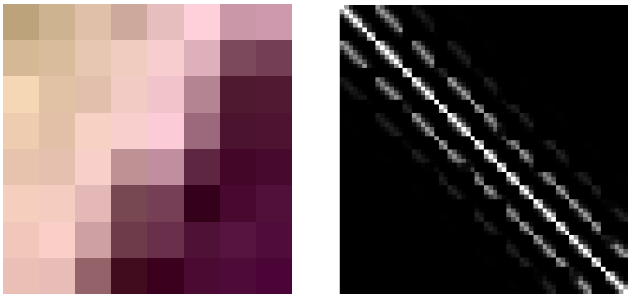


Fig. 3. The image block(left) and the graph built from it(shown in adjacant matrix, black means high edge weight).

$$w_{i,j} = exp(-\delta_{i,j}) \tag{2}$$

where

$$\delta_{i,j} = \frac{(x_i - x_j)^2}{\sigma_x^2} + \frac{(l_i - l_j)^2}{\sigma_y^2} \tag{3}$$

denotes the estimated feature distance between pixel i and j[14], which follows the design of bilateral filtering, taking into account both the pixel intensity and the location. In the equation above, x_i represents the pixel value, l_i represents the location of node i, and σ_x, σ_y denote the corresponding coefficient of these two factors respectively.

The graph constructed from a image using above formula has the property that, edges between close pixels with similar values have large weights, while edges between pixels with large distance or dissimilar values have small weights. In this way, the *flat* area of the image has large edge weights in the graph. A visualization of this graph construction process is shown in Figure 3. For a 8*8 block example shown in the figure, the graph has 64 nodes where each node denotes a pixel. The relationship between adjacent pixels in the original image is represented in the adjacency matrix by contiguous cells, while adjacent pixels in columns are depicted as parallel lines. The lighter-colored cells in the adjacency matrix correspond to flat regions in the image, while the breakpoints in the white slashes indicate the edges.

As we now have the ability to represent *flat* areas in the image, the next step is to design a loss function that ensures the corresponding areas in the watermarked image remain *flat*. Consider the definition of Graph Laplacian Regularizer(GLR)[15] which is commonly used in works with similar purpose:

$$S_{\mathcal{G}}(\mathbf{u}) = \mathbf{u^T L u} = \sum w_{i,j}(\mathbf{u}_i - \mathbf{u}_j)^2 \tag{4}$$

The GLR represents the smoothness of a graph signal $\mathbf{u}$ with respect to graph G. For our purpose, we want the graph signal of watermarked image is smooth with respect to the graph of the original image. Therefore he formula of our loss is similar to the one above that

$$Loss_{GLR} = \sum_{blocks} \mathbf{u_B}^{\mathbf{T}} \mathbf{L_A u_B} = \sum w_{\mathbf{A}i,j}(\mathbf{u_B}_i - \mathbf{u_B}_j)^2 \tag{5}$$

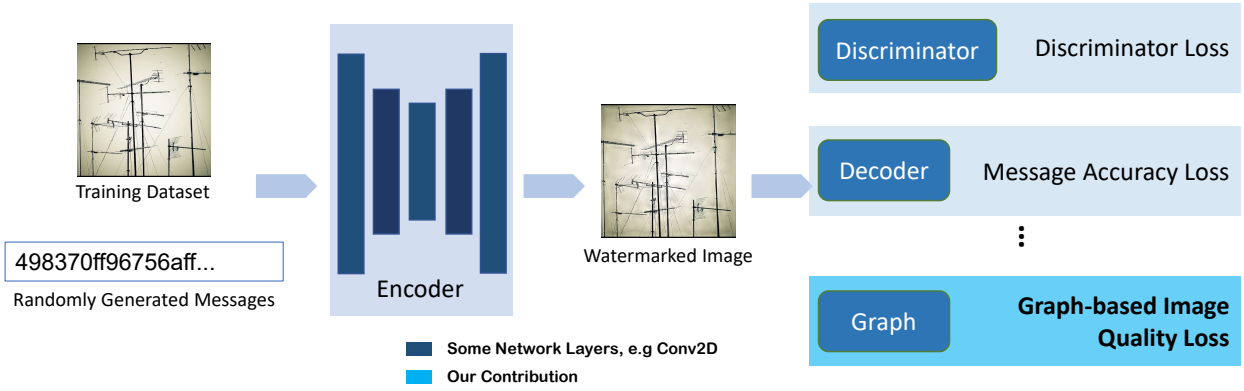


Fig. 4. Sturcture of deep learning based watermarking scheme.

where the variables with subscript **A** denote the original image and the ones with subscript **B** denote the watermarked image.

With this loss function, we want to learn the structure-adaptive watermarked images that's as similar as possible to the original one in *flat* areas. For the watermark generated during training process, if it's still *dirty* in flat areas as in Figure 2, this loss function will punish the intermediate results and let the watermark be more concentrated on areas with complex textures where the variation is less likely to be recognized by the human eye. By minimizing this loss, we can enforce the watermarked image to have graph signal smooth with respect to the original graph.

IV. Experiment

Since our optimization method is general for all deep learning based watermark schemes, we need to utilize it in a specific frame work for experiment. We choose the popular StegaStamp[4] here. In the following subsections, we outline the typical architecture of deep learning-based watermarking and explain how our approach fits into it. Then, we provide a brief overview of the experimental setup. Finally, we present the results of image quality and watermark invisibility improvement through both quantitative metrics and visual examples, and analyze the robustness against various common attacks after our optimization.

A. System Overview

Our method can be applied to most deep learning-based robust watermarking schemes. The typical structure of these schemes is illustrated in Figure 4. In this architecture, images and messages are fed into an encoder-decoder neural network for training. The normal loss during training typically includes the message cross-entropy loss, the discriminator loss of the original image and the watermarked image, and so on. Our main contribution, the GLR loss, is added to this structure to explicitly enhance the visual quality of the watermarked images.

Let's take StegaStamp, the one we choose to embed our method for experiments, as an example.The structure of StegaStamp end-to-end system can be summarized as *Encoder-Perturbation-Decoder*. The *encoder* is a U-Net style architecture that receives a four channel 400 × 400 pixel input (input image RGB channels plus one for the message) and outputs a three channel RGB residual image. The *perturbation* is a combination of distortion simulation functions that are derivable. The *decoder* is a spatial transform network that accepts a transformed image and outputs the decoded message with the same length of original input. The system is trained end-to-end and the loss follows the formula below:

$$L = \lambda_R L_R + \lambda_P L_P + \lambda_C L_C + \lambda_M L_M \tag{6}$$

where L_R, L_P, L_C, L_M denote the L_2 regularization loss, LPIPS[16] loss, critic loss and the cross entropy loss of the message respectively, and the λs denotes the coefficients. Our loss can be directly added here as:

$$L_{new} = L + \lambda_G Loss_{GLR} \tag{7}$$

where $Loss_{GLR}$ is described in Formula 5.

B. Experiment Setup

We choose the same training and testing dataset as the original paper. The training dataset is mirflickr dataset[17]. The testing dataset used for visualization and robustness experiment consists of 100 images randomly sampled from ImageNet[18]. For robustness tests, the messages is encoded with the same BCH error-correcting code as in the original paper, with 56 bits valid information and 100 bits total length, which is sufficient for general purpose. For the screen-capture and print-capture experiments, we use the HUAWEI Mate40 Pro cellphone to take photos. The display is SANC N50 Plus and the printer is Canon LBP9100C color.

C. Results

First we show our image quality improvement by both metrics and visual examples. Table I showcases the significant improvements we achieved in terms of visual quality-related metrics. PSNR and SSIM data of original StegaStamp are obtained from their paper. When the length of the embedded message was the same, our PSNR and SSIM were considerably

TABLE I
IMAGE QUALITY METRICS COMPARISON. FOR BOTH PSNR AND SSIM, HIGHER IS BETTER.

	PSNR↑		SSIM↑	
Message Length	StegaStamp	Optimized	StegaStamp	Optimized
50 bits	29.88	35.18	0.930	0.988
100 bits	28.50	35.42	0.905	0.987
150 bits	26.47	32.67	0.876	0.983

higher than those of the original method. Even when embedding messages with three times the length (150 bits compared with 50 bits), the model with our optimization performs still better than the original one. Some image quality comparison examples of watermarked images are shown in Figure 5. For the images containing faces and skin(the first two baby images and the Lenna example in Figure 2), our method significantly improves the quality; for images with relatively flat/simple backgrounds(the chandelier and the golf course), our optimized images show a slight improvement; for images with complex textures(the insect amd the animal), although there is also some improvement on the blurred background areas, the improvement in these areas is not easily observable by the naked eye, meaning the visual difference is not significant for these images.

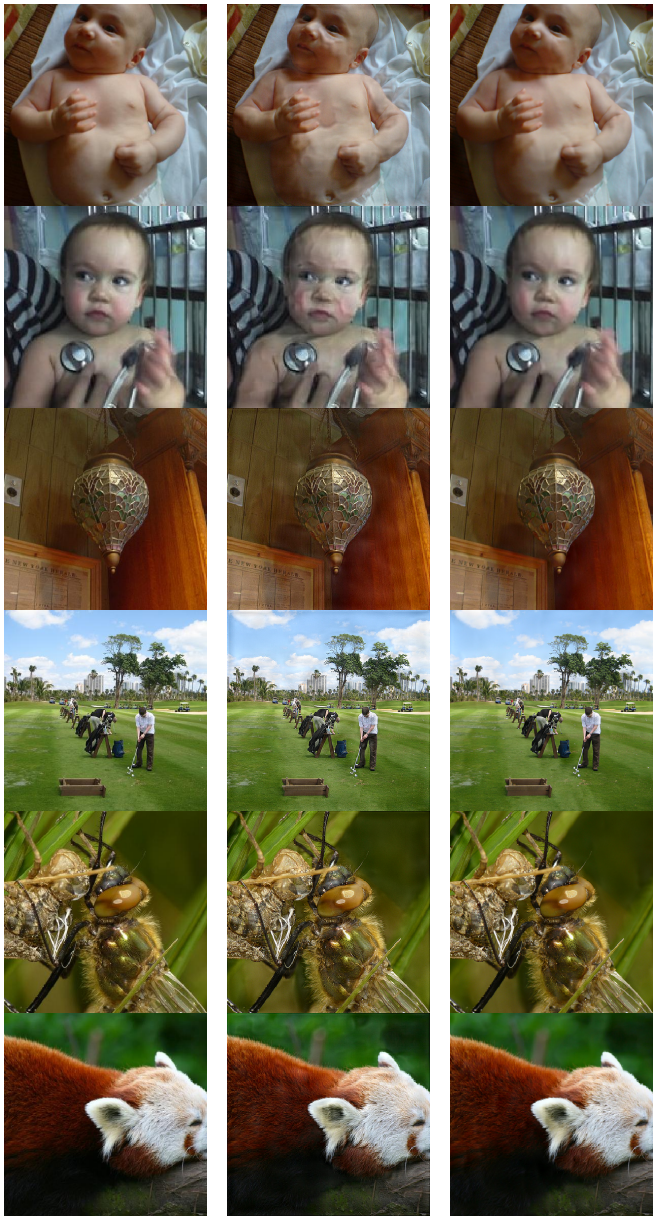
Improving visual quality alone is not sufficient for real-world applications, as we still need to keep sufficient robustness to resist various attacks. Therefore, we also conducted a series of experiments in various scenarios to prove the robustness of our method. Table II and III shows the robustness against screen-caption attacks, from varying horizontal angles and distances between the camera and the screen. Besides, we also show our strong robustness against JPEG compression attacks in Table IV.

TABLE II
COMPARISON OF BIT ACCURACY UNDER SCREEN-CAPTURE ATTACKS FROM DIFFERENT HORIZONTAL ANGLES.

		Bit Accuracy		Decoding success rate	
		StegaStamp	Optimized	StegaStamp	Optimized
left	10°	98.28%	98.74%	94%	97%
	20°	98.64%	98.82%	96%	97%
	30°	98.46%	98.64%	96%	97%
	40°	98.10%	98.62%	93%	95%
right	10°	98.02%	98.43%	93%	95%
	20°	98.62%	98.43%	95%	92%
	30°	98.24%	98.50%	93%	94%
	40°	98.02%	98.67%	90%	99%

TABLE III
COMPARISON OF BIT ACCURACY UNDER SCREEN-CAPTURE ATTACKS FROM DIFFERENT DISTANCES.

	Bit accuracy		Decode success rate	
	StegaStamp	Optimized	StegaStamp	Optimized
10 cm	94.83%	94.21%	63%	69%
20 cm	97.72%	97.93%	86%	91%
30 cm	98.68%	98.24%	98%	93%



(a) Original image (b) StegaStamp (c) Ours

Fig. 5. Watermarked Image Quality Comparison.

Inevitably, improving the visual quality of watermarked images will almost certainly lead to a decrease in robustness. Although it performs well or even better in most of the scenarios mentioned above, we found that there is a slight loss of robustness when suffering severe distorion and when the images are too simple to embed something at the same time. As shown in Table V, optimization seems to result in some decrease in decoding success rate under print-capture and print-scanner attacks, although the bit accuracy does not decrease significantly.

After careful study and comparison, we found that the images that are difficult to recognize after optimization are almost all images of simple objects placed on a white back-

TABLE IV
COMPARISON OF BIT ACCURACY UNDER JPEG COMPRESSION ATTACKS WITH DIFFERENT QUALITY.

JPEG Quality	Bit accuracy		Decode success rate	
	StegaStamp	Optimized	StegaStamp	Optimized
75	99.79%	99.92%	100%	100%
60	99.72%	99.84%	100%	100%
45	99.78%	99.55%	100%	100%
30	99.66%	98.72%	100%	95%

ground(some of them are shown in Figure 6), which are already very difficult to embed a 100-bit watermark into, and are common in the ImageNet dataset that's normally used for object detection and image classification. The original StegaStamp actually embed the watermark at the boundary of these images, and the shape was similar. Such a watermark is not only easily recognizable by the naked eye, but also meaningless. After excluding these images, our optimized method also achieve similar accuracy in print-related attacks. Since these images are rarely or cannot be used to embed watermarks in the real world, and our optimization method can still successfully decode these images under screen-capture and JPEG compression attacks, we believe that the differences in the decoding success rates shown in Table 6 is acceptable, and our optimization does improve the image quality in a way that is applicable to real-world scenarios.

TABLE V
COMPARISON OF BIT ACCURACY UNDER PRINT-CAPTURE AND PRINT-SCANNER ATTACKS.

	Bit accuracy		Decode success rate	
	StegaStamp	Optimized	StegaStamp	Optimized
print-capture	98.41%	96.73%	98%	85%
print-scanner	97.21%	95.62%	90%	80%

V. CONCLUSION

In this paper, we present a novel graph-based optimization for improving the quality of watermarked images of deep learning-based watermarking systems. The optimization is demonstrated through its deployment on the StegaStamp framework. The experimental results indicate that our method maintains similar robustness under various types of attack and noticeably improves the image quality as evidenced by both objective metrics and subjective observation.

REFERENCES

[1] W. M. Khedr and M. W. Abo Elsoud, "A novel blind and robust watermarking technique of multiple images," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 595–599. DOI: 10.1109/CSCI49370.2019.00112.

[2] S. Lagzian, M. Soryani, and M. Fathy, "A new robust watermarking scheme based on rdwt-svd," *International Journal of Intelligent Information Processing*, vol. 2, no. 1, pp. 22–29, 2011.



(a) Original image (b) StegaStamp (c) Ours

Fig. 6. Comparison of some undecodable images after our optimization under attacks shown in Table V.

[3] H. Fang, W. Zhang, H. Zhou, H. Cui, and N. Yu, "Screen-shooting resilient watermarking," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1403–1418, 2019. DOI: 10.1109/TIFS.2018.2878541.

[4] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2114–2123. DOI: 10.1109/CVPR42600.2020.00219.

[5] J. Pang and G. Cheung, "Graph laplacian regularization for image denoising: Analysis in the continuous domain," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1770–1785, 2017. DOI: 10.1109/TIP.2017.2651400.

[6] Y. Bai, G. Cheung, X. Liu, and W. Gao, "Graph-based blind image deblurring from a single photograph," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1404–1418, 2019. DOI: 10.1109/TIP.2018.2874290.

[7] C. Dinesh, G. Cheung, and I. V. Bajić, "Point cloud denoising via feature graph laplacian regularization," *IEEE Transactions on Image Processing*, vol. 29, pp. 4143–4158, 2020. DOI: 10.1109/TIP.2020.2969052.

[8] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, *Hidden: Hiding data with deep networks*, 2018. DOI: 10.48550/ARXIV.1807.09937. [Online]. Available: https://arxiv.org/abs/1807.09937.

[9] K. Hao, G. Feng, and X. Zhang, "Robust image watermarking based on generative adversarial network," *China Communications*, vol. 17, no. 11, pp. 131–140, 2020. DOI: 10.23919/JCC.2020.11.012.

[10] T. Huynh-The, C.-H. Hua, N. A. Tu, and D.-S. Kim, "Robust image watermarking framework powered by convolutional encoder-decoder network," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*, 2019, pp. 1–7. DOI: 10.1109/DICTA47822.2019.8945866.

[11] E. Wengrowski and K. Dana, "Light field messaging with deep photographic steganography," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1515–1524. DOI: 10.1109/CVPR.2019.00161.

[12] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph spectral image processing," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 907–930, 2018. DOI: 10.1109/JPROC.2018.2799702.

[13] J. Pang, G. Cheung, A. Ortega, and O. C. Au, "Optimal graph laplacian regularization for natural image denoising," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2294–2298. DOI: 10.1109/ICASSP.2015.7178380.

[14] W. Hu, X. Gao, G. Cheung, and Z. Guo, "Feature graph learning for 3d point cloud denoising," *IEEE Transactions on Signal Processing (TSP)*, Mar. 2020.

[15] J. Pang and G. Cheung, "Graph laplacian regularization for image denoising: Analysis in the continuous domain," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1770–1785, 2017. DOI: 10.1109/TIP.2017.2651400.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068.

[17] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, Vancouver, Canada: ACM, 2008.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A print-scan-resilient watermarking scheme for trademark images

Wenfa Qi* and Jiameng Wang† and Zichen Yuan‡ and Xiaolong Li§
* Wangxuan Institute of Computer Technology, Peking University, Beijing, China
E-mail: qiwenfa@pku.edu.cn
† Institute of Information Science, Beijing Jiaotong University, Beijing, China
E-mail: 22120328@bjtu.edu.cn
‡ Institute of Information Science, Beijing Jiaotong University, Beijing, China
E-mail: 22120347@bjtu.edu.cn
§ Institute of Information Science, Beijing Jiaotong University, Beijing, China
E-mail: lixl@bjtu.edu.cn

***Abstract*—Currently, print-scan-resilient watermarking algorithms remain a challenging research direction. In the industrial field, there is a huge demand for copyright protection of trademark images, which are particularly vulnerable to attacks such as printing and scanning for counterfeiting purposes. Therefore, this paper proposes a print-scan-resilient watermarking scheme based on the human eye's insensitivity to changes in image brightness, using circular masks and linear interpolation to modify the brightness component and achieve local gradients in colour images for embedding copyright information. Then the information is extracted by comparing the mean brightness values of the modified local region and its concentric outer ring. Because the watermark information is distributed across various local regions of the image and integrated with local brightness gradients, it is less susceptible to damage or forgery even after printing and scanning, ensuring the robustness of the algorithm. Experimental results demonstrate that this method shows excellent robustness when the embedding strength surpasses 5 and the modified area radius is well-suited.**

I. Introduction

Digital watermarking is commonly employed in various domains such as digital copyright management, information security, and privacy preservation [1][2][3][4]. It aims to embed specific identification information into digital media such as images, audio, and video without compromising quality. This allows for effective identification of media ownership and tracking of its usage while maintaining the integrity and quality of the media itself. In practical scenarios, digital watermarks often encounter various attacks and threats, including image processing attacks, cropping attacks, and print-scan attacks, with the latter being the most common and challenging attack method [5][6]. Adversaries print materials with digital watermarks and then scan or photograph them again, attempting to remove the digital watermark or forge a new watermark to deceive detection systems. This attack is not only easy to implement, and cost-effective, but also has the potential to spread digital content to uncontrolled environments, posing a significant threat to copyright protection [7][8][9]. Correspondingly, there is an enormous economic demand for utilizing digital watermarking technology that is resistant to print-scan attacks to safeguard the copyright of trademark images.

Most existing anti-print scan algorithms typically convert the original image into the frequency or wavelet domain for processing. For instance, K. Solanki et al. proposed an algorithm called 'Selective Embedding in Low Frequencies' (SELF), which can effectively resist print-scan attacks [10]. R. Riad and colleagues utilized the improved spread spectrum technique to embed the watermark within a middle frequency band of an image's discrete Fourier transform [11] . Ruanaidh et al. present a watermarking method based on the log polar map of DFT magnitudes [12]. The method suggested in [13] also utilizes DFT magnitudes, but the watermark is designed with circular symmetry to avoid the need for log-polar coordinate transformation. Tang and Huang inserted a set of peaks into selected DFT coefficients to achieve watermark embedding [14]. In addition, others also embed watermarks by modifying the polarity of discrete cosine transform (DCT) coefficients [15] or by efficiently integrating the three components, red, green and blue [16]. Some composite watermarking schemes have also been proposed, such as using the discrete wavelet transform-discrete cosine transform [17]. However, these algorithms involve complex processes, are limited by certain assumptions, and are better suited for carrier images with complex texture information, making them less applicable to copyright protection for trademark images in practical scenarios.

Considering the aforementioned circumstances, this paper presents a print-scan-resilient watermarking scheme designed specifically for trademark images. The scheme achieves watermark embedding by modifying brightness values in a dispersed manner within the spatial domain, creating a localized colour gradient in the image. Experiments demonstrate that the proposed method effectively withstands print-scan attacks while ensuring the concealment of the watermark.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Proposed scheme

This scheme utilizes the brightness gradient between monochromatic smooth regions in trademark images to embed and extract watermarks in local areas. Specifically, the watermark information is embedded in an image by adjusting the brightness values and utilizing error diffusion techniques for seamless integration into neighbouring pixels. During the extraction process, circular masks are employed to sample and compare local areas of the image, improving the scheme's resistance against attacks such as printing and scanning. In addition, robustness has been enhanced through the implementation of BCH error correction coding techniques.

A. Embedding Process

This method modifies the brightness component and achieves local gradients in colour images to embed watermarks. The process of modifying the brightness involves several steps, as illustrated in Fig.1. First, locations that meet the brightness modification condition are selected as centre coordinates. Next, the original RGB image is converted to the YCbCr format and the Y component is separated from the Cb and Cr channels. Then, based on the copyright information bits encoded by BCH, the brightness value gradually varies from the centre to the outside of the circular area determined by each centre, forming a transition effect from dark to light or from light to dark, i.e., a radiation-like brightness gradient. The modified Y component is ultimately combined with the Cb and Cr channels and transformed back into a colour image, resulting in the watermarked trademark image.

To illustrate the key steps of implementing the local gradient effect, i.e., the specific operations within the loop in Fig.1, more clearly, this paper takes a certain central coordinate as an example to describe how to modify the brightness component in the circular regions determined at that position to achieve the embedding operation. This process consists of two parts: mask operation and linear interpolation.

1) Mask operation: A grid of the same size as the original trademark image is constructed to enable pixel-level operations on the separated Y component. First, the Euclidean distance of each pixel point to the centre is calculated, and those with distances less than or equal to the radius of the modifiable region are marked as 1 while the rest are marked as 0. Here, the radius denoted as r_1 is an adjustable parameter used to control the size of the circular region that can be modified. The result of the above operation is a two-dimensional distance mask matrix, referred to as DK_{r_1}, which is shown in Fig.2. Subsequently, a binary matrix is obtained to mark the circular region where brightness pixel values can be modified. Specifically, for each nonzero element in the DK_{r_1} matrix, the minimum distance is subtracted and then divided by the distance difference, obtaining distance parameters within the range of [0, 1], which are stored in the distances map denoted as DM_{r_1}.

2) Linear interpolation: According to the BCH-encoded binary number corresponding to the circle's centre, a new pixel average value μ_2 is calculated based on the original pixel average value μ_1 and embedding strength dif within a modifiable circular region, as shown in Eq. (1). Subsequently, an evenly spaced sequence of length r_1 is constructed from μ_2 to μ_1, referred to as the gradient vector. Finally, the distance parameter DM_{r_1} is mapped into the range of the gradient vector gradient using linear interpolation, and the resulting values are assigned to the corresponding pixels inside the circular region to achieve a gradient effect.

$$\begin{cases} \mu_2 = \mu_1 + dif & binary\ number = 0 \\ \mu_2 = \mu_1 - dif & otherwise \end{cases} \tag{1}$$

Specifically, the principle of linear interpolation is to estimate the function value y_q at an unknown position x_q, given a set of n known data points (x_i, y_i), where $X = \{x_i\}$ represents the positions of the data points on a one-dimensional coordinate axis and $Y = \{y_i\}$ represents their corresponding function values. On a one-dimensional coordinate axis, the set of known data points $\{(x_i, y_i)\}$ can be viewed as a series of ordered line segments connecting each point. When estimating the function value y_q at position x_q, we first find the two nearest data points (x_k, y_k) and (x_{k+1}, y_{k+1}) to x_q in terms of their positions along the axis. Assuming $x_k < x_{k+1}$, then x_q lies between x_k and x_{k+1}. We can use these two data points to calculate the interpolated function value y_q for x_q using the linear interpolation formula:

$$y_q = y_k + (x_q - x_k)\frac{y_{k+1} - y_k}{x_{k+1} - x_k} \tag{2}$$

Here, the factor $\frac{y_{k+1}-y_k}{x_{k+1}-x_k}$ represents the slope of the line segment between (x_k, y_k) and (x_{k+1}, y_{k+1}), and $(x_q - x_k)$ represents the distance between x_q and x_k.

Once the principle of interpolation is known, the mapping process described above becomes clearer. Specifically, the gradient is represented as the function values corresponding to the end points of an interval, which indicates the range of the gradient effect. This effect involves a transition in pixel values from μ_2 to μ_1 within a circular region. Using DM_{r_1}, the distance between each pixel point in the circular region and the centre can be calculated to determine the modified pixel value based on distance. The resulting values are assigned to the corresponding pixel points within the circular region, causing the brightness values within the region to gradually change from the centre outwards, ultimately forming a radiation-like gradient effect.

B. Extraction Process

The flowchart for watermark extraction is shown in Fig. 3. Prior to this step, it may be necessary to correct for any perspective distortion in the obtained colour image with watermark using specialized image processing software. This technology is well-established and will not be the focus of this discussion. Corresponding to the embedding process, it is necessary to convert the original image and the watermarked image to the YCbCr colour space to obtain the Y component.

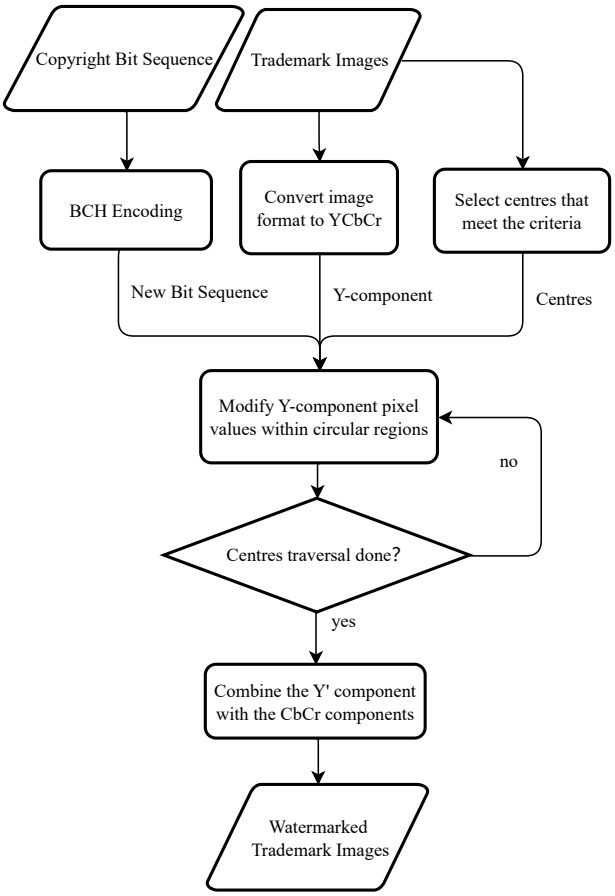


Fig. 1. Watermark embedding processes

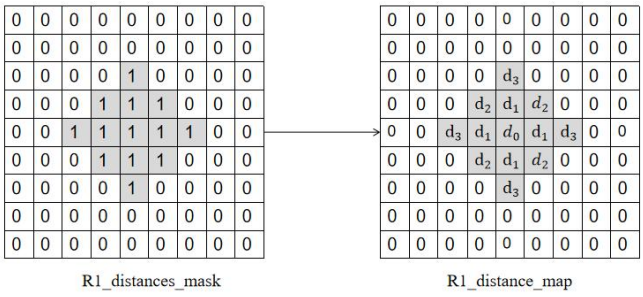


Fig. 2. The DK_{r_1} and DM_{r_1} of a circular region

Then, a mask is applied to extract the annular region for each known centre coordinate, and the average difference between the pixel values within the annular region and the average pixel value within the inner circular region with the same centre is calculated. After iterating through all predefined centre coordinates and performing the aforementioned operations, the extraction results from each local region are concatenated to form a binary code sequence. Subsequently, these sequences are used as input to the BCH decoder for error correction and decoding, finally yielding the corrected watermark information.

To provide a clearer description of the process for masking

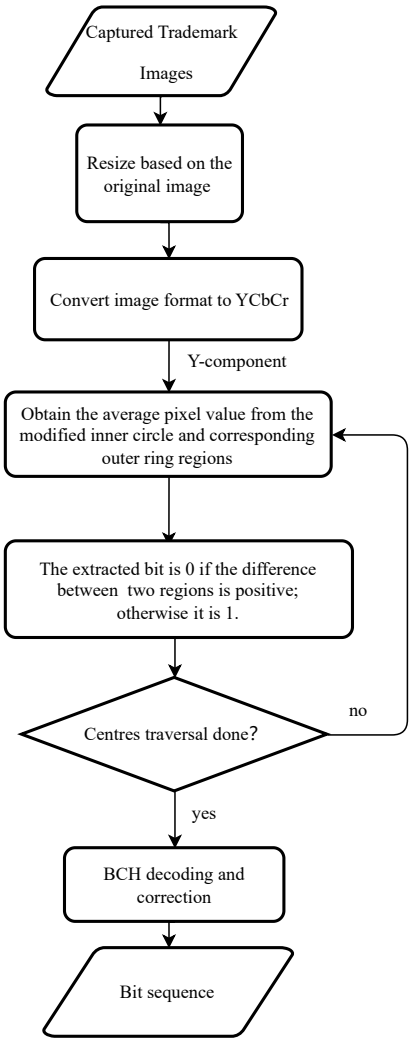


Fig. 3. Watermark extraction process

operation and calculate the difference, for the same center coordinate in the embedding method, the specific extracting process is as follows:

After constructing the grid, perform a masking operation on the Y component of the watermarked image. This mask matrix is referred to as DK_{r_2}, which extracts information from the grey annular region marked as 1 shown in Fig. 4. The limited rule for the area range states that the distance between a pixel and the centre of a circle should be greater than the modifiable circle radius r_1 and less than the user-defined radius r_2. r_2 is also manually specified, and it is required that the pixel values within the circular area formed by the centre and r_2 must be similar.

Next, the areas marked as 1 in theDK_{r_2} and DK_{r_1} matrices are calculated, i.e., the pixel mean of the blue area and grey area in Fig .4, and their difference is used to extract information. If the difference value is greater than 0, the extraction result is considered as 0, otherwise, it is considered as 1.

0	0	0	0	1	0	0	0	0
0	0	0	1	1	1	0	0	0
0	0	1	1	0	1	1	0	0
0	1	1	0	0	0	1	1	0
1	1	0	0	0	0	0	1	1
0	1	1	0	0	0	1	1	0
0	0	1	1	0	1	1	0	0
0	0	0	1	1	1	0	0	0
0	0	0	0	1	0	0	0	0

R2_distances_mask

Fig. 4. The DK_{r_2} of a circular area

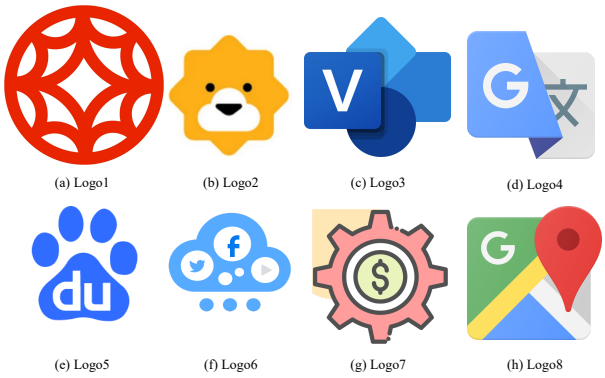


Fig. 5. Trademark images

III. Experimental Results

To demonstrate the effectiveness of the proposed method, as shown in Fig. 5 ,eight different trademark images are selected as experimental samples, and the optimal size of the modifiable circular area radius is determined for each one. The original watermark (15 bits, including 11 information bits) is then embedded into these eight images with embedding strengths of $dif = 5$, $dif = 10$, $dif = 15$, and $dif = 20$, respectively. Refer to Fig. 6 for the embedding effect of logo1. Fig. 7 displays the effect of logo1 after being printed and scanned, as well as the result after correction, with a partial enlargement of the image to better showcase the gradient effect. Extraction experiments are conducted for each embedding strength, and the bit correctness rate is used as the evaluation metric. The bit correctness rate (BCR) is defined as the ratio of the number of correctly extracted bits to the total number of extracted bits. This metric provides a measure of the accuracy of the watermark extraction process.

To increase the accuracy and credibility of the results, each combination is repeated at least 20 times during the experiments. Additionally, other factors that may affect the experimental results, such as image clarity and noise, are carefully controlled during the experiment.

According to the TABLE 1, when the embedding strength

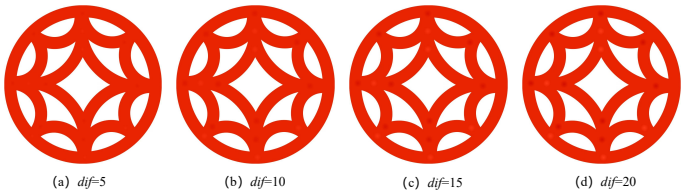


Fig. 6. Results of Logo1 with different embedding strengths.

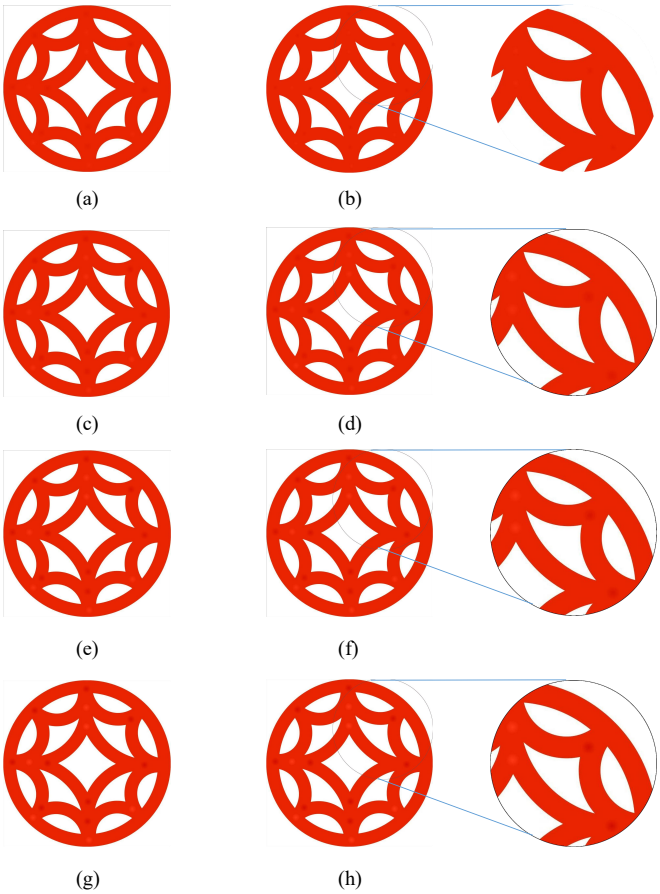


Fig. 7. (a), (c), (e), and (g) represent logos that have been printed and scanned with dif values of 5, 10, 15, and 20, respectively. (b), (d), (f), and (h) are the corrected logos corresponding to (a), (c), (e), and (g), respectively, and have been selectively enlarged.

is greater than 5 and the inner circle radius is appropriate, the probability of successful extraction is higher, which demonstrates excellent robustness.

IV. Conclusions

In this paper, we fully consider the graphic characteristics of trademark images and propose a new print-scan-resilient watermarking scheme that adopts a watermark embedding method based on the local circular region gradient while ensuring the robustness and invisibility of the watermark. This method can effectively avoid interference from the quality of printing ink and includes steps for correcting perspective distortion and using BCH encoding for error correction to enhance the water-

TABLE I
BIT CORRECTNESS RATE OF 8 TRADEMARK IMAGES

Trademark Images	Embedding Strength dif (%)			
	5	10	15	20
Logo1	98.7	99.7	100	100
Logo2	80.3	87.2	99.0	100
Logo3	100	100	100	100
Logo4	86.1	86.5	88.0	89.2
Logo5	100	100	100	100
Logo6	97.6	99	100	100
Logo7	100	100	100	100
Logo8	95.1	99.4	100	100

mark's resistance to attacks. The performance of this method is verified through experiments. However, there is still room for further improvement in the performance of this method. In future work, we will not be limited to simple trademark images and will continue to explore the application of this print-scan-resilient watermarking scheme in other scenarios.

REFERENCES

[1] K. Solanki, O. Dabeer, U. Madhow, B. Manjunath, and S. Chandrasekaran, "Robust image-adaptive data hiding: Modeling, source coding and channel coding," in *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, vol. 41, 2003, pp. 829–838.

[2] P. V. Sanivarapu, K. N. Rajesh, K. M. Hosny, and M. M. Fouda, "Digital watermarking system for copyright protection and authentication of images using cryptographic techniques," *Applied Sciences*, vol. 12, no. 17, p. 8724, 2022.

[3] A. Mohanarathinam, S. Kamalraj, G. Prasanna Venkatesan, R. V. Ravi, and C. Manikandababu, "Digital watermarking techniques for image security: A review," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 3221–3229, 2020.

[4] P. Kadian, S. M. Arora, and N. Arora, "Robust digital watermarking techniques for copyright protection of digital data: A survey," *Wireless Personal Communications*, vol. 118, pp. 3225–3249, 2021.

[5] X. Zhao, H. Lin, X. Xu, and Q. Zeng, "Resilient watermarking algorithm to printing and scanning combining local feature regions and DWT-SVD," in *Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence*, 2023, pp. 1191–1195.

[6] G. Yuan and Q. Hao, "Digital watermarking secure scheme for remote sensing image protection," *China communications*, vol. 17, no. 4, pp. 88–98, 2020.

[7] L. Yu, X. Niu, and S. Sun, "Print-and-scan model and the watermarking countermeasure," *Image and Vision Computing*, vol. 23, no. 9, pp. 807–814, 2005.

[8] A. H. Allaf and M. A. Kbir, "A review of digital watermarking applications for medical image exchange security," in *Innovations in Smart Cities Applications Edition 2: The Proceedings of the Third International Conference on Smart City Applications*, 2019, pp. 472–480.

[9] M. K. Harahap and N. Khairina, "Copyright protection of scientific works using digital watermarking by embedding doi QR code," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 3, no. 2, pp. 234–240, 2021.

[10] K. Solanki, U. Madhow, B. Manjunath, and S. Chandrasekaran, "Estimating and undoing rotation for print-scan resilient data hiding," in *2004 International Conference on Image Processing*, vol. 1, 2004, pp. 39–42.

[11] R. RIAD, D. Hassan, M. El Hajji, R. Harba, and F. Ros, "A print-scan resilient watermarking based on fourier transform and image restoration," *International Journal of Computer Applications*, vol. 128, pp. 13–17, 2015.

[12] J. J. Ò. Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal processing*, vol. 66, no. 3, pp. 303–317, 1998.

[13] V. Solachidis and L. Pitas, "Circularly symmetric watermark embedding in 2-d DFT domain," *IEEE Transactions on Image Processing*, vol. 10, no. 11, pp. 1741–1753, 2001.

[14] Y.-L. Tang and Y.-T. Huang, "Print-and-scan resilient watermarking for authenticating paper-based certificates," in *2010 First International Conference on Pervasive Computing, Signal Processing and Applications*, 2010, pp. 357–361.

[15] Y.-L. Tang and S.-Y. Tseng, "Print-and-scan resilient watermarking through polarizing DCT coefficients," in *2012 Computing, Communications and Applications Conference*, 2012, pp. 411–414.

[16] P.-Y. Chen and C.-L. Lin, "Print and scan resistant watermarking scheme for colour images," *The Imaging Science Journal*, vol. 63, no. 5, pp. 273–284, 2015.

[17] S. H. Amiri and M. Jamzad, "Robust watermarking against print and scan attack through efficient modeling algorithm," *Signal Processing: Image Communication*, vol. 29, no. 10, pp. 1181–1196, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network

Ziyi Xi, Wenmin Huang, Kangkang Wei, Weiqi Luo and Peijia Zheng
GuangDong Province Key Laboratory of Information Security Technology, Sun Yat-sen University, GuangZhou, China
E-mail: luoweiqi@mail.sysu.edu.cn xiziyi@mail2.sysu.edu.cn

***Abstract*—With the rapid evolution of AI Generated Content (AIGC), forged images produced through this technology are inherently more deceptive and require less human intervention compared to traditional Computer-generated Graphics (CG). However, owing to the disparities between CG and AIGC, conventional CG detection methods tend to be inadequate in identifying AIGC-produced images. To address this issue, our research concentrates on the text-to-image generation process in AIGC. Initially, we first assemble two text-to-image databases utilizing two distinct AI systems, DALL·E2 and DreamStudio. Aiming to holistically capture the inherent anomalies produced by AIGC, we develope a robust dual-stream network comprised of a residual stream and a content stream. The former employs the Spatial Rich Model (SRM) to meticulously extract various texture information from images, while the latter seeks to capture additional forged traces in low frequency, thereby extracting complementary information that the residual stream may overlook. To enhance the information exchange between these two streams, we incorporate a cross multi-head attention mechanism. Numerous comparative experiments are performed on both databases, and the results show that our detection method consistently outperforms traditional CG detection techniques across a range of image resolutions. Moreover, our method exhibits superior performance through a series of robustness tests and cross-database experiments. When applied to widely recognized traditional CG benchmarks such as SPL2018 and DsTok, our approach significantly exceeds the capabilities of other existing methods in the field of CG detection.**



(a) DALL·E2 (b) DreamStudio

Fig. 1. Generated images utilizing two different AI systems, DALL·E2 and DreamStudio, respectively. The prompt for the first row is:"The lake is like a mirror, reflecting the trees and mountains surrounding it." The prompt for the second row is: "The charming old-fashioned streetcars that trundled along the tracks added a touch of nostalgia to the scene."

I. Introduction

AI-generated content (AIGC), which pertains to the production or creation of content through artificial intelligence systems, has exerted a significant influence across various application scenarios, particularly with the emergence of large-scale models. The content produced by generative AI models comprises text, image, audio, video and cross-modal transformations like text-to-image, text-to-audio and more [1]. Among these, text-to-image has garnered enormous attention in social media platforms, primarily owing to its potential to supplant human visual design. In this field, several representative models have emerged, one such model is DALL·E, which is trained using Variational Autoencoders, another important model is VQGAN-CLIP based on Generative Adversarial Networks (GANs). Nowadays the diffusion model has become the core approach in text-to-image generation, with popular models like Stable Diffusion, Disco Diffusion, MidJourney, and DALL·E2 etc [2]. By typing the prompt containing expected image concepts, attributes and styles, users can cultivate an realistic or artistic image which is significantly correlative with the prompt. Some image examples are shown in Fig.1. Compared to conventional Computer-generated Graphics (CG) technology, more ease of operation, low capitalized cost and high-quality image feedback have propelled text-to-image generation into the forefront of internet. Therefore how to distinguish text-to-image graphics (denoted as T2I for short) from Photographs (PG) generated by a digital camera has emerged as a pressing issue that requires resolution.

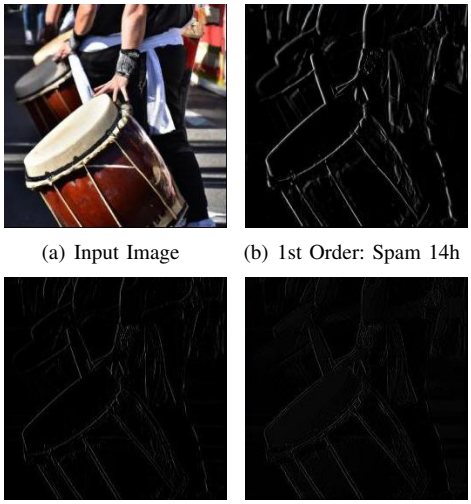
To our knowledge, only a limited number of works have focused on distinguishing between T2I and PG. Certain methods capable of distinguishing between these two image types capitalize on familiar classification techniques prevalent in computer vision, such as ResNet [3], [4] and EfficientNet [5]. However, these methods often fall short in providing a comprehensive analysis of inherent artifacts in T2I images and usually necessitate pre-training. As an alternative, we could harness contemporary methods established for distinguishing between CG and PG, given that these methods are adept at detecting variations in lighting, color, texture, and other

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

attributes present in PG. To date, several conventional methods have been proposed for CG and PG detection. For instance, Quan et al. [6] employed a 7×7 convolutional kernel in the initial layer to adaptively extract noise information. Yao et al. [7] designed three fixed high-pass filtering kernels to capture high-frequency information from images. Zhang et al. [8] developed a module comprising plain convolutional structures to extract correlations between adjacent pixels in RGB color space. He et al. [9] applied a Gaussian filter as a preprocessing step on the input image to achieve a robust network. Quan et al. [10] leveraged the SRM [11] to extract residual information from different channels. Other prevalent solutions include strategies based on transfer learning [12]–[15], the integration of Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN) [16], and those based on self-supervised learning [17], among others.

While the methods mentioned above prove effective for CG and PG detection, directly applying these techniques to T2I and PG detection might not yield optimal results. This limitation is due to the unique generative mechanisms present in T2I and CG. Inspired by successful approaches in CG detection, such as those presented in [9], [10], we propose a novel cross-attention enhanced dual-stream network specifically designed for T2I and PG detection. Our method incorporates two streams: a residual stream and a content stream. The residual stream employs a SRM [11] residual extraction module, which meticulously extracts texture information from the images. Conversely, the content stream focuses on the low-frequency aspects of the images, utilizing a dedicated network structure. The information extracted from both streams is downscaled using CNN modules, and then fused via a cross multi-head attention mechanism situated in the network's middle layer. After independent learning, the features from both streams are combined through channel concatenation, and a classifier is subsequently utilized to discern whether the image is T2I or PG. To ascertain the efficacy of our proposed method, we generate two T2I databases. Comprehensive comparative and ablation experiments have affirmed the superiority of our method. In summary, our contributions in this paper include:

- We propose a novel dual-stream framework that combines a residual stream and a content stream to comprehensively explore the generation traces of T2I. Moreover, we introduce a cross multi-head attention mechanism to enhance information exchange between the two streams.
- To assess the performance of our approach, we construct two T2I databases using two AI systems: DALL·E2 and DreamStudio. Each database consists of 20,000 generated images covering various scenarios.
- Our method excels in detection performance on our custom databases (DALL·E2 and DreamStudio) and outperforms comparable methods on established CG detection datasets, specifically DsTok and SPL2018. Moreover, we have conducted several ablation experiments to validate the effectiveness of our model design.



(a) Input Image (b) 1st Order: Spam 14h

(c) 2nd Order: Spam 12h (d) EDGE 3×3: Spam 14h

Fig. 2. An example image and its corresponding residual information obtained using three different high-pass filters.

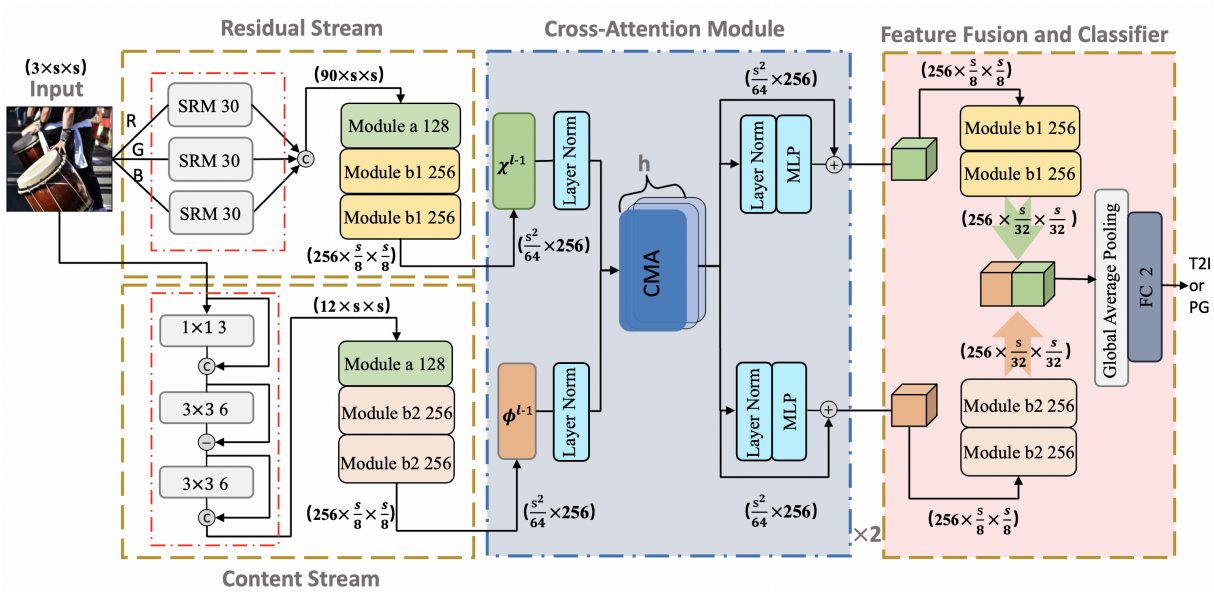
II. Proposed Method

The network architecture of the proposed model is visualized in Fig. 3. In the following subsections, we will provide a detailed description of the residual stream, content stream, cross-attention module, feature fusion and classifier utilized in the proposed model, respectively. Additionally, we will demonstrate the loss function employed in our model.

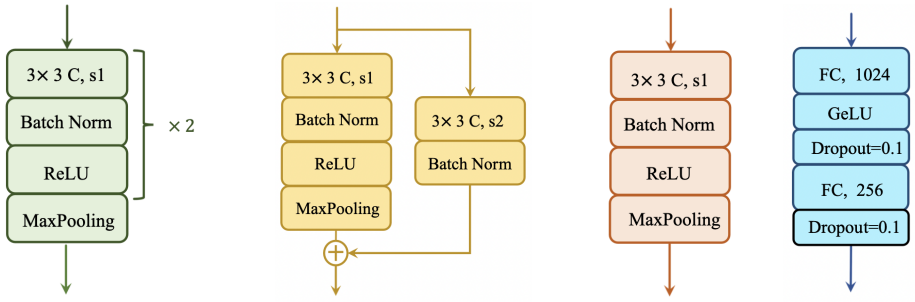
A. Residual Stream

Previous studies in CG detection [7], [10] and image steganalysis [18], [19] have shown that forgery artifacts are often present in the high-frequency components of images. To better detect these artifacts, several studies have employed multiple high-frequency filters from the SRM during image preprocessing to generate corresponding residual images. In our proposed model, we adopt a similar approach inspired by the methodologies presented in papers [10], [19]. As shown in the Fig. 3, the dimension of the input image is $3 \times s \times s$, where 3 means the color channels and $s \times s$ represents the resolution of input image. We begin by applying 30 SRM kernels individually to each of the RGB color channels in the input images. In this way, the input image is transformed into the residual domain for further feature analysis. Fig. 2 illustrates an example of the resulting residual images after filtering with three typical SRM kernels. This approach generates a total of 90 different residual images, aiming to capture a comprehensive representation of the multi-channel residual information. To facilitate further learning, we downscale 90 residual images using a custom-designed stacked CNN modules, which consist of one Module a and two Module b1. Module a and Module b1 are defined as follows:

- Module a: As shown in Fig. 3-(b), it comprises two consecutive groups of 3×3 convolutional layer, batch normalization (BN), ReLU, finally followed by a maximum pooling layer of size 3×3 with a stride of 2.



(a) The network architecture of the proposed model



(b) Module a (c) Module b1 (d) Module b2 (e) MLP

Fig. 3. The network architecture of the proposed model, and four modules used in the proposed model. The variable h corresponds to the number of headers. FC means fully connected layer. C represents the number of output channels, s1 means the stride of convolution kernel is 1.The symbol $\bigoplus$ denotes element-wise summation, and $\ominus$ means element-wise subtraction, ⓒ means channel-wise concatenation.

- Module b1: As shown in Fig. 3-(c), it includes two parts: convolutional downsampling and pooling downsampling. The pooling downsampling involves a sequence of 3×3 convolutional layer, BN, and a ReLU, followed by a maximum pooling layer of size 3×3 with a stride of 2. In parallel, the input feature maps are downscaled in another branch using a 3×3 convolutional layer with a stride of 2. The downscaled feature maps from both branches are fused together through element-wise addition, then activated by ReLU function.

After 90 residual feature maps through one Module a and two Module b1 in sequence, we can obtain $\chi^{l-1} \in R^{256 \times \frac{s}{8} \times \frac{s}{8}}$, as labeled in Fig. 3.

B. Content Stream

As mentioned earlier, the residual stream primarily focuses on image texture while disregarding the underlying content information. Unlike traditional steganography, where the modifications are usually minimal and do not disrupt the mid-to-low frequency information of the original image, images generated by T2I or CG often exhibit significant statistical differences



Fig. 4. The visualization of $\tilde{\ell}$ with 6 channels.

compared to images captured by a camera. Therefore, to address the limitations of relying solely on the residual stream, we propose a content stream that specifically targets forgery traces in the mid-to-low frequency range.

In the proposed content stream, we first utilize a 1×1 convolution to generate three composite images which integrate the R, G, and B channels. This step is performed to enhance

the breadth of channel information. Next, we concatenate the above three channels and the RGB image to obtain feature maps with 6 channels(denoted as ℓ). Then, pass the ℓ through a 3×3 convolution, and $conv_{3\times3}(\ell)$ is subtracted from ℓ to obtain the feature maps $\tilde{\ell}$, as Eq. (1). Subsequently, apply 3×3 convolution to $\tilde{\ell}$ for further learning, and then concatenate the $\tilde{\ell}$ and $Conv_{3\times3}(\tilde{\ell})$ along the channel dimension to obtain ℓ_{out} with 12 channels, as Eq. (2). Fig. 4 visualizes $\tilde{\ell}$, where it can be observed that four out of the six channels exhibit distinct content variations in the image, while the remaining two channels primarily capture edge information.

$$\tilde{\ell} = \ell - conv_{3\times3}(\ell) \tag{1}$$

$$\ell_{out} = Concat(\tilde{\ell}, conv_{3\times3}(\tilde{\ell})) \tag{2}$$

Following that, ℓ_{out} are downscaled through the stacked CNN modules which involve one Module a and two Module b2. Finally get the feature map $\phi^{l-1} \in R^{256\times\frac{s}{8}\times\frac{s}{8}}$. As shown in Fig. 3-(d), the Module b2 consists of 3×3 convolutional layer, BN layer, ReLU layer, and Maximum pooling layer of size 3×3 and a stride of 2.

C. Cross-Attention Module

To facilitate the interactive fusion of dual-stream, we propose a cross-attention module to exchange information between dual-stream. The details are as follows:

First, flattening the output of two streams χ^{l-1} and ϕ^{l-1} respectively, and exchange the last two dimensions to obtain $\chi^{l-1}, \phi^{l-1} \in R^{\frac{s^2}{64}\times256}$, followed by a layer normalization. Then, send them to the subsequent cross multi-head attention module (CMA), as shown in Fig. 5. Its principle and implementation details are as follows:

The ViT [20] is a Transformer-based model for image classification. It treats the input image as a sequence of patches and maps each patch to a single vector using linear mapping. Self-attention (SA) mechanism, the core of transformer, utilizes three variables: Q (Query), K (Key), and V (Value). In essence, it computes the attention weights between Query token and Key token, and multiplies them with the corresponding Value token associated with each Key. The formula for SA can be expressed as follows, d_k means the dimension of token:

$$SA = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Denotes the number of heads is h. We split χ^{l-1}, ϕ^{l-1} into h parts, attaining $\chi_i^{l-1}, \phi_i^{l-1} \in R^{\frac{s^2}{64}\times\frac{256}{h}}$, where $i \in \{1, 2, ..., h\}$. Different from SA, CMA calculates the degree of correlation between different vectors at each space position. In each head i, We can obtain the Query $Q_{\chi_i^{l-1}} \in R^{\frac{s^2}{64}\times\frac{256}{h}}$ and Key $K_{\phi_i^{l-1}} \in R^{\frac{s^2}{64}\times\frac{256}{h}}$ by linear projection respectively. Following that, ϕ_i' is obtained by multiplying the attention matrix with ϕ_i^{l-1}. d_k is $\frac{256}{h}$ here.

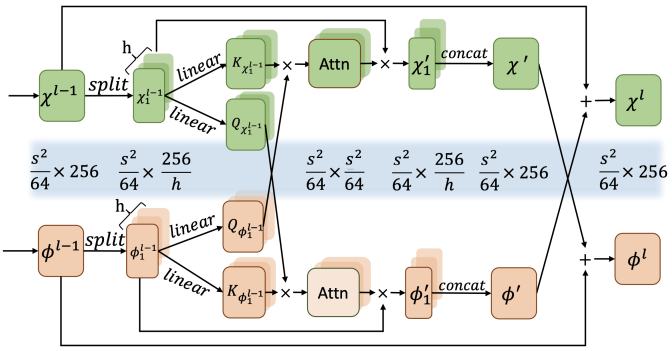


Fig. 5. The specific calculation process of the proposed CMA mechanism.

$$\phi_i' = softmax(\frac{Q_{\chi_i^{l-1}} \times K^T_{\phi_i^{l-1}}}{\sqrt{d_k}})\phi_i^{l-1} \tag{4}$$

Subsequently, the ϕ_i' obtained from each head are concatenated together to obtain the ϕ' relevant to χ^{l-1}, and add it to the χ^{l-1} to get the output χ^l. Similarly, we can obtain ϕ^l. After that, a layer normalization and an MLP module are applied, as Fig. 3. The cross-attention module utilizes transformer encoder structure in general as paper [20] and repeats it for twice. After through the cross-attention module, we can obtain the outputs $\in R^{\frac{s^2}{64}\times256}$.

$$\chi^l = \chi^{l-1} + Concat(\phi_1', \phi_2', ..., \phi_h') \tag{5}$$

D. Feature Fusion and Classifier

After through the cross-attention module, we need to convert the output from cross-attention module back to the feature maps with its original dimensions, swap the last two dimensions before reshaping and obtain the $\phi^l, \chi^l \in R^{C\times\frac{s}{8}\times\frac{s}{8}}$. The downsampling process further progresses as the feature maps are passed through a series of stacked CNN modules which involve two Module b1/b2. Afterwards, the feature maps are concatenated together by channel. Finally, classification is performed utilizing global average pooling and fully connected layer.

E. Loss Function

The loss function used in the proposed model is the cross entropy loss, which is defined as follows.

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log p\left(y_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right)\right] \tag{6}$$

Here, N stands for the number of samples in a batch. The symbol y_i denotes the true label of the i-th sample, where $y_i \in \{0, 1\}$. On the other hand, $p(y_i)$ indicates the predicted probability associated with the true label y_i.



(a) Flowers (b) Animals (c) Hill (d) Buildings (e) Clouds (f) Sculptures (g) Streets

Fig. 6. Some image samples from ALASKA, DALL·E2, and DreamStudio under seven different scenarios.

III. Experiments Result

In this section, we begin by providing a detailed explanation of the process for creating two T2I databases, along with the implementation details of our model. Subsequently, we conduct comparative experiments on T2I databases, comparing our approach with seven contemporary CG detection methods and ResNet18[4]. Additionally, we assess the robustness of our model against various post-processing operations and employ it for CG detection. Finally, we design ablation experiments to illustrate the validity of the proposed model.

A. Constructing Databases and Experimental Setup

Firstly, we randomly select 20,000 uncompressed PG images of size 256×256 from the ALASKA database[1], which is widely used in image steganalysis and forensics. To construct the T2I dataset, we start by extracting specific spots or tangible objects from the selected PG images, such as sports fields, lakes, the sun, indoor scenes, Gothic architecture, and more. We then utilize ChatGPT to expand these keywords into 5,000 prompts. For each prompt, we use the APIs of two modern AI systems, namely DALL·E2[2] and DreamStudio[3], to generate four different images. Finally, we create two databases, referred to as DALL·E2 and DreamStudio, each containing 20,000 PG images and their corresponding T2I images with similar scenes. Some image examples are shown in Fig. 6.

It is important to note that the original sizes of the DALL·E2 and DreamStudio images are 256×256 and 512×512, respectively. To ensure uniformity, we resize all PG and T2I images to three different resolutions: 256×256, 128×128, and 64×64. Following the resizing, we apply JPEG compression to the images using a randomly selected quality factor ranging from 75 to 95. Consequently, we have a total of six distinct databases generated by the two AI systems, each incorporating the three resolutions. For each database, we partition the 20,000 PG-T2I image pairs into three mutually exclusive sets: 12,000 pairs for training, 3,000 pairs for validation, and the remaining 5,000 pairs for testing purposes.

We employ an Adam optimizer with an initial learning rate of 2×10^{-4} and a batch size of 64. The learning rate is reduced by a factor of 0.1 every 30 epochs, and the total number of training epochs is set to 120. The number of heads (i.e. h) is set as 8. For training, we utilize 4 NVIDIA TITAN Xp GPUs. During evaluation, we consider the True Positive Rate (TPR) and the True Negative Rate (TNR) as our performance indicators. Once this manuscript is accepted, we will provide the source code of our model online [4], allowing readers to replicate our experimental results.

B. Comparative Experiments

To highlight the efficacy of our model, we conduct comparative studies against eight relevant techniques, including ResNet18 [4] and seven other contemporary methods for CG detection. These experiments are carried out on DALL·E2 and DreamStudio across three distinct resolutions. The results are presented in Table I. A review of these findings indicates that, in most instances, our proposed model outperforms the alternatives. T2I poses a significant challenge when it comes to visual discernment. Despite this, our model, along with several CG detection algorithms such as CGNet[15], has demonstrated the ability to detect T2I with considerable effectiveness, especially when the image resolution is high. For instance, at the resolution of 256×256, the accuracies of several methodologies exceed 98.5%. Remarkably, even when the resolution drops to 64×64, our model maintains an impressive performance, achieving 93.1% and 97.8% accuracy for the two T2I databases, respectively. It's worth noting that,

[1]https://alaska.utt.fr/
[2]https://openai.com/product/dall-e-2
[3]https://beta.dreamstudio.ai/generate
[4]https://github.com/zoie-ui/AI-Generated-Image-Detection

TABLE I

COMPARATIVE RESULTS FOR DIFFERENT METHODS ON DALL·E2 AND DREAMSTUDIO AT DIFFERENT RESOLUTIONS. THE HIGHLIGHTED VALUES DENOTED THE BEST RESULT IN THE CORRESPONDING CASE. THE HORIZONTAL LINE MEANS THAT THE METHOD HAS NO OUTPUT AT THIS RESOLUTION.

Methods	DALL·E2									DreamStudio								
	256×256			128×128			64×64			256×256			128×128			64×64		
	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC
ResNet18[4]	96.3	94.8	95.6	94.0	92.3	93.2	87.4	85.8	86.6	97.8	96.3	97.1	97.5	97.0	97.2	93.3	93.7	93.5
Quan[6]	97.6	97.4	97.5	96.8	95.9	96.3	–	–	–	98.2	98.6	98.4	98.5	97.3	97.9	–	–	–
Yao[7]	96.3	94.5	95.3	96.3	94.2	95.3	86.1	88.3	87.2	97.1	97.5	97.3	97.6	97.0	97.3	91.3	92.3	91.8
SPL2018[16]	98.3	98.1	98.2	97.6	96.8	97.2	92.2	90.3	91.3	98.9	99.0	99.0	99.3	98.8	99.1	96.5	95.3	95.9
He[9]	98.4	98.1	98.3	97.2	97.1	97.2	93.1	90.4	91.8	99.0	98.1	98.5	99.2	98.7	99.0	97.2	96.2	96.7
HcNet[8]	98.6	98.6	98.6	97.5	96.0	96.8	90.9	88.1	89.5	98.8	99.0	98.9	99.2	97.9	98.5	96.0	95.9	96.0
QuanNet[10]	98.6	98.4	98.5	98.5	97.2	97.9	**93.2**	87.9	90.6	98.3	98.4	98.4	99.2	98.9	99.1	97.6	97.0	97.3
CGNet[15]	98.4	98.1	98.3	97.9	**98.0**	98.0	92.5	**93.1**	92.8	99.4	99.2	99.3	99.3	98.9	99.1	**98.0**	97.6	**97.9**
Ours	**99.3**	**99.1**	**99.2**	**98.6**	97.9	**98.3**	93.1	**93.1**	**93.1**	**99.5**	**99.6**	**99.5**	**99.4**	**99.5**	**99.5**	97.7	**97.8**	97.8

TABLE II

DETECTION ACCURACY FOR DIFFERENT POST-PROCESSING ON THE DALL·E2 TEST DATASET AT A RESOLUTION OF 256 × 256. THE NUMBERS IN PARENTHESES REPRESENT THE RANKING AMONG ALL METHODS FOR THE SAME POST-PROCESSING.

DALL·E2	Chromaticity	Brightness	Contrast	Sharpness	Rotation	Gaussian Blur	Mean Blur	Average Acc
ResNet18[4]	94.0 (8)	91.3 (8)	82.3 (9)	94.8 (3)	53.4 (9)	86.6 (6)	71.5 (3)	82.0
Quan[6]	96.7 (6)	94.5 (6)	92.7 (4)	96.9 (1)	63.1 (6)	89.9 (3)	60.4 (8)	84.9
Yao[7]	92.1 (9)	86.9 (9)	86.9 (8)	77.9 (9)	68.8 (2)	59.2 (9)	60.4 (8)	76.0
SPL2018[16]	95.7 (7)	93.9 (7)	90.7 (6)	91.3 (7)	69.9 (1)	86.2 (7)	69.1 (5)	85.3
He[9]	97.3 (4)	94.8 (4)	90.3 (7)	93.6 (5)	58.1 (8)	89.4 (4)	71.5 (3)	84.9
HcNet[8]	97.5 (3)	94.8 (4)	91.3 (5)	93.9 (4)	64.1 (5)	92.7 (1)	85.1 (1)	**88.5**
QuanNet[10]	97.8 (1)	95.4 (1)	94.5 (2)	96.5 (2)	64.7 (4)	87.7 (5)	67.8 (6)	86.3
CGNet[15]	97.2 (5)	95.4 (1)	94.0 (3)	86.4 (8)	61.5 (7)	81.3 (8)	67.7 (7)	83.4
Ours	97.7 (2)	95.2 (3)	95.7 (1)	91.7 (6)	67.4 (3)	90.9 (2)	80.6 (2)	**88.5**

TABLE III

DETECTION ACCURACY FOR DIFFERENT POST-PROCESSING ON THE DREAMSTUDIO TEST DATASET AT A RESOLUTION OF 256 × 256. THE NUMBERS IN PARENTHESES REPRESENT THE RANKING AMONG ALL METHODS FOR THE SAME POST-PROCESSING.

DreamStudio	Chromaticity	Brightness	Contrast	Sharpness	Rotation	Gaussian Blur	Mean Blur	Average Acc
ResNet18[4]	95.9 (9)	91.7 (7)	85.3 (9)	89.5 (3)	56.2 (9)	67.1 (1)	66.7 (1)	78.9
Quan[6]	97.5 (5)	92.4 (6)	92.7 (6)	88.8 (4)	78.8 (2)	57.2 (8)	57.8 (8)	80.7
Yao[7]	96.3 (8)	94.9 (2)	93.2 (3)	83.5 (9)	72.9 (4)	50.2 (9)	50.5 (9)	77.4
SPL2018[16]	97.1 (7)	92.9 (5)	91.3 (7)	85.9 (7)	85.0 (1)	57.4 (7)	59.7 (5)	81.3
He[9]	97.3 (6)	89.4 (8)	90.2 (8)	83.9 (8)	63.1 (6)	60.3 (3)	61.8 (2)	78.0
HcNet[8]	98.2 (3)	94.7 (3)	92.8 (5)	87.8 (6)	62.4 (7)	57.9 (6)	58.3 (7)	78.9
QuanNet[10]	97.9 (4)	89.1 (9)	93.1 (4)	88.3 (5)	58.5 (8)	59.7 (4)	61.5 (3)	78.3
CGNet[15]	98.7 (2)	95.8 (1)	94.4 (1)	90.9 (2)	63.8 (5)	58.1 (5)	59.2 (6)	80.1
Ours	98.8 (1)	93.0 (4)	94.0 (2)	91.8 (1)	73.1 (3)	67.0 (2)	60.6 (4)	**82.6**

although ResNet18 was advocated for T2I detection in the research [4], our empirical results suggest that its detection performance falls short when compared to certain classic CG detection methods.

C. Robustness against Post-processing

In this section, we assess the robustness of our proposed model in the face of seven post-processing techniques, encompassing adjustments to chromaticity, brightness, contrast, sharpness, rotation, and the application of Gaussian blur and mean blur. Chromaticity, brightness, contrast, and sharpness modifications are facilitated by the ImageEnhance function from the Pillow library, while the remaining alterations are executed using the OpenCV library. To create a more realistic simulation of complex real-world scenarios, we've incorporated randomness into the parameters controlling the image alterations. For instance, the factors governing the degree of image manipulation (chromaticity, brightness, contrast) are randomly selected from a range of 0.5 to 2.5 for each image in the test dataset. Similarly, the factor controlling image sharpness is an arbitrary integer within the range of 0 to 4. Rotation degrees range from 0 to 360, and the kernel size for both Gaussian and mean filters is 5×5. In our experimental procedures, we utilize the models obtained from the previous subsection III-B to evaluate the post-processed test images directly. To keep the process simple, we restrict the detection accuracy display to image sizes of 256×256. The results of these experiments are illustrated in Table II and Table III for DALL·E2 and DreamStudio, respectively.

Observations from these tables reveal that our proposed model exhibits superior performance in terms of average accuracy when compared to current methods. Additionally, we find that among all post-processing operations, current methods demonstrate comparatively lower robustness against rotation

TABLE IV

COMPARATIVE RESULTS OF DIFFERENT METHODS ON DSTOK AND SPL2018 AT VARYING IMAGE RESOLUTIONS. THE HIGHLIGHTED VALUES INDICATE THE BEST RESULTS IN EACH CASE. THE HORIZONTAL LINE INDICATES THAT THE METHOD DID NOT PRODUCE ANY OUTPUT AT THAT RESOLUTION.

Methods	DsTok									SPL2018								
	224×224			112×112			56×56			224×224			112×112			56×56		
	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC
ResNet18[4]	77.1	73.9	75.5	65.4	71.4	68.4	62.2	57.3	59.7	85.3	85.3	85.3	65.8	70.6	68.2	68.6	58.4	63.5
Quan[6]	80.7	84.9	82.8	72.6	75.6	74.1	–	–	–	89.5	89.6	89.6	85.9	83.2	84.6	–	–	–
Yao[7]	86.5	91.6	89.1	87.8	84.2	86.0	79.6	74.5	77.1	90.1	87.4	88.7	89.4	84.1	86.8	88.0	79.9	83.9
SPL2018[16]	91.8	94.9	93.3	84.2	84.1	84.2	78.3	65.2	71.7	94.3	88.6	91.4	86.4	85.7	86.1	77.7	72.7	75.2
He[9]	91.8	91.6	91.8	82.4	79.9	81.1	76.6	73.6	75.1	92.1	87.1	89.6	75.4	82.6	79.0	75.4	82.6	79.0
HcNet[8]	92.6	93.7	93.1	79.9	83.7	81.8	74.9	74.2	74.5	94.4	90.3	92.3	77.4	86.2	81.8	71.2	73.9	72.5
QuanNet[10]	92.4	95.4	93.9	91.7	91.0	91.3	87.0	87.7	87.4	**94.7**	91.3	93.0	90.6	88.6	89.6	88.1	87.4	87.7
CGNet[15]	93.2	91.4	92.3	86.9	88.8	87.8	81.9	81.6	81.7	93.7	93.4	93.5	**92.5**	89.2	90.8	86.8	**88.4**	87.6
Ours	**98.1**	**96.9**	**97.5**	**95.3**	**93.6**	**94.4**	**92.8**	**90.4**	**91.6**	94.4	**94.2**	**94.3**	92.2	**90.1**	**91.2**	**89.2**	87.6	**88.4**

TABLE V

ABLATION STUDY ABOUT RESIDUAL STREAM, CONTENT STREAM AND CROSS-ATTENTION MODULE IN THE PROPOSED MODEL. A CHECKMARK DENOTES THE INCLUSION OF THE CORRESPONDING COMPONENT IN OUR MODEL, WHILE A CROSS INDICATES ITS REMOVAL. THE HIGHLIGHTED VALUES DENOTED THE BEST RESULT IN THE CORRESPONDING CASE.

Ablation settings			DALL·E2			DreamStudio			DsTok			SPL2018		
Residual stream	Content stream	Cross-Attention	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC	TPR	TNR	ACC
✓	✗	✗	99.1	99.0	99.0	98.9	**99.6**	99.3	93.2	**98.6**	95.9	92.2	93.3	92.8
✗	✓	✗	98.9	98.4	98.6	99.2	99.0	99.1	95.6	92.9	94.2	93.5	93.3	93.4
✓	✓	✗	**99.3**	98.8	99.0	99.4	99.4	99.4	95.9	96.9	96.4	92.3	**94.2**	93.2
✓	✓	✓	99.2	**99.2**	**99.2**	**99.5**	**99.6**	**99.5**	**98.1**	96.9	**97.5**	**94.4**	**94.2**	**94.3**

and blurring post-processing.

D. CG Detection

In this section, we assess the performance of our model on two widely used datasets in the field of CG detection, namely SPL2018 [16] and DsTok [21]. The SPL2018 dataset consists of a comprehensive collection of 13,600 images. It includes both CG images obtained from over 50 game rendering software and professionally captured photorealistic PG images using various camera models under diverse environmental conditions. To create training, validation, and testing sets, we partitioned the dataset following a ratio of 10:3:4, respectively. The DsTok dataset comprises a total of 9,700 images sourced from the internet. We partition them into training, validation, and testing sets using a ratio of 3:1:1. All test images are processed using center cropping with dimensions of 224×224, 112×112, and 56×56.

We conduct comparative experiments on both datasets at different resolutions, as presented in Table IV. The results demonstrate that our method consistently achieves the highest accuracy for both DsTok and SPL2018 at varying image resolutions. The improvement is particularly significant for DsTok. For example, our proposed method achieves accuracy improvements of 3.6%, 3.1%, and 4.2% compared to the sub-optimal method QuanNet for the three resolutions. In the case of SPL2018, our accuracy outperforms CGNet by 0.8%, 0.4%, and 0.8%, respectively.

E. Ablation study

The proposed model, as depicted in Fig. 3, incorporates two streams, namely the residual stream and the content stream, along with the cross-attention module to enhance information exchange between the two streams. In this section, we conduct an ablation study to investigate the impact of the residual stream, content stream, and cross-attention module in our proposed model for both the T2I and CG detection tasks. Among them, the residual stream and content stream are followed by the corresponding parts of feature fusion and classifier to achieve the task goal of binary classification. The image resolutions used in our experiments are 256×256 and 224×224 for the respective tasks. The results of the ablation study are presented in Table V. From the results presented in Table V, we observe that employing all three components (i.e., both streams and the cross-attention module) in the proposed model consistently achieves the highest detection accuracy across all cases. When any of the components is removed, the performance of the corresponding model experiences varying degrees of decline in most cases. These results strongly demonstrate the rationale behind the proposed model and reinforce its effectiveness for both the T2I and CG detection tasks.

IV. CONCLUSION

In this study, we have directed our attention towards the Test-to-Image (T2I) generation process inherent in AIGC, adopting an innovative, cross-attention enhanced dual-stream network for T2I and PG detection. In an endeavor to appraise the effectiveness of our proposed method, we generate two distinct T2I graphics databases, leveraging the capabilities of DALL·E2 and DreamStudio systems. Through exhaustive comparative and ablation experiments, our method affirms its superiority in detection performance, repeatedly surpassing traditional CG detection techniques across a diverse range of image resolutions. Importantly, our method demonstrate exceptional performance not just on our novel databases, but also excelled when applied to mainstream CG detection

datasets, specifically DsTok and SPL2018.

This research heralds a substantial advancement in the field of AI-generated image forgery detection. Our innovative, cross-attention enhanced dual-stream network offers a promising solution to the escalating challenges introduced by AIGC. The effectiveness of our approach lays a compelling groundwork for further exploration and enhancement in this domain, with the potential to evolve into more sophisticated and dependable techniques for image forgery detection amid the dynamic environment of AI-generated content. Looking towards the future, we plan to further fine-tune this dual-stream network approach and broaden our investigations to its applicability in detecting other types of AI-generated forgeries, thus bolstering our comprehensive response to the expanding realm of artificial intelligence in content creation.

REFERENCES

[1] Y. Cao, S. Li, Y. Liu, *et al.*, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," *arXiv preprint arXiv:2303.04226*, 2023.

[2] C. Zhang, C. Zhang, M. Zhang, and I. Kweon, "Text-to-image diffusion model in generative AI: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image diffusion models," *arXiv preprint arXiv:2210.06998*, 2022.

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.

[6] W. Quan, K. Wang, D. Yan, and X. Zhang, "Distinguishing between natural and computer-generated images using convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2772–2787, 2018.

[7] Y. Yao, W. Hu, W. Zhang, T. Wu, and Y. Shi, "Distinguishing computer-generated graphics from natural images based on sensor pattern noise and deep learning," *Sensors*, vol. 18, no. 4, pp. 1296–1296, 2018.

[8] R. Zhang, W. Quan, L. Fan, L. Hu, and D. Yan, "Distinguishing computer-generated images from natural images using channel and pixel correlation," *Journal of Computer Science and Technology*, vol. 35, pp. 592–602, 2020.

[9] P. He, H. Li, H. Wang, and R. Zhang, "Detection of computer graphics using attention-based dual-branch convolutional neural network from fused color components," *Sensors*, vol. 20, no. 17, pp. 4743–4743, 2020.

[10] W. Quan, K. Wang, D. Yan, X. Zhang, and D. Pellerin, "Learn with diversity and from harder samples: Improving the generalization of CNN-based detection of computer-generated images," *Forensic Science International: Digital Investigation*, vol. 35, pp. 301 023–301 023, 2020.

[11] M. Goljan, J. Fridrich, and R. Cogranne, "Rich model for steganalysis of color images," in *IEEE International Workshop on Information Forensics and Security*, 2014, pp. 185–190.

[12] E. De Rezende, G. Ruppert, and T. Carvalho, "Detecting computer generated images with deep convolutional neural networks," in *SIBGRAPI Conference on Graphics, Patterns and Images*, 2017, pp. 71–78.

[13] Q. Cui, S. McIntosh, and H. Sun, "Identifying materials of photographic images and photorealistic computer generated graphics based on deep CNNs," *Computers, Materials & Continua*, vol. 55, no. 2, pp. 229–241, 2018.

[14] H. Nguyen, T. Tieu, H. Nguyen-Son, V. Nozick, J. Yamagishi, and I. Echizen, "Modular convolutional neural network for discriminating between computer-generated images and photographic images," in *International Conference on Availability, Reliability and Security*, 2018, pp. 1–10.

[15] Y. Yao, Z. Zhang, X. Ni, Z. Shen, L. Chen, and D. Xu, "CGNet: Detecting computer-generated images based on transfer learning with attention module," *Signal Processing: Image Communication*, vol. 105, pp. 116 692–116 692, 2022.

[16] P. He, X. Jiang, T. Sun, and H. Li, "Computer graphics identification combining convolutional and recurrent neural networks," *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1369–1373, 2018.

[17] K. Wang, "Self-supervised learning for the distinction between computer-graphics images and natural images," *Applied Sciences*, vol. 13, no. 3, pp. 1887–1887, 2023.

[18] X. Deng, B. Chen, W. Luo, and D. Luo, "Fast and effective global covariance pooling network for image steganalysis," in *ACM workshop on information hiding and multimedia security*, 2019, pp. 230–234.

[19] K. Wei, W. Luo, S. Tan, and J. Huang, "Universal deep network for steganalysis of color image based on channel representation," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3022–3036, 2022.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] E. Tokuda, H. Pedrini, and A. Rocha, "Computer generated images vs. digital photographs: A synergetic feature and classifier combination approach," *Journal of Visual Communication and Image Representation*, vol. 24, no. 8, pp. 1276–1292, 2013.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

ResNet-Based Camera Model Identification with Adaptive Preprocessing Module and Weight Fusion of Global Information

Boru Chen* and Waleed Abdulla†
* University of Auckland, New Zealand
E-mail: bche455@aucklanduni.ac.nz Tel/Fax: +86-15813005335
† University of Auckland, New Zealand
E-mail: w.abdulla@auckland.ac.nz Tel/Fax: +64 93737599 Ext.88969

Abstract—Identifying camera models used in digital images is an essential component of image forensics, which plays a vital role in preventing image-related crimes. While deep learning techniques, especially Convolutional Neural Networks (CNNs), are currently the most commonly used approach for Camera Model Identification (CMI) due to their high efficiency and accuracy, the performance of the CNN-based method will be adversely affected by the image content due to its structure.

In this study, we introduce a novel approach to CMI, which utilizes the Residual Neural Network (ResNet), Adaptive Preprocessing Module (APM), and Weight Fusion of Global Information (WFGI) to overcome this issue. The results of the experiment indicate that our proposed method has achieved an accuracy exceeding 97.5% when tested on a large-scale dataset that consists of 45 different camera models based on the Vision and Dresden database. Furthermore, our approach exhibits robustness under different attack scenarios, thereby highlighting its potential for practical use in image forensics applications.

I. Introduction

Images have been an essential medium for conveying information in human society since their inception due to their efficient and vivid characteristics. Smart devices such as smartphones, tablets, digital cameras, and smart wearables generate vast images daily and distribute them worldwide, making images increasingly crucial as information carriers. However, technological advances have also led to security risks, as image editing software has made it easy to modify image information which brings up the issue of credibility [1]. False images disseminated widely through the internet may cause a series of social security problems, making it vital to confirm the source of images to prevent harm to society. Camera Model Identification (CMI) studies can provide ideas to solve this problem.

A. Existing Methods

Fig. 1 illustrates the primary pipeline of the digital image imaging process. Almost every step in the imaging process leaves corresponding features and traces, and many successful works have taken advantage of these features to solve the problem of CMI. Generally, existing approaches to CMI can be categorized into three categories: those based on differences in camera hardware, those based on differences in software, and those based on deep learning.

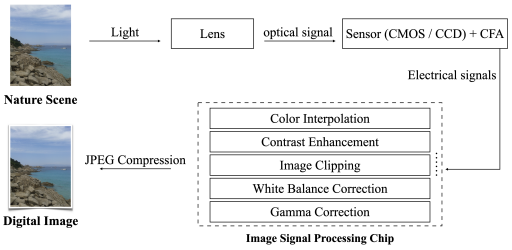


Fig. 1. The pipeline of image acquisition.

1) Hardware-based approaches involve utilizing differences in camera hardware, such as camera component attributes and differences in manufacturing processes. For instance, some researchers have used dust particles in front of the camera lens [2], lens distortion [3], and Photo Response Non-Uniformity (PRNU) [4] [5] [6] [7] to identify individual cameras.
2) Software-based approaches, on the other hand, determine the device through the differences in software, the unique features left by the imaging process of an image can be traced back to the corresponding camera model. Such as through the CFA interpolation [8], Automatic White Balance(AWB) [9] and compression statistics [10].
3) Deep learning-based approaches represented by convolutional neural networks (CNNs) has significantly improved the accuracy and scalability compared to traditional methods. With advanced neural networks such as GoogleNet [11], ResNet [12], and DenseNet [13], researchers have started applying these high-performance networks to CMI [14] [15] and furtherly demonstrating their superior performance compared to traditional CNNs.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

B. Research Contributions

Although CNN-based solutions have become mainstream in CMI due to their excellent performance, they extract deep-level information related to devices, not content information. However, due to structural characteristics, image content still affects CNN-based approaches' performance. Therefore, scholars have proposed techniques to suppress content interference, such as preprocessing with denoising filters or algorithms and patch selection strategies. However, both techniques have areas for improvement. Preprocessing images with denoising filters or algorithms can enhance noise information and improve identification performance, but it can also cause the loss of partial original image information and retain "fingerprints" [16] that may help in identification. The patch selection strategy increases training data, improving training efficiency and accuracy. Still, since the resolution of selected patches is smaller than the complete image, original image information is insufficient, making it difficult to obtain good robustness [17]. This paper proposes a new method that performs well in large-scale datasets and maintains robustness in different manipulation scenarios by combining both techniques.

This paper is organized as follows: Section 2 presents related work briefly. Section 3 describes the method based on ResNet, APM, and WFGI. Section 4 introduces experimental settings, datasets, and experimental results analysis. Additionally, the method's robustness is verified by considering different attack scenarios. Finally, the paper concludes with a summary and brief introduction to future work.

II. Related Work

A. Patch Selection Strategy

Patch selection strategy is regarded as an effective and straightforward method that can improve the accuracy of the CNN-based method in CMI, and it is often used with the majority voting rule. In [18], Bondi et al. first proposed using the patch selection strategy to solve the problem of CMI and achieved 93% identification accuracy in the data set containing 18 camera models.

By extracting patches from the image, the interference of CNN by the image's content can be avoided to a certain extent because the smaller resolution will force the CNN to extract the deep-level information instead of content information. However, this also brings another problem: how to confirm the number of patches and extract the patches. In [19], the authors proposed selecting patches using a fixed-size window to slide over the completed images. In contrast, the authors randomly selected patches from the original images in [20]. In [17], Yunxia et al. proposed to select representative patches based on multiple criteria after enhancing the training data and conducted experiments on the Dresden database [21], including 6 camera brands, 18 camera models, and 74 camera instances, achieving an average performance of 98.14%, 92.62%, and 41.54% respectively.

B. Model Fusion

Model fusion usually refers to combining the logits of two or more different models differently. Among many methods of model fusion, the representative ones are voting, averaging, bagging, and boosting.

In [22], Aniket et al. proposed a method based on AdaBoost and achieved 99.1% average identification accuracy in the experiment of 10,507 images collected by ten cameras in the Dresden database. In [23], Anselmo et al. proposed a data-driven ensemble approach consisting of two models and reported 92.51% identification accuracy in the data set consisting of 2740 JPEG images from 10 camera models.

You et al. in [24] proposed Multiscale Content-Independent Feature Fusion Network (MCIFFN). MCIFFN is composed of three parallel networks. The experimental results show that MCIFFN achieves 98.51% accuracy in the data set containing 23 camera models from the Dresden database.

C. Preprocessing image by filter or algorithm

To maximize the extraction ability of the convolutional neural network, many scholars try to use different techniques to preprocess the image before feeding the original images into CNN.

In [4], Lukas et al. mentioned that each camera's sensor has a unique pattern noise. They use a wavelet-based denoising filter to extract the pattern noise from the image and then use the noise to determine the different cameras. Amel et al. in [25] respectively applied a High pass filter [26] and wavelet-based denoising filter [27] to preprocess images and compared the influence of these two different filters on identification accuracy.

III. Proposed Method

Figure 2 shows the pipeline of our method, which includes two stages: the training and testing stages.

During the training stage, two steps are executed in order. Firstly, the APM divides the training set into two blocks: the first is the Median Filter Residual (MFR) obtained by applying a median filter on the original image, and the second consists of patches obtained using the Elastic Patch Selection (EPS) algorithm. Secondly, we use two independent base models to process the MFR and patches obtained in the first step separately. Finally, we obtain one model trained on MFR, which we call the MFR model, and one model trained on patches, which we call the patch model.

During the testing stage, three steps are executed in order. Firstly, the APM divides the test set into MFR and patches. Secondly, we use the MFR and patch models for processing the MFR and patches separately, resulting in two independent logits, as shown in the lower part of the

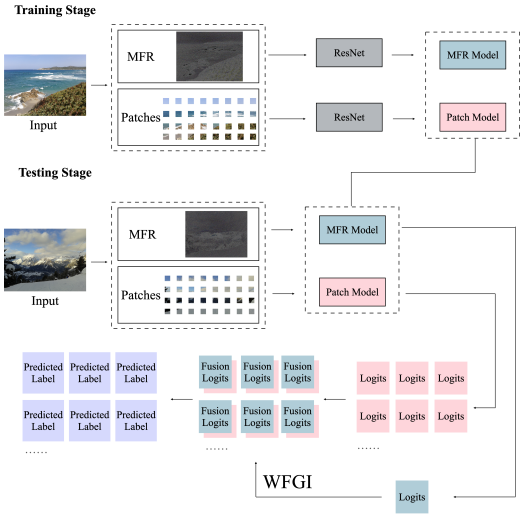


Fig. 2. The structure of our proposed method.

Fig. 2. Thirdly, the two independent logits are fused using the WFGI.

In this paper, we would use the ResNet with eighteen layers structure which was first proposed by Kaiming et al. in [12] as the base model due to its excellent performance under the deep structure situation compared with traditional neural networks.

A. Adaptive Processing Module

The first step of the proposed method in both the training and testing stages is to preprocess the dataset using APM, which consists of MFR extraction and patch extraction by EPS algorithm.

- MFR Extraction: The extraction process is as follows: assuming that the size of the median filter kernel is $k \times k$, and the original image is $O(i,j)$, then the output processed image by the median filter is $P(i,j)$. The MFR is

$$\begin{aligned} MFR &= MedianFilter_{k\times k}(O(i,j)) - O(i,j) \\ &= P(i,j) - O(i,j) \end{aligned} \tag{1}$$

  The median filter size will be determined by experiment in section 4.
- Patch Extraction: The patches are selected from the original images using the EPS algorithm, as shown in Figure3. The patch selection strategy has been demonstrated in many studies to be effective in improving accuracy. However, the methods of patch selection, size, and the number of patches significantly affect the model's performance.
  We propose an EPS algorithm to maximize the efficiency of model training and preserve the image's global information. EPS algorithm automatically calculates the step size of the horizontal and vertical movement of the crop window based on the patch selection number, crop window size, and original image size. This processing ensures that non-overlapping patches cover the whole image to the maximum extent possible. However, getting overlapping patches is unavoidable for images with small resolutions since the total area of patches may already exceed its pixel sum, as shown in Fig. 3.

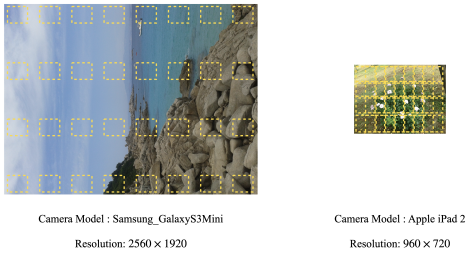


Fig. 3. The operation of EPS algorithm.

B. Weight Fusion of Global Information

The WFGI, the core of our proposed method, is a model fusion. Generally speaking, the conventional model fusion can only fuse different logits obtained by different models for processing the input with the same quantitative dimension, which is a one-to-one fusion approach. Our proposed WFGI is a one-to-many approach, which can fuse the logits obtained by the MFR model processing the MFR extracted from the image and multiple logits obtained by the patch model processing the patches of the image.

In Fig. 2, the MFR model outputs logits of one image while the patch model outputs N logits since we have divided a complete image into N patches by the EPS algorithm. Due to the number of logits output by the two models being different in dimension, it cannot be operated like the conventional model fusion. Therefore, we will weigh and fuse the logits output by the MFR model with the logits of each patch in a one-to-one way. Finally, each patch will get new fusion logits after WFGI. The process of WFGI is shown in the following formula:

$$L_F^{N_i} = L_M \times W_M + L_P^{N_i} \times W_P \tag{2}$$

where W_M and W_P represent the weights of MFR model and patch model respectively. L_M denotes the logits output by the MFR model and $L_P^{N_i}$ is the ith logits of the N patch of patches output by patch model. In this way, each patch can obtain complete image information through fusion with the output of the MFR model, which can effectively enhance its confidence. We will evaluate the identification performance of WFGI in experiment 2 in section 4. In this paper, we set both W_M and W_P to 0.5.

C. Performance Evaluation

In the early stages of CMI development, as with other classification tasks in the field of computer vision, the evaluation of accuracy was based on the following formula:

$$Accuracy = \frac{No.\ of\ correctly\ classified\ images}{No.\ of\ total\ images\ for\ testing} \times 100\% \quad (3)$$

However, the image-level accuracy can not well evaluate the classification ability of the model due to the majority voting operation has good fault tolerance to the prediction results of patches. Therefore, some researchers have proposed a more stringent evaluation protocol [28] [17], directly using the prediction results of patches to calculate the accuracy, which is called patch-level accuracy and the formula is as follows:

$$Accuracy = \frac{No.\ of\ correctly\ classified\ patches}{No.\ of\ total\ patches\ for\ testing} \times 100\% \quad (4)$$

In this paper, we would use patch-level accuracy to evaluate the performance of the proposed method unless otherwise specified.

IV. Experiments

For experiments in this research, we would use two databases: the Vision database [29] and the Dresden database.

For the Vision database, we will use all camera models inside the database, selecting the model instance with the most significant number of natural images for those models with more than one instance. We will consider Galaxy S III Mini GT-I8190N and Galaxy S III Mini GT-I8190 as the same camera model, as suggested in [20]. Therefore, the Vision database provides a total of 29 camera models. For the Dresden database, we will merge Nikon D70s and Nikon D70 into Nikon D70, as suggested in [17] [30], and randomly choose 16 camera models. Lastly, obtaining a large-scale data set containing 45 camera models. Detailed information is shown in Table I.

For our paper's experiment, we randomly chose 140 images from each camera model as the database, achieving a total of 6300 images. Thirty images from each camera model are used for the test set. At the same time, the remaining part is the development set (train set and validation set), and the ratio between the train set and validation set is set to 9:1. Unless otherwise specified, the following experiments were performed on the database constructed above.

All experiments were conducted on the PyTorch [31] and using the SGD optimization algorithm with initialization value as 0.001 and momentum as 0.9. The step size and gamma value of the learning rate decay are set to 50 and 0.1, respectively, while the maximum epoch setting is 100.

TABLE I
Camera models used in this experiment.

No.	Camera Model	Brand	Resolution
0	Galaxy S III Mini GT-I8190N	Samsung	2560 × 1920
1	iPhone 4S	Apple	3264 × 2448
2	P9 EVA-L09	Huawei	3968 × 2976
3	LG D290	LG	3264 × 2448
4	iPhone 5c	Apple	3264 × 2448
5	Lenovo P70-A	Lenovo	4784 × 2704
6	Galaxy Tab 3 GT-P5210	Samsung	2048 × 1536
7	iPhone 4	Apple	2592 × 1936
8	Galaxy S3 GT-I9300	Samsung	3264 × 2448
9	Xperia Z1 Compact D5503	Sony	5248 × 3936
10	iPad 2	Apple	960 × 720
11	iPhone 6	Apple	3264 × 2448
12	P9 Lite VNS-L31	Huawei	4160 × 3120
13	Lumia 640 LTE	Microsoft	3264 × 1840
14	iPhone 6 Plus	Apple	3264 × 2448
15	iPad mini	Apple	2592 × 1936
16	Ridge 4G	Wiko	3264 × 2448
17	Galaxy Trend Plus GT-S7580	Samsung	2560 × 1920
18	Zenfone 2 Laser	Asus	3264 × 1836
19	Redmi Note 3	Xiaomi	4608 × 2592
20	A3000	OnePlus	4640 × 3480
21	Galaxy S5 SM-G900F	Samsung	5312 × 2988
22	P8 GRA-L09	Huawei	4160 × 2336
23	iPhone 5	Apple	3264 × 2448
24	Honor 5C NEM-L51	Huawei	4160 × 3120
25	Galaxy S4 Mini GT-I9195	Samsung	3264 × 1836
26	A3003	OnePlus	4640 × 3480
27	Ascend G6-U10	Huawei	2448 × 3264
28	Galaxy Tab A SM-T555	Samsung	2592 × 1944
29	Canon Ixus70	Canon	3072 × 2304
30	Casio EX-Z150	Canon	3264 × 2448
31	FujiFilm FinePixJ50	Fujifilm	3264 × 2448
32	Kodak M1063	Kodak	3664 × 2748
33	Nikon CoolPixS710	Nikon	4352 × 3264
34	Nikon D200	Nikon	3872 × 2592
35	Nikon D70	Nikon	3008 × 2000
36	Olympus mju-1050SW	Olympus	3648 × 2736
37	Panasonic DMC-FZ50	Panasonic	3648 × 2736
38	Pentax OptioA40	Pentax	4000 × 3000
39	Ricoh GX100	Ricoh	3648 × 2736
40	Rollei RCP-7325XS	Rollei	3072 × 2304
41	Samsung NV15	Samsung	3648 × 2736
42	Sony DSC-H50	Sony	3456 × 2592
43	Sony DSC-T77	Sony	3648 × 2736
44	Sony DSC-W170	Sony	3648 × 2736

A. Experiment 1: Determined the parameter settings

In this experiment, we use ResNet to carry out experiments on the first 18 camera models in Table I to determine the parameter settings of the APM, including the size of the median filter kernel and the number of patches in the EPS algorithm introduced in section 3.

To determine the size of the median filter kernel, we firstly use median filters (implemented with the "cv2.medianblur" built-in function of the OpenCV library [32]) with different sizes ranging from 3 to 9 as shown in Table II to extract the MFR from the images, then fed MFR into ResNet and observe the impact of different sizes of median filters on the identification accuracy.

As the experiment result shows in Table II, using the median filter to preprocess the input image can effectively improve the identification accuracy. Taking everything into consideration, the size of the kernel will be set as 3.

Next, we would determine the number of patches selected by the EPS algorithm. We set the number of patches

TABLE II
Comparison of identification accuracy in different sizes of the median filter kernel.

Median Kernel Size	Original	3	5	7	9
Training Time (min)	11.81min	11.19min	12.73min	13.62min	13.71min
Accuracy (100%)	69.81%	92.96%	91.11%	90.74%	87.59%

ranging from 8 to 64 as shown in Table III.

The experiment results in Table III illustrate that the identification accuracy increases with the increase of the number of patches. However, it should be noted that the training time of the neural network will become longer at the same time, and the computational cost will also increase. Considering the balance between training efficiency and identification accuracy, the number of patches of our proposed EPS algorithm will be set to 32.

B. Experiment 2: Validation the performance of WFGI

This experiment aims to validate the performance of WFGI, which is the core of our proposed method. In this experiment, we would use the first 18 camera models from Table I as the experiment data set.

For ID1 in Table IV, we would only evaluate the performance of the patch model since we are now assessing the identification performance of the proposed method without WFGI. For ID2 in table IV, we would fuse the MFR model with the patch model by WFGI to obtain the identification accuracy after fusion.

The result in Table IV reveals that the fusion of the MFR model and patch model through WFGI can effectively improve identification performance with a slight increase in computational cost. Although ResNet can achieve 96.06% accuracy in 18 camera models, after WFGI, it can still achieve more than 2.5% improvement to 98.92% high accuracy, which demonstrates the potential of WFGI.

C. Experiment 3: Results on Vision database

In this experiment, we would use all the camera models from the Vision database, which is the first 29 camera models in Table I. We would compare our accuracy with Marra's method [20], which is currently the method with the most significant number of camera models and the highest identification accuracy on the Vision database to our best knowledge.

All the camera model we use in this experiment are precisely the same as [20], except we used the same numbers of test images while using fewer train and validation images, so this is a stricter setting technically. Therefore, it is reasonable to directly reuse the experimental results reported in [20]. In addition, [20] provides both patch-level and after majority voting image-level identification accuracy, so for comparison, we will also provide image-level accuracy in this experiment.

We report the identification accuracy of our proposed method in Table V. In [20], Marra et al. used DenseNet-40, DenseNet-121 and XceptionNet[33] with patch selection strategy to implement CMI respectively, where XceptionNet-based method obtained 95.15% patch-level accuracy and 99.31% image-level accuracy respectively. Compared with them, the experiment result demonstrates that our method achieved 98.41% patch-level accuracy and 99.54% image-level accuracy, respectively, which is highest in Table V.

D. Experiment 4: Results in large-scale data set

This experiment is performed on all camera models shown in Table I to evaluate the performance of our proposed method on a large-scale data set containing 45 camera models.

As the experiment result shows in Table VI, our proposed method shows good identification performance on a large scale of different camera models, achieving 97.54% accuracy across 45 camera models.

At the same time, we compare our proposed method with the situation using GoogleNet or AlexNet [34] as the base model. The result in Table VI reveals that whether using ResNet, GoogleNet, or AlexNet, WFGI can significantly improve the accuracy of identification, which demonstrates the generality of the WFGI.

E. Experiment 5: Result in manipulated data sets

In the actual application of CMI, the images we need to identify may have been variously processed or contaminated and do not have the original ideal quality.

Therefore, whether the proposed method has good robustness is very important for practical application. In this experiment, we selected 5 camera models from the Vision database. Each camera model provides 200 images, of which 80 are used as the test set, and the remaining 120 are divided into the train set and validation set in a ratio of 9 to 1, i.e. 108 and 12. The total number of the train set, validation set, and test set is 540, 60, and 400, respectively. Details of the data set is shown in Table VII. To test our proposed method's robustness, we propose four different simulation scenarios:

- JPEG Compression: JPEG compression by using the "cv2.imwirte" built-in function of the OpenCV library with two quality factors 70%, and 90%.
- Gaussian noise: Adding Gaussian noise by using the "imnoise" built-in function of the Matlab [35] with two variances settings, 0.1 and 0.01.
- Salt and Pepper noise: Adding salt and pepper noise by using the "imnoise" built-in function of the Matlab with two density settings: 0.1 and 0.01.
- Resize Operation: Resize the data set by using the "resize" built-in function of the Pytorch by two rescale factors, 0.9 and 1.1.

TABLE III
Comparison of identification accuracy in different numbers of patches selected by EPS

No. of Patches	8	16	32	64
Training Time (min)	29.50min	49.49min	93.75min	170.49min
Accuracy (100%)	93.91%	95.08%	96.06%	96.63%

TABLE IV
Comparison of identification accuracy with/without the WFGI.

ID	Method	Accuracy (100%)	Training Time (min)
1	ResNet	96.06%	93.75min
2	ResNet-WFGI	98.92%	104.94min

TABLE V
Comparison of identification accuracy in Vision data base including 29 camera models.

Method	Patch-Size	Pacth-Level Accuracy	Image-Level Accuracy
		%	
DenseNet-40-Based Method [20]	32 × 32 × 3	87.96	97.80
DenseNet-121-Based Method [20]	224 × 224 × 3	93.88	99.10
XceptionNet-Based Method [20]	299 × 299 × 3	95.15	99.31
Proposed Method	224 × 224 × 3	98.41	99.54

TABLE VI
Identification accuracy in the data set containing 45 camera models.

Method	Accuracy (100%)	Training Time (min)
AlexNet [34]	85.62	103.81
AlexNet-WFGI	94.28	114.26
GoogleNet [11]	95.07	292.49
GoogleNet-WFGI	97.44	310.33
ResNet	95.15	212.24
ResNet-WFGI	97.54	227.48

TABLE VII
Camera models used in testing robustness experiments.

ID	Camera Model	Brand	Resolution
1	P70A	Lenovo	1280 × 720
2	Xperia Z1 Compact	Sony	1920 × 1080
3	iPhone6 Plus	Apple	1920 × 1080
4	P8 GRA-L09	Huawei	1920 × 1080
5	Galaxy S4 Mini GT-19195	Samsung	1920 × 1080

Based on the above, we now have eight different data sets obtained from different manipulation of the original data set. In the training stage, the train set and validation set of both the original data set and manipulated data set are combined to generate a new data set called integrated data set. In this way, it increases the volume of the train set and validation set by 9 folds to 4860 and 540, respectively.

In the testing stage, the test set of the original data set and the test set of each manipulated test set will be counted for accuracy separately, as shown in Fig. 4.

In scenario one, we first test the performance of our proposed method under JPEG compression with different quality factors. In addition to the above-mentioned quality factors of 70% and 90%, we use the JPEG compression function with quality factors of 50%, and 80% to manipulate the test set of the original data set, as we did above, and then count and compare the accuracy.

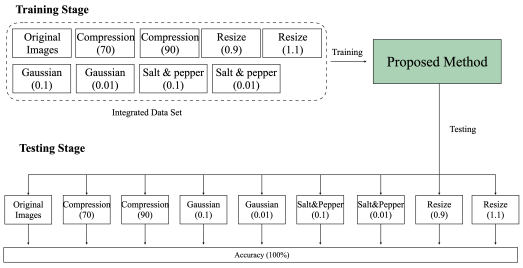


Fig. 4. The experiment setting of robustness validation.

TABLE VIII
Experiment results of our method under the scenario of JPEG compression.

Camera Models	Quality Factors				
	Original	50%	70%	80%	90%
	%				
P70A	99.88	99.10	98.95	99.22	99.69
Xperia Z1.	99.92	93.44	99.80	99.88	99.92
iPhone6 Plus	99.69	91.45	99.45	99.61	99.88
P8 GRA-L09	98.91	97.19	98.52	98.36	98.67
Galaxy S4.	99.77	96.33	98.67	99.06	99.41
Average	99.63	95.5	99.08	99.23	99.52

As seen in Table VIII, compared with original images, images after JPEG compression are slightly lower but still maintain a high identification accuracy. It is worth noting that the images with quality factors of 50% and 80% have not appeared in the train set and validation set. However, our proposed method can still achieve good accuracy, indicating that our method is generalized for JPEG compression and does not biased to the specific quality factor.

In scenario two, we evaluate the performance of the proposed method under the scenario contaminated by Gaussian noise. To further study the influence of Gaussian

noise on identification accuracy, we manipulate the test set with variances of 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, and 0.09 to validate whether our method can maintain good robustness under the noise of different variance which does not appear in the train set and validation set, and observe the trend that the identification accuracy with the gradual increase of noise variances.

As seen in Table IX, the accuracy of different variances shows certain dependence. The closer the variance of the test image is to 0.1 or 0.01, both of these settings appear in the train set and validation set, the higher the identification accuracy will be. Based on the analysis, we can conclude that the proposed method has a good generalization for Gaussian noise with different variances, even for the variance settings that do not appear in the train set and validation set.

In scenario three, we investigate the robustness of the proposed method under the scenario contaminated by adding salt and pepper noise.

As we did in scenario 2, we also manipulated the test set with densities of 0.02 to 0.09 to validate whether our method can maintain good robustness under the noise of different densities which does not appear in the train set and validation set. Table X presents the results. On the whole, our method can maintain good robustness. When the density is 0.03, it has the most significant influence on the accuracy, which is more than 7% lower than the original test set but still more than 90%.

In scenario four, we examine the performance of our proposed method in resisting resize operation. Table XI illustrates that our proposed method is robust to the resize operation of different re-scale factors. Although the factor of 0.7 has the greatest impact on the accuracy, it still maintains accuracy of 92.73%.

V. Conclusion

In this paper, we propose a camera model identification method based on ResNet, APM, and WFGI to improve the identification performance of the CNN-based method on CMI. Compared to other CNN-based methods, our proposed method fuses logits of different dimensions output from two models using WFGI. This approach saves global information using the patch selection strategy, allowing our method to maintain robustness while ensuring high accuracy. The experimental results show that the proposed method achieves more than 98% identification accuracy using a data set containing 29 camera models based on the Vision database.

In addition, we proposed four simulation scenarios to ensure that our method is robust under different adversarial conditions. These are JPEG compression, adding Gaussian noise, adding salt and pepper noise, and resizing operation. The experimental results show that our method can perform well in different scenarios. In future work, we aim to explore a particular fusion structure that can combine the patch selection strategy and global information of complete images without using model fusion.

References

[1] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, pp. 74–90, 2008.

[2] A. E. Dirik, H. T. Sencar, and N. Memon, "Digital single lens reflex camera identification from traces of sensor dust," IEEE Transactions on Information Forensics and Security, vol. 3, no. 3, pp. 539–552, 2008.

[3] K. Choi, E. Lam, and K. Wong, "Source camera identification using footprints from lens aberration," Proceedings of SPIE - The International Society for Optical Engineering, vol. 6069, pp. 60690J–60690J–8, 2006.

[4] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 205–214, 2006.

[5] C. Mo, J. Fridrich, M. Goljan, and J. Lukáš, "Source digital camcorder identification using sensor photo response nonuniformity," Proceedings of SPIE - The International Society for Optical Engineering, vol. 6505, pp. 65051G–65051G–12, 2007.

[6] A. J. Cooper, "Improved photo response non-uniformity (prnu) based source camera identification," Forensic Science International, vol. 226, no. 1–3, pp. 132–141, 2013.

[7] D. Valsesia, G. Coluccia, T. Bianchi, and E. Magli, "Image retrieval based on compressed camera sensor fingerprints," in IEEE International Conference on Multimedia and Expo, 2015.

[8] S. Bayram, H. Sencar, N. Memon, and I. Avcibas, "Source camera identification based on cfa interpolation," in Image Processing, 2005. ICIP 2005. IEEE International Conference on, 2005.

[9] Z. Deng, A. Gijsenij, and J. Zhang, "Source camera identification using auto-white balance approximation," in IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, 2011.

[10] S. C. Kai, E. Y. Lam, and K. K. Y. Wong, "Source camera identification by jpeg compression statistics for image forensics," in Tencon IEEE Region 10 Conference, 2007.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, "Going deeper with convolutions," IEEE Computer Society, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269, 2017.

[14] Y. Chen, H. Yue, and X. Ding, "Camera model identification with residual neural network," in 2017 IEEE International Conference on Image Processing (ICIP), 2017.

[15] U. Kamal, A. M. Rafi, R. Hoque, S. Das, and M. K. Hasan, "Application of densenet in camera model identification and post-processing detection," 2018.

[16] X. Ding, Y. Chen, Z. Tang, and Y. Huang, "Camera identification based on domain knowledge-driven deep multi-task learning," IEEE Access, vol. PP, pp. 1–1, 2019.

[17] Y. Liu, Z. Zou, Y. Yang, N.-F. B. Law, and A. A. Bharath, "Efficient source camera identification with diversity-enhanced patch selection and deep residual prediction," Sensors, vol. 21, no. 14, 2021.

[18] Bondi, Luca, Gueera, David, Bestagini, Paolo, Delp, Edward, J., and B. and, "First steps toward camera model identification with convolutional neural networks," IEEE Signal Processing Letters, 2017.

[19] H. Na, J. He, N. Zhu, X. Xuan, G. Liu, and C. Chang, "Identification of the source camera of images based on convolutional neural network," Digital Investigation, vol. 26, pp. S1742287618302664–, 2018.

TABLE IX
Experiment results of our method under the scenario of adding Gaussian noise with different variance.

Camera Models	Variances										
	Original	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
	%										
P70A	99.88	99.22	98.98	98.87	98.48	98.75	98.55	98.36	98.44	98.44	98.01
Xperia Z1.	99.92	99.30	99.34	98.83	98.55	97.93	97.81	97.89	98.01	97.70	98.20
iPhone6 Plus	99.69	95.16	87.30	81.91	78.63	81.52	82.89	87.30	90.20	92.46	92.58
P8 GRA-L09	98.91	94.69	91.64	91.05	92.70	93.75	94.80	94.84	95.98	96.37	96.29
Galaxy S4.	99.77	98.16	97.30	96.72	96.45	97.11	97.70	98.20	98.63	98.71	98.87
Average	99.63	97.30	94.91	93.48	92.96	93.81	94.35	95.32	96.25	96.73	96.79

TABLE X
Experiment results of our method under the scenario of adding salt and pepper noise with different density.

Camera Models	Density										
	Original	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
	%										
P70A	99.88	98.59	87.15	86.09	89.53	95.63	97.77	97.77	97.54	97.81	97.62
Xperia Z1.	99.92	99.77	98.95	98.71	98.83	99.10	99.22	99.45	99.49	98.98	99.14
iPhone6 Plus	99.69	98.09	97.73	90.63	94.73	97.38	95.31	95.74	94.77	95.23	94.92
P8 GRA-L09	98.91	98.16	81.91	87.19	91.99	92.85	93.13	94.10	94.26	95.12	95.43
Galaxy S4.	99.77	99.14	99.10	98.87	97.66	97.93	96.88	97.03	96.80	96.76	96.21
Average	99.63	98.75	92.97	92.30	94.55	96.58	96.46	96.82	96.57	96.78	96.66

TABLE XI
Experiment results of our method under the scenario of resize with different re-scale factors.

Camera Models	Rescale Factors				
	Original	0.7	0.9	1.1	1.3
	%				
P70A	99.88	92.70	99.45	99.92	99.84
Xperia Z1.	99.92	96.21	99.77	99.96	99.96
iPhone6 Plus	99.69	87.38	99.77	99.65	99.34
P8 GRA-L09	98.91	87.34	99.02	99.49	99.38
Galaxy S4.	99.77	100	99.88	99.65	98.28
Average	99.63	92.73	99.58	99.73	99.36

[20] F. Marra, D. Gragnaniello, and L. Verdoliva, "On the vulnerability of deep learning to adversarial attacks for camera model identification," Signal Processing Image Communication, p. S0923596518303217, 2018.

[21] T. Gloe and R. Bhme, "The dresden image database for benchmarking digital image forensics," ACM, 2010.

[22] A. Roy, R. S. Chakraborty, U. Sameer, and R. Naskar, "Camera source identification using discrete cosine transform residue features and ensemble classifier," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[23] A. Ferreira, H. Chen, B. Li, and J. Huang, "An inception-based data-driven ensemble approach to camera model identification," in 2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018, pp. 1–7, IEEE, 2018.

[24] C. You, H. Zheng, Z. Guo, T. Wang, and X. Wu, "Multiscale content-independent feature fusion network for source camera identification," Applied Sciences, vol. 11, no. 15, 2021.

[25] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," 2016 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6, 2016.

[26] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," in Media Watermarking, Security, and Forensics 2015 (A. M. Alattar, N. D. Memon, and C. D. Heitzenrater, eds.), vol. 9409, pp. 171 – 180, International Society for Optics and Photonics, SPIE, 2015.

[27] J. Fridrich, "Digital image forensic using sensor noise," IEEE Signal Processing Magazine, 2009.

[28] Z. Zou, Y. Liu, W. Zhang, and Y. Chen, "Camera model identification based on residual extraction module and squeezenet," in Proceedings of the 2nd International Conference on Big Data Technologies, ICBDT2019, (New York, NY, USA), p. 211–215, Association for Computing Machinery, 2019.

[29] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "Vision: a video and image dataset for source identification," Eurasip Journal on Information Security, vol. 2017, no. 1, p. 15, 2017.

[30] A. M. Rafi, RemNet: Remnant Convolutional Neural Network for Camera Model Identification and Image Manipulation Detection. PhD thesis, 2021. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-05-11.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library. Red Hook, NY, USA: Curran Associates Inc., 2019.

[32] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807, 2017.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol. 60, p. 84–90, may 2017.

[35] MATLAB, version 7.10.0 (R2022a). Natick, Massachusetts: The MathWorks Inc., 2022.

Structural Quality Assured Global Optimization for CTU-Level Rate Control of Screen Content Coding

Tong Tang[1,2,3,*], Yuan Tan[1,2,3], Shihang Ding[1,2,3], Zhidu Li[1,2,3]
[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications
[2] Advanced Network and Intelligent Interconnection Technology Key Laboratory of Chongqing Education Commission, China
[3] Chongqing Key Laboratory of Ubiquitous Sensing and Networking
E-mail: tangtong@cqupt.cn

***Abstract*—Screen content video (SCV) has been applied to various scenes, e.g. online meeting, remote desktop, cloud game, AR/VR and so on, generating massive screen content data and bringing huge pressure on transmission bandwidth. Rate control is an effective technique for improving bandwidth utilization, but existing rate control methods of SCV has not exploited characteristics of human visual system (HVS) fully, leaving space to improve coding efficiency. Therefore, this paper studies the characteristics of HVS and proposed a subjective rate control method for screen content coding (SCC). Concretely, first, a novel coding tree unit (CTU) level subjective rate-distortion (R-D) model is built by utilizing structural feature. Then, on this basis, a global optimization scheme of CTU-level bit allocation is derived. Finally, experimental results show that compared to the default rate control method in HM-16.20+SCM-8.8, the proposed method could averagely save 2.64% bit cost under similar subjective quality.**

I. Introduction

Screen content video has been applied to various scenes, e.g. online meeting, remote desktop, cloud game, AR/VR and so on, generating massive screen content data and bringing huge pressure on transmission bandwidth. To relieve the pressure of efficient transmission and storage of screen content, HEVC (high efficiency video coding)-SCC was formulated to improve compression efficiency of screen content coding (SCC) [1], plenty of new techniques (intra block copy, palette etc.) were designed by considering the features of screen content. At the same time, rate control (RC) as a crucial technique in enhancing the efficiency of compression, with its primary goal being to reduce distortion as much as possible within the constraints of bitrate limitations. General RC methods (e.g. HM-SCM) use objective quality metrics to measure image distortion, e.g. PSNR (Peak Signal to Noise Ratio), MSE (Mean Squared Error), ignoring the subjective perceptual characteristics of human visual system (HVS), leaving visual redundancy.

Hence, in recent years, there has been a tendency among researchers to utilize the perceptual attributes of the HVS to optimize RC and numerous Subjective Bit Allocation methods at the Coding Tree Unit (CTU) level have been introduced. Li et al. [2] found a linear relationship between SSIM and MSE by testing and proposed an CTU-level DSSIM -DMSE model to guide SSIM-based bit allocation. Zhou et al. modeled the pixel-level JND factor as distortion weights to optimize distortion at the frame level and CTU level [3]. Xiang et al. [4] proposed a fresh CTU-level perceptual distortion model using the JND factor to represent masking effects and then optimized the λ for each CTU. However, the above subjective bit allocation methods are designed for natural content video. Contrasting natural content video, screen content video possesses distinct features, for example, screen content image is usually mixed with graphics and text, containing plenty of sharp edges. Therefore, directly applying the rate control method of natural content video to screen content video will degrade encoding efficiency seriously.

Nowadays, researchers proposed many CTV-level rate control methods for screen content video. Yang et al. adopted three independent parameter updating methods for three different types of CTUs according to their respective rate-distortion (R-D) relationships [5]. Sanchez et al. [6] utilized the rate distortion information from coded CTUs to predict the current coding unit, due to the presence of numerous recurring words and patterns in the same screen content image. Chen et al [7] estimated the complexity of the remaining CTUs by a specially designed pre-analysis to allocate bitrate to each CTU. Esmaeeli et al. [8] proposed a content-based HEVC-SCC R-D model for different modes and sizes of CTUs. In conclusion, these RC techniques leveraged the unique attributes of screen content, thereby enhancing the efficiency of the coding process.

To the best of our understanding, there presently isn't a CTU-level subjective RC method specifically for screen content coding. This paper investigates the attributes of the HVS and puts forth a subjective rate control method designed for screen content coding. Experimental findings demonstrate that the suggested method could improve visual quality of screen content coding under similar bit consumption compared with the state-of-the-art method.

The remainder of this paper is organized as follows. Section II introduces the proposed subjective RC method. Section III describes the experimental design and results. Lastly, Section IV provides the conclusion for this paper.

II. Proposed Subjective Rate Control Method

A. Subjective R-D Modeling

The efficiency of bit allocation depends on the correctness of its R-D model. To choose a proper distortion model for screen content coding, A number of models for image quality (PSNR, SSIM, GMSD [9], GFM [10], ESIM [11], GSIM [12] and VSI [13]) were tested for HEVC-SCC distortion. The model

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

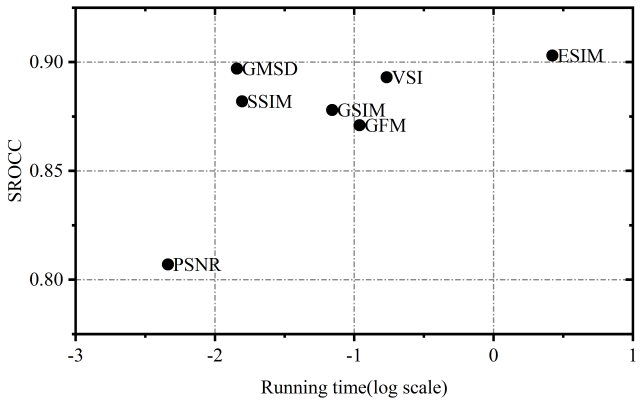


Fig. 1. Accuracy and complexity of the quality evaluation model.

accuracy and complexity of all models are shown in Fig.1. The accuracy of the quality models is assessed using the Spearman Rank-Order Correlation Coefficient (SROCC). Running time is used to evaluate computational complexity.

As evident from Fig. 1, GMSD is characterized by its precision and low computational complexity. In other words, GMSD achieves good compromise between model accuracy and computational complexity. Therefore, GMSD is applied to assess the subjective quality in this paper.

As mentioned above, screen content image is usually mixed with graphics and text, containing plenty of sharp edges. Thus structural information is very sensitive for HVS. To assure structural quality, gradient information is used to build the R-D model. The paradigm of the proposed R-D model is inspired by [14], to verify the performance of our model, we conducted an experiment, where a screen content sequence (*SlideEditing*) was randomly selected and encoded using the reference software HM-SCM, configured with low delay encoding mode and the encoded QP is 22. The bit cost and GMSD of each CTU are extracted and the R-D relationship are shown in Fig. 2, the fitting accuracy is over 96.04%. Thus the R-D model is built as

$$\frac{RG}{\mathrm{GMSD}} = \kappa R + \eta, \tag{1}$$

where κ and η represented factors associated with the screen contentt, R is the actual coded bitrate, G refers to the gradient magnitude and GMSD denotes the distortion.

B. Bit Allocation Optimization

From the R-D model expressed in (1), the value of λ can be deduced as follows:

$$\lambda = -\frac{\partial D}{\partial R} = -\frac{\eta G}{(\kappa R + \eta)^2}. \tag{2}$$

R can be written as

$$R = \frac{\sqrt{-\eta G/\lambda} - \eta}{\kappa}. \tag{3}$$

The distortion D can be calculated as follows:

$$D = \frac{G(\sqrt{-\eta G/\lambda} - \eta)}{\kappa\sqrt{-\eta G/\lambda}}. \tag{4}$$

Given target bitrate, the Lagrange multiplier optimized target is formulated as

$$\begin{aligned} \{\lambda_1^*, \lambda_2^*, \ldots, \lambda_M^*\} &= \arg\min \textstyle\sum_{i=1}^{M} D(\lambda_i) \\ &\text{s.t. } \textstyle\sum_{i=1}^{M} R(\lambda_i) \le R_c, \end{aligned} \tag{5}$$

where λ_i^* denotes the optimal value of the ith CTU, M represents the count of CTU in a single fram, while R_c denotes the target bits of one frame.

According to Karush-Kuhn-Tucker (KKT) condition, (5) could be transformed into unconstrained optimization problems as

$$\begin{aligned} L = \sum_{i=1}^{M}\left(\frac{G_i}{\kappa_i} - \frac{\eta_i G_i\sqrt{-\lambda_i/\eta_i G_i}}{\kappa_i}\right) \\ + \mu\left(\sum_{i=1}^{M}\frac{\sqrt{-\eta_i G_i/\lambda_i} - \eta_i}{\kappa_i} - R_c\right), \end{aligned} \tag{6}$$

where the variables satisfy

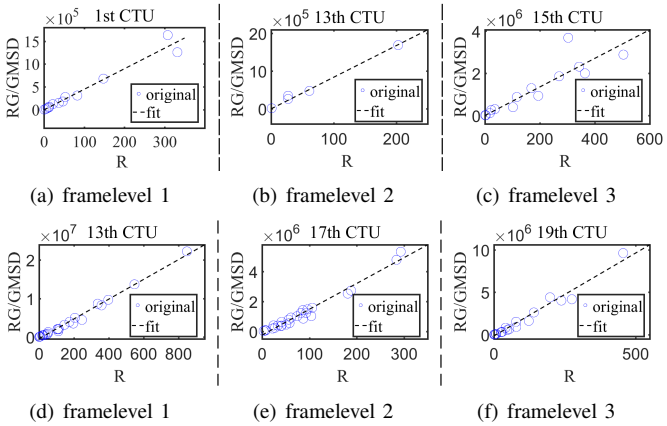
$$\begin{cases} \kappa_i > 0, i = 1, \ldots, M \\ \eta_i < 0, i = 1, \ldots, M \\ \lambda_i > 0, i = 1, \ldots, M \\ \dfrac{\partial L}{\partial \lambda_i} = -\dfrac{\eta_i G_i}{\kappa_i}\cdot\sqrt{-\dfrac{1}{\eta_i G_i}}\cdot\frac{1}{2}\lambda_i^{-\frac{1}{2}} - \mu\cdot\dfrac{\sqrt{-\eta_i G_i}}{\kappa_i}\cdot\frac{1}{2}\lambda_i^{-\frac{3}{2}} \\ \mu\left(\sum_{i=1}^{M}\dfrac{\sqrt{-\eta_i G_i/\lambda_i} - \eta_i}{\kappa_i} - R_c\right) = 0. \end{cases} \tag{7}$$

The optimized Lagrange multiplier can be derived as

$$\lambda_i^* = \left(\frac{\sum_{i=1}^{M}(\sqrt{-G_i\eta_i}/\kappa_i)}{R_c + \sum_{i=1}^{M}(\eta_i/\kappa_i)}\right)^2. \tag{8}$$

The optimized bit allocation at CTU-level can be written as

$$R_i = \frac{\sqrt{-\eta_i G_i/\lambda_i^*} - \eta_i}{\kappa_i}, \tag{9}$$



(a) framelevel 1 (b) framelevel 2 (c) framelevel 3

(d) framelevel 1 (e) framelevel 2 (f) framelevel 3

Fig. 2. The fitted R-D model for (a)-(c):*SlideEditing* and (d)-(f):*Desktop*.

where R_i denotes the bits allocated to the ith CTU, κ_i and η_i are model parameters associated with screen content, G_i is the gradient magnitude of the ith CTU.

C. Model Parameter Update

Parameter Initialization: In the proposed R-D model, κ and η are two parameters pertaining to the screen content. Parameters of the first CTU at every frame level are initialized according to the average model parameter fitting results. In our experiment, κ and η are set to 7410 and -1000, respectively.

Parameter update: Model parameters will be updated after encoding one CTU, the updated parameters will be used for the next co-located CTU at the same frame level. Referring to method in [15], κ_{new} and η_{new} are updated as in (10) and (11)

$$\kappa_{new} = \frac{\lambda_{old} R_{real} G_{old} + D_{real} G_{old}}{D_{real}^2}, \tag{10}$$

$$\eta_{new} = -\frac{\lambda_{old} R_{real}^2 G_{old}}{D_{real}^2}, \tag{11}$$

where λ_{old} and G_{old} are the parameter of the encoded CTU, R_{real} and D_{real} are the actual consumption bitrate and actual distortion of the encoded CTU, respectively.

D. Summary

The method we proposed could be outlined as follows:

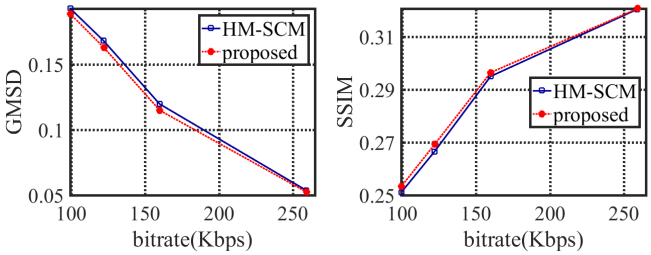
1) Initialize R-D model parameters;
2) Calculate gradient magnitude of the current CTU;
3) Allocate bits to the current CTU according to (8) and (9);
4) Finish encoding the current CTU;
5) Update model parameters according to (10) and (11);
6) Return to 3) until finishing encoding of the current frame.

III. Experimental Result

In this section, we initially present the design of the experiment, followed by a comparison of the R-D performance, and finally, we present the subjective quality comparison.

A. Experimental Design

In order to confirm the effectiveness of our method, we implement it in HM-SCM which refers to HM-16.20+SCM-8.8 [16] in this paper and the default RC method in HM-SCM is compared with our method. Test screen content sequences are randomly selected from the Common Test Conditions proposal [17], containing TGM (Text and graphics with motion) type sequences (*ChineseDocumentEditing*, *CircuitLayoutPresentation*, *ClearTypeSpreadSheet*, *Desktop*, *EnglishDocumentEditing*, *SlideEditing*, *WordEditing*), M (mixed content) type sequences (*BasketballScreen*, *KristenAndSaraScreen*, *video_conferening_doc_sharing*) and A (animation) type sequences (*BigBuck*, *ChinaSpeed*). These sequences are all encoded with low delay B (LDB) configuration. The sequences are encoded utilizing four distinct QPs (22, 27, 32, 37) to generate the target bitrate of RC methods. The testing platform is a computer with hardware device Intel (R) Core (TM) i3-10100 @ 3.60 GHz.



(a) in terms of GMSD (b) in terms of SSIM

Fig. 3. R-D curve comparison for *ClearTypeSpreadsheet*.

TABLE I
Rate Distortion Performance Compared with HM-SCM

Type	Sequence	BD-rate(%)		
		GMSD	PSNR	SSIM
TGM	*ChineseDocumentEditing*	-3.53	-2.03	-3.00
	CircuitLayoutPresentation	-2.47	-0.99	-2.36
	ClearTypeSpreadsheet	-2.57	-1.84	-2.02
	Desktop	-4.40	-4.45	-5.94
	EnglishDocumentEditing	-4.55	-1.01	0.74
	SlideEditing	-1.33	-2.18	-6.36
	WordEditing	-3.69	-4.29	-3.40
M	*BasketballScreen*	-3.02	-7.03	-4.24
	KristenAndSaraScreen	-0.35	-2.79	-1.18
	video_conferencing_doc_sharing	-4.85	-8.46	-2.68
A	*BigBuck*	-2.85	-5.45	-3.09
	ChinaSpeed	-1.02	-2.42	-2.06
	Average	**-2.64**	**-3.58**	**-2.97**

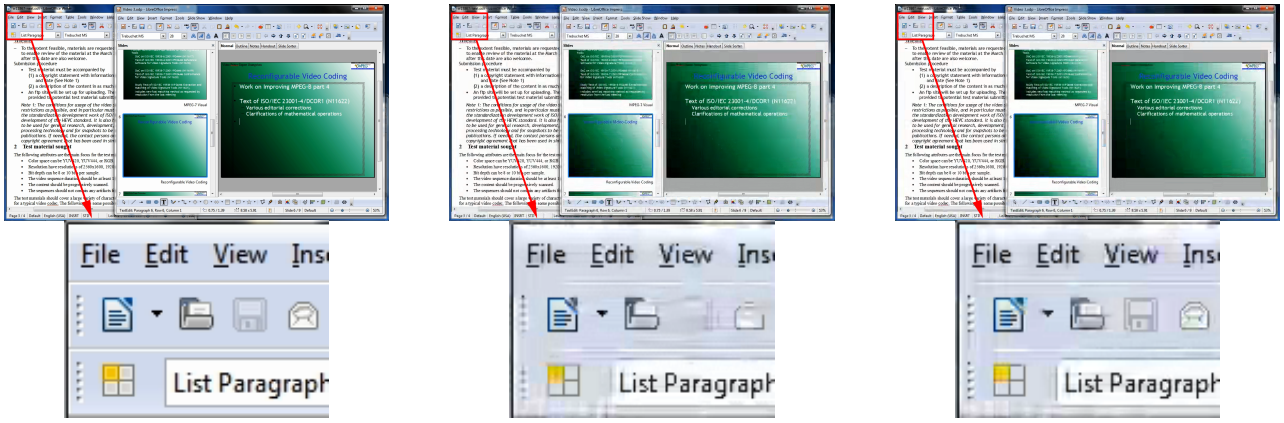
B. R-D Performance Comparison

The encoding efficiency is evaluated using the BD-rate (Bjøntegaard delta bit rate), which represents the bitrate saving under the same quality. Table I displays the BD-rate results (using three quality metrics) of the method we proposed in comparison to HM-SCM. As evident from Table I, firstly, compared with HM-SCM, the proposed method could averagely achieve BD-rates of 2.64%, 3.58% and 2.97% in terms of GMSD (the smaller GMSD is, the higher the quality is), PSNR and SSIM respectively. Secondly, the method we proposed could realize a performance gain for all the sequences. Thirdly, the method proposed could attain performance benefits for all three type sequences (TGM, M, A).

Moreover, Fig. 3 provides a comparison of the R-D curves for *ClearTypeSpreadsheet* using GMSD, PSNR, and SSIM as evaluation metrics. It can be observed that the method proposed delivers the highest quality at all the bitrate consumption.

C. Subjective Quality Comparison

Fig. 4 shows a comparison of subjective quality for *SlideEditing*, contrasting the proposed method with HM-SCM. We can see that under similar bit consumption, the proposed method could preserve more structural information than HM-SCM for



(a) Original

(b) HM-SCM (bit: 232, GMSD: 0.078, PSNR: 29.065, SSIM: 0.314)

(c) Propose (bit: 224, GMSD: 0.071, PSNR: 29.637, SSIM: 0.317)

Fig. 4. Subjective quality comparison for 186th frame of *SlideEditing* encoded with LDB configuration and QP 37.

the graphics content, thereby has better visual quality than HM-SCM. The key factor is that the proposed method allocate more bits for CTUs with more sharp edges.

D. Encoding Complexity

$$\Delta T_{all} = \frac{T_P - T_C}{T_C} \times 100\%, \tag{12}$$

$$\Delta T_G = \frac{T_G}{T_C} \times 100\%, \tag{13}$$

Table II presents the increase in encoding complexity when comparing the proposed method with HM_SCM, where ΔT_{all} represents the total encoding complexity increase and is defined as in (12), ΔT_G represents the encoding complexity increase caused by gradient calculation and is defined as in (13), T_P represents the the encoding time of the proposed method, T_C denotes the encoding time of HM_SCM, T_G is the time cost of gradient calculation. We can see that firstly, the proposed method averagely increases encoding time by 9.06% compared with HM_SCM. Then, the encoding complexity increase is mainly caused by gradient calculation. And the computational time of gradient calculation could be further reduced through code optimization (such as parallelism). Finally, for sequences of TGM type, ΔT_G is larger than ΔT_{all}, because the proposed bit allocation scheme allocates more bits for key frames, which could reduce quality refinement and encoding time [18].

TABLE II
ENCODING COMPLEXITY INCREASE

Type	ΔT_{all}	ΔT_G
TGM	11.23%	11.36%
M	11.06%	10.89%
A	4.89%	4.31%
Average	9.06%	8.85%

E. Conclusion

In this paper, we propose an optimization scheme for subjective bit allocation at the CTU-level, aimed at boosting the compression efficiency of screen content coding. First, a novel R-D model is constructed grounded on the perceptual characteristic of HVS, measured by the gradient magnitude similarity deviation value. Then, the CTU-level subjective bit allocation for screen content coding is derived, using the subjective R-D model as a foundation. Finally, experimental results indicate that, the proposed method could obviously improve the subjective quality under similar bit consumption compared to existing methods.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China under grant U20A20157, 61771082, 62271096 and 61871062, the General Program of Chongqing Natural Science Foundation under grant cstc2021jcyj-msxmX0032, the Natural Science Foundation of Chongqing, China (cstc2020jcyj-zdxmX0024), the Science and Technology Research Program of Chongqing Municipal Education Commission under grant KJQN202300632, and the University Innovation Research Group of Chongqing (CXQT20017).

REFERENCES

[1] A. Heindel, B. Prestele, A. Gehlert and A. Kaup, "Enhancement Layer Coding for Chroma Sub-Sampled Screen Content Video," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 788-801, Feb. 2022.

[2] Y. Li and X. Mou, "Joint Optimization for SSIM-Based CTU-Level Bit Allocation and Rate Distortion Optimization," in IEEE Transactions on Broadcasting, vol. 67, no. 2, pp. 500-511, June 2021.

[3] T. Wang, F. Li and P. C. Cosman, "Learning-Based Rate Control for Video-Based Point Cloud Compression," in IEEE Transactions on Image Processing, vol. 31, pp. 2175-2189, 2022.

[4] G. Xiang, X. Zhang, X. Huang, et al., "Perceptual Quality Consistency Oriented CTU Level Rate Control for HEVC Intra Coding," in IEEE Transactions on Broadcasting, vol. 68, no. 1, pp. 69-82, March 2022.

[5] M. Zhou, X. Wei, C. Ji, T. Xiang and B. Fang, "Optimum Quality Control Algorithm for Versatile Video Coding," in IEEE Transactions on Broadcasting, vol. 68, no. 3, pp. 582-593, Sept. 2022.

[6] V. Sanchez, “Rate Control for Predictive Transform Screen Content Video Coding Based on RANSAC,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 11, pp. 4422-4438, Nov. 2021.
[7] Y. Chen, M. Wang, S. Wang, Z. Ni and S. Kwong, “A CTU-level Screen Content Rate Control for Low-delay Versatile Video Coding,” in IEEE Transactions on Circuits and Systems for Video Technology, 2023.
[8] Esmaeeli, H., Rezaei, M. “A content-based intra rate-distortion model for HEVC-SCC,” in Multimed Tools Appl 81, pp. 16515–16536, 2022.
[9] Y. Cao, X. Min, W. Sun and G. Zhai, “Attention-Guided Neural Networks for Full-Reference and No-Reference Audio-Visual Quality Assessment,” in IEEE Transactions on Image Processing, vol. 32, pp. 1882-1896, 2023.
[10] M. Li and S. Li, “Quality Assessment of Screen Content Images Based on Multi-Pathway Convolutional Neural Network,” in IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 1-5, 2022.
[11] Y. Cao, X. Min, W. Sun and G. Zhai, “Attention-Guided Neural Networks for Full-Reference and No-Reference Audio-Visual Quality Assessment,” in IEEE Transactions on Image Processing, vol. 32, pp. 1882-1896, 2023.
[12] H. Zeng, H. Huang, J. Hou, J. Cao, Y. Wang and K. -K. Ma, “Screen Content Video Quality Assessment Model Using Hybrid Spatiotemporal Features,” in IEEE Transactions on Image Processing, vol. 31, pp. 6175-6187, 2022.
[13] H. Duan, X. Min, Y. Zhu, G. Zhai, X. Yang and P. Le Callet, “Confusing Image Quality Assessment: Toward Better Augmented Reality Experience,” in IEEE Transactions on Image Processing, vol. 31, pp. 7206-7221, 2022.
[14] X. Wu, H. Wang, S. Hu, S. Kwong and C. . -C. J. Kuo, “Perceptually Weighted Mean Squared Error Based Rate-Distortion Optimization for HEVC,” in IEEE Transactions on Broadcasting, vol. 66, no. 4, pp. 824-834, Dec. 2020.
[15] Y. Li, Z. Chen, et al., “Rate Control for VVC, JVET-K0390,” Ljubljana, 2018. 10-18 July.
[16] https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.2+SCM-8.8/.
[17] Yu. H, Cohen. R, Rapaka. K and Xu. J, “Common Conditions for Screen Content Coding Tests,” JCTVC-S1015, October. 2014.
[18] T. Tang, L. Li and J. Li, “Improved hierarchical quantisation parameter setting method for screen content coding in high efficiency video coding,” in IET Image Processing, vol. 13, pp. 1382-1390, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multimodal Emotion Recognition based on 2D Kernel Density Estimation for Multiple Labels Fusion

Zhaojie Luo* and Kazunori Komatani *
* SANKEN, Osaka University, Japan
E-mail: luo@sanken.osaka-u.ac.jp; komatani@sanken.osaka-u.ac.jp

***Abstract*—Psychological emotion can be typically categorized into discrete emotion states, such as anger, happiness, and neutrality labels, as well as estimated as degrees of a 2D continuous valence-arousal (VA) space. Previous studies on multimodal emotion recognition have employed fusion mechanisms for multiple modalities but treated emotion labels and VA degrees as separate recognition tasks. By modeling the relationship between these different labels, it becomes possible to leverage training datasets with different types of labels for improving multimodal emotion recognition. In this study, we explore the utilization of multiple labels by employing a 2D Kernel Density Estimation (2D-KDE) method to mathematically model their relations. Subsequently, we propose a label fusion layer (LFL) based on these relations to adjust the predicted probabilities of emotion states obtained from existing baselines of multimodal emotion recognition networks. Through extensive experiments, we demonstrate the effectiveness of our proposed model in improving emotion recognition performance and achieving superior results on the IEMOCAP and OMG-Emotion datasets.**

I. Introduction

The recognition of emotions in human-to-human communication is greatly influenced by the underlying emotional aspects embedded in our language. The emerging field of automatic emotion recognition (AER) aims to identify emotions from various input signals, such as audio, text, or facial image signals. Previous studies have shown that humans rely on multiple modalities to comprehend emotions, rather than relying solely on a single modality [1]. For example, research highlighted in [2] indicates that voice calls convey more information than text messages, suggesting that the effective prosody of audio can provide additional cues for emotion recognition. Similarly, studies have indicated that face-to-face communication is more effective in conveying emotions compared to voice calls alone [3], suggesting that visual cues play a significant role in emotion classification.

Given the importance of multiple modalities in emotion recognition, fusion methods that combine different modalities have gained substantial popularity in the field. These fusion methods can be broadly categorized as early fusion [4], late fusion [5], [6], or hybrid fusion approaches [7], [8].

On the other way, in the realm of emotions, previous research has put forth two primary perspectives on their classification. The first perspective posits that emotions are distinct and fundamentally different constructs [9]. According to this view, emotions are discrete categories that are not directly comparable to one another. The second perspective suggests that emotions can be characterized in terms of dimensional groupings [10]. This perspective proposes that emotions can be represented along continuous dimensions, specifically valence (measuring positivity or negativity) and arousal (quantifying the intensity or strength of the emotional response). In light of these perspectives, psychological labels for emotions can typically be divided into two categories. The first category encompasses discrete emotion states such as anger, happiness, and neutrality, which represent specific emotions. The second category involves a continuous two-dimensional (2D) space that encompasses degrees of valence and arousal, providing a more nuanced representation of emotions.

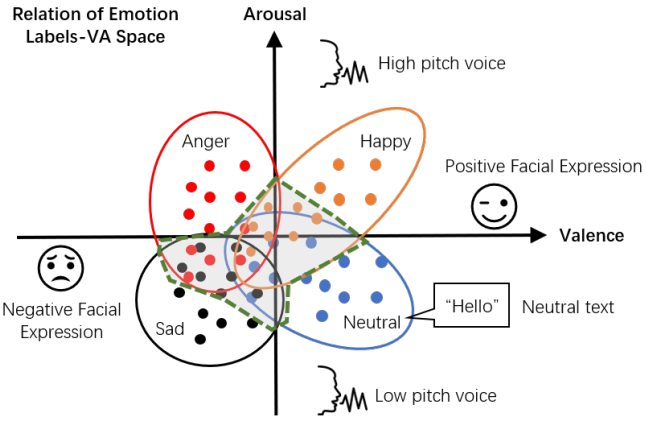


Fig. 1. Relations between emotion states and Valence-Arousal space.

However, previous studies have predominantly treated emotion state labels and valence-arousal degrees as separate entities, utilized for distinct tasks such as emotion classification and valence-arousal degree estimation, respectively. Limited efforts have been made to integrate these different label types into a unified task. Recent studies have proposed that employing the valence-arousal (VA) space to represent both emotion dimensions and emotion states is a more efficient and effective approach for solving multimodal-related problems than relying solely on discrete emotion states [11].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

To illustrate this integration, Figure 1 exemplifies how different modalities can be associated with the VA space. The pitch feature of audio, recognized as one of the most reliable emotion features [12], can be easily annotated with arousal degrees. Visual features, on the other hand, are well-suited for recognizing valence degrees. The text modality provides a crucial basis for recognizing emotion states. By utilizing the VA space, a more comprehensive and holistic representation of emotional information can be achieved. This, in turn, leads to improved performance in multimodal emotion recognition tasks.

In this study, we aim to improve the accuracy of emotion recognition by leveraging the relationships between emotion states and the valence-arousal (VA) space. To achieve this, we propose a network architecture that integrates VA-degree estimation networks (VA-networks) with emotion state recognition networks (ER-networks) into a single task. The objective is to enhance the overall emotion recognition accuracy by considering both emotion states and VA degrees simultaneously.

Figure 1 illustrates that emotion states and VA degrees exhibit overlapping and non-overlapping areas in the VA space. This indicates that they are neither entirely independent nor completely dependent on each other. To address this relationship, we utilize a 2D kernel density estimation (KDE) technique to calculate the probability area in the VA space for each emotion state. This KDE-based estimation allows us to quantify the association between emotion states and their corresponding VA regions. We refer to these KDE probability estimates as Label Fusion Layers (LFL). The LFL acts as a mechanism to adjust the predicted results for the basic emotion state recognition task by considering the probability distribution of VA degrees associated with each emotion state. By incorporating this information, we aim to refine and improve the accuracy of the emotion state predictions.

To evaluate the effectiveness of our proposed LFL approach, we compare its performance against several baseline models. These baselines include the BC-LSTM model, which is based on Long Short-Term Memory (LSTM) networks [5], the MMAN model, which is based on Transformer networks [8], and the HFGCN model, which utilizes Graph Convolutional Networks (GCN) [13]. We experiment with different kernel sizes in the 2D KDE to examine their impact on the performance of the emotion recognition task.

To conduct our evaluation, we utilize publicly available datasets, namely the IEMOCAP dataset [14] and the OMG-Emotion dataset [15]. These datasets provide a diverse range of multimodal data, including audio, visual, and textual modalities, along with annotations for emotion states and VA degrees. By employing these datasets, we can assess the performance of our proposed LFL approach in comparison to the baseline models and gain insights into its effectiveness in enhancing emotion recognition accuracy.

The rest of the paper is organized as follows. Section II presents the details of the related works. Section III introduces the proposed labels fusion based network. Section IV describes the experimental setup. Section V reports the results and analysis. Finally, conclusions are drawn in Section VI.

II. Related Work

A. Valence Arousal Space and Discrete Emotion States

In the field of emotion research, the Valence-Arousal space is a commonly used two-dimensional representation that captures the emotional dimensions of valence and arousal. Valence represents the positivity or negativity of an emotion. It ranges from negative or unpleasant emotions (e.g., sadness, anger) to positive or pleasant emotions (e.g., happiness, joy). The valence dimension is typically represented on a continuum, with negative emotions on one end and positive emotions on the other. Arousal, on the other hand, represents the level of physiological and psychological activation associated with an emotion. It ranges from low arousal (e.g., calmness, relaxation) to high arousal (e.g., excitement, fear). The arousal dimension is also represented on a continuum, with low arousal at one end and high arousal at the other.

By combining the valence and arousal dimensions, the Valence-Arousal space allows for a comprehensive representation of emotions. Different emotions can be positioned within this space based on their valence and arousal values. For example, emotions with high positive valence and high arousal, such as excitement, would be located in the upper-right quadrant of the Valence-Arousal space. Discrete emotion states, on the other hand, refer to specific emotions that are categorized and labeled separately. Examples of discrete emotion states include happiness, sadness, anger, fear, surprise, disgust, and neutral. These discrete emotions are distinct and have their own unique characteristics and patterns of valence and arousal within the Valence-Arousal space. In research and analysis, the Valence-Arousal space and discrete emotion states are often used to understand and classify emotions, study emotional responses, and develop models and algorithms for emotion recognition and analysis

B. 2D Kernel Density Estimation

Two-Dimensional Kernel Density Estimation (2D KDE) is a statistical technique used to estimate the probability density function (PDF) of a two-dimensional random variable based on observed data points. It extends the concept of Kernel Density Estimation (KDE), which is primarily used for univariate data, to handle data in two dimensions. The kernel function determines the shape and width of the smoothing window around each data point. It is typically a symmetric, positive-valued function, such as the Gaussian kernel. The bandwidth parameter, denoted by h, controls the width of the kernel and influences the smoothness of the estimated density. A larger bandwidth results in a smoother estimate but may oversmooth the data, while a smaller bandwidth captures more local details but may introduce more noise.

To estimate the 2D PDF, the kernel function is placed at each data point, denoted by (x_i, y_i), and the contributions from all kernels are summed up. The result is a density estimate, denoted by $\hat{f}(x, y)$, that assigns higher density values

to regions with a higher concentration of data points and lower density values to sparser regions. Mathematically, the 2D KDE can be represented as:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}, \frac{y-y_i}{h}\right)$$

where n is the number of observed data points, and $K(u,v)$ is the kernel function, typically symmetric and positive-valued. The choice of kernel function, $K(u,v)$, and bandwidth, h, are crucial in 2D KDE. Different kernel functions, such as the Gaussian kernel [16], Epanechnikov [17], biweight [18], or triangular kernels [19], can be used depending on the desired properties of the estimate. Bandwidth selection methods, such as rule-of-thumb bandwidth estimator (RT) [20], automatic bandwidth selectors (Unbiased Cross Validation [21], Biased Cross Validation [22], or Complete Cross-Validation [23]), help determine an optimal bandwidth that balances the trade-off between bias and variance.

In this study, to capture the relationships between emotion states and the VA space, we employ a 2D kernel density estimation (KDE) technique. The KDE estimates the probability density function for each emotion state in the VA space, considering both overlapping and non-overlapping areas. The KDE provides a comprehensive representation of the association between emotion states and their corresponding VA regions.

III. Method

A. Framework

Our proposed architecture, depicted in Figure 2, utilizes emotion recognition networks (ER-networks) and Valence Arousal degree estimation networks (VA-networks) with integrated label fusion layers. Both the ER-networks and VA-networks share identical architectures, which are built upon baseline models such as LSTM, Transformer, and GCN. The detailed explanation of these baseline models can be found in Section IV-C. To enhance emotion recognition using valence arousal degrees, we employ a label fusion layer, as described in Section III-B, which transforms valence arousal degrees into the probability density of emotional states denoted as $\hat{f}_h^{\mathbf{e}}(v,a)$. This transformation is accomplished through the use of 2D kernel density estimation (2D-KDE), a technique explained in Section II-B. The predicted probability distribution ($\mathcal{P}^e(X)$) of emotion labels obtained from the ER-networks is combined with the transferred probability distribution ($\hat{f}_h^{\mathbf{e}}(v,a)$) of emotion labels derived from the VA-networks(consist of valence degrees network and arousal degrees network). This fusion process yields the final probability distribution ($\mathcal{G}^e(X)$) of emotions. For the task of emotion recognition, we output the emotional state with the highest probability.

B. Labels Fusion Layer

To perform the multi-label fusion, we applied 2D kernel density estimation (2D-KDE) to determine the "centered" area of each emotion states $\mathbf{e} \in \{happy, sad, neutral, angry\}$, based on their corresponding valence and arousal degrees

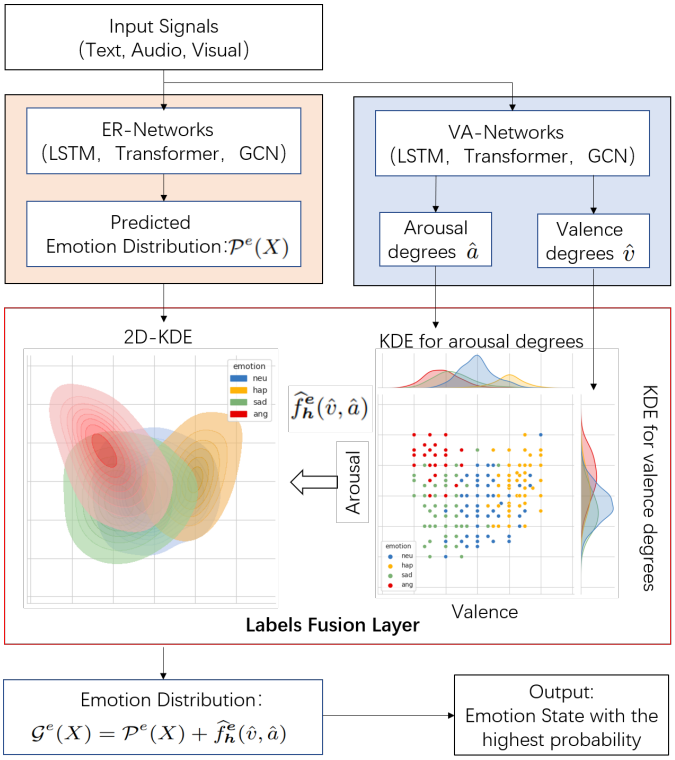


Fig. 2. Architecture for models

$(V^{\mathbf{e}}, A^{\mathbf{e}})$. The 2D kernel density estimation is then applied as follows:

$$\hat{f}_h^{\mathbf{e}}(v,a) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_v^{\mathbf{e}}}\frac{1}{h_a^{\mathbf{e}}}K(\frac{v^{\mathbf{e}}-V_i^{\mathbf{e}}}{h_v^{\mathbf{e}}})K(\frac{a^{\mathbf{e}}-A_i^{\mathbf{e}}}{h_a^{\mathbf{e}}}) \quad (1)$$

Each of the n observations is of the form $(V_i^{\mathbf{e}}, A_i^{\mathbf{e}})$, where the $V_i^{\mathbf{e}}$ indicates the value that the random variable valence degrees $V^{\mathbf{e}}$ takes on at the i th observation, and the $A_i^{\mathbf{e}}$ represents the same for $A^{\mathbf{e}}$. ($h_v^{\mathbf{e}}$, $h_a^{\mathbf{e}}$) are the bandwidths employed in the 2D-KDE for valence degrees and arousal degrees. K is the kernel function, which are described in the 2D-KDE of the related work (Section II-B). Among the Kernels, the Gaussian kernel [16] and the Epanechnikov kernel [17] are often used in previous media data density estimation works. As proved in [24], the Gaussian kernel is quick and easier to compute, thus we choose Gaussian Kernel $K(x) = \frac{1}{\sqrt{2\pi}}exp\left(\frac{1}{2}(-x^2)\right)$ as Kernel function. Then the Density function $\hat{f}_h^{\mathbf{e}}(x)$ can be calculated as

$$\begin{aligned}\hat{f}_h^{\mathbf{e}}(v,a) &= \hat{f}_{h_v^{\mathbf{e}},h_a^{\mathbf{e}}}(v^{\mathbf{e}},a^{\mathbf{e}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_v^{\mathbf{e}}}\frac{1}{h_a^{\mathbf{e}}}\cdot\frac{1}{\sqrt{2\pi}}exp\left\{\frac{1}{2}\left(-(\frac{v^{\mathbf{e}}-V_i^{\mathbf{e}}}{h_v^{\mathbf{e}}})^2\right)\right\} \\ &\quad\cdot\frac{1}{\sqrt{2\pi}}exp\left\{\frac{1}{2}\left(-(\frac{a^{\mathbf{e}}-A_i^{\mathbf{e}}}{h_a^{\mathbf{e}}})^2\right)\right\}\end{aligned} \quad (2)$$

For the bandwidth ($h_v^{\mathbf{e}}$, $h_a^{\mathbf{e}}$), as described in Related work (Section II-B), the previous works tried objective rule-of-thumb bandwidth estimator (RT) [20] and automatic bandwidth

selectors (Unbiased Cross Validation [21], Biased Cross Validation [22], and Complete Cross-Validation [23]) to minimize the estimation error of the target density $\hat{f}_h^{\mathbf{e}}()$. Among them, the RT ($h_{RT} = (\frac{4\hat{\sigma}^5}{3n})^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}$) is a quick, simple, and inexpensive selector for the optimal choice for normal-like data. As shown in Fig. 2, the Valence and Arousal degrees trend to be normal-like data, the bandwidth in $\hat{f}_h^{\mathbf{e}}()$ can be calculated as

$$h_a^{\mathbf{e}} = (\frac{4\hat{\sigma_a^{\mathbf{e}}}^5}{3n})^{\frac{1}{5}} \approx 1.06\hat{\sigma_a^{\mathbf{e}}}n^{-1/5}; \\ h_v^{\mathbf{e}} = (\frac{4\hat{\sigma_v^{\mathbf{e}}}^5}{3n})^{\frac{1}{5}} \approx 1.06\hat{\sigma_a^{\mathbf{e}}}n^{-1/5} \tag{3}$$

where, $\hat{\sigma_a^{\mathbf{e}}}$ and $\hat{\sigma_v^{\mathbf{e}}}$ are the standards of valence degrees and arousal degrees for emotion $\mathbf{e} \in \{neutral, happy, angry, happy\}$.

Note that the kernel weights correspond to equal bandwidth in each direction, i.e. $\boldsymbol{h} = (h_a^{\mathbf{e}}, h_a^{\mathbf{e}})^\top = (1,1)^\top$. When we use different bandwidths, the observations around (v, a) in the density estimate $\widehat{f}_{\boldsymbol{h}}^{\boldsymbol{e}}(v, a)$ will be used with different weights in both dimensions.

C. Classification

For the classification stage, the multimodal emotion features X are fed into three fully connected networks: an emotion classifier, a valence classifier, and an arousal classifier. Each of these networks includes a fully connected layer with a softmax function to produce the probability distributions of the emotion labels $\mathcal{P}^e(X)$, arousal degrees $\mathcal{P}^a(X)$, and valence degrees $\mathcal{P}^v(X)$. To incorporate the valence and arousal degrees into the emotion probability, we extract the maximum outputs of the arousal degrees $\hat{a} = argmax(\mathcal{P}^a(X))$ and valence degrees $\hat{v} = argmax(\mathcal{P}^v(X))$. These values are then transferred to the emotion probability distribution $\widehat{f}_{\boldsymbol{h}}^{\boldsymbol{e}}(\hat{v}, \hat{a})$ using the Label fusion Layer described in Section III-B. Finally, the distribution $\mathcal{G}^e(X)$ of the emotion labels can be adjusted as follows:

$$\mathcal{G}^e(X) = \mathcal{P}^e(X) + \widehat{f}_{\boldsymbol{h}}^{\boldsymbol{e}}(\hat{v}, \hat{a}) \tag{4}$$

D. Training Steps

According to the VA model [25]–[27] based on psychology, emotions can be categorized or described using a two-dimensional VA subspace. In light of this, we have designed a special loss function $L_{e,v,a}$ to jointly predict emotion labels and VA degrees. This loss function employs categorical cross-entropy $\mathcal{L}$ for both emotion classification and VA classification. The formulation is as follows:

$$L_{e,v,a} = W_1\mathcal{L}(\mathcal{P}_i^e, y_i^e) + W_2\mathcal{L}(\mathcal{P}_i^v, y_i^v) + W_3\mathcal{L}(\mathcal{P}_i^a, y_i^a), \\ \mathcal{L}(\mathcal{P}_i, y_i) = -\sum_{i=1}^{C} y_i \log(\mathcal{P}_i) \tag{5}$$

In this equation, C represents the number of input data, and $\mathcal{P}_i^e, \mathcal{P}_i^v, \mathcal{P}_i^a$ are the model's output probability distributions for the emotion labels, valence degrees, and arousal degrees, respectively, for the i-th data sample. The expected labels for emotion, valence, and arousal are denoted as y_i^e, y_i^v, y_i^a, respectively. The weights W_1, W_2, W_3 are hyperparameters that determine the importance of each loss term and can be set manually or optimized using methods such as the Tree-structured Parzen Estimator (TPE) [28], which is an automatic hyperparameter tuning method. In our experiments, we found that setting $W_1 = 0.6$, $W_2 = 0.2$, and $W_3 = 0.2$ yielded the best results for the weights.

As the VA degrees are discrete but ordinal, we experimented with different types of loss functions in our multitask loss formulation, including both regression-type and classification-type losses. However, the cross-entropy loss was found to be significantly better than other types of losses.

TABLE I
DISTRIBUTION OF TRAINING AND TEST SETS.

Dataset	Split	Utterance
IEMOCAP	Train/Test	5810/1623
OMG-Emotion	Train/Test	2441/2229

IV. EXPERIMENTAL SETUP

A. Dataset

In this work, we utilizes the relation between emotion states and the Valence-Arousal space, which requires datasets (such as IEMOCAP and OMG-Emotion) that include both valence and arousal labels, as commonly used in multimodal emotion recognition.

The **IEMOCAP** [14] dataset comprises 10,000 videos consisting of 5-minute dyadic conversations, specifically collected for human emotion analysis. Each conversation involves two speakers, and the dialogues are further segmented into individual spoken utterances. Each utterance within every dialogue is annotated with an emotion label. To align with previous works, we consider the emotion classes of angry, happy (excited), sad (frustrated), and neutral. Moreover, in addition to emotion labels, each utterance was assessed by two evaluators for its valence and arousal degrees, which range from one to five. The final valence and arousal (VA) degrees were obtained by averaging the scores from the two evaluators.

The **OMG-Emotion** [15] dataset were collected from YouTube with no environmental constraints. The dataset is annotated with seven discrete emotion states: *disgust, fear, happiness, surprise, sadness, anger,* and *neutral*. Additionally, it includes two continuous emotion labels: *arousal* in the range [0,1] and *valence* in the range [-1,1]. This study specifically focuses on four emotion classes: *anger, happiness (including surprise), sadness (including disgust and fear)*, and *neutral*, aligning with previous research in the field.

Speakers in the training and test sets were kept disjoint from each other for both datasets under the assumption of being unknown at inference time. We used 10% of the training dialogues as the validation set and followed the previous work [8], [15] to set distribution of training and test sets shown in Table I

TABLE II
RESULTS OF BASELINE MODELS (BC-LSTM, MMAN, AND HFGCN) AND PROPOSED MODELS (APPLYING THE LFL WITH DIFFERENT BANDWIDTH (H, H/2 AND 2H) IN CORRESPONDING BASELINE MODELS) OVER 5-FOLDS CROSS-VALIDATION

Methods	IEMOCAP					OMG-Emotion				
	happy	sad	neutral	angry	avg. F1 %	happy	sad	neutral	angry	avg. F1 %
BC-LSTM	57.43	77.87	65.29	72.73	71.78	60.23	**57.56**	65.93	62.33	61.43
BC-LSTM ($LFL_{h/2}$)	65.66	78.148	**66.25**	73.62	72.15	62.21	56.24	62.45	63.14	60.25
BC-LSTM (LFL_{h})	**72.32**	**78.74**	65.23	**74.43**	**74.95**	**63.23**	57.49	**66.13**	**64.23**	**62.25**
BC-LSTM (LFL_{2h})	69.23	77.87	64.38	70.39	73.13	60.23	56.68	65.93	62.33	61.43
MMAN	63.32	**78.37**	68.24	74.29	73.94	62.52	59.45	67.23	61.45	64.53
MMAN ($LFL_{h/2}$)	68.22	76.21	65.19	74.73	74.34	61.25	**60.06**	66.59	63.33	64.24
MMAN(LFL_{h})	**73.34**	77.45	**69.93**	**76.82**	**76.11**	**63.23**	58.46	**67.93**	**65.23**	**65.05**
MMAN (LFL_{2h})	69.13	76.23	68.29	73.68	74.91	62.93	59.56	67.13	62.36	64.77
HFGCN	67.14	**79.87**	69.11	72.45	75.01	59.04	60.45	69.23	66.32	67.43
HFGCN ($LFL_{h/2}$)	71.42	77.17	68.23	73.28	75.31	60.24	**61.56**	69.89	68.23	67.03
1st HFGCN (LFL_{h})	**74.43**	78.35	**69.25**	**75.13**	**77.91**	**61.33**	60.69	**70.13**	**69.23**	**68.24**
HFGCN (LFL_{2h})	72.34	75.54	68.54	71.51	76.81	60.23	59.65	67.93	68.33	67.37

B. Features Extraction

We followed *Poria et al.* [5] for the extraction of low-level features. To extract these features, we initially separate the input video of an utterance into its corresponding text, video frames, and speech modalities. Then, individual pre-trained networks, transferred from other tasks, are employed for the extraction process. For each modality, the feature of each utterance is extracted as a fixed-length vector."

Speech: We utilized the OpenSMILE toolkit [29] in combination with the IS13-ComParE feature set [30] to extract various features from the speech waveform, including pitch and voice intensity. The extraction process involved a frame-rate of 30 Hz and a sliding window of 100 ms. Prior to feature extraction, the raw waveform underwent normalization using Z-standardization, and a voice intensity threshold was applied to identify segments with vocal expressions. The extracted features encompassed Mel Frequency Cepstral Coefficients (MFCC), spectral centroid, spectral flux, beat histogram, beat sum, voice intensity, pitch, mean, root quadratic mean, among others [31].

Text: Word2Vec [32] has been widely adopted as a pre-trained word vector representation for text due to its effectiveness in sentiment classification. In our approach, each word in an utterance transcript is initially embedded into a 300-dimensional Word2Vec vector. These embedded word vectors are then concatenated, padded if necessary, and further standardized into a 1-dimensional vector by passing through a CNN [33].

Visual frames: To extract the body language of the speaker, we employ a pre-trained 3D-CNN [34] originally designed for human action recognition. The 3D-CNN is applied to the consecutive visual frames capturing the speaker's upper body. By analyzing the relevant features within each frame and the changes across a given number of consecutive frames, known as motion cues, the 3D-CNN effectively captures the body language information.

C. Baseline Systems

We evaluated the proposed labels fusion layers using the following multimodal emotion recognition baseline models over 5-folds cross-validation: BC-LSTM [5], MMAN [8], and HFGCN [13]. BC-LSTM and MMAN are baseline and state-of-the-art (SOTA) methods in multimodal emotion recognition, respectively, and they were trained using only emotion state labels. HFGCN, on the other hand, was trained using both emotion state labels and valence arousal degrees, employing cross-entropy as the multitask loss. In cases where the baseline models did not report certain results on the datasets, we re-implemented the models to ensure a comprehensive evaluation. In the context of 2D Kernel Density Estimation, the bandwidth plays a crucial role. Therefore, we experimented with different bandwidth sizes: normal size (h), half size ($h/2$), and double size ($2h$). The choice of bandwidth size is essential for achieving accurate 2D Kernel Density Estimation results.

BC-LSTM : BC-LSTM (Bidirectional Convolutional LSTM) is the LSTM baseline method, in which context-aware utterance representations are generated by capturing the contextual content from the surrounding utterances using a Bi-directional LSTM network. It is commonly used for sequential data analysis, such as natural language processing and time series prediction tasks.

MMAN: MMAN represents a recent advancement in multimodal AER, leveraging the power of the Transformer architecture, hybrid fusion LSTM network, and multimodal attention mechanisms. By effectively integrating multimodal information and capturing long-range dependencies, MMAN aims to improve the accuracy and performance of multimodal emotion recognition tasks with cross attention for multi modalities.

HFGCN : This is a recent state-of-the-art GCN-based multimodal AER method that uses GCN to fuse multimodal information with the intermodality interactions. In HFGCN, the multimodal information is represented as a graph, where each modality is considered as a node. The GCN layers capture intermodality interactions by propagating information through the graph. This allows the model to leverage the relationships and dependencies between different modalities to enhance the fusion process.

LFL variants: This is our proposed method that applies labels fusion layers based on 2D kernel density estimation with different bandwidth (h, $h/2$, and $2h$) in baseline models as (LFL_h), ($LFL_{h/2}$) and (LFL_{2h}).

TABLE III
RESULTS OF HFGCN AND PROPOSED MODEL HFGCN (LFL_h) EVALUATING WITH DIFFERENT MODALITIES (TEXT (T), AUDIO (A), AND VISUAL (V)) IN IEMOCAP DATASET (AVG. F1 %) OVER 5-FOLDS CROSS-VALIDATION.

Modalites	T	A	V	T+A	T+V	A+V	T+A+V
HFGCN	71.28	65.24	67.35	74.78	73.41	70.38	75.01
HFGCN (LFL_h)	72.29 (+1.01)	70.45 (**+5.21**)	69.47 (+2.12)	77.79 (+3.01)	74.23 (+0.82)	71.69 (+1.31)	77.91 (+2.90)

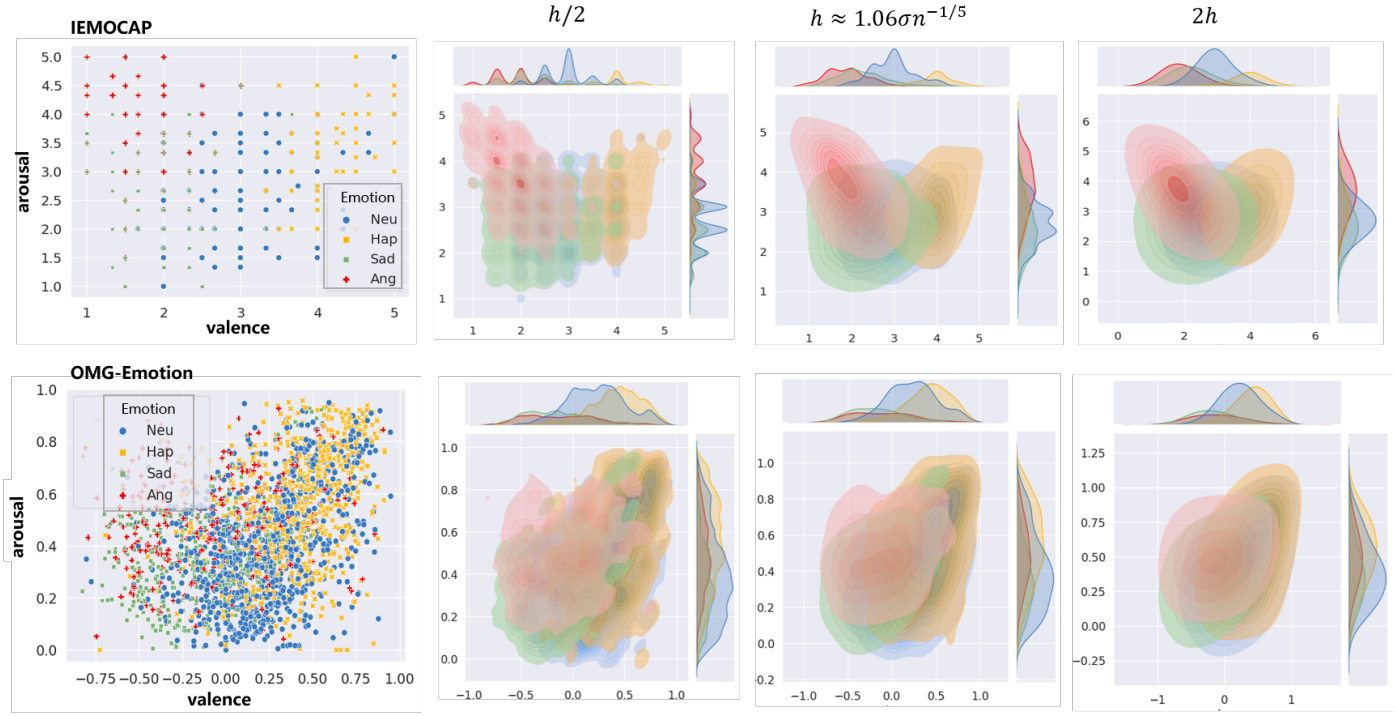


Fig. 3. 2D Kernel Density Estimation with different bandwidths for datasets IEMOCAP and OMG-Emotion.

V. RESULTS AND ANALYSIS

A. Comparison with baselines

Table II presents the results, which demonstrate that the application of the label fusion layers (LFL) to the baseline models leads to improved average F1 scores. This improvement indicates that the 2D Kernel Density Estimation (2D-KDE) technique for label fusion is effective in enhancing emotion recognition performance. It is worth noting that the F1 scores for angry and happy emotions showed larger improvements compared to sad and neutral emotions. This observation aligns with the fact that the overlap areas of the former emotions were smaller, as illustrated in Figure 3. However, it is important to note that for some baseline models, there was a slight decrease in F1 scores specifically for sad and neutral emotions. This suggests that the efficacy of the 2D-KDE approach depends on the relationship between the annotated VA degrees and the emotion labels. The relationship between the two factors can influence the performance of the label fusion process.

B. Comparison with different bandwidth

In our evaluation, we considered different bandwidth values ($h/2$, h, and $2h$) for the density estimation in the proposed label fusion layers (LFL). Comparing the results, we found that the Gaussian Kernel with a normal size (h) generally produced the best outcomes for most emotions. Figure 3 visually illustrates that using a normal size h allows for appropriate positioning of density centers for each emotion, resulting in more effective minimization of overlap areas compared to other bandwidth sizes. Smaller overlap areas generally correspond to better results, as demonstrated in this study. Based on these findings, we conclude that the Gaussian Kernel is an effective option for emotion label density estimation in the LFL. The bandwidth can be set using the rule-of-thumb bandwidth estimator, which provides a reliable estimation for achieving optimal performance.

C. Comparison with different datasets

The results presented in Table II and Figure 3 reveal that when comparing the performance of LFL with the same bandwidth across different datasets, there were greater improvements observed in the IEMOCAP dataset compared to The OMG-Emotion dataset. This difference in performance can be attributed to the nature of the data in each dataset. The OMG-Emotion dataset consists of wild data collected from YouTube videos, where the valence degrees are annotated on

a scale from -1 to 1 and arousal is annotated on a scale from 0 to 1. In contrast, the IEMOCAP dataset is an acted emotion database where valence arousal degrees are annotated on a scale from 1 to 5 levels. The less separated valence and arousal annotations in The OMG-Emotion dataset compared to the clearer distinctions in the IEMOCAP dataset can explain the difference in LFL performance. The LFL approach tends to be more efficient when there is a clear distinction between emotional data in the datasets. Therefore, the LFL approach is particularly effective in datasets where the emotional categories are well-separated and distinct.

D. Comparison with different modalites

As mentioned in the introduction, different modalities can provide complementary information due to variations in labeling styles. Therefore, we extended our comparison by applying the label fusion layers (LFL) to the HFGCN model, which demonstrated the best results among different modalities in Table II. Table III presents the results, indicating that utilizing all modalities together leads to improved performance. Furthermore, when comparing the uni-modality results obtained with and without LFL, we observed that the visual and speech modalities showed greater improvements compared to the text modality. These findings suggest that individuals express their valence and arousal more prominently through speech and visual cues rather than text. The richer and more expressive nature of speech and visual modalities allows for capturing and conveying emotional information more effectively. Incorporating these modalities in multimodal emotion recognition models, particularly with the aid of LFL, contributes to better performance.

VI. Conclusion

In this work, we have proposed a label fusion layer (LFL) that utilizes a 2D Gaussian Kernel Density Estimation to adjust predicted probabilities of emotion states using valence-arousal (VA) degrees. The main objective was to enhance emotion state recognition by exploiting the correlation between emotion states and VA degrees. Through experiments, we have demonstrated the effectiveness of our proposed method across different baseline models, including LSTMs, Transformer, and GCN. We have also observed that the efficiency of the method can be influenced by the choice of dataset, such as IEMOCAP and OMG-emotion, as well as the modality involved, including text, audio, and visual. Overall, our work showcases the potential of employing multiple label fusion techniques for improving the accuracy of emotion recognition systems. By incorporating VA degrees and leveraging the correlation between VA degrees and emotion states, we have achieved promising results in enhancing emotion recognition performance.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Numbers JP19H05692 and JP22H00536.

References

[1] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: Plasticity and interactions," *Current opinion in neurobiology*, vol. 11, no. 4, pp. 505–509, 2001.

[2] N. Sebe, I. Cohen, and T. S. Huang, "Multimodal emotion recognition," in *Handbook of pattern recognition and computer vision*, World Scientific, 2005, pp. 387–409.

[3] A. L. Drolet and M. W. Morris, "Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts," *Journal of Experimental Social Psychology*, vol. 36, no. 1, pp. 26–50, 2000.

[4] J. Sebastian and P. Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," in *Proc. Interspeech*, 2019, pp. 51–55.

[5] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 873–883.

[6] J. Xue, Z. Luo, K. Eguchi, T. Takiguchi, and T. Omoto, "A bayesian nonparametric multimodal data modeling framework for video emotion recognition," in *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 601–606.

[7] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, *Multimodal sentiment analysis using hierarchical fusion with context modeling*, 2018. arXiv: 1806.06228 [cs.CL].

[8] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," *arXiv preprint arXiv:2009.04107*, 2020.

[9] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[10] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[11] R. G. Praveen, P. Cardinal, and E. Granger, "Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.

[12] Z. Luo, S. Lin, R. Liu, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Decoupling speaker-independent emotions for voice conversion via source-filter networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 11–24, 2022.

[13] S. Tang, Z. Luo, G. Nan, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Fusion with hierarchical graphs for multimodal emotion recognition," in *2022 Asia-Pacific*

Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2022, pp. 1288–1296.

[14] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[15] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2018, pp. 1–7.

[16] M. Kristan, A. Leonardis, and D. Skočaj, "Multivariate online kernel density estimation with gaussian kernels," *Pattern recognition*, vol. 44, no. 10-11, pp. 2630–2642, 2011.

[17] S. Weglarczyk, "Kernel density estimation and its application," in *Proceedings of the ITM Web of Conferences*, EDP Sciences, vol. 23, 2018, p. 00 037.

[18] M. C. Jones, "Simple boundary correction for kernel density estimation," *Statistics and computing*, vol. 3, pp. 135–146, 1993.

[19] M. Cuturi, "Fast global alignment kernels," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 929–936.

[20] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[21] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[22] S. R. Sain, K. A. Baggerly, and D. W. Scott, "Cross-validation of multivariate densities," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 807–817, 1994.

[23] M. D. Mullin and R. Sukthankar, "Complete cross-validation for nearest neighbor classifiers.," in *Proceedings of the ICML*, 2000, pp. 639–646.

[24] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, Springer, 2012, pp. 702–715.

[25] L. F. Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition and Emotion*, vol. 12, no. 4, pp. 579–599, 1998. DOI: 10.1080/026999398379574.

[26] L. F. Barrett, "Solving the emotion paradox: Categorization and the experience of emotion," *Personality and social psychology review*, vol. 10, no. 1, pp. 20–46, 2006.

[27] h. Yang and H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, May 2012. DOI: 10.1145/2168752.2168754.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, ACM, 2010, pp. 1459–1462.

[30] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[31] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2019.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[34] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

RobustL2S: Speaker-Specific Lip-to-Speech Synthesis exploiting Self-Supervised Representations

Neha Sahipjohn*, Neil Shah*†, Vishal Tambrahalli*, Vineet Gandhi*
* CVIT, Kohli Centre for Intelligent Systems, IIIT Hyderabad, India
† TCS Research, Pune, India

***Abstract*—Significant progress has been made in speaker-dependent Lip-to-Speech synthesis, which aims to generate speech from silent videos of talking faces. Current state-of-the-art approaches primarily employ non-autoregressive sequence-to-sequence architectures to directly predict mel-spectrograms or audio waveforms from lip representations. We hypothesize that the direct mel-prediction hampers training/model efficiency due to the entanglement of speech content with ambient information and speaker characteristics. To this end, we propose RobustL2S, a modularized framework for Lip-to-Speech synthesis. First, a non-autoregressive sequence-to-sequence model maps self-supervised visual features to a representation of disentangled speech content. A vocoder then converts the speech features into raw waveforms. Extensive evaluations confirm the effectiveness of our setup, achieving state-of-the-art performance on the unconstrained Lip2Wav dataset and the constrained GRID and TCD-TIMIT datasets. Speech samples from RobustL2S can be found at https://neha-sherin.github.io/RobustL2S/**

I. Introduction

Understanding lip movements offers a distinct advantage in situations where auditory cues are unavailable. It proves particularly valuable for individuals with hearing impairments, and speech disorders and aids in speech rehabilitation by providing visual feedback [1]. The synthesis of accurate speech from lip movements can assist in tasks such as movie dubbing [2], language learning, forensic investigations [3], video conferencing in noisy conditions, voice inpainting [4] or giving artificial voice to people who cannot produce intelligible sound.

The problem of Lip-to-Speech synthesis is inherently ill-posed because a sequence of lip movements can correspond to multiple possible speech utterances [5]. Additional challenges arise from factors such as head pose movements, non-verbal facial expressions, variations in capture quality, and ambient noise, which further complicate the problem. Reliance on contextual information, such as environment, place, topic, etc., can help alleviate the Lipreading challenges [6], [7]; however, such information may not always be available.

Most existing approaches constitute an encoder-decoder architecture; the encoder maps the lip sequence to intermediate representations, which are then directly decoded into mel-spectrograms. The major drawback of this approach is that apart from speech content, the decoder is also forced to predict the time-varying speaker and ambient noise characteristics present in the ground truth Mel. We hypothesize that this dependence hurts the model's performance in terms of speech intelligibility, reducing its usability for various downstream applications [8]. Our work addresses these limitations by taking a modularized approach, exploiting the advances in Self-supervised learning (SSL) in audio and audio-visual scenarios.

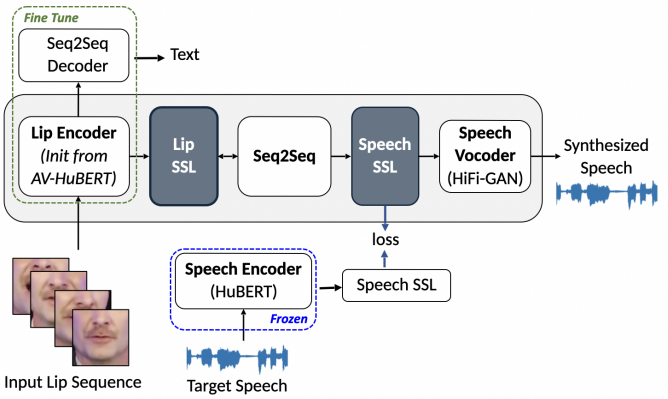


Fig. 1. The proposed RobustL2S model utilizes lip encoder and speech encoder to extract SSL representations from lip sequences and their corresponding speech. A Seq2Seq model maps the lip representations to speech representations, which are then decoded to synthesize speech.

Fig. 1 illustrates the proposed RobustL2S framework. In contrast to direct mel prediction from lip features, we take a two-staged approach. The first step extracts SSL representations of lip sequences and maps them to corresponding speech SSL representations using a sequence-to-sequence (Seq2Seq) model. The key idea is to use speech embeddings that disentangle the content from the speaker and ambient information. The second stage maps the content-rich speech embeddings to raw speech using a speaker-conditioned vocoder. The proposed RobustL2S framework simplifies training and brings robustness to variations in head-pose, ambient noise, and time-varying speaker characteristics, leading to significant gains in speech intelligibility. To validate the efficacy of our approach, we perform comprehensive experiments on GRID [9], TCD-TIMIT [10] and Lip2Wav [5] datasets. The quantitative measures and MOS scores show that the synthesized speech generated by our method accurately represents the intended content and improves on the intelligibility/naturalness compared to current state-of-the-art methods [11], [12] on all three datasets.

More formally, our work makes the following contributions: (1) We propose a novel modularized framework for Lip-to-Speech synthesis exploiting self-supervised embeddings for

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

both lip and speech sequences (2) A Seq2Seq network for cross-modal knowledge transfer to map lip SSL representations to speech SSL representations; and (3) Thorough experimental results demonstrating that RobustL2S is capable of synthesizing high-quality speech, achieving state-of-the-art results in objective and subjective evaluation without requiring additional data augmentation [11].

II. Related Work

A. Lip-to-Speech synthesis

1) Constrained Lip-to-Speech synthesis: Constrained lip-to-speech synthesis tackles speech generation from videos with limited vocabulary and minimal head movement [9], [10]. Ephrat *et al.*[13] introduced a CNN-based approach to predict Linear Predictive Coding features from silent talking videos. They later enhanced their model to a two-tower CNN-based encoder-decoder architecture [14], encoding raw frames and optical flows separately. [15] propose a combination of an autoencoder for extracting bottleneck features from audio spectrograms and a lipreading network comprising CNN, LSTM, and fully connected layers, for visual feature extraction. On the other hand, [16] utilizes a stochastic modeling approach employing a variational autoencoder. Other methods [17]–[20] employ GANs to synthesize speech from video frames. [21] train an attention-based encoder-decoder model to reconstruct speech from silent facial movement sequences without human annotations. [22] performs multi-modal supervision, using text and audio, to complement the insufficient word representations to reconstruct speech with correct contents from the input lip movements. [23] achieves zero-shot Lip-to-Speech synthesis using variational autoencoder to disentangle speaker and content information and a face identity encoder for unseen speakers.

2) Unconstrained Lip-to-Speech synthesis: Unconstrained lip-to-speech synthesis focuses on generating speech from real-world videos of individuals talking. This approach considers videos with an extensive vocabulary and significant head movements. Seminal work by Prajwal *et al.* [5] introduced a 3D-convolution plus autoregressive Seq2Seq model, adapted from Tacotron2 [24] for single speaker Lip-to-Speech synthesis. Their model generates mel-spectrograms based on the input video frames. In contrast, [25] employed a non-autoregressive architecture to expedite the inference process. It uses 3D-convolution blocks, a transformer condition module, and a Glow decoder module [26] for refining the mel-spectrograms. [27] expanded on [5] by training a transformer model to learn a joint latent distribution for speech generation. The VV-Memory architecture [28] combines audio and visual information using a key-value memory structure to enable video-to-speech reconstruction and speaker-independent speech retrieval. A recent work [11] proposed an end-to-end non-autoregressive transformer for synthesizing speech directly from unconstrained talking videos. Their approach involves a visual encoder, an acoustic decoder, a linear layer, and a GAN-based vocoder operating on Mel-spectrograms. LipSound2 [21] investigates cross-modal self-supervised pre-training of an encoder-decoder architecture with a location-aware attention mechanism to map face image sequences to mel-scale spectrograms. In contrast, our proposed RobustL2S differs by using content-rich SSL representations [29] and learning target speech representations from lip sequences.

B. Self-supervised representation for Lip-to-Speech synthesis

Traditional works on Lip-to-Speech synthesis encode lip or face sequences to hidden states, followed by a decoder to generate Mel-spectrograms. An independently trained vocoder transforms them into time-domain waves. However, these Mel frames are highly correlated along both time and frequency axes which may degrade the performance of entire Lip-to-Speech synthesis [30]. Moreover, these Mel-spectrograms have higher variance than that of quantized speech SSL representations increasing the complexity of training a Seq2Seq model. Due to its recent emergence, the utilization of SSL representation in Lip-to-Speech synthesis remains limited.

In VCVTS [12], vector quantized contrastive predictive coding (VQCPC) units are extracted from lip movements, and a speaker encoder, pitch predictor, and decoder are used to infer Mel frames. A separate voice conversion model and vocoder are employed for speaker representation learning and speech synthesis.

Revise [31] employs a combined audio-visual speech recognition module (P-AVSR - initialized with AV-HuBERT - an audio-visual SSL model) and a modified text-to-speech synthesis module (P-TTS) for generalized speech enhancement tasks. P-AVSR predicts discrete units derived from a self-supervised speech model, and P-TTS converts these units into speech using a modified HiFi-GAN [29] trained on the LJSpeech [32] dataset. Our work closely relates with [31] utilizing SSL representations for Lip-to-Speech synthesis. However, we utilize disentangled AV-HuBERT and HuBERT [33] features and a decoupled training procedure to train a Seq2Seq model. Additionally, the existing works have not yet fully capitalized on SSL representations for speaker-specific Lip-to-Speech generation, and our efforts aim to address this gap.

III. Method

A. Preliminaries

RobustL2S consists of four modules: two encoders - an encoder that extracts lip representation from video sequences and a pre-trained HuBERT model that generates speech features for corresponding speech sequences, a Seq2Seq model that maps lip representations to speech representations, and a modified HiFi-GAN vocoder that synthesize speech using the speech representations. We introduce four functions as follows:

- $f_{\mathrm{l}} : L^{T \times W \times H} \mapsto L_{\mathrm{ssl}}$, which maps the input lip sequence to its corresponding SSL representation. Here, T represents the number of time steps (frames), and H and W correspond to the spatial dimensions of the frames.
- $f_{\mathrm{s}} : X \mapsto S_{\mathrm{ssl}}$, which maps the ground-truth raw speech to its corresponding SSL representation.

- $f_{\text{s2s}} : L_{\text{ssl}} \mapsto S_{\text{ssl}}$, which maps the lip representation to its corresponding speech representation.
- $f_{\text{voc}} : S_{\text{ssl}} \mapsto \widehat{X}$, which maps the speech representation to the synthesized speech $\widehat{X}$.

B. Encoder

We use HuBERT [33] for f_{s}, to extract speech representation of target speech signals. HuBERT representation is content rich [29] and agnostic to other variations. Although our framework is compatible with various off-the-shelf SSL models, we specifically utilize AV-HuBERT [34] for f_{l}, our video encoder to extract lip representations. This choice is based on the similarity in training methods between AV-HuBERT and HuBERT, ensuring content-rich representations for improved intelligibility, which aligns with our goals. The HuBERT and AV-HuBERT models are trained using a masked-prediction loss to predict cluster IDs, which are learned using k-means clustering. The labels for the first iteration are derived by clustering MFCC features derived from acoustic frames. For subsequent iterations, more complex features derived from an audio or audio-visual encoder, depending on the specific model being used, are clustered. AV-HuBERT incorporates a modified ResNet [35], [36] as its frontend, coupled with a transformer encoder. We also finetune the pretrained AV-HuBERT using an attention-based Seq2Seq cross-entropy loss as in [34].

C. Seq2Seq model

In recent years, Seq2Seq models have gained significant attention in the field of cross-domain generation. The core concept of our approach is to align representations from two different domains - visual and audio, that share a common generating process. By recovering correspondences between these domains, we facilitate the transfer of knowledge from one domain to the other.

Our Seq2Seq model, denoted as f_{s2s}, adopts a non-autoregressive-based encoder-decoder architecture to map lip representations to their corresponding speech representations. The encoder and decoder consist of feed-forward transformer blocks with self-attention [37], along with 1-dimensional convolutions inspired by Fastspeech2 [38]. A transposed convolution layer is used at the encoder to match the rate of video and audio representations. The encoder takes the lip representation L_{ssl} and encodes it into a sequence of fixed-dimensional vectors. The decoder generates predictions for all representations of S_{ssl} simultaneously. We train three versions of the Seq2Seq model:

- $f_{\text{s2s-units}}$: This encoder-decoder architecture utilizes Cross-Entropy (CE) loss to train the model on the decoded speech units. The input to the architecture consists of cluster IDs from the video encoder, and the decoder predicts the corresponding HuBERT cluster IDs for the audio. The objective can be written as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} S_{\text{ssl_units_i}} \log(\hat{S}_{\text{ssl_units_i}}), \tag{1}$$

  where $S_{\text{ssl_units}}$ are the ground-truth speech units, $\hat{S}_{\text{ssl_units}}$ are the decoded speech units, and N are the number of HuBERT units.
- $f_{\text{s2s-features}}$: Here the model learns mapping from audio-visual feature to corresponding speech feature vectors. This model utilizes L1 loss, quantifying the difference between the decoded features and ground-truth speech features. The objective can be written as:

$$\mathcal{L}_{\text{L1}} = \frac{1}{T}\sum_{i=1}^{T} |S_{\text{ssl_features_i}} - \hat{S}_{\text{ssl_features_i}}|, \tag{2}$$

  where $S_{\text{ssl_features}}$ are the ground-truth speech features, $\hat{S}_{\text{ssl_features}}$ are the decoded speech features, and T is the time-steps.
- $f_{\text{s2s-features-ctc}}$: This architecture follows the same structure as $f_{\text{s2s-features}}$, but also includes an additional fully connected linear head to predict CTC tokens after the encoder layer. For given input lip representation $L_{\text{ssl}} \in \mathbb{R}^{TxD}$ of length T and dimension D, let Enc_{ssl} be the output of the encoder. The goal is to minimize the negative log-likelihood by using $P_{CTC}(S_{\text{ssl}}|Enc_{\text{ssl}})$ to train the model effectively using the CTC approach and is defined as:

$$\mathcal{L}_{CTC} := -\log P_{CTC}(S_{\text{ssl}}|Enc_{\text{ssl}}). \tag{3}$$

  By weighted summing the L1 and CTC loss functions, the objective function can be formulated as:

$$\mathcal{L}_{\text{Tot}} = \alpha_{CTC} * \mathcal{L}_{\text{CTC}} + \alpha_{L1} * \mathcal{L}_{\text{L1}}, \tag{4}$$

  where $\alpha_{\text{CTC}} \in \mathbb{R}$ and $\alpha_{\text{L1}} \in \mathbb{R}$ are the hyperparameter that balances the influence between two loss.

D. Speech Vocoder

We use a modified version of HiFiGAN-v2 [39] to synthesize speech. It has a generator G and a discriminator D. G runs S_{ssl} through transposed convolutions for upsampling to recover the original sampling rate followed by residual block with dilations to increase the receptive field to synthesize the signal, $\widehat{X} := G(S_{\text{ssl}})$.

The discriminator in our model has the task of distinguishing the synthesized signal $\widehat{X}$ from the original signal X. It is evaluated using two sets of discriminator networks. The multi-period discriminators operate on equally spaced samples of the signals, focusing on capturing temporal patterns and characteristics. On the other hand, the multi-scale discriminators analyze the input signal at different scales, enabling the model to capture both fine-grained details and global structure. The primary objective of the model is to minimize the discrepancy, measured by $D(X, \widehat{X})$, between the original signal and the synthesized signal. This optimization process applies to all the parameters of the speech decoder, improving its overall performance and fidelity.

IV. Experiments

A. Datasets

1) Lip2Wav: The Lip2Wav dataset [5] is a large, person-specific, unconstrained dataset, commonly used for learning Lip-to-Speech synthesis for individual speakers. It consists of real-world lecture videos featuring 5 different speakers. Each speaker has approximately 20 hours of video data, and the vocabulary size exceeds 5000 words for each speaker. We do experiments on all five speakers: Chess Analysis (chess), Chemistry Lectures (chem), Hardware Security (hs), Deep Learning (dl), and Ethical Hacking (eh).

2) GRID-4S: The GRID-4S is a subset of the GRID audio-visual dataset [9] specifically designed for constrained Lip-to-Speech synthesis. This subset includes two male speakers ($s1$, $s2$) and two female speakers ($s4$, $s29$), which are frequently used in the literature [5], [19]. The videos in the dataset were captured in an artificial environment. The vocabulary used in GRID-4S is limited to only 51 words. The sentences in the dataset follow a restricted grammar, with each sentence containing 6 to 10 words.

3) TCD-TIMIT-3S: The TCD-TIMIT-3S is a subset of the TCD-TIMIT dataset, which comprises recordings of 62 speakers captured under studio conditions. Among these speakers, three are trained lip-speakers. The primary objective of selecting this subset was to enable comparison with previous studies [5], [19]. Our focus was solely on the audio-visual data generated by these three lip-speakers. Each lip-speaker delivers 375 distinct sentences that exhibit phonetic diversity. Additionally, two sentences are spoken by all three lip-speakers.

B. Implementation details

1) Data preparation: For the GRID-4S and TCD-TIMIT-3S datasets, we adhere to the convention of randomly selecting 90% data for training, 5% for validation, and 5% for testing, as established in previous works [5], [11], [17], [40]. For Lip2Wav, we adopt the official data split [5]. For consistency, in line with previous works, we evaluated RobustL2S on the Lip2Wav dataset using a speaker-dependent setting [5], [14], [28]. This involved training the network separately with individual speakers. However, for the GRID-4S and TCD-TIMIT-3S datasets, we evaluated RobustL2S in a constrained (seen) speaker setting [5], [17], [40]. The video sequences are resampled to a frame rate of 25 frames per second (fps), while the raw audio is sampled at 16 kHz. We utilize the SFD [41] face detector to detect 68 key points, allowing us to crop a mouth-centered region-of-interest measuring 96×96 pixels. In order to solely assess the advantages of using SSL representation in our proposed setup, we opt not to employ any data augmentation techniques to enhance the quality of synthesized speech. The Lip2Wav dataset does not provide transcripts, so we rely on the Whisper *small* model [42] to extract transcripts. These transcripts are then used for finetuning the AV-HuBERT model.

2) SSL representation: We utilize the official fairseq repository implementation of the BASE models AV-HuBERT [34] and HuBERT [33] for our experiments. We fine-tune the AV-HuBERT pre-trained model with an attention-based STS cross-entropy loss for visual speech recognition [34]. To achieve this, a transformer decoder is added to the pre-trained model, which autoregressively decodes the AV-HuBERT features to target character probabilities. The fine-tuned AV-HuBERT model extracts SSL representations for lip sequences, while HuBERT is employed to extract representations from speech signals. Both models provide 768-dimensional features. For $f_{\text{s2s-units}}$ model, following the approach in [33], [34], the lip features are clustered into 2000 AV-HuBERT units, while the speech features are clustered into 100 HuBERT units, using k-means clustering. For $f_{\text{s2s-features}}$ model, the output features from HuBERT and AV-HuBERT models are used directly.

3) Seq2Seq model: Our model comprises a 6-layer transformer encoder and decoder with a hidden dimension of 512 and 2 attention heads. we set the batch size to 32 and the maximum number of steps to $20,000$. We employ the Adam optimizer with an initial learning rate of 4.4 x 10^{-2}, along with an annealing rate of 0.3 and annealing steps at $[3000, 4000, 5000]$. The HuBERT model encodes speech into features at a frame rate of 50 Hz, while the SSL unit from AV-HuBERT is encoded at 25 Hz. To match these rates, we incorporate a lightweight transposed convolution layer with a kernel size of $[4, 3]$ and a stride length of $[2, 1]$. We set α_{CTC} and α_{L1} mentioned in (4) to 0.001 and 1, respectively.

4) Speech Vocoder: We train modified HiFiGAN-v2[1] to generate audio from speech SSL representations, for all speakers. This model employs encoding of raw audio into a sequence of discrete tokens from a set of 100 possible HuBERT tokens, with a code hop size of 160 raw audio samples. We set the batch size to 16, the learning rate to $2\text{x}10^{-4}$, the number of embeddings to 100, the embedding dimension to 128, and the model input dimension to 256. We train the model up to $300k$ steps. Following the approach in [43], F0 is not used as a feature in our training process. The aforementioned vocoder configuration is effective for speech units. However, for our investigated feature-based models, $f_{\text{s2s-features}}$ and $f_{\text{s2s-features-ctc}}$, we apply a pre-trained k-means[2] clustering model trained on HuBERT features. During the inference phase, the generated features undergo k-means clustering to obtain discrete speech units, which are then passed through the speech vocoder. We train the vocoders for the three datasets separately.

C. Evaluation metric

During our evaluation, we employ several metrics to assess the quality of the synthesized speech. These include: Word Error Rate (WER), Short-Time Objective Intelligibility (STOI) [44], and Extended Short-Time Objective Intelligibility (ESTOI) [45]. Additionally, we conduct subjective evaluations using Mean Opinion Score (MOS), where human evaluators rate

[1]https://github.com/facebookresearch/speech-resynthesis

[2]https://github.com/facebookresearch/fairseq/tree/main/examples/textless_nlp/gslm/speech2unit

TABLE I
PERFORMANCE COMPARISON: SEQ2SEQ MODEL VS. EVALUATED VARIATIONS VS. NO SEQ2SEQ MODEL EMPLOYED ON CHEMISTRY SPEAKER OF LIP2WAV DATASET.

Baseline (Ours)	**STOI ↑**	**ESTOI ↑**
f_l(pre-trained) + f_{voc}	0.447	0.22
f_l(finetuned) + f_{voc}	0.50	0.27
f_l(finetuned) + $f_{\text{s2s-units}}$ + f_{voc}	0.18	0.013
f_l(**finetuned**) + $f_{\textbf{s2s-features}}$ + f_{voc}	**0.583**	**0.397**
f_l(finetuned) + $f_{\text{s2s-features-ctc}}$ + f_{voc}	0.557	0.368

the quality, intelligibility, and naturalness of the synthesized speech based on their subjective perception.

V. RESULTS

A. Need for Seq2Seq model

We tested our hypothesis of using a Seq2Seq model on the Lip2Wav dataset, specifically for a chemistry speaker, and report our findings in Table I. Fine-tuning the AV-HuBERT model using transcripts consistently improved objective metrics by approximately 0.05 units on both the metrics compared to the pre-trained version. Deploying the Seq2Seq model on the finetuned AV-HuBERT features (f_l(finetuned) + $f_{\text{s2s-features}}$ + f_{voc}) resulted in an increase of approximately 0.08 and 0.12 units in STOI and ESTOI metrics, respectively, compared to not using the Seq2Seq model (f_l(finetuned) + f_{voc}). These results highlight the effectiveness of our Seq2Seq approach using SSL representations for Lip-to-Speech synthesis. The significant performance gap (approximately 0.39 units on both metrics) between the Seq2Seq model using SSL features and the model using SSL units on both evaluated metrics approximates the amount of information lost in speech reconstruction when audio-visual sequences are represented as SSL units instead of SSL features. From now on, we will refer to f_l(finetuned) + $f_{\text{s2s-features}}$ + f_{voc} as RobustL2S. The inclusion of CTC loss in our f_l(finetuned) + $f_{\text{s2s-features-ctc}}$ model resulted in a statistically insignificant decrease of approximately 0.02 units in STOI compared to the model without CTC loss, f_l(finetuned) + $f_{\text{s2s-features}}$. The decrease may be due to the lack of ground-truth transcripts in the Lip2Wav dataset. However, when evaluating our model using CTC loss on datasets (GRID-4S and TCD-TIMIT-3S) with ground-truth transcripts, we observed a slight increase of 0.02 units. Nevertheless, our focus is on working with datasets in the wild that generally lack ground-truth transcripts, so we proceed with experiments excluding the CTC loss.

B. RobustL2S in Constrained settings

Table II and III summarizes the performance of our RobustL2S in the context of Lip-to-Speech synthesis using constrained datasets: GRID-4S and TCD-TIMIT-3S. We compare our results with existing Lip-to-Speech synthesis works, including state-of-the-art approaches. We report the mean test scores on all four speakers of the GRID-4S dataset and all three speakers of the TCD-TIMIT-3S dataset, as documented in previous works. Remarkably, our RobustL2S approach demonstrates significant improvements in terms of STOI and WER

TABLE II
PERFORMANCE COMPARISON IN CONSTRAINED-SPEAKER SETTING ON GRID-4S DATASET

Method	**STOI ↑**	**ESTOI ↑**	**WER ↓**
Vid2speech [13]	0.491	0.335	44.92 %
Lip2AudSpec [15]	0.513	0.352	32.51 %
1D GAN-based [17]	0.564	0.361	26.64 %
Vocoder-based [40]	0.648	0.455	23.33 %
Ephrat *et al.* [14]	0.659	0.376	27.83 %
Lip2Wav [5]	0.731	0.535	14.08 %
VAE-based [16]	0.724	0.540	-
VCA-GAN [19]	0.724	**0.609**	12.25 %
kim *et al.* [28], [46]	0.738	0.579	-
RobustL2S	**0.754**	0.571	**11.21** %

TABLE III
PERFORMANCE COMPARISON IN CONSTRAINED-SPEAKER SETTING ON TCD-TIMIT-3S DATASET

Method	**STOI ↑**	**ESTOI ↑**	**WER ↓**
Vid2speech [13]	0.451	0.298	75.52 %
Lip2AudSpec [15]	0.450	0.316	61.86 %
1D GAN-based [17]	0.511	0.321	49.13 %
Ephrat *et al.* [14]	0.487	0.310	53.52 %
Lip2Wav [5]	0.558	0.365	31.26 %
VCA-GAN [19]	0.584	0.401	-
RobustL2S	**0.596**	**0.452**	**29.03** %

metrics when compared to other approaches. This improvement is particularly noticeable on the TCD-TIMIT-3S dataset, which contains a larger number of novel words that were unseen during training. This observation highlights the ability of our RobustL2S to accurately pronounce new words and effectively capture semantic information from lip movements, resulting in the generation of more intelligible speech.

TABLE IV
PERFORMANCE COMPARISON IN SPEAKER-DEPENDENT SETTING ON LIP2WAV DATASET

Speaker	**Method**	**STOI ↑**	**ESTOI ↑**
Chemistry Lectures (chem)	Ephrat *et al.* [5]	0.165	0.087
	GAN-based [47]	0.192	0.132
	Lip2Wav [5]	0.416	0.284
	Hong et al. [28]	0.566	**0.429**
	RobustL2S	**0.583**	0.397
Chess Analysis (chess)	Ephrat *et al.* [5]	0.184	0.098
	GAN-based [47]	0.195	0.104
	Lip2Wav [5]	0.418	0.290
	Hong et al. [28]	0.506	0.334
	RobustL2S	**0.517**	**0.340**
Deep Learning (dl)	Ephrat *et al.* [5]	0.112	0.043
	GAN-based [47]	0.144	0.070
	Lip2Wav [5]	0.282	0.183
	Hong et al. [28]	0.576	0.402
	RobustL2S	**0.627**	**0.419**
Hardware Security (hs)	Ephrat *et al.* [5]	0.192	0.064
	GAN-based [47]	0.251	0.110
	Lip2Wav [5]	0.446	0.311
	Hong et al. [28]	0.504	0.337
	RobustL2S	**0.511**	**0.337**
Ethical Hacking (eh)	Ephrat *et al.* [5]	0.143	0.064
	GAN-based [47]	0.171	0.089
	Lip2Wav [5]	0.369	0.220
	Hong et al. [28]	0.463	**0.304**
	RobustL2S	**0.493**	0.277

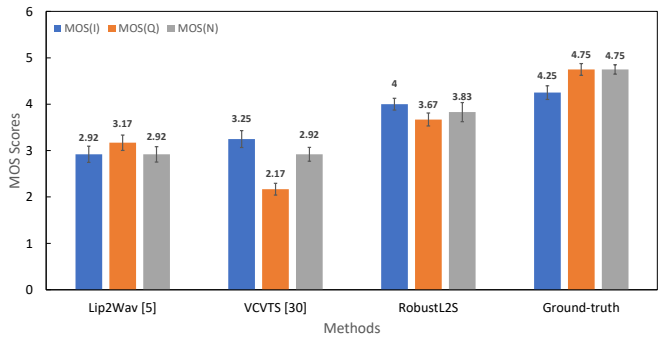


Fig. 2. MOS scores on Intelligibility, Quality, and Naturalness with their 95 % confidence interval computed from their t-distribution on GRID-4S dataset

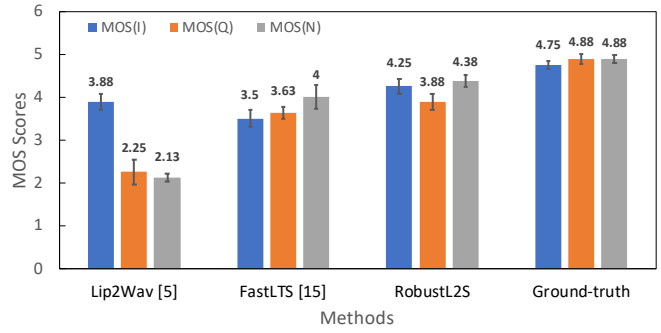


Fig. 3. MOS scores on Intelligibility, Quality and Naturalness with their 95 % confidence interval computed from their t-distribution on Lip2Wav dataset

C. RobustL2S in Unconstrained settings

Table IV provides a synopsis of RobustL2S's performance on the Lip2Wav dataset. This dataset includes a significant amount of silences between words, and RobustL2S shows a notable improvement across all metrics. Despite Lip2Wav's data asynchrony issues, which may affect the quality of the generated speech, RobustL2S demonstrates substantial performance gains in objective metrics, highlighting its overall superiority in producing intelligible speech. However, it is worth noting that RobustL2S performs similarly or slightly worse than [28] on the *hs* and *eh* ESTOI metric. This could potentially be attributed to the poor resolution ($480p$ and $360p$) of the original videos, making it challenging to accurately recognize lip regions.

D. Subjective evaluation

Fig. 2 and Fig. 3 shows the MOS scores on intelligibility (MOS(I)), quality (MOS(Q)), and naturalness (MOS(N)) of synthesized speech from evaluated methods on Grid-4S and Lip2Wav datasets. We requested ten English-proficient subjects to score five randomly selected samples from different methods on the Lip2Wav and GRID-4S datasets. It can be observed that our model outperforms the evaluated methods, exhibiting higher Mean Opinion Score (MOS) values. This demonstrates that the proposed approach inherits the advantages of disentangled SSL features and the mapping of lip sequences to content-specific information. As a result, our model not only inherently improves the intelligibility aspect of synthesized speech but also generates speech that is highly natural and of high quality.

VI. Conclusions

We propose a novel framework for Lip-to-Speech system, called RobustL2S, which accurately synthesizes spoken content from silent videos. This is accomplished by utilizing a non-autoregressive based sequence-to-sequence model to establish an inter-modality mapping, allowing us to learn a suitable decoding space from the lips' self-supervised (SSL) representations. We further demonstrate the effectiveness of mapping SSL features rather than SSL units for synthesizing intelligible speech. Both quantitative and qualitative results showcase state-of-the-art performance in constrained settings (such as GRID and TCD-TIMIT) and unconstrained settings (like Lip2Wav). In our future work, we aim to introduce emotive effects in the synthesized speech, considering that HuBERT embeddings are known to lack prosody information. Additionally, we plan to explore diffusion-based speech vocoders and their application in a multi-lingual setup.

References

[1] Z. Su, S. Fang, and J. Rekimoto, "Liplearner: Customizable silent speech interactions on mobile devices," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[2] G. Cong, L. Li, Y. Qi, *et al.*, "Learning to dub movies via hierarchical prosody models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 687–14 697.

[3] R. El-Bialy, D. Chen, S. Fenghour, *et al.*, "Developing phoneme-based lip-reading sentences system for silent speech recognition," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 129–138, 2023.

[4] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 283–292.

[5] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 796–13 805.

[6] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[7] L. E. Bernstein, N. Jordan, E. T. Auer, and S. P. Eberhardt, "Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training," *American Journal of Audiology*, vol. 31, no. 2, pp. 453–469, 2022.

[8] J. Choi, M. Kim, and Y. M. Ro, "Intelligible lip-to-speech synthesis with speech units," *arXiv preprint arXiv:2305.19603*, 2023.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[10] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[11] Y. Wang and Z. Zhao, "Fastlts: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5678–5687.

[12] D. Wang, S. Yang, D. Su, X. Liu, D. Yu, and H. Meng, "Vcvts: Multi-speaker video-to-speech synthesis via cross-modal knowledge transfer from voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7252–7256.

[13] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 5095–5099.

[14] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 455–462.

[15] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 2516–2520.

[16] R. Yadav, A. Sardana, V. P. Namboodiri, and R. M. Hegde, "Speech prediction in silent videos using variational autoencoders," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7048–7052.

[17] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," *arXiv preprint arXiv:1906.06301*, 2019.

[18] R. Mira, K. Vougioukas, P. Ma, S. Petridis, B. W. Schuller, and M. Pantic, "End-to-end video-to-speech synthesis using generative adversarial networks," *IEEE Transactions on Cybernetics*, 2022.

[19] M. Kim, J. Hong, and Y. M. Ro, "Lip to speech synthesis with visual context attentional gan," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2758–2770, 2021.

[20] R. Mira, A. Haliassos, S. Petridis, B. W. Schuller, and M. Pantic, "Svts: Scalable video-to-speech synthesis," *arXiv preprint arXiv:2205.02058*, 2022.

[21] L. Qu, C. Weber, and S. Wermter, "Lipsound2: Self-supervised pre-training for lip-to-speech reconstruction and lip reading," *IEEE transactions on neural networks and learning systems*, 2022.

[22] M. Kim, J. Hong, and Y. M. Ro, "Lip-to-speech synthesis in the wild with multi-task learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[23] Z.-Y. Sheng, Y. Ai, and Z.-H. Ling, "Zero-shot personalized lip-to-speech synthesis with face image based voice control," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[24] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[25] J. He, Z. Zhao, Y. Ren, J. Liu, B. Huai, and N. Yuan, "Flow-based unconstrained lip to speech generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 843–851.

[26] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[27] M. Varshney, R. Yadav, V. P. Namboodiri, and R. M. Hegde, "Learning speaker-specific lip-to-speech generation," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 491–498.

[28] J. Hong, M. Kim, S. J. Park, and Y. M. Ro, "Speech reconstruction with reminiscent sound via visual voice memory," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3654–3667, 2021.

[29] A. Polyak, Y. Adi, J. Copet, *et al.*, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[30] C. Du, Y. Guo, X. Chen, and K. Yu, "Vqtts: High-fidelity text-to-speech synthesis with self-supervised vq acoustic feature," *arXiv preprint arXiv:2204.00768*, 2022.

[31] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 795–18 805.

[32] K. Ito and L. Johnson, *The lj speech dataset*, https://keithito.com/LJ-Speech-Dataset/, 2017.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[34] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 513–520.

[37] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[39] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[40] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, "Vocoder-based speech synthesis from silent videos," *arXiv preprint arXiv:2004.02541*, 2020.

[41] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[42] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *OpenAI Blog*, 2022.

[43] A. Lee, P.-J. Chen, C. Wang, *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[44] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[45] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[46] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Multi-modality associative bridging through memory: Speech sound recollected from face video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 296–306.

[47] D. Burnham, R. Campbell, G. Away, and B. Dodd, *Hearing eye II: the psychology of speechreading and auditory-visual speech*. Routledge, 2013.

Realizing Nipple in Profile Recognition and Nipple Detection Using a Single Classification

Yi-Chong Zeng
Institute for Information Industry, Taiwan, R.O.C.
E-mail: yichongzeng@iii.org.tw Tel: +886-66072958

***Abstract* —Nipple in Profile (NIP) is one of the positioned criteria for mammography quality evaluation. The current evaluation work relies on manual reviewing, which is subjective and time-consuming. This paper introduces a novel NIP recognition based on a convolutional neural network (CNN) without referring to prior information, which outputs an objective recognition result in real-time. Amazingly, the added value associated with the proposed method is nipple detection which analyzes the extracted gradient-weighted class activation mapping (Grad-CAM) based on three-category classification. In this work, the defined categories are Nipple in Profile, Nipple not in Profile, and non-nipple. For the NIP recognition, we only exploit the comparison between the Grad-CAMs of Nipple in Profile and Nipple not in Profile. Our method is more creative than the typical classification approaches to realize multiple tasks with a single classification. The experiment results demonstrate that the three-category classification with Densenet-121 achieves better performance in both NIP recognition and nipple detection.**

I. INTRODUCTION

Mammography uses a low-dose X-ray to generate projected images and is a common way of screening breast lesions. Then, approved radiologists screen mammograms for detecting suspicious findings, e.g., mass, calcifications, architecture distortion, and asymmetry. When radiologists have difficulty screening inadequate-quality mammograms, hospitals recall patients to retake mammograms. However, it's been a while since the last shot. Furthermore, the retaking task becomes costly and time-consuming for patients living in remote areas. To solve such problems, radiographers need a skill in quality evaluation, and they decide whether retake mammograms depending on their domain knowledge and experiences. Up to now, mammography quality evaluation still relies on manual reviewing. In [1, 2], the documents list evaluation criteria defined over four quality levels: Perfect, Good, Moderate, and Inadequate. For example, one of the positioned criteria addressed in [3-10], the difference of posterior nipple lines (PNL) on Craniocaudal (CC)-view and Mediolateral-Oblique (MLO)-view mammograms at the same laterality should be less than 1 cm.

Recently, researchers developed deep learning techniques for assessments of the mammogram positioned quality. In [11], Watanabe et al. proposed positioned quality classifications for inframammary fold and nipple profile based on the deep convolutional neural network (DCNN). In the previous works, we presented convolutional neural network (CNN)-based mammography quality evaluation methods for skinfolds [12] and inframammary fold [13]. Those two methods aimed at the criteria of absence of skin folds and well-demonstrated inframammary fold addressed in [1, 2, 9]. Brahim et al. [14] analyzed the nipple position, pectoralis muscle angle and level using the self-designed CNN models. To highlight the relevance of the model prediction, they extracted gradient-weighed class activation mapping (Grad-CAM) for nipple position detection and then determined whether the nipple was in profile. However, their method downscaled mammograms resulting in small nipple vanishing and misdetermined the nipple was not in profile.

This paper presents a CNN-based method to recognize Nipple in Profile (NIP). The NIP is one of the positioned criteria for mammography quality evaluation, which associates with retro-areolar tissue well-separated. Our method performs NIP recognition without prior information, such as nipple position. Amazingly, the added value associated with the proposed method is a nipple detector, which analyzes the extracted Grad-CAM based on a three-category classification. In this work, the three categories represent Nipple in Profile, Nipple not in Profile, and non-nipple. The samples extracted from the breast area and background belong to the non-nipple.

The training dataset consists of 224×224 samples annotated with one of three labels in advance. During the training process, we train three-category classifiers by seven CNN models, including, Squeezenet, Densenet-121, Resnet-18,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Resnet-101, Alexnet, Vgg-16, and Vgg-19. During the inference process, the trained model applies to entire mammograms. For nipple detection, our method analyzes the Grad-CAMs of all categories. For NIP recognition, we only exploit the comparison between the Grad-CAMs of Nipple in Profile and Nipple not in Profile. Therefore, the proposed method realizes NIP recognition and nipple detection using a single classification model. The main contributions of our work can be summarized as follows:

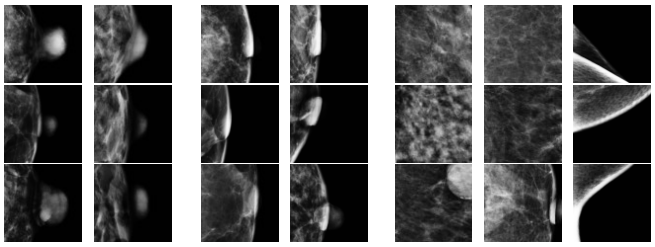
- We construct a CNN-based three-category classification for NIP recognition that performs without referring to prior information.
- We implement nipple detection based on the constructed classification by analyzing the Grad-CAMs.
- Extensive experiments verify the effectiveness and efficiency of the constructed classifications, and Densenet-121 is the better choice by summarizing the performance of NIP recognition, nipple detection, and computation complexity.

The rest of this paper is organized as follows: the related work and the proposed method are introduced in Sections II and III, respectively. The experiment results are revealed in Section IV, and the conclusion is drawn in Section V.

II. The Related Work

A. Nipple Detection

There were plenty of nipple detection approaches investigated for various applications, such as mammography quality evaluation [13-19]. However, those approaches were workable under the condition of a visible nipple which presents a unique shape. For example, Lin et al. [15] used Digital Database for Screening Mammography (DDSM), and nipples were in the profile of the attained mammograms. Another researcher developed techniques to estimate invisible nipples. In [16], Jas et al. estimated the farthest pixel from the base to the breast boundary and then found the position of an invisible nipple. In [17], Chae et al.'s method analyzed the fibroglandular tissue. In [13, 18], the previous work computed the curvature values of pixels on the breast boundary, and the pixel with the highest curvature value determined the position of a visible nipple. Moreover, we assumed that an invisible nipple overlapped breast tissue as a superposed area, and this area presented brighter intensity than the surrounding pixels on mammograms. Therefore, in [18], the previous work employed U-Net to detect a bright-intensity region and treated the detection result as a superposed area.



(a) C_0 (b) C_1 (c) C_2

Fig.1. Samples of (a) Nipple in Profile, (b) Nipple not in Profile, and (c) non-nipple.

B. Nipple in Profile

Breast nipple is a significant reference for mammography quality evaluation. For example, for an adequately positioned breast [1, 2, 19], the difference between PNLs on CC-view and MLO-view mammograms should be within 1cm. The PNL is a line started at the posterior nipple perpendicularly towards a pectoral muscle. Furthermore, the Nipple in Profile is the positioned criterion associated with retro-areolar tissue well-separated. Fig.1(a) and Fig.1(b) illustrate the samples with NIP and Nipple not in Profile (NnIP), respectively.

In [10], Bassett et al. found a large proportion of mammograms with NIP in their experiments. They claimed NIP is a secondary consideration for positioning a breast in mammography. However, while nipples were unseen in both CC-view and MLO-view mammograms, a third-view mammogram was required in screening mammography. Similarly, in [7], Huppe et al. presented a report about mammography against various positioning criteria for full-field digital mammography and digital breast tomosynthesis. They showed results that nipples were visible in the profiles of most mammograms. In [20], Holen et al. concluded that no evidence supports the correlation between NIP and odds of screen-detected breast cancer and recall. However, they did not confirm whether screening mammography encounters a negative effect while the criterion of NIP is removed or relaxed.

In summary, there is no direct evidence of whether removing the criterion of NIP negatively influences mammogram screening. However, occurring unseen nipples in both CC-view and MLO-view mammograms can be costly in mammogram screening. Consequently, NIP analysis is essential for mammography quality evaluation.

III. The Proposed Method

A. Problem Statement

NIP recognition determines whether a nipple is in profile or not in a mammogram. An intuitional way is to analyze a

detected nipple sample by a two-category classification. However, an inevitable issue is that the efficiency of nipple detection affects the performance of NIP recognition. As we mentioned in Section II, most nipple detection approaches were available for visible nipples and had difficulty detecting invisible/inverted nipples. There is a potential risk in nipple detectors. For risk avoidance, the proposed method recognizes NIP based on an entire mammogram rather than a nipple sample.

B. Three-category Classification

The proposed method implements a three-category classification rather than a two-category one. The three categories represent Nipple in Profile (denoted as C_0), Nipple not in Profile (denoted as C_1), and non-nipple (denoted as C_2). Fig.1 shows the training samples of those categories. The non-nipple consists of samples extracted from the breast area and background, including fibroglandular tissue, fatty tissue, pectoralis muscle, and breast boundary.

During the training process, we train three-category classifiers by seven CNN models, including, Squeezenet, Densenet-121, Resnet-18, Resnet-101, Alexnet, Vgg-16, and Vgg-19. The training dataset consists of 224×224 samples in three categories. During the inference process, the trained model applies to an entire mammogram. Subsequently, the method extracts Grad-CAMs over categories and estimates an index image. Let $s_i(x,y)$ be the score of the Grad-CAM at the coordinate (x,y) in C_i, where $i=\{0, 1, 2\}$. The index $d(x,y)$ denotes the category corresponding to the maximum score at (x,y) estimated by,

$$d(x,y) = \arg\max_{i=\{0,1,\ldots,N\}} s_i(x,y) \tag{1}$$

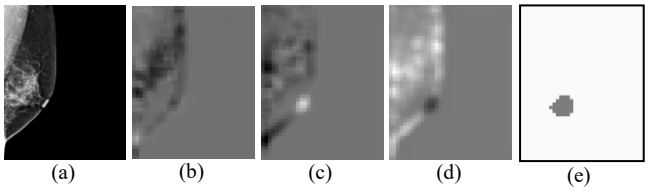
For instance, Fig. 2 shows a mammogram and its Grad-CAMs. The bright-tonal pixel represents a high score in the Grad-CAMs of Figs.2(b)-(d). Fig.2(e) is the index image constructed by (1) with N=3, and the black, gray, and white pixels in the index image represent C_0, C_1, and C_2, respectively.

C. Nipple Detection and NIP Recognition

Analyzing the Grad-CAM in C_2, we identify the pixels with higher scores in C_2 than those in C_0 and C_1 as a portion of non-nipple. Therefore, excluding the pixels corresponding to C_2 in the index image, the rest constitutes candidates of nipple areas. The proposed method determines the detected nipple area (denoted as A) according to,

$$\begin{aligned} \mathrm{A} &= \Omega_{j^*}, \\ j^* &= \arg\max_{j=\{0,1,\ldots,N_C\}} \left\{ \arg\max_{i=\{0,1\}} \left(s_i(x,y) \middle| (x,y) \in \Omega_j \right) \right\}, \end{aligned} \tag{2}$$

where Ω_j represents the jth candidate, and N_c is the total number of candidates. In Fig.2(e), the gray pattern is the detection result. In the case of Fig.2, the mammogram is NnIP by manual reviewing. Analyzing the Grad-CAMs in C0 and C_1, the nipple area in C_1 (as shown in Fig.2(c)) expresses brighter tones than that in C_0 (as shown in Fig.2(b)). It signifies that our method determines that the nipple is not in profile in the mammogram.



(a) (b) (c) (d) (e)

Fig.2. Three-category classification: (a) a testing mammogram with Nipple not in Profile, the Grad-CAMs in (b) C_0, (c) C_1, (d) C_2, and (e) the index image

D. Performance Evaluation

This work utilizes ten measures to evaluate the performance of the proposed method and the compared ones. Those measures correspond to three aspects: the accuracy of NIP recognition, the efficiency of nipple detection, and the computational complexity of a classification model. In NIP recognition, we compute sensitivity (SEN), specificity (SPC), accuracy (ACC), and Cohen's kappa coefficient [21] based on a confusion matrix. Landis and Koch characterized the kappa coefficient into six magnitudes guidelines [22]. The negative coefficient represents no agreement, and the kappa coefficient ranging from 0.8 to 1 is an almost perfect agreement for the model. In nipple detection, two measures, Intersection over Union (IoU) and Dice coefficient (Dice), compute the similarity of nipple area between the ground truth and the detection result.

The parameter size of a classification model and the processing time of an inference process are two bases. Then, we compute the size-reducing rate (SRR) and time-reducing rate (TRR) based on those two bases for evaluating computational complexity, which is formulated as,

$$SRR_k = \frac{(\max_i V_i) - V_k}{(\max_i V_i)}, \tag{3}$$

$$TRR_k = \frac{(\max_i T_i) - T_k}{(\max_i T_i)}, \tag{4}$$

where Vk and Tk denote as the parameter size and the processing time, respectively. The subscript k represents a CNN model, that is k={Squeezenet, Densenet-121, Resnet-18, Resnet-101, Alexnet, Vgg-16, Vgg-19}.

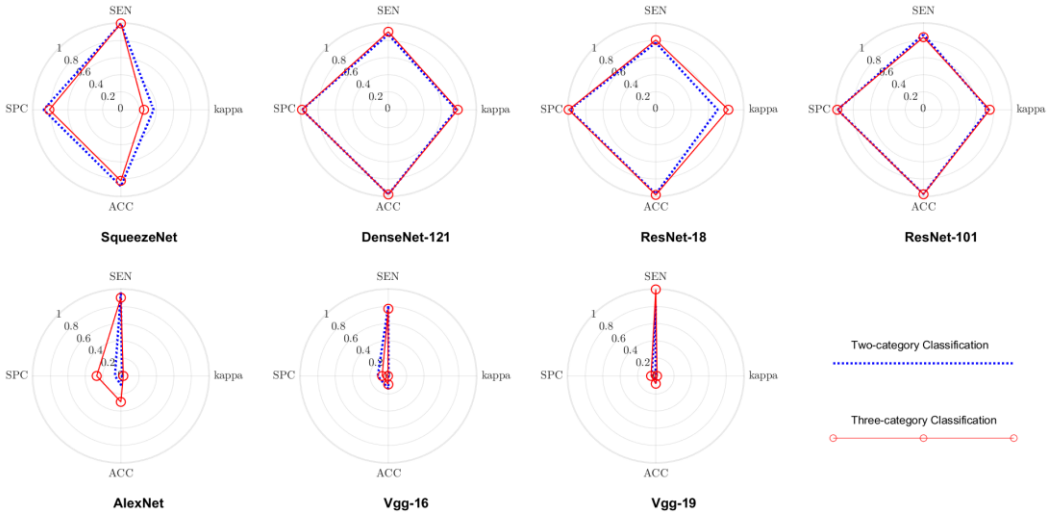


Fig.3. Radar charts against four measures for two-category and three category classifications with seven CNN models for NIP recognition

Additionally, a radar chart is an approach to visualize the performance against eight measures for the seven CNN models. The selected eight measures include SEN, SPC, ACC, kappa, IoU, Dice, SRR, and TRR. The radar chart assists us to choose the best combination according to the aspects.

IV. The Experiment Results

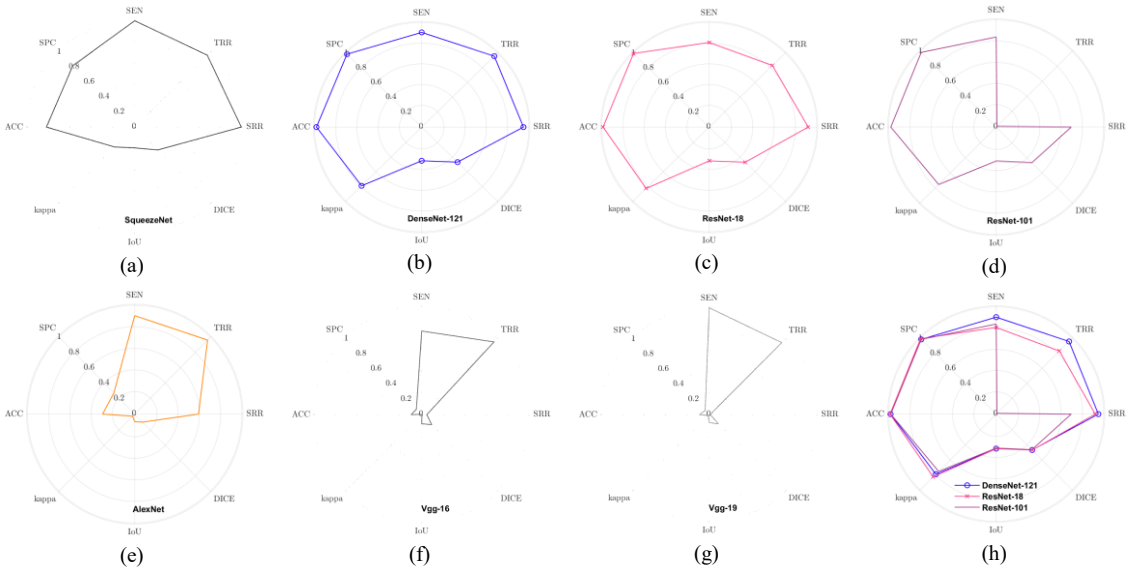
In the experiments, we emphasized two tasks: NIP recognition and nipple detection. The proposed method implemented multi-category classifications with various CNN models for finding the proper combination. The training dataset consisted of 224×224 samples extracted from 1004 CC-view and 1006 MLO-view mammograms. The testing dataset consisted of 392 CC-view and 394 MLO-view mammograms. All programs run on a personal computer with NVIDIA GeForce GTX 1660 and RAM 32 GB.

A. NIP Recognition

In the first experiment, two-category and three-category classifications with seven CNN models were implemented, including Squeezenet, Densenet-121, Resnet-18, Resnet-101, Alexnet, Vgg-16, and Vgg-19. This work defined three categories: Nipple in Profile (C_0), Nipple not in Profile (C_1), and non-nipple (C_2). The CC-view training dataset consisted of 1896, 1873, and 3042 samples in C_0, C_1, and C_2. Similarly, the MLO-view training dataset consisted of 2889, 2958, and 7098 samples in C_0, C_1, and C_2. During the model training process, two-category classification only utilized the datasets of C_0 and C_1. During the recognition process, we extracted the Grad-CAMs in C_0 and C_1 and then determined the category of NIP. Summarizing the recognition results of CC-view and MLO-view mammograms, Fig.3 illustrates radar charts against four measures for the two-category and three-category classifications with CNN models, and those measures are sensitivity (SEN), specificity (SPC), accuracy (ACC), and Cohen's kappa coefficient (kappa). Compared two classifications with the same model, the performance were similar in Densenet-121 and Resnet-101. Furthermore, three-category classifications achieved better performance than two-category ones with CNN models, except for Squeezenet and Vgg-16. The Densenet-121, Resnet-18, and Resnet-101 were more outstanding than the others performed in three-category classification, and the kappa coefficients were 0.791, 0.827, and 0.754, respectively. According to [21], the three-category classification with Resnet-18 was almost perfect, and the classifications with Densenet-121 and Resnet-101 were substantial.

B. Nipple Detection

Two-category classification got worse detection results in C1 due to misrecognition. The texture of non-nipple is similar to the samples in C_1 but different from the ones in C_0. Therefore, the two-category classifier recognized the non-nipple around a nipple as C_1 rather than C_0. It resulted in inaccurate detection results. For example, applying a two-category classification to Fig.2(a), the extracted Grad-CAMs in C_0 and C_1 were respectively in Figs.4(a) and 4(b), and the index image was in Fig.4(c). The black and white pixels in the index image represent C_0 and C_1, respectively. Therefore, Fig.4(c) expresses the misrecognition results in non-nipple in two-category classification and shows none of the nipple areas. We computed IoU and Dice between the ground truth and the detected nipple area. TABLE I lists the measures of nipple area detection using three-category



(a) (b) (c) (d) (e) (f) (g) (h)

Fig.5. Radar charts against eight measures for (a) Squeezenet, (b) Densenet-121, (c) Resnet-18, (d) Resnet-101, (e) Alexnet, (f) Vgg-16, (g) Vgg-19, and (h) the superposition of (b), (c), and (d).

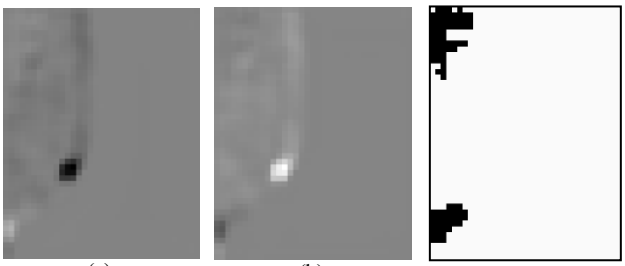
classifications with CNN models. Resnet-18, Densenet-121, and Resnet-101 had the similar performance.

C. Computation Complexity

We further considered the computation complexity of three-category classifications by calculating two self-defined measures: SRR and TRR. Vgg-19 required the largest parameter size among all models, and Resnet-101 took much time in the inference process. Therefore, the SRR of Vgg-19 and the TRR of Resnet-101 were zeros. TABLE II lists the SSR and TRR of CNN models.

D. Combination Selection

Figs.5(a)-(g) illustrate the radar charts against eight measures for seven CNN models, and those measures are SEN, SPC, ACC, kappa, IoU, Dice, SRR, and TRR. Summarized the experiment results among NIP recognition, nipple detection, and computation complexity, the top 3 models were Densenet-121, Resnet-18, and Resnet-101 in order. Fig.5(h) shows a superposition of the radar charts of those models. Densenet-101 and Resnet models were neck and neck in NIP recognition and nipple detection. Densenet-101 needed less computation complexity in the inference process, especially processing time. Consequently, a three-category classification with Densenet-121 is the best choice for this work. In NIP recognition, this combination achieved an accuracy of 0.982 and a kappa coefficient of 0.791. In nipple detection, the Dice coefficient and IoU were 0.468 and 0.315, respectively. The average time in mammogram inference was 2.75 sec.



(a) (b) (c)

Fig.4. Two-category classification: the Grad-CAMs in (a) C_0, (b) C_1, and (c) the index image

TABLE I. Measures of nipple area detection using three-category classifications with seven CNN models

	Squeezenet	Densenet-121	Resnet-18	Resnet-101	Alexnet	Vgg-16	Vgg-19
IoU	0.187	0.315	**0.316**	0.31	0.06	0.076	0.066
Dice	0.297	0.468	**0.469**	0.464	0.095	0.123	0.113

TABLE II. SSR and TRR of the seven CNN models

	Squeezenet	Densenet-121	Resnet-18	Resnet-101	Alexnet	Vgg-16	Vgg-19
SRR	**0.995**	0.949	0.920	0.694	0.592	0.038	0.000
TRR	0.957	0.958	0.830	0.000	**0.961**	0.954	0.953

V. CONCLUSION

This paper presents the CNN-based 'Nipple in Profile' recognition method using a three-category classification. The proposed method is implemented without prior nipple information. By contrast, the added value associated with our method is nipple detection by analyzing the extracted Grad-CAMs. We experimented with multi-category classifiers

integrated with seven CNN models, and Densenet-121 was the best choice. The experiment results demonstrated that the proposed method was superior to the other models. It achieved an accuracy of 0.982 and a kappa coefficient of 0.791 in NIP recognition; the Dice coefficient and IoU were respectively 0.468 and 0.315 in nipple detection. Our method is more creative than the typical classification approaches to realize multiple tasks with a single classification.

REFERENCES

[1] Supporting Women with Breast Cancer – Today and Every Day, Chapter 3: Radiography, *Westmead Breast Cancer Institute*, 2016. https://www.bci.org.au/wp-content/uploads/2016/10/Radiography-Review-2016-v3.pdf

[2] PGMI evaluation system, Radiopaedia. https://radiopaedia.org/articles/pgmi-evaluation-system

[3] S. Deepthi, and R. Kavitha, "Breast cancer diagnosis using both MLO and CC view mammogram", *IRJMETS*, vol.2, no.9, pp.299-306, Sept. 2020.

[4] I. T. Nakano, G. L. D. O. Salvador, H. R. Schelin, V. Denyak, H. J. Khoury, and S. Tacara, "Image evaluation and breast density categories as a function of mammary positioning in full-field digital mammography", *Acta Radiologica*, vol.61, no.7, pp.868-874, Nov. 2019.

[5] M.B. Popli, R. Teotia, M. Narang, and H. Krishna, " Breast positioning during mammography: mistakes to be avoided", *Breast cancer: basic and clinical research*, 8, BCBCR-S17617, pp.119-124, 2014.

[6] S. Pal, D. M. Ikeda, R. A. Jesinger, L. J. Mickelsen, C. A. Chen, D. B. Larson, "Improving performance of mammographic breast positioning in an academic radiology practice", *American Journal of Roentgenology (AJR)*, vol.210, no.4, pp.807-815, April 2018.

[7] A. I. Huppe, K. L. Overman, J. B. Gatewood, J. D. Hill, L. C. Miller, and M. F. Inciardi, "Mammography positioning standards in the digital era: is the status quo acceptable?", *American Journal of Roentgenology (AJR)*, vol.209, no.6, pp.1419-1425, 2017.

[8] R.-J. I. Sweeney, S. J Lewis, P. Hogg, and M. F. McEntee, "A review of mammographic positioning image quality criteria for the craniocaudal projection", *The British Journal of Radiology (BJR)*, vol.91, 1082, Feb. 2018.

[9] Mammography accreditation clinical image review sheet, *American College of Radiology*, Jan. 2020. https://www.acraccreditation.org/-/media/ACRAccreditation/Documents/Mammography/Clinical-Image-Review-Sheet---MAP.pdf

[10] L. W. Bassett, I. A. Hirbawi, N. DeBruhl, and M. K. Hayes, "Mammographic positioning: evaluation from the view box", *Radiology*, vol.188, no.3, pp.803-806, 1993.

[11] H. Watanabe, S. Hayashi, Y. Kondo, E. Matsuyama, N. Hayashi, T. Ogura, and M. Shimosegawa, "Quality control system for mammographic breast positioning using deep learning", *Research Square*, pp.1-17, 2022. https://assets.researchsquare.com/files/rs-2288325/v1/b118cdb0-9cd8-49a2-89c3-827fa803a6a7.pdf?c=1669736584

[12] Y.-C. Zeng, and Y.-H. Chen, "Skinfold recognition for positioning quality evaluation on mammography", *IEEE GCCE*, pp.890-892, Osaka, Japan, Oct. 18-21, 2022.

[13] Y.-C. Zeng, Y.-C. Wu, C.-Y. Yeh, S.-C. Li, T.-H. Chou, Y.-W. Huang, G.-C. Hsu, and H.-H. Hsu, "Mammography quality evaluation and model interpretation based on CNN-based inframammary fold classification", *APSIPA ASC*, pp.1259-1264, Chiang Mai, Thailand, Nov. 7-10, 2022.

[14] M. Brahim, K. Westerkamp, L. Hempel, R. Lehmann, D. Hempel, and P. Philipp, "Automated assessment of breast positioning quality in screening mammography", *Cancers*, vol.14, no.19, 4704, pp.1-27, 2022.

[15] Y. Lin, M. Li, S. Chen, L. Yu, and F. Ma, "Nipple detection in mammogram using a new convolutional neural network architecture", *CISP-BMEI*, pp.1-6, Oct. 2019.

[16] M. Jas, S. Mukhopadhyay, J. Chakraborty, A. Sadhu and N. Khandelwal, "A heuristic approach to automated nipple detection in digital mammograms", *Journal of Digital Imaging*, vol. 26, no. 5, pp. 932-940, Oct. 2013.

[17] S.-H. Chae, J.-W. Jeong, J.-H. Choi, E. Y. Chae, H. H. Kim, Y.-W. Choi, et al., "Fully automated nipple detection in digital breast tomosynthesis", *Computer Methods and Programs in Biomedicine*, vol. 143, pp. 113-120, May 2017.

[18] Y.-C. Zeng, "Automatic detections of nipple and pectoralis major in mammographic images", *IEEE ICCE-TW*, pp.1-2, May 2018

[19] Posterior Nipple Line, https://radiopaedia.org/articles/posterior-nipple-line

[20] Å. S. Holen, M. Larsen, N. Moshina, G. G. Wåade, I. Sechopoulos, B. Hanestad, L. Tøsdal, and S. Hofvind, "Visualization of the Nipple in Profile: Does It Really Affect Selected Outcomes in Organized Mammographic Screening?", *Journal of Breast Imaging*, vol.3, no. 4, pp.427-437, July/August 2021.

[21] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, vol.20, no.1, pp.37–46, 1960.

[22] J.R. Landis; and G.G. Koch, "The measurement of observer agreement for categorical data", *Biometrics*, vol.33, no.1, pp.159–174, 1977.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Exploring a CLIP-Enhanced Automated Approach for Video Description Generation

Siang-Ling Zhang, Huai-Hsun Cheng, Yen-Hsin Chen, Mei-Chen Yeh
Department of Computer Science and Information Engineering
National Taiwan Normal University, Taiwan
E-mails: {sss36592580,zsisc.huaish,yenhsinchen09}@gmail.com
myeh@csie.ntnu.edu.tw

***Abstract*—Visual storytelling is a learned ability that humans have developed through the course of their evolution. In contrast to written records and descriptions, people are now sharing glimpses of their lives through short videos. Human conversations largely rely on visual and auditory inputs, followed by corresponding feedbacks. For machines, the conversion of images to texts serves as a bridge between visual and linguistic information, enabling machine-human interactions more naturally. Video captioning—involving automatically generating textual descriptions from videos—is one of the core technologies that enable such applications. In this work, we present CLIP-CAP, an automatic method for transforming visual contents to concise textual descriptions. We investigate the CLIP pretraining model as well as its potential in this task. Through experiments on the ActivityNet Captions dataset, we show that the proposed CLIP-CAP model outperforms existing video captioning methods in terms of several different metrics.**

I. Introduction

With the rapid development of information technology, robots have become an integral part of our lives, enhancing productivity and convenience. Powered by Artificial Intelligence (AI), robots have the potential to capture the surrounding environment through cameras and provide descriptive text. Video captioning—which automatically translates video clips into natural language sentences—enables natural communication between humans and machines. This technology can be leveraged to enhance human-robot interaction, as well as enable applications such as improved accessibility for visually or hearing impaired individuals, automated video indexing, content search on social media, and quick multimedia information retrieval (see Fig. 1).

As reported in [1], traditional supervised learning methods require a large amount of labeled data, which can be costly and may suffer from issues such as label inconsistency and bias. Label inconsistency refers to the variation in labels assigned to the same sample by different annotators, which can lead to inconsistent model performance during training and testing. On the other hand, bias refers to the potential influence of subjective factors on labelers, resulting in inaccurate or unfair labeling of certain categories. These challenges and limitations in data annotation pose difficulties for traditional supervised learning approaches.

In recent years, a significant breakthrough has been witnessed in the field of natural language processing (NLP) with the emergence of foundation models. Foundation models are large-scale neural network architectures that are pre-trained on vast amounts of text data, enabling them to learn rich and diverse language representations. Additionally, the advent of multimodal foundation models, such as the Contrastive Language-Image Pre-training (CLIP) model [2], has further expanded the capabilities of foundation models. CLIP integrates visual and textual information, allowing the model to understand and generate coherent descriptions for images and videos. The success of foundation models, including CLIP, can be attributed to their ability to capture complex language patterns and visual representations, thereby enabling them to tackle challenging tasks. Furthermore, their pre-training and fine-tuning framework allows for efficient transfer learning, enabling adaptation to specific downstream tasks with relatively small amounts of task-specific data.

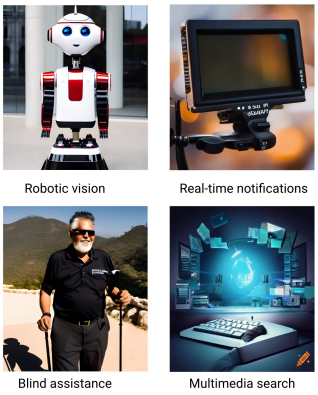


Fig. 1. Converting visual into textual information for human-machine interaction. Examples include robots using this technology to enhance visual perception and interact with users, real-time notification based on surveillance footage, assisting blind individuals in understanding their surroundings, and conducting diverse video searches.

Traditional video captioning approaches often struggle to capture the nuanced and complex relationships between visual and textual modalities, limiting their descriptive capabilities. By leveraging the power of CLIP, which has demonstrated exceptional performance in multimodal tasks, in this study we seek to enhance the accuracy and contextual relevance of video captions using CLIP. The motivation is based on two observations. First, CLIP's pre-training on massive amounts of image and text data enables it to learn joint representations

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

that effectively bridge the gap between visual and textual information. Second, CLIP's contrastive learning framework encourages the model to understand the underlying semantic relationships between images and their corresponding captions. The potential of CLIP for video captioning lies in its ability to leverage this learned multimodal understanding to generate captions that not only describe visual content accurately but also capture the context and intention behind the scenes.

In this paper, we develop an effective video captioning system based on the CLIP model [2] and an encoder-decoder framework named VLCap [3]. The experimental results on the ActivityNet Captions dataset [4] demonstrate the effectiveness of the proposed system in improving video captioning performance. By leveraging CLIP's description capabilities and the novel loss functions from [3], our system offers promising advancements in video-to-text translation tasks. This study contributes to the field of video captioning and opens up avenues for further exploration and improvement in this domain.

II. Related Work

Image analysis and natural language processing are among the core techniques in vision-language tasks. With the emergence of deep neural networks, most video captioning methods employ Recurrent Neural Networks (RNN) [5] to perform sequence modeling and utilize one or more Convolutional Neural Networks (CNN) [5] to analyze visual content. For instance, CNN is used as the encoder and the Long Short-Term Memory (LSTM) model is used as the decoder in [6]. Specifically, this approach employs three modules to process the video: (i) LOCATE for object detection in the video, (ii) RELATE for action detection in the video, and (iii) FUNC for natural language processing. However, in the RNN-based text generation models trained using the Teacher-Forcing method, there is a problem known as exposure bias. During training, the model uses previously generated words to predict the next word, leading to accumulated errors.

To address such issues, the Self-Critical Sequence Training (SCST) approach is proposed [5], utilizing reinforcement learning. SCST combines a sequence training model with Actor-Critic [7] to overcome the aforementioned problem and improve computational efficiency. In recent years, a novel architecture named Transformer has been introduced, which incorporates self-attention mechanisms to explore overall context in a sequence. Unlike the network architectures of RNN or CNN, Transformer performs better in training long sentences or videos. However, the use of attention mechanisms also increases its computational cost. Recent studies use Transformer as a foundational framework and optimize it with various methods [8]–[12]. For example, a new cross-model module is proposed in [13] to address the issues of poor computational efficiency and information asymmetry. Additionally, the skip-connected network is incorporated to further enhance the performance.

The proposed CLIP-CAP network is developed based on [3], which has two notable aspects. First, the fundamental Transformers heavily rely on self-attention mechanisms, which enable machines to learn the important relationships between words, similar to how humans understand words in a sentence. However, many existing video captioning methods use CNN as the encoder, which lacks self-attention. Therefore, an Adaptive Attention Mechanism (AAM) [14] is used in the encoding module, incorporating self-attention into the neural network. Second, most existing methods rely on Maximum Likelihood Estimation (MLE) for loss computation. However, MLE only considers the loss for the entire video and does not evaluate the learning progress of the model. To address this issue, a novel loss function, named the VL loss, is introduced in [3]. This method divides the video into multiple events and computes the VL Loss for each event, resulting in a more accurate error assessment.

III. Method

As displayed in Fig. 2, the model architecture consists of two main modules: the Encoder and the Decoder. The Encoder is responsible for processing the input video and generating visual features. On the other hand, the Decoder contains an attention mechanism and creates sentences based on the visual features. In this work, we rely on one of the powerful foundation models, CLIP [2], to compute effective visual features for this task. We now describe the background of CLIP, as well as the technical details of our video captioning system.

A. CLIP Pre-training

The core concept of CLIP [2] is to map text and images into the same space. Two encoders are involved in the training of CLIP—text encoder and image encoder. The text encoder is used to map a piece of text to a specific point in the space. The image encoder is used to map an image to a specific point in the same space.

Fig. 3 illustrates the training process of CLIP. It takes one batch at a time with n pairs of texts and images. A batch of texts involving n segments is fed into the text encoder to obtain text features $\{T_1, T_2, T_3, \ldots, T_n\}$. Similarly, the batch of n images is fed into the image encoder to obtain image features $\{I_1, I_2, I_3, \ldots, I_n\}$.

The CLIP model is optimized by using the loss: for the i-th pair of features (T_i, I_i) describing the same instance, the features should be as similar as possible, with a larger dot product between them indicating better similarity. This refers to the light blue area in Fig. 3. For the non-corresponding pairs in the batch, we want them to be as dissimilar as possible, ideally resulting in a dot product of zero. By using this contrastive loss, paired images and texts are mapped to surrounding points in the space, while non-matching images and texts will be mapped to different points that are far away from each other.

B. Encoder

We apply the encoder-decoder framework described in [3]. The video is first segmented into multiple small clips, referred

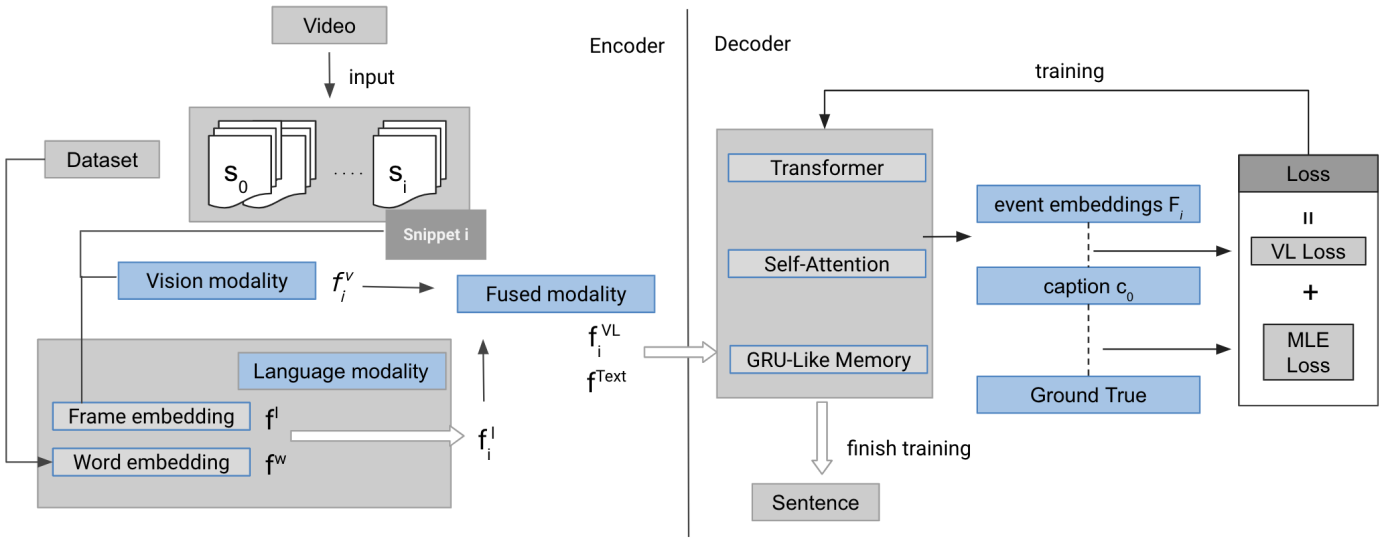


Fig. 2. Overall architecture of the proposed video captioning model. It consists of two modules: (i) CLIP-CAP Encoder takes a snippet S as input and returns VL feature as output (left). (ii) CLIP-CAP Decoder takes a list of VL features and returns a predicted caption (right).

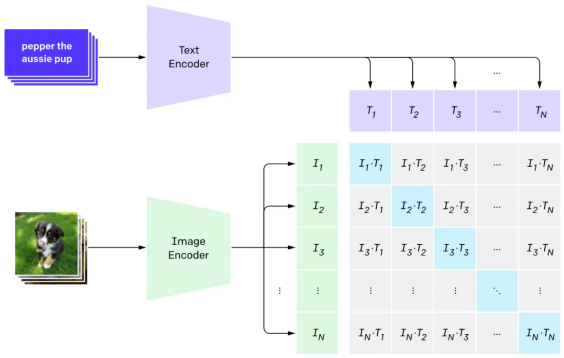


Fig. 3. Illustration of CLIP. The model trains an image encoder and a text encoder to predict image-text pairings. (Source: OpenAI - https://openai.com/research/clip)

to as events V_e. Each event is assigned a start time e_i^s and an end time e_i^e, i.e., $V_e = \{V_i | e_i^s < i < e_i^e\}$. The events are then fed into the Encoder as individual units. We then divide each input event V_e into L snippets, i.e., $V_e = \{S_i\}_{i=1}^{L}$. As depicted in Fig. 2 (left), the encoder employs three modalities to extract features from each snippet. The workflow of these modalities is illustrated in Fig. 4.

1) Vision Modality: The vision modality module is responsible for processing visual features. In this work, we apply CLIP [2] to compute visual features f_i^v for each snippet.

2) Language Modality: The language modality module is responsible for handling language features and consists of two stages: frame embedding and word embedding. In the frame embedding stage, the input snippet is transformed into visual features f^l using Vision Transformer [15]. In the word embedding stage, the language vocabulary is transformed into word features f^w. Both features (f^l and f^w) are then normalized using the pre-trained matrix [16] W from CLIP. Finally, the Adaptive Attention Mechanism (AAM) [14] is applied to filter out the language features f_i^l that are most relevant to the visual features.

3) Fused Modality: The fuse modality module combines the features generated by the two modules, namely the visual feature f_i^v and the language feature f_i^l, using self-attention mechanisms. This module generates the overall feature vector f_i^{VL} for the given snippet. Finally, L snippet feature vectors are combined to form the event feature vector f_i^{VL} and are sent to the Decoder.

C. Decoder

The input of the Decoder consists of the features F_i^{VL} generated by the Encoder for each event and the sentence obtained from the previous training session. The Transformer technique is used to construct the relationships between words in the sentence (intra-event coherence), while the GRU-Like Memory (historical information) handles the temporal aspects between sentences. At the end of each training session, a predicted caption c_i is generated, and then the Maximum Likelihood Estimation (MLE) is used to calculate the error, resulting in L_{MLE}. As mentioned earlier, using MLE alone cannot evaluate the current learning progress of the model. Therefore, the VL loss proposed in [3] is used in this work. This loss function involves using the event embeddings F_i generated by the Decoder and the sentences c_i obtained during each training iteration. These are processed through the Text Transformer to obtain the features f^T. The loss L_{vl} is computed as follows:

$$L_{vl} = -\sum_{i,j=1}^{N} 1_{i=j} \log e^p(f_i \cdot f_j^T) + 1_{i \neq j}(1 - \log e^p(f_i \cdot f_j^T)). \quad (1)$$

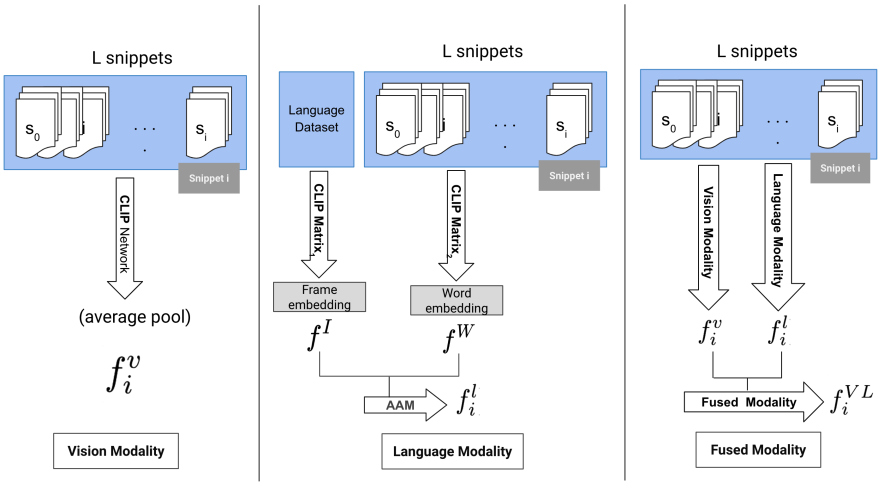


Fig. 4. Workflow of the vision modality (left), language modality (middle), and fused modality (right) in the encoder.

Finally, the training loss is a combination of L_{MLE} and L_{vl}:

$$L_{VL} = L_{MLE} + L_{vl}. \quad (2)$$

The Encoder and the Decoder are trained in an end-to-end manner.

IV. Experiments

The proposed CLIP-CAP model was trained and evaluated using the ActivityNet Captions dataset [4]. We used the pre-trained CLIP [17] model to compute visual features and did not fine-tune the model. The training samples were used to train the Encoder and the Decoder.

We evaluated the proposed CLIP-CAP model using three standard accuracy metrics, including BLEU@4 [18], METEOR [19], CIDEr [20], and one diversity metric, R@4 [21]. It is worth noting that we had fewer training samples to train CLIP-CAP due to the unavailability of approximately 3,000 videos in this dataset. Table I shows the comparison results with the state-of-the-art methods. Bold indicates the best result and underlined indicates the second best result. Although we used fewer training samples to train CLIP-CAP, it has the best performance in all metrics in comparison with other methods. Fig. 5 showcases a few results by comparing the descriptions generated by our model with the ground truth. The descriptions generated by our model are accurate and reasonable.

Next, we investigated the effectiveness of the CLIP features. In this experiment, we replaced the CLIP features with the C3D features, as shown to be a powerful representation for video captioning in [3]. As these features have different dimensions, we reduced the 100x2048 C3D features to 100x512 by using an autoencoder for a fair comparison. We used a smaller dataset comprising 16,580 samples in this experiment. The first two rows in Table II show the result. The CLIP features surpass the reduced C3D features in all metrics.

Finally, we reported the performance of the model using the original C3D features (2048-d) and that of using the concatenation of the C3D and the CLIP features (2560-d). The results are shown in the last two rows in Table II. These features are not complementary. Using the combined features achieved a comparable performance with C3D in BLEU and METOR, and slightly improved C3D in CIDEr and R@4.



Fig. 5. Descriptions generated by our method. (Left) A man is mowing the lawn on his lawn; the ground truth is "A man is seen pushing a lawn mower across a lawn to cut the grass." (Right) He starts mowing the lawn on the lawn; the ground truth is "The man continues moving the machine all along the yard while the camera captures his movements."

V. Conclusions

In this work, we have developed CLIP-CAP, a system for automated video description generation. The CLIP-CAP approach takes advantage of the powerful foundation model, CLIP. The pre-trained CLIP [17] model brings several advantages to a video captioning system, including semantic consistency and generality. It significantly simplifies the training pipeline, improves model performance, and enhances efficiency. Experiments on the ActivityNet Captions dataset demonstrate the effectiveness of CLIP-CAP, which outperforms state-of-the-art methods in four metrics.

Currently, our model focuses on processing visual content alone. In the future, we aim to integrate additional modalities such as audio information to further optimize our model's capabilities. Additionally, we aspire to implement dense video captioning as part of our ongoing advancements.

TABLE I
COMPARISON OF METHODS ON ACTIVITYNET CAPTIONS *ae-val*.

Metric	BLEU@4↑	METOR↑	CIDEr↑	R@4 ↓
Vanilla Transformer [22]	9.75	15.64	22.16	7.79
AdvInf [23]	10.04	<u>16.60</u>	20.97	5.76
GVD [24]	11.04	15.71	21.95	8.76
Transformer-XL [25]	10.39	15.09	21.67	8.54
Transformer-XLRG [26]	10.17	14.77	20.40	8.85
MART [26]	10.33	15.68	23.42	<u>5.18</u>
PDVC [27]	<u>11.80</u>	15.93	<u>27.27</u>	-
CLIP-CAP (Ours)	**12.29**	**17.19**	**28.31**	**3.21**

TABLE II
ANALYSIS OF CLIP AND C3D FEATURES ON ACTIVITYNET CAPTIONS *ae-val*.

Metric	Feature Size	BLEU@4↑	METOR↑	CIDEr↑	R@4 ↓
reduced c3d [3]	100x512	8.39	12.65	17.44	2.88
clip [17]	100x512	10.00	14.10	19.77	2.72
c3d [3]	100x2048	10.24	14.14	23.17	2.93
c3d+clip	100x2560	10.16	14.12	23.41	2.64

VI. ACKNOWLEDGMENT

This work was supported by the National Science and Technology Council of Taiwan (111-2221-E-003-016-MY2, 112-2813-C-003-020-E).

REFERENCES

[1] R. Alec, K. Jong Wook, H. Chris, *et al.*, "Learning transferable visual models from natural language supervision," *in Proc. of ICML*, 2021.

[2] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," *arXiv:2103.00020v1*, 2021.

[3] K. Yamazaki, S. Truong, K. Vo, *et al.*, "Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning," in *in Proc. of IEEE ICIP*, 2022.

[4] K. Ranjay, H. Kenji, R. Frederic, L. Fei-Fei, and N. Juan Carlos, "Dense-captioning events in videos," *in Proc. of ICCV*, 2017.

[5] R. Steven J, M. Etienne, M. Youssef, R. Jerret, and G. Vaibhava, "Self-critical sequence training for image captioning," *in Proc. of CVPR*, 2017.

[6] T. Gancho, L. Daqing, W. Meng, and Z. Zheng-Jun, "Learning to discretely compose reasoning module networks for video captioning," *in Proc. of IJCAI*, 2020.

[7] S. Richard S and B. Andrew G, "Reinforcement learning: An introduction," *Second Edition, MIT Press, Cambridge, MA*, 2018.

[8] L. Junnan, L. Dongxu, S. Silvio, and H. Steven, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *in Proc. of ICML*, 2022.

[9] C. Marcella, S. Matteo, B. Lorenzo, and C. Rita, "Meshed-memory transformer for image captioning," *in Proc. of CVPR*, 2020.

[10] I. Vladimir and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," *in Proc. of BMVC*, 2020.

[11] B. Max, N. Arsha, V. Gul, and Z. Andrew, "Frozen in time: A joint video and image encoder for end-to-end retrieval," *in Proc. of ICCV*, 2021.

[12] W. Jianfeng, Y. Zhengyuan, H. Xiaowei, *et al.*, "Git: A generative image-to-text transformer for vision and language," *in TMLR*, 2022.

[13] L. Chenliang, X. Haiyang, T. Junfeng, *et al.*, "Mplug: Effective and efficient vision-language learning by cross-modal skip-connections," *in EMNLP*, 2022.

[14] K. Vo, H. Joo, K. Yamazaki, *et al.*, "Aei: Actors- environment interaction with adaptive attention for temporal action proposals generation," *in Proc. of BMVC*, 2021.

[15] D. Alexey, B. Lucas, K. Alexander, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *in Proc. of ICLR*, 2021.

[16] R. Alec, K. Jong Wook, Chris, *et al.*, "Learning transferable visual models from natural language supervision," *in Proc. of ICML*, 2021.

[17] I. Vladimir, "Video features," unpublished. [Online]. Available: https://github.com/v-iashin/video_features.

[18] P. Kishore, R. Salim, W. Todd, and Z. Wei-Jing, "Bleu: A method for automatic evaluation of machine translation," *in Proc. of ACL*, 2022.

[19] D. Michael and L. Alon, "Meteor universal: Language specific translation evaluation for any target language," in *ACL Workshop on Statistical Machine Translation*, 2014.

[20] V. Ramakrishna, Z. C. Lawrence, and P. Devi, "Cider: Consensus-based image description evaluation," *in Proc. of IEEE CVPR*, 2015.

[21] X. Yilei, D. Bo, and L. Dahua, "Move forward and tell: A progressive generator of video descriptions," *in Proc. of ECCV*, 2018.

[22] Z. Luowei, Z. Yingbo, J. C. Jason, S. Richard, and X. Caiming, "End-to-end dense video captioning with masked transformer," *in Proc. of CVPR*, 2018.

[23] P. Jae Sung, R. Marcus, D. Trevor, and R. Anna, "Adversarial inference for multi-sentence video description," *in Proc. of CVPR*, 2019.

[24] Z. Luowei, K. Yannis, C. Xinlei, J. C. Jason, and R. Marcus, "Grounded video description," *in Proc. of CVPR*, 2019.

[25] D. Zihang, Y. Zhilin, Y. Yiming, C. Jaime, L. Quoc, and S. Ruslan, "Transformer-xl: Attentive language models beyond a fixed-length context," *in Proc. of ACL*, 2019.

[26] L. Jie, W. Liwei, S. Yelong, Y. Dong, B. Tamara, and B. Mohit, "Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning," *in Proc. of ACL*, 2020.

[27] W. Teng, Z. Ruimao, L. Zhichao, Z. Feng, C. Ran, and L. Ping, "End-to-end dense video captioning with parallel decoding," *in Proc. of CVPR*, 2021.

3D Point Cloud Denoising Based on Color Attribute

Wei-Chi Lin 1, Ming-Zhan Lee 2, He-Sheng Chou 3, Yuan-Jin Lin 4,
Kuo-Chen Li 5, Ting-Lan Lin 6, and Shin-Lun Chen 7*
Department of Electronic Engineering, Chung Yuan Christian University, Taiwan
E-mail: g11176043@cycu.edu.tw Tel: +886-3-265-4644
Department of Electrical Engineering National Cheng Kung University, Taiwan
E-mail: q16101282@gs.ncku.edu.tw Tel: +886-3-265-4644
Department of Electronic Engineering, Chung Yuan Christian University, Taiwan
E-mail: g10776018@cycu.edu.tw Tel: +886-3-265-4644
Department of Electrical Engineering National Cheng Kung University, Taiwan
E-mail: m26124069@gs.ncku.edu.tw Tel: +886-3-265-4644
Information Management Department Chung Yuan Christian University, Taiwan
E-mail: kuochen@cycu.edu.tw Tel: +886-3-265-5414
Department of Electronic Engineering National Taipei University of Technology, Taiwan
E-mail: tinglan@ntut.edu.tw Tel: +886-2-2771-2171 #2278
*Department of Electronic Engineering, Chung Yuan Christian University, Taiwan
E-mail: chrischen@cycu.edu.tw Tel: +886-3-265-4610

***Abstract* —Collecting or transmitting point cloud data is often subject to noise, which can potentially affect the accuracy of geometry and color representation in different spatial domains. This study addresses the denoising problem specifically for 3D point cloud color data and proposes two dedicated denoising algorithms based on the characteristics of noise in different spatial domains. In the denoising process, the RGB color space is first transformed into the YUV color space for further denoising operations. Surface smoothing is achieved by employing either a median filter or a bilateral filter based on the impact of noise on spatial information. These algorithms, built upon 2D image processing techniques, offer two key contributions: 1) color correction on spatial points to enhance denoising performance, and 2) the use of low-complexity filters while maintaining comparable filtering effectiveness, resulting in nearly a twofold reduction in processing time.**

Keywords – Median Filter, Bilateral Filter, Point Cloud.

I. INTRODUCTION

Point cloud is an unordered collection of points in 3D space. Information such as color, normals, transparency, position, or light intensity can be obtained through various acquisition methods such as scanning or capturing reconstruction. Point clouds have been increasingly applied in various fields, including medical modeling, heritage reconstruction, autonomous vehicle navigation, urban planning, and more. Compared to traditional 2D images that can only present a single perspective, point cloud data avoids issues such as occlusion and viewpoint transformation. However, during the acquisition process, point cloud data may suffer from geometric and color noise due to sensor errors, environmental interference, improper operation during acquisition, or noise introduced during transmission and compression, resulting in a degradation of its original quality.

Denoising has always been one of the significant challenges in point cloud technology. Unlike one-dimensional signals or two-dimensional images, point clouds do not possess explicit structural data. Therefore, the problem is primarily defined using Graph Signal Processing (GSP) [1] techniques. In previous literature, numerous denoising algorithms have been proposed [2]-[3], but most of them focus on addressing noise affecting the 3D positions. However, the color attributes of point clouds often come with noise, and even the noise in the positional attributes can indirectly impact the color information.

This paper focuses on the color denoising problem in point cloud space. Assuming the geometric space is either noise-free or has been corrected by previous denoising algorithms in the position space, the objective is to reduce noise in the color attributes without decreasing the number of points or altering the position properties of the point set. By adopting the concept of graph signal processing, each point in the point cloud is treated as a node in an undirected graph. Building upon well-established filtering algorithms developed for two-dimensional

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

images, the proposed denoising algorithm aims to remove noise from the color attributes without affecting the spatial positions.

II. PROPOSED ALGORITHM

After observing point cloud data affected by noise interference, this paper proposes different algorithms for processing color data based on specific circumstances, as illustrated in Fig. 1. Initially, a K Nearest Neighbors (KNN) search is performed to achieve a smoothing effect using the color information of neighboring points. Followed by the data preprocessing, the denoising process is applied using the proposed algorithms.

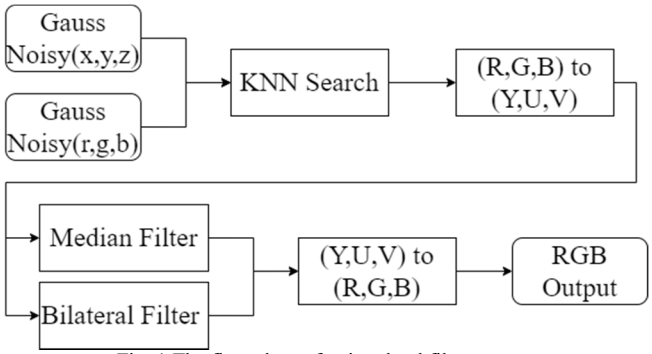


Fig. 1 The flow chart of point cloud filters .

A. Model Definition

A point cloud is composed of discrete points in 3D space, where each point can be represented as $p_i = [q_i, c_i]$. Here, p_i encompasses six pieces of information, namely the positional attribute represented by $q_i = [x_i, y_i, z_i]$, indicating the Cartesian coordinate information of point i. All positional attributes can be collectively represented as $q = [q_1, q_2, \ldots\ldots, q_N]^T$, where N denotes the total number of points in the point set. Based on the preliminary assumption , the positional coordinates are assumed to be noise-free or have already been corrected by other positional denoising algorithms. The other three attributes representing color are denoted as $c_i = [r_i, g_i, b_i]$. The color information in the point cloud data is stored in the RGB format. All color attributes can be collectively represented as $c = [c_1, c_2, \ldots\ldots, c_N]$. Therefore, the entire point cloud collection can also be represented as $p = \{q, c\}$.

Traditional two-dimensional image restoration refers to the problem of restoring a degraded or interfered image to its original form. Building upon this foundation, the color attribute model affected by noise can be simply defined as (1) :

$$d = c + n \quad (1)$$

B. Graph Construction

Due to the discrete and unordered nature of point clouds, mathematical operations can be highly challenging. Therefore, it is necessary to assign a representation that captures the correlation between points. An undirected graph $G = \{\nu, \varepsilon\}$ is defined, where ν represents the set of vertices in the graph, and ε represents the edges that form each connection within the graph.

The K NN method is employed to find neighboring points. This approach is based on the nearest Euclidean distance. By calculating the distances, nine closest points to point p_i are identified and considered as its neighboring points. This method treats each calculation as constructing a new undirected graph $G\{\nu, \varepsilon\}$, where ν represents the nine neighboring points along with point p_i, and ε represents the connections between each point and p_i.

C. Color Space Conversion

A color space is a standardized representation that defines how colors are displayed and typically consists of three parameters. The RGB model is a commonly used color space, where each color is represented based on corresponding spectral components of red, green, and blue. However, color hue, saturation, and brightness cannot be directly obtained from the RGB model. In contrast, the YUV model uses luminance and chrominance as parameters. One advantage of the YUV model is that luminance (Y) and chroma apperception (U,V) are separated, resulting in lower correlation compared to the RGB approach. In typical situations, luminance and chrominance exhibit piecewise smoothness (PWS) compared to the RGB color space. It can use the following formulas for conversion [4] :

$$[Y\ U\ V] = [0.299\ 0.587\ 0.114\ -0.147 \\ -0.289\ 0.436\ 0.614\ -0.515 \\ -0.100\ ][R\ G\ B] \quad (2)$$

$$[R\ G\ B] = [1.000\ 0.000\ 1.140\ 1.000\ -0.369 \\ -0.581\ 1.000\ 2.029\ 0.000\ ][Y\ U\ V] \quad (3)$$

The relationship between the YUV channels and RGB is linear, and the Y channel represents the additive synthesis of RGB, preserving the original high-frequency signal. The U,V channels, on the other hand, are obtained through subtraction operations from RGB. In addition to separating luminance and chrominance, the original colors are also divided into high-frequency and low-frequency components. Therefore, it is easier to achieve noise removal in image processing using this approach.

D. Bilateral Filtering

Bilateral filtering is a spatial filter that has been well-developed and commonly used to address denoising problems in 2D images [6]. It is based on the notion of spatial proximity and color similarity, using a Gaussian kernel function as weights. Points that are close in space and similar in color will have a larger influence on the new color attributes of the current point. The weight formula is as shown in (4) :

$$W_j = exp(-\frac{\|q_i - q_j\|^2}{2\sigma_q^2})exp(-\frac{(d_i - d_j)^2}{2\sigma_d^2}) \quad (4)$$

where q_i represents the coordinates of the current point, q_j denotes the coordinates of the neighboring points obtained through KNN search, d_i represents the color information of the current point, d_j represents the color value of the neighboring points, and σ_q and σ_d are the standard deviations of position and color, respectively.

The formula for bilateral filtering is defined in (5):

$$\hat{c}_i = \sum_{j=1}^{k} \frac{1}{w_j} d_j * w_j \quad (5)$$

Based on the bilateral filtering formula, this paper calculates the product of the neighboring weights obtained through KNN search and uses filters to obtain new values for the Y, U, and V channels. The color-related computations for each channel are performed independently, while the point positions remain fixed. Due to the characteristics of bilateral filtering, the newly obtained values can preserve image features and edges.

E. *Median Filtering*

While processing the noise for the positional properties, it may indirectly introduce errors in the color properties. P oints that have moved to incorrect positions can be seen as color attributes being disturbed by impulse signals. In 2D images, filters can effectively eliminate such noise while preserving the original image details and features.

One of the filters that effectively removes this type of noise is the median filter [5], which is computationally simple and fast compared to other filters. The median filter is defined in (6):

$$Median\,(\hat{c}_i) = c_j\left(\frac{k+1}{2}\right), where\ k = 9 \quad (6)$$

where c_j represents the set of neighboring points obtained through KNN search, and the YUV values of c_j are arranged in ascending or descending order. After sorting, the median of each YUV channel is taken as the new YUV value.

III. EXPERIMENTAL RESULTS

To evaluate the denoising results objectively, this study uses the Peak Signal-to-Noise Ratio (PSNR) as a standard metric. In order to calculate the PSNR values before and after denoising, this study first performed calculations on the color channels of the original image and the comparison image at the corresponding positions. Considering that the denoising results might lead to positional mismatches, if there is no corresponding point at the same position in both images, the nearest neighboring point is selected as the corresponding point. The Mean Squared Error (MSE) is then calculated based on the color differences between the two points. The calculation formulas for PSNR and MSE are as follows in (7) and (8):

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \|c_i - \hat{c}_i\|^2 \quad (7)$$

$$PSNR = 10log_{10}\left(\frac{255^2}{MSE}\right) \quad (8)$$

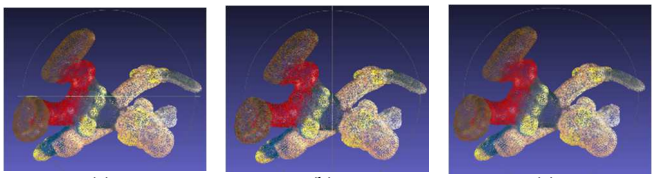
Where c_i and $\hat{c}_i$ respectively represent the color information of the original point cloud and the color information of the denoised or noise-added points, N is the total number of points in the point cloud collection. The software and hardware platform used in this study are shown in Table 1.

Table 1 The hardware and software platform

Hardware platform	Version
CPU	Intel Core i7-7700
DRAM	DDR4 2666 16G
Software platform	Version
MATLAB	R2019a

A. *Performance of Bilateral Filtering*

In this section, the denoising method is applied to the Greyc dataset of synthetic point clouds to validate the effectiveness of the proposed methods. From Fig. 2, it can be observed that Fig. 2-a and Fig. 2-b depict the original point cloud and the point cloud with color attributes affected by Gaussian noise. After applying the proposed denoising processes, the color noise has been significantly removed, as shown in Fig. 2-c. Furthermore, the positional properties of the entire point cloud data have not been altered, and there has been no increase or decrease in the number of points.



(a) (b) (c)

Fig. 2 Asterix in point cloud model (a) Ground-truth (b)Noisy (c) Result.

Fig. 3 illustrates the appearance of the point cloud after surface reconstruction using the triangulation method. The result shown in Fig. 3-c provides a more intuitive and clear visualization than the image affected by Gaussian noise as shown in Fig. 3-b. The colors on the surface are obtained by averaging the color attributes of the connected three points.



(a) (b) (c)

Fig. 3 Asterix in triangulation method (a) Ground-truth (b)Noisy (c) Result.

In the Greyc dataset, the color attributes of the point cloud were simulated to be affected by Gaussian noise with a noise level of $\mu = 0$ and $\sigma = 20$. The proposed color denoising algorithm in this paper was then applied to restore the point cloud. Table 2 and Table 3 provide a comparison of the MSE and PSNR between the method proposed in this paper and the Tikhonov, TV, and SGW [7] methods.

Table 2 MSE Comparison of Color Denoising Algorithm for Greyc Dataset

Model	Noisy	Proposed	SGW[7]	Tikhonov	TV
Asterix	361.25	**76.51**	98.2	103.56	279.02
Green Dinosaur	395.78	63.36	**63.03**	63.04	301.5
Jaguar	370.16	**71.29**	76.68	85.83	279.21
Pokemon	311.15	**83.75**	156.1	149.53	241.15
Red Horse	376.13	100.90	**80.64**	82.52	293.32
Average	362.894	**79.16**	94.93	96.896	278.84

Table 3 PSNR Comparison of Denoising Algorithm for Greyc Dataset

Model	Noisy	Proposed	SGW[7]	Tikhonov	TV
Asterix	22.56	**29.31**	28.21	27.98	23.67
Green Dinosaur	22.16	30.12	**30.13**	**30.13**	23.34
Jaguar	22.46	**29.62**	29.28	28.79	23.67
Pokemon	23.21	**28.93**	26.2	26.38	24.31
Red Horse	22.38	28.13	**29.07**	28.97	23.46
Average	22.554	**29.22**	28.578	28.45	23.69

These two indicators objectively demonstrate that the proposed method in this paper performs well in restoring the point cloud compared to the current algorithms. In order to compare the performance of this study, this study compares with other methods namely SGW, GLR, and GTV, which have shown good performance in denoising algorithms. As shown in Table 4 and Table 5, this study slightly outperforms the other algorithms in terms of average denoising performance. It is particularly notable in terms of speed, with an average execution time (AET) that is twice as fast as SGW. As a result this study maintains the denoising performance while improving the speed.

Table 4 Comparison of Color Denoising Based on PSNR

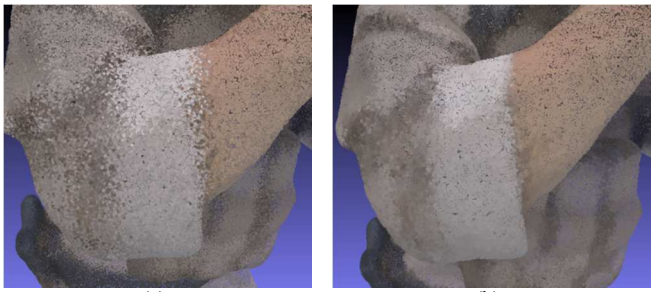
Model	Noisy	Proposed	SGW[7]	GLR[8]	GTV[8]
Asterix	22.56	**29.31**	28.21	28.10	27.56
Green Dinosaur	22.16	30.12	30.13	30.30	**30.34**
Red Horse	22.38	28.13	**29.07**	27.85	28.57
Average	22.37	**29.19**	29.14	28.75	28.82

Table 5 Comparison of Color Denoising Time Based on AET (s)

Model	Proposed	SGW[7]	GLR[8]	GTV[8]
Asterix	0.58	1.06	5.20	211.00
Green Dinosaur	1.06	1.82	11.20	389.00
Red Horse	1.84	3.48	19.50	851.00
Average	1.16	2.21	11.97	483.67

B. Performance of Median Filtering

From Fig. 4-a, after applying position-based denoising, some points affected by Gaussian noise still appear in incorrect positions. These erroneous points on the segmented smooth surface resemble impulse interference.



(a) (b)

Fig. 4 The cuff is disturbed by Gaussian noise (a) two stage filter (b) Result.



(a) (b)

Fig. 5 The Gaussian noise data (a) two stage filter (b) Result

Fig. 4-a and Fig. 5-a demonstrate the results obtained after applying the two-stage filter. Median filtering effectively corrects the color attributes of erroneous points, restoring the surface color of the point cloud to a smoother state, as shown in Fig. 4-b and Fig. 5-b.

Table 6 The comparison of PSNR

	Two Stage Filter	This Proposal
PSNR	24.04	27.85

The PSNR values provided in Table 6 indicate that the position-based denoising algorithm still significantly affects the PSNR in terms of color. In other words, this study effectively addresses the color errors caused by noise interference in the position.

IV. CONCLUSIONS

This study proposes a low-complexity and effective method for denoising point cloud colors and classifying noise, as noise can interfere with color space when acting in different spaces. The proposed algorithms in this paper are based on two-

dimensional image processing techniques, namely median filtering and Gaussian bilateral filtering, both of which are developed specifically for point cloud data under the assumption of position-wise stationary (PWS) properties. By employing K-nearest neighbors (KNN) search to identify a specific number of neighboring points and defining them as correlated data, the algorithm facilitates subsequent filtering operations. In both subjective and objective evaluations, the proposed algorithm demonstrates promising results, surpassing the state-of-the-art techniques in terms of computational speed. Extensive quantitative analysis is conducted on the Grey dataset, where evaluations based on mean squared error (MSE) and peak signal-to-noise ratio (PSNR) achieve satisfactory outcomes. Unlike existing point cloud denoising approaches, this paper focuses on denoising the color attributes of point clouds while preserving the integrity of their positional attributes. In future work, the integration of graph theory and triangulated meshes will be explored to achieve a more accurate representation of neighboring points. This alternative approach, as opposed to KNN, is expected to enhance the precision of the denoising process.

ACKNOWLEDGMENT

This research is supported by the Ministry of Science and Technology, Taiwan, under Grant numbers of MOST-111-2221-E-033-041, MOST-111-2823-8-033-001, MOST-111-2622-E-131-001, MOST-110-2223-8-033-002, MOST-110-2221-E-027-044-MY3, MOST-110-2218-E-035-007, MOST-110-2622-E-131-002, MOST-109-2622-E-131-001-CC3, MOST-109-2221-E-131-025, MOST-109-2410-H-197-002-MY3, and in part by the National Chip Implementation Center, Taiwan.

REFERENCES

[1] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph Spectral Image Processing," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 907–930, May 2018, doi: 10.1109/JPROC.2018.2799702.

[2] Z. Xu and A. Foi, "Anisotropic Denoising of 3D Point Clouds by Aggregation of Multiple Surface-Adaptive Estimates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 6, pp. 2851–2868, Jun. 2021, doi: 10.1109/TVCG.2019.2959761.

[3] H. Chen, M. Wei, Y. Sun, X. Xie, and J. Wang, "Multi-Patch Collaborative Point Cloud Denoising via Low-Rank Recovery with Graph Constraint," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3255–3270, Jan. 2020, doi: 10.1109/TVCG.2019.2920817.

[4] R. J. Patan, M. Chakraborty, and S. S. Devi, "Blind color image de-convolution in YUV space," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2016, pp. 1551–1555. doi: 10.1109/ICCSP.2016.7754419.

[5] G. George, R. M. Oommen, S. Shelly, S. S. Philipose, and A. M. Varghese, "A Survey on Various Median Filtering Techniques For Removal of Impulse Noise From Digital Image," in *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, Mar. 2018, pp. 235–238. doi: 10.1109/ICEDSS.2018.8544273.

[6] B.-H. Chen, H.-Y. Cheng, Y.-S. Tseng, and J.-L. Yin, "Two-Pass Bilateral Smooth Filtering for Remote Sensing Imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2020.3048488.

[7] M. A. Irfan and E. Magli, "Joint Geometry and Color Point Cloud Denoising Based on Graph Wavelets," *IEEE Access*, vol. 9, pp. 21149–21166, 2021, doi: 10.1109/ACCESS.2021.3054171.

[8] C. Dinesh, G. Cheung, and I. V. Bajić, "3D Point Cloud Color Denoising Using Convex Graph-Signal Smoothness Priors," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, Sep. 2019, pp. 1–6. doi: 10.1109/MMSP.2019.8901695.

The DSP and DDR4 VLSI Design for Multi-Sensor in Biomedical System

Jia-Sheng Zhang[1],Chiung-An Chen[1], Shih-Lun Chen[2],and Yao-Zhong Zhang[1], *Member, IEEE*

[1]Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City, Taiwan

[2]Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan City, Taiwan

DDR4 is a high-speed memory structure commonly employed in modern medical devices. The benefit of DDR4 architecture enables fast data transmission and efficient data processing, it can effectively reduce the time for responding to medical devices and improve their performance. Therefore, our proposal is to design under Verilog and implement a DDR4 structure in order to transfer the biomedical signals of temperature, electrocardiogram(ECG), and electroencephalogram(EEG) data, furthermore, using ModelSim to simulate and verify the accuracy of the system.

Keywords—DDR4, ECG, EEG, temperature, multi-sensor Verilog, Modelsim

I. INTRODUCTION

With the continuous advancement of modern medical equipment and technological capabilities [1], we now have a plethora of devices capable of detecting human physiological signals. However, these devices are faced with escalating demands such as high-speed data transmission and reduced power consumption. Hence, our research endeavors to design a system that can efficiently receive multiple signals, operate at high speed, and minimize power consumption. This system is intended to address the evolving challenges posed by contemporary medical devices.

Double Data Rate 4 Synchronous Dynamic Random-Access Memory (DDR4 SDRAM) is a type of synchronous dynamic random-access memory with a high bandwidth ("double data rate") interface. The main advantages of DDR4 compare with its predecessor, DDR3, include higher module density and lower voltage requirements, moreover, it has better speed with data transmission. The standard of DDR4 allows for DIMMs with up to 64 GB in capacity. DDR4 works at a voltage of 1.2 V and a frequency between 800 to 1600 MHz. [2]

Electrocardiography is the process of producing an electrocardiogram (ECG), it records the heart's electrical activity through repeated cardiac cycles. Three main components are composed of an ECG: the P wave, which represents depolarization of the atria; the QRS complex, which represents depolarization of the ventricles; and the T wave, which represents repolarization of the ventricles.[3]

Electroencephalography (EEG) is a method to record an electrogram for the spontaneous electrical activity of the brain. EEG can detect abnormal electrical discharges such as sharp waves, spikes, or spike-and-wave complexes in epilepsy patients, thus it is often used to inform the medical diagnosis. EEG can detect epilepsy patients for spatiotemporal (location and time) evolution of seizures and the continuous status of epileptics. It is also used to help diagnose sleep disorders, depth of anesthesia, coma, encephalopathies, cerebral hypoxia after cardiac arrest, and brain death.[4]

II. SYSTEM ARCHITECTURE

In this paper, our objectives are threefold: signal differentiation, data compression, and high-speed transmission. To achieve these goals, our adopted architecture consists of three main components: an adaptive filter, entropy processor, and a DDR4 controller. As depicted in Fig. 1, this diagram illustrates the system architecture of our study. In the subsequent sections, each of these components will be introduced and explained in detail.

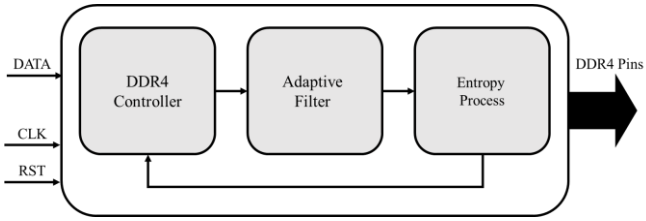


Fig. 1. The system structure of the frame diagram

A. Adaptive Filter

Considering the diverse input sources encompassing body temperature, electroencephalogram (EEG), and electrocardiogram (ECG) signals in this paper, individualized compression processing for distinct signal sources is aimed to be achieved. This approach offers the advantage of enhancing data compression efficiency while reducing system power consumption. Hence, the role of the adaptive filter in the preprocessing stage is of paramount importance. Given the varying bit lengths within our data, we can classify signals by detecting the bit length of the input source, thus achieving improved compression results.

B. Entropy processor

Entropy coding is a data compression technique that represents data efficiently by utilizing probability models and the concept of entropy. The objective of this approach is to reduce redundant information in data, thereby occupying less storage space or transmission bandwidth.[5]

In entropy coding, symbols with higher probabilities of occurrence are assigned shorter codes, while symbols with lower probabilities receive longer codes. The aim is to save space when encoding common symbols and ensure that codes for different symbols do not overlap. The two primary methods of entropy coding are Huffman coding and arithmetic coding. In this paper, we adopt Huffman coding due to its simplicity and good compression efficiency.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In this paper, we deal with three types of biomedical signals: body temperature, electroencephalogram (EEG), and electrocardiogram (ECG). Since body temperature exhibits relatively small variations within each time unit, we choose to employ the differential coding method for compression. Conversely, EEG and ECG display a high degree of temporal redundancy within each time unit, making Huffman coding an effective choice for achieving compression. Through these selections, we ensure data integrity while achieving efficient compression.

C. DDR4 Controller

To achieve the goal of high-speed transmission, we adopt the DDR4 architecture.[6] To accomplish this objective, designing a DDR4 controller is essential for managing data read and write operations. Throughout this design process, referencing the DDR4's Data Sheet is crucial to ensure accurate operation.

The primary functionalities of DDR4 revolve around executing distinct actions through command instructions, including activating, reading/writing, precharging, and refreshing. In our design, we need to tailor the DDR4 controller based on these fundamental operations to ensure our system executes these actions accurately.

Fig. 2 illustrates a simplified state diagram for DDR4, aiding our comprehension of its operational processes. We will utilize this FSM diagram to craft our DDR4 controller, ensuring smooth transitions between different states to accomplish high-speed data read and write operations.

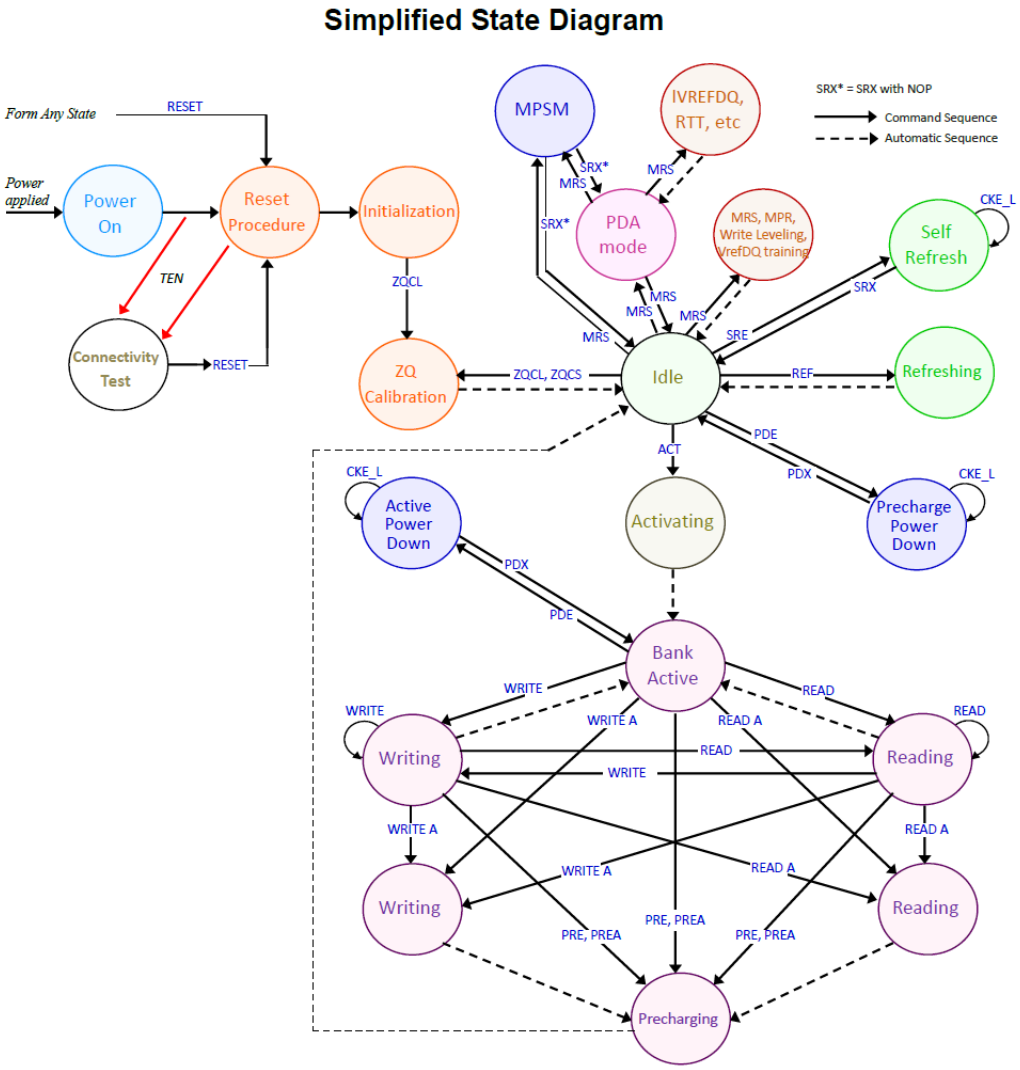


Fig. 2. The DDR4 FSM diagram

III. Simulation Environment and Results

In this study, we opted for the Windows operating system as the platform for execution and simulation, utilizing ModelSim SE 10.5 to create our simulation environment. The choice of the Windows operating system offers a user-friendly graphical interface and an extensive software ecosystem, facilitating development, simulation, and analysis. Selecting ModelSim SE 10.5 as our simulation tool enables us to conduct in-depth analysis of our design.

In the upcoming sections, we will present simulation results for the adaptive filter, entropy processor, and DDR4 controller components. Additionally, relevant data will be introduced to provide detailed explanations and analyses.

Table I presents the power consumption, number of Look-Up Tables (LUTs), and number of flip-flops (FFs) utilized by each component. Additionally, Figs. 3, 4, and 5 display the simulated waveform of each component, providing a clear representation of the relationship between input signals and filtered outputs.

TABLE I. Hardware Resources Consumed by Each Component

	Power	LUT	FF	Gate Counts
Adaptive filter	0.2 W	6	17	0.36 k
Entropy processor	0.94 W	45	35	0.27 k
DDR4 Controller	3.81 W	208	306	1.25 k
Total	4.95 W	259	358	1.55 k

Figure 3 illustrates the output and input of the Adaptive Filter. In this simulation, "data_in" represents the input signal, and for this particular simulation, it uses EEG signals. "state_out" and "data_out" are the output signals. In this design, "state_out" signals 0 to 3 correspond to "IDLE," "body temperature," "ECG," and "EEG" respectively, in sequence. From the simulation diagram, it is evident that the simulation is accurate and correctly represents the intended functionality

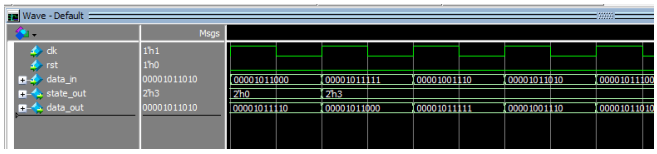


Fig. 3. ModelSim waves of Adaptive Filter processor

Figure 4 displays the output and input of the Entropy processor. In this illustration, "data_in" and "state_in" represent the input signals. "state_in" is derived from the previous stage, the Adaptive Filter, and it indicates the source of the input signal, selecting the corresponding entropy processing method. "data_out" and "length" are the output signals.

For this Entropy processor design, it first subtracts the two consecutive data points. In the mode where the input source is "body temperature," it directly outputs the difference between these two data points. In the modes where the input source is "ECG" or "EEG," the Entropy processor transforms this difference into the corresponding Huffman-encoded value. Table II and III describes the Huffman encoding methods for "ECG" and "EEG."

For this simulation, EEG signals are used. Therefore, it can be observed from Figure 4 that the "state_in" signal is 3, indicating EEG signals. From the simulation diagram, it is evident that the simulation is accurate and correctly represents the intended functionality.

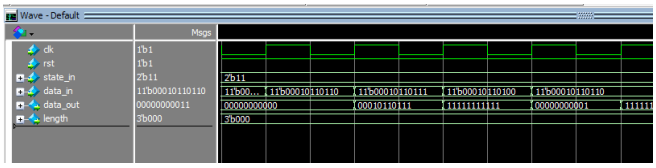


Fig. 4. ModelSim waves of Entropy processor

TABLE II. The Huffman Encoding Methods for ECG.

The Huffman encoding methods for ECG			
value	frequency	symbol	length
0	226	000	10
1	118	010	11
-1	119	011	11
2	49	100	11
-2	21	101	11
-3	24	110	11
3	15	111	11

TABLE III. The Huffman Encoding Methods for EEG.

The Huffman encoding methods for EEG			
value	value	value	value
0	0	0	0
1	1	1	1
-1	-1	-1	-1
2	2	2	2
-2	-2	-2	-2

Figure 5 presents the input and output signals of the DDR4 controller section. Signals like "cke," "actn," "bank_add," "rasn," "casn," "bank_add," and "addr" play critical roles in DDR4 control, and Table IV lists some key DDR4 commands. The design of the DDR4 controller is primarily based on the signal timing provided in the datasheet, with many design details referencing datasheet information. The signals in Figure 5 illustrate the functionality of DDR4 initialization.

In this study, the Terasic HAN Pilot Platform development board was used to test the functionality of the DDR4 controller. This development board offers Nios II software program control, enabling quick hardware testing through software. Figure 6 displays the test results

for data write and read operations, demonstrating successful completion of both write and read functions. This indicates that DDR4 initialization and read/write functions operate as expected.

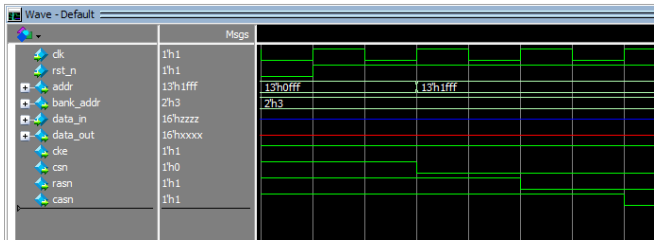


Fig. 5. ModelSim waves of DDR4 Controller processor

TABLE IV. THE DDR4 COMMANDS.

Symbol	CKE		CS	ACT	RAS	CAS	WE
MRS	H	H	L	H	L	L	L
REF	H	H	L	H	L	L	H
PRE	H	H	L	H	L	H	L
ACT	H	H	L	L	Row Address		
WR	H	H	L	H	H	L	L
RD	H	H	L	H	H	L	H
NOP	H	H	L	H	H	H	H

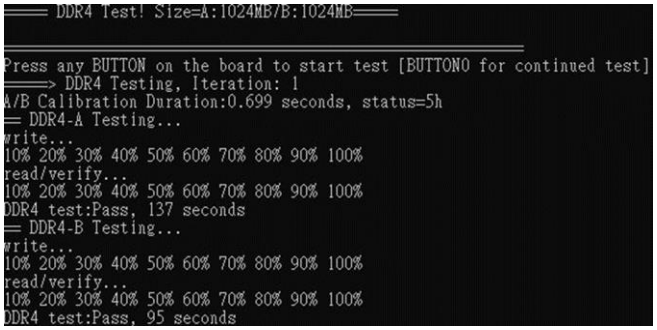


Fig. 6. The DDR4 Controller testing result

IV. CONCLUSION

In this paper, we utilized ModelSim for circuit simulation and verification, achieving DDR4 high-speed transmission, automatic classification of sensor signals by the adaptive filter, and corresponding compression. The compression ratios achieved are TEMP:3.6, ECG:2.2, EEG:2.5.

Compared to the work by Y. -J. Huang et al., their system can adaptively handle four different input signals, which is superior to our paper's three signals. However, our research introduced innovations in high-speed data transmission and data compression technologies, providing significant advantages to our system, including reduced power consumption and improved system performance. This makes our system more suitable for addressing modern and diverse application requirements while delivering higher efficiency.

In comparison to the paper by T. -H. Tsai and M. A. Hussain, although their work focused on lossless compression of ECG, our research introduced compression for multiple types of signals, making our system more versatile and functional.

Compared to the paper by L. Kuang, J. Zeng, and P. Georgiou, while their paper also utilized DDR4 high-speed storage, our system excels in terms of functionality and versatility.

In summary, the comparison in Table V clearly outlines the strengths of our paper. Our system excels in processing multiple sensor signals, adaptive filtering, and high-speed transmission capabilities.

TABLE V. COMPARISON WITH OTHER ARTICLES

	This paper	2013[8]	2020[9]	2021[10]
Number of sensors	3	4	1	1
Compression ratios	Temp: 3.6 ECG: 2.2 EEG: 2.5	×	2.77	×
Adaptive filter	✓	✓	×	×
High-speed transmission	DDR4	×	×	DDR4

REFERENCES

[1] Abid Haleem, Mohd Javaid, Ravi Pratap Singh, Rajiv Suman,Medical 4.0 technologies for healthcare: Features, capabilities, and applications,Internet of Things and Cyber-Physical Systems,Volume 2,2022,Pages 12-30,ISSN 2667-3452,

[2] Wiki, USA. DDR4 SDRAM. (2023) [Online]. Available: https://en.wikipedia.org/wiki/DDR4_SDRAM. Accessed on: Mar.12, 2023

[3] Wiki, USA. Electrocardiography. (2023) [Online]. Available: https://en.wikipedia.org/wiki/Electrocardiography. Accessed on: Mar.12, 2023

[4] Wiki, USA. Electroencephalography. (2023) [Online]. Available: https://en.wikipedia.org/wiki/Electroencephalography. Accessed on: Mar.12, 2023

[5] Huffman, David (1952). "A Method for the Construction of Minimum-Redundancy Codes". Proceedings of the IRE. Institute of Electrical and Electronics Engineers (IEEE). 40 (9): 1098–1101. doi:10.1109/jrproc.1952.273898

[6] M. A. Islam, M. Y. Arafath and M. J. Hasan, "Design of DDR4 SDRAM controller," 8th International Conference on Electrical and Computer Engineering, 2014, pp. 148-151, doi: 10.1109/ICECE.2014.7026950.

[7] "8Gb_DDR4_C_Die_component_Datasheet", Nanya Technology, New Taipai City. Taiwan, 2023.

[8] Y. -J. Huang et al., "A self-powered CMOS reconfigurable multi-sensor SoC for biomedical applications," 2013 Symposium on VLSI Circuits, Kyoto, Japan, 2013, pp. C248-C249.

[9] T. -H. Tsai and M. A. Hussain, "VLSI Implementation of Lossless ECG Compression Algorithm for Low Power Devices," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 12, pp. 3317-3321, Dec. 2020, doi: 10.1109/TCSII.2020.2978554.

[10] L. Kuang, J. Zeng and P. Georgiou, "Live Demonstration: Real-Time and High-Speed Ion Imaging Using CMOS ISFET Arrays," 2021 IEEE Biomedical Circuits and Systems Conference (BioCAS), Berlin, Germany, 2021, pp. 01-01, doi: 10.1109/BioCAS49922.2021.964

Identification of Victims Wearing Vibrant Clothing using MATLAB

HAO-CHENG,LU/CHIUNG-AN,CHEN/Jia-Sheng, Zhang/Yao-Zhong, Zhang

Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City, Taiwan

M11128006@o365.mcut.edu.tw, joannechen@mail.mcut.edu.tw, M11128005@o365.mcut.edu.tw ,M12128014@o365.mcut.edu.tw

***Abstract*—This paper presents an image-based target detection and integration algorithm for victim wearing vibrant clothing. The algorithm utilizes image processing and analysis techniques to automatically identify and locate victims, and combines the detected bounding boxes into a larger box. Experimental results validate the effectiveness and accuracy of the algorithm in various environments. The proposed algorithm can be applied in emergency response, disaster management, and surveillance systems, enhancing the efficiency of victim search and rescue operations**

Keywords—victim search, vibrant clothing

I. INTRODUCTION

While mountaineering is undoubtedly a highly enjoyable recreational activity, it also comes with various risks. Adverse weather conditions, slips and falls, rockfalls, and altitude sickness, among others, can lead to mountain accidents. With the increasing popularity of mountaineering in recent years, the number of mountain incidents has also been on the rise. Mountain rescue is an extremely challenging and resource-intensive task. Hindered by terrain, signal, and visibility issues, mountain rescue operations are highly demanding. The dense vegetation and foliage not only obstruct the rescuers' line of sight but also make aerial rescue difficult. Rescue missions are a race against time because once night falls, the rescue operation becomes even more daunting. G. Seeja et al [1] mentioned that the possibility of saving a victim is high only during the first 48 hours, and then the prospect tends to zero. So they developed a Transfer Learning-based Deep Learning approach to identify human victims under collapsed building environments by integrating machine learning classification algorithms. To ensure higher accuracy within a shorter timeframe.

Image processing is a field of study that focuses on analyzing and manipulating digital images using computer algorithms. It involves various techniques and methods to enhance, transform, or extract information from images for a wide range of applications. In this area, Xuewen Wang et al [2] used an algorithm for extracting characters from text regions of natural scene images with complex backgrounds. W. Hoshino et al [3] mounted a wide-angle camera on the drone, which is easy to control in a narrow building space. In some experiments, the drone was equipped with a thermographic camera to facilitate detection of victims. They investigated the effects of each and looked for ways to detect victims more accurately.

H. Liu et al [4] created an innovative air-space integrated system that combines the Remote Satellite System (RSS) and Unmanned Aerial Vehicles (UAVs) to enhance disaster response. Utilizing advanced image analysis and UAV control techniques, they employ fusion algorithms for victim detection, satellite navigation for patrolling, and camera-based navigation for precise aid delivery.

D. Yanguas-Rojas et al [5] presents a control strategy for multiple heterogeneous robots oriented to the assistance in situations of search and rescue from two complementary perspectives, the real time control and the discrete task allocation. This approach takes into account not only the current conditions of the environment and victims but also estimates the future positions of victims and the potential locations of remaining victims.

Under the comprehensive framework outlined in numerous research papers, the potential for implementing the entire system has been enhanced. This includes aspects such as search and localization, path planning, as well as visual analysis techniques, and more. The focal point of this article will be on feature recognition, aiming to address the system's limitations and enhance its capabilities.

The article proposes a method that can quickly search for mountaineers in images, utilizing the functionalities provided by MATLAB. The application areas include computer vision, autonomous driving [6] , medical imaging and many other fields.

II. IMAGE PROCESSING

As shown in Figure1., our proposed system consists of six-part: Input, Pre-Processing, Target Detection, Region Labeling, Region Selection, and Integration of Bounding Boxes. Next, each part will be described in detail.

A. Color Space Conversion

In digital images, colors are typically represented using different color models, such as RGB, HSV, CMYK, and YUV. Each color model has its own advantages and characteristics, and converting between them enables us to extract specific color information or enhance certain features of an image.

C. Liu et al [7] mentioned a fog level detection method based on image HSV color histogram and shows that the HSV model is very suitable for image algorithm processing color perception based on the human visual system. The conversion from RGB to HSV is based on the transformation of the cube model in RGB color space to the cylindrical model in HSV color space is obtained by the following formula:

$$H = \begin{cases} 0, & if\ Min = Max \\ 60 \times \frac{(G-R)}{(Max-Min)} + 60\ (if\ Min = B) \\ 60 \times \frac{(B-G)}{(Max-Min)} + 180\ (if\ Min = R) \\ 60 \times \frac{(R-B)}{(Max-Min)} + 300\ (if\ Min = G) \end{cases} \quad (1)$$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

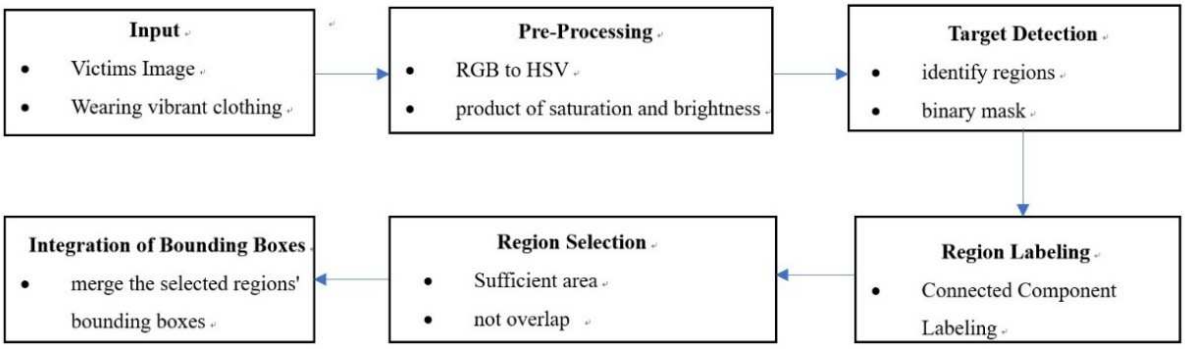


Figure1.The System Flowchart

$$S = \begin{cases} 0, & if\ Max = 0 \\ \frac{Max-Min}{Max} \times 255, & otherwise \end{cases} \quad (2)$$

$$V = Max \times 255 \quad (3)$$

Min is the color channel in RGB where the minimum value is found. It is used to determine the basic color tone of the color. Max is the color channel in RGB where the maximum value is found. It is used to calculate H and determine the color's saturation and brightness.

B. Feature Extraction

H. Xu et al [8] designed a feature extraction method of ophthalmic image. The feature blocks of ophthalmic image are divided, and the window of threshold selection is determined according to the comparison of pixel and threshold range.

The paper uses the product of saturation and brightness of an image as a feature, where the highlighted region represents the location of the victim in the image. This characteristic is characterized by higher feature values

A threshold is set based on this feature map. Pixels with a product greater than the threshold are set to 1 with red color, while pixels below the threshold are set to 0 with the green color, creating a binary mask showing in Figure2. By labeling the connected regions in the mask, the pixels belonging to the same region are grouped together and can be segmented into different areas.


Figure2. Binary Mask

C. Expansion Factor

H. Tian et al [9] mentioned the goal of expansion step is to compute a support plane for each pixel in the image. The idea behind expansion is the fact that the neighboring pixels are very likely to have similar planes. And is able to expand its good plane to its neighboring pixels especially for high resolution images because of the local similarity.

The expansion factor is a parameter used to adjust the size of the detected regions during the victim search algorithm. By increasing the expansion factor, the algorithm enlarges the bounding boxes around the detected regions. This expansion helps to include more surrounding pixels and provide a larger margin of error, increasing the chances of capturing the complete area of the victim.

D. Connected Component Labeling

J. G. Pandey et al [10] present a field-programmable gate array (FPGA) platform based approach for implementing an efficient and improved two-scan equivalence-based connected component labeling algorithm. With the development J. G. Pandey et al [10] present, we can further validate the feasibility of the connected component labeling algorithm and assess its accuracy for application in hardware showing in Figure3.

To remove a lot of noise, the "bwlabel" function is used to calculate the connected components of a binary image, identifying and labeling connected regions showing in Figure4. The binary mask (MASK) is generated based on the product of saturation and brightness and a threshold value called "vividness_threshold". MASK marks which pixels are considered to have a sufficiently high saturation (color vividness) and brightness (luminance), i.e., falling above the threshold. Since victims are defined as relatively larger objects, the "bwlabel" function is used for connected region analysis, where each connected group of pixels is assigned the same label, and the number of labels corresponds to the total number of these connected regions.

After labeling each connected region, we check whether its area is greater than "min_area" and examine if the new connected region overlaps with the previously identified regions. If the overlap exceeds a certain threshold (in this case, 50%), it is considered the same region. Otherwise, the new connected region is added to the list of already identified regions, and those with sufficient area are selected for expansion. The expanded regions are merged into a single bounding box.

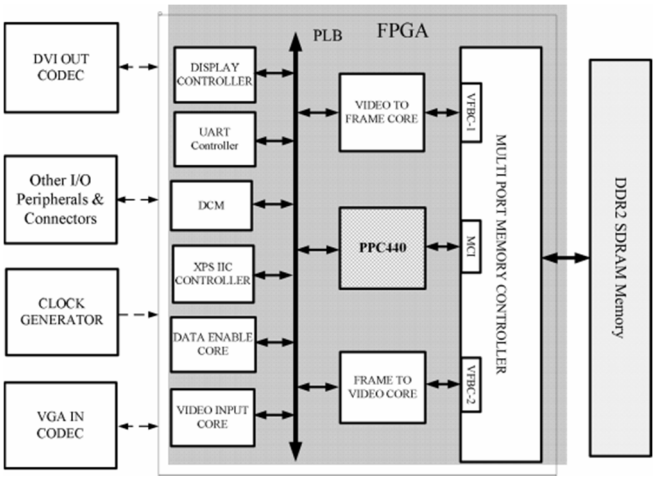


Figure3. System architecture for image acquisition for connected component labeling algorithm[8]

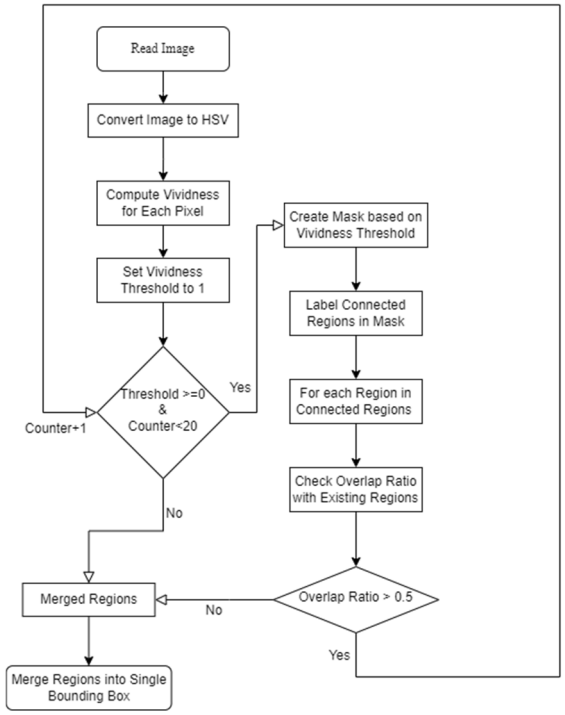


Figure4. Identifying And Labeling Connected Regions

E. *Intergration Of Bounding Box*

We have identified multiple areas with high saturation (color vividness) and brightness in our findings. We generated corresponding bounding boxes for these areas. However, there may be some scattered or overlapping boxes, leading to judgment errors. Therefore, we attempted to merge these boxes into a larger one to highlight the overall region.

Calculated the average position of all box centers and the distance from each box center to the average position are obtained by the following formula:

$$Sum = \sum_{x=x_{min}}^{x_{max}} \sum_{y=y_{min}}^{y_{max}} I(x,y) \quad (4)$$

$$Mean = \frac{1}{NumPixels} \sum_{x=x_{min}}^{x_{max}} \sum_{y=y_{min}}^{y_{max}} I(x,y) \quad (5)$$

I(x, y) represents the pixel value at position (x, y) in the image and "NumPixels" represents the number of pixels within the bounding box.

Then, we excluded boxes that were far from the average position and retained only those close to it. This was done to remove scattered boxes and keep only those in concentrated areas. We used the 'insertShape' function to draw the merged large box on the original image. Additionally, we inserted text next to the large box to display the value of the average threshold. This text serves to provide additional information about the average value of the threshold for identifying these areas with high saturation (color vividness) and brightness (luminance).

III. EXPERIMENTAL RESULTS

Ž. Marušić et al [11] examine the effectiveness of visual search of human observers for the purpose of finding missing/lost persons on aerial imagery. On average they were able to detect 84 target objects and the average recall was 80.43%. Significantly better results were achieved in precision that takes into account false alarms or false detections. Average precision was 90.98% shows in Figure5.

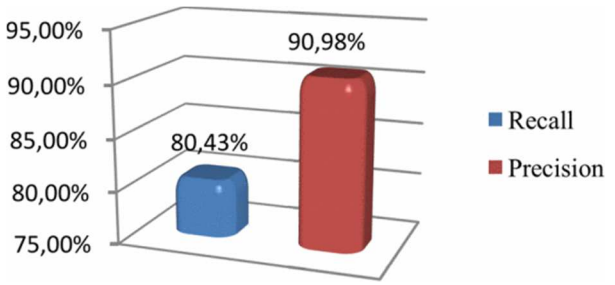


Figure5. Average recall and precision[9]

We present the experimental results of the proposed image-based target detection and integration algorithm for victim search. We evaluated the performance and accuracy of the algorithm using various image datasets and scenarios. The following are our experimental results.

In Figure 5, it is clearly visible to the naked eye that the victim standing in the center of the image is wearing vibrant clothing, making it relatively easier for search and rescue operations as there are no significant obstacles obstructing the rescuers' line of sight. Conversely, in Figure 6, the victim is more likely to be overlooked by the rescuers due to the presence of numerous interfering factors that affect their judgment. However, the experimental results accurately capture the vibrant clothing worn by the victim.



Figure6. Mountain Accident Victim



Figure7. Mountain Accident Victim

IV. CONCLUSION

In this paper, the position of victims in the image is identified using MATLAB. The test results show that the vibrant clothing of the victims can be accurately recognized. However, in cases where the surrounding environment is too complex or the surroundings are more vibrant than the clothing of the victims, the recognition may be distorted. furthermore, due to a lack of photo support and an insufficient number of victim photos available, there is not enough data to populate all the elements of the confusion matrix, including true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

In the future, training on the images will be conducted to achieve higher accuracy in identification.

REFERENCES

[1] G. Seeja, A. S. A. Doss and V. B. Hency, "A Novel Approach for Disaster Victim Detection Under Debris Environments Using Decision Tree Algorithms With Deep Learning Features," in IEEE Access, vol. 11, pp. 54760-54772, 2023, doi: 10.1109/ACCESS.2023.3281461.

[2] Xuewen Wang, Xiaoqing Ding and Changsong Liu, "Character extraction and recognition in natural scene images," Proceedings of Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 2001, pp. 1084-1088, doi: 10.1109/ICDAR.2001.953953.

[3] W. Hoshino, J. Seo and Y. Yamazaki, "A study for detecting disaster victims using multi-copter drone with a thermographic camera and image object recognition by SSD," 2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Delft, Netherlands, 2021, pp. 162-167, doi: 10.1109/AIM46487.2021.9517524.

[4] H. Liu et al., "The design of air-space integrative calamity information analysis and rescue system," 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Shenyang, China, 2015, pp. 1997-2001, doi: 10.1109/CYBER.2015.7288254.

[5] D. Yanguas-Rojas, G. A. Cardona, J. Ramirez-Rugeles and E. Mojica-Nava, "Victims search, identification, and evacuation with heterogeneous robot networks for search and rescue," 2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC), Cartagena, Colombia, 2017, pp. 1-6, doi: 10.1109/CCAC.2017.8276486.

[6] T. Okuyama, T. Gonsalves and J. Upadhay, "Autonomous Driving System based on Deep Q Learnig," 2018 International Conference on Intelligent Autonomous Systems (ICoIAS), Singapore, 2018, pp. 201-205, doi: 10.1109/ICoIAS.2018.8494053.

[7] C. Liu, X. Lu, S. Ji and W. Geng, "A fog level detection method based on image HSV color histogram," 2014 IEEE International Conference on Progress in Informatics and Computing, Shanghai, China, 2014, pp. 373-377, doi: 10.1109/PIC.2014.6972360.

[8] H. Xu, "Feature extraction of ophthalmic image based on machine learning," 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 2021, pp. 613-616, doi: 10.1109/TOCS53301.2021.9688641.

[9] H. Tian and F. Li, "Accurate multi-view stereo by selective expansion," 2017 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Copenhagen, Denmark, 2017, pp. 1-4, doi: 10.1109/3DTV.2017.8280405.

[10] J. G. Pandey, A. Karmakar, A. K. Mishra, C. Shekhar and S. Gurunarayanan, "Implementation of an improved connected component labeling algorithm using FPGA-based platform," 2014 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 2014, pp. 1-6, doi: 10.1109/SPCOM.2014.6983911.

[11] Ž. Marušić, D. Zelenika, T. Marušić and S. Gotovac, "Visual Search on Aerial Imagery as Support for Finding Lost Persons," 2019 8th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2019, pp. 1-4, doi: 10.1109/MECO.2019.8760201.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Point Cloud Inpainting Based on Delaunay Triangulation

Yu-Lin Liu 1, He-Sheng Chou 2, Ming-Zhan Lee 3, Mei-Ling Chan 4, Ting-Lan Lin 5, Chiung-An Chen 6,and Shin-Lun Chen 7*

Department of Electronic Engineering, Chung Yuan Christian University, Taiwan
E-mail: g11176050@cycu.edu.tw Tel: +886-3-265-4644
Department of Electronic Engineering, Chung Yuan Christian University, Taiwan
E-mail: g10776018@cycu.edu.tw Tel: +886-3-265-4644
Department of Electrical Engineering National Cheng Kung University, Taiwan
E-mail: q16101282@gs.ncku.edu.tw Tel: +886-3-265-4644
College of Physical Education, Jiaying University, Meizhou City College of Physical Education
E-mail: Lynn202207017@jyu.edu.cn
Department of Electronic Engineering National Taipei University of Technology, Taiwan
E-mail: tinglan@ntut.edu.tw Tel: +886-2-2771-2171 #2278
Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 243303, Taiwan
E-mail: joannechen@mail.mcut.edu.tw Tel: +886-2-2908-9899 #5219
*Department of Electronic Engineering, Chung Yuan Christian University, Taiwan
E-mail: chrischen@cycu.edu.tw Tel: +886-3-265-4610

***Abstract*— Point cloud is a point-based representation of three-dimensional data used to describe the geometry of objects or scenes However, due to its large data volume, a point cloud needs to be compressed to reduce the data volume. After point cloud compression, some points may be lost, causing the loss of edge details and the appearance of point cloud holes and outliers, leading to a significant subjective quality decrease and affecting point cloud development in practical applications. In this paper, an effective point cloud repair method is proposed. First, use the mean filter to remove outliers, so that the subsequent point-filling algorithm can avoid errors caused by outliers. Secondly, the point cloud is analyzed by the Delaunay triangulation method. Perform statistical analysis on the areas of all triangles. If an individual triangle has an area larger than the average, it is marked. After comparing the triangles, the coordinates of all the holes in the point cloud can be found, so as to identify the holes in the point cloud. In the hole-filling strategy, the interpolation method is used to average the coordinates and color information of the three vertices of the triangle to obtain the filling point, that is, to complete the restoration of the point cloud data. When evaluating the results, in addition to PSNR, the GraphSim index proposed by Q. Yang was also referred to [1]. This indicator is closer to the subjective evaluation of people and better reflects the subjective evaluation of people. According to the experimental results, the GraphSim of the proposed method can reach 0.363, which significantly improves the point cloud data after V-PCC compression [2].**

I. INTRODUCTION

Point cloud has become an important research direction in computer vision, robotics, autonomous driving, and other fields. The subjective visual quality of the point cloud is crucial in its applications. Therefore, point cloud repair has become a popular research topic. Two common reasons for point cloud holes are identified. Firstly, specialized scanning devices usually acquire point cloud data using 3D laser scanning to obtain 3D data on object surfaces. However, occlusions or scanning surface reflection errors can cause holes in the collected point cloud [3]. Secondly, point clouds typically have a large amount of data and require compression, such as the Motion Picture Experts Group (MPEG) established video-based point cloud compression (V-PCC) standard [2]. The compression standard employed in this study involves storing point clouds in 2D image format to reduce data size, and decoding them back to 3D point clouds when necessary. However, this process will cause a large loss of edge details, which will lead to the loss of information at the edge.

In recent years, several methods have been proposed for repairing holes in point clouds. For instance, Keming Cao et al. [4] utilized V-PCC to annotate the edges of 2D images and then mapped them back to the 3D point cloud to obtain the hole positions. They then mapped the holes onto the XYZ plane and performed a closing operation on each of the three planes to repair the holes. This paper refers to the hole-filling strategy used by Erjie Nie [5], who preprocessed the point cloud using

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Delaunay triangulation. Instead of interpolating the midpoint of adjacent points as in the reference paper, this paper calculated the sum of the areas of all triangles and used the average area as the threshold. When the triangle area is larger than the average area, it is defined as a hole-filling point. The centroid coordinates of the triangle are used as the interpolation point. The process repeats until the area is smaller than the average, indicating that hole filling is complete. This method solves the problem of sparse insertion points in the repair process.

Research structure flow chart: First, denoise the point cloud, and then perform Delaunay Triangulation and interpolation calculation. Finally, analyze the experimental results of the point cloud Fig. 1.

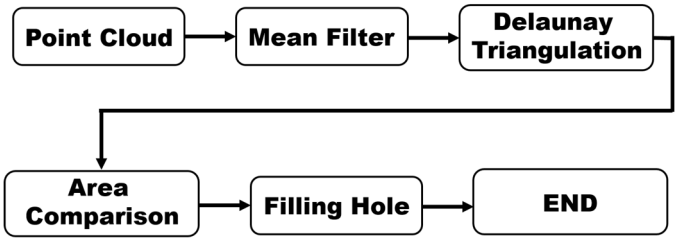


Fig. 1 Research structure Flow chart

II. EXPERIMENTAL METHOD

A. *Denoising for Point Cloud*

The point cloud with holes in this paper was obtained by referring to the method in [6], which uses the MPEG V-PCC's r1 file. Compared to the r5 file, which has better subjective visual performance after reconstruction, the geometry QP and texture QP values in the r1 file were increased from 16 and 22 to 32 and 42, respectively. Increasing the QP values means that the distortion of point cloud compression increases, resulting in poorer performance of details. As shown in Fig. 2 there are significantly many outlier points on the edges of the point cloud, and these outlier points will cause the Delaunay triangulation effect to deteriorate in the subsequent processing. Therefore, point cloud denoising preprocessing is needed to filter out the outlier points.

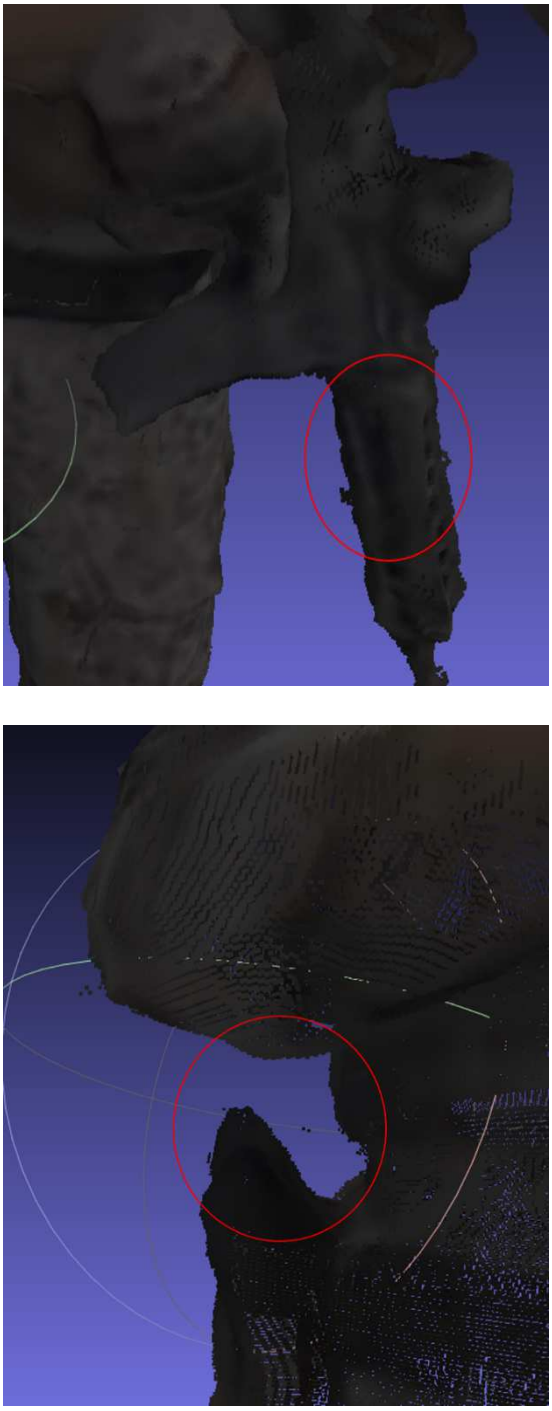


Fig. 2 Point cloud after V-PCC r1 file

Point cloud denoising is a highly popular area in computer vision. The main filtering methods can be classified into several categories[7]. The first type is based on neighborhood analysis, which estimates the similarity of a point and its neighbors. Examples include bilateral filtering, mean filtering, and non-local mean filtering. The second type is based on statistics, such as local median filtering, local averaging, and Bayesian statistics. The third type is based on partial differential equations (PDEs). Since this paper only aims to filter out outlier points instead of complex Gaussian noise, we adopt mean filtering. Compared to other methods, it has lower

computational complexity and is faster. By performing a K-Nearest-Neighbors (KNN) search to find k nearest points(P_i) of the target point P, we obtain the positions of k points and then average them by adding their coordinates. Finally, we replace the coordinates of the target point P with the calculated average value (1). This process moves the outlier points closer to the main body of the point cloud. As shown in Fig. 3, the denoised point cloud has significantly fewer outlier points compared to Fig. 2.

$$P(x) = mean(\sum_{i=1}^{k} P_i(x))$$
$$P(y) = mean(\sum_{i=1}^{k} P_i(y)) \quad (1)$$
$$P(z) = mean(\sum_{i=1}^{k} P_i(z))$$



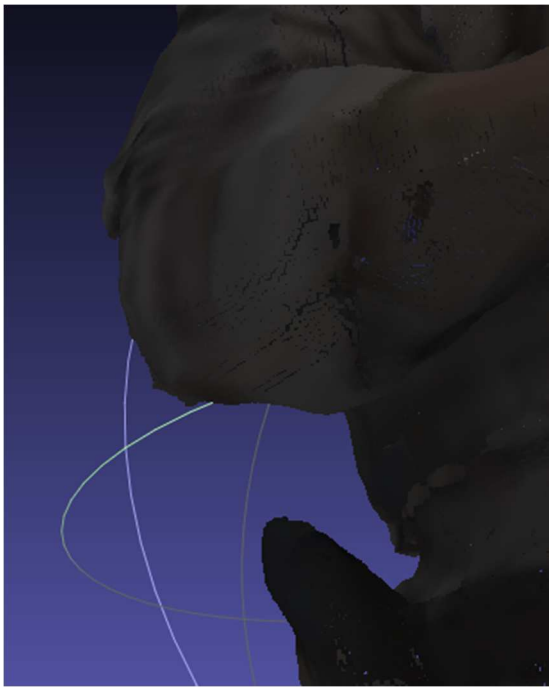


Fig. 3 Point cloud after mean filter

B. Point Cloud Delaunay Triangulation

The purpose of this research is to obtain the geometric relationship between points through the Delaunay triangulation of the point cloud. Unlike 2D images with inherent order, 3D point clouds are disordered, so a method is needed to obtain the computational correlation. As mentioned in [6], slicing the point cloud along the y-axis to obtain a 2D image facilitates computation. In this study, Delaunay triangulation is adopted, which is an important preprocessing technique for numerical analysis and computer graphics. Delaunay triangulation has two advantages: first, it maximizes the minimum angle, resulting in a triangle mesh that is closest to being regular; second, no four points can be cocircular. This method is particularly effective for dense point clouds such as Soldier, as shown in Fig. 4 which displays the results of Delaunay triangulation of the Soldier point cloud.

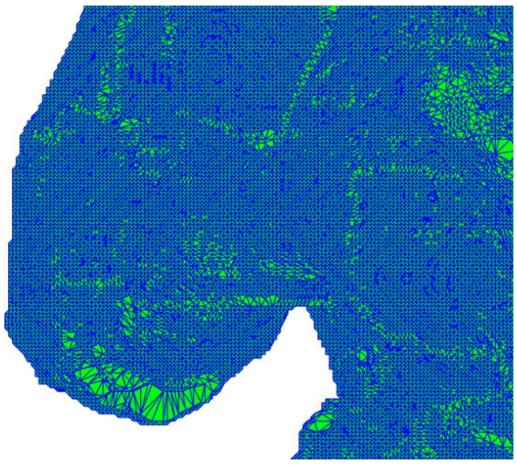


Fig. 4 Point cloud after Delaunay Triangulation

C. Hole Detection

After converting the point cloud into a large number of three-point matrices through Delaunay triangulation, the next step is to detect holes. For a triangle with three points A(x1,y1,z1), B(x2,y2,z2), and C(x3,y3,z3), this paper calculates the area of each triangle and sums them (2), then divides by the total number of triangles to obtain the average area. This value is used as a threshold u and individual triangles with areas greater than the threshold are marked, allowing the coordinates of the holes in the point cloud to be obtained. Fig. 5 shows the hole triangles calculated after comparison.. After multiple attempts, a threshold value of 2 times the average area was found to be most effective, as it can identify large holes while avoiding marking holes that are too small compared to the average.

$$Mean\ Area = Mean(\sum_{i=1}^{n} \frac{1}{2} \left|\overrightarrow{A_iB_i} \times \overrightarrow{A_iC_i}\right|) \quad (2)$$

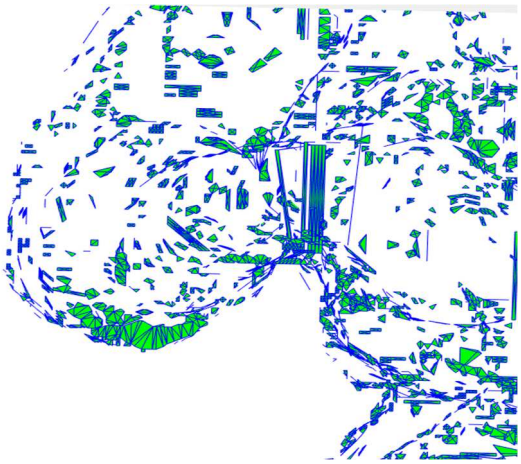


Fig. 5 Hole triangle after area comparison

D. *Hole Inpainting*

In the hole-filling strategy, the coordinates and color information of the three vertices of each triangle are averaged using interpolation to obtain information about the hole-filling points. After adding the hole-filling points, the triangle areas are recalculated and compared with the average area. If any triangle has an area greater than the threshold value, interpolation is continued and the iteration is repeated until all triangles have an area smaller than the threshold, which is considered as the end of hole-filling. By using the above algorithm, the damaged point cloud data can be repaired, and a visually better point cloud can be obtained subjectively. The result is shown in Fig. 6

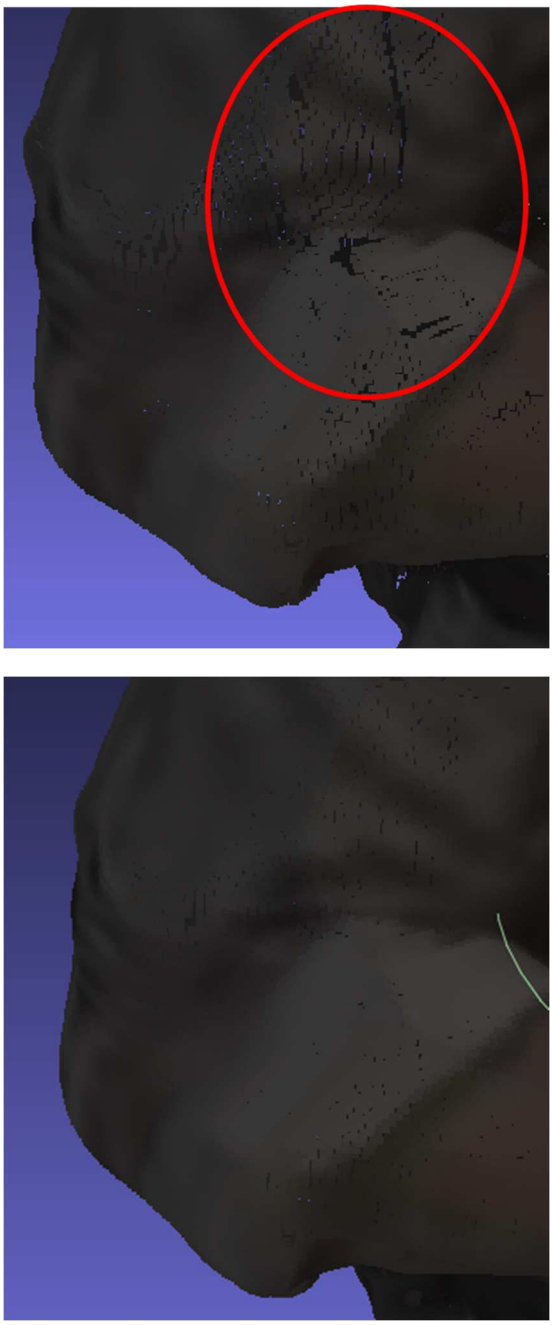


Fig. 6 The inpainting result is compared with the point cloud after denoising in the previous step

III. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the algorithm proposed in this paper, the original data compressed by V-PCC and the data after the algorithm in this paper are calculated and evaluated. In addition to comparing the traditional PSNR, we also referred to the GraphSIM index adopted by [1] for evaluation. The GraphSIM metric takes into account that the human visual system is more responsive to contours and edges. Unlike PSNR, which compares individual points, GraphSIM better balances the expression of structural changes and subjective visual perception. In other words, GraphSIM has significant effects in evaluating subjective visual quality. Table

1 shows the experimental results. Although our proposed algorithm leads to a marginal decrease in PSNR, it notably attains a GraphSIM index of 0.363, surpassing the reference paper by 0.012. This index, being more representative of subjective evaluation, highlights the significant enhancement in visual perception resulting from our approach. As shown in Fig. 6 in the previous section, the proposed algorithm can significantly improve the completeness of the point cloud data in subjective observation.

Table 1: Comparison table with reference studies.

	V-PCC r1 [4]	Proposed
PSNR(Point to Point)	65.9417	64.588
PSNR(Point to Plane)	67.6579	67.4547
Y-PSNR	29.4043	28.86
U-PSNR	43.0417	43.0816
V-PSNR	43.2131	43.2503
GraphSIM	0.349	**0.361**

IV. CONCLUSION

This paper addresses edge damage and outliers that occur during V-PCC by removing outliers and repairing holes. Experimental results show that although there is a slight decrease in PSNR, there is a significant improvement in the subjective indicator GraghSIM. Since the Delaunay triangulation is used in this paper's analysis, the algorithm's performance may decrease when the hole is too large and has a high gradient. Therefore, the proposed method is more suitable for dense point clouds, where the gradient of holes is smaller and direct triangulation interpolation works better. Overall, this paper proposes a new process for point cloud repair, and we will continue to work towards ensuring that the repair method is more complete and improves human subjective visual evaluation in future research.

V. ACKNOWLEDGMENT

First of all, the author would like to thank the organizer for organizing this conference and thank He-Sheng Chou and Ming-Zhan Lee for their guidance. Most importantly, I would like to express my thanks to the three professors Shin-Lun Chen, Ting-Lan Lin, and Mei-Ling Chan for their guidance and suggestions.

REFERENCES

[1] Q. Yang, Z. Ma, Y. Xu, Z. Li, and J. Sun, “Inferring Point Cloud Quality via Graph Similarity,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3015–3029, Jun. 2022, doi: 10.1109/TPAMI.2020.3047083.

[2] ISO/IEC FDIS, 23090-5, "Coded Representation of Immersive Media Part 5: Visual Volumetric Video-Based Coding (V3C) and Video-Based Point Cloud Compression (V-PCC)", 2020.

[3] C. Dinesh, I. V. Bajic, and G. Cheung, “Exemplar-based framework for 3D point cloud hole filling,” 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4, Dec. 2017, doi: 10.1109/VCIP.2017.8305070.K. Elissa, “Title of paper if known,” unpublished.

[4] K. Cao and P. Cosman, “Denoising and Inpainting for Point Clouds Compressed by V-PCC,” IEEE Access, vol. 9, pp. 107688–107700, 2021, doi: 10.1109/ACCESS.2021.3102029.

[5] E. Nie, J. Li, R. Zhang, L. Cao, L. Sun, and H. Liu, “Three dimensional point cloud hole repairing strategy for binocular stereo reconstruction,” in 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Feb. 2017, pp. 2456–2461. doi: 10.1109/ROBIO.2017.8324788.

[6] H.-B. Bu, C.-H. Lin, C.-F. Liang, T.-L. Lin, and Y.-S. Chiou, “Quality enhancement for point cloud using morphology methods,” in 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Sep. 2021, pp. 1–2. doi: 10.1109/ICCE-TW52618.2021.9603255.

[7] A. Yerokhin, V. Semenets, A. Nechyporenko, O. Turuta, and A. Babii, “F-transform 3D Point Cloud Filtering Algorithm,” in 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Aug. 2018, pp. 524–527. doi: 10.1109/DSMP.2018.8478581.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Dense Three-Dimensional Color Reconstruction for Large-Scale Outdoor Scenes

Zixiao Liu*, Sheng Guo† and Man-On Pun*
*The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
†Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China

***Abstract*—Accurate three-dimensional (3D) color models play a crucial role in various emerging applications, such as geographic surveying and mapping, smart cities, and digital twin cities. To achieve high-fidelity representations of large-scale scenes, recent advancements in machine learning techniques and LiDAR-based Simultaneous Localization and Mapping (SLAM) have been instrumental. Leveraging these cutting-edge technologies, this paper presents a sensor-fusion framework that integrates a solid-state LiDAR, an inertial measurement unit (IMU), and a monocular camera. Additionally, a point cloud upsampling technique is introduced into LiDAR SLAM to enhance the density of the point clouds. Within this framework, the LiDAR-IMU odometry accurately estimates the positions and poses of the collected point clouds. Simultaneously, synchronized images are captured and utilized to incorporate texture information into the point clouds. The proposed framework demonstrates its capability to generate highly detailed and dense 3D color models for large-scale outdoor scenes, all within a limited on-site scanning time. Extensive experimental results validate the effectiveness and efficiency of the proposed approach.**

I. Introduction

Accurate reconstruction of large-scale outdoor scenes in three-dimensional (3D) color models is essential for various applications, including geographic surveying, urban planning, smart cities, digital twin cities, and environmental protection. However, outdoor scenes present unique challenges due to their complex and dynamic nature. One of the challenges arises from the sheer size and complexity of outdoor environments. These scenes often consist of numerous objects with diverse shapes and textures, making it difficult to capture and reconstruct them accurately. Moreover, outdoor scenes are subject to changing illumination conditions, which can affect the color and appearance of objects. Additionally, the presence of occlusions caused by trees, buildings, or other obstacles further complicates the reconstruction process.

To address these challenges in dense 3D color reconstruction, researchers have made significant progress by developing advanced algorithms and techniques. In the literature, two main approaches have been proposed: image-based and LiDAR-based methods. In image-based methods, [1] focused on 3D color reconstruction using images, including surround photos of objects or buildings and aerial drone remote sensing images. However, these methods have demonstrated poor performance in larger-scale scenes, limiting their practicality. Recent developments introduced NICE-SLAM [2], which combines Simultaneous Localization and Mapping (SLAM) with Neural Radiance Fields (NeRF) and has shown success in indoor scene reconstruction. Another method, SensatUrban [3], specifically designed for large-scale 3D reconstruction of buildings and cities using aerial drone images, has also been reported. Despite these advancements, image-based methods face challenges in practical deployment, such as the complex planning process for data collection and high computational complexity.

On the other hand, recent breakthroughs in solid-state LiDAR technology have enabled the generation of dense and accurate point clouds for 3D urban reconstruction. For example, the CamVox framework [4] integrates the Livox solid-state LiDAR with a monocular camera using the ORB-SLAM2 Visual SLAM technique. Similarly, R3LIVE [5] utilizes LiDAR Odometry and Mapping (LOAM) and its variants to render model textures. However, the LiDAR-based approach faces challenges in generating realistic 3D models due to the limited scanning time windows in practical applications like post-disaster rescue missions. To tackle this, pioneering works have integrated LiDAR SLAM with deep learning techniques [6] to enhance point cloud density using machine learning-based methods.

In summary, the combination of LiDAR-based approach and point cloud upsampling techniques holds immense promise for achieving large-scale 3D color reconstruction. This approach offers enhanced accuracy in representing geometric structures and effectively captures intricate textures, resulting in the creation of realistic and high-quality 3D city models.

This paper introduces a novel sensor-fusion system that combines both hardware and software components to address the challenges in generating denser 3D color models for large-scale outdoor scenes. The main contributions of this paper are as follows:

1) A novel sensor-fusion system that integrates a solid-state LiDAR, an inertial measurement unit (IMU), and a monocular camera is proposed. This combination of hardware components allows for comprehensive data capture, enabling the generation of more detailed 3D

This work was supported in part by the Shenzhen Science and Technology Innovation Committee under Grant No. JCYJ20190813170803617, the National Key R&D Program of China under grant 2018YFB1800800, the Basic Research Project under Grant HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Shenzhen Outstanding Talents Training Fund 202002, Guangdong Research Projects under Grant 2017ZT07X152 and 2019CX01X104, the Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001, the National Natural Science Foundation of China under Grant 41801323. *(Corresponding authors: Man-On Pun (SimonPun@cuhk.edu.cn); Sheng Guo (guogis@126.com)).*

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

color models.

2) A novel method for upsampling point clouds is introduced. This technique involves dividing the point cloud into patches and representing these patches as neural fields. By integrating these local neural fields into a global surface, the proposed technique achieves upsampling of arbitrary ratios.
3) To validate the effectiveness of the proposed system, a proof-of-concept prototype platform is constructed and tested on the campus of the Chinese University of Hong Kong, Shenzhen. Through extensive data collection and scene reconstruction tests, the experimental results demonstrate that the proposed system can obtain dense 3D color models of large-scale outdoor scenes within limited scanning time windows.

II. Proposed Hardware Framework

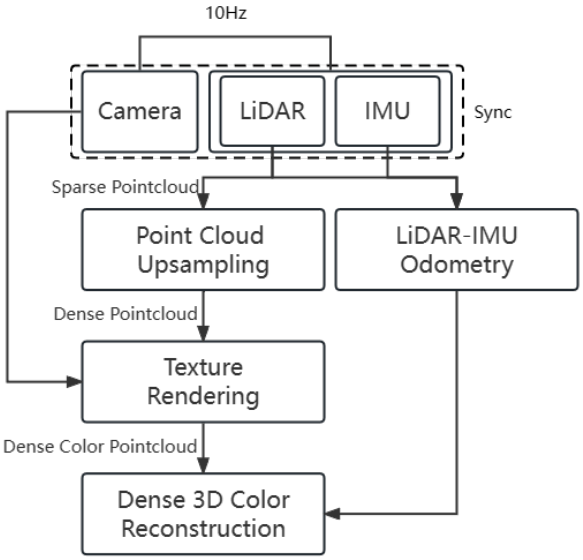


Fig. 1: The proposed framework.

The block diagram of the proposed framework is shown in Fig. 1. Compared to the Visual SLAM backbone of CamVox, the proposed framework is based on LiDAR SLAM that is less susceptible to ambient light, and has higher precision and smaller cumulative errors. The three main components of the proposed framework are the LiDAR-IMU odometry, the texture rendering system and the point cloud upsampling system, respectively. In the proposed framework, data are first collected by the camera and the IMU-empowered LiDAR. After that, the localization information of the point cloud, such as pose and orientation, is estimated using LOAM. In the meantime, the point cloud upsampling system performs upsampling on the point cloud to obtain denser point cloud. After the dense point cloud is generated, the texture rendering module renders the color of each point by matching each point in the point cloud with its counterpart in the image using intrinsic and extrinsic parameters. Finally, the proposed framework reconstructs the dense 3D color model by exploiting the resulting dense colored point cloud and the localization information.

A. LiDAR-IMU-Camera Calibration

1) Intrinsic Calibration: To fuse the images from the camera with the point cloud from the LiDAR, it is critical to align the coordinate systems of the camera and the LiDAR. This paper utilized a chessboard to calibrate the intrinsic and distortion parameters of the monocular camera. More specifically, more than 20 photos of the chessboard from various angles were first taken, covering all the fields of view (FOV). These photos together with the actual size of the black and white squares on the chessboard were used by the cameraCalibrator module from MATLAB to derive the intrinsic and distortion parameters.

We denote by $\mathbf{P}$ and P_d the intrinsic matrix of the camera and the distortion parameters, respectively. $\mathbf{P}$ takes the following form as shown in Eq. (1).

$$\mathbf{P} = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where $\{f_u, f_v\}$, $\{c_u, c_v\}$ are the focal length set in pixels and the optical center set along the u and v axes in pixels, respectively.

Furthermore, P_d contains five parameters intended for distortion correction as shown in Eq. (2).

$$P_d = (k_1, k_2, p_1, p_2, k_3), \tag{2}$$

where k_i for $i = 1, 2, 3$ are the radial distortion parameters while p_j for $j = 1, 2$ are the tangential distortion parameters. By exploiting P_d, image distortion in the photos can be corrected. Next, the extrinsic matrix will be introduced to perform coordinate system conversion between the camera and the LiDAR.

2) Extrinsic Calibration: We denote by $\mathbf{T}$ the extrinsic matrix that transforms the coordinate system of the LiDAR to that of the camera. The calibration method reported in [7] is employed to derive $\mathbf{T}$ as it is capable of automatic extrinsic calibration in targetless environments by extracting depth-continuous edges and aligning point cloud projections and image edges. Let $\mathbf{X} = [x, y, z, 1]^T$ and $\mathbf{Y} = [u, v, 1]^T$ stand for the coordinates of a point in the coordinate systems of the LiDAR and the camera, respectively. Furthermore, we denote by H_{depth} the z-axis value of a given point in the camera coordinate system, i.e. the depth of this point in the image. The extrinsic matrix $\mathbf{T}$ consists of a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$ as shown in Eq. (3).

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}, \tag{3}$$

where

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}. \tag{4}$$

Given a pair of intrinsic matrix $\mathbf{P}$ and extrinsic matrix $\mathbf{T}$, 3D point cloud can be projected onto the corresponding 2D image plane. An extrinsic calibration example is shown in Fig. 2 in which the point cloud is projected onto the image plane with edges being well aligned.

$$H_{depth}\mathbf{Y} = H_{depth}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (5)$$

$$= \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (6)$$

$$= \mathbf{PTX}. \quad (7)$$

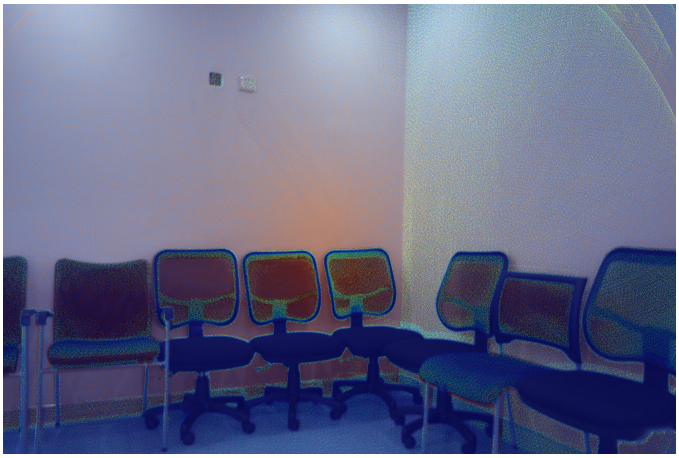


Fig. 2: Visual illustration of an extrinsic calibration example.

B. LiDAR-IMU Odometry

In this paper, the solid-state LiDAR called Livox Avia from DJI is used in our system. We utilize the software package provided by DJI to obtain pose and orientation information from the point cloud. More specifically, the odometry uses an iterative pose optimization method to estimate the LiDAR pose. When an operator performs the feature extraction and selection, it extracts surface and line features using an iterative closest point (ICP) algorithm. Let P_i be the i-th nearest points of P_w from the set of feature points of the w-th frame. We can calculate the edge-to-edge and plane-to-plane residuals according to Eq. (8) and Eq. (9).

$$r_e(P_w) = \frac{|(P_w - P_5) \times (P_w - P_1)|}{|P_5 - P_1|}, \quad (8)$$

$$r_p(P_w) = \frac{(P_w - P_1)^T((P_3 - P_5) \times (P_3 - P_1))}{|(P_3 - P_5) \times (P_3 - P_1)|}. \quad (9)$$

Using the internal IMU of Livox Avia, the odometry is able to eliminate the rotation distortion.

C. Texture Rendering

Next, we proceed to generate dense point cloud. Given the coordinate $\mathbf{X} = [x, y, z, 1]^T$ of a point in the point cloud coordinate system, the coordinate $\mathbf{Y} = [u, v, 1]^T$ of this point in the image coordinate system can be derived from Eq. (5). Since u and v may not be integers, the system will use the nearest integer point as its projection point. In this way, we are able to find the corresponding pixel of this point as well as its color. After each point in the dense point cloud is indexed by the coordinates (x, y, z), the final step is to add the (r, g, b) color information to each point in the point cloud. The resulting data point is recorded in the format of (x, y, z, r, g, b). In order to obtain accurate color information, the LiDAR and camera are synchronized to collect data at a constant frequency, e.g. 10 Hz. Thus, all data samples in the point cloud and images are time-stamped with the dense point cloud inheriting the timestamp of the raw point cloud from the LiDAR, which enables us to match point cloud and images of similar timestamps by using the nearest time registration algorithm. The texture rendering system matches each data point in the point cloud with one pixel in the image before adding the color information (r, g, b) to the matched data point. After rendering the texture of all points in one frame, we can obtain a colored point cloud frame. The rendering procedures discussed above are then repeated for each pair of point cloud and images to generate a sequence of dense colored point cloud frames.

D. Dense 3D Color Reconstruction

Finally, the proposed framework splices the resulting sequence of dense color point clouds into a global map. Let $\mathbf{X_w} = [x_w, y_w, z_w]^T$ and $\mathbf{X_l} = [x_l, y_l, z_l]^T$ be the coordinates in the global and local coordinate systems, respectively. By exploiting the pose and orientation information obtained from the odometry, we can convert the coordinate of a given point in the local coordinate system to the global coordinate system using Eq. (10).

$$\mathbf{X_w} = \mathbf{QX_l} + \mathbf{m}, \quad (10)$$

where $\mathbf{Q}$ is the rotation matrix and $\mathbf{m}$ is the transition vector given as follows:

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}. \quad (11)$$

By performing the transformation in Eq. (10) on each point in every frame of the dense colored point cloud, the resulting colored points are properly positioned in a global map, i.e. the dense 3D color map.

III. Point Cloud Upsampling

A. Overall Structure

Inspired by the Neural Points reported in [8], the proposed point cloud upsampling module consists of three main sub-modules as shown in Fig. 3.

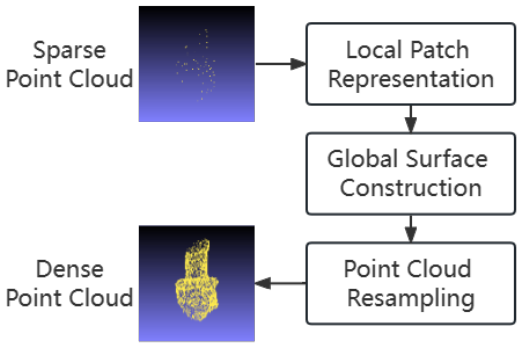


Fig. 3: The structure of the point cloud upsampling module.

B. Local Patch Representation

First, the local patch representation sub-module divides the input, i.e. the sparse point cloud, into patches. Specifically, we denote by $X = \{x_i\}_{i=i}^{I}$ the input, where x_i is the center point. The input is divided into overlapping patches $\{P_i\}_{i=i}^{I}$. For each center point x_i, together with its neighbor points, they are processed by the DGCNN backbone for feature extraction. The DGCNN contains several EdgeConv layers equipped with dynamic graph updates. The resulting local feature f_i is a part of the input into the neural field. In the meantime, the patches are mapped to 2D parametric domain $\{D_i\}_{i=i}^{I}$, which will be used in neural field representation. This mapping is bidirectional and is denoted by $\phi_i : D_i \to P_i$ and $\phi_i^{-1} : P_i \to D_i$. For each 2D point (u_r, v_r), the normal of its corresponding 3D point x_i^r is defined as Eq. (12).

$$Normal_{x_i^r} = \nabla_u \phi_i \times \nabla_v \phi_i. \tag{12}$$

Subsequently, the local patches are converted into local neural fields by using an MLP-based network with the ReLU activation layers to represent ϕ_i. For each 2D parametric coordinates (u_r, v_r) passed into ϕ_i, the local patch representation sub-module performs the positional encoding γ on (u_r, v_r). The resulting positional code and the local feature are fed into the neural implicit function Φ. Mathematically, Φ is shown in Eq. (13).

$$\Phi((u_r, v_r), f_i, x_i) = x_i + \theta_\Phi(\gamma(u_r, v_r) \oplus f_i), \tag{13}$$

where θ_Φ is the MLP stated above. In this way, these discontinuous patches are represented as local neural fields that are continuous.

C. Global Surface Construction

After representing the local neural fields, the global surface construction sub-module combines all these local neural fields into one global continuous surface. Specifically, each input point x in the local neural fields needs to be mapped into the global coordinate system, i.e. the global continuous surface. First, for each x, this sub-module finds its nearest center points denoted by x_k in $\mathcal{X} = \{x_i\}_{i=i}^{I}$. According to the distance between x and x_k, a weighting coefficient c_k can be computed as shown in Eq. (14).

$$c_k = e^{-\alpha_1 ||x - x_k||_2^2}. \tag{14}$$

After obtaining c_k, we can compute the global coordinates of the given point x as shown in Eq. (15).

$$\rho(x) = (\sum c_k \cdot Projection(x, \phi_k))/(\sum c_k), \tag{15}$$

where $Projection(x, \phi_k)$ is the nearest point in ϕ_k to x. After that, the global continuous surface is constructed by applying Eq. (15) to each input point x.

D. Point Cloud Resampling

Finally, the point cloud resampling sub-module is designed to perform point cloud upsampling at any ratio by exploiting resampling on the global continuous surface. Specifically, this sub-module performs resampling uniformly in a 2D parametric domain D_i. Subsequently, each 2D point (u_r, v_r) are mapped onto the 3D local patch as x_i^r. After resampling all the local patches, the union of these sampled points is formed and denoted by $\mathcal{X}_R$ as shown in Eq. (16).

$$\mathcal{X}_R = \bigcup x_i^r. \tag{16}$$

After that, we perform uniform sampling on $\mathcal{X}_R$ and convert them to global coordinate system. The resulting upsampling point cloud is denoted by $\mathcal{Y}_R = \{y_{R_j}\}_{j=1}^{J}$.

IV. Experiments and Results

A. Hardware Prototype

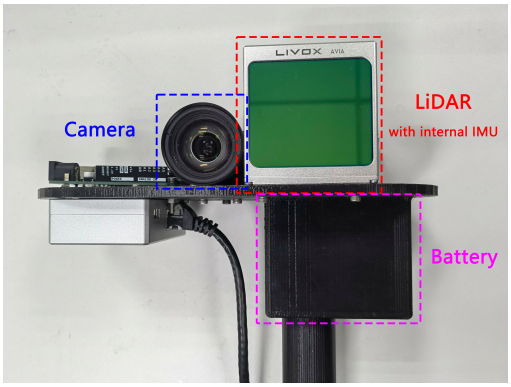


Fig. 4: Prototype of the proposed handheld system.

A handheld system for data collection has been constructed as shown in Fig. 4. The mechanical module of this system is 3D printed. This system includes a power supply unit, a MV-CE060-10UC rolling shutter camera and a Livox Avia LiDAR with an internal IMU BMI088, whose range and angular precision is reported as 2 cm at 20 m and less than 0.05°, respectively. The LiDAR and camera are synchronized before

being deployed to collect data at a frequency of 10 Hz. Furthermore, the internal IMU of LiDAR outputs at a frequency of 200 Hz synchronized with the other two sensors. Thanks to the lightweight of the LiDAR and the camera, the whole system can be easily mounted onto UAVs for geographic surveying and mapping.

The LiDAR is connected to a computer through the Ethernet interface whereas the camera is connected through the USB3.0 interface. The computer used in our experiments runs on the Ubuntu 18.04 64-bit operating system with ROS Melodic installed. In particular, ROS plays the role of data communication and processing center. Data from the LiDAR, the IMU and the monocular camera are received as ROS topics. Because every message in ROS is equipped with a detailed timestamp, these messages can be matched by aligning their timestamps. The proposed system subscribes these topics and processes these synchronized messages to achieve dense 3D color reconstruction.

The experiments of point cloud upsampling were implemented on the PyTorch framework with an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of memory. The point cloud upsampling model is trained and tested on Sketchfab dataset [9], including 90 training and 13 testing models. In order to avoid GPU memory shortage, the number of points in the input point cloud of each point cloud upsampling operation is set to 256. We divide each input large point cloud into small point clouds of 256 points on which the point cloud upsampling is performed. Finally, we re-fused the resulting upsampled small point clouds into one large point cloud as the final output.

B. Experiment Results

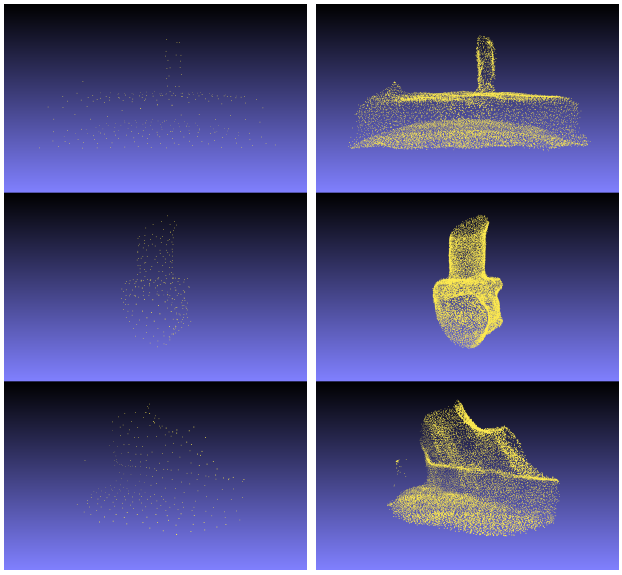


Fig. 5: The results of point cloud upsampling on Sketchfab. From left to right: (a) input point clouds, (b) 64x upsampled point clouds.

Fig. 5 shows experimental results derived on three examples taken from Sketchfab dataset [9] in which the left column shows the input sparse point clouds and the right column the 64x upsampled point clouds. Inspection of Fig. 5 suggests that the point cloud density has been greatly improved with additional points being generated along the surface of the object, preserving its original shape.

Extensive experiments were conducted in the outdoor environment on the Chinese University of Hong Kong, Shenzhen (CUHKSZ) campus. The collected point clouds were upsampled by a factor of 64. Fig. 6 shows a color picture of the test site of about $120 \times 120m^2$ with multiple buildings of $5-6$ floors. The test site is highlighted with the red boxes in Fig. 6.



Fig. 6: Test site with multiple buildings of $5-6$ floors.

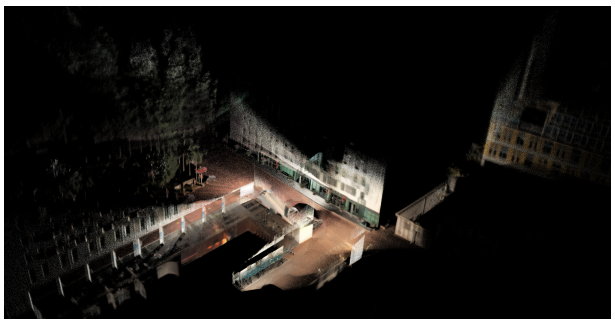
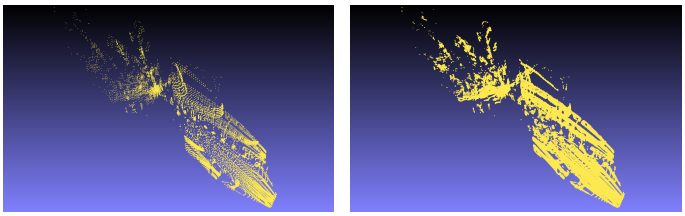


Fig. 7: The raw 3D model without upsampling.



Fig. 8: The corresponding site image captured by the camera.

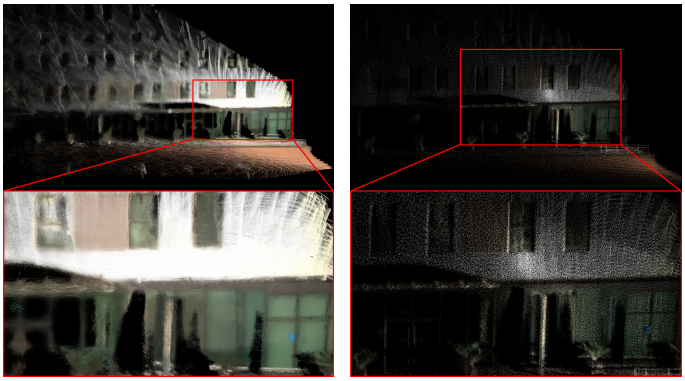
We limited the scanning time to be less than 100 s that is considered short for such a large scene. The resulting *raw* 3D model *without* upsampling is shown in Fig. 7 whereas the site image captured by the camera over the same area is shown in Fig. 8. Comparison of Fig. 7 and Fig. 8 suggests that the quality of the raw 3D model is far from being satisfactory.



(a) Sparse Point Cloud (b) Dense Point Cloud

Fig. 9: Comparison of 64x point cloud upsampling for a single frame.

Fig. 9 shows the point cloud upsampling result for a single point cloud frame. It is evidenced from Fig. 9 that the number of points has been greatly increased with the original structure of the object being preserved.



(a) With 64x point cloud (b) Without upsampling

Fig. 10: Comparison of 64x point cloud upsampling for a part of the large-scale outdoor scene.

Finally, Fig. 10(a) depicts the results derived from the proposed system. As compared to Fig. 10(b) showing the same area obtained from the conventional methods without upsampling, Fig. 10(a) demonstrates noticeable point cloud density improvement. In particular, if we focus on the zoom-in area inside the red box in both figures, points on the wall in Fig. 10(a) are closely packed without any gaps, forming a continuous surface. This confirms that the proposed point cloud upsampling system performed well with the LiDAR SLAM. It should be emphasized that the amount of input data in both Fig. 10(a) and Fig. 10(b) is identical.

V. Conclusion

In this paper, a sensor-fusion system has been proposed by integrating a solid-state LiDAR (Livox Avia), an IMU and a monocular camera. In particular, a novel point cloud upsampling technique has been developed to obtain denser 3D color maps. By exploiting the LiDAR-IMU-Camera fusion, the resulting reconstruction 3D color models are more precise and informative, i.e. accurate depth and fine texture. Furthermore, compared to the conventional SLAM method, the proposed framework is capable of generating point cloud of much higher density with less scanning time. Extensive experiments have confirmed that the proposed framework is capable of fast and dense 3D color reconstruction, which can be applied in a wide range of applications, such as geographic surveying and mapping, urban planning, smart cities, digital twin cities, and environmental protection.

References

[1] Y. Hu, T. Fu, G. Niu, Z. Liu, and M.-O. Pun, "3D map reconstruction using a monocular camera for smart cities," *The Journal of Supercomputing*, vol. 78, no. 14, pp. 16 512–16 528, 2022.

[2] Z. Zhu, S. Peng, V. Larsson, *et al.*, "Nice-slam: Neural implicit scalable encoding for slam," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 786–12 796.

[3] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, "Sensaturban: Learning semantics from urban-scale photogrammetric point clouds," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 316–343, 2022.

[4] Y. Zhu, C. Zheng, C. Yuan, X. Huang, and X. Hong, "Camvox: A low-cost and accurate lidar-assisted visual slam system," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5049–5055.

[5] J. Lin and F. Zhang, "R 3 live: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package," in *Proc. 2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 672–10 678.

[6] Z. Liu, Y. Hu, T. Fu, and M.-O. Pun, "Dense three-dimensional color reconstruction with data fusion and image-guided depth completion for large-scale outdoor scenes," in *Proc. 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 3468–3471.

[7] C. Yuan, X. Liu, X. Hong, and F. Zhang, "Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7517–7524, 2021.

[8] W. Feng, J. Li, H. Cai, X. Luo, and J. Zhang, "Neural points: Point cloud representation with neural fields for arbitrary upsampling," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 633–18 642.

[9] Y. Qian, J. Hou, S. Kwong, and Y. He, "Pugeo-net: A geometry-centric network for 3d point cloud upsampling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*, Springer, 2020, pp. 752–769.

Safety Enhancement for Mobility Scooter with Rule-Based Danger Prevention

Yan-Ru Chen, Shih-Wei-Chen Tseng, Yu-Chi Chen, and Yeong-Hwa Chang
Department of Electrical Engineering, Chang Gung University, Taiwan
E-mail: {b0921251, b0921217, b1021140, yhchang}@cgu.edu.tw

***Abstract* — This study proposes an assistive system to improve the safety of mobility scooter users. The system autonomously detects potential dangers and avoids accidents using multi-sensors, a single-board computer and active control mechanisms. The working principle includes braking assistance and throttle limitation and the potential benefits of the proposed system are identified. A vector table is constructed by experiment to determine the appropriate distance in which to safely stop the mobility scooter to prevent a collision. The assistive mobility scooter system is tested in the real world to demonstrate its robustness and safety.**

Index terms — Mobility Scooter, Danger Prevention, Safety Enhancement

I. INTRODUCTION

A mobility scooter, a vehicle that allows elderly or handicapped persons to move around. Most mobility scooters are simple to use, slow and suitable for short-distance travel [1, 2]. The demand for mobility scooters among the elderly is increasing worldwide and in Taiwan, the United States and Europe as the elderly population increases [3-6]. Many mobility scooter users view the ability to move as a symbol of health and independence but some elderly users have physical limitations, cognitive impairment, visual impairment or side effects from medication, and much more [2, 3]. The reasons above make users prone to accidents while they drive mobility scooters, epically a high risk of single-vehicle accidents [7-9]. Therefore, a mobility scooter with assistive functions, as shown in Fig. 1, is an important option to increase safety during daily driving [3, 10].

In terms of predicting danger for mobility scooters, many factors are relevant, such as the location, the direction of movement and the relative speed of an obstacle or pedestrian. Moreover, speed control must allow the vehicle to slow down smoothly [11, 12].



Fig. 1. Schematic diagram of a mobility scooter with assistive function modules.

This study collects data for these factors to determine the magnitude of danger using a rule-based model. To uncomplicatedly increase the safety of the mobility scooter user, this study adds multi-sensors and a single-board computer (SBC) with a danger estimation model to create a safer environment for users. The assistive mobility scooter system helps the user by slows the vehicle, stops the vehicle, and alerts the driver.

This paper is organized as follows. Sec. 2 reviews related studies of mobility scooters. The subsystems for the proposed mobility scooter are described in Sec. 3. Sec. 4 presents the rule-based braking method and the results of the experimental tests. Finally, the concluding remarks are given in Sec. 5.

II. RELATED STUDIES

Different types of intelligent mobility scooters have been proposed by previous studies. One study [11] presented a retrofitted mobility scooter for autonomous driving in a closed field for elderly or blind persons. A survey showed that after testing in a closed field, 95 of 99 users were willing to use the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

intelligent-assisted mobility scooter [11]. This study used two LIDAR systems to detect obstacles, map the local area and traced a dynamic virtual bumper for static obstacles and moving pedestrians. Another study [12] based on this study [11] involved driving in an environment with many pedestrians. This system also used two LIDAR systems, but a more complex model was used to prevent the mobility scooter from bumping into pedestrians.

One study [13] used a braking system for an electric scooter to prevent accidents. Ultrasonic sensors detected the distance from an obstacle ahead. The motor was reserved and the rear brake was applied when the distance between the scooter and the obstacle was less than a meter. Another study [14] showed that vehicles with an automatic emergency braking (AEB) have a lower accident rate, and another study [15] showed how the AEB works in a timing table. It also considered speed, force and front distance to increase driver safety. Another study [16] used AEB on a motorcycle, but the brake is currently out-triggered.

This study uses a system that is similar to those previous studies [13] and [15], but instead the static distance is not used to trigger the brake: this study uses a vector table to determine when to activate the braking system. Multi-sensor modules to gather information on distance, speed and throttle position uses a system that is similar to those of previous studies [11] and [15]. The proposed mobility scooter better gather information by integrating sensor modules in a mobility scooter that is currently commercially available.

III. System Overview of Mobility scooter

A. *Specifications*

The base platform for the proposed assistive mobility scooter system is the LY-EW401 that is manufactured by Comfort Orthopedic Co. Ltd. The stock-rated unladen weight of the mobility scooter is 68 kg with a 25 kg battery. The maximum payload is 120 kg and the average range is 30 km. This study adds sensors, microcontroller units (MCUs), a SBC and batteries to increase user safety. The specifications for the mobility scooter before and after the retrofit are summarized in Table 1.

The assistive mobility scooter system is non-destructive: all modules are connected to the mobility scooter using existing joints or glue. Fig. 1 shows all additional modules, such as the power-on module, the motor module, the distance modules, the speed module and. In practice, these function modules can be easily installed on other mobility scooters and calibration is easy.

Table 1. Mobility scooter specification before and after retrofit.

Items	Before Retrofit	After Retrofit
Dimensions (L × W × H)	132 × 62 × 94 cm	
Empty Weight	93 kg	98 kg
Maximum Payload	120 kg	115 kg
Maximum Speed	11 km/h	10 km/h
Range	36 km	30 km
Slope	10 degrees	10 degrees
Working voltage	24 volts	24 and 5 volts

B. *Architecture*

The assistive mobility scooter system uses a master-slave architecture: the master is a SBC and the slaves are MCUs. The SBC is a Raspberry Pi 3B+, which communicates with all MCUs over Wi-Fi via the MQTT protocol. The MCU uses a ATmega328PU as the main processor core. Each module has a MCU, and the system has five modules: a power-on module, a motor module, a front distance module, a rear distance module and a speed module. The system is also connected to the Internet for time calibration and downloads local weather for the user. The architecture for the assistive system is shown in Fig. 2. The SBC and MCUs are described later.

C. *Power-on module*

The power-on module is connected to the battery and all modules. The module detects the battery life and turns the power on or off for each MCU. When the user switches on the mobility scooter, this module is automatically activated. All other modules are powered when the power-on module is activated to ensure that all modules are switched on when the mobility scooter is switched on and switched off when the mobility scooter is switched off.

D. *Motor module*

This module is plugged into the vehicle's throttle to record the behavior of users when driving the mobility scooter. An electromagnetic brake is applied in the mobility scooter, where the braking is automatically activated when the throttle position is set to zero. Therefore, motor module decreases the throttle or provides active braking in an emergency. Throttle adjustment and active braking prevent injury in unforeseen circumstances.

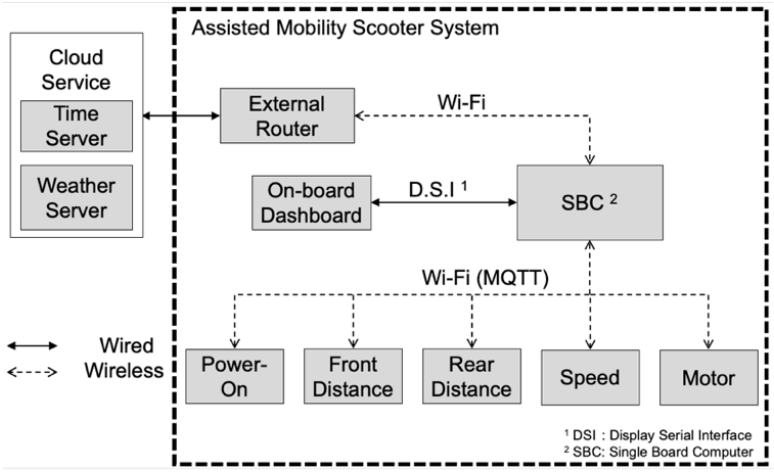


Fig. 2. The architecture for the mobility scooter.

E. Front and rear distance modules

These two modules perceive distance information using multiple HC-04 ultrasonic sensors and a time-of-flight (TOF) infrared sensor. The ultrasonic sensor emits signals and if an obstacle is less than 3 meters ahead, the signal is reflected to the sensor. Once the signal is received, the distance between the sensor and the obstacle could be discovered from the time duration that the signal travel. The front distance module uses five HC-04 ultrasonic sensors and one TOF sensor to detect obstacles in front of the mobility scooter. The effective range of the HC-04 is relatively short so a TOF sensor is additionally installed to extend the effective sensing range ahead to 5 meters. The rear distance sensor uses only three HC-04 ultrasonic sensors because the mobility scooter moves slowly in reverse.

F. Speed module

This module comprises two Hall sensors with ultra-strong magnets (two on the left and four on the right) on the inner side of the vehicle's rear wheels. The vehicle's speed is calculated using the angular velocity in (1):

$$v = \omega \times r \tag{1}$$

where v is the linear speed of the vehicle, ω is the angular velocity, and r is the radius of the rear wheel ($r = 16.5$ cm).

The angular velocity is calculated using the duration between two passing magnets on the rear wheels. The speed is calculated for the right and left rear wheels to ensure an accurate result.

G. Single board computer

The Raspberry Pi acts as the SBC. It has an embedded Wi-Fi chip that functions as a wireless access point (WAP) that allows the MCUs to connect and communicate via MQTT protocol. The Raspberry Pi receives all information that is transmitted from the MCUs and verifies that there is a connection between the SBC and the MCUs. Information is saved in the Raspberry Pi and is viewed on the dashboard or is used for calculations by the danger prevention model.

The onboard dashboard, which is shown in Fig. 3, is 800 × 480 pixels, which accommodates a Raspberry Pi 7-inch touchscreen. This display shows the Wi-Fi status, the time and the weather information on the top left, the brake status on the top right and the distance warning on the bottom right. The lock-unlock button is on the bottom of the dashboard and the speed is shown in the middle of the dashboard.

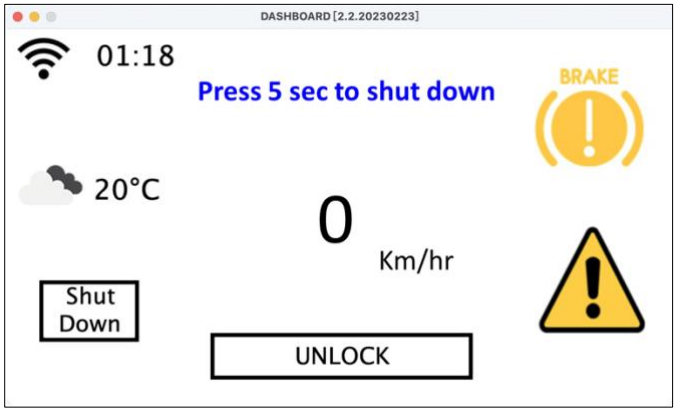


Fig. 3. Onboard dashboard for the mobility scooter.

A throttle lock prevents the user from accidentally activating the throttle. The user locks or unlocks the vehicle by pressing a button and a signal is then transmitted to the motor module via the MQTT protocol. This function ensures the safety of the driver. Most mobility scooters do not display the vehicle's speed, but this system displays the speed on the dashboard.

The warning light on the bottom right activates when an obstacle is close to the front or rear of the mobility scooter. If an obstacle is detected, the throttle is limited via the motor module. The braking indicator on the top right illuminates when the brake is activated by the braking assistance system, which activates when an obstacle is too close to the obstacle. The brake and throttle limiter are discussed in the next section.

IV. Braking Assistance

A. Working principle

When a danger situation is detected, the braking assistant autonomously slows or stops the mobility scooter to prevent an accident. The flow chart for the system is shown in Fig. 4. When the mobility scooter is powered on, the assistive system monitors the speed of the mobility scooter, the front distance and rear distance from obstacles.

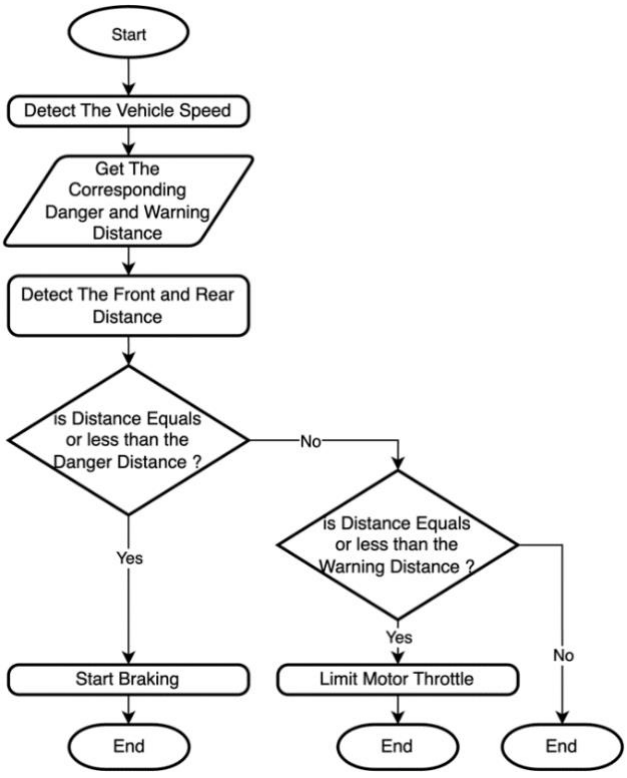


Fig. 4. Flow chart for the mobility scooter.

The assistive system determines the danger distance to a obstacle ahead using the information from Table 2 with the instantaneous speed. If the distance from the obstacle is less than the danger distance, the system activates the brake to stop the mobility scooter. The throttle is limited when the distance to the obstacle exceeds the warning distance. The warning distance is 1 meter longer than the danger distance, as listed in Table 2. The throttle limit value is set to 80% of the current throttle value. The percentage for limited the throttle can be adjust on the setting page, to satisfied the comfort of the user.

Table 2. Danger distance and warning distance versus vehicle speed.

Vehicle Speed	Danger Distance	Warning Distance
0 ~ 1 km/h	32 cm	132 cm
1 ~ 2 km/h	41 cm	141 cm
2 ~ 3 km/h	64 cm	164 cm
3 ~ 4 km/h	76 cm	176 cm
4 ~ 5 km/h	91 cm	191 cm
5 ~ 6 km/h	105 cm	205 cm
6 ~ 7 km/h	136 cm	236 cm
7 ~ 8 km/h	171 cm	271 cm
8 ~ 9 km/h	212 cm	312 cm
9 ~ 10 km/h	258 cm	358 cm

B. Working theory

To determine the braking distance for the mobility scooter, this study uses a table. The equation is derived using the resistance force. The resistance force is listed in (2) for a dynamic equation for a car using D'Alembert's principle [17]:

$$F_t = F_f + F_i + F_w + F_j \tag{2}$$

where F_t is the total resistance, F_f is the rolling resistance, F_i is the resistance due to a gradient, F_w is the air resistance and F_j is the acceleration resistance. In most cases, a mobility scooter moves slowly; therefore, F_f is regarded as a constant and F_w and F_j are considered as zeros. F_i is defined in (3) as:

$$F_i = m \times g \times \sin\theta \tag{3}$$

where m is the mass of the vehicle and user, g is the acceleration due to gravity, and $\sin\theta$ is the vertical component of the force, where θ is positive for travel up a gradient. The total braking force is expressed as (4):

$$F_{total} = F_f + m \times g \times sin\theta + m \times \beta = m \times a \quad (4)$$

where F_{total} is the total force when braking, β is the vehicle's deceleration under braking and a is the total deceleration due to braking. Using the value for the total braking force that is calculated using Equation (4), the kinematic equation is written as (5):

$$V_1{}^2 = V_0{}^2 - 2 \times a \times S \quad (5)$$

where V_1 is the final velocity, which is 0, V_0 is the initial velocity and S is the braking distance. Using (4) and (5), the braking distance is calculated as (6):

$$S = \frac{-V_0{}^2}{2\times\frac{F_{total}}{m}} = \frac{-V_0{}^2}{2\times\left(\frac{F_f}{m} + g\times sin\theta + \beta\right)} \quad (6)$$

The braking distance is proportional to the square of the initial velocity and is inversely proportional to the sine of the gradient. Mobility scooters are mostly used on flat roads so the resistance due to any gradient is constant at 20 cm. This distance ensures the safety of a mobility scooter that has no gyro sensor.

C. Experimental results

An experiment is shown in Fig. 5 to calculate the distance table for the assistive mobility scooter system. Each distance was measured multiple times, and the experimental procedure is as follows: The mobility scooter was parked a few meters behind an infrared sensor and accelerated to a constant velocity that increased in 1 km/h increments. The mobility scooter then drove at a constant speed until it passed the infrared sensor, and the mobility scooter was then immediately stopped by resetting the throttle. These distances were recorded and used to construct the distance table (Table 2).

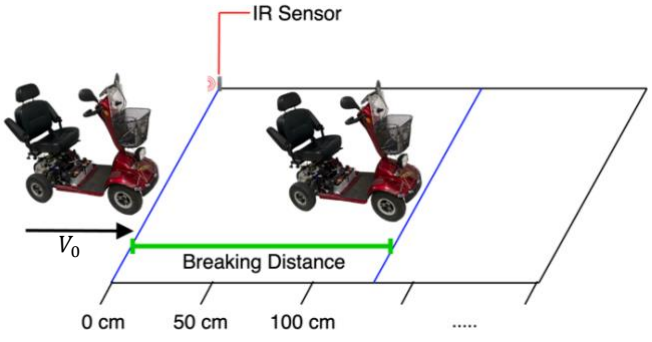


Fig. 5. Schematic diagram of the experiment.

The results of the experiments demonstrate that the assistive mobility scooter safely brakes from 1 to 10 km/h, as shown in Table 2. The experiment was repeated in real-world environments to verify the vehicle's capabilities. The results of one experiment are shown in Fig. 6, for an initial speed of 5 km/h, which is a typical speed for a mobility scooter.

Initially, the vehicle's speed is 5 km/h, the throttle magnitude was 55% and the distance of any obstacles ahead the vehicle is more than 5 meters. The sensor only responds at 5 meters due to the limitations of the detection distance. After a few seconds, the front of the scooter is 191 cm from the obstacle, which is within the warning distance. The system then initiates throttle limitation and the throttle value decreases from 55% to around 44% so the vehicle's speed decreases from 5 km/h to 4.1 km/h. The vehicle does not stop and the front distance quickly decreases to 91 cm from the obstacle, which is less than the safe distance. The system immediately sets the throttle to zero, which activates the brake mechanism and completely stops the vehicle. These results show that the assistive mobility scooter activates the brake in both experimental and real-world environments to prevent a collision.

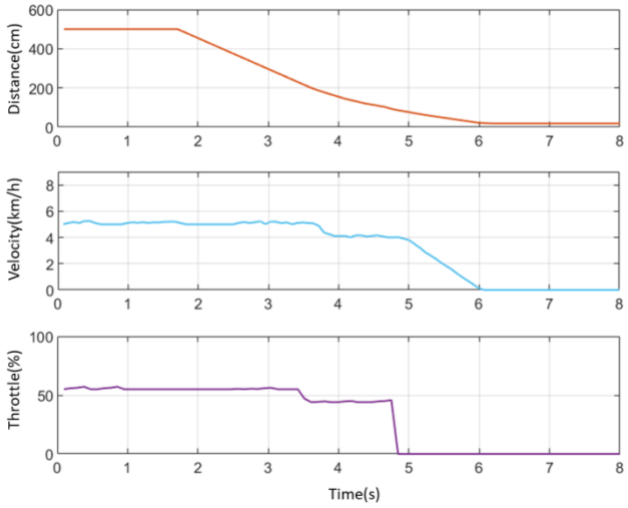


Fig. 6. Timing table for mobility scooter in a real-world scenario.

V. CONCLUSIONS

The assistive mobility scooter system that is proposed by this study enhances safety for users. The system autonomously activates the braking assistant by continuously monitoring information from sensors and comparing them to the corresponding danger distance to prevent accidents. Limiting the throttle when the vehicle approaches the warning distance. This comprehensive approach addresses any physical

limitations or cognitive impairment that users might experience and which increase the risk of accidents for mobility scooters.

This study presents a practical and effective solution to increase the safety of users and increase the sense of independence and well-being. The use of advanced technology and the focus on danger prediction using relevant factors reduces the risk of accidents. Further improvements and real-world evaluations are necessary to optimize the system's performance and to ensure its effectiveness for different environments and scenarios. Mobility scooters are increasingly used by the elderly population and this type of assistive systems significantly improves the quality of life and promotes a sense of health and independence.

Future study will allow participants to experience the assistive mobility scooter and determine the user's reaction, in terms of the warning distance and the braking resistance at high speed. The danger prediction model will also be improved by adding more target obstacles to make the system safer.

REFERENCES

[1] K. Liu, "Smart and Connected Scooter for People with Mobility Challenges," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019: IEEE, pp. 1-6.

[2] M. Isaacson and D. Barkay, "Mobility scooters in urban environments: A research agenda," *Journal of transport & health,* vol. 18, p. 100917, 2020.

[3] M. Hirai, T. Tomizawa, S. Muramatsu, M. Sato, S. Kudoh, and T. Suehiro, "Development of an intelligent mobility scooter," in *2012 IEEE International Conference on Mechatronics and Automation*, 2012: IEEE, pp. 46-52.

[4] National Development Council. "Human Resources Development of Population aging." https://www.ndc.gov.tw/Content_List.aspx?n=D527207EEEF59B9B

[5] J. Vespa, L. Medina, and D. M. Armstrong, "Demographic Turning Points for the United States: Population Projections for 2020 to 2060," *U.S. Census Bureau,* pp. 1-5, 2018.

[6] eurostat. "Ageing Europe - statistics on population developments." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_population_developments

[7] A. Peters, M. G. J. Versteegen, F. van Osch, H. M. J. Janzing, and D. G. Barten, "Mechanism and severity of mobility scooter–related injuries," *Traffic Injury Prevention,* vol. 23, no. 2, pp. 112-117, 2022/02/17 2022, doi: 10.1080/15389588.2021.1998469.

[8] O. Krahelski, S. Sivarajah, W. Eardley, T. Smith, and C. Hing, "Major trauma associated with mobility scooters: An analysis of the trauma audit research network," *Injury,* vol. 53, no. 9, pp. 3011-3018, 2022.

[9] T. Zanotto, "Characteristics and consequences of falls among people with multiple sclerosis who use wheelchairs or scooters: Differences between injurious and non-injurious falls," *Multiple sclerosis and related disorders,* vol. 73, p. 104631, 2023.

[10] H. Leijdesdorff, J. van Dijck, P. Krijnen, and I. Schipper, "Accidents involving a motorized mobility scooter: a growing problem," *Nederlands Tijdschrift Voor Geneeskunde,* vol. 158, pp. A7858-A7858, 2014.

[11] H. Andersen, "Autonomous personal mobility scooter for multi-class mobility-on-demand service," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016: IEEE, pp. 1753-1760.

[12] Y. Luo, P. Cai, A. Bera, D. Hsu, W. S. Lee, and D. Manocha, "Porca: Modeling and planning for autonomous driving among many pedestrians," *IEEE Robotics and Automation Letters,* vol. 3, no. 4, pp. 3418-3425, 2018.

[13] P. Athul., "Microcontroller based Intelligent Braking System in Electric Scooters," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020: IEEE, pp. 188-192.

[14] J. B. Cicchino, "Effects of automatic emergency braking systems on pedestrian crash risk," *Accident Analysis & Prevention,* vol. 172, p. 106686, 2022.

[15] S. Kristalniy, N. Popov, V. Gaevskiy, A. Andreev, and M. Toporkov, "Experience of using a moving pedestrian model in tests of automatic emergency braking system," in *2021 Systems of Signals Generating and Processing in the Field of on Board Communications*, 2021: IEEE, pp. 1-5.

[16] S. Piantini, V. N. Aathresh, G. Savino, and M. Pierini, "Assessment of the effect of motorcycle autonomous emergency braking (MAEB) based on real-world crashes," *Traffic Injury Prevention,* vol. 23, no. sup1, pp. S174-S180, 2022/10/10 2022, doi: 10.1080/15389588.2022.2117983.

[17] R. Zhai, P. Xiao, R. Zhang, and J. Ju, "In-wheel motor control system used by four-wheel drive electric vehicle based on whale optimization algorithm-proportional–integral–derivative control," *Advances in Mechanical Engineering,* vol. 14, no. 6, p. 16878132221104574, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Dictionary-driven Chinese ASR Entity Correction with Controllable Decoding

Rongjun Li* and Wei Peng†
* IT Innovation and Research Center, Huawei Technologies Co., Ltd., China
E-mail: lirongjun3@huawei.com
† IT Innovation and Research Center, Huawei Technologies Co., Ltd., China
E-mail: peng.wei1@huawei.com

***Abstract*—The mainstream ASR error correction system is mainly based on the encoder-decoder structure, by learning the mapping of incorrect text to correct text. Although this approach showed good results in ASR error correction, it mainly solves the spelling errors that occur at high frequencies, while struggling with low-frequency ASR entity errors, especially those entities unseen in the training set. Another downside of the structure is that it may produce undesirable hallucination entities. In this paper, we introduce a DiCODER model for Chinese ASR entity error correction. Specifically, we first retrieve relevant entities from the entity dictionary based on the similarity of the text and phonetic information and then fuse them with ASR text for joint encoding. Next, a controlled decoding strategy utilizing dynamic vocabulary is proposed to generate more reasonable results. Experimental results on three publicly available datasets demonstrate the effectiveness of our proposed method.**

Index Terms: ASR error correction, dictionary knowledge, controllable decoding

I. Introduction

Deep learning-based ASR recognition systems have made significant progress in real-world applications. However, there are still various types of ASR errors in practice, especially entity errors. For example, the named entity "马克黄"("Mark Huang") may incorrectly be identified as "马克谎"("Mark Clung") in "公司分析师马克黄说"("Analyst Mark Huang said"), which will impair downstream tasks such as semantic understanding [1], [2]. Therefore it is necessary to design ASR error correction methods that can effectively handle these errors.

ASR error correction methods can generally be divided into cascaded and end2end categories. The cascaded method first detects the incorrect words and then fixes these words [3]. This method can easily perform error analysis, but it requires both error detection and correction annotations, which is costly in terms of labor. Instead, the end2end method is more labor-efficient through learning the mapping between incorrect and correct text directly. Since its training data is easy to produce (e.g. through TTS-ASR simulations), it is currently the mainstream ASR error correction method [4]–[6].

The end2end approach can give good results for ASR spelling errors, but it is insufficient for ASR entity errors [7]. The reason is that the incorrect entities are often not appearing in the training dataset. Recently, Amazon proposed a KNN memory-based method to address this problem, which shows promising results [8]. However, it requires a large space for storing the episodic memory. [9] incorporates candidate entities into the correction model with a shared context encoder and shows excellent performance for out-of-vocabulary person names. However, It utilizes text edit distance to obtain candidate entities and fails to effectively recall erroneous entities with large textual differences but phonetic similarity. Further, the ASR text and the candidate entities are encoded independently, making them lack direct mutual interactions. On the other hand, the decoder of these end2end model relies on the language model for text generation, which may generate hallucination entities.

To address the above two challenges, we propose a controllable ASR error correction method based on entity-dictionary knowledge. Specifically, we first retrieve the relevant entities from the entity lexicon based on the similarity of character and phonetic information, and later jointly encode the ASR text and the relevant entities, followed by a controlled decoding strategy to fix entity errors and avoid producing unreasonable decoding results. We also propose a hallucination rate metric for measuring the ratio of hallucination entities in correction results. This evaluation metric shows that the quality of error correction can be further improved by our proposed decoding strategy. The experimental results of three publicly available datasets indicate that our method can effectively improve speech recognition quality, especially entity correction performance.

II. Related Work

ASR error correction and ASR entity error correction are two closely related research topics. The former aims to enhance the accuracy of speech recognition by correcting contextual errors in the ASR outputs, which also implicitly fixes some entity errors. The latter, on the other hand, concentrates on rectifying the entity errors to improve the overall quality of ASR transcriptions.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

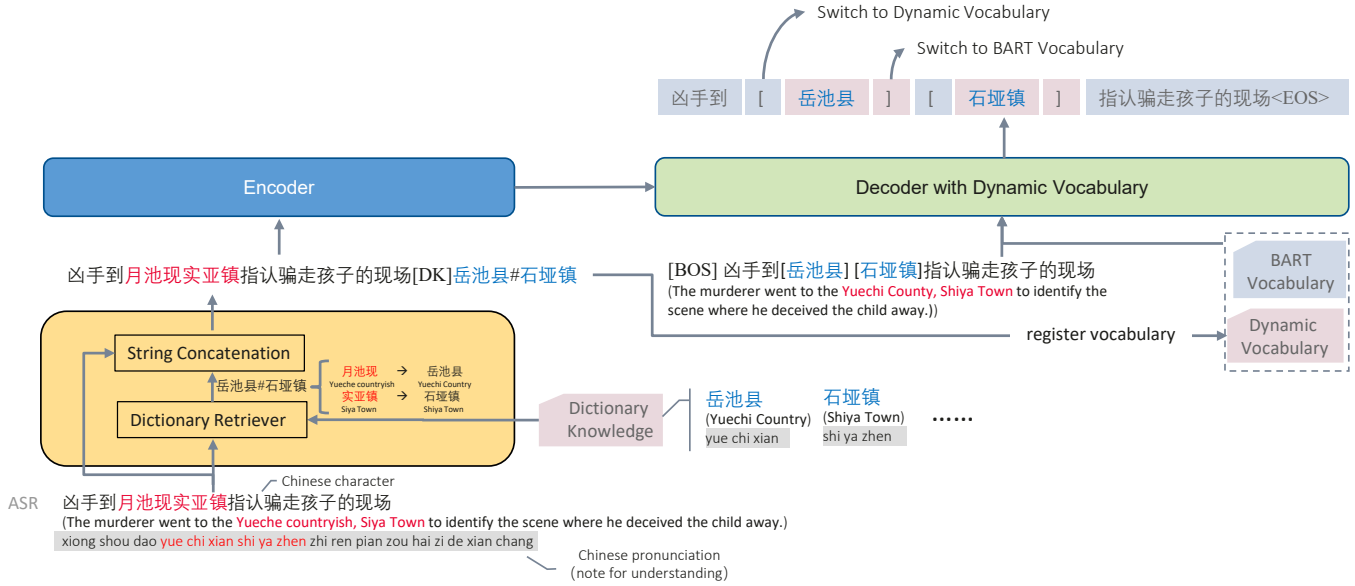


Fig. 1. The overall architecture of the proposed method.

A. ASR Error Correction

ASR error correction methods have mainly adopted a deep learning-based framework. [10] treats ASR error correction as a machine translation task and models it based on the RNN encoder-decoder. [11] replaced the RNN structure with the Transformer structure to further improve the results. Although the above two methods achieved good results on the ASR error correction task, the training of such models relies on a large amount of training corpus. [5] and [12] enhance the generalization performance of such error correction models by initializing the model parameters with BERT or BART pre-trained models, respectively. [7] tackles the problem of misrecognition of near-phonetic characters in English by adding phonetic information to the textual information. [13] solves the problem of recognizing near-phonetic characters in Chinese by adding pinyin information. Further, [14] and [15] introduce phonetic-related pre-training tasks to solve the ASR error correction problem in English and Chinese, respectively. In addition, [16] improves the effectiveness of ASR error correction by modeling the N-best results of ASR output.

B. ASR Entity Error Correction

The topic of this paper is knowledge-enhanced ASR error correction, which is just beginning to be investigated. [8] improves the likelihood of correct entities being decoded by registering text fragments containing entities with external memory, and later using KNN retrieval to obtain candidate entities from memory at each moment of decoding. [17] proposes an approach that draws on existing graph knowledge and finds relevant entities by subgraph matching, followed by ASR entity error correction. While both methods achieve good results in ASR entity correction, such text fragments and related graphs that contain entities are often difficult to obtain in real application scenarios, leading to their limited application. [9] uses two shared-parameter encoders to encode an ASR text and candidate entities such as personal names respectively, then the decoder attends to the encoders by attention mechanism to generate correction results.

III. Approach

The task of our ASR entity error correction can be treated as a functional mapping from X, D to Y, where X denotes text outputs from ASR containing errors, and D is an entity dictionary that contains W entity words. Y is the ground-truth text. As shown in Figure 1, the framework contains the following three parts, including 1) dictionary retriever module: that is a dictionary fuzzy matching module. 2) context encoder module: which encodes ASR text and matching entities corresponding. 3) decoder module: which decodes the character guided by matching entities.

A. Dictionary Retriever

Different from general errors, entity errors cannot be determined according to their context, because entities are very changeable. This is also an important reason why we want to introduce external dictionaries. For example, the model can easily correct “The whether is nice today!” as “The weather is nice today!”, but it is difficult to determine “The murderer went to Siya Town.” or “The murderer went to Shiya Town.”, unless there is a “Shiya” in your mind. Since we cannot determine the wrong location in advance, in order to recall possible right entities, we propose a fuzzy matching method that considers both character and sound shape information to recall all possible entities.

The algorithm 1 summarizes the whole process. At first, a character trie and a sound shape[1] trie are constructed for the entity dictionary. Second, for each text span in the sentence, we retrieve its matching entity words with confidence scores at the character level or pronunciation level. At last, for matching accuracy and computational efficiency considerations, we limit the number of matching results by their confidence scores. In more detail, we constrain the number of matching results from two levels: the number of matching results per span and the total number of matching results. In Figure 1, the output result C of the dictionary retriever is "岳池县" 和 "石垭镇" because they have similar Chinese pronunciations to "月池现" and "实亚镇" in X, respectively.

Algorithm 1: Dictionary Knowledge Retrieval

input : X, an ASR text.
D, an entity lexicon list.
output: C, a retrieved entity list.

```
// Initialization.
charTrie = InitCharTrie()
sscTrie = InitSoundTrie()
for d ∈ D do
    charTrie.add(d)
    sscTrie.add(GetSoundShape(d), d)
// search the character and sound shape
   tries.
C = List()
foreach substring s ∈ X do
    if charTrie.has(s) then
        score = len(s)
        C.add((s, score))
    else if sscTrie.has(s) then
        score = 0
        e = sscTrie.getEntity(s)
        foreach character sc, ec ∈ zip(s, e) do
            sp = GetPinYin(sc)[2]
            ep = GetPinYin(ec)
            if sc == ec then
                score += 1
            else if sp == ep then
                score += 0.6
            else
                score += 0.1
        C.add((e, score))
C.sort(key=lambda x: x[1], reverse=True)
return C
```

[1]https://github.com/qingyujean/ssc
[2]https://pypi.org/project/pypinyin/

B. Context Encoder

We employ a context encoder to encode X and C jointly. In particular, X and C are concatenated into the form X [DK] C, where [DK] denotes special tokens that serve to separate ASR text from retrieved entities. For example, in Figure 1, the fused input is "凶手到月池现实亚镇指认骗走孩子的现场 [DK] 岳池县 # 石垭镇". This concatenated input is subsequently fed into the encoder.

C. Decoder with Dynamic Vocabulary

The commonly used seq2seq model is prone to generate hallucination entities. In Figure 1, the location entity "实亚镇" (Siya Town) is wrongly corrected as "实亚锦" (Siya towne) by baseline method but the ideal entity should be "石垭镇" (Shiya Town). In this case, "实亚锦" (Siya towne) is a hallucination entity, which is created under the language modeling of the decoder.

To alleviate this problem, we propose a controlled decoding strategy to help the decoder ease this kind of miss-corrections. As shown in Figure 1, two different vocabularies were used, one is a common dictionary (here refers to the vocabulary of BART), and the other is the dynamically built vocabulary which only contains the fused input tokens of the context encoder. We do two tasks at the same time when decoding, one is entity recognition, and the model output will enclose the location of the corresponding entity in square brackets; the second task is entity correction. Specifically, at the initial stage of decoding, the decoder uses the common vocabulary to decode until generating the entity start position indicator "[". Next, the decoder uses the dynamic vocabulary to generate the corrected entity followed by the end position indicator "]" generated. After that, it switches back to using the common vocabulary, and this process will repeat until the end of the generation.

IV. Experiments

A. Datasets

TABLE I
Statistics of the datasets. Sents, Avg.L and Dict are the sentence number, sentence average length and the dictionary size of the dataset, respectively.

Dataset	Type	Sents	Avg.L	Dict
Aishell	train	120,098	15.41	
	dev	14,326	15.33	
	test	7,176	15.60	16,470
$\text{MAGICDATA}_{\text{SMALL}}$	train	16,000	10.90	
	dev	2,000	10.69	
	test	1,775	10.74	4,591
$\text{MAGICDATA}_{\text{OOV}}$	train	17,433	10.89	
	dev	1,290	10.87	
	test	1,052	10.85	4,591

In our experiment, we employ three datasets that leverage distinct ASR engines.

1) **Aishell**[3] is an open-source Mandarin Chinese speech corpus that offers audio recordings and their corresponding manual transcriptions [18]. The AISHELL_NER [19] supplies related entity annotations for this dataset. We employ an open-source conformer ASR model [20] to transcribe the audio recordings into ASR texts. These texts are then paired with their manual transcriptions to create an error correction dataset.

2) **MAGICDATA**[4] is an additional Mandarin Chinese read speech corpus. We randomly select a subset of samples and annotate their entities to create the MAGICDATA$_{\text{SMALL}}$ dataset. Subsequently, we employ our commercial ASR service to transcribe the audio recordings into texts.

3) **MAGICDATA$_{\text{OOV}}$** is created by re-splitting MAGICDATA$_{\text{SMALL}}$ into new training, development, and test sets that contain distinct entities. The objective of constructing this dataset is to assess the efficacy of our model on entities that are not present in the vocabulary.

We add "[" and "]" symbols for each entity to indicate their start and end positions within the sentence, then extract these entities from every dataset to form their corresponding entity dictionaries. Table I shows the statistics information of the datasets.

B. Baselines

We compare our methods with the following baselines:

1) **Transformer** [21]: We implement a Chinese version Transformer model for ASR error correction[5]. It is a standard Transformer model with 6 encoder and 6 decoder blocks, trained from scratch.

2) **SC_BART** [5]: BART is a seq2seq language model that has been pre-trained on various denoising tasks [22]. It initially corrupts text using noise functions and then learns to reconstruct the original text. This makes BART well-suited for error correction tasks. The SC_BART uses BART as its base model and fine-tunes it to achieve state-of-the-art results on Chinese ASR error correction tasks. We select a pre-trained BART[6] with 6 encoder and 6 decoder layers as the base model and re-implement SC_BART.

3) **CSC_BART**: Similar to [9], we replicate an entity error correction model that encodes an ASR text and entity candidates separately with two encoders that shared parameters. The model is also initialized with the pre-trained BART.

C. Model training details

The maximum number of matching results per span is set to 1, and the maximum total number of matching results is set to 3. In line with the baseline methods, our model is initialized using BART. The maximum length of the decoder output is set to 64. The initial learning rate is 2e-5. We use the Adam optimizer [23] with a batch size of 32 (the same as the compared methods) and train for 10 epochs. The patience for early stopping is set to 3.

D. Evaluation Metrics

For the error correction task, we evaluate using the sentence error rate (SER) and character error rate (CER). For the entity correction task, after ASR correction, entities are first recognized by a NER tool[7], and then the recall of the entity is calculated using the ground truth.

Furthermore, we have found that hallucinated entities are more disruptive to users and downstream systems than original ASR error entities. This is because a user with background knowledge can infer the correct entity from an ASR error entity, but cannot do so from a hallucinated entity. For instance, if a user already knows the "石垭镇" (Shiya Town) is a town name, he will be aware that the incorrectly recognized "实亚镇" (Siya Town) refers to it due to their similar Chinese pronunciation, but it's hard for him to make this judgment from the generated hallucinated entity"实亚锦"(Siya towne). To measure the degree of hallucination in error correction results, we propose a new metric called the hallucination rate (HR), which is defined as follows:

$$\text{HR} = \frac{\text{number of hallucination entities}}{\text{number of corrected entities by the model}}. \quad (1)$$

The lower value of the HR means less hallucination entities generated and vice versa.

E. Results

1) **Evaluation of ASR error correction**

As shown in Table II, both the SC_BART and CSC_BART methods have varying capabilities for correcting spelling errors. Compared to the results of No correction, the average SER/CER of SC_BART and CSC_BART decreased by 25.31%/2.79% and 26.75%/2.64%, respectively. Compared to SC_BART and CSC_BART, the DICODER$_{\text{CD}}$ further reduces the average SER/CER by 5.95%/0.51% and 4.51%/0.65%, respectively. It is noteworthy that our methods achieve consistent results on datasets produced by different ASR techniques, including the conformer ASR model and our commercial ASR service, demonstrating a reasonable level of robustness.

2) **Evaluation of ASR entity error correction**

On average, the CSC_BART method achieves competitive improvements in entity recall across the datasets, outperforming SC_BART. Compared to CSC_BART, DICODER$_{\text{CD}}$ further improves the average recall by 7.98%, benefiting from the joint encoding of ASR text and candidate entities. Furthermore, the controllable decoding

[3]https://openslr.org/33

[4]https://openslr.org/68

[5]github.com/pytorch/fairseq/tree/master/examples/translation

[6]huggingface.co/fnlp/bart-base-chinese

[7]Use the text alignment tool systran-align to align the ASR text with the entities of the correct text to get the entities in the ASR text. Tool link is https://pypi.org/project/systran-align/

TABLE II

THE ASR ERROR CORRECTION AND ENTITY CORRECTION RESULTS OF COMPARED METHODS AND OUR APPROACHES. THE DICODER IS THE METHOD WITHOUT CONTROLLABLE DECODING MECHANISM. HR STANDS FOR HALLUCINATION DEGREE.

Method	Aishell				MAGICDATA$_{SMALL}$				MAGICDATA$_{OOV}$			
	SER↓	**CER↓**	**Recall↑**	**HR↓**	**SER↓**	**CER↓**	**Recall↑**	**HR↓**	**SER↓**	**CER↓**	**Recall↑**	**HR↓**
No correction	37.05	4.31	77.36	-	100.00	13.29	61.80	-	100.00	13.05	31.56	-
Transformer	36.46	4.46	78.04	65.35	39.10	9.06	74.31	42.78	97.25	30.39	10.55	98.58
SC_BART	30.74	3.64	81.37	48.17	38.82	5.95	77.01	31.49	91.55	12.69	29.56	92.27
CSC_BART	30.65	3.67	81.32	50.53	37.46	6.27	78.65	30.64	88.69	12.77	32.32	63.21
DiCODER	**28.34**	**3.35**	**90.24**	**25.34**	**36.00**	**5.86**	**81.86**	25.86	76.42	11.34	47.34	43.78
DiCODER$_{CD}$	28.82	3.45	88.99	28.59	38.20	6.12	79.61	**25.38**	**76.24**	**11.18**	**47.62**	**35.71**

TABLE III

EXAMPLES OF DIFFERENT METHODS. WE HIGHLIGHT THE ERRONEOUS/CORRECT CHARACTERS IN RED/BLUE FONTS.

Method	**Transcript**	**Retrieved Entity**
No correction	给我拨一个庚橙庆的歌曲。 (Ply a song of Geng Orange Qing.)	
SC_BART	给我播一个庚橙庆的歌曲。 (Play a song of Geng Orange Qing.)	
DiCODER$_{CD}$	给我播一个庚橙庆的歌曲。 (Play a song of Geng Orange Qing.)	
No correction	男一忍受股价被低估。 (Clan stand undervalued shares.)	
SC_BART	难以忍受股价被低估。 (Can't stand undervalued shares.)	
DiCODER$_{CD}$	难以忍受股价被低估。 (Can't stand undervalued shares.)	魏楠; 蚁人 (Kenstan; Ant-Man)
No correction	但直到一二年后的法国阿尔贝为尔冬奥会上。 (But not until the 2012 Winter Olympics in Albertwell, France.)	
SC_BART	但直到一二年后的法国阿尔贝塞冬奥会上。 (But not until the 2012 Winter Olympics in Ashville, France.)	
DiCODER$_{CD}$	但直到一二年后的法国阿尔贝维尔冬奥会上。 (But not until the 2012 Winter Olympics in Albertville, France.)	阿尔贝维尔; 韦尔 (Albertville; Vance)
No correction	凶手到月池现实亚镇指认骗走孩子的现场。 (The murderer went to Yueche countryish, Siya Town to ...)	
SC_BART	凶手到月池现实亚锦指认骗走孩子的现场。 (The murderer went to Yueche countryish, Siya towne to ...)	
DiCODER$_{CD}$	凶手到岳池县石垭镇指认骗走孩子的现场。 (The murderer went to Yuechi County, Shiya Town to ...)	岳池县; 石垭镇 (Yuechi County; Shiya Town)
No correction	涨去科技年内涨超二倍。 (RocketFunTech rose more than two times in a year.)	
SC_BART	思科技年内涨超二倍。 (ThinkTech rose more than two times in a year.)	
DiCODER$_{CD}$	掌趣科技年内涨超二倍。 (PocketFunTech rose more than two times in a year.)	掌趣科技 (PocketFunTech)

strategy can effectively suppress the appearance of hallucinated entities. The average HR on three datasets for DiCODER$_{CD}$ is 27.41%, 18.23% and 1.77% lower than those of SC_BART, CSC_BART, and not using a controllable decoding strategy, respectively. It should be noted that the DiCODER result is from an ablation study in which the controllable decoding strategy was removed from DiCODER$_{CD}$. The DiCODER method adopts a more aggressive approach to correct entity errors, resulting in the improved recall at the expense of HR.

Additionally, we select some noisy entities from the MSRA dataset[8] and add them to our entity dictionary to observe the effect of our method. Figure 2 shows that our proposed method exhibits a certain degree of robustness to these noise.

F. Case Study

To demonstrate the properties of our model, we show several running examples in Table III. From these cases, we can tell that our model is capable to correct both

[8]https://github.com/OYE93/Chinese-NLP-Corpus/tree/master/NER/MSRA

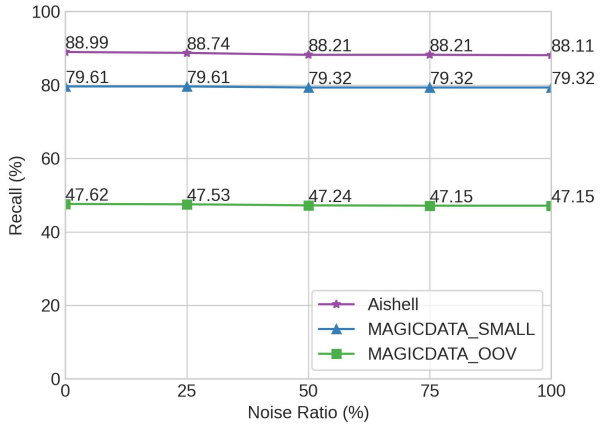


Fig. 2. The entity recall under different noise ratio, which is defined as the ratio of number of noise entities to the number of entities in original dictionary knowledge.

spelling errors and entity errors. The first two examples contain the spelling errors, such as "拨"("Ply") 和 "男一"("Clan"). Benefiting from the language model, both the baseline and our method can deal with this problem by changing "拨"("Ply") to "播"("Play"), "男一"("Clan") to "难以"("Can't")。

Moreover, the last three examples, containing the entity errors, such as "阿尔贝为尔"("Albertwell"), "月池现实亚镇"("Yueche countryish, Siya Town") 和 "涨去科技"("RocketFunTech"). Through leveraging the retrieved entity knowledge, our model can greatly correct these errors. Although the retrieved entity knowledge may contain noise, as seen in the case, the model has some resistance to noise and can tolerate retrieval of entity errors. In sentence "但直到一二年后的法国阿尔贝为尔冬奥会上" ("But not until the 2012 Winter Olympics in Albertwell, France.") the retrieved entities are "阿尔贝维尔; 韦尔" ("Albertville; Vance") by query "阿尔贝为尔; 为尔", where "韦尔" is noise, but it does not affect our error correction results. However, SC_BART relies on the language model alone, generating the hallucination entity "阿尔贝塞"("Ashville") in the results.

V. Conclusions

In this paper, we propose a method to solve the ASR entity error correction problem using entity dictionary knowledge. Experiments on three publicly available datasets show that the method can reduce SER/CER while improving entity error correction performance. However, because the knowledge retrieval component is specially designed for Chinese, it limits the application of our method. In the future, we will extend the retriever to adapt our method to more languages.

References

[1] C. Huang and Y. Chen, "Learning asr-robust contextualized embeddings for spoken language understanding," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, IEEE, 2020, pp. 8009–8013.

[2] Y. Weng, S. S. Miryala, C. Khatri, *et al.*, "Joint contextual modeling for ASR correction and language understanding," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, IEEE, 2020, pp. 6349–6353.

[3] A. Sarma and D. D. Palmer, "Context-based speech recognition error detection and correction," in *Proceedings of HLT-NAACL 2004: Short Papers, Boston, Massachusetts, USA, May 2-7, 2004*, The Association for Computational Linguistics, 2004.

[4] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, IEEE, 2019, pp. 5651–5655.

[5] Y. Zhao, X. Yang, J. Wang, Y. Gao, C. Yan, and Y. Zhou, "BART based semantic correction for mandarin automatic speech recognition system," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, ISCA, 2021, pp. 2017–2021.

[6] J. Yang, R. Li, and W. Peng, "ASR error correction with constrained decoding on operation prediction," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds., ISCA, 2022, pp. 3874–3878.

[7] H. Wang, S. Dong, Y. Liu, J. Logan, A. K. Agrawal, and Y. Liu, "ASR error correction with augmented transformer for entity retrieval," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, ISCA, 2020, pp. 1550–1554.

[8] D. Bekal, A. Shenoy, M. Sunkara, S. Bodapati, and K. Kirchhoff, "Remember the context! ASR slot error correction through memorization," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, IEEE, 2021, pp. 236–243.

[9] X. Wang, Y. Liu, S. Zhao, and J. Li, "A light-weight contextual spelling correction model for customizing transducer-based speech recognition systems," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno,*

Czechia, 30 August - 3 September 2021, ISCA, 2021, pp. 1982–1986.

[10] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, IEEE, 2019, pp. 5651–5655.

[11] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of automatic speech recognition with transformer sequence-to-sequence model," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, IEEE, 2020, pp. 7074–7078.

[12] W. Li, H. Di, L. Wang, K. Ouchi, and J. Lu, "Boost transformer with BERT and copying mechanism for ASR error correction," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, IEEE, 2021, pp. 1–6.

[13] D. Duan, S. Liang, Z. Han, and W. Yang, "Pinyin as a feature of neural machine translation for chinese speech recognition error correction," in *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, Springer, 2019, pp. 651–663.

[14] P. Serai, V. Sunder, and E. Fosler-Lussier, "Hallucination of speech recognition errors with sequence to sequence learning," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 890–900, 2022.

[15] R. Zhang, C. Pang, C. Zhang, *et al.*, "Correcting chinese spelling errors with phonetic pre-training," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, Association for Computational Linguistics, 2021, pp. 2250–2261.

[16] Y. Leng, X. Tan, R. Wang, *et al.*, "Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Association for Computational Linguistics, 2021, pp. 4328–4337.

[17] N. Das, D. H. Chau, M. Sunkara, S. Bodapati, D. Bekal, and K. Kirchhoff, "Listen, know and spell: Knowledge-infused subword modeling for improving ASR performance of OOV named entities," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, IEEE, 2022, pp. 7887–7891.

[18] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017*, IEEE, 2017, pp. 1–5.

[19] B. Chen, G. Xu, X. Wang, P. Xie, M. Zhang, and F. Huang, "AISHELL-NER: named entity recognition from chinese speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, IEEE, 2022, pp. 8352–8356.

[20] Z. Yao, D. Wu, X. Wang, *et al.*, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, ISCA, 2021, pp. 4054–4058.

[21] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of automatic speech recognition with transformer sequence-to-sequence model," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, IEEE, 2020, pp. 7074–7078.

[22] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Association for Computational Linguistics, 2020, pp. 7871–7880.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Method of Efficient Synthesizing Post-disaster Remote Sensing Image with Diffusion Model and LLM

Ruizhe Ou, Haotian Yan, Ming Wu, Chuang Zhang
Beijing University of Posts and Telecommunications, China
E-mail: ouruizhe@bupt.edu.cn

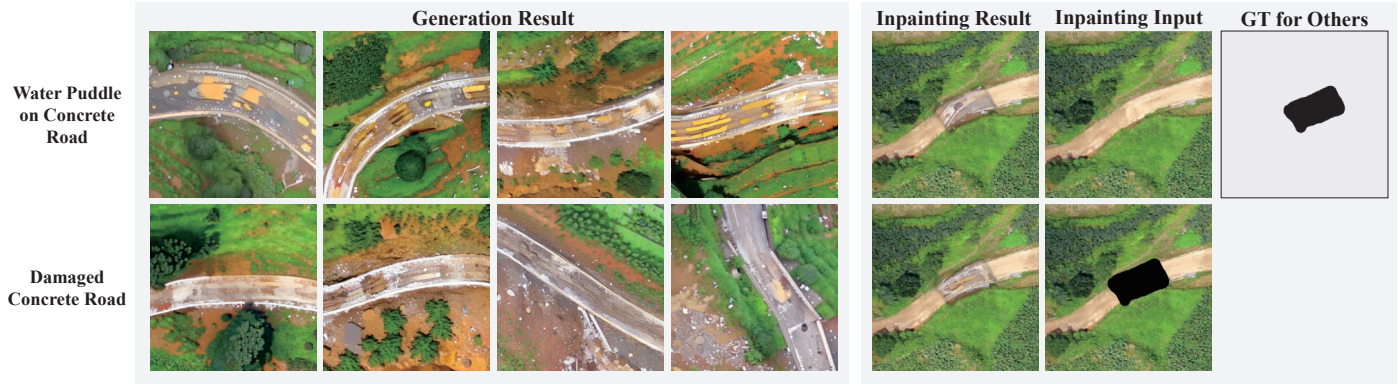


Fig. 1: **Synthesizing Remote Sensing Images Affected by Disaster under Natural Language Guidance with Diffusion Model and LLM.** (1)The images in left block are synthesized in generation manner. The text to the left of the image is condition promtps for synthesizing by the row of images. (2)The images in right block are synthesized in inpainting manner under the same condition of the generation. In addition, synthesizing the images with inpainting manner can obtain the ground truth of disaster as the position label for other interpretation model learning.

***Abstract*—Due to the fact that current deep learning models are typically driven by big data, existing interpretation models for emergency management lack relevant learning data. However, existing pre-trained image generative models cannot directly generate post-disaster remote sensing images without fine-tuning. In this paper, we demonstrate the ability of natural language guidance synthesizing remote sensing imagery affected by disaster by pre-trained image generative model fine-tuned with very few unlabelled images (i.e., less than 100 fine-tuning images) at very low training cost (i.e., one 2080Ti GPU). To trade for lower cost, we embrace the trend of large model, leveraging a pre-trained caption model, GPT-4 and a pre-trained text-to-image Stable Diffusion model for this task. The Stable Diffusion Model, fine-tuned with our method, successfully synthesizes remote sensing images affected by disasters using natural language guidance in both image inpainting and image generation tasks. In addition, the ground truth for other interpretation models learning. With this achievement, our method can synthesize a large amount of data for the emergency management interpretation model to learn when there is less existing data, only unlabelled data and less time, so as to achieve better interpretation performance. Furthermore, our approach highlights the significant of combining human feedback with large models in synthesizing data which is out of the prior knowledge of large model, especially when there is less data available and less computational power available.**

I. Introduction

Due to the unpredictability of disasters and the reliance of existing algorithms on large-scale datasets, there is a lack of learning data for interpretation models[1] serving in emergency management. The pressing issue is how to utilise rapidly advancing image generative models to synthesize post-disaster remote sensing images with limited available data, providing high-quality remote sensing data for interpretation models. Previous disaster remote sensing image synthesis methods primarily include generating disaster-affected remote sensing images by image generative models and inpainting masked position on the image[2] that is not affected by the disaster by image translation model[3]. Nevertheless, the previous methods suffer from several drawbacks: (i) the computational cost of training or fine-tuning an model is exceedingly high; (ii) they required extensive labelled (e.g., mask, class or prompt) data. In this paper, we propose a remote sensing imagery affected by disaster synthesizing method that mitigates all the above problems. Given only a few of unlabelled post-disaster for fine-tuning image generative model, our method can synthesizing high-quality disaster remote sensing imagery under natural language guidance in inpainting way and in

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

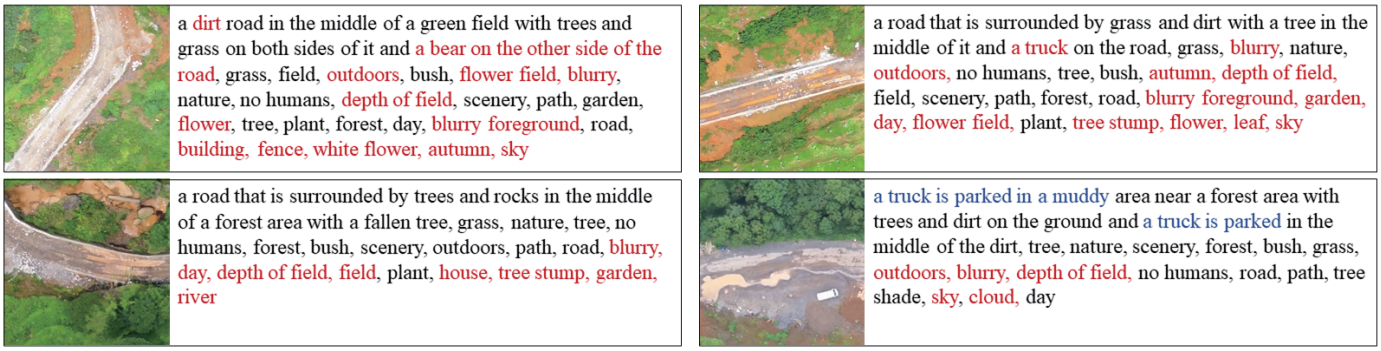


Fig. 2: **The Prompts Captioned by BLIP.** The red words are wrongly captioned by BLIP. There is not semantic information about "remote sensing image" in the prompts captioned by BLIP. Interestingly, The BLIP succeed in captioning the semantic information of small object (e.g, a truck) in a aerial view, which shows BLIP is a very powerful captioning model to caption the semantic information of the object which is not related to disaster.

generation way, as demonstrated at Fig.1. In inpainting way, the label of image which is the mask for inpainting can be the ground truth for other interpretation models training. In generation way, numerous of unlablled post-disaster remote sensing images can be synthesized under natural language guidance.

To achieve this feat, embracing the trend of large mode, we divide this task into three steps:

1) Caption textual prompts remote sensing image affected by disaster to obtain the semantic information of the image.
2) Optimize prompt with human feedback and GPT-4[4] to supply the semantic information of disaster.
3) Fine-tuning Stable Diffusion Model[5] with optimized prompts and post-disaster remote sensing image to capture the image-specific appearance.

Finally, we achieve synthesizing remote sensing image affected by disaster with fine-tuned Stable Diffusion Model under natural language guidance in both image inpainting manner and image generation manner. These results showcase the generality, versatility, and quality of our method in synthesizing out-of-domain data of the large model. We additionally conduct an ablation study, highlighting the effect of each element of our method. We summarize our main contributions as follows:

1) Illustrating a pioneering method for synthesizing post-disaster remote sensing images.
2) Presenting a novel way to cost-effective data augmentation using pre-trained large model.

II. Related Work

A. Image Generative Models and Fine-tuning Strategy

1) Diffusion Models (DMs): DMs[6]–[9] have become state-of-art image generative image. The core premise of Diffusion Model is to initialize synthesizing with a randomly sampled noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ then iteratively denoise it in a learned fashion, until it is synthesized into a learning image $\mathbf{x}_0$ intermediate step satisfies:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \tag{1}$$

where α_t being the hyperparameter of noise-adding process t (i.e., diffusion process), $\boldsymbol{\epsilon}$ is the standard Gaussian noise.

However, these models evaluate and optimize in pixel space, result in slow inference speed and excessively high computational costs.

2) Latent diffusion model (LDM): LDM[5] is one of the DMs, which performs in the latent space, while incorporating a cross-attention[10] mechanism into the denoising process for conditioning the synthesize process. LDM achieves powerful and flexible synthesis conditioning input by introducing a domain specific encoder τ_θ to project conditioning input y to an intermediate representation $\tau_\theta(y)$, which augmenting backbone via a cross-attention layer.

3) Stable Diffusion Model (SDM): SDM, a LDM trained on the LAION-5B[11] dataset that contains 5.85 billion high-quality image-text pairs, has exhibited impressive results in various generative tasks. Users can synthesize realistic high-resolution images using only a text prompt describing the desired scene. However, for synthesizing post-disaster remote sensing imagery, Stable Diffusion requires fine-tuning on post-disaster remote sensing images to achieve better performance.

4) SDM Adpator - Hypernetworks: Hypernetworks[12] involves training an auxiliary network to generate the weights of main network. In 2021, NovelAI introduced a new approach to control model generation, which apply a single small neural network without changing any other part of generation model, either a linear layer or multi-layer perceptron, at cross-attention layer of the larger network to modify hidden states. Since the Hypernetworks is smaller, training is fast and demands limited computational resources. Training Hypernetworks as adaptor inserting into Stable Diffusion model instead of directly fine-tuning it, the computational overhead can be significantly reduced.

B. Iamge Captioning

BLIP: Bootstrapping Language Image Pre-training[13], a novel vision-language pre-training (VLP) framework for vision-language understanding and generation, including image-text retrieval and image captioning. The pre-trained

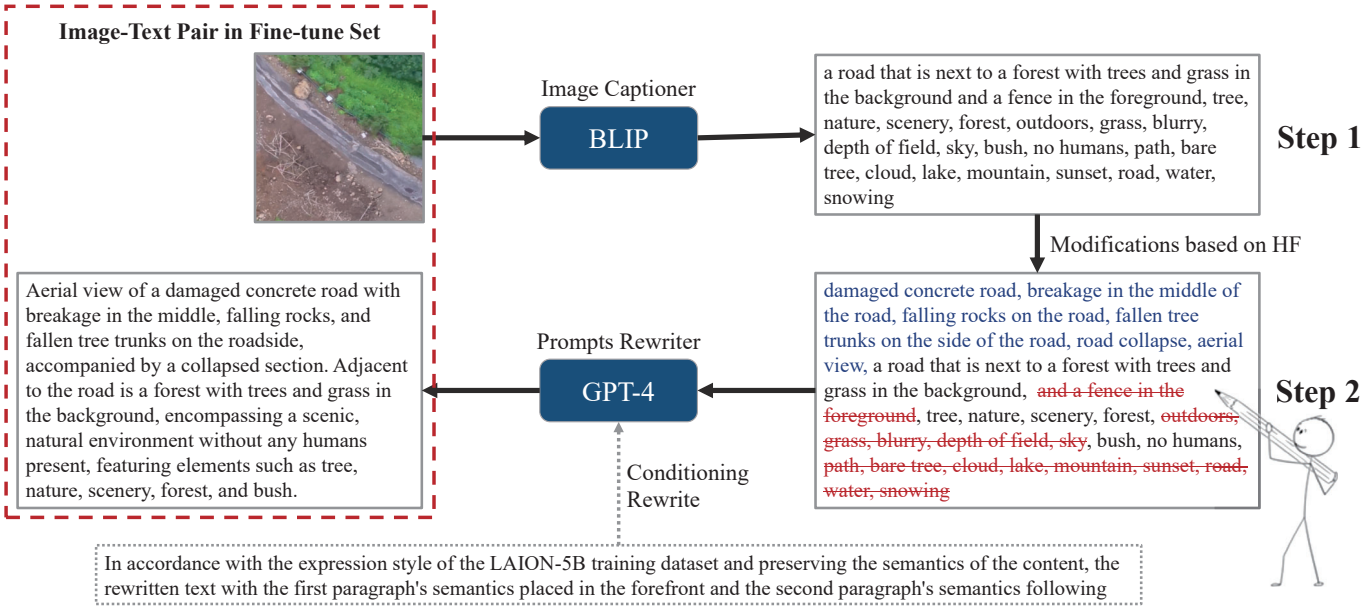


Fig. 3: **The Process of Image Captioning of Our Approach.** In step 1, the semantic information of the object in the image is captioned by BLIP. In step 2, with the help of human feedback, the errors in captioned prompts are removed and supplement the semantic information of disaster at the beginning of the prompts. The bottom statement circled in a box is the prompts for GPT-4 to rewrite prompts for the image according with the writing style of the Stable Diffusion Model training dataset.

model demonstrated poor performance in post-disaster remote sensing image captioning tasks, as evidenced by the results presented in Fig.2. Given the limited availability of post-disaster remote sensing images and the absence of any textual annotations within these images, we must consider the potential detrimental effects of catastrophic forgetting when directly fine-tuning a pre-trained BLIP model to gain a better captioning performance.

In light of the challenge, we propose an alternative approach in this paper, wherein we incorporate validated human feedback to tune the captioned prompts. This enables us to retain the semantic information derived from the BLIP while introducing prompts that align with the specific semantic context of post-disaster remote sensing image.

III. OUR METHOD

In order to achieve efficient post-disaster remote sensing image synthesis, we choose to embrace the large models, that is, to use pre-trained large models as much as possible, with as little fine-tuning as possible. We divide synthesizing task into three steps:

1) Captioning semantic information of the image by BLIP to get prompt words that contain as much semantic information as possible, even if they contain errors.
2) Optimize prompts with human feedback and GPT-4. Remove the errors in captioned prompts and supple the semantic information of disaster. GPT-4 is then used to rephrase the prompts using the writing style of the LAION-5B dataset while maintaining semantic integrity.
3) ine-tuning Stable Diffusion Model with Hypernetworks to minimize fine-tuning cost. The fine-tuning data contains optimized prompts and the images

A. Captioning

In captioning, the pre-trained model we used is a open-source *BLIP w/ ViT-L* model that trained with 129 million images, with ViT-L[14] as the vision backbone.

B. Optimize prompts with Human Feedback and GPT-4

The process is shown in Fig.3. We optimize the captioned prompts based on human feedback (HF). Human experts can directly remove prompts they perceive as defective and supplement with concise terms or natural language expressions to describe disaster-related information they deem missing. When adding the missing disaster-related information, human experts should first describe the characteristics of the disaster-affected locations, and then, centred around these damaged locations, gradually expand their descriptions outward. Examples of human experts supplementing missing disaster descriptions can be found in Fig.3. The prompts composed by human are positioned at the beginning of the captioned prompts to increasing their weight.

To mitigate potential biases in image descriptions due to language usage issues on the part of human feedback, we utilise the GPT-4 as a rewriter for the human-introduced prompts. The GPT-4 maps its original expression to that of the Stable Diffusion model training dataset, ensuring that the adjusted prompts comply with the prior knowledge of the pre-trained Stable Diffusion Model. By adopting this method, we generate the prompts that combine the rich semantic

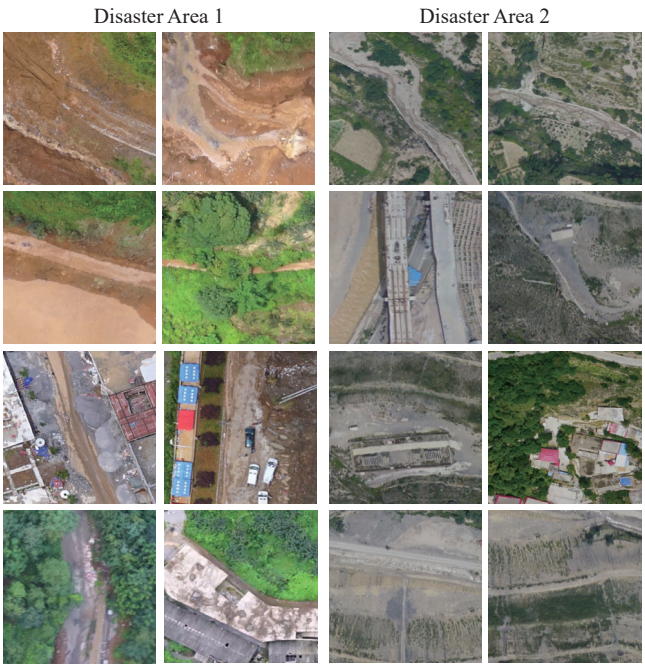


Fig. 4: **The Images that were Filtered Out.** A. The images that unable to identify the road; B. The images that unable to identify damage; C. The images with cluttered backgrounds; D. Blurry and low quality. Moreover, The geographical features of remote sensing images in different regions are different (i.e., the color, the distribution, etc.)

information provided by the BLIP model and the natural language information introduced through human expert feedback. Furthermore, these prompts are articulated in a manner that is readily comprehensible by the Stable Diffusion model.

C. Fine-tuning Stable Diffusion model with Hypernetworks

We utilize a simplistic three layers fully connected neural network, designated as an adaptor for Stable Diffusion to lower fine-tuning cost. The Hypernetworks are inserted into the cross-attention layers of the Stable Diffusion model, where the dimensions of key and value vector are 640, 768, 1024 and 1280. The dimension of the hidden layer of Hypernetworks we used is three times of the dimension of the input layer and output layer, as it called $[1, 3, 1]$ Hypernetworks. We initialized it with XavierNormal[15] which initializes weights with Gaussian noise centred at 0 and sets biases to 0.

IV. EXPERIMENT

A. Dataset

In this study, we utilized drone-captured remote sensing imagery from two regions, namely Disaster Area 1 and Disaster Area 2, as the dataset for our study. In order to improve the quality of fine-tune data, the image that unable to identify the road, unable to identify damage, cluttered background and blurry will be removed. Fig.4 is the example of images that have been filtered out. After filtering, compared to the number of images in the Stable Diffusion model training set, only very little data is available for our fine-tuning. The size of fine-tuning set is showed in Table.I.

TABLE I: Size of Fine-tuning Set

Area No.	Scale	With disaster	Without disaster	Tot.
1	Std.	37	61	98

To explore the generalization of our model, we added total 170 remote sensing images without disaster that contains 49 images with smaller scale from Disaster Area 1 and 60 images with same scale from Disaster Area 2 to the inpainting set. The size of inpainting set is showed in Table.II

TABLE II: Size of Inpainting Set

Area No.	Scale	Without disaster	Tot.
1	Std.	61	170
1	Smaller	49	
2	Std.	60	

B. Implementation details

The $[1, 3, 1]$ Hypernetworks was trained using 98 remote sensing images affected by disaster for 20,000 steps. When training, we unload VAE[16] and CLIP[17] from VRAM. Experimentally, We employed a training strategy of gradually decreasing the learning rate in stages, where a larger learning rate was used at the beginning to activate learning and then lowered artificially as the training progressed to ensure model convergence. Specifically, the learning rate adjustment strategy was as follow:

$$lr_s = \begin{cases} 5e-5 & 0 \le s \le 100 \\ 5e-6 & 100 \le s \le 1,500 \\ 5e-7 & 1,500 \le s \le 10,000 \\ 5e-8 & 10,000 \le s \le 20,000 \end{cases}$$

where s stands for step. The version of pre-trained Stable Diffusion is 2.1. All the experiments are conducted on ONLY ONE GPU - NVIDIA 2080Ti.

C. Ablation Study

1) Prompts optimization: To validate the effectiveness of the prompts optimization based on human feedback, we conducted three sets of ablation experiments. In these experiments, the remote sensing images affected by disaster are synthesized by the Stable Diffusion Model that is fine-tuned with $[1, 3, 1]$ Hypernetworks based on the prompts synthesized in three ways. In the results, as shown in Fig.5 of these three ablation experiments, it can be observed:

- Only using captioned prompts for fine-tuning, the fine-tuned Stable Diffusion Model do not support natural language guidance, as the expression format of captioned prompts significantly differs from that of natural language.

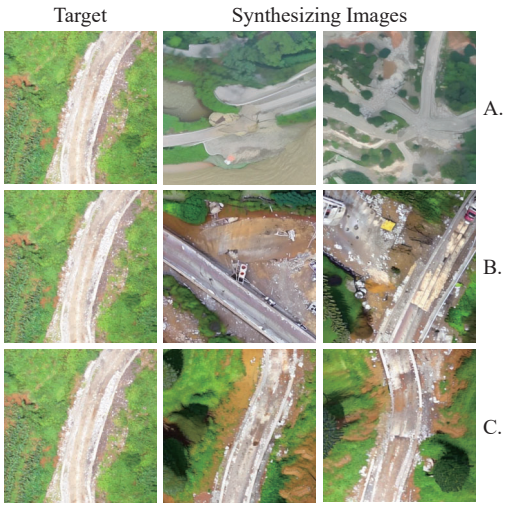


Fig. 5: **Effectiveness of prompts optimization.** A. The images were synthesizing by the model fine-tuned with captioned prompts; B. The images were synthesizing by the model fine-tuned with hand-writing prompts; C. The images were synthesizing by the model fine-tuned with the prompts with our prompts optimization method.

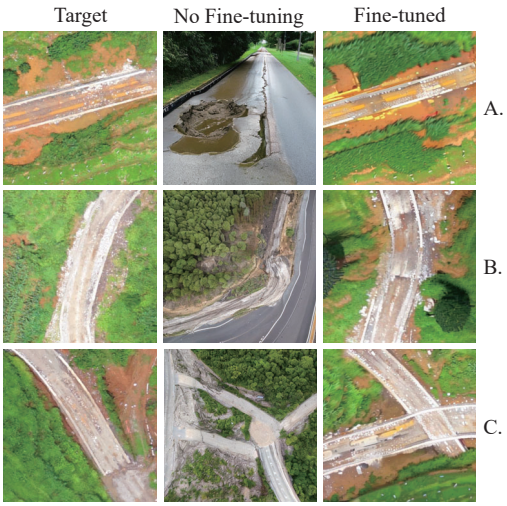


Fig. 6: **Effectiveness of fine-tuning.** A. The images synthesized under "water puddle on concrete road" guidance; B. The images synthesized under "damage concrete road" guidance; C. The images synthesized under "the concrete road is broken in the middle".

- Only using human-composed prompts for fine-tuning, the synthesized images are relatively chaotic in areas other than the damaged parts, as the inability of human-composed prompts to comprehensively describe semantic information of the images used for fine-tuning.

The model that fine-tuned using prompts that synthesized with our method produces best results. It successfully synthesizes remote sensing images affected by disasters similar to the expected images, and also realizes the synthesis guided by natural language.

TABLE III: Fine-tuning Comparison

Fine-tune	FID	SSIM
No Fine-tuning	83.22	0.18
Fine-tuned with $[1, 3, 1]$ Hypernetworks	83.42	**0.31**

D. Fine-tuning

To demonstrate the effect of fine-tuning on the output images, we synthesize images using the Stable Diffusion Model with or without the fine-tuning. we calculated the Fréchet Inception Distance (FID) and Structural Similarity Index (SSIM) between the synthesized images and those containing the same damage semantics. The FID assesses the distribution similarity between two images, with a lower value indicating better. The SSIM measures the similarity in structure between two images, where a higher value suggests better similarity.

With the SSIM value in Table.III almost doubling after fine-tuning, which indicates that fine-tuning brings an enhancement of structural similarity between the synthesizing images and the reference images.

In addition, we observed that the pre-trained Stable Diffusion model struggles to effectively differentiate between semantic information related to specific damage forms, such as collapse and breakage, resulting in the generation of relatively similar images. A comparison between rows B. and C. in Fig.6 substantiates our conjecture regarding the model's inability to accurately distinguish between the distinct types of road damage.

E. Generalization of Synthesising by Inpainting

We conducted inpainting experiments on the images with small-scale from same area and and the images with scale same as learning image from different area to assess the generalization capability of the model.

Observing Fig.7 to analyze the generalization of the model. When using smaller-scale remote sensing images containing more information and applying the same textual prompts as used for large-scale images, the model is struggle for inpainting as the condition. It may simply inpainting the mask with surrounding pixels. For Inpainting on the image of the same scale but from different area, the synthesized image aligns the guidance prompts, but there are some conflicts between the edges of the synthesized content and the input remote sensing images.

F. Limitations

Trading for low fine-tuning cost, all the processes in our method perform at word level or pixel level, the model only learned small amount of semantic information and the feature information of the learning data. Furthermore, based on the assumption that there is less data available for the task,

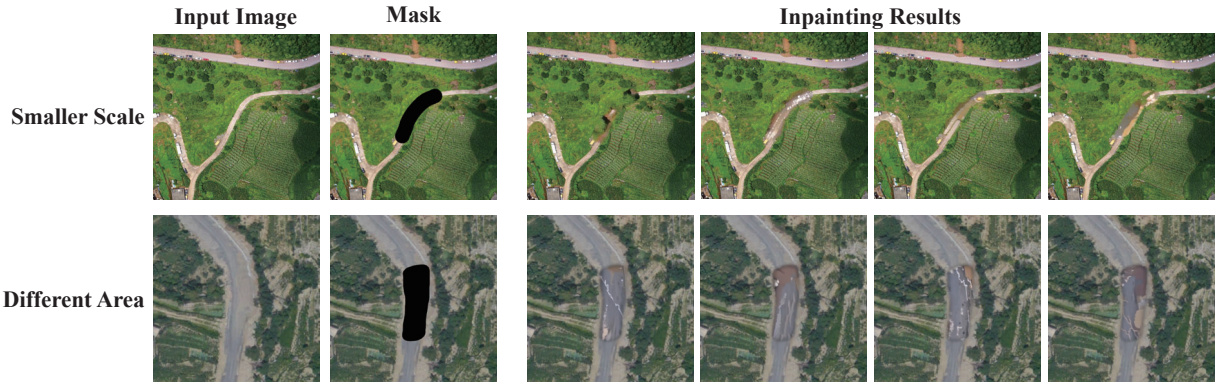


Fig. 7: **Results of Synthesis in the Context of Our Method, Demonstrating Model Generalization Capability.** The synthesized results showcased in this figure are generated using an inpainting technique on images that are devoid of any disaster-related semantics information. The images in the upper row are the outcomes of synthesis performed within the same geographical area, albeit at smaller scales. Conversely, the images in the lower row are the results of synthesis carried out within distinct geographical areas, while maintaining a consistent scale.

the result is that the distribution of synthesizing remote sensing images affected by disaster is quite similar to the learning ones. Considering these two limitations, the model fine-tuned with our method struggles in synthesizing the image via inpainting manner that inpaint on the image containing more semantic information or different semantic information.

V. Conclusions

In this paper, we propose a method to synthesize remote sensing image affected by disaster with Stable Diffusion and GPT-4 under natural languadge guidance in the condition that the data is limited and the data is unlabelled.

To achieve this feat, we firstly synthesize the textual prompts for the image and treat it as the label for training. In our method, a text prompt is captioned by BLIP at the beginning, then by introducing hand-writing prompts through human feedback, we supplement the disaster impact semantics missing in the captioned prompts. GPT-4 is finally used to map the human-feedback-modified prompts to the expression format of the pre-trained Stable Diffusion model's training dataset. After obtaining textual prompts, the images and the prompts are combined to fine-tune Stable Diffusion Model to capture the appearance of remote sensing image affected by disaster. The model fine-tuned with our method succeeds in synthesizing remote sensing images under natural language conditioning in both image inpainting manner and image generation manner.

The model fine-tuned with our method only takes one 2080Ti-GPU and less than 100 images to complete fine-tuning. However, trading to low training cost, the model struggles in synthesizing the image with different scale or containing different senmantic information from the learning dataset.

Based on the fact that interpretation models in emergency management often need to interpret specific types of disasters with only a small amount of data for learning and in a short time, our method can synthesize a large amount of data for the emergency management interpretation model to learn when there is less existing data and time, so as to achieve better interpretation performance. With this achievement, our approach highlights the significant of combining human feedback with large models in out-of-domain data synthesis tasks (e.g., prompt synthesis and image synthesis), especially when there is less data available and less computational power available.

Acknowledgment

I would like to give my sincere thanks to my supervisor, with patience, gave me lots of valuable advice. Then, I also appreciate my family and friends, they support me a lot during my research.

References

[1] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 182–186.

[2] Y. Huang, H. Wei, J. Yang, and M. Wu, "Damaged road extraction based on simulated post-disaster remote sensing images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE, 2021, pp. 4684–4687.

[3] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5143–5153.

[4] OpenAI, *Gpt-4 technical report*, 2023. arXiv: 2303.08774 [cs.CL].

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[7] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[8] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] C. Schuhmann, R. Beaumont, R. Vencu, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[12] D. Ha, A. Dai, and Q. V. Le, *Hypernetworks*, 2016. arXiv: 1609.09106 [cs.LG].

[13] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, PMLR, 2022, pp. 12 888–12 900.

[14] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Privacy-oriented Coded Caching in Mobile Information-centric Networking

Binchen Yang, Yu Guo and Xingyan Chen
Financial Intelligence and Financial Engineering Key Laboratory of Sichuan Province
Southwestern University of Finance and Economics
E-mail: {yangbc, guoyugy}@smail.swufe.edu.cn, xychen@swufe.edu.cn

***Abstract*—Wireless network communication has the characteristics of broadcasting and bandwidth limitation, which makes it difficult for devices in the wireless network to obtain high-quality privacy protection and service. The difficulty of achieving both privacy and service quality exacerbates the challenge of privacy-oriented content caching optimization. To solve this problem, coded caching technology offers a new way, but there is little research on the integration of this technology with privacy protection. We propose a privacy-protecting coded caching framework named PCC for mobile Information-Centric Networks (ICNs). Firstly, we discuss the privacy leakage problems existing in the uncoded caching techniques of ICNs and propose an effective adversary model and a privacy protection strength evaluation model for caching in ICNs based on the Shannon cipher system. Secondly, we design a content encoding mechanism based on machine learning to solve the privacy leakage issue in caching, relying on a framework that predicts changes in network fluid nodes. We further propose a joint optimization problem for content caching and privacy protection and propose a privacy-oriented coded caching algorithm called PCC. Compared with existing solutions, our solution not only provides an improvement in privacy protection, but also achieves performance improvements in latency, throughput, and bandwidth savings.**

I. Introduction

Thanks to advances in wireless communication and AI, we've seen a surge in smart mobile devices and apps in wireless networks. These services aim for high-performance, secure, privacy-oriented digital manufacturing, causing massive traffic on the networks. Cisco's annual report[1] forecasts that global wireless network device connections will rise from 8.8 billion in 2018 to 13.1 billion by 2023, reflecting a 1.48-fold overall growth. This surge in mobile services and Device-to-Device(D2D) connections pose twin challenges of cache efficiency and data privacy preservation due to limited bandwidth and poor privacy protection. To mitigate these issues, content caching has been widely researched to lower delay and network traffic.

The ICNs [2] architecture has been proposed to address these limitations. In ICNs, nodes assume varying roles including data consumer, content provider, and forward relay, supporting content-centric services in mobile networks [3]. Content request involves a data consumer seeking content by name, the network identifying providers and maintaining a route between consumer and providers via the forward relay, and providers returning the requested content. ICNs provide a seamless user experience during location changes and allow proactive content caching, reducing content access latency. However, the design incurs privacy issues [4], [5]. Therefore, devising a cache mechanism preventing privacy leaks while accommodating mobile communication demand is optimal.

Traditional uncoded caching strategies cache uniformly-sized content chunks at the network edge. However, their inability to adapt to network dynamics results in inefficient cache resource utilization and compromised privacy [6]. Coded content caching, a potential alternative, has spurred research in mobile services [6]–[9] and data privacy [10]–[13]. Coded caching employs network coding to encode multiple content chunks into packets for distribution [6]. Unlike uncoded chunks, coded packets can simultaneously satisfy multiple users' requests via a broadcast channel. By increasing transmission capacity and reducing delivery latency, coded caching has gained traction in wireless mobile networks [7], [8]. Specifically, [14] proposed the MATO-CAD algorithm, which utilizes user encoding caching, for the delivery of AR video content in the network. Yet, existing coded caching studies have limitations. For example, [6] leaves cached content uncoded until transmission, causing low cache utilization due to redundancy. [9] equally treats content in the coding phase, overlooking content features and relevance, leading to inefficient cache resource utilization.

Unencrypted content broadcasted over wireless networks faces significant privacy risks. V. Leung et al. [11] suggest a new eavesdropping scheme and propose a cooperation-based passive eavesdropping approach that goes unnoticed by victims. In response, authors in [10] propose a strategy for selecting interference jammers and relay nodes from intermediate nodes to improve eavesdropping attack security. In [12], they suggest XOR-based secure coded caching in wireless Ad Hoc networks, demonstrating that coded caching simplifies reactive routing and achieves near-perfect secrecy for D2D file transmissions. Moreover, [13] presents secure coded caching for cloud network storage. However, these strategies overlook network dynamics, impeding practical deployment in D2D-based networks, and require trusted nodes and secure channels to transmit decoding coefficients—a strict prerequisite for D2D-based mobile ICNs.

To counter these issues, we suggest a new coded caching mechanism aimed at jointly optimizing service performance and privacy protection. The main contributions are as follows:

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

1) We present an adversary model for XOR-secured caching, proving that with lesser cached coded content, privacy equals uncoded policy. We also offer a threat-quantifying privacy model.
2) We address data privacy by carefully devising a machine learning-based content coding architecture that integrates representational learning and deep belief networks(DBN).
3) Leveraging proposed architectures, we formulate the joint optimization problem of performance and privacy, developing a privacy-focused coded caching algorithm for optimization and protection without trusted nodes or secure channels.
4) We design a series of simulations to compare our algorithm with the state-of-the-art approaches [15][16]. Results demonstrate our algorithm outperforms state-of-the-art solutions in terms of privacy protection intensity, latency, throughput and resource consumption.

II. Related Works

Several papers have contributed to uncoded and coded caching mechanism designs. [17] introduced a novel uncoded caching network topology that focuses on a combination network with end-user caches. [9] examined decentralized coded content cache placement in heterogeneous networks, highlighting the throughput advantages of coded strategies. [18] proposed novel caching schemes based on erasure codes to achieve low latency in distributed coded storage systems.

Comparatively, [19]proposed a joint MDS codes and weighted graph based coded caching policy to minimize the fronthaul load. [20] suggested a knapsack optimization-based content caching protocol for intelligent traffic networks, enhancing cache utilization of roadside units. [21] proposed a matrix completion-based method for accurate content popularity prediction and cache placement. [12] discussed a decentralized coded caching technique using XOR operations for improved secrecy and simplified routing. [22] proposed a privacy-preserving design to resist quantum attacks in intelligent traffic networks. [4] proposed a privacy-aware caching solution to counter timing attacks in ICNs, albeit sacrificing service performance.

In a prior work [16], we proposed a novel fluid-based structure for 5G ICNs integrated with Device-to-Device communications. We employed an optimal centralized algorithm aimed at minimizing content access latency and caching resources for multimedia services. Further, an optimal coded content caching solution was proposed, leveraging a DBN [23]. This fluid-based model helped simulate state transitions of nodes under coded content caching. A cloud-assisted caching algorithm was designed, encompassing a code content selection mechanism and an optimal coded content distribution, to enable low-latency, cost-effective caching for mobile nodes. However, many studies, including ours, overlooked privacy performance for data and users in the mobile environment.

III. System, Adversary and Privacy Model

A. Preparations

In our model, each mobile node is equipped with D2D interface and caching capabilities, which makes it possible to build a cache-enable D2D network [24]. We consider unicast-based request packet forwarding strategy [25]. From a service perspective, according to [16], we assume that such mobile nodes play any of the three roles: *consumer*, *producer* and *forwarder*, each role follows the same definition in [16]. Besides, we assume the contents are divided into multiple chunks with equal size [16]. The requesting rate $q(i)$ of each chunk follows the Zipf distribution with parameter ρ. The Random Way Point (RWP) model is employed for mobility description. We assume that an adversary can eavesdrop and decipher all of the ciphertexts in the consumer request packet.

B. File Recover with Coded Packets

For capacity analysis, we represent the library of contents by a set $\mathscr{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_m\}$, where m is the total amount of content. We assume that each content f_i contains k number of chunks as $\boldsymbol{f}_i = \{p_{i,1}, p_{i,2}, ..., p_{i,k}\}$. Since the content is rich in the real-world system and the popularity of different content is disparate, coding everything together is infeasible. Therefore, we consider a classification-based coding strategy and divide all of these chunks into K different categories. Besides, we denote $\mathscr{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, ..., \boldsymbol{l}_K\}$ as the set of all coded categories and for the category j we have $\boldsymbol{l}_j = \{l_j^1, l_j^2, ..., l_j^{N_j}\}$, $j \in [1, K]$ as the set of chunks in class j where $N_j = |\boldsymbol{l}_j|$ is the size of $\boldsymbol{l}_j$. According to [9], for each class, we randomly select chunks from set $\boldsymbol{l}_j$ with probability 0.5 and then mix the selected chunks via the XOR operation to create a coded packet. The total amount of coded packets is $2^{N_j} - 1$. The i-th coded packet of category j can be represented as:

$$g_j^i = \sum_{n=1}^{N_j} a_j^n l_j^n = \boldsymbol{a}_j \boldsymbol{l}_j^T \tag{1}$$

where $\boldsymbol{a}_j = [a_j^1, a_j^1, .., a_j^{N_j}]$ is a vector, and each of its components follows a binary random distribution where $\boldsymbol{a}_j$ is a uniformly distributed vector of coding counters with binary elements and the summation is carried over Galois Field (GF) of two elements (0,1). In this way, if the mobile nodes find a set of nodes in the forward path with enough coded packets containing N linearly independent coded vectors, they can collaborate to decode the chunk. As shown in Fig. 1, considering the content set of the current network is $\{A, B, C, D, E\}$, we assume that node n_0 contains the coded content $\{B \oplus C \oplus E\}$. If n_0 requests content A, it will need uncoded content packet A or $\{A \oplus B \oplus C \oplus E\}$ for decoding A. Thus, n_0 sends out the request packet containing the *Counter Vector* $(1,0,0,0,0)$ and $(1,1,1,0,1)$ to the neighbor n_1. The receiving node n_1 does not contain the asked packets. Since n_1 contains $\{B \oplus E\}$, decoding the requested packet will only need any one of the packets A, $\{A \oplus B \oplus C \oplus E\}$, $\{A \oplus B \oplus E\}$ and $\{A \oplus C\}$. Thus, n_1 forwards the request with vectors $(1,0,0,0,0)$, $(1,1,1,0,1)$, $(1,1,0,0,1)$ and $(1,0,1,0,0)$ to next hop n_2 which has coded packet $\{A \oplus C\}$. n_2 passes the $\{A \oplus C\}$ to n_1, and n_1 obtains the

$\{A\oplus B\oplus C\oplus E\}$ by combining the $\{A\oplus C\}$ and $\{B\oplus E\}$. Then, n_1 returns the asked $\{A\oplus B\oplus C\oplus E\}$ to n_0 and n_0 decodes the A with $\{A\oplus B\oplus C\oplus E\}$ and $\{B\oplus C\oplus E\}$.

Based on above discussion, each mobile node can decodes the required p-th chunks via N_j linear independent packets. The decoding process can be represented by a equation $\sum_{i=1}^{2^{N_j}} b_j^i g_j^i$, where $b_j^i \in GF(2)$ denotes whether to perform the XOR operation with coded packet g_j^i over the decoding process. Then, we substitute the equation (1) into $\sum_{i=1}^{2^{N_j}} b_j^i g_j^i$ and we have $\sum_{i=1}^{2^{N_j}} b_j^i \boldsymbol{a}_j \boldsymbol{l}_j^T = \boldsymbol{v}_p \boldsymbol{a}_j$. Intuitively, we have vector $\boldsymbol{v}_p = \sum_{i=1}^{2^{N_j}} b_j^i \boldsymbol{l}_j^T$ and requiring 1 only for p-th component in vector $\boldsymbol{v}_p$ and 0 for other components.

C. Adversary Model

In this subsection, we will introduce the privacy security of the coded caching and then describe the adversary model of coded caching privacy. We can consider the XOR coded process as the encoding function of the Shannon cipher system. According to [12], if a consumer caches k coded packets, that means there are at most 2^k possible random keys (*Counter Vectors*) can be built. So the XOR coded caching strategy provides asymptotic perfect secrecy.

Unfortunately, this security mechanism will be broken by adversaries under our previous assumptions. We assume that the adversary can eavesdrop on the *Counter Vectors* which is sent by the consumers. The exposed *Counter Vectors* can help the adversary infer what the user requested. As shown in Fig. 1, the adversary eavesdrop that the *Counter Vectors* of the request packet are $(1,0,0,0,0)$ and $(1,1,1,0,1)$. Based on this, the adversary can directly infer that the content requested by the user is A. Let's consider a more general case, suppose the consumer caches k linearly independent coded packets and generates a request for the chunk $p_{i,j}$ which cannot be decoded by local caching. The set of *Counter Vectors* sent by the consumer must contain uncoded packets vector, such as $(1,0,0,0,0)$ or $(0,1,0,0,0)$. The adversary can filter out these uncoded vectors from the set of *Counter Vectors*. If the adversary gets only one uncoded vector, the consumer request can be inferred directly. If there are multiple uncoded vectors obtained by the adversary, the adversary can still narrow down user requests to a small range. According to linear algebra theory, the average number of uncoded vectors sent by the consumers is $1+\frac{N-1}{2^{N-k+1}}$ where N represents the total number of chunks encoded together. This means that encoding more chunks together (larger N) will result in a smaller number of uncoded vectors under a given caching capacity of each node.

Since the uncoded packet directly expose the privacy of user request, we believe that users should avoid requesting uncode packets directly and instead request coded content.

D. Privacy Model

Under the new adversary model, we find that the crux of privacy protection depends on the number of uncoded vectors. The more chunks that are encoded together, the better the privacy protection performance. So, we can assume that for the adversary, different coded vectors are equivalent. The adversary cannot distinguish which chunk is requested by the consumer based on the coded vector.

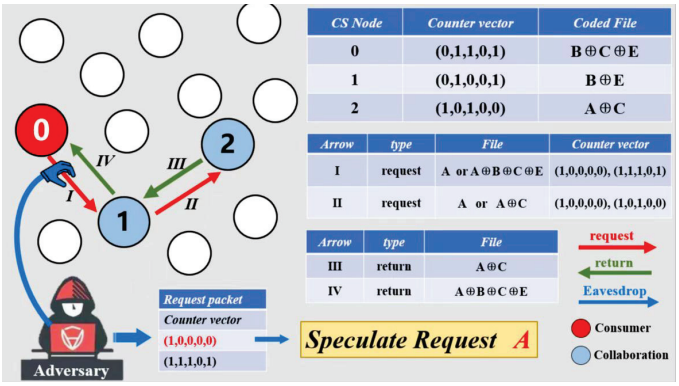


Fig. 1. D2D network decoding example and potential adversary model

We use entropy as the quantitative performance of privacy intensity which is positively correlated with the number of uncoded vectors $|\mathscr{A}|$. We denote $H(\mathscr{A})$ as the function of privacy intensity which is represented as:

$$H(\mathscr{A}) = -\sum_{i=1}^{|\mathscr{A}|} p_i(\mathscr{A}) \log p_i(\mathscr{A}) \tag{2}$$

where the $p_i(\mathscr{A}) = 1/|\mathscr{A}|$ and the privacy intensity can be rewritten as $H(\mathscr{A}) = \log|\mathscr{A}|$. The privacy intensity of the original XOR mechanism is $\log N$. This indicates that when we encode more chunks together, the privacy intensity $\log N$ is better. Thus, the issues of coded caching deployment and coded content selection have a significant impact on data privacy.

IV. The Architecture of Coded Content Caching

A. Fluid-based Architecture

We discuss the basic concepts of the fluid-based model for a given class $\boldsymbol{l}_j$. For detailed information, please refer to our early work [16], we have four roles for mobile nodes: **Consumer**, **Relay**, **Provider** and **Ordinary**. We use the following four bits, 1 represents existence:

- **Request bit** (R): consumer for the chunk
- **Forward bit** (F): relay for the chunk
- **Spread bit** (S): able to spread the request to neighbor
- **Have bit** (H): capable of decoding the corresponding chunk

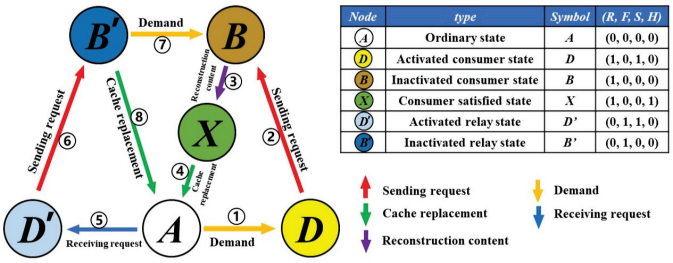
Unlike the [16] which considered the non-coded scenarios, we in this paper model six possible node states:

- **Ordinary state** $\boldsymbol{A}$ (R=0, F=0, S=0, H=0)
- **Activated consumer state** $\boldsymbol{D}$ (R=1, F=0, S=1, H=0)
- **Inactivated consumer state** $\boldsymbol{B}$ (R=1, F=0, S=0, H=0)
- **Consumer satisfied state** $\boldsymbol{X}$ (R=1, F=0, S=0, H=1)
- **Activated relay state** $\boldsymbol{D'}$(R=0, F=1, S=1, H=0)
- **Inactivated relay state** $\boldsymbol{B'}$ (R=0, F=1, S=0, H=0)

Each node in network is one of the six states, hence the sum of all nodes of each state is constantly equal to the total number of nodes in networks, namely:

$$A(t)+D(t)+B(t)+X(t)+D'(t)+B'(t)=1 \tag{3}$$

Further, we build a fluid-based model to describe the dynamics of those nodes. The transition of 6 states and 8 possible



Node	type	Symbol	(R, F, S, H)
A	Ordinary state	A	(0, 0, 0, 0)
D	Activated consumer state	D	(1, 0, 1, 0)
B	Inactivated consumer state	B	(1, 0, 0, 0)
X	Consumer satisfied state	X	(1, 0, 0, 1)
D'	Activated relay state	D'	(0, 1, 1, 0)
B'	Inactivated relay state	B'	(0, 1, 0, 0)

Fig. 2. Possible state transitions of nodes in coded content caching

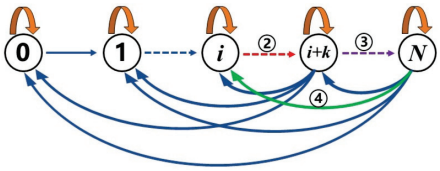


Fig. 3. The state space of the Markov chain for decoding

types of state transitions (represented by the dotted/solid line with an arrow) can be interpreted as Fig. 2.

The conversion rate of the above 8 transition processes is summarized in Table 1. According to Fig. 2 and Table 1, we can obtain the following O.D.E functions:

$$\dot{A}_j = X_j(t) + 2B'_j(t)/\mathbb{H} - \beta_j A_j(t) - A_j(t)\left(D_j(t) + D'_j(t)\right) \quad (4)$$

$$\dot{D}_j = \beta_j A_j(t) - D_j(t) \quad (5)$$

$$\dot{B}_j = D_j(t) - B_j(t)/\mathbb{H} + \beta_j B'_j(t) \quad (6)$$

$$\dot{X}_j = B_j(t)/\mathbb{H} - X_j(t) \quad (7)$$

$$\dot{D'}_j = A_j(t)\left(D_j(t) + D'_j(t)\right) - D'_j(t) \quad (8)$$

$$\dot{B'}_j = D'_j(t) - \beta_j B'_j(t) - 2B'_j(t)/\mathbb{H} \quad (9)$$

$$\boldsymbol{U}_j|_{t=t_0} = U_{j,t_0} \quad (10)$$

Where $\boldsymbol{U}_j$ represents the system state of the class j. Given the initial value U_{j,t_0} by

$$U_{j,t_0} = \left(A_j(t_0), D_j(t_0), B_j(t_0), X_j(t_0), D'_j(t_0), B'_j(t_0)\right)$$

B. Accuracy of the O.D.E Approximation

To validate the accuracy of proposed O.D.E function, we use the ndnSIM based on NS3 to simulate the real NDN environments and conduct a series of tests. Specifically, we build an NDN network with 1000 mobile nodes moving arbitrarily over a 3000 × 3000 square area. The RWP mobility model is used and its velocity range is set to $[3, 10]$m/s. Consider 10 different contents which are 10s long and consist of 5 chunks. All chunks are encoded randomly. The caching capacity of each mobile node in NDN is set to 5 chunks large.

TABLE I
STATE UPDATE AND CONVERSION RATE OF TRANSITIONS

Transition	State update	Conversion rate
1	(0,0,0,0) → (1,0,1,0)	$\beta_j A_j(t)$
2	(1,0,1,0) → (1,0,0,0)	$D_j(t)$
3	(1,0,0,0) → (1,0,0,1)	$B_j(t)/\mathbb{H}$
4	(1,0,0,1) → (0,0,0,0)	$X_j(t)$
5	(0,0,0,0) → (0,1,1,0)	$A_j(t)\left(D_j(t) + D'_j(t)\right)$
6	(0,1,1,0) → (0,1,0,0)	$D'_j(t)$
7	(0,1,0,0) → (1,0,0,0)	$\beta_j B'_j(t)$
8	(0,1,0,0) → (0,0,0,0)	$2B'_j(t)/\mathbb{H}$

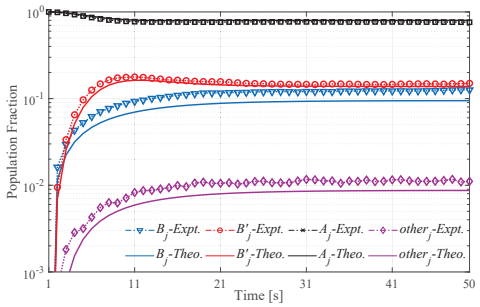


Fig. 4. Population fraction evolution in experimental and theoretical contexts

To eliminate the randomness, we repeat 20 times with different random seeds and take the average of the results. Fig.4 shows the comparison between theoretical and experimental values of the system evolution. Our model converges well with the simulation results. Hence, we can conclude that our model indeed describes the realistic system evolving.

C. Learn-based Content Coding Architecture

In this subsection, we first design a representation learning method based on word2vec to complete the embedding process, and then propose a four layers DBN for coded caching selection. We consider the wireless mobile network as shown in Fig. 5(a). We collect the chunk request records of all nodes and sort the records of each node in chronological order as the representation learning training set as shown in Fig. 5(b). We consider the records sequence of each node as the *text* and each record as the *word* of the *text*. Thus, we use the representation learning in Fig. 5(c). The representation learning model has three layers, the input layer selects z records before and after a specific record p, with each record corresponding to a request of a specific chunk. For all chunks, we initialize them with a N-dimensional vector $\boldsymbol{v}(p)$ with random value. Hence, the input layer contains $2z$ vectors with N-dimensional. In the projection layer, we calculate the sum of $2z$ vectors as the projection which can be expressed as $\boldsymbol{w_v} = \sum_{i=1}^{z}\left[\boldsymbol{v}_i^-(p) + \boldsymbol{v}_i^+(p)\right]$ where $\boldsymbol{v}_i^-(p)$ represent the i-th chunk in front of p and $\boldsymbol{v}_i^+(p)$ represent the i-th chunk behind. The output layer is a Huffman tree which is constructed based on the request frequency of each chunk. In the Huffman tree, each leaf node corresponds to a specific chunk, and each non-leaf node corresponds to a softmax function. This tree-based structure is a basic hierarchical softmax. For any $\boldsymbol{w_v}$, there always exists a path in the Huffman tree. The principle of this process is to calculate the probability of chunk p in a given context $\boldsymbol{w_v}$ which is derived as

$$P(p|w_v) = \prod_{x=1}^{n} P\left(d_x^w | \theta_{x-1}^w\right) \quad (11)$$

where $\theta_{x-1}^{\boldsymbol{w}}$ is the vector of x-th softmax function on the path and $d_x^{\boldsymbol{w}}$ is a binary result which is calculate by input $\theta_{x-1}^{\boldsymbol{w}}$ into x-th softmax function. If the reach leaf p'=p, give positive feedback on the path; otherwise, give negative feedback. As shown in Fig. 5(d), once the representations of chunks are learned by CBOW model, we use k-means clustering to divide the contents into K categories.

When the clustering is completed, we can obtain the coded packet by XOR process. We consider category $\boldsymbol{l}$ has N different

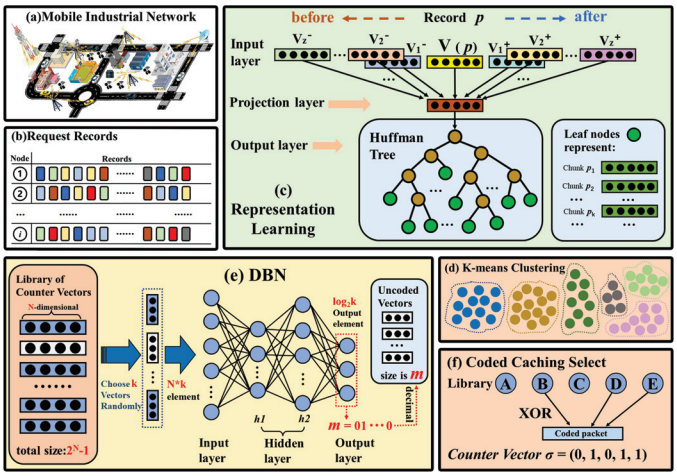


Fig. 5. The learning-based content coding architecture

chunks and each node caching k number of coded packets in this category. The total number of coded packets in this category is 2^N, how to choose k independent packets from such a large number of libraries to ensure privacy performance, is tough work. Thus, we propose a four layers DBN for coded caching selection. Fig. 5(e) illustrates the structure of DBN. The input layer consists of $k*n$. h_1 and h_2 consist n neurons and k neurons, respectively. And the size of output layer is $\lceil \log_2 k \rceil^+$, where $\lceil \cdot \rceil^+$ represents ceil. We randomly select k independent coded packets from the library as input and take the number of uncoded vectors as the label. Thus, we can obtain a large number of (input, label) pairs as our training set. We use this training set to train DBN in the forward and backward directions. Finally, as shown in Fig. 5(f), we use the trained DBN to guide decisions such as coded caching selection and replacement.

V. Joint Optimization of Performance and Privacy

A. Problem Formulation

Coded content caching balances the popularity of different contents and decouples the associativity of each chunk. Therefore, we need to consider "how many independent coded packets need to be cached?". For i-th mobile node, we denote $\eta_i = (k_1^i, k_2^i, ..., k_j^i, ..., k_{K_c}^i)$ as caching policy, where k_j^i indicates the number of independent coding packets in category j. The average number of coded packets cached in all nodes is equal to $\sum_{j=1}^{K_c} \bar{k}_j$. According to our previous assumptions, the more linear independent packets, the higher performance of data privacy and the less hop of content retrieve. However, we can't infinitely cache the coded packets. Thus, the objective function of the joint optimization problem for coded caching can be formulated as the tradeoff between delay, privacy and resource consumption, which can be represented as:

$$J_\eta = \sum_{j=1}^{K_c} J_\eta^j = \sum_{j=1}^{K_c} \left(\phi B_\eta^j \left(T_\eta^j \right) - \varphi H \left(\mathscr{A}_\eta^j \right) + \psi \bar{k}_j \right) \tag{12}$$

where $B_\eta^j(T_\eta^j)$ represents the value of $B_j(t)$ under the coded caching policy η at time T_η^j, T_η^j is the time when $B_\eta^j(t)$ reaches the peak value, namely $\{T_\eta^j | B_\eta^j(T_\eta^j) = \max_t B_\eta^j(t)\}$. k_j is the total scale of coded content in category j. ϕ, φ and ψ are the weight parameters where $\phi + \varphi + \psi = 1$. Thus, we can formulate the optimization as follows.

$$\boldsymbol{min} \quad J_\eta \tag{13}$$

$$\boldsymbol{s.t.} \quad \sum_{j=1}^{K_c} k_j^i \le C_i \quad \forall i \tag{14}$$

where C_i indicates the caching capacity of node i.

B. Optimal Control

In this subsection, we will discuss how to minimize the objective J_η by regulating the coded caching policy η. According to equation (12), J_η is separable for the caching category, we can consider the sub-problem J_η^j instead.

$$J_\eta^j = \phi B_\eta^j \left(T_\eta^j \right) - \varphi H \left(\mathscr{A}_\eta^j \right) + \psi \bar{k}_j \tag{15}$$

According to equation (15), we have the following lemma.

Lemma 5.1: When $B_\eta^j(T_\eta^j)$ and $H(\mathscr{A}_\eta^j)$ are continuous with respect to $\bar{k}_j$, we have J_η^j continuous with respect to $\bar{k}_j$.

The proof of Lemma 1 is intuitive. Since J_η^j is a linear function of $B_\eta^j(T_\eta^j)$ and $H(\mathscr{A}_\eta^j)$, we have J_η^j is continuous with respect to $\bar{k}_j$ based on the above condition.

Since the continuity of $H(\mathscr{A}_\eta^j)$ can be obtained from equation (2), here we focus on the continuity of $B_\eta^j(T_\eta^j)$. To simplify the description of O.D.E function (6)-(12), we rephrase the population fraction of each state with corresponding lowercase letter, such as b for $B(t)$, $\dot{b}$ for $\dot{B}(t)$.

$$\begin{aligned} \dot{a}_j = {} & x_j(t) + \alpha_j b'_j(t) - \beta_j a_j(t) \\ & - a_j(t) \left(d_j(t) + d'_j(t) \right) \end{aligned} \tag{16}$$

$$\dot{d}_j = \beta_j a_j(t) - d_j(t) \tag{17}$$

$$\dot{b}_j = d_j(t) - \gamma_j b_j(t) + \beta_j b'_j(t) \tag{18}$$

$$\dot{x}_j = \gamma_j b_j(t) - x_j(t) \tag{19}$$

$$\dot{d'}_j = a_j(t) \left(d_j(t) + d'_j(t) \right) - d'_j(t) \tag{20}$$

$$\dot{b'}_j = d'_j(t) - \beta_j b'_j(t) - \alpha_j b'_j(t) \tag{21}$$

where $\alpha_j = (\mathbb{H} - 1) / \left(\sum_{i=1}^{\mathbb{H}-1} i \right)$ and $\gamma_j = 1/\mathbb{H}$. Let $\boldsymbol{f}_j$ denotes as $(f_{a_j}, f_{d_j}, f_{b_j}, f_{x_j}, f_{d'_j}, f_{b'_j})$, which are the right side of the (15)-(20). According to (3), we have $(a_j, d_j, b_j, x_j, d'_j, b'_j) \in [0,1]^6$. Therefore the Jacobian of $\boldsymbol{f}_j$ is bounded and satisfy the Lipchitz condition. According to (14), we can limit our focus on state evaluation of $(b_j(t), b'_j(t))$.

$$\dot{b}_j = d_j(t) - \gamma_j b_j(t) + \beta_j b'_j(t) \tag{22}$$

$$\dot{b'}_j = d'_j(t) - \beta_j b'_j(t) - \alpha_j b'_j(t) \tag{23}$$

where β_j is constant and only related to the popularity of $\boldsymbol{l}_j$. As we can see, $b_j(t)$ and $b'_j(t)$ are the monotonically decreasing in terms of γ_j and α_j respectively. Hence, we know that $b_j(t)$ and $b'_j(t)$ are monotonically decreasing with the respect to $\bar{k}_j$.

Assume $\boldsymbol{\omega}_j^{(1)}(t) = (b_j^{(1)}(t), b_j'^{(1)}(t))$ and $\boldsymbol{\omega}_j^{(2)}(t) = (b_j^{(2)}(t), b_j'^{(2)}(t))$ are the trajectories of equation (24) and (25) under different caching policies η_1 with $\bar{k}_j^{(1)}$ and η_2 with $\bar{k}_j^{(2)}$. With loss of generality, we assume $\bar{k}_j^{(1)} > \bar{k}_j^{(2)}$, and denote $\Delta\gamma_j = \gamma_j^{(1)} - \gamma_j^{(2)}$ and $\Delta\alpha_j = \alpha_j^{(1)} - \alpha_j^{(2)}$. According to (17), we have following inequality.

$$||\boldsymbol{\omega}_j^{(1)} \left(T_\eta^j \right) - \boldsymbol{\omega}_j^{(2)} \left(T_\eta^j \right)||_2 \le C T_\eta^j \sqrt{\Delta\gamma_j^2 + \Delta\alpha_j^2}$$

where $C = \sqrt{2}(2+\beta_j)$. Since $\boldsymbol{f}_j$ is bounded, and T_η^j is finite, we have $b_j(T_\eta^j)$ continuous with respect to $\bar{k}_j$, where γ and α are continuous with respect to $\bar{k}_j$.

Fig. 6 shows the variation of $B_\eta^j(T_\eta^j)$, $H(\mathscr{A}_\eta^j)$ and J_η^j with respect to $\bar{k}_j$, where ϕ=0.1, φ=0.5 and ψ=0.4. In addition, the setting of other experimental parameters is the same as in *Section 3.B*. We can see $B_\eta^j(T_\eta^j)$ is monotonically decreasing and $H(\mathscr{A}_\eta^j)$ is monotonically increasing, which also proves the previous conclusion with $\bar{k}_j$. This explains the fact that the more coded packets cached in the mobile nodes, the fewer *consumers* are in the networks, and the better privacy performance of each node. For the value of objective function J_η^j, we can see the corresponding curve firstly experience a decreasing trend and then increase with respect to the $\bar{k}_j$. This is because as the cache fraction increases, the cost of networks cache increases linearly and other terms decrease exponentially, which results in the total increase. Therefore, we can find the optimal $\bar{k}_j$ and J_η^j through numerical calculation. Based on the numerical results, the optimal caching strategy is to let each mobile node cache 5 coded packets under the parameter settings.

Algorithm 1: PCC Algorithm

Input:
Choose the condition numbers ϕ, φ, $\psi \in (0,1)$;
Cloud Center Side:
foreach *time-slot T* **do**
  Collect the records of mobile nodes and $\boldsymbol{U}$;
  Chunks embedding, get K_c categories;
  foreach *category j* **do**
    Calculate the optimal cache policy $\boldsymbol{\eta}_j$;
    Train the DBN$_j$;
  end
  Broadcasts $(\boldsymbol{\eta}, \textbf{DBN})$ to all mobile nodes;
end
Mobile Nodes Side:
foreach *time-slot T* **do**
  Upload $\boldsymbol{U}$ and request records;
  Deployment the coded caching based on $(\boldsymbol{\eta}, \textbf{DBN})$;
end
final

C. Practical Algorithm

We assume that there is a cloud center (e.g. data center) that periodically (the update interval is T) collects the request records of each mobile node in the wireless mobile network system. The cloud maintains request records of each node and orders them by time. Further, the request history set is divided into a series of parts $\{data, label\}$ in a maximum interval $2z$ which contains the context records as the data and predicted record as the label. Therefore, the cloud will use the CBOW model to complete the embedding and get the representation vector of each chunk. Then, we use the learned vectors and K-means to divide the chunks into K_c categories. According to current network status $\boldsymbol{U}_j$, the cloud calculates the coded cache policy η at set intervals. The cloud needs to train the corresponding DBN for each category. After obtaining the optimal η and DBNs, the cloud server broadcasts them to all nodes through the access point (e.g. base station), and then the mobile nodes perform the deployment of coded caching based on η and DBNs. The above process is implemented by the pseudo code of **Algorithm 1**.

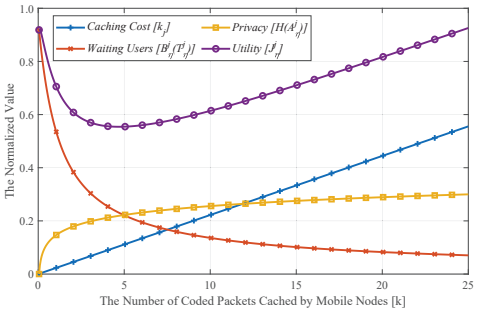


Fig. 6. The normalized value varies with the number of coded packets k_j^i

VI. Performance Evaluation

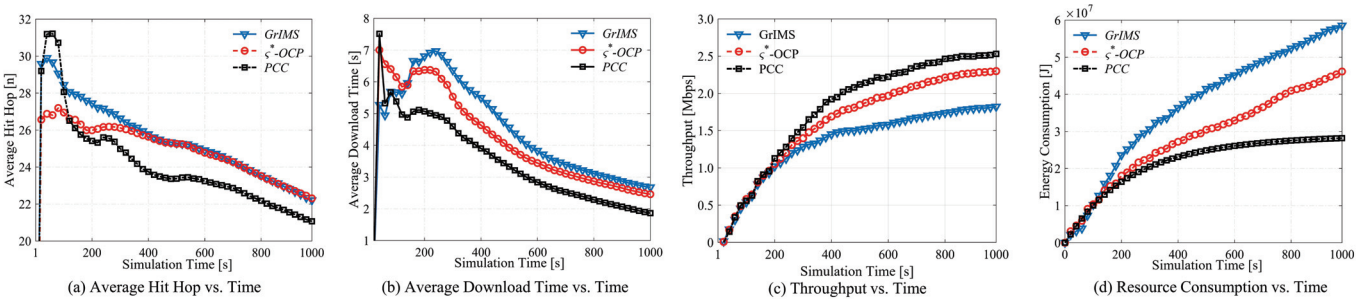
A. Experimental Settings

The parameters setting of RWP model are the same with *Section 3.B*. We use network simulator 3[1] which is a discrete-event network simulator for Internet systems to drive our simulation. We deploy a 3000*3000 m^2 quadrate simulation scenario with 200 mobile nodes and set the simulation time to 1000s. Besides, we adopt 1000 different contents and each of the contents contains 10 chunks. The playback length of each video chunk is 2 seconds. The user's request for content follows the Zipf distribution with parameter ρ = 0.8. The caching size of each node is set to 1×10^5 MTUs, where each MTU equals to 1500B. In such cases, we set that the mobile node can store 300 content chunks at most. We employ the K-means to cluster the content into 100 categories, and a central server broadcast the categories and coding strategy η to all mobile nodes. We assign random initialization coded caching to each mobile node. We also deploy two network providers that contain all of the coded packets. The simulation is deployed on a full-fledged laptop equipped with AMD R9 5900HX, 32G, RTX 3080.

B. Network Performance

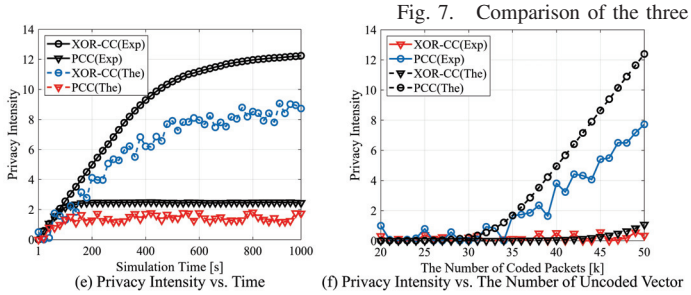
We evaluated the performance of our solution in terms of average hit hop, average downloading time, throughput, and energy consumption. We compared the performance with two state-of-the-art solutions: ς^*-OCP [16], and GrIMS [15] in the wireless mobile network system. Where ς^*-OCP is a strategy that uses a centralized server to adjust the deployment of network cache on mobile nodes. By setting a "cache time" for each piece of content, the strategy optimizes network cache deployment, ensuring that cache content that exceeds the set time is replaced. GrIMS is a centralized network cache deployment strategy that takes into account energy consumption. This strategy, through a cloud controller, adjusts

[1]https://www.nsnam.org/



(a) Average Hit Hop vs. Time (b) Average Download Time vs. Time (c) Throughput vs. Time (d) Resource Consumption vs. Time

Fig. 7. Comparison of the three algorithms on AHH, ADT, TP, RC



(e) Privacy Intensity vs. Time (f) Privacy Intensity vs. The Number of Uncoded Vector

Fig. 8. Comparison of the PCC and XOR algorithms

network cache based on the cache energy consumption and resource utilization rate of each mobile node, thus optimizing deployment. It prioritizes the replacement of cache content that has high energy consumption and low utilization. The comparison metrics are shown as follows:

Average hit hop (**AHH**): The average hop of finding all packets for decoding. **Average downloading time** (**ADT**): The average time interval between sending out requests and obtaining enough coding packets. **Throughput** (**TP**): The amount of data transmitted successfully from one node to another in a given time period. **Resource Consumption** (**RC**): The total resource consumption of all forwarding and caching operations during the simulation.

According to Fig7(a), we can see that *PCC* works best on **AHH**, and the **AHH** decreases by about 2 hops. The superiority of *PCC* is because the coded content caching reduces the likelihood of redundant replicas. Assuming five distinct content chunks with each node caching one, the likelihood of two nodes caching identical content is 1/5 in uncoded caching, whereas in coded caching with approximately 2^5 linearly independent coded contents, it drops to $1/2^5$.

Fig7(b) shows *PCC* is the lowest **ADL** of three caching solutions varying with the simulation time. All curves begin with a brief period of shock and then slowly decrease over time. This is because the node movement causes the system to be unstable at the beginning. As time pass by, the cache space is gradually optimized for full utilization and cache configuration, which reduces the **ADT**.

As shown in Fig7(c), the **TP** of three solutions first go through a rapid ascent and then stabilize. we can see *PCC* is superior to the other two algorithms in the whole time domain. Our method can greatly improve the throughput of the network, and its performance is improved by 10% and 38% respectively compared with the other two methods.

In addition, Fig7(d) shows *PCC* outperforms than ς^*-OCP and GrIMS at 1000s by about 24% and 47%. Because the decrease in **AHH** reduces the resource consumed by forwarding, the cache space can be more effectively utilized by caching encoded content.

C. Privacy Performance

For privacy protection, we test the variation of privacy intensity in two dimensions which are simulation time and the number of coded packets t. We test both theoretical and experimental results of XOR-CC [12] and our algorithm. XOR-CC designs a network coding-based coded caching scheme for distributed storage, using XOR operations to code all network content. However, lacking pre-encoding content classification, its service performance degrades significantly with a large amount of content.

Fig8(e) shows experimental and theoretical privacy intensity experience rapid growth, and then tend to stabilize around 800 seconds. While our algorithm deviates more from theoretical results in experimental cases, it still demonstrates an impressive improvement in privacy intensity, achieving nearly a 400% enhancement, which is almost an order of magnitude increase.

As per Fig8(f), our method's privacy protection rapidly increases when coded packets reach about 30. Despite the large deviation between theoretical and experimental results, it outperforms XOR-CC across all time frames, enhancing data privacy intensity by roughly 400% compared to XOR-CC.

VII. Conclusion

This paper proposed a coded caching scheme for joint optimization of coded caching deployment and data privacy. We propose an effective adversary model for the current XOR-based secure coded caching and present a privacy model to quantify its threat. We based on the *fluid-based architecture* proposed a *learning-based content coding architecture* which integrates representational learning and DBN. We design a privacy-oriented coded caching algorithm that jointly optimizes performance and privacy, ensuring data privacy protection and caching optimization without relying on trusted nodes or secure channels. Our algorithm's excellence is confirmed by numerous simulation results. With our model's ability to depict data propagation under diverse transmission strategies, our future work plans to develop efficient ICN forwarding strategies to reduce transmission delays and save energy costs associated with transmission and caching.

References

[1] Cisco, "Cisco annual internet report (2018–2023) white paper," Cisco Annual Internet Report, Tech. Rep. white-paper-c11-741490, 2020.

[2] L. Zhang *et al.*, "Named data networking," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 66–73, Jul. 2014, ISSN: 0146-4833.

[3] G. Deng *et al.*, "Distributed probabilistic caching strategy in vanets through named data networking," *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 314–319, 2016.

[4] G. Acs *et al.*, "Privacy-aware caching in information-centric networking," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 2, pp. 313–328, 2019.

[5] X. Zhang *et al.*, "Data-driven caching with users' content preference privacy in information-centric networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 9, pp. 5744–5753, 2021.

[6] Y. Fadlallah *et al.*, "Coding for caching in 5g networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 106–113, 2017.

[7] M. K. Kiskani and H. Sadjadpour, "Application of index coding in information-centric networks," in *2015 International Conference on Computing, Networking and Communications (ICNC)*, 2015, pp. 977–983.

[8] J. Saltarin *et al.*, "Netcodccn: A network coding approach for content-centric networks," *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, 2016.

[9] M. K. Kiskani and H. R. Sadjadpour, "Throughput analysis of decentralized coded content caching in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 663–672, 2017.

[10] K. Wang *et al.*, "Strategic antieavesdropping game for physical layer security in wireless cooperative networks," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 9448–9457, 2017.

[11] N. Zhao *et al.*, "Collusive eavesdropping in interference alignment based wireless networks," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 5549–5561, 2017.

[12] M. K. Kiskani and H. R. Sadjadpour, "Secure coded caching in wireless ad hoc networks," *2017 International Conference on Computing, Networking and Communications (ICNC)*, pp. 387–391, 2017.

[13] M. K. Kiskani and H. R. Sadjadpour, "Secure and private information retrieval (sapir) in cloud storage systems," *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 12 302–12 312, 2018.

[14] L. Zhong *et al.*, "A multi-user cost-efficient crowd-assisted vr content delivery solution in 5g-and-beyond heterogeneous networks," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2022.

[15] C. Xu *et al.*, "Grims: Green information-centric multimedia streaming framework in vehicular ad hoc networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 483–498, 2018.

[16] C. Xu *et al.*, "Optimal information centric caching in 5g device-to-device communications," *IEEE Transactions on Mobile Computing*, vol. 17, pp. 2114–2126, 2018.

[17] K. Wan *et al.*, "Combination networks with end-user-caches: Novel achievable and converse bounds under uncoded cache placement," *IEEE Transactions on Information Theory*, vol. 68, no. 2, pp. 806–827, 2022.

[18] K. Liu *et al.*, "Optimal caching for low latency in distributed coded storage systems," *IEEE/ACM Transactions on Networking*, vol. 30, no. 3, pp. 1132–1145, 2022.

[19] Y. Jiang *et al.*, "Joint mds codes and weighted graph-based coded caching in fog radio access networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 6789–6802, 2022.

[20] M. Rahim *et al.*, "An intelligent content caching protocol for connected vehicles," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 4, e4231, 2021.

[21] J. Tan *et al.*, "A high-accurate content popularity prediction computational modeling for mobile edge computing using matrix completion technology," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 6, e3871, 2021.

[22] D. Dharminder *et al.*, "Post quantum secure conditional privacy preserving authentication for edge based vehicular communication," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 11, e4346, 2021.

[23] X. Chen *et al.*, "Optimal coded caching in 5g information-centric device-to-device communications," *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, 2018.

[24] M. Meddeb *et al.*, "Cache coherence in machine-to-machine information centric networks," *2015 IEEE 40th Conference on Local Computer Networks (LCN)*, pp. 430–433, 2015.

[25] B. Liu *et al.*, "Malware propagations in wireless ad hoc networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, pp. 1016–1026, 2018.

MKTformer: Fine-grained Meter Classification Based on Multi-modal Knowledge Transfer

Zhaoye Zheng[1], Ke Zhang[1,2*], Chaojun Shi[1,2], Fei Zheng[3]
[1]Department of Electronic and Communication Engineering, North China Electric Power University, Baoding, China
[2]Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding, China
[3]Beijing China-Power Information Technology Co., Ltd, Beijing, China
E-mail: zyezheng@163.com, zhangkeit@ncepu.edu.cn, scj@ncepu.edu.cn, zhengfei3@sgitg.sgcc.com.cn

***Abstract* —Meter classification in converter station lays the foundation for the subsequent detection-related tasks, however, training excellent meter classification models takes a large quantity of data and computing resources. Aiming to solve this problem, a fine-grained meter classification method based on multi-modal knowledge transfer is proposed. Firstly, the multi-modal knowledge of the Contrastive Language-Image Pre-training (CLIP) model is transferred to provide a more general visual representation for fine-grained features extraction. Secondly, a Task Space Mapping Unit (TSMU) is designed to improve the transfer ability of the multi-modal knowledge. Finally, a new transfer learning strategy is proposed on this basis to achieve a better transfer performance. The experimental results show that our method can achieve higher accuracy than its counterpart in significantly less train time under both fully supervised and few-shot settings, which verifies the its superiority in capturing fine-grained features and reducing training cost.**

I. INTRODUCTION

With the construction and development of power system, the importance of intelligent detection and analysis of power transmission and transformation equipment is increasing [1]. The converter station is used to complete the conversion of alternating current and direct current, which plays an important role in the high voltage direct current system. At present, with the popularization of automatic inspection technology such as panoramic monitoring system and inspection robot, converter station has gradually shifted to unattended mode [2]. However, because of the significance of converter station in the power grid, the inspection requirements for converter station are becoming much higher [3].

There are various types of meters in the converter station, such as oil level gauge, lightning arrester and SF6 meter. These meters are the focus of the converter station inspection which is the foundation of subsequent meter recognition and defect detection [4]. However, training a fine-grained meter recognition model is still a challenging problem. On the one hand, the similar appearance of different meters in converter station poses a challenge for their recognition which requires the model to have a strong ability to capture the fine-grained characteristics about meters. On the other head, influenced by the traditional pre-training paradigm, most existing mainstream fine-grained classification models are pretrained on ImageNet datasets and takes a large quantity of data and computing resources to transfer to downstream tasks of other domains.

In recent years, multi-modal pre-training technology has developed rapidly which effectively improves the performance of the model and reduces the training cost in downstream tasks [5]. Inspired by the application of multi-modal pre-training representation and the thought of deep transfer learning, this paper proposes a fine-grained meter classification method based on Multi-modal Knowledge Transfer with transformer architecture (MKTformer) to leverage the more universal multi-modal representation for fine-grained feature extraction of meters in converter station. The main contribution of this paper can be summarized in three folds:

1. We propose MKTformer, a fine-grained classification method which transfer the underlying general multi-modal knowledge of CLIP for fine-grained feature extraction and interaction in meter classification.
2. We design a TSMU module based on the CLIP representation which provides a latent space between source domain and target domain to improve transfer

*Corresponding author: Ke Zhang (e-mail: zhangkeit@ncepu.edu.cn).
Project supported by the National Natural Science Foundation of China (62076093, 62206095) and the Fundamental Research Funds for the Central Universities (2020MS099, 2020YJ006, 2022MS078).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

efficiency of multi-modal knowledge.

3. We propose a new transfer learning strategy which further boosts the performance of multi-modal knowledge transfer in fine-grained meter classification.

II. Related Works

A. Fine-grained Visual Classification

The purpose of Fine-Grained Visual Classification (FGVC) is to distinguish subclasses from the coarse-grained category. Early works mainly adopt sophisticated Convolutional Neural Network (CNN) architecture for fine-grained feature extraction. Lin et al. [6] used two CNNs as feature extractors and provided a more complex feature interaction by introducing a bilinear feature fusion. Krause et al. [7] achieved a promising result with a large amount of noisy data for training the model. In recent years, with the popularity of transformer architecture in Computer Vision (CV), the self-attention mechanism has shown its superiority in fine-grained feature extraction and interaction. In [8], He et al. proposed part selection module and introduced contrastive feature learning to enhance the fine-grained feature extraction ability of transformer. Chen et al. [9] proposed a deformable patch module to address the corrupted semantics of objects and improve the effect of fine-grained recognition. The latest study by Wang et al. [10] of fine-grained meter classification adopted the transformer architecture and the above schemes of deformable patch as well as contrastive feature learning in fine-grained meter classification.

To sum up, these models rely on more complex structures or need more training data for fine-grained feature extraction which make the training process expensive. Meanwhile, [10] only improve the meter classification model on the level of general architecture while have not fully considered the characteristics of meter recognition itself. Based on the fact that distinctions of meters lie in the scale marks and text on the meter's dial, we transfer the multi-modal knowledge of CLIP which has shown potential in Optical Character Recognition (OCR) to boost the performance and save the cost of training fine-grained meter classification model.

B. Transfer Learning with Multi-modal Knowledge

In order to reduce the training cost and the resource wastes caused by training models repeatedly in downstream tasks, researchers began to use multi-modal data such as natural language for pre-training general representation, which is called Vision-Language Pre-training (VLP) [5]. In recent years, VLP has begun to influence the traditional pre-training paradigm of CV models. Sariyildiz et al. [11] introduced a proxy task in which visual representation is learned as a by-product and can transfer to different CV tasks. A novel VLP method VirTex was proposed by Desai et al. [12], they used image captioning for visual representation learning and transferred the pre-trained image encoder to downstream tasks. However, due to the limitation of pre-training scale, the performance of the above models in downstream tasks still lags behind the most advanced CV models. In order to further improve the performance of VLP models, CLIP [13] was pre-trained on 400 million image-text pairs which enables the task-agnostic encoders of CLIP to learn universal knowledge representation and powerful ability for transferring. A series of subsequent works leverage the task-agnostic encoders for different downstream tasks. [14] and [15] studied the effect of introducing CLIP features into the task of image captioning and show the superiority of it. Conde et al. [16] fine-tuned the CLIP model and used it to perform fine-grained art classification.

The most related works to this paper are [14] and [16]. Different from [14], this paper studies the transfer learning for fine-grained classification instead of image captioning. Meanwhile, this paper focus on leveraging and enhancing the transfer performance of multi-modal knowledge in the domain of meter recognition which makes it distinguish from [16].

III. Method

we elaborate the MKTformer model in this section and the overview of our method is shown in Fig. 1.

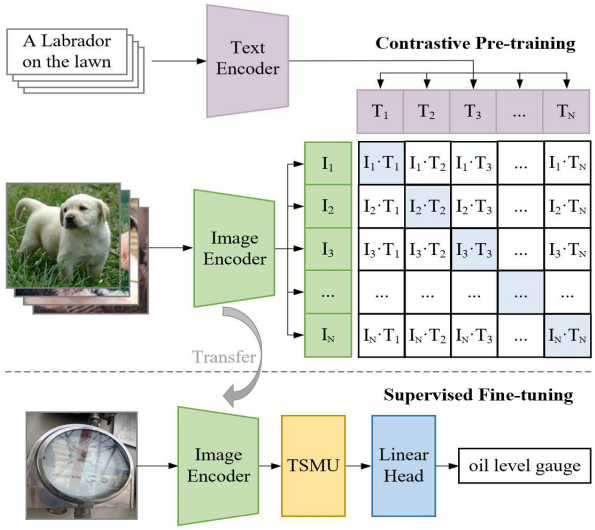


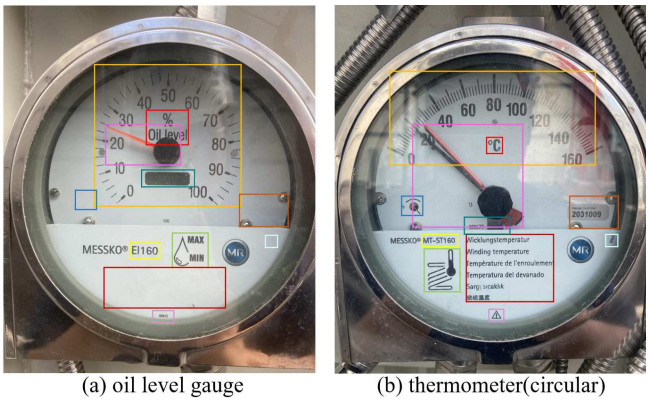
Fig. 1 Overview of the MKTformer.

A. Knowledge Representation of CLIP

CLIP jointly learns the image and text encoder using a large-scale dataset containing 400 million image-text pairs by pairing the correct image-text pairs [13]. Thanks to the large scale pretraining, the model of CLIP has obtained the abundant multi-modal knowledge about various visual and linguistic

conceptions, which enables both the image and text encoder can adapt to a wide range of downstream tasks across different modalities. In this paper, we focus on transferring the image encoder to leverage the multi-modal knowledge.

The recent work has shown the image encoder of CLIP can effectively save the training cost including training data and computing power [14]. In addition, the distinction of meters usually lies in the scale marks and text on the dial as shown in Fig. 2, while CLIP has achieved promising results in OCR which also benefits the extraction of text-related fine-grained features [13]. These factors inspire us to transfer multi-modal knowledge in the image encoder of CLIP for better extracting the fine-grained features of meters in converter station.



(a) oil level gauge (b) thermometer(circular)

Fig. 2 the distinction of different meters is on the dials.

B. *Multi-modal Knowledge Transfer*

Since the multi-head attention has brought advantage to transformer-based models in fine-grained classification in recent years, we use the ViT-B/16 image encoder of CLIP as the backbone of the MKTformer model to leverage the underlying multi-modal knowledge for fine-grained feature extraction. The structure of the backbone is shown in Fig. 3.

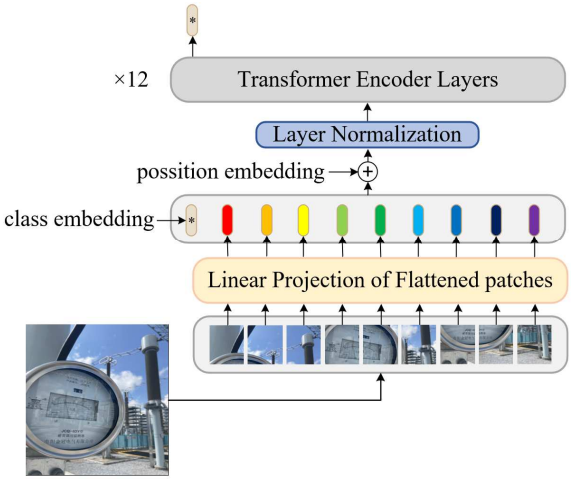


Fig. 3 Structure of the ViT image encoder in CLIP.

Different from the vanilla ViT in [17], the ViT encoder in CLIP adds a layer normalization after position encoding and uses a different initialization scheme. The ViT encoder firstly divides the input image into several image patches and embeds these patches. After that, class embedding and position embeddings are added. Then, these embeddings are passed to a series of transformer encoder layers composed of multi-head self-attention and feed-forward network for feature extraction and interaction, the calculating process of each transformer encoder layer can be written as:

$$z_l' = \mathrm{LN}\left(\mathrm{Multihead}\left(z_l\right)\right) + z_l \tag{1}$$

$$z_{l+1} = \mathrm{LN}\left(\mathrm{FFN}\left(z_l'\right)\right) + z_l' \tag{2}$$

Where z_l is the input of lth transformer encoder layer; LN(·) is layer normalization; FFN(·) is a feed-forward network; multihead(·) is multi-head self-attention which is calculated as:

$$\mathrm{Multihead}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \mathrm{Concat}\left(\mathbf{H}_1,...,\mathbf{H}_\mathbf{h}\right)\mathbf{W}^\mathbf{O} \tag{3}$$

$$\mathbf{H}_\mathbf{i} = \mathrm{Attention}\left(\mathbf{Q}\mathbf{W}_\mathbf{i}^\mathbf{Q},\mathbf{K}\mathbf{W}_\mathbf{i}^\mathbf{K},\mathbf{V}\mathbf{W}_\mathbf{i}^\mathbf{V}\right) \tag{4}$$

Where $\mathbf{W}^\mathbf{O}$ is the linear transformation matrix, $\mathbf{H_i}$ is the result of self-attention in the ith head; $\mathbf{W}_\mathbf{i}^\mathbf{Q}$, $\mathbf{W}_\mathbf{i}^\mathbf{K}$ and $\mathbf{W}_\mathbf{i}^\mathbf{V}$ are the projection matrices of the ith head; Attention(·) is scaled dot-product attention which can be calculated as:

$$\mathrm{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{5}$$

Where $\mathbf{Q}$,$\mathbf{K}$ and $\mathbf{V}$ represent the query, key and value respectively; d is the dimension of $\mathbf{Q}$ and $\mathbf{K}$.

In this way, each image is encoded with multi-modal knowledge. Finally, the class embedding output by the last transformer encoder layer containing global information is used as the feature vector to represent the image and pass to the TSMU module for classification, which transfers the multi-modal knowledge into the task of meter classification.

C. *Task Space Mapping Unit*

For fine-tuning the image encoder of CLIP, the source domain of the pre-training task and the target domain of the downstream task correspond to different feature spaces, which can be unfavorable for the transfer performance when we used it directly for meter classification [14]. Therefore, in order to make full use of the multimodal knowledge in CLIP, inspired by recent research achievements in transfer learning [18], we design a TSMU module to smooth the mapping process and improve the performance of multi-modal knowledge transfer as shown in Fig. 4.

The TSMU module learns to map the output of the task-agnostic image encoder from the feature space of the source domain into that of the target domain. Firstly, in order to realize

the basic mapping, TSMU provides a latent vector space between the feature space of source domain and target domain by using a Fully Connected layer (FC layer). Secondly, we introduce Dropout regularization to alleviate the overfitting problem and further improve the classification accuracy. Thirdly, the activation function is introduced to increase the nonlinearity of the mapping, which can be written as:

$$\mathrm{QuickGELU}(x)=\frac{x}{1+e^{-1.702x}} \tag{6}$$

Where x is a feature vector representing the input image.

Meanwhile, the output of the network is adjusted to a suitable range by using layer normalization in order to stabilize the training process of the network. Finally, the effect of TSMU is further improved by selecting the appropriate length of latent feature vector and the number of FC layers.

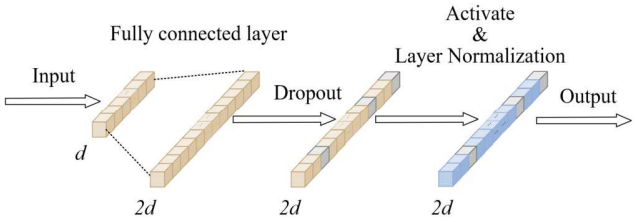


Fig. 4 Structure of the TSMU module.

D. Transfer Learning Strategy

In order to reduce the training cost and improve the transfer efficiency of the model, we design a new transfer learning strategy for fine-tuning based on multi-modal knowledge and TSMU module.

Firstly, we use the ViT-type CLIP image encoder to construct the basic model. The image encoder can encode the images with multi-modal knowledge with its multi-head self-attention mechanism. In addition, a task-specific head is added for meter classification.

Secondly, we insert the TSMU module between the backbone and the task-specific head. TSMU module takes the output of the image encoder and maps it into the latent vector space to raise the transfer efficiency.

Finally, we optimize the parameters in groups. The parameters of CLIP image encoder with a higher level of optimization is optimized with a smaller learning rate, and the random initialized parameters of TSMU and classification head are optimized with a larger learning rate.

IV. Experiments and Analysis

In this section, we evaluate the effectiveness of MKTformer model and its key components on the Converter Station Meter Classification Dataset (CSMC Dataset).

A. Dataset and Settings

CSMC Dataset: The images of the converter station meters in this dataset were collected manually and included different light and weather condition such as full light during daytime, insufficient light at dusk, illumination of the station at night, and others. The dataset consists of 4550 images including five types of meters: thermometer (circular), thermometer (square), lighting arrester, oil level gauge and SF6 meter. The ratio of training set and test set is 9:1, and the distribution of samples of training set and test set is shown in Fig. 5.

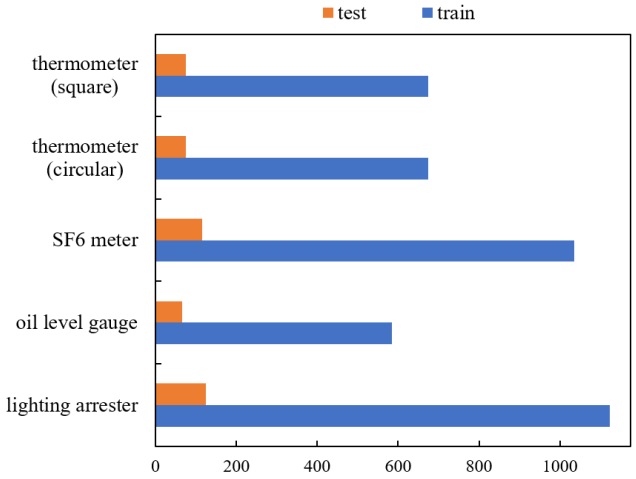


Fig. 5 The distribution of samples in CSMC Dataset.

Experiment settings: We used the PyTorch 1.10.0 framework for experiment and used a single NVIDIA GeForce RTX 3060 Laptop GPU to train and test our model. We trained our model with a batch size of 8 for 30 epochs and Adam optimizer with initial learning rate 10^{-5} for backbone's parameters and 10^{-3} for parameters of TSMU and head.

Evaluation metrics: In this work, we explore better trade-off between accuracy and training cost for fine-grained classification, so we adopt top-1 accuracy (Acc) and training time (Time) as our primary metrics. Meanwhile, we also provide parameters (Params) and floating point operations (FLOPs) to measure the complexity of different models.

B. Comparison Experiment

We train different models under the fully supervised setting to verify the effectiveness of MKTformer, in which models are trained with full training data. The comparison result is shown in Table 1. The experimental result shows our model has a similar complexity to ViT while the accuracy of MKTformer can reach 99.3%, which is significantly higher than other models. In addition, the MKTformer only need significantly less training time to achieve a higher accuracy. The result verifies that our method can effectively improve the

classification accuracy while significantly reduce the computing resources consumed during the training process.

TABLE I PERFORMANCE COMPARISON OF FULLY SUPERVISED SETTING

Methods	*Param*	*FLOPs*	*CSMC Dataset*		*BDC Dataset*	
	(M)	*(G)*	*Acc(%)*	*Time(h)*	*Acc(%)*	*Time(h)*
Szegedy et al. [19]	25.1	17.7	89.2	2.3	78.1	1.8
Howard et al. [20]	4.2	0.2	95.9	1.2	84.3	1.0
Dosovitskiy et al. [17]	88.1	17.6	97.4	1.5	88.9	1.3
He et al. [8]	80.5	14.2	97.8	1.6	89.5	1.2
Chen et al. [9]	46.1	6.9	98.1	1.8	88.7	1.3
Wang et al. [10]	72.3	12.5	98.0	1.6	89.8	1.4
Our method	87.4	17.6	99.3	0.4	90.6	0.5

Then in order to further test the potential and applicability of our method in the domain of electric power vision [1], we evaluate our method on the Bolt Defect Classification Dataset (BDC Dataset) which represents another type of fine-grained classification problem in the power system, the result is also shown in Table 1. The result has shown that the MKTformer can also significantly shorten the training time while achieve higher accuracy, which verifies its effectiveness in the task of bolt defect classification where the pictures have lower quality and the categories are more difficult to recognize.

In order to study the data requirement of training MKTformer, we fine-tune the MKTformer and vanilla ViT under the few-shot setting for comparison, in which models are trained with limited data, the result is shown in Table 2. The accuracy of MKTformer under three settings is significantly higher than vanilla ViT pre-trained on ImageNet. Specially, our model can achieve an accuracy of 97.1% with only 0.5% of the training data, which is 6.4% higher than vanilla ViT. This result verifies MKTformer can effectively reduce data requirements for training fine-grained meter classification model.

TABLE II PERFORMANCE COMPARISON UNDER LIMITED TRAINING DATA

Methods	*Data Proportion(%)*	*Acc(%)*	*Time(h)*
Dosovitskiy et al.	0.1	75.8	0.8
	0.2	82.8	0.6
	0.5	90.7	0.6
Our method	0.1	80.8	0.3
	0.2	87.5	0.2
	0.5	97.1	0.1

C. Ablation Experiment

Before the ablation experiment, we first evaluate the design of the TSMU module. The performance of different TSMU settings is shown in Table 3. Firstly, it can be seen that different settings have little influence on the complexity of the model. Secondly, the settings of Dropout, layer normalization, and activation function can effectively improve the accuracy of MKTformer in downstream tasks. Thirdly, the length of latent vector has different degrees of influence on the effect of TSMU, which reaches the best when the length of the feature vector is twice the length of the input vector d. Finally, when there are more FC layers, the model is prone to overfitting, so the TSMU in this paper contains only one FC layer.

TABLE III PERFORMANCE COMPARISON OF DIFFERENT TSMU SETTINGS

TSMU settings	*Param(M)*	*FLOPs(G)*	*Acc(%)*	*Time(h)*
Default	87.4	17.6	99.3	0.4
No dropout	87.4	17.6	99.1	0.6
No layer normalization	87.4	17.6	98.2	0.6
No activation function	87.4	17.6	99.0	0.6
Length of latent vector: 0.5d	86.5	17.6	98.6	0.8
Length of latent vector: d	86.8	17.6	98.9	0.5
Length of latent vector: 2d	87.4	17.6	99.3	0.4
Length of latent vector: 4d	88.6	17.6	99.2	0.5
Number of FC layer: 1	87.4	17.6	99.3	0.4
Number of FC layer: 2	89.7	17.6	97.8	0.8

Then in order to study the effect of each component of our method, we implement ablation experiment, the result is shown in Table 4. Firstly, after introducing CLIP features containing multi-modal knowledge, the training time is greatly shortened and classification accuracy is raised. Secondly, adding TSMU module on the basis of multi-modal knowledge can further shorten the training time and improve the accuracy. Finally, optimizing the parameters in groups can further improve the accuracy of meter classification, which verifies the effectiveness of our method and indicates that MKTformer achieves a better precision-resource consumption balance.

TABLE IV PERFORMANCE COMPARISON OF MKTFORMER

Multi-modal knowledge	*TSMU*	*Transfer learning strategy*	*Acc(%)*	*Time(h)*
-	-	-	81.7	6.3
√	-	-	98.7	0.8
√	√	-	99.0	0.4
√	√	√	**99.3**	**0.4**

D. Qualitative Experiment

In order to show the performance of MKTformer intuitively, we use Grad-CAM [21] technology to visualize the heat map of different models, as shown in Fig. 6. The first column shows the original images, the second column shows the heat map generated by vanilla ViT and the last column shows the heat map generated by MKYformer. The result shows that

MKTformer can capture more valid fine-grained features of meters especially about text or scale marks on the dial, which verifies our model has the ability to learn key fine-grained features of meters.

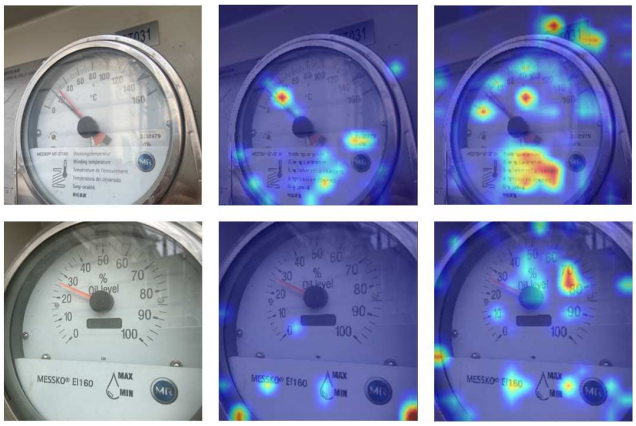


Fig. 6 The distribution of samples in CSMC Dataset.

V. CONCLUSION

This paper proposes a fine-grained meter classification method based on multi-modal knowledge transfer which is called MKTformer. The core idea of MKTformer is leveraging the more universal underlying knowledge learned in multi-modal pre-training to enhance the ability of fine-grained features extraction. We also design an auxiliary TSMU module and a new transfer learning strategy to further boost the transfer performance. The MKTformer has strong ability to capture fine-grained features of meters and can effectively reduce the training cost of fine-grained meter classification model. We will further explore meter recognition and their defect detection with multi-modal knowledge in the future.

REFERENCES

[1] Z. Zhao, W. Zhang, Y. Zhai, et al. “Concept, re-search status and prospect of electric power vision technology,” Electric Power Science and Engineering, vol. 36, no. 1, pp. 1-8, 2020.

[2] M. Xie, Q. Shao, W. Wang, et al. “Research on su-per-resolution reconstruction method of panorama monitoring im-ages of extra-high voltage converter station protection system,” Guangdong Electric Power, vol. 35, no. 5, pp. 101-109, 2022.

[3] X. Quan, H. Yang, and Y. Qiao, “Design of a vis-ualized management system for full coverage inspection of robots in converter stations,” Automation & Instrumentation, vol. 265, no. 11, pp. 172-175+187, 2021.

[4] H. Yang, R. Yao, and N. He, “Research and practice of converter station safety supervision based on intelligent inspection robot,” Information Technology, vol. 371, no. 10, pp. 70-75+84, 2022.

[5] H. Zhang, T. Wang, M. Li, et al. “Comprehensive review of visual-language-oriented multimodal pre-training methods,” Journal of Image and Graphics, vol.27, no. 9, pp. 2652-2682, 2022.

[6] T. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for fine-grained visual recognition,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 1449-1457, 2015.

[7] J. Krause, B. Sapp, A. Howard, et al. “The unreasonable effectiveness of noisy data for fine-grained recognition,” In Proceedings of the European Conference on Computer Vision, pp. 301-320, 2016.

[8] J. He, J. Chen, S. Liu, et al. “Transfg: A transformer architecture for fine-grained recognition,” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 852-860, 2022.

[9] Z. Chen, Y. Zhu, C. Zhao, et al. “Dpt: Deformable patch-based transformer for visual recognition,” In Proceedings of the 29th ACM International Conference on Multimedia, pp. 2899-2907, 2021.

[10] N. Wang, T. Yang, J. Shao, et al. “Fine-grained classification of substation meters based on vision transformer,” In Proceedings of the 3rd International Conference on Power Data Science, pp. 52-56, 2021.

[11] M. Sariyildiz, J. Perez, and D. Larlus, “Learning visual repre-sentations with caption annotations,” In Proceedings of the European Conference on Computer Vision, pp. 153-170, 2020.

[12] K. Desai and J. Johnson, “Virtex: learning visual representations from textual annotations,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11162-11173, 2021.

[13] A. Radford, J. Kim, C. Hallacy, et al. “Learning transferable visual models from natural language supervision,” In Proceedings of the International conference on machine learning, pp. 8748-8763, 2021.

[14] R. Mokady, A. Hertz, and A. BERMANO, “Clipcap: clip prefix for image captioning,” arXiv 2021, arXiv: 2111.09734.

[15] M. Barraco, M. Cornia, S. Cascianelli, et al. “The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4662-4670, 2022.

[16] M. Conde and K. Turgutlu, “CLIP-Art: Contrastive pre-training for fine-grained art classification” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3956-3960, 2021.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv 2020, arXiv:2010.11929.

[18] M. Sariyildiz, Y. Kalantidis, K. Alahari, et al. “Improving the generalization of supervised models,” arXiv 2022, arXiv: 2206.15369.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, et al. “Rethinking the inception architecture for computer vision,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818-2826, 2016.

[20] A. Howard, M. Sandler, G. Chu, et al. “Searching for mobilenetv3,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314-1324, 2019.

[21] R. Selvaraju, M. Cogswell, A. Das, et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626, 2017.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Feature Augmentation Reconstruction Network for Few-Shot Image Classification

Zhen Li†, Lang Wang*, Wenjuan An*, Song Qi*, Xiaoxu Li*, Xuezhi Fei†
* Lanzhou University of Technology, Lanzhou, China
† Machinery Industry Shanghai Lanya Petrochemical Equipment Testing Institute Co., Ltd.
E-mail: feixuezhi@163.com

***Abstract*—In recent years, deep learning has achieved significant success in various computer vision tasks, notably in image classification. Nonetheless, labeled data is frequently insufficient in real-world applications, presenting significant impediments to the efficacy of deep learning models. This issue is especially pronounced in few-shot image classification tasks. To alleviate the above problem, we propose to augment features in few-shot image classification to increase sample diversity. This is achieved through learning and fitting the features' distribution, generating pseudo-data based on the distribution, and ultimately augmenting the features. Specifically, the adherence of the sample to the Gaussian distribution is learned using both local and global information, and then feature data is generated based on the distribution for feature augmentation. We evaluated the proposed method on four benchmark datasets, and the experimental results show that our method achieves state-of-the-art performance compared to prevailing few-shot image classification methods in most experimental settings.**

I. Introduction

With the widespread application of deep learning technology, image classification has garnered significant attention in the field of computer vision. However, traditional image classification tasks typically necessitate a large amount of labelled data to train the model, which is infeasible for many practical applications. Particularly in some few-shot application scenarios, such as medical image diagnosis and military target recognition, there is often a scarcity of labelled data available for training. Moreover, obtaining and annotating data is also highly time-consuming and costly. Therefore, the accurate classification of images in situations with few samples has evolved into a pivotal research issue. More and more researchers are exploring how to use a small number of labelled samples to learn discriminative information and generalise it to novel classes. To this end, researchers have proposed many methods for few-shot learning, including meta-learning, transfer learning, metric learning, feature learning, feature augmentation, etc. These methods can help the model learn more feature information from a limited quantity of labelled data, thereby improving classification accuracy.

Metric-based methods are a classic approach for few-shot image classification. These methods learn a metric that effectively reduces the distance or increases the similarity between the same class, while increasing it between different classes. However, the effectiveness of the metric is hampered by insufficient feature information extraction or misaligned features between the two ends of the metric.

Due to the limited number of images and classes in few-shot datasets, the diversity of tasks during training is greatly restricted, which has a significant impact on the model's performance. One clear manifestation of this phenomenon is that the model's performance improves significantly when the experimental setting changes from 1-shot to 5-shot. To alleviate the problem of scarce samples in classification tasks, feature augmentation-based methods have received widespread attention. Traditional data augmentation produces new samples by transforming existing data through techniques like flipping, scaling, and cropping. Some newer methods of data augmentation employ generative adversarial models and encoder-decoders to train networks and generate new data. However, the generated data is not ideal due to the influence of data and augmentation methods, which affects the model's generalization ability.

The Feature Map Reconstruction Network (FRN) [1] proposed by Wertheimer et al. is one of the classic few-shot algorithms based on metric learning. By placing support features of the same class in a feature pool, the FRN uses the support features in the pool to reconstruct unseen query features. The reconstructed query features are then evaluated with the original, and their distance is used as the classification criterion. The FRN offers the benefits of a straightforward model structure, commendable performance, swift convergence, and a limited number of parameters. Consequently, we present our suggested feature augmentation network, which employs this technique to improve model performance by augmenting sample features, increasing sample diversity and complementing the existing features.

To summarise, our work makes a triple contribution:

- We propose a feature augmentation network that learns both global and local information of sample features to fit the feature distribution and achieve feature augmentation.
- We verify the effectiveness of the proposed method on three fine-grained datasets and one coarse-grained dataset. The results show that our method achieves state-of-the-art performance in most cases.
- We validate the generalisability of the proposed approach through comparative experiments of adding feature augmentation networks in different network structures, and

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

consistently improve the network performance.

II. Related Work

A. Data Augmentation-Based Methods

In deep learning models, data augmentation is a highly utilized approach that can effectively enlarge the dataset's sample classes and quantities to a certain extent. In the early stages, data augmentation methods used simple geometric transformations to create new samples from the original sample data [2], which helped to prevent the network model from overfitting during the training process. Traditional methods of data augmentation comprise image rotation, mirror flipping, scaling, cropping, deformation, and adding noise, etc [3], [4], [5]. These approaches can generate varied effects by adjusting parameters such as rotation angle, scaling ratio and cropping size.

With the ongoing advancements in few-shot learning, numerous data augmentation algorithms are continually being proposed. These can be broadly categorised into two groups: data synthesis and feature augmentation. Data synthesis techniques employ generative networks to create fresh data, which is then used to expand the sample space. One extensively studied method for data generation is the use of Generative Adversarial Networks (GANs). This approach trains the network to generate samples that more closely resemble the real data, thereby increasing the size and diversity of the existing dataset and improving the generalisation performance of the model [6], [7].

Feature augmentation methods augment the feature space of samples by utilizing a feature generator, thereby improving the generalization performance of feature extractors. Dixit et al. [8] introduced the Attribute-Guided Augmentation (AGA) model for image attribute-guided synthesis. This model projects image samples onto an attribute space and trains the encoder-decoder of the model to generate images with varying depths and poses. Schwartz et al. [9] proposed a data augmentation method based on an enhanced auto-encoder. This approach extracts variations and disparities among instances of training classes and employs these distinctions to generate new samples. These newly generated samples are then used to train a linear classifier. Chen et al. [10] proposed the TriNet model based on an encoder-decoder architecture, which maps samples to a semantic space and learns the concepts of samples within this semantic space. By introducing noise and finding nearest neighbors in the semantic space, the model performs sample augmentation, and then maps back to the visual space to obtain more augmented samples.

Although some current methods for data augmentation have been developed, they mostly concern general feature creation rather than focused global and local feature information generation. Therefore, we propose our approach to address this gap.

B. Metric-Based Methods

1) Distance/Similarity-based methods: This method assumes that there is a way to measure the distance or similarity between samples in the feature space. The classifier's parameters are learned by reducing the distance between samples belonging to the same class and increasing the distance between samples from different classes. Vinyals et al. [11] proposed a matching network model based on memory and attention mechanisms, which learns the nearest neighbor classifier in an end-to-end manner and uses cosine similarity as the final classification criterion. Li et al. [12] proposed a deep nearest neighbor neural network based on the recent scenario training mechanism. The network replaces the image-level feature metric in the last layer with image-to-class measures based on local descriptors and trains them end-to-end. BSNet [13]combines two types of similarity measures, cosine similarity measure and relation measure, to form a bipartite similarity measure module. It calculates similarity scores separately for each type of measure and then takes the average as the final metric to learn more discriminative features in fine-grained images.

2) Prototype-Based Methods : This method regards each class as a prototype (such as class center or mean value), and learns the parameters of the classifier by minimizing the distance between each sample and its corresponding class prototype. A pioneering metric-based approach is the prototype network model proposed by Snell et al. [14]. This model uses a convolutional neural network to nonlinearly map sample features to an embedding space, and uses the mean of embedding vectors of support set samples as class prototypes. It predicts the sample class by calculating the distance between the embedding vector of the query set sample and the class prototype. In the latest metric-based few-shot learning methods, COMET [15] learns multiple embedding functions, each corresponding to an image segment or concept, and constructs multiple concept prototypes accordingly to adapt to fine-grained classification. CovaMNet [16] uses the second-order feature representation of samples in each class to calculate prototype vectors. PHR [17] uses three levels of feature embeddings to represent each sample: local features, global features, and semantic features, and updates prototypes by superimposing the three features with different weights.

3) Network Metric-Based Methods: This method uses labeled samples to learn the parameters of the classifier and metric network, and improves the performance of the network through iterative learning strategies. Sung et al. [18] proposed a relational network based on deep neural networks, which includes two modules: the embedding module and the relation module. The embedding module is responsible for extracting and encoding the image features of the support set and the query set, and the relation module takes these two features as inputs and output the matching score between two images. Hilliard et al. [19] introduced a new network structure consisting of different stages trained end-to-end, including the feature stage, relation stage, and classifier. This method replaces the static measurement method and compares classes by training the network. LCCRN [20] for few-shot fine-grained image classification. This method can distinguish crucial discriminative local information with high similarity by

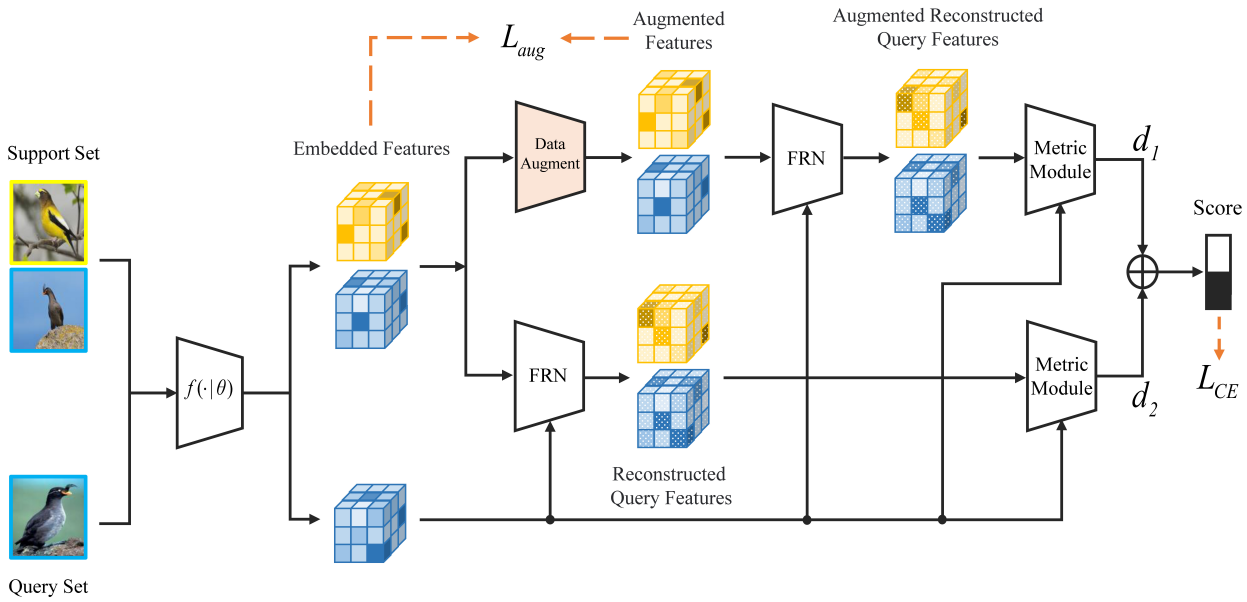


Fig. 1. Architecture diagram of the feature augmentation network

measuring cross reconstruction of different features.

III. Proposed Method

A. Problem Formulation

Few-shot image classification problem usually divides the dataset into training set D_{train}, validation set D_{val} and testing set D_{val}, which are used for the training of the model, the evaluation of the model in the training stage, and the final trained model test.

The N-way K-shot is a common setup in few-shot image classification tasks, where N denotes the number of classes in a task and K denotes the number of labeled samples in each class. For different stages, the corresponding classification tasks are constructed using the above sub-datasets, and each classification task contains a support set and a query set. The number of support set images is $N \times K$. N classes are randomly selected from the sub-dataset and then K images are randomly selected from each class. The number of query set images is $N \times M$, where M is the number of images randomly sampled from the remaining images in the category corresponding to the supporting set. Taking the metric learning-based few-shot training process as an example, during the training stage, the model learns sample features and model parameters from labeled support samples on the training set D_{train} and performs classification on unlabeled query samples, optimizing the model based on the classification loss. The learned model is evaluated using the validation set D_{val} to select the model's hyperparameters. Finally, the optimal model is saved. In the testing stage, the model's performance is tested on the testing set D_{test} to obtain the final classification accuracy.

B. Architecture Overview

As shown in Figure 1, the proposed Feature Augmentation Network consists of five main components. The first component is the Feature Embedding Module, denoted as $f(\cdot \mid \theta)$, which aims to extract deep features from the image samples to capture rich information. The second component is the Feature Reconstruction Module, denoted as FRN, which is used to generate reconstructed query features. The third component is the Feature Augmentation Module proposed in this context. In this module, the learned sample features are fitted to a Gaussian distribution and generate augmented features for subsequent training. The fourth component is the Metric Module, which measures the distance between the reconstructed features and query features as the final criterion for classification. The final component is the Network Loss, which includes the calculation of cross-entropy loss L_{CE} to measure the difference between the true labels and predicted labels, and an additional loss L_{aug} that computes the difference between the augmented features and original support features to constrain the augmented features.

C. Feature Augmentation Module

The proposed Feature Augmentation Module, as illustrated in Figure 2. The embedded features of the samples obtained from the embedding module $f(\cdot \mid \theta)$. The dimension of the features is $[K, 640, 5, 5]$, where K represents the number of samples from the same class, corresponding to the shot number in the N-way K-shot few-shot classification task. The value 640 denotes the number of channels, and [5, 5] represents the width and height of the features. Then, the input features are divided into two parts and processed by two different structures to generate features. The adoption of both fully connected networks and convolutional networks is motivated by their respective capabilities in capturing global and local

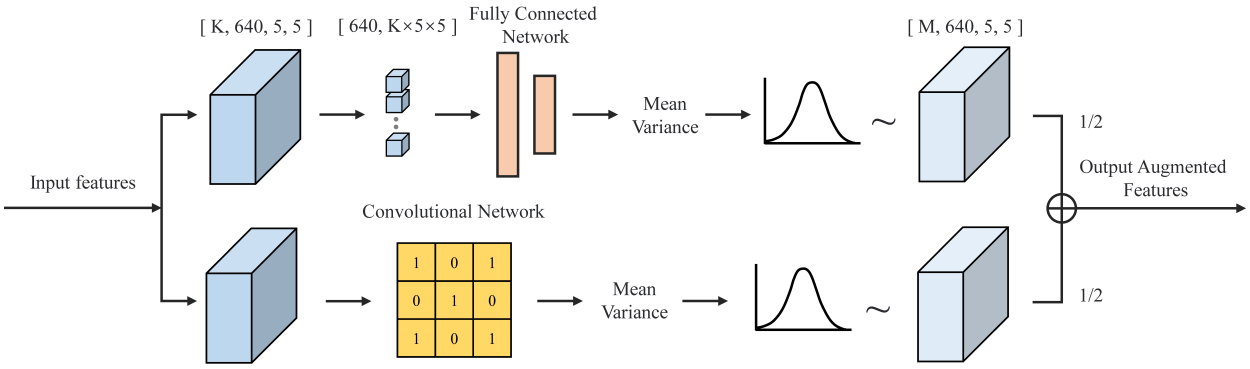


Fig. 2. Architecture diagram of the feature augmentation module

information. The fully connected network establishes global connections, capturing global information, while the convolutional network enables local connections, capturing local information. Therefore, by utilizing both network structures, we can consider both global and local information in the generated features, which better conforms to the distribution requirements.

For the upper part of the structure, the input features undergo dimension transformation, where features from different images belonging to the same class are gathered into a feature pool. This aggregation collects feature vectors across positions and images into a single feature pool, resulting in a dimension change from [K, 640, 5, 5] to [640, K×5×5]. This transformation has the benefit of alleviating the impact of K-shot on the classification task, especially when the shot is 1. The fully connected network in this part consists of two fully connected blocks, each containing two fully connected layers. ReLU activation functions are applied after each fully connected block to introduce non-linearity to the structure. The two fully connected blocks are responsible for learning the mean and variance of the feature distribution for this specific class:

$$\mu_{FC}^{i} = h_{\phi}^{1}\left(X_i\right), \tag{1}$$

$$\sigma_{FC}^{i} = h_{\phi}^{2}\left(X_i\right), \tag{2}$$

where μ_{FC}^{i} represents the mean of the features obtained by passing the transformed features of class i through the fully connected network, σ_{FC}^{i} represents the variance of the feature obtained by passing the transformed features of class i through the fully connected network, $h_{\phi}^{n}(\cdot)$ denotes the corresponding fully connected block, and X_{i} represents the transformed features of class i.

Then, we use the mean and variance to fit the Gaussian distribution to the features of this class, and finally, we randomly generate sample features based on this distribution:

$$q\left(\mathcal{A}_{FC}^{i} \mid X_i\right) = \mathcal{N}\left(\mu_{FC}^{i}, \sigma_{FC}^{i}\right), \tag{3}$$

Where $q\left(\cdot \mid \cdot\right)$ is the distribution function, $\mathcal{A}_{FC}^{i}$ represents the class distribution fitted based on the mean and variance obtained from the features of the i-th class after passing through the fully connected layers, and $\mathcal{N}\left(\cdot \mid \cdot\right)$ denotes the Gaussian distribution function.

$\mathcal{A}_{FC}^{i}$ can be generated using the following equation:

$$\mathcal{A}_{FC}^{i} = \mu_{FC}^{i} + \sigma_{FC}^{i} * \varepsilon, \varepsilon \sim \mathcal{N}(0, I), \tag{4}$$

where $\mathcal{N}(0, I)$ is a standard normal distribution.

The lower part consists of two convolutional blocks, each containing two convolutional layers followed by a batch normalization layer and a ReLU activation function. This structure is used to normalize the data and introduce non-linearity to the features. The two convolutional blocks learn the mean and variance of the distribution for the samples of the same class:

$$\mu_{CN}^{i}{=}g_{\varphi}^{1}(X_i), \tag{5}$$

$$\sigma_{CN}^{i}{=}g_{\varphi}^{2}(X_i), \tag{6}$$

where μ_{CN}^{i} represents the mean of the i-th class sample features after passing through the convolutional network, σ_{CN}^{i} represents the variance of the i-th class sample features after passing through the convolutional network, and $g_{\phi}^{n}(\cdot)$ represents the corresponding convolutional block.

Then, the feature distribution is modeled by fitting a Gaussian distribution based on the variance and mean. Finally, the sample features are randomly generated using this distribution:

$$q(\mathcal{A}_{CN}^{i}|X_i) = \mathcal{N}(\mu_{CN}^{i}, \sigma_{CN}^{i}), \tag{7}$$

$\mathcal{A}_{CN}^{i}$ represents the class distribution obtained by fitting a Gaussian distribution based on the mean and variance of the features from the i-th class after passing through the convolutional layers.

$\mathcal{A}_{CN}^{i}$ can be generated using the following equation:

$$\mathcal{A}_{CN}^{i} = \mu_{CN}^{i} + \sigma_{CN}^{i} * \varepsilon, \varepsilon \sim \mathcal{N}(0, I), \tag{8}$$

After the aforementioned two parts of the network structure, two sets of fitted features are generated, each with a feature dimension of [M, 640, 5, 5], where M represents the number of augmented features generated for the current class. At this point, the two sets of features respectively possess the global and local information of the current class features. Subsequently, the two types of features are combined through

TABLE I
RESULTS OF 5-WAY 1-SHOT AND 5-WAY 5-SHOT CLASSIFICATION WITH THE CONV-4 BACKBONE ON THE CUB, DOGS, AND FLOWERS DATA. MEAN ACCURACY AND 95% CONFIDENCE INTERVALS ARE REPORTED.

Method	CUB		Dogs		Flowers	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet [11]	60.06±0.88	74.57±0.73	46.10±0.86	59.79±0.72	71.89±0.90	85.46±0.59
ProtoNet [14]	63.64±0.23	84.23±0.15	45.12±0.21	69.16±0.16	54.23±0.23	84.97±0.16
RelationNet [21]	63.94±0.92	77.87±0.64	47.35±0.88	66.20±0.74	-	-
Baseline++ [22]	62.36±0.84	79.08±0.61	44.49±0.70	64.48±0.66	70.54±0.84	86.63±0.58
DeepEMD [23]	64.08±0.50	80.55±0.71	46.73±0.49	65.74±0.63	-	-
DSN [24]	71.57±0.92	83.51±0.60	44.33±0.81	60.04±0.68	67.71±0.92	84.58±0.70
TOAN [25]	65.34±0.75	80.43±0.60	49.30±0.77	67.16±0.49	-	-
BSNet [26]	62.84±0.95	85.39±0.56	43.42±0.89	71.90±0.68	72.79±0.91	84.93±0.64
FRN [27]	73.63±0.21	**88.50±0.13**	**58.23±0.22**	76.50±0.17	73.61±0.22	88.93±0.14
VFD [28]	68.42±0.92	82.42±0.61	57.03±0.86	73.00±0.66	-	-
Ours	**74.02±0.22**	88.37±0.13	58.19±0.22	**76.73±0.17**	**74.58±0.21**	**89.75±0.13**

summation and averaging, resulting in a set of augmented features that encompass both global and local information:

$$\mathcal{A}^i = \frac{\mathcal{A}^i_{FC} + \mathcal{A}^i_{CN}}{2}, \tag{9}$$

After the feature augmentation module, the augmented features are obtained for each support class. The augmented features of the support set and the query features are fed into the FRN reconstruction module for feature reconstruction, resulting in augmented reconstructed query features $\bar{Q}^{aug}_c$. Simultaneously, the original support features and query features are fed into the FRN module for feature reconstruction, resulting in reconstructed query features $\bar{Q}_c$. The detailed reconstruction process of FRN can be referred to in reference [1].

After obtaining the reconstructed features, the Euclidean distance is used to measure the distance between the two types of support features and the original query feature separately:

$$d_{c,1} = \left\|\bar{Q}^{aug}_c - Q\right\|^2_F, \tag{10}$$

$$d_{c,2} = \left\|\bar{Q}_c - Q\right\|^2_F, \tag{11}$$

$$d_c = d_{c,1} + d_{c,2}. \tag{12}$$

By using the aforementioned distances, the final classification probabilities can be obtained:

$$P\left(y_i = c|x_i\right) = \frac{\exp\left(-\tau d_{c,i}\right)}{\sum\limits_{c' \in C} \exp\left(-\tau d_{c',i}\right)}, \tag{13}$$

Where τ is a learnable hyperparameter that adjusts the range of the metric distances.

During the training phase, the cross-entropy (CE) loss is utilized as the loss function to optimize the model:

$$L_{CE} = -\frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{y}_i^T \log\left(p_i\right)\right), \tag{14}$$

where y_i represents a one-hot vector, p_i represents the predicted probability vector, and m represents the number of query samples.

In addition, an augmented feature constraint loss L_{aug} is introduced to enforce the similarity between the augmented features and the original features:

$$L_{aug} = \left\|\mathcal{A}^i_p - X^i_p\right\|^2_F, \tag{15}$$

where $\mathcal{A}^i_p$ represents the prototype of the augmented features generated from the support features of class i, and X^i_p represents the prototype of the support features of class i.

The final loss is obtained by combining the loss L_{CE} and the loss L_{aug}:

$$Loss = L_{CE} + kL_{aug}, \tag{16}$$

where k is used to adjust the impact of the loss L_{aug} on the overall loss.

IV. EXPERIMENTS

A. Datasets and Preprocessing

We conduct experiments on four popular benchmark datasets: CUB-200-2011 (CUB), Stanford Dogs (Dogs), Flowers, and mini-ImageNet. The CUB dataset is consistent with the dataset used in FRN [1], where images are pre-cropped to the provided bounding boxes. All datasets were divided into training, validation, and testing sets in a ratio of 2:1:1.

B. Implementation Details

All our experiments follow the standard settings for few-shot image classification and utilize the Conv-4 and ResNet-12 backbone networks.

The Conv-4 [31] network consists of four convolutional blocks, each containing a convolutional layer, a batch normalization layer, a ReLU non-linear activation layer, and a max pooling layer. The input image size is 3×84×84, and the output feature size is 64×5×5. The ResNet-12 [31] architecture

TABLE II
RESULTS OF 5-WAY 1-SHOT AND 5-WAY 5-SHOT CLASSIFICATION WITH THE RESNET-12 BACKBONE ON THE CUB, DOGS, AND FLOWERS DATA. MEAN ACCURACY AND 95% CONFIDENCE INTERVALS ARE REPORTED.

Method	CUB		Dogs		Flowers	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [14]	79.64±0.20	91.15±0.11	72.85±0.22	86.54±0.13	75.41±0.22	89.46±0.14
DeepEMD [23]	71.11±0.31	86.30±0.19	67.59±0.30	81.13±0.20	70.00±0.35	83.63±0.26
RENet [29]	79.49±0.44	91.11±0.24	71.69±0.47	85.60±0.30	79.91±0.42	92.33±0.22
MixFSL [30]	67.86±0.94	82.12±0.66	67.26±0.90	82.05±0.56	72.60±0.91	86.52±0.65
FRN [27]	83.16±0.19	92.59±0.11	**76.49±0.21**	**88.22±0.12**	81.07±0.20	92.52±0.11
VFD [28]	79.12±0.83	91.48±0.39	76.24±0.87	88.00±0.47	-	-
Ours	**83.35±0.20**	**92.77±0.12**	76.25±0.21	88.09±0.12	**81.93±0.20**	**93.27±0.11**

TABLE III
RESULTS OF 5-WAY 1-SHOT AND 5-WAY 5-SHOT CLASSIFICATION WITH THE RESNET-12 BACKBONE ON THE MINI-IMAGENET DATA. MEAN ACCURACY AND 95% CONFIDENCE INTERVALS ARE REPORTED.

Method	mini-ImageNet	
	1-shot	5-shot
ProtoNet [14]	62.39±0.21	80.53±0.14
MatchingNet [11]	65.64±0.20	78.72±0.15
Meta-Baseline [22]	63.17±0.23	79.26±0.17
DeepEMD [23]	65.91±0.82	82.41±0.56
DSN [24]	62.64±0.66	78.83±0.45
FEAT [32]	**66.78±0.20**	82.05±0.14
FRN [27]	66.45±0.19	82.83±0.13
Ours	66.59±0.19	**83.15±0.13**

TABLE IV
PERFORMANCE IMPROVEMENT COMPARISON OF FEATURE AUGMENTATION ON DIFFERENT NETWORKS WITH 5-WAY 1-SHOT AND 5-WAY 5-SHOT CLASSIFICATION ON CONV-4 BACKBONE NETWORK.

Method	CUB		Dogs	
	1-shot	5-shot	1-shot	5-shot
ProtoNet	63.64	84.23	45.12	69.16
ProtoNet + M_{aug}	**64.48**	**84.91**	**46.35**	**69.74**
RelationNet	63.94	77.87	47.35	66.20
RelationNet + M_{aug}	**64.51**	**78.49**	**47.85**	**66.54**
Baseline++	62.36	79.08	44.49	64.48
Baseline++ + M_{aug}	**63.14**	**79.74**	**44.87**	**64.91**

includes four residual blocks, each composed of a convolutional layer, a batch normalization layer, a ReLU non-linear activation layer, and a max pooling layer. The input dimension is 3×84×84, and the output feature dimension is 640×5×5. We adopt the 10-way 5-shot setting for training, 5-way 5-shot setting for validation, and 5-way 1-shot and 5-way 5-shot settings for testing. The SGD optimizer is used with an initial learning rate of 0.1 and momentum of 0.9. The models are trained for a total of 1200 epochs. During the validation phase, the model is validated every 20 epochs, and the parameters achieving the best performance are saved. In the testing phase, the final test accuracy is obtained by averaging the accuracy over 10,000 randomly selected tasks.

C. Experimental Results

To validate the effectiveness of our proposed method, we conduct experiments and compare with state-of-the-art methods under the 5-way 1-shot and 5-way 5-shot experimental settings. Table 1 presents the experimental results on the CUB, Dogs, and Flowers datasets using the Conv-4 backbone network. Table 2 shows the experimental results using the ResNet-12 backbone network. Since some methods did not evaluate the Flowers dataset in their original papers, certain cells in the table are left blank and marked with "-". Table 3 displays the experimental results on the mini-ImageNet dataset using the ResNet-12 backbone network. The highest value in each column is indicated in bold, and the second highest value is underlined.

From Table 1, it can be observed that the proposed method achieved the best or second-best performance in most cases, except for the 5-way 5-shot setting on CUB and the 5-way 1-shot setting on Dogs, where it performed slightly lower than FRN. On the CUB dataset, the proposed method outperformed the second-best FRN by 0.97% in the 1-shot setting and by 0.82% in the 5-shot setting. According to the comparative experiments in Table 2, the proposed method achieved the maximum accuracy in all cases except for the Dogs dataset, where it slightly underperformed compared to FRN. On the Flowers dataset, compared to the second-best FRN, the proposed method improved the accuracy by 0.86% in the 1-shot setting and by 0.75% in the 5-shot setting. Furthermore, in Table 3, the proposed method demonstrated excellent performance compared to state-of-the-art methods in the mini-ImageNet dataset experiments under the same settings.

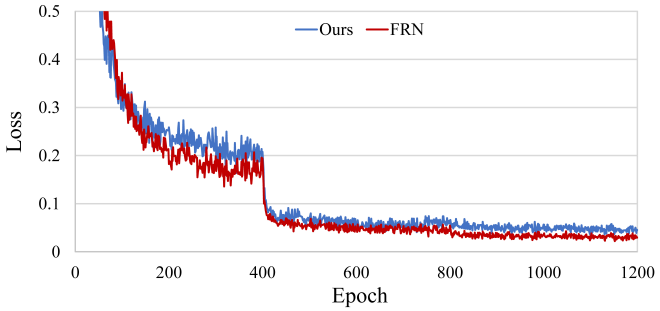


Fig. 3. The training loss of our method and FRN on the CUB dataset.

Overall, the comparative experiments with current state-of-the-art methods in few-shot image classification demonstrate that the proposed method achieved the best or second-best performance in nearly all cases, confirming its effectiveness.

As shown in Table 4, to demonstrate the effectiveness of the proposed feature augmentation module, the module M_{aug} was added to three network architectures: ProtoNet, RelationNet, and Baseline++. Comparative experiments were conducted on the Conv-4 backbone network under the 5-way 1-shot and 5-way 5-shot settings. The three network models were adjusted accordingly following the structure shown in Figure 1. The generated augmented features and the original support features were processed in separate branches, and the scores from both branches were combined as the final score. Additionally, the loss function L_{aug} was added to encourage the similarity between the prototypes of the generated augmented features and the original support features. The comparative experiments on the three datasets revealed consistent performance improvement after incorporating the feature augmentation module, further confirming the effectiveness of feature augmentation.

To demonstrate how our proposed data augmentation network enhances the model, Figure 3 presents the loss alteration between ours and the FRN during the training stage. In the figure, our approach shows a higher loss in contrast to FRN. When considered together with the experimental results, our approach becomes more difficult to train due to the data augmentation and yields a slightly higher loss than FRN. However, the generalisation of the model improves and performs better on novel class.

V. CONCLUSION

In this paper, we propose a new feature augmentation network structure that learns to fit the distribution of features from both global and local information of the image in a learnable way, and then increase the diversity of samples by randomly generating sample features. Numerous experiments demonstrate that our proposed method performs competitively or outperforms to existing few-shot classification methods on four popular benchmark image datasets.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62176110, the Key Research and Development Program of Gansu Province under grant 22YF7GA130, Young Doctoral Fund of Education Department of Gansu Province under Grant 2021QB-038, Hong-liu Distinguished Young Talents Foundation of Lanzhou University of Technology.

REFERENCES

[1] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8008–8017, 2020.

[2] Sebastien C. Wong, Adam Gatt, Victor Stamatescu, and Mark D. McDonnell. Understanding data augmentation for classification: When to warp? *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, 2016.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[4] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *ArXiv*, abs/1708.04896, 2017.

[5] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2013.

[6] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2020.

[7] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1847–1856, 2021.

[8] Mandar Dixit, Roland Kwitt, Marc Niethammer, and Nuno Vasconcelos. Aga: Attribute-guided augmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3328–3336, 2016.

[9] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Neural Information Processing Systems*, 2018.

[10] Z. Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, X. Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 28:4594–4605, 2018.

[11] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, 2016.

[12] Wenbin Li, Lei Wang, J. Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7253–7260, 2019.

[13] Xiaoxu Li, Jijie Wu, Z. Sun, Zhanyu Ma, Jie Cao, and Jing-Hao Xue. Bsnet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, 30:1318–1331, 2020.

[14] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *ArXiv*, abs/1703.05175, 2017.

[15] Kaidi Cao, Maria Brbić, and Jure Leskovec. Concept learners for few-shot learning. *arXiv: Learning*, 2020.

[16] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. Distribution consistency based covariance metric networks for few-shot learning. In *AAAI Conference on Artificial Intelligence*, 2019.

[17] Yuanen Zhou, Yanrong Guo, Shijie Hao, and Richang Hong. Hierarchical prototype refinement with progressive inter-categorical discrimination maximization for few-shot learning. *IEEE Transactions on Image Processing*, 31:3414–3429, 2022.

[18] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2017.

[19] Nathan Hilliard, L. Phillips, Scott Howland, Artëm Yankov, Court D. Corley, and Nathan Oken Hodas. Few-shot learning with metric-agnostic conditional embeddings. *ArXiv*, abs/1802.04376, 2018.

[20] Xiaoxu Li, Qi Song, Jijie Wu, Rui Zhu, Zhanyu Ma, and Jing-Hao Xue. Locally-enriched cross-reconstruction for few-shot fine-grained image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[21] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *CVPR*, pages 1199–1208, 2018.

[22] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Y. Wang, and Jia-Bin Huang. A closer look at few-shot classification. *ArXiv*, abs/1904.04232, 2019.

[23] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12200–12210, 2020.

[24] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Tafazzoli Harandi. Adaptive subspaces for few-shot learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4135–4144, 2020.

[25] Huaxi Huang, Junjie Zhang, Litao Yu, Jian Zhang, Qiang Wu, and Chang Xu. TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):853–866, 2022.

[26] Xiaoxu Li, Jijie Wu, Zhuo Sun, Zhanyu Ma, Jie Cao, and Jing-Hao Xue. BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, 30:1318–1331, 2021.

[27] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8012–8021, 2021.

[28] Jingyi Xu, Hieu M. Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. *IEEE/CVF International Conference on Computer Vision*, pages 8792–8801, 2021.

[29] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[30] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagn'e. Mixture-based feature space learning for few-shot image classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9021–9031, 2021.

[31] Itamar Arel, Derek C. Rose, and Thomas P. Karnowski. Deep machine learning - a new frontier in artificial intelligence research [research frontier]. *IEEE Computational Intelligence Magazine*, 5:13–18, 2010.

[32] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Dual Feature Reconstruction Network For Few-shot Image Classification

Xiaowei Guo*, Jijie Wu*, Kai Ren*, Qi Song*, Xiaoxu Li*
* Lanzhou University of Technology, Lanzhou, China
E-mail: gxw111ut@163.com

***Abstract*—Few-shot image classification aims to provide accurate predictions for novelty by learning from a limited number of samples. Classical few-shot image classification methods usually use data augmentation and self-supervision to compensate for the lack of training sample, and introduce migration learning and meta-learning to pre-train the model or accelerate the model optimization, which improves the classification performance of the model. However, with a small amount of labeled sample data, these methods cannot meet the requirements of the model's ability to characterize sample features, resulting in a model that is highly susceptible to overfitting problems. In this paper, we propose a Dual Feature Reconstruction Network (DFRN) for few-shot image classification. The network constructs the double feature vector by two modules, in which the first-level feature module generates an attention mask based on the image to make the feature vector characterize more of the target region, and the secondary feature module interferes with the feature vector to improve its generalization performance. In addition, the network also enhances the classification performance of the model by considering the contextual information of the support classes through an auxiliary loss function. Through extensive experiments, the network proposed in this paper achieves excellent performance on Flowers, CUB and Cars datasets and outperforms other reference fine-grained image classification methods such as FRN.**

I. Introduction

Although traditional deep learning models have surpassed human recognition levels in image classification tasks [1], [2], [3], e.g. residual network [4] have outperformed humans in recognition accuracy on ImageNet [5] datasets. But it also exposes the potential problem of image classification models relying on large-scale training data. In contrast, humans make effective use of prior knowledge to learn new things quickly. Drawing on the human approach to rapidly learn new things, few-shot image classification has received widespread attention.

Few-shot image classification aims to provide accurate predictions for new things by learning from a few labeled examples. The most frequently used of the existing few-shot image classification methods is the metric-based learning approaches. Metric learning calculates the distance between two samples by using a given distance function to obtain the similarity. Commonly used distance metric functions are Euclidean distance, cosine distance and Manhattan distance. Existing metric learning-based methods for classifying few-shot images are divided into the following three main parts.

Improving feature embedding space. Deep nearest neighbor neural network [6] uses local features of samples to construct a local feature space of classes and predict the class to which a query sample belongs by the similarity between the query sample and the class. The neural network-based task adaptive projection enhancement proposed by Yoon et al. [7] extracts sample features through an embedding module, learns a class reference vector for each class, and projects the sample features and the class reference vector using a mapping matrix to project into a task-relevant classification space. Wu et al. [8] proposed a location-aware network to collect information from different locations in a location-aware manner through a double correlation attention mechanism to obtain global information about the features.

Improving class prototypes. In the few-shot image classification task, Snell et al. [9] proposed a prototype network in which the support sample features of the same class are averaged by means of an averaging operation to obtain the class prototype of that class. Kim et al. [10] proposed a method for generating class prototypes using a variational autoencoder, which removes the distracting noise from the samples in the original dataset, retains the most valuable information of the images, and classifies the generated class prototypes using the k-Nearest Neighbor algorithm. Wu et al. [11] proposed a few-shot image classification method based on a capsule network, in which the weights of the supported samples are obtained by a classifier, the class prototypes are the weighted averages of the supported samples, and finally the similar samples are placed close to each other and the different samples are placed far away from each other using a triadic loss function.

Sung et al. [12] proposed a relational network that outlawed the traditional distance function and used a neural network to construct a relational module to measure the distance between each query sample and the support sample. Deep Earth Mover's Distance [13] split the whole image into multiple blocks and used Earth Mover's Distance to calculate the distance between the query sample and the support sample blocks. Qiao et al. [14] proposed a transductive scenario adaptive metric to construct pairwise constraints and regularization priors for each task, converting the task optimization process into a standard semi-definite programming problem, and obtaining its closed solution. For the transduced features, compare the similarity between the query sample and each support sample, and the similarity between the support sample and each query sample, which is a two-way similarity measure to complete the prediction of the query sample. Existing methods based on few-shot image classification ignore class-level knowledge. To address this problem, Chen et al. [15]

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

proposed an explicit class knowledge propagation network, which uses an instance graph to learn into a relational representation of samples through a comparison module, compresses the instance-level graph to generate a class graph, models the relationship between different classes, and describes the relationship between class through a correction module to obtain a more discriminative. Finally, the class knowledge representation is combined with the instance-level samples for query sample inference.

However, existing methods for few-shot image classification only convert features into feature vectors by means of an averaging operation on information at different locations in the image, without taking into account the fact that a simple averaging operation can make the feature vectors susceptible to the influence of the irrelevant background of the image during the transform process. In this paper, we propose a dual feature reconstruction network. The network takes into account real-world interference situations and uses attention masks to identify distinguishable regions of the target, thus learning more diverse feature information. And the risk of overfitting is further avoided by considering contextual information of the support classes through an auxiliary loss function.

II. Related Work

A. Metrics Learning

In the same space, samples with small differences are closer to each other, and samples with large differences are farther away from each other, and it is through this feature that metric learning determines the category to which the query sample belongs. Commonly used metrics in few-shot image classification include cosine distance, Euclidean distance, Manhattan distance, etc. In addition to the above commonly used metrics, more effective metrics have emerged in recent years to further improve the classification performance of the model. Das et al. [16] proposed a multilayer neural network, which performs knowledge transfer in two phases: feature extraction and distribution. In the first stage, the relative features are used to obtain the discriminant space, and the model is utilized to predict the variance of each class. Then the Mahalanobis distance of the mean class representation is computed to perform the categorization. In the second stage, a category agnostic transformation from the mean class-mean representation to the class prototype representation is learned. Few-shot Learning with Graph Neural Network [17] is the first application of the concept of graph neural network in few-shot image classification, which contains two parts, the embedding module and the GNN module, where the embedding module extracts the deep features of the samples. The nodes in the GNN module are the deep features and the edges represent the degree of similarity between the nodes, according to the degree of similarity the corresponding adjacency matrix is computed, and finally the classification of query samples is achieved. Li et al. [18] proposed Deep Nearest Neighbor Neural Networks to classify the samples using the local features of the samples, which form a local feature space of the classes, and classify the query samples by comparing the similarity relationship between the query samples and the classes.

To the best of our knowledge, metric learning nowadays mainly focuses on a single similarity metric, which is highly susceptible to metric bias and leads to poor classification performance of the model. Therefore, in the metric module of DFRN, we utilize the query feature vector and the secondary feature vector obtained from the dual feature module to compute the similarity, while considering the influence of the first-level feature vector obtained from the dual feature module, which mitigates the risk of overfitting and further improves the classification accuracy of the model.

B. Loss Function

The loss function usually calculates the difference between the predictions and ground truths, and the neural network is trained by minimizing the difference obtained from the loss function. Li et al. [19] proposed a double-crossing loss function by adding a penalty term on the basis of the cross-entropy loss function, which reduces the probability of the category assignment error and mitigates the effect of the gradient vanishing on classification. Wu et al. [20] proposed a triple loss function that reduces the distance between similar samples and increases the distance between dissimilar samples. The loss function in MeTAL [21] is modified in the inner loop under the guidance of the current task requirements, which enables the loss function to adapt to each task more efficiently and improves the generalization ability of the model.

Different from the above methods, we have multiple loss functions in our proposed DFRN, in which the classification loss of the first-level feature vector is only used to compute the attention mask, and throughout the training process, we update the network parameters by using the combined back-propagation of the classification loss of the secondary feature vector and the ternary auxiliary loss to further optimize the embedding module.

III. Theoretical Basis

In this section, we first define the few-shot learning problem. Then, we describe the DFRN in detail.

A. Problem Definition

The dataset for a few-shot image classification task is usually divided into a train set T_{train}, a test set T_{test} and a validation set T_{val}, which are mutually exclusive to each other. Their corresponding label sets are denoted as $\mathcal{L}_{\text{train}}$, $\mathcal{L}_{\text{test}}$ and $\mathcal{L}_{\text{val}}$, respectively.

In the training phase, C different classes are randomly selected in the training set and K different images are randomly selected from each class to form the support set $S = \{(x_e, y_e)_{e=1}^{N}$, where x_e is the e image in the support set and y_e is its label. Then, a portion of the remaining training samples in each class is randomly selected as the query set $Q = \{(x_p, y_p)_{p=1}^{M}$, where x_p is the p image in the query set and y_p is its label. Thus, the few-shot image classification task is often referred to as a "C-way K-shot" task.

B. Dual Feature Reconfiguration Network

In this paper, we propose a Dual Feature Reconstruction Network (DFRN) for few-shot classification with Feature Map Reconstruction Network (FRN) [22]. The details of the DFRN module are as follows:

1) Embedded Module: The embedding module is used to generate the deep features of the input support samples and query samples. The support sample features are aggregated to generate the support feature matrix. Given the support sample x_e and the query sample x_p, the support sample feature t_e and the query sample feature h_p are obtained through the embedding module f_θ the calculation formula is shown as follows:

$$t_e = f_\theta(x_e) \in \mathbb{R}^{a\times h\times w} \tag{1}$$

$$h_p = f_\theta(x_p) \in \mathbb{R}^{a\times h\times w} \tag{2}$$

where a denotes the number of channels, and h and w denote the height and width of the feature map. Then, the support sample features t_e are reconstructed as $d = h \times w$ dimensional local descriptors, and for each class, the k support sample features from class $c \in C$ are aggregated together and represented by the support feature matrix $S_c \in \mathbb{R}^{kd\times a}$.

2) Feature Reconstruction Module: The feature reconstruction module generates support feature reconstruction by reconstructing the support feature matrix in a ridge regression manner. The support feature matrix is fed into the feature reconstruction module and a closed-loop solution are found by means of ridge regression to reconstruct the support feature matrix into support feature reconstruction, which is calculated as shown below:

$$t_e^{'} = \gamma Q S_c^T (S_c S_c^T + \lambda I)^{-1} S_c \tag{3}$$

$$t_e^{'} = \gamma Q (S_c^T S_c + \lambda I)^{-1} S_c^T S_c \tag{4}$$

where γ and λ are learnable parameters, γ and λ are set to e^α and e^β to ensure their non-negativity, and α and β are learnable hyperparameters, which are initialized to zero. The computational equation is shown below:

$$\gamma = \frac{kd}{c} e^\alpha a \tag{5}$$

$$\lambda = e^\beta \tag{6}$$

We can find that Eq.(3) and Eq.(4) are equivalent in terms of results, but we still compute the support feature reconstruction in these two ways. The reason for this is that the performance consumption caused by computing the transpose of the matrix and the inverse of the matrix is different. When $a > kd$, it is recommended to use Eq.(3). This is because the most performance-consuming step is the computation of the transpose of the $kd \times kd$ matrix. When $a < kd$, it is recommended to use Eq.(4). because the most performance-consuming step in this case is solving the inverse of the $d \times d$ matrix. The support feature reconstruction is fed into the dual feature module to further explore and exploit the space of feature maps to mine richer information. For the query sample feature, the query feature vector $h_p^{'}$ is obtained by global average pooling.

3) Dual Feature Module: The dual feature module extracts a double feature vector for support feature reconstruction and consists of a first-level feature module and a secondary feature module. The first-level feature module is used to generate attention masks and thus identify important regions, while the secondary feature module is used to simulate real-world interference situations, so that the classifier can focus on non-comfortable regions, dig deeper and utilize the feature maps, and learn more diverse discriminative information. It is worth noting that the first-level feature vector is only used for model training. Based on the support feature reconstruction $t_e^{'}$, two copies of it are fed into the first-level feature module and the secondary feature module, and two feature vectors are output. The first-level feature vector $t_{e-1}^{'}$ is obtained through the first-level feature module, and the secondary feature vector $t_{e-2}^{'}$ is obtained through the secondary feature module. First, an attempt is made to obtain the attention mask T_a of the first-level feature module, which is calculated as shown below:

$$T_a = arg\max_T(E(F(t_e^{'}, T), y)) \tag{7}$$

where y is the true label supporting the feature reconstruction, T is the entire region of the sample, F is the classifier and $E(\cdot,\cdot)$ is a measure of the classifier prediction accuracy. L_{flat} is used to express the accuracy of the classifier prediction and is calculated based on the average mask $T_0 = 1/h \times w$ updated attention mask T_a of the secondary feature module, as shown below:

$$T_a = T_0 + \mu \frac{\partial L_{flat}}{\partial T}|_{T_0} \tag{8}$$

where μ is the step size. Note that the attention mask T_a is obtained by back propagation. The average mask T_0 is applied to the support feature reconstruction to obtain the secondary feature vector. The calculation formula is shown below:

$$t_{e-2}^{'} = \textstyle\sum_{i,j} t_e^{'} T_0 \tag{9}$$

where (i, j) represents the spatial location of the support feature reconstruction. The first-level feature vector can be calculated directly based on the attention mask and the support feature reconstruction, and the calculation formula is shown below:

$$\begin{aligned} t_{e-1}^{'} &= \textstyle\sum_{i,j} T_a t_e^{'} \\ &= \textstyle\sum_{i,j} (T_0 + \mu \frac{\partial L_{flat}}{\partial T}|_{T_0}) t_e^{'} \\ &= \textstyle\sum_{i,j} T_0 t_e^{'} + \sum_{i,j} \mu \frac{\partial L_{flat}}{\partial T}|_{T_0} t_e^{'} \\ &= t_{e-2}^{'} + \mu \triangle t_{e-2}^{'} \end{aligned} \tag{10}$$

From Eq.(10), it is clear that the first-level feature vector is a small perturbation combination of the secondary feature vector and noise, which helps to mitigate the risk of overfitting. It is worth noting that the first-level feature vector is only used for the training of the model.

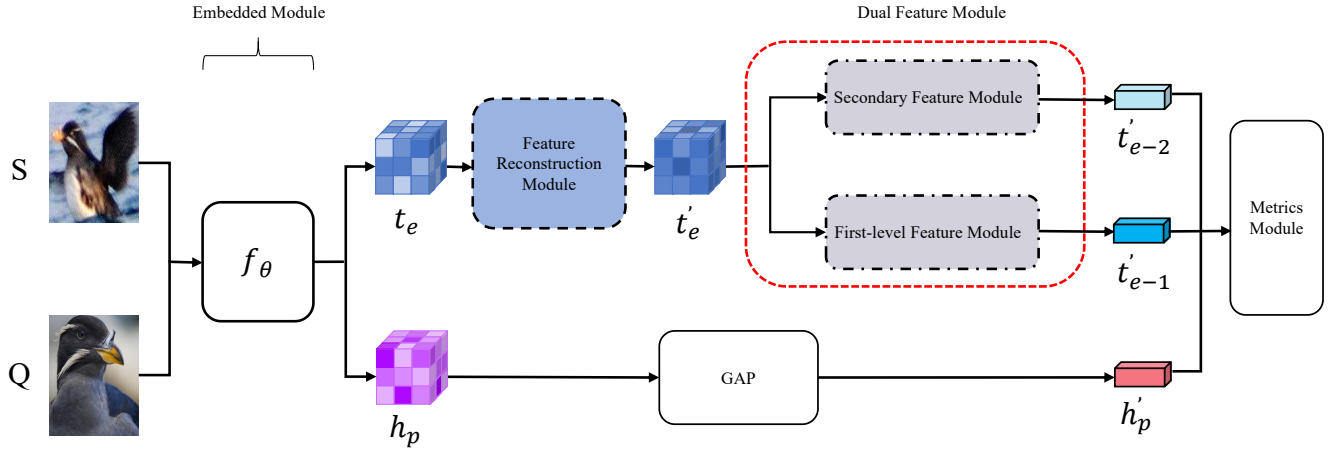


Fig. 1. DFRN network. The proposed network consists of an embedded module, feature reconstruction module, dual feature module and metrics module. The embedded module extracts the features t_e and h_p of the support sample and query sample respectively. The support sample feature t_e is passed through the feature reconstruction module to get the reconstructed feature of the support sample $t_e^{'}$. The feature vector $h_p^{'}$ of the query sample is the query sample feature h_p obtained by the Global Average Pooling(GAP). $t_e^{'}$ undergoes a dual feature module to obtain the first-level feature vector $t_{e-1}^{'}$ and the secondary feature vector $t_{e-2}^{'}$. The metrics module uses the cosine distance to calculate the similarity score of the secondary feature vector $t_{e-2}^{'}$ and the feature vector of the query sample $h_p^{'}$, so as to realize the classification of the query sample. This example illustrates the 1-way 1-shot learning process.

4) Metrics Module: The metric module outputs the similarity scores of the query samples and the support samples. The query feature vector and the secondary feature vector are fed into the metric module, and the cosine similarity is used to measure the similarity score between them to achieve the classification of the query samples.

5) Loss Function: The formula for calculating the classification loss of the first-level feature vector is shown below:

$$L_{level-1} = -\log p(y = c|x)t_{e-1}^{'} \tag{11}$$

where y is the label of the support sample and $p(y = c|x)$ is the Softmax probability. It is important to note that $L_{level-1}$ is only used to calculate the attention mask and is not used for parameter updates. The formula for calculating the classification loss of the secondary eigenvector is shown below:

$$L_{level-2} = E(F(t_{e-2}^{'}, T), y) = \sum_{i=1}^{n} -p\log(p_i) \tag{12}$$

where p_i is the predicted probability of the i class and n is the number of classes. In addition, an auxiliary loss function is invoked to further learn a more compact metric space by considering the contextual information of the support classes. We design a triple $(q, S_e, S_e^{'})$, where q denotes the distribution of a query, S_e is the distribution of a class matching q in the support set and $S_e^{'}$ is the distribution of the remaining classes, and the auxiliary loss function is calculated as shown below:

$$L_{aux-KL} = \sum_{p=1}^{n}\sum_{e=1}^{m}[(q\log\frac{q}{S_e}) - (q\log\frac{q}{S_e^{'}})] \tag{13}$$

Throughout the training process, the loss function is a combination of the classification loss of the secondary feature vector and the auxiliary loss function, which is calculated as shown below:

$$L_{train} = L_{level-2} + L_{aux-KL} \tag{14}$$

In this paper, gradient descent is used to optimize the L_{train}, and the L_{train} is back-propagated to update the network parameters and further optimize the embedding module.

IV. Experiments Result

A. Dataset

For the experiments in this chapter, three commonly used fine-grained datasets were selected: Flowers-102 (Flowers), Stanford-Cars (Cars), and CUB-200-2011 (CUB). In the remainder of this chapter, short symbols in parentheses are used to indicate.

Flowers dataset [23] contains 102 classes of flowers, and the categories of train/validation/test are divided into 60/20/22 in this paper. Cars dataset [24] contains 196 classes of cars, in which the train/validation/test categories are divided into 130/17/49 in this paper. The CUB dataset [25] contains 200 species of birds, and the categories of train/validation/test in this paper are 100/50/50.

B. Implementation Details

To evaluate the method in this paper, 5-way 1-shot and 5-way 5-shot classification tasks were performed, respectively, with Conv-4 as the backbone network. Conv-4 consists of 4 contiguous 64 channels convolution blocks, each consisting of a 3×3 convolution, a batch normalization, and ReLU nonlinear layers. The input image size is 84×84 and the output feature map size is 64×5×5. The experiments in this chapter train the model by SGD optimizer with an initial learning rate set to 0.1. During the training period, standard data enhancement methods are used. The number of iterations for the experiments is 800

under Conv-4. The final test results were obtained with 10,000 experiments during the testing period and 95% confidence intervals were reported.

C. Classification Performance

To verify the effectiveness of DFRN, the experiments use Conv-4 as the backbone to demonstrate the classification performance of each few-shot image classification method on Flowers, CUB and Cars datasets regarding 5-way 1-shot and 5-way 5-shot settings. As shown in Table I and II, the accuracy of the DFRN method is improved over all other methods.

TABLE I
CLASSIFICATION PERFORMANCE OF EACH METHOD AT 5-WAY 1-SHOT ON FINE-GRAINED DATASETS. THE BOLDED PARTS ARE THE METHODS PROPOSED IN THIS PAPER.

Method	5-Way 1-shot Accuracy(%)		
	Flowers	CUB	Cars
PrototypeNet	64.22±0.25	62.83±0.25	48.43±0.22
DN4	70.41±0.95	54.23±0.61	59.80±0.71
DSN	67.65±0.91	71.48±0.92	49.21±0.85
BSNET	66.62±1.03	62.85±0.95	40.88±0.75
FRN	73.68±0.22	73.56±0.21	63.41±0.22
DFRN(Ours)	**73.99±0.23**	**74.90±0.21**	**64.77±0.22**

TABLE II
CLASSIFICATION PERFORMANCE OF EACH METHOD AT 5-WAY 5-SHOT ON FINE-GRAINED DATASETS. THE BOLDED PARTS ARE THE METHODS PROPOSED IN THIS PAPER.

Method	5-Way 5-shot Accuracy(%)		
	Flowers	CUB	Cars
PrototypeNet	84.93±0.17	70.63±0.15	71.40±0.16
DN4	84.52±0.79	74.70±0.43	87.60±0.48
DSN	80.41±0.73	83.52±0.58	60.71±0.75
BSNET	85.09±0.61	85.40±0.55	84.68±0.51
FRN	88.43±0.14	88.31±0.13	83.18±0.14
DFRN(Ours)	**88.95±0.14**	**89.34±0.12**	**84.75±0.13**

D. Ablation Study

In this section, an ablation study is performed to test the effectiveness of the dual feature module and the auxiliary loss function on the CUB and Cars datasets using different backbone networks, and the experimental results are reported in Table III. The composition of the ablation methods in this section is set up as follows:

NAFRN: In the DFRN method, only the dual feature module is used for the experiments, and the other parts are set up in the same way as the FRN method settings.

AFRN: In the DFRN method, only the auxiliary loss function is used for the experiments, and the settings of the other parts are the same as those of the FRN method.

From the experiment results in Table III, it can be seen that the DFRN method achieves the best classification performance and that NAFRN and AFRN achieve the best classification performance compared to the baseline method FRN in 5-way 1-shot and 5-way 5-shot.

TABLE III
CLASSIFICATION PERFORMANCE OF ABLATION EXPERIMENTS UNDER THE CONV-4 SETTING ON FINE-GRAINED DATASET.

Method	5-Way 1-shot Accuracy(%)		5-Way 5-shot Accuracy(%)	
	CUB	Cars	CUB	Cars
FRN	73.56±0.21	63.41±0.22	88.31±0.13	83.18±0.14
NFRN	74.21±0.22	64.35±0.22	88.98±0.12	84.45±0.13
AFRN	73.93±0.21	63.97±0.22	88.85±0.12	83.88±0.14
DFRN(Ours)	**74.90±0.21**	**64.77±0.22**	**89.34±0.12**	**84.75±0.13**

V. CONCLUSION

In this paper, a dual feature reconstruction network is proposed. The network takes into account information that could lead to inaccurate predictions by using attention masks to further identify distinguishable regions of the target, forcing the classifier to leave its comfort zone and focus its attention on other regions, thus mining more diverse discriminative information. In this paper, we test the classification performance of DFRN on Flowers, CUB and Cars datasets, and the experimental results show that the classification performance of DFRN is significantly higher than that of FRN.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62176110, the Key Research and Development Program of Gansu Province under grant 22YF7GA130, Young Doctoral Fund of Education Department of Gansu Province under Grant 2021QB-038, Hong-liu Distinguished Young Talents Foundation of Lanzhou University of Technology.

REFERENCES

[1] Xiaoxu Li, Zhuo Sun, Jing-Hao Xue, and Zhanyu Ma. A concise review of recent few-shot meta-learning methods. *Neurocomputing*, 456:463–468, 2021.

[2] Jun Shu, Zongben Xu, and Deyu Meng. Small sample learning in big data era. *arXiv preprint arXiv:1808.04572*, 2018.

[3] Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang. Learning from very few samples: A survey. *arXiv e-prints*, pages arXiv–2009, 2020.

[4] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[6] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7260–7268, 2019.

[7] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *International conference on machine learning*, pages 7115–7123. PMLR, 2019.

[8] Ziyang Wu, Yuwei Li, Lihua Guo, and Kui Jia. Parn: Position-aware relation networks for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6659–6667, 2019.

[9] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[10] Junsik Kim, Tae-Hyun Oh, Seokju Lee, Fei Pan, and In So Kweon. Variational prototyping-encoder: One-shot learning with prototypical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9470, 2019.

[11] Fangyu Wu, Jeremy S Smith, Wenjin Lu, Chaoyi Pang, and Bailing Zhang. Attentive prototype few-shot learning with capsule network-based embedding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 237–253. Springer, 2020.

[12] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[13] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020.

[14] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3603–3612, 2019.

[15] Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, and Zhe Ma. Eckpn: Explicit class knowledge propagation network for transductive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6596–6605, 2021.

[16] Debasmit Das and CS George Lee. A two-stage approach to few-shot learning for image recognition. *IEEE Transactions on Image Processing*, 29:3336–3350, 2019.

[17] Cen Chen, Kenli Li, Wei Wei, Joey Tianyi Zhou, and Zeng Zeng. Hierarchical graph neural networks for few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):240–252, 2021.

[18] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7260–7268, 2019.

[19] Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao. Dual cross-entropy loss for small-sample fine-grained vehicle classification. *IEEE Transactions on Vehicular Technology*, 68(5):4204–4212, 2019.

[20] Fangyu Wu, Jeremy S Smith, Wenjin Lu, Chaoyi Pang, and Bailing Zhang. Attentive prototype few-shot learning with capsule network-based embedding. In *European Conference on Computer Vision*, pages 237–253. Springer, 2020.

[21] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Lee. Meta-learning with task-adaptive loss function for few-shot learning. pages 9445–9454, 10 2021.

[22] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8012–8021, 2021.

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

A Cloud-based Data Platform for Efficient EEG Data Management, Collaboration, and Analysis

Qi Tian*, Wen Wu*, Qin Zhu*, Tao Cai*, Siyi Jiang*, Yaqing Li*, Jinrun Zhou*, Nan Zhu*, Yina Wei*,†, Tao Tang*,†, Kedi Xu‡, Feng Lin* and Linqing Feng*,†,1

* Research Center for Augmented Intelligence, Institute of Artificial Intelligence, Zhejiang Lab, Hangzhou, China
E-mail: tianq@zju.edu.cn, fenglq@zhejianglab.com
†College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China
‡Qiushi Academy for Advanced Studies (QAAS), Zhejiang University, Hangzhou, China

Abstract —**Electroencephalography (EEG) data plays a crucial role in brain science research as it provides valuable insights into the mechanisms of brain functions. Typical EEG research involves conducting experiments, collecting data, and comprehensive data analysis. The success of such complex studies truly hinges on the collective efforts of multiple researchers. An efficient tool is required to facilitate such collaborative research. While many related tools exist, none specifically aim to enhance efficiency across data management, data analysis, and collaboration. To address this issue, we developed and deployed an integrated cloud-based platform for EEG experiment management, collaboration, and data analysis. Verified by practice, our platform facilitates collaboration, simplifies EEG data analysis for researchers not proficient in programming, and boosts the efficiency of overall EEG research.**

I. INTRODUCTION

Electroencephalography (EEG) is a technique that records the electrical activity between neurons in the brain. It provides valuable insights into the mechanisms of various brain activities, such as cognition, memory, sleep, and emotion, and helps neuroscientists unravel the secret of the human brain [1], [2]. As a real-time technique that directly reflects the ongoing conditions inside the human brain, EEG has been widely used in clinical diagnosis and treatment of brain disorders, including epilepsy, Alzheimer's disease, and anxiety [3], [4], [5]. EEG provides valuable information for monitoring and measuring changes in brain activity related to different tasks and contexts, which assist in investigating the mechanisms, structure, and function underlying cognition and behavior.

Both the process of conducting experiments to collect data, and the subsequent analysis of EEG data, are important in neuroscience research. They underpin hypothesis testing and enable the uncovering of new scientific discoveries [6], [7]. Various works are involved in these two procedures, such as experimental design, subject recruitment, data collection, data cleaning, data preprocessing, and data analysis. To ensure the overall success of a neuroscience study, multiple researchers may collaborate concurrently to accelerate the progress of the study in different tasks [8]. Therefore, a platform that can improve the efficiency of each task is critical for accelerating the entire research project.

The collection of EEG data in neuroscience research incurs significant financial costs, including those for equipment acquisition and participant recruitment [7], [9]. Consequently, the collected data represent valuable resources that warrant careful management for data security and to support subsequent analyses. A variety of data management tools have been developed specifically for EEG research [6], [10]. In EEG data analysis, there is an ongoing need for user-friendly, accurate, and efficient tools capable of handling large amounts of EEG data. A range of tools has been developed to assist researchers in effective EEG analysis, including MATLAB-based tools like EEGLAB and FieldTrip [11], [12], and the Python-based package MNE-Python library [13]. These tools enable researchers to tailor their data analysis to their needs, providing a convenient way to accomplish analysis tasks, and given familiarity with the respective programing language. They are less user-friendly and less effective for researchers lacking programming proficiency [14]. Additionally, researchers' varying programming skills could influence their choice of analytical methods. Moreover, the availability of hardware resources, such as computing power, could also limit researchers in their choice of specific analysis strategies. As a

[1] Corresponding Author: Linqing Feng, fenglq@zhejianglab.com

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

result, the introduction of diverse method choices in research complicates the comparison of results across different research projects [15]. Addressing this issue, Li Dong et al. developed a cloud-based brainformatics platform(WeBarin) [14] designed to store EEG data on a cloud server and provide online computing services, thereby enhancing the efficiency and consistency of EEG analysis. Similar tools are also reported in other studies [16], [17]. In summary, while a host of tools and platforms exist to aid in various aspects of EEG research, the need for a comprehensive solution that combines data management, collaboration, and data analysis remains a key challenge. Addressing this challenge could significantly enhance the efficiency and reproducibility of EEG studies.

Beyond analyzing EEG data, a neuroscience research project involves tasks like experiment arrangement and data collection, which typically require the collaboration of multiple researchers [6], [9]. Providing support to researchers in data analysis alone is insufficient to advance neuroscience research. Boosting the efficiency of other tasks, such as data management, is of equal importance. However, existing EEG analysis tools rarely offer adequate support for these tasks. To bridge this gap, our study presents a cloud-based data analysis collaboration platform designed to streamline EEG data management, enhance research collaboration, and simplify EEG data analysis. The platform provides cloud-based services encompassing data storage, computing, analysis, sharing, and collaboration. This facilitates a more efficient development of EEG-related experiments and aids in the analysis of EEG data to decode the connections between data and brain activity.

II. METHODS

A. Architecture design

The system architecture of this platform, structured on a browser-server architecture, is illustrated in Fig. 1. The backend of the platform resides on a cloud server, allowing users to interact with cloud-based data via a front-end browser. The cloud server comprises an administrative database, an EEG file database, and a computation server, which are used for storing administrative data, storing EEG files, and analyzing and computing large amounts of EEG data, respectively. The platform mainly consists of four core modules: data management, data visualization, data query, and data analysis. Users can upload collected EEG data wirelessly or through a hard drive, with commonly used EEG storage formats in neuroscience research, such as EDF, BDF, and CNT. Following the upload, the data undergoes a simple quality check, whereby accepted data is stored on the file server. For uploaded data, Vue.js [18] front-end display frameworks are used for data visualization. The search function of the platform is built on the Elasticsearch [19] engine, which constructs retrieval indexes based on uploaded files and sorts them based on the similarity of each signal. The platform's analysis function utilizes the MNE algorithm library and enables researchers to upload and share their custom algorithm files for EEG analysis purposes.

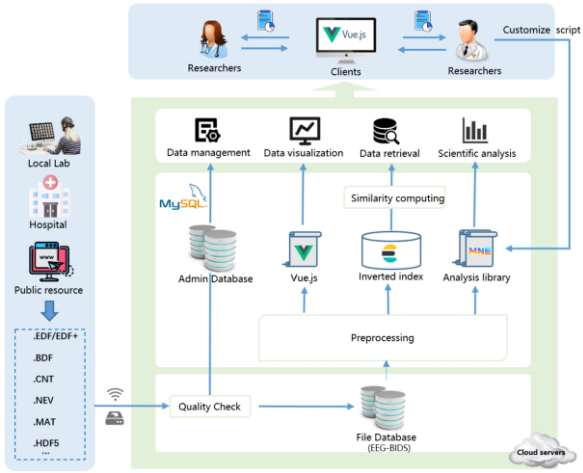


Fig. 1 The architecture of the platform

B. Workflow design

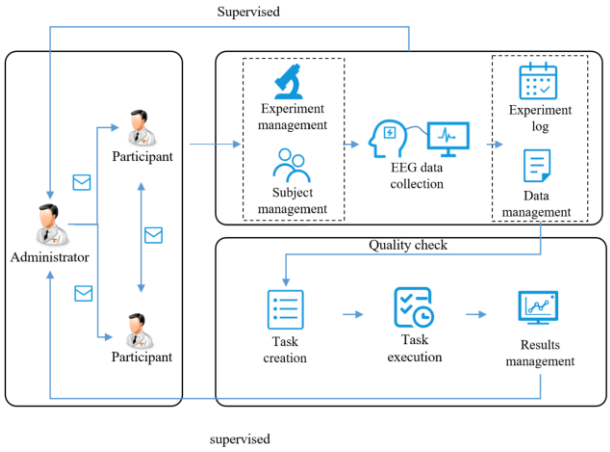


Fig. 2 The workflow of the platform.

The platform is designed with a workflow that encompasses all research processes, assisting researchers in their corresponding tasks. At the initial research stage, the research administrator can assemble their team, with all participants able to coordinate and record their work on the platform. In the

data collection phase, researchers can use the platform to create a project to record relevant information about an EEG experiment, such as the subject's details, and schedule the experiment accordingly. All information linked to the experiment, such as its location, duration, participants, and devices used, can be recorded to the platform for management and retrieval. The data generated from the experiment will also be uploaded to the platform and associated with details such as the experiment operator, subject, and device, making subsequent data management easy.

In the data analysis phase, researchers responsible for data analysis can create corresponding analysis tasks. The platform can provide users with commonly used data analysis support, such as EEG data filtering, time-frequency analysis, and EEG topomap. Researchers can also upload their custom EEG analysis code files for data analysis. The defined analysis tasks will be executed sequentially by the computation server, and the data and results generated in each step will be stored on the platform. Researchers can check them online or download them for further analysis on a local computer.

The experiment administrators can assign tasks to each participant and monitor the progress of each task. They also have operational authority over data, algorithms, and execution results involved in their project. This enables the administrator to ensure the experiments and data acquisition proceeds as planned or choose to share data, algorithms, and results on the platform after the research for communication and comparison.

To boost data analysis efficiency and minimize the impact of individual researchers' programming skills, our platform provides a user-friendly interface (Fig. 4) for data analysis functions. This obviates the need for researchers to write and execute code for relevant algorithms; instead, they simply fill in the corresponding parameters in the analysis algorithm's interface for data analysis. Additionally, researchers can save the defined parameters as fixed schemes, so they do not have to redefine them when performing the same analysis in the future but can instead select the scheme. Researchers can choose to make custom schemes public or private to facilitate other researchers' studies.

III. Results

A. *The EEG platform*

The screenshot of the platform is shown in Fig. 3. Corresponding to the structure diagram, the platform's four core modules can be accessed using the three labels located on the navigator's left-hand side in Fig. 3. The "Experiment" label provides access to the data management and visitation modules, enabling researchers to review and manage the information and data relevant to their EEG experiments. The "TaskList" label links to the data analysis module, allowing neuroscientists to create data processing tasks, such as filtering and time-frequency analysis. Finally, the "SignalSearch" label links to the data retrieval module, permitting users to retrieve EEG files that contain fragments matching their selected target signal.

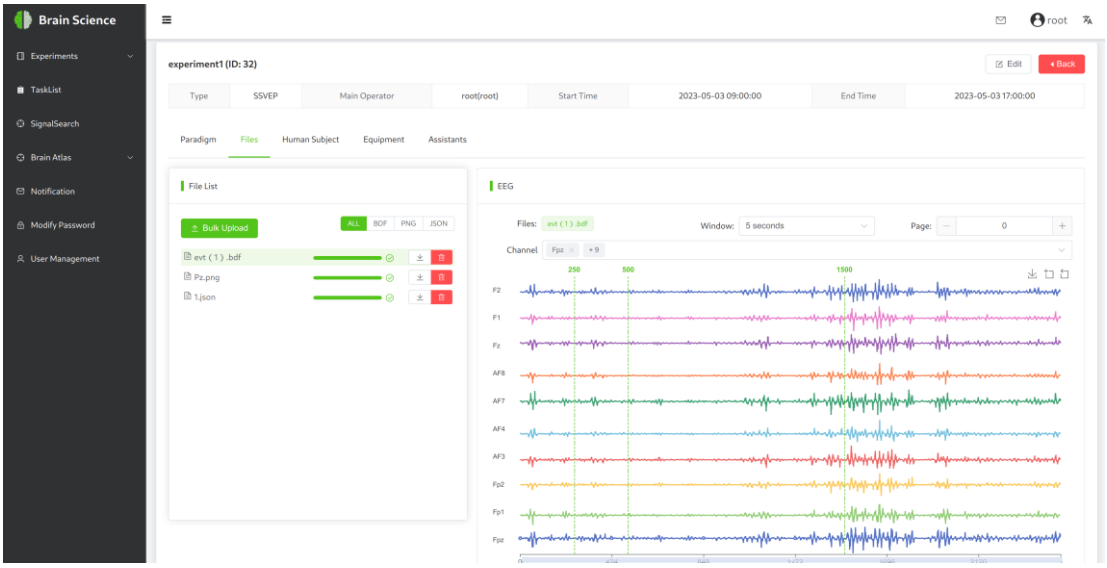


Fig. 3 Screenshot of this platform

C. *EEG analysis design*

Other labels in the platform's navigator are maintenance

modules or functionalities that will be developed in the next version of the platform.

Currently, the platform has been deployed on the server of Zhejiang Lab, for internal testing. Three research teams, totaling 10 researchers, are utilizing the platform to conduct research in a variety of areas, including human steady state visually evoked potential(SSVEP), motor imagery, human sleep, and EEG related to some animal experiments.

B. Comparison with other EEG tools

We conducted a comprehensive comparison of the functional characteristics of the platform developed herein with existing EEG research-related tools, focusing on three key aspects of EEG research: data management (Table 1), collaborative work (Table 2), and data analysis (Table 3). The first rows in each of these tables display the name of the tools. For tools that were not explicitly named in the corresponding research, we represented them by their author's name. Throughout the tables, we referred to the platform developed in this study as "brain science".

Specifically, in the domain of data management, we compared the tools based on their provision of services such as data storage, data retrieval, EEG experiment management, subject management in experiments, and data sharing capabilities with other researchers.

Table 1. Comparison form data management aspect

	Storage	Retrieval	Experiment	Subjects	Share
Musestudio [7]					
mEEG[10]	✓			✓	
DRT[20]					
BrainBase[6]	✓	✓	✓	✓	✓
WeBrain[14]	✓	✓			✓
ReportFlow [21]	✓			✓	✓
BEAPP[16]					
Alphonse's [17]					
Sans Tracas[9]	✓		✓	✓	
Petr's [22]	✓				
brainscience	✓	✓	✓	✓	✓

Regarding workflow support, our analysis focused on whether these tools offered assistance in areas like authority management, data collection, message exchange, and other workflow-related functionalities crucial for conducting EEG research.

Table 2. Comparison form work flow support aspect

	Authority management	Data collection	Messages exchange	Workflow
Musestudio		✓		
mEEG				
DRT				
BrainBase		✓		
WeBrain	✓		✓	
ReportFlow	✓	✓	✓	✓
BEAPP				
Alphonse's				
Sans Tracas	✓	✓		✓
Petr's				
platform				
brainscience	✓	✓	✓	✓

In the context of data analysis, we thoroughly examined whether the analyzed tools provided support for common analysis requirements encountered in EEG research. These requirements encompassed tasks such as format transformation of EEG files, visualization, filtering, classification, and time-frequency analysis. It is important to note that the needs for EEG analysis are diverse, and each tool has its own specific design intentions. Thus, certain less common analysis functions were not exhaustively outlined in the table.

Table 3. Comparison form data analysis aspect

	Transformation	Visualization	Filtering	Classification	time-frequency analysis
Musestudio	✓	✓			
mEEG		✓	✓	✓	
DRT	✓				
BrainBase					
WeBrain		✓	✓		✓
ReportFlow					
BEAPP	✓	✓	✓		
Alphonse's		✓	✓	✓	✓
Sans Tracas		✓			
Petr's		✓	✓	✓	
brainscience		✓	✓		✓

Overall, the majority of existing tools that primarily focus on providing functional support in one or two specific aspects, while few tools offer comprehensive support across all three aspects. DRT, Musestudio, Alphonse's, BEAPP, and Petr's primarily concentrate on data analysis. ReportFlow primarily emphasizes workflow support, whereas mEEG contributes to both data analysis and storage functionalities. WeBrain demonstrates the ability to provide support in multiple aspects, albeit lacking in experiments and subject management.

It is important to notice that some tools are not exclusively designed for the three aspects mentioned above. For instance, ReportFlow primarily focuses on the secure transmission of encrypted EEG reports, while the main purpose of DRT is to transform EEG data structures into a standardized format. Both

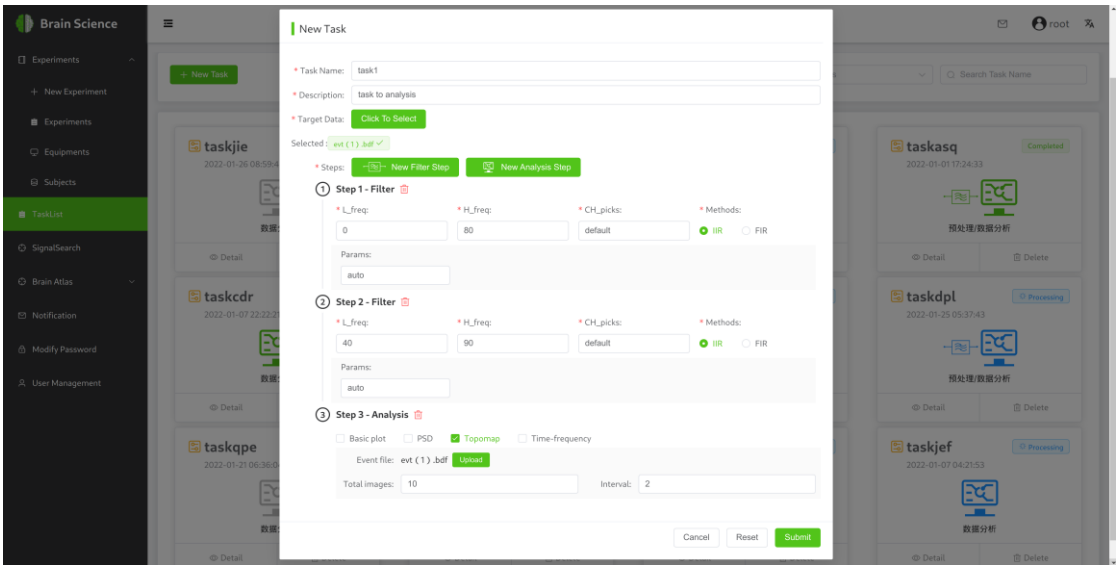


Fig. 4 The UI design of EEG analysis tasks

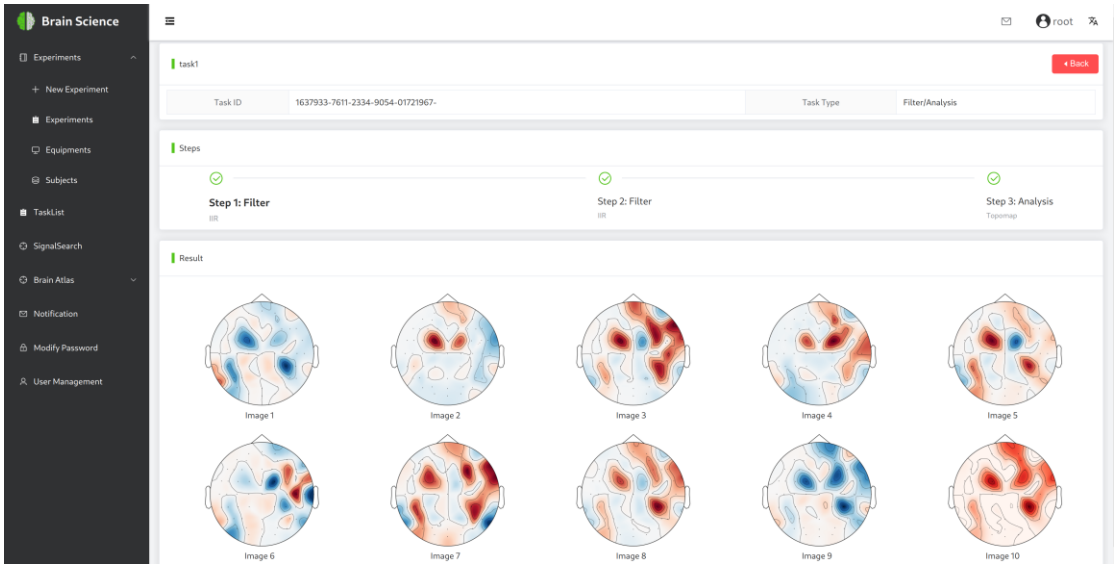


Fig. 5 The results of an EEG topomap task

of these tools, however, are useful for promoting EEG research by addressing specific areas of improvement.

C. An example of an EEG analysis task

Here, we use an EEG topomap task as an example to demonstrate how to analyze EEG data on the platform. Initially, the researcher creates a task in the task module, customizing the task's detailed content through the platform's user interface (Fig. 4). They can then select the "Target data" option and choose the data to be processed by the task. In this demonstration, we use data obtained from an SSVEP experiment conducted by our research team. The signals were captured using a 64-channel EEG recording device with a sampling frequency of 1,000 Hz

Next, filtering operations are added to the task. In this task, we implement an 80 Hz low-pass filter and a 40-90 Hz band-pass filter. Researchers can submit the parameters of two sets of filters in the form to achieve filtering (Fig. 4).

Subsequently, researchers can customize the parameters of the topomap task. These parameters include uploading event-related files, drawing intervals, and the number of image displays (Fig. 4).

Finally, the defined task can be submitted to the computing server, and results will be displayed on the platform in a few seconds (Fig. 5).

IV. Discussion

EEG data collection and analysis are two important steps in neuroscience. While many EEG data analysis tools have been designed and implemented to improve research efficiency, critical tasks such as data collection and experiment management, often involving multiple researchers, require equal attention. Therefore, enhancing collaboration among multiple researchers in different stages of neuroscience research is a fundamental requirement for present-day neuroscience researchers.

This study aims to design and implement a cloud-based EEG data management and analysis platform, providing researchers with cloud storage, and computing capabilities, and fostering multi-person collaboration to enhance neuroscience experiments. The platform facilitates the collection, management, and computation of large batches of EEG data, ultimately enhancing the efficiency of EEG experiments and promoting the development of neuroscience research.

In this study, we designed a workflow that aligns with the process of neuroscience research to facilitate collaborations among researchers. After actual deployment and testing, the platform effectively assisted researchers with their daily EEG experiment work and improved data management efficiency. By designing user-friendly EEG analysis functions, we reduce the need for advanced programming skills in EEG analysis tasks, making these tasks more accessible to new researchers.

Nevertheless, there remains room for further improvement in platform's functionality. Through the internal testing process, we received feedback indicating that the data analysis functions need to be expanded. Despite the implementation of interface operations for data analysis, the platform still needs enhancements to fully meet researchers' analysis requirements. Additionally, the current version of the platform lacks compatibility with the Brain Imaging Data Structure (BIDS) [23] data specification, which will be fixed before it is released to public in the near future. Lastly, to enable analysis of multi-modal neuroscience data, this platform needs to support more types of neuroscience data.

Currently, the platform is only available for internal testing. Nonetheless, we plan to release a public version with enhanced data security modules in the near future, aiming to assist researchers worldwide in their neuroscience research, facilitating academic communications and fostering progress in the field of neuroscience.

V. Conclusions

In this study, we have developed a cloud-based platform dedicated to EEG data management and analysis. This platform equips researchers with cloud computing and storage capabilities, streamlining their work process. In practical application, our platform not only diminishes the complexities of EEG data analysis but also enhances efficiency in data collection and data management. With these benefits, we believe our platform holds substantial potential in propelling the advancement of EEG-related neuroscience research.

Acknowledgment

This work is supported by the Research Project of Zhejiang Laboratory (No. 2022KI0AC02, 2022KI0AC01, K2023KA1BB01, K2023KI0AA04).

References

[1] S. Asadzadeh, T. Yousefi Rezaii, S. Beheshti, and S. Meshgini, "Accurate emotion recognition using Bayesian model based EEG sources as dynamic graph convolutional neural network

nodes," *Sci. Rep.*, vol. 12, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41598-022-14217-7.

[2] N. Robitaille, R. Marois, J. Todd, S. Grimault, D. Cheyne, and P. Jolicœur, "Distinguishing between lateralized and nonlateralized brain activity associated with visual short-term memory: fMRI, MEG, and EEG evidence from the same observers," *NeuroImage*, vol. 53, no. 4, pp. 1334–1345, Dec. 2010, doi: 10.1016/j.neuroimage.2010.07.027.

[3] P. M. Rossini et al., "Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis. Report from the IFCN-sponsored panel of experts," *Clin. Neurophysiol.*, vol. 131, no. 6, pp. 1287–1310, Jun. 2020, doi: 10.1016/j.clinph.2020.03.003.

[4] J. Ha, D.-H. Kim, and L. Kim, "An approach for assessing stroke motor function ability using the similarity between electroencephalographic power spectral densities on both motor cortices," in *2018 6th International Conference on Brain-Computer Interface (BCI)*, IEEE, 2018, pp. 1–3.

[5] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals," *IEEE Access*, vol. 6, pp. 25399–25410, 2018.

[6] T. Thiagarajan, "Brainbase: a research and data management platform for human EEG," in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, IEEE, 2017, pp. 1–5.

[7] M. Á. Sánchez-Cifo, F. Montero, and M. T. López, "MuseStudio: Brain activity data management library for low-cost EEG devices," Appl. Sci., vol. 11, no. 16, p. 7644, 2021.

[8] A. S. Malik and H. U. Amin, Designing EEG experiments for studying the brain: Design code and example datasets. *Academic Press*, 2017.

[9] R. Desai, R. Orji, E. Roudaia, and A. B. Sekuler, "Sans Tracas: A Cross-platform Tool for Online EEG Experiments.," in *Companion of the 2022 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 2022, pp. 37–43.

[10] D. Mirarchi, P. Vizza, P. Cinaglia, G. Tradigo, and P. Veltri, "meeg: A system for electroencephalogram data management and analysis," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2017, pp. 1642–1646.

[11] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[12] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Comput. Intell. Neurosci.*, vol. 2011, pp. 1–9, 2011.

[13] "MNE — MNE 1.2.2 documentation." https://mne.tools/stable/index.html (accessed Nov. 13, 2022).

[14] L. Dong et al., "WeBrain: A web-based brainformatics platform of computational ecosystem for EEG big data analysis," *NeuroImage*, vol. 245, p. 118713, Dec. 2021, doi: 10.1016/j.neuroimage.2021.118713.

[15] P. Isager, "Variability in the analysis of a single neuroimaging dataset by many teams," 2019.

[16] A. R. Levin, A. S. Méndez Leal, L. J. Gabard-Durnam, and H. M. O'Leary, "BEAPP: the batch electroencephalography automated processing platform," *Front. Neurosci.*, vol. 12, p. 513, 2018.

[17] J. Alphonse and S. Diwakar, "Deploying a web-based electroencephalography data analysis virtual laboratory," *Procedia Comput. Sci.*, vol. 171, pp. 2420–2425, 2020.

[18] "Vue.js - The Progressive JavaScript Framework | Vue.js." https://vuejs.org/ (accessed Jun. 07, 2023).

[19] G. Amato, P. Bolettieri, F. Carrara, F. Falchi, and C. Gennaro, "Large-scale image retrieval with elasticsearch," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 925–928.

[20] L. Dong et al., "DRT: A new toolbox for the Standard EEG Data Structure in large-scale EEG applications," SoftwareX, vol. 17, p. 100933, 2022.

[21] S. Bertuccio et al., “ReportFlow: an application for EEG visualization and reporting using cloud platform,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–9, 2021.

[22] P. Jezek and L. Vareka, “Cloud Infrastructure for Storing and Processing EEG and ERP Experimental Data.,” in *ICT4AWE*, 2019, pp. 274–281.

[23] C. R. Pernet et al., “EEG-BIDS, an extension to the brain imaging data structure for electroencephalography,” *Sci. Data*, vol. 6, no. 1, p. 103, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids

Xiajie Zhou, Candy Olivia Mawalim, Benita Angela Titalim, Masashi Unoki
* Japan Advanced Institute of Science and Technology, Japan
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: {s2210112, candylim, titalim, unoki}@jaist.ac.jp

***Abstract*—Recent studies in speech processing often utilize sophisticated methods for solving a task to obtain high-accuracy results. Although high performance could be achieved, the methods are too complex and require high-performance computational power that might not be available for a wide range of researchers. In this study, we propose a method to incorporate the low dimensional and the recent state-of-the-art acoustic features for speech processing to predict the speech intelligibility in noise for hearing aids. The proposed method was developed based on the stack regressor on various traditional machine learning regressors. Unlike other existing works, we utilized the results of the digit triplet test, which is usually used to measure the hearing ability in the existence of noise, to improve the prediction. The evaluation of our proposed method was carried out by using the first Clarity Prediction Challenge dataset. This dataset is utilized for speech intelligibility prediction that consists of speech signals output of hearing aids that were arranged in various simulated scenes with interferers. Our experimental results show that the proposed method could improve speech intelligibility prediction. The results also show that the digit triplet test results are beneficial for speech intelligibility prediction in noise.**

I. Introduction

Hearing loss is a widespread and significant problem that affects people's lives. With age and other factors, an individual's hearing function gradually deteriorates, decreasing speech comprehension and communication skills [1], [2]. Hearing loss directly impacts the clarity of an individual's speech. Hearing loss at specific frequencies makes it difficult for individuals to discriminate certain speech sounds, reducing speech intelligibility and comprehension [3]. This effect is further exacerbated by background noise and complex listening environments [4], [5]. As individuals' hearing function gradually deteriorates, they often suffer from physical burdens, further exacerbated by emotional depression due to the inability to communicate effectively with others.

A significant goal of hearing aids is to improve the speech intelligibility of individuals with hearing loss to better participate in social activities, communicate with others, and receive and understand verbal information effectively. Evaluation of hearing abilities takes place through various methods such as audiograms, speech tests, auditory questionnaires, auditory neurophysiological measures, and cognitive tests [6]–[8]. Existing methods for speech intelligibility prediction in hearing aids often rely only on audiograms to represent the status of hearing loss, which has certain limitations.

The traditional audiogram assessments only provide information about hearing sensitivity and frequency characteristics. In contrast, speech comprehension involves more complex auditory and cognitive processes, and audiograms do not fully reflect changes in these aspects [9]. Moreover, audiograms do not provide information about an individual's performance in a given speech environment, which is essential when assessing the speech intelligibility of hearing aids. The effects of environmental noise, speech intelligibility, and individual auditory processing abilities prevent conventional audiograms from providing detailed information on these aspects. To obtain more comprehensive and accurate hearing assessment results in assessing the speech intelligibility of hearing aids, we need to combine other assessment methods, such as the digit triplet test (DTT) as a method of assessing hearing ability in noise.

The DTT has essential applications in assessing speech perception in complex listening environments, such as background noise or competing sounds [10], [11]. The test uses digit triplets to measure an individual's ability to perceive and discriminate speech sounds in the presence of hearing loss, especially in the high-frequency range [12]. By measuring the speech reception threshold (SRT) obtained from the DTT, we can reveal an individual's ability to comprehend speech under adverse conditions and correlate it with the overall speech intelligibility in the real world. Lower SRT values indicate that individuals are more capable of successfully recognizing and discriminating speech in the DTT, which implies a higher level of discourse intelligibility. In addition, the DTT provides practical guidance for hearing aid assessment, selection, and individualized rehabilitation planning by measuring an individual's ability to perceive and discriminate speech at lower levels. It can also predict their overall speech intelligibility in the real world.

The purpose of this paper is to propose a speech intelligibility prediction method that considers the results of the DTT for hearing aids in complex listening environments. Our approach builds on a stacked regressor that combines the functionality of traditional machine learning regressors, including linear regressor, support vector regressor, and random forest regressor. Since the DTT can assess the hearing ability in the presence of noise, we hypothesize that incorporating the results of the DTT into our prediction model has the potential to improve the prediction further.

The paper next consists of five sections. Section 2 includes details of the related work, i.e., hearing condition assessment, baseline of the modified Binaural Short-time Objective Intelligibility (MBSTOI), and speech intelligibility prediction.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I
SUMMARY OF SPEECH INTELLIGIBILITY ASSESSMENT METHODS

Assessment Methods	Phoneme Tests	Word Recognition Tests	Sentence Intelligibility Tests	Digit Triplet Test
Purpose	Assess discrimination of phonemes, syllables	Evaluate recognition of individual words	Measure understanding of sentences	Assess speech perception in connected speech
Response	Discrimination or identification	Identification or repetition	Repetition or understanding	Identification or repetition
Stimuli	Individual speech sounds or minimal pairs	Isolated words	Sentences	Triplet of connected speech sounds
Types of Tests	Phoneme Discrimination Tests, Phoneme Recognition Tests	Consonant-Vowel Tests	Speech Intelligibility Index, Hearing in Noise Test	Digit Triplet Test
Measurement Outcome	Discriminative thresholds, discrimination accuracy	Word recognition score	Threshold level for speech intelligibility	Accuracy or threshold level for speech intelligibility

Section 3 presents the proposed method, i.e., feature extraction and regression analysis. Section 4 describes the experiments, such as the dataset, evaluation metrics, and results. Section 5 shows the conclusion of our study.

II. Related Work

A. Hearing Condition Assessment

An audiogram measures the sensitivity of an individual's hearing at different frequencies. The threshold of the different sound intensities they can hear is measured by performing a hearing test and drawing an audiogram. The shape and results of the audiogram can provide detailed information about an individual's hearing status. A physician or audiologist can use the audiogram to help determine if there is a hearing loss and its degree and type. Recording the hearing thresholds of individuals at different frequencies plots the audiogram. These hearing thresholds represent the minimum sound intensity that an individual can hear. The pure tone average (PTA) provides the audiogram with a measure of the degree of hearing loss. It summarizes an individual's hearing sensitivity over a specific frequency range. A common way to calculate the PTA is to select frequency points. These frequency points are the most critical frequency ranges for speech and communication. A higher PTA value indicates a higher hearing threshold, i.e., an individual has a lower auditory sensitivity to sound, suggesting the presence of hearing loss.

Existing methods for assessing speech intelligibility with hearing aids include a series of tests to assess an individual's ability to perceive and understand language [11], [13]–[16]. These assessments aim to quantify the degree of speech intelligibility and provide valuable information for diagnosing hearing impairment, assessing treatment effectiveness, and guiding intervention strategies. All of the methods assess an individual's ability to perceive and comprehend language. The acceptable discrimination required by phoneme tests allows for detailed speech analysis but may only partially reflect the challenges faced in complex listening environments. The word recognition tests provide insight into recognition at the lexical level but may not reflect an understanding of communication at the sentence level. The sentence intelligibility tests present higher cognitive and linguistic demands, potentially affecting an individual's ability to understand and accurately respond to the entire sentence, which places a higher cognitive burden on individuals with hearing loss.

The DTT has demonstrated reasonable validity and reliability in assessing speech intelligibility [11]. The DTT presents a series of digraph triplets consisting of three speech sounds [10]. During the test, individuals recognized different digit triplets at different signal-to-noise ratios. By setting different signal-to-noise ratios in the test environment, the DTT provides insight into a person's ability to perceive speech in complex listening environments and helps to assess their overall communication ability more accurately. Compared to isolated phonemes, words, or sentences, the DTT uses a triplet of digital speech with specific phonetic features, and the test results more closely mimic real-life speech intelligibility.

The advantages of the DTT in complex listening environments and for individuals with hearing loss make it a promising option for assessing speech intelligibility in challenging listening conditions. The DTT has higher sensitivity and specificity than other testing methods, allowing for a more accurate assessment of an individual's ability to perceive and discriminate speech sounds. Moreover, the DTT results correlate strongly with audiogram PTA measurements, which serve as an index to assess the degree of hearing loss and reflect an individual's hearing threshold at different frequencies. The correlation with the PTA suggests that the DTT can be used as one indicator to complement the hearing status assessment, providing more comprehensive information for a hearing assessment. The DTT is the ability to capture the individual's ability to perceive the nuances of speech features, thus providing a more comprehensive assessment of the individual's speech intelligibility.

B. Baseline Modified Binaural Short-time Objective Intelligibility (MBSTOI)

The MBSTOI is an improved binaural short-time objective intelligibility index for speech intelligibility prediction in hearing aids. The short-time objective intelligibility (STOI) measures have successfully predicted the intelligibility of noisy speech processed by time-frequency (TF) weighting methods [17]. However, a limitation of STOI is its reliance on monophonic speech signals, which ignores valuable dual-channel

cues available to listeners.

The binaural short-time objective intelligibility (BSTOI) is based on the principles of STOI but introduces some modifications [18]. In BSTOI, the equalisation cancellation (EC) model uses information from binaural cues to simulate how the human auditory system works [19]. During the EC phase, the linear combination of the left and right ears is described by Eq.(1). $x_{k,m}^{(l)}$ is the signal of the TF unit corresponding to the left ear in the m-th time period and k-th frequency bin. The factor λ implements the equalization step in the EC stage.

$$\hat{x}_{k,m} = \lambda \cdot \hat{x}_{k,m}^{(l)} - \lambda^{-1} \cdot \hat{x}_{k,m}^{(r)} \tag{1}$$

The MBSTOI performs independent EC stages at each time range and each frequency band to estimate EC parameters. The MBSTOI improves binaural processing and optimized equalization stages over the BSTOI to more accurately predict speech intelligibility. Such improvements allow MBSTOI to better capture and utilize binaural information in the speech signal and improve the performance of speech intelligibility assessment.

C. Speech Intelligibility Prediction

In the field of hearing aids, predicting the speech intelligibility of listeners is an important research task. Such prediction can be achieved by objective measurements, guiding audiologists in selecting appropriate signal processing algorithms when fitting hearing aids, and providing valuable tools in developing machine learning-based hearing aid algorithms and other speech enhancement methods [20]. The objective measurement of speech intelligibility refers to analyzing various features and attributes of the speech signal. These features can include spectral characteristics of speech, time domain characteristics, and acoustic parameters. By modeling the relationship between these characteristics and listener comprehension, it is possible to predict the speech intelligibility of listeners with hearing loss under listening conditions.

The widespread use of machine learning methods in speech intelligibility prediction is due to their ability to learn patterns and correlations from large-scale speech data and apply them to new data samples for prediction. These methods use large datasets and features to train models to establish associations between features and speech intelligibility. Standard machine learning algorithms include support vector machines, random forests, and deep learning models. Through objective speech intelligibility measures and machine learning methods, audiologists and researchers can better understand listeners' speech intelligibility and select appropriate signal processing algorithms to improve listeners' listening experience.

III. Proposed Method

As the proposed model, the noisy speech after hearing aid processing is taken as input speech for feature extraction. The result of extracted speech features also combines with the bilateral audiogram and the DTT results. The aims to incorporate both audiogram and the DTT is to account for the individual listener's hearing profile and tailor the prediction accordingly. Besides, we use this holistic approach to enhance the accuracy and relevance of speech intelligibility predictions by taking into account the listener's individual auditory capability. After the feature extraction stage, the proposed model concludes with regression analyses to predict speech intelligibility.

A. Feature Extraction

In the feature extraction, we corporate various feature extraction process, including Geneva minimalistic acoustic parameter set (GeMAPS) and its extended version (eGeMAPS), wav2vec2, hidden unit bidirectional encoder representation from transformer (HuBERT), and waveform language model (wavLM). The details of each feature extraction are described in the following section.

1) GeMAPS and eGeMAPS: The open-source toolkit openSMILE incorporates to perform extraction of the GeMAPS and eGeMAPS parameter sets. Our proposed method includes these sets to extract features related to signal-to-noise ratio, spectral envelope, pitch, and temporal dynamics. The valuable information about speech characteristics might capture the relevant factors affecting speech intelligibility in the context of hearing loss perception under complex listening environments. The GeMAPS is a minimalistic parameter set that incorporating essential prosodic, excitation, vocal tract, and spectral features [21], [22]. Meanwhile, the eGeMAPS, as an extended version, includes additional cepstral features to enhance recognition accuracy beyond using only prosodic and spectral parameters.

In detail, the set of features extracted by the GeMAPS includes frequency-related, amplitude-related, and spectral features. Frequency-related features include pitch (measured in semitones), jitter (deviations in F0 period lengths), and formant frequencies. Amplitude-related features include a shimmer (difference in peak amplitudes), loudness (perceived intensity), and harmonics-to-noise ratio (HNR). Finally, Spectral features include:

- Alpha ratio: energy ratio between specific frequency ranges
- Hammarberg Index: ratio of energy peaks
- Spectral slope: regression slopes in frequency bands
- Formant relative energy
- Harmonic difference of the first harmonic to the second harmonic (H1-H2) and the third formant range (H1-A3)

These parameters undergo smoothing using a moving average filter and various functions to generate a set of parameters. These functionals include arithmetic mean, coefficient of variation, percentiles, range, and slope-related measures. Additionally, temporal features, such as the rate of loudness peaks and statistics of voiced and unvoiced regions, provide a comprehensive analysis of speech characteristics. In total, 62 parameters were extracted by GeMAPS.

As aforementioned, the extension of GeMAPS, the eGeMAPS, includes cepstral and dynamic parameters to model the affective states in speech analysis. The eGeMAPS introduces additional parameters, including Mel-frequency cepstral coefficients (MFCCs), spectral flux difference, and formant bandwidth. Functionals like arithmetic mean and coefficient

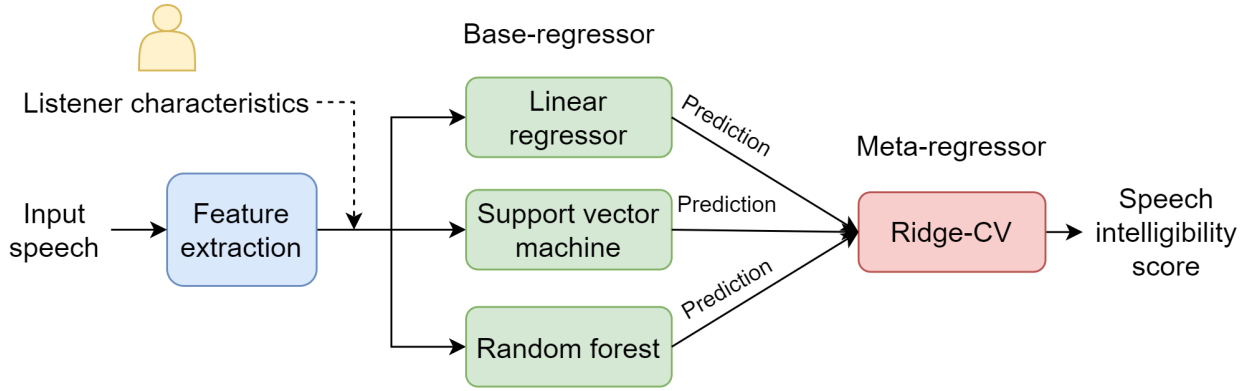


Fig. 1. Block diagram of our proposed method

of variation are applied to these parameters. The eGeMAPS set, combined with the GeMAPS, generates 88 parameters.

2) wav2vec2: The wav2vec2 is a framework of self-supervised learning in representing speech audio that is fine-tuned on transcribed speech [23]. We chose self-supervised learning because predicting speech intelligibility for hearing loss conditions using a speech recognition model only applies in real applications if it can deal with unlabeled data or explicit annotation. Besides, wav2vec2 extracts high-level acoustic features that capture important aspects of the speech signal useful for downstream tasks of predicting speech intelligibility.

The speech features extracted by wav2vec2 include the phonetic and linguistic properties by learning to model the temporal relationship between different phonemes or speech units. These representations encode information about the distinct speech and their temporal patterns, which are crucial for speech intelligibility. The wav2vec2 also captures the prosodic cues such as pitch variation, rhythm, and intonation, which play a role in conveying meaning and syntactic structure. Lastly, wav2vec2 extracts spectral information related to the frequency content, indicating phonetic distinctions and acoustic cues relevant to intelligibility.

3) HuBERT: Another self-supervised learning framework we consider is HuBERT [24]. HuBERT consists of three main components: the feature encoder, transformer encoder, and classification head, which appeared as the extension of wav2vec2. The feature encoder extracts low-level acoustic features from the input signal and converts them into time-frequency representation. Then, the transformer encoder captures the contextual information through unidirectional and bidirectional transformers. Finally, the classification head maps the acoustic features to the predicted speech intelligibility.

Although HuBERT shares the same overall architecture as the pre-trained model, HuBERT introduces enhancements and modifications to improve speech representation learning. It also utilizes larger-context modeling by increasing the context window size during pre-training to capture more long-range dependencies and contextual information in the speech signal. Unlike the unidirectional transformer layers used in wav2vec2, HuBERT incorporates both unidirectional and bidirectional transformer layers. This combination enables the model to effectively capture the local and global context in the speech signal, improve the ability to understand the intricacies of speech, and extract relevant features for speech intelligibility prediction.

4) wavLM: The last self-supervised learning model to learn the representation for speech-related tasks that we consider is the wavLM [25]. It is introduced as the extension of the HuBERT framework that enables the pre-trained model to work on speech recognition and related tasks. We include the wavLM because of the ability to extract fine-grained details within an audio signal. Besides, the wavLM trained using a language modeling objective, resulting in rich representations of speech and its linguistic properties, which can be beneficial for tasks related to speech intelligibility.

In terms of feature extraction to predict speech intelligibility, the wavLM model does not explicitly extract handcrafted features. Instead, it directly operates on the raw waveform data, allowing it to capture complex patterns and dependencies present in the speech signal. The wavLM model learns to encode various aspects of speech intelligibility within its internal representations and capture acoustic features such as pitch, spectral characteristics, phonetic content, and temporal dynamics, among others. In addition, by training on a large corpus of speech data, the WavLM model learns to implicitly extract relevant features for the task at hand, making it a powerful tool for speech intelligibility prediction.

B. Regression analysis

The extracted features were fed to the base-regressor and the meta-regressor to obtain the final speech intelligibility score. While speech is indeed a complex and multidimensional signal, we use the base-regressor, which consists of linear regressor, support vector machines, and random forest, to make predictions. Although linear regression models have been successfully used in various speech-related tasks, in the case of complex listening environments and listener individuals that affect speech intelligibility, the pattern captured by single linear model cannot be adequate. Therefore, we utilize support vector machines and random forests, alongside linear regressor, to capture broader range of relationships and better handle non-linearities in the data. Finally, each of the prediction results

was fed to the Ridge-CV as the meta-regressor and ensemble model to generate the final speech intelligibility score.

IV. Experiment

A. Dataset

The dataset utilized in the experiment is associated with 6 speakers, 10 hearing aid systems from the entrants of the first Clarity Enhancement Challenge, and 27 listeners [26]. The noisy speech in the dataset was generated from the various living room scenarios of sound propagation through the room and interaction to the human head. The target sentences consist of 7-10 words in length with the subset of 1500 utterances [27]. On the other hand, the interferer was recorded from sounds of daily electronic appliances, such as washing machines, vacuum cleaners, and kettles, to demonstrate the non-impulsive noise in the real world condition. Next, to create the consistence scenarios, the position of target speaker, listeners, and interferer is adjusted. For each scene, input signals are convoluted with appropriate geometric room acoustic model and head related transfer function (HTRF) database, which include measurements for hearing aid microphones to create suitable inputs to hearing aid system.

The information about the listeners is also given in the dataset and characterized by the bilateral pure-tone audiograms. The audiogram is specified with average hearing loss in dB in frequency between 2 to 8 kHz. Based on the measurement obtained, there is 1 listener with mild hearing loss (15–35 dB), 9 listeners with moderate hearing loss (35–56 dB), and 17 listeners with severe hearing loss (>56 dB) [28]. The other hearing ability measurement is also provided, that is, the DTT, Glasgow hearing-aid benefit profile questionnaire (GHABP), and Speech, Spatial, & Qualities of Hearing questionnaire (SSQ12). In this paper, more concentration will be devoted to the analysis of audiogram and the DTT results. Since the results of the DTT are not available for several listeners, we remove those data in investigating the speech intelligibility.

There are two tracks provided in the dataset. Track 1 is a closed-set track with the known listeners and hearing aid processors in the training set (4812 responses) and test set (2421 responses). Track 2 is an open set track with one hearing aid processor and five listeners unseen in the training set. The remaining 22 listeners and 9 hearing aid processors are available in both training and test sets, generated (3545 responses). The purpose of including track 2 is to demonstrate how well the method generalizes the prediction to the unseen data. Since the DTT results are unavailable for all listeners, we remove all the data with missing values. The total number of listeners is nineteen people with the characteristics detailed in Table III. Note that due to this removal process, the total number of data for this experiment is less than those in our prior works and the existing works in the Clarity Prediction Challenge [29].

B. Evaluation Metrics

Five universal metrics used for regression task are utilized to assess the results, including Pearson correlation coefficient (ρ_p), Spearman correlation coefficient (ρ_s), root mean square error (RMSE), standard error (SE), and the coefficient of determination (R^2).

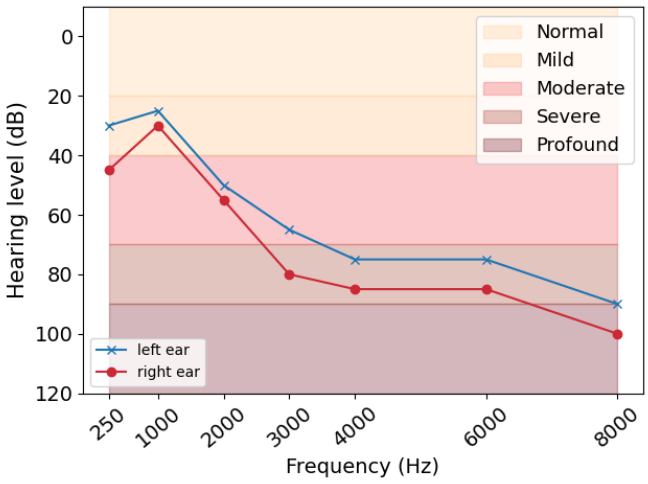


Fig. 2. The audiogram of listener L239. The hearing ability is categorized into severe hearing loss since the average hearing loss level in the audiogram with frequency in {250, 500, 1000, 2000, 3000, 4000, 6000, 8000} Hz is $>$ 56 dB. Additionally, the SRT obtained from the DTT can be categorized as high, which indicates the minimum hearing level to recognize 50% of the digit tests.

The ρ_p assesses the linear correlation between two variables. In the context of speech intelligibility prediction, ρ_p is used to determine the relationship between the actual and predicted speech intelligibility score. The ρ_s is also corporated to measure the association between the actual and predicted speech intelligibility score, even if they have a non-linear relationship. To assess the prediction accuracy, RMSE is corporate. In addition, SE helps to determine the precision of the prediction and reliability of the experiment results. Moreover, R^2 represents the goodness of fit between the predicted and actual intelligibility scores.

C. Results

In this section, we present the results of the experiment using the proposed lightweight machine learning method with several speech features for predicting speech intelligibility. Table II shows the experiment results of the proposed method compared to the baseline MBSTOI [29] in both Track 1 and Track 2. The baseline MBSTOI was developed based on the MBSTOI metric [17] with the hearing loss model developed by Nejime et al. [30]. The speech intelligibility model was built by fitting the output MBSTOI score of the training data with a logistic mapping model with a sigmoid function.

Since we aim at reducing the computational complexity, we experiment on several low dimensional acoustic features (i.e., eGeMAPS) and the embedding features in the state-of-the-art features for automatic speech recognition (ASR) (i.e., wav2vec2, HuBERT, and wavLM). The listener characteristics, such as audiogram and the DTT, were also simplified into categorical data. For audiograms, we use three categories: mild, moderate, and severe. Meanwhile, based on the SRT for the DTT results, we use two categories: low and high. These

TABLE II
EXPERIMENT RESULTS OF THE PROPOSED METHOD WITH SEVERAL ACOUSTIC FEATURES COMPARE TO THE BASELINE MBSTOI.

Method	**Feature dim.**	**Track 1 (closed-set)**					**Track 2 (open-set)**				
		$\rho_p \uparrow$	$\rho_s \uparrow$	**RMSE** $\downarrow$	**SE** $\downarrow$	$R^2 \uparrow$	$\rho_p \uparrow$	$\rho_s \uparrow$	**RMSE** $\downarrow$	**SE** $\downarrow$	$R^2 \uparrow$
Baseline MBSTOI [29]	-	0.61	0.53	28.92	0.71	0.37	0.54	0.53	35.51	1.53	0.05
Proposed method (mean value)											
GeMAPS	62	0.71	0.58	25.77	0.63	0.50	0.58	0.53	30.85	1.36	0.29
eGeMAPS	88	0.72	0.60	25.26	0.62	0.52	0.60	0.54	30.09	1.34	0.32
wav2vec2	1024	0.66	0.55	27.48	0.67	0.43	0.51	0.41	31.56	1.45	0.25
HuBERT	1024	0.70	0.59	26.08	0.64	0.49	0.55	0.47	30.76	1.40	0.29
wavLM	1024	**0.75**	**0.63**	**24.30**	**0.59**	**0.56**	**0.64**	**0.58**	**28.25**	**1.29**	**0.40**
Proposed method (concatenate)											
GeMAPS	62 x #sec.	0.71	0.59	25.68	0.63	0.50	0.60	0.53	29.59	1.34	0.34
eGeMAPS	88 x #sec.	0.72	0.60	25.49	0.62	0.51	0.60	0.53	29.67	1.34	0.34
wav2vec2	1024 x #sec.	0.63	0.52	28.39	0.69	0.39	0.50	0.41	31.72	1.45	0.25
HuBERT	1024 x #sec.	0.68	0.57	26.85	0.66	0.46	0.50	0.39	32.20	1.46	0.22
wavLM	1024 x #sec.	**0.74**	**0.62**	**24.54**	**0.60**	**0.55**	**0.59**	**0.54**	**29.51**	**1.35**	**0.35**
Proposed method (concatenate + feature selection (512))											
wav2vec2	512	0.64	0.53	28.11	0.69	0.41	0.37	0.23	34.47	1.57	0.11
HuBERT	512	0.69	0.57	26.55	0.65	0.47	0.55	0.46	30.73	1.39	0.29
wavLM	512	**0.74**	**0.62**	**24.52**	**0.60**	**0.55**	**0.60**	**0.55**	**29.40**	**1.35**	**0.35**

TABLE III
LISTENER CHARACTERISTICS FROM THE DIGIT TRIPLET TEST RESULTS AND AUDIOGRAM

Listener	L200	L201	L202	L209	L215	L216	L218	L219	L220	L222
SRT	-6.7	-8.8	-7.6	-6.1	-8.5	-6.3	-6.4	-6.1	-17.0	-8.2
DTT category	High	Low	High	High	High	High	High	High	Low	High
Hearing loss level	Severe	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Severe	Moderate	Moderate
Listener	L225	L227	L229	L231	L235	L239	L241	L242	L243	
SRT	-8.3	-6.5	-11.4	-11.0	-10.0	-6.1	-7.1	-6.9	-12.1	
DTT category	High	High	Low	Low	Low	High	High	High	Low	
Hearing loss level	Moderate	Severe	Moderate	Moderate	Mild	Severe	Moderate	Severe	Moderate	

categories were based on the calculation of the average SRT of all listeners that represents a collective level of speech intelligibility. If an individual's SRT value is significantly higher than the mean value, the individual needs a higher signal-to-noise ratio to discriminate speech accurately. The individual has poor speech intelligibility under specific hearing conditions and may have a hearing loss or hearing impairment. Conversely, an individual with a significantly lower than average SRT value indicates that the individual can discriminate speech accurately at a relatively low signal-to-noise ratio, which implies that the individual has better speech intelligibility and auditory discrimination and may have better hearing ability. Table III shows the categorization of the listener characteristics.

Table II indicates the results from the baseline MBSTOI and the proposed method with three ways to reduce the dimension of the features. For wav2vec2, HuBERT, and wavLM, we extracted the output of the intermediate layer for each second of the speech signal. For instance, if a signal has a 5-second length, then we could extract 1024 $\times$ five dimensional feature. We utilized these features in three ways by 'mean value', 'concatenate', and 'concatenate + feature selection (512)'. The 'mean value' is performed by averaging the extracted feature based on the time stamp. Hence, we only have a vector with 1024 dimensions. The 'concatenate' means that we use all the extracted features, so a vector with 1024 $\times$ number of seconds (#sec.) dimension can be obtained. Lastly, the 'concatenate + feature selection (512)' means after the concatenation, we perform a univariate feature selection by selecting the 512 highest-scoring features based on the training data.

In most cases, the results in the table II have a higher correlation, lower RMSE, and lower SE, which indicates that the proposed method could achieve better results than the baseline MBSTOI. In addition to the higher correlation and lower RMSE, the proposed method could be regarded as a blind or non-intrusive method because the reference speech signal is not required for predicting speech intelligibility. Although the GeMAPS only consists of 62 dimensions, it could be utilized to achieve higher performance compared to the baseline method. The additional 26 extended parameters also positively contribute to improving the prediction accuracy. Comparing the performance of these low dimensional features to the recent state-of-the-art features, the wavLM could achieve the highest performance almost in all evaluation metrics used in the experiment. This finding aligns well with the prior research, which also showed that the wavLM performs better than wav2vec2 and HuBERT on all downstream speech processing tasks [25].

Observing the experiment results we conducted to reduce the feature dimension, we found that averaging the vectors obtained from each speech segment could achieve a higher

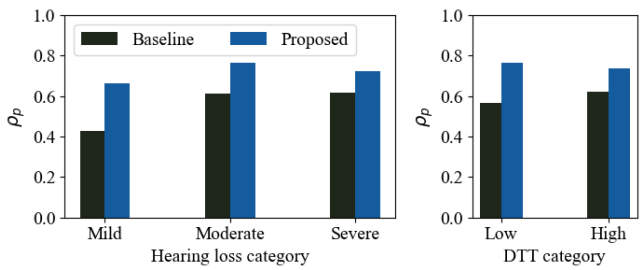


Fig. 3. Correlation analysis on the actual correctness and prediction results based on the listener characteristics: (Left) audiogram, (Right) digit triplet test (DTT)

prediction accuracy. For instance, by using the wavLM feature, the ρ_p of the proposed method (mean value) is 5% higher than the proposed method (concatenate). Similarly, the RMSE is smaller, and the R^2 is higher. The feature selection method in our experiment could not improve the prediction performance. For instance, the prediction by the feature selection using wav2vec2 and HuBERT caused a reduction in the correlation. At the same time, the results show no significant difference ($\rho > 0.05$) compared to those without feature selection.

An in-depth understanding of the hearing status allows for assessing the degree and type of hearing loss. An audiogram is a standard assessment tool that provides detailed information about hearing loss by plotting an individual's hearing levels at different frequencies and volume levels. Meanwhile, the SRT obtained from the DTT provides a direct and operational indicator of an individual's level of speech intelligibility in complex listening environments. Figure 3 shows the results grouped by the audiogram and DTT results. These results indicate that the proposed method could more accurately predict speech intelligibility in all categories of hearing loss conditions. Compared to the analysis of the results by audiogram and the DTT categories, the prediction using the DTT category shows a more stable speech intelligibility prediction for listeners with mild levels of hearing loss than the prediction grouped by the hearing loss category from audiogram.

While this research provides valuable insights into speech intelligibility prediction, we acknowledge certain limitations inherent in the study. First, the dataset used for training and evaluation may be limited and not fully represent the wide range of real-world listener conditions and noisy scene scenarios. To our knowledge, the CPC1 dataset is the only data that includes the DTT results. In the future, it is important to carry out more data collection for additional listeners on a broader range of scenes. Second, the study primarily focuses on the features used in common speech processing, such as in automatic speech recognition or emotion recognition, overlooking potential contributions from other important acoustic features or cues in speech intelligibility perception. For instance, the binaural hearing cues naturally utilized by human ears to recognize speech have yet to be considered. Future research should explore the incorporation of these cues to enhance the accuracy and robustness of speech intelligibility prediction models.

V. Conclusion

We proposed a method incorporating low-dimensional and state-of-the-art acoustic features for speech processing to predict speech intelligibility in noise for hearing aids. The prediction method was developed based on a stack regressor using various traditional machine learning regressors, including linear regressor, support vector regressor, and random forest regressor. Based on the overall results, the proposed method outperformed the previous method, particularly when utilizing wavLM features. The better performance is achieved with the help of the DTT input that is categorized based on the calculation of the average SRT of all listeners, representing speech intelligibility level. Since the proposed method does not require high computational power due to the low dimensional feature and relatively simple machine learning model, it can be considered as a baseline method before going through a further experiment using extensive machine learning models and sophisticated features for speech intelligibility prediction. In future work, we plan to further analyze the phenomena in hearing loss that both successfully and unsuccessfully restored by hearing aids and the effect on speech intelligibility perception.

Acknowledgment

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (No. 201605002), a Grant-in-Aid for Scientific Research (B) (No. 21H03463), and the Japan Society for the Promotion of Science (JSPS) KAKENHI grant (No. 22K21304).

References

[1] D. M. Jayakody, O. P. Almeida, C. P. Speelman, *et al.*, "Association between speech and high-frequency hearing loss and depression, anxiety and stress in older adults," *Maturitas*, vol. 110, pp. 86–91, 2018.

[2] T. L. G. Health, "Amplifying the global issue of hearing loss," *The Lancet. Global health*, vol. 10(10), e1360, 2022.

[3] J. Löhler, M. Cebulla, W. Shehata-Dieler, S. Volkenstein, C. Völter, and L. E. Walther, "Hearing impairment in old age," *Deutsches Ärzteblatt International*, vol. 116, no. 17, pp. 301–310, 2019.

[4] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, 2001.

[5] L. Gwilliams and M. H. Davis, "Extracting language content from speech sounds: The information theoretic approach," in *Speech perception*, vol. 74, 2022, pp. 113–139.

[6] H. Meister, S. Rählmann, M. Walger, S. Margolf-Hackl, and J. Kießling, "Hearing aid fitting in older persons with hearing impairment: The influence of cognitive function, age, and hearing loss on hearing aid benefit," *Clinical Interventions in Aging*, pp. 435–443, 2015.

[7] R. M. Cox, G. C. Alexander, and I. M. Rivera, "Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 4, pp. 904–915, 1991.

[8] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103 204, 2022.

[9] D. Byrne and N. Murray, "Predictability of the required frequency response characteristic of a hearing aid from the pure-tone-audiogram," *Ear and hearing*, vol. 7, no. 2, pp. 63–70, 1986.

[10] E. Van den Borre, S. Denys, A. van Wieringen, and J. Wouters, "The digit triplet test: A scoping review," *International journal of audiology*, vol. 60, no. 12, pp. 946–963, 2021.

[11] A. Heinrich, H. Henshaw, and M. A. Ferguson, "The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests," *Frontiers in psychology*, vol. 6, p. 782, 2015.

[12] M. S. Vlaming, R. C. MacKinnon, M. Jansen, and D. R. Moore, "Automated screening for high-frequency hearing loss," *Ear and hearing*, vol. 35, no. 6, p. 667, 2014.

[13] D. Ertmer, "Relationships between speech intelligibility and word articulation scores in children with hearing loss," *Journal of speech, language, and hearing research*, vol. 53, pp. 1075–86, 2010.

[14] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2365–2375, 2007.

[15] A. J. Bosmana and G. F. Smoorenburg, "Intelligibility of dutch cvc syllables and sentences for listeners with normal hearing and with three types of hearing impairment," *Audiology*, vol. 34, no. 5, pp. 260–284, 1995.

[16] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 259–271, 2002.

[17] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[18] A. H. Andersen, J. M. d. Haan, Z.-H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] P. Guiraud, A. H. Moore, R. R. Vos, P. A. Naylor, and M. Brookes, "Machine learning for parameter estimation in the mbstoi binaural intelligibility metric," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[20] T. H. Falk, V. Parsa, J. F. Santos, *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[22] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features," in *Interspeech 2009*, 2009, pp. 1999–2002.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[25] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] S. Graetzer, J. Barker, T. J. Cox, *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Interspeech 2021*, vol. 2, 2021, pp. 686–690.

[27] S. Graetzer, M. A. Akeroyd, J. Barker, *et al.*, "Dataset of british english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data in brief*, vol. 41, p. 107 951, 2022.

[28] T. Cox, M. Akeroyd, J. Barker, *et al.*, "Predicting speech intelligibility for people with a hearing loss: The clarity challenges," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 265, 2023, pp. 4599–4606.

[29] J. Barker, M. Akeroyd, T. J. Cox, *et al.*, "The 1st clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Interspeech 2022*, 2022, pp. 3508–3512.

[30] Y. Nejime and B. C. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, 1997.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Deep Learning-based MRI Super-Resolution Using Non-uniform Segmented Phase-Scrambling Fourier Transform Signals

Kazuki Yamato*, Shuntaro Fujisawa†, and Satoshi Ito†
* Department of Fundamental Engineering, School of Engineering, Utsunomiya University
E-mail: yamato@is.utsunomiya-u.ac.jp Tel/Fax: +81-286896277
† Graduate School of Regional Development and Creativity, Utsunomiya University
E-mail: {mc236894@s., itohst@is}.utsunomiya-u.ac.jp Tel/Fax: +81-286896277

***Abstract*—In the previous study, we proposed image acquisition and deep learning reconstruction using segmented phase-scrambling Fourier transform signals to improve the resolution of magnetic resonance images. Although the previous method achieved an effective resolution improvement for the reconstructed images, aliasing artifacts were generated in the reconstructed images because the segment width and segment skipping space of the previous segmented sampling were uniformly arranged. In this paper, we propose a new segmented signal sampling method in which the segment width and segment skipping space are non-uniformly arranged. Simulation experiments showed that the sampled signals in the proposed method were reconstructed in the deep learning reconstructions and that the qualities of the reconstructed images were compared with those of a uniform segmented signal quantitatively as well as perceptually to verify whether aliasing artifacts occurred. From the experimental results, it was confirmed that the proposed non-uniform segmented sampling could suppress aliasing artifacts while maintaining the reconstruction quality.**

I. Introduction

The magnetic resonance imaging (MRI) is an in-vivo imaging technique that can safely acquire intravital information at high resolution. However, the imaging procedure by MRI takes a long time, such as several tens of minutes. Therefore, accelerating the imaging process while maintaining high resolution is a critical issue for MRI. A super-resolution MRI method using phase-scrambling Fourier transform (PSFT) imaging was proposed by Ito et.al [1], [2]. PSFT [3], [4] is an imaging technique for MRI that applies quadratic phase modulation to the object and has useful features that are not found in Fourier transform (FT) imaging generally used in MRI, such as alias-free reconstruction [5], [6], superior image quality for compressed sensing (CS) [7]–[9]. The iterative algorithm of this super-resolution method is based on one proposed by Gerchberg [10] and estimates high-frequency signals by iterative extrapolation to a band-limited PSFT signal. This method does not sample the signal using a specific sampling pattern but instead restricts the sampling area of the signal domain. Therefore, this super-resolution method is considered to be a kind of acceleration of MRI.

Our previous study introduced a segment-wise signal sampling and skipping approach in which PSFT signals were divided into sub-bands and under-sampled in units of sub-bands to improve the super-resolution method [11]. Under-sampled signals called segmented PSFT signals could be interpolated between segments as well as extrapolated outside of segments. In addition, segmented sampling can obtain higher frequency sub-bands than band-limited sampling such as [1]. Thus, this method can be regard as CS-MRI [12], [13] using segmented sampling, and unrolling model-based deep learning reconstruction [14]–[17] can be used to improve the reconstruction quality. As a result, it was confirmed that the resolution improvement of the segmented sampling was higher than the band-limited sampling. However, aliasing artifacts were generated in the reconstructed images from segmented PSFT signals because the previous segmented sampling uniformly arranged the segment widths and skipping spaces and did not satisfy the incoherent sampling required for CS.

To solve this problem, a non-uniform segmented sampling method in which segment widths and skipping spaces are non-uniformly arranged is proposed to suppress the occurrence of aliasing artifacts in the reconstructed image. In this paper, three non-uniform segmented sampling patterns are described: a non-uniform segment width pattern, a non-uniform skipping space pattern, and non-uniform segment width and skipping space pattern. To verify the effectiveness of the proposed method, simulation experiments were conducted. From the simulation results, it was confirmed that the quality of reconstructed images without aliasing artifact was achieved by using non-uniform segmented sampling.

II. Phase-scrambling Fourier transform imaging

PSFT [3], [4] is a technique in which a quadratic field gradient $\Delta B = b(x^2 + y^2)$ is added to the pulse sequence of general-purpose FT imaging in MRI by synchronizing with the field gradient for phase encoding. The PSFT signal $v(k_x, k_y)$ is represented as follows:

$$v(k_x, k_y) = \iint_{-\infty}^{\infty} \left\{ \rho(x, y) e^{-j\bar{c}(x^2+y^2)} \right\} e^{-j(k_x x + k_y y)} dx dy, \tag{1}$$

where $\rho(x, y)$ represents the spin density distribution in the subject. $\bar{c}$ is the maximum phase modulation value. $\rho(x, y)$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

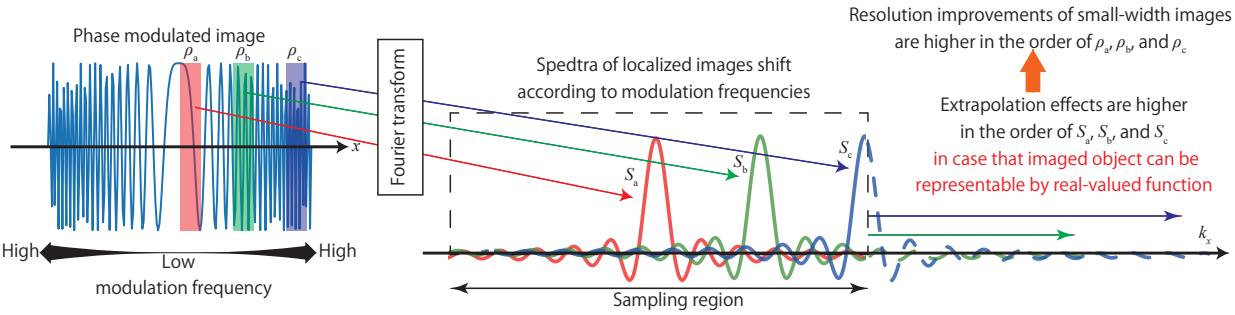


Fig. 1: Super-resolution principle of PSFT signal. ρ_a, ρ_b, and ρ_c are localized small-width images with quadratic phase modulation. The modulation frequencies are higher in the order of ρ_a, ρ_b, and ρ_c. The spectra obtained by applying FT to those small-width images are S_a, S_b, and S_c. Each spectrum is shaped symmetrically with respect to the peak, and the peak position of each spectrum shifts with modulation frequency. The spectra outside the sampling region can be estimated by extrapolation using the symmetrical feature of the spectra. The extrapolation effect is higher for larger shifts, and it is higher in the order of S_a, S_b, and S_c in this figure. Accordingly, the resolution improvements of small-width images are also higher in the order of ρ_a, ρ_b, and ρ_c.

can be obtained by taking the inverse FT (IFT) of $v(k_x, k_y)$ and multiplying the signal obtained by applying IFT by the quadratic phase term $e^{j\bar{c}(x^2+y^2)}$. Since the phase of the object shifts quadratically in the x and y directions, the modulation frequency on the image function increases as the position from the origin of the image space increases, i.e.

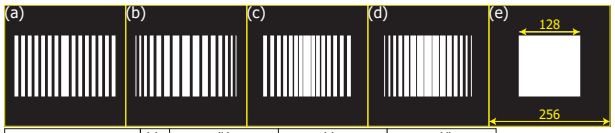
$$\rho(x)e^{-j\bar{c}(x^2+y^2)} = \rho(x)e^{-j\{\omega(x)x+\omega(y)y\}}, \tag{2}$$

where $\omega(x) = \bar{c}x$ and $\omega(y) = \bar{c}y$. $\bar{c}$ can be written as $\bar{c} = \pi/(N\Delta x^2)$ from the relation $\Delta x(\partial\omega_x(x)/\partial x)|_{x=N\Delta x/2} = \pi$. In this paper, the phase modulation value c is controlled as follows:

$$c = h\bar{c}, \tag{3}$$

where h is the phase-scrambling parameter. Note that when $h = 0$, the PSFT signal $v(k_x, k_y)$ is equivalent to the FT imaging signal because $e^{-jhc(x^2+y^2)} = 1$. Two methods have been reported so far to acquire a PSFT signal from MRI. One is the addition of coils that generate a quadratic field gradient [6], and the other is the application of frequency modulation to radio frequency pulses while applying a linear field gradient [3].

Consider the phase modulated 1D image and the localized small-width images ρ_a, ρ_b, and ρ_c on its 1D image, as shown in Fig. 1. The spectra of localized images S_a, S_b, and S_c shift in k-space in accordance with the modulation frequency $\omega(x)$, and the shape of each spectrum is symmetric with respect to its peak. $\omega(x)$ is the lowest in the center of the image and increases with increasing distance from the center. When it is assumed that the imaged object can be represented as a real-valued function, the spectrum can be extrapolated to make it symmetric with its peak, which is similar in the half Fourier encoding method [18]. Consequently, the extrapolation effects are higher in the order of S_a, S_b, and S_c, and the resolution improvements are also higher in the order of ρ_a, ρ_b, and ρ_c.



	(a)	(b)	(c)	(d)
number of segments N	15			
width of center segment w_0	16			
widths n-th segment w_n	8	$w_0 - 2n\ (1 \le n \le 7)$	8	$w_0 - 2n\ (1 \le n \le 7)$
n-th skipping spaces s_n	6	6	$n\ (1 \le n \le 7)$	$n\ (1 \le n \le 7)$

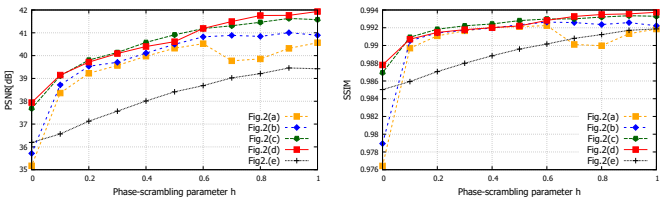
Fig. 2: Uniform and non-uniform segment-based sampling patterns. (a) is the uniform segmented pattern proposed in previous study. (b), (c), and (d) are non-uniform segmented patterns proposed in this study. (b) has non-uniform widths of segments and (c) has non-uniform skipping spaces between segments. (d) is non-uniform segmented pattern both in segment widths and skipping spaces. (e) is the band-limited pattern proposed by Ito et al. [1].

The pattern sizes are 256×256, and the number of under-sampled signals on each pattern is $128 \times 128 = 16384$ points, i.e. the sampling rate is 25%. The phase encoding direction is the horizontal direction.

III. Non-uniform segmented sampling

In the previous study [11], an image reconstruction method that uses segmented PSFT signals was proposed. This method regarded the reconstruction from segmented PSFT signal as an optimization problem of CS. To enhance the resolution improvement effect, Generic-ADMM-Net [14] unrolling alternating direction method of multipliers (ADMM) [19] was used as the deep learning based reconstruction method instead of iterative reconstruction methods. As a result, the resolution of the reconstructed image was higher than that of a band-limited PSFT signal. However, aliasing artifacts were generated in the reconstructed image in our previous study [11] when the phase-scrambling parameter h was the specified values because the segmented sampling pattern was uniform, as show in Fig. 2(a).

To solve this problem, 3 kinds of non-uniform segment-based sampling patterns, as shown in Fig. 2(b) to (d), are



(a) PSNR (b) SSIM

Fig. 3: Evaluation scores of reconstructed images by Generic-ADMM-Net. Abscissa and ordinate are the phase-scrambling parameter h for PSFT and evaluation score of each graph, respectively. Evaluation scores of $h = 0$ on the graphs indicate the results of reconstructing FT imaging signal.

TABLE I: PSNRs[dB] of reconstructed images by ISTA-Net and Generic-ADMM-Net. FT and PSFT denote reconstructions from FT imaging and PSFT signals ($h = 1.0$), respectively. The higher image reconstructed from the FT and PSFT signals under-sampled by the same pattern is underlined. For each reconstruction method, the highest PSNR is bolded.

Usage	ISTA-Net		Generic-ADMM-Net	
patterns	FT	PSFT	FT	PSFT
Fig.2(a)	31.91	31.58	35.17	40.57
Fig.2(b)	31.17	31.82	35.71	40.89
Fig.2(c)	33.13	35.17	37.68	41.57
Fig.2(d)	35.14	**35.49**	37.94	**41.93**

proposed. Those non-uniform segment patterns are determined to be non-uniform segment widths and skipping spaces under the condition that the sampling rate, number of segments, and width of the center segment will be the same as Fig. 2. Let w_0 be the width of center segment and N the number of segments. When the right and left segments are symmetrical in segment widths and skipping spaces with the center segment and the number of right segments is equal to the number of left segments, as shown in Fig. 2, the number of segments on one side $\hat{N} = \frac{N-1}{2}$ is an even number. To satisfy the above conditions, the width of n-th segment w_n and n-th skipping space s_n ($1 \leq n \leq \hat{N}$) are determined as follows:

$$w_n = w_0 - \alpha n, \tag{4}$$

$$s_n = \beta n, \tag{5}$$

where α and β are constants for determining the segment width and skipping spacing, respectively. Note that s_n is a space between the n-th segment and the $n-1$-th segment. In this study, we set $w_0 = 16$, $N = 15$, $\alpha = 2$, and $\beta = 1$.

IV. Experimental Results and Discussion

To verify the effectiveness of the proposed non-uniform segmented patterns, simulation experiments were conducted. In this simulation, PSFT signals generated from the MR images were regarded as the signals acquired from the MRI and were under-sampled by using the sampling patterns of Fig. 2. Note that $\bar{c}$ of Eq. 1 is set to 0.614 in this paper because this simulation assumes super-resolution from the under-sampled 128×128 data to 256×256 data. The used images were 570 T2-weighted MRI images of the head (of volunteers who gave consent) from the IXI Dataset [20] and 400 and 170 images were randomly selected for training and testing, respectively. The deep learning reconstruction methods were two unrolling-based deep learning methods, namely, Generic-ADMM-Net [14] and ISTA (Iterative Shrinkage-Thresholding Algorithm)-Net [16]. For training parameters of the two methods, the input size, the number of epochs, the activation function, and the optimizer were set to 256×256, 300, Rectified Linear Unit (ReLU), and Adam [21], respectively. In evaluating the quality of the reconstructed images, the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) [22] were used. The simulation was conducted on a computer with an Intel® Core™ i9-11900K CPU (3.5 GHz), 64 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU running Windows 10. The software used was Python 3.7.9, PyTorch 1.11.0, and CUDA ToolKit 11.3. The training times under the above conditions were about 45 minutes on Generic-ADMM-Net and about 2 hours on ISTA-Net.

As shown in Fig. 3, PSNRs and SSIMs of the reconstructed images by using the proposed segmented patterns shown in Fig. 2(b)-(d) are higher than the uniform segmented patterns of Fig. 2(a) in any h. Comparing Fig. 2(b) and (c), (c) has higher scores than (b). Therefore, the reconstruction is improved with non-uniform skipping spaces rather than widths. Furthermore, it is effective to make both widths and skipping spaces non-uniform.

Table I shows the PSNRs using another deep learning method, i.e. ISTA-Net. Even if ISTA-Net is used, the reconstructed images from the under-sampled PSFT signal is also better than under-sampled FT signal except for Fig 2(a). It could be confirmed that non-uniform skipping spaces produced a more effective reconstruction than non-uniform segment widths, and the reconstructed image from the under-sampled PSFT signal by Fig. 2(d) had the highest PSNR. Hence, the non-uniform segmented patterns proposed in this paper can be effective regardless of the unrolling-based deep learning methods.

Fig. 4 shows the reconstructed images and their error maps. The reconstructed images of Fig. 4(c)-(f) had better qualities than the zero-filling image of (b). However, there was a smoothing effect in low luminance regions compared to the full-data, as shown in the enlarged views of Fig. 4(c) and (e). In contrast, the reconstructed image in Fig. 4(d) and (f) does not show smoothing effects like (c) and (e) and are similar to the full-data. In comparison to the error maps in Fig. 4(g)-(k), the error maps of (i) and (k) have smaller errors than (g), (h), and (j). Comparing error maps of Fig. 4(i) and (k), the errors in the central region of (i) are comparable to those of (k), whereas the errors in the outside region of (i) are high. Meanwhile, it is clear that the errors in the outside region of Fig. 4(k) are smaller than other error maps. In other words, the non-

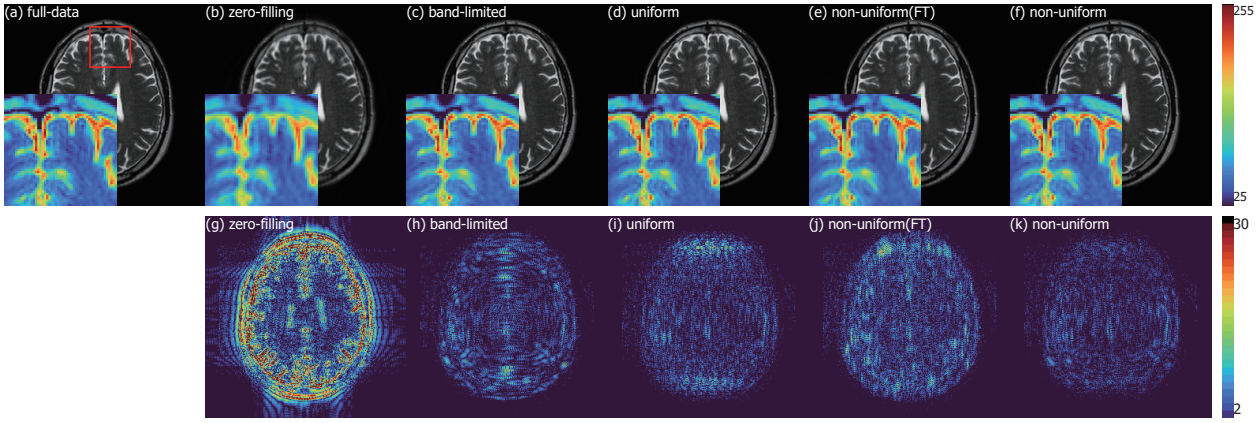


Fig. 4: Reconstructed images by Generic-ADMM-Net and their error maps. Upper row images show full-size images (256×256 pixels) and enlarged views of areas indicated by a red rectangle. Bottom row images show error maps of (b)-(f). (a) shows full-data image. (b) and (c) are zero-filling and reconstructed images from a band-limited PSFT signal, respectively. (d) shows a reconstructed image from under-sampled PSFT signal by the uniform segmented pattern shown in Fig. 2(a). (e) and (f) show reconstructed images from under-sampled FT and PSFT signals by the non-uniform segmented pattern of Fig. 2(d), respectively. h of (b), (c), (d), and (f) is set to 1.0. In (g)-(k), pixels of each error map show the absolute difference between pixels at the same location in the full-data and reconstructed images.

uniform segmented pattern could improve the resolution of the outside region of the reconstructed image without sacrificing the resolution of the central region of the reconstructed image.

Fig. 5 shows profiles of reconstructed PSFT signals and their errors. Comparing the full-data and each reconstructed signal, both Fig. 5(d) and (f) are almost reconstructed to the same profile as the full-data, and it is possible to estimate the unsampled signals from the under-sampled signal shown in Fig. 5(b) and (c). In contrast, the reconstructed signal using the uniform segmented pattern of Fig. 5(d) has large errors at regular intervals, as shown in Fig. 5(e), with particularly large errors in the low-frequency bands such as when the signal index is between 9 and 14. These errors which are regularly generated correspond to the part of the signals not sampled, as shown in Fig. 5(b). The reasons for these errors are considered. The under-sampled signals using the uniform segmented pattern have high coherence, thus introducing aliasing artifacts. In this study, no visible artifacts were observed in the reconstructed images from uniform segmented under-sampled signals by using the deep learning reconstruction method. However, the reason for the errors shown in Fig. 5(e) in the reconstructed signals is considered to be that the aliasing artifacts could not be completely removed even with the deep learning reconstruction method.

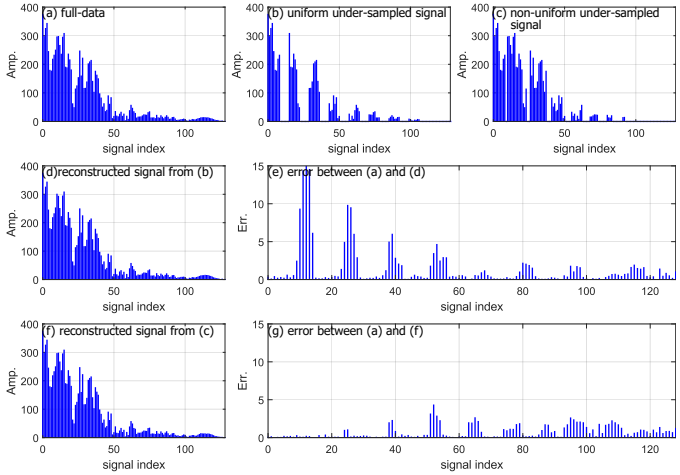


Fig. 5: PSFT signals of reconstruct images. Each profile shows the 127-th row of PSFT signal spectra calculated from Fig. 4(a), (d), and (f). (a) shows full-data signal and (b) and (c) show under-sampled signals using uniform and non-uniform patterns, respectively.(d) and (f) are the reconstructed signals from (b) and (c), respectively. (e) and (g) are errors between full-data and each reconstructed signal.

On the other hand, the errors of the reconstructed signals using the non-uniform segmented pattern shown in Fig. 5(g) were less than 5. In particular, the errors in the low frequency band with a signal index of 20 or less were very small. This is because the non-uniform segmented pattern samples the low-frequency band more densely than uniform segmented pattern. In addition, since the under-sampled signal using the non-uniform segmented pattern is incoherent, the unrolling model-based deep learning reconstruction such as Generic-ADMM-Net seems to perform effectively. However, the errors in the high-frequency band of Fig. 5(g) are slightly larger than those of (e) because the sampling rate in the high-frequency band of the non-uniform segmented pattern is less than that of the uniform segmented pattern. To solve this problem, it is necessary to construct a pattern that can increase the amount of under-sampled signals from the high-frequency band while maintaining the non-uniformity of the segment width and skipping spacing.

V. Conclusion

In this paper, we proposed a segmented sampling pattern in which both the segment widths and skipping spaces were non-uniform to suppress the occurrence of aliasing artifacts and improve the image reconstruction by deep learning reconstruction methods. From the experimental results, it was confirmed that the reconstructed images from under-sampled PSFT signals according to the proposed non-uniform segmented pattern had higher quality than those of the previous uniform segmented pattern regardless of the specific deep learning reconstruction method. Future work will include a validation of the effectiveness of the proposed pattern for complex images including a phase component.

VI. acknowlegements

The authors thank the Information eXtraction from Images project for creating the IXI Dataset.

References

[1] S. Ito and Y. Yamada, "Improvement of spatial resolution in magnetic resonance imaging using quadratic phase modulation," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2497–2500. DOI: 10.1109/ICIP.2009.5413931.

[2] S. Ito and Y. Yamada, "Compressed sensing with super-resolution in magnetic resonance using quadratic phase modulation," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–4.

[3] A. Maudsley, "Dynamic range improvement in nmr imaging using phase scrambling," *Journal of Magnetic Resonance (1969)*, vol. 76, no. 2, pp. 287–305, 1988, ISSN: 0022-2364. DOI: https://doi.org/10.1016/0022-2364(88)90111-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022236488901114.

[4] V. J. Wedeen *et al.*, "Dynamic range compression in mri by means of a nonlinear gradient pulse," *Magnetic Resonance in Medicine*, vol. 6, no. 3, pp. 287–295, 1988. DOI: https://doi.org/10.1002/mrm.1910060306. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.1910060306. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910060306.

[5] S. Ito and Y. Yamada, "Alias-free image reconstruction using fresnel transform in the phase-scrambling fourier imaging technique," *Magnetic Resonance in Medicine*, vol. 60, no. 2, pp. 422–430, 2008. DOI: https://doi.org/10.1002/mrm.21672. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.21672. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21672.

[6] M. Zaitsev, G. Schultz, and J. Hennig, "Extended anti-aliasing reconstruction for phase-scrambled mri with quadratic phase modulation," in *Proc. ISMRM*, 2009, p. 2859.

[7] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. DOI: 10.1109/TIT.2006.871582.

[8] G. Puy *et al.*, "Spread spectrum magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 31, no. 3, pp. 586–598, 2012. DOI: 10.1109/TMI.2011.2173698.

[9] Y. Wiaux *et al.*, "Spread spectrum for compressed sensing techniques in magnetic resonance imaging," in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010, pp. 756–759. DOI: 10.1109/ISBI.2010.5490066.

[10] R. Gerchberg, "Super-resolution through error energy reduction," *Optica Acta: International Journal of Optics*, vol. 21, no. 9, pp. 709–720, 1974. DOI: 10.1080/713818946. eprint: https://doi.org/10.1080/713818946. [Online]. Available: https://doi.org/10.1080/713818946.

[11] K. Yamato and S. Ito, "Super-resolution magnetic resonance imaging using segmented signals in phase-scrambling fourier transform imaging and deep learning," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2551–2555. DOI: 10.1109/ICIP46576.2022.9897337.

[12] M. Lustig *et al.*, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007. DOI: https://doi.org/10.1002/mrm.21391. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.21391. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21391.

[13] M. Lustig *et al.*, "Compressed sensing mri," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008. DOI: 10.1109/MSP.2007.914728.

[14] Y. Yang *et al.*, "Admm-csnet: A deep learning approach for image compressive sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 521–538, 2020. DOI: 10.1109/TPAMI.2018.2883941.

[15] K. Hammernik *et al.*, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, 2018. DOI: https://doi.org/10.1002/mrm.26977. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.26977. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.26977.

[16] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[17] H. K. Aggarwal *et al.*, "Modl: Model-based deep learning architecture for inverse problems," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 394–405, 2019. DOI: 10.1109/TMI.2018.2865356.

[18] J. Finsterbusch and J. Frahm, "Half-fourier single-shot steam mri," *Magnetic Resonance in Medicine*, vol. 47, no. 3, pp. 611–615, 2002. DOI: https://doi.org/10.1002/mrm.10090. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.10090. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.10090.

[19] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976, ISSN: 0898-1221. DOI: https://doi.org/10.1016/0898-1221(76)90003-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0898122176900031.

[20] *IXI dataset*, https://brain-development.org/ixi-dataset/(Accessed: Jan.31, 2023).

[21] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. DOI: 10.48550/ARXIV.1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980.

[22] Z. Wang *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861.

An Extreme Gradient Boosting-based Prediction for Depression

Ahmed Ibrahum*, Kwang Ho Park*, Jang-Eui Hong*, Van-Huy Pham†, and Keun Ho Ryu†
* Chungbuk National University, Republic of Korea
Tel/Fax: +82-43-267-2254 / +82-43-275-2254
† Ton Duc Thang University, Vietnam
E-mail: {Ahmed_ibrahum, khblack, jehong}@chungbuk.ac.kr, {phamvanhuy, khryu}@tdtu.edu.vn

***Abstract* — Depression, a pervasive mental health ailment impacting millions globally, necessitates precise and intelligible predictive methodologies. Physicians encounter formidable hurdles when screening large populations. Despite a plethora of proposed Artificial Intelligence (AI) techniques for depression prediction, creating transparent models remains a formidable challenge. In this investigation, we confront this challenge by constructing prediction models grounded in diverse machine-learning algorithms and employing the Hybrid Feature Selection (HFS) approach, fusing statistical hypothesis tests and LASSO. This process facilitates the identification of pivotal features enhancing the predictive model's efficacy. Consequently, the chosen features serve as inputs for our predictive model, enabling us to harness their predictive potential and delve into their connection with the depression target variable. The ensemble of models encompassed Extreme Gradient Boosting (XGBoost), Light Gradient Boosting, Gradient Boosting, CatBoost, Ada Boost, Linear Discriminant Analysis, Ridge, Random Forest, Extra Trees, Logistic Regression, Decision Tree, K-Neighbors, Naive Bayes Gaussian, and Naive Bayes Bernoulli. These models underwent rigorous evaluation utilizing the Korea National Health and Nutrition Examination Survey (KNHANES) datasets. Our findings reveal that the XGBoost model outperformed its counterparts, achieving the highest accuracy (0.871), AUC (0.927), and F1 Score (0.673). Furthermore, our model sheds light on the intricate relationship between input variables and the predicted outcome, quantifying the influence of each input variable on the result.**

Keywords— machine-learning, Depression, Hybrid Feature Selection, LASSO, XGBoost

I. INTRODUCTION

Depression is a debilitating mental health disorder that affects millions of individuals worldwide. There is a wide range of physical and cognitive symptoms associated with depression, including persistent feelings of sadness and loss of interest or pleasure in activities. Depression can have a significant impact on an individual's daily functioning, quality of life, and overall well-being. Understanding the prevalence, causes, and consequences of depression is crucial for effective prevention, early intervention, and treatment strategies. Depression is a widespread global health issue, with substantial prevalence rates across different populations. According to the World Health Organization (WHO), an estimated 264 million people of all ages suffer from depression worldwide [1]. This staggering figure highlights the magnitude of the problem and the urgent need for effective interventions. Anal the impact the epidemiological data further reveals important insights into the distribution and impact of depression. It is worth noting that depression is more commonly diagnosed in women than in men [2]. Additionally, depression often manifests in early adulthood, with the median age of onset being around 32 years [3]. The disorder can persist chronically or recur episodically throughout an individual's lifespan, leading to long-term impairment and reduced productivity. Research on depression has provided valuable insights in to its multifactorial etiology. The disorder is believed to result from a complex interplay of genetic, biological, environmental, and psychosocial factors. Genetic studies have identified certain genetic variations associated with an increased risk of developing depression [4]. Furthermore, imbalances in neurotransmitters such as serotonin, norepinephrine, and dopamine have been implicated in the pathophysiology of depression [5]. Psychosocial factors, such as stressful life events, childhood trauma, and chronic psychosocial stress, can also contribute to the development and exacerbation of depression [6]. Socioeconomic factors, including low income, unemployment, and limited access to mental healthcare, further compound the burden of depression [7]. In terms of treatment, a range of interventions is available, including psychotherapy, pharmacotherapy, and lifestyle modifications. However, the effectiveness of these treatments can vary depending on individual characteristics and the severity of the depressive symptoms [8]. The experimental results show that the proposed XGBoost model predictive performance while also providing interpretability between input and output. Our model was evaluated against various high-performing machine learning benchmarks including Extreme Gradient Boosting (XGBoost), Light Gradient Boosting, Gradient Boosting, CatBoost, Ada Boost, Linear Discriminant Analysis, Ridge, Random Forest, Extra Trees, Logistic Regression, Decision T

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

ree, K-Neighbors, Naive Bayes Gaussian, and Naive Bayes Bernoulli.

The main contributions of this work are the proposal of a framework that combines Extreme Gradient Boosting (XGBoost) with Hybrid Feature Selection (HFS) to explore the input-target relationship for depression prediction. Our model demonstrated superior predictive performance, outperforming all evaluated models, with an accuracy of approximately 0.892, AUC of 0.9268, and F1 Score of 0.6729 on the KNHANES datasets.

The findings presented in this study carry significant implications, especially concerning public health outcomes associated with depression. This paper is structured as follows: Section 1 introduces depression, Section 2 discusses related work, Section 3 presents the materials and methods employed, providing comprehensive details on the dataset and overall experimental results, including the configuration and construction process. Section 4 elaborates on the procedure, and Section 5 concludes the paper.

II. RELATED WORK

Artificial Intelligence (AI) has long been proposed as a valuable tool for clinical decision-making in healthcare. In recent years, the use of AI in the healthcare industry has been on the rise, particularly in the development of predictive models for diagnosing and forecasting depression illnesses and diseases. One important area of investigation is the identification of factors associated with depression. Several studies have highlighted the role of genetic factors in susceptibility to depression. For instance, a study by Sullivan et al. (2018) identified specific genetic variations that contribute to an increased risk of developing depression [9]. These findings have shed light on the biological underpinnings of the disorder and have the potential to inform personalized treatment approaches. Psychotherapy, including Cognitive-Behavioral Therapy (CBT) and InterPersonal Therapy (IPT), has also demonstrated effectiveness in the treatment of depression [10]. These therapeutic approaches focus on identifying and modifying maladaptive thought patterns and improving interpersonal relationships, respectively. Additionally, alternative, and complementary therapies, such as mindfulness-based interventions and exercise, have shown promise as adjunctive treatments for depression [11]. Treatment optimization has also been an area of interest. AI and ML algorithms have been used to personalize treatment plans by analyzing patient characteristics, treatment history, and response data. These models can assist in selecting the most suitable intervention, adjusting medication dosages, and optimizing therapy regimens to improve treatment outcomes [12,13]. Additionally, Natural Language Processing (NLP) techniques have been employed to analyze text data, such as social media posts, online forums, and electronic medical records, to gain insights into individuals' experiences with depression and inform mental health interventions [14,15]. By extracting sentiment, linguistic patterns, and thematic information from text, researchers have explored novel ways of understanding the lived experiences of individuals with depression and tailoring interventions accordingly.

III. MATERIALS AND METHODS

The procedures are applied to meet the objectives of this study into three phases (Figure 1). The first phase is concerned with the observation and collection of data and generating training and testing patterns to build features. The second phase is about designing building the prediction model. The last phase includes the predictive performance evaluation measurements and comparison between other benchmark machine learning models.

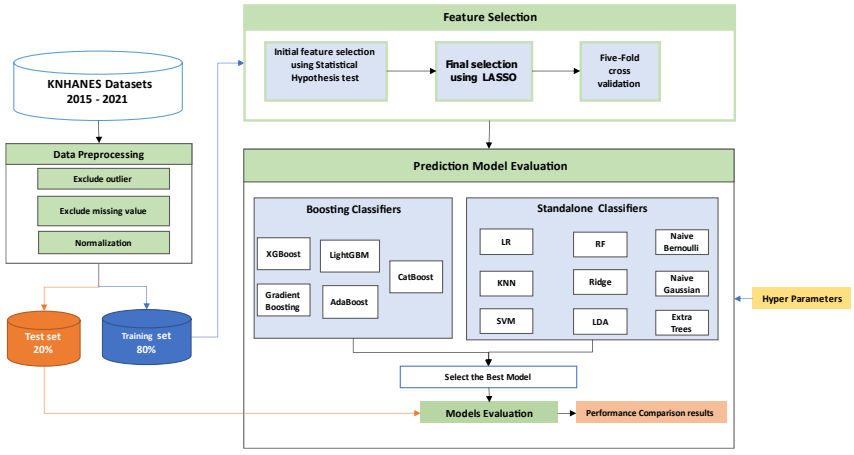


Fig. 1 Experimental design of comparative for hypertension diseases prediction.

A. Data and Preprocessing

The Korea National Health and Nutrition Examination Surv ey (KNHANES) dataset is administered by the Korea Centers for Disease Control and Prevention (KCDC), which publish th e Korea health statistics annually, and microdata are publicly a vailable through the KNHANES website (http://knhanes.cdc.g o.kr) [16]. The dataset includes a broad range of health examin ations of various numbers of diseases, health interviews, nutrit ion surveys, demographic, socioeconomic, dietary, and health- conditions, totaling over 361 features. In this study, we distrib uted across all features from a total of 51,570 individuals' data collected between 2015 and 2021. However, the dataset holds several missing values and outliers that demand attention. Stat istical power is reduced by missing values, and parameters are inaccurately estimated as a result. To produce complete data f or analysis, there are several methods to deal with missing dat a. To prevent model complications, we excluded all the missin g values and outliers. Based on the interquartile range, outliers were removed from the data. apply Hybrid Feature Selection (HFS) to selecting significant features, Furthermore, these sele cted features serve as inputs for the predictive model. To ident ify the most relevant features, we applied the Hybrid Feature S election (HFS) method, which combines statistical hypothesis tests and the least absolute shrinkage and selection operator (L ASSO). This process helped us select significant features that contribute to the predictive model's performance. Subsequentl y, the selected features were utilized as inputs for our predictiv e model, allowing us to leverage their predictive power and ex plore the relationship between these features and the target var iable of depression.

B. Classification Algorithms

The experimental results evaluate various high-performing machine learning benchmarks, including Extreme Gradient Bo osting (XGBoost), Light Gradient Boosting, Gradient Boostin g, CatBoost, Ada Boost, Linear Discriminant Analysis, Ridge, Random Forest, Extra Trees, Logistic Regression, Decision Tr ee, K-Neighbors, Naive Bayes Gaussian, and Naive Bayes Ber noulli. Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm that combines the advantages of gradient boosting and random forests. It iteratively trains an ensemble of weak decision tree models, where each subsequent model corrects the errors made by the previous ones. XGBoost is known for its scalability, speed, and ability to handle large and complex datasets [17]. Light Gradient Boosting Machine (LightGBM) is another gradient boosting framework that aims to achieve high efficiency and low memory usage. It utilizes a novel tree-building algorithm called Gradient-based One-Side Sampling (GOSS), which selects the most informative samples for growing trees. LightGBM is particularly suitable for large-scale and high-dimensional problems [18]. Gradient Boosting is a general machine learning algorithm that builds an ensemble of weak models in a sequential manner. Each model focuses on correcting the mistakes made by the previous ones. By combining multiple weak models, gradient boosting can create a strong predictive model. It is widely used in various domains and has proven to be effective in many applications [19]. CatBoost is a gradient boosting algorithm that is specifically designed to handle categorical features effectively. It uses a novel approach called ordered boosting, which combines a set of decision trees in a particular order to exploit the inherent order in categorical features. CatBoost is known for its robustness, accuracy, and ability to handle high-cardinality categorical variables [20]. Adaptive Boosting (AdaBoost) is an ensemble learning algorithm that iteratively combines multiple weak models to create a strong model. It assigns weights to the training instances and adjusts them based on the performance of the previous models. AdaBoost focuses on misclassified instances, making it particularly suitable for tasks with imbalanced classes [21]. Linear Discriminant Analysis (LDA) is a statistical technique used for dimensionality reduction and classification. It seeks to find a linear combination of features that maximally separates different classes while minimizing the within-class scatter. LDA assumes that the data follows a Gaussian distribution and that the classes have equal covariance matrices. It is widely used for pattern recognition and feature extraction [22]. Ridge regression is a linear regression technique that incorporates a penalty term to regularize the model. It adds a term proportional to the square of the coefficients to the loss function, which encourages smaller and more balanced coefficient values. Ridge regression helps prevent overfitting and can handle multicollinearity in the input features [23]. Random Forest is an ensemble learning method that constructs a multitude of decision trees and combines their predictions through voting or averaging. Each tree is trained on a random subset of the training data, and the final prediction is determined by aggregating the individual tree predictions. Random Forests are robust against overfitting, can handle high-dimensional data, and provide feature importance measures [24]. Extra Trees (Extremely Randomized Trees) is another ensemble learning method similar to Random Forests. It builds multiple decision trees using random subsets of features and random thresholds for splitting nodes. Extra Trees

further randomizes the tree construction process by selecting the splitting thresholds at random, without optimization. This randomness often leads to faster training times compared to Random Forests [25]. Logistic Regression is a linear classification algorithm used to model the probability of binary or multiclass outcomes. It uses the logistic function to map the linear combination of features to a probability value. Logistic Regression is widely used in various fields, particularly in cases where interpretability and simplicity are desired [26]. Decision Trees are non-parametric machine learning algorithms that create a flowchart-like model of decisions and their possible consequences. Each internal node represents a feature or attribute, and each leaf node represents a class label or a decision outcome. Decision Trees are interpretable, handle both numerical and categorical data, and can capture non-linear relationships [27]. K Nearest Neighbors (KNN) is a non-parametric classification algorithm that classifies new instances based on their similarity to known instances in the training data. It assigns the class label based on the majority class among its K-nearest neighbors in the feature space. KNN is simple, flexible, and often used for pattern recognition and recommendation systems [28]. Naive Bayes Gaussian is a variant of the Naive Bayes algorithm that assumes the continuous features in the dataset follow a Gaussian (normal) distribution. It uses the Gaussian probability density function to estimate the likelihood of each feature value given in the class. Naive Bayes Gaussian is particularly useful when dealing with continuous data and can handle high-dimensional feature spaces efficiently [29]. Naive Bayes Bernoulli is another variant of the Naive Bayes algorithm that assumes the features are binary or follow a Bernoulli distribution (taking values of 0 or 1). It models the likelihood of each feature value being 1 or 0 independently for each class. Naive Bayes Bernoulli is commonly used in text classification tasks, where binary features represent the presence or absence of certain words in documents [30].

C. Experimental Setup

To optimize the hyper-parameters of various machine learning classifiers, a random search approach was employed using 5-fold cross-validation. The resulting optimized hyper-parameter settings for the baseline classifiers are summarized in Table 1.

D. Evaluation Metrics

For evaluating the NCDs prediction model were trained and evaluated using five-fold cross-validation in this study. The performance of different algorithms for the testing dataset is

Table 1. Searching Space of Hyper-parameters

Model	Parameters	Search Space
Random Forest	adept	(2, 8)
	min_samples_split	(1, 8)
	min_samples_leaf	(1, 8]
	criterion	{'gini,' 'entropy'}
	bootstrap	{True, False}
AdaBoost	learning_rate	(0.1, 1)
	algorithm	{'SAMME.R,' 'SAMME'}
XGBoost	min_child_weight	(1, 10)
	gamma	{0, 0.1, 0.5, 0.8, 1}
	subsample	{0.5, 0.75, 0.9}
	colsample_bytree	{0.5, 0.6, 0.7, 0.8, 0.9, 1}
	max_depth	{2, 8}
	learning_rate	{0.01, 0.1, 0.2, 0.3, 0.5}
LightGBM	min_child_samples	(10, 60)
	reg_alpha	{0, 0.1, 0.5, 0.8, 1}
	subsample	{0.5, 0.75, 0.9}
	colsample_bytree	{0.5, 0.6, 0.7, 0.8, 0.9, 1}
	max_depth	(2, 8)
	learning_rate	{0.01, 0.1, 0.2, 0.3, 0.5}
CatBoost	min_child_samples	(10, 60)
	subsample	{0.5, 0.75, 0.9}
	colsample_bytree	{0.5, 0.6, 0.7, 0.8, 0.9, 1}
	max_depth	(2, 8)
	learning_rate	{0.01, 0.1, 0.2, 0.3, 0.5}

evaluated after the completion of the training phase and was compared using the following performances the confusion matrix for a binary classification problem. As a consideration of this provided information of confusion matrix, performance metric such as accuracy, precision, recall, Specificity, F-measure, and ROC area [31] and [32] which are computed by Eq.1-4. Accuracy is a measure that indicates the proportion of correct results, which includes both true positive and true negative results, in a selected population.

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad (1)$$

Sensitivity is a measure that represents the probability of a diagnostic test correctly identifying patients who have the disease.

$$Sensitivity\ (Recall) = \frac{TP}{(TP + FN)} \quad (2)$$

Specificity is a measure that represents the probability of a diagnostic test correctly identifying individuals who do not have the disease without giving false-positive results.

$$Specificity = \frac{TN}{(TN + FP)} \quad (3)$$

Precision represents positive predictive value.

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

ROC graphs are two-dimensional graphs in which TP rate is plotted on the Y axis and FP rate is plotted on the X axis. An ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives) [32].

IV. EXPERIMENTAL RESULT AND ANALYSIS

As illustrated in Figure 1, the proposed experimental design involves three distinct components: (I) data preprocessing (II) features selection and cross validation (III) comparison of machine learning classifiers for the prediction models of Depression. To demonstrate the effectiveness of our proposed model, we compare it to high-performing machine learning benchmarks. We evaluate the model's performance on the test set using three key metrics: area under the curve (AUC) and accuracy and F1 score.

A. The Results of Hybrid Feature Selection

The initial stage of the Hybrid Feature Selection (HFS) process involves utilizing t-test statistical hypothesis tests to determine the null or alternative hypothesis, thereby selecting significant features. Moving on to the second stage of HFS, the least absolute shrinkage and selection operator (LASSO) plays a crucial role in enhancing the predictive capability of models by eliminating irrelevant features that hold no relevance to the target classes. Consequently, these features were eliminated by assigning them to a coefficient of zero. By employing the second stage HFS method, we were able to select a representative set of 32 features from the KNHANES dataset. Furthermore, these selected features serve as inputs for the predictive model.

B. Comparison Results

In this section, the efficient An Extreme Gradient Boosting-based Approach (XGBoost) framework for Depression Disease Prediction is compared with baseline models including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes Gaussian, Naive Bayes Bernoulli, Random Forest, Gradient Boosting, AdaBoost, and LightGBM. The performance of these models is evaluated using the KNHANES dataset and the evaluation metrics used are classification accuracy and area under the Receiver Operating Characteristic Curve (AUC). The data is split into 80%for the training set and 20% for the evaluation set. A 5-fold stratified cross-validation procedure is applied to the training set to prevent overfitting. In this procedure, the dataset is randomly partitioned into five parts, and four folds of the training sets are used to train classification models while the remaining fold is used to validate the model. Table 2 shows the performance of all predictive models in the KNHANES dataset. The XGBoost model achieved the highest accuracy of 0.871 AUC 0.927, and F1 Score 0.673, Naive Bayes Gaussian had lower accuracy scores but still performed reasonably well with accuracy.

Table 2. Performance Comparison of Machine Learning Models

Model	Accuracy	AUC	F1
Extreme Gradient Boosting	**0.871**	**0.927**	**0.673**
LightGBM	0.869	0.925	0.668
Gradient Boosting	0.866	0.923	0.655
CatBoost	0.862	0.918	0.649
Ada Boost	0.861	0.918	0.642
Linear Discriminant Analysis	0.859	0.913	0.649
Ridge	0.858	0.661	0.619
Random Forest	0.855	0.913	0.609
Extra Trees	0.852	0.907	0.599
Logistic Regression	0.838	0.885	0.551
Decision Tree	0.808	0.711	0.541
K-Neighbors	0.784	0.705	0.550
Naive Bayes Gaussian	0.762	0.671	0.424
Naive Bayes Bernoulli	0.776	0.509	0.344

C. Model Interpretability

In this section, we will delve into the interpretation of the XGBoost model employed in our study. Our model enables us to determine the impact of each variable on the model's output, as illustrated in Figure 2. For instance, if age increases, the probability of depression also tends to increase. The features are ordered based on their effects on the prediction, from the highest to the lowest. To ensure interpretability of the XGBoost model in predicting depression among the Korean population, the features were sorted in ascending order of their importance scores. Figure 3 showcases the most useful features identified for depression prediction in the Korean population, including age, age of onset of smoking (BS2_1), frequency of female bullying (BD2_32), body mass index (HE_BMI), and gender. Furthermore, we explored the relationship between these features and the target variable of depression. It is evident that age and gender have a significant association with the target variable. Specifically, being an older male is associated with higher values of the target variable, indicating a higher likelihood of depression. These findings provide valuable insights into the factors that contribute to depression prediction in the Korean population. The identified features shed light on the demographic, behavioral, and physical factors that play a crucial role in understanding and predicting depression.

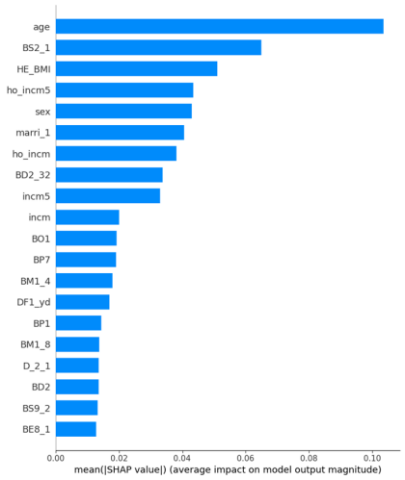


Fig. 2 Illustrates the 20 features that significantly contribute to the models.

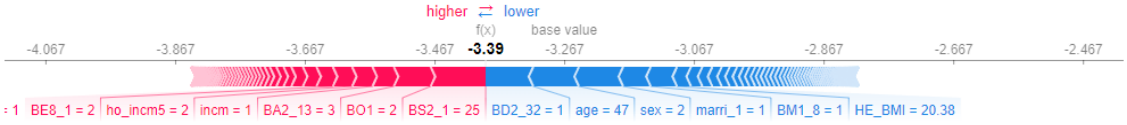


Fig. 3 Features and influence on predicted label for depression.

CONCLUSION

In conclusion, depression presents itself as a significant global public health concern and a primary contributor to suicide rates. Despite the introduction of artificial intelligence techniques for predicting depression risk, the integration of interpretable models remains a formidable hurdle. Our study has illuminated the XGBoost model's capacity to achieve superior predictive performance in comparison to numerous models using the expansive KNHANES dataset, while also offering partial interpretability between input and output variables. Nonetheless, working with the extensive KNHANES datasets required employing a hybrid feature selection (HFS) approach to pinpoint vital features, limited it may still fall short in comprehensively unraveling the intricate nature of depression. Depression's roots are entwined with numerous factors, including genetics, environment, and cultural influences, which these datasets may not fully encompass. Consequently, the model's applicability across diverse populations and contexts may be limited. Furthermore, there exists an opportunity to augment the model's interpretability to delve deeper into the relationship between input and output variables. The proposed XGBoost model harbors the potential to refine the accuracy of depression diagnosis and contribute to a more profound comprehension of this vital domain in public health. Nevertheless, it remains crucial to acknowledge these limitations as we persist in advancing our understanding and management of depression.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (RS-2023-00237203).

REFERENCES

[1] World Health Organization (WHO). (2017). Depression and Oth er Common Mental Disorders: Global Health Estimates.

[2] Kessler, R. C., et al. (2005). Prevalence, severity, and comorbidi ty of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6), 617-627.

[3] Kessler, R. C., et al. (2005). Age of onset of mental disorders: a review of recent literature. Current Opinion in Psychiatry, 20(4), 359-364.

[4] Sullivan, P. F., et al. (2018). Psychiatric genomics: an update an d an agenda. American Journal of Psychiatry, 175(1), 15-27.

[5] Belmaker, R. H., & Agam, G. (2008). Major depressive disorder. New England Journal of Medicine, 358(1), 55-68.

[6] Kendler, K. S., et al. (2003). Life events and depressive sympto ms: evidence for shared genetic etiology. Psychological Medicin

e, 33(06), 1131-1141.

[7] Lorant, V., et al. (2003). Socioeconomic inequalities in depressi on: a meta-analysis. American Journal of Epidemiology, 157(2), 98-112.

[8] Sullivan, P. F., et al, Psychiatric genomics: an update and an age nda. American Journal of Psychiatry, 175(1), 15-27, 2018.

[9] Cuijpers, P., et al. (2013). Psychological treatment of depression: a meta-analysis. World Psychiatry, 12(3), 225-230.

[10] Cooney, G. M., et al, Exercise for depression. Cochrane Databas e of Systematic Reviews, 9(9), 2013.

[11] Dwyer, D. B., et al, Machine learning approaches to depression: A survey. IEEE Transactions on Biomedical Engineering, 66(3), 685-698, 2018.

[12] Bzdok, D., & Meyer-Lindenberg, Machine learning for precision psychiatry: Opportunities and challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3(3), 223-230, 201 8.

[13] Guntuku, S. C., et al, Detecting depression and mental illness on social media: An integrative review. Current Opinion in Behavi oral Sciences, 18, 43-49, 2017.

[14] Coppersmith, G., et al Quantifying mental health signals in Twit ter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology, 51-60, 2014.

[15] Korea Centers for Disease Control & Prevention. [Online]. Avai lable: http://knhanes.cdc.go.kr. Accessed on: Jan. 2, 2023.

[16] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

[17] Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting Sy stem. In Proceedings of the 22nd ACM SIGKDD International C onference on Knowledge Discovery and Data Mining (pp. 785-7 94), 2016.

[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Y e, Q. LightGBM: A Highly Efficient Gradient Boosting Decisio n Tree. In Advances in Neural Information Processing Systems (pp. 3146-3154), 2017.

[19] Friedman, J. H. Greedy function approximation: A gradient boos ting machine. Annals of Statistics, 29(5), 1189-1232, 2001.

[20] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. CatBoost: unbiased boosting with categorical features. In Proceedings of the 32nd AAAI Conference on Artificial Intel ligence (pp. 5568-5575), 2018.

[21] Freund, Y., & Schapire, R. E. Experiments with a new boosting algorithm. In Proceedings of the 13th International Conference o n Machine Learning (pp. 148-156), 1996.

[22] Fisher, R. A. The use of multiple measurements in taxonomic pr oblems. Annals of Eugenics, 7(2), 179-188, 1936.

[23] Hoerl, A. E., & Kennard, R. W. Ridge regression: Biased estima tion for nonorthogonal problems. Technometrics, 12(1), 55-67, 1 970.

[24] Breiman, L. Random forests. Machine Learning, 45(1), 5-32, 20 01.

[25] Geurts, P., Ernst, D., & Wehenkel, L. Extremely randomized tre es. Machine Learning, 63(1), 3-42, 2006.

[26] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. Applied l ogistic regression. John Wiley & Sons, 2013.

[27] Quinlan, J. R. Induction of decision trees. Machine Learning, 1 (1), 81-106, 1986.

[28] Cover, T., & Hart, P, Nearest neighbor pattern classification. IE EE Transactions on Information Theory, 13(1), 21-27, 1967.

[29] John, G. H., & Langley, P. Estimating continuous distributions i n Bayesian classifiers. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (pp. 338-345), 1995.

[30] McCallum, A., & Nigam, K. A comparison of event models for naive Bayes text classification. In AAAI-98 Workshop on Learn ing for Text Categorization (Vol. 752, No. 1, pp. 41-48), 1998.

[31] W. Zhu, N. Zeng, and N. Wang, “Sensitivity, specificity, accura cy, associated confidence interval and ROC analysis with practic al SAS implementations,” in NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland, 2010.

[32] T. Fawcett, “An introduction to ROC analysis,” Institute for the Study of Learning and Expertise, Palo Alto, CA, USA, Dec. 19, 2005. [Online]. Available: https://www.ics.uci.edu/~welling/teac hing/271fall09/AreaUnderCurveRoc/fawcett.pdf.

An Improved Check Digit-based Participant Identification System for Human Biorepositories

Minseok Chu*, Gilwon Kang** and Keun Ho Ryu***
* Korea Disease Control and Prevention Agency, Korea
E-mail: msc202@korea.kr Tel: +82-43-719-6512
** Chungbuk National University, Korea
E-mail: gilwon67@chungbuk.ac.kr
*** Ton Duc Thang University, Vietnam
E-mail: khryu@tdtu.edu.vn Tel: +82-43-267-2254

***Abstract* — Bioresearch data primarily contain private identifiable information. Thus, to ensure personal information protection, anonymization is required throughout the data collection and storage phases, which necessitates significant efforts to prevent errors. This paper proposes a code system ideally suited for replacing personal identification numbers with unique identifiers. The study further explores a check digit algorithm optimized for minimizing identification errors. The participant identification number, which plays a crucial role in the integrated management of bioresearch data, comprises 10 digits, with the last one serving as the check digit for validating the accuracy of the identification number. Among check digit algorithms available for participant identification number systems, the Verhoeff algorithm has demonstrated superior error detection performance compared to other algorithms. Simulations and applications of the coding system confirmed that the proposed method effectively anonymizes personal identification information with improved accuracy. The method is highly applicable to various scenarios involving the collection and storage of bioresearch data.**

I. INTRODUCTION

Personal data and identity protection are globally recognized as crucial issues. In particular, bioresearch data usually contain personally identifiable information, which when inappropriately disclosed can result in privacy infringement or criminal misuse, with significant socioeconomic impact on individuals [1].

In South Korea, these issues are addressed under the Bioethics and Safety Act. This legislation regulates the permanent deletion of personal information from human biological material donors, promoting anonymization through replacement with an institution-specific identifier (Article 2, Clause 19). Anonymization is mandated throughout all data processing phases when the donor does not provide consent for the inclusion of personal information (Articles 38 and 42–44). Similarly, the Personal Information Protection Act defines "pseudonymized information" as "personal information that is pseudonymized, making it impossible to identify a particular individual without the use/combination of additional information for its restoration to the original state" (Article 2, Clause 1). Anonymized or pseudonymized information serves as a legal and institutional framework for replacing personally identifiable information with unique identifiers either entirely or partially, to prevent donor identification. Within this legal framework, Korean biobanks issue identification numbers generated by anonymization algorithms to ensure personal information protection while collecting and storing human biological materials, genetic information, epidemiological information, and clinical data. These unique identification numbers can be reverse-converted to participant identification numbers by authorized personnel or committee(s). Such an identification number system plays a vital role in data sharing and linkage among healthcare providers and government entities, when providing human biological materials and data to researchers.

Although the accurate identification of each participant is crucial when collecting biological specimens and conducting epidemiological surveys, errors cannot be completely avoided despite careful research design and strict adherence to protocols [2]. The Laboratory Accreditation Checklists of the College of American Pathologists in the United States mandate two patient identifiers prior to specimen collection [3]. However, patient misidentification during laboratory testing remains a significant cause of medical errors [4].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Identification errors can occur at any stage of a diagnostic testing procedure [5]. Although electronic systems can effectively reduce error occurrence, cohort study data are often collected from participants' handwritten notes, resulting in frequent errors during the transformation of data from paper into electronic format by data entry personnel [6].

Given these circumstances, check digits can be utilized in computer data processing to automatically detect potential errors in identification numbers as an effective method to reduce the risk of participant misidentification. A check digit is a mathematically calculated digit added to the data to verify whether a valid value has been correctly submitted to the recipient without incurring any alterations to the original data during transmission [7]. The check digit acts as a checksum or error control, and various algorithms have been introduced depending on the purpose and components of the data (e.g., binary, alphanumeric, or decimal). For instance, data composed of decimal numbers (e.g., resident registration numbers, bank account numbers, and product codes) utilize algorithms such as Modular, Luhn, Damm, and Verhoeff to detect errors in the input values.

This study examined the application of personal identification numbers for participant identification in the bioresearch environment and related error detection algorithms. The remainder of this paper is organized as follows: Section 2 provides a brief review of previous studies on the application of check digits. Chapter 3 presents a case study on the use of check digits in composing bioresearch data identification numbers, explaining the identification number system and error detection algorithms. Chapter 4 presents the experimental results obtained from applying check digits. Finally, in Chapter 5, related issues are discussed and examined.

II. Related Research

The check digit automatically verifies the correct arrangement and accuracy of sequentially inputted numbers in complex and dynamic work environments and is widely recognized as a technique that effectively detects errors during manual recording or data transmission.

Error-detection algorithms for input data are widely used and continually improved upon across various fields, including the production and distribution of goods. In particular, check digit systems based on the Luhn algorithm are commonly used because of their simplicity. However, although the Luhn algorithm is effective in detecting transposition errors (e.g., $12 \rightarrow 21$), it is less sensitive to other types of errors because of the simple repetition pattern of the weights [8]. Consequently, alternative algorithms such as the International Standard Book Number (ISBN) [9] and Universal Product Code (UPC) [10] have been developed to address this limitation. More complex algorithms, such as Verhoeff [11] and Damm [12] detect errors in a more precise manner. Moreover, with advancements in computer processing speed, the calculation speed of these algorithms has reached a level that does not significantly impact workflow efficiency.

With the internationalization of distribution systems, UPC has been expanded to the European Article Number (EAN)-13, EAN-8, UPC-A, and UPC-E, depending on the intended use and required number of digits, as part of the global standard (GS1) of barcoding systems, which encompasses various standardized formats [13]. ISBN, a unique identifier for books originally proposed by Foster and Smith in 1966, was officially adopted as the "International Standard Book Number" by the International Organization for Standardization in 1970, as ISO 2108 [9], and is currently used for all published books [14]. It consists of a numeric code broken down into numbers representing the country, publisher, and book number, with the final digit serving as a check digit. To generate the check digit for a 10-digit ISBN, the ISBN Modulus 11 Check is used, which involves a mathematical manipulation, whereby each digit is multiplied by a weight ranging from 1 to 10, to ensure that the resulting number is divisible by 11 without a remainder. Check digits are used in credit cards worldwide. A credit card number typically comprises 10–19 digits that contain account information, along with a single check digit to identify the card issuer. The Luhn algorithm is employed to calculate the check digit [15]. The credit card issuer identification number has been standardized by ISO/IEC 7812-1, and this standard has also been adopted by the International Mobile Equipment Identity (IMEI) for check digit calculation [16].

The National Provider Identifier (NPI), which was adopted in 1996 pursuant to the Health Insurance Portability and Accountability Act (HIPAA), is a 10-digit numeric identifier consisting of a nine-digit serial number and one digit check. It is used to identify healthcare providers in the United States [17]. The check digit of the NPI is calculated using the Luhn algorithm, which is also employed in credit cards and mobile equipment, based on the ISO/IEC 7812-1 standard mentioned above. In the GenomeEUtwin project, the cohort study focuses on twin genetic analysis in Europe, and a numeric code is used to differentiate individual participants. This code

consists of a three-digit country code, serial number, and one check digit [18]. The Generalized Universal Mathematical Method (GUMM) is utilized to calculate the check digit in the GenomeEUtwin project [19].

In summary, the identification numbers used to detect errors that may occur during manual recording and data transmission, are generated by a coding system suitable for the given environment, with the last digit serving as the check digit. The suitability of different check digit algorithms should be evaluated on a case-by-case basis, considering their applicability for the intended use. This study proposes a code system for identifying participants who have donated human biological materials for research in the healthcare field, further exploring the most effective check-digit algorithm to reduce misidentification rates.

III. Materials and Methods

A. *Integrated management of bioresearch data*

This process focuses on the unique identifiers used to anonymize the participants' personal information throughout the process of bioresearch data collection, storage, and utilization. Fig. 1 provides a brief overview of this scenario.

Large-scale biodata collection projects typically exhibit

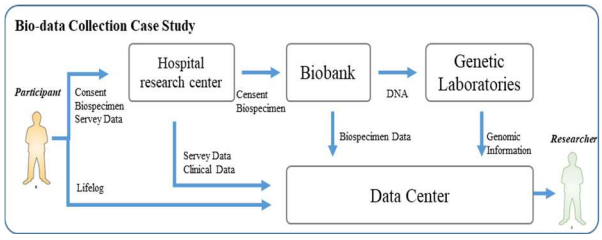


Fig. 1 Flowchart of the integrated system for bioresearch data under study. The data donated by each participant go through various processing stages within multiple organizations before being provided to researchers.

similar organizational structures and roles. The left side of Fig. 1 illustrates the data generation process, in which a voluntary participant visits a hospital or research center to complete a consent form, initiating the data generation process. The participant undergoes an epidemiological survey and clinical examination conducted by the hospital or research center staff and provides biological materials, such as blood or tissues. These donated biological materials are stored in a biobank, whereby genetic laboratories utilize DNA to generate genomic information. The bioresearch data produced through this process are integrated at the data center (bottom right of Fig. 1), to provide convenient access to researchers.

The workflow depicted in Fig. 1 demonstrates that biodata are generated at four different locations, with each location generating and storing different types of data. Because researchers must be able to identify the participant from the data, all bioresearch data and the corresponding participant identification numbers must remain identifiable throughout the data generation and transmission processes. According to the Bioethics and Safety Act, participant identification numbers should be replaced with unique identifiers to ensure anonymization rather than using personally identifiable information directly.

Errors in participant identification numbers can significantly affect subsequent processes within the workflow. These errors often remain undetected until the data reach the final datacenter for integration. Indeed, with data generated and modified at each stage, small errors in the identification numbers from the previous stage are not easily detectable, making arbitrary modifications impossible. This can result in complex quality-control issues.

B. *Participant identification number*

The protection of participants' biodata-related personal information is a major issue being actively researched from various perspectives [20, 21]. The issue of personal information protection in biobanking and health data is not only emphatically covered in mainstream media [22] but also emphasized in the academic literature [23–25]. Specifically, the potential to identify an individual's DNA within a DNA mixture [26] is a significant concern in terms of participant identification within a biobank.

To address these concerns, rigorous anonymity of participant identification numbers is essential. In this context, a participant identification number is defined as "a unique identifier assigned by the data center under the law to replace personally identifiable information, such as resident registration, passport, driver's license, and foreigner registration numbers, to prevent easy recognition when combined with other information." Furthermore, a participant identification number should serve as a mediator to enable bidirectional communication between the participant and both the biobank and bioresearch data for authorized users (or information systems).

The GenomeEUtwin project [18] added specific personal information to participant identifiers, such as singleton, twin, triplet, and quadruplet information, during the collection of genotypic and phenotypic information for twin research in Europe. This was to address end users' preference for identifiers containing eyeball data that they could verify

visually. Literature being scarce on participant identifiers, the current academic literature does not reflect other cohort situations. Korean research projects that collect human-derived materials often include code for specific personal information, such as region, gender, mother-child relationships, and the presence of disease. Including specific personal information in the participant identification number may be convenient for the workflow; however, there is the disadvantage of partial disclosure of personal information to anyone. The Personal Information Protection Act defines personal information as “information that can be used to identify a specific individual when combined with other information, even if the individual cannot be identified with this information alone.” Under the Resident Registration Act, the resident registration number contains unique personal identification information and is composed of 13 digits that disclose personal information such as the holder’s birth date, gender, and region of first issuance,. In response to human rights issues arising from excessive exposure to sensitive personal information, this ID card system was revised in October 2020, with the numbers in the latter part of the resident registration number representing the region replaced by random numbers. Therefore, it is desirable to remove as much identifiable information as possible from the participant identification number, as it is visible to anyone during biodata processing. Even at the expense of convenient visual identification of data during processing, it is necessary to ensure personal information protection.

This paper proposes the use of participant identification numbers that employ random instead of serial numbers, without encoding specific personal information, to minimize the risk of personal information disclosure (Fig. 2). The level

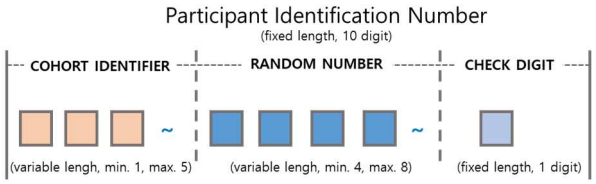


Fig. 2 Composition of the participant identification number scheme.

of personal data protection is enhanced by using numbers that do not follow a specific order or rule. To easily identify the cohort, which refers to the entire process from biodata collection to utilization, only a cohort identifier assigned by the data center is used as a prefix. The data center issues a unique cohort identifier upon request from the cohort research project manager. However, random numbers belonging to each cohort are assigned and managed by the corresponding institution, which further strengthens anonymity. Thus, the proposed system enables the efficient use of the necessary research data while protecting each participant's personal information.

The participant identification number is composed of 10 digits broken down into three parts as follows:

- Cohort identifier: 1 to 5 digits
- Random number: 4 to 8 digits
- Check digit: 1 digit

The All of Us Research Program [27], the world's largest biodata collection project, reportedly has more than one million voluntary participants across the United States. However, given the Korean population of 50 million [28], it is unrealistic to conduct research projects targeting hundreds of millions. Accordingly, to register all domestic bioresearch data in one location, it is reasonable to use nine digits that cover cohort projects with thousands of participants. That is, a 9-digit system that combines a cohort identifier (initial numbers) and random numbers (subsequent numbers) can be used, with the digits for the cohort identifier appropriately designated according to the cohort size. To ensure the accuracy of this 9-digit number, a check digit is added to verify the correct assignment of the identification number. An appropriate check digit algorithm is presented and explained in the following section.

C. Comparison of check digit methods

Table 1 presents common types of digit errors that occur during the generation of numeric code through manual input, as identified by Siddharth [29].

Table 1. Types of errors likely to occur during number code generation process

Category	Error Type	Example
1	Single digit error	1 → 2
2	Single transposition error	12 → 21
3	Twin error	11 → 22
4	Jump transposition error	546 → 645
5	Jump twin error	545 → 343
6	Phonetic error	13 (thirteen) → 30 (thirty)

A single-digit error refers to the case in which only one digit of a number sequence is incorrectly entered. A single transposition error occurs when the nth and (n+1)th digits are replaced by the (n-1)th and nth digits, respectively. A twin error occurs when two consecutive identical digits are

mistakenly entered as two other identical digits. A jump transposition error occurs when the nth and (n+2)th digits are replaced by the (n+2)th and nth digits, respectively. A jump-twin error refers to a case in which the nth and (n+2)th digits are identical and incorrectly entered as two other identical digits. Errors may also occur because of similar sounding numbers in some languages. Examples include the English numbers 16 (sixteen) and 60 (sixty) and German numbers 15 (fünfzehn), and 50 (fünfzig) [11]. In Korean, number pairs that are distinguished by the presence or absence of the final consonant of a syllable, such as 1 (일, il) and 2 (이, i) or 3 (삼, sam) and 4 (사, sa), may sound similar and be entered incorrectly. However, because these cases can be classified as single-digit errors, phonetic errors are not considered separately in this study. In addition to error detection during number code generation, there are also issues concerning data standard compliance, such as inputting nonnumeric characters or values with ten places. These issues can be easily resolved through exception handling during the information system implementation process; hence, they were not considered in this study.

The ISBN-13, Damm, and Verhoeff algorithms, which consider possible errors when generating numerical codes, can be used to generate participant identification numbers. The remainder of this section explains these three check digit algorithms used to generate participant identification numbers.

1) ISBN-13

Every publication is assigned a unique ISBN number. The ISBN was originally introduced by Foster and Smith in 1966 [30] and published as ISO 2108 by the International Organization for Standardization in 1970. An ISBN consists of the numbers α1, α2... α9, and the check digit α10 is chosen to satisfy the weight coefficients W1, W2, W3....W9 (10,9,8,7,6,5,4,3,2). The Luhn algorithm, developed by IBM scientist Luhn [31], assigns weights to the given digits and uses the remainder of the total as the check digit, similar to ISBN-13.

2) Damm algorithm

Unlike the Luhn algorithm and ISBN-13, the Damn algorithm uses an approach based on a more complex mathematical principle [12] of symmetric 10 × 10 two-dimensional array in its calculations.

3) Verhoeff algorithm

Verhoeff conducted an empirical analysis of user-input errors when processing decimal input values in the Dutch postal system. The analysis revealed that 89.26% of the errors were due to transcription (ab) or transposition (ab → ba), accounting for 79.05% and 10.21%, respectively [11]. Based on these analytical results, Verhoeff proposed an algorithm to detect single-digit and single-transposition errors. This method is similar to the Damm algorithm because it uses a two-dimensional array structure.

IV. EVALUATION

The participant identification number, designed for the integrated management of bioresearch data, comprises 9 digits and a check digit, thus allowing the generation of one billion unique identifiers. After generating all types of errors (Table 1) that could occur for these identifiers, we assessed the ability of the ISBN-13, Damm, and Verhoeff algorithms to accurately detect errors in randomly selected identifiers.

The error detection test model was implemented using Python 3.9 and implementation modules were incorporated for each algorithm. The model is executed in a computing environment powered by cloud-based Infrastructure as a Service (IaaS). The system running the model was equipped with an Intel Xeon 2 CPU, 4 GB of memory, and a 128 GB Non-Volatile Memory express (NVMe) storage device.

The comparative test results for approximately 3.5 billion identifiers are summarized in Table 2. All three check digit algorithms detected 100% of the single-digit errors. While the Damm and Verhoeff algorithms detected 100% of the single-transposition errors, ISBN-13 detected only 88.89%. Fenwick noted that ISBN-13 cannot detect adjacent transpositions of two digits with a difference of 5, such as 1 ↔ 6 or 8 ↔ 3 [8]. Given that the weights of all odd and even digits of ISBN-13 are equivalent to 1 and 3, respectively, approximately half of the jump transposition errors (even ↔ even, odd ↔ odd position) cannot be detected. The Damm and Verhoeff algorithms, which use a more complex mathematical approach than the weight-based check digit approach for ISBN-13, exhibited high mean error detection rates of approximately 96%.

For the proposed system, which combines a 9-digit identifier with a check digit to ensure the anonymity of personal information, the Verhoeff algorithm appears to be the most suitable. This method demonstrated a higher detection rate than the other two methods. However, considering the minimal difference in error detection performance between the Verhoeff and Damm algorithms, the latter can also be employed as an alternative depending on specific needs.

Table 2. Error detection rates of the check digit methods for integrated management of bioresearch data.

Error Type	ISBN-13	Damm Algorithm	Verhoeff Algorithm
Single digit error	100%	100%	100%
Single transposition error	88.89%	100%	100%
Twin error	88.89%	96.99%	91.90%
Jump transposition error	0%	92.23%	93.65%
Jump twin error	88.89%	89.67%	93.65%
Mean error detection rate	73.33%	95.78%	95.84%

V. Discussion

Small errors that may occur in identifiers during the collection of health-related information for research purposes (e.g., human biological materials, genetic data, epidemiological information, and clinical examination data) can be difficult to detect, potentially leading to complex quality management issues. Using check digits with participant identification numbers obviates the need for database lookups for verification, enabling easy detection of identifier-related errors throughout the various data processing stages. This approach is highly effective for quality management of bioresearch data. The methods presented in this paper can be applied not only to participant identification numbers, but also to appropriate numbering systems designed to detect input errors in the bio-data management process, such as donation consent forms, epidemiological questionnaires, clinical records, and human biological materials.

The process of collecting bioresearch data involves various stages: obtaining the participant's consent for donation, conducting subsequent clinical examinations, and collecting biospecimens (such as blood or tissue), which are then used to generate human biological materials, such as cells and DNA. DNA is prepared for genomic analysis data and supplied to researchers. The proposed participant identification number, used throughout this complex process of biodata collection, enables the anonymous management of human bio-specimens and associated data, thereby allowing operators at each stage to accurately identify and efficiently verify any input errors.

The research utility of biomedical data has increased in recent years amid the burgeoning big data market and increasing interdisciplinary integration. The establishment of registration systems for searching human biospecimens and the associated data necessary for research, is actively underway. Both academia and government entities are intensely promoting data-driven innovative growth while striving for a balanced tradeoff between data utilization and information protection. Confronted with a surge in information linkage between bioresearch data and collaborative research, the participant identification system for information management was proposed, alongside research on encryption and anonymization technologies, which are expected to provide significant assistance.

VI. Conclusion

This study investigated the use of identification numbers for the tracking of participants who provide bioresearch data, as well as applications of error-detection models. To minimize the risk of personal information exposure, we proposed a method that uses random numbers instead of sequential numbers, obviating the encoding of specific participant information in identification numbers. A check digit was added as the last digit of the participant's identification number to detect errors that may have occurred during manual input or data transmission. In this context, three different check digit methods: the ISBN-13, Luhn, and Verhoeff algorithms were compared. The results show that both the Verhoeff and Luhn algorithms effectively detected errors, significantly outperforming ISBN-13.

Although the check digit is indeed a very useful tool for error detection, it has limited capability for restoring the original information in the event of errors. To overcome this limitation, additional research on the application of Error Correction Codes (ECC) is required. Future research should aim to apply ECC, a concept that has been extensively studied in the field of computer science, to healthcare data. This would help prevent data corruption and alteration, ensure the accuracy and integrity of the data, and provide an important direction for efficient data management.

References

[1] T. Song, "Efficient utilization of health and welfare big data," *KIHAS Health and Welfare Data Portal*, vol. 193, pp. 68-76, 2012.

[2] J. Van den Broeck, S. Argeseanu Cunningham, R. Eeckels, et al., "Data cleaning: detecting, diagnosing, and editing data abnormalities," *PLoS Med.*, vol. 2, pp. e267, 2005.

[3] A. Rabinovitch, *The College of American Pathologists (CAP) Laboratory Accreditation Checklists*, College of American Pathologists, 2000.

[4] P.J. Howanitz, S.W. Renner, and M.K. Walsh, "Continuous wristband monitoring over 2 years decreases identification errors: a College of American Pathologists Q-Tracks Study," *Arch. Pathol. Lab. Med.*, vol. 126, pp. 809-815, 2002.

[5] G. Lippi, G.L. Salvagno, M. Montagnana, et al., "Phlebotomy issues and quality improvement in results of laboratory testing," *Clin. Lab.,* vol. 52, pp. 217-230, 2006.

[6] R. Sakurai, M. Ueki, S. Makino, et al., "Outlier detection for questionnaire data in biobanks," *Int. J. Epidemiol.*, vol. 48, pp. 1305-1315, 2019.

[7] J.F. Ludvigsson, P. Otterblad-Olausson, B.U. Pettersson, et al., "The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research," *Eur. J. Epidemiol.*, vol. 24, pp. 659-667, 2009.

[8] P. Fenwick, *Introduction to Computer Data Representation*, Bentham Science Publishers, 2014.

[9] International Organization for Standardization, "ISO 2108 Information and Documentation: International Standard Book Number (ISBN)," 2005. Available at: https://www.iso.org/standard/65483.html (Accessed: September 11, 2022).

[10] D. Savir and G.J. Laurer, "The characteristics and decodability of the Universal Product Code symbol," *IBM Syst. J.,* vol. 14, pp. 16-34, 1975.

[11] J. Verhoeff, "Error detecting decimal codes," *MC Tracts*, 1969.

[12] H. M. Damm, "Total anti-symmetrische Quasigruppen," Ph.D. Philipps-Universität Marburg, 2004.

[13] J. G. Keogh, A. Rejeb, N. Khan, et al., "Blockchain and GS1 standards in the food chain: A review of the possibilities and challenges," in *Building the Future of Food Safety Technology*, 2020, pp. 171.

[14] W. Kamaku, C. Mwathi, and B. Kivunge, "Limitations in the Convectional ISBN-10 Code," *Am. Int. J. Contemp. Res*, vol. 2, pp. 553-569, 2012.

[15] International Organization for Standardization, "ISO/IEC 7812-1:2017 Identification cards - Identification of issuers - Part 1: Numbering system," 2017. Available at: https://www.iso.org/standard/70484.html (Accessed: September 11, 2022).

[16] GSM Association, "IMEI Allocation and Approval Guidelines Version 16.0," 2019. Available at: https://www.gsma.com/newsroom/wp-content/uploads/TS.06-v16.0.pdf (Accessed: September 11, 2022).

[17] Centers for Medicare & Medicaid Services, "Standard Unique Health Identifier for Health Care Providers," 2004. Available at: https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand (Accessed: September 11, 2022).

[18] J. Litton, J. Muilu, A. Björklund, et al., "Data modeling and data communication in GenomEUtwin," *Twin Res. Hum. Genet.,* vol. 6, pp. 383-390, 2003.

[19] H. Gumm, "A new class of check-digit methods for arbitrary number systems (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 31, pp. 102-105, 1985.

[20] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nat. Rev. Genet.*, vol. 15, pp. 409-421, 2014.

[21] N. Bandeira, E.W. Deutsch, O. Kohlbacher, et al., "Data management of sensitive human proteomics data: current practices, recommendations, and perspectives for the future," *Mol. Cell. Proteomics*, vol. 20, 2021.

[22] The Washington Post, "Newborns' Blood Samples Are Used for Research Without Parents' Consent," Available at: http://www.washingtonpost.com/wp-dyn/content/article/2009/06/29/AR2009062903118.html (Accessed: September 11, 2022).

[23] K. Hens, C.E. Van El, P. Borry, et al., "Developing a policy for paediatric biobanks: principles for good practice," *Eur. J. Hum. Genet.*, vol. 21, pp. 2-7, 2013.

[24] D. Budimir, O. Polasek, A. Marusic, et al., "Ethical aspects of human biobanks: a systematic review," *Croat. Med. J.*, vol. 52, pp. 262-279, 2011.

[25] J.E. Lunshof, R. Chadwick, D.B. Vorhaus, et al., "From genetic privacy to open consent," *Nat. Rev. Genet.*, vol. 9, pp. 406-411, 2008.

[26] N. Homer, S. Szelinger, M. Redman, et al., "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genet.*, vol. 4, pp. e1000167, 2008.

[27] NIH All of Us Research Program, All of Us Research Program Begins Beta Testing of Data Platform, May 27, 2020; Biotech Policy Research Center, *BioINwatch,* 18-41, The US NIH embarks on a study of a million people, June 5, 2018.

[28] KOSIS (Korean Statistical Information Service), "Main Demographic Indicators, 2021." Available at: https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1BPA002&checkFlag=N (Accessed: September 11, 2022).

[29] A. Siddharth, “Error detection in numeric codes,” *Resonance*, Heidelberg, 2012, pp. 653.

[30] ISBN.org, “ISBN History,” 2022. Available at: http://www.isbn.org/ISBN_history (Accessed: September 11, 2022).

[31] H.P. Luhn, “U.S. Patent No. 2950048 - Computer for Verifying Numbers,” 1960.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhancing Snoring Detection with Statistical Analysis of Audio Features

Suphachok Buaruk[1], Somrudee Deepaisarn[2,*]
Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
[1]E-mail: d6522300067@g.siit.tu.ac.th
[2]E-mail: somrudee@siit.tu.ac.th
[*] Corresponding Author

***Abstract*—This research investigates classification models and data representation techniques to distinguish snoring from non-snoring sounds. Spectrogram images were fed into a pre-trained ResNet18 model, attaining a classification accuracy of 99.00%. The mean and standard deviation of the acoustic feature profiles were extracted from audio features, including spectral features, zero crossing rate, root mean square, and Mel Frequency Cepstral Coefficients (MFCCs). Different combinations of the extracted features were utilized for training various machine learning classifiers to detect whether the audio samples contained snoring sounds. The MFCC features play a significant role in capturing snoring patterns against non-snoring sounds, attaining the highest classification accuracy of 98.50% when coupled with a random forest classifier. In addition, the feature importance list is presented for guiding feature engineering and providing data insights for greater interpretability. These findings emphasize the effectiveness of using statistical values derived from audio signals as features for machine learning-based classification. This work contributes to the field of snoring detection and classification with potential applications in sleep disorder monitoring and diagnosis.**

I. Introduction

According to a statistical analysis conducted by N. Bhattacharyya in 2015, snoring is a widespread phenomenon worldwide, affecting a substantial proportion of the adult population. Specifically, the study revealed that approximately 59.0% of adult men and 46.9% of women experience snoring on at least an occasional basis [1]. The snoring phenomenon can signify a more severe condition in several life aspects, such as health status and relationship strain [2]. The cause of snoring is the relaxation of the throat, narrowing the airway during sleep [3]. The vibration of the soft tissues in the airway creates the sound of snoring. It can be a sign of a serious condition called obstructive sleep apnea (OSA), which is characterized by pauses in breathing while asleep and a significant decrease in blood oxygen levels. OSA is linked to different health issues, such as high blood pressure, heart disease, stroke, and diabetes [4], [5]. In addition, it could also affect daytime fatigue, reduce cognitive function, and decrease quality of life. Snoring and OSA are more common among men, overweight people, and those over 40 years old. Other factors contributing to snoring and OSA include smoking, alcohol consumption, nasal congestion, and sleeping position [6]. The classification task in snoring detection enables the development of automated systems for snoring analysis. The INTERSPEECH 2017 ComParE Snoring Challenge [7] involved the classification of snoring sounds, aiming to explore and understand the different mechanisms and characteristics of snoring related to various sources of OSA.

The ability to accurately identify the sound of snoring and non-snoring is also very important for early diagnosis in the medical process. Several tools are used to diagnose snoring conditions, including questionnaires, physical examinations, and sleep tests. The results provide crucial information for developing appropriate treatment plans. The majority of snoring and non-snoring classification research over the recent decade has been around the applications of artificial intelligence (AI) and machine learning (ML) techniques. Adesuyi *et al.* represented sound signals by spectrogram images and input into a 2D-CNN model for the classification of snoring and non-snoring sounds but the classification accuracy was poor [8]. The same group later proposed the 1D-CNN network incorporated with the numerical extracted features for the same classification task, resulting in a high-performance accuracy of 99.7% [9]. A number of published works describe the importance of signal processing in extracting features from snoring sound for classification purposes. The selection of appropriate features is critical as it directly impacts the performance and accuracy of the classification models [10]. Recent evidence has indicated that sound signal features in the domain of music information retrieval (MIR) garnered significant interest in the feature extraction process [9]. These features include the mel-frequency cepstral coefficients (MFCCs), short-time Fourier transform (STFT), zero-crossing rate (ZCR), and spectral features.

In this study, a range of audio features commonly employed in the analysis of snoring sounds is provided with data representation to capture the sound characteristics. The binary classification of snoring and non-snoring is carried out using two approaches of data representation as inputs for deep learning and classical machine learning models, respectively, detailed as follows:

- Spectrogram images are the data represented inputs for the pre-trained ResNet18 model.
- Statistical values of extracted audio features are the data represented inputs for machine learning classification

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

models.

The remaining sections of this paper are structured as follows. Section II outlines the methods used in this study, which include details on the dataset, data representation techniques, and classification models. Section III presents the results and discussion regarding the classification performance of each data representation technique and classification model. Lastly, Section IV summarizes the study and explores potential directions for future research.

II. Method

A. Snoring and Non-snoring Dataset

There are 1,000 sound samples in the dataset, separated into two groups: snoring sounds and non-snoring sounds [11]. Each class consists of 500 samples. These sounds were collected from various online sources and were previously used to create a binary classification model of snoring and non-snoring sounds. The length of each pre-processed sound sample is one second. Among the 500 snoring sound samples, there are 363 sound samples without any background noise acquired from children, adult men, and adult women. The remaining 137 samples include snoring sounds accompanied by other non-snoring sounds in the background. The 500 non-snoring samples were made up of different ambient noises typically heard around a person who snores. These sounds were gathered from ten different categories, containing 50 samples per category. The categories include sounds of a baby crying, a clock ticking, a door opening and closing, complete silence or the faint sound of a vibrating gadget, a toilet flushing, emergency vehicle sirens, rain and thunderstorms, streetcar sounds, people talking, and background television news. Fig. 1 presents example visualizations of the waveforms for snoring and non-snoring sounds.

In this work, the dataset was divided into three subsets by ratios of 60:20:20 for training, validating, and testing, respectively. The same training and testing datasets were utilized in the machine learning and deep learning experiments to ensure comparable evaluation of their performance. For deep learning with spectrogram representations, 60% of the data were assigned for training (600 samples), 20% for validating (200 samples), and the remaining 20% for testing (200 samples). For machine learning using extracted audio features, the same split ratios of 60:20:20 were applied.

B. Data Representation

1) Spectrogram: Transforming a sound signal into a spectrogram involves applying a Fourier transform to the signal. During this process, the time-domain signal is transformed to the frequency domain. As a result, the signal is represented as a sequence of frequencies along with their corresponding magnitudes over time. The resulting spectrogram can be visualized as a 2D plot, with the intensity of each pixel indicating the magnitude of the frequency component at that point in time. The spectrogram images of snoring sound in Fig. 1 are illustrated in Fig. 2.

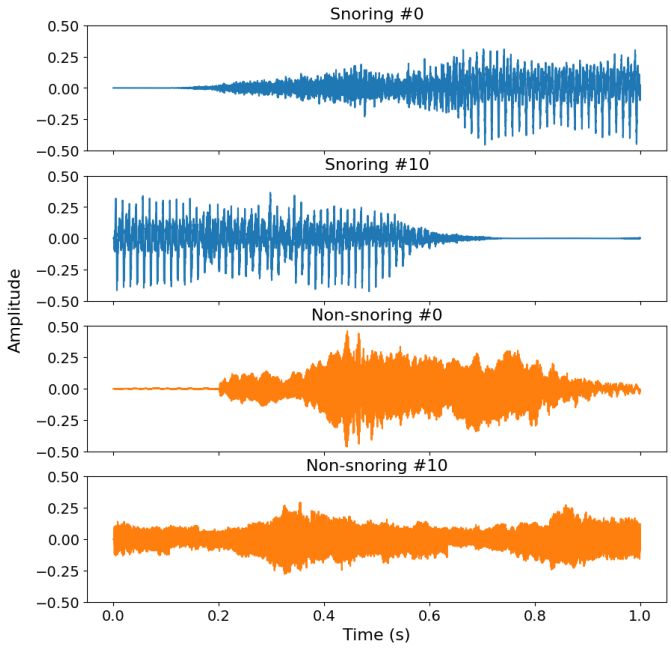


Fig. 1. Examples of snoring waveforms (blue) and non-snoring waveforms (orange), providing the visualization of physical features underlying the audio data

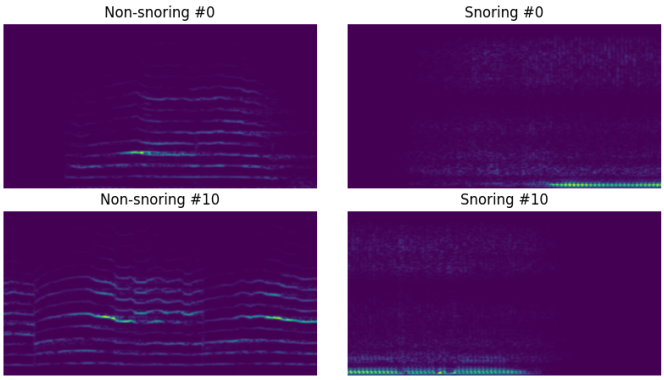


Fig. 2. Spectrogram images generated from snoring sounds (right) and non-snoring sounds (left).

2) Audio Features: Signal data processing typically involves identifying and extracting physical features from the signals, such as frequency, amplitude, and phase. Selecting and applying appropriate techniques for feature extraction in signal data can provide insight into the underlying patterns that can be used for further analysis or modeling.

- *Mel-frequency cepstral coefficients (MFCCs)*: The MFCCs are a common feature extraction technique used to analyze sound signals. The process involves breaking audio into small frames and then applying mathematical operations to each frame [12]. These operations involve multiplying a window function, followed by a Fourier transform. This sequence of steps converts the signal from the time domain to the frequency domain. The resulting spectrum is passed through a set of Mel filters,

and the logarithm of the Mel-filter bank energies is taken to produce Mel frequency. A Discrete Cosine Transform (DCT) is applied to the Mel frequency coefficients to obtain a set of cepstral coefficients representing the spectral envelope of signals. Previous studies have reported that adopting MFCCs as extracted features works well for snoring identification [11], [13], [14].

- *Root Mean Square (RMS)* is often used to characterize the waveform of audio amplitude. The RMS of an audio signal intensity, I, is defined in (1):

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|I_i|^2} \tag{1}$$

where n is the number of signal data points in an individual window, which was set to 2,048 with the sliding size of 512 [9]. $i = 1, ..., n$ is the index of signal intensity across a window. In the case of snoring, RMS can be used to identify the intensity representing the loudness of the snoring signal.

- *Zero Crossing Rate (ZCR)* is the rate at which the signal crosses the zero level [15]. In the case of snoring, ZCR can be used to identify the periodicity of the snoring sound.
- *Spectral Centroid* is a measure of the center of mass of the spectrum. It provides information about the frequency at which the energy of the signal is concentrated [15]. In the case of snoring, spectral centroid can be used to identify the dominant frequency or range of frequencies associated with a snoring sound.
- *Spectral Bandwidth* is a measure of the spread of the spectrum around the spectral centroid. It provides information about the width of the frequency distribution of the signal [15]. In the case of snoring, spectral bandwidth can be used to identify the overall spectral content of snoring sound.
- *Spectral Rolloff* is a measure of the frequency at which the spectrum falls to a certain percentage of its maximum value. It provides information about the high-frequency content of the signal [15]. In the case of snoring, spectral rolloff can be used to identify the upper-frequency limit of snoring sound.

In summary, the feature extraction of snoring sound can provide useful information about the characteristics of the sound, which can be helpful in the diagnosis of sleep disorders and other health conditions.

C. Classification Models for Snoring Sound

This work aims to develop classification models to distinguish between snoring and non-snoring sounds. Two different approaches for data representation and subsequent classification approaches were investigated. ResNet-18 was employed to demonstrate the power of deep learning, particularly by highlighting its capability to eliminate the need for any feature extraction process. Furthermore, it also aimed to compare with traditional machine learning, which required feature extraction steps. The explanation of these two approaches was detailed as follows:

1) Deep Learning with Image Representation Inputs: The snoring sound dataset was transformed into spectrograms to extract visual representations of the audio signals. Image data representations of snoring and non-snoring sounds were fed into the Fast.ai deep learning framework [16]. The pre-trained ResNet18 model is used as a feature extractor [17], employed the Adam optimization function with the defaulted learning rate of 10^{-7}. The initial learning rate was determined using the **`lr_find`** function provided in the Fast.ai framework to optimize the classification performance. The model is fine-tuned in the training process by adjusting the learning rate based on the initial learning rate. Specifically, a fine-tuning step consisting of 10 epochs at a learning rate of 2×10^{-6} is implemented. This fine-tuning process was configured to automatically stop and save the model when the classification performance discontinue improvement. This fine-tuning step aimed to optimize the model performance and enable it to improve the classification between snoring versus non-snoring sounds.

2) Machine Learning with Feature Extracted Inputs: Statistical features are extracted from the snoring sound samples to capture relevant information. Specifically, the mean and standard deviation of the extracted audio features were calculated. These statistical features, along with their corresponding labels (snoring and non-snoring), are used to construct a dataset for training the machine learning models. Different classification algorithms available in PyCaret [18], including Random Forest (RF), Support Vector Machine (SVM), and *k*-Nearest Neighbor (KNN) are utilized. The dataset is split into training and testing sets same as the dataset for the deep learning approach. Model hyperparameters are optimized using the technique of 20 folds cross-validation with randomized parameters.

III. Results and Discussion

In this work, the performance of the classification for snoring and non-snoring sounds is presented. The classification models, developed using both machine learning and deep learning approaches, are evaluated using the accuracy metric. Spectrogram images and numerical extracted features are used as input data for deep learning and machine learning, respectively. This study aims to explore the strengths and capabilities of various data representations. The accuracy of classification performance in various scenarios of models and features is investigated as shown in Table I. In addition, the advantages and disadvantages of each approach for snoring and non-snoring sound classification are discussed in Table II.

A. Image Data Representation

The conversion of audio signals into spectrograms enables the capture of significant patterns and variations associated with snoring sounds. The classification model for snoring and non-snoring sounds, based on the ResNet18 pre-trained model, achieved an astonishing accuracy of 99% using spectrogram

TABLE I
CLASSIFICATION PERFORMANCE OF VARIOUS MODEL PIPELINES WITH VARYING DATA REPRESENTED FEATURE SETS AND CLASSIFICATION MODELS

Data Represented Features	Models	Accuracy
Spectrogram images	ResNet18	**99.00%**
Spectral (Centroid+Bandwidth+RollOff)	RF	89.00%
	SVM	61.50%
	KNN	87.00%
Spectral+RMS+ZCR	RF	94.00%
	SVM	73.50%
	KNN	89.50%
MFCCs	RF	**98.50%**
	SVM	90.50%
	KNN	96.50%
MFCCs+Spectral+RMS+ZCR	RF	97.50%
	SVM	94.50%
	KNN	98.00%

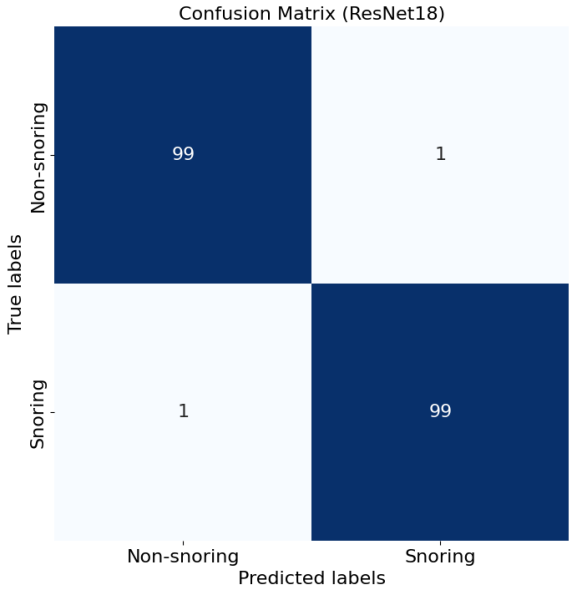


Fig. 3. Confusion Matrix of the snoring and non-snoring sound classification accuracy using the ResNet18 pre-trained model. The spectrogram images, serving as the input training dataset, are processed to predict the class labels.

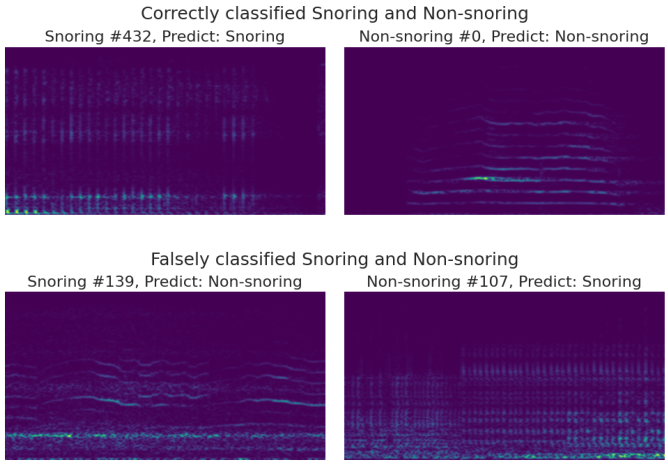


Fig. 4. Correctly (top) and falsely (bottom) classified image data representations for snoring (left) sound and non-snoring (right) sounds.

images as input data. The experimental results showed a significant improvement over the previous study conducted by Adesuyi *et. al.* that achieved an accuracy of 86.80% using their proposed 2D-CNN model [9]. The combination of fine-tuning and learning rate searching on the Fast.ai platform contributed to the higher classification accuracy achieved in this study. In addition, the pre-trained model has shown to be very helpful, as it provides a starting point by using knowledge gained from training on large datasets. It can also serve as a feature extractor which can save computational costs and time [19]. The confusion matrix is visualized to compare the predicted classes with the actual classes, as shown in Fig. 3. These results highlight the potential of deep learning and the effectiveness of spectrogram representation in accurate separation. However, there are missing predictions in classifying between snoring and non-snoring, possibly due to the represented image characteristics. The correct classification results are illustrated in the top part of Fig. 4, while the false classification results are shown in the bottom part of Fig. 4. This struggle can occur due to similarities in patterns, noise, or artifacts in the spectrogram.

B. Extracted Feature Representation

This section focuses on classifying snoring and non-snoring sounds using various audio features. These features include Mel-frequency cepstral coefficients (MFCCs), zero-crossing rate (ZCR), root mean square (RMS), and spectral features. The collected features are first extracted in the form of series data and are commonly used in audio classification. Especially for MFCC features, there are 20 orders of MFCC features extracted from the raw audio. These features were used as input series data in a 1D-CNN model, which resulted in a classification accuracy of 99.7% [9]. In this work, statistical values of the mean and standard deviation of those features are computed to create representative data inputs for classification. The classification of snoring and non-snoring is performed using three machine-learning models, including Random Forest (RF), Support Vector Machine (SVM), and *k*-Nearest Neighbor (KNN). The input datasets are separated into four scenarios, composed of four different sets of features. The performance of the classification models is evaluated using the accuracy score as shown in Table I. The results indicate the role of feature selection in audio classification. The feature set of spectral features shows that the RF model achieves the highest accuracy of 89.00%. In contrast, the SVM model exhibits lower accuracy at 61.50%, while the KNN model performs well with an accuracy of 87.00%. These findings suggest that the RF model is good at using spectral features to classify, but the SVM model may struggle because of the

TABLE II
COMPARISON OF ADVANTAGES AND DISADVANTAGES OF SNORING AND NON-SNORING CLASSIFICATION APPROACHES.

Approach	Advantages	Disadvantages
2D-CNN model using spectrogram images	- Captures temporal and frequency information. - Provides visual representation for pattern identification.	- Increases computational complexity and memory requirements for generating spectrogram images. - Lacks interpretability in terms of the physical features of the audio waveforms.
1D-CNN model using extracted audio features	- Extract features containing information on spectral, temporal, and amplitude characteristics. - Eliminates the need for additional pre-processing steps.	- Obtains limited capacity to capture complex spectral patterns compared to spectrogram-based approaches.
Machine learning models using statistical input features	- Simplifies input representation and reduces dimensionality. - Enhances efficiency for computational processing. - Utilizes the physical characteristics extracted from the audio waveforms.	- Potentially loses information due to the summarization process. - Not fully capture the dynamic and temporal nature of snoring and non-snoring sounds via the acquisition of statistical features alone.

linear separation assumption. Expanding the feature set by including Root Mean Square (RMS) and Zero Crossing Rate (ZCR) results in improved performance for all models. The higher accuracy demonstrates the power of combined features of spectral, temporal, and amplitude-based features of audio classification. When considering the Mel Frequency Cepstral Coefficients (MFCCs) as the features, the RF model achieves an impressive accuracy of 98.50%. This suggests that extracted MFCC features capture important characteristics of the audio signals that enable the classification model to achieve high-performance accuracy. Furthermore, combining MFCCs with spectral, RMS, and ZCR features leads to further improvement in classification performance. This combination is specifically employed to investigate the importance of different features in distinguishing between snoring and non-snoring sounds. The MFCC features capture the spectral characteristics of the audio signals. The spectral features provide additional insights into the distribution of energy across different frequency bands. The integration of RMS and ZCR features allows the consideration of the temporal and amplitude-related properties of the sound signals. The list of feature importance, as ordered in Fig. 5, was used for training an RF model where the predominant features were found based on the average values of MFCC features. Those MFCC features seem significant in capturing patterns related to snoring and non-snoring sounds. Other features, such as the standard deviation of spectral rolloff and ZCR, also enhance the classification performance.

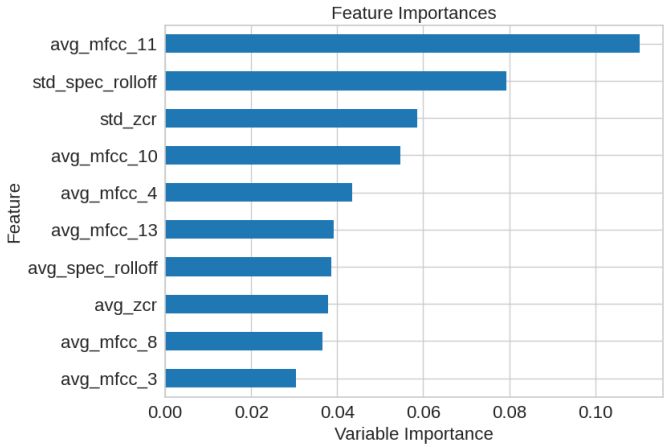


Fig. 5. The feature importance for training the Random Forest Classifier model is listed in descending order. The most important statistical features of the extracted series include the average of 11th MFCC, the standard deviation of spectral rolloff, the standard deviation of ZCR, the average of 10th MFCC, the average of 4th MFCC, the average of 13th MFCC, the average of spectral rolloff, the average of ZCR, the average of 8th MFCC, and the average of 3rd MFCC. For the hyperparameter tuning process, a 20-fold cross-validation approach is employed using randomized parameter sets.

IV. CONCLUSION

The outcomes of snoring and non-snoring sound classification using different data representations and classifiers are examined. This work contributes to understanding feature selection and suggesting appropriate model pipelines for this binary classification task. The deep learning model yields an accuracy of 99.00% using spectrogram images as the data represented input along with the application of the ResNet18 pre-trained model. For classical machine learning approaches, the features are extracted from audio waveforms, and statistical values are computed accordingly for use as input data. The best results are obtained when the MFCC feature is used in combination with the RF model, achieving an accuracy of 98.50%. This approach achieves comparable performance to a pre-trained deep learning model while offering the advantage of reduced computational resources and time in the data preparation and training processes. In addition, the results emphasize the benefit of statistical feature extraction from audio waveforms over the image-based data representation, which is mainly the ability to interpret the results and algorithmic simplicity. Further studies are demanded to explore powerful feature engineering techniques and classification models to improve the solutions to snoring detection and other audio signal classification problems.

ACKNOWLEDGMENT

The first author (S.B.) acknowledges the Graduate Scholarship Program for Excellent Thai Students, awarded by Sirindhorn International Institute of Technology, Thammasat University, Thailand, for his doctoral study.

References

[1] N. Bhattacharyya, "Sleep and health implications of snoring: A populational analysis," *The Laryngoscope*, vol. 125, no. 10, pp. 2413–2416, 2015.

[2] T. Young, L. Finn, H. Kim, *et al.*, "Nasal obstruction as a risk factor for sleep-disordered breathing," *Journal of Allergy and Clinical Immunology*, vol. 99, no. 2, S757–S762, 1997.

[3] B. Hofauer, B. Braumann, C. Heiser, *et al.*, "Diagnosis and treatment of isolated snoring—open questions and areas for future research," *Sleep and Breathing*, vol. 25, pp. 1011–1017, 2021.

[4] P. J. Strollo Jr and R. M. Rogers, "Obstructive sleep apnea," *New England Journal of Medicine*, vol. 334, no. 2, pp. 99–104, 1996.

[5] S. R. Surani, "Diabetes, sleep apnea, obesity and cardiovascular disease: Why not address them together?" *World journal of diabetes*, vol. 5, no. 3, p. 381, 2014.

[6] C. Sjöström, E. Lindberg, A. Elmasry, A. Hägg, K. Svärdsudd, and C. Janson, "Prevalence of sleep apnoea and snoring in hypertensive men: A population based study," *Thorax*, vol. 57, no. 7, pp. 602–607, 2002.

[7] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[8] T. A. Adesuyi, B. M. Kim, and Y. S. Shin, "A brief on snoring data and classification methods," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 426–432, 2020.

[9] T. A. Adesuyi, B. M. Kim, and J. Kim, "Snoring sound classification using 1d-cnn model based on multi-feature extraction," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 22, no. 1, pp. 1–10, 2022.

[10] B. Arsenali, J. van Dijk, O. Ouweltjes, *et al.*, "Recurrent neural network for classification of snoring and non-snoring sound events," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 328–331.

[11] T. Khan, "A deep learning model for snoring detection and vibration notification using a smart wearable gadget," *Electronics*, vol. 8, no. 9, p. 987, 2019.

[12] K. S. Rao and K. Manjunath, *Speech recognition using articulatory and excitation source features*. Springer, 2017.

[13] W. Duckitt, S. Tuomi, and T. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," *Physiological measurement*, vol. 27, no. 10, p. 1047, 2006.

[14] K. Qian, Z. Xu, H. Xu, Y. Wu, and Z. Zhao, "Automatic detection, segmentation and classification of snore related signals from overnight audio recording," *IET Signal Processing*, vol. 9, no. 1, pp. 21–29, 2015.

[15] T. Giannakopoulos and A. Pikrakis, *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014.

[16] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] M. Ali, *Pycaret: An open source, low-code machine learning library in python*, PyCaret version 1.0, Apr. 2020. [Online]. Available: https://www.pycaret.org.

[19] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Un-Rectifying in ReLU Networks and Applications

Shih-Shuo Tung,* Ming-Yu Chung,† Jinn Ho,* and Wen-Liang Hwang *
* Institute of Information Science, Academia Sinica, Taiwan
† National Taiwan University, Taiwan

***Abstract*—Most learning algorithms for deep neural networks (DNNs) employ gradient descent or block coordinate descent, which involves the direct application of a non-linear activation function throughout the course of optimization, such that only the input and output of the activation function are optimized. By contrast, the "un-rectifying" technique expresses a non-linear point-wise (non-smooth) activation function as a data-dependent variable (activation variable), which means that the activation variable along with its input and output can all be employed in optimization. The fact that the ReLU network in this study was un-rectified means that the activation functions could be replaced with data-dependent activation variables in the form of equations and equality constraints. We also demonstrate it is possible to optimize a network based on the augmented Lagrangian approach. In experiments and comparisons with other methods addressing classification and compressed sensing recovery problems, the proposed approach makes it easier to learn a network to a given degree of accuracy within a small number of epochs.**

I. Introduction

DNN training methods include the popular back-propagation algorithm and its variants, block coordinate descent (BCD), and the alternating direction method of multipliers (ADMM). The BCD approach involves decomposing a problem into sub-problems to be solved cyclically with the remaining blocks fixed according to their most recent updates. The ADMM approach searches for saddle points of augmented Lagrangian functions via primal and dual variable updates, wherein the updating of primal variables is based on the BCD approach.

In this study, we employed un-rectifying in our implementation of the augmented Lagrangian method in order to investigate the issue of convergence in learning an L-layer ReLU network, represented as follows:

$$\mathcal{M}_L = M_L \rho M_{L-1} \dots \rho M_1, \tag{1}$$

where ρ refers to ReLU and M_i is an affine transform, the linear portion of which is $\mathbf{W}_i$ and the bias is $\mathbf{b}_i$ [1], [2].

The learning problem in (1) can be formulated in a manner that allows the application of alternating optimization whenever a single block of variables is minimized, while leaving the other variables fixed. Below, we illustrate this approach using an example involving a three-layer DNN $\mathcal{M}_3 = M_3 \rho M_2 \rho M_1$ with the following three-splitting formulation:

$$\begin{cases} \min_{\mathbf{U}_1,\mathbf{V}_1,M_1,M_2} \ell(\mathbf{Y}, M_3\mathbf{V}_2) + S(\mathbf{V}_1,\mathbf{V}_2,\mathbf{U}_1,\mathbf{U}_2) \\ +T(M_1,M_2,M_3) \\ \mathbf{V}_1 = \rho\mathbf{U}_1, \mathbf{V}_2 = \rho\mathbf{U}_2 \\ \mathbf{U}_1 = M_1\mathbf{X}, \text{ and } \mathbf{U}_2 = M_2\mathbf{V}_1, \end{cases} \tag{2}$$

where ℓ is the loss function, ρ is the non-linear activation function, and S and T are regularization functions. The equality constraint is approximated using the penalty method with penalty parameter γ.

Nearly all approaches to DNN learning treat the non-linear activation ρ as a function, such that ρ is not directly involved in optimization [3], [4]. Most theoretical studies on convergence in DNN learning have assumed that if ρ is not a smooth activation function, then it must be approximated. By contrast, the un-rectifying technique [5] expresses point-wise CPWL function ρ as a data-dependent variable. For example, $a = \rho z$ becomes $a = dz$, where d is a data-dependent variable (i.e., its value is dependent on the value of z). Un-rectifying makes it possible to use d in optimization in the same way as other variables. It is from this perspective that we explore DNN learning using the augmented Lagrangian approach under the framework of the Conn, Gould, and Toint algorithm [6].

Un-rectifying ReLUs introduces an input-dependent diagonal matrix of entries $\{0, 1\}$, such that for input $\mathbf{x}$,

$$\mathbf{D}_\mathbf{x}\mathbf{x} = \rho\mathbf{x} \tag{3}$$

with the following constraints:

$$\begin{cases} \mathbf{0} \le \mathbf{D}_\mathbf{x}\mathbf{x}, \\ (\mathbf{I} - \mathbf{D}_\mathbf{x})\mathbf{x} \le \mathbf{0} \\ \mathbf{D}_\mathbf{x} = \mathrm{diag}(\mathbf{d}) \text{ with } \mathbf{d} = [d_i \in \{0,1\}]_i. \end{cases} \tag{4}$$

In other words, $\rho\mathbf{x}$ can be equivalently expressed as $\mathbf{D}_\mathbf{x}\mathbf{x}$, where $\mathbf{D}_\mathbf{x}$ satisfies (4). In [7], [8], the learning model in (1) using the un-rectifying approach achieved global convergence by relaxing the domain of the un-rectifying variables (variables replacing the activation functions) from integer to real values. This transforms a combinatorial problem into a real-domain problem, thereby making it possible to solve the problem of learning an un-rectifying network via constrained optimization. Meanwhile, applying the CGT-algorithm involves introducing data-dependent slack variables to convert inequality constraints into equality constraints.

This paper presents an alternative approach to the conversion of inequalities into equations using non-linear operations without introducing data-dependent slack variables. The first two contributions of this paper pertain to technical enhancement of the CGT-algorithm for ReLU networks:

- We amend the assumption of the CGT-algorithm (referring to as CGT-extension) in coping with continuous constraints (with mild assumptions pertaining to the first and second derivatives on constraint functions). We

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

demonstrate that the CGT-extension can reach global convergence to critical points of problems.
- We demonstrate that the equality constraints introduced by un-rectifying ReLUs satisfy the CGT-extension, which means that ReLU networks can be trained to critical points. Replacing ReLUs with equality constrained functions avoids the need to introduce massive data-dependent slack variables. This is a dramatic improvement over the un-rectifying representation of ReLUs in [7] when dealing with large datasets and deep networks.

We compared the performance of the proposed CGT-extension, BCD, ADMM, and Adam (state-of-the-art back-propagation algorithm) in training fully-connected networks in terms of classification and compressed sensing recovery using the MNIST, fashion-MNIST, and set11 datasets:

- The CGT-extension method was more robust than BCD and ADMM in selecting initial weights and the number of layers within a fixed number of epochs. The CGT-extension method generally required more time than did back-propagation methods in the execution of a single epoch; however, CGT-extension provided a considerable advantage in terms of accuracy within the first few epochs, while achieving average execution times lower than those of the Adam algorithm.

The remainder of the paper is organized as follows. Section II briefly reviews related works. Section III presents the un-rectifying representation of a ReLU network and solutions by which to learn the networks. Section IV outlines experiments and comparisons with other methods and provides a demonstration of numerical convergence by the proposed algorithm. Concluding remarks are presented in Section V.

II. Related Works

A. Deep learning algorithms

Our review of deep learning algorithms deals with the gradient descent approach, alternating minimization (block coordinate descent), and the alternating direction method of multipliers (ADMM), as these are most widely used methods with theoretical support. The back-propagation algorithm [9] and its variants are the most popular deep learning algorithms, due to their adaptability to all types of feed-forward neural networks. However, this approach is susceptible to gradient instability and most of the methods designed to deal with this issue necessitate setting an appropriate range of values for the updating of parameters (particularly the learning rate). Other techniques that counter gradient instability are reviewed in [10].

Deep learning problems can be formulated using the two-splitting and three-splitting approaches, which allows the application of alternating optimization whenever a single block of variables is minimized, while the other variables remain fixed [3]. The fundamental principle underlying two-splitting and three-splitting is the introduction of auxiliary variables to separate highly-coupled DNN structures into separate sub-components and equality constraints [11]–[13]. The penalty method is then used to transform the constrained problem into an un-constrained problem, in which equality constraints are replaced as penalty terms in the objective function.

The ADMM approach, which is closely related to the augmented Lagrangian approach, has recently received considerable attention for DNN learning, due to its ease of implementation and proven effectiveness in optimizing sums of simple but non-smooth convex functions in situations where high accuracy is not a strict requirement [14]. The convergence of non-convex ADMM under a 2-block linear constrain is outlined in [15]. Recent breakthroughs in convergences of non-convex ADMMs under multi-block linear and multi-affine constraints are respectively presented in [16] and [4]. Thus, the DNN problem can be formulated/approximated as a non-convex optimization problem with multi-affine constraints [4] [17].

B. CGT-algorithm

Conn, Gould, and Toint [6] considered the problem of finding a local minimizer for the following function:

$$f(\mathbf{x}^1, \mathbf{x}^2)\colon \mathbb{R}^n \to \mathbb{R}, \text{ where } \mathbf{x}^1 \in \mathbb{R}^{n_1} \text{ and } \mathbf{x}^2 \in \mathbb{R}^{n_2} \quad (5)$$

where $n = n_1 + n_2$ and $(\mathbf{x}^1, \mathbf{x}^2)$ must satisfy

$$c_i(\mathbf{x}^1, \mathbf{x}^2) = 0,\ 1 \le i \le m. \quad (6)$$

There is no constraint on $\mathbf{x}^1$, and any component x_j of $\mathbf{x}^2$ must satisfy the following simple bounds:

$$l_j \le x_j \le u_j \quad (7)$$

where $u_j > l_j$, $l_j \in \mathbb{R}$, and $u_j \in \mathbb{R} \cup \{\infty\}$. Here, B is used to denote the convex set of $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2)$, wherein $\mathbf{x}^2$ satisfies (7). The CGT-algorithm is a global convergence algorithm[1], which solves problems involving equality constraints and variables with simple bounds, in situations where the objective and all of the constraints are twice-continuously-differentiable (that is, f and c_i are in $\mathcal{C}^2$).

The Lagrangian function and augmented Lagrangian function of (5)-(7) with respect to constraints $\{c_i\}$ are respectively written as follows:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i c_i(\mathbf{x}), \quad (8)$$

and

$$L_\mu(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i c_i(\mathbf{x}) + \frac{1}{2\mu} \sum_{i=1}^{m} c_i(\mathbf{x})^2, \quad (9)$$

where $\{\lambda_i\}$ are the Lagrangian multipliers.

Global convergence of the CGT-algorithm can be achieved for problems in which the following three conditions are assumed:
(AS1) Objective $f(\mathbf{x})$ and constraints $\{c_i\}_i$ are twice-continuously-differentiable for $\mathbf{x} \in B$.

[1]Global convergence means that the iterators generated by the algorithm from any initial point converge to limit points for which necessary conditions of optimality hold.

(AS2) The iterators $\{\mathbf{x}^{(k)}\}$ under consideration lie within a closed, bounded domain, $\Omega \subset B$.
(AS3) J denotes the variables (with indices) in $\mathbf{x}^1$ and the variables (with indices) in $\mathbf{x}^2$, such that $i \in J$ implies that variable x_i is in $\mathbf{x}^1$ or $\mathbf{x}^2$ with a value strictly in the interior of the closed interval $[l_i, u_i]$.

The template of the CGT-algorithm is given in Algorithm 1. The CGT-algorithm solves (5)-(7), provided that the primal-

Algorithm 1 CGT-algorithm solving (5)-(7)

Input parameters: the Lagrangian multipliers $\mu^{(0)}$ of equality constraints and stopping conditions ω^* and η^*, $Iter$ (the maximum number of iterations), the value of ρ^0.
Let α denote primal variables and let $\mathbf{c}$ denote the vector of all equality constraints.
Set $k = 0$, $\omega_0 = \omega^{(0)} = 1$, $\eta_0 = \eta^{(0)} = 1$, and $\tau = 0.01$, as suggested in [6].
repeat
    repeat
        $\alpha^{(k)} \leftarrow$ Update primal variables α with fixed Lagrangian multipliers $\bar{\mu}^{(k)}$.
    until $\|\mathbf{P}(\alpha^{(k)}, \nabla_\alpha L^{(k)}_{\rho^{-1}})\| \leq \omega^{(k)}$
    if $\|\mathbf{c}(\alpha^{(k)})\| \leq \eta^{(k)}$ **then**
        [Test for convergence and update Lagrangian multiplier estimates].
        if $\|\mathbf{P}(\alpha^{(k)}, \nabla_\alpha L^{(k)}_{\rho^{-1}})\| \leq \omega^*$ and $\|\mathbf{c}(\alpha^{(k)})\| \leq \eta^*$ **then**
            Return.
        end if
        [Update dual variables by using (10)].
        $\rho^{(k+1)} = \rho^{(k)}$; $\beta = \min((\rho^{(k+1)})^{-1}, 0.1)$; $\omega^{(k+1)} = \omega^{(k)}\beta$; $\eta^{(k+1)} = \eta^{(k)}\beta^{0.9}$.
    else
        [Reduce the penalty parameter ρ^{-1} (or increase ρ)].
        $(\rho^{(k+1)})^{-1} = \tau(\rho^{(k)})^{-1}$.
        $\bar{\mu}^{(k+1)} = \bar{\mu}^{(k)}$; $\beta = \min((\rho^{(k+1)})^{-1}, 0.1)$.
        $\omega^{(k+1)} = \omega_0\beta$ and $\eta^{(k+1)} = \eta_0\beta^{0.1}$.
    end if
    Increase k by one.
until $k \geq Iter$

Remark 1. The value of $\rho^{(k)}$ is a non-decreasing function of k. The updating of dual variable μ associated with constraint $c(\alpha) = 0$ is performed as follows:

$$\mu^{(k+1)} \leftarrow \mu^{(k)} + \rho^{(k)} c(\alpha^{(k)}). \tag{10}$$

Remark 2. The initial values are as follows: ρ_1 and ρ_2 (1), ρ_3 and ρ_4 (100), c_1 (0.001), and c_2 (10^{-6}).

variable update converges to a critical point of (9) over B.

C. Learning using data-dependent slack variables

When employing the un-rectifying approach, the ReLU function becomes a data-dependent variable, which can be analyzed in the same manner as other variables. It is represented through the introduction of activation variable d with a value in $\{0, 1\}$. Variable d can be used to indicate the activeness of the ReLU. The value of d is set to 1 when the input is a positive scalar, and otherwise equal to 0. Note that using discrete values for the activation variable could hinder optimization, due to the fact that the discrete domain can render a combinatorial problem for which solutions cannot be efficiently derived. In the context of optimization, variable relaxation refers to the relaxation of discrete variables by replacing discrete values with continuous values. This kind of relaxation can be applied to (4), as long as the solutions to the original problem are equivalent to the problem with relaxed variables. Below, we review [7], in which it is asserted that relaxing the domain of the activation variable from discrete to a continuous closed interval $[0, 1]$ enables the application of the CGT-algorithm to train DNNs to achieve globally convergence to critical points of (5)-(7).

The inequalities in (4) can be converted into equations satisfying (6) through the introduction of data-dependent slack variables, which would double the number of variables for a given input (compared to the size of ReLUs). For each i, we obtain

$$\begin{cases} d_i x_i - s_i = 0, \\ (1 - d_i)x_i + t_i = 0, \\ d_i \in \{0, 1\}, \\ t_i, s_i \geq 0. \end{cases} \tag{11}$$

Slack variables t_i and s_i are associated with component x_i of input $\mathbf{x}$. Constraint $d_i \in \{0, 1\}$ is then relaxed to $d_i \in [0, 1]$, yielding

$$\begin{cases} d_i x_i - s_i = 0, \\ (1 - d_i)x_i + t_i = 0, \\ d_i \in [0, 1], \\ t_i, s_i \geq 0. \end{cases} \tag{12}$$

Un-rectifying involves replacing the ReLU function with a variable in a discrete domain. Relaxation involves relaxing the domain of the variable, which allows it to be treated as a continuous variable. This cannot be regarded as a smoothing of the ReLU function, because most smoothing technique involve a trade-off between accuracy and computational efficiency. From another perspective, unless un-rectification and relaxation are applied to a ReLU, the status of the function is discrete (i.e., active or in-active), with a value depending exclusively on the input of the function. Un-rectifying and relaxing a ReLU transforms the function into a constrained continuous variable. This facilitates optimization by allowing continuous variables to be optimized in the same way as other variables, based on its most recent input and output values.

III. Learning un-rectified networks

We first consider the regression problem used to learn a 2-layer representation $\mathcal{M}_2$ from N training data $\{(\mathbf{x}_j \in \mathbb{R}^{N_0}, \mathbf{y}_j \in \mathbb{R}^{N_2})\}_{j=1}^N$. Imposing the Frobenius norm on the

weight matrices to stabilize the outputs with respect to input perturbation [5] renders the least squares regression problem as follows:

$$\sum_j \|\mathbf{y}_j - \mathcal{M}_2\mathbf{x}_j\|^2 + \frac{c_1}{2}(\|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2), \tag{13}$$

where $c_1 > 0$ is a parameter, $\mathbf{W}_i \in \mathbb{R}^{N_i \times N_{i-1}}$ are weight matrices, and $\mathbf{x}_j \in \mathbb{R}^{N_0}$ is the input. We un-rectify the ReLU in $\mathcal{M}_2$ by replacing ρ with data-dependent diagonal matrices $\mathbf{D}_j$ and add the term $\frac{c_2}{2}\sum_j \|\mathbf{D}_j\|_F^2$ to (13), in which the value of c_2 is small enough that its effect on the solution of (13) is negligible. We should see that including the term produces a unique closed-form solution of variables in $\mathbf{D}_j$, where $\mathbf{D}_j$ is defined as

$$\mathbf{D}_j M_1\mathbf{x}_j = \rho M_1\mathbf{x}_j \tag{14}$$

and the i-th diagonal entry of $\mathbf{D}_j$ is denoted as

$$d_{ji} = \begin{cases} 1 & \text{if } (M_1\mathbf{x})_i = (\mathbf{W}_1\mathbf{x}_j + \mathbf{b}_1)_i > 0, \\ 0 & \text{else.} \end{cases} \tag{15}$$

By introducing a small deviation of $\epsilon > 0$ from 0, we then formulate the constrained optimization problem with the aim of obtaining $\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, \{d_{ji}\}, \{z_{ji}\}$ at critical points of the following:

$$\text{(P1)} \begin{cases} \sum_j \|\mathbf{y}_j - M_2\mathbf{D}_j M_1\mathbf{x}_j\|^2 + \frac{c_1}{2}(\|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2) \\ +\frac{c_2}{2}\sum_j \|\mathbf{D}_j\|_F^2, \\ \text{subjected to} \begin{cases} -d_{ji}(M_1\mathbf{x}_j)_i + \max(0, z_{ji}) = \epsilon, \\ z_{ji} = (M_1\mathbf{x}_j)_i, \\ d_{ji} \in [0,1]. \end{cases} \end{cases}$$

The data-independent variables are $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, and $\mathbf{b}_2$, whereas the data-dependent variables are d_{ji} and z_{ji}, where $j = 1, \cdots N$ and $i = 1, \cdots, N_1$. If $\bar{\lambda}_{ji} \in \mathbb{R}$ and $\bar{\mu}_{ji} \in \mathbb{R}$ respectively denote the Lagrangian multipliers of the first and second constraints in (P1), then the augmented Lagrangian function for (P1) can be written as follows:

$$\begin{aligned} &\sum_j \|\mathbf{y}_j - M_2\mathbf{D}_j M_1\mathbf{x}_j\|^2 + \frac{c_1}{2}(\|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2) \\ &+ \frac{c_2}{2}\sum_j \|\mathbf{D}_j\|_F^2 + \sum_{ji}[\bar{\lambda}_{ji}(-d_{ji}(M_1\mathbf{x}_j)_i + \max(0, z_{ji}) - \epsilon] \\ &+ \frac{\rho_1^2}{2}\| - d_{ji}(M_1\mathbf{x}_j)_i + \max(0, z_{ji}) - \epsilon\|^2] \\ &+ \sum_{ji}[\bar{\mu}_{ji}(z_{ji} - (M_1\mathbf{x}_j)_i) + \sum_j \frac{\rho_2^2}{2}\|z_{ji} - (M_1\mathbf{x}_j)_i\|^2] \\ &+ \sum_{ji} \iota_{[0,1]} d_{ji}, \end{aligned} \tag{16}$$

where $\rho_1^2 > 0$ and $\rho_2^2 > 0$ are parameters and $\iota_{[0,1]}$ is the indicator function of $[0,1]$. To simplify the notation, we let $\bar{\theta} = \{c_1, c_2, \rho_1^2, \rho_2^2\}$, $\bar{d} = \{d_{ji}\}$, $\bar{z} = \{z_{ji}\}$, $\bar{\lambda} = \{\bar{\lambda}_{ji}\}$ and $\bar{\mu} = \{\bar{\mu}_{ji}\}$ and let $L_{\bar{\theta}}$ denote the above augmented Lagrangian function.

The primal variables in $L_{\bar{\theta}}$ include unbounded variables $\mathbf{W}_2$, $\mathbf{b}_2$, $\mathbf{W}_1$, $\mathbf{b}_1$, $\bar{z}$, where $\mathbf{b}_i$ are bias terms, and simple bounded variables $\bar{d}$. All primal variable except $\bar{z}$ can be alternatively updated to reach a local minimum of the augmented Lagrangian function in (reverse) from layer 2 to layer 1 when dual variables are fixed. This is evidenced by the fact that the augmented Lagrangian function is a strongly quadratic convex function when viewed as a function of any of the above variables except for $\bar{z}$ with the other variables fixed.

We can streamline the derivation of learning an L-layer ReLU network $\mathcal{M}_L\colon\ \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ from input data, as in the 2-layer case. Consider learning an L-layer ReLU network $\mathcal{M}_L\colon \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ based on $\{(\mathbf{x}_j, \mathbf{y}_j) | \mathbf{x}_j \in \mathbb{R}^{N_0}, \mathbf{y}_j \in \mathbb{R}^{N_L}\}_{j=1}^N$ input data. After un-rectifying ρs using data-dependent diagonal matrices $\{\mathbf{D}_j^1, \cdots, \mathbf{D}_j^L\}$ with constraints for input $\mathbf{x}_j$, we impose regularization on the Frobenius norm of weighting matrices, and introduce the quadratic term $\sum_{j,k} \|\mathbf{D}_j^k\|^2$, such that a unique closed-form solution to $\mathbf{D}_j^k$ can be obtained. Thus, the problem of learning (1) can be formulated as the process of seeking $\{\mathbf{W}_k\}_k, \{\mathbf{b}_k\}_k, \{\mathbf{D}_j^k\}_{k,j}$, and $\{z_{ji}^k\}_{i,j,k}$ as critical points for

$$\text{(PL)} \begin{cases} \sum_{j=1}^N \|\mathbf{y}_j - \mathcal{M}_L\mathbf{x}_j\|^2 + \frac{c_1}{2}\sum_{l=1}^L \|\mathbf{W}_l\|_F^2 \\ \qquad + \frac{c_2}{2}\sum_{j=1}^N \sum_{k=1}^{L-1} \|\mathbf{D}_j^k\|_F^2 \\ \text{subjected to} \begin{cases} -d_{ji}^k(M_k\mathbf{D}_j^{k-1}\cdots\mathbf{D}_j^1 M_1\mathbf{x}_j)_i \\ \qquad + \max(0, z_{ji}^k) = \epsilon, \\ z_{ji}^k = (M_k\mathbf{D}_j^{k-1}\cdots\mathbf{D}_j^1 M_1\mathbf{x}_j)_i, \\ d_{ji}^k \in [0,1], \end{cases} \end{cases}$$

where $j = 1, \cdots, N$, $k = 1, \cdots, L-1$, i is layer-dependent where $i = 1, \cdots, N_k$, c_2 is a very small value, and $\mathbf{D}_j^k = \mathrm{diag}([d_{ji}^k]_i)$. The augmented Lagrangian function can be obtained from the objective function and the first two sets of constraints. Analogous to the 2-layer case, the global convergence of L-layer ReLU networks learning can be derived by alternately updating the primal variables to reach critical points of the augmented function of (PL) with Lagrangian multipliers fixed.

Analogous to (2), problem (PL) can be formulated using the three-splitting approach. The CGT- algorithm that is used to solve the three-splitting formulation is similar to its counterpart algorithm in [7], omitted in the current study.

IV. Experiments

In Section IV-A, we choose the initial weights and numbers of layers by which to train networks (1) effectively. In Sections IV-B and IV-C, we compare the respective performance of classification and compressed sensing recovery problems involving a three-layer ReLU network $\mathcal{M}_3$ (fully-connected network with hidden layers comprising $1,000$ hidden units), using full-batch training, including MNIST, fashion-MNIST, and set11 (natural images) datasets as well as the initial weights and number of layer discussed in Section IV-A.

In this section, CGT-extension denotes the proposed method; BCD denotes the block-coordinate-descent method with three-

splitting formulation [3][2]; dlADMM denotes the ADMM method proposed for deep learning [18][3]; and Adam denotes the Adam optimizer (Pytorch optimization package), which is a state-of-the-art implementation of the back-propagation algorithm. The parameters unique to each method were carefully selected or set at default values to enable fair comparisons.

A. Initialization phase

Initial weights are typically obtained via sampling from an independently identically distributed (i.i.d) zero-mean Gaussian random variable of given variance. To provide a fair basis for comparison, we first conducted experiments aimed at determining the variance of an i.i.d. zero-mean Gaussian random variable for initial weights that work well with all selected methods. To access robustness against the number of hidden layers, we also conducted experiments using various numbers of hidden layers using the following simulation:

Let $Y \in \mathbb{R}^{784\times 60{,}000}$ denote the output of a two-layer ReLU network using the MNIST dataset as an input. Let $X \in \mathbb{R}^{784\times 60{,}000}$ denote the input and let σ denote the ReLU activation functions. Note that $Y = \bar{W}_2\sigma\bar{W}_1X$, where $\bar{W}_1 \in \mathbb{R}^{784\times 784}$ and $\bar{W}_2 \in \mathbb{R}^{784\times 784}$ are Gaussian matrices. Each element in a matrix was sampled from an i.i.d. Gaussian random variable with the mean set at zero and standard deviation set at one.

We then trained three-layer and eight-layer ReLU networks, $\mathcal{M}_3$ and $\mathcal{M}_8$, with $1,000$ hidden units in each hidden layer based on objective functions $\|X-\mathcal{M}_3Y\|_F^2$ and $\|X-\mathcal{M}_8Y\|_F^2$, where the training input is Y and output is X. We then compared the performance of the resulting networks when the initial weights were sampled from the i.i.d. zero-mean Gaussian random variable with the standard deviation of the weights respectively set at 0.1, 0.05, and 0.01.

Figures 1 and 2 compare the reconstruction quality of an image from the MNIST dataset using various initial weights and numbers of hidden layers.

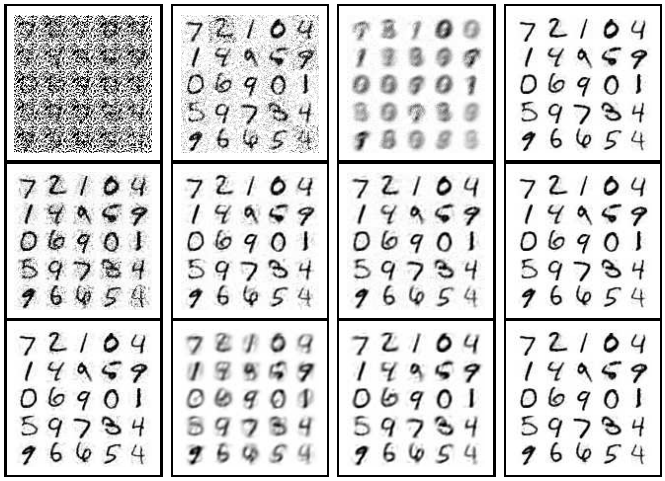
B. Classification

Both the MNIST [19] and Fashion-MNIST datasets contain ten classes of images, including $60,000$ training samples and $10,000$ testing samples with 784 features each. The loss function was deterministic cross-entropy loss. The number of learning epochs was set at 200 and the batch size was set at $60,000$. The learning rate of the Adam optimizer was at 0.001 for both MNIST and Fashion-MNIST datasets.

Figures 3 and 4 respectively present curves indicating the testing accuracy of the baselines and CGT-extension method. The CGT-extension method outperformed the baseline values for both datasets. Table I lists the time required to complete classification of the MNIST and Fashion-MNIST datasets. The CGT-extension method achieved high accuracy classification in fewer epochs than was required by the other methods. Note that

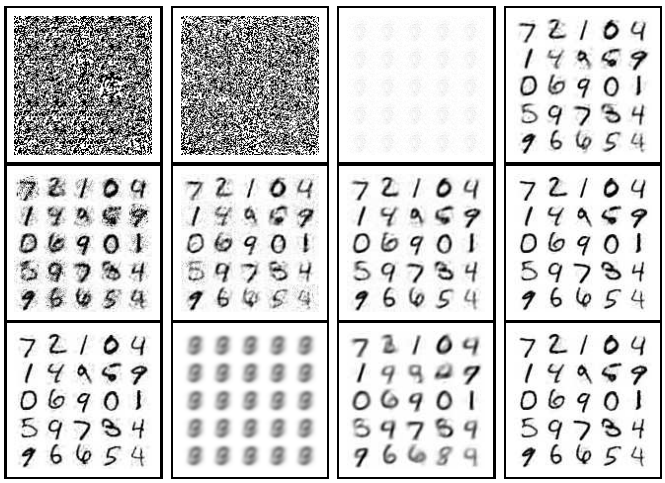
[2]The code of BCD is publicly available at "https://github.com/timlautk/BCD-for-DNNs-PyTorch"

[3]The code of dlADMM is publicly available at "https://github.com/xianggebenben/dlADMM"



(a) BCD (b) dlADMM (c) Adam (d) CGT-ext

Fig. 1: Image recovery performance based on reconstruction using three-layer network $\mathcal{M}_3$ based on (a) BCD, (b) dlADMM, (c) Adam, and (d) CGT-extension methods. The standard deviations in the initial weights were as follows: (first row) 0.1, (second row) 0.05, and (third row) 0.01. Only the CGT-extension method generated recognizable images regardless of the standard deviation. Standard deviation of 0.01 was deemed acceptable for the BCD method and 0.05 was acceptable for the others.



(a) BCD (b) dlADMM (c) Adam (d) CGT-ext

Fig. 2: Image recovery performance in terms of reconstruction using eight-layer network $\mathcal{M}_8$ based on (a) BCD, (b) dlADMM, (c) Adam, and (d) CGT-extension methods. The standard deviation in the initial weights were as follows: (first row) 0.1, (second row) 0.05, and (third row) 0.01. Only the CGT-extension method generated recognizable images regardless of the standard deviation. Standard deviation of 0.01 was deemed acceptable for the BCD method and 0.05 was acceptable for the others.

CGT-extension required far more time than did the Adam in increasing accuracy from 0.96 to 0.97. This indicates that the learning times could be reduced by implementing two methods within a hybrid system in which CGT-extension is used in the early stage and the Adam optimizer is used at later stage.

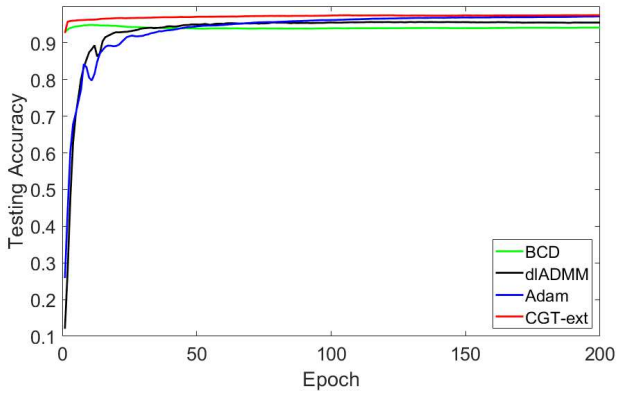


Fig. 3: Classification performance when applied to the MNIST dataset. BCD and CGT-extension methods both achieved accuracy of 0.9 in the early stage. In subsequent iterations, CGT-extension increased steadily towards 1.0; however, the BCD method improved little. Note that the curves of dlADMM and Adam rose slowly. The CGT-extension method converged far more quickly than did the Adam optimizer; however, the ultimate performance was similar.

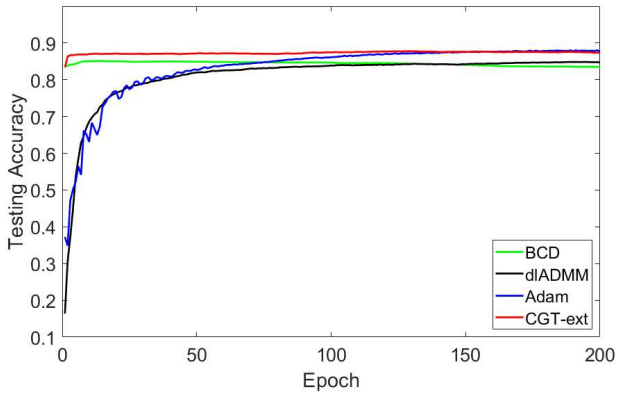


Fig. 4: Classification performance when applied to the Fashion-MNIST dataset. BCD and CGT-extension methods both exceeded 0.8 in the early stage. In subsequent iterations, CGT-extension increased steadily towards 0.9; however, the BCD method improved little. Note that the curves of the dlADMM and Adam methods did not rise as quickly as those of the CGT-extension and BCD methods. The CGT-extension method converged far more quickly than did the Adam optimizer; however, the ultimate performance was similar.

C. Compressed sensing recovery

For a given $m \times n$ sensing matrix $\mathbf{A}$ in which $m < n$ and measurement $\mathbf{y} \in \mathbb{R}^m$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$, the compressed sensing recovery problem involves recovering the sparsest signal $\mathbf{x} \in \mathbb{R}^n$ from $\mathbf{y}$. This is an ill-posed inverse problem, as $n > m$, where the CS ratio is defined as $\frac{m}{n}$. In the current study, we adopted the approach proposed in [20]–[22] in which a DNN learns a deep ReLU network mapping sparse vectors from measurements, where sensing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a Gaussian random matrix with elements sampled from i.i.d. Gaussian random variables with zero mean and variance $\frac{1}{m}$. Our training data was $(\mathbf{A}^+\mathbf{Y}, \mathbf{X})$, where $\mathbf{X}$ was an array with sparse vectors in the columns, $\mathbf{A}^+$ was the pseudo-inverse of $\mathbf{A}$, $\mathbf{A}\mathbf{A}^+ = \mathbf{I}$, and arrays $\mathbf{Y}$ and $\mathbf{X}$ were related to each other via $\mathbf{Y} = \mathbf{A}\mathbf{X}$. The ReLU network $\mathcal{M}_3$ (three-layer, each hidden layer with 1000 hidden units) was learned with the aim of optimizing the least-squares regression problem as follows:

$$\|\mathbf{X} - \mathcal{M}_3\mathbf{A}^+\mathbf{Y}\|_F^2. \tag{17}$$

$\mathcal{M}_3$ is un-rectified and then learned using Algorithm 1. In testing, we obtain the vector of $\mathbf{y}$ with respect to $\mathbf{A}$ via forward-inference using $\hat{\mathbf{x}} = \mathcal{M}_3\mathbf{A}^+\mathbf{y}$. We compared $\mathcal{M}_3$ derived using various learning methods in terms of compressed sensing recovery performance using experiments based on the MNIST and set11 datasets.

In the MNIST dataset, n was set at 784 and the value of m varied (10, 25, 100, or 200). Note that these values respectively correspond to CS ratios of 1%, 3%, 13%, and 26%. The number of learning epochs was set at 300 and the batch size was set at $60,000$. Figure 5 illustrates the reconstruction errors in terms of MSE versus CS ratio in the recovery of MNIST testing data. Figure 6 illustrates the perceptual quality of the reconstructed images. Consistent with the numerical results shown in Figure 5, the visual quality of the reconstructed images improved with an increase in the number of measurements.

Experiment using Set11 used natural images for CS recovery. The 91 images in [22] were used for training[4]. Patches (32×32 pixels) were extracted from the luminance components of the images using a stride of 4 in each dimension. The total number of training patches was $117,242$. The 11 images in Set11 were used for testing. $58,523$ testing patches were obtained in a manner similar to that used in the extraction of training patches. The size of image patch n was set at 1024 px and the size of m varied (10, 51, 102, 153, 256, or 512 px). Note that these values respectively correspond to CS ratios of 1%, 5%, 10%, 15%, 25%, and 50%. Let $\mathbf{X}$ and $\mathbf{Y}$ denote the sets of training patches and their respective measurements. Three-layer ReLU network $\mathcal{M}_3$ was learned using $\mathbf{Y}$ as the input and $\mathbf{X}$ as the desired output. The number of epochs was set at 500 and the batch size was set at $117,242$. Reconstructed (testing) images were obtained by averaging the overlapped patches and then measuring the

[4]The images are publicly available from "SRCNN-Tensorflow/Train at master · tegg89/SRCNN-Tensorflow-GitHub"

TABLE I: Average classification times (in seconds) when applied to the MNIST dataset and Fashion-MNIST datasets (best performance in bold). In the table, "x" indicates that a given method was unable to reach that level of accuracy.

Method	MNIST				Fashion-MNIST			
	0.94	0.95	0.96	0.97	0.84	0.85	0.86	0.87
BCD	**7.0**	x	x	x	**3.5**	**17.1**	x	x
dlADMM	276.2	421.4	x	x	1032.0	x	x	x
Adam	55.4	76.4	111.8	**209.8**	74.6	95.4	115.0	**146.8**
CGT-ext	48.1	**71.2**	**94.2**	1082.6	50.8	50.8	**73.9**	441.8

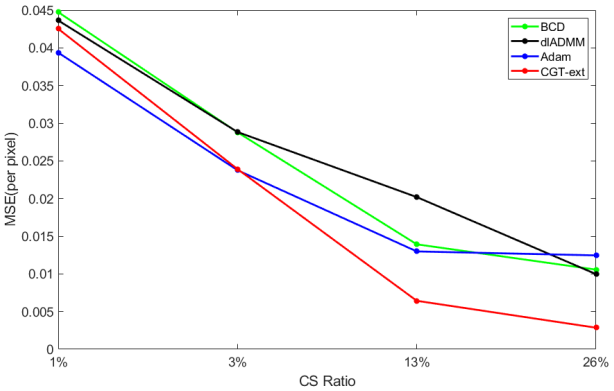
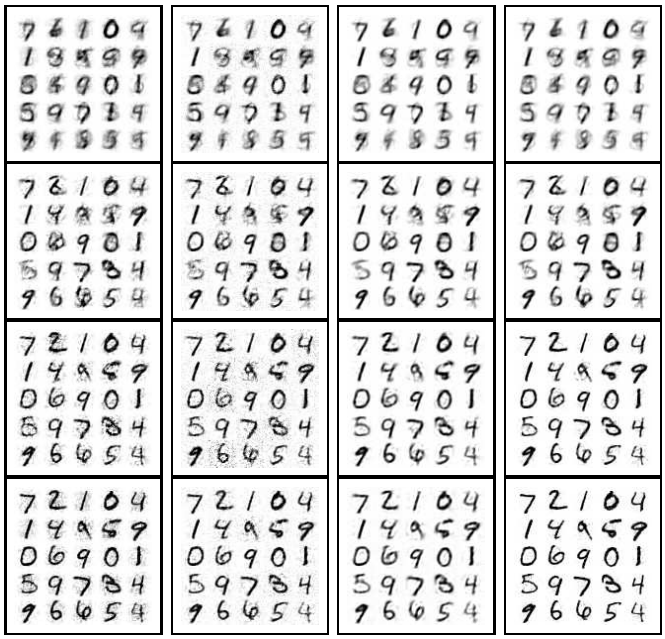


Fig. 5: Comparison of average MSE (per pixel) from MNIST testing images versus CS ratio for a compressed sensing recovery problem with the learning rate of the Adam optimizer set at 0.001, the BCD and dlADMM parameters set at default values, and the initial CGT-extension parameters set int accordance with Remark 2 in Algorithm 1.



(a) BCD (b) dlADMM (c) Adam (d) CGT-ext

Fig. 6: Quality of images reconstructed from (a) BCD, (b) dlADMM, (c) Adam, and (d) CGT-extension methods. CS-ratio: (first row) 1%, (second row) 3%, (third row) 13%, and (fourth row) 26%. When the CS ratio fell below 26% (corresponding to $m = 200$), the Adam optimizer and the CGT-extension methods achieved perceptual quality superior to that of the BCD and dlADMM. Nonetheless, at a CS ratio 26%, all four methods yielded images of consistently high perceptual quality.

average peak-to-noise ratio (PSNR) in Table II. CGT-extension achieved the best performance regardless of CS ratio. Figure 7 presents images ($Parrot$) reconstructed with the CS ratio set at 10% and 50%. The perceptual quality matched the results in Table II. Overall, the CGT-extension method (CS ratio = 50%) preserved more of the details, while providing sharper edges. At a CS ratio of 10%, the perceptual quality of the BCD and proposed methods were roughly the same.

TABLE II: Average PSNR(dB) in the recovery of images in Set11 using varying CS ratios (best performances in bold).

Method	CS ratio					
	1%	5%	10%	15%	25%	50%
BCD	19.58	23.60	25.71	27.50	28.94	30.75
dlADMM	19.59	23.48	24.65	26.15	25.99	25.94
Adam	19.67	22.67	23.52	23.79	23.76	23.92
CGT-ext	**19.71**	**23.67**	**25.94**	**27.67**	**29.99**	**34.22**

V. Conclusions

This paper demonstrates that deep learning problems can be solved using the augmented Lagrangian approach, based on our extension of the algorithm originally developed by Conn, Could, and Toint. Note that this conclusion is based on the application of un-rectification without introducing additional data-depependent slack variables, which would otherwise hamper the applications of this technique to large datasets and deep neural networks. Un-rectifying a ReLU network makes it possible to formulate a learning problem as a non-convex optimization problem with activation variables corresponding to un-rectifying ReLUs relaxed to data-dependent variables within domains confined to closed segments, and constraints expressed as continuous functions. We demonstrate that un-rectifying networks allows global convergence to local optima when the updating of primal variables is subjected to alternating minimization. We compared the performance of the proposed scheme with other state-of-the-art deep learning

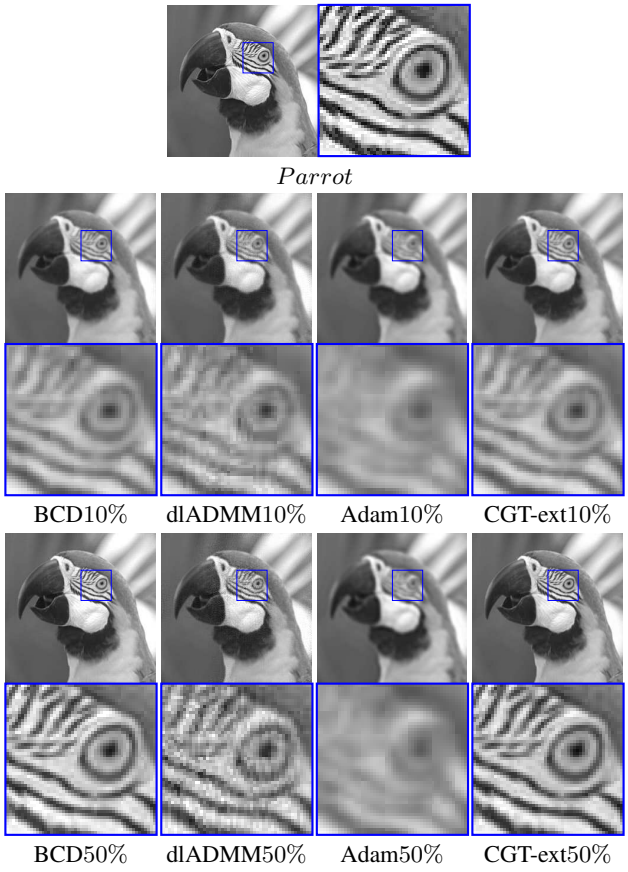


Fig. 7: Perceptual quality of reconstructed images in Set11 dataset with CS ratio set at 10% and 50% with the learning rate of the Adam optimizer set at 0.001 and the BCD and dlADMM parameters set at default values.

methods for classification and compressed sensing recovery problems. Our theoretical results assume a batch input. The question of whether theoretical results could be derived using mini-batch inputs is a subject worthy of further study.

REFERENCES

[1] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

[2] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, PMLR, 2011.

[3] J. Zeng, T. T.-K. Lau, S. Lin, and Y. Yao, "Global convergence of block coordinate descent in deep learning," *arXiv preprint arXiv:1803.00225*, 2018.

[4] W. Gao, D. Goldfarb, and F. E. Curtis, "Admm for multiaffine constrained optimization," *Optimization Methods and Software*, vol. 35, no. 2, pp. 257–303, 2020.

[5] W. L. Hwang and A. Heinecke, "Un-rectifying non-linear networks for signal representation," *IEEE Trans. on Signal Processing*, vol. 68, pp. 196–210, 2020.

[6] A. R. Conn, N. I. Gould, and P. Toint, "A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds," *SIAM Journal on Numerical Analysis*, vol. 28, no. 2, pp. 545–572, 1991.

[7] W.-L. Hwang and S.-S. Tung, "Learning dnn networks using un-rectifying relu with compressed sensing application," *arXiv preprint arXiv:2101.06940*, 2021.

[8] S.-S. Tung and W.-L. Hwang, "Un-rectifying relu deep neural network in compressed sensing recovery," in *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 1–2, 2021.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[10] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[11] Z. Zhang and M. Brand, "Convergent block coordinate descent for training tikhonov regularized deep neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1721–1730, 2017.

[12] F. Gu, A. Askari, and L. E. Ghaoui, "Fenchel lifted networks: A lagrange relaxation of neural network training," *International conference on machine learning (ICML)*, 2018.

[13] T. T.-K. Lau, J. Zeng, B. Wu, and Y. Yao, "A proximal block coordinate descent algorithm for deep neural network training," *arXiv preprint arXiv:1803.09082*, 2018.

[14] J. Eckstein and W. Yao, "Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives," *Pac. J. Optim.*, vol. 11, no. 4, pp. 619–644, 2015.

[15] S. Magnússon, P. C. Weeraddana, M. G. Rabbat, and C. Fischione, "On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 3, pp. 296–309, 2015.

[16] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.

[17] J. Zeng, S.-B. Lin, and Y. Yao, "A convergence analysis of nonlinearly constrained admm in deep learning," *arXiv preprint arXiv:1902.02060*, 2019.

[18] J. Wang, F. Yu, X. Chen, and L. Zhao, "Admm for efficient deep learning with global convergence," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2019.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] A. Mousavi and R. G. Baraniuk, "Learning to invert: Signal recovery via deep convolutional networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2272–2276, IEEE, 2017.

[21] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International Conference on Machine Learning*, pp. 537–546, PMLR, 2017.

[22] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1828–1837, 2018.

OpenPose Based Yoga Poses Difficulty Estimation for Dynamic and Static Yoga Exercises

Wan-Chia Huang*, Cheng-Liang Shih*, Irin Tri Anggraini#, Yanqi Xiao#, Nobuo Funabiki#, and Chih-Peng Fan*
* Department of Electrical Engineering, National Chung Hsing University, Taichung, Taiwan
# Department of Electrical and Communication Engineering, Okayama University, Okayama, Japan
E-mail: a0908652289@gmail.com, as5577557@gmail.com, pq7n2xt6@s.okayama-u.ac.jp, p8126fpo@s.okayama-u.ac.jp, funabiki@okayama-u.ac.jp, cpfan@dragon.nchu.edu.tw

***Abstract* — Assessing the difficulty of a Yoga pose is a challenging task, often dependent on individuals' subjective judgement or consultation with professionals to choose poses that fit their needs. In this study, a preliminary difficulty calculation method for estimating the difficulty of Yoga poses is proposed, and it is applied to the Yoga poses in the proposed system in order to provide a user-friendly difficulty reference. OpenPose is applied in the system to detect user's movements and extract the key-points information. In the experimental results, seven users are inquired to participate in practicing Yoga by the proposed system, and they are asked to provide the difficulty levels rating from 1 to 5. These subjective difficulty levels are used to compare the estimated difficulty levels. The results indicate a positive correlation between the subjective and estimated difficulty results.**

***keywords* — Yoga, difficulty, complexity, duration time, stability, angle difference, Euclidean distance, level.**

I. INTRODUCTION

Yoga is one of the most popular exercises nowadays. It requires less time and doesn't need to have large space, making it convenient for daily exercise at home. Yoga helps people maintain physical health and is suitable for all ages of users. There are numerous poses in Yoga, and each pose may vary in difficulty. However, the difficulty of Yoga poses is subjective and varies from different people. In this research, a preliminary method for estimating the difficulty of Yoga poses is proposed and applied to the Yoga poses in this system. It is used to provide a more objective difficulty assessment, allowing users to choose Yoga poses are more suitable for them in a more straightforward way.

In the system, the OpenPose [1-2] is applied for human body detection. OpenPose is applied in a wide research fields, such as self-learning system [3-4], exercise analysis [5], and medical applications [6]. This paper focuses on the design of a self-learning system for Yoga training. OpenPose is an open source library [7] provides various pose models for body detection, and the BODY25 model is used in this system. The 25 key-points of the human body can be extracted through the BODY25 model as shown in the Fig. 1, the key-points and skeleton information are rendered on the image of instructor.

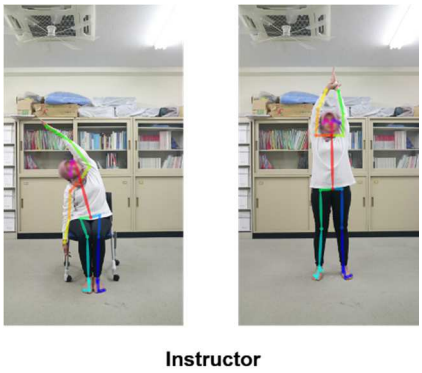


Fig. 1. Examples of BODY25 model detection

II. RELATED WORK

The five Yoga poses are provided from [3]. For each pose, only the main key-points is selected called selected key-points in [8]. The selected frames such as initial frame, keyframes and sync frames in each pose are defined in [9]. This research is an extended work of [10], the standard duration time, time distribution, the definition of static and dynamic regions of each pose is proposed. In [11], the researchers classified ten Yoga poses on six models, and they utilized angle information by using the tf-pose algorithm to extract the key-points of the poses. In [12], the researchers proposed a Yoga pose correction system, gave users suggests to help them correct their pose.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

III. PROPOSED METHOD

In the method, we use instructor's video to analyze the difficulty of each Yoga pose.

A. Elements of Estimation

In the proposed estimation method, there are three elements are considered to measure the difficulty of Yoga poses: complexity (**CP**), duration time (**DT**), and stability (**SB**).

From the 25 key-points extracted by OpenPose, key-points 3, 4, 6, 7 (upper body) and 10, 11, 13, 14 (lower body) [8] are selected as target key-points to observe the movement variations. CP is defined based on the number of the target key-points required in the pose, as shown in TABLE I. The more key-points are needed, the more body parts need attention during the Yoga.

TABLE I. COMPLEXITY OF EACH YOGA POSE

Pose	Complexity (CP)
Mountain	4
Seat_1	4
Seat_2	4
Side_bend	4
Warrior	8

In Yoga, we have divided the sequence into dynamic and static regions, and defined the movement time for each movement region as duration time [10]. When the required DT is longer, the user needs more concentration.

The stability is analyzed separately for the upper and lower body. In the upper body, the angle difference (AD) is calculated for each selected key-point, while in lower body, the Euclidean distance (ED) is calculated for each selected key-point as follows:

$$AD = \theta^{C}_{T(\alpha)} - \theta^{C}_{T(\beta)} \tag{1}$$

$$ED = \sqrt{\left(x_\alpha - x_\beta\right)^2 + \left(y_\alpha - y_\beta\right)^2} \tag{2}$$

In Eq. (1), the angel of the selected key-point (T) is relative to the center key-points (C) which are 2nd, 5th (upper body) and 9th, 12th (lower body). In Eq. (2), the x, y are the coordinates of the selected target key-points. In both equations, the subscripts α and β are the index of the instructor's frames will be discussed in the following chapters.

The angle difference values in the upper body and the distance values in the lower body of each selected key-point are summed up, referred to the total angle difference (TAD_{upper}) and the total distance (TD_{lower}) as follows:

$$TAD_{upper} = \sum_{i=0}^{N} \left| AD_{selected[i]} \right| \tag{3}$$

$$TD_{lower} = \sum_{i=0}^{N} ED_{selected[i]} \tag{4}$$

In Eq. (3) and Eq. (4), the N is the number of selected key-points in each Yoga pose.

According to different considerations regarding static poses and dynamic poses, the estimation method is divided into static poses difficulty estimation and dynamic poses difficulty estimation.

B. Static Poses Difficulty Estimation

The difficulty estimation of each static pose in the Yoga is as follows:

$$Difficulty_S = CP * \sqrt{DT_{Si}} * \begin{cases} SB_{s,upper} \\ SB_{s,lower} \end{cases} \tag{5}$$

In Eq. (5), the DT_{Si} is the duration time of each static region, the stability calculation is mapped to two situations (upper and lower body).

Fig. 2 shows an example of evaluating the stability of a static pose in the Warrior pose. In static poses, the stability is calculated by comparing the keyframes of the instructor with the 0th frame of the Yoga pose, to assess the stability of the static pose in the Yoga, the stability of upper body ($SB_{s,upper}$) and stability of lower body ($SB_{s,lower}$) can be obtained from Eq. (1) and Eq. (2) respectively, which α is 0 and β is 280. In a Yoga video, since there are multiple static regions, there will be multiple static difficulty values and stored into a list after calculation.

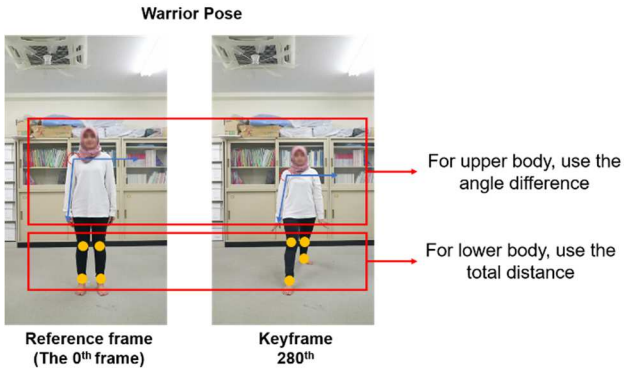


Fig. 2. Schematic of static poses difficulty estimation

C. *Dynamic Poses Difficulty Estimation*

The difficulty estimation of each dynamic pose in the Yoga is as follows:

$$Difficulty_D = CP * \sqrt{DT_{Di}} * \begin{cases} SB_{D,upper} \\ SB_{D,lower} \end{cases} \quad (6)$$

In Eq. (6), the DT_{Di} is the duration time of each dynamic region, the stability calculation is different from static poses, the pose velocity in the dynamic regions is considered as shown in Fig. 3.

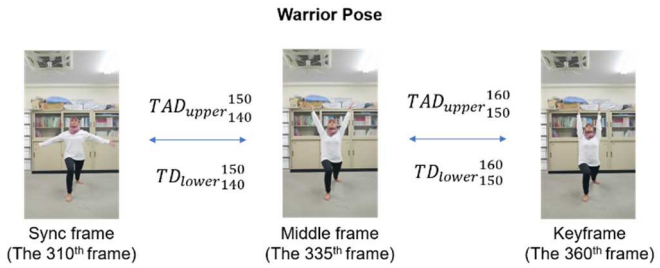


Fig. 3. Schematic of dynamic poses difficulty estimation

The specific frames (initial/key/sync frames) have defined in [9]. In a dynamic region, an additional frame is considered by including a middle frame to measure the pose speed as the dynamic stability. In Fig. 3, there is a dynamic region from the 310th frame to the 360th frame of Warrior pose. In this region, the total angle difference of the upper body and the total distance of the lower body are separately calculated between the sync frame and the middle frame, as well as between the middle frame and the keyframe. Then, these values are used to determine the stability of the upper and lower body in the dynamic region as the formula follows:

$$SB_D = \begin{cases} \frac{TAD_{upper}{}^{Middle}_{Start} + TAD_{upper}{}^{End}_{Middle}}{frame\ number}, upper\ body \\ \frac{TD_{lower}{}^{Middle}_{Start} + TD_{lower}{}^{End}_{Middle}}{frame\ number}, lower\ body \end{cases} \quad (7)$$

In Eq. (7), the frame number is the index difference of selected frame plus 1, in Fig. 3, the frame number is 51. $Start$ is the index of the initial frame or sync frames, $Middle$ is the index of the middle frame, and End is the index of the keyframes.

In a Yoga video, since there are multiple dynamic regions, there will be multiple dynamic difficulty values and stored into a list after calculation.

D. *Difficulty Normalization and Level Transformation*

Because the upper and lower body have different considerations and calculation methods for static and dynamic poses. After difficulty calculation, it needs to do the data normalization. The difficulty data normalization flow is as follows:

Step 1: Merge the body difficulty lists of static and dynamic poses in each Yoga pose, resulting in four merged lists as shown in Fig. 4.

Step 2: Normalize the difficulty values in each merged list to a range of 0 to 1.

Step 3: Divide each normalized merged list into segments based on the number of static and dynamic regions in each Yoga pose.

Step 4: Calculate the sum and average of each list.

Step 5: Finally, each Yoga pose will have four floating-point difficulty values representing the overall upper and lower body static difficulty, overall upper and lower body dynamic difficulty respectively.

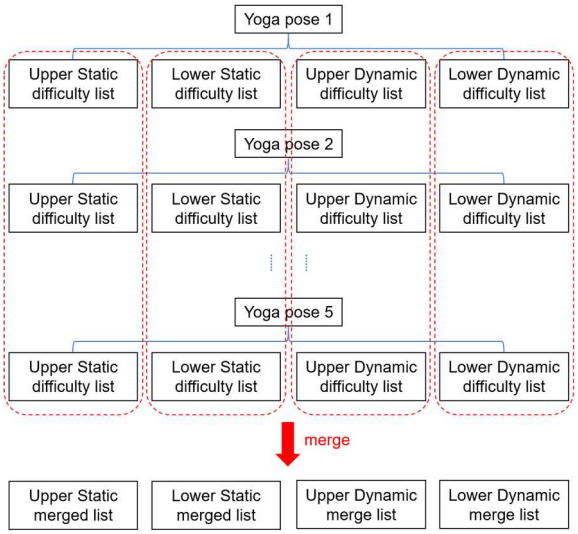


Fig. 4. Difficulty data merge flow

However, not all Yoga poses perform both upper and lower body parts. According to the TABLE II, the overall difficulty of body parts is selected in each Yoga pose. For example, the

overall upper body static and dynamic difficulty is selected in Mountain pose.

TABLE II. SELECTED OVERALL BODY DIFFICULTY OF EACH POSE

Pose	Selected Overall Body Difficulty
Mountain	Upper
Seat_1	Upper
Seat_2	Upper
Side_bend	Upper
Warrior	Upper and Lower

After selecting the corresponding body part difficulty, the overall body static and dynamic difficulty are averaged as pose difficulty, then, the pose difficulty is transformed into pose level ranging from 1 to 5 according to the TABLE III. This provides a more straightforward difficulty reference for users to choose the Yoga poses.

TABLE III. POSE DIFFICULTY TO LEVEL TRANSFORMATION

Pose	Pose Level
$0 \leq$ pose difficulty < 0.2	1
$0.2 \leq$ pose difficulty < 0.3	2
$0.3 \leq$ pose difficulty < 0.4	3
$0.4 \leq$ pose difficulty < 0.5	4
pose difficulty ≥ 0.5	5

IV. EXPERIMENTAL RESULTS

In the experimental results, five Yoga poses named Mountain (arm stretching and bending forward), Seat_1 (lateral stretching with both hands combined performed on chair), Seat_2 (lateral stretching with one arm performed on chair), Side_bend (long-time held side bend pose) and Warrior (full-body stretching movement) are used to estimate the difficulty. For Mountain, Seat_1, Seat_2 and Side_bend, they focus on the upper body movements, only the upper body difficulty is selected. However, for Warrior, it requires the coordination between the upper and lower body, both upper and lower body difficulties are selected.

In TABLE IV, it presents the dynamic and static difficulty for each Yoga pose, categorized into upper body (**U.**), lower body (**L.**) and the sum of the difficulty (**S.**). In the dynamic analysis of Warrior pose, the sum of the body difficulties exceeds 1, but the limit is 1. Therefore, it is reassigned as 1.

After obtaining the sum of body difficulties, the average difficulty of static and dynamic analysis is calculated to obtain the pose difficulty (**PD.**). Then, according to the TABLE III, the pose levels for each Yoga pose are determined as 2, 2, 2, 3, and 5 respectively.

TABLE IV. ESTIMATED DIFFICULTY OF EACH POSE

Estimated Difficulty							
Pose	Static			Dynamic			PD.
	U.	L.	S.	U.	L.	S.	
Mountain	0.12		0.12	0.33		0.33	0.225
Seta_1	0.23		0.23	0.36		0.36	0.295
Seat_2	0.17		0.17	0.41		0.41	0.29
Side_bend	0.45		0.45	0.17		0.17	0.31
Warrior	0.28	0.36	0.64	0.74	0.43	1	0.82

To compare the estimated pose level in this research, seven male users (M1~ M7) are invited to perform the Yoga poses and provide the subjective pose level from 1 to 5 as shown in TABLE V. According to the survey, the user's subjective pose levels are 2, 2, 2, 2, 3.

TABLE V. USER'S SUBJECTIVE DIFFICULTY LEVEL

Pose	Subjective Difficulty (1 ~ 5)				
	Mountain	Seat_1	Seat_2	Side_bend	Warrior
M1	2	2	3	4	4
M2	3	2	2	1	3
M3	2	2	2	1	3
M4	1	2	2	1.5	1.5
M5	3	2	2	2	2
M6	2	3	2	3	4
M7	2	2	3	1	4
Avg.	2.14	2.14	2.29	1.93	3.07
Round	2	2	2	2	3

To better observe the estimated pose levels and user's subjective pose levels, Fig. 5 shows the trend of the two data. It can be seen that in Mountain, Seat_1, and Seat_2 poses, the estimated pose levels are equal to the user's subjective experience. In Side_bend and Warrior poses, although there is some difference, the trends of the two data are similar.

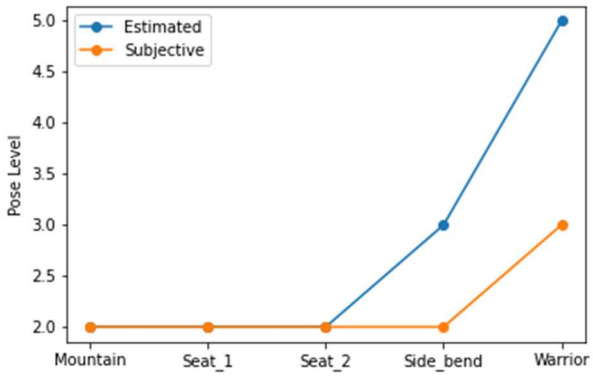


Fig. 5. Pose level trend between estimated and subjective results

V. Conclusions and Future Work

This study proposes a preliminary Yoga pose difficulty estimation method by considering key-points, movements duration time and the stability of the Yoga poses. The estimated pose levels are provided as a reference for users. The experimental results compare the estimated pose levels with the users' subjective experience, and it show the similarity in each other.

The difficulty estimation in this study applies on the Yoga instructor videos in the system, and the estimated difficulty results are relative. It will be better to improve the estimation method to be a more comprehensive approach and apply on more external Yoga data. In the user's survey, there are all young people, which may lead to the user survey a bias. It will be better to invite more middle-aged or elderly users to take the survey, make the user survey more diverse and comprehensive.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[2] Z. Cao, T. Simon, S. -E. Wei and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1302-1310, doi: 10.1109/CVPR.2017.143.

[3] I.T. Anggraini, A. Basuki, N. Funabiki, X. Lu, C. Fan, Y. Hsu, C. Lin "A Proposal of Exercise and Performance Learning Assistant System for Self-Practice at Home", 103 Advances in Science, Technology and Engineering Systems Journal, 2020, vol. 5, no. 5, pp. 1196-1203.

[4] M. C. Thar, K. Z. N. Winn and N. Funabiki, "A Proposal of Yoga Pose Assessment Method Using Pose Detection for Self-Learning," 2019 International Conference on Advanced Information Technologies (ICAIT), Yangon, Myanmar, 2019, pp. 137-142, doi: 10.1109/AITC.2019.8920892.

[5] Y. -C. Li, C. -T. Chang, C. -C. Cheng and Y. -L. Huang, "Baseball Swing Pose Estimation Using OpenPose," 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI), Hong Kong, Hong Kong, 2021, pp. 6-9, doi: 10.1109/RAAI52226.2021.9507807.

[6] Y. -G. Wu et al., "Interactive Intelligent Exercise System Design for the Elderly," 2022 International Conference on Engineering and Emerging Technologies (ICEET), Kuala Lumpur, Malaysia, 2022, pp. 1-5, doi: 10.1109/ICEET56468.2022.10007223.

[7] G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, H. Joo, and Y. Sh eikh, "OpenPose library," [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[8] S. -W. Shen, W. -C. Huang, I. T. Anggraini, N. Funabiki and C. -P. Fan, "Exercise and Performance Learning Assistant System for Self-Practice Dynamic Yoga by OpenPose and Fuzzy Based Design," 2022 10th International Conference on Information and Education Technology (ICIET), Matsue, Japan, 2022, pp. 16-21, doi: 10.1109/ICIET55102.2022.9778954.

[9] S. -W. Shen, W. -C. Huang, I. T. Anggraini, N. Funabiki and C. -P. Fan, " Design of OpenPose-Based of Exercise Assistant System with Instructor-User Synchronization for Self-Practice Dynamic Yoga," The 10th International Conference on Computer and Communications Management (ICCCM), 2022.

[10] W. -C. Huang, C. -L. Shih, I. T. Anggraini, N. Funabiki and C. -P. Fan, "OpenPose Technology Based Yoga Exercise Guidance Functions by Hint Messages and Scores Evaluation for Dynamic and Static Yoga Postures," 2023 4th International Conference on Information Technology and Education Technology (ITET 2023), May 2023.

[11] Y. Agrawal, Y. Shah and A. Sharma, "Implementation of Machine Learning Technique for Identification of Yoga Poses," 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), Gwalior, India, 2020, pp. 40-43, doi: 10.1109/CSNT48778.2020.9115758.

[12] A. Chaudhari, O. Dalvi, O. Ramade and D. Ambawade, "Yog-Guru: Real-Time Yoga Pose Correction System Using Deep Learning Methods," 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 2021, pp. 1-6, doi: 10.1109/ICCICT50803.2021.9509937.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multimodal Multifaceted Music Emotion Recognition Based on Self-Attentive Fusion of Psychology-Inspired Symbolic and Acoustic Features

Jiahao Zhao and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Kyoto, Japan
E-mail: {jzhao, yoshii}@sap.ist.i.kyoto-u.ac.jp

***Abstract*—This paper describes automatic music emotion recognition (MER) that aims to estimate the valence and arousal (V/A) scores of a piece of piano music. The emotion is multifaceted in nature; it is rendered by various features that are often mutually dependent and inherent in music composition and performance. A basic approach to MER is to train a deep neural network (DNN) that extracts latent features representing the emotion as a whole and estimates the V/A scores, using only a limited amount of audio data with imbalanced V/A annotations. Such a black-box approach, however, suffers from limited performance and interpretability. To overcome these limitations, in this paper we propose a multimodal multifaceted MER method that fuses various kinds of musically-meaningful symbolic and acoustic features extracted from both MIDI and audio data, respectively, based on the expert knowledge of musical psychology. More specifically, our method separately extracts the affective features representing the rhythm, dynamics, melody, harmony, and tone color of a piano piece as the main factors affecting the emotion and integrates them with a self-attention mechanism that can learn the complicated cross-modal relationships. The experiments using the common EMOPIA dataset showed that the proposed model achieved the state-of-the-art V/A classification accuracy of 69.2% and that the multimodal and multifaceted feature fusions contributed to the performance improvement.**

I. Introduction

Music emotion refers to the affective information conveyed through music and is considered the essence of music itself. It is thus necessary to analyze music data from both objective and subjective viewpoints. As well as automatic music transcription (AMT), music emotion recognition (MER) has been a long-standing fundamental task in the field of music information retrieval (MIR) [1], because it plays a crucial role in emotion-based personalized music retrieval and recommendation systems. The abstract nature of the emotion and the expensive cost of collecting subjective annotations, however, pose a challenge to the development of MER methods.

Music emotion is multifaceted and multimodal in nature; various features of music composition and performance may implicitly interact to render an emotion in a sophisticated manner unique to the musical piece. This makes deep learning (DL) methods [2]–[4] hard to learn a consistent mapping from a specific type of music representation (e.g., score, lyrics, or audio) to the overall emotion of the piece, resulting in the limited performance of the valence and arousal (V/A) estimation. To mitigate this problem, for instance, Berardinis et al. [5] separately extracted affective acoustic features from the vocal, bass, drum, and the other parts of a music signal obtained with a source separation method. Zhao et al. [6] separately extracted text and acoustic features from the text (i.e., lyrics and track and artist names) and audio data and fused them one by one with a cross-modal attention mechanism. However, these existing methods still have some problems: on the one hand, these methods have high requirements for experimental data and can only be applied to specific music; on the other hand, the performance and interpretability of these methods are still unsatisfactory.

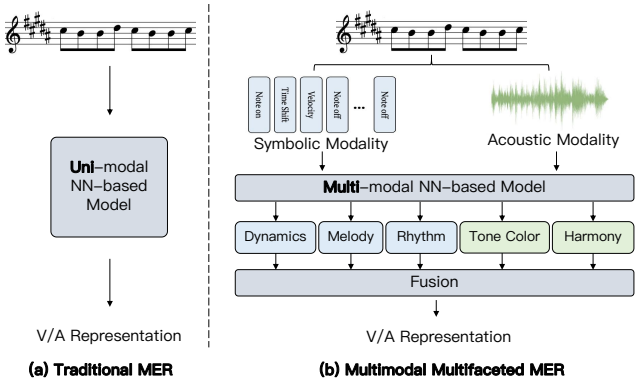


Fig. 1. Comparison between the conventional and proposed approaches.

For interpretable MER, we refer to the investigations of affective features in the fields of music psychology and affective computing. Laukka et al. [7], for example, proposed a human perception model for instrumental music and found that six factors, i.e., dynamics, rhythm, timbre, register, tonality, and structure, characterize the emotional expression of music. Renato [8] revisited this issue from a computational perspective and found eight factors useful for MER, i.e., melody, harmony, rhythm, dynamics, tone color, expressivity, texture, and form. These findings were supported by several empirical studies and experimental results [9], [10]. The superiority of a multifaceted approach was also shown in [11], [12].

In this paper, we propose a multimodal multifaceted MER method for piano music that fuses various kinds of musically-meaningful symbolic, and acoustic features separately extracted from MIDI and audio data, respectively (Fig. 1). Using the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

findings of musical psychology for the network design, a deep neural network (DNN) is expected to be trained effectively with a limited amount of multimodal data. Specifically, our method separately extracts affective features regarding particularly important five factors, i.e., rhythm, dynamics, melody, harmony, and tone color. At the heart of our method is to use a self-attention mechanism that can learn the complicated relationships between different factors and modalities for integrating the extracted features in a weighted manner depending on the musical piece. We show that the proposed approach can achieve the state-of-the-art performance of V/A estimation on the public dataset called EMOPIA [13].

II. Music Emotion Revisited

This section describes five major factors that are considered to play key roles in emotional expression in piano music and which aspects of emotional content are affected. Our theory is based on the findings of recent MER studies and the knowledge in music psychology.

Rhythm refers to the temporal organization and pattern of notes and rests. It determines the timing and duration of each note, creating a sense of pulse, meter, and musical flow. Rhythmic information is usually described as beat, tempo, or accentuation. Rhythm is considered one of the most important elements in emotional expression, it is mainly associated with both the excitement and the happiness of the music.

Dynamics refers to the variations in volume and intensity. It involves the manipulation of loudness and softness to create expressive and emotional effects. It adds depth, contrast, and nuance to the musical interpretation, allowing the pianist to shape the phrases, highlight melodic lines, and create a dramatic impact. Dynamics is also related to the power and intensity of the music and the variation of dynamics may influence the positiveness of the music.

Melody refers to the main musical line or theme that carries the expressive and melodic content of the composition. The prominent sequence of notes stands out and captures the listener's attention. The melody is usually played in the right hand and encompasses the memorable aspect of the music. It carries the emotional and narrative essence of the piece, often conveying a sense of beauty, expression, and storytelling. The melodic elements therefore influence the valence and arousal scores of the emotional expression.

Harmony refers to the simultaneous sounding of multiple notes to create chords and chord progressions. It involves the vertical aspect of music, where different pitches are combined to form harmonious sounds. Harmony provides the foundation and support for the melody, adding depth, richness, and complexity to the overall musical texture. It establishes the tonal framework and creates a sense of stability, tension, and resolution. Harmonic arrangement significantly contributes to the mood, character, and emotional impact of the piece. Major chords are usually related to positive emotions, while minor chords to negative emotions. The complexity of harmony can also be related to the tension and instability of the emotional expression.

Tone color or timbre refers to the unique character of a sound. It is a distinct combination of harmonics, overtones, and how sound is shaped and perceived. Tone color adds richness, variety, and individuality to the musical expression. It can vary according to the specific piano used, the technique employed by the pianist, and the musical context. For example, playing the same note on different parts of the piano keyboard or using different levels of touch can produce different tone colors. Tone color plays a significant role in conveying emotions, creating contrasts, and highlighting musical ideas. Brighter tone color is related to positive or exciting emotions.

Based on the research of Renato et al. [8], we selected some music features that are most suitable for representing the five major factors and designed our model for interpretable piano music emotion recognition.

III. Proposed Method

This section describes the proposed MER method that can consider the multifaceted and multimodal nature of music emotion (Fig. 2). The key feature of the proposed method is to separately analyze the major factors that form the emotion and integrate them in a data-adaptive manner. Our method is implemented on a standard multi-branch DNN consisting of three separate feature extraction branches, the feature fusion module, and the classifier.

A. Melody, Rhythm, and Dynamics Analysis for MIDI Data

We assume that the melody, rhythm, and dynamics factors are originally encapsulated in the symbolic domain (musical score), where these factors have close connections to the distributions of note pitches, locations, and velocities, respectively. Most existing MER methods, however, have dealt with only audio data, where such score-originated factors are treated together with audio-originated factors in a mixed way. In this study, we aim to explicitly learn the statistical patterns of notes from musical scores in a musically-meaningful manner faithful to the findings of music psychology.

As a score representation that plays a crucial role in MER, we use the standard MIDI-like representation. A musical score is serialized as a token sequence, where each note consists of the "note-on", "note-off", and "velocity" tokens and the "time shift" token represents the time interval between notes.

To learn contextual relationships and structural information from MIDI data, we use a recurrent neural network (RNN). Because symbolic music datasets with emotion are severely limited in size, we use as a relatively simpler structured model the bidirectional gated recurrent unit (Bi-GRU) instead of the bidirectional long short-term memory (Bi-LSTM) network to avoid overfitting. Our model consists of two stacked units with 256 hidden layer units each. This computation process can be simply described as follows:

$$\overrightarrow{h_t} = \mathrm{GRU}(\mathrm{token}_t, \overrightarrow{h}_{t-1}), \tag{1}$$

$$\overleftarrow{h_t} = \mathrm{GRU}(\mathrm{token}_t, \overleftarrow{h}_{t+1}), \tag{2}$$

$$h_t = \mathrm{Concat}(\overrightarrow{h_t}, \overleftarrow{h_t}). \tag{3}$$

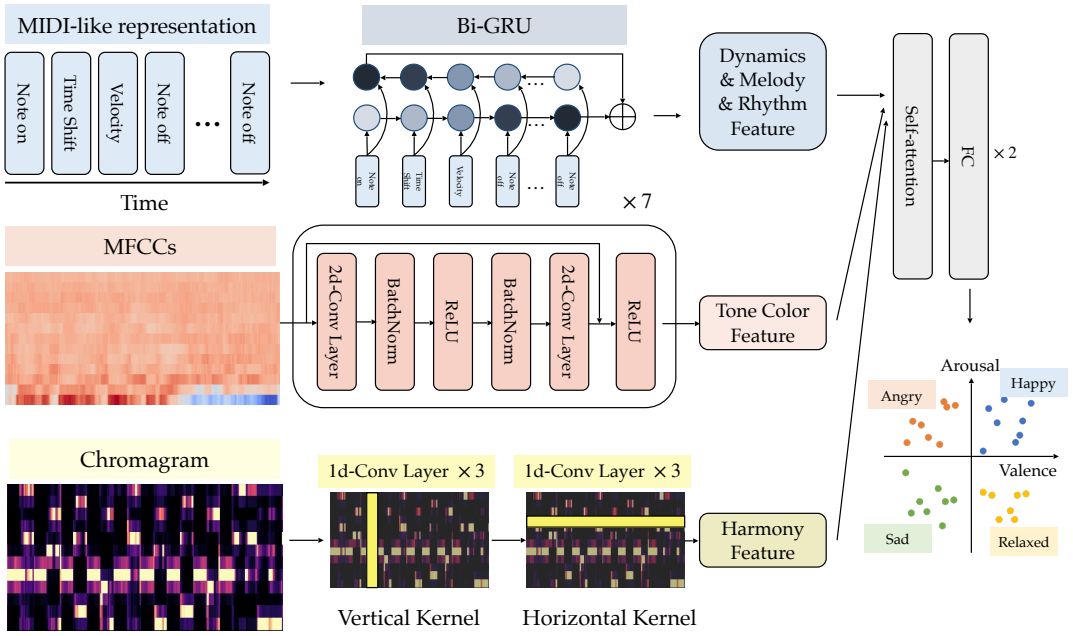


Fig. 2. The proposed network architecture consisting of symbolic and acoustic feature extraction modules and an adaptive feature fusion module with a self-attention mechanism for multimodal multifaceted music emotion recognition.

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ refer to the forward and backward hidden states at time step t respectively, token_t refers to the input MIDI-like token at time step t, h_t is the output hidden state of the Bi-GRU at time step t, which is obtained by concatenating the hidden states in both directions.

After being processed by the trained Bi-GRU, the MIDI-like token sequence is encoded into symbolic domain features encompassing melody, rhythm, and dynamics information. These features can be directly fed into a fully connected classifier for emotion prediction.

B. Harmony Analysis for Chromagrams

We assume that the harmony factor can be effectively extracted from the chromagram of a music signal, which represents the time-varying distributions of twelve pitch classes ranging from C to B in one octave. The chromagram can be obtained from the short-time Fourier transform (STFT) of a music signal with librosa.

To simultaneously learn harmonic structures over frequency and their progressions over time, we use a convolutional neural network (CNN) with a stack of vertical (frequency-directional) and horizontal (time-directional) convolutions on the chromagram. More specifically, we use three Conv1D layers with a vertical convolution kernel to learn the instantaneous information (pitch combination), followed by three Conv1D layers with a horizontal convolution kernel to aggregate the temporal information. The sizes of the vertical and horizontal kernels are (12, 1) and (1, 48), respectively, in this study. Although the pitch or melody information is extracted from the chromagram as well as the harmony information, the self-attention mechanism in the feature fusion module helps this model recognize the essential part of the feature.

C. Tone Color Analysis for MFCCs

We assume that the tone color factor has a close connection to the mel frequency cepstral coefficients (MFCCs) of a music signal, inspired by many studies on various classification tasks (e.g., genre, instrument, and singer classification) including MER. The MFCCs are obtained by applying the discrete cosine transform (DCT) to a log-scale mel-spectrogram. In general, about the first twenty coefficients are used as acoustic features related to the tone color representing the characteristics of the energy distribution and spectral envelope. In this study, the music signal is resampled at 22.05 kHz and the number of mel filterbanks is set to 23, and the 25-dimensional MFCCs are extracted.

To extract the tone color features from the MFCCs, we use a CNN consisting of seven convolutional blocks as in [14]. Each block consists of two Conv2D layers, followed by a 2D batch normalization (BN) layer and a ReLU layer. The kernel size is set to (3, 3) and the stride is set to 2 for every Conv2D layer. We also use the residual connection for efficient learning.

D. Adaptive Fusion of Symbolic and Acoustic Features

We aim to fuse the symbolic and acoustic features separately exacted from the three modules. Since each feature represents only a partial aspect of music emotion, the relative weights of the three features have a large impact on the performance of MER. The basic approach to feature fusion is to design a loss function for multiple features, where the feature weights are treated as hyperparameters that should be optimized in advance based on prior knowledge or with grid search. The performance of this approach based on the fixed weights, however, is limited because the importance of each factor on the overall emotion of the piece can vary considerably.

To adaptively adjust the feature weights according to the characteristics of the musical piece, we use a multi-head self-attention (MHSA) mechanism. Let F be a concatenated feature given by

$$F = \mathrm{Concat}(F_{\mathrm{sym}}, F_{\mathrm{chroma}}, F_{\mathrm{mfcc}}), \tag{4}$$

where F_{sym}, F_{chroma}, F_{mfcc} denote the outputs of the symbolic, harmony, and tone color analysis modules, respectively. A four-headed self-attention is then computed on F to capture the intricate dependencies within F as follows:

$$\mathrm{MHSA}(F) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_4)W_O, \tag{5}$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{6}$$

where head_i denotes an attention head i, W_O is a linear projection matrix, W_i^Q, W_i^K, and W_i^V denote projection matrices applied to Q, K, and V in the attention head i, respectively, and $\mathrm{Attention}$ is defined as

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{7}$$

where Q, K, V represent the query, key, and value matrices, respectively, and d_k represents the dimensionality of K. The output of the MHSA is flattened and fed into two fully-connected layers, followed by a sigmoid function that predicts the V/A probabilities.

IV. Evaluation

This section reports a comparative experiment conducted to evaluate the proposed and conventional MER methods.

A. Dataset and Metrics

We used a piano music dataset named EMOPIA [13] that contains 1087 piano clips from 387 songs, along with their corresponding MIDI files and emotion labels. Music recordings were collected from YouTube by their corresponding IDs with the 'youtube-dlp' package. We used the same train-validation-test split presented in [13]. Due to the unavailability of several pieces on YouTube, approximately 5% of the original data could not be included in our experiment.

The EMOPIA dataset was annotated with four-quadrant (4Q) emotion labels, where music emotions are categorized into four classes in terms of V/A levels, i.e., "happy" for high arousal and high valence, "relaxed" for low arousal and high valence, "angry" for high arousal and low valence, and "sad" for low arousal and low valence. We used the 4Q classification accuracy as the primary evaluation metric. We also focused on the binary classification accuracies in the V/A estimation. Note that these metrics indicate the performance for recognizing a single aspect of the emotion.

As MIDI-based baseline methods, we tested the symbolic linear regression model (Symbolic-LR) and the LSTM model with an attention mechanism (LSTM-Attn) [15] on the MIDI-like or REMI representation implemented in [13]. As audio-based baseline methods, we tested the acoustic linear regression model (Acoustic-LR) [13] and the short-chunk ResNet model [14]. These uni-modal methods are all lightweight models trained and evaluated only on the EMOPIA dataset.

TABLE I
COMPARISON OF THE PROPOSED AND BASELINE MER METHODS ON EMOPIA DATASET.

Method	4Q	Arousal	Valence
Ours	**0.692**	0.884	**0.869**
LSTM-Attn [15]+MIDI-like [19]	0.684	0.882	0.833
LSTM-Attn [15]+REMI [20]	0.615	0.890	0.746
Symbolic-LR [13]	0.581	0.849	0.651
Audio-LR [13]	0.523	**0.919**	0.558
Short-chunk ResNet [14], [21]	0.677	0.887	0.704

TABLE II
COMPARISON OF PRE-TRAINED MUSIC LANGUAGE MODELS ON EMOPIA DATASET.

Method	4Q Accuracy
Ours	0.692
MidiBERT [16]	0.634
MT-MidiBERT [16], [17]	0.676
QM-MidiBERT [16], [18]	**0.719**

We also tested a large-scale music language model called MidiBERT-Piano [16] and its variants called MT-MidiBERT [16], [17] and QM-MidiBERT [16], [18]. Such models were pre-trained on large-scale MIDI data. Since the V/A estimation was not evaluated on these models, we only used the 4Q metric for comparative evaluation.

B. Results and Discussions

The performances of the proposed and conventional methods are shown in Table. I and Table. II. To the best of our knowledge, our method is the first to combine acoustic and symbolic analyses on MER for piano music. We thus compared only uni-modal methods.

As shown in Table. I, the proposed method achieved the best score of 0.692 in the 4Q metric while keeping the computational cost comparable with the other methods. In arousal estimation, the proposed method performed comparably with the other methods. Since arousal is mainly attributed to low-level acoustic features (e.g., tempo and intensity), even basic audio-based models such as Audio-LR can recognize it accurately. In valence estimation, the proposed method also achieved the best score of 0.869. We found that audio- and MIDI-based methods tended to be better at arousal and valence estimations, respectively. This tendency is consistent with our hypothesis described in Section 1. Overall, by taking into account all factors, the proposed method achieved a better-balanced performance of MER. As shown in Table. II, the proposed model showed a performance comparable to the state-of-the-art large pre-trained model QM-MidiBERT [18], while keeping the computational cost small.

C. Ablation Study

To evaluate the effectiveness of the proposed multimodal multifaceted approach, we conducted an ablation study, where four ablated versions of the proposed method were made by removing two branches out of the three branches in the feature extraction stage or the MHSA mechanism in the feature fusion stage. As shown in Table. III, the full version performed best. This experimental result was consistent with our hypothesis.

TABLE III
ABLATION STUDY OF THE PROPOSED METHOD ON EMOPIA DATASET.

Method	4Q	Arousal	Valence
Full model	**0.692**	**0.884**	**0.869**
Symbolic analysis branch	0.690	0.833	0.744
Harmony analysis branch	0.514	0.726	0.685
Tone color analysis branch	0.649	0.871	0.711
Without MHSA	0.639	0.858	0.753

Since the music score includes information about rhythm, dynamics, and melody related to both valence and arousal, the symbolic analysis branch performed well in all evaluation metrics. Although harmony plays an essential role in emotional expression, it is hard to recognize the overall emotion of a piece only from its harmony information. The harmony analysis branch thus performed worst in all metrics. The tone color analysis branch showed a promising performance, especially in arousal estimation, because the energy distribution and tone color information obtained from MFCCs significantly contributes to the arousal aspect and the overall emotion. By integrating these three branches, the proposed method performed better, especially in valence estimation. We also confirmed the effectiveness of the MHSA-based adaptive feature fusion. The proposed full model outperformed its ablated version by 1.2 pts in the 4Q metric and achieved better balanced V/A estimation performance.

V. CONCLUSION

In this paper, we proposed a multimodal multifaceted MER method for piano music. Inspired by music psychology, we designed a DNN with three branches that separately analyze five main factors of music emotion, i.e., 1) rhythm, dynamics, and melody, 2) harmony, and 3) tone color, from the score, chromagram, and MFCCs of a piece, respectively. Our model fuses the extracted musically-meaningful symbolic and acoustic features with the MHSA mechanism in a data-adaptive manner. In the comparative experiment using the public dataset EMOPIA, our model achieved the state-of-the-art performance in the 4Q metric and a well-balanced performance in the V/A classification metric with a smaller computational cost. The ablation study showed the effectiveness of the multimodal multifaceted approach and that of the MHSA-based adaptive feature fusion. As a future direction, we plan to explore music embedding for MER, which can be used to design a unified MER model that can better capture multiple interdependent factors.

ACKNOWLEDGMENT

This study was partially supported by JSPS KAKENHI Nos. 21K12187, 21K02846, 22H03661, 20H00602, and 21H03572, JST PRESTO No. JPMJPR20CB, and JST FOREST No. JPMJPR226X.

REFERENCES

[1] Juan Sebastián Gómez Cañón, Estefanía Cano, Tuomas Eerola, Perfecto Herrera, Xiao Hu, Yi-Hsuan Yang, and Emilia Gómez. Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Process. Mag.*, 38(6):106–114, 2021.

[2] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition. *IEEE Transactions on Multimedia*, 21(12):3150–3163, 2019.

[3] Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. Transformer-based approach towards music emotion recognition from lyrics. In *European Conference on Information Retrieval*, pages 167–175. Springer, 2021.

[4] Ziyu Wang and Gus Xia. Musebert: Pre-training music representation for music understanding and controllable generation. In *ISMIR*, pages 722–729, 2021.

[5] Jacopo de Berardinis, Angelo Cangelosi, and Eduardo Coutinho. The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability. In *Proceedings of the 21st international society for music information retrieval conference*, pages 310–317, 2020.

[6] Jiahao Zhao, Ganghui Ru, Yi Yu, Yulun Wu, Dichucheng Li, and Wei Li. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[7] Petri Laukka, Tuomas Eerola, Nutankumar S Thingujam, Teruo Yamasaki, and Grégory Beller. Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion*, 13(3):434, 2013.

[8] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 2020.

[9] Xin Gu, Yinghua Shen, and Jie Xu. Multimodal emotion recognition in deep learning: A survey. In *2021 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 77–82. IEEE, 2021.

[10] Shreyan Chowdhury and Gerhard Widmer. On perceived emotion in expressive piano performance: Further experimental evidence for the relevance of mid-level perceptual features. *arXiv preprint arXiv:2107.13231*, 2021.

[11] X Hu, K. Choi, and J. S. Downie. A framework for evaluating multimodal music mood classification. *Journal of the American Society for Information Science*, 68(2):273–285, 2017.

[12] Yahan Yu, Bojie Hu, and Yu Li. GHAN: Graph-based hierarchical aggregation network for text-video retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5547–5557, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. *arXiv preprint arXiv:2108.01374*, 2021.

[14] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of CNN-based automatic music tagging models. *arXiv preprint arXiv:2006.00751*, 2020.

[15] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[16] Yi-Hui Chou, I Chen, Chin-Jui Chang, Joann Ching, and Yi-Hsuan Yang. MidiBERT-piano: Large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv:2107.05223*, 2021.

[17] Jibao Qiu, CL Chen, and Tong Zhang. A novel multi-task learning method for symbolic music emotion recognition. *arXiv preprint arXiv:2201.05782*, 2022.

[18] Zhexu Shen, Liang Yang, Zhihan Yang, and Hongfei Lin. More than simply masking: Exploring pre-training strategies for symbolic music understanding. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 540–544, 2023.

[19] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32(4):955–967, 2020.

[20] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1180–1188, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Learned String Quartet Music with Variational Auto Encoder

Young-Long Chen*, Hsin-I Huang and Tzu-Te Yen
* National Taichung University of Science and Technology, Taiwan
E-mail: ylchen66@nutc.edu.tw Tel: +886-4-22196621

Abstract — **Music holds a distinct place within the realm of human artistic. Music generation is also considered as one of the highly challenging domains, demanding significant expertise and innovation artificial intelligence of in the field. In this paper, we propose a novel architecture with multi-track encoder for multi-decoder network. Furthermore, we present a novel music dataset specifically designed for string quartet compositions. Experimental results demonstrate that our architecture is capable of generating very rich musical compositions.**

I. INTRODUCTION

With the continuous development and growth of artificial intelligence, music generation is considered highly challenging and promising within the field of artificial intelligence [1] and Music Information Retrieval (MIR) tasks [2]. In the domain of music generation, there are various branches, including automatic music generation [3], jazz music composition [4], accompaniment generation and polyphonic melody generation [5]. In a broader sense, music can be seen as a form of language, both being systems that can be recorded symbolically in written form, and both having order and structure. Just as language has grammar, music also has its own set of norms and regulations.

The natural language processing (NLP) of field has achieved remarkable success when applied to MIR tasks. For instance, the application of generative adversarial networks [6] (GANs) in MIR has shown great potential. MuseGAN [7] is a GAN-based model that adopts the concept of an orchestra. It utilizes both conductor and musician architectures for music generation. There are also approaches which use conditional convolutional GANs to generate melodies following given prior chord sequences [8], either starting from scratch or conditioning on melodies from previous measures. Currently, most research directions in this domain employ generative adversarial networks for music generation.

In this paper, our goal develops a polyphonic music generator with a variational autoencoder (VAE) [9] architecture. Specifically designed for string quartet compositions. We shall refer to it as Multi-Muse variational autoencoder (MMVAE). To achieve this, we first create a music dataset for string quartets, which serves as the training data for our model. In a string quartet, the instruments are mostly divided into two parts violins, viola and cello. Inspired by the content-style disentanglement idea [10] and [11], we utilize a variational autoencoder as the main framework in this research. This architecture effectively learns the latent factors of music, such as texture. We propose a novel composite architecture that employs a multi-track encoder to a multi-layered decoder structure. By sharing parameters, the architecture enables the generation of distinctive melodic lines while maintaining their coordination.

In this paper, we have made the following primary contributions:

1) We have curated a novel MIDI dataset specifically designed for string quartets.

2) We have proposed a multi-layered variational encoder architecture for multi-track sequence generation.

II. MULTI-MUSEVAE METHOD

We have proposed a novel multi-track sequence generation model called MMVAE. In this section, we provide a detailed account of the architecture for generating string quartet music based on an encoder framework. The ultimate goal generates musical compositions specifically tailored for string quartets. This architecture primarily learns from musical compositions comprising 4 measures, totaling 16 beat in length.

Firstly, let's introduce the overall structure, with the Musical Instrument Digital Interface (MIDI) format of string quartets serving as the input format. The input MIDI files are segmented and extracted based on their tracks. The segmented tracks consist of four distinct instruments: Violin I, Violin II, Viola, and Cello. Furthermore, the four musical semantic information are extracted and compressed. By compressing the musical semantic information, the sparsity of the data is effectively

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

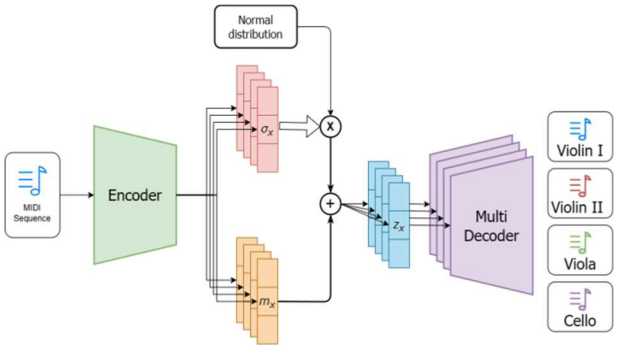


Fig. 1 System diagram of the proposed MMVAE model for string quartet data generation.

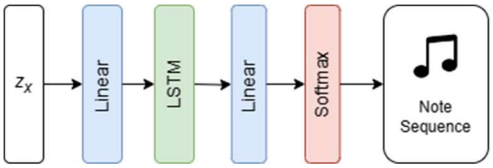


Fig. 3 The Decoder architecture of MMVAE.

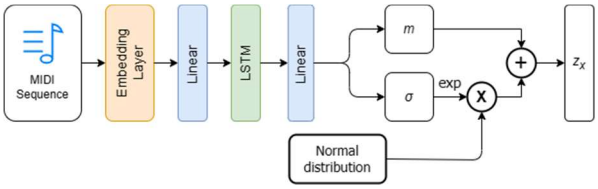


Fig. 2 The Encoder architecture of MMVAE.

reduced. After the extraction and segmentation process is completed, the data undergoes preprocessing to transform it into the desired input format for the model. Finally, the processed data is fed into the VAE architecture corresponding to each section, that it allows for separate processing and generation within each component. As VAE adheres to the structure of encoding and decoding, it enables a comparison between the generated outcomes and the original results. Furthermore, VAE can more directly generate content that closely approximates the original data.

A. Music Embedding

The MIDI files is inputted into the model, data preprocessing is required to transform the MIDI data into a suitable format. Due to the nature of MIDI files, each musical note can be viewed as a switch event. Additionally, MIDI files often represent these switch events using absolute timestamps. This can result in inconsistent durations of musical semantic information if the timing is not uniformly processed. In this paper, all MIDI files are standardized to a single tempo according to music theory and the characteristics of the music. By utilizing the library "pretty_midi" [12] to extract information from MIDI files, the beats per minute (BPM) of all MIDI files is standardized to 120 beats per minute. By standardizing the BPM to 120 beats per minute, any influence introduced by time can be eliminated during subsequent data processing.

Additionally, our proposed method enhances the model's versatility and generalizability. According to [13], we transpose all MIDI files into 11 different keys to enhance the model's versatility (A major would be transposed to G major). Indeed, this approach significantly expands the melodic possibilities. We transposing MIDI files to different keys that we obtain a broader range of musical expression and variations.

Finally, all transposed MIDI files are segmented into individual tracks based on MMVAE model architecture. Additionally, the tempo and key information are preserved during the segmentation process, we can ensure that the speed and tonality of the music are retained in the individual tracks. We obtained a rich collection of segmented MIDI tracks for string quartets, including Violin I, Violin II, Viola, and Cello.

B. Data Representation

In order to train the variational autoencoder model for polyphonic music, we convert the MIDI files into the Piano-Roll representation. The Piano-Roll representation is a binary data format that uses a matrix to record musical score information. The Piano-Roll representation can be regarded as a two-dimensional matrix. In the Piano-Roll representation, the horizontal axis corresponds to time, while the vertical axis represents pitch. The presence of notes at different time steps is represented in a two-dimensional matrix format.

The incongruity in the lengths of MIDI files within the dataset hampers the training process. According to [7], we set the temporal resolution to one beat consisting of 24-time steps and we can allow for an effective coverage of common rhythmic patterns, such as triplets and 16th notes. The size of the input tensor is hence size 4 (bar) × 96 (time steps) × 36 (range of pitches) × 4 (number of tracks).

C. Multi-MuseVAE

Inspired by MuseGAN, this study incorporates the concepts of interference and synthesis present in MuseGAN. Due to the

distinct textures exhibited by different sections, the requirements of the decoder will vary accordingly. However, the model inputs and parameters are shared among them. The overall architecture of MMVAE is depicted in the Fig. 1.

The overall model architecture follows a one-encoder to multi-decoder approach. By employing this approach, it allows the encoder to encode the latent space (z_x) with a certain degree of consistency. Based on the requirements of different decoders (D), we define four types of latent spaces generated by the encoder as follows: $z_{Violin\ I}$, $z_{Violin\ II}$, z_{Viola} and z_{Cello}.

Music is a temporal art form with inherent dependencies and contextual relationships. In previous approaches, there have been implementations of melody generation based on Recurrent Neural Network (RNN) [14]. There have also been implementations of melody generation based on Long Short-Term Memory (LSTM). Based on [15], we used LSTM as a crucial component in our proposed MMVAE model.

D. Encoder

The MIDI is transformed into the latent space through the encoder. The encoder architecture is shown in Figure 2. The model takes segmented MIDI short sequences as input. Because of the sequential nature of music, the input sequence of musical notes convert into events by the encoder. Through the encoding layer, the data is extracted and transformed into tensors representing the pitch and duration of musical notes. The encoded pitch tensor is then passed through a linear layer (Linear) to flatten it into a one-dimensional embedding tensor, which serves as the feature representation. The input size of the linear layer depends on the pitch range (typically 36 notes). The output feature size is setting to 300.

After passing through the LSTM network, the resulting tensor is further processed by a linear layer. The mean (m) and standard deviation (σ) of the variational autoencoder are obtained. The exp () operation is performed on the normal distribution with σ, and then the resulting values are combined with m. The final output of the encoder is the latent space z_x. The z_x latent space is defined in four types based on the input of the encoder: $z_{Violin\ I}$, $z_{Violin\ II}$, z_{Viola} and z_{Cello}. By adding normal distribution constraints, the encoder can learn the fuzzy structure of music data that it is helpful for generating rich music.

E. Decoder

We inspire by the content-style disentanglement idea. The decoder is required to reconstruct the data using the latent space that it is generated by the encoder. Due to the distinct characteristics of each instrument in a string quartet, the decoder needs to undergo separate training for each instrument. By utilizing different decoders to reconstruct the latent space, the goal of disentanglement and generation can be achieved. The decoder model architecture is shown in Fig. 3.

TABLE I
CORRESPONDING TO THE DECODER PARAMETER.

Decoder	*Vocabulary size*	*Pitch range*	*MIDI Pitch Number*
Violin I	*36*	*(55 , 91)*	*G3 – G6*
Violin II	*24*	*(55 , 79)*	*G3 – G5*
Viola	*36*	*(48 , 84)*	*C3 – C6*
Cello	*36*	*(36 , 72)*	*C2 – C5*

The working principle of the Multi-MuseVAE decoders is as follows. Firstly, the latent space is transformed into features that are suitable for the LSTM network through a linear layer. The LSTM network is utilized to decode the latent space at each time step. The decoded tensor is then passed through another linear layer to transform it into a latent tensor representing pitch and duration. Finally, the latent tensor representing each note and duration is reconstructed into a note sample by a SoftMax layer. Different parameters are adjusted for each instrument in a string quartet to align with the unique characteristics of each instrument. Tt is shown in Table I.

Due to the common practice in most string quartet compositions, the Violin I and Violin II parts are typically used as the melody lines. In a string quartet, the Violin I part typically encompasses a higher range of pitches compared to the Violin II part. Therefore, the pitch range of the Violin II is often limited or constrained to 24 notes to accommodate the higher range typically that it is utilized by the Violin I. They utilize a pitch range of 36 notes and spanning across 3 octaves.

Finally, we can obtain four sets of note samples by the decoder that each note sample corresponds to one instrument in the string quartet. Assign specific instruments to the four samples according to their respective tracks. MIDI post-processing is applied to the samples after they have been generated. We can obtain MIDI sequences for each of the four parts, representing the final compositions for the string quartet. Through the utilization of MIDI processing, amalgamate the acquired four musical samples into a singular MIDI file.

III. EXPERIMENTAL AND RESULT

A. Lakh MIDI Dataset

The Lakh MIDI dataset is a large collection consisting of 176,581 unique MIDI files. Among them, 115,160 MIDI files have been matched and aligned with the entries listed in the Million Song Dataset. We utilize the LPD-5 dataset according

TABLE II

INTRA-TRACK EVALUATION

(VALUES CLOSER TO THE DATASET ARE BETTER).

Part / Metric	Dataset				Generation				Ablated
	Violin I	Violin II	Viola	Cello	Violin I	Violin II	Viola	Cello	All
EB	8.9	11.3	12.9	13.9	6.8	10.1	10.3	9.1	87.3
UP	26.8	21.9	21.1	23.6	22.9	25.4	16.1	17.4	89.2
UPC	4.2	3.6	3.4	3.9	3.9	3.2	3.1	3.9	11.1
PR	57.9 – 89.4	56.4 – 82.1	50.0 – 74.3	38.1 – 66.9	57.5 – 88.42	54.6 – 76.8	48.7 – 69.2	36.4 – 63.4	10.4 – 99.6
QN	94.5	92.4	88.7	91.3	97.4	95.3	86.7	92.5	41.3

to [7]. Furthermore, we compress and merge the Lakh MIDI dataset into five common categories: Drums, piano, guitar, bass, and strings. The dataset consists of three types: LPD-5-full, LPD-5-matched, and LPD-5-cleansed. We utilize the LPD-5-cleansed dataset as our pretraining dataset. There is a total of 21,425 MIDI files.

B. String Quartet Dataset

To specifically train music generation models for string quartets, we have curated a new dataset. The String Quartet dataset consists of 110 standard classical string quartet compositions. The dataset was sourced from professional musicians and created by reproducing sheet music from renowned classical composers such as Mozart, and Bach. The sheet music was obtained from the international music score library project (IMSLP). After transposing the pieces into 12 different keys, we obtain a collection of 1,320 musical compositions. After organizing and segmenting the data, we obtained a total of 71,280 musical phrases that it consists of four measures.

C. Objective Metrics

To effectively evaluate the performance of the model, the comparison is made between the original musical phrase data and the generated melodies. This evaluation aims assess the authenticity of the melodies that are generated by the model. We randomly generated 800 musical phrases for comparison. In [7], a series of evaluation metrics is proposed, such as the number of pitch repetitions, rhythm variations, and other related evaluation metrics. To evaluate the four tracks of the string quartet, a total of five evaluation metrics is used:

1) Empty bar (EB): Proportion of empty measures, expressed as a percentage.
2) Used pitch (UP): Number of notes used.
3) Used pitch classes (UPC): Number of distinct note categories used, ranging from 0 to 12.
4) Pitch range (PR): Range of pitch used, represented by the highest and lowest notes.

TABLE III

COMPARISON WITH OTHER APPROACHES

(VALUES CLOSER TO THE DATASET ARE BETTER).

Metric / Method	EB	UP	UPC	PR	QN
Lpd5- string Dataset	8.62	27.68	3.92	50.41 – 78.23	87.42
MuseGAN [7]	6.10	30.43	4.09	47.32 – 79.67	72.12
Our	9.37	26.32	4.13	49.43 – 77.61	86.30

5) Qualified note (QN): Percentage of qualified notes, representing the proportion of notes that meet certain criteria. This metric evaluates whether the generated music by the model is too sparse.

By the aforementioned objective evaluation metrics, we can effectively analyze whether the model has learned effectively. Through this analysis, we can assess the differences between the generated samples and the training data. As training progresses, the distribution of generated samples should become increasingly closer to the distribution of the training samples.

We apply evaluation metrics to the training data to analyze various indicators of the training data. Comparing the generated results with the training data, we can effectively assess the learning performance of the model. The experiment involved splitting the training data and generated data according to the sections of the string quartet. The results are presented in Table II. From the training data, the value of EB is very low that there are very few empty measures in the dataset. From UP, we can observe that the training data exhibits a rich variety of pitch usage. From UPC, we can see that all four string section in the training data. The training data scores is higher than 3.0 that the dataset is a rich variety of chords. From PR, we observe that the pitch ranges is used in the four string sections of the training data which are generally very close to the pitch range designed for the model. This indicates

that the PR setting of the model is reasonable. From QN, we can observe that the values in the training data are very high that the training data is not overly fragmented.

To examine the results of each model's generation, we generated 800 music segments for each model based on the experimental design. According to the objective evaluation metrics, the results are shown in Table II. The values closer to the dataset are better. We found that the EB values for all four sections are very high that the model has successfully learned to generate filled melodies. In UP, the values for all four sections are very close to the dataset. It indicates that the model has successfully learned the rich musical melodies from the dataset. From the UPC, we have discovered that the four divisions have also acquired a profound understanding of harmonic variations. Furthermore, the PR values indicate that the model has adeptly grasped the tonal ranges of the individual sections within the string quartet. Lastly, with regard to the QN values, it is evident that the generated data exhibits a notable absence of fragmentation.

Furthermore, we conducted a comparison with MuseGAN, and the results are presented in Table III. It is evident that VAE's architecture exhibits superior performance in the QN category compared to GAN networks. In the PR category, the utilized pitch range is relatively close. However, in the UPC category, our proposed method slightly lags behind GAN-based approaches. In the UP category, our method aligns more closely with the dataset. Finally, in the EB category, our scores closely match those of the dataset. From these results, we discern that the VAE architecture, in contrast to the GAN framework, possesses superior reconstructive capabilities, indicating a better understanding of the foundational musical elements.

D. Subjective Evaluation

To assess our system from a user's perspective, we conduct tests involving audiences with diverse musical backgrounds and educational levels. The participants were categorized into three user groups: professional musicians, amateur musicians, and music enthusiasts.

All melodies will be rendered into music files in the same manner. The test consists of a total of 12 song segments, comprising 4 segments generated by our method, 4 segments from the original dataset, and 4 segments from the single-layer encoder architecture (ablation experiment). All participants are required to rate each of the 12 songs based on the following questions by a Likert scale ranging from 1 to 5. Users should rate the songs without knowing which ones are generated and which ones are from the original dataset.

TABLE IV
SUBJECTIVE EVALUATION
(M: Melody, R: Rhythm, MS: Musical structure, V: Variations, OR: Overall rating).

	M	R	MS	V	OR
Dataset	**4.33**	3.73	**4.23**	**3.55**	**4.19**
Generation	3.54	**3.76**	3.63	3.15	3.82
Ablated	2.26	3.28	2.36	2.96	2.30

The questions are as follows:

1) How pleasant does this melody sound?
2) How pleasant does this rhythm sound?
3) Does this melody have a clear musical structure?
4) Does this melody exhibit rich variations?
5) Overall rating.

We extended invitations to a total of 42 test participants to engage in the evaluation process. They were 24 professional musicians, 8 amateur musicians, and 10 music enthusiasts. Experimental results is presented in Table IV. Tt is evident that the compositions is generated by our proposed multi-track decoder model closely align with the ratings of the original dataset. Furthermore, the ablation experiment conducted on the single-track decoder model yielded significantly lower scores. Henceforth, it can be inferred that our proposed model has approached the realm of human artistic creation.

E. Ablation Experiment

To evaluate whether the multi-track decoder architecture has an advantage, we designed a single-track decoder as an ablation experiment. The four tracks are compressed into a single common track, and the single-track decoder needs to independently generate the four separate melodies for each section. We set the pitch range of the decoder model to the standard MIDI pitch range, which spans from C-2 to G8, comprising a total of 128 pitch values. The experimental results are presented in Table II and IV.

Due to the encoder compressing four tracks into a single track, theoretically, the decoder needs to learn more complex latent features. This can lead to suboptimal learning outcomes. From the results of ablation experiments, it is observed that in Table II, the proportion of EB is notably high. We attribute this to the scarcity of qualified notes generated, resulting in insufficient eligible notes to compose melodies for bars. Furthermore, in the UP and PR categories, the loss of decoder specificity leads to the almost comprehensive utilization of the entire tonal range. In the QN category, the notes generated by

a single encoder and decoder pair are significantly less qualified compared to others, resulting in an overly sparse generation.

IV. Conclusions

In this paper, we established a novel dataset from scratch. Furthermore, we proposed a novel music generation model. We utilize a variational encoder architecture for the generation of string quartets. By employing LSTM (Long Short-Term Memory) architecture, we implemented a model to generate multi-track sequences. Objective and subjective evaluations indicate that our proposed model demonstrates initial proficiency in generating string quartets. Although it still lags behind the creative level of human musicians in terms of music theory and human aesthetic assessment. The proposed to model already possesses certain desirable characteristics.

References

[1] Fernández, Jose D., & Francisco Vico. "AI methods in algorithmic composition: A comprehensive survey." *Journal of Artificial Intelligence Research* 48, pp. 531-582, 2013.

[2] Oramas, S., Espinosa-Anke, L., Zhang, S., Saggion, H., & Serra, X. "Natural language processing for music information retrieval." *In Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016),* 2016.

[3] Liang, Feynman T., M., Johnson, M., & Shotton, J. "Automatic Stylistic Composition of Bach Chorales with Deep LSTM." *In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR2017)*, 2017.

[4] Herremans, Dorien, Ching-Hua C., & Elaine C. "A functional taxonomy of music generation systems." *ACM Computing Surveys (CSUR)* 50.5, pp. 1-30, 2017.

[5] Dong, H. W., Yang, Y. H. "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation." *In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR2018)*, 2018.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. "Generative adversarial networks." *Communications of the ACM 63.11*, pp.139-144, 2020.

[7] Dong, H. W., Hsiao, W. Y., Yang, L.C., & Yang, Y. H. "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment." *In Proceedings of the Association for the Advancement of Artificial Intelligence Conference (AAAI) on Artificial Intelligence,* 2018.

[8] Yu, Y., Srivastava, A. & Canales, S. "Conditional LSTM-GAN for Melody Generation from Lyrics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* pp. 1-20, 2021.

[9] Kingma, Diederik P., & Welling, M. "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning,* pp. 307-392, 2019.

[10] Trieu, N., & Keller, R. "JazzGAN: Improvising with Generative Adversarial Networks." *In Proceedings of the Musical Metacreation Musical Metacreation,* 2018.

[11] Wang, Z., Wang, D., Zhang, Y., & Xia, G. "Learning interpretable representation for controllable polyphonic music generation." *arXiv preprint arXiv:2008.07122,* 2020.

[12] Raffel, C., & Ellis, D. P. "Intuitive analysis, creation and manipulation of midi data with pretty_midi." *In Proceedings of the 15th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers (ISMIR 2014),*, 2014.

[13] Ycart, A., & Benetos, E. "Learning and Evaluation Methodologies for Polyphonic Music Sequence Prediction With LSTMs." *IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28,* pp. 1328-1341, 2020.

[14] Boulanger-Lewandowski, N., Bengio, Y. & Vincent, P. "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription." arXiv preprint arXiv:1206.6392, 2012.

[15] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music." *In Proceedings of the 35th International Conference on Machine Learning,* pp. 4364-4373, 2018.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

SOAda-YOLOR: Small Object Adaptive YOLOR Algorithm for Road Object Detection

Yu-Fang Huang*, Tsung-Jung Liu*, Chun-An Lin* and Kuan-Hsien Liu†
* National Chung Hsing University, Taiwan
E-mail: lin12099@yahoo.com.tw; tjliu@dragon.nchu.edu.tw; julian135707@gmail.com Tel: +886-4-22851549
† National Taichung University of Science and Technology, Taiwan
E-mail: khliu@nutc.edu.tw Tel: +886-4-22196320

***Abstract*—Nowadays, road object detection has become a challenging topic and a long-standing problem. In previous algorithms, traditional convolutional neural networks (CNNs) have almost dominated computer vision and achieved considerable successes in different detection tasks. However, Transformer-based models in the past two years have shown impressive performance, and some of which even surpass CNN-based methods and become the state-of-the-art for advanced vision tasks. In this paper, we propose a detection model named SOAda-YOLOR, which uses the current mainstream Bi-FPN concept with the innovative Optimized-NMS. We also propose a novel ShuffleSimAM module to the backbone, and apply some bag-of-freebies such as SIoU loss to road object detection, and finally achieve competitive results.**

I. Introduction

Object detection is a relatively complex image processing task in computer vision related applications, such as IQA [1], [2], SIQA [3], [4], VQA [5]–[7], Superresolution [8], [9]. Traditional object detection methods are usually based on algorithms, and since current object detection algorithms are not very accurate, research and applications in this field are still innovating. As mentioned earlier, although most convolutional neural network (CNN) based methods such as YOLOv3 [10], YOLOv4 [11], YOLOv5, Scaled-YOLOv4 [12], and other different types of methods such as DINO [13] achieve excellent performance, they all have one common disadvantage. That is most close objects can be detected accurately, but far objects such as distant pedestrians, intersection signs and oncoming vehicles are hard to detect. Due to the convolutional operations required by the algorithm, the target to be detected is often compressed to less than 3 pixels. This compression impedes the model's ability to learn effectively and results in poorer quality output.

To ameliorate the issue of inadequate detection of diminutive objects, certain techniques are conventionally employed, namely: 1) enhancing the resolution of image acquisition; 2) increasing the depth of the model; 3) locating the desired image area; 4) expanding the training dataset via data augmentation; 5) learning model anchors automatically; and 6) filtering out irrelevant categories as many as possible. Of course, through these methods, a large amount of hardware resources and time investment are bound to be required. DINO is an example which is based on Swin Transformer, so the ordinary computer is not enough to train such a large model.

In addition, YOLOR [14] demonstrated models incorporating implicit knowledge can achieve impressive performance in real-time object detection. While the accuracy is slightly increased, the time required for training and detection can be greatly reduced, making it a leader in current one-stage object detection algorithms.

Overall, the principal contributions of this work can be encapsulated as follows:

- We use YOLOR [14] as our main backbone and deepen its modules and enhance the connectivity between layers. By importing the concept of Bi-FPN into the model, we create a tighter connection with the backbone layer, and use Dropout in the first layer to avoid the model learning bad features.
- We perform low-light enhancement (LLE) [15], [16] to improve the low-light problem encountered in the current road dataset.
- We propose a novel ShuffleSimAM module to the backbone.
- We propose optimized-NMS to improve our results.
- We use some bag-of-freebies (e.g., SIoU [17] loss) and apply it to our model.

By systematically integrating these components into the YOLOR framework, SOAda-YOLOR achieves improved efficiency and accuracy in small object detection for complex road scenes.

II. Related Work

The authors of YOLOR adopted Scaled-YOLOv4 as its original architecture. It demonstrated the YOLOv4 object detection neural network is based on the CSP method, which can maintain the best speed and accuracy, and can also be scaled up and down (i.e., suitable for small and large targets). This methodology can be modified not solely in terms of depth, width, and resolution, but additionally in regard to the architecture of the network. Further extending on the basis of Scaled-YOLOv4, and adding the concept of tacit knowledge, YOLOR is proposed to become a unified network that can handle multiple tasks. With the same accuracy, the inference speed of YOLOR is 88 % faster than its previous version (i.e., Scaled-YOLOv4 [12]).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

A. Backbone Network

CSPDarknet53 [12] derives its architectural foundation from Darknet53 [11] and adopts the structure of CSPNET [12]. It is mainly composed of five CSP modules, which can reduce the computational overhead. Because of the continuous superposition of CSP modules, it makes the extraction of features by the network vulnerable to the interference of background noise. To further optimize backbone performance, modifications to the native CSP modules have been instituted by the original authors.

B. Path Aggregation Network

PANet [18] was proposed in 2018 and is still used by the public. PANet can accurately preserve spatial information, which helps to correctly locate pixels to form masks. With some minor modifications in the YOLO series, PANet usually adds adjacent layers together to use self-adapt feature pooling for mask prediction. However, when using PANet [18] in YOLOv4 [11], this approach is slightly distorted. Therefore, instead of adding adjacent layers, a concatenation operation is applied to them, improving the prediction accuracy.

C. Bidirectional Feature Pyramid Network

Bidirectional feature pyramid network (Bi-FPN) [19] was proposed as a weighted two-way feature pyramid network in EfficientDet [19], which allows simple and fast multi-scale feature fusion. The purpose of the authors is to pursue more efficient multi-scale fusion methods. In the past, feature fusion was to treat different scale features equally, and the authors introduced weights (similar to Attention) to better balance the feature information of different scales.

D. Spatial Pyramid Pooling

The CSPSPP block is formed by adding the Spatial Pyramid Pooling (SPP) [20] structure to the CSP block. SPP comes from SPPNet proposed in 2015, and fast-RCNN [21] was also proposed on its basis. At that time, the neural network contained a fully connected layer, which led to the fact that the size of the input image to the network must be a fixed size. The SPP structure was originally proposed in SPPNet to solve this problem. It can be understood as the process of generating a fixed-size input image of any scale. The SPP module consists of four max-poolings of different sizes stacked in the channel dimension. In this model, the pooling kernel sizes are set to 13×13, 9×9, 5×5, and 1×1. By incorporating the SPP architecture into YOLOv4, expansion of the receptive field is attainable, facilitating isolation of the most salient contextual elements without slowing down the detection speed.

III. Proposed Method

The proposed Small Object Adaptive-YOLOR (SOAda-YOLOR) architecture (Fig. 1) is based on the YOLOR model, and the details of each component are explained as follows:

A. CSPSE-D Block

Squeeze-and-Excitation Networks (SENet) [22] is actually a lightweight block that can be universally improved for all networks. Compared with the original network, the amount of parameters for the block is very small. In this way, the accuracy is greatly improved without additional burden, and its output is calibrated through the module. We add this module at each CSP convolution stage and include dropout in the first module to drop different neurons at each training to ensure bad information is not learned.

B. Adp-FPN

The weighted Bi-FPN[19] is the pursuit of a more efficient multi-scale fusion method. We refer to this idea and include it into our model, called Adp-FPN. Deepening between layers and connecting each other will create more features.

C. SPP

In order to obtain richer feature information, we add two additional SPP modules to our model, it can increase the most important contextual feature of accepting fields and isolation without reducing the detection speed.

D. ShuffleSimAM Attention

The attention mechanism empowers neural networks to achieve precise focus on all relevant elements of the input data, and it has gradually become an important part of improving the performance of deep neural networks. At present, there are two main attention mechanisms widely used in this field, namely spatial attention and channel attention, which are used to capture pixel-level pairwise correlations and inter-channel dependencies, respectively. Although fusing them together may achieve better performance than their respective implementations, it inevitably increases the computational overhead. We referred to the method of Shuffle Attention (SA) [23] and introduced SimAM [24] to replace the spatial attention part of SA. The resultant diagram of ShuffleSimAM is shown in Fig. 2. Although the computational overhead has increased a little, the final result is slightly improved. In order to overcome the drawbacks in the Squeeze-and-Excitation (SE) channel attention, SA adopts an optional method which first uses global average pooling (gap) to generate channel-wise data, and its formula can be expressed as:

$$s = F_{gap}(X_{k1}) = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{k1}(i, j), \qquad (1)$$

where X_{k1} represents the part of sub-feature X_k split into the "channel attention" branch, and s denots channel-wise statistics. The above formula means that, for X_{k1} along the dimension of the channel, the feature map of each channel is summed to get a value, and then divided by the number of pixels in this map, and use this value to replace the feature information of this channel. Then through F_c operation and

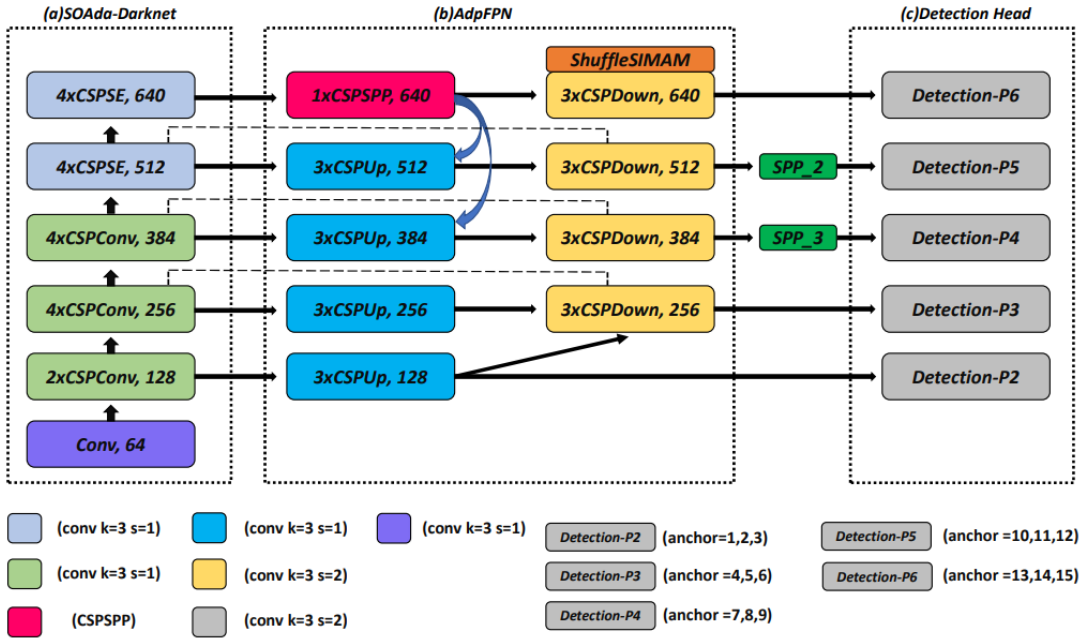


Fig. 1. The structure of SOAda-YOLOR

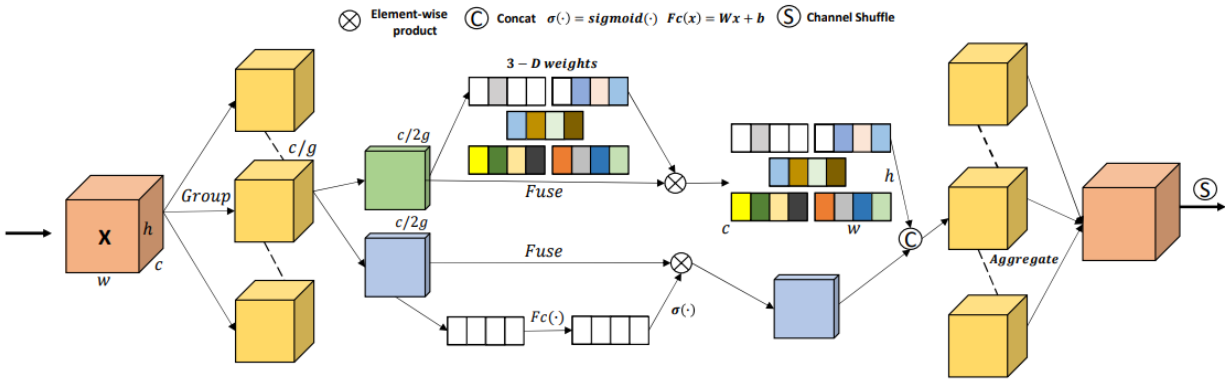


Fig. 2. The structure of ShuffleSimAM Attention.

sigmoid activation to generate the attention weight corresponding to each channel, the output of the channel attention can be obtained by

$$X'_{k1} = \sigma(F_c(s)) * X_{k1} = \sigma(W_1 s + b_1) * X_{k1}, \quad (2)$$

The F_c operation is to multiply s by a matrix $W_1 \in R^{C/2G\times1\times1}$ plus a bias $b_1 \in R^{C/2G\times1\times1}$, whose function is to scale and shift s.

Then we will mention the spatial attention part. According to the SimAM formula, the author defines the following energy function for each neuron:

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2. \quad (3)$$

Here, $\hat{t} = w_t t + b_t$ and $\hat{x}_i = w_t x_i + b_t$ are linear transforms of t and x_i, where t and x_i are the target neuron and other neurons in a single channel of the input feature $X \in R^{C\times H\times W}$. i is index over spatial dimension and $M = H \times W$ is the number of neurons on that channel. w_t and b_t are weight and bias of the transform. The Eq.(3) attains the minimal value when $\hat{t}$ equals to y_t, and all other $\hat{x}_i$ are y_o. Minimizing the above formula is equivalent to train the linear separability between the target neurons t and other neurons in the same channel. In order to simplify the presentation, the authors adopt binary labels and add a regularization term in Eq.(3). The final energy function is defined as follows:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_t x_i + b_t))^2 + (1 - (w_t t + b_t))^2 + \lambda w_t^2. \quad (4)$$

Therefore, the minimum energy can be obtained by the fol-

lowing formula:

$$X'_{k2} = e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{u}^2) + 2\sigma^2 + 2\lambda}, \tag{5}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu}^2)$. Next, we concatenate two features X'_{k1}, X'_{k2} and aggregate all sub-features to complete this task. Similar to ShuffleNetv2 [25], the "channel shuffle" operator is used to realize cross-group information flow along the channel dimension. The final output of SA module retains identical dimensionality to X, thereby facilitating seamless integration with contemporary architectures.

E. Non-maximum suppression (NMS) optimization

On the basis of Soft-NMS [26] and DIoU-NMS [27], the two NMS methods are fused and fine-tuned to detect small objects. When the IoU is less than the chosen threshold, the original Soft-NMS score penalty mechanism is added to the DIoU-NMS. Then the IoU value and the distance between the two boxes are modified and evaluated. The optimized NMS can be expressed as the following formula:

$$TS_i = \begin{cases} S_i, IoU - RDIoU(M, b_i) < Ni \\ S_i(1 - RDIoU(M, b_i)), IoU - RDIoU(M, b_i) \geq Ni, \end{cases} \tag{6}$$

where S_i represents the score of each box, M is the box with the highest score, b_i is one of the remaining boxes, Ni is the threshold, and $RDIoU$ is denoted as follows:

$$RDIoU = \frac{p^2(b^g, b)}{c^2}, \tag{7}$$

where b and b^g represent the center points of the two rectangular boxes, p represents the Euclidean distance between the two rectangular boxes, and c represents the diagonal distance of the closure area of the two rectangular boxes. The comparison of detection results with NMS optimization and without NMS optimization is in Fig. 3.

F. Data Augmentation

For data augmentation [28]–[30], initial conditions delineated in Table I were established, with 500 images excerpted from the validation set for image enhancement via low-light enhancement (LLE) [31], [32]. Detection outcomes post-LLE are visualized in Fig. 4. Furthermore, adhering to small object augmentation methodology, labeling software was utilized for cropping small target labels and duplicating generation, with insertion into 4,000 images from the validation set, comprising 2,000 daytime and 2,000 nighttime images selected by the same criteria in TableI. This subset of 4,000 images was exclusively for small object label augmentation, independent of the previous 500 images augmented via LLE. Additionally, only the following 4 categories were chosen: 1) Traffic signs; 2) Traffic lights; 3) Pedestrians; 4) Cars, with small target labels in each category copied twice, thereby introducing 8 small target labels per image to enrich small object diversity and mitigate original target label imbalance.

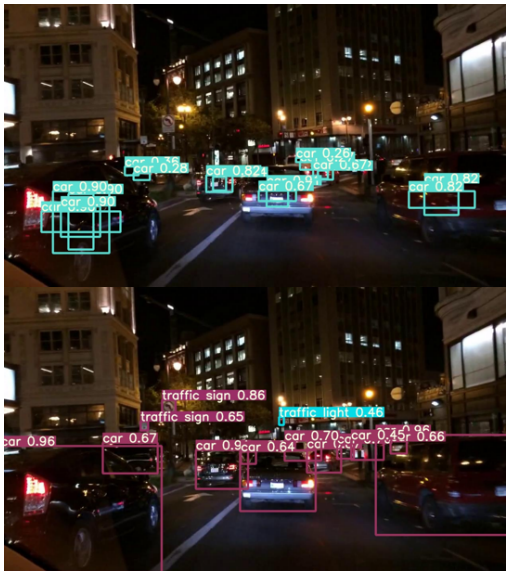


Fig. 3. Top) Detection results without NMS optimization. Bottom) Detection results with NMS optimization.

TABLE I
SELECTION CRITERIA FOR THE AUGMENTATION DATASET

Condition (1).	After the height and width are multiplied, it is less than 10 to 20 pixels
Condition (2).	The photo was taken at night or rainy day
Condition (3).	The category focus on traffic lights, traffic signs, cars, pedestrians
Condition (4).	In a single photo, there are more than or equal to 3 targets that meet the criteria of condition (1)

G. Bag of freebies

The variation of activation functions is a part of the Bag of freebies. For distinct models and task modalities, different activation function imports can lead to remarkable improvements. Accordingly, we conducted comparative analyses of different activation functions, scrutinizing their traits and limitations, and finally chose to import the Cyclemoid activation function into the model. Although it will bring a little calculation, the final result shows that Cyclemoid has its advantages, where the Cyclemoid activation function can be defined as:

$$Cyclemoid = tanh(\pi x) * tanh(\pi x^2 - 0.95)^2 * cos(-3 \leq x \leq 3)^2. \tag{8}$$

Regarding loss functions, SIoU loss [17] was appended to the native loss function. This can take into account the vector angle between the required regressions and redefine the penalty index. Thus, there can be slight improvements in training performance.

IV. EXPERIMENTS

For experimentation, we use YOLOR [14] as the baseline, and combine the methods described in Section III, called SOAda-YOLOR. Training and evaluation were conducted on the BDD100k dataset [33], one of the largest and most diverse

TABLE II
OBJECT DETECTION RESULTS ON BDD100K DATASET [33] AND KITTI DATASET [34]

Models	BDD100k [33] *Precision*	*Recall*	*mAP@0.5*	*mAP@.5 : .95*	*Time*	KITTI [34] *Precision*	*Recall*	*mAP@0.5*	*mAP@.5 : .95*	*Time*	*Params*	*FLOPs*
YOLOR [14]	47.7%	60.9%	59.7%	35.1%	101hr	85.1%	**96.9%**	97.7%	83.7%	13.3hr	37.2M	80.6G
YOLOv3 [10]	76.1%	55.4%	59.5%	31.6%	133hr	93.7%	91.3%	94.1%	70.6%	20.1hr	61.6M	155.2G
YOLOv4 [11]	33.5%	49.3%	39.8%	21.2%	132hr	91.4%	86.9%	92.2%	77.5%	28.3hr	64.3M	142.8G
YOLOv5	76.6%	54.5%	59.9%	34%	111hr	**96.4%**	90.8%	95.2%	76.1%	20.8hr	20.9M	48.2G
Scaled-YOLOv4 [12]	36.1%	62.4%	54.7%	28.6%	118hr	73.4%	94.7%	94.6%	**95%**	19.2hr	52.5M	119.4G
YOLOv7 [35]	**78.4%**	57.8%	**64.9%**	**37.1%**	125hr	93.3%	92.1%	94.4%	72.9%	18.4hr	37.2M	104.9G
Tph-YOLOv5 [36]	44.6%	38.4%	37.7%	18.2%	**89.3hr**	92%	88.5%	93.5%	66.4%	**11.6hr**	17.7M	4.2G
SOAda-YOLOR (Ours)	52.5%	**63.1%**	62.6%	36.5%	105hr	94.5%	96.3%	**98.1%**	86.3%	16.5hr	38.3M	85.3G

TABLE III
ABLATION STUDY ON BDD100K DATASET [33]

Model	BDD100k [33] CSPSE-D	Adp-FPN	ShuffleSIMAM	Cyclemoid	SIoU	Data Aug	Opt-NMS	mAP@.5	mAP @.5:.95	Precision	Recall
SOAda-A	✓	-	-	-	-	-	-	60.4%	35.5%	48.6%	61.5%
SOAda-B	✓	✓	✓	-	-	-	-	60.9%	35.6%	49.9%	62.2%
SOAda-C	✓	✓	-	✓	-	-	-	60.9%	35.6%	50.1%	62.2%
SOAda-D	✓	✓	-	✓	✓	-	-	61.4%	35.6%	50.9%	61.7%
SOAda-E	✓	✓	✓	✓	✓	-	✓	62.4%	36.2%	52.3%	62.8%
SOAda-F	✓	-	✓	✓	-	✓	✓	61.1%	35.9%	52.1%	62.2%
SOAda-G	✓	✓	-	-	✓	✓	✓	62.3%	36%	52.1%	62.7%
SOAda-H	-	-	✓	✓	✓	✓	✓	61.3%	35.7%	51.4%	61.8%
SOAda-I	✓	✓	✓	✓	✓	✓	✓	62.6%	36.5%	52.5%	63.1%



Fig. 4. Top) Detection results before LLE. Bottom) Detection results after LLE.

TABLE IV
COCOAPI ON BDD100K DATASET [33]

Model	BDD100k [33] AP_S	AP_M	AP_L	Data Aug
YOLOR [14]	39.1%	57.4%	68.9%	✓
YOLOv3 [10]	35.5%	55.1%	63.4%	✓
YOLOv4 [11]	32.9%	54.2%	62.8%	✓
YOLOv5	39.9%	59.3%	70.2%	✓
Scaled-YOLOv4 [12]	33.7%	55.7%	65.2%	✓
YOLOv7 [35]	41.4%	59.3%	69.8%	✓
Tph-YOLOv5 [36]	38.2%	54.9%	66.4%	✓
SOAda-YOLOR (Ours)	40.6%	59.1%	69.2%	✓

autonomous driving datasets to date, released by UC Berkeley. After adding the augmentation data to the existing training set, the final number of training set is 74,500 (70000+4000+500), the number of validation set is 15,500 (20000-4000-500), and the number of test set is 10,000. The training epoch number is set to 100, the number of anchors is increased from 9 to 12, and the input image resolution is 1280×1280 pixels. The time to complete the training is 51 hours.

KITTI [34] is another dataset that covers most categories for autonomous driving detection. In Table II, we compared the precision, recall, mean Average Precision (mAP), and Time results with other SOTA methods, including YOLOR [14], YOLOv3 [10], YOLOv4 [11], YOLOv5, Scaled-YOLOv4 [12], YOLOv7 [35], and Tph-YOLOv5[36]. We can see our proposed model has the highest or the 2nd highest mAP among all models in both BDD100k [33] abd KITTI [34]. Also, the Time, parameters, and FLOPs of our approach are smaller than Scaled-YOLOv4 and YOLOv7. To prove the effectiveness of each component in Section III, we perform the ablation study in Table III. We can conclude the performance will be the best when using SOAda-I model (i.e., using all the components). As evident in Fig. 5, we can see that our model can detect more small objects, and increase more confidence for other objects in the road images at night. It proves that our model can have high accuracy in this environment. In addition, we compare the test results of AP_S, AP_M, and AP_L for each model in BDD100k. The experimental results can be found in Table IV.



Fig. 5. Detection results from SOAda-YOLOR in BDD100k dataset.

V. Conclusions

In this paper, we propose a model tailored for small object detection in the context of road driving. First, We optimize the original CSP module to become CSPSE-D Block. Secondly, we take YOLOR as our backbone and deepen its modules and enhance the connectivity between layers by proposing a new model similar to Bi-FPN, and use Dropout in the first layer to avoid the model learning bad features. Thirdly, we propose a novel ShuffleSimAM module for the backbone, and improve the problem of poor detection for small targets. Moreover, the traditional NMS [37] is optimized, so that it can detect objects and reduce redundant detection frames as many as possible. Then for data augmentation, we propose a solution targeting the imbalance between large and small object labels. Through the LLE and small object label augmentation, the final trained model has achieved performance improvements in detection results. Finally, we apply some bag-of-freebies to our model to further improve the detection performance.

References

[1] T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Comparison of subjective viewing test methods for image quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3155–3159.

[2] T.-J. Liu, "Study of visual quality assessment on pattern images: Subjective evaluation and visual saliency effects," *IEEE Access*, vol. 6, pp. 61 432–61 444, 2018.

[3] T.-J. Liu, C.-T. Lin, H.-H. Liu, and S.-C. Pei, "Blind stereoscopic image quality assessment based on hierarchical learning," *IEEE Access*, vol. 7, pp. 8058–8069, 2019.

[4] T.-J. Liu, K.-H. Liu, and K.-H. Shen, "Learning based no-reference metric for assessing quality of experience of stereoscopic images," *Journal of Visual Communication and Image Representation*, vol. 61, pp. 272–283, 2019.

[5] T.-J. Liu, K.-H. Liu, and H.-H. Liu, "Temporal information assisted video quality metric for multimedia," in *IEEE International Conference on Multimedia and Expo*, 2010, pp. 697–701.

[6] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Recent developments and future trends in visual quality assessment," *APSIPA ASC*, 2011.

[7] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "A fusion approach to video quality assessment based on temporal decomposition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–5.

[8] B.-X. Chen, T.-J. Liu, K.-H. Liu, H.-H. Liu, and S.-C. Pei, "Image super-resolution using complex dense block on generative adversarial networks," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2866–2870.

[9] T.-J. Liu and Y.-Z. Chen, "Satellite image super-resolution by 2d rrdb and edge-enhanced generative adversarial network," *Applied Sciences*, vol. 12, no. 23, p. 12 311, 2022.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled yolov4 scaling cross stage partial network," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2021, pp. 13 029–13 038.

[13] H. Zhang, F. Li, S. Liu, *et al.*, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[14] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv:2105.04206*, 2021.

[15] T.-J. Liu, K.-H. Liu, J. Y. Lin, W. Lin, and C.-C. J. Kuo, "A paraboost method to image quality assessment," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 1, pp. 107–121, 2017.

[16] T.-J. Liu and K.-H. Liu, "No-reference image quality assessment by wide-perceptual-domain scorer ensemble method," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1138–1151, 2018.

[17] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022.

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[19] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[21] R. Girshick, "Fast rcnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[23] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 2235–2239.

[24] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*, PMLR, 2021, pp. 11 863–11 874.

[25] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[26] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

[27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 12 993–13 000.

[28] K.-Y. Wen, T.-J. Liu, K.-H. Liu, and D.-Y. Chao, "Identifying poultry farms from satellite images with residual dense u-net," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 102–107.

[29] K.-C. Chang, T.-J. Liu, K.-H. Liu, and D.-Y. Chao, "Locating waterfowl farms from satellite images with parallel residual u-net architecture," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 114–119.

[30] L.-C. Lan, T.-J. Liu, and K.-H. Liu, "Age regression with specific facial landmarks by dual discriminator adversarial autoencoder," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 2718–2722.

[31] M. Tanaka, T. Shibata, and M. Okutomi, "Gradient-based low-light image enhancement," in *IEEE International Conference on Consumer Electronics (ICCE)*, 2019, pp. 1–2.

[32] T.-J. Liu, H.-H. Liu, S.-C. Pei, and K.-H. Liu, "A high-definition diversity-scene database for image quality assessment," *IEEE Access*, vol. 6, pp. 45 427–45 438, 2018.

[33] F. Yu, H. Chen, X. Wang, *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[34] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[35] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[36] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2778–2788.

[37] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "Nms by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 750–10 759.

Badminton Self-Training System Based on Virtual Reality

Wei-Shen Tai*, Kuan-Hsien Liu
*National Taichung University of Science and Technology, Taiwan
E-mail: wstai@nutc.edu.tw Tel: +886-4-22196368
National Taichung University of Science and Technology, Taiwan
E-mail: khliu@nutc.edu.tw Tel: +886-4-22196627

***Abstract* —In sports games, quick sensory processing and attention are crucial for athletes to win. Therefore, players need to strengthen their perceptual response and movement coordination through iterative training to increase their chances of winning. However, conventional badminton training is limited by the availability of training venues and coaches. Additionally, training materials are imprecise and unstable due to human factors, and players cannot effectively analyze their strengths and weaknesses from training results.**

This study proposes a virtual reality badminton self-training system to solve these training problems, including the lack of training venues and coaches, imprecise training materials, and ineffective analysis of strengths and weaknesses. As a result, the proposed system provides players with an effective evaluation of their training performance under stable and precise training processes. Players can identify their shortcomings and receive constructive suggestions for improvement.

I. INTRODUCTION

Virtual Reality (VR) applies 3D computer rendering to construct a virtual environment that resembles the actual world. Users can directly interact with virtual objects through user interfaces or sensing devices such as helmets and controls in an immersive environment [1, 2]. VR can simulate interactive effects between objects, like those in the physical environment, making it widely applicable in several domains, such as training, education, and entertainment [3, 4].

The advantages of VR for educational training include:

i. 3D visualization can present complex, abstract problems in a visual interface to facilitate observation and interpretation.

ii. Learners can attempt to solve problems differently without operating restrictions.

iii. First-hand hypotheses and test results can be acquired through the user's direct interaction with virtual objects.

iv. The VR environment can enable learners to acquire knowledge more efficiently through repeated practice [5, 6].

II. VR AND BADMINTON

Recently, VR has been adopted in badminton training systems. This includes watching detailed movements closely via VR helmet [7], using motion capture equipment to perform full-body badminton combat training [8, 9], recording eye tracking and smash actions to verify hand-eye coordination between beginners and experienced players [10], and using Kinect and an inertial measurement unit (IMU) to capture hand tracks and evaluate the effectiveness of badminton players' movements [11, 12].

Badminton Kings VR and Space Badminton are two 3D VR games on the Steam VR website. Players can challenge virtual opponents and judge effectiveness based on their scores. However, these games still focus on entertainment and lack further analysis to identify weaknesses in player training.

During the training process, objective information collected through information technology can more effectively assist athletes in understanding their own shortcomings [13-16]. These shortcomings may include the lack of coordination in certain body parts or insufficient proficiency in certain movements. Furthermore, this information can help coaches better understand athletes' performances and provide more personalized advice. For example, when coaches understand that athletes are performing poorly in certain areas, they can focus on specialized training in those areas. In addition, by collecting and analyzing information technology, coaches can better evaluate athletes' progress and adjust training plans for better results.

In summary, badminton training studies mainly focus on learner operation and battle practice but lack the ability to analyze learner problems during training processes effectively. We also plan to incorporate machine learning or deep learning

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

methods to help training process such as facial age estimation [17, 18], facial expression recognition [19], clothing retrieval [20, 21], segmentation of remote sensing image [22], and visual quality assessment [23]. This study aims to address this issue by building a VR badminton self-training system that provides players with precise and stable training materials during badminton training. The system will also provide players with practical analysis and suggestions to help them identify their own shortcomings more effectively.

III. VR Badminton Self-training

The badminton self-training system uses HTC VIVE devices to perform badminton training settings and actions. A virtual badminton court is modeled in 3D (see Fig. 1). The system has four functions: posture demonstration, serve practice, hitting practice, and effectiveness analysis (see Fig. 2).

A. Posture Demonstration

On the left side of the court (Fig. 3), badminton swing, forehand serve and backhand serve are demonstrated via teaching animations. Players can imitate and adjust their posture in accordance with demonstration contents.

B. Serve Practice

In this stage, the focal point is to improve the precision of the serve. Players can set the target position through the setting panel on the right side of the court as shown in Fig. 4. Then players must hit the target accurately by manipulating their own serve. All hitting data will be recorded and analyzed in the effectiveness analysis.

C. Hitting Practice

In Fig. 4, players set the serve point, landing point, height, and speed of the serve machine according to their distinct needs. Players can overcome their difficulties of specified hitting positions by means of iterative practices in this stage. These hitting data will also be recorded and analyzed in the effectiveness analysis.

D. Effectiveness Analysis

Players can turn on/off the traveling track to observe badminton trajectories during the training process. The analyzed results from serve and hitting records are presented in the effectiveness analysis panel. They show hit rate, hitting deviation, strength, and speed to assist players in effectively understanding the strengths and weaknesses of the training.



Fig. 1 VR scene in the badminton self-training system.

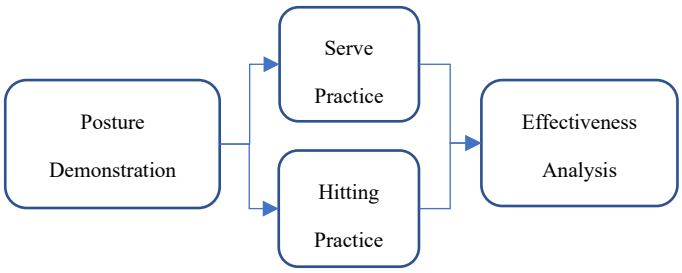


Fig. 2 Functions of the badminton self-training system.



Fig. 3 Posture demonstration (backhand serve).

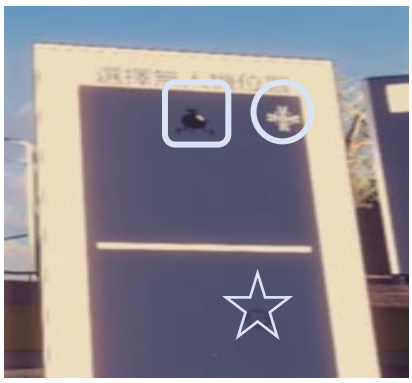


Fig. 4 Serve and hitting practice settings(circle, rectangle and star denote the target position, serve point and landing point, respectively).



Fig. 5 Effectiveness analysis panel.

IV. Training Suggestion

During badminton training, analyzing the results of serving and hitting for the player's reactions is one of the important issues for understanding and improving the player's weaknesses. With the assistance of artificial intelligence algorithms, the historical records of the player's training process can be analyzed, and individual analysis can be carried out for the serving and hitting status.

For example, when serving, whether the player has a blind spot on the landing point and cannot let the ball fall in a scoring position. In the process of hitting, whether the player has a high turnover rate in a specific position, the contact position between the ball and the racket face, the acceleration force, and the information about the landing position after hitting are analyzed.

The training system finds the focus that needs to be strengthened during training by analyzing the player's problems. For example, when the distance between the two landing points of the hit ball is too large and the player cannot react to the hit in time, the player's movement speed can be strengthened, or other hitting methods can be used to improve the problem of the distance between the landing points being too large.

The virtual coach in the system can give players suggestions on ball route patterns based on training analysis results, including improving the player's movement training method and hitting method, which can help players overcome their weaknesses in subsequent practices. In the future, artificial intelligence may be integrated to correct the target position of the badminton serve and the hitting trajectory of the serve.

For example, if the player has specific routes that they are not good at, the number of serves and hitting trajectories can be increased for individual weaknesses to enhance the training intensity of individual players' serving and hitting weaknesses. This can lead to a better understanding of the player's strengths and weaknesses, as well as the potential adjustments that can be made to improve their overall performance.

Moreover, the analysis of serving and hitting can also help players to understand the importance of positioning and timing, which are crucial factors in badminton. By analyzing the player's movement and reaction time, the system can provide targeted training suggestions to help players improve their positioning and timing, which in turn can improve their overall performance.

V. Conclusion

In conventional badminton training, players are restricted by the availability of a venue and coach. This badminton self-training system overcomes these limitations by constructing a realistic 3D VR environment. Players can practice as much as they need according to their individual requirements without being limited by time, space, or the availability of a coach.

This system provides stable and precise training materials without human intervention. Players can set various training conditions based on their own weaknesses. Finally, the training results are analyzed and presented in the effectiveness analysis panel. The analysis of serving and hitting in badminton training is a crucial aspect of understanding and improving a player's weaknesses, and the use of artificial intelligence algorithms can greatly enhance this process. By providing targeted training suggestions and automatic corrections, the system can help players to overcome their weaknesses and improve their overall performance effectively through the proposed system.

REFERENCES

[1] M. Hussein and C. Nätterdal, "The benefits of virtual reality in education-A comparision Study," Bachelor, Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Göteborg, Sweden, 2015.

[2] R. Yung and C. Khoo-Lattimore, "New realities: a systematic literature review on virtual reality and augmented reality in tourism research," *Current Issues in Tourism,* vol. 22, no. 17, pp. 2056-2081, 2019, doi: 10.1080/13683500.2017.1417359.

[3] M. Vesisenaho *et al.*, "Virtual Reality in Education: Focus on the Role of Emotions and Physiological Reactivity," *Journal of Virtual Worlds Research,* vol. 12, no. 1, pp. 1-15, 2019, doi: 10.4101/jvwr.v12i1.7329.

[4] K.-T. Huang, C. Ball, J. Francis, R. Ratan, J. Boumis, and J. Fordham, "Augmented Versus Virtual Reality in Education: An Exploratory Study Examining Science Knowledge Retention When Using Augmented Reality/Virtual Reality Mobile Applications," *Cyberpsychology, Behavior, and Social Networking,* vol. 22, no. 2, pp. 105-110, 2019, doi: 10.1089/cyber.2018.0150.

[5] T. Rose, C. S. Nam, and K. B. Chen, "Immersion of virtual reality for rehabilitation-Review," *Applied ergonomics,* vol. 69, pp. 153-161, 2018, doi: 10.1016/j.apergo.2018.01.009.

[6] M. Javaid and A. Haleem, "Virtual reality applications toward medical field," *Clinical Epidemiology and Global Health,* vol. 8, no. 2, pp. 600-605, 2020, doi: 10.1016/j.cegh.2019.12.010.

[7] H.-Y. Lee, Chung, Chi-Yang, Yang, Fan, "Research on Virtual Reality-Based Badminton Teaching in Physical Education Courses," presented at the 2019 Hawaii University International Conferences Science, Technology& Engineering, Arts, Mathematics & Education, Honolulu, Hawaii, 2019.

[8] G. A. Ginja, "Applications of Virtual Reality in the Practice of Para-Badminton," in *2018 20th Symposium on Virtual and Augmented Reality (SVR)*, 28-30 Oct. 2018 2018, pp. 230-232, doi: 10.1109/SVR.2018.00042.

[9] M. D. B. M. J. a. D. B. A. Ghani, "Badminton: To Promote Badminton Sport Using Virtual Reality Among Younger Generations," *Journal of Engineering and Applied Sciences,* vol. 13, no. 20, pp. 8535-8552, 2018, doi: 10.36478/jeasci.2018.8535.8552.

[10] K. Ishibe, S. Aihara, Y. Hayashi, and H. Iwata, "The development of an immersive three-dimensional virtual reality system for identifying hand–eye coordination in badminton," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020: IEEE, pp. 1778-1784.

[11] Z. Taha, M. S. S. Hassan, H. J. Yap, and W. K. Yeo, "Preliminary investigation of an innovative digital motion analysis device for badminton athlete performance evaluation," *Procedia engineering,* vol. 147, pp. 461-465, 2016, doi: 10.1016/j.proeng.2016.06.341.

[12] Z. Taha, M. Y. Wong, H. J. Yap, A. Abdullah, and W. K. Yeo, "Evaluation of real time motion tracking accuracy of customised IMU sensor for application in a mobile badminton virtual reality training system," *Movement, Health & Exercise,* vol. 7, no. 1, pp. 201-209, 2018, doi: 10.15282/mohe.v7i1.185.

[13] J. Liu, G. Huang, J. Hyyppä, J. Li, X. Gong, and X. Jiang, "A survey on location and motion tracking technologies, methodologies and applications in precision sports," *Expert Systems with Applications,* vol. 229, p. 120492, 2023/11/01/ 2023, doi: 10.1016/j.eswa.2023.120492.

[14] T. Li *et al.*, "Design and feasibility study of a HCPS framework-based VR alpine skiing decision-making training system," *Computers & Graphics,* vol. 114, pp. 138-149, 2023/08/01/ 2023, doi: 10.1016/j.cag.2023.06.007.

[15] Y. Tong, N. Li, and Q. Wang, "Low energy split-cluster wireless acceleration sensors for motion and posture change recognition in sports," *Alexandria Engineering Journal,* vol. 76, pp. 179-192, 2023/08/01/ 2023, doi: 10.1016/j.aej.2023.06.039.

[16] J. Yin and X.-q. Wang, "Study on safety mode of dragon boat sports physical fitness training based on machine learning," *Safety Science,* vol. 120, pp. 1-5, 2019/12/01/ 2019, doi: 10.1016/j.ssci.2019.06.028.

[17] K. H. Liu, C. H. Chiu, and T. J. Liu, "Fusion of Triple Attention to Residual in Residual Dense Block to Attention Based CNN for Facial Expression Recognition," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 9-12 Oct. 2022 2022, pp. 1045-1050, doi: 10.1109/SMC53654.2022.9945323.

[18] K. H. Liu, Y. H. Wu, and T. J. Liu, "Clothing Retrieval from Channel Attention to KN Loss Learning," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 13-16 Dec. 2022 2022, pp. 1-5, doi: 10.1109/VCIP56404.2022.10008803.

[19] K. H. Liu, Y. H. Wu, and T. J. Liu, "Clothing Retrieval from Vision Transformer and Class Aware Attention to Deep Embedding Distance Learning," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 9-12 Oct. 2022 2022, pp. 2378-2383, doi: 10.1109/SMC53654.2022.9945226.

[20] K.-H. Liu and B.-Y. Lin, "MSCSA-Net: Multi-Scale Channel Spatial Attention Network for Semantic Segmentation of Remote Sensing Images," *Applied Sciences,* vol. 13, no. 17, p. 9491, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/17/9491.

[21] K. H. Liu and T. J. Liu, "A Structure-Based Human Facial Age Estimation Framework Under a Constrained Condition," *IEEE Transactions on Image Processing,* vol. 28, no. 10, pp. 5187-5200, 2019, doi: 10.1109/TIP.2019.2916768.

[22] K. H. Liu, S. Yan, and C. C. J. Kuo, "Age Estimation via Grouping and Decision Fusion," *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 11, pp. 2408-2423, 2015, doi: 10.1109/TIFS.2015.2462732.

[23] T. J. Liu and K. H. Liu, "No-Reference Image Quality Assessment by Wide-Perceptual-Domain Scorer Ensemble Method," *IEEE Transactions on Image Processing,* vol. 27, no. 3, pp. 1138-1151, 2018, doi: 10.1109/TIP.2017.2771422.

ROTATION ANGLE DETECTION USING A PILOT SIGNAL FROM ROTATED STEGO-IMAGE

Rinka Kawano and Masaki Kawamura
Yamaguchi University, Yamaguchi, Japan
E-mail: {m.kawamura,rinka1005}@m.ieice.org Tel: +81-83-933-5701

***Abstract*—Robust watermarking methods are necessary to resist attacks. With conventional watermarking methods, watermarks are extracted and their errors are corrected regardless of the type of attack. For example, error-correcting codes and spread spectrum are used to correct watermark errors. To reduce watermark errors, embedding methods in the frequency domain, such as DCT and DFT, and the use of scale invariant feature transform have been proposed. These methods are effective for non-geometric attacks such as additive noise and distortion. However, they do not work well against geometric attacks due to the lack of synchronization. It is necessary to perform optimal synchronization for geometric attacks such as scaling, rotation, cropping, etc. We focus on the rotation attack and propose a method of estimating the rotation angle using a grid-shaped pilot signal. While suppressing the original image-derived slant, the slant of the pilot signal can be emphasized and detected. Thus, the rotation angle of a stego-image can be accurately estimated.**

I. Introduction

Watermarking is a technique for embedding secret information in digital content such as still images. The embedded information is called a watermark, and the image into which the watermark is embedded is called a stego-image. Image-watermarking methods can be used to manage digital content and prevent unauthorized use of images. For example, by embedding an invisible signature as a watermark in an image, the author can claim their copyright when the image is illegally used. However, many of the images that are misused in social media are processed by scaling, rotating, cropping, and so on. The processed images are also always compressed when they are saved. That is, published images may be vulnerable to different types of attacks. Attacks on these images may cause loss of the watermark. Therefore, robust watermarking methods are necessary against various attacks.

The simplest method is to embed the watermark directly into the pixel values. However, it is easy to detect the position of the watermark, and the watermark can be easily removed by an attacker. To make watermarks imperceptible, they are embedded in the frequency domain by using discrete cosine transforms (DCT), discrete Fourier transforms (DFT), discrete wavelet transforms (DWT), and so on [1], [2], [3]. The watermark in the frequency domain also tends to be insensitive to errors even if the image is degraded by compression or other processing.

Watermark errors are caused by degradation of the image due to attacks. There are two main types of attacks: geometric and non-geometric. Non-geometric attacks are those that change pixel values, such as JPEG compression and noise addition. While they degrade the watermark, its coordinates remain unchanged. Therefore, conventional error-correcting methods can be applied to the watermark. These methods include using error-correcting codes [4] and spread-spectrum methods [5], [6]. However, if the watermark was not extracted, no error correction is conducted. Geometric attacks are those that change the positions of pixels due to scaling, rotation, or clipping of the image. Since geometric attacks change the positions of the watermark, error-correcting methods cannot be applied without synchronization. With watermarking methods using scale invariant feature transform (SIFT) features [7], synchronization can be achieved at SIFT feature points. Watermarks are embedded around the feature points, allowing the watermark to resist geometrical attacks [8], [9], [10].

With many conventional watermarking methods, the redundancy of the watermark is used to recover the watermark even if it is distorted by an attack. However, if the type and strength of the attack could be estimated, it would be possible to extract the watermark more accurately. Therefore, we focused on reducing errors by estimating the type and strength of attacks on the basis of the framework of communication-channel models in which a message degraded through the channel is received. Hyper-parameters, such as noise strength of the channel, are estimated by a signal called a pilot signal, which is distinct from the message. Therefore, the pilot signal is introduced into a watermarking method to estimate an attack. If the embedding position of the watermark can be estimated correctly, watermarking errors would be reduced. Assuming that an attack can be estimated, the embedding strength of the watermark would be reduced, resulting in improved image quality.

Su *et al.* [11] proposed a watermarking method using pilot signals. With this method, SIFT feature points are first detected, then pilot signals around the points are detected to discover the embedding positions of watermarks. The embedding regions are rotated so that the gradient direction of the SIFT feature is the reference angle. In other words, the image quality and watermarks degrade due to the rotation of the image during the embedding process. Note that both the watermark and pilot signal are embedded in the same domain. This method doesn't use a pilot signal for estimation of the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

attack parameters.

A novel approach is required to estimate the type and strength of attacks. The optimal shape of the pilot signal for each attack may be different, and its embedding technique may also be different. It is difficult to determine a type of attack due to there being many types. Therefore, we fix the type of attack and focus on estimating the strength of the attack. In our previous study, we estimated the magnification ratio of the image using a grid-shaped pilot signal [12]. Therefore, in this paper, we propose a method of estimating the rotation angle using the pilot signal.

This paper is organized as follows. In Section II, we present the proposed method for estimating the rotation angle using the pilot signal. In Section III, we present the results from computer simulations conducted to verify the proposed method. We conclude our study in Section IV.

II. Estimation of Rotation Angle Using Pilot Signal

The proposed method estimates the rotation angle of a stego-image using the pilot signal. The advantage of this method is that the estimated rotation angle can be used to synchronize the embedding regions, potentially reducing watermarking errors. Eventually, watermarks will be able to be extracted with smaller embedding strengths, improving the quality of stego-images.

Similar to our previous study on magnification-ratio estimation [12], we focused on the embedding of a grid-shaped pilot signal. When the stego-image is rotated, orthogonal lines corresponding to the rotation angle appear in the DFT coefficients. Therefore, the rotation angle can be estimated by calculating the slant of the orthogonal lines.

A. embedding method

The original color image is decomposed into its YUV components, and the watermark is embedded in the Y component, the luminance value. To avoid affecting the watermark, the pilot signal is embedded in the V or U component. In the following explanation, the U component is used. The size of the image is $L_w \times L_h$ pixels, and the pixel value at coordinate (x, y) is denoted as $U(x, y)$.

A grid-shaped pilot signal is embedded into $U(x, y)$ by using quantization index modulation (QIM) [13]. Figure 1 shows the shape of the pilot signal. The grid interval γ and grid width β shown in Fig. 1 are $\gamma = 50$ and β pixels, respectively. The pilot signal takes the value 0 or 1. The yellow area is embedded with the value 0, and the blue area with the value 1. The green regions, which are the intersections of the two types of grids, are embedded with alternating values of 0 and 1. Other non-colored regions are left unchanged.

With QIM [13], the pilot signal is embedded by quantizing the original value with a step size Δ_r. The pilot signal $p \in \{0, 1\}$ is embedded in $U(x, y)$ by using QIM, and the embedded pixel value $U'(x, y)$ is obtained by

$$U'(x,y) = 2\Delta_r \left(\left\lfloor \frac{U(x,y)}{2\Delta_r} - \frac{p}{2} + 0.5 \right\rfloor + \frac{p}{2} \right). \quad (1)$$

In this study, Δ_r of QIM was set to 10.

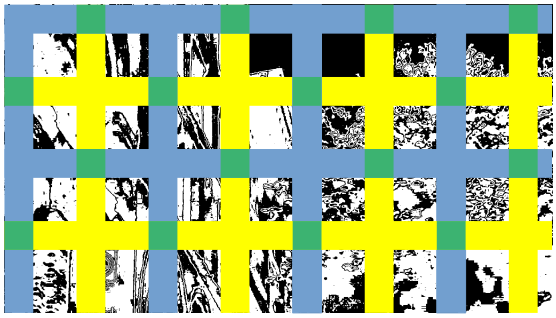


Fig. 1. Shape of pilot signal. Pilot signal is embedded in colored regions by using QIM. Non-colored regions are left unchanged.

B. Detection of rotation angle

We describe the procedure for estimating the rotation angle on the basis of the pilot signal. We assume that the stego-image has been rotated by an angle of $-180 \leq \theta < 180$ degrees. Since the pilot signal has a grid shape, slant components appear as orthogonal lines in the DFT coefficients when the stego-image is rotated. Therefore, the rotation angle can be estimated by measuring the slant of these lines.

Figure 2 shows the procedure for estimating the rotation angle. Each processing step is described below.

1) Color-space separation: An image is decomposed into its YUV components, then the U-component image is divided into N_s blocks of $L_r \times L_r$ pixels. We set the length of one side of the block to $L_r = 300$ pixels. If the block size is too small, the period of the pilot signal cannot be detected. The block size L_r should be several times larger than the grid interval γ.

2) Extraction of embedded information: Binary embedded information $\hat{p}^l(x, y)$ is extracted from the l-th block $U^l(x, y)$ by using QIM. Embedded information can be obtained by

$$\hat{p}^l(x,y) = \left\lfloor \frac{U^l(x,y)}{\Delta_r} + 0.5 \right\rfloor \bmod 2. \quad (2)$$

The pilot signal or original image-derived information is extracted as $\hat{p}^l(x, y)$.

3) Amplitude-spectrum calculation: When each block is transformed by DFT, the DFT coefficients are given by

$$F^l(k_1,k_2) = \frac{1}{L_r} \sum_{x=0}^{L_r-1} \sum_{y=0}^{L_r-1} \hat{p}^l(x,y) e^{-2\pi j \left(\frac{k_1 x}{L_r} + \frac{k_2 y}{L_r} \right)}, \quad (3)$$

where j is an imaginary unit. The amplitude spectrum $|F^l|$ is calculated from the DFT coefficients $F^l(k_1, k_2)$. Denoting the real and imaginary parts of the DFT coefficients as $\mathrm{Re}\ F^l(k_1, k_2)$ and $\mathrm{Im}\ F^l(k_1, k_2)$, respectively, the amplitude spectrum is given by

$$|F^l(k_1,k_2)| = \sqrt{(\mathrm{Re}\ F^l(k_1,k_2))^2 + (\mathrm{Im}\ F^l(k_1,k_2))^2}. \quad (4)$$

As mentioned above, when the stego-image is rotated, the slants due to the pilot signal appear as orthogonal lines in the

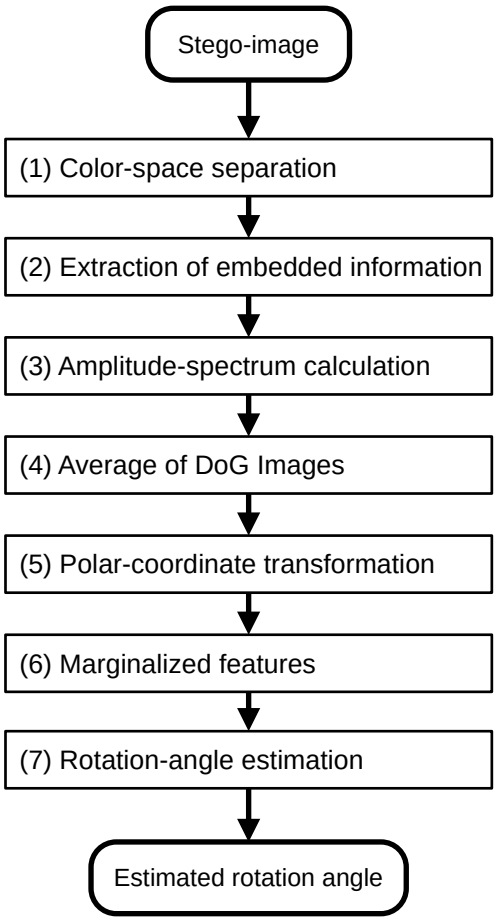


Fig. 2. Procedure for estimating rotation angle

amplitude-spectrum $|F^l|$. In addition, horizontal and vertical lines through the center point appear in a few cases. Figure 3 is the amplitude spectrum when the image is rotated 30 degrees. We can find that orthogonal lines appear through the center point corresponding to the rotation angle.

4) Average of DoG Images: The orthogonal lines may be difficult to see depending on the content of the images. To improve line detection, the Difference of Gaussian (DoG) filter [7] is applied to $|F^l|$. Let $\mathrm{DoG}^l(k_1, k_2)$ be the l-th DoG image generated by the filter. The average DoG image $\overline{\mathrm{DoG}}(k_1, k_2)$ over N_s blocks is computed by

$$\overline{\mathrm{DoG}}(k_1, k_2) \quad = \quad \frac{1}{N_s} \sum_{l=1}^{N_s} \mathrm{DoG}^l(k_1, k_2). \tag{5}$$

By dividing the image into blocks and averaging the DoG images of the blocks, the components derived from the original image are averaged out, and the components derived from the pilot signal can be emphasized.

5) Polar-coordinate transformation: To extract the slant, $\overline{\mathrm{DoG}}(k_1, k_2)$ is converted to polar coordinates. The coefficients converted to polar coordinates are defined as feature $G(\theta, \rho)$. The range of the radius ρ is $0 \leq \rho \leq \sqrt{2}L_r/2$. The radius ρ and angle θ are quantized to integers. Note that the

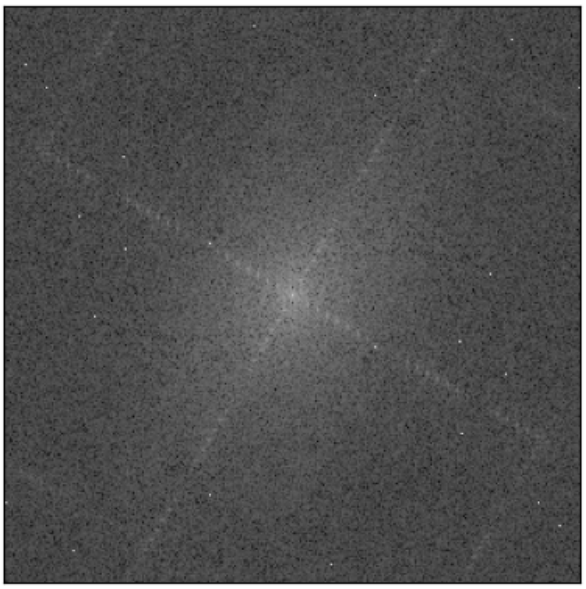


Fig. 3. Amplitude spectrum at 30-degree rotation. Slanted orthogonal lines appear.

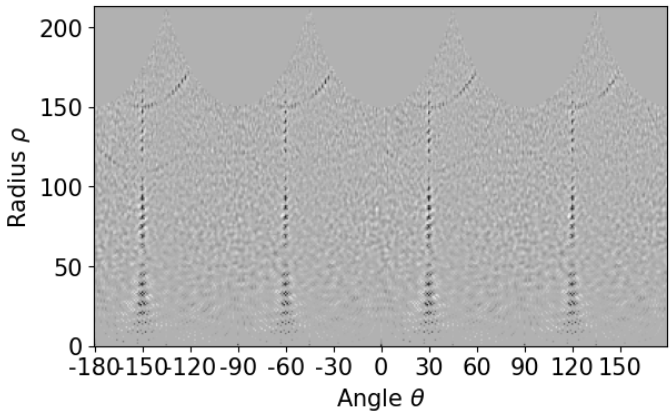


Fig. 4. Feature space. Minima appear at position of rotation angle.

components of the feature corresponding to the regions outside the square block are assumed to be $G(\theta, \rho) = 0$.

Figure 4 shows $G(\theta, \rho)$ when the stego-image is rotated 30 degrees. The horizontal axis represents θ, and the vertical axis represents ρ. We can see that minima appear at $30 + 90n, n = -2, -1, 0, 1$ degrees.

6) Marginalized features: The features are marginalized for the radial direction to determine the rotation angle. The marginal distribution $S(\theta)$ of the features can be given by

$$S(\theta) \quad = \quad \sum_{\rho=0}^{\sqrt{2}L_r/2} G(\theta, \rho). \tag{6}$$

The distribution $S(\theta)$ is also normalized so that it has mean 0 and variance 1. Letting the mean of $S(\theta)$ be μ_S and the standard deviation be σ_S, the Normalized feature intensity

$z(\theta)$ is given by

$$z(\theta) \quad = \quad \frac{S(\theta) - \mu_S}{\sigma_S}. \tag{7}$$

Figure 5 (a) shows the normalized feature intensity for Fig. 4. The peaks appear at $30 + 90n$ degrees. Angles that coincide by rotating $90n$ degrees are defined as one angle group. For example, if $(30 + 90n)$-degrees angles appear, they are called a 30-degree group.

7) Rotation-angle estimation: The rotation angle is estimated from the peak of $z(\theta)$. The angles with peaks below the threshold Θ are detected. When the rotation angle is θ, four angles $\theta+90n, n = -2, -1, 0, 1$ of the θ-degree group are detected. Angles of $-180, -90, 0, 90$ degrees may be detected regardless of the rotation angle. Thus, four or eight peaks will be detected. Figure 5 (b) shows $z(\theta)$ when the image is rotated 25 degrees. In this case, eight peaks of $25 + 90n$ and $0 + 90n, n = -2, -1, 0, 1$ degrees appear. When eight peaks are detected, a non-zero-degree group angle is chosen as the estimated rotation angle. Therefore, 25 degrees is chosen as the estimated rotation angle. In summary, the proposed method can determine the estimated rotation angle from the amplitude spectrum, excluding the $90n$-degree differences. At most four extraction processes are required to extract the correct watermark. One property of DFT is that the DFT coefficients and their point symmetries are complex conjugate. Hence, the conjugate-amplitude spectrum satisfies

$$|F(k_1, k_2)| \quad = \quad |F(-k_1, -k_2)|. \tag{8}$$

Thus, if the image is rotated 180 degrees, the rotated amplitude spectrum $|F(-k_1, -k_2)|$ is equal to the original spectrum before rotation, $|F(k_1, k_2)|$. If watermarks are embedded in the amplitude-spectrum, they must also be embedded in the conjugate amplitude-spectrum, so that the same watermarks can be extracted even if the stego image is rotated 180 degrees. Therefore, if the watermarks are extracted after the rotation angle is estimated in advance using this method, the number of extraction processes can be reduced to a maximum of two, since the 180-degree rotation need not be taken into account.

III. COMPUTER SIMULATION

We evaluated the proposed method in terms of the percentage of successful detections of the rotation angle.

A. experimental conditions

The images used for evaluation were six IHC (Information Hiding and its Criteria for evaluation) standard images (4608×3456 pixels) [14]. The pilot signal is embedded in the U component of the image. A rotation attack on the stego-image is conducted with the rotation-angle range $\theta \in [-90, 90]$. After scaling each stego-image, four HDTV-sized (1920×1080 pixels) images are cropped out, centered at the coordinates $(x \pm 700, y \pm 500)$. The watermark is then estimated from the total $24 = 6 \times 4$ cropped images.

The length of one side of the block is set to $L_r = 300$ pixels, and the U-component image is divided into $N_s = 18$ blocks. The threshold for detecting peaks from the normalized feature intensity $z(\theta)$ is set to $\Theta = -4$.

B. Percentage of successful detections of rotation angle

Although this method can determine the rotation-angle group, it cannot uniquely determine the estimated rotation angle. Therefore, if the rotation-angle group is correct, we assume that the detection was successful. Figure 6 shows the successful-detection rate of rotation angles. The horizontal and vertical axes represent the rotation angle of θ and success rate, respectively. For rotation angles other than 0 degrees, the detection was 100% successful. We examined the 0-degree case with several detection failures. Figure 5 (c) shows the intensity $z(\theta)$ for the detection failures. The red line represents the threshold $\Theta = -4$. Less than four peaks below the threshold were detected. Since four peaks must be detected when detecting the rotation angle, detection failed.

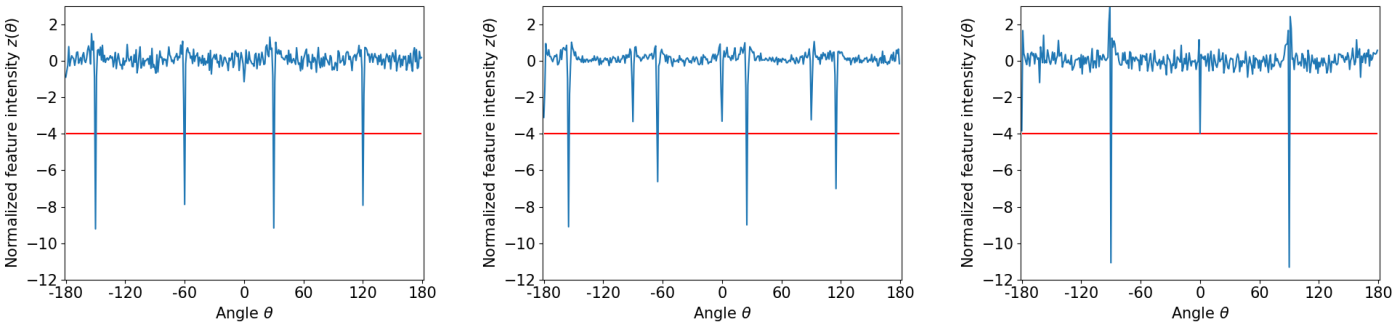
C. Validity of averaging process

If the original image contains slant, the binary embedded information may include the slant derived from the original image, resulting in false detection of the slant. Figure 7 shows the stego-image rotated -90 degrees. If the angle is estimated without averaging ($N_s = 1$), only the solid red block is used to estimate the rotation angle. This block obviously contains the original image-derived slant. Figure 8 (a) show the intensity $z(\theta)$ obtained without averaging ($N_s = 1$). The peaks appear at $\theta = -180, -167, -90, 0, 90$ degrees. Here, the peak at $\theta = -167$ degrees is not a pilot signal but a slant derived from the original image. If the angle is estimated with the average of five blocks ($N_s = 5$), the rotation angle is estimated from the average of the five blocks consisting of the red square and the four yellow squares in Fig. 7. Figure 8 (b) shows the intensity $z(\theta)$ averaged over $N_s = 5$ blocks. The averaging process reduces the original image-derived slant and emphasizes the pilot signal peaks, thus detecting very clear peaks. Figure 9 shows the detection rate when the number of blocks is $N_s = 1, 5$. The average detection rate for $N_s = 1$ was about 89% and for $N_s = 5$ about 99%. We found that the detection rate was not good without averaging, but averaging over at least five blocks resulted in almost 100%.

IV. CONCLUSIONS

To extract a watermark from a rotated stego-image, it is necessary to synchronize the position of the watermark with respect to the geometric transformation. Therefore, watermarks have been embedded using invariants such as SIFT feature points. With conventional watermarking methods using SIFT feature points [11], the watermark is embedded on the basis of the SIFT feature points, and the correct embedding position can be detected. However, even if the position can be estimated, a slanted embedding region is detected. Hence, the rotation angle must be searched by brute force to extract the watermark.

With the proposed method, the pilot signal is used as the synchronization. As in our previous study [12], we introduced



(a) $\theta = 30$ degrees (b) $\theta = 25$ degrees (c) $\theta = 0$ degrees

Fig. 5. Normalized feature intensity $z(\theta)$ at rotation angle of θ

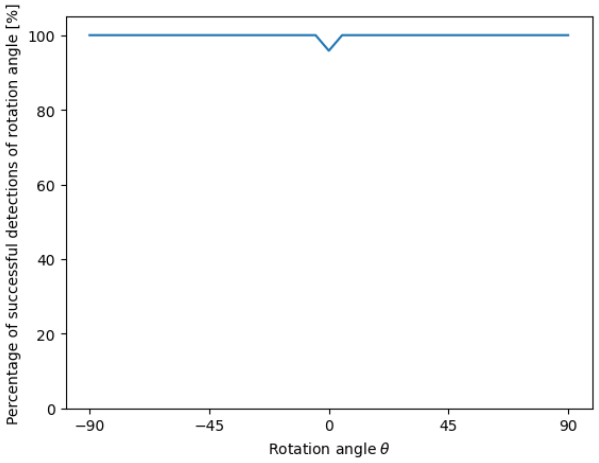


Fig. 6. Successful-detection rate of rotation angle

a grid-shaped pilot signal. The rotation angle is estimated by detecting the pilot signal in a rotated stego-image. From computer simulations, we verified that the rotation angle can be accurately detected. If the rotation angle can be estimated, the embedding strength of the watermark will be reduced, which is expected to improve image quality. For future work, the embedding strength will be optimized and the bit error rate of the watermark will be evaluated. As future work, we will examine methods for estimating combination attacks. Also, the embedding strength will be optimized and the bit error rate of the watermark will be evaluated.

ACKNOWLEDGMENTS

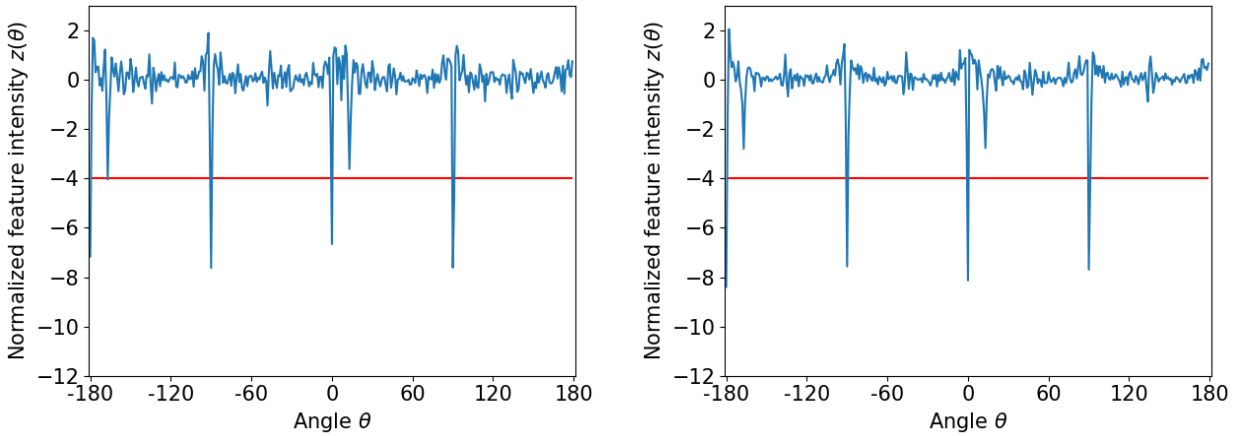
This work was supported by JSPS KAKENHI Grant Number JP20K11973, Yamaguchi University Fund and the Cooperative Research Project Program, R04/B09 Research on Multifunctional Multimedia Production, of the RIEC, Tohoku University. Some of the computations were performed using the supercomputer facilities at the Research Institute for Information Technology, Kyushu University.

REFERENCES

[1] H. Luo, X. Sun, H. Yang, Z. Xia, “A Robust Image Watermarking Based on Image Restoration Using SIFT,” Radioengineering, vol.20, no.2, pp. 525-532, 2011
[2] L. Li, B. Guo, J. Pan, ”Feature-Based Image Watermarking Resisting Geometric Attacks, “ 2008 3rd International Conference on Innovative Computing Information and Control, Dalian, Liaoning, p. 18, 2008
[3] M. Beguma, J. Ferdushb, M. S. Uddinc, “A Hybrid robust watermarking system based on discrete cosine transform, discrete wavelet transform, and singular value decomposition,” Journal of King Saud University - Science, vol. 34, Issue 8, part B, pp. 5856-5867, 2022
[4] H. Kang, K. Iwamura, “Watermarking based on the difference of discrete cosine transform coefficients and an error-correcting code,” Proc. of IWIHC, pp. 9-17, 2014
[5] I. J. Cox, J. Kilian, F. T. Leighton and T. Shamoon, “ Secure spread spectrum watermarking for multimedia,” in IEEE Transactions on Image Processing, vol. 6, no. 12, pp. 1673-1687,1997
[6] J. J.K.O. Ruanaidh, T. Pun, “ Rotation, scale and translation invariant spread spectrum digital image watermarking,” Signal Process., vol. 66, Issue 3, pp. 303-317, 1998
[7] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints, ” International J. Compute Vision, vol. 60, no. 2, pp. 91-110, 2004
[8] M. Hayashi, M. Kawamura, “Improved SIFT feature-based watermarking method for IHC ver. 5,” Proc. of APSIPA ASC 2018, 2018
[9] V. Solachidis, I. Pitas, “ Circularly Symmetric Watermark Embedding in 2-D DFT Domain,” IEEE Trans Image Process, vol. 10, no.11, pp.1741-1752, 2001
[10] J. S. Seo, C. D. Yoo, “Localized image watermarking based on feature points of scale-space representation,” Pattern Recognition, vol. 37, Issue 7, pp. 1365-1375, 2004
[11] P. C. Su, Y. C. Chang and C. Y. Wu, “ Geometrically Resilient Digital Image Watermarking by Using Interest Point Extraction and Extended Pilot Signals,” IEEE Trans. Inf. Forensics Secur., vol. 8, no. 12, pp. 1897-1908, 2013,
[12] R. Kawano and M. Kawamura, “ Detection of periodic pilot signal in image watermarking,” 2021 APSIPA ASC, SS-IVM-1.6, pp.1647-1652, 2021
[13] B. Chen, G. W. Wornell, “ Quantization index modulation: A class of provably good methods for digital watermarking and information embedding,” IEEE Trans. Inf. Theory, vol.47, no.4, pp.1423-1443, 2001
[14] Information Hiding and its Criteria for evaluation, https://www.ieice.org/iss/emm/ihc/en/image/image.php (20th February, 2023 access)



Fig. 7. Stego-image at -90-degree rotation. The red and yellow squares are the blocks used for angle estimation, respectively.



(a) $N_s = 1$ (b) $N_s = 5$

Fig. 8. Normalized feature intensity $z(\theta)$ at rotation angle of θ in the case of N_s divisions

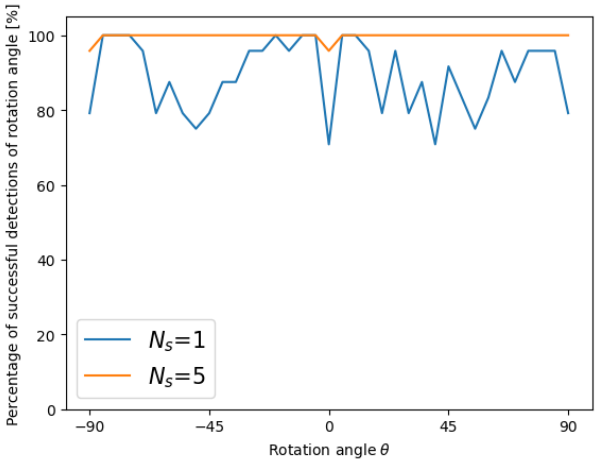


Fig. 9. Successful-detection rate of rotation angle in the case of N_s divisions

Application for generating re-accessible screenshots of web pages using histogram shrinkage

Ayaka Sakamoto*, Rinka Kawano† and Masaki Kawamura†
* Yamaguchi University, Yamaguchi, Japan
E-mail: d023vbu@yamaguchi-u.ac.jp
† Yamaguchi University, Yamaguchi, Japan
E-mail: {m.kawamura, rinka1005}@m.ieice.org

***Abstract*—We developed an application that can generate a screenshot image of a website and redisplay the website with the URL extracted from the image. Since URLs contain multibyte characters and have variable length, they are converted using Base64 and QR code. The QR code allows error correction. The converted bit sequence is then embedded as a watermark. Unlike natural images, screenshot images are artificial. In many screenshot images, luminance values tend to be distributed around 0 or 255. When embedding, many watermarks are dropped due to overflow. Therefore, we also propose a method of shrinking the luminance distribution for screenshot images. Compared with the case in which the luminance distribution was left unchanged, the proposed method had a smaller bit error rate and was able to redisplay websites with the URLs extracted from the screenshot images.**

I. Introduction

Smartphones now have a screenshot function that allows users to save information displayed on the screen as an image. The screenshot is very useful since a user can always refer back the information viewed in the browser. However, there are often cases in which the information one wants to see is outside the scope of the screenshot image. In this case, the user needs to search the website again. To mitigate this inconvenience, we developed an application that enables re-access to a website from a screenshot image and applies watermarking.

Watermarking is a technique of secretly embedding in or extracting information from content, such as still images, video, and audio [1]. For example, watermarking is used to detect and track unauthorized use of content and prevent tampering. Information embedded into an image is called a watermark, and the embedded image is called a stego-image. Typical watermarking methods include additive embedding [2] and quantization index modulation (QIM) [3] to embed the watermark into the content. Since screenshot images may be shared or transferred between devices, QIM is used as a robust embedding method [3]. Processing to the content, such as compression and cropping, causes errors in the watermark. Therefore, error-correcting codes [4] and spread-spectrum technique [5] are used to correct these errors.

We developed an application that can generate a screenshot image of a website and redisplay the website with the URL extracted from the image The images used in our application are screenshots of websites displayed in a browser; thus, smartphone-application screenshots are out of the scope of this study. Screenshot images are artificial and most of their luminance values tend to be close to 0 or 255. Hence, unlike natural images, conventional watermarking methods often fail to embed watermarks in screenshot images correctly. Therefore, we propose a method for shrinking the luminance distribution for screenshot images.

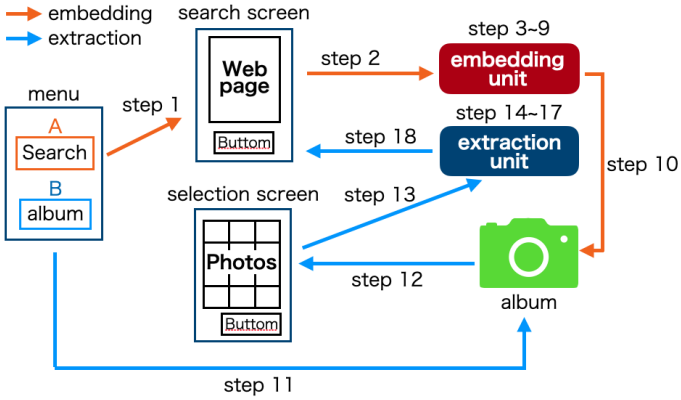


Fig. 1: Overview of application

This paper is organized as follows. In Section II, we present our developed application and proposed method. In Section III, we discuss the evaluation of the proposed application on extracting URLs. We conclude the paper in Section IV.

II. Developed application and proposed method

We describe the overall concept of our application and method suitable for screenshots. Since our study is aimed at developing applications for the iPhone, the following sample codes are written in Swift.

A. Process flow of application

The main functions of the application are to save a screenshot image with a URL and display the website with the URL extracted from the stego-image in a browser. Figure 1 shows an overview of the application, which includes embedding and extraction procedures.

1) Embedding Procedure: Steps 1–10 (steps 1, 2 and 10 in Fig. 1 and steps 3–9 in Fig. 2) are those involved in the embedding procedure.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

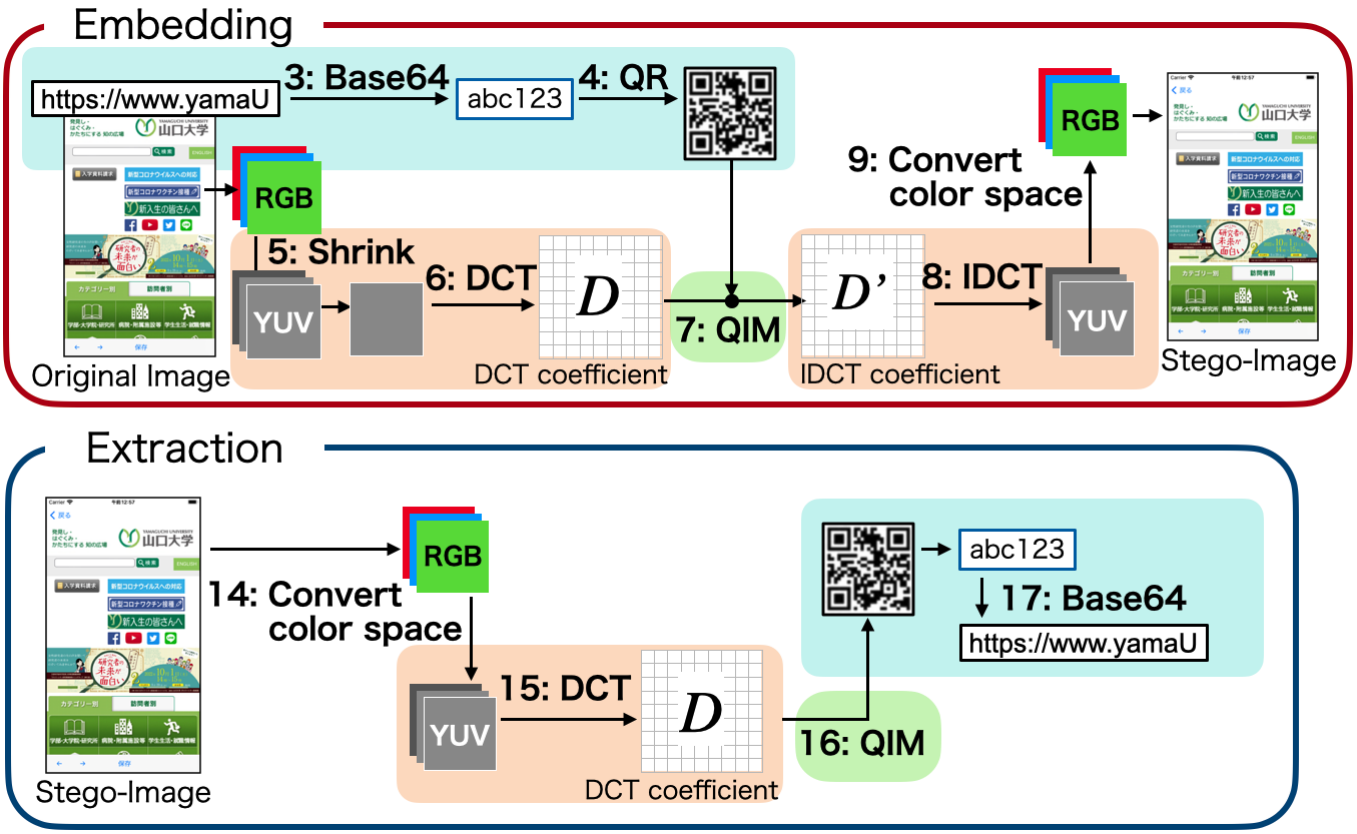


Fig. 2: Processing in the embedding unit and extraction unit. The light blue area shows the conversion process between URL and QR Code, the orange area the DCT process, and the green area the QIM process.

step 1: When the user presses the ”search” button on the menu, the screen changes to the search screen, as shown in Fig. 3 (b).

step 2: When the user presses the ”screenshot” button at the bottom of the screen shown in Fig. 3 (b), both the URL as a watermark and screenshot image as an original image are sent to the embedding unit.

step 3: In the embedding unit, the URL is encoded using Base64 since multibyte characters, such as Japanese, are often included in the URL. Hereafter, an encoded string is referred to as a Base64-encoded URL. The number of letters in a Base64-encoded URL should be a constant length.

step 4: The Base64-encoded URL is converted into a QR code. The bit sequence binarized from the black-and-white pixels of the generated QR code is called a watermark.

step 5: The original image is converted from the RGBA color space to YUV color space. The distribution of luminance values is shrunk.

step 6: Luminance values are converted to the frequency domain by the two-dimensional (2D) discrete cosine transform (DCT).

step 7: The watermark is embedded in the DCT coefficients by QIM.

step 8: The Y component of the stego-image is generated by the inverse DCT.

step 9: The generated Y component is combined with the U and V components of the original image and converted to the RGBA color space to generate a stego-image.

step 10: The generated stego-image is saved to the photo folder in the device.

2) Extraction Procedure: Steps 11–18 (step 11-13 and 18 in Fig. 1 and steps 14–18 in Fig. 2) are involved in the extraction procedure.

step 11: When the ”album” button on the menu screen shown in Fig. 3 (a) is pressed, the album of the device is displayed, as shown in (c).

step 12: When a user selects an image from the album, the selected image is displayed, as shown in (d).

step 13: When the ”Extract URL” button in the lower right corner of the screen is pressed, the selected stego-image is sent to the extraction unit.

step 14: A stego-image is converted from the RGBA color space to the YUV color space.

step 15: The Y component of the stego-image is transformed to the frequency domain by 2D DCT.

step 16: Watermark, which is a binarized QR code, is extracted from the DCT coefficients by using QIM.

step 17: The extracted watermark is converted to a black-and-white image and decoded into a string by using the



(a) menu (b) search screen

(c) album (d) selection screen

Fig. 3: Application screen on simulator

QR-code-reader function. Since this string is a Base64-encoded URL, it is further decoded using Base64.

step 18: The decoded string is converted to a URL by removing the nil character at the end of the decoded string. The website with the URL is then displayed.

B. QR Code and Base64

When embedding a watermark, the watermark should be of fixed length to simplify the algorithm. A fixed length is often assumed with conventional watermarking methods. Watermarks must also be robust against errors. Our application converts a variable-length URL into a QR code, making the watermark fixed-length and providing error-correction capability.

A fixed-length QR code can be generated if the version can be specified. However, the QR-code-generation method `QRCodeGenerater` [6] in the Swift standard library cannot specify the version. Therefore, the length of the QR code varies depending on the length of the URL. In addition, alphanumeric characters, special symbols, and multibyte characters in URLs are assigned different bit lengths in the QR code, so the length of the QR code may differ even for the same number of letters. Therefore, it is necessary to encode multibyte characters into alphanumeric characters only. A method `base64EncodedString` [7] in the Swift standard library was used for encoding. It is a method that returns a Base64-encoded string when binary data are given. The sample code for `base64EncodedString` is shown in Fig. 4. Since the URL is string type, it is converted to data type. As the argument `data` is given to `base64EncodedString`, the encoded string is returned. As a result, the URL can be converted to an alphanumeric-only Base64-encoded URL.

The terminating character in Base64 is ”=” and can be decoded even if there are two or more ”=”s at the end. In other words, it is possible to make the length of a string constant by adding an ”=” at the end. A sample code for adjusting the length of a string is shown in Fig. 5. The argument `length` is the desired number of letters. The variable `url_ad` is assigned the string `url`, and ”=” is added until its length becomes the target number of letters. Thus, an encoded fixed-length URL is generated.

Watermarks must be robust against errors. In our application, QR codes are used for error correction. To convert a Base64-encoded URL to a QR code, it is necessary to specify `"CIQRCodeGenerater"` [6] to `CIFilter` in the Swift standard library. This method generates a QR code when a string and error-correction level are given as arguments. A sample code of `QRCodeGenerater` is shown in Fig. 6. The encoded URL and error-correction level ”H” (high) are given. Accordingly, the generated QR code is returned. Since the returned QR code is an image, the pixel values are scanned and converted to binary values. That is, white pixels are converted to 1 and black pixels are converted to 0. This binarized code is used as a watermark.

C. Histogram Shrinkage

In our application, the original image in which the URL is embedded is a screenshot image of a website. Figure 7 shows histograms of luminance values, where (a) and (b) represent histograms of luminance values of a screenshot image and natural image, respectively. This indicates that, unlike natural images, screenshot images have a sparse luminance distribution. The screenshot image is an artificial image, and many of the background colors of websites take luminance values close to 0 or 255. Therefore, embedding the watermark causes an overflow of luminance values, and the watermark is lost due to the rounding process. In other words, embedding watermarks in screenshot images often fails. Therefore, our proposed method shrinks the luminance distribution.

The luminance components are separated from the original image. Let $Y(i,j)$ be the luminance value at coordinate (i,j). If there is a pixel of luminance value $Y(i,j) < 14$ or $Y(i,j) > 241$, the distribution is shrunk. Let $Y'(i,j)$ be the luminance

```
//Encode URLs in Base64
func Enc64(url: String)->String{
    let data = url.data(using: .utf8)
    let EncStr: String! = data?.base64EncodedString()

    return (EncStr)
}
```

Fig. 4: Sample code of base64EncodedString

```
//Adjust the number of characters
func adjust(length: Int, url: String)->String{
    var url_ad: String = url
    //Add "=" to adjust the number of characters to the variable length
    for _ in 0...length-url.count-1{
        url_ad += "="
    }

    return (url_ad)
}
```

Fig. 5: Sample code for adjusting length of string

```
//Generate QR Code
func generateQR(url: String)->(uiQR:UIImage, cgQR:CGImage){
    let url = url

    //Convert url from String type to Data type
    let data = url.data(using: .utf8)!
    //inputCorrectionLevel specifies the level of error correction
    let qr = CIFilter(name: "CIQRCodeGenerator", parameters: ["inputMessage":data,
        "inputCorrectionLevel": "H"])!
    //put a white border
    let sizeTransform = CGAffineTransform(scaleX: 1, y: 1)

    //Convert QR code to CIImage type
    let ciQR = qr.outputImage!.transformed(by: sizeTransform)

    //CI to (UI & CG)
    let uiQR = UIImage(ciImage: ciQR)
    let OpCG = context.createCGImage(ciQR, from: ciQR.extent)
    let cgQR = OpCG!

    return (uiQR,cgQR)
}
```

Fig. 6: Generation of QR code

value after shrinkage, which is given by

$$Y'(i,j) \quad = \quad rY(i,j) + (1-r)m, \tag{1}$$

where $m = 255/2 = 127.5$, and $r \leq 1.0$ is the magnification rate.

The embedding domain is the frequency domain of the DCT. Thus, the 2D DCT is used on $Y'(i,j)$. When using the DCT in Swift, the DCT class of the vDSP framework is used. However, this DCT class implements only the 1D DCT. Therefore, we implemented the 2D DCT by using the 1D DCT twice.

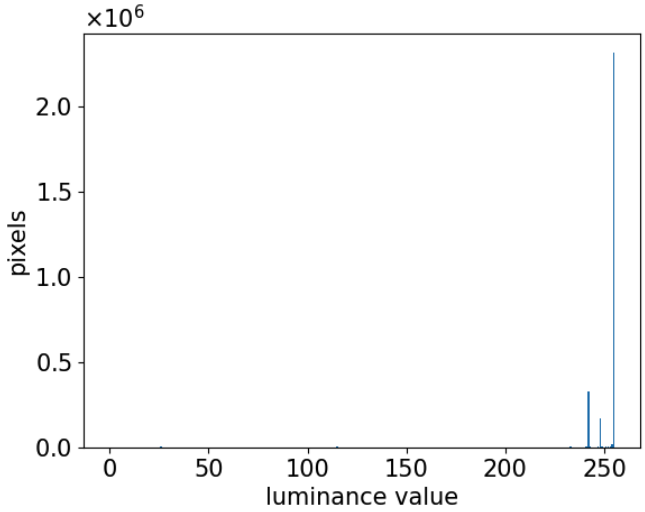
D. Embedding and Extraction with QIM

There are two types of watermark embedding and extraction methods: non-blind that requires the original image for extraction and blind that does not require the original image. Our application uses a blind method, i.e., QIM [3]. First, we explain the embedding process by QIM. Let a K-bit watermark be $\boldsymbol{m} = (m_1, m_2, ..., m_K)^\top, m_i \in \{0,1\}$. The watermark m_i is embedded into the DCT coefficients D_i by using QIM. The embedded DCT coefficient D'_i is given by

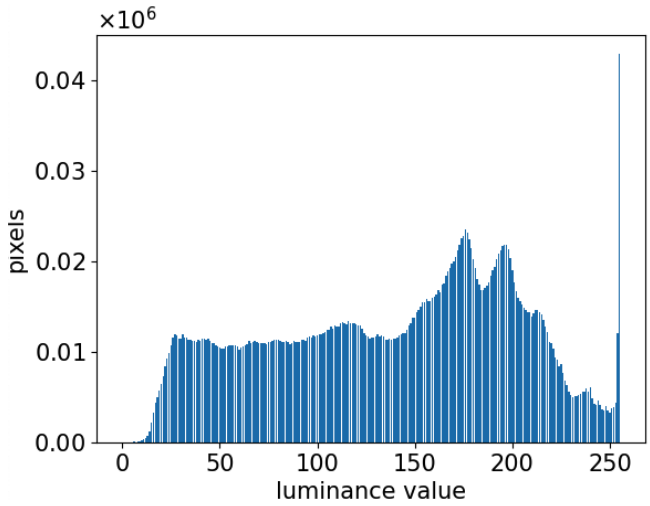
$$D'_i \quad = \quad 2\Delta \left(\left\lfloor \frac{D_i}{2\Delta} - \frac{m_i}{2} + 0.5 \right\rfloor + \frac{m_i}{2} \right), \tag{2}$$

where Δ is a quantization step size [3] and $\lfloor x \rfloor$ represents the floor function.

We next explain the extraction process by QIM. The watermark $\hat{m}_i$ is extracted from D'_i. The extracted watermark



(a) Histograms for screenshot image (No. 2 in Table I)



(b) Histograms for natural image

Fig. 7: Histograms of luminance values

$\hat{\boldsymbol{m}} = (\hat{m}_1, \hat{m}_2, ..., \hat{m}_K)^\top$ is given by

$$\hat{m}_i \quad = \quad \left\lfloor \frac{|\hat{D}_i|}{\Delta} + 0.5 \right\rfloor \bmod 2. \tag{3}$$

III. Evaluation

We describe the evaluation method of this application and present the results.

A. Bit Error Rate

We evaluated whether URLs were correctly extracted from the screenshot images. The evaluation was based on the bit error rate (BER) for the watermark. Note that a watermark is a binary bit sequence of a QR code containing a Base64-encoded URL. The BER is defined as

$$\text{BER} \quad = \quad \frac{1}{K} \sum_{i=1}^{K} m_i \oplus \hat{m}_i, \tag{4}$$

where $\oplus$ is an exclusive OR.

B. Simulation Conditions

We ran our application on the iPhone SE (3rd generation) simulator on Xcode. The screen size of the iPhone SE is 750×1334 pixels. The method that leaves the luminance distribution unchanged is called an original embedding, in contrast to the proposed method. Screenshots were taken at the 15 different websites shown in Table I. For the screenshot image of each site, the same URL was embedded with both the proposed method and the original embedding. The error-correction level of the QR code is set to "H". The desired number of letters in the Base64-encoded URL was set to 1000. The magnification rate r of the luminance distribution was set to $r = 0.95$. The step size of QIM was set to $\Delta = 5000$.

C. Evaluation Results

The results of the computer simulation are listed in Table I, where the URLs were the addresses of 15 different websites, including the number of letters of the URLs. The table shows whether the websites could be displayed again by using the extracted URLs with the proposed method and the original embedding, and the BERs of the watermark. Figure 8 shows the BERs. In the original embedding, only Nos. 4, 6, and 10 could be redisplayed. For all other websites, the QR codes were not decoded. A zero-length string representing decoding failure was returned. Figure 8 shows that the BERs with the original embedding were larger than 0.14, except for website Nos. 4, 6, and 10. This is considered due to overflow of luminance values and loss of the watermark.

As shown in Table I, all websites could be redisplayed by using the proposed method. The table also shows that BER< 0.015 for all websites. To confirm that no overflow occurred, the shrunken luminance distribution is shown in Fig. 9. Compared with Fig. 7 (a), the luminance values were distributed inward, mostly within the range $[0 : 255]$. In other words, the shrinkage of the luminance distribution was effective for embedding into the screenshot images.

IV. Summary

We developed an application that can embed the URL of a website into a screenshot image, then redisplay the website from the image. Since many background colors have luminance values close to 0 or 255, embedding a watermark in a screenshot image will cause an overflow of such values. To solve this problem, we also proposed a method of shrinking the luminance distribution.

From computer simulations involving our application, we confirmed that it can be operated from the user's viewpoint without crashes or other problems. As a result of URL extraction using the proposed method and the original embedding,

TABLE I: Results of application validation

No.	URL	number of letters	Original Embedding		Proposed Method	
			reaccess	BER	reaccess	BER
1	https://m.yahoo.co.jp/	22	Failure	0.2252	Success	0.0132
2	https://www.google.co.jp/	25	Failure	0.4293	Success	0.0134
3	https://www.applicationle.com/jp/	25	Failure	0.2141	Success	0.0133
4	https://www.yamaguchi-u.ac.jp/	30	Success	0.0283	Success	0.0135
5	https://www.instagram.com/yucoop_favo/	38	Failure	0.3057	Success	0.0137
6	https://www.is.sci.yamaguchi-u.ac.jp/Japanese/	46	Success	0.0142	Success	0.0134
7	https://m.facebook.com/YamaguchiUniversityOfficial/	51	Failure	0.2898	Success	0.0138
8	https://m.youtube.com/channel/UCJoJPkjqd8KS3FRczCLq1Vg	54	Failure	0.2297	Success	0.0135
9	https://www.yamaguchi-u.ac.jp/faculty/science/index.html	56	Failure	0.2226	Success	0.0134
10	https://www.nintendo.co.jp/zelda/totk/special/index.html	56	Success	0.0134	Success	0.0134
11	https://ja.m.wikipedia.org/wiki/%E5%B1%B1%E5%8F%A3%E5%A4%A7%E5%AD%A6	68	Failure	0.2703	Success	0.0135
12	https://www.asahi.com/sp/articles/ASR5D4RW8R5DUTIL018.html?iref=sp_new_news_list_n	82	Failure	0.2233	Success	0.0145
13	https://twitter.com/YamaguchiU_koho?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor	90	Failure	0.3332	Success	0.0132
14	https://www.google.co.jp/maps/place/%E3%80%92753-0841+%E5%B1%B1%E5%8F%A3%E7%9C%8C%E5%B1%B1%E5%8F%A3%E5%B8%82%E5%90%89%E7%94%B0%EF%BC%91%EF%BC%96%EF%BC%97%EF%BC%97%E2%88%92%EF%BC%91+%E5%B1%B1%E5%8F%A3%E5%A4%A7%E5%AD351 letters%A6/@34.1503459,131.465216,12z/data=!4m6!3m5!1s0x354490cee731f3f1:0x4794e014e0c1e7da!8m2!3d34.1486699!4d131.46792!16s%2Fm%2F03gzkc0?hl=ja	351	Failure	0.1634	Success	0.0137
15	https://www.google.co.jp/maps/place/%E3%80%92753-0302+%E5%B1%B1%E5%8F%A3%E7%9C%8C%E5%B1%B1%E5%8F%A3%E5%B8%82%E4%BB%81%E4%BF%9D%E4%B8%AD%E9%83%B7%EF%BC%91%EF%BC%92%EF%BC%93+KDDI%E3%83%91%E3%83%A9%E3%83%9C%E3%83%A9%E9%A4%A8/@34.2254738,131.5570624,11z/data=!4m9!1m2!2m1!1z6Zu75rOi5pyb6YGg6Y-h!3m5!1s0x3544be4eadec6305:0x6ba91f5f78ceb437!8m2!3d34.2150789!4d131.555713!16s%2Fg%2F1ym82tm4x?hl=ja	390	Failure	0.1498	Success	0.0137

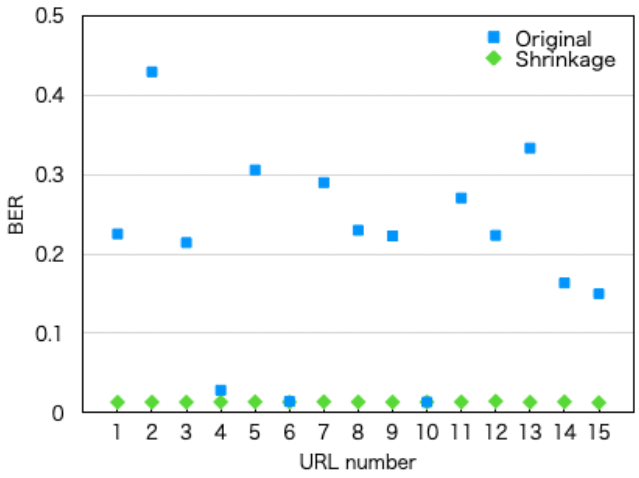


Fig. 8: BERs of watermarks from extracted QR code. Blue and green points represent BERs of original embedding and proposed method, respectively.

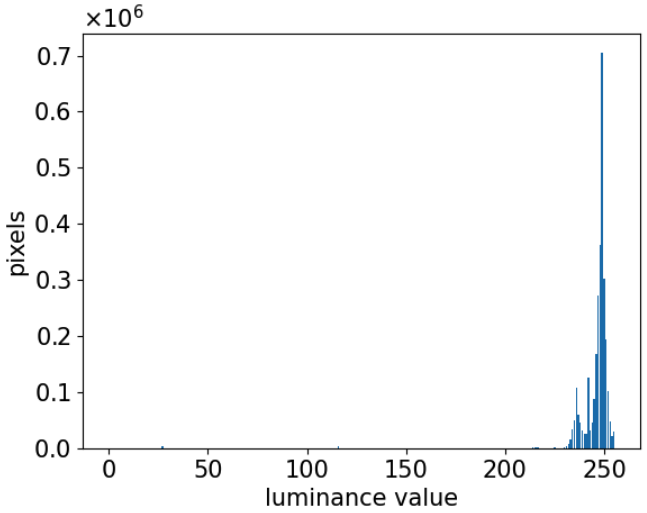


Fig. 9: Histogram of luminance values after embedding into screenshot image with proposed method. (No. 2 in Table I)

the original embedding had large BERs and failed to redisplay in 12 out of 15 cases. The proposed method had small BERs and was able to redisplay all websites. Therefore, shrinking the luminance distribution is effective for embedding a watermark in a screenshot image.

The current proposed application does not re-access dynamic pages or web pages that use cookies for session management. It also does not support screenshots in other applications that cannot capture URLs. This application only provides the function to retrieve a URL.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP20K11973.

REFERENCES

[1] F. A. Petitcolas, R. J. Anderson, M. G. Kuhn, " Information hiding-a survey, " Proceedings of the IEEE, vol. 87, no. 7, pp.1062–1078, 1999

[2] A. Nikolaidis, I. Pitas, "Asymptotically Optimal Detection for Additive Watermarking in the DCT and DWT Domains," IEEE Transactions on Image Processing, vol. 12, no. 5, pp.563– 571, 2003

[3] B. Chen, G. W.Wornell, "Quantization index modulation : A class of provably good methods for digital watermarking and information embedding," IEEE Transactions on Information Theory, vol. 47, no. 4, pp. 1423-1443, 2001.

[4] N. Hirata, T. Nozaki, M. Kawamura, "Image Watermarking Method Satisfying IHC by Using PEG LDPC Code," IEICE Transactions on Information and Systems, vol. E100- D, no. 1, pp. 13-23, 2017

[5] I.J. Cox, J. Kilian, F.T. Leighton, T. Shamoon, "Secure spread spectrum watermarking for multimedia, " IEEE Transactions on Image Processing, vol. 6, no. 12, pp.1673– 1687, 1997

[6] QRCodeGenerator, https://developer.applicationle.com/documentation/coreimage/cifilter/3228262-qrcodegenerator, (2023/2/6 access)

[7] base64EncodedString(options:), https://developer.apple.com/documentation/foundation/data/2142853-base64encodedstring, (2023/5/31 access)

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Domain Adaptation for Efficiently Fine-tuning Vision Transformer with Encrypted Images

Teru Nagamori*, Sayaka Shiota* and Hitoshi Kiya*
* Tokyo Metropolitan University, Japan
E-mail: nagamori-teru@ed.tmu.ac.jp, sayaka@tmu.ac.jp, kiya@tmu.ac.jp

***Abstract*—In recent years, deep neural networks (DNNs) trained with transformed data have been applied to various applications such as privacy-preserving learning, access control, and adversarial defenses. However, the use of transformed data decreases the performance of models. Accordingly, in this paper, we propose a novel method for fine-tuning models with transformed images under the use of the vision transformer (ViT). The proposed domain adaptation method does not cause the accuracy degradation of models, and it is carried out on the basis of the embedding structure of ViT. In experiments, we confirmed that the proposed method prevents accuracy degradation even when using encrypted images with the CIFAR-10 and CIFAR-100 datasets.**

I. Introduction

Deep neural networks (DNNs) have been deployed in many applications including security-critical ones such as biometric authentication and medical image analysis. In addition, training a deep learning model requires a huge amount of data and fast computing resources, so cloud environments are increasingly used in various applications of DNN models. However, since cloud providers are not always reliable in general, privacy-preserving deep learning has become an urgent problem [1], [2].

One of the privacy-preserving solutions for DNNs is to use encrypted images to protect visual information on images for training models and testing. In this approach, images are transformed by using a learnable encryption method, and encrypted images are used as training and testing data. The use of models trained with encrypted images enables us to use state-of-the-art learning algorithms without any modification. However, all conventional learnable encryption methods [3]–[8] have the problem that is the performance degradation of encrypted models compared to models without encryption.

Traditional cryptographic methods such as homomorphic encryption [9], [10] are one of the other privacy-preserving approaches, but the computation and memory costs are expensive, and it is not easy to apply these methods to state-of-the-art DNNs directly. Federated learning [11] allows users to train a global model without centralizing the training data on one machine, but it cannot protect privacy during inference for test data when a model is deployed in an untrusted cloud server.

Accordingly, we focus on the problem of degraded model performance when models are trained with encrypted images. In this paper, we consider reducing the performance degradation of encrypted models when using encrypted images in the vision transformer (ViT) [12] which has high performance in image classification, and we propose a domain adaptation method to reduce the influence of encryption. In experiments, the proposed method is demonstrated to maintain almost the same performance as models trained with plain images in terms of image classification accuracy even when using encrypted images.

II. Related work

A. Image Encryption for Deep Learning

Various image transformation methods with a secret key, often referred to as perceptual image encryption or image cryptography, have been studied so far for many applications [1]. In this paper, we focus on learnable images transformed with a secret key, which have been studied for deep learning. Learnable encryption enables us to directly apply encrypted data to a model as training and testing data. Encrypted images have no visual information on plain images in general, so privacy-preserving learning can be carried out by using visually protected images. In addition, the use of a secret key allows us to embed unique features controlled with the key into images. Adversarial defenses [13] and access control [14] are carried out with encrypted data using unique features.

Tanaka first introduced a block-wise learnable image encryption method (LE) with an adaptation layer [3], which is used prior to a classifier to reduce the influence of image encryption. Another encryption method is a pixel-wise encryption (PE) method in which negative-positive transformation (NP) and color component shuffling are applied without using any adaptation layer [4]. However, both encryption methods are not robust enough against ciphertext-only attacks, as reported in [15], [16]. To enhance the security of encryption, LE was extended to ELE by adding a block scrambling step and a pixel encryption operation with multiple keys [5]. However, ELE still has an inferior accuracy compared with using plain images, although an additional adaptation network (denoted as ELE-AdaptNet hereinafter) is applied to reduce the influence of the encryption. Moreover, images large in size cannot be applied to ELE because of the high computation cost of ELE-AdaptNet.

Recently, some methods have been proposed to use these block-wise encryptions combined with isotropic networks such as ViT to prevent performance degradation [6], [7], but this problem has not been completely solved. Another encryption method, a neural network-based image encryption method called NeuraCrypt [17], has been proposed, but it has problems

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

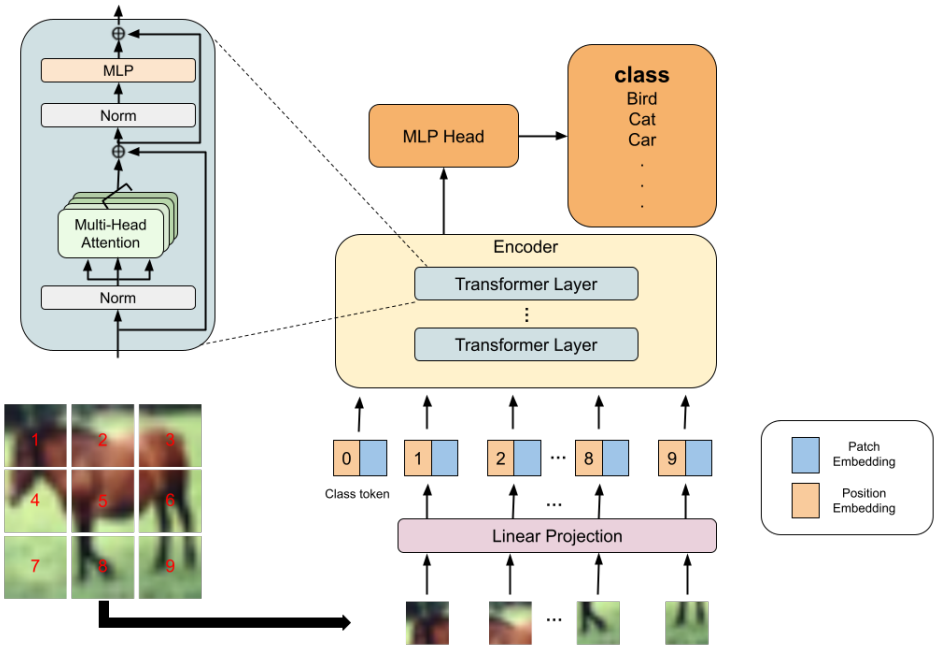


Fig. 1: Architecture of Vision Transformer

such as difficulty in using it for color images and difficulty with existing models because the input is not an image. To solve the problem, Color-NeuraCrypt [8], which is an extension of NeuraCrypt, was proposed, but this method also has the same performance degradation problem as conventional block-wise encryption methods.

Accordingly, we focus on the performance degradation problem when combining block-wise encryption and an isotropic network, and we propose a domain adaptation method to solve the problem.

B. Vision Transformer

The Vision Transformer (ViT) [12] is commonly used in image classification tasks and is known to provide a high classification performance. As shown in Fig. 1, in ViT, an input image is divided into patches, and each patch is transformed into a one-dimensional trainable vector. Afterward, they are input into the transformer encoder. The encoder outputs only the class token, which is a vector of condensed information about the entire image, and it is used for classification. Also, the architecture of ViT has two types of embedding: positional embedding, which stores the location information of cropped patches in an image, and patch embedding, which transforms each patch into a trainable vector.

Previous studies have indicated that when DNN models are trained with encrypted images, the performance of the models is degraded compared to models trained with plain images. In contrast, it has been pointed out that block-wise encryption has a high affinity with models that have embedding structures such as ViT [6], [18]. Accordingly, we focus on the affinity between block-wise image encryption and the embedding structure of ViT to reduce the influence of image encryption even when encrypted images are used for training models.

III. Proposed Method

A. Overview

Figure 2 shows the procedure for fine-tuning ViT models with encrypted images and the proposed domain adaptation. Since the pre-trained model as in [12] was prepared in advance, the area surrounded by a red box is discussed in this paper.

In ViT, an input image $x \in \mathbb{R}^{h \times w \times c}$ is divided into N patches with a size of $p \times p$, where h, w and c are the height, width, and number of channels of the image. Also, N is given as hw/p^2. Afterward, each patch is flattened into $x_p^i = [x_p^i(1), x_p^i(2), \ldots, x_p^i(L)]$. Finally, a sequence of embedded patches is given as

$$z_0 = [x_{class}; x_p^1\mathbf{E}; x_p^2\mathbf{E}; \ldots x_p^i\mathbf{E}; \ldots x_p^N\mathbf{E}] + \mathbf{E_{pos}}, \quad (1)$$

where

$$\mathbf{E_{pos}} = ((e_{pos}^0)^\top (e_{pos}^1)^\top \ldots (e_{pos}^i)^\top \ldots (e_{pos}^N)^\top)^\top,$$
$$L = p^2 c, x_{class} \in \mathbb{R}^D,\ x_p^i \in \mathbb{R}^L,\ e_{pos}^i \in \mathbb{R}^D,$$
$$\mathbf{E} \in \mathbb{R}^{L \times D},\ \mathbf{E_{pos}} \in \mathbb{R}^{(N+1) \times D}.$$

x_{class} is a class token, $\mathbf{E}$ is an embedding (patch embedding) to linearly map each patch to dimensions D, $\mathbf{E_{pos}}$ is an embedding (position embedding) that gives position information to patches in the image, e_{pos}^0 is position information of a class token, e_{pos}^i is position information of each patch, and z_0 is a sequence of embedded patches.

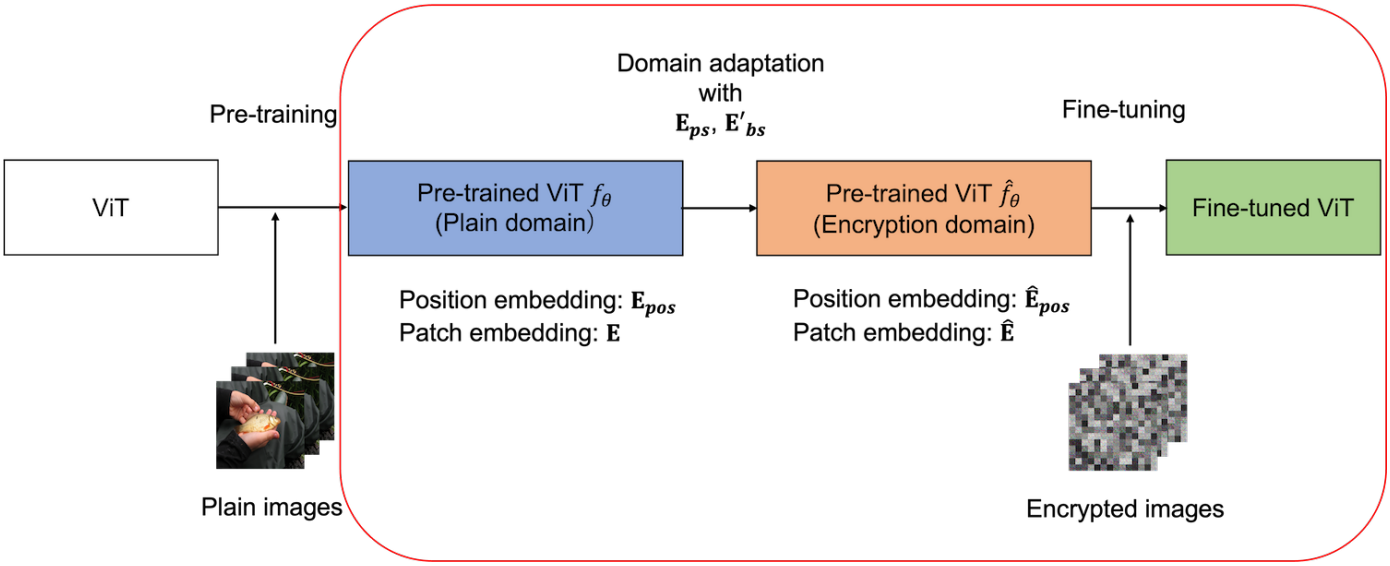


Fig. 2: Fine-tuning procedure using image encryption and domain adaptation

In this paper, we use block scrambling and pixel shuffling used in ELE for image encryption [5], where two keys are commonly applied to all images and blocks, respectively. In block scrambling, since an image is divided into non-overlapped blocks (patches) and the divided blocks are randomly permutated with a key, position embedding $\mathbf{E_{pos}}$ is affected by the position scrambling in general. In contrast, since pixel shuffling randomly permutates the position of pixels in each block, patch embedding $\mathbf{E}$ is affected by the pixel shuffling. As a result, models fine-tuned with encrypted images have a lower performance than models fine-tuned with plain images due to the influence of encryption. In other words, since pre-training a model is carried out by using plain images, the domain of fine-tuned models is different from that of pre-trained ones.

Accordingly, to reduce the influence of encryption, we propose a domain adaptation method, which is carried out on the basis of the embedding structure of ViT, for efficiently fine-tuning models.

B. Image Encryption

Encrypted images are used to fine-tune a ViT model f_θ as shown in Fig. 2. Figure 3 illustrates the encryption procedure used in this paper. This procedure can be expressed as follows.

(a) Divide an image $x \in \mathbb{R}^{h \times w \times c}$ into non-overlapped blocks with a size of $p \times p$ such that $B = \{B_1, \ldots, B_N\}$.

(b) Permutate blocks with secret key k_1 such as

$$\hat{B} = \mathbf{E_{bs}} B^\top, \ \hat{B} \in \mathbb{R}^N, \tag{2}$$

where $\mathbf{E_{bs}}$ is generated using the key k_1 as follows.

1) Using key k_1, generate a random integer vector with a length of N as

$$l_e = [l_e(1), l_e(2), \ldots, l_e(i) \ , \ldots, l_e(N)] \ , \tag{3}$$

where

$$l_e(i) \in \{1, 2, ..., N\} \, ,$$
$$l_e(i) \neq l_e(j) \text{ if } i \neq j$$
$$i, j \in \{1, \ldots, N\} \, .$$

2) Given $k_{(i,j)}$ as

$$k_{(i,j)} = \begin{cases} 0 & (j \neq l_e(i)) \\ 1 & (j = l_e(i)) \end{cases} . \tag{4}$$

3) Define $\mathbf{E_{bs}}$ as

$$\mathbf{E_{bs}} = \begin{bmatrix} k_{(1,1)} & k_{(1,2)} & \cdots & k_{(1,N)} \\ k_{(2,1)} & k_{(2,2)} & \cdots & k_{(2,N)} \\ \vdots & \vdots & \ddots & \vdots \\ k_{(N,1)} & k_{(N,2)} & \cdots & k_{(N,N)} \end{bmatrix}, \tag{5}$$

$$\mathbf{E_{bs}} \in \mathbb{R}^{N \times N}.$$

(c) Flatten each block $\hat{B}_i$ into a vector $\hat{b}_i$ such that

$$\hat{b}_i = [\hat{b}_i(1), \ldots, \hat{b}_i(L)]. \tag{6}$$

where $\hat{b}_i$ is the same as x_p^i in Eq. (1).

(d) Shuffle pixels in $\hat{b}_i$ with secret key k_2 such as

$$b'_i = \mathbf{E}_{ps} \hat{b}_i^\top, \ b'_i \in \mathbb{R}^L, \tag{7}$$

where $\mathbf{E_{ps}}$ is a matrix generated by the same procedure as $\mathbf{E_{bs}}$ (N is replaced with L).
Also, Eq. (7) is expressed as

$$b'_i = \mathbf{E}_{ps} \hat{b}_i^\top = \mathbf{E}_{ps} x_p^i \top = x_p^{'i}. \tag{8}$$

(e) Concatenate the encrypted vectors $b'_i (i \in 1, \ldots, N)$ into an encrypted image x'.

Figure 4 shows an example of images encrypted with this procedure, where block size p was 16.

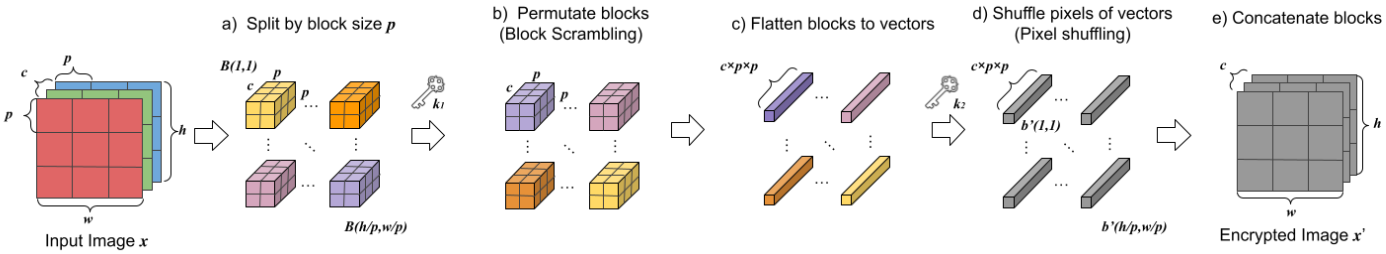
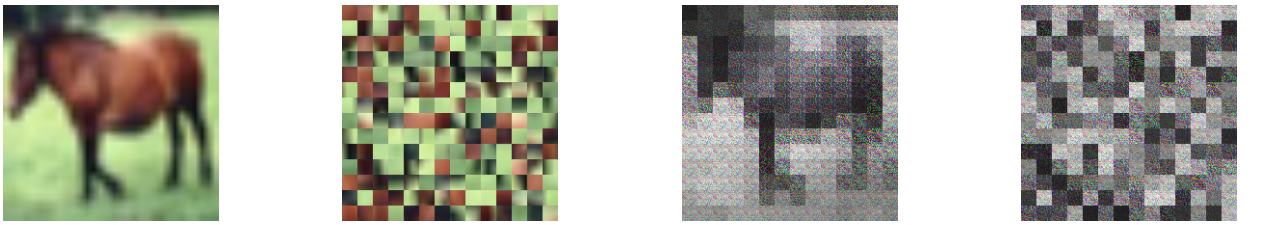


Fig. 3: Procedure of image encryption



(a) Original ($224 \times 224 \times 3$)
(b) Block scrambling (block size = 16)
(c) Pixel shuffling (block size = 16)
(d) Block scrambling + Pixel shuffling (block size = 16)

Fig. 4: Example of encrypted images

C. Domain Adaptation

As shown in Fig. 2, a model f_θ pre-trained in the plain domain is transformed into a model $\hat{f}_\theta$ to reduce the influence of encryption prior to the fine-tuning of the pre-trained model by using a novel domain adaptation method. The procedure of the proposed adaptation is summarized here.

The domain adaptation is carried out in accordance with the embedding structure of ViT. $\mathbf{E_{bs}}$ in Eq. 5 is used to permutate blocks in an image, so it has a close relationship with position embedding $\mathbf{E_{pos}}$. $\mathbf{E_{pos}}$ includes the position information of a class token x_{class}. In contrast, $\mathbf{E_{bs}}$ does not consider the information of x_{class}. To fill the gap, $\mathbf{E_{bs}}$ is extended as

$$\mathbf{E'_{bs}} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & k_{(1,1)} & k_{(1,2)} & \dots & k_{(1,N)} \\ 0 & k_{(2,1)} & k_{(2,2)} & \dots & k_{(2,N)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & k_{(N,1)} & k_{(N,2)} & \dots & k_{(N,N)} \end{bmatrix}, \tag{9}$$

$$\mathbf{E'_{bs}} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

In the domain adaptation, $\mathbf{E_{pos}}$ is transformed as in

$$\hat{\mathbf{E}}_{\mathbf{pos}} = \mathbf{E'_{bs}}\mathbf{E_{pos}}. \tag{10}$$

By using the adaptation, the position embedding of the pre-trained model can be adapted to the position information permutated by the block scrambling of images. Similarly, patch embedding $\mathbf{E}$ is transformed to adapt the position information of every pixel randomly replaced by pixel shuffling, as in

$$\hat{\mathbf{E}} = \mathbf{E_{ps}}\mathbf{E}. \tag{11}$$

TABLE I: Classification accuracy (%) with and without proposed method (CIFAR-10)

	Domain Adaptation	
Encryption	Yes	No
Pixel shuffling	**99.05**	97.78
Block scrambling	**98.93**	94.25
Pixel shuffling + Block scrambling	**98.98**	72.86
Plain	99.00	

From Eqs. 2, 8, 10, and 11, under the use of image encryption and domain adaptation, an adapted sequence of embedded patches is given by

$$\begin{aligned} \hat{z}_0 =& \mathbf{E'_{bs}}[x_{class}; \mathbf{E_{ps}}x_p^1\mathbf{E_{ps}}\mathbf{E}; \dots \mathbf{E_{ps}}x_p^N\mathbf{E_{ps}}\mathbf{E}]^\top + \mathbf{E'_{bs}}\mathbf{E_{pos}} \\ =& \mathbf{E'_{bs}}[x_{class}; \hat{x}_p^1\hat{\mathbf{E}}; \dots \hat{x}_p^N\hat{\mathbf{E}}]^\top + \mathbf{E'_{bs}}\mathbf{E_{pos}}. \end{aligned} \tag{12}$$

Finally, fine-tuning with encrypted images is applied to the pre-trained model corrected with the proposed adaptation.

IV. Experiments

A. Setup

We conducted image classification experiments on the CIFAR-10 [19] and CIFAR-100 datasets [19]. CIFAR-10 consists of 60000 images with 10 classes (6000 images for each class), and CIFAR-100 consists of 60000 images with 100 classes (600 images for each class), where 50000 images were used for training and 10000 images were used for testing. Also, we fine-tuned a ViT-B_16 model pre-trained with the ImageNet21k dataset, so we resized the image size from $32 \times 32 \times 3$ to $224 \times 224 \times 3$. Training parameters were a batch size of 32, a learning rate of 0.001, a momentum of 0.9, and

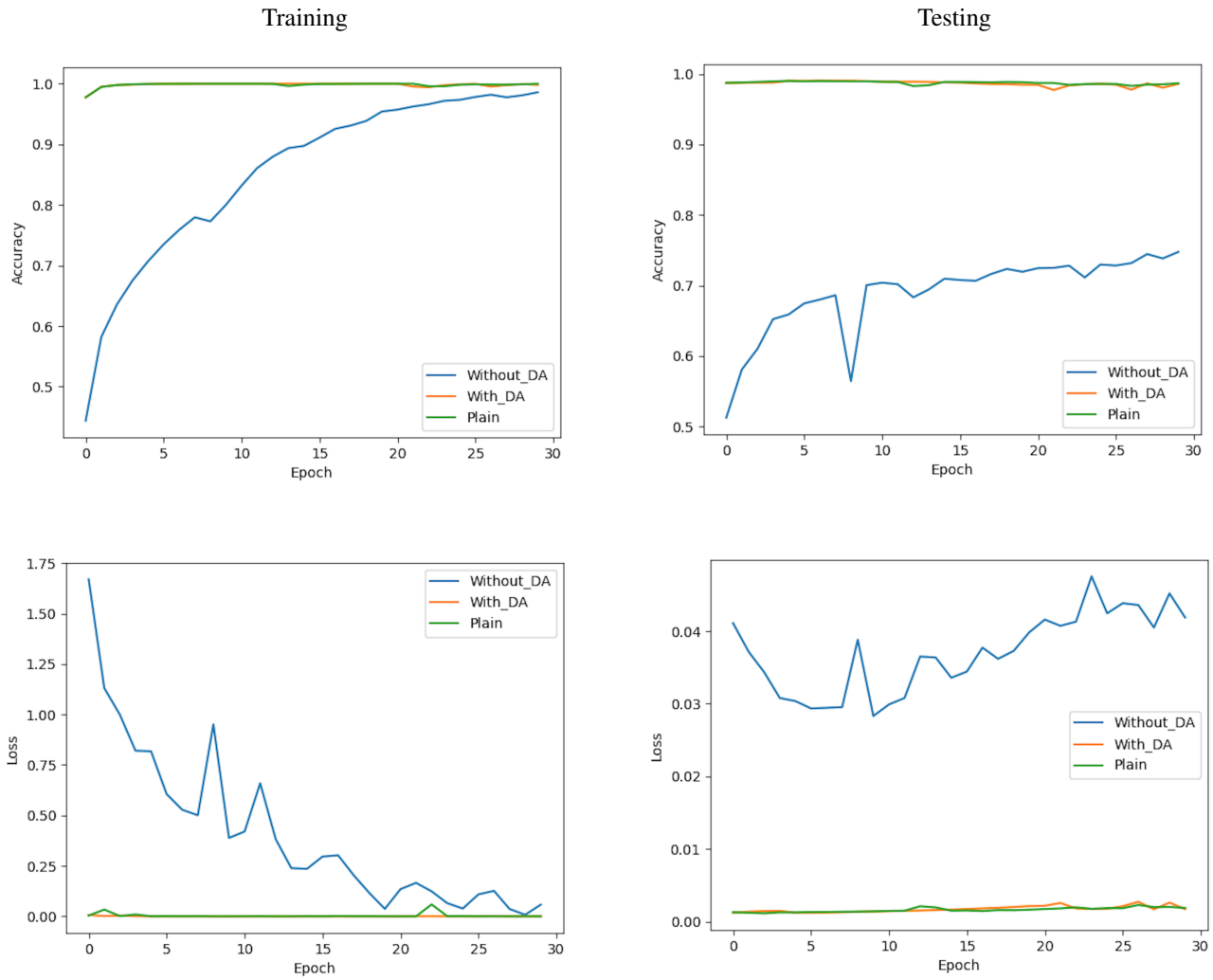


Fig. 5: Efficiency of model training
(Top: classification accuracy per epoch, Bottom: loss value per epoch)

a weight decay of 0.0005 on the stochastic gradient descent (SGD) for 15 epochs. A cross-entropy loss function was used as the loss function. The block size of the encryption was set to 16 to match the ViT patch size.

B. Effects of Domain Adaptation

1) Classification Performance: Table I shows the classification accuracy of models trained with encrypted images. When training models without the proposed domain adaptation, the accuracy decreased compared with that of using plain images. In particular, when applying the combined use of pixel shuffling and block scrambling, the classification accuracy decreased significantly. In contrast, when training models with the domain adaptation, the accuracy was almost the same as that of using plain images. Accordingly, we confirmed that the proposed domain adaptation was effective in fine-tuning pre-trained models with encrypted images.

2) Training Efficiency: Figure 5 shows the classification accuracy and loss value per epoch during training and testing where the number of epochs was increased to 30.

For training, when using the domain adaptation during training, models fine-tuned with encrypted images (with_DA) achieved almost the same training efficiency as those trained with plain images. In contrast, models fine-tuned without the domain adaptation (without_DA) needed the larger number of epochs to achieve the same performance as plain models. For testing, models with the domain adaptation also achieved the same performance as those with plain images, but models without the domain adaptation were overtrained.

Accordingly, the domain adaptation allows us not only to train a high-performance model without increasing training time even when using encrypted images, but to also avoid overtraining during test time.

C. Comparison with Conventional Methods

In Table II, the proposed method is compared with the state-of-the-art methods for privacy preserving image classification with encrypted images such as LE [3], PE [4], and ELE [5]. From the table, the proposed method not only has the lowest performance degradation on both CIFAR-10 and CIFAR-100

TABLE II: Comparison with existing methods

Encryption	Model	Accuracy (%)	
		CIFAR-10	CIFAR-100
LE [3]	AdaptNet + Shakedrop	94.49	75.48
ELE [5]	AdaptNet + Shakedrop	83.06	62.97
PE [4]	ResNet-18	92.03	-
EtC [6]	ViT-B_16	87.89	-
	ConvMixer-256/8	92.72	-
Color-NeuraCrypt [8]	ViT-B_16	96.20	-
Block scrambling + Pixel position shuffling [7]	ViT-B_16	96.64	84.42
Block scrambling + Pixel shuffling + Domain adaptation (Proposed)	ViT-B_16	**98.98**	**92.58**
Plain	ShakeDrop	96.70	83.59
Plain	ResNet-18	95.53	-
Plain	ViT-B_16	99.00	92.60
Plain	ConvMixer-256/8	96.07	-

datasets but also outperforms all existing methods in terms of the accuracy of image classification.

V. Conclusion

In this paper, we proposed a domain adaptation method to reduce the influence of image encryption for ViT models. The method is applied to a model pre-trained in the plain domain prior to the fine-tuning of the pre-trained model. In experiments, the proposed adaptation was demonstrated not only to reduce the performance degradation of models but to also achieve the highest classification accuracy in conventional methods on the CIFAR-10 and CIFAR-100 datasets.

Acknowledgment

This study was partially supported by JSPS KAKENHI (Grant Number JP21H01327) and JST CREST (Grant Number JPMJCR20D3).

References

[1] H. Kiya, M. AprilPyone, Y. Kinoshita, S. Imaizumi, and S. Shiota, "An overview of compressible and learnable image transformation with secret key and its applications," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, e11, 2022. DOI: 10.1561/116.00000048.

[2] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1515–1525, 2019. DOI: 10.1109/TIFS.2018.2881677.

[3] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2018, pp. 1–2. DOI: 10.1109/ICCE-China.2018.8448772.

[4] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, "Pixel-based image encryption without key management for privacy-preserving deep neural networks," *IEEE Access*, vol. 7, pp. 177 844–177 855, 2019. DOI: 10.1109/ACCESS.2019.2959017.

[5] K. Madono, M. Tanaka, O. Masaki, and O. Tetsuji, "Block-wise scrambled image recognition using adaptation network," in *Workshop on Artificial Intelligence of Things (AAAI WS)*, 2020.

[6] M. AprilPyone and H. Kiya, "Privacy-preserving image classification using an isotropic network," *IEEE MultiMedia*, vol. 29, no. 2, pp. 23–33, 2022. DOI: 10.1109/MMUL.2022.3168441.

[7] Z. Qi, M. AprilPyone, Y. Kinoshita, and H. Kiya, "Privacy-preserving image classification using vision transformer," *EURASIP European Signal Processing Conference, Belgrade, Serbia, August 31*, pp. 543–547, 2022.

[8] Z. Qi, M. AprilPyone, and H. Kiya, *Color-neuracrypt: Privacy-preserving color-image classification using extended random neural networks*, 2023. arXiv: 2301.04875.

[9] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2018. DOI: 10.1109/TIFS.2017.2787987.

[10] Y. Wang, J. Lin, and Z. Wang, "An efficient convolution core architecture for privacy-preserving deep learning," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5. DOI: 10.1109/ISCAS.2018.8350963.

[11] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv*, 2016. [Online]. Available: https://arxiv.org/abs/1610.05492.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, OpenReview.net, 2021.

[13] M. AprilPyone and H. Kiya, "Block-wise image transformation with secret key for adversarially robust defense," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2709–2723, 2021. DOI: 10.1109/TIFS.2021.3062977.

[14] H. Kiya, R. Iijima, M. AprilPyone, and Y. Kinoshita, "Image and model transformation with secret key for vision transformer," *IEICE Transactions on Information and Systems*, vol. E106.D, no. 1, pp. 2–11, 2023. DOI: 10.1587/transinf.2022MUI0001.

[15] A. H. Chang and B. M. Case, *Attacks on image encryption schemes for privacy-preserving deep neural networks*, 2020. arXiv: 2004.13263 [cs.CR].

[16] H. Ito, Y. Kinoshita, M. Aprilpyone, and H. Kiya, "Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks," *IEEE Access*, vol. 9, pp. 64 629–64 638, 2021. DOI: 10.1109/ACCESS.2021.3074968.

[17] A. Yala, H. Esfahanizadeh, R. G. L. D. Oliveira, *et al.*, *Neuracrypt: Hiding private health data via random neural networks for public training*, 2021. arXiv: 2106.02484.

[18] H. Kiya, T. Nagamori, S. Imaizumi, and S. Shiota, "Privacy-preserving semantic segmentation using vision transformer," *Journal of Imaging*, vol. 8, no. 9, 2022, ISSN: 2313-433X. DOI: 10.3390/jimaging8090233. [Online]. Available: https://www.mdpi.com/2313-433X/8/9/233.

[19] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Study on Face Landmark-based Analysis for Synthetic Media Identification Generated by Adversarial Generative Networks

Akinobu Ura*[1] and Minoru Kuribayashi* and Nobuo Funabiki*
* Graduate School of Natural Science and Technology, Okayama University, Japan
E-mail: [1] pl843doc@s.okayama-u.ac.jp

***Abstract*—With the development of deep learning technology, artificially created images and videos, known as deepfakes, have become a problem. It is very difficult for human visual and auditory senses to distinguish them from real ones. Therefore, there is a growing risk of them being used for crimes such as bullying, defamation, extortion, and fraud on social media. StyleGAN is one of the generators for such fake content and is known for its ability to generate high-quality images with a great degree of realism. This study aims to distinguish between real and fake images generated by StyleGAN. We introduce effective image processing techniques into the detector which is CNN model designed by fine-tuning the XceptionNet. We examined two methods: one that extracts color attributes from the image and the other that distinguishes based on landmarks. We evaluated the accuracy of distinguishing between real and fake images by computer simulation.**

I. Introduction

Using deep learning-based editing techniques in multimedia content, it has become easier to create scenes with potential risks or high production costs. These techniques are widely utilized in various fields such as television programs, film production, medical applications, and educational media materials. However, the advancement of such sophisticated content editing techniques has also brought about new challenges, such as the emergence of fake content which is artificially created by using a deep neural network. One prominent example of fake content is deepfakes, a term combining ”deep learning” and ”fake”. Deepfakes refer to a type of synthesized photos or videos created using deep learning techniques, characterized by their high quality, making it difficult to distinguish them from genuine content at first glance. This pose concerns as it enables the production of artificially created videos or images featuring influential individuals, such as leaders of nations or political parties, which could be used for manipulating public perception and damaging trust. Additionally, there are issues concerning the abuse of deepfakes for cyberbullying, defamation, threats, fraud, and other criminal activities on social media.

One of the techniques used to create such fake content is Generative Adversarial Networks (GANs) [1], which can generate entirely sysnthesized content by learning features from given datasets. Currently, various GAN derivatives, such as DCGAN [2], BigGAN [3], and PGGAN [4], have been developed. StyleGAN [5], capable of generating high-resolution images, can produce highly realistic fake images to the extent that it has been said, ”The era when photographs can serve as evidence is over”.

To address these fake content challenges, several countermeasures have been developed, including tools like Microsoft Video Authenticator[1] and initiatives like the DeepFake Detection Challenge[2] [6]. Microsoft Video Authenticator, released by Microsoft in September 2020, is a deepfake detection tool that analyzes images and videos to identify any anomalies and provides a confidence score indicating the likelihood of the content being a deepfake. The DeepFake Detection Challenge, launched in September 2019 by major companies and research institutions, is a contest aimed at developing technologies for detecting deepfake videos. While the detection rate for the provided deepfake videos exceeded 80%, the detection rate for different deepfake videos was around 65%, highlighting the challenge of detecting unknown data.

A problem in conventional study is that the performance of a detector significantly decreases when it is trained on images generated by one version of StyleGAN and tested on images generated by a different version. Our objective in this study is to achieve a universal detector that is not affected by the differences in the generator. We begin with a basic detector by designing a fine-tuning model of XceptionNet [7], which is customized for a binary classifier to distinguish between real and fake images. Since StyleGAN artificially combines characteristics from multiple training images to create its output, we propose a method for classifying local regions. In order to extract the features that reflect the characteristics of artificially generated images, we conducted to evaluate the classification accuracy based on different color spaces as a preliminary experiment. The results showed that the classification based on images represented in the HSV color space was effective. Therefore, we developed a method to classify the hue component, saturation component, and value component, which are the color attributes of the target images in the HSV color representation system. We evaluate the

[1] https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/
[2] https://www.kaggle.com/c/deepfake-detection-challenge

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

identification accuracy of real and fake images using these methods and an effective combination of them.

II. Related Study

In this section, we describe the technologies related to our study.

A. Generative Adversarial Networks (GANs)

GANs are a type of generative model that can learn features from data and generate new data or transform existing data along learned feature directions. GANs have attracted attention as an unsupervised learning method that can learn features without the need for labeled data.

GANs consist of two neural networks: a Generator and a Discriminator. The Generator generates data by adding features to random noise inputs and transforming them into data such as images. The Discriminator receives both the fake data generated by the Generator and the real data, and estimate the probability that a sample comes from the Generator. The Generator attempts to create fake samples, while the Discriminator attempts to determine which samples are fake and which are real. If the two models are allowed to compete for a long time, they will ultimately improve.

B. StyleGAN

StyleGAN is one type of generative adversarial network that can generate high-resolution images. It has been continuously improved and extended, including StyleGAN [5], StyleGAN2 [8], StyleGAN2-ADA [9], and StyleGAN3 [10]. Fig. 1 shows the architecture of StyleGAN, where the left side represents GAN, and the right side represents StyleGAN. StyleGAN consists of two networks: the Mapping network and the Synthesis network. Unlike conventional GANs, StyleGAN generates images from a fixed $4 \times 4 \times 512$ tensor instead of directly from the latent variable z. The Mapping network performs a non-linear transformation of the latent variable z into the latent space w using 8 fully connected layers. This transforms a sequence of random numbers (latent variable z) that has no semantic meaning at input time into a multidimensional space representing style information such as age, gender, and expression, which is used as style information in the Synthesis network. The Synthesis network performs image upsampling, merges style and noise information, and performs various manipulations such as adjusting the overall color and facial expressions by inserting style information at each upsampling step. AdaIN merges the latent space w obtained from the Mapping network as style information for each resolution. Noise introduces stochastic variation in fine visual details such as hair flow, beard, and freckles. If the noise is removed, the image will have a featureless, painting-like appearance.

StyleGAN2 addressed the problem of distortion on droplets, which was an issue in StyleGAN image generation, by modifying the method of applying style to each resolution. StyleGAN2-ADA applies a data augmentation technique called data augmentation to the Discriminator to prevent overfitting. StyleGAN3 resolved issues such as the teeth position, hair, and hair patterns being unnaturally fixed in absolute coordinates.

C. Synthetic Media Recognition

To analyze the slight unnatural signals present in GAN-generated images, identification methods using convolutional neural networks (CNNs) have been studied. When training CNN models from scratch, a huge dataset and computational resources are required, so fine-tuning of pre-trained models for image recognition is mainly studied [11]–[13]. Xuan et al. [14] use Gaussian filters in preprocessing to extract stable components that exclude unstable distorted signals and obtain stable identification accuracy. McCloskey et al. [15] consider a method that extracts feature components by weighting the color space based on the ratio of red to green in the RGB color space and identifies them using a support vector machine (SVM). Wang et al. [16] achieve high-accuracy identification by augmenting the data through techniques such as blurring and compression, thereby enhancing generalization and robustness.Methods using co-occurrence matrices [17] and analysis techniques using forensics [18], [19] have also been studied.

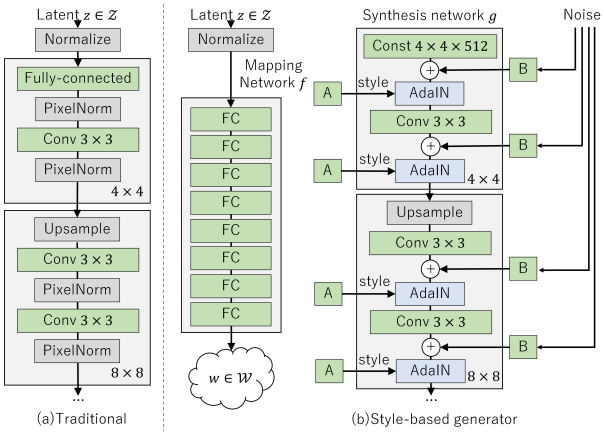


Fig. 1. Architecture of StyleGAN

D. Fine-Tuning Model

Fine-tuning is a technique where the fully connected layers of a pre-trained model are replaced, and the last part of the convolutional layers is retrained along with them. By using this technique, the pre-trained model is trained as a different model, allowing it to be utilized as a feature extractor. Fine-tuning offers several advantages, such as the ability to create high-accuracy models even with limited datasets and significant reduction in training time. As a result, it has become an essential technique in the development of CNN systems.

In this study, we employ fine-tuning on the pre-trained model, XceptionNet, to create a binary classifier capable of distinguishing between real and fake images. For training, we use the SGD gradient-based optimization scheme with a learning rate of 0.001, a batch size of 64, and momentum of 0.9. The network is trained until the loss, evaluated on the validation data, does not improve for 5 consecutive epochs.

III. PROPOSED METHOD

In this study, we propose a method that combines two preprocessing methods with the basic approach of fine-tuning the XceptionNet [7] model for identifying images generated by StyleGAN. The first method involves identifying segmented image blocks corresponding to local regions, namely the nose, right mouth corner, left mouth corner, right eye, and left eye (referred to as landmarks). The second method involves identifying images based on their color attributes (hue, saturation, and brightness).

A. Basic Detector

The basic detector is a binary classifier that fine-tunes the XceptionNet to distinguish between real and fake images. Without loss of generality, a target image is a RGB color image of $M \times N$ pixels, and it is divided into non-overlapping blocks B_i of $s \times s$ pixels for $1 \leq i \leq n$, where $n = \lfloor M/s \rfloor \times \lfloor N/s \rfloor$. To determine if the target image is a fake, each block B_i is checked by the detector, and the confidence level is calculated as the probability p_i that the block is fake. The final decision for the target image is determined by the decision rule given by the following equation:

$$P_b = \frac{1}{n} \sum_{i=1}^{n} (2p_i - 1) \tag{1}$$

- If $P_b \geq 0$, the image is classified as fake.
- Otherwise the image is classified as real.

As the input size of XceptionNet is 299×299 pixels, we use $s = 299$ as the input to the basic detector in this study.

B. Local Regions

When using StyleGAN to generate fake images, the characteristics of multiple training images are mixed to create artificially synthesized images. Since each local region has different characteristics, it is suitable to customize the detector to account for these differences. Therefore, we focus on five landmarks in a face and cut out five blocks that take into account the overlap of these blocks. We then use these landmark-specific blocks to train five fine-tuning models of the same architecture. Table I presents the five binary classifiers corresponding to each landmark. The final classification is determined by majority voting based on the results of these five blocks.

The MTCNN [20] is used to detect the five landmarks: nose, right mouth corner, left mouth corner, right eye, and left eye.

the coordinate points of each landmark are denoted as (x_ℓ, y_ℓ), $\ell \in \mathcal{L}$, where $\mathcal{L} = \{N : nose, MR : mouth_right, ML : mouth_left, RE : right_eye, LE : left_eye\}$. To determine if the target image is a fake, each block B_ℓ is checked by the detector, and the confidence level is calculated as the probability p_ℓ that the block is fake. The final decision for the target image is determined by the decision rule given by the following equation:

$$P_f = \frac{1}{5} \sum_{\ell \in \mathcal{L}} (2p_{\ell-1}) \tag{2}$$

TABLE I
BINARY CLASSIFIERS CORRESPONDING TO EACH LANDMARK

Binary classifier	landmark
model N	nose area
model MR	mouth_right area
model ML	mouth_left area
model RE	right_eye area
model LE	left_eye area

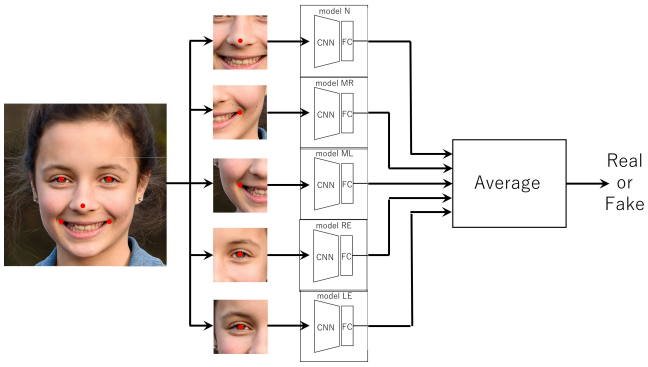


Fig. 2. Local regions processing flow

- If $P_f \geq 0$, the image is classified as fake.
- Otherwise the image is classified as real.

The processing flow of this method is as shown in Fig. 2.

C. Color Attributes

In order to extract the features that reflect the characteristics of artificially generated images, we conducted a preliminary experiment to evaluate the classification accuracy based on the differences in color spaces. Among the four color spaces tested, namely RGB, HSV, Lab, and YCbCr, it was found that the images represented in the HSV color space showed effective classification results. Fig. 3 illustrates real images and fake images generated by StyleGAN2-ADA in the RGB color space, HSV color space, and the Hue component. In the RGB color space, when the brightness of a certain color varies, the detectable range of that color may also change. However, in the HSV color space, the hue of a color remains consistent even when the brightness changes, allowing for the detection of the same color regardless of its brightness. This is a characteristic of the HSV color space. In convolutional processes like CNN, spatial features such as shape and position can be captured. However, these processes are unable to capture features resulting from differences in color spaces. Therefore, techniques such as color space conversion and extraction of color attributes are necessary to handle such differences in color spaces.

We aim to investigate the effective color attributes among the hue, saturation, and value components. Hue represents the type of color and is represented within a range of 0 to 360 degrees. Saturation refers to the intensity or vividness of a color and is represented within a range of 0 to 100%. Value represents the brightness of a color and is also expressed within a range of 0 to 100%. To convert from the RGB color space to the HSV



Fig. 3. The differences in color representation between real images and fake images vary depending on the color space used: the top images represent real images, while the bottom images represent fake images generated by StyleGAN2-ada. From left to right, the images correspond to the RGB color space, HSV color space, and the hue component.

color space, you can follow the steps below. However, note that among the three values R, G, and B, the maximum value should be denoted as MAX, and the minimum value should be denoted as MIN:

$$H = \begin{cases} \lfloor 60 \times ((G-B)/(MAX-MIN)) \rfloor & \text{if } R = MAX, \\ \lfloor 60 \times ((B-R)/(MAX-MIN)) + 120 \rfloor & \text{if } G = MAX \\ \lfloor 60 \times ((R-G)/(MAX-MIN)) + 240 \rfloor & \text{if } B = MAX \\ 0 & \text{if } MIN = MAX \end{cases}$$

$$S = \lfloor (MAX - MIN)/MAX \rfloor$$

$$V = MAX$$

We use blocks from the hue component to train fine-tuning model in this method, and the classification approach is the same as the basic detector.

IV. SIMULATION

In this section, we describe the simulation environment, evaluation metrics, and experimental results.

A. Simulation Environment

We used various datasets and set hyperparameters for the identifier in this study. We conducted the study under four different conditions, using a training dataset and a testing dataset. The breakdown of these cases is summarized in Table II.

The dataset used with StyleGAN2-ADA consists of two types of images: real images used to train StyleGAN2-ADA, taken from the FFHQ dataset, and fake images generated by StyleGAN2-ADA. The same applies to the dataset used with StyleGAN3. The original image size is 1024×1024 pixels, while the block size after division is set to 299×299 pixels. In the "Local Region" approach, 10,000 blocks each of fake and real images are used per landmark. In the "Basic Detector" and "Color Attributes" approaches, the image is divided into 9 segments, resulting in a total of 90,000 blocks (10,000 blocks per block for fake and real images).

TABLE II
BREAKDOWN OF DATASET

	Train dataset	Test dataset
Pattern 1	StyleGAN2-ADA	StyleGAN2-ADA
Pattern 2	StyleGAN2-ADA	StyleGAN3
Pattern 3	StyleGAN3	StyleGAN2-ADA
Pattern 4	StyleGAN3	StyleGAN3

TABLE III
HYPERPARAMETERS IN FINE-TUNING MODEL

Batch size	64		
Epoch	50		
Input size	299×299		
Fully connected layer	Dense	Nodes	256
		Activation function	relu
	Dropout	Node ratio	0.5
	Dense	Nodes	1
		Activation function	sigmoid
Optimizer	SGD(learning rate = 1e-3,momentum=0.9)		
Loss function	binary_crossentropy		

Additionally, we summarize the hyperparameters set for the detector used in this study in Table III.

B. Evaluation Metrics

In this study, as it involves binary classification of fake and real images, there are four possible outcomes for correctness assessment: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Let's explain each of them:

TP Represents the number of true (positive) items correctly identified as positive. In this study, it refers to the number of correctly identified real images as real.

TN Represents the number of false (negative) items correctly identified as negative. In this study, it refers to the number of correctly identified fake images as fake.

FP Represents the number of false (negative) items incorrectly identified as positive. In this study, it refers to the number of fake images mistakenly identified as real.

FN Represents the number of true (positive) items incorrectly identified as negative. In this study, it refers to the number of real images mistakenly identified as fake.

Accuracy
It is the ratio of the correctly identified images to the total number of images. This metric indicates how accurate the predictions are. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision
It is the ratio of the number of images correctly identified as real to the total number of images identified as real. This metric indicates the level of

misidentification. The formula for precision is:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall

It is the ratio of the number of real images correctly identified as real to the total number of actual real. This metric indicates the level of under-detection. The formula for recall is:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

C. Difference of Local Regions

We compared the classification accuracy of the basic detector with the classification accuracy of each local region to examine the effect of local regions.

First, We present the comparison results of the classification accuracy for the five detectors trained on local regions in a Table IV. Each detector was trained using blocks that consider the overlapping landmarks. Specifically, if a single block contains multiple landmarks (e.g. nose, right_eye, left_eye), that block is used for training multiple detectors (model N, model RE, and model LE). However, images that failed to detect any of the five landmarks are excluded from the training data.

Next, we present the total result of the majority voting based on the five detectors in the Table V.

According to Table IV, when the training and testing datasets differ, the landmark-specific classifiers showed approximately 10-20% improvement in Accuracy and Precision compared to the basic detector. There was no significant difference in performance among the classifiers based on landmarks. Similarly, according to Table V, the overall results also showed approximately 10-20% improvement in Accuracy and Precision compared to the basic detector.

D. Difference of Color Attribute

To examine the effect of color attributes, we compared the classification accuracy of the classification accuracy of images where the Hue (H), Saturation (S), and Value (V) components were extracted separately. The experimental results are shown in the Table VI.

The results showed that both Accuracy and Precision were highest for the H component, with approximately 13-45% improvements compared to the basic detector. The S and V components showed similar accuracy to the basic detector. Recall was high for all cases.

E. Combination Approach

Based on the results of the proposed methods so far, an effective combination approach was performed by conducting local region-based classification on the Hue component of the target image. The comparison results of the classification accuracy for the five detectors are shown in Table VII, and the results of majority voting for the five detectors are shown in Table VIII.

TABLE IV
COMPARISON OF ACCURACY BETWEEN THE BASIC DETECTOR AND THE FIVE DETECTORS IN IDENTIFYING THE TARGET IMAGE IN THE RGB COLOR SPACE

(A) Accuracy

dataset		Accuracy					
train	test	basic	N	MR	ML	RE	LE
2-ADA	2-ADA	1.000	0.996	0.993	0.994	0.992	0.992
2-ADA	3	0.502	0.636	0.641	0.659	0.724	0.655
3	2-ADA	0.512	0.736	0.672	0.711	0.721	0.738
3	3	1.000	0.999	0.999	0.998	0.989	0.998

(B) Precision

dataset		Precision					
train	test	basic	N	MR	ML	RE	LE
2-ADA	2-ADA	1.000	0.994	0.989	0.991	0.989	0.988
2-ADA	3	0.501	0.579	0.582	0.594	0.645	0.592
3	2-ADA	0.506	0.655	0.604	0.634	0.642	0.657
3	3	1.000	1.000	0.999	1.000	0.999	1.000

(C) Recall

dataset		Recall					
train	test	basic	N	MR	ML	RE	LE
2-ADA	2-ADA	1.000	0.999	0.998	0.998	0.996	0.997
2-ADA	3	1.000	0.998	0.998	0.998	0.996	0.997
3	2-ADA	1.000	0.998	0.998	0.997	0.997	0.995
3	3	1.000	0.998	0.998	0.997	0.997	0.995

TABLE V
TOTAL RESULT OF FIVE DETECTORS BASED ON THE TARGET IMAGE IN THE RGB COLOR SPACE

dataset		Accuracy	Precision	Recall
train	test			
2-ADA	2-ADA	0.999	0.999	1.000
2-ADA	3	0.638	0.580	1.000
3	2-ADA	0.712	0.635	0.999
3	3	1.000	1.000	0.999

Based on the results, as shown in Table VII, the detectors for MR and ML exhibited higher Accuracy and Precision compared to the detectors for other landmarks. When comparing Table VIII and Table VII, it can be observed that the overall results of the five detectors did not show any improvement in accuracy compared to their individual results. One possible reason for this is that when classifying fake images, if one out of the five landmarks is determined to be real with high probability while the others are determined to be fake with subtle probabilities, the final result may be erroneously classified as real due to being influenced by the result of that one landmark.

F. Dependency on generator

According to Table IX, the detector trained with the older version of StyleGAN exhibited higher accuracy in image classification compared to the detector trained with the newer version of StyleGAN. Therefore, it was hypothesized that detectors trained with relatively newer versions of StyleGAN would show a decrease in classification accuracy. To test this hypothesis, an experiment was conducted.

The detector used in this experiment was trained on the Hue component of the target images, as mentioned in the previous

TABLE VI
COMPARISON OF ACCURACY BETWEEN RGB COLOR SPACE IMAGES AND IMAGES WITH HUE, SATURATION, AND VALUE COMPONENT USING THE BASIC DETECTOR

(A) Accuracy

dataset		Accuracy			
train	test	RGB	H	S	V
2-ADA	2-ADA	1.000	1.000	1.000	1.000
2-ADA	3	0.502	0.953	0.500	0.501
3	2-ADA	0.512	0.713	0.500	0.500
3	3	1.000	1.000	1.000	1.000

(B) Precision

dataset		Precision			
train	test	RGB	H	S	V
2-ADA	2-ADA	1.000	1.000	1.000	1.000
2-ADA	3	0.501	0.914	0.500	0.500
3	2-ADA	0.506	0.635	0.500	0.500
3	3	1.000	1.000	1.000	1.000

(C) Recall

dataset		Recall			
train	test	RGB	H	S	V
2-ADA	2-ADA	1.000	1.000	1.000	1.000
2-ADA	3	1.000	1.000	1.000	1.000
3	2-ADA	1.000	1.000	1.000	1.000
3	3	1.000	1.000	1.000	1.000

TABLE VII
COMPARISON OF ACCURACY BETWEEN THE BASIC DETECTOR AND THE FIVE DETECTORS IN IDENTIFYING THE HUE COMPONENT OF THE TARGET IMAGE

(A) Accuracy

dataset		Accuracy					
train	test	basic(H)	N	MR	ML	RE	LE
2-ADA	2-ADA	1.000	0.988	0.986	0.987	0.992	0.985
2-ADA	3	0.953	0.609	0.764	0.851	0.705	0.701
3	2-ADA	0.713	0.678	0.752	0.714	0.740	0.745
3	3	1.000	0.996	0.996	0.995	0.997	0.995

(B) Precision

dataset		Precision					
train	test	basic(H)	N	MR	ML	RE	LE
2-ADA	2-ADA	1.000	0.980	0.978	0.980	0.988	0.975
2-ADA	3	0.914	0.561	0.680	0.773	0.623	0.627
3	2-ADA	0.635	0.609	0.670	0.638	0.659	0.664
3	3	1.000	0.998	0.999	0.999	0.999	0.999

(C) Recall

dataset		Recall					
train	test	basic(H)	N	MR	ML	RE	LE
2-ADA	2-ADA	1.000	0.996	0.994	0.995	0.996	0.995
2-ADA	3	1.000	0.996	0.994	0.995	0.996	0.995
3	2-ADA	1.000	0.994	0.993	0.991	0.995	0.991
3	3	1.000	0.994	0.993	0.991	0.995	0.991

section on "Color Attribute Effects". The test dataset consisted of real images and fake images generated from both StyleGAN and StyleGAN2. The original image size was 1024×1024 pixels, and the block size after division was set to 299×299 pixels. Since there were 9 divisions, a total of 400 images $\times$ 9 divisions = 3600 blocks were used.

The experimental results are shown in Table IX. The results show that the discriminator that used StyleGAN3 during training had lower Accuracy and Precision when using StyleGAN than when using StyleGAN2 during testing. In addition, discriminators that used StyleGAN2-ADA during training had lower Accuracy and Precision when using StyleGAN than when using StyleGAN3 during testing (refer to the previous section on "Color Attribute Effects"). This indicates that performance tended to be higher when using images generated with a lower version of StyleGAN than when training with images generated with a newer version of StyleGAN.

TABLE VIII
TOTAL RESULT OF FIVE DETECTORS BASED ON THE HUE COMPONENT OF THE TARGET IMAGE

dataset		Accuracy	Precision	Recall
train	test			
2-ADA	2-ADA	0.997	0.995	1.000
2-ADA	3	0.707	0.631	1.000
3	2-ADA	0.731	0.651	1.000
3	3	1.000	1.000	1.000

TABLE IX
RESULTS OF IDENTIFYING THE HUE COMPONENT OF IMAGES GENERATED BY STYLEGAN AND STYLEGAN2 USING THE BASIC DETECTOR

dataset		Accuracy	Precision	Recall
train	test			
3	2	0.740	0.658	1.000
3	1	0.509	0.504	1.000
2-ADA	2	0.999	1.000	0.998
2-ADA	1	0.713	0.635	0.998

V. CONCLUSIONS

In this study, we evaluated the classification accuracy of two methods: one that identifies images based on local regions and another that extracts the color attributes of the images. We also proposed an effective combination of these methods. The method of using the basic detector to identify the Hue component of the target image was the most effective. The proposed method achieved the highest accuracy of approximately 95% when distinguishing between fake images trained with a different GAN. This demonstrates the effectiveness of our approach. In the future, we plan to further investigate the classification of other GANs, ensemble learning through bagging to determine the majority vote of the five classifiers, and delve into the impact of generator differences.

ACKNOWLEDGMENT

This study was supported by the JSPS KAKENHI (22K19777), JST SICORP (JPMJSC20C3), JST CREST (JPMJCR20D3), Japan.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Proc. NeurIPS'14*, vol. 27, 2014, pp. 2672–2680.

[2] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR'16*, 2016.

[3] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR'19*, 2019.

[4] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. ICLR'18*, 2018.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021. DOI: 10.1109/TPAMI.2020.2970919.

[6] B. Dolhansky, J. Bitton, B. Pflaum, *et al.*, "The deepfake detection challenge dataset," *CoRR*, vol. abs/2006.07397, 2020. arXiv: 2006.07397.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016. arXiv: 1610.02357. [Online]. Available: http://arxiv.org/abs/1610.02357.

[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR'19*, 2020, pp. 8107–8116. DOI: 10.1109/CVPR42600.2020.00813.

[9] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. NeurIPS'20*, 2020, pp. 12 104–12 114.

[10] T. Karras, M. Aittala, S. Laine, *et al.*, "Alias-free generative adversarial networks," in *Proc. NeurIPS'21*, 2021, pp. 852–863.

[11] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *CoRR*, vol. abs/1812.02510, 2018. arXiv: 1812.02510.

[12] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 384–389. DOI: 10.1109/MIPR.2018.00084.

[13] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot...for now," in *Proc. CVPR'20*, 2020.

[14] X. Xuan, B. Peng, W. Wang, and J. Dong, "On the generalization of gan image forensics," in *Proc. CCBR'19*, Berlin, Heidelberg: Springer-Verlag, 2019, pp. 134–141, ISBN: 978-3-030-31455-2. DOI: 10.1007/978-3-030-31456-9_15.

[15] S. McCloskey and M. Albright, "Detecting GAN-generated imagery using color cues," *CoRR*, vol. abs/1812.08247, 2018. arXiv: 1812.08247. [Online]. Available: http://arxiv.org/abs/1812.08247.

[16] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, *Cnn-generated images are surprisingly easy to spot... for now*, 2020. arXiv: 1912.11035 [cs.CV].

[17] L. Nataraj, T. M. Mohammed, B. S. Manjunath, *et al.*, "Detecting GAN generated fake images using co-occurrence matrices," in *Proc. MWSF'19*, vol. 5, 2019, pp. 532.1–532.7.

[18] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" In *Proc. MIPR'19*, 2019, pp. 506–511. DOI: 10.1109/MIPR.2019.00103.

[19] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. ICCV'19*, 2019, pp. 7555–7565. DOI: 10.1109/ICCV.2019.00765.

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. DOI: 10.1109/LSP.2016.2603342.

HDR Image Watermarking based on Saliency Detection and Quantization Index Modulation

Ahmed Khan*, Minoru Kuribayashi†, KokSheik Wong*, Vishnu Monn Baskaran*
* School of Information Technology, Monash University Malaysia, Malaysia.
†Graduate School of Natural Science and Technology, Okayama University, Japan.
E-mail: *{ahmed.khan1, wong.koksheik, vishnu.monn}@monash.edu, †kminoru@okayama-u.ac.jp

***Abstract*—High-dynamic range (HDR) images are circulated rapidly over the internet with risks of being exploited for unauthorized usage. To protect these images, some HDR image-based watermarking methods were put forward. However, they inherited the same problem faced by the conventional IW methods for standard dynamic range images, where only trade-offs among conflicting requirements are managed instead of simultaneous improvement. In this paper, a novel saliency (eye-catching object) detection based trade-off independent HDR-IW is proposed, to simultaneously improve robustness, imperceptibility and payload. First, the host image goes through our proposed salient object detection model to produce a saliency map, which is, in turn, exploited to segment the foreground and background of the host image. Therefore, a convolutional deep learning model is trained with various datasets containing multiple salient objects. Next, the binary watermark is partitioned into the foregrounds and backgrounds using the same mask and scrambled using a random permutation algorithm. Finally, the watermark segments are embedded into the selected bit-plane of the corresponding host segments using quantized indexed modulation. Experimental results suggest that the proposed work outperforms state-of-the-art methods in terms of concurrently improving the aforementioned conflicting requirements.**

I. Introduction

High dynamic range (HDR) imaging allows for better image quality and visual perception because it can represent a wider range of illumination in comparison to standard dynamic range (SDR) imaging [1]. Consequently, HDRI is deployed for various applications, including computer graphics, computer vision, broadcasting, surveillance, to name a few [2]. Often these images are of commercial, political, educational, or sentimental value. Hence they must be secured from unauthorized usage or illegal copying during transmission or in storage. Image watermarking (IW) is an art and science that injects copyright data into an image (host) to guard the host image. The inserted data can be retrieved later from the potentially modified image for copyright claim, authentication, tampering detection, etc. [3]–[5].

In the case of HDR-IW, various non-blind and blind state-of-the-art (SOTA) IW methods are proposed, but often only a trade-off among capacity, quality, and robustness (especially against the tone-mapping operations (TMOs)) is achieved. Huang et al. [6] designed a zero-bit IW method to protect the HDR image using dual-tree complex wavelet transform (DCWT) and Quaternion Fast Fourier transform (QFFT). First, DCWT and QFFT are sequentially applied to each color channel. Next, singular value decomposition (SVD) is applied to each LL subband, and finally, the watermark is embedded into the diagonal of the S matrices. Perez et al. [7] presented an HDR-IW method using Luma Variation Tolerance. Specifically, the scrambled watermark is embedded into the selected HDRI regions with less luma variation, to maintain the image quality. Maiorana et al. [8], [9] and Shiju et al. [10] designed an HDR-IW system primarily based on Guerrini et al.'s method [4]. However, the watermarked image quality is better due to the utilization of the radon-DCT. Autrusseau et al. [11] proposed an HDR-IW to improve the robustness feature. First, RGB channels are extracted from the HDRI using the normalization process. Fianlly, DWT is applied to each color channel to obtain the LH sub-band for the watermark embedding using α-blending.

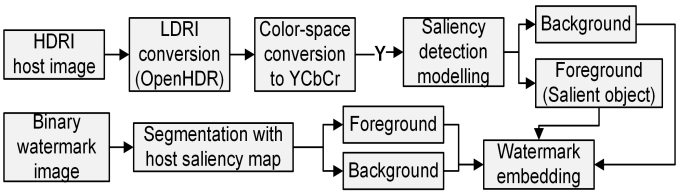


Fig. 1: Proposed framework of HDR-IW method based on salient object detection. Detailed flow-diagrams of the proposed saliency detection model and the watermark embedding procedure are shown in Fig. 3 and 2, respectively.

In analyzing the aforementioned methods, it is evident that capacity is sacrificed in an attempt to improve the other two aspects of quality and robustness [6], [8]–[19]. To simultaneously achieve high embedding capacity, image quality, and robustness, one approach is to leverage on different saliency regions of an image for a watermarking process [12], [20]. This can be achieved by using salient object detection (SoD) to extract the salient regions of an image to embed the watermark. With this approach, the critical regions of an image are used to embed the watermark without compromising the apparent visual quality. While subjectively reducing any observed visual distortions by the human eye, the IW can be made more resilient against different attacks. Motivated by this approach, we propose a novel framework for a reliable and invisible HDR-IW using SoD (see Fig 1 and Fig 3). The proposed method ensures that the quality of the watermarked image remains intact while protecting the watermark (i.e., by providing an extra layer of security).

The rest of the paper is organized as follows, Section II

©2023 APSIPA

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

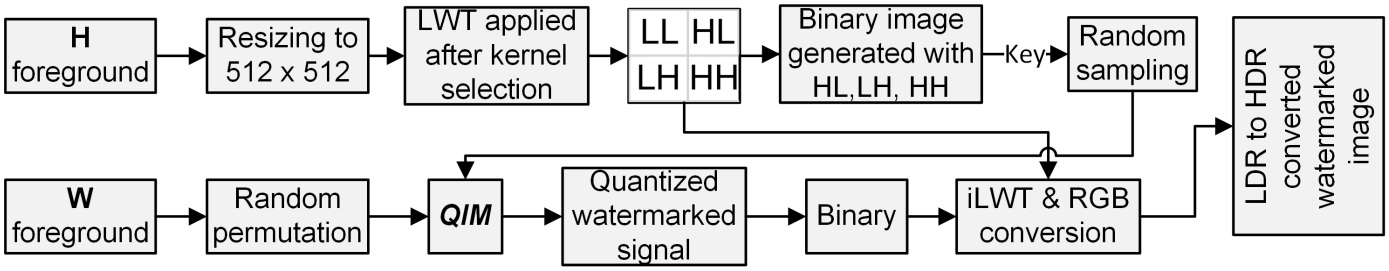


Fig. 2: Proposed framework of our HDR-IW design

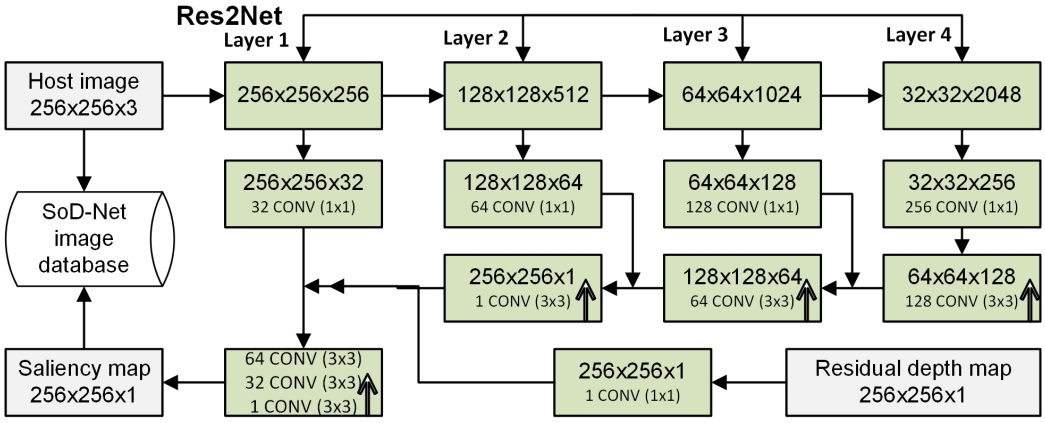


Fig. 3: Proposed framework of SoD-Net

details the proposed method, SoD-Net, watermark encryption, embedding methods. Experimental results are discussed in Section III, followed by the conclusion in Section IV.

II. Methodology

First, the host HDR image is tone-mapped to an SDR image H. From H, the foreground H_f and background H_b are determined using our proposed SoD-Net method (see Fig 3). Based on the features of H, the watermark W is then divided into two partitions W_f and W_b such that the $W = W_f \bigcup W_b$, and both W_f and W_b are shuffled before being embedded into the H (see Fig. 1). Details are provided in the following subsections.

A. Res2Net based Salient Object Detection

SoD-Net aims to detect salient objects in H (see Fig. 3). First, HDR to SDR conversion is performed using normalized tone-mapping operation

$$\tau = \lceil \log_2[\max(c) + 128] \rceil \tag{1}$$

to get

$$H'_c = \lfloor \frac{c \cdot 256}{2^{(\tau - 128)}} \rfloor \tag{2}$$

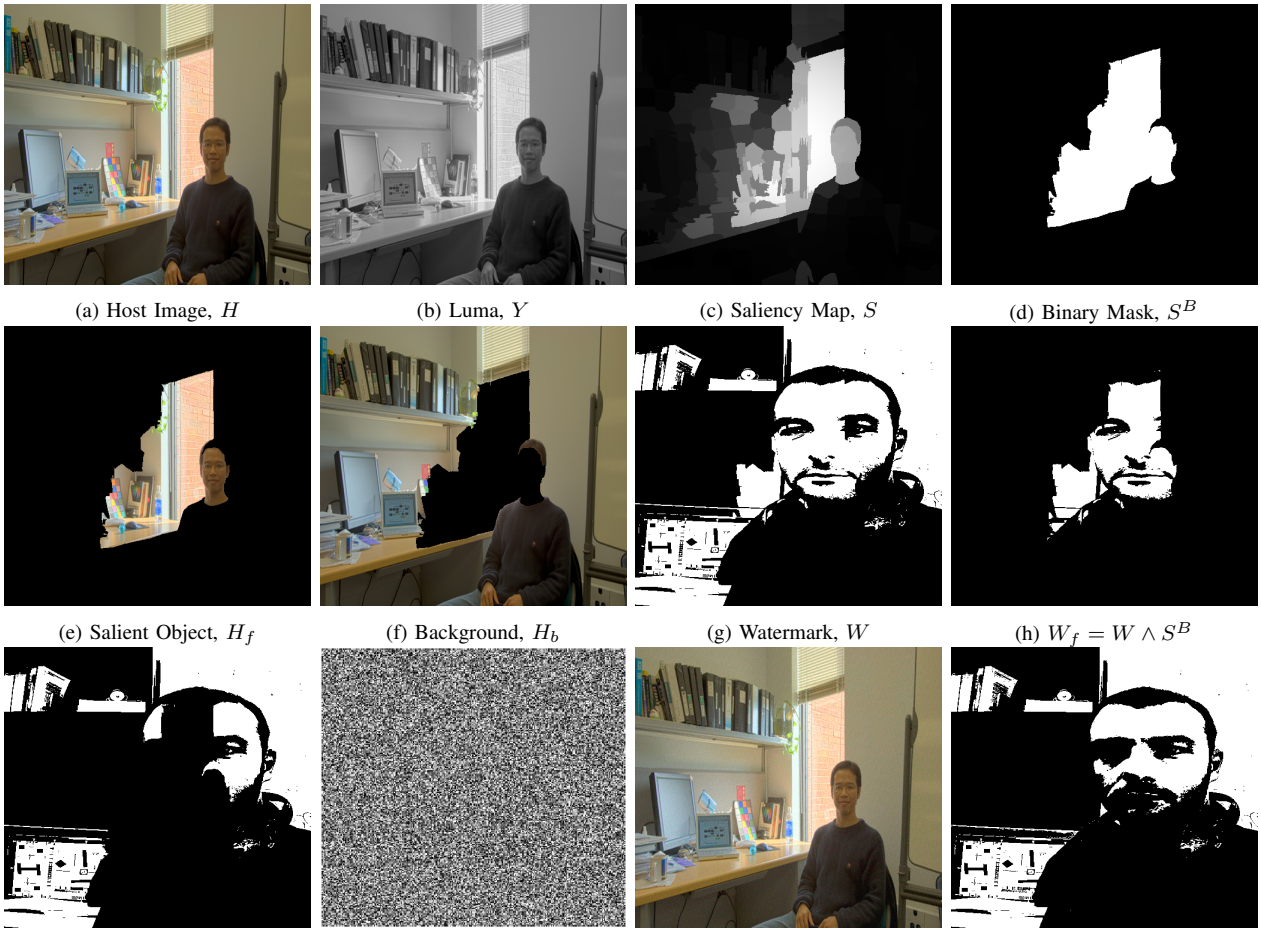
and

$$H''_c = \lfloor \frac{H'_c + 0.5 + 2^{(\tau - 128)}}{256} \rfloor \tag{3}$$

where $c \in \{R, G, B\}$. Here, $\{H''_c, H'_c\}$ are the corresponding color channels c in SDR and HDR, respectively. Next, the color and edge features are extracted from H' using Res2Net [21].

In SoD-Net, to verify that the output of Res2Net is an image, we eliminate all the completely linked layers. We remove the initial max pooling layer and, we set CONVs stride to 1 (instead of 2) to avoid excessive down-sampling ($\downarrow$). As a result, sparse feature loss is avoided, while retaining the saliency details. Res2Net offers four outputs through the $\downarrow$ phases, namely, low-level features extracted at Layer 1, middle-level features at Layers 2 and 3, and sparse features at Layer 4, which we receive at $\downarrow$ phase. One crucial use of 1×1 CONVs is as a dimension reduction module to eliminate computational constraints that would restrict the size of SoD-Net. It expands our networks' depth and width without suffering performance degradations. The number of channels is then reduced to 1/8 of the original number using four 1×1 CONVs. The downsizing is performed because the original data is large in size, hence consuming more time for the training and testing phase.

Similarly, to create the depth saliency map, we erase the first max pooling layers, set the CONVs stride to 1, and utilize 3×3 CONVs. The depth error in non-salient objects causes the disparity map's nearest objects to have either the lowest or highest intensities. These objects have the potential to deceive neural networks (NN) when NN models seek for salient objects in depth maps. As a result, a single-modal saliency map that simply includes depth information is flawed. By building a hybrid saliency map using the additive RGB and depth information, we can get rid of depth inaccuracy.



(a) Host Image, H (b) Luma, Y (c) Saliency Map, S (d) Binary Mask, S^B

(e) Salient Object, H_f (f) Background, H_b (g) Watermark, W (h) $W_f = W \wedge S^B$

(i) $W_b = W \wedge (1 - S^B)$ (j) Encrypted, W^e (k) Watermarked, H^W (l) Extracted, W'

Fig. 4: Intermediate images produced by proposed image watermarking method

Specifically, the weighted saliency map created by combining the RGB and depth modes contains the likelihood information of each pixel to be part of a certain object. We design a decoder network that consists of 3×3 CONVs and up-sampling ($\uparrow$) to produce the cumulative saliency map. We concatenate the lower-level and current features following each $\uparrow$. The decoder block D^n is as follows

$$S = \bigcup D^n = \sum_{i=1}^{k} \uparrow (CONV(D_i^{n-1} \bigoplus d_i^n.\Theta)), \quad (4)$$

where i refers to the i^{th} feature channel, $CONV$ is the convolution with value Θ, and $\bigoplus$ is the *concatenation* operator. The binary mask S^B is produced by thresholding S [12]. The loss function is $L(S, g) = \sum\sum(S \log g) + (1 - S) \log(1 - g)$, where g is the ground truth. Finally, H_b is obtained by $H_b = H' \wedge (1 - S^B)$. Likewise, H_f is obtained by $H_f = H' \wedge S$. Subsequently, W_f and W_b are calculated (see Fig. 4).

B. QIM based Watermark Embedding

We embed the divided watermark $\{W_f, W_b\}$ correspondingly into $\{H_f, H_b\}$ after shuffling them according to a secret key (see Fig. 2). Details are provided in the following steps.

Step 1: First, $\{H_f, H_b\}$ are resized (cubic interpolation) to 512×512, followed by the YC_bC_r conversion. Next, lifting wavelet transform (LWT) using *Daubechies cdf2.2-kernel* is applied to H_f of Y to produce the sub-bands $[LL, HL, LH, HH]_{\downarrow}$. Similarly, LWT using *Haar-kernel* is applied to H_b of Y. Note that we apply the same method for both $\{H_f, H_b\}$ to embed $\{W_f, W_b\}^e$, which is of size $H/2$.

Step 2: Subsequently, the shuffled binary $\{W_f, W_b\}^e$ is generated by applying random permutation to $\{W_f, W_b\}$. Next, a binary image $B^n_{\kappa,(i,j)} \in \{0, 1\}$ is constructed from the n-th bit plane of $\kappa \in \{HL, LH, HH\}$ sub-band. In order to generate the sampled host signal $\varphi_{(i,j)}$, a random sampling operation is applied to $B^n_{\kappa,(i,j)}$ according to a secret key.

Step 3: The quantization modulation operation is applied to embed $W^e_{(i,j)}$, which is the (i, j)-th element of $\{W_f, W_b\}^e$, in $\varphi_{(i,j)}$. The watermarked signal $\varphi^W_{(i,j)}$ is computed as:

$$\begin{aligned} \varphi^W_{(i,j)} &= \text{QIM}(\varphi_{(i,j)}, W^e_{(i,j)}, \Delta) \\ &= \begin{cases} \Delta \cdot \tilde{\varphi}_{(i,j)} & \text{if } W^e_{(i,j)} = \tilde{\varphi}_{(i,j)} \pmod 2; \\ \Delta \cdot \tilde{\varphi}_{(i,j)} + 1 & \text{otherwise,} \end{cases} \end{aligned} \quad (5)$$

where Δ is the quantization step within the range of [0.01,



(a) Exploratorium (b) BMW 507 (c) Frontier

Fig. 5: Additional host images for experiments

1.00], and $\tilde{\varphi}_{(i,j)} = \lfloor \frac{\varphi_{(i,j)}}{\Delta} \rfloor$. Next, τ is embedded into LL (Eq. 5) for synchronized robust TMO.

Step 4: The binary form of $\varphi^{W}_{(i,j)}$ is produced in order to combine it with the binary $B^{n}_{\kappa(i,j)}$ to perform the smooth inverse operation. Finally, inverse LWT followed by YC_bC_r to *RGB* conversion are computed to achieve H'^{W}. Finally, inverse-TMO is applied to H'^{W} to produce the HDR watermarked H^{W}. Intermediate results are shown in Fig. 4.

C. Watermark Extraction

It is noteworthy that the watermark can be extracted as W' by reversing the embedding steps. Note that the SoD-Net process is independent of the QIM based watermark embedding. Specifically, SoD-Net builds a model of the saliency maps information of the earlier used host images. On a later stage, at the time the watermark extraction, the same saliency maps are used to segment the foreground and background of the host images. That means, the host images are not going to pass through the SoD-Net model again to evaluate the saliency maps. Therefore, SoD-Net is completely stable with the QIM based watermark embedding. Next, inverse QIM is applied to extract the watermark images.

In addition, the watermark is extracted in the SDR domain, and hence robustness against tone-mapping is important and is of our interest. Intermediate results are shown in Fig. 4.

III. Experiments and Results

Our SoD-Net HDR-IW method is implemented in Matlab 2020 (Windows 10) running on a PC with an AMD Ryzen 5 3500U, Radeon Vega Gfx, and 16GB-RAM. For experiment and comparison purposes, we used the RIT-MCSL [22] and MSRA datasets[1] [24]. Additionally, we set the $\Delta = 0.50$, background to be slightly hazy ($\beta = 0.04$) and salient object to be entirely invisible ($\alpha = 0.02$). Visual results are shown in Fig. 4. The PSNR and SSIM of H^{W} are evaluated [19]. Results suggest that H^{W} produced by our method is of high quality, i.e., $>$ 59dB in terms of PSNR, and it is consistently higher in comparison to SOTA methods (see Table I). The SSIM values also indicate that our watermarked images are of high quality, which agrees with the PSNR value.

Next the normalized correlation (NC) [6] $NC \in [0, 1]$ and bit error rate (BER), is considered to assess the quality of W'.

[1]For each SDR image, we produce 6 images each of different exposure, which are then combined using Durand's inverse TMO method [23].

TABLE I: Analyzing quality of W and H^{W} for our and SOTA methods

Image	Proposed	[25]	[22]	[1]	[18]	[6]
	Man - Fig. 4(a)					
PSNR (dB)	**59.85**	49.60	48.97	50.43	46.22	45.04
SSIM	**0.9999**	-	0.9874	-	-	0.9867
NC for W'	0.9999	0.9865	**1**	**1**	0.9820	0.9761
	Exploratorium - Fig.5(a)					
PSNR (dB)	**60.22**	48.33	47.30	49.46	45.20	45.37
SSIM	**0.9999**	-	0.9869	-	-	0.9866
NC for W'	0.9999	0.9835	**1**	0.9989	0.9630	0.9761
	BMW 507 - Fig.5(b)					
PSNR (dB)	**59.92**	48.78	48.11	48.99	45.32	45.11
SSIM	**0.9999**	-	0.9927	-	-	0.9892
NC for W'	0.9999	0.9769	0.9908	**0.9999**	0.9750	0.9761
	Frontier - Fig.5(c)					
PSNR (dB)	**59.81**	49.78	46.30	49.08	46.05	45.19
SSIM	**0.9999**	-	0.9811	-	-	0.9890
NC for W'	0.9999	0.9862	0.9992	**0.9999**	0.9790	0.9761

TABLE II: Results NC after applying watermark attack

Attack	NC_{W_1}	NC_{W_2}	[25]	[22]	[23]	[18]	[6]
Mean-filter	0.9861	0.9699	**0.9872**	0.9718	0.9748	-	0.9710
Median-filter	0.9848	0.9683	**0.9911**	-	-	0.9840	-
Shearing	**0.9864**	0.9675	-	-	0.9513	0.9658	-
Noise	**0.9920**	0.9792	0.8931	0.9680	0.9905	0.9850	0.9178
Rotation	**0.9791**	0.9518	0.9090	0.9055	0.6335	0.9498	0.8799
Cropping	**0.9893**	0.9681	-	0.9552	0.9602	-	0.9685
JPEG-Comp	0.9672	0.9446	-	**0.9688**	0.9559	-	0.9220

Mathematically,

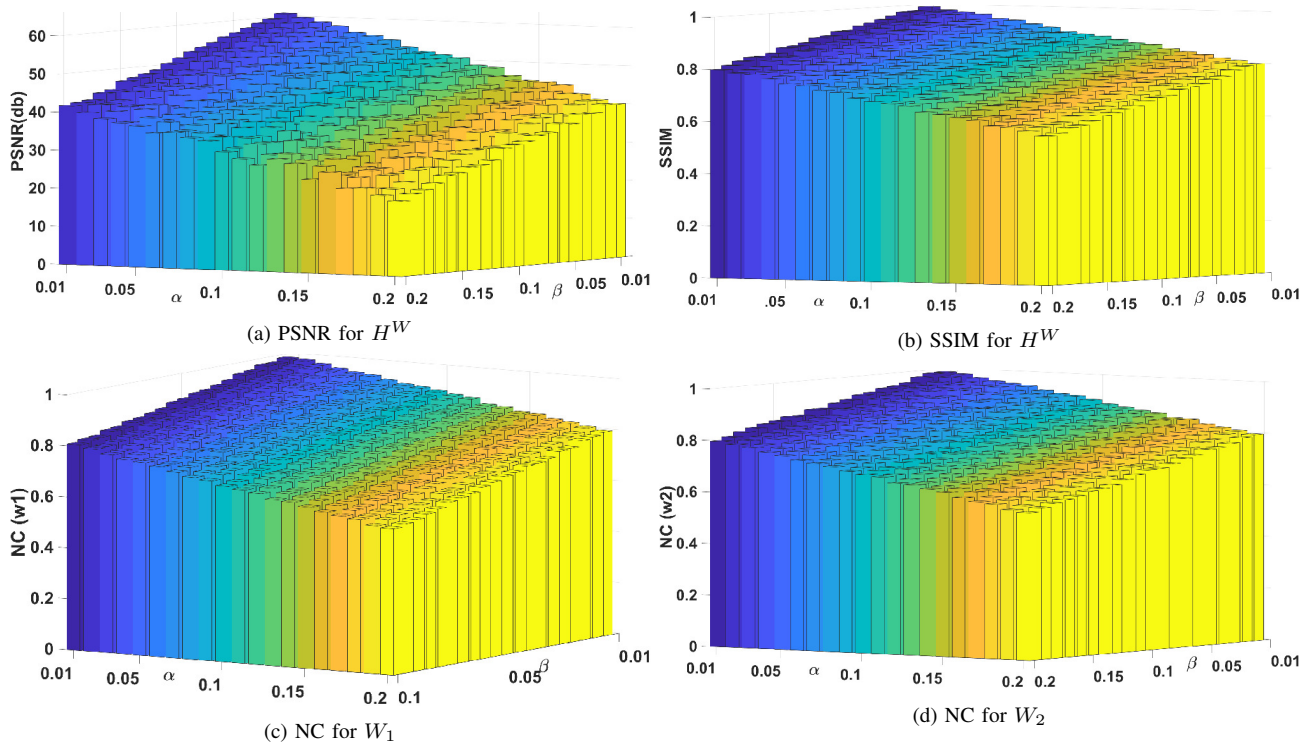
$$NC = \frac{\sum_{(i,j)}(W_{(i,j)} - \mu_W)(W'_{(i,j)} - \mu_{W'})}{\sqrt{\sum_{(i,j)}(W_{(i,j)} - \mu_W)^2}\sqrt{\sum_{(i,j)}(W'_{(i,j)} - \mu_{W'})^2}} \tag{6}$$

and

$$BER = \frac{\sum_x \sum_y |W'_{(i,j)} - W_{(i,j)}|}{512 \times 512}, \tag{7}$$

where, μ_W is average value of the W. As expected, for the case of not undergoing any attack, the NC value is high as reported in Table I (see row NC for W'). The results here are collected using Durand's TMO.

To evaluate robustness, H^{W} is subjected to the mean and median filterings, affine shearing $[1, 0, 0; 0.5, 1, 0; 0, 0, 1]$, additive noise (0.01), rotation of 45^o, cropping with $[75 : 424, 68 : 424]$, and JPEG compression to assess the robustness of our HDR-IW method. Here, all attacks take place in the tone-mapped SDR images produced by using Durand's TMO. The results are recorded in Table II. Note that two cases are considered here, namely, using the proposed SoD-Net approach (NC_{W_1}) and embedding equally throughout the host image (NC_{W_2}). First, results show that the $NC_{W_1} > NC_{W_2}$ for all of the mentioned attacks in Table II. This outcome suggest that embedding W into salient regions of the host image leads to better robustness against attacks. It is because $\{W'_f, W'_b\}$ are embedded (spreaded) into different regions (details) of the H by using different embedding methods. Besides robustness, the SoD based embedding also ensures the imperceptibility feature. Hence, this leads to the concurrent improvement in watermarking conflicting requirement. On the other hand, when embedding W evenly throughout the image



(a) PSNR for H^W

(b) SSIM for H^W

(c) NC for W_1

(d) NC for W_2

Fig. 6: The graphs of image quality and NC for various α and β values. The NC value shown is the average NC for the 7 attacks mentioned in Table II.

H (without SoD), H^W may undergo the attacks evenly and hence attaining the low robustness.

When compared to the SOTA methods, the proposed SoD-Net HDR-IW method achieves the high robustness for four out of seven attacks. It should be noted that, although our method does not achieve the highest NC value for the Mean-filter, Median-filter, and JPEG-Comp attacks, our NC values are only marginally below the best performing methods (i.e., [22], [25], respectively). Note that the BER values are calculated for all attacks mentioned in the Table II, that is 0.09 on average which is significantly better then the SOTA methods [6], [18], [22], [23]. Our HDR-IW performs better against noise attack. This result may have resulted from W being shuffled before embedding into H, which spreads W, as opposed to the typical SOTA approaches that do not encrypt the W prior to embedding. In addition, for the same H, the SOTA methods can embed less than 1 bit/pixel (bpp) (e.g., [18], [25] can only embed 128×128 and 32×32 bits, respectively), while our method can diffuse 256×256, which means 1 bit of W into 1 pixel of H, i.e., 1 bpp.

Interestingly, despite embedding significantly more data, our method produces H^W of the highest quality when compared to the SOTA approaches. Furthermore, we also analyzed the quality of watermarked H^W and the extracted watermark W' with different (α, β) values as shown in Fig. 6. Results suggest that the increment in (α, β) values produces lower quality H^W, which eventually generates low quality (less robust) W', and vice versa. However, for security purposes, we set $\alpha \neq \beta$.

All in all, we conclude that our HDR-IW method outperforms the conventional HDR-IW methods in terms of imperceptibility, robustness, and payload capability, hence achieving simultaneous improvement, viz., trade-off independence.

IV. Conclusion & Future Work

In this study, a novel trade-off independent HDR image watermarking based on saliency detection is proposed, which simultaneously achieves higher robustness, image quality and payload. First, the HDR image is tone-mapped to an SDR image, which is then segmented into the background and foreground regions based on a mask produced by the proposed neural-network based SoD method called SoD-Net. Prior to embedding, bit-plane of the watermark are partitioned into foregrounds and backgrounds using the same mask derived from the host and scrambled using the pseudo-random permutation algorithm. Next, the watermark segments are embedded into the chosen bit-planes of the host image segments using the quantized indexed modulation. Results suggest that our method is resilient to the non-malicious attacks (TMOs) and outperforms SOTA IW methods in terms of image quality, robustness, and payload.

In future, we analyze the effect of different wavelets and decomposition levels on proposed HDR-IW method and investigate the robustness for the different tone mapping operators.

Acknowledgment

This work is supported by Advanced Engineering Platform's Funding (account number AEP-2023-Cluster-04), Monash University Malaysia.

REFERENCES

[1] X. Xue, M. Okuda, and S. Goto, "Bilateral filtering based watermarking for high dynamic range image," in *2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)*. IEEE, 2011, pp. 1–5.

[2] X. Xue, T. Jinno, X. Jin, M. Okuda, and S. Goto, "Watermarking for HDR image robust to tone mapping," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 94, no. 11, pp. 2334–2341, 2011.

[3] V. Solachidis, E. Maiorana, P. Campisi, and F. Banterle, "HDR image watermarking based on bracketing decomposition," in *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–6.

[4] F. Guerrini, M. Okuda, N. Adami, and R. Leonardi, "High dynamic range image watermarking robust against tone-mapping operators," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 283–295, 2011.

[5] M. Liu, Y. Chen, Y. Zhou, Z. Dong, and H. Wu, "Cost assignment based adaptive active forensics scheme for image tampers," *Journal of Electronic Imaging*, vol. 32, no. 3, pp. 033 022–033 022, 2023.

[6] J. Huang, S. Shi, Z. He, and T. Luo, "A novel zero watermarking based on DT-CWT and quaternion for HDR image," *Electronics*, vol. 10, no. 19, p. 2385, 2021.

[7] K. R. Perez-Daniel, F. Garcia-Ugalde, and V. Sanchez, "Watermarking of HDR images in the spatial domain with HVS-imperceptibility," *IEEE Access*, vol. 8, pp. 156 801–156 817, 2020.

[8] E. Maiorana and P. Campisi, "High-capacity watermarking of high dynamic range images," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, pp. 1–15, 2016.

[9] E. Maiorana, V. Solachidis, and P. Campisi, "Robust multi-bit watermarking for HDR images in the radon-DCT domain," in *2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2013, pp. 284–289.

[10] K. Shiju and P. T. Selvi, "Performance analysis of high dynamic range image watermarking based on quantization index modulation," in *2012 International Conference on Power, Signals, Controls and Computation*. IEEE, 2012, pp. 1–5.

[11] F. Autrusseau and D. Goudia, "Non linear hybrid watermarking for high dynamic range images," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 4527–4531.

[12] A. Khan and K. Wong, "High payload watermarking based on enhanced image saliency detection," *Multimedia Tools and Applications*, pp. 1–19, 2022.

[13] E. Maiorana and P. Campisi, "Multi-bit watermarking of high dynamic range images based on perceptual models," *Security and Communication Networks*, vol. 9, no. 8, pp. 705–720, 2016.

[14] V. Solachidis, E. Maiorana, and P. Campisi, "HDR image multi-bit watermarking using bilateral-filtering-based masking," in *Image Processing: Algorithms and Systems XI*, vol. 8655. SPIE, 2013, pp. 29–40.

[15] Y. Bai, G. Jiang, H. Jiang, M. Yu, F. Chen, and Z. Zhu, "Novel robust high dynamic range image watermarking algorithm against tone mapping," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 9, pp. 4389–4411, 2018.

[16] F. Y. Bakhsh and M. E. Moghaddam, "A robust HDR images watermarking method using artificial bee colony algorithm," *Journal of Information Security and Applications*, vol. 41, pp. 12–27, 2018.

[17] T. Luo, G. Jiang, M. Yu, H. Xu, and W. Gao, "Robust high dynamic range color image watermarking method based on feature map extraction," *Signal Processing*, vol. 155, pp. 83–95, 2019.

[18] M. Du, T. Luo, H. Xu, Y. Song, and C. Wang, "Robust HDR video watermarking method based on saliency extraction and T-SVD," *The Visual Computer*, vol. 38, no. 11, pp. 3775–3789, 2022.

[19] G. Anbarjafari and C. Ozcinar, "Imperceptible non-blind watermarking and robustness against tone mapping operation attacks for high dynamic range images," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 24 521–24 535, 2018.

[20] A. Khan, K. Wong, and V. M. Baskaran, "Image watermarking based on saliency detection and multiple transformations," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1514–1518.

[21] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, "Res2Fusion: Infrared and visible image fusion based on dense Res2Net and double nonlocal attention models," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[22] Y. Bai, G. Jiang, Z. Zhu, H. Xu, and Y. Song, "Reversible data hiding scheme for high dynamic range images based on multiple prediction error expansion," *Signal Processing: Image Communication*, vol. 91, p. 116084, 2021.

[23] S. Shi, T. Luo, J. Huang, and M. Du, "A novel HDR image zero-watermarking based on shift-invariant shearlet transform," *Security and Communication Networks*, vol. 2021, 2021.

[24] S. K. Singh and R. Srivastava, "A robust salient object detection using edge enhanced global topographical saliency," *Multimedia Tools and Applications*, pp. 1–18, 2020.

[25] X. Gao, Z. Pan, E. Gao, and G. Fan, "Reversible data hiding for high dynamic range images using two-dimensional prediction-error histogram of the second time prediction," *Signal Processing*, vol. 173, p. 107579, 2020.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Quick Response (QR) codes embedding in VVC using Quantisation Parameter Manipulation

Hau Joan*, Li Peng Tan† and Yiqi Tew‡
Faculty of Computing & Information Technology
Tunku Abdul Rahman University of Management and Technology
Kuala Lumpur, Malaysia
Email: {*joanh-wm18@student., †tanlp@, ‡yiqi@}tarc.edu.my

***Abstract*—This study presents an innovative approach for concealing textual content within Versatile Video Coding (VVC), H.266 by utilising the Quantisation Parameter (QP) and modifying the QP index within the existing quantisation framework. The primary objective is to seamlessly integrate concealed textual content into video streams while minimising any noticeable impact on visual fidelity and compression efficiency. Additionally, the study investigates the capability of the proposed method to embed Quick Response (QR) code into the video to enhance interactivity and provide viewers with quick access to supplementary information. The proposed method focuses on the encoder aspect, precisely adjusting the QP index to incorporate the concealed data. By strategically manipulating the index values, the encoded video frames maintain their visual fidelity while accommodating the concealed textual content. The embedding process seamlessly integrates into the pre-existing quantisation framework, providing an effective and efficient solution for hiding information in video streams.**

I. Introduction

In today's modern culture, videos have become an integral part of daily life, serving as a powerful means of communication and information dissemination across various platforms, such as social media, education, and entertainment industries [1]. The power of videos lies in their ability to captivate audiences and present information engagingly, thus encouraging participation and enhancing knowledge transfer. With the widespread accessibility of high-speed internet and mobile devices, videos have become an increasingly effective medium for spreading awareness, educating on critical topics, promoting products and services, and fostering customer loyalty and interaction.

As digital communication advances, the importance of information security has reached unprecedented levels [2]. Within this context, the utilisation of QR code has gained significant prominence. QR code serve as a versatile tool with diverse applications in our daily lives. They enable contactless payments, provide quick access to product information and promotions, facilitate ticketing and boarding processes, enhance authentication and security measures, and empower digital marketing campaigns [1]. The ability to store and convey information in a compact and machine-readable format makes QR code indispensable to our modern society.

Meanwhile, embedding information within videos has emerged as a critical method for delivering various video services, including authentication, copyright protection, and the concealment of sensitive data [3], [4]. By seamlessly integrating QR code into videos, new opportunities arise for utilising these codes to enhance interactivity and engagement. The integrated framework for encryption and decryption of digital data, including 21 x 21 QR code and text, is depicted in Figure 1. During the embedding process, it may be necessary to assign a secret key to ensure that only authorised recipients can access the concealed data. All digital data is transformed into binary form to facilitate the encryption process. QR code embedded in videos can provide viewers instantaneous access to additional content, resources, or promotions, enriching the video experience and facilitating knowledge dissemination.

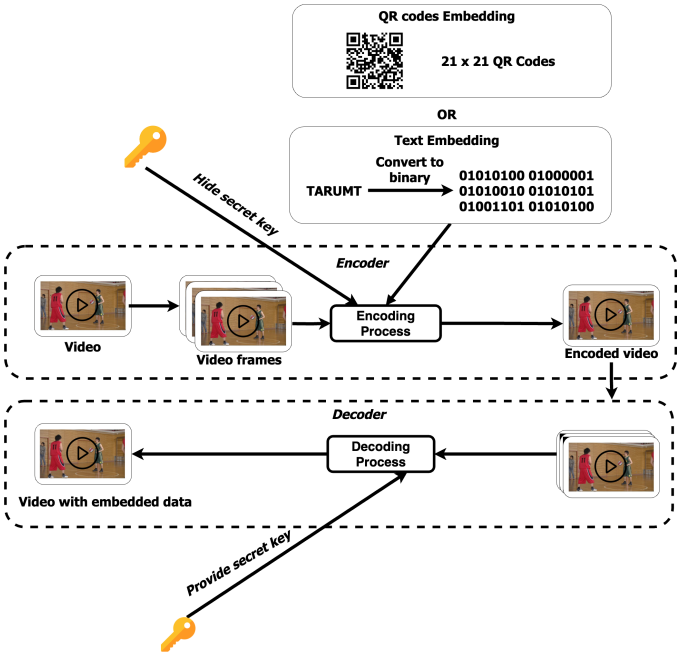


Fig. 1. An integrated framework for encryption and decryption of digital data

However, alongside the advancements in video technology, ensuring the integrity and authenticity of video content has become increasingly paramount. The rise of deepfake technology [5], [6] has underscored the importance of secure and responsible video content manipulation. has underscored the importance of secure and responsible video content ma-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

nipulation. Deepfakes, which employ sophisticated machine learning algorithms to replace or superimpose one person's likeness onto another's in video or audio, have raised concerns regarding the potential misuse of such technology for deceptive or malicious purposes. While our primary focus in this research is on concealing textual content and the potential to hide QR code within videos using QP manipulation, we recognise the broader context of deepfake-related challenges. By developing techniques for information hiding within video streams, our work aligns with the broader objective of ensuring the integrity and authenticity of multimedia content in an era where deepfakes pose significant threats to trust and credibility. Therefore, the ability to hide and reveal information within videos, such as through QR code, becomes a valuable means of securing the trustworthiness of video content.

The embedding of QR code within videos brings forth a range of usability benefits [7], [8], [9], enhancing user engagement and interaction. QR code offer an interactive element, allowing viewers to access additional information or resources by scanning the embedded QR code. This seamless integration ensures that the QR code does not detract from the video content but enriches it by providing supplementary information or functionalities. Viewers can easily access detailed product information, relevant web pages, or social media profiles by scanning the QR code, facilitating efficient information retrieval. Moreover, embedded QR code are a powerful marketing tool, enabling businesses to promote products, services, or special offers, fostering customer engagement and driving conversions. By monitoring QR code scans and engagements, content creators and marketers can gather valuable analytics data, gaining insights into viewer preferences and campaign effectiveness.

Furthermore, QR code provide contextual information within videos, enhancing the viewing experience by offering translations, definitions, or supplementary resources. Additionally, incorporating QR code in videos allows for gamified experiences, where viewers can scan codes to unlock rewards, access hidden content, or participate in interactive challenges. Overall, the usability of QR code embedded in videos extends beyond their conventional applications, offering enhanced interactivity, easy access to information, marketing opportunities, tracking capabilities, contextual information, and gamified experiences. Leveraging the power of QR code in videos elevates the viewer experience, drives engagement, and unlocks new dimensions of user interaction and involvement.

Within the realm of VVC, this paper introduces a novel technique for embedding QR code into video using QP Manipulation. The proposed method modifies the QP index within the quantisation layer, specifically in the quantisation framework. The primary objective is to seamlessly integrate hidden QR code into video frames while minimising any perceptible impact on visual quality and compression efficiency. By addressing the demand for efficient and imperceptible QR code embedding in VVC, this research significantly contributes to the advancement of video coding and information security.

Here is an outline of the paper's structure: Section II offers an exhaustive review of pertinent literature in video coding and information hiding. It delves into the fundamental concepts and techniques that form the foundation of our approach, with a particular focus on manipulating the QP index and its seamless integration into the existing quantisation framework. Section III presents a comprehensive exposition of our proposed compressed video information hiding algorithm. Section IV presents our experimental results, which vividly demonstrate the effectiveness of our approach concerning both visual fidelity and compression efficiency. In conclusion, we summarise our findings and discuss the broader implications of our work, highlighting its significance in interactive video content and information dissemination, and suggest avenues for future research.

II. VVC Video Encoder Architecture and Functionality

VVC, also known as H.266 [10], [11], is the latest international video coding standard introduced by the Joint Video Experts Team (JVET). Building upon its predecessor, High-Efficiency Video Coding (HEVC) [12], VVC aims to improve video compression efficiency, enabling higher resolutions significantly, enhanced visual quality, and improved streaming capabilities. During the general compression process, before moving on to the first stage of the video encoder, the input video is often pre-processed to reduce noise and artefacts and to improve the video quality. Following that, each video frame is split up into several coding units (CUs) [13], which are blocks of varying sizes and shapes that depend on the complexity and spatial properties of the image. The next step is to do intra [14] and inter-prediction [15], in which each CU's pixels are predicted based on those of its neighbours. The prediction error (or residual), which would otherwise be redundant, is then encoded using a transform [16], [17] (such as the Discrete Cosine Transform (DCT) or Discrete Sine Transform (DST)).

Additionally, motion estimation and compensation may lower the prediction error for frames with a strong temporal association with their preceding frames. The motion vector and residual are then encoded due to the encoder's search for the best match between the current block and a block in the reference frame [18]. Each block will go through transformation and quantisation [19], where the transform coefficients are reduced in bit depth and enhanced for compression. Depending on the frame's content, the quantisation step size may be changed to achieve the desired trade-off between bit rate and video quality. The quantised coefficients and other syntactic components (such as motion vectors and intra/inter coding modes) are further reduced using context-adaptive binary arithmetic coding (CABAC) to exploit their statistical properties.

A. Quantisation

Quantisation [20], [21] is a fundamental process that determines the level of compression and quality achieved in

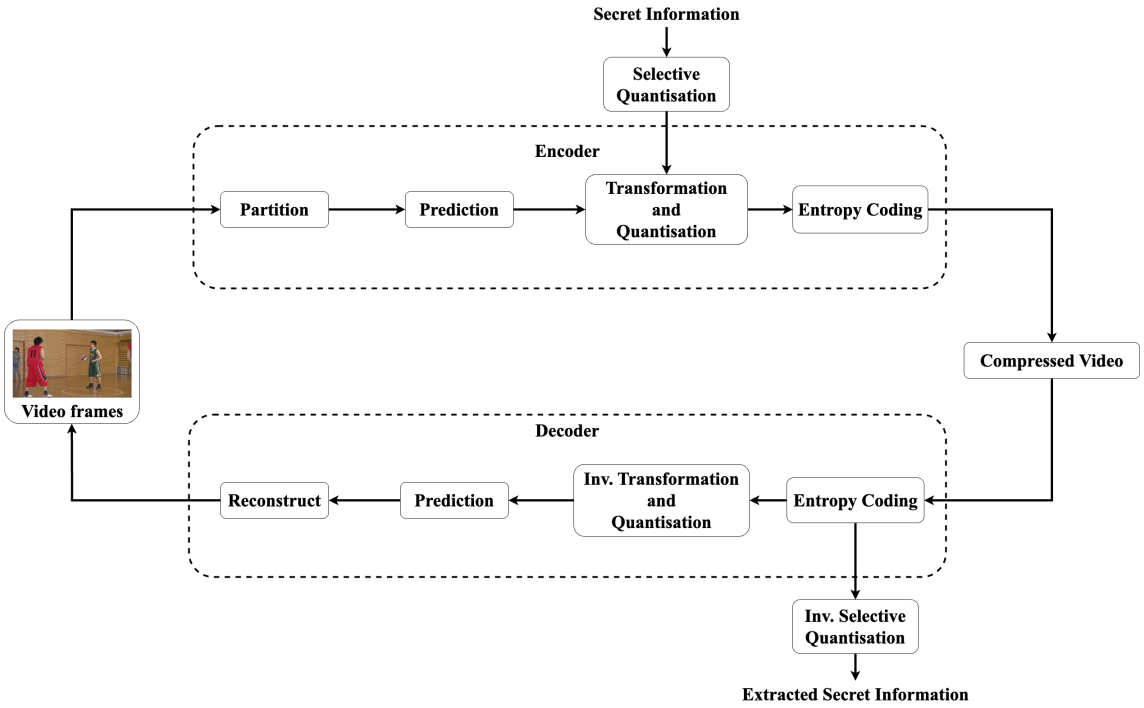


Fig. 2. Selective Quantisation-based Information Hiding Technique in Versatile Video Coding.

the encoded video. The goal of quantisation is to reduce the precision of video data, allowing for more efficient storage and transmission. VVC employs a flexible and adaptive quantisation scheme to improve compression efficiency and video quality compared to previous video coding standards. The quantisation process in VVC involves the following steps:

1) Transform Coding [17]
   First, the input video frame is divided into blocks or transform units (TUs), such as square blocks of 4x4, 8x8, or 16x16 pixels. Transform coding is applied to these TUs to convert them from the spatial domain to the frequency domain using mathematical transformations like the DCT or the Integer Transform.
2) Quantisation [22], [19]
   After the transformation, quantisation is performed on the transformed coefficients. The quantisation process reduces the precision of these coefficients by dividing them by a quantisation step size. The quantisation step size is determined by a QP specified for each TU.
3) Adaptive Quantisation [23]
   VVC introduces adaptive quantisation, where the quantisation step size can vary within a frame or a TU. Adaptive quantisation allows for more precise control over the level of quantisation applied to different parts of the video, enabling better quality preservation in essential regions and more significant compression in less critical areas.
4) Rate Distortion Optimization (RDO) [24]
   A RDO technique is employed during the quantisation process to find the optimal QP for each TU. RDO considers the trade-off between the compression efficiency (bitrate) and the encoded video's visual quality (distortion). It searches for the best QP that minimises the distortion while achieving the desired bitrate.
5) In-loop Filtering [25]
   After quantisation, in-loop filtering techniques are applied to the quantised coefficients to refine the quality further and reduce artefacts caused by quantisation. These filtering operations can alleviate blocking artefacts and enhance the visual quality of the reconstructed video.

VVC allows for a flexible balance between compression efficiency and video quality by controlling the quantisation process. The QP is a critical parameter that determines the strength of quantisation applied to the video, and its adaptive usage in VVC contributes to improved compression efficiency and enhanced video coding performance.

III. Proposed Information Hiding Technique

In the proposed information hiding framework depicted in Figure 2, a selective quantisation-based approach is adopted to conceal data within video frames during encoding. This process occurs precisely during the transformation and quantisation stages, making it an integral part of the video processing pipeline. During the selective quantisation, the algorithm focuses on modifying the QP values of video slices based on the assigned information at the beginning of the information-hiding process. By switching between odd and even QP values according to the detected program bits, the algorithm enables embedding secret information into video frames. The

variations in QP values are utilised to distinguish the embedded data within the video frames. This algorithm presents a robust approach for concealing information while preserving video quality.

A. Quantisation-Based Text Concealment Algorithm

This algorithm 1 aims to integrate concealed text into a video stream while ensuring minimal impact on the visual fidelity. The steps of the algorithm are as follows:

1) If the **enable hide text** (h) flag is true, proceed with the hiding process.
2) Check if the **index** (i) is within the valid range of **length of the binary text** (γ).
3) If the current **binary digit** (β_i) is '1':
   - Make sure the slice's **QP** (α_i) is even. Otherwise, increment it by 1 to ensure it becomes even.
   - Increment the i to move to the next binary digit.
4) If β_i is '0':
   - Make sure the α_i is odd. Otherwise, increment it by 1 to make it odd.
   - Increment the i to move to the next binary digit.
5) If the hiding process reaches the maximum γ:
   - Set the α_i of the video slice to its current α_i.
   - Set the h flag to false to indicate that no text is being hidden.

This approach focuses on modifying the video slice's QP depending on the binary text that has to be concealed. The program ensures the hidden text is merged while preserving visual accuracy by carefully altering the QP values.

B. Quantisation Parameter Analysis Algorithm for Concealed Text Extraction

Algorithm 2 provided above focuses on extracting hidden information from the video. The steps of the algorithm are as follows:

1) If the last eight elements of β_i do not match the **end of the text** (e) condition and the h flag is set to true, the algorithm proceeds with the extraction process.
2) In this process, the algorithm examines the parity of the α_i. If α_i is an even value, it implies that the hidden binary digit appended to β_i is '1'.
3) Conversely, if α_i is an odd value, it signifies that the hidden binary digit appended to β_i is '0'.
4) If the conditions are not met, indicating the completion of the extraction process, the algorithm updates the h flag to false, signifying the termination of any further information extraction.

By analysing the parity of the QP, the algorithm selectively extracts hidden information from β_i while adhering to specific conditions. This approach ensures accurate retrieval of concealed data in a professional and controlled manner.

Both the information concealing and information extraction procedures include a reset mechanism at the end of each separate operation to guarantee adherence to accepted norms.

Algorithm 1 Quantisation-Based Text Concealment Algorithm

```
Initialise x, h
Assign n = x + e
Convert n to β, where γ is the length of β
if h is true then
    if i < γ then
        if β_i = 1, where i ∈ {0, 1, ..., γ − 1} then
            if α_i mod 2 ≠ 0 then
                α_i += 1
                i++
            end if
        else if β_i = 0, where i ∈ {0, 1, ..., γ − 1} then
            if α_i mod 2 = 0 then
                α_i += 1
                i++
            end if
        end if
    end if
else
    α_i = α_i
    h = false
end if
```

Algorithm 2 QP Analysis Algorithm for Concealed Text Extraction

```
Initialise h
if β_[−8:] ≠ e and h is true then
    if α_i%2 == 0, where i ∈ {0, 1, ..., γ − 1} then
        β_i+ = 1
    else if α_i%2 ≠ 0, where i ∈ {0, 1, ..., γ − 1} then
        β_i+ = 0
    end if
else
    h = false
end if
Insert β_i to arr_i
Convert arr to s
Display s
```

This technique works to stop any tweaks or modifications that are unnecessary or without a valid reason. The algorithms guarantee a regulated and dependable execution, safeguarding the integrity of the data while respecting professional standards by reverting to the usual standard procedure after completion.

IV. Performance Evaluation

Video compression techniques play a pivotal role in optimising the storage and transmission of video data by achieving high compression ratios without compromising visual quality. This study evaluates a proposed technique's compression efficiency and performance within VTM 18.1, employing a range of video sequences.

The assessment encompasses two fundamental aspects: compression efficiency and visual fidelity. Compression efficiency is measured through Rate-Distortion (RD) curves, which depict the trade-off between bitrate (in kilobits per second) and Peak Signal-to-Noise Ratio (PSNR) in decibels (dB). Higher PSNR values indicate superior visual quality. Visual fidelity is evaluated using Bjontegaard Delta-Peak Signal-to-Noise Ratio (BD-PSNR) and Bjontegaard Delta-Rate (BD-BR) metrics. The proposed technique is compared against three scenarios: embed words within specific frames (M_1), embed bit '0' in all frames (M_2), and embed bit '1' in all frames (M_3).

The video sequences selected for evaluation include BasketballPass (BP), BasketballDrill (BD), FourPeople (FP), and ParkScene (PS), ensuring a diverse range of content for a comprehensive analysis of the proposed technique's performance across different scenarios. By examining the PSNR values and bitrates, we can assess the effectiveness of the proposed technique in enhancing compression efficiency while preserving visual quality.

The RD curves presented in Figure 3 depict the relationship between PSNR and bitrate for the original video sequences and the proposed technique across various scenarios. These curves illustrate the impact of the proposed technique on compression efficiency and its ability to maintain visual fidelity. Notably, PSNR values across various video sequences and resolutions at each bitrate are similar for the original and proposed technique's video. This can be attributed to the proposed technique's utilisation of a modified QP value, minimising the impact on video quality. The RD curves demonstrate a consistent trend between the original and proposed video when encoded using identical video sequences in different scenarios. The maximum degradation observed between the two versions is less than or equal to 0.05 dB in terms of PSNR, indicating minimal impact on visual quality. The minimum degradation observed among the four encoded videos under different scenarios is less than 0.05 dB.

Table I comprehensively evaluate the PSNR performance, comparing the original VVC standard with the proposed technique implemented in VTM 18.1. The assessment encompasses various bitrates and three distinct scenarios. The analysis of PSNR values allows for an accurate assessment of the proposed technique's effectiveness in enhancing video quality compared to the original standard.

For the BP sequence, at a bitrate of 100 kbps, the proposed technique achieves a higher PSNR of 32.3890 dB than the original standard's PSNR of 32.3793 dB. This trend continues at higher bitrates, demonstrating improved PSNR values for the proposed technique. Specifically, the average PSNR difference amounts to 0.0496 dB in M_1. When considering the M_2, the average PSNR difference is 0.0137 dB, and for the M_3, the average PSNR difference is 0.0161 dB.

Likewise, for the BD sequence, at a bitrate of 100 kbps, the proposed technique achieves a PSNR of 28.9460 dB, while the original standard yields a lower PSNR of 28.8720 dB. The average PSNR difference is 0.0149 dB in M_1. When considering the M_2, the average PSNR difference is 0.0058 dB, and for the M_3, the average PSNR difference is 0.0388 dB.

As we go on to the FP sequence, the video with the proposed technique exhibits the highest quality at 500 kbps in M_2 compared to the original video out of all the scenarios with a difference of 0.0454 dB. However, the maximum degradation of the FP video sequences under all the scenarios is 0.5747 dB, which occurs in M_3 at 100 kbps.

In addition, for the PS video sequence, the PSNR is improved by a maximum of 0.011 dB at 2000 kbps in the M_1 condition, and the maximum degradation of this video from the original video is 0.1742 dB at 100 kbps in the M_2 condition.

The table II showcases the BD-BR and BD-PSNR values obtained for all video sequences, indicating the quality gains and bitrate savings achieved by the proposed video compared to the original video. BD-PSNR quantifies the quality improvement achieved by the proposed video at equivalent bitrates. At the same time, BD-BR reflects the bitrate savings required to maintain the same quality level, as denoted by the BD-Rate value.

For the BP video sequence, the BD-PSNR values indicate the quality improvement achieved by the proposed technique in scenarios M_1, M_2, and M_3. In scenario M_1, the proposed technique yields a BD-PSNR improvement of 0.0414 dB, demonstrating enhanced visual quality. However, it is essential to note that the BD-BR value is negative (-0.7661%), indicating that the proposed technique's video needs fewer bits to achieve the original video quality. In contrast, in scenario M_2, the BD-PSNR improvement is smaller (0.0020 dB), with a negligible BD-BR change (0.0961%). Scenario M_3 exhibits a decrease in visual quality (-0.0485 dB) with a moderate BD-BR reduction (0.5906%). When employing the proposed technique, these results highlight the trade-off between visual quality and bitrate reduction.

For the BD video sequence, similar trends are observed. In scenario M_1, the proposed technique achieves a BD-PSNR improvement of 0.0287 dB, albeit with a slight increase in bitrate (-0.6779%). Scenario M_2 shows a BD-PSNR improvement of 0.0309 dB, accompanied by a BD-BR reduction of -0.7192%. In scenario M_3, the proposed technique demonstrates the highest BD-PSNR improvement of 0.0399 dB, with a notable BD-BR reduction of -0.9439%. These findings emphasise that scenario M_3 provides the best trade-off between visual quality improvement and bitrate reduction for the BD video sequence.

Moving on to the FP video sequence, scenario M_1 shows a marginal decrease in visual quality (-0.0129 dB) with a small BD-BR increase (0.8474%). In scenario M_2, the BD-PSNR change is minimal (0.0020 dB) with a BD-BR reduction of -1.0316%, indicating a balance between quality and bitrate. Scenario M_3 yields a slight BD-PSNR improvement (-0.0094 dB) with a BD-BR increase of 1.6120%. These results suggest that scenario M_2 strikes a reasonable compromise between visual quality and bitrate reduction for the FP video sequence.

Moreover, scenario M_1 exhibits a slight BD-PSNR decrease (-0.0040 dB) with a minor BD-BR increase (0.1489%) for the PS video sequence. In scenario M_2, the BD-PSNR change is -0.0089 dB, and the BD-BR increase is 0.2642%. Scenario M_3 shows a similar BD-PSNR decrease (-0.0088 dB) with a BD-BR increase of 0.2745%. These findings indicate that all three scenarios have a relatively small impact on visual quality and bitrate for the PS video sequence.

Furthermore, the BD-BR values range from -1.0316% to 1.6120%, while the BD-PSNR values span from -0.0485 dB to 0.0399 dB. Notably, a BD-BR value of -1.0316% indicates that the original video requires 1.0316% more bits than the

TABLE I
COMPARISON OF PSNR PERFORMANCE IN VVC: ORIGINAL STANDARD VS. PROPOSED TECHNIQUE IN VTM 18.1

Video Sequences	Bitrate (kbps)	Original (PSNR)	Proposed Technique (PSNR)		
			M_1	M_2	M_3
BP (416 x 240)	100	32.3793	32.3890	32.2780	32.3798
	500	38.5145	38.5721	38.5454	38.4089
	1000	42.8721	42.8662	42.8602	42.8867
	2000	46.6592	46.7237	46.6604	46.7077
	3000	48.7468	48.7964	48.7605	48.7629
BD (832 x 480)	100	28.8720	28.9460	28.8871	28.7883
	500	34.6994	34.6866	34.7049	34.7311
	1000	37.3463	37.3538	37.3685	37.4070
	2000	40.0756	40.1195	40.0807	40.3391
	3000	42.3053	42.2979	42.3016	42.3179
	4000	43.3967	43.3939	43.3917	43.4010
	5000	44.1699	44.1719	44.1710	44.1527
FP (1280 x 720)	100	33.1226	32.7599	32.9430	32.5479
	500	40.1076	40.1127	40.1530	40.1323
	1000	42.2928	42.2912	42.3129	42.2974
	2000	44.0585	44.0106	43.9908	44.0690
	3000	44.7891	44.7554	44.7405	44.7750
	4000	45.2169	45.1590	45.2000	45.2029
	5000	45.5294	45.5045	45.5203	45.5061
PS (1920 x 1080)	100	29.5055	29.5111	29.3313	29.4069
	500	33.5555	33.5346	33.5408	33.5434
	1000	35.5030	35.5021	35.5123	35.4966
	2000	37.3981	37.4091	37.3993	37.3885
	3000	38.4732	38.4796	38.4608	38.4761
	4000	39.3680	39.3512	39.3593	39.3375
	5000	40.1047	40.0986	40.1044	40.1019

TABLE II
EVALUATION OF COMPRESSION EFFICIENCY IN VTM 18.1 USING PROPOSED TECHNIQUE

Video Sequences	M_1		M_2		M_3	
	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)	BD-PSNR (dB)	BD-BR (%)
BP (416 x 240)	0.0414	-0.7661	0.0020	0.0961	-0.0485	0.5906
BD (832 x 480)	0.0287	-0.6779	0.0309	-0.7192	0.0399	-0.9439
FP (1280 x 720)	-0.0129	0.8474	0.0020	-1.0316	-0.0094	1.6120
PS (1920 x 1080)	-0.0040	0.1489	-0.0089	0.2642	-0.0088	0.2745

proposed video to achieve an equivalent quality level. Additionally, the proposed video showcases a maximum quality improvement of 0.0399 dB compared to the original video at equivalent bitrates.

Overall, the experimental results demonstrated the proposed technique's effectiveness in achieving high compression efficiency while maintaining acceptable visual quality. The RD curves illustrated the trade-off between bitrate and PSNR, indicating the impact of the proposed technique on compression efficiency and its ability to preserve visual fidelity. Comparing the proposed technique to the original video, the average differences in PSNR were relatively small, indicating that the proposed technique minimised the impact on video quality. Moreover, the evaluation of BD-PSNR and BD-BR values further emphasised the improvements achieved by the proposed video in terms of both quality gains and bitrate savings compared to the original video. Across different video sequences and scenarios, the proposed technique consistently outperformed the original standard, yielding higher PSNR values at equivalent bitrates. The experimental results showcased the potential of the proposed technique in enhancing video quality and compression efficiency across a range of content.

The experimental findings affirm the efficacy of the proposed techniques for concealing QR codes within video se-

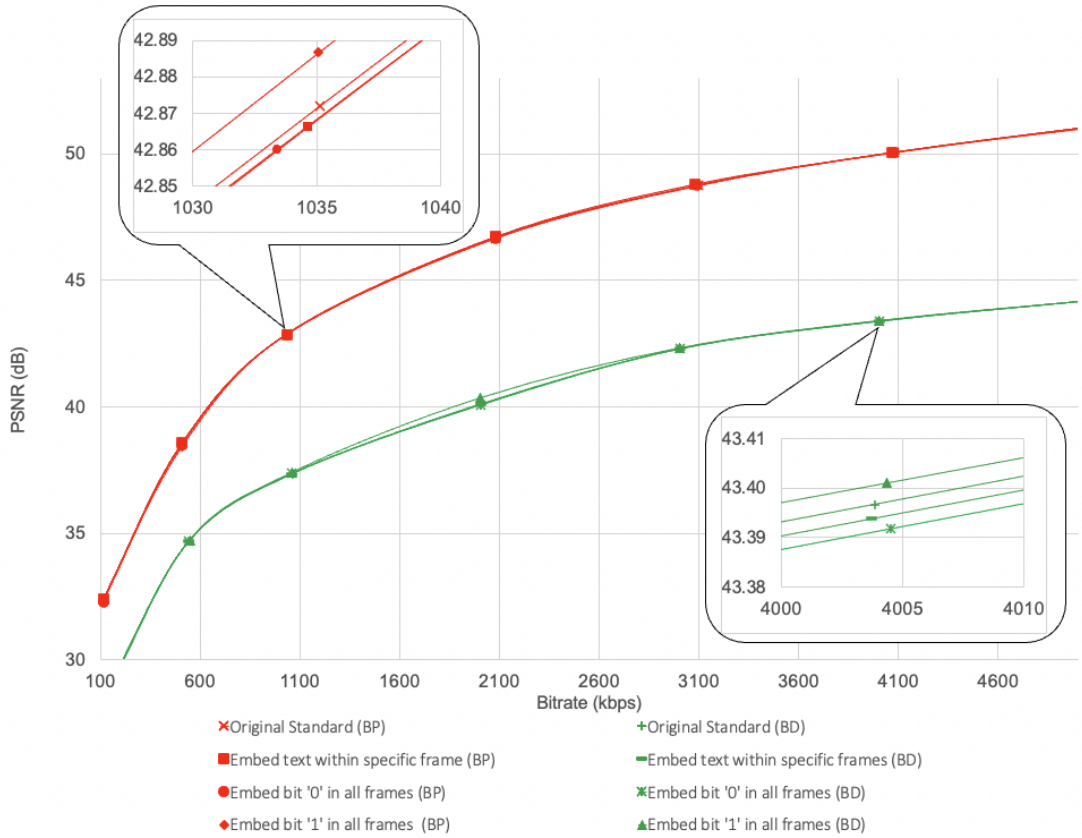
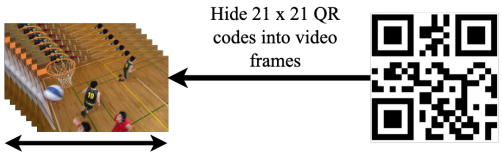


Fig. 3. RD Analysis for BP and BD Video Sequence: Original vs. Proposed Technique in Various Scenarios

TABLE III
QR DATA PAYLOAD FOR DIFFERENT QR VERSIONS AND ERROR CORRECTION LEVELS [26]

QR Version	Number of QR data (bit/modules)			
	L	M	Q	H
1	152	128	104	72
5	864	688	496	368
10	2,192	1,728	1,232	976
15	4,184	3,320	2,360	1,784
20	6,888	5,352	3,880	3,080
25	10,208	8,000	5,744	4,304
30	13,880	10,984	7,880	5,960
35	18,448	14,496	10,288	7,888
40	23,648	18,672	13,328	10,208

quences. Among the various versions of QR code as shown in Table III, the embedding process primarily focuses on QR Code Version 1, which features a matrix grid size of 21 x 21 modules comprising black and white squares. This version boasts a data capacity ranging from 72 (with a high Error Correction (ECC) Level) to 152 (with a low ECC level) bits. In Figure 4, a 21 x 21 QR code is embedded within 68 video frames, each running at 24 frames per second (fps). The content, 'TARUMTMYKL', is stored in the QR code with 68 bits. High-level error correction has been employed, allowing



Content	TARUMTMYKL
Error Correction Level	30%
Version	1
Size	21 x 21
Data Bits	68 bits

Fig. 4. Procedure for Integrating QR code into Video Frames using VVC

recovery of up to 30% of data in the event of errors. According to the proposed techniques present in this paper, we can store the bits in the video depending on how many video frames are in the sequences. Therefore, it can be concluded that, to hide a 21 x 21 QR code with a high ECC level (72 bits), 72 frames of video sequences are sufficient.

The evaluation of compression efficiency and visual fidelity, including the analysis of RD curves and BD-PSNR/BD-BR values, serves as a basis for assessing the feasibility of embedding QR codes in video. The results obtained from this study demonstrate that the proposed technique effectively maintains video quality while achieving high compression

efficiency. This indicates that the technique can be adapted to seamlessly embed QR codes within video frames, allowing for their hidden presence without compromising visual quality. In summary, the concepts and techniques presented in this study hold promise for integrating QR codes within video sequences, thereby offering new avenues for information dissemination, interactive media experiences, and enhanced video content.

V. Conclusion

In conclusion, this study evaluated the compression efficiency and visual fidelity of a proposed video compression technique implemented in VTM 18.1 in VVC. The experimental results demonstrated that the proposed method achieved successful information hiding and showed the capability to embed QR code with minimal impact on video quality. However, it is essential to acknowledge a limitation of the proposed technique. It is currently limited in its ability to hide information, with the total number of bits hidden dependent on the video frames. To address this limitation and further enhance the technique's versatility, future work should explore more sophisticated approaches and further investigate the security of the information hiding technique proposed. This could involve investigating the feasibility of embedding more complex QR code with larger data payloads. This direction of research aims to overcome the current constraint and unlock new possibilities for information hiding within video sequences.

References

[1] Falon Fatemi, "How the pandemic has changed video content and consumption," February 2021.

[2] Wisam Abed Shukur, Wathiq Najah Abdullah, and Luheb Kareem Qurban, "Information hiding in digital video using dct, dwt and cvt," *Journal of Physics: Conference Series*, vol. 1003, no. 1, pp. 012035, may 2018.

[3] Yiqi Tew and KokSheik Wong, "An overview of information hiding in h.264/avc compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 305–319, 2014.

[4] Mukesh Dalal and Mamta Juneja, "A survey on information hiding using video steganography," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5831–5895, 2021.

[5] Yisroel Mirsky and Wenke Lee, "The creation and detection of deepfakes: A survey," Sep 13 2020, Copyright - © 2020. This work is published under http://arxiv.org/licenses/nonexclusive-distrib/1.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2020-09-24.

[6] Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, and Ahmad Neyaz Khan, "Deepfake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.

[7] Tina Zurbi and Diana Gregor-Svetec, "Use of qr code in dairy sector in slovenia," *SAGE open*, vol. 13, 2023.

[8] Shintaro Okazaki, Hairong Li, and Morikazu Hirose, "Benchmarking the use of qr code in mobile promotion: Three studies in japan," *Journal of advertising research*, vol. 52, no. 1, pp. 102–117, 2012.

[9] Eunji Lee and Jikyung Jang, "Research trend analysis for sustainable qr code use - focus on big data analysis," *KSII transactions on Internet and information systems*, vol. 15, no. 9, pp. 3221–3242, 2021.

[10] ITU-T, "Versatile video coding," https://www.itu.int/rec/T-REC-H.266, 2022.

[11] Tung Nguyen, Xiaozhong Xu, Felix Henry, Ru-Ling Liao, Mohammed Golam Sarwer, Marta Karczewicz, Yung-Hsuan Chao, Jizheng Xu, Shan Liu, Detlev Marpe, and Gary J. Sullivan, "Overview of the screen content support in vvc: Applications, coding tools, and performance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3801–3817, 2021.

[12] ITU-T, "High efficiency video coding," https://www.itu.int/rec/T-REC-H.265, 2021.

[13] Yu-Wen Huang, Jicheng An, Han Huang, Xiang Li, Shih-Ta Hsiang, Kai Zhang, Han Gao, Jackie Ma, and Olena Chubach, "Block partitioning structure in the vvc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3818–3833, 2021.

[14] Jonathan Pfaff, Alexey Filippov, Shan Liu, Xin Zhao, Jianle Chen, Santiago De-Luxán-Hernández, Thomas Wiegand, Vasily Rufitskiy, Adarsh Krishnan Ramasubramonian, and Geert Van der Auwera, "Intra prediction and mode coding in vvc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3834–3847, 2021.

[15] Shiba Kuanar, Dwarikanath Mahapatra, Vassilis Athitsos, and K. R Rao, "Gated fusion network for sao filter and inter frame prediction in versatile video coding," 2021.

[16] Werda Imen, Belghith Fatma, Maraoui Amna, and Nouri Masmoudi, "Dct -ii transform hardware-based acceleration for vvc standard," in *2021 IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS)*, 2021, pp. 1–5.

[17] Xin Zhao, Seung-Hwan Kim, Yin Zhao, Hilmi E. Egilmez, Moonmo Koo, Shan Liu, Jani Lainema, and Marta Karczewicz, "Transform coding in the vvc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3878–3890, 2021.

[18] Thomas Wiegand and Heiko Schwarz, "Video coding: Part ii of fundamentals of source and video coding," *Found. Trends Signal Process.*, vol. 10, no. 1–3, pp. 1–346, dec 2016.

[19] Heiko Schwarz, Muhammed Coban, Marta Karczewicz, Tzu-Der Chuang, Frank Bossen, Alexander Alshin, Jani Lainema, Christian R. Helmrich, and Thomas Wiegand, "Quantization and entropy coding in the versatile video coding (vvc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3891–3906, 2021.

[20] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Inf. Theor.*, vol. 35, no. 5, pp. 1020–1033, sep 2006.

[21] Benjamin Bross, Jianle Chen, Jens-Rainer Ohm, Gary J. Sullivan, and Ye-Kui Wang, "Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

[22] Gary J. Sullivan, "Efficient scalar quantization of exponential and laplacian random variables," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1365–1374, 1996.

[23] Guoqing Xiang, Huizhu Jia, Mingyuan Yang, Yuan Li, and Xiaodong Xie, "A novel adaptive quantization method for video coding," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14817–14840, June 2018.

[24] G.J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[25] Ye-Kui Wang, Robert Skupin, Miska M. Hannuksela, Sachin Deshpande, Hendry, Virginie Drugeon, Rickard Sjöberg, Byeongdoo Choi, Vadim Seregin, Yago Sanchez, Jill M. Boyce, Wade Wan, and Gary J. Sullivan, "The high-level syntax of the versatile video coding (vvc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3779–3800, 2021.

[26] Pei-Yu Lin and Yi-Hui Chen, "High payload secret hiding technology for qr codes," *EURASIP Journal on Image and Video Processing*, vol. 2017, 02 2017.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

CPIPS: Learning to Preserve Perceptual Distances in End-to-End Image Compression

Chen-Hsiu Huang and Ja-Ling Wu
CMLab, CSIE, National Taiwan University, Taiwan
E-mail: {chenhsiu48,wjl}@cmlab.csie.ntu.edu.tw

***Abstract*—Lossy image coding standards such as JPEG and MPEG have successfully achieved high compression rates for human consumption of multimedia data. However, with the increasing prevalence of IoT devices, drones, and self-driving cars, machines rather than humans are processing a greater portion of captured visual content. Consequently, it is crucial to pursue an efficient compressed representation that caters not only to human vision but also to image processing and machine vision tasks. Drawing inspiration from the efficient coding hypothesis in biological systems and the modeling of the sensory cortex in neural science, we repurpose the compressed latent representation to prioritize semantic relevance while preserving perceptual distance. Our proposed method, Compressed Perceptual Image Patch Similarity (CPIPS), can be derived at a minimal cost from a learned neural codec and computed significantly faster than DNN-based perceptual metrics such as LPIPS and DISTS.**

***Index Terms*—End-to-end learned compression, image quality assessment, perceptual distance, coding for machines.**

I. Introduction

The concept of *efficient coding* [1], [2] in early biological sensory processing systems hypothesized that the internal representation of images in the human visual system is optimized to encode the visual information it processes efficiently. In other words, the brain effectively compresses visual information.

The field of neural science has made discoveries regarding modeling neural single-unit and population responses in higher visual cortical areas using goal-driven hierarchical convolutional neural networks (HCNNs) [3]. The sensory cortex's fundamental framework models the visual system through encoding, the process by which stimuli are transformed into patterns of neural activity, and decoding, the process by which neural activity generates behavior. In their work [3], HCNNs have successfully described the mapping of stimuli to measured neural responses in the brain.

In recent years, the rapid advancement of deep neural network techniques has significantly improved computer vision tasks [4]–[6] and image processing tasks [7]–[9]. Neural compression [10], an end-to-end learned image compression method [11]–[17], has also gained significant attention and has been shown to outperform traditional expert-designed image codecs. Traditionally, most image processing algorithms cannot be directly applied to hand-crafted image codecs like JPEG [18]. As a result, the first step before further image processing or analysis is typically decompressing the image into raw pixels. With the evolution of neural compression, there is a growing trend to apply CNN-based methods directly to the compressed latent space [19]–[22], leveraging the advantages of joint compression-accuracy optimization [21] and eliminating the need for decompression. Consequently, international standards such as JPEG AI [23] and MPEG VCM (Video Coding for Machines) [24] have been initiated to bridge data compression and computer vision, catering to both human and machine vision needs.

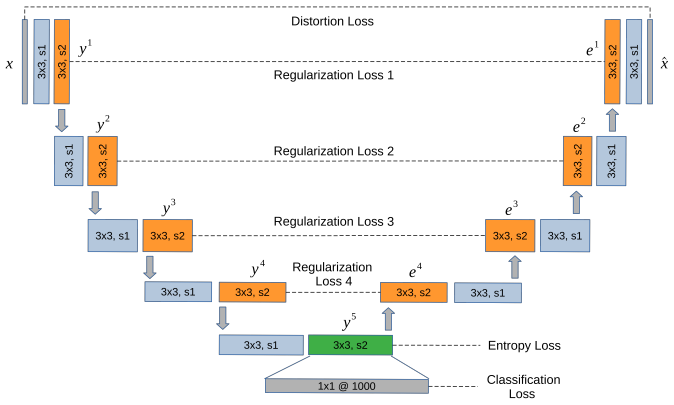


Fig. 1. The architecture we proposed for conducting perceptual distance preserving image compression. The innermost green convolution output y^5 represents the compressed latents to be further entropy-coded. The orange layer outputs y^l and e^l, trained with the image classification task, contain semantic features that preserve perceptual differences in human vision.

Drawing inspiration from sensory cortex modeling [3] and the efficient coding hypothesis employed in information-theoretic perceptual quality metrics [25], we aim to develop an end-to-end learned image compression method jointly trained with the ImageNet classification task as the goal-driven HCNN. Fig. 1 illustrates our proposed architecture, which resembles a UNet network. The compressed latent representations and the intermediate decoder output layers are mapped to a semantic space that preserves the perceptual distance between two different images. We name our method the *Compressed* Perceptual Image Patch Similarity (CPIPS), which utilizes the entropy-coded bitstream and intermediate decoder output to measure the perceptual distance between

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

images. In the context of Coding for Machines, the compressed image bitstream transmitted by IoT devices can be readily utilized by machines to assess perceptual distortions resulting from image operations.

Our contributions can be summarized as follows:

- We demonstrate the utilization of a goal-driven HCNN as an auxiliary task to map the latent space of the end-to-end learned image compression method to a space with semantic meaning.
- We provide guidance and insights on designing the network architecture when a high-level computer vision task is jointly trained with a variational autoencoder network.
- The proposed perceptual metric, CPIPS, is lightweight compared to other CNN-based perceptual metrics, such as LPIPS [26] and DISTS [27]. Computing CPIPS is significantly faster than LPIPS, with an acceleration of approximately 50 times.

II. Related Works

A. Learned Image Compression

The field of learned image compression has witnessed significant advancements with the introduction of convolutional neural networks. Several approaches have been proposed in the literature, starting with Ballé et al. [12] that surpassed traditional codecs like JPEG [18] and JPEG 2000 [28] in terms of PSNR and SSIM metrics. Minnen et al. [13] further improved coding efficiency by employing a joint autoregressive and hierarchical prior model, surpassing the performance of the HEVC [29] codec. More recently, Cheng et al. [15] developed techniques that achieved comparable performance to the latest coding standard VVC [30]. Several comprehensive survey and introduction papers [10], [31], [32] have summarized these advancements in end-to-end learned compression.

Currently, there are two remaining challenges [33] in this field: computational complexity and subjective image quality. The neural compressor employs high-capacity networks to end-to-end model data dependency in exchange for better bitrate-distortion (BD) efficiency. The channel-conditional method proposed by Minnen et al. [14] achieves performance close to VVC but at the cost of high computational complexity (600K FLOPS/pixel). Regarding image quality, Valenzise et al. [34] conducted subjective tests on DNN-based methods and observed that these methods produce artifacts that are difficult to evaluate using traditional metrics like PSNR. They concluded that PSNR is inadequate for evaluating DNN-based methods. Upenik et al. [35] benchmarked a set of DNN-based image codecs using a crowdsourcing-based subjective quality evaluation procedure with Differential Mean Opinion Scores (DMOS). Their results demonstrate that learning-based approaches can achieve promising bitrate-DMOS performance compared to HEVC. However, despite their superior subjective scores, these DNN-based image codecs are optimized with pixel difference-based distortion functions.

B. Perceptual Quality Metrics

The evaluation of image codec quality traditionally relies on full-reference image quality assessment (FR-IQA) metrics, which measure the similarity between the reconstructed image and the original image as perceived by human observers. In addition, to mean square error (MSE) or PSNR, various FR-IQA metrics, such as SSIM variants [36], [37], PIM [25], and DISTS [27], have been proposed to predict subjective image quality judgments. Johnson et al. [38] proposed using the feature vector distance from the VGG network [39] as a perceptual loss for image transformation tasks based on the hypothesis that the same image features used for image classification are also helpful for other tasks.

Zhang et al. [26] introduced the BAPPS dataset, which includes a large-scale collection of human judgments on image pairs, and trained the Learned Perceptual Image Patch Similarity (LPIPS) metric. LPIPS was found to be more aligned with human judgments than traditional quality metrics such as L2, PSNR, and SSIM. Ding et al. [40] conducted an interesting study to evaluate whether DNN-based quality metrics can be used as objectives for optimizing image processing algorithms. Developing effective perceptual quality metrics for image tasks remains a challenging problem.

C. Coding for Machines

Lossy image coding standards such as JPEG and MPEG have primarily focused on achieving high compression rates for human consumption of multimedia data. However, with the rise of IoT devices, drones, and self-driving cars, there is a growing need for efficient compressed representations that cater not only to human vision but also to image processing and machine vision tasks. Techniques such as image data hiding [19], image denoising [20], and image super-resolution [41] have been developed to operate directly on neural compressed latent spaces.

Le et al. [21] proposed an inference-time content-adaptive fine-tuning scheme that optimizes the latent representation to improve compression efficiency for machine consumption. Duan et al. [22] employed transfer learning to perform semantic inference directly from quantized latent features in the deep compressed domain without pixel reconstruction. Choi et al. [42] introduced scalable image coding frameworks based on well-developed neural compressors, achieving up to 80% bitrate savings for machine vision tasks.

III. Proposed Methods

To enable joint training of the image compression network and an image classification task, one has to design a suitable network architecture that can be shared between a variational encoder network $\mathcal{G}_e$ and a DNN feature extraction network $\mathcal{F}$. We leverage the successful UNet [5] and VGG [39] networks and propose a Left-UNet. Our Left-UNet consists of $L = 5$ downsampling convolution layers, each with two convolution blocks. As shown in Fig. 1, the first orange block from the top-left represents the intermediate encoder output feature y^1 from the second convolution block of the first layer, denoted

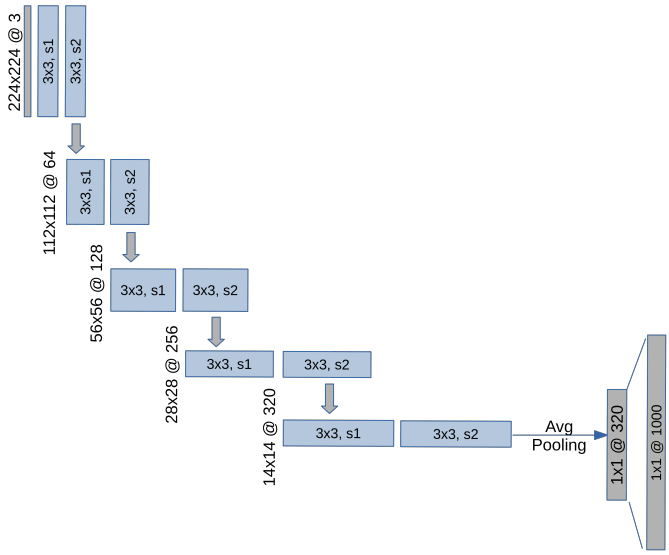


Fig. 2. The Left-UNet feature extraction network $\mathcal{F}$ and the classifier we proposed for the image classification task.

as **conv_1_2**. In Fig. 1, the innermost latent vector y^5, colored green, is outputted from **conv_5_2**. This vector is subject to quantization, resulting in an approximation $\hat{y}^5 = \mathsf{Q}(y^5)$, which is then entropy-coded.

A. Image Classification

We illustrate the Left-UNet architecture in Fig. 2 and provide details in Table I. The feature extraction network $\mathcal{F}$ for the image classification task uses the parameterized ReLU as the activation function for all layers, while the encoder network $\mathcal{G}_e$ employs Generalized Divisive Normalization (GDN) at the end of each downsampling layer. GDN, proposed by Ballé et al. [43], is inspired by modeling neurons in biological visual systems and has been proven effective in Gaussianizing image densities for a superior rate-distortion trade-off.

The extracted image features are then average pooled and connected to a linear layer with 1,000 neurons to optimize the classification loss $\mathcal{L}_C$ using cross-entropy:

$$\mathcal{L}_C = -\sum_i t_i \log(\mathcal{F}(x)_i) \tag{1}$$

It is known that a high-capacity neural network trained for a high-level vision task implicitly learns to reason about relevant semantics [38]. Our goal is not to solve the classification problem directly. Instead, we aim to design a moderately-sized network that can learn semantic features without significantly increasing the encoder-decoder complexity.

B. Image Compression Network

A typical learned neural codec consists of an encoder-decoder pair, a quantization module, and an entropy coder. Given an input image $x \in \mathcal{X}$, the neural encoder $\mathcal{G}_e$ transforms x into a latent representation $y = \mathcal{G}_e(x)$, which is later quantized to a discrete-valued vector $\hat{y}$. The discrete probability distribution $P_{\hat{y}}$ is estimated using a neural network and then encoded into a bitstream using an entropy coder. The *rate* of this discrete code, R, is lower-bounded by the entropy of the discrete probability distribution $H(P_{\hat{y}})$. On the decoder side, we decode $\hat{y}$ from the bitstream and reconstruct the image $\hat{x} = \mathcal{G}_d(\hat{y})$ using the neural decoder. The *distortion*, D, is measured by a perceptual metric $d(x, \hat{x})$. Overall, we optimize the network parameters for a weighted sum of the rate and distortion, $\mathsf{R} + \lambda\mathsf{D}$, over a set of images.

Table II illustrates the decoder network $\mathcal{G}_d$, which is designed to complement the encoder. In the generic neural codec concept, the innermost latent vector $\hat{y}^5$ is equivalent to the discrete-valued vector $\hat{y}$. During image reconstruction, the intermediate output vectors e^l from each upsampling layer **conv_l_2** play a crucial role because they represent learned multi-scale semantic layers, which are equivalent to the feature layers of a VGG-16 network.

TABLE I
LEFT-UNET ARCHITECTURE FOR $\mathcal{G}_e$ AND $\mathcal{F}$

Layer	Kernel	Stride	In	Out	Output
conv_1_1	3	1	3	32	
PReLU					
conv_1_2	3	2	32	32	
PReLU or GDN					y^1
conv_2_1	3	1	32	64	
PReLU					
conv_2_2	3	2	64	64	
PReLU or GDN					y^2
conv_3_1	3	1	64	128	
PReLU					
conv_3_2	3	2	128	128	
PReLU or GDN					y^3
conv_4_1	3	1	128	256	
PReLU					
conv_4_2	3	2	256	256	
PReLU or GDN					y^4
conv_5_1	3	1	256	320	
PReLU					
conv_5_2	3	2	320	320	y^5

TABLE II
DECODER NETWORK ARCHITECTURE $\mathcal{G}_d$

Layer	Kernel	Stride	In	Out	Output
deconv_5_1	3	2	320	320	
PReLU					
conv_5_2	3	1	320	256	
GDN					e^4
deconv_4_1	3	2	256	256	
PReLU					
conv_4_2	3	1	256	128	
GDN					e^3
deconv_3_1	3	2	128	128	
PReLU					
conv_3_2	3	1	128	64	
GDN					e^2
deconv_2_1	3	2	64	64	
PReLU					
conv_2_2	3	1	64	32	
GDN					e^1
deconv_1_1	3	2	32	32	
PReLU					
conv_1_2	3	1	32	3	$\hat{x}$

Like [12], we employ kernel density estimation with a neural network to obtain the probability distribution $P_{\hat{y}}$. The rate loss R is computed as follows:

$$\mathsf{R} = -\mathbb{E}[\log_2 P_{\hat{y}}] \tag{2}$$

In our experiments, we utilize the MSE as the distortion function. However, alternative quality metrics such as SSIM variants [36], [37] can be employed to fit perceptual quality better. The distortion loss D is defined as:

$$\mathsf{D} = \mathbb{E}[d(x, \hat{x})] = \mathbb{E}||x - \hat{x}||_2^2 \tag{3}$$

C. Joint Compression-Classification Learning

Although the intermediate convolution output features are seldom used in most machine learning tasks, these features, which are tuned to be predictive of essential structures, exhibit a high correlation with human perceptual similarity [26]. However, storing intermediate latent features in the context of data compression becomes impractical if the final bottleneck layer contains sufficient information for the decoder to reconstruct the image. Another approach to mitigate storage waste is to reduce the number of downsampling layers. However, modeling the sensory cortex in the visual system [3] requires at least five layers of feature extraction to generate neural responses, a finding that our experiments validate as well. Consequently, we utilize the intermediate output e^l from the decoder as a proxy for multi-scale semantic features and apply a regularizer to constrain the decoder. Specifically, we employ the l_1 distance to define our regularization loss:

$$\mathcal{L}_R = \sum_{l=1}^{4} ||e^l - y^l||_1 \tag{4}$$

To initialize the Left-UNet encoder $\mathcal{G}_e$ and an auxiliary classifier, we utilize pre-trained semantic features from the image classification task mentioned in Section III-A. Subsequently, we train an end-to-end image compression network using the overall loss function:

$$\mathcal{L} = \mathsf{R} + \lambda \mathsf{D} + \alpha \mathcal{L}_C + \beta \mathcal{L}_R \tag{5}$$

The hyper-parameter λ represents the rate-distortion trade-off, which can be adjusted according to the desired image quality factor Q. We set $\alpha = 0.3$ and $\beta = 1.0$ for our experiments.

Through joint compression-classification training, the weights of the Left-UNet encoder are initially initialized with pre-trained semantic features. Subsequently, the gradient descent optimizer updates the encoder-decoder weights to analyze and synthesize the image while improving classification accuracy.

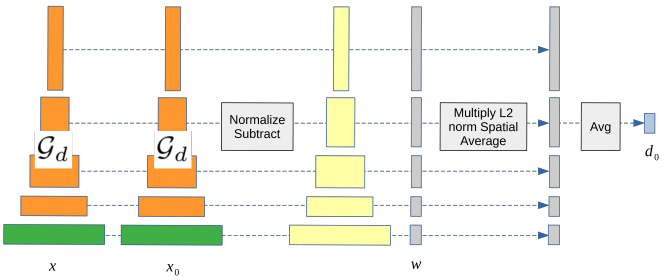


Fig. 3. Computing Euclidean distances from feature outputs e^l and $\hat{y}^5$ between images x and x_0.

D. Compressed Perceptual Image Patch Similarity

To obtain the distance between two images, denoted as x and x_0, we follow the same procedure as LPIPS [26] by learning a linear layer w on the BAPPS dataset. This linear layer assigns weights to the compressed latents and intermediate decoder outputs. Fig. 3 illustrates the process of obtaining the distance using entropy-decoded $\hat{y}^5$ and feature outputs e^l from our decoder network $\mathcal{G}_d$. We extract feature maps $\hat{y}^5, e^l \in \mathbb{R}^{C_l \times H_l \times W_l}$ for all layers l and normalize them in the channel dimension. The activations are then scaled channel-wise using the vector $w^l \in \mathbb{R}^{C_l}$, and the l_2 distance is computed. Finally, we average across the spatial dimensions and all layers to obtain the following:

$$d(f^l) = \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (f^l_{hw} - f^l_{0hw})||_2^2 \tag{6}$$

Eq. (7) calculates the final distance between image x and x_0, that is:

$$d_0 = \sum_{l=1}^{4} d(e^l) + d(\hat{y}^5) \tag{7}$$

Furthermore, we train another smaller network, denoted as $\mathcal{D}$, to predict perceptual judgments h from the distance pair (d_0, d_1) on the BAPPS 151k patches 2AFC (two alternative forced choice) dataset.

IV. Experimental Results

A. Experimental Settings

To implement our CPIPS, we utilize the CompressAI[1] [44] implementation of the hyperprior neural compressor [12] and the official release[2] of LPIPS. We pre-train the image classification task on the ImageNet dataset, which consists of 1.2 million images. The training is performed using the PyTorch Adam optimizer with a learning rate 0.0001 for 120 epochs. Following that, we jointly train the compression-classification task with the pre-trained weights for 150 epochs, employing the Adam optimizer with a learning rate of 0.0001.

[1]https://github.com/InterDigitalInc/CompressAI
[2]https://github.com/richzhang/PerceptualSimilarity

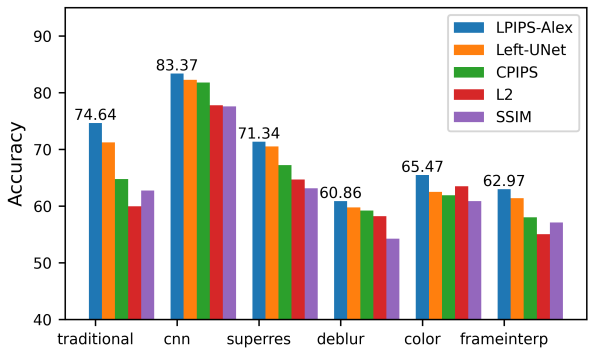


Fig. 4. Comparison of 2AFC accuracy against human ratings on the BAPPS dataset.

Regarding CPIPS weights w and the judgment network $\mathcal{D}$, we train them for ten epochs using the BAPPS 2AFC dataset, as mentioned in the original LPIPS paper.

B. Left-UNet Image Classification

Table III displays our top-1 and top-5 accuracy compared to high-capacity deep networks such as VGG-16. The achieved top-1 accuracy of 60.11% is considered favorable, indicating that the pre-trained weights can serve as a suitable initialization for the Left-UNet encoder $\mathcal{G}_e$.

TABLE III
IMAGENET CLASSIFICATION ACCURACY

Network	**Top-1 Acc.**	**Top-5 Acc.**
AlexNet	56.52%	79.06%
Left-UNet	60.11%	81.95%
ResNet18	69.36%	89.03%
VGG-16	71.51%	93.38%

C. Human Judgment Accuracy

We compare our method with LPIPS and traditional L2 and SSIM metrics, in terms of the accuracy of image judgments against human ratings on the BAPPS dataset. Table IV and Fig. 4 present the results.

TABLE IV
2AFC JUDGMENT ACCURACY

Method	Trad.	CNN	S.Res	DeBlur	Color	F.Interp
LPIPS-Alex	74.64	83.37	71.34	60.86	65.47	62.97
Left-UNet	71.23	82.27	70.51	59.74	62.50	61.39
CPIPS	64.77	81.77	67.21	59.20	61.91	58.00
L2	59.94	77.76	64.67	58.19	63.50	55.02
SSIM	62.73	77.59	63.13	54.23	60.88	57.10

Evidently, the metrics incorporating learned semantic features, such as LPIPS, Left-UNet, and CPIPS, exhibit a higher correlation with human judgments compared to L2 and SSIM. While Left-UNet does not achieve the same level of accuracy as LPIPS, it serves as an upper bound for our proposed CPIPS since they share the same feature extraction convolution layers. Our CPIPS achieves similar accuracy to Left-UNet in the CNN, DeBlur, and Color subsets but experiences a more considerable drop in accuracy in the Traditional, Super-Res, and Frame-Interp subsets. We attribute this drop to two factors: 1) the rate-distortion optimization process influencing the semantic properties of the latent vectors, thereby affecting the perceptual representation, and 2) the multi-scale feature maps e^l serving as proxies for the feature extraction vectors y^l reconstructed in the decoding stages through the regularization loss. Investigating and improving upon these factors are left as future work.

Qualitatively, we select some sample image patches from the BAPPS dataset and present their different judgments in Fig. 5. We can see that the L2 and SSIM cannot reflect human perceptual preferences. At the same time, the CPIPS and LPIPS align with the ground truth better. The second image pair in Fig. 5 demonstrates that the SSIM has a strong bias with structures and tends to be impacted by additive noises.

D. Computational Complexity

We assessed the computation time of the metrics on an Intel i7-9700K workstation with an Nvidia GTX 3090 GPU. To compare our CPIPS metric with LPIPS and DISTS[3], we used the Kodak dataset [45] and calculated the average time cost, as shown in Table V. Due to utilizing of a less complex neural network that only requires decoding the bitstream and intermediate features, our CPIPS method is approximately 50 times faster.

TABLE V
METRIC COMPUTATION TIME ON KODAK

Method	Avg. Time (secs.)
CPIPS	**0.0205**
LPIPS-Alex	1.0681
DISTS	1.0373

V. CONCLUSIONS

In this work, we have introduced an end-to-end learned approach for image compression that aims to preserve perceptual distances. By leveraging pre-training on an image classification task and joint compression-classification training, we initialize the parameters of a learned image coding model with semantic features and guide the gradient descent process to emphasize semantic relevance. We have proposed a UNet-inspired network architecture Left-UNet, shared between the image classifier and the image encoder. Our approach calculates the difference in feature vectors between rate-distortion optimized compressed latents and intermediate decode outputs of two images, providing a perceptual distance preserving metric. We refer to this metric as CPIPS, derived from a learned image codec bitstream at no additional cost. Our experimental results demonstrate that CPIPS aligns more with human subjective judgments than traditional distortion metrics such as L2 and SSIM.

[3]https://github.com/dingkeyan93/DISTS



Fig. 5. The qualitative comparison of selected samples from the BAPPS dataset. The image sets are from the Traditional subset for the first two, CNN and Color subset for the third and last. The orders in each image set are the reference image, the distorted patch-0, and the distorted patch-1.

ACKNOWLEDGMENT

The authors would like to thank the NSTC of Taiwan and CITI SINICA for supporting this research under the grant numbers 111-2221-E-002-134-MY3 and Sinica 3012-C3447.

REFERENCES

[1] F. Attneave, "Some informational aspects of visual perception.," *Psychological review*, vol. 61, no. 3, p. 183, 1954.

[2] H. B. Barlow *et al.*, "Possible principles underlying the transformation of sensory messages," *Sensory communication*, vol. 1, no. 01, pp. 217–233, 1961.

[3] D. L. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature neuroscience*, vol. 19, no. 3, pp. 356–365, 2016.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, *Yolov4: Optimal speed and accuracy of object detection*, 2020. arXiv: 2004.10934 [cs.CV].

[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, Springer, 2014, pp. 184–199.

[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[10] Y. Yang, S. Mandt, and L. Theis, "An introduction to neural data compression," *arXiv preprint arXiv:2202.06533*, 2022.

[11] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[12] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[13] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Neural Information Processing Systems*, vol. 31, pp. 10 771–10 780, 2018.

[14] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 3339–3343.

[15] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[16] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Transactions on Image Processing*, vol. 30, pp. 3179–3191, 2021.

[17] Z. Duan, M. Lu, Z. Ma, and F. Zhu, "Lossy image compression with quantized hierarchical vaes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 198–207.

[18] G. K. Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[19] H. Chen-Hsiu and W. Ja-Ling, "Image data hiding in neural compressed latent representations," *unpublished*, 2023.

[20] M. Testolina, E. Upenik, and T. Ebrahimi, "Towards image denoising in the latent space of learning-based compression," in *Applications of Digital Image Processing XLIV*, SPIE, vol. 11842, 2021, pp. 412–422.

[21] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, H. R. Tavakoli, and E. Rahtu, "Learned image coding for machines: A content-adaptive approach," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2021, pp. 1–6.

[22] Z. Duan, Z. Ma, and F. Zhu, "Unified architecture adaptation for compressed domain semantic inference," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[23] J. Ascenso and E. Upenik, "White paper on jpeg ai scope and framework v1. 0," *ISO/IEC JTC 1/SC 29/WG1 N90049*, 2021.

[24] "Call for evidence for video coding for machines," *ISO/IEC JTC 1/SC 29/WG 2*, 2020.

[25] S. Bhardwaj, I. Fischer, J. Ballé, and T. Chinen, "An unsupervised information-theoretic perceptual quality metric," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13–24, 2020.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[27] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[28] M. D. Adams, "The jpeg-2000 still image compression standard," in *ISO/IEC JTC 1/SC 29/WG 1 N 2412*, Citeseer, 2001.

[29] J. Lainema, M. M. Hannuksela, V. K. M. Vadakital, and E. B. Aksu, "Hevc still image coding and high efficiency image file format," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 71–75.

[30] J.-R. Ohm and G. J. Sullivan, "Versatile video coding–towards the next generation of video compression," in *Picture Coding Symposium*, vol. 2018, 2018.

[31] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wanga, "Image and video compression with neural networks: A review," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[32] D. Mishra, S. K. Singh, and R. K. Singh, "Deep architectures for image compression: A critical review," *Signal Processing*, vol. 191, p. 108 346, 2022.

[33] J. Ballé, *Dcc 2023 - perception: The next milestone in learned image compression*, Apr. 2023. [Online]. Available: https : / / www . youtube . com / watch ? v = Y3ySwIhwvTE.

[34] G. Valenzise, A. Purica, V. Hulusic, and M. Cagnazzo, "Quality assessment of deep-learning-based image compression," in *2018 ieee 20th international workshop on multimedia signal processing (mmsp)*, IEEE, 2018, pp. 1–6.

[35] E. Upenik, M. Testolina, J. Ascenso, F. Pereira, and T. Ebrahimi, "Large-scale crowdsourcing subjective quality evaluation of learning-based image coding," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, 2021, pp. 1–5.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[37] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Ieee, vol. 2, 2003, pp. 1398–1402.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, Springer, 2016, pp. 694–711.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *International Journal of Computer Vision*, vol. 129, pp. 1258–1281, 2021.

[41] E. Upenik, M. Testolina, and T. Ebrahimi, "Towards super resolution in the compressed domain of learning-based image codecs," in *Applications of Digital Image Processing XLIV*, SPIE, vol. 11842, 2021, pp. 531–541.

[42] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," *IEEE Transactions on Image Processing*, vol. 31, pp. 2739–2754, 2022.

[43] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium (PCS)*, IEEE, 2016, pp. 1–5.

[44] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "Compressai: A pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.

[45] *Kodak photocd dataset*, http://r0k.us/graphics/kodak/.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Task-Specific Pruning: Efficient Parameter Reduction in Multi-task Object Detection Models

Wei-Hsun Ke*, Yu-Wen Tseng† and Wen-Huang Cheng‡
* National Yang Ming Chiao Tung University, Taiwan
E-mail: arch1n1my.ee11@nycu.edu.tw Tel: +886-3-5712121(Ext.)54208
† National Taiwan University, Taiwan
E-mail: ywt@cmlab.csie.ntu.edu.tw Tel: +886-2-33664888(Ext.)506
‡ National Taiwan University, Taiwan
E-mail: wenhuang@csie.ntu.edu.tw Tel: +886-2-33664888(Ext.)420

***Abstract*—This study explores the optimization effect of task-specific pruning in object detection models. Object detection models are designed to handle classification and regression tasks, which are fundamentally distinct. However, the relationship between these two tasks, particularly regarding loss calculation, remains an unresolved issue in both model training and pruning. To address this, we propose a task-specific pruning method that combines the advantages of pruning-during-training with established sparsity learning techniques. Through our experiments, we demonstrate that task-specific pruning is the most effective approach for reducing parameter count while maintaining high accuracy in multi-task deep neural networks, across different configurations. Specifically, when applied to the Scaled-Yolov4-p5 model using the MS COCO 2017 dataset, our approach achieves a target sparsity of 90% with only a marginal 1% drop in mAP.**

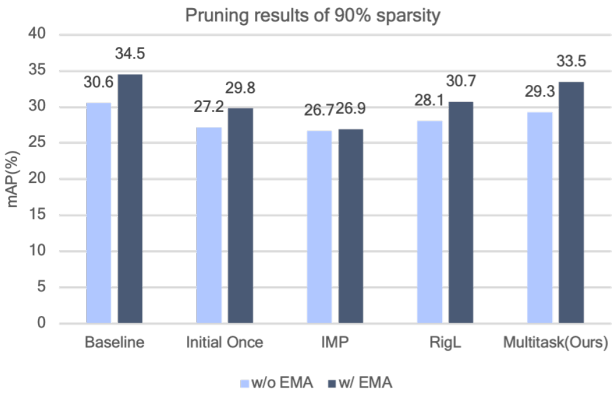


Fig. 1. Our experimental results show that the multi-task(Ours) setting produces the most promising results on the Scaled-Yolov4-p5 object detection model. It achieves a target sparsity of 90% with only a 1% drop in mAP on the MS COCO 2017 dataset. EMA (Exponential Moving Average) is a method of calculating an average that assigns higher weights to recent parameters. This technique can improve training performance. The experimental results indicate that both with and without EMA, the pruning method exhibits the same trend in accuracy after pruning.

I. Introduction

In recent years, there has been a significant increase in the number of edge devices. To achieve faster computation, deep neural networks are often installed directly on the edge device instead of transmitting the computation information over the network. However, these edge devices have limited storage and less computational capabilities compared to general computers. Therefore, techniques such as model compression and multi-task learning have become increasingly important in order to make these models more efficient for practical applications.

A current mainstream approach is to use a multi-task model architecture to implement multi-task learning. For example, DeepLab [8], originally designed for semantic segmentation, can be modified with some output structure adjustments to simultaneously handle tasks such as key-point detection and depth prediction. These types of model architectures have the characteristic of sharing task information within a single model, which results in savings in storage space and faster computational speed compared to using multiple single-task models. The architecture can be divided into two parts: shared layers, which store generic information shared by all tasks, and task-specific layers, which handle the specific information required by each task. To prevent conflicts between task-specific information, model designers often design multiple task-specific heads to differentiate between different tasks.

Object detection models, which are widely used in computer vision, can also be classified as multi-task models as they handle different tasks such as classification and bounding box regression within a single model. However, the inherent differences between these two tasks, with one focused on differentiating between object classes and the other focused on the distribution trend of object positions, have always posed a challenge in loss calculation during model training [14] [15].

Among various model compression approaches, pruning is an efficient and practical technique that compresses models by determining the importance of model parameters and removing unnecessary ones. The mainstream pruning methods can be categorized based on their timing relative to training into three types: train-prune-finetune, prune-at-initial, and prune-during-training. Another approach involves repeatedly pruning during training to optimize sparse models. Sparse learning [5] methods enable models to learn how to prune themselves during

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

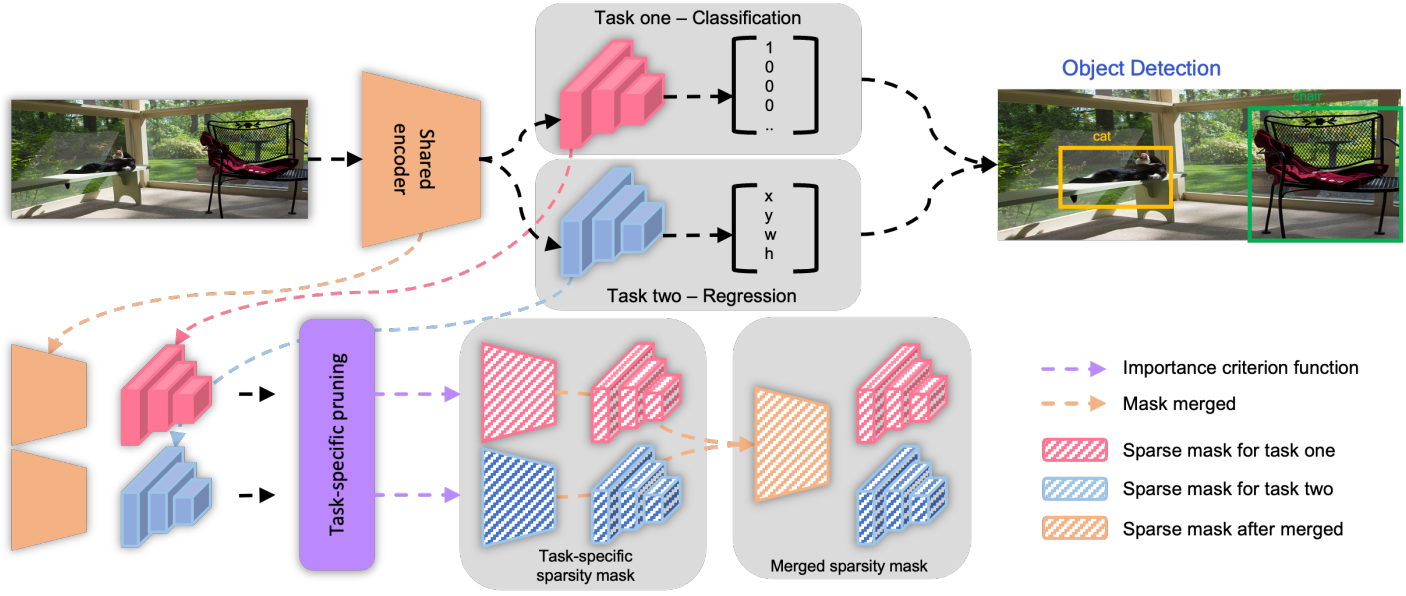


Fig. 2. Our proposed multi-task pruning for object detection.

training, which helps the models acquire additional knowledge that normal training processes may not capture, such as the importance of weights or gradients within individual layers. Prune-during-training [10] attempts to find the optimal trade-off between train-prune-finetune and prune-at-initial methods by determining the stability of the model during training and selecting the best timing for pruning.

In general, multi-task models often handle multiple closely related tasks to improve accuracy and memory efficiency, such as object classification, object detection and segmentation, joint scene geometry, and key point detection. However, applying pruning techniques commonly used for classification tasks to these models may not yield the best results. The DiSparse [7] method introduces a multi-task pruning technique that can be applied to multiple tasks, such as semantic segmentation and depth prediction. This technique allows for pruning across multiple tasks simultaneously. On the other hand, the MTP [9] method addresses the implicit multi-task learning problem in semantic segmentation models by combining the tasks of classification and segmentation for pruning. By leveraging the relationship between these two tasks, these method improves the performance of semantic segmentation models.

Inspired by such pruning methods, our work categorizes object detection models, which traditionally handle a single object detection task, as a type of multi-task model (Fig. 2). We apply the same concept to help the model learn how to differentiate shared and task-specific information. By evaluating the model with the loss functions specific to classification and regression tasks, we obtain task-specific gradients as the basis for pruning importance. However, calculating these task-specific gradients consumes a significant amount of computation time, which increases with each iteration. Therefore, we combine the advantages of sparse learning and prune-during-training methods to determine the optimal timing for pruning. This approach allows us to save considerable computation time while learning to distinguish tasks.

II. Related Works

A. Network Pruning

Currently, network pruning methods can be classified into two categories based on their approach to processing the model architecture: unstructured pruning and structured pruning. Unstructured pruning [5], [13], [19] removes unimportant parameters or neuron connections to achieve computational resource savings, while structured pruning directly removes neurons, channels, filters, or other structured parts of the network to achieve model size reduction and computational resource savings. Unstructured pruning is more precise and intuitive in pruning strategy, but due to its unstructured nature, it usually requires special hardware or software support. Structured pruning [10], [16], [17], on the other hand, presents a greater challenge as it requires careful consideration of pruning positions to mitigate potential accuracy degradation. This is due to the dependencies among the pruned elements, making it more complex to implement.

There are two commonly used methods for determining the importance of network elements for pruning: magnitude-based [16], [17] and gradient-based [18]–[20] methods. Magnitude-based methods typically involve ranking network parameters according to their absolute values, while gradient-based methods compute the gradient of the loss function with respect to each parameter and use this information to rank them. It should be noted that implementing the pruning method mentioned above requires additional computational resources. Therefore, the timing and duration of performing pruning are factors that need to be considered when designing pruning algorithms. The Prune-during-Training [10]–[12] method is suitable for such adjustments, and it is also the method we have

adopted. By finding the optimal timing and duration, a trade-off between saving computational resources and achieving the best accuracy can be achieved.

B. Object detection tasks

Object detection can be widely used in various applications, such as security monitoring, autonomous driving, and crowd recognition. It is mainly used to identify objects in images or videos and label their positions and types. Modern object detection technology is mainly based on Convolutional Neural Networks (CNN) [21], [23], [24] and Transformer [25]–[28] models. The model processes the image multiple times to obtain different features. These features are then used to detect objects in the image through two tasks, object localization and object classification. Object localization generates a bounding box for each object, indicating its position and size, while object classification classifies the detected objects and labels them with their respective types.

Among CNN-based object detection models, there are two types based on their detection method: two-stage and one-stage. Two-stage methods first identify potential object locations as proposals and then perform feature recognition and classification on those proposals, such as the RCNN [4], [29], [30] family, including Mask R-CNN [4]. One-stage methods, on the other hand, simultaneously process both bounding box regression and object classification to identify the object's position. While the accuracy of one-stage methods is not as high as two-stage methods, they are faster. One-stage models can further be divided into coupled-head [1], [2], [21] and decoupled-head [22], [24] based on how they handle the two tasks in the model architecture. Decoupled-head is a commonly seen model structure in Multi-task-learning. It will perform classification and bounding box regression separately in two model heads for final calculation and output. Coupled-head is a more commonly used architecture in one-stage models today, where both tasks are combined and calculated in the same head to achieve smaller model size.

III. Method

A. Decouple the tasks

To begin with, we decompose the original coupled loss function ℓ into task-specific loss functions ℓ_k, depending on the task k we wish to address. We calculate the task-specific loss L_k for each task separately using the current N sets of predictions x_i and targets y_i.

In addition, we define task-specific loss functions ℓ_1 and ℓ_2 for the classification and bounding-box regression tasks, respectively. These loss functions will be used to compute the importance criteria $\mathbb{I}_k$ for multi-task pruning.

$$L_k = \frac{1}{N}\sum_{i=1}^{N} \ell_k(y_i, f_p(x_i)). \tag{1}$$

$$\ell = \ell_{cls} + \ell_{obj} + \ell_{box}. \tag{2}$$

$$\ell_1 = \ell_{cls} + \ell_{obj}. \tag{3}$$

$$\ell_2 = \ell_{box} + \ell_{obj}. \tag{4}$$

Algorithm 1 Task-specific-Pruning During Training

Input: $Loss\ Functions\ of\ each\ Task\ \ell_k$,
$Sparsity\ Distribution: \mathbb{S} = \{s^1, ..., s^L\}$
$Schedule: T_i, T_e, \Delta T$
$Untrained\ Network\ f_w$,
$Parameter\ Sets\ of\ each\ Layer\ W_l\ in\ f_w$
$W_l \leftarrow Random\ sparsify\ W_l\ using\ \mathbb{S}$
while $T_i \leq epoch\ t \leq T_e$ **do**
  if epoch t $(\mathrm{mod}\ \Delta T) == 0$ **then**
    $copy\ network\ f_p \leftarrow W_l\ of\ pruning\ target\ f_w$
    for `each task k` **do**
      $L_k = \frac{1}{N}\sum_{i=1}^{N} \ell_k(y_i, f_p(x_i))$
      $\mathbb{I}_k = ArgTopK(-|\nabla_{W_l} L_k|, s^l)$
      if k == 0 **then**
        $\mathbb{I}_{prune} \leftarrow \mathbb{I}_k$
      else
        $\mathbb{I}_{prune} \leftarrow \mathbb{I}_{prune} \cap \mathbb{I}_k$
      end if
    end for
    $W_l \leftarrow Update\ connections\ W_l\ using\ \mathbb{I}_{prune}$
  end if
end while

Before completing the pruning process, we create a copy of the target model f_w's weights and optimizer in a new model f_p. We then pass the task-specific loss L_k to this copy model to obtain the task-specific gradient magnitudes. These gradient magnitudes are used to determine the importance $\mathbb{I}_k$ of each parameter for pruning. Finally, we prune the top s^l smallest unimportant parameters in the dictionary $\mathbb{I}_{prune}$ for each layer l such that $1 \leq l \leq L$, to achieve the desired sparsity distribution $\mathbb{S}$ for each layer. We use RigL [5]'s Erk distribution as the sparsity calculation metric for both CNN and linear layers. The sparsity distribution $\mathbb{S}$ represents the ratio of zero parameters to the total parameter count in each layer, resulting in an overall model sparsity of S.

$$\mathbb{S} = \{s^1, ..., s^L\} \tag{5}$$

$$S = \frac{1}{L}\sum_{l=1}^{L} s^l. \tag{6}$$

$$\mathbb{I}_k = ArgTopK(-|\nabla_{W_l} L_k|, s^l). \tag{7}$$

B. Prune a stable model

After training the model for a period, it will reach a state of stability where pruning can be viewed as beneficial noise

in the training process. Various metrics, such as the number of remaining neurons after pruning [10], flatness [6], and Stochastic Gradient Descent(SGD) noise [13], are employed by different studies to assess stability. However, a common observation is that the model tends to become more stable after a certain number of training epochs, making it conducive for pruning. Pruning a stable model can lead to better results, where stability refers to the point in time where the pruned model structures, obtained from evaluating the model using the same importance criteria over several consecutive training epochs, exhibit little variation.

To prune a stable model, we use a schedule that defines T_i as the epoch when the first pruning begins, ΔT as the required number of training epochs for the model to reach stability, and T_e as the latest possible time for pruning. We utilize a grid-search approach to identify the optimal pruning start time T_i with a slight delay, ensuring that the model reaches stability before initiating the pruning process. Additionally, we adjust the end epoch T_e for pruning to reduce the computational overhead associated with the pruning process.

IV. Experiment

A. Experiment settings

In this section, we present a comprehensive experiment on model pruning for object detectors. Considering the practical application ability, we choose the object detector from the YOLO series [1]–[3], which has a better trade-off between inference time and detection performance. In this experiment, we use Scaled-YOLOv4-p5 [1] as our base detector. If there are no other mentions, all object detectors are trained in one Nvidia 2080ti GPU for 300 epochs from scratch. Because of the limitation of the hardware resources, we use a lower input image size(256) and lower batch size(32) in our training process. We adopt the SGD optimizer with the parameters as the original Scaled-YOLOv4 settings: a momentum of 0.937, an initial learning rate of 0.01, and a weight decay of 0.0005. At last, we evaluate our model on the most general and challenging object detection dataset, the MS COCO dataset. The following experiments show our proposed method can achieve a competitive performance with other state-of-the-art pruning methods.

We utilize "mean average precision (mAP)" as the performance evaluation metric. It is calculated by averaging the precision scores across all 80 categories in the MS COCO dataset, considering 10 Intersection over Union (IoU) thresholds ranging from .50 to .95 with an increment of .05. Averaging over IoUs rewards detectors with better localization, making mAP a comprehensive measure of both detection accuracy and localization performance.

B. Performance

Main result. To present a fair comparison, we re-implement our baseline, IMP [13] and RigL [5] with the same settings as our training process. Tab. I reports the performance of our pruning method for object detection and other baselines for the MS COCO dataset. Initial once means the model will be randomly pruned at the beginning of the training process. In IMP, model weight magnitudes are used for importance evaluation during the pruning interval. In RigL, an additional metric based on model gradient magnitudes is incorporated following the weight magnitude importance evaluation. In the multi-task settings, we calculate the losses of the two tasks separately for task-specific gradient calculation, and we utilize their intersection as the final evaluation metric for pruning. This represents the primary distinction between the multitask method and RigL.

As the results shown in Tab. I, our method surpasses all the other methods and loses only 1.0 mAP (-3%) with 90% sparsity while other methods get a performance loss of at least 3.8 mAP (-11%). This excellent result validates our perspective that the important weights have some nuances between the regression task and classification task, which can be addressed through the multi-task settings (ours).

Pruning with different pruning intervals. RigL [5] suggests pruning the first three-quarters of all training epochs, using an

TABLE I
Detection accuracy comparisons in terms of mAP on MS COCO 2017 Dataset

Pruning method	mAP(%)	Input size	Batch size	Sparsity(%)
Baseline	34.5	256×256	32	0
Initial Once	29.8 (-4.7)	256×256	32	90
IMP [13]	26.9 (-7.6)	256×256	32	90
RigL [5]	30.7 (-3.8)	256×256	32	90
multi-task(Ours)	**33.5 (-1.0)**	256×256	32	90

TABLE II
Results with varying pruning interval

Pruning Method	mAP(%)	ΔT	T_i - T_e	Sparsity(%)
Baseline	34.5	-	-	0
Initial Once	29.8	-	-	90
IMP [13]	26.9	1	0 - 225	90
IMP	29.6	1	10 - 20	90
IMP	28.6	1	100 - 110	90
multi-task	**33.5**	1	10 - 20	90
multi-task	31.3	1	100 - 110	90

TABLE III
Results with different pruning settings

Pruning Method	mAP(%)	ΔT	T_i - T_e	Sparsity(%)
None(baseline)	34.5	-	-	0
Initial Once	29.8	-	-	90
IMP [13]	26.9	1	0 - 225	90
RigL [5]	30.7	1	0 - 225	90
IMP	29.6	1	10 - 20	90
multi-task (head only)	30.7	1	10 - 20	90
multi-task (backbone + head)	**33.5**	1	10 - 20	90

update interval of 100 iterations and an initial drop fraction of 0.5 for the best results. We follow the settings mentioned above, as well as the dynamic pruning method of DiSparse [7]. However, we found that continuously pruning from the beginning until 3/4 of training epochs is inefficient and unnecessary. Instead, we reduced the pruning epochs from 225 to 10, with an update interval of $\Delta T = 1$ epoch.

Referring to Tab. II, our experimental results show that pruning only at these 10 epochs can still yield good or even better pruning results. When to Prune [10] states that pruning slightly after the start of training, when the model becomes more stable, can lead to better pruning results. We set the pruning start epoch T_i as epoch 10 out of the total 300 training epochs. Furthermore, we conducted pruning at different intervals during epochs 10-20 and epochs 100-110 separately within the same total 300 training epochs. The results showed that pruning during the interval 10-20 produced better outcomes. This suggests that pruning after a short initial training phase is more effective and provides additional epochs for fine-tuning after pruning the target model.

Task-specific pruning for the partial model. To validate the need for task-specific pruning in coupled-head models, our method assumes that performing task-specific pruning on the target model will lead to better performance of the pruned model. In our experiments, we explore multi-task processing by separately applying task-specific pruning to the backbone + head and head-only configurations of the model. Among them, the backbone + head configuration shows the best performance, indicating that for a coupled-head model, performing multi-task processing on all layers except the input and output layers yields the best results. This approach effectively addresses the conflict between the two tasks in the coupled head.

V. Conclusions and Future Work

In this work, we applied task-specific pruning to address the challenges of pruning in the object detection task. By adjusting the timing of the pruning start and interval, we achieved improved results compared to other methods that require more effort. Our findings highlight the importance of considering the two tasks of the object detection model separately and ensuring model stability during the initiation of the pruning process.

The concept of this work is model-agnostic, which means it can be applied to various object detection models. In the future, we plan to validate this by applying the proposed method to other models such as transformers and two-stage object detection models.

Acknowledgment

This work was supported in part by Qualcomm through a Taiwan University Research Collaboration Project, as well as the National Science and Technology Council of Taiwan under Grants: NSTC-109-2223-E-002-005-MY3, NSTC-112-2628-E-002-033-MY4 and NSTC-112-2634-F-002-002-MBK.

References

[1] Wang, C., Bochkovskiy, A. & Liao, H. Scaled-YOLOv4: Scaling Cross Stage Partial Network. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 13029-13038 (2021)

[2] Redmon, Joseph and Divvala, Santosh and Girshick, Ross and Farhadi, Ali. You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)

[3] Alexey Bochkovskiy and Chien-Yao Wang and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv:2004.10934*. (2020)

[4] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 2961-2969 (2017)

[5] Evci, U., Gale, T., Menick, J., Castro, P. & Elsen, E. Rigging the lottery: Making all tickets winners. *International Conference On Machine Learning*. pp. 2943-2952 (2020)

[6] Bartoldson, B., Morcos, A., Barbu, A. & Erlebacher, G. The generalization-stability tradeoff in neural network pruning. *Advances In Neural Information Processing Systems*. **33** pp. 20852-20864 (2020)

[7] Sun, X., Hassani, A., Wang, Z., Huang, G. & Shi, H. DiSparse: Disentangled Sparsification for multi-task Model Compression. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 12382-12392 (2022)

[8] Chen, L., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. *ArXiv Preprint ArXiv:1706.05587*. (2017)

[9] Chen, X., Zhang, Y. & Wang, Y. MTP: multi-task pruning for efficient semantic segmentation networks. *2022 IEEE International Conference On Multimedia And Expo (ICME)*. pp. 1-6 (2022)

[10] Shen, M., Molchanov, P., Yin, H. & Alvarez, J. When to prune? a policy towards early structural pruning. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 12247-12256 (2022)

[11] Lym, S., Choukse, E., Zangeneh, S., Wen, W., Sanghavi, S. & Erez, M. Prunetrain: fast neural network training by dynamic sparse model reconfiguration. *Proceedings Of The International Conference For High Performance Computing, Networking, Storage And Analysis*. pp. 1-13 (2019)

[12] Alvarez, J. & Salzmann, M. Learning the number of neurons in deep networks. *Advances In Neural Information Processing Systems*. **29** (2016)

[13] Frankle, J., Dziugaite, G., Roy, D. & Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. *International Conference On Machine Learning*. pp. 3259-3269 (2020)

[14] Song, G., Liu, Y. & Wang, X. Revisiting the sibling head in object detector. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 11563-11572 (2020)

[15] Jiang, X., Li, Z., Tian, M., Liu, J., Yi, S. & Miao, D. Few-Shot Object Detection via Improved Classification Features. *Proceedings Of The IEEE/CVF Winter Conference On Applications Of Computer Vision*. pp. 5386-5395 (2023)

[16] Fang, G., Ma, X., Song, M., Mi, M. & Wang, X. DepGraph: Towards Any Structural Pruning. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 16091-16101 (2023,6)

[17] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S. & Zhang, C. Learning efficient convolutional networks through network slimming. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 2736-2744 (2017)

[18] Tanaka, H., Kunin, D., Yamins, D. & Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances In Neural Information Processing Systems*. **33** pp. 6377-6389 (2020)

[19] Wang, C., Zhang, G. & Grosse, R. Picking winning tickets before training by preserving gradient flow. *ArXiv Preprint ArXiv:2002.07376*. (2020)

[20] De Jorge, P., Sanyal, A., Behl, H., Torr, P., Rogez, G. & Dokania, P. Progressive skeletonization: Trimming more fat from a network at initialization. *ArXiv Preprint ArXiv:2006.09081*. (2020)

[21] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. & Berg, A. Ssd: Single shot multibox detector. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. pp. 21-37 (2016)

[22] Lin, T., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 2980-2988 (2017)
[23] Wang, C., Yeh, I. & Liao, H. You only learn one representation: Unified network for multiple tasks. *ArXiv Preprint ArXiv:2105.04206*. (2021)
[24] Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. Yolox: Exceeding yolo series in 2021. *ArXiv Preprint ArXiv:2107.08430*. (2021)
[25] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. End-to-end object detection with transformers. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. pp. 213-229 (2020)
[26] Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. & Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 13619-13627 (2022)
[27] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. & Shum, H. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ArXiv Preprint ArXiv:2203.03605*. (2022)
[28] Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q. & Liu, Y. DETRs Beat YOLOs on Real-time Object Detection. (2023)
[29] Girshick, R. Fast r-cnn. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 1440-1448 (2015)
[30] Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances In Neural Information Processing Systems*. **28** (2015)

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Transformer-based Image Compression with Variable Image Quality Objectives

Chia-Hao Kao*, Yi-Hsin Chen*, Cheng Chien, Wei-Chen Chiu, and Wen-Hsiao Peng
National Yang Ming Chiao Tung University, Taiwan

***Abstract*—This paper presents a Transformer-based image compression system that allows for a variable image quality objective according to the user's preference. Optimizing a learned codec for different quality objectives leads to reconstructed images with varying visual characteristics. Our method provides the user with the flexibility to choose a trade-off between two image quality objectives using a single, shared model. Motivated by the success of prompt-tuning techniques, we introduce prompt tokens to condition our Transformer-based autoencoder. These prompt tokens are generated adaptively based on the user's preference and input image through learning a prompt generation network. Extensive experiments on commonly used quality metrics demonstrate the effectiveness of our method in adapting the encoding and/or decoding processes to a variable quality objective. While offering the additional flexibility, our proposed method performs comparably to the single-objective methods in terms of rate-distortion performance.**

I. Introduction

Compared to traditional image codecs, learned image codecs can be easily adapted to any differentiable quality metrics. When a learned image codec is trained separately with different quality metrics, the decoded images may exhibit distinctive visual characteristics. For instance, involving mean squared error (MSE) in the training objective usually leads to blurry reconstructed images, particularly at low bitrates. On the other hand, when trained to minimize the Learned Perceptual Image Patch Similarity (LPIPS) [1], the codec typically yields visually pleasing images, but may sometimes introduce undesirable artificial patterns [2]. By supporting multiple quality objectives, a compression system can cater to a wide range of application requirements and user preferences. It enables the user to adapt the quality objective to their needs, whether it is to maximize the perceptual quality or reconstruction fidelity of the decoded image, or a weighted combination of these objectives. Therefore, developing a flexible compression system that can support multiple quality objectives is of importance in meeting the diverse demands of different image compression applications.

Although end-to-end learned image compression has made significant progress [3], [4], [5], most existing methods have to train separate models for different quality objectives. Normally, each model is optimized for one specific quality metric, which can be MSE, multi-scale structural similarity (MS-SSIM), or a weighted combination of these metrics. However, this approach is impractical in terms of computational resources and storage requirements.

*Equal contribution.

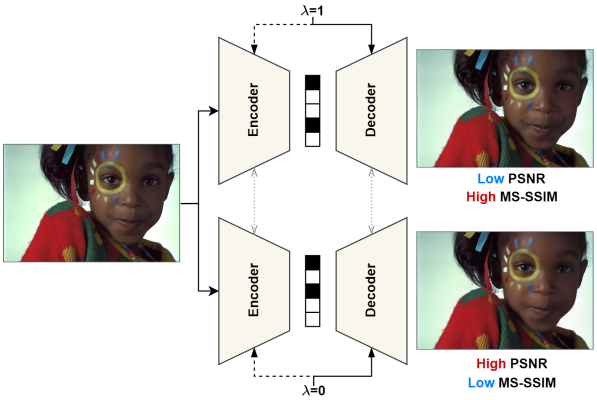


Fig. 1. Illustration of our proposed framework. By adjusting the condition parameter λ, the reconstructed images exhibit distinct characteristics, enabling customizable quality balance based on specific preferences.

Recently, Agustsson *et al.* [6] propose a compression system with a conditional decoder for adaptive reconstruction of images to trade off high fidelity (i.e. low MSE) for realism. From the same bitstream, it is able to decode an image with variable quality objectives. However, they [6] only address the trade-off between realism and MSE, disregarding other metrics. Furthermore, their conditioning mechanism is performed on the decoder side only, without exploring the potential benefits of incorporating it into both the encoder and decoder.

In this work, we propose a Transformer-based learned image codec that is able to trade off one image quality objective for another (e.g. Peak-Signal-to-Noise-Ratio versus MS-SSIM). Transformer-based image compression systems [7], [5], [8] have attracted lots of attention recently for their comparable or even superior coding efficiency to CNN-based methods. Fig. 1 illustrates the main idea of our framework, which is designed to adapt the encoding and/or decoding processes to optimize the decoded image quality according to a given trade-off between two selected quality objectives. Motivated by the recent success of prompt tuning techniques [9], [10], we propose a prompt generation network to adapt the coding process according to the specified quality objective. Specifically, we explore two variations. The first applies prompt tuning to both encoder and decoder, specializing the bitstream for a specific quality objective. In comparison, the second adapts only the decoder, allowing a single bitstream to be decoded into images that meet variable quality objectives.

The contributions of this paper are as follows:

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

- We introduce a Transformer-based image compression framework that is able to adapt the decoded image quality to a variable objective.
- To our best knowledge, this is the first work that leverages prompt-tuning techniques to trade off one image quality objective for another. Additionally, we explore two variants of the proposed method for different application scenarios.
- Extensive experiments confirm the effectiveness of our proposed method. Specifically, when both encoder and decoder are adapted, our method is able to trade off one quality objective for another effectively while achieving similar rate-distortion performance to the case when the codec is optimized solely for one of the two objectives.

II. Related Works

A. Learned Image Compression

End-to-end learned image compression methods [11], [12], [3], [4] have made significant amount of progress over the past few years, even outperforming traditional codecs, such as intra coding in HEVC [13] and VVC [14]. While most of the existing literature in this field adopt CNN-based autoencoder as their foundation, Transformer-based compression systems [5], [7] have emerged as promising alternatives, particularly for their low computational complexity and outstanding coding efficiency. Zhu *et al.* [5] adopt Swin-Transformer [15] architecture for image compression, showing comparable performance to its CNN-based counterparts with significantly reduced decoder-side kMACs (multiply-accumulate-operations). Similarly, Lu *et al.* [7] utilize Swin-Transformer blocks but substitute the patch merging/splitting operations with more general convolution and transposed convolution layers. More recently, Liu *et al.* [8] propose a mixed Transformer-CNN image compression system that further improves the coding efficiency.

B. Conditional Compression Systems

Conditional compression systems designs are commonly adopted for variable rate coding, enabling a single compression model to encode input images at various bit-rates based on the given trade-off of rate and distortion. In this line of works, the condition is often applied through altering the distribution of the feature maps. Cui *et al.* [16] introduce gain units that perform channel-wise scaling on the image latents for continuous rate adaptation. Song *et al.* [17] use spatial feature transform (SFT) layers [18] to perform element-wise affine transformation on the intermediate feature maps of the autoencoder, achieving both variable rate and ROI coding simultaneously. In a similar vein, Wang *et al.* [19] target Transformer-based models and perform channel-wise scaling on value matrices in all self-attention layers for variable rate control. Recently, Agustsson *et al.* [6] extend the idea to trade off realism for fidelity (i.e. MSE) of reconstructed images, introducing a conditional decoder that performs channel-wise shifting on the intermediate feature maps inside each residual blocks. The condition is applied to adapt the decoder for different levels of realism and produce the reconstructed images accordingly. However, this method only addresses the trade-off between realism and fidelity, and limits the adoption of conditioning mechanism on the decoder side alone.

C. Prompt Tuning

Prompting techniques [9], [20], [21] were first introduced in natural language processing (NLP) field for efficiently adapting large pre-trained Transformer models to various downstream tasks. The main idea is to keep the entire large pre-trained Transformer frozen, and introduce additional input tokens, known as prompts, for adaptation. These tokens are involved in the self-attention and act as a guidance for the current task. Recently, Jia *et al.* [10] extend the idea to Vision Transformers [22] and computer vision tasks. They prepend learnable prompts to the image tokens at each Transformer blocks, achieving comparable or even superior performance than fine-tuning the entire model on downstream task while only learning less than 1% of the total parameters.

III. Proposed Method

A. Preliminary

Most of the learning-based image compression models [11], [12], [3], [4] share a similar coding architecture, consisting of a main autoencoder and a hyperprior autoencoder [11]. As illustrated in Fig. 2, our framework also adopts this design. Specifically, given an input image $x \in \mathbb{R}^{3\times H\times W}$ of height H and width W, a learned non-linear analysis transform N_{enc} encodes the image into its latent representation $y \in \mathbb{R}^{192\times \frac{H}{16}\times \frac{W}{16}}$. To efficiently transmit y, it is first uniformly quantized into $\hat{y}$, which is then entropy coded using a learned prior distribution $p(\hat{y})$. This prior distribution is modeled using the hyperprior autoencoder in a content-adaptive manner [12]. As shown, the hyper-analysis transform H_{enc} converts the quantized latent $\hat{y}$ into the side information $\hat{z} \in \mathbb{R}^{128\times \frac{H}{64}\times \frac{W}{64}}$ and the quantized $\hat{z}$ is transmitted for generating the learned prior distribution for entropy coding $\hat{y}$. Lastly, the synthesis transform N_{dec} decodes $\hat{y}$ from the bitstream to obtain the reconstructed image $\hat{x} \in \mathbb{R}^{3\times H\times W}$.

B. System Overview

Our work proposes an image compression system capable of producing reconstructed images that reflect variable image quality objectives. Specifically, given two selected quality objectives, the coding process is adapted to trade off one objective for the other. Fig. 2 illustrates the overall framework of our proposed method, which is built upon TIC [7], a Transformer-based image compression system. For the analysis transform N_{enc}, synthesis transform N_{dec}, and hyper-autoencoder H_{enc}, H_{dec} of TIC, the commonly-seen CNN-based design is replaced with interleaving convolutional layers and Swin-Transformer blocks (STB). Note that we discard the context model in favor of a simpler Gaussian prior for entropy coding. To achieve variable image quality objectives, our system incorporates an additional input, a uniform lambda

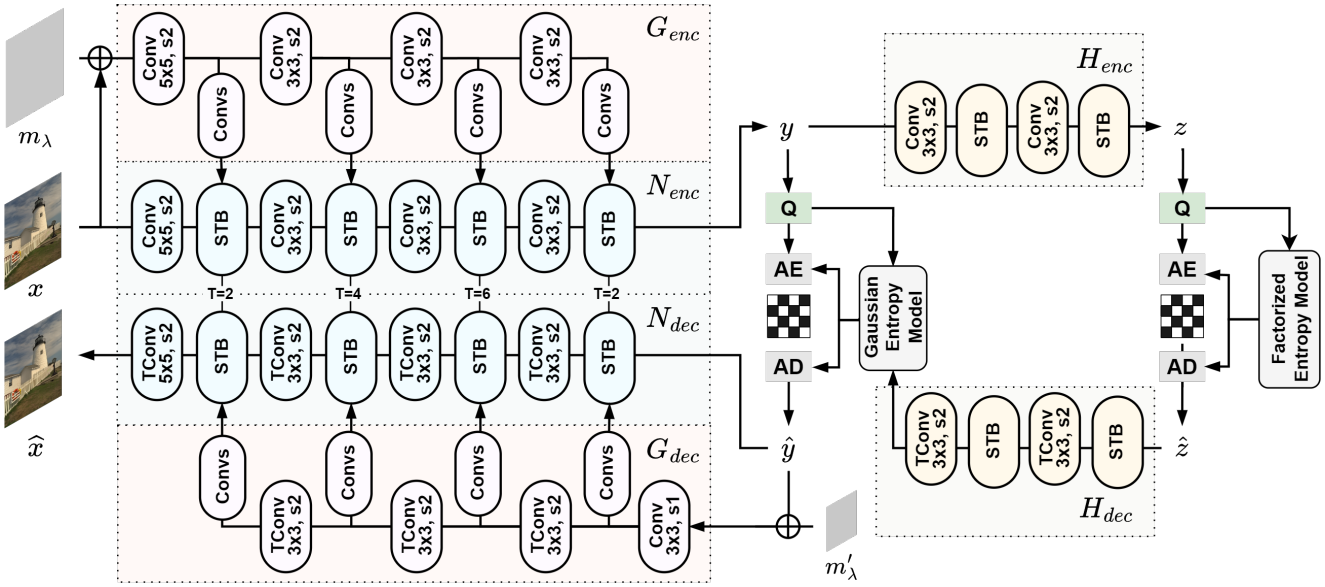


Fig. 2. Overall architecture of our proposed method. We introduce prompt generation networks G_{enc}, G_{dec} to generate prompt tokens as additional input for adapting to variable image quality objectives.

map $m_\lambda \in \mathbb{R}^{1\times H\times W}$ of the same resolution as the input image on the encoder side, and a smaller-scale uniform lambda map $m'_\lambda \in \mathbb{R}^{1\times \frac{H}{16}\times \frac{W}{16}}$ on the decoder side. Both m_λ and m'_λ are populated with the same parameter $\lambda \in [0, 1]$. By adjusting the value of $\lambda \in [0, 1]$, users can specify their desired trade-off between the two quality objectives. At the two extreme ends, where $\lambda = 0$ or $\lambda = 1$, the system reduces to simply optimizing for one specific quality objective. Inspired by the success of prompt tuning [10], we leverage the idea of prompting to condition our Swin-Transformer based codec for adaptation. In this work, we explore two variants of our conditioning method for two different application scenarios. The first variant incorporates the conditioning mechanism into both the encode and decoder, specializing the resulting bitstream for a specific quality objective. The other variant adopts only the decoder-side adaptation, enabling a single compressed bitstream to be decoded into images with variable quality objectives. The details of these variants are presented in the last paragraph of Section III-C.

C. Prompt-based Conditioning

Drawing inspiration from prompt tuning methods [10], we employ additional learnable tokens, termed prompts, to condition our Transformer-based compression model for variable image quality objectives. We introduce prompt generation networks G_{enc} and G_{dec} as add-on modules on top of our main autoencoder. As shown in Fig. 2, G_{enc}, which consists of several convolutional layers, is introduced to the encoder. It takes as input the concatenation of the input image x and lambda map m_λ to generate prompt tokens for each STB in N_{enc}. Similarly, the decoder incorporates G_{dec}, which shares the same architecture as G_{enc} but replaces convolutions with transposed convolutions for upsampling. The input to G_{dec} is the concatenation of the quantized latent $\hat{y}$ and a smaller-resolution lambda map m'_λ. What sets our method apart from [10], in which prompts are directly learned parameters, is the introduction of prompt generation networks for generating prompt tokens adaptively according to the given λs and the input image or latent.

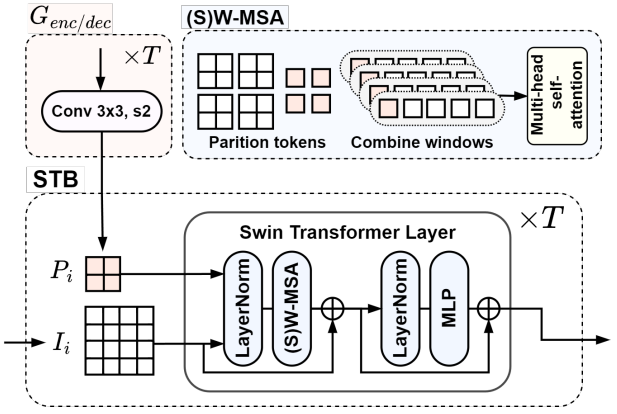
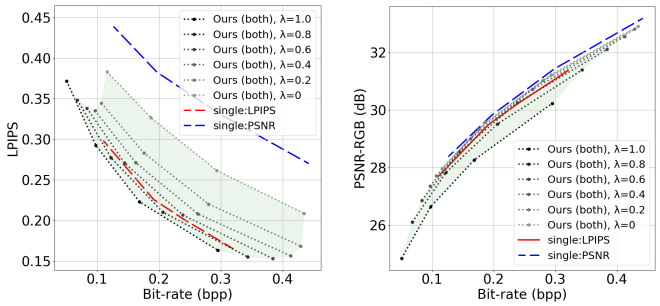


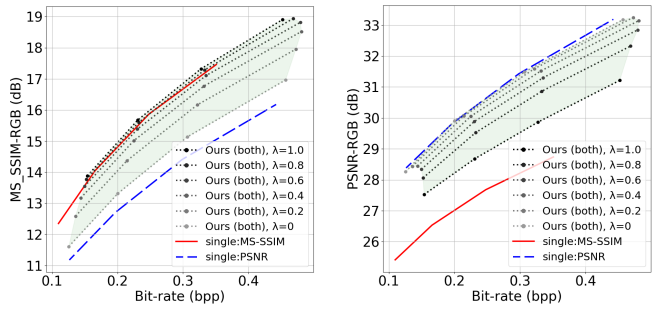
Fig. 3. Detailed architecture of Swin-Transformer Block (STB).

Fig. 3 illustrates the structure of STB and how we incorporate prompt tokens. A regular STB is comprised of multiple Swin-Transformer layers, with T denoting the number of layers and the image tokens I_i being the input to the i^{th} layer. Each Swin-Transformer layer consists of layer normalization, window-based multi-head self-attention (W-MSA), and a MLP layer to aggregate neighboring information. Note that the window for W-MSA is shifted every other layer for tokens across windows to exchange information. To perform prompting, the input to i^{th} Swin-Transformer layer additionally includes a set of prompt tokens P_i obtained through the prompt generation network. In our design, the number of tokens for P_i is only a quarter of that of I_i for reducing computational complexity.



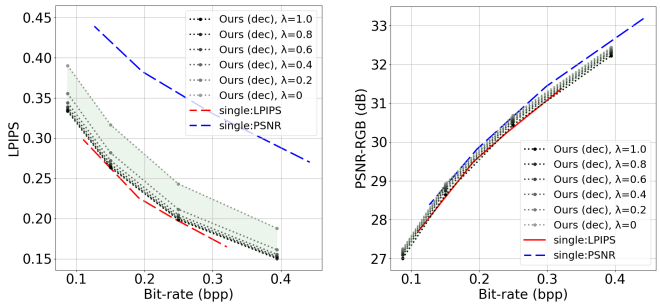
(a) LPIPS (b) PSNR

Fig. 4. Rate-distortion performance under (LPIPS, PSNR) pair with both-side prompts. Green shaded area indicates the adaptive range of our method.



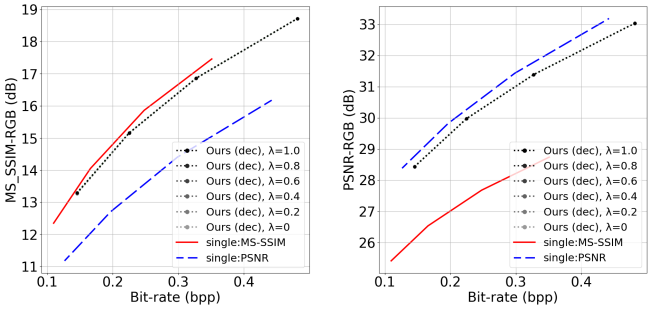
(a) MS-SSIM (b) PSNR

Fig. 5. Rate-distortion performance under (MS-SSIM, PSNR) pair with both-side prompts. Green shaded area indicates the adaptive range of our method.



(a) LPIPS (b) PSNR

Fig. 6. Rate-distortion performance under (LPIPS, PSNR) pair with decoder-side only prompts. Green shaded area indicates the adaptive range of our method.



(a) MS-SSIM (b) PSNR

Fig. 7. Rate-distortion performance under (MS-SSIM, PSNR) pair with decoder-side only prompts. Green shaded area indicates the adaptive range of our method.

As depicted in Fig. 3, in W-MSA or SW-MSA, both I_i and P_i are arranged into windows of tokens, and the collocated windows are combined in the first step of the (S)W-MSA. Specifically, both the prompt and image tokens are involved in formulating the key and value matrices, whereas the query matrix is derived solely from the image tokens only. The query matrix remains unchanged since only the output of image tokens is propagated and not that of prompt tokens. In symbols, we have

$$\begin{aligned} Q &= I\mathbf{W}_Q \\ K &= [P, I]\mathbf{W}_K \\ V &= [P, I]\mathbf{W}_V, \end{aligned} \tag{1}$$

where P, I collectively refer to the prompt and image tokens in the windows, and $[\cdot, \cdot]$ denotes concatenation along the token dimension. $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are learnable weights projecting the input into query $Q \in \mathbb{R}^{N\times d}$, key $K \in \mathbb{R}^{M\times d}$, and value matrices $V \in \mathbb{R}^{M\times d}$, where d is the number of channels of each token, N is the number of query tokens, and $M = N + \frac{N}{4}$ is the number of key (or value) tokens. Then we perform canonical self-attention as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^\top/\sqrt{d} + B)V, \tag{2}$$

where B is the learned relative positional bias matrix.

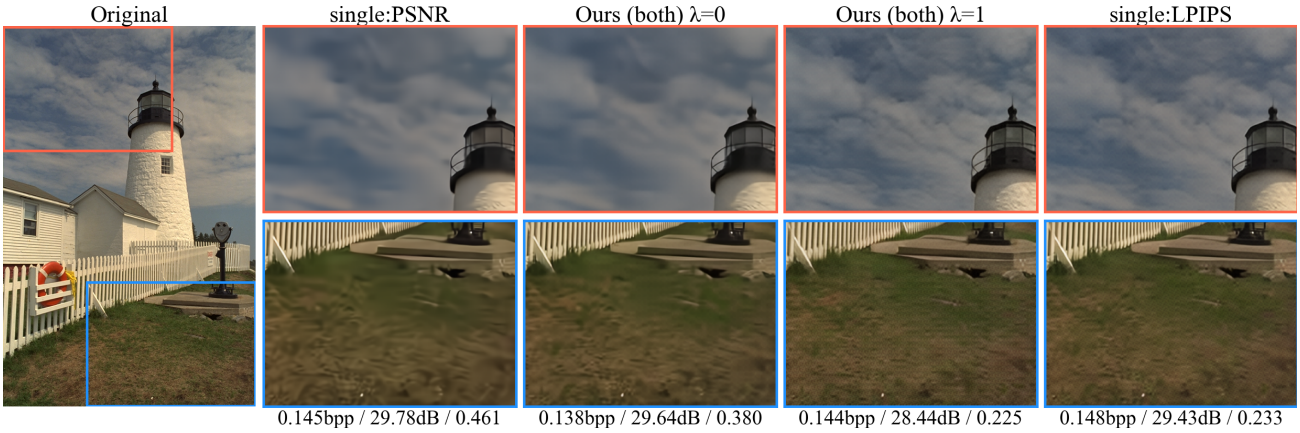
In this work, we explore two variants of the prompt-based conditioning methods for different application scenarios. The first one, both-side prompts, applies prompting to both the analysis and synthesis transforms by introducing G_{enc} and G_{dec}. In this way, the latent y is specialized for one specific trade-off λ. Although the λ needs to be specified before encoding, this approach achieves better rate-distortion performance because of the fully-customized bitstream. The other variant conditions only the decoder, i.e. no G_{enc}, resulting in a shared latent y that can be decoded with different λs. However, it has certain limitations, including showing inferior rate-distortion performance or not being able to adapt on certain cases, which are discussed and compared with both-side conditioning in Section IV.

D. Training Objective

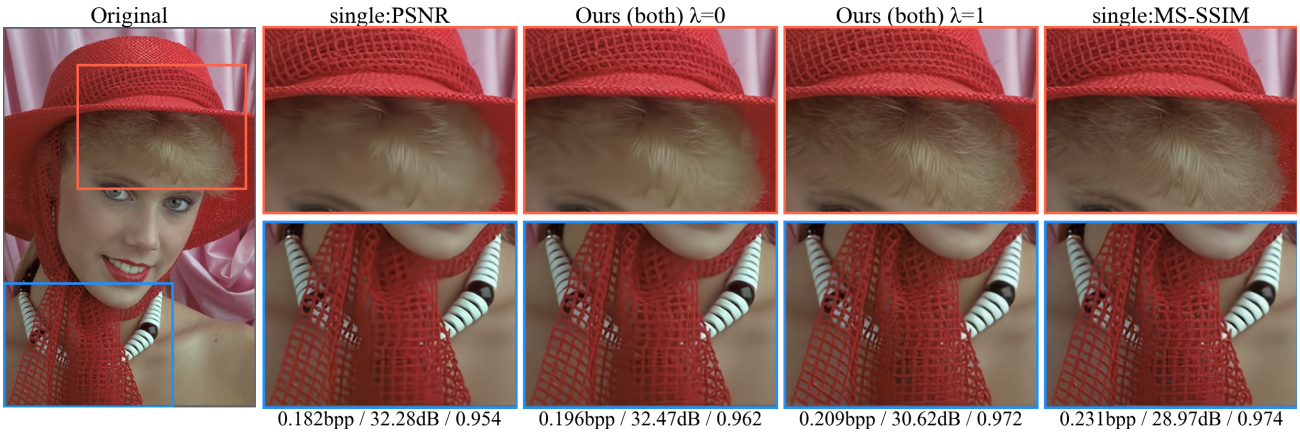
We follow the conventional training objective of learning-based image compression, which involves minimizing a rate-distortion cost, to train our model. In order to adapt between two selected quality objectives, we formulate the loss function $\mathcal{L}$ as

$$\mathcal{L} = R_\lambda \underbrace{(-\log p(\hat{z}) - \log p(\hat{y}|\hat{z}))}_{Rate} + \underbrace{\lambda d_A + (1-\lambda) d_B}_{Distortion}, \tag{3}$$

where d_A, d_B denote the distortion terms corresponding to two selected quality objectives, respectively, and R_λ controls the trade-off between distortion and rate. When λ is set to one, the distortion d_A of the quality objective A dominates the distortion term, resulting in the optimization for the quality objective A only, and vice versa. During training, λ is uniformly sampled



(a) (LPIPS, PSNR), the text below each image indicates its BPP/PSNR/LPIPS



(b) (MS-SSIM, PSNR), the text below each image indicates its BPP/PSNR/MS-SSIM

Fig. 8. Subjective quality comparison of single-metric methods and our method with both-side prompts on two pairs of quality objectives. Zoom in for better visualization.

so that the model is able to adapt the coding process to variable image quality objectives.

IV. Experiments

A. Settings

Quality Objectives. In this paper, the two quality objectives selected for adaptation are commonly-used image quality metrics. Specifically, we demonstrate the effectiveness and generalization of our proposed method on two selected quality objective pairs, (LPIPS, PSNR) and (MS-SSIM, PSNR). Note that the two objectives in either pair exhibit distinct differences from each other. Optimizing for PSNR, which is simply minimizing the mean squared error between the reconstructed and original images, leads to blurry decoded images. On the other hand, MS-SSIM considers luminance, contrast, and structure distortions all at once, and LPIPS utilizes a pre-trained VGG network [23] to measure the image distortion in the feature domain. Both MS-SSIM and LPIPS tend to produce subjectively more pleasing results than PSNR. Each quality metric has its own distortion function d, as defined below:

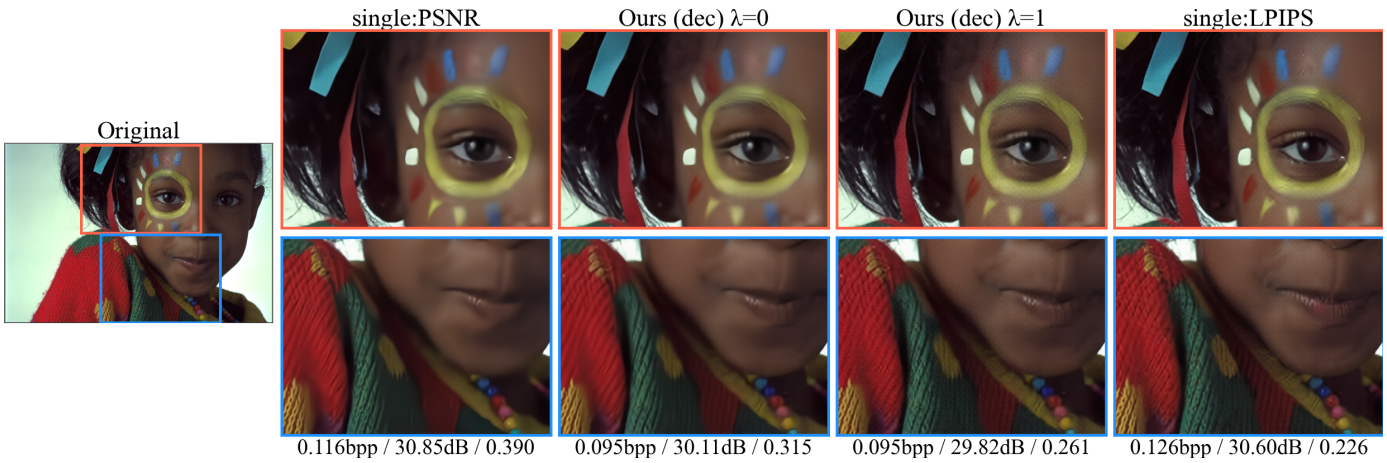
$$\begin{aligned} d_{\text{PSNR}}(x,\hat{x}) &= 0.01 \times \text{MSE}(x,\hat{x}), \\ d_{\text{MS-SSIM}}(x,\hat{x}) &= 25 \times (1 - \text{MS-SSIM}(x,\hat{x})), \\ d_{\text{LPIPS}}(x,\hat{x}) &= 2.5 \times \text{LPIPS}(x,\hat{x}). \end{aligned} \tag{4}$$

By plugging these distortion functions into Eq. (3) as d_A and d_B, we obtain the final training loss function for each objective pair.

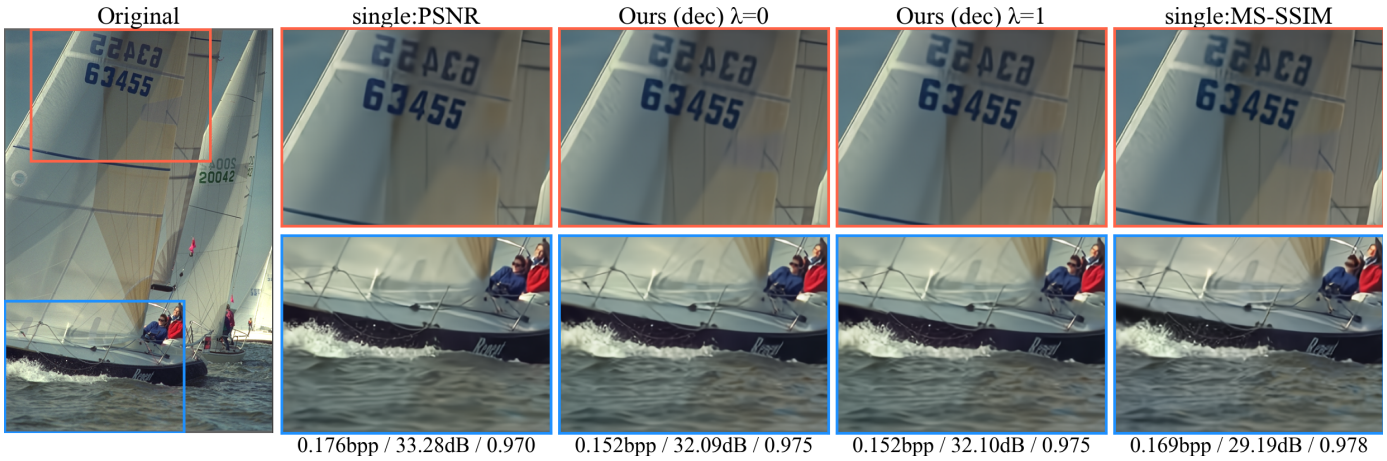
Training. For both objective pairs, we first pre-train the base codec (i.e. $N_{enc}, N_{dec}, H_{enc}, H_{dec}$) by optimizing it only for PSNR following [7] on Flicker2W dataset [24]. The resulting PSNR model is then combined with the prompt generation networks G_{enc} and G_{dec} for the next training stage, which jointly trains the entire model end-to-end using the loss function in Eq. (3) and Eq. (4). During training, we choose $R_\lambda = \{5.556, 2.857, 1.493, 0.769\}$, which is the reciprocal of the Lagrange multiplier used in [7] times 0.01. Note that R_λ is associated with the rate term. The trade-off λ is uniformly sampled between 0 and 1.

Evaluation. We evaluate the performance of our proposed method on Kodak dataset [25], which consists of 24 uncompressed images with resolution 512x768 or 768x512. The evaluation is done for both pairs of quality metrics (i.e. PSNR versus LPIPS, and PSNR versus MS-SSIM). When testing our proposed method, we choose $\lambda = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ to evaluate the capability of the proposed method in fulfilling various quality objectives.

Baselines. We compare the performance of our proposed



(a) (LPIPS, PSNR), the text below each image indicates its BPP/PSNR/LPIPS



(b) (MS-SSIM, PSNR), the text below each image indicates its BPP/PSNR/MS-SSIM

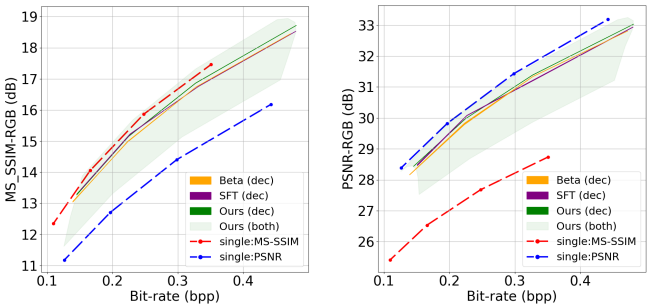
Fig. 9. Subjective quality comparison of single-metric methods and our method with decoder-side prompts on two pairs of quality objectives. Zoom in for better visualization.

method with various baselines. Firstly, we train our base codec (i.e. $N_{enc}, N_{dec}, H_{enc}, H_{dec}$) by optimizing for a single quality metric, which is denoted as *single:*$\{$*metric*$\}$ (e.g. *single:PSNR*). Note that for *single:LPIPS*, we also include a MSE term during training to prevent instability when solely optimizing for LPIPS. In addition, we compare our prompt-based conditioning method with the existing methods: **SFT** [17] and **beta conditioning** [6]. For a fair comparison, both conditioning methods are implemented on our Swin-Transformer base codec and trained following the same procedure as ours. In the case of SFT [17], we adopt the conditioning networks from [17] and apply affine transformation to the output feature maps of every STB in both the encoder and decoder following [17]. As for the beta conditioning [6], we employ the mapping function and projection to perform channel-wise shifting on the output feature maps of every Swin-Transformer layer in the decoder following [6]. In order to compare with our both-side and decoder-side conditioning methods, we construct a variant of SFT with only the decoder-side condition. Likewise, we develop a variant of the beta conditioning method by applying it to both the encoder and decoder.

B. Performance Comparison for Both-side Prompts

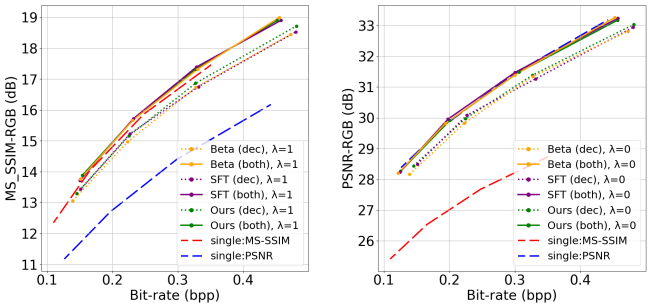
We first examine the compression performance of our proposed method with both-side prompts, where both G_{enc} and G_{dec} are utilized. For the (LPIPS, PSNR) objective pair, Fig. 4 compares the rate-distortion performance of our method with that of the single-metric methods. By adopting different λ values, our method is able to adapt to a wide range of quality objectives (see the green shaded area). When $\lambda = 0$, which translates to optimizing solely for PSNR, our method is able to reach the same performance as the single-metric method *single:PSNR*. Notably, when $\lambda = 1$, our method even outperforms marginally *single:LPIPS*, which is due in part to the increased model capacity. These results suggest that our proposed method has the capability of fulfilling a wide range of image quality objectives with a single model, while performing comparably to or even better than the single-metric methods at the two extreme λs.

Similarly, Fig. 5 shows that our method is capable of effectively trading off MS-SSIM for PSNR. Similar to (LPIPS, PSNR), our method performs comparably or superior to *single:PSNR* on PSNR with $\lambda = 0$ and *single:MS-SSIM* on



(a) MS-SSIM (b) PSNR

Fig. 10. Adaptive range comparison under (MS-SSIM, PSNR) pair with various conditioning methods.



(a) MS-SSIM (b) PSNR

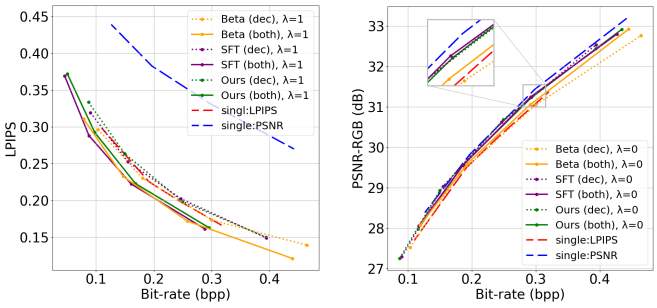
Fig. 11. Rate-distortion performance comparison under (MS-SSIM, PSNR) pair with various conditioning methods at $\lambda = 1$ or $\lambda = 0$.

MS-SSIM with $\lambda = 1$. This demonstrates the versatility of our proposed method, which is able to handle various quality objective pairs with ease and achieve consistently good performance.

Fig. 8 visualizes the decoded images produced by our method and the single-metric methods for both pairs. The variable quality objectives are also reflected on the reconstructed images. For LPIPS and PSNR (Fig. 8(a)), the two methods with higher PSNR (*single:PSNR*, Ours $\lambda = 0$) have smoother images with fewer details, while the other two with low LPIPS (*single:LPIPS*, Ours $\lambda = 1$) shows clearer edges and more refined structures. However, the latter two, when examined closely, have some artificial patterns. This may be attributed to the characteristics of the LPIPS metric. With our proposed method, the undesirable patterns can be reduced through adjusting λ to strike a balance between the two quality objectives, while the single-metric methods have to re-train the entire codec for different objectives. With MS-SSIM and PSNR (Fig. 8(b)), the results are similar. It is observed that ours with $\lambda = 0$ produces blurry reconstructed images with less distinct edges and details. In contrast, ours with $\lambda = 1$ and *single:MS-SSIM* both exhibit sharper and more visually pleasing reconstructed images despite their lower PSNR.

C. Performance Comparison for Decoder-side Prompts

Next, we analyze the compression performance of our method with decoder-side prompts alone, where only G_{dec} is utilized. Fig. 6 presents the rate-distortion results of the



(a) LPIPS (b) PSNR

Fig. 12. Rate-distortion performance comparison under (LPIPS, PSNR) pair with various conditioning methods at $\lambda = 1$ or $\lambda = 0$.

(LPIPS, PSNR) objective pair. While our method is still able to adjust the reconstructed quality with different λ values, the range of adaptation is relatively smaller in comparison with the both-side results shown in Fig. 4. At the two extreme λ values, the adapted results barely match the rate-distortion performance of the single-metric methods. This observation may be attributed to the removal of G_{enc}, thereby producing a single bitstream that needs to be decoded at different λ values.

Fig. 7 presents the rate-distortion results on the (MS-SSIM, PSNR) objective pair. Contrary to the previous results with both-side prompts, the rate-distortion performance remains somewhat fixed regardless of the given λ, with all six curves coinciding with each other. This indicates that when relying only on the decoder for adaptation, our method is unable to trade off PSNR for MS-SSIM. This result suggests that there exist some metric pairs where the metrics involved are incompatible with each other in the sense that when a single bitstream is used, the decoder-side only adaptation is not able to fulfill variable quality objectives. We further verify this through experiments with the other conditioning methods in the next section.

Fig. 9 presents the subjective quality comparison of decoded images obtained through the single-metric methods and our method on both pairs of metrics. For LPIPS and PSNR, our method with $\lambda = 1$, which corresponds to optimizing for LPIPS, displays richer details but also exhibits more prominent artificial patterns, the result of which is similar to *single:LPIPS* (Fig. 9(a)). In contrast, $\lambda = 0$ and $\lambda = 1$ produce nearly identical images on MS-SSIM and PSNR (Fig. 9(b)).

D. Comparison with Competing Methods

In this section, we compare our proposed method with the other conditioning methods, namely SFT [17] and beta conditioning [6]. First, we focus on the failure case with decoder-side prompts on MS-SSIM and PSNR, and test both baselines under the same setting of decoder-side only conditioning. Fig. 10 illustrates the effective range of adaptation. It is apparent that all methods using decoder-side only conditioning are unable to display any substantial adaptation range, which is a direct contrast to our variant with both-side prompts. This further indicates that the failure to adapt between MS-SSIM and PSNR lies in the combination of these objectives rather

than the proposed conditioning method.

Fig. 11 and Fig. 12 further compare the rate-distortion performance of all the methods at $\lambda = 1$ and $\lambda = 0$ under the both-side and decoder-side settings. For the (MS-SSIM, PSNR) objective pair, Fig. 11 shows that while all three methods with the both-side conditioning show similar performance on both metrics, our proposed prompt-based method outperforms SFT [17] and the beta conditioning [6] when using the decoder-side only conditioning. This suggests that prompting offers a better means to condition Swin-Transformer-based models. As for the (LPIPS, PSNR) metric pair, Fig. 12 shows that the beta conditioning performs the best in terms of LPIPS; however, it exhibits a significant drop in PSNR under the both-side and decoder-side settings. On the other hand, although SFT achieves a slight gain over our method in terms of LPIPS with the both-side prompts, we note that SFT introduces considerably higher complexity than our method, adding nearly 15 million model parameters as opposed to 6 million with our scheme. Overall, our proposed method strikes a better balance among the adaptation capability, rate-distortion performance, and complexity.

V. Conclusions

In this paper, we propose a practical yet efficient solution for adapting to variable quality objectives in image compression, which is achieved through inserting model-generated prompts into the Swin-Transformer based codec. It offers superior performance compared to existing methods and demonstrates consistent results across different quality objective pairs. We additionally explore two variants of our proposed method and compare them in terms of performance and applications. The ability to trade off one quality objective for another using a single model, without the need for retraining, makes our method highly versatile and applicable to a wide range of scenarios.

Acknowledgement

This work is supported by National Science and Technology Council, Taiwan under Grants NSTC 111-2634-F-A49-010- and MOST 110-2221-E-A49-065-MY3, MediaTek, and National Center for High-performance Computing.

References

[1] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[2] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11913–11924, 2020.

[3] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020.

[4] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5718–5727, 2022.

[5] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.

[6] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-realism image compression with a conditional generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2023.

[7] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," in *Data Compression Conference (DCC)*, IEEE, 2022.

[8] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[9] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[10] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 709–727, Springer, 2022.

[11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proceedings of International Conference on Learning Representations*, 2018.

[12] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[13] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[14] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[16] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10532–10541, 2021.

[17] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[18] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 606–615, 2018.

[19] X. Wang, M. Lu, and Z. Ma, "Block-level rate control for learnt image coding," in *2022 Picture Coding Symposium (PCS)*, IEEE, 2022.

[20] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 3045–3059, Association for Computational Linguistics, Nov. 2021.

[21] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 4582–4597, Association for Computational Linguistics, Aug. 2021.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] J. Liu, G. Lu, Z. Hu, and D. Xu, "A unified end-to-end framework for efficient deep image compression," *arXiv preprint arXiv:2002.03370*, 2020.

[25] E. Kodak, "Kodak lossless true color image suite (PhotoCD PCD0992)."

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

From Synthetic To Real: Enhancing Deep Learning Models With Generative Adversarial Networks For Efficient Data Utilization In Automatic Retail Stores

Cong-Ty Dang*, Vu-Hoang Tran†, Ngoc-Hoang-Lam Le‡, and Ching-Chun Huang §
* † ‡ Ho Chi Minh City University of Technology and Education, Vietnam
* E-mail: 19142415@student.hcmute.edu.vn
† E-mail: hoangtv@hcmute.edu.vn
‡ E-mail: lengochoanglam@gmail.com
§ National Yang Ming Chiao Tung University, Taiwan
E-mail: chingchun@nycu.edu.tw

***Abstract*—One of the major challenges in developing deep learning models for automatic retail stores is the availability of diverse and accurate data. Achieving high accuracy requires a large amount of data that is specifically collected and labeled for the task at hand. However, this process is time-consuming and prone to potential errors during labeling, which can affect model training. To mitigate this challenge, data augmentation methods have been widely adopted to train models when data is limited. However, the effectiveness of data augmentation is not always guaranteed, as it still relies on the original data. To address this limitation, a new approach has been developed that utilizes synthetic data generated through 3D scanning of objects. However, the challenge lies in ensuring that this synthetic data accurately represents real-world data to maintain model accuracy. In this paper, we propose an approach that integrates adversarial learning into the YOLOv8 model and trains it using synthetic data. We then apply this method to automatic retail stores to demonstrate its efficacy in real-life scenarios. Experimental results on the AIC-22 dataset reveal significant improvements compared to the original YOLOv8 version. The proposed approach achieves around 6.7% increase in precision, a 7.1% improvement in recall, and a 10.3% enhancement in mAP while maintaining the same inference speed of 127 fps on the RTX 3090Ti.**

Index Terms—object detection, domain adaptation, automatic retail stores

I. Introduction

Currently, traditional methods such as bar-code or QR-code scanning are still widely used at payment counters, but they are not highly efficient or fast. This is evident from the long queues of customers waiting at supermarkets and convenience stores. To address this issue, various studies have been conducted, including the use of image processing to recognize products within enclosed chambers [1]. However, this method is not fast as it can only identify one product at a time, as depicted in Figure. 1. Moreover, it does not provide a user-friendly experience when implemented at payment counters, resulting in its limited popularity.

Another approach involves leveraging deep learning models and cameras to identify products while customers are still holding them. This solution focuses on developing the model and dataset rather than requiring complex hardware setups. However, deep learning models rely on large and accurate datasets, which can be time-consuming and labor-intensive to create. As an alternative, synthetic data [2] has been considered, as it can be automatically generated and is independent of the original data. Synthetic data enables the creation of a sufficiently diverse and minimally erroneous dataset. However, effectively utilizing synthetic data poses a challenge due to the differences between synthetic and real-world images, particularly in terms of domain dissimilarity, which affects accuracy.



Fig. 1. The isolated condition (left) and the real-life condition (right)

To overcome this limitation, a novel approach is proposed in this paper, which incorporates domain adaptation techniques. The proposed method combines the GAN idea [3] with the YOLOv8 framework [4] during the training process to enhance object detection accuracy using synthetic data. The key idea is to identify the shared domain, which represents the common features between synthetic and real data, and enable the model to extract real features within the synthetic domain. In this approach, GAN is responsible for determining these domains, which subsequently guides YOLOv8 to learn from the identified shared feature domain. As a result, the model is trained solely on synthetic data but performs effectively on real-life data through unsupervised domain adaptation for object detection.

By utilizing synthetic data, the need for a large labeled dataset is eliminated, reducing the time and cost associated with data collection and labeling. The combination of GAN

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

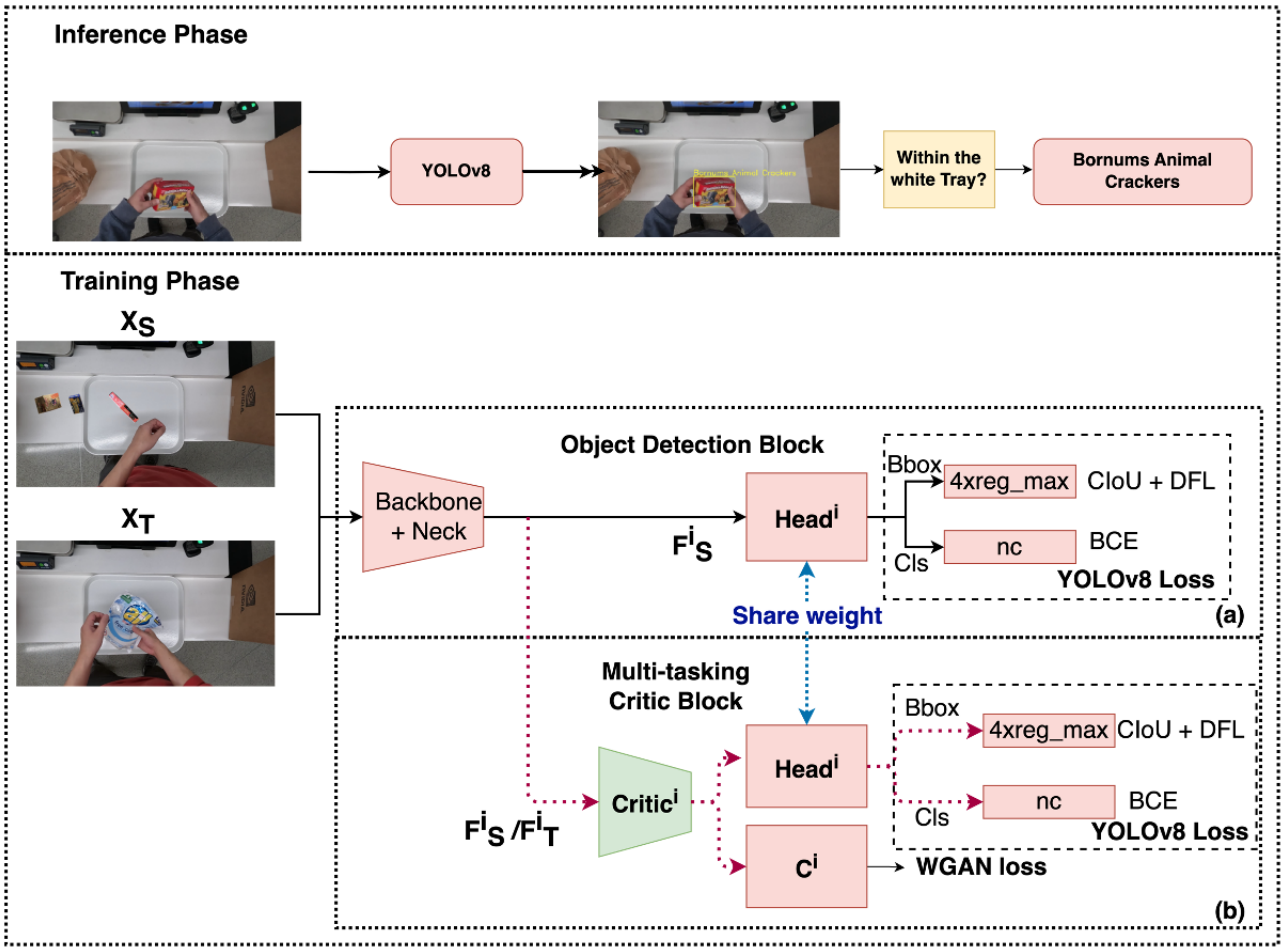


Fig. 2. The overall system

and YOLOv8 during training enhances the accuracy of object detection, thereby improving the reliability of the system. Additionally, the use of unsupervised domain adaptation ensures that the model performs well on real-life data, further enhancing its efficiency in a retail environment. In summary, our contributions can be summarized as follows: (1) integrated adversarial learning into YOLOv8 framework for automatic retail system; (2) proposed multi-critic framework to embed the idea of WGAN into existing multi-head detection networks; (3) Proposed a multi-tasking Critic structure to solve the problem of domain adaptation in object detection and recognition applications.

II. Proposed Method

As shown in Fig. 2, our proposed system utilizes the YOLOv8-based Detection model, which is then enhanced through the application of a domain adaptation technique. Due to the inherent differences between the synthetic and real feature domains, if the detection model is only trained on the synthetic domain, it cannot perform well on the real domain. Domain adaptation allows for the transfer of knowledge from a well-trained source domain to a related target domain. To achieve this, we employ an adversarial learning strategy inspired by generative adversarial networks (GANs) to minimize the domain discrepancy. This strategy aims to generate source and target feature distributions that are indistinguishable from a well-trained domain discriminator. By incorporating these techniques, our method not only improves the performance of the YOLOv8 model during training with synthetic data but also maintains the inference speed. It is important to note that the training phase and the inference phase have different architectures. In order to simplify the process, the region of interest (ROI) is fixed. During inference, if any product is detected by YOLOv8 within the ROI, the system generates an output containing the product name. In the following sections, we will provide a detailed explanation of each block in the system design.

A. Object Detection Block with Multi-Critic based Adversarial Learning

As mentioned previously, during the training phase, only the synthetic dataset is utilized. To address the domain discrepancy between real and synthetic images, an adversarial learning technique, specifically WGAN [5] [6], is employed. This tech-

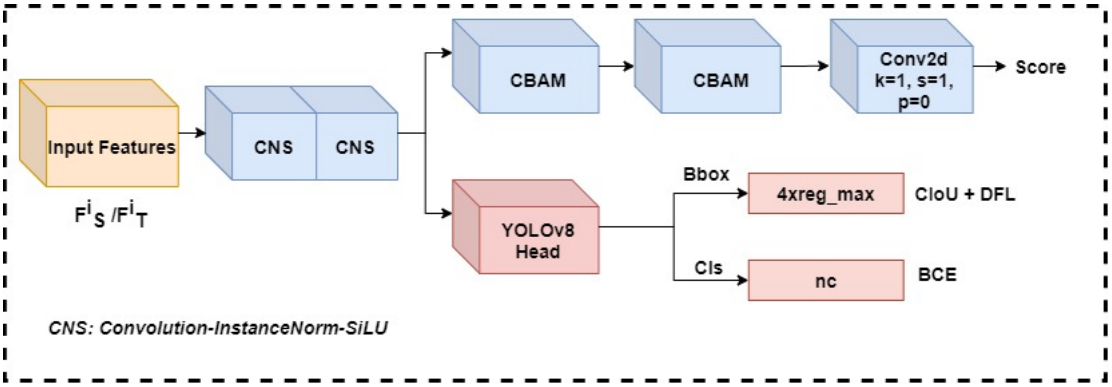


Fig. 3. The Multi-tasking Critic architecture

nique aims to minimize the feature domain differences between the real and synthetic images during training. The training architecture is illustrated in Fig. 2. In the architecture, X_T and X_S represent real images (target domain) and synthetic images (source domain), respectively. F_T^i and F_S^i correspond to the features extracted from the Backbone and Neck of YOLOv8 for the real and synthetic images. The variable i denotes the serial number of YOLOv8's Head. In this paper, in order to both minimize the difference on feature domains with different scales and have a framework suitable for multi-head detection system, we apply adversarial learning on all Heads of YOLOv8. Therefore, each Head will have a corresponding Critic module. The structure of these Critic modules is the same as shown in Fig. 3, so for the sake of simplicity, we only present one Critic module and one YOLOv8's Head in Fig. 2. With the incorporation of the WGAN technique, the total losses of YOLOv8 are transformed as shown in (1). YOLOv8 network will be trained to minimize this Loss function in order to both solve the detection and recognition problem, and to be able to extract common features between the two domains of synthetic and real images.

$$Det_{Loss} = YOLOv8_{losses} + \gamma_1 * Adv_{loss}, \quad (1)$$

where, $YOLOv8_{losses}$ is the original loss of YOLOv8 [4] as in (2), it is used to help the model still work on the detection task. Adv_{loss} is the loss of WGAN as calculated in (3) to make sure features extracted by YOLOv8 from real and synthetic images are not so different, γ_1 is the balance factor between $YOLOv8_{losses}$ and Adv_{losses}.

$$YOLOv8_{losses} = L_{CIOU} + \alpha_1 * L_{DLF} + \alpha_2 * L_{BCE}, \quad (2)$$

where, L_{CIOU} [7] is used to guide the predicted bounding box towards the ground truth, while L_{DLF} [8] is introduced to encourage the network to build a general distribution for that bounding box. This means introducing uncertainty about the boundaries of the object in the dataset into the learning process. L_{BCE} is the binary cross-entropy loss classification loss proposed in [9] to establish information connection between the classification task and the object detection task during training. α_1 and α_2 are the balance weights.

$$Adv_{loss} = \frac{1}{3}\sum_{i=1}^{3}[\mathbb{E}_{F_S^i} C^i(F_S^i) - \mathbb{E}_{F_T^i} C^i(F_T^i)]. \quad (3)$$

In (3), $C^i(F_S^i)$ and $C^i(F_T^i)$ are the scores, evaluated by i^{th} Critic model, assigned to synthetic and real features extracted from the i^{th} YOLOv8's Head, respectively. YOLOv8 has 3 Heads corresponding to 3 different scales, so our network architecture will also have 3 Critics and we will average the scores evaluated from these critics.

B. Multi-tasking Critic Block

The Critic will work opposite of YOLOv8, it will take the features from YOLOv8 and try to evaluate whether the feature is from real or synthetic images. Therefore, the objective of Critic will be opposite of YOLOv8 as shown in (4), the plus sign in formula (1) is replaced by a minus sign to maximize the difference between real and synthetic features. Thus, if the YOLOv8 can extract features that fool the Critic, those features will be able to represent both image domains, and when the YOLOv8 performed well on synthetic data, it will also work well on real data. The Critic model is designed to be compatible with the speed and learning capabilities of YOLOv8 as shown in Fig. 3, which include 2 blocks CNS (Convolution - InstanceNorm - SiLU) for extract features, then those features are fed into 2 branches for two different tasks. Therefore, we have named our Critic model as the multi-tasking Critic to reflect its ability to perform multiple tasks simultaneously. The first one consists of 2 CBAM Block [10] (Convolution Block Attention Module) and a final Convolution layer with a kernel size of 1, stride of 1 and no padding. The second brand is copied directly from YOLOv8's Head. Our design is quite different from the traditional Critic design, which has only a first branch, because the nature of our problem is different from the traditional problem. In the traditional problem, the feature of the whole image is taken into account when trying to minimize the difference between the two regions. But in our problem, the synthetic image is created by stitching the synthetic objects into the real background. Therefore, to orient the Critic to focus only on evaluating features extracted from

objects, a second branch is added. This results in the Loss Function for the multi-tasking Critic still consisting of two components as shown in (4). With a small network capacity as shown in Fig. 3, in order to perform well on both tasks in the loss function (4), the multi-tasking Critic is forced to focus solely on extracting and evaluating object features without considering the presence of the background.

$$Critic_{Loss} = YOLOv8_{losses} - \gamma_2 * Adv_{loss}, \quad (4)$$

C. Inference

During the training, the Critic will be added to the system to train the YOLOv8 on synthetic data. However, during inference, the Critic is removed from the system, and only the detection block is retained, as shown in Fig. 2. This ensures that the system maintains the fast inference speed of the original YOLOv8 while also improving performance. During testing, a real image is input into the system, and after passing through YOLOv8, the object's location is obtained. If the object falls within the fixed ROI, the system outputs the object's name.

III. Experiment Results

A. Dataset details

The training data used in this study was provided by NVIDIA as part of the AIC22 Track 4 dataset [2] [11]. This dataset consists of 116,000 images, each accompanied by its corresponding mask. Each image represents a single object and is classified into one of 116 classes. These images are generated from 3D scanned models of objects, as illustrated in Fig. 4. The dataset covers a wide range of conditions, including variations in brightness, resolution, orientation, and environmental noise.

However, the provided synthetic data is not directly suitable for training an object detection model since the objects are presented separately. To address this, preprocessing steps were performed to generate a dataset suitable for object detection. First, product images and their corresponding masks were randomly selected and merged onto backgrounds obtained from video clips that did not contain any products. This integration process resulted in a final dataset image, as shown in Fig. 4. Additionally, histogram equalization was applied to address brightness variations, ensuring consistent lighting conditions across the images.

To address class imbalance, the number of occurrences for each class was limited to 100. This approach prevents certain classes from dominating the training data and helps maintain class distribution balance. As a result, the training dataset consists of approximately 7,500 images, with each image containing 3 to 5 products. The validation set followed the same approach and consists of around 2,000 images.

In contrast, during the testing phase, a real-world dataset extracted from video clips of payment scenes is utilized. This dataset, also provided by NVIDIA, encompasses over 100 different types of products. In the video clips, the camera is positioned above the checkout counter, facing downwards, while a customer mimics the checkout process by naturally "scanning" objects placed in front of the counter. To represent the testing area for the AI model, a white tray is placed beneath the camera. Each video clip captures one or more complete scanning actions involving multiple items. For the testing set, frames were extracted from the video clips and manually labeled. This process ensures that each image in the testing set is annotated with the corresponding groundtruth labels. The testing set consists of approximately 1,500 images.



Fig. 4. The provided dataset (left) and the dataset for detection (right)

B. Implementation details

The outstanding dedication of the Ultralytics [4] team has played a significant role in achieving the impressive performance of the YOLOv8 model. Therefore, the author has selected this model as the target for improvement in terms of accuracy and as the baseline for modifying the training architecture to align with the proposed method.

To enhance the training process, several techniques are employed with the Critic model. These include the use of the RMSprop optimizer with an initial learning rate of 1e-3, an alpha value of 0.99, and weight sharing between the Head of YOLOv8 and the Head of Critic. Besides, the proposed loss is balanced by assigning specific balance weights: γ_1 is given a weight of 10, γ_2 a weight of 100, α_1 a weight of 0.2, and α_2 a weight of 0.06. The training procedure of the proposed network is followed as in [5].

C. Results

The final results are shown in Table I, which used Precision, Recall, and mAP50-95 for evaluation. The term 'Source only' refers to the model training without adversarial learning, the symbol w means that the weight between YOLOv8's Head and the critic's Head is shared, m means that the mixup augmentation method is implemented to augment the training data. It is easy to see that training with adversarial learning greatly enhances precision and mAP. However, it is important to note that the system starts to focus more on minimizing the difference between the features of the objects, instead of the background, only when the head weights of YOLOv8 and Critic are shared and the data is augmented. This leads to a significant increase in the Recall rate. Overall, the proposed system achieves an increase of approximately 6.7% in precision, a 7.1% improvement in recall, and a 10.3% enhancement in mAP compared to the baseline model with the same settings, this demonstrates the effectiveness of our approach in improving object detection accuracy. Moreover,

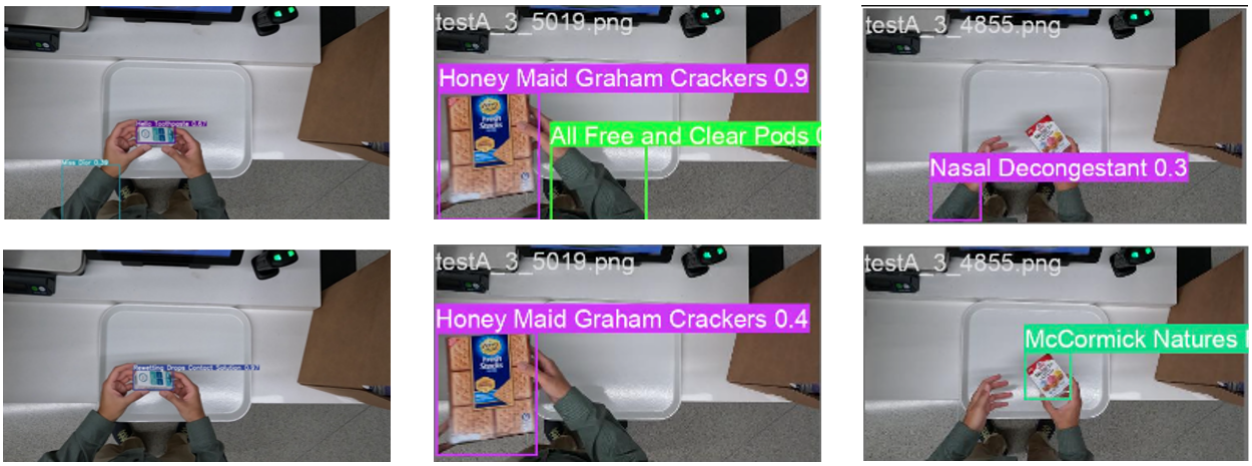


Fig. 5. Compare results between original YOLOv8 (first row) and Our method (second row)

our system is fully trainable as semi-supervised learning by adding some labeled real data to the training dataset. Table II demonstrated that by adding only 10% of the real labeled data, the system performance was increased dramatically. This proves a problem that, the proposed method helps to build the automatic retail stores system more effectively, we can completely pre-train the system with sythetic data, then it only takes a very small amount of effort to label the real data and we can have a system with very high accuracy.

TABLE I
COMPARISON OF PROPOSED METHOD WITH BASELINES USING UNSUPERVISED TRAINING

Method	Learning style	Precision	Recall	mAP50-95
YOLOv8	Source only	0.660	0.496	0.409
YOLOv8^{m}	Source only	0.708	0.578	0.470
Our	Unsupervised	0.723	0.501	0.451
Our^{w}	Unsupervised	0.762	0.539	0.495
Our$^{w\ m}$	Unsupervised	**0.775**	**0.649**	**0.573**

TABLE II
COMPARISON OF PROPOSED METHOD WITH BASELINE USING SEMI-SUPERVISED LEARNING

Method	Learning style	Precision	Recall	mAP50-95
YOLOv8^{m}	10% labeled real data	0.924	0.826	0.784
Our$^{w\ m}$	10% labeled real data	**0.935**	**0.887**	**0.831**

For an intuitive look, the results of the visual comparison between the proposed method and the baseline are shown in Fig. 5. It is easy to see that, when adversarial learning is not integrated, the baseline often confuses the objects to be detected with the human hand. This is understandable, because the baseline is only learned with synthetic data, and it is possible that the human hand in the test image is quite similar to the characteristics of a certain object in the synthetic data. By using adversarial learning, the proposed system tries to find the most common features between real and synthetic objects, thereby excluding these cases. However, our system still has some cases of mis-classification and mis-detection as shown in Fig. 6. In the image on the top, the shopper's hand accidentally came close to the shoes, resembling the act of holding an object, and the shape of the shoes now bears a striking resemblance to Advil Liquid Gel, leading to a false detection by our system. However, this problem can be completely eliminated at the post-processing stage, since the detected object is outside the ROI (white tray) area. In the image on the bottom, the shopper's right hand covers most of the object when passing it from one hand to the other, resulting in the system not being able to detect the object. And this can also be solved because in the next few frames, the shopper has to show the object to the camera.



Fig. 6. Some failure cases

IV. Conclusion

In conclusion, this paper addresses the challenge of limited data for training deep learning models by proposing a novel approach that integrates Adversarial Learning into the YOLOv8 model. The approach utilizes synthetic data generated through 3D scanning of objects, aiming to overcome the limitations of traditional data augmentation methods. The study demonstrates the effectiveness of the proposed method in improving model accuracy and performance. Experimental results on the AIC-22 data-set showcase a 6.7% increase in precision, a 7.1% improvement in recall, and a 10.3% enhancement in mAP compared to the original YOLOv8 version. Importantly, these improvements are achieved without sacrificing inference speed, with the model reaching an impressive 127 fps on the RTX 3090Ti. The practical application of this method in automatic retail stores further validates its real-life efficacy. By leveraging synthetic data and integrating Adversarial Learning, this approach presents a promising solution to the data limitations faced by deep learning models, paving the way for enhanced accuracy and performance in various applications.

References

[1] S. T. Bukhari, A. W. Amin, M. A. Naveed, and M. R. Abbas, "Arc: A vision-based automatic retail checkout system," 2021.

[2] Y. Yao, L. Zheng, X. Yang, M. Napthade, and T. Gedeon, "Attribute descent: Simulating object-centric datasets on the content level and beyond," *arXiv preprint arXiv:2202.14034*, 2022.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[4] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[5] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[6] W.-Z. Zheng, V.-H. Tran, and C.-C. Huang, "D2na: Day-to-night adaptation for vision based parking management system," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1573–1577.

[7] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," 2019.

[8] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020.

[9] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 3490–3499.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018.

[11] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, S. Prajapati, A. Li, S. Li, K. Kunadharaju, S. Jiang, and R. Chellappa, "The 7th AI City Challenge," in *CVPRW*, June 2023.

Virtual Garment Fitting Through Parsing and Context-Aware Generative Adversarial Networks with Discriminator Group

Wei-Hong Su, Sze-Ann Chen, Chen-I Chin, Hsu-Feng Hsiao
Department of Computer Science
National Yang Ming Chiao Tung University, Hsinchu, Taiwan
{piercesu.cs08, dudhfysibw1027.cs08, harry1024.cs08}@nycu.edu.tw, hillhsiao@cs.nycu.edu.tw

Abstract — **Owing to the rapid growth of the e-commerce industry, image-based virtual try-on has emerged as a popular research topic in recent years. Despite the introduction of multiple approaches to achieve this concept, there remains ample scope for research and improvement. In this regard, Generative Adversarial Networks (GANs) demonstrate a framework possessing immense potential for subsequent development. Nonetheless, the generated images reported in the literature often manifest blurred edges between semantic regions, thereby diminishing the credibility of results. Furthermore, the generation of try-on images may retain the original shape of the upper body clothing on the model mistakenly, such as the length and tightness of the torso, rather than adapting to the shape of the target clothing. In this paper, we propose a more comprehensive architecture to overcome the limitations of GAN-based approaches, which includes the following contributions. First, we introduce a new parsing and context generator that takes into account the warped binary mask of the geometric matching image of the target clothing. The outputs of this generator incorporate the generation of human parsing images that correspond to the generated try-on images. Moreover, we have designed a novel discriminator group that is specifically focused on judging whether the generated image is a reasonable representation of the specific clothing being worn. According to the experimental results, our method effectively exhibits better synthesis quality and remedies the common challenges encountered while using GANs for virtual try-on.**

I. INTRODUCTION

In the passage of recent years, the e-commerce industry has shown a swift and pronounced expansion, characterized by a growing number of individuals migrating towards online outlets for their routine acquisitions, including apparel [1]. Yet, online clothing purchases harbors a glaring disadvantage when compared with conventional retail experiences: the lack of physical interaction with prospective purchases. This absence of tangible evaluation escalates the likelihood of consumers seeking refunds upon realizing that their ordered garments fail to fulfill their initial visualized expectations. Such circumstances incited the inception of "Virtual Fitting Rooms," a concept offering the potential for online consumers to visualize their appearance in a selected item whilst navigating digital catalogues.

Image-based virtual try-on techniques have emerged as an exceptionally viable alternative [2], outshining methods like 3D modeling [3, 4, 5], owing to their non-reliance on costly hardware and graphics rendering equipment. These methodologies predominantly aim to fabricate try-on images by utilizing two inputs: an image of the potential wearer and another of the selected attire, endeavoring for maximal authenticity. The assessment of a virtually generated try-on image's efficacy fundamentally revolves around three cornerstones: 1) the ability of the image to retain the target individual's identity adequately, 2) the faithfulness of the texture or patterns of the synthesized garment to the target item, and 3) the degree to which the shape of the generated clothing aligns with the target item. Overlooking any of these critical elements would substantially jeopardize the integrity of the resulting images.

Particularly, C-VTON [6] demonstrates proficiency in preserving garment textures and patterns, achieved through the incorporation of an image warping phase within its proposed architecture, and context aware normalization—a technique that empowers the generator in the GAN [7] structure to utilize image context and furnish it with essential contextual insights. Nonetheless, the clothing shapes in its outcomes frequently fall short of authenticity as they tend to echo the shapes of the garments worn in the target person image, demonstrating less

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

sensitivity to distinct shape attributes such as fit, length, and neckline.

In response to these challenges, this paper introduces the Parsing and Context Driven Virtual Try-On Network (PC-VTON), an architectural innovation inspired by C-VTON. This model encompasses an advanced parsing and context generator that factors in the warped binary mask of the geometric matching image of the target attire. The outputs of this generator facilitate the creation of human parsing images that synchronize with the generated try-on images. Additionally, we have proposed an unprecedented group of discriminators, purposefully dedicated to determining whether the generated image accurately mirrors the specific attire depicted. As affirmed by our experimental findings, our methodology exhibits superior synthesis quality, effectively counteracting the commonplace obstacles faced when leveraging GANs for virtual try-on applications.

II. Related Work

This section provides an overview of significant contributions in the field, classifying them into three categories: techniques involving a garment warping stage, methods utilizing latent space interpolations, and approaches employing separate encoding of distinct features. By analyzing these categories, we aim to recognize both the strengths and weaknesses that define existing methodologies, setting the stage for the introduction of our novel approach.

A. Techniques Involving a Garment Warping Stage

SwapNet [8] embodies a two-tier process. Initially, a warping network is employed to align the garments with the pose of the potential wearer. Subsequently, a U-net architecture is deployed to generate the virtual try-on image. Although it manages to spawn a variety of poses, the system is constrained to a coarse generation of clothing content, compromising the precise replication of patterns or designs.

ACGPN [9] enhances the performance of the generator network by comparing network-generated garments with their warped counterparts, utilizing pixel-to-pixel L1 loss as a better reference. This, coupled with perceptual losses such as VGG loss, results in improved outputs. The dataset requires both pose information and product images of the garments. The technique manages to faithfully recreate clothes with text or patterns, but its effectiveness decreases in other domains, such as the replication of body parts and clothing shapes.

CP-VTON [10] incorporates a fully learnable Thin-Plate-Spline transformation [11] in its geometric matching stage. This approach demonstrates better alignment with the target body pose, especially when compared to prior works that often relied on hand-crafted keypoint-based shape matching. However, in its try-on module, which employs a U-Net architecture, the system does not directly produce an image of the target person wearing the desired clothing. Instead, it pastes the try-on clothing onto the target person using a generated clothing mask.

C-VTON [6] leverages a two-tier generative adversarial network structure to enable virtual try-on functionality. The first stage involves a Body-Part Geometric Matcher (BPGM) mapping product images of clothes onto a given human pose via an unconventional geometric method. The subsequent stage's generator utilizes the output of BPGM, densepose information of the human pose, and the original image (with the upper garment region excised) to generate a virtual try-on image. While this technique succeeds in accurately reproducing patterns and texture of the try-on items, its performance is less than satisfactory in terms of shape restoration (e.g., fit and length).

B. Methods Utilizing Latent Space Interpolations

VOGUE [12] constructs try-on images by interpolating and generating images in latent space of a StyleGAN [14] network with the assistance of produced human parsing images, based on the images of the individual and the clothing being tried on. Despite its potential, this approach could lead to an amalgamation of the target clothing with the attire originally worn by the model, thereby inducing inaccuracies in texture and pattern details.

FashionTex [13] seeks to condition virtual try-on image synthesis using both text and texture. By harnessing the generative capabilities of a pretrained StyleGAN [14], it maps input images into the model's latent space using an e4e encoder [15]. The system further introduces a latent code editor conditioned by text and texture embeddings. While it can generate fashion images influenced by text and texture, its omission of clothing product images as a synthesis constraint prevents it from serving as a comprehensive virtual try-on solution.

C. Approach Employing Separate Encoding of Distinct Features

ADGAN [16] segments and concatenates the encoded style codes of diverse parts from the human parsing. Upon applying

these codes to a particular human pose via an AdaIN layer, it spawns images of the person adopting the specified posture. By swapping the encoded part of the upper garment in the style code, the effect of changing clothes is enabled. However, it can only roughly replicate the shape and pattern of the attire.

III. The Proposed Methods

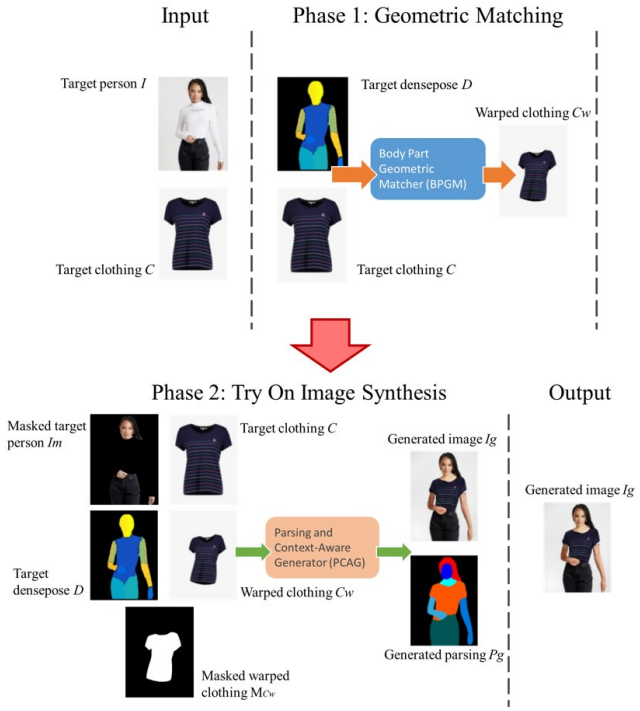


Fig. 1 Overview of the proposed PC-VTON architecture.

This research introduces the Parsing and Context Driven Virtual Try-On Network (PC-VTON), a dual-phase virtual fitting framework, consisting of Phase 1: Clothing Image Warping and Phase 2: Virtual Try-On Generation. The process flow is depicted in Fig. 1. In Phase 1, the targeted attire's image is warped to align with the densepose representation of the prospective wearer's body posture. Subsequently, the warped garment image, obtained from Phase 1, is processed through a generator, accompanied by auxiliary information, including a masked RGB image of the target person, a densepose image of the target person, an unwarped frontal image of the clothing item, and the binary mask image of the warped clothing item, in Phase 2. Ultimately, the generator produces a virtual try-on image showcasing the subject wearing the targeted attire.

A. Phase 1: Clothing Image Warping

The proposed PC-VTON architecture employs the BPGM [6] for the first phase. The BPGM, shown in Fig. 2, processes a densepose image and a frontal view clothing item image, encoding them with two dedicated encoders, E_1 and E_2, into feature representations $\psi_{E_1} \in \mathbb{R}^{d_f \times w_f \times h_f}$ and $\psi_{E_2} \in \mathbb{R}^{d_f \times w_f \times h_f}$, where w_f and h_f represent the spatial dimensions of the feature representations, and d_f signifies the number of output channels. The feature representations are subsequently normalized channel-wise into a unit L2 norm, spatially flattened, and restructured into a matrix $\psi_{\{E_1,E_2\}} \in \mathbb{R}^{d_f \times w_f h_f}$, serving as the basis for computing the correlation *Corr* between the two feature representations.

$$Corr = \psi_{E_1}^T \psi_{E_2} \in \mathbb{R}^{(w_f h_f) \times (w_f h_f)} \quad (1)$$

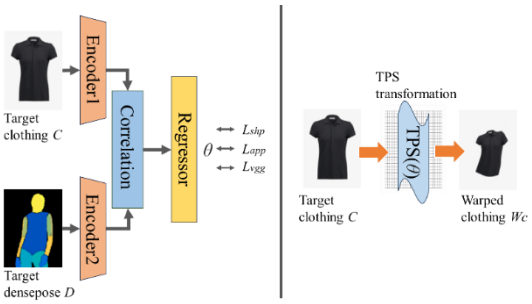


Fig. 2 Workflow of the Body-Part Geometric Matcher.

The correlation matrix is then processed through a regressor that predicts the Thin-Plate-Spline (TPS) transformation [11] parameters θ, a $2n^2$ dimensional vector accounting for the x and y offsets of an $n \times n$ grid, symbolizing how the frontal view clothing item image should be warped to conform to the provided densepose image. The BPGM's parameters are trained according to the following metrics:

- The shape loss L_{shp} calculates the L1 loss between the mask image of the warped frontal view clothing item and the clothing region's mask of the source person image I, i.e.:

  $$\mathcal{L}_{shp} = \|M_w - M_c\|_1 = \|T_\theta(M_f) - M_c\|_1 \quad (2)$$

  where M_c is the mask of the clothing item region in I; M_w is the mask of the warped frontal view of the clothing item; M_f is the mask of the frontal view of the clothing item and T_θ is the TPS transformation with parameters θ.

- The appearance loss L_{app} compares the image content of the torso area in the warped clothing item image C_w and the source person image I, i.e.:

  $$\mathcal{L}_{app} = \|C_w \odot M_t - I \odot M_t\|_1 \quad (3)$$

  where M_t is the binary mask of the torso, and $\odot$ is the Hadamard product operation.

- The perceptual loss L_{vgg} compares the high-level differences, such as image content, between the warped

clothing item image C_w and the source person image I, i.e.:

$$\mathcal{L}_{vgg} = \sum_i^m \lambda_i \|\Phi_i(C_w \odot M_t) - \Phi_i(I \odot M_t)\|_1 \quad (4)$$

where Φ_i is a feature map generated before each max-pooling layer (before $m = 5$) of a pretrained VGG19 model, and λ_i is its corresponding weight.

Thus, the BPGM is optimized to minimize the following joint loss function:

$$\mathcal{L}_{BPGM} = \lambda_{shp}\mathcal{L}_{shp} + \lambda_{app}\mathcal{L}_{app} + \lambda_{vgg}\mathcal{L}_{vgg} \quad (5)$$

where λ_{shp}, λ_{app}, λ_{vgg} represent the respective weights of the three loss functions in order.

B. Phase 2: Virtual Try-On Generation

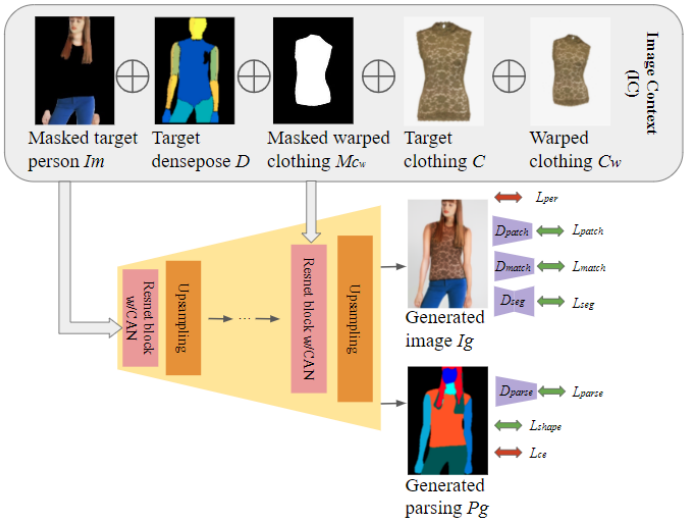


Fig. 3 Overview of the proposed PCAG in phase 2.

Phase 2, the proposed try-on image synthesis phase, adopts a GAN structure comprising a single generator, the Parsing and Context Aware Generator (PCAG), and four discriminators, which forms a discriminator group: the segmentation discriminator (D_{seg}), the match discriminator (D_{match}), the patch discriminator (D_{patch}), and the human parsing discriminator (D_{parse}).

PCAG design. The proposed PCAG, shown in Fig. 3, takes as input an RGB image of the target person where the upper body and background are masked, a corresponding densepose image, a frontal image of the targeted attire, a warped image of the same (output of Phase 1), and an additional binary mask of the transformed garment. It then produces a virtual try-on image of the target person adorned in the target attire along with a corresponding 16-channel human parsing image. The proposed PCAG incorporates context aware normalization (CAN) layers, introduced in [6], within a series of modified ResNet blocks and 2× upsampling layers allows the generator to integrate essential contextual information from the input images through normalization at various resolutions. Each ResNet block repeats the processes of batch normalization and CAN operations twice. The concept of generating human parsing images alongside RGB try-on images draws inspiration from [12], which demonstrates how the application of losses—based on human parsing images that are generated simultaneously and serve in an auxiliary capacity—can enhance the performance of try-on image generation, particularly in aspects related to semantic regions and context.

Training of the PCAG. The PCAG training involves losses that can be categorized into two groups:

- **Losses that do not involve discriminators:**

1. The perceptual loss L_{per} compares the high-level differences, such as image content, between the generated try-on image and its corresponding ground truth image, i.e.:

$$\mathcal{L}_{per} = \sum_i^m \lambda_i \|\Phi_i(I_g) - \Phi_i(I)\|_1 \quad (6)$$

where Φ_i denotes a feature map, generated before each max-pooling layer (before $m = 5$) of a pre-trained VGG19 model. λ_i is the corresponding weight. I_g and I symbolize the synthesized image and the ground truth image, respectively. This loss is calculated solely in instances where the ground truth is extant, which applies to the generation of virtual try-on images that depict the original attire worn by the target individual.

2. The shape loss L_{shape} quantifies the L2 loss between the upper body clothing region in the generated parsing image and its counterpart in the image warped via BPGM, i.e.:

$$\mathcal{L}_{shape} = \|M_{gc} - M_w\|_2 \quad (7)$$

where M_{gc} designates the region encompassing the upper-body attire within the generated parsing image, while M_w signifies the analogous segment in the BPGM-warped clothing item image. Through our experimental results, we determined that applying L2 loss yields better outcomes compared to L1 loss. The latter appears to overly constrain the generation of garment parsing by adhering too strictly to the input BPGM-warped mask, often neglecting the nuances of the human body pose.

3. The cross-entropy loss, L_{ce}, evaluates the dissimilarity between the probability distributions of the generated human parsing image and the corresponding ground truth. Formally, it is articulated as:

$$\mathcal{L}_{ce} = -\sum_k^d \sum_i^p log\left(\frac{\exp\left(P_g^{(k,i)}\right)}{\sum_h^d \exp\left(P_g^{(h,i)}\right)} P^{(k,i)}\right) \quad (8)$$

Herein, d represents the number of labels, while p denotes pixels in the image. P_g corresponds to the predicted human parsing, and P is the ground truth human parsing. This loss is computed when ground truth is available, which is when virtual try-on images that depict the original attire worn by the target individual are generated.

- **Adversarial losses based on a respective discriminator:**

1. The segmentation discriminator, denoted as D_{seg}, derived from [6], examines whether an image is real by predicting per-pixel segmentation maps S and their origin. Given an authentic image I, or a synthesized image I_g, the discriminator predicts a tensor of dimensions $w \times h \times (d+1)$, where the initial d channels encompass information about the segmentation of bodily regions, while the terminal channel signifies the authenticity of each pixel (i.e., whether the pixel is genuine or fabricated). The loss function for the discriminator is formulated as:

$$\mathcal{L}_{D_{seg}} = -\mathrm{E}_{(I,S)}\left[\sum_{k=1}^d a_k\left(S_k \odot logD_{seg}(I)_k\right)\right] \quad (9)$$
$$-\mathrm{E}_{I_g}\left[logD_{seg}\left(I_g\right)_{d+1}\right]$$

, where a_k symbolizes the balancing weight associated with each term. Subsequently, the loss function for the generator, specifically in the context of the Parsing and Context-Aware Generator (PCAG), is articulated as:

$$\mathcal{L}_{seg} = -\mathrm{E}_{(I_g,S)}\left[\sum_{k=1}^d a_k\left(S_k \odot logD_{seg}\left(I_g\right)_k\right)\right] \quad (10)$$

2. The patch discriminator, symbolized as D_{patch}, aims to ensure the verisimilitude of body-part generation. It accomplishes this by focusing its attention on body-part patches P, defined for 5 distinct body parts, such as the torso, as well as the left and right upper arms and forearms. The main objective of this discriminator is to identify whether a body part is generated or real. The loss function for this discriminator is as follows.

$$\mathcal{L}_{D_{patch}} = -\mathrm{E}_{P_{real}}\left[logD_{patch}(P_{real})\right] \quad (11)$$
$$-\mathrm{E}_{P_{fake}}\left[log\left(1 - D_{patch}\left(P_{fake}\right)\right)\right]$$

where P_{real} and P_{fake} represent the patches extracted from authentic and synthesized images, respectively. The corresponding loss function for the generator is formulated as:

$$\mathcal{L}_{patch} = -\mathrm{E}_{P_{fake}}\left[logD_{patch}\left(P_{fake}\right)\right] \quad (12)$$

3. The match discriminator, denoted as D_{match}, is engineered to impel the proposed PCAG to synthesize output images wherein the fidelity of the attire adorning the subject is in close approximation to the target clothing item, symbolized as C. This discriminator is calibrated using authentic (I) and synthesized (I_g) images. The loss function for D_{match} is expressed as:

$$\mathcal{L}_{D_{match}} = -\mathrm{E}_{(I,C)}[logD_{match}(I,C)] \quad (13)$$
$$-\mathrm{E}_{(I_g,C)}\left[log\left(1 - D_{match}(I_g,C)\right)\right]$$

, while the generator loss it provides to the PCAG is designed as the following:

$$\mathcal{L}_{match} = -\mathrm{E}_{(I_g,C)}\left[logD_{match}\left(I_g,C\right)\right] \quad (14)$$

4. The parsing discriminator, denoted as D_{parse}, is entrusted with the discerning task of evaluating the verisimilitude of a human parsing image in aptly representing a specified attire donned by a particular individual. The goal of this discriminator is to encourage the PCAG to generate human parsing images that scrupulously preserve the morphological characteristics of the target clothing item, denoted as C, as well as the posture information in the densepose. The loss function attributed to this discriminator is as follows.

$$\mathcal{L}_{D_{parse}} = -\mathrm{E}_{(H,M_c,Dp)}\left[logD_{parse}(H,M_c,Dp)\right] \quad (15)$$
$$-\mathrm{E}_{(H_g,M_c,Dp_g)}\left[log\left(1 - D_{parse}\left(H_g,M_c,Dp_g\right)\right)\right]$$

, where H_g and H represent the synthesized and authentic human parsing images of C in wear respectively; M_c, symbolizing the binary mask of C; and Dp_g and Dp, representing the densepose images upon which H_g and H are based on. That leads us to the following generator loss for the PCAG:

$$\mathcal{L}_{parse} = -\mathrm{E}_{(H_g,M_c,Dp_g)}\left[logD_{parse}\left(H_g,M_c,Dp_g\right)\right] \quad (16)$$

In summary, the joint training objective for the PCAG can be denoted as:

$$\mathcal{L}_{PCAG} = \lambda_{per}\mathcal{L}_{per} + \lambda_{shape}\mathcal{L}_{shape} + \lambda_{ce}\mathcal{L}_{ce} + \lambda_{seg}\mathcal{L}_{seg} \quad (17)$$
$$+\lambda_{patch}\mathcal{L}_{patch} + \lambda_{match}\mathcal{L}_{match} + \lambda_{parse}\mathcal{L}_{parse}$$

where the lambda terms are the coefficients of the respective generator losses.

IV. Experimental Results

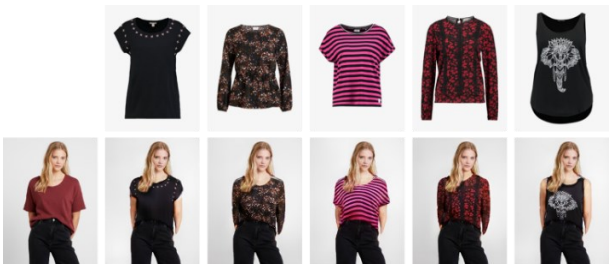


Fig. 4 Exemplar demonstrations of garment adaptability across diverse styles using the proposed method.

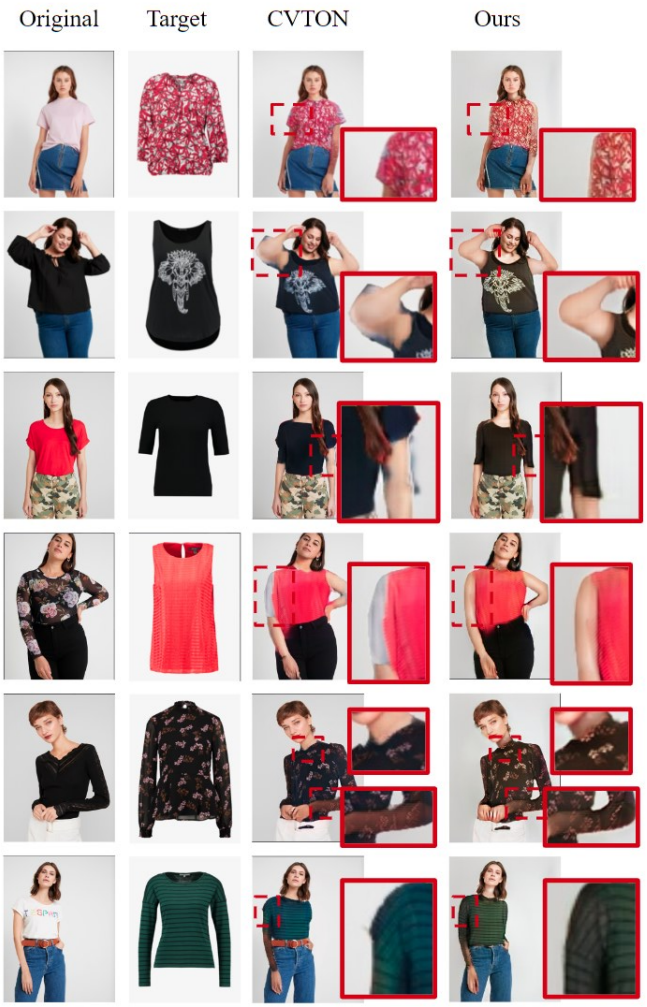


Fig. 5 Comparison of the results utilizing the VITON-HD dataset. A discernible enhancement in image quality is observed with our approach, particularly in areas that have been emphasized through highlighting.

Training was rigorously conducted by harnessing the computational prowess of an Nvidia A100 GPU, graciously provisioned by Google Colab. To facilitate periodic monitoring and potential recovery from unintended interruptions, the models were checkpointed at intervals of 2,000 iterations. Subsequent to each training session, we employed an adaptive approach to parameter tuning; we fine-tuned the hyperparameters such as the learning rate in accordance with the emergent patterns in the convergence profiles of the various loss functions. The models were subjected to intensive training on the VITON-HD dataset [17], endowed with images having a resolution of 256×192 pixels.

In Figure 4, we unveil the capabilities of our method in orchestrating a virtual try-on of an assortment of garments on an identical individual. The outcomes conspicuously illustrate that our model is adept at gracefully navigating scenarios characterized by drastic shape changes.

Figure 5 presents a comparison of images engendered by our model vis-à-vis those generated by the C-VTON model. A meticulous examination reveals that the images produced by C-VTON exhibit a predilection for maintaining the morphological integrity of the original attire mistakenly. For instance, when simulating the try-on of a long-sleeved garment, vestiges of the original short-sleeved cuff remain discernible, with the model exhibiting palpable difficulties in accommodating variations in sleeve length. On the other hand, our proposed model evinces a superlative acuity in accurately reproducing the salient features of the try-on attire, including but not limited to sleeve length, cuff, and collar.

V. Conclusions

In this research, we have unveiled PC-VTON, a dual-phase virtual try-on image generation architecture, which is bolstered by a Generative Adversarial Network (GAN) paradigm and targets the harmonious fusion of clothing morphology and contextual preservation. This architecture is complemented by an array of purpose-crafted loss functions and an innovative discriminator group design. Our work has addressed the weakness often observed in extant virtual try-on methodologies –the incapacity to faithfully render the shapes of clothing during image synthesis. Our empirical evidence compellingly attests to the prowess of the proposed framework, which not only retains the laudable capacity for the precise depiction of textures and patterns, reminiscent of its predecessor, the CVTON framework, but also excels in the reproduction of structural attributes of clothing with remarkable finesse. The cornerstone of this marked enhancement is the incorporation of human parsing generation into our schema, which is further augmented by the proposed loss functions and discriminators.

REFERENCES

[1] Rose Francoise Bertram and Ting Chi.“A study of companies’ business responses to fashion e-commerce’s environmental impact”, International Journal of Fashion Design, Technology and Education, 2018

[2] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. “Fashion Meets Computer Vision: A Survey”, ACM Computing Surveys, 2021

[3] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. “Markerless garment capture”, ACM Transactions on Graphics, 2008

[4] Stefan Hauswiesner, Matthias Straka, and Gerhard Reitmayr. “Virtual Try-On through Image-Based Rendering”, IEEE Transactions on Visualization and Computer Graphics (T-VCG), 2013

[5] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. “ClothCap: Seamless 4D Clothing Capture and Retargeting”, Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), 2017

[6] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. “C-VTON: Context-Driven Image-Based Virtual Try-On Network”, IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Networks”, Advances in Neural Information Processing Systems (NIPS), 2014

[8] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. “SwapNet: Image Based Garment Transfer”, European Conference on Computer Vision (ECCV), 2018

[9] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. “Towards Photo-Realistic Virtual Try-On by Adaptively Generating↔Preserving Image Content”, IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2020

[10] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, Meng Yang. “Toward Characteristic-Preserving Image-based Virtual Try-On Network”, European Conference on Computer Vision (ECCV), 2018

[11] Jean Duchon. “Splines Minimizing Rotation-Invariant SemiNorms in Sobolev Spaces.”, Constructive Theory of Functions of Several Variables, 1977.

[12] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. “VOGUE: Try-On by StyleGAN Interpolation Optimization” , IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2021

[13] Anran Lin, Nanxuan Zhao, Shuliang Ning, Yuda Qiu, Baoyuan Wang and Xiaoguang Han. “FashionTex: Controllable Virtual Try-on with Text and Texture”, Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), 2023

[14] Tero Karras, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”, IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2019

[15] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. “Designing an Encoder for StyleGAN Image Manipulation”, Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), 2021

[16] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. “Controllable Person Image Synthesis with Attribute-Decomposed GAN”, IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2020

[17] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo “VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization”, IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2021

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Sparse Tensor-based point cloud attribute compression using Augmented Normalizing Flows

Tzu-Po Lin[1], Monyneath Yim[1], Jui-Chiu Chiang[1], Wen-Hsiao Peng[2], and Wen-Nung Lie[1]

[1]Department of Electrical Engineering, National Chung Cheng University, Taiwan

[2]Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan

***Abstract*—The large amount of data of point cloud poses challenges for efficient storage and transmission. To address this problem, various learning-based techniques, in addition to rule-based solutions, have been developed for point cloud compression. While many previous works employed the variational autoencoder (VAE) structure, they have failed to achieve promising performance at high bitrates. In this paper, we propose a novel point cloud attribute compression technique based on the Augmented Normalizing Flow (ANF) model, which incorporates sparse convolutions where a sparse tensor is used to represent the point cloud attribute. The invertibility of the NF model provides better reconstruction compared to VAE-based coding schemes. ANF provides a more flexible way to model the input distribution by introducing additional conditioning variables into the flow. Not only comparable to G-PCC, the experimental results demonstrate the effectiveness and superiority of the proposed method over several learning-based point cloud attribute compression techniques, even without requiring sophisticated context modeling.**

I. INTRODUCTION

In recent decades, point clouds have garnered significant attention in the field of computer vision, particularly in human-machine collaboration. They have found wide-ranging applications in areas such as autonomous driving, augmented reality, and robotics. Unlike images, point cloud is a set embedded in a continuous space. It comprises the 3D coordinates and corresponding attributes such as color, normal and reflectiveness. Due to the unordered and non-uniform distribution of points in the point cloud, processing them poses significant challenges.

Efficient point cloud compression has recently gained considerable attention and the MPEG has developed two point cloud coding standards: Video-based Point Cloud Compression (V-PCC) [1] and Geometry-based Point Cloud Compression (G-PCC) [2]. In V-PCC, 3D patches are formed by grouping points with similar normals. The geometry and the attribute are then projected and packed into 2D images, and an occupancy map is built to guide the patch packing. The geometry, attribute and occupancy map are encoded by conventional video codecs. In G-PCC, point cloud geometry and attribute are separately compressed. The geometry is encoded first, followed by attribute compression which relies on the decoded geometry.

Graph Fourier Transform (GFT) has been utilized in numerous studies for point cloud attribute compression [3, 4]. Besides, sparse representation is typically leveraged to model point clouds [5, 6]. While popular transform, such as DCT, DFT and GFT are commonly used, [7] performs multi-scale structured dictionary learning to more effectively represent the sparsity of point clouds. In [8], geometry-guided sparse representation was proposed for compressing the voxelized point cloud attribute. This method uses a virtual adaptive sampling technique to model irregular point clouds.

In G-PCC, geometric information is expressed using octree representation, which recursively divides point cloud into eight sub-cubes based on root cube occupancy. Attributes are compressed using three coding methods: Region-Adaptive Hierarchical Transform (RAHT) [9], Prediction Transform and Lifting Transform. RAHT is a variant of the Haar wavelet transform that converts attributes into frequency domain coefficients. The upsampled transform domain prediction further improves coding efficiency by using decoded RAHT coefficients in lower resolution to predict coefficients in higher resolution. The Prediction Transform first constructs level of details (LOD) by arranging the points into a set of refinements levels, followed by distance-based prediction. The Lifting transform is built on top of the Prediction Transform to further improve coding efficiency, employing an update operator and adaptive quantization in additional to the prediction operation. The G-PCC encoding platform, TMC13, demonstrates state-of-the-art performance in point cloud attribute compression.

Learning-based techniques have been successful in point cloud geometry compression and can be classified into point-based and voxel-based methods. Point-based methods, such as those presented in [10] and [11], use PointNet as the encoder backbone. A fully-connected network is used in the decoder in [10] while a graph neural sampling module is incorporated to gather neighborhood features in [11]. Voxel-based methods first partition the point cloud into non-overlapping 3D cubes and employ 3D convolutional neural networks (CNNs) to

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

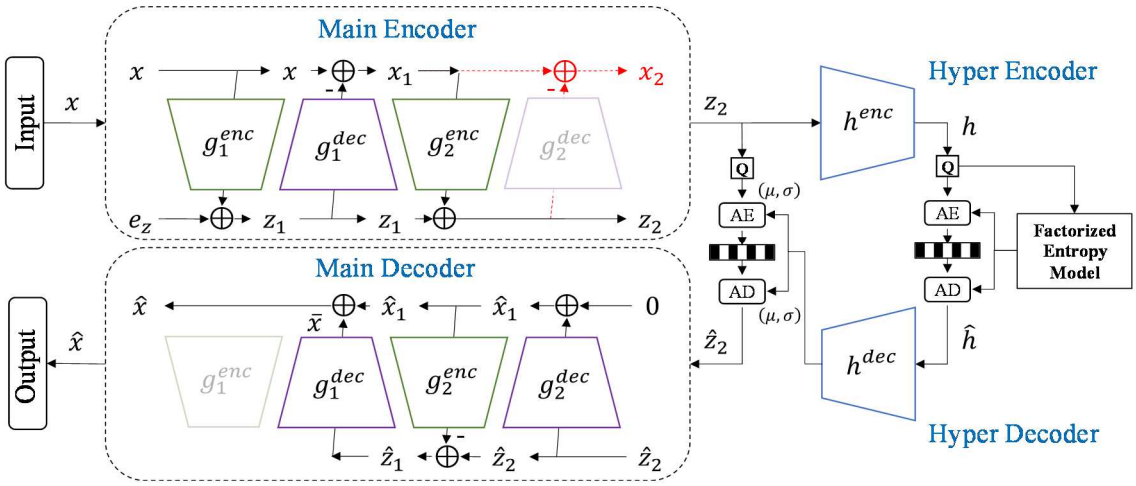


Fig. 1. The network architecture of our proposed ANF-PCAC. The red lines indicate the paths used during the training phase.

compress the geometry [12]. In CNN-based methods, the input data is transformed into a lower dimensional latent space. For the voxel-based approach, the geometry reconstruction is represented as a binary classification. In [13], a VAE structure with hyperprior is proposed. Later, sparse convolution on a multi-scale structure is presented in [14] to further improve coding efficiency. This architecture processes the downsampled latent in a lossy manner, while preserving the coarse geometry in a lossless manner. A hierarchical reconstruction is then conducted from low resolution to high resolution.

The network architectures used for point cloud geometry coding can be also applied to point cloud attribute coding. Many works have adopted VAE-based architectures, such as Deep-PCAC [15], which is the first end-to-end learned point cloud attribute compression using multilayer perceptrons (MLPs). In addition, sparse convolutions have been used in [16-17] to decrease computation cost. These methods apply an adaptive entropy model with the joint utilization of a hyper prior and autoregressive neighbors [18] to better estimate the bit rate. In [19], point cloud attributes are represented with volumetric functions modeled by local coordinate-based networks. They adopt a hierarchical structure of latent codes, which are represented by RAHT. In [20], folding-based image compression is proposed. The method parameterizes a point cloud as a 2D manifold and folds the geometry into 2D grids. Then, an image is formed by projecting the associated attribute in the grid, followed by conventional image compression.

Flow-based models have emerged as a promising alternative to VAE-based models for compression task due to their ability to ensure perfect reconstruction with quantization [21-22]. Affine coupling layers are typically used to establish a bijective mapping between the input and its latent features. In [23], a point cloud attribute compression method using normalizing flow (NF) was proposed, which included a 3D voxel squeeze layer to process irregular point clouds as regular 3D data. However, this method had a large number of model parameters.

In [24], an augmented normalizing flow (ANF) was used for point cloud geometry compression (ANFPCGC), focusing on voxelized point cloud geometry and leveraging 3D CNNs to extract the latent features. Built on top of the flow model, an independent noise input is coupled in ANF to enable improved transformation capability [25]. ANFPCGC comprises a stack of 2 VAEs and superior coding performance over several VAE-based designs [13-14] are demonstrated.

In this paper, we present ANF-PCAC, an ANF-based point cloud attribute coding scheme that utilizes a sparse tensor architecture with sparse convolution instead of 3D CNN, which distinguishes it from ANFPCGC. The main contributions of this paper are as follow:

- We introduce ANF-PCAC, a learning-based framework for compressing point cloud attributes. ANF-PCAC utilizes the ANF structure and incorporates sparse convolution.
- Experimental results provide validation for the effectiveness of the proposed technique, showcasing competitive performance with G-PCC.
- Our method surpasses several learning-based solutions, even when no context model is applied.

II. Proposed Method

The sparse convolution takes into account sparsity in CNNs by only computing outputs on predefined voxels and storing them as compact sparse tensors. It is efficient for irregular point clouds and can significantly reduce computation and memory costs. In our proposed network, we use sparse convolution as the basic operation. The sparse tensor $\{\vec{C}, \vec{F}\}$ is represented by a set of coordinates $\vec{C} = \{(x_i, y_i, z_i) \mid i \in [0, N-1]\}$ of the point cloud and the corresponding features $\vec{F} = \{(R_i, G_i, B_i) \mid i \in [0, N-1]\}$, where N is the total number of occupied points. We assume that the point cloud geometry is available and can be losslessly coded, which is consistent with other learning-based point cloud attribute compression methods. In the following, we introduce our ANF model and the internal VAE structure in detail, as well as our training strategy.

A. ANF-PCAC

Following the design in ANFPCGC, we proposed ANF-based point cloud attribute compression. Fig. 1 shows the network architecture of the proposed ANF-PCAC, which is built upon SparsePCAC [16] while discarding the context model. ANF-PCAC consists of two parts: the main encoder/decoder and the hyper encoder/decoder. We employ two layers of autoencoding transform in the main encoder/decoder, resulting in a two-step ANF that balances model capacity and complexity. The two autoencoding transforms have the same design with g_1^{enc} and g_2^{enc} sharing the same structure, and so do g_1^{dec} and g_2^{dec}. The detail of the autoencoding transforms are shown in Fig. 2. Similar to [16], we increase the channel size from 3 to 128 and perform three-time downsampling in g_1^{enc} and g_2^{enc} to make the features smaller and more compact. In g_2^{dec} and g_1^{dec}, we perform the reverse operation to reconstruct the point cloud attributes.

During encoding, we input the point cloud attribute and an augmented noise e_z, which follows a Gaussian distribution with zero mean. In the experiments, e_z is set to zero. During the training, we obtain an additive coupling output z_1 by feeding the input to the first encoding transform g_1^{enc}. Then we get the difference between x and the reconstructed one, denoted as x_1, after performing the first decoding transform g_1^{dec}. In the next step, x_1 and z_1 are served as inputs to the second autoencoder, and we get x_2 and z_2 as the final output of the main encoder, as shown in (1) and (2):

$$(x_1, z_1) = (x - g_1^{dec}(z_1), g_1^{enc}(x) + e_z) \quad (1)$$

$$(x_2, z_2) = (x_1 - g_2^{dec}(z_2), g_2^{enc}(x_1) + z_1) \quad (2)$$

As indicated in (2), x_2 is the difference between the original x_1 and the reconstructed one and is expected to be close to zero if the encoding and decoding are well matched to each other. It implies that the input information x is expected to be completely transferred to the latent z_2. It this way, we can save bits if x_2 is not transmitted. In inference, g_2^{dec} is not executed during encoding.

After quantization, we perform decoding by reversing the operation of encoding. Following the design in ANFPCGC, we do not transmit x_2 and the main decoder performs reconstruction by setting x_2 to zero. To reduce the mismatch between the true value of x_2 and zero. We encourage x_2 to be close to zero by including a loss term. We obtain $\hat{x}_1$ by feeding the quantized $\hat{z}_2$ to g_2^{dec}. Then we input $\hat{x}_1$ and $\hat{z}_2$ to the bottom autoencoder in a reverse direction, and we obtain the final output $\hat{x}$, as shown in (3)(4):

$$(\hat{x}_1, \hat{z}_1) = (g_2^{dec}(\hat{z}_2), \hat{z}_2 - g_2^{enc}(\hat{x}_1)) \quad (3)$$

$$\hat{x} = \hat{x}_1 + g_1^{dec}(\hat{z}_1) = \hat{x}_1 + \bar{x} \quad (4)$$

To further reduce bit consumption, we use a conditional entropy model based on hyperpriors [18]. The hyper encoder learns the probability distribution of the latent z_2 and

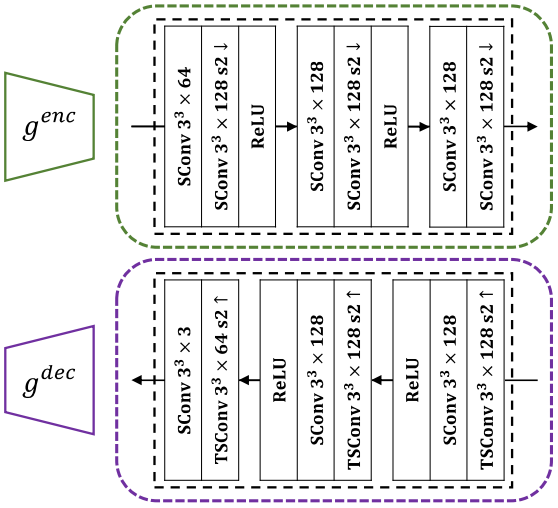


Fig. 2. Encoding transform and decoding transform of the ANF-PCAC transform architecture.

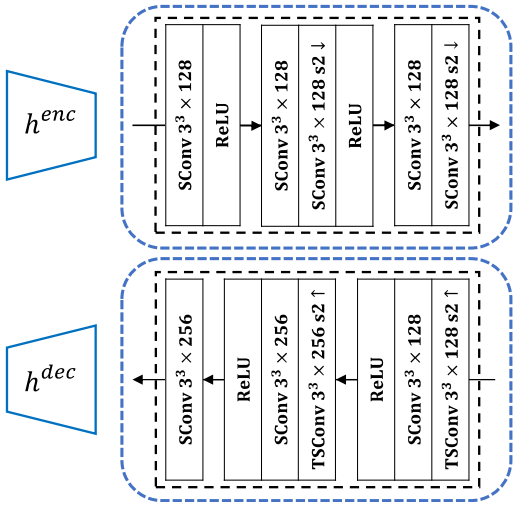


Fig. 3: The hyperprior transform

generates a hyper prior h, which is then quantized as $\hat{h}$ and described by a factorized entropy model. Then $\hat{h}$ passes through the hyper decoder to generate the mean and variance (μ, σ) which are used to describe the distribution of the latent z_2. With this conditional Gaussian distribution, the bits consumption of the latent z_2 can be estimated. The detail of the hyper encoder/decoder can be seen in Fig. 3.

B. Training strategy

To train the proposed ANF-PCAC, we use Lagrangain loss, as shown in (5). For the rate, there are two terms, denoted as $R_{\hat{z}_2}$ and $R_{\hat{h}}$, presenting the rate of the latent $\hat{z}_2$ and the hyperprior $\hat{h}$, respectively. Three terms are involved in the computation of the distortion. The first term $d(x, \hat{x})$ computes the mean squared error (MSE) between the input x and the reconstructed signal. The second term $d(x, \bar{x})$ measures the MSE between the input x and $\bar{x}$ with the aim of enabling better reconstruction for the autoencoding. As mentioned earlier, x_2 represents the residual information that is expected to be



Fig. 4: **Training Dataset.** Some high-quality version point cloud samples of ScanNet Dataset.



Fig. 5: **Testing Dataset.** 8iVFB dataset, from left to right are longdress, loot, redandblack, and soldier.

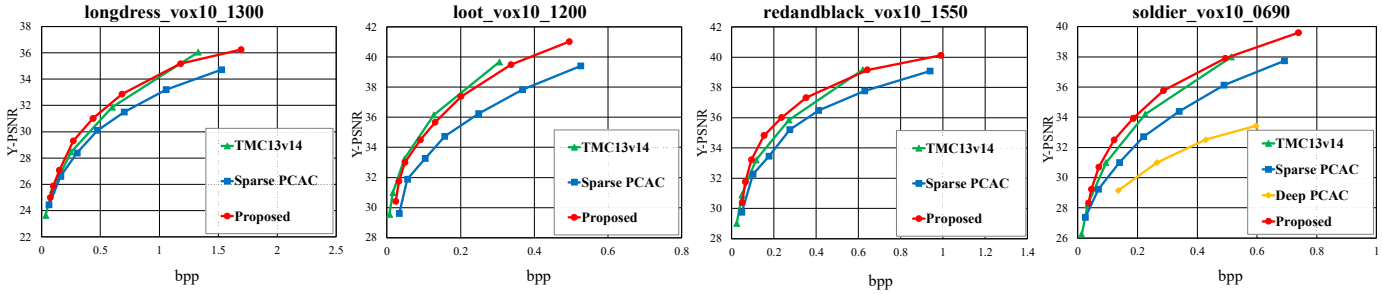


Fig. 6: R-D curves for longdress, loot, redandblack, and soilder.

approximated by zero. This task can be accomplished by including the third term.

$$J = R_{\hat{z}_2} + R_{\hat{h}} + \lambda \cdot \left(\left(d(x, \hat{x}) + d(x, \bar{x}) \right) + \gamma \cdot (x_2 - 0)^2 \right) \quad (5)$$

III. Experimental Results

In this section, we will provide the information of our training and testing datasets, as well as the anchor schemes. We will also describe our training details, followed by the presentation of quantitative and qualitative comparisons to demonstrate the superiority of our proposed method.

A. Dataset and Anchor

For our training dataset, we utilize the ScanNet [26] dataset, which is a large open-source dataset of indoor scenes. The dataset contains both high-quality and low-quality reconstructed meshes, and we used the high-quality version, as shown in Fig. 4. We converted the mesh into point cloud and scaled the geometry to 10 bits, followed by cube division with size $64 \times 64 \times 64$. We randomly selected 50,000 cubes and used them for training. To make a fair comparison, the test dataset utilized in SparsePCAC [16] has been assessed. Four point clouds in the 8i Voxelized Full Body(8iVFB) [27] are used, as shown in Fig. 5. Notably, our training and testing datasets are entirely different, enabling us to demonstrate the generalization ability of our proposed scheme. We compared our approach with three methods, including two learning-based methods: Deep PCAC [15] and SparsePCAC [16], and a traditional point cloud compression standard G-PCC (TMC13v14).

Table 1: BD-Rate (%) and BD-PSNR (dB) with respect to abchors

Point Cloud	TMC13 v14		SparsePCAC		Deep PCAC	
	BD-Rate (%)	BD-PSNR(Y) (dB)	BD-Rate (%)	BD-PSNR(Y) (dB)	BD-Rate (%)	BD-PSNR(Y) (dB)
longdress	-6.42	0.28	-26.92	1.09	-	-
loot	23.78	-0.56	-39.75	1.73	-	-
redandblack	-10.99	0.42	-33.65	1.28	-	-
soldier	-11.58	0.46	-37.63	1.64	-70.52	4.45
Average	-1.30	0.15	-34.49	1.44	-70.52	4.45

B. Training details

We set λ to 0.05, 0.025, 0.01, 0.005, 0.0025, 0.001, 0.0005 and 0.00025 to obtain models with different bitrates. γ is set as 0.01 to balance the distortion. We trained our models for 500 epochs, and all models use the initial learning rate 1×10^{-4} which gradually decreases to 1×10^{-6}. We use Nvidia GeForce RTX 2080 Ti for training and testing.

C. Objective Quality Comparison

We present the R-D (rate-distortion) performance in Fig.6, in which the bitrate is measured in bits per point (bpp) and the distortion is measured by the Peak Signal-to-Noise Ratio (PSNR) of the Y channel. Deep-PCAC evaluates only the sequence of soldiers in the 8iVFB dataset, while it is trained with others sequence in the same dataset. Fig. 6 shows that our proposed scheme outperforms Deep-PCAC and SparsePCAC. Additionally, our method is comparable with TMC13v14. To assess the average R-D performance, we used the Bjøntegaard

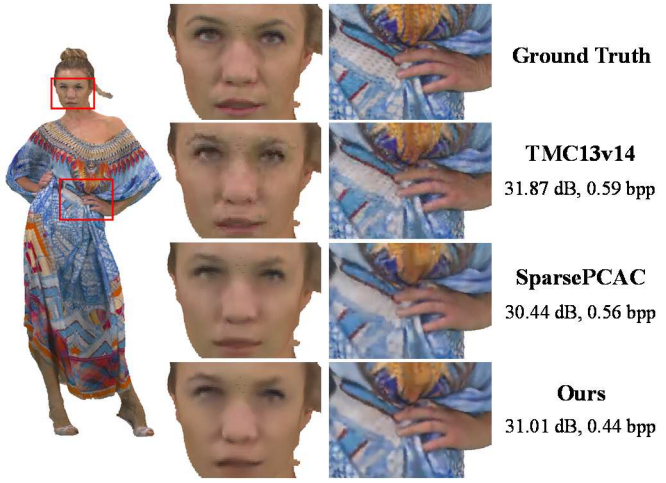


Fig. 7: Qualitative comparison of longdress.

Table 2: Average runtime for 8iVFB dataset

	TMC13v14	SparsePCAC	Ours
Enc (s)	1.17	122.92	0.45
Dec (s)	0.84	835.63	0.29

metric, including BD-PSNR and BD-Rate, which is summarized in Table 1. The results indicate that our method achieves, on average 70.52%, 34.49% and 1.3% rate reduction when compared to Deep PCAC, SparsePCAC and TMC13v14, respectively.

D. Subjective Quality Comparison

For quality visualization, we compare our method with TMC13v14 and SparsePCAC in Fig. 7. The reconstruction by TMC13v14 has noticeable blurs and noise, while our method generally delivers a more visually pleasing result with smoother textures.

E. Runtime Comparison

To specifically demonstrate the efficiency of our method where only a hyperprior is used to build the entropy model of the latent, we compare our method with SparsePCAC and TMC13v14. Since SparsePCAC relies on both autoregressive prior and hyperprior, it requires a more expensive inference time, as shown in Table 2.

IV. Conclusions

In this paper, we propose a learning based point cloud attribute compression using ANF. We use sparse convolutional layers as our basic layer in the two-step VAE network. Our proposed method outperforms Deep PCAC, SparsePCAC and TMC13v14 with 70.52%, 34.49% and 1.3% rate reduction, respectively, and shows competitive qualitative performance compared to state-of-the-art TMC13v14. For the runtime of inference, our method saves significant time consumption compared to SparsePCAC, as our method works well even without an autoregressive context model.

References

[1] V-PCC Codec Description, ISO/IEC JTC 1/SC 29/WG 7 N00100, 2021.

[2] G-PCC codec description, ISO/IEC JTC 1/SC 29/WG 7 N00271, 2022.

[3] C. Zhang, D. Florêncio and C. Loop, "Point cloud attribute compression with graph transform," in *IEEE International Conference on Image Processing*, 2014, pp. 2066-2070.

[4] F. Song, G. Li, W. Gao and T. H. Li, "Rate-distortion optimized graph for point cloud attribute coding," *IEEE Signal Processing Letters*, vol. 29, pp. 922-926, 2022.

[5] J. Hou, L. -P. Chau, Y. He and P. A. Chou, "Sparse representation for colors of 3D point cloud via virtual adaptive sampling," in *IEEE International Conference on Acoustics, Speech and Signal Processing,* 2017, pp. 2926-2930.

[6] S. Gu, J. Hou, H. Zeng, H. Yuan and K. -K. Ma, "3d point cloud attribute compression using geometry-guided sparse representation," *IEEE Transactions on Image Processing,* vol. 29, pp. 796-808, 2019.

[7] S. Yangmei, W. Dai, C. Li, J. Zou, and H. Xiong, "Multi-scale structured dictionary learning for 3-D point cloud attribute compression," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, no. 7 pp. 2792-2807, 2020.

[8] H. Liu, H. Yuan, Q. Liu, J. Hou, H. Zeng and S. Kwong, "A hybrid compression framework for color attributes of static 3d point clouds," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1564-1577, March 2022

[9] D. Queiroz, Ricardo L., and P. A. Chou, "Compression of 3D point clouds using a region-adaptive hierarchical transform," *IEEE Transactions on Image Processing* vol. 25, no. 8, pp. 3947-3956, 2016.

[10] W. Yan, S. Liu, T. H. Li, Z. Li, and G. Li, "Deep autoencoder-based lossy geometry compression for point clouds," *arXiv preprint arXiv:1905.03691*, 2019

[11] L. Gao, T. Fan, J. Wan, Y. Xu, J. Sun and Z. Ma, "Point cloud geometry compression via neural graph sampling," in *IEEE International Conference on Image Processing* , 2021, pp. 3373-3377.

[12] M. Quach, G. Valenzise, and F. Dufaux, "Learning convolutional transforms for Lossy Point Cloud Geometry Compression," in *IEEE International Conference on Image Processing*, 2019.

[13] J. Wang, H. Zhu, H. Liu and Z. Ma, "Lossy point cloud geometry compression via end-to-end learning," *IEEE Transactions on Circuits and Systems for Video Technology* vol, 31, no. 12, pp. 4909-4923, 2021.

[14] J. Wang, D. Ding, Z. Li and Z. Ma, "Multiscale point cloud geometry compression," in *Data Compression Conference*, 2021, pp. 73-82.

[15] X. Sheng, L. Li, D. Liu, Z. Xiong, Z. Li and F. Wu, "Deep-pcac: an end-to-end deep lossy compression framework for point cloud

attributes," *IEEE Transactions on Multimedia*, vol. 24, pp. 2617-2632, 2021.

[16] J. Wang and Z. Ma, "Sparse tensor-based point cloud attribute compression." In *EEE International Conference on Multimedia Information Processing and Retrieval*, 2022, pp. 59-64.

[17] X. Huo, S. Zhang, and F. Yang, "Variable rate point cloud attribute compression with non-local attention optimization. *Applied Sciences*, *12*(16), 8179, 2022.

[18] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprio," *International Conference on Learning Representations,* 2018.

[19] B. Isik, P. Chou, S. J. Hwang, N. Johnston, and G. Toderici, "Lvac: Learned volumetric attribute compression for point clouds using coordinate based networks," *Frontiers in Signal Processing*, 2022.

[20] M. Quach, G. Valenzise and F. Dufaux, "Folding-based compression of point cloud attributes," in *IEEE International Conference on Image Processing*, 2020, pp. 3309-3313.

[21] L. Helminger, A. Djelouah, M. Gross, and C. Schroers , "Lossy image compression with normalizing flows,"in *International Conference on Learning Representations Workshop*, 2021.

[22] Y. -H. Ho, C. -C. Chan, W. -Hi. Peng, H. -M. Hang, and M. Domański, "ANFIC: Image compression using augmented normalizing flows," IEEE Open Journal of Circuits and Systems, No. 2, pp. 613-626, 2021.

[23] R. B. Pinheiro, J.-E. Marvie, G. Valenzise, F. Dufaux, "NF-PCAC-deep point cloud attribute compression with normalizing flow, ISO/IEC JTC 1/SC 29/WG 7 m61142, 2022.

[24] S. -Y. Li, J. -J. Chiu, J. -C. Chiang, W. -H. Peng and W. -N. Lie, "Augmented normalizing flow for point cloud geometry coding," in *IEEE International Conference on Visual Communications and Image Processing*, 2022.

[25] C. -W. Huang, L. Dinh, and A. C. Courville, "Augmented normalizing flows: bridging the gap between generative flows and latent variable models, CoRR, 2020.

[26] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser and M. Nießner, "ScanNet: richly-annotated 3d reconstructions of indoor scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2432-2443.

[27] E. d'Eon et al. 8i voxelized full bodies - a voxelized point cloud dataset. ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) m38673/M72012, May 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Toward Leveraging Pre-Trained Self-Supervised Frontends for Automatic Singing Voice Understanding Tasks: Three Case Studies

Yuya Yamamoto
University of Tsukuba, Tsukuba, Japan
E-mail:s2130507@u.tsukuba.ac.jp

***Abstract*—Automatic singing voice understanding tasks, such as singer identification, singing voice transcription, and singing technique classification, benefit from data-driven approaches that utilize deep learning techniques. These approaches work well even under the rich diversity of vocal and noisy samples owing to their representation ability. However, the limited availability of labeled data remains a significant obstacle to achieving satisfactory performance. In recent years, self-supervised learning models (SSL models) have been trained using large amounts of unlabeled data in the field of speech processing and music classification. By fine-tuning these models for the target tasks, comparable performance to conventional supervised learning can be achieved with limited training data. Therefore, in this paper, we investigate the effectiveness of SSL models for various singing voice recognition tasks. We report the results of experiments comparing SSL models for three different tasks (i.e., singer identification, singing voice transcription, and singing technique classification) as initial exploration and aim to discuss these findings. Experimental results show that each SSL model achieves comparable performance and sometimes outperforms compared to state-of-the-art methods on each task. We also conducted a layer-wise analysis to further understand the behavior of the SSL models.**

I. Introduction

The singing voice plays an important role in the music. It provides emotional expressions for us through its melody and lyrics. Computational understanding tasks of singing voices, such as singer identification, singing voice transcription, and singing expression identification, are beneficial for many applications such as music discovery [1], pedagogy [2], musicological analysis [3], etc. Processing the singing voices is a long-running challenge in the music information retrieval (MIR) field [4] due to its wide variation. Recently, the methods based on deep learning outperformed the conventional methods based on the hand-crafted manner in many singing voice understanding tasks thanks to its intense expressiveness [5], [6], [7], [8]. However, deep learning approaches typically necessitate extensive datasets comprising sung tracks with high-quality labels, entailing substantial costs for both data collection and annotation.

Transfer learning is one of the techniques to alleviate the requirements of large-scale datasets for the low-resource situation. It is based on the transfer of the knowledge derived from high-resource upstream pre-training tasks to target downstream tasks. In the audio domain, there are many works on audio classification tasks that utilize transfer learning of PANNs [9] and VGGish [10], which are pre-trained on a large-scale audio dataset. Notably, transfer learning of the model that is pre-trained by self-supervised learning (SSL) fashion is rapidly emerging. SSL models are leveraging a vast amount of unlabeled data, several notable models have been developed in both the speech and music domains. In the speech domain, these models include Wav2Vec2.0 [11], HuBERT [12], and WavLM [13]. Meanwhile, in the music domain, models such as MERT [14], and MapMusic2Vec [15] have been introduced.

Singing voice encompasses characteristics of both speech and music, and researchers have explored leveraging pre-trained SSL models in singing voice understanding tasks [16], [17], [18], [19] for each. The transfer learning of pre-trained SSL models from the speech or music domains to the singing domain holds potential. However, there is still ample room for investigating the utility of each pre-trained SSL model. This includes investigating which domains can contribute to singing voice understanding tasks, how the model extracts valuable features for the target tasks, and how to fully utilize the potential of the SSL models, and so on.

In this study, our objective is to examine the usefulness of SSL models pre-trained on the speech or music domains for singing voice understanding tasks. To achieve this, we employ pre-trained SSL models as a front-end to our voice identification model and evaluate their performance through fine-tuning.

We present the following contributions in this study:

1) Comparative analysis of SSL models: We compare multiple SSL models across three distinct three tasks: singer identification, singing voice transcription, and singing technique classification, corresponding to ”Who sings?”, ”What song?”, and ”How to sing?”, respectively.
2) Comparison with SoTA models: We also compared the SSL models with state-of-the-art (SoTA) models. Through fine-tuning the pre-trained SSL models, we demonstrate that they achieve performance comparable to that of current state-of-the-art methods in several singing voice understanding tasks.
3) Investigation of layer-wise behavior: Additionally, we delve deeper into the behavior of the model’s layers and analyze their characteristics across different tasks,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

by utilizing learnable weight for each layer.

II. Related Works

Various machine learning approaches have been explored for low-resource problems in the singing voice understanding tasks. Semi-supervised learning [20], [21], data augmentation [22], [23], and self-supervised learning on singing voices [24], [25], leveraging speech data [26], [27] etc. have been proposed to mitigate the drawbacks of supervised-learning fashion.

More recent, pre-trained self-supervised models are used for singing understanding tasks. Ou et al. leveraged Wav2Vec 2.0 [11] model that is pre-trained using Librispeech corpus [28] and fine-tuned automatic speech recognition for lyric transcription. They achieved comparable performance with the state-of-the-art performance of lyric transcription methods [29], [30], which are learned from over 150 hours of data, by using its 10% of amount (i.e., 15 hours) [16]. Gu et al. leveraged Wav2Vec2.0 for singing voice transcription and outperformed conventional works [18]. Heydari et al. tackled singing beat tracking, which is a beat tracking method when the input is only a singing voice. They leverage WavLM [13] and DistilHuBERT [31] for the frontend of the model, and they outperformed the model that adapts only the spectrogram for the frontend feature. Donahue et al. proposed melody transcription from the musical mixture utilizing codified music representation [32] derived from the hidden representation of JukeBox [33], which is originally proposed for musical audio generation.

III. Methods

We compare four SSL models: 1) Wav2Vec2.0 [11], 2) WavLM [13], 3) MERT [14], and 4) MapMusic2Vec [15]. Each model takes a raw waveform as input and employs convolutional layers for feature extraction along with 12 Transformer encoder layers. The output of each model consists of a 768-dimensional vector per frame. We utilized these models as the frontend for each singing voice understanding task.

A. SSL models

1) Wav2Vec2.0: Wav2Vec2.0 [11] is the model that has convolutional and Transformer Encoder layers and acquired speech representation through contrastive learning and masking. It takes raw speech waveforms as input, and the initial convolutional layers produce latent representations denoted as z. To facilitate contrastive learning, the quantization module is applied to convert z into discrete representations denoted as Q. Simultaneously, z is fed into the Transformer layers after applying random masking to several frames. The output feature C is then derived from the Transformer layers. Finally, contrast learning is performed masked time step in C and Q. Here, the same time steps are considered as positive examples and different time steps are considered as negative examples. Wav2Vec2.0 has shown its effectiveness in several downstream speech tasks by fine-tuning [34], [35], [36]. We used the Wav2Vec2.0 Base model[1].

[1]https://huggingface.co/facebook/wav2vec2-base-960h

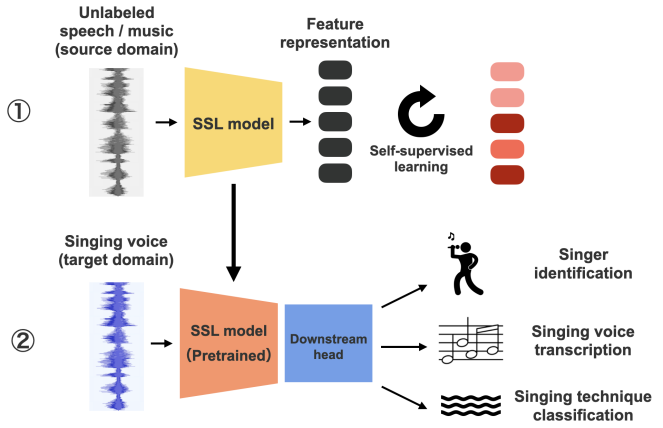


Fig. 1. The concept of this paper. 1) Pre-training on upstream task: Utilizing a vast amount of unlabeled data (either speech or music) and pre-training the model in a self-supervised fashion. 2) Transfer learning for downstream task: Leveraging pre-trained model and solving the target task (i.e., singing voice understanding tasks).

2) WavLM: WavLM [13] is a large-scale pre-trained model with 94k hours of speech data as input that can treat full-stack speech processing. It adopts masked prediction of hidden units like HuBERT [12] and denoising of the input speech as self-supervised pretraining. The diversity of the pre-trained corpus of WavLM is wider than that of Wav2Vec2.0; Gigaspeech [37], a collection of audiobooks, podcasts, and YouTube and VoxPopli[38], a collection of European Parliament, in addition to Librispeech. We used the WavLM Base plus model, which demonstrates better performance than the Base model[2].

3) MERT: MERT [14] is a large-scale pre-trained model using unlabeled music data. It is also inspired by masked prediction of hidden units as WavLM while it uses CQT spectrogram as its target in addition to quantized acoustic features with the purpose of enhancing the pitch representative power. It can achieve the parameter efficiency compared to the JukeMIR [32], the musical audio representation derived from JukeBox. We used the public-v0 model[3], which is only trained on a public music dataset.

4) MapMusic2Vec: MapMusic2Vec[4] [15] is a model that is pre-trained using BYOL [39] with 1k hours of music data. It relies on two neural networks that have the same structure as each other, reffed to the teacher model and the student model. The parameters of the teacher model are updated according to the exponential moving average (EMA) of the student model. The student model takes partially masked audio raw waveform as input while the teacher model takes unmasked one and outputs the prediction of hidden outputs from the last K layers of the teacher model.

[2]https://huggingface.co/microsoft/wavlm-base-plus
[3]https://huggingface.co/m-a-p/MERT-v0-public
[4]https://huggingface.co/m-a-p/music2vec-v1

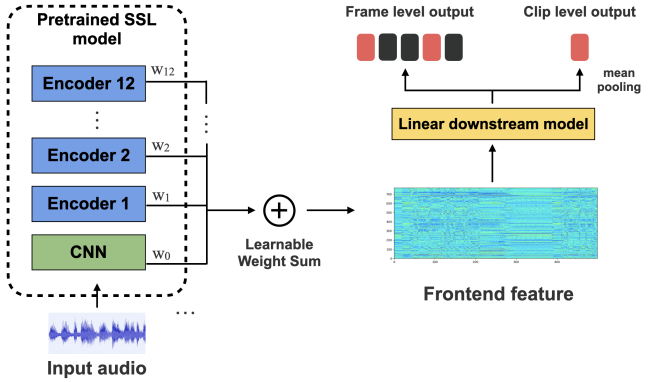


Fig. 2. The overview of the whole model to solve the target tasks.

B. Downstream model

We describe the overview of how to use the SSL model and the downstream models in Figure 2.

1) Weighted sum: We employed a weighted sum of the outputs from each Transformer encoder layer, including the input of the first layer, as the input for the downstream model. This approach was motivated by previous works (e.g.,[40], [41], [42], [43]) that have demonstrated different aspects of the input being captured by early, intermediate, and late layers of Transformer models. By using a weighted sum, we aimed to fully leverage the potential of SSL models. We set 13 learnable weight values, each corresponding to the weight value assigned to the output of each layer.

2) Classifier: Given the aforementioned features, we use a linear downstream model to predict each. The way of deriving the output of the target is depending on the tasks; For singing voice transcription, we directly used the frame-wise output. For the classification tasks (i.e., singer identification and singing technique classification), the feature is mean-pooled over the time axis to derive a clip-wise output.

C. Fine-tuning

In order to solve the downstream tasks, we fine-tune the models. There are various strategies for fine-tuning large-scale pre-trained models, we follow the two-stage training as Gu et al. [18] did. First, we freeze the parameter of the SSL models and make them learnable only on downstream models (i.e., the value of the weighted sum and the linear model). After several epochs, we unfreeze the Transformer encoders and fine-tune them.

IV. Experiments and Results

A. Experiment on Singer Identification

Singer identification is the classification task that identifies who is singing in a given sung audio clip.

1) Experimental Condition: We demonstrate 20-way singer identification using Artist20 [44] dataset, which collects 20 singers' music tracks. The dataset contains six albums for each singer in the data set, for a total of 1,413 songs. We split the dataset per album in order to avoid the leakage of production information about an album over the training and test set. We assign four albums for *train*, one album for *validation*, and the rest one album for *test* subset. Since the audio clips of the dataset include accompaniment of musical instruments, we applied vocal separation using Demucs V4 [45]. Then, we split them into five-second chunks without overlapping at a sample rate of 16kHz and discarded non-vocal chunks by RMS filtering [17].

For training the models, we used Adam optimizer [46] with 30 epochs. We set the learning rate of 3×10^{-3} for the first stage on the first six epochs and 5×10^{-5} for the second stage with the remaining epochs. The batch size is set to 32.

We evaluated the models by the following metrics: F1-score, Top-2 accuracy, and Top-3 accuracy. For the baseline for the comparison, we adopt the **CRNN** model by Hsieh et al.[23]. The model takes a 128-dimensional mel spectrogram as input and consists of four convolutional layers and two GRU [47] layers. We re-implemented the model in order to measure Top-2 and Top-3 accuracy since [23] only reported F1-score.

2) Results: Table I shows the results of singer identification. SSL models are demonstrating superior performance compared to the conventional SoTA model (i.e., CRNN) of singer identification. Notably, WavLM exhibited the highest performance in terms of F1 score, while MapMusic2Vec excelled in achieving the highest accuracy for Top-2 and Top-3 accuracy. These outcomes collectively suggest that the pre-training of SSL models with either music or speech data is leveraged in encoding the information associated with the singers.

TABLE I
THE RESULTS OF SINGER IDENTIFICATION.

Methods	F1-score	Top-2	Top-3
Wav2Vec2.0	60.0	70.7	76.3
WavLM	**61.9**	70.2	76.4
MERT	56.8	68.4	75.6
MapMusic2Vec	59.6	**71.5**	**77.0**
CRNN [23]	49.5	63.4	71.3

B. Experiment on Singing Voice Transcription

Singing voice transcription refers to the process of converting sung audio signals into corresponding musical notes. In this paper, we employed the piano roll representation as the target for the singing voice transcription.

1) Problem Definition: We follow the settings of Wang et al. [48]. The target has four attributes: *onset*, *silence*, *pitch class*, and *octave*. The beginning of silence is considered as the offset instead of a direct estimation of them due to its difficulty. We set the pitch range from C2 (MIDI number 36, 65.41Hz) to B5 (MIDI number 83, 987.77Hz), therefore the target of *octave* is four classes (i.e., 2-5.) In addition, the class of inactive is added on *octave* and *pitch class*, respectively. Eventually, each frame contains 20-dimensional vectors as a target (i.e., *onset* and *silence* are binary, *pitch class* is five classes, and *octave* is 13 classes).

2) Experimental Condition: We use MIR-ST500 [48], which consists of 500 Chinese pop songs with manually annotated vocal melody notes. The authors of [48] provide the official data split that allocates 400 songs for the training set and 100 songs for the test set. Therefore, we followed the split as is, using the training set for the model training and the test set for the evaluation. Since the audio clips of the dataset include accompaniment of musical instruments, we applied vocal separation using Demucs V4 [45]. During the training of the model, we split the input audio into five-second chunks without overlaps at a sample rate of 16kHz. Given the input, the model outputs the prediction of the aforementioned 20-dimensional target vector per frame. The frame length is about 20ms. Suppose the prediction for *onset*, *silence*, *pitch class* and *octave* are $\hat{O}, \hat{S}, \hat{P}, \hat{V}$, the objective functions are as follows:

$$\mathcal{L}_{svt} = \frac{1}{T}\sum_{t=1}^{T}\Big[BCE(\sigma(\hat{O}_t), O_t, w_o) + BCE(\sigma(\hat{S}_t), S_t, w_s) + CE(\hat{P}_t, P_t) + CE(\hat{V}_t, V_t)\Big] \tag{1}$$

Where T denotes the number of frames in the input, $\sigma(\cdot)$ denotes the sigmoid function, BCE denotes binary cross-entropy loss, and CE denotes cross-entropy loss. Considering the imbalance between positive and negative samples, the weight is applied to the binary cross entropy loss. The values are $w_o = 15.0$ for onset, and $w_s = 1.0$ for silence, respectively. For training the models, we used Adam optimizer [46] with 30 epochs. We set the learning rate of 3×10^{-3} for the first stage and 5×10^{-5} for the second stage.

We follow the strategy of postprocessing as Gu et al. [18] for deriving the actual estimation. Briefly, each of the note properties is determined as follows:

- onset: Setting a threshold value of 0.4. If the prediction value is higher than the threshold and the local maximum, the frame is set to onset.
- offset: $\arg\min(\hat{S}_t > 0.5)$ of the estimated silence sequence and after the onset time.
- pitch class and octave (i.e., midi number): Assign the mode of the estimated value between the onset and the offset time.

To evaluate the performance of our model, we adopted three evaluation metrics proposed in [49], namely F1-score of COn (correct onset), COnP (correct onset and pitch), and COnPOff (correct onset, offset, and pitch). We utilized the mir_eval library to calculate these metrics, using the default parameters: 50 cents for pitch tolerance, 50 ms for onset tolerance, and the larger value between 50 ms and 0.2 of the note duration for offset tolerance.

In order to establish baselines for comparison, we considered several conventional works:

1) **EfficientNet-b0**: This approach is based on utilizing the EfficientNet-b0 model [50], which was originally proposed as the baseline for the MIR-ST500 task [48].
2) **JDC$_{note}$**: This model was trained on pseudo-labeled data obtained through quantization of automatically detected vocal melody contours [51].
3) **Wav2Vec2-Large**: This model employed the Wav2Vec2.0-Large model for the frontend and feeds only the last layer's output to the downstream model [18].

3) Results: Table II shows the results of singing voice transcription. MERT achieved the best score on COn and COnP among the four SSL models, while MapMusic2Vec demonstrated the best score on COnPOff. We observed that the speech models (i.e., Wav2Vec2.0 and WavLM) exhibit lower COnP compared to the music models (i.e., MERT and MapMusic2Vec). It suggests that the music models have already acquired the representation related to musical notes while the speech models lack such incorporation due to the gap between speech and singing voice (i.e., the length of stable pitch region, musical expression, etc.). In terms of performance comparisons with conventional works, every SSL model showed comparable performance with [48] and [51]. In addition, MERT outperforms Wav2Vec2Large in the COnP with fewer parameters.

TABLE II
RESULTS OF SINGING VOICE TRANSCRIPTION. ALL VALUES ARE EXPRESSED IN %. THE BEST-PERFORMING CONDITION AMONG ALL CONDITIONS IS HIGHLIGHTED IN BOLD, AND THE BEST-PERFORMING CONDITION AMONG SSL MODELS IS UNDERLINED.

Methods	COnPOff	COnP	COn
Wav2vec2.0	44.8	67.0	76.3
WavLM	44.4	67.3	76.9
MERT	46.7	**71.6**	78.2
MapMusic2Vec	50.7	70.0	77.9
EfficientNet-b0 [48]	45.8	66.6	75.4
JDC$_{note}$ [51]	42.2	69.7	76.2
Wav2Vec2-Large [18]	**52.4**	70.7	**78.3**

C. Experiment on Singing Technique Classification

Singing technique classification is the task that identifies a singing technique that appeared in a given input audio clip.

1) Experimental condition: We used VocalSet [52], which is a publicly available dataset that annotated singing techniques. VocalSet contains singing voices by 20 different professional singers (9 female and 11 male), performing 17 different singing techniques in various contexts. We selected ten techniques ("belt," "breathy," "inhaled singing," "lip trill," "spoken excerpt," "straight tone," "trill," "trillo," "vibrato," and "vocal fry") by all singers for the classification. We used the officially provided data split: 15 singers for the training set and 5 singers for the test set. We trimmed the silence from the audio and split it into non-overlapping chunks of 3 seconds. Typically the sampling rate of the audio tracks of VocalSet is 44.1kHz, the audio is resampled to 16kHz. For training the models, we used Adam optimizer [46] with 20 epochs. We set the learning rate of 3×10^{-3} for the first stage on the first five epochs and 5×10^{-5} for the second stage with the remaining epochs. The batch size is set to 16.

TABLE III
SELECTED SAMPLES FROM VOCALSET.

Label name	Type of fluctuation	Samples #
straight	None	1241
belt	Timbre	423
breathy	Timbre	455
vocal fry	Timbre, Modulation	587
vibrato	Modulation	1034
trill	Modulation	323
trillo	Modulation	242
lip trill	Modulation	376
inhaled	Other	151
spoken	Other	73

The data distribution of VocalSet is imbalanced over the classes that affect the classification performance [53]. Therefore, we adopt the inverse frequency weight for the loss function. The typical weight value w_c for the class c is as follows:

$$w_c = \frac{1}{(n_c)^\alpha} \tag{2}$$

where n_c is the number of training samples in class c, and α is the smoothing factor, which controls the smoothing of the loss weights. Note that α = 0 corresponds to the value of 1 (i.e., no weighting) and α = 1 corresponds to a reciprocal number (i.e., weighting by the inverse class frequency). We set α = 0.2, which performed the best score in [53]. We evaluated the models by the following metrics: F1-score, Accuracy, Balanced accuracy, Top-2 accuracy, and Top-3 accuracy. We consider several conventional works as baselines.

1) **1DCNN**: This model utilizes a CNN architecture that directly takes raw waveform as inputs. It serves as the official baseline model of the Vocalset dataset [52].
2) **OblongCNN**: This model employs a CNN architecture that takes a multi-resolution spectrogram, consisting of stacked representations with three different time-frequency resolutions as input. Additionally, it incorporates four convolutional layers of varying shapes [8].
3) **D-CNN-cRT**: This model replaces the standard convolutional layers with Deformable convolution and employs Classifier Retraining (cRT) [54] for training with a focus on addressing class imbalance [53].

2) Results: Table IV shows the results of singing technique classification. MapMusic2Vec exhibits the best performance in four SSL models and comparable performance to other conventional approaches. It also achieved higher accuracy compared to the best-performing method among the conventional works, D-CNN-cRT.

D. Layer-wise contribution analysis

We further analyze the weight of each encoder layer's output for each SSL model. Figure 3, 4, and 5 show the weights after training on singer identification, singing voice transcription, and singing technique classification, respectively. In each figure, 'Ln' denotes the weight of the n-th encoder layer's output (i.e., L0 is for the input of the first layer.).

TABLE IV
THE RESULTS OF SINGING TECHNIQUE CLASSIFICATION.

Methods	F1	Acc	BAcc	Top-2	Top-3
Wav2Vec2.0	56.1	58.3	60.1	74.8	82.1
WavLM	55.6	60.8	57.9	75.3	83.8
MERT	54.1	58.5	58.5	76.1	85.3
MapMusic2Vec	<u>60.8</u>	**<u>66.0</u>**	<u>62.1</u>	<u>79.0</u>	<u>86.9</u>
1DCNN [52]	48.8	58.4	48.4	76.4	86.3
OblongCNN [8]	51.3	55.4	57.5	74.3	85.8
D-CNN-cRT [53]	**62.0**	65.6	**65.5**	**81.5**	**88.7**

For singer identification, the strong contribution lies in the early layers of each SSL model. This finding aligns with previous research investigating the layer-wise contribution in various speech-related tasks, such as those presented in several works [13], [55]. These studies suggest that speaker-related information tends to be captured in the early layers of the models. Consequently, it can be inferred that a similar pattern holds true for singers, indicating that the crucial information for singer identification is also encoded in the early layers of the SSL models.

For singing voice transcription, except for MERT, the contribution also tends to lie in the early layers of each SSL model. According to [41], low-level prosodic features such as pitch or loudness tend to locate the early layers of speech SSL models. Since pitch information is the most important for singing voice transcription, the results that align with the work are unsurprising. In the case of MERT, CQT spectrogram is utilized as a prediction target. Therefore, it is plausible that the last layers of MERT contain pitch-related information. As several works [56], [57] reported, singing voice transcription is relating the information of phonemes as well as the pitch (i.e., onsets are often accompanied by the transition of the lyrics). It might cause this "unsmooth-shaped" distribution compared to other tasks.

In the case of singing technique classification, the early layers exhibit greater values compared to other layers in each model. Specifically, the first layer holds the greatest significance. It might be because singing techniques are influenced by pitch modulation and timbre variation. However, there is also a possibility that it is affected by the data imbalance where vibrato and straight, whose characteristics are the status of pitch modulation, are the majority classes.

V. CONCLUSION

In this study, we address the challenge of limited data availability in singing voice understanding tasks by leveraging transfer learning of pre-trained self-supervised (SSL) models. We employ four different SSL models as the frontend of our target task model. The models are subjected to comprehensive experiments encompassing singer identification, singing voice transcription, and singing technique classification. Our experimental results demonstrate that each SSL model achieves comparable and in some cases superior, performance when compared to conventional state-of-the-art models. Additionally, we conduct layer-wise analysis to inspect the behaviors, specif-

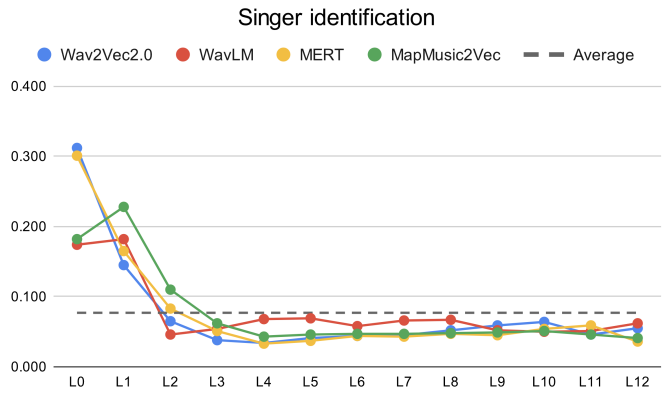


Fig. 3. The weights after training on singer identification.

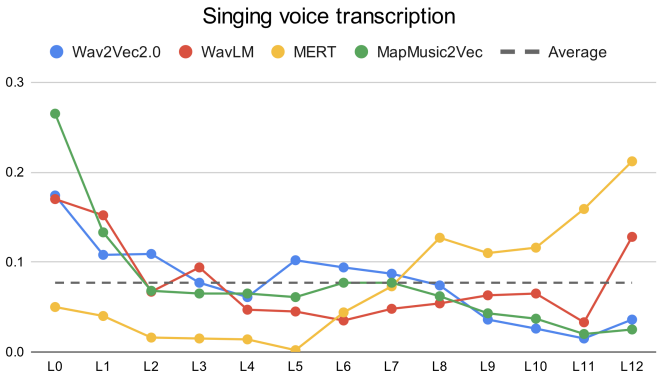


Fig. 4. The weights after training on singing voice transcription.

ically analyzing the weights of each layer. Moving forward, our future research endeavors will focus on further investigating the impact of feature representation in addressing the data scarcity issue in automatic singing voice understanding tasks. Furthermore, we propose that future studies can explore other singing voice understanding tasks, such as vocal melody extraction [58], lyric transcription [59], and singer diarization [60], among others, to expand the scope of research in this domain.

VI. Acknowledgements

I am grateful to Hiroko Terasawa and Juhan Nam, for their fruitful comments on the methods and evaluation of models. I also appreciate Jun-You Wang from National Taiwan University, for assisting in the preparation of *MIR-ST500*.

This work was supported by JST SPRING, Grant Number JPMJSP2124.

References

[1] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: Musical information retrieval in an audio database," in *Proceedings of the third ACM international conference on Multimedia*, 1995, pp. 231–236.

[2] T. Nakano, M. Goto, and Y. Hiraga, "Mirusinger: A singing skill visualization interface using real-time feedback and music cd recordings as referential data," in *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*. IEEE, 2007, pp. 75–76.

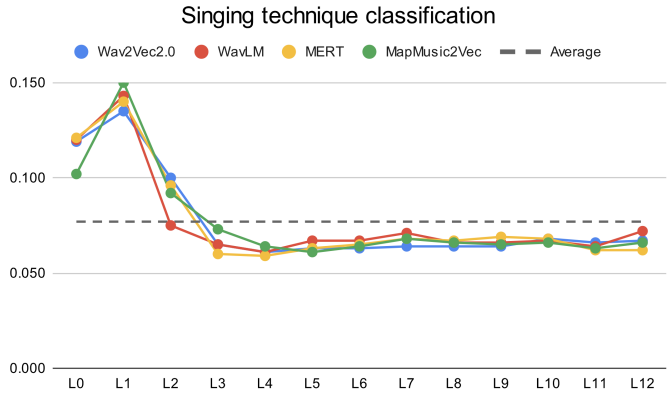


Fig. 5. The weights after training on singing technique classification.

[3] M. Panteli, R. Bittner, J. P. Bello, and S. Dixon, "Towards the characterization of singing styles in world music," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 636–640.

[4] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Kruspe, and L. Yang, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 82–94, 2018.

[5] C. Gupta, H. Li, and M. Goto, "Deep learning approaches in topics of singing information processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2422–2451, 2022.

[6] Z.-S. Fu and L. Su, "Hierarchical classification networks for singing voice segmentation and transcription," in *The 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[7] Z. Nasrullah and Y. Zhao, "Music artist classification with convolutional recurrent neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[8] Y. Yamamoto, J. Nam, H. Terasawa, and Y. Hiraga, "Investigating time-frequency representations for audio feature extraction in singing technique classification," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 890–896.

[9] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[10] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[14] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[15] Y. Li, R. Yuan, G. Zhang, Y. Ma, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He *et al.*, "Map-music2vec: A simple and effective baseline

for self-supervised music audio representation learning," *arXiv preprint arXiv:2212.02508*, 2022.

[16] L. Ou, X. Gu, and Y. Wang, "Transfer learning of wav2vec 2.0 for automatic lyric transcription," in *The 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[17] M. Heydari and Z. Duan, "Singing beat tracking with self-supervised front-end and linear transformers," in *The 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[18] X. Gu, W. Zeng, J. Zhang, L. Ou, and Y. Wang, "Deep audio-visual singing voice transcription based on self-supervised learning models," *arXiv preprint arXiv:2304.12082*, 2023.

[19] C. Donahue, J. Thickstun, and P. Liang, "Melody transcription via generative pre-training," in *The 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[20] S. Kum, J.-H. Lin, L. Su, and J. Nam, "Semi-supervised learning using teacher-student models for vocal melody extraction," in *The 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[21] J.-Y. Hsu and L. Su, "Vocano: A note transcription framework for singing voice in polyphonic music." in *The 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 293–300.

[22] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." in *The 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 121–126.

[23] T.-H. Hsieh, K.-H. Cheng, Z.-C. Fan, Y.-C. Yang, and Y.-H. Yang, "Addressing the confounds of accompaniments in singer identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5.

[24] H. Yakura, K. Watanabe, and M. Goto, "Self-supervised contrastive learning for singing voices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1614–1623, 2022.

[25] C. Noufi and P. Verma, "Self-supervised learning of context-aware pitch prosody representations," *arXiv preprint arXiv:2007.09060*, 2020.

[26] S. Basak, S. Agarwal, S. Ganapathy, and N. Takahashi, "End-to-end lyrics recognition with voice to singing style transfer," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2021, pp. 266–270.

[27] C. Zhang, J. Yu, L. Chang, X. Tan, J. Chen, T. Qin, and K. Zhang, "Pdaugment: Data augmentation by pitch and duration adjustments for automatic lyrics transcription," in *The 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[29] E. Demirel, S. Ahlbäck, and S. Dixon, "Automatic lyrics transcription using dilated convolutional neural networks with self-attention," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[30] ——, "Mstre-net: Multistreaming acoustic modeling for automatic lyrics transcription," in *The 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[31] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[32] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *The 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[33] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[34] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[35] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971.

[36] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[37] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[38] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[39] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[40] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 251–16 265, 2021.

[41] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1104–1111.

[42] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[43] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[44] D. P. Ellis, "Classifying music audio with timbral and chroma features," in *The 8th International Conference for Music Information Retrieval Conference (ISMIR)*, 2007.

[45] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[47] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[48] J.-Y. Wang and J.-S. R. Jang, "On the preparation and validation of a large-scale dataset of singing transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 276–280.

[49] E. Molina, A. M. Barbancho-Perez, L. J. Tardon-Garcia, I. Barbancho-Perez *et al.*, "Evaluation framework for automatic singing transcription," in *The 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[50] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[51] S. Kum, J. Lee, K. L. Kim, T. Kim, and J. Nam, "Pseudo-label transfer from frame-level to note-level in a teacher-student framework for singing transcription from polyphonic music," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2022, pp. 796–800.

[52] J. Wilkins, P. Seetharaman, A. Wahl, and B. A. Pardo, "Vocalset: A singing voice dataset," in *The Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 468–474.

[53] Y. Yamamoto, J. Nam, and H. Terasawa, "Deformable CNN and Imbalance-Aware Feature Learning for Singing Technique Classification," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022, pp. 2778–2782.

[54] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *International Conference on Learning Representations (ICLR)*, 2020.

[55] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proceedings of the IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2022, pp. 6147–6151.

[56] S. Yong, L. Su, and J. Nam, "A phoneme-informed neural network model for note-level singing transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2023.

[57] T. Deng, E. Nakamura, and K. Yoshii, "End-to-end lyrics transcription informed by pitch and onset estimation," in *The 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[58] K. S. Rao, P. P. Das *et al.*, "Melody extraction from polyphonic music by deep learning approaches: A review," *arXiv preprint arXiv:2202.01078*, 2022.

[59] X. Gao, "Automatic lyrics transcription of polyphonic music," Ph.D. dissertation, National University of Singapore (Singapore), 2022.

[60] H. Suda, D. Saito, S. Fukayama, T. Nakano, and M. Goto, "Singer diarization for polyphonic music with unison singing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1531–1545, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Out-of-Vocabulary Word Detection in Spoken Dialogues Based on Joint Decoding with User Response Patterns

Miki Oshio*, Hokuto Munakata*, Ryu Takeda* and Kazunori Komatani*
* Osaka University, The Institute of Scientific and Industrial Research (SANKEN) Osaka, Japan
E-mail: {m-oshio, h_munakata}@ei.sanken.osaka-u.ac.jp, {rtakeda, komatani}@sanken.osaka-u.ac.jp

***Abstract*—We detect out-of-vocabulary (OOV) words from spoken utterances. A straightforward cascade model, where syllable recognition and word segmentation are simply concatenated, often fails to correctly segment syllables around OOV words. Because OOV words in spoken dialogues with systems do not appear in the training data, some reasonable constraints in such situations are necessary. In this study, we propose a method combining two ideas to detect OOV words. First, we jointly decode syllable sequences and word boundaries in them using user response patterns, which exist in the user responses to explicit confirmations from a spoken dialogue system. These patterns serve as constraints on vocabulary in utterances and the positions of OOV words. We represent OOV words as arbitrary syllable sequences and use finite-state transducers (FSTs) containing both OOV words and in-vocabulary words as constraints during decoding. Joint decoding with these constraints can also reduce syllable recognition errors. Second, we automatically select the better results between the baseline cascade model and the joint model through model selection. This effectively improves the detection performance as not all user utterances follow the patterns. Experimental results show improvement in the F-score for OOV word detection.**

I. Introduction

A. Motivation

It is important for spoken dialogue systems to recognize out-of-vocabulary (OOV) words. These are words not included in a vocabulary list utilized by automatic speech recognition (ASR) in a dialogue system. Traditional ASR having a finite set of words as its vocabulary cannot recognize OOV words. If OOV words are not recognized, the system cannot respond appropriately. Therefore, we need an ASR that can recognize OOV words. This process also leads to acquiring them as a new vocabulary [1].

Syllable recognition [2] is a promising approach for recognizing OOV words instead of word-based ASRs. Since the performance of syllable recognition by a deep neural network (DNN) has improved [2], [3], we use it in this study. A straightforward approach for recognizing OOV words is to simply concatenate syllable recognition and word segmentation [4]. We call this approach the cascade model. This model is effective for segmenting frequent syllable patterns in the training data. However, OOV words in spoken dialogue systems do not generally appear in large training data. Thus, some reasonable constraints are required on the appearance positions of OOV words and vocabulary in user utterances.

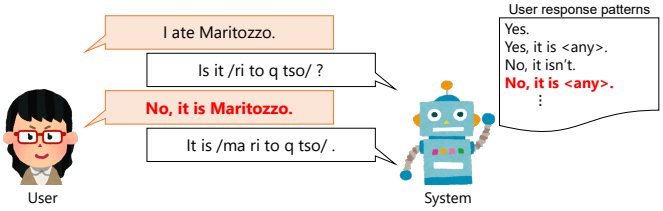


Fig. 1. Ideal example of OOV word detection using explicit confirmation from a spoken dialogue system. Although the first user's utterance is incorrectly recognized as "/ri to q tso/," the correct recognition of "/ma ri to q tso/" is obtained by using the user response pattern.

We focus here on a situation after receiving explicit confirmations from spoken dialogue systems. An example is shown in Fig. 1. In this situation, after recognizing the user's first utterance "I ate Maritozzo," the system explicitly confirms it by repeating recognized syllable sequences. In situations where a system asks a question as an explicit confirmation, the user often responds using specific patterns. These patterns can be used as reasonable constraints for OOV word detection. We aim to correctly detect OOV words by using these user response patterns as constraints (shown on the right in Fig. 1). In Fig. 1, although the user's utterance is initially misrecognized as "/ri to q tso/," the correct recognition of "/ma ri to q tso/" is obtained thanks to the use of the user response patterns.

In this study, we propose a method that combines two ideas for OOV word detection. The first is a joint decoding of syllable recognition and word segmentation with the user response patterns. The patterns are described as a finite-state transducer (FST) [5], [6] and are used during decoding. In addition, the joint decoding of syllable recognition and word segmentation can reduce errors in syllable recognition.

The second is automatic model selection, that is, to select either of the recognition results obtained from two models: the cascade or joint model. Because users can speak arbitrarily, not all user responses can be included in the FST. OOV word detection with model selection is expected to yield better results for user utterances that do not match the user response patterns. We then automatically select an OOV word detection result

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

from the two, leading to improved performance. We compared the logarithmic scores of syllable sequences recognized by the two models during the model selection.

B. Related Works

Recently, there have been many studies on end-to-end speech recognition for OOV word detection. Each of these studies has varied in terms of approaches [7]. One example is a study that accurately recognizes OOV words using external information such as knowledge graphs [7]. However, this method assumes that OOV words are included in the entities of the knowledge graph, making it unsuitable for scenarios involving entirely new words not present in the existing knowledge. Another approach is to use a hybrid system that combines multiple models, such as word, subword, and phoneme models [8]–[11]. In particular, [9] takes advantage of the trend to repeat OOV words in dialogue.

There are also several studies that aim to acquire OOV words from speech without using end-to-end speech recognition [12], [13]. Neubig et al. created a language model by extracting word-phoneme sequences and transition probabilities from spoken utterances without using any prior linguistic information [12]. Shinozaki et al. extracted phoneme sequences of OOV words from spoken utterances and expanded the system's internal pronunciation dictionary [13]. Their study assumes the system has an existing pronunciation dictionary in advance.

In addition, there are several studies on unsupervised word segmentation directly from speech [14]. Kamper et al. focused on infants' language acquisition from speech in their native language. In their experiment, only 11 vocabulary words, from "zero" to "nine" and "oh", were used in the evaluation. However, this is a very primitive method that uses clustering to identify words, which makes it very difficult to recognize thousands of words in real dialogues.

The biggest difference between the aforementioned studies and our study is the situation of asking back to OOV words. Unlike ASR systems, spoken dialogue systems can ask questions to the user as needed.

II. Preliminaries

A. Problem Settings

Our goal is to detect OOV words contained in a user's speech. Specifically, we estimate a syllable sequence and word boundaries from the user's speech signal, to obtain a word sequence. We regard words that are not included in the vocabulary list as OOV words. Syllable sequences are represented in katakana. The vocabulary list is a set of unique words, which are registered in the form of syllable sequences.

We treat the simultaneous estimation of syllable sequences and word boundaries in OOV word detection as a search problem to maximize the posterior probability.

$$\hat{\mathbf{c}}, \hat{\mathbf{z}} = \mathrm{argmax}_{\mathbf{c},\mathbf{z}} p(\mathbf{c}, \mathbf{z}|\mathbf{X}) \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T], \mathbf{x}_t \in \mathbb{R}^D$ $(t = 1, ..., T)$ is the input speech feature sequences of length T, $\mathbf{c} = [c_1, ..., c_L]$ is a corresponding syllable sequence of length L, and $\mathbf{z} = [z_1, ..., z_L], z_l \in \{0, 1\}$ $(l = 1, ..., L)$ denotes a sequence of binary indicator variables of length L representing word boundaries. If a word boundary immediately follows a syllable c_l, $z_l = 1$; otherwise, $z_l = 0$. For example, if $\mathbf{c} =$ [a, me, ka, ze] , and $\mathbf{z} = [0, 1, 0, 1]$, we will obtain two word tokens, "ame" and "kaze". If there is no entry for "kaze" in the vocabulary list V, "kaze" is recognized as an OOV word.

For simplicity, we approximate the posterior probability by the following conditional probability. We then reformulate it using syllable recognition.

$$p(\mathbf{c}, \mathbf{z}|\mathbf{X}) \approx p(\mathbf{c}|\mathbf{X}) p(\mathbf{z}|\mathbf{c}), \tag{2}$$

where we define two models. We define $p(\mathbf{c}|\mathbf{X})$ as the syllable recognition model and $p(\mathbf{z}|\mathbf{c})$ as the word segmentation model. Our goal is to find $\hat{\mathbf{c}}, \hat{\mathbf{z}}$ that maximizes the posterior probability.

B. Simple OOV Word Detection Method: Cascade Model

The cascade model is that in which syllable recognition and word segmentation are concatenated sequentially. Syllable recognition and word segmentation are performed in accordance with the formula as:

$$\hat{\mathbf{c}} = \mathrm{argmax}_{\mathbf{c}} p(\mathbf{c}|\mathbf{X}), \text{and} \tag{3}$$

$$\hat{\mathbf{z}} = \mathrm{argmax}_{\mathbf{z}} p(\mathbf{z}|\hat{\mathbf{c}}) \tag{4}$$

In syllable recognition, we estimate a syllable sequence $\hat{\mathbf{c}}$ from a sequence of speech features $\mathbf{X}$. In word segmentation, we estimate a word boundary variable sequence $\hat{\mathbf{z}}$ from the estimated syllable sequence $\hat{\mathbf{c}}$. We then obtain a word sequence from the estimated syllable sequence $\hat{\mathbf{c}}$ and the word boundary variable sequence $\hat{\mathbf{z}}$, and OOV word judgment is performed. After that, we check if these words are included in the vocabulary list, and regard words that are not included as OOV words.

III. Joint Decoding with the User Response Pattern

OOV word detection using the cascade model is often incorrect because this model does not utilize information about the context of the dialogue and the location of OOV words. Therefore, we use the joint model. In this model, we utilize a specific situation where a user responds to an explicit confirmation from the system. During decoding, we jointly decode syllable sequences and word boundaries. By using a model specialized for this situation based on the user response pattern, we aim to detect OOV words more accurately than the cascade model.

User responses to explicit confirmations from the system show several patterns, particularly regarding the position of OOV words and the vocabulary included in the user's responses. To construct the user response pattern model, we first prepare a set of user response patterns U from the user response data (syllable sequences and word boundaries) provided, which are represented using an FST. Subsequently, we use the FST for joint decoding of syllable recognition and word segmentation.

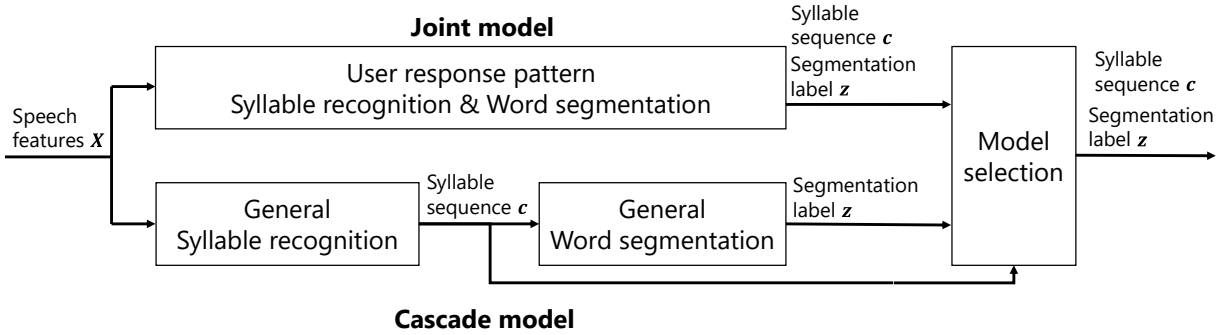


Fig. 2. Framework for our proposed method. The cascade and joint models are used together with model selection.

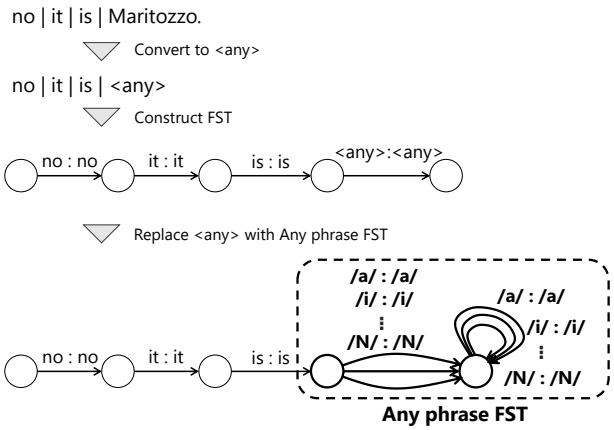


Fig. 3. How to construct FST from user utterances data.

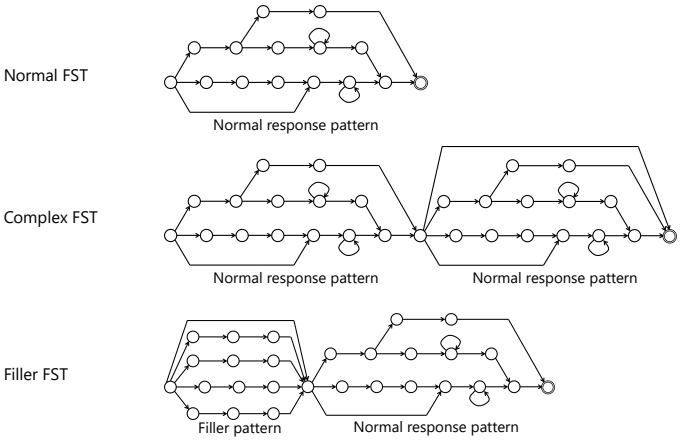


Fig. 4. Three kinds of FST.

Because not all user utterances follow the patterns, we use the model selection to flexibly detect OOV words for utterances that do not match the user response patterns. We select either of the recognition results obtained from the two models as shown in Fig. 2.

A. Joint Model

1) Formulation: In joint decoding, in accordance with Eq. (2), we jointly estimate syllable sequences $\mathbf{c}$ and binary indicator variable sequences $\mathbf{z}$ from speech features $\mathbf{X}$. We incorporate the user response pattern FST into an end-to-end ASR recognition model that uses language model scores, such as the hybrid CTC/Attention model [2]. $p(\mathbf{z}|\mathbf{c})$ corresponds to the language model.

By replacing the language model with the following function $f(\mathbf{z}, \mathbf{c})$, we can express the equivalent of excluding syllable sequences that are not accepted by FST from the candidate recognition results.

$$f(\mathbf{z}, \mathbf{c}) = \begin{cases} \text{const.} & \text{if } [\mathbf{z}, \mathbf{c}] \in U \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

2) Construction of User Response Pattern FST: In this method, a user response pattern FST is constructed from the user response data to explicit confirmations from the system. First, candidate OOV words, which can be replaced with OOV words, in the user response data are converted into an "any tag": <any>. In Fig. 3, Maritozzo is replaced with "any tag": <any> as a candidate OOV word. Then, an FST is constructed from the converted data to ensure only the response data can be recognized. Next, "any tags" are converted to Any phrase FST, which accepts an arbitrary syllable sequence in accordance with the following two rules to make it a natural Japanese word. In syllable recognition, if the Any phrase FST is generated without any constraints, syllable sequences that do not follow Japanese syllable chain rules may be generated, so it is necessary to add rules to make it natural Japanese.

The first rule is to reject only single-syllable words because a word corresponding to an OOV word rarely has only one syllable. The second rule is to reject sequences of syllables that cannot exist in the Japanese language. This FST only accepts patterns that are contained in the training data.

In addition to the normal response pattern FST, we construct two FST: one for complex responses (Complex FST) and one for responses with a filler (Filler FST), which increases the number of utterances that the FST accurately matches (in Fig. 4). This is because the normal response pattern FST accepts only the response patterns included in the user response data used to construct the FST. There will be cases in which a

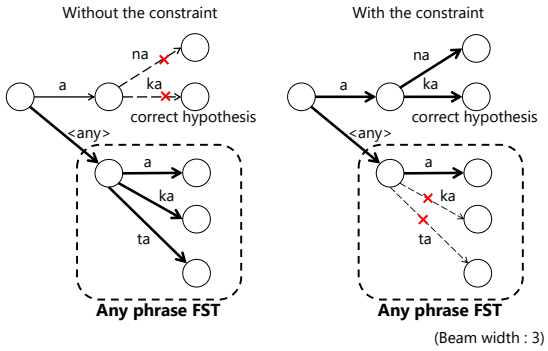


Fig. 5. The constraint in which only one hypothesis in the Any phrase FST is retained in order not to overflow the correct hypothesis. In this figure, all candidates except for the best candidate ("a") are excluded.

response is given in complex sentences, such as "No, it isn't. It is Maritozzo." There will also be cases in which a reply is given in sentences with a filler, such as "Uh, it is Maritozzo." Therefore, we construct three kinds of FST in Fig. 4 to accept such utterances. The Complex FST includes additional patterns that repeat the normal response pattern twice. The Filler FST includes additional patterns that contain filler at the beginning of a sentence in Fig. 4.

3) Decoding with User Response Pattern FST and post-processing: In decoding, hypotheses are searched for using beam search. Since the FST includes the Any phrase FST, we need to apply certain constraints. The correct hypothesis out of the Any phrase FST will overflow from the beam width without any constraints. Consequently, the correct hypothesis may not be recognized.

We set a constraint where only one hypothesis is retained in the Any phrase FST (Fig. 5). In Fig. 5, only the best hypothesis "a" is retained, while the other hypotheses are rejected. This reduction in the Any phrase FST can help streamline the beam search process.

After obtaining the word sequence by using the FST, word segmentation is performed again for "any phrase", which is compatible with the Any phrase FST (post-processing). This meant that "any phrase" may include known words in this constraint. By post-processing, our approach can handle cases where "any phrase" contains a known word.

B. Model Selection

It is possible to flexibly recognize utterances that do not match the pattern by selecting which recognition results to adopt, either from the cascade model or the joint model. The joint model is inflexible and does not accurately recognize utterances that do not match the FST. Thus, we recognize such utterances by using the cascade model.

Model selection is determined by the score of the syllable recognition model $p(\mathbf{c}|\mathbf{X})$. This is because utterances that do not match the FST are recognized as syllable sequences that are significantly different in the joint model. Thus, we measure the difference between the log scores of the results obtained from the cascade model and those from the joint model to determine the more suitable recognized results on the basis of threshold θ. The search for $\hat{\mathbf{c}}, \hat{\mathbf{z}}$ is performed as:

$$\hat{\mathbf{c}}, \hat{\mathbf{z}} = \begin{cases} \mathbf{c}_d, \mathbf{z}_d & \text{if } \log(p(\mathbf{c}_g|\mathbf{X})) - \log(p(\mathbf{c}_d|\mathbf{X})) \leq \theta \\ \mathbf{c}_g, \mathbf{z}_g & \text{otherwise} \end{cases} \quad (6)$$

where $\mathbf{c}_g, \mathbf{z}_g$ and $\mathbf{c}_d, \mathbf{z}_d$ are the syllable sequences and the sequences representing the word boundaries of the recognition results using the cascade and joint models.

TABLE I
AMOUNT OF DATA

Data	CSJ full	CSJ core	User responses
Dialogues	380,872	63,387	553
Syllables	14,789,778	1,074,846	6,337
Words	7,251,909	499,452	2,835
Vocabulary size	54,463	13287	266
OOV words	0	0	169

IV. EXPERIMENTS

We examined the effect of the proposed method on OOV word detection. The cascade model described in Section 2 was used as a baseline for comparison. We discussed the detection of OOV words on the basis of the experimental results using the proposed method.

A. Configuration

In syllable recognition, we used ESPnet [15], in which we used the hybrid CTC/Attention model and Transformer for the encoder and decoder. In word segmentation of the cascade model and post-processing of the joint model, we used the Nested Pitman-Yor Language Model (NPYLM) [4]. To construct the FSTs, we used OpenFst [16]. Then, the nouns corresponding to foreign foods such as "Maritozzo" were assumed to be OOV words and replaced with "any tag" <any>. The value of the constant in Eq. (5) was set to 1.0. The model selection's threshold θ in Eq. (6) was set to 0.5 as the best threshold by experimenting with several values.

The cascade model was used as the baseline in the experiment. We compared the performance of the proposed method, which involved the model selection methods of the cascade and joint models. To compare FSTs, OOV word recognition was performed for the methods using the normal FST, Complex FST, and Filler FST. In addition, we also experimented with the joint model only without model selection, to confirm the effect of model selection.

The evaluation metrics used were precision, recall, and F-score of OOV word detection. We also evaluated the character error rate (CER) of syllable recognition, and precision, recall, and F-score of word segmentation to assess the performance of syllable recognition and word segmentation. For OOV word detection, we evaluated the correctness of the start and end positions of OOV words in a syllable sequence. For OOV word recognition, we evaluated whether the estimated OOV words matched the OOV words in the reference data.

TABLE II
EVALUATION FOR EACH MODEL

	Configuration		Word segmentation			OOV word detection			OOV word recognition		
	With FST	Model selection	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Baseline			0.901	0.953	0.926	0.350	0.354	**0.352**	0.281	0.285	0.283
Proposed	√	√	0.909	0.959	0.933	0.388	0.342	**0.364**	0.309	0.272	0.289
Only FST	√		0.748	0.925	0.827	0.289	0.277	**0.283**	0.197	0.189	0.193

TABLE III
COMPARISON OF THE THREE TYPES OF FST

	Configuration	Word segmentation			OOV word detection			OOV word recognition		
	With FST	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Proposed	Normal	0.910	0.938	0.924	0.378	0.342	**0.359**	0.301	0.272	0.286
	Complex	0.909	0.959	0.933	0.388	0.342	**0.364**	0.309	0.272	0.289
	with Filler	0.921	0.874	0.897	0.295	0.373	**0.330**	0.210	0.266	0.235

TABLE IV
CER OF SYLLABLE RECOGNITION

		Baseline	Proposed	Only FST
Configuration	FST		√	√
	Model selection		√	
Correctness		**0.935**	**0.938**	**0.907**
Substitution error rate		**0.041**	**0.039**	**0.060**
Deletion error rate		0.024	0.023	0.032
Insertion error rate		0.012	**0.012**	**0.164**

B. Train Data and Test Data

The Corpus of Spontaneous Japanese (CSJ) [17] full dataset was used to train syllable recognition models for ESPnet. The CSJ core dataset was used to train the word segmentation model using NPYLM. CSJ full dataset contains 660 hours of Japanese speech data from individuals of various ages and genders. CSJ core dataset is part of the CSJ full dataset, about 500,000 words. In addition, spoken dialogues between users and the system were used for the construction of the user response patterns FST and test data. These dialogues were originally recorded in our laboratory and were conducted on the theme of food in foreign countries, which was decided in advance. The system prompted the user for utterances containing OOV words in dialogues. In this experiment, we used only the responses in the dialogue data and performed 4-fold cross-validation. The data was divided into four blocks, with one block as the test data, and the remaining for constructing the FST for the user response patterns.

Table I shows the size of the data used in this experiment. The set of unique words in CSJ was used as the vocabulary list for the system, resulting in 54,463 words. The row of OOV words in Table I indicates the number of words in the data that are not included in the vocabulary list.

C. Results

Table II shows the results of OOV word detection for three models. In the table, the Complex FST was used in the proposed method, which had the highest F-score for OOV word detection among the three FSTs. The F-score serves as the main evaluation criterion, while recall and precision indicate a reference. A comparison of each FST will be presented later.

First, for the F-score in OOV word detection, the proposed method increased by 0.012 points from 0.352 to 0.364, compared with the baseline. Second, the proposed method also increased the F-scores for both word segmentation and OOV word recognition. The proposed method increased the F-score for word segmentation by 0.007 points from 0.926 to 0.933, and that for OOV word recognition by 0.006 points from 0.283 to 0.289, compared with the baseline. Third, model selection is an effective method. For comparison, we also experimented with the method of decoding with only the FST without model selection in the row "Only FST." This method yielded lower values than the proposed method and the baseline for the F-score in OOV word detection. The F-score decreased by 0.081 points from 0.364 to 0.283, compared with the proposed method. This lower performance is likely due to the usage of the FST for utterances that did not match the FST.

Table III shows the results of each FST model: normal FST, Complex FST, and Filler FST. Among these models, the Complex FST model achieved the highest F-score for OOV word detection. The Complex FST model increased the F-score by 0.005 points from 0.359 to 0.364, compared with the normal FST model. In addition, the Complex FST model also has the highest F-score among the three models for word segmentation and OOV word recognition. The Complex FST model increased the F-score of word segmentation by 0.009 points from 0.924 to 0.933 and the F-score of OOV word recognition by 0.003 points from 0.286 to 0.289, compared with the normal FST model.

Next, we discussed the performance of syllable recognition. Table IV shows the CER for each model. The correctness in the first row represents the rate of syllables recognized without character errors. The proposed method reduced the error rate by 0.002 from 0.041 to 0.039 and increased the correctness by 0.003 points from 0.935 to 0.938, compared with the baseline. This indicates that the proposed method improves the performance of syllable recognition. We also compared the method of decoding with only FST without model selection. In the only FST model, the insertion error

TABLE V
RECOGNITION EXAMPLE. THE FST MODEL CAN SPLIT CORRECTLY. THE CASCADE MODEL RECOGNIZED A KNOWN WORD AS AN OOV WORD.

Model	Recognized text	Predicted OOV word	OOV word detection Precision
Reference	o ra N da | no | ta be mo no | da | yo	-	-
Cascade model	o | ra N da | no | ta be mo no | da | ya	/ra N da/	-
FST model	o ra N da | no | ta be mo no | da | ya	-	$\surd$

TABLE VI
RECOGNITION EXAMPLE. THIS UTTERANCE DOES NOT FIT THE NORMAL FST, BUT FITS THE COMPLEX FST.

Model	Recognized text	OOV word
Reference	ma ri to q tso | de su | i ta ri a | no | ta be mo no | de su	/ma ri to q tso/
Cascade model	ma | ri to q cho: | de su | i ta ri a | no | ta be mo no | de su	/ri to q cho:/
Normal FST model	ma | ri to q cho: | de su | i ta ri a | no | ta be mo no | de su | mi su | mi ta i | na | ka N ji	/ri to q cho:/
Complex FST model	ma | ri to q cho | de su | i ta ri a | no | ta be mo no | de su	/ri to q cho:/

TABLE VII
RECOGNITION EXAMPLE. THERE ARE MANY INSERTION ERRORS IN THE RESULT OF THE USER RESPONSE PATTERN MODEL

Model	Recognized text	Syllable recognition score
Reference	e: q to | ne | ne pa: ru | no | ta be mo no | de | da ru ba: to | da | yo	-
Cascade model	e: q to | ne | ne pa: ru | no | ta be mo no | de | da ru ba: to | da | yo	-0.119
FST model	e: q to | ne | ne pa: ru | no | ta be mo no | q te | yu: | o | ka shi | na | N | da | ke do | ne pa: ru | no | ta be mo no | da | yo	-4.308

increased significantly by 0.152 points from 0.012 points to 0.164 points, compared with the proposed method. In addition, only the FST model reduced the correctness by 0.031 points from 0.938 to 0.907. Thus, model selection is effective for reducing errors in syllable recognition.

An example of OOV word detection using user response patterns is shown in Table V. The cascade model split "/o ra N da/ (Netherlands)" into "/o/" and "/ra N da/," and "/ra N da/" was incorrectly regarded as an OOV word. On the other hand, the user response pattern model did not split "/o ra N da/ (Netherlands)" and guessed that it is a known word.

Next, we compared the models using the normal FST and the Complex FST. Table VI shows an example of the recognition results. For utterances that do not match the user response pattern, the normal FST model recognition result differs significantly from the reference. On the other hand, the Complex FST model can correctly recognize them because they correspond to the response pattern. However, in the situation of Fig. VI, "any phrase" is "/ma ri to q cho:/" in accordance with the response pattern, but the post-processing split it into "/ma/" and "/ri to q cho:/". Therefore, the Complex FST model resulted in incorrect OOV word detection.

Table VII shows increasing insertion error of recognition results. The syllable recognition score in the table is the difference between the syllable recognition scores of the two models. In this example, this utterance matches the Complex FST but does not match normal FST. When the FST model was used for utterances that did not match the FST, many insertion errors "/mi su/ /mi ta i/ /na/ /ka N ji/" occurred in the normal FST. The increase in insertion errors results in a lower likelihood of syllable recognition. Therefore, by setting an appropriate threshold, we could choose the appropriate model between the cascade model and the user response pattern model.

V. CONCLUSION

In this paper, we focused on OOV word detection of user responses during explicit confirmations. We proposed a method that contains two ideas, the joint model and model selection. In the joint model, we constructed the user response pattern FST and applied it to syllable recognition and word segmentation. In model selection, we choose the recognition results of the joint model and those of the cascade model. The proposed method improved the F-score of OOV word detection. The use of the user response pattern model in syllable recognition also improved its performance.

In addition to the normal response pattern FST, we also constructed the Complex FST and Filler FST, and compared their OOV word detection performance. The performance of the Complex FST model was higher than that of the normal FST model. The Filler FST model did not perform better than the normal pattern FST model.

One of the issues to be addressed in the future is the consideration of post-processing of the "any phrases." Currently, word segmentation is performed for all phrases identified as "any phrases". However, this method can lead to over-segmentation splitting OOV words that could have been correctly extracted. We believe that post-processing with proper conditions can reduce over-segmentation and achieve higher performance in OOV word detection.

Another issue is how to construct the FST. In the proposed method, we recognized utterances that did not match the FST by using the cascade model. However, it often failed to correctly detect OOV words. Although we constructed the normal FST that only accepts patterns that the user response data contains, the performance of OOV word detection may be improved if we construct an FST that accepts more user response patterns.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Numbers JP23H03457 and JP22H00536.

REFERENCES

[1] K. Komatani, K. Ono, R. Takeda, E. Nichols, and M. Nakano, "User impressions of system questions to acquire lexical knowledge during dialogues," *Dialogue and Discourse*, vol. 13, no. 1, pp. 96–122, 2022.

[2] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *Journal of Selected Topics in Signal Processing (JSTSP)*, vol. 11, no. 8, pp. 1240–1253, 2017.

[3] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *Journal of the IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 10, pp. 1533–1545, 2014.

[4] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, 2009, pp. 100–108.

[5] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[6] M. Mohri, "Finite-state transducers in language and speech processing," *Computational linguistics*, vol. 23, no. 2, pp. 269–311, 1997.

[7] N. Das, M. Sunkara, D. Bekal, D. H. Chau, S. Bodapati, and K. Kirchhoff, "Listen, know and spell: Knowledge-infused subword modeling for improving ASR performance of OOV named entities," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7887–7891.

[8] E. Egorova, H. K. Vydana, L. Burget, and J. Cernockỳ, "Out-of-vocabulary words detection with attention and CTC alignments in an end-to-end ASR system.," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 2901–2905.

[9] E. Egorova and L. Burget, "Out-of-vocabulary word recovery using FST-based subword unit clustering in a hybrid ASR system," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5919–5923.

[10] L. Qin and A. I. Rudnicky, "Finding recurrent out-of-vocabulary words.," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 2242–2246.

[11] S. Ueno, H. Inaguma, M. Mimura, and T. Kawahara, "Acoustic-to-word attention-based model complemented with character-level ctc-based model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5804–5808.

[12] G. Neubig, M. Mimura, S. Mori, and T. Kawahara, "Learning a language model from continuous speech," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010.

[13] T. Shinozaki, S. Watanabe, D. Mochihashi, and G. Neubig, "Semi-supervised learning of a pronunciation dictionary from disjoint phonemic transcripts and text.," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2546–2550.

[14] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *Journal of the IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 4, pp. 669–679, 2016.

[15] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 2207–2211.

[16] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library: (extended abstract of an invited talk)," in *Proceedings of Implementation and Application of Automata*, 2007, pp. 11–23.

[17] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of Language Resources and Evaluation Conference*, vol. 6, 2000, pp. 1–5.

Synthetic Data Augmentation for ASR with Domain Filtering

Tuan Vu Ho*, Shota Horiguchi*, Shinji Watanabe†, Paola Garcia‡, and Takashi Sumiyoshi*
* Hitachi, Ltd. Research & Development Group, Japan
† Carnegie Mellon University, USA
‡ Johns Hopkins University, USA

***Abstract*—Recent studies have shown that synthetic speech can effectively serve as training data for automatic speech recognition models. Text data for synthetic speech is mostly obtained from in-domain text or generated text using augmentation. However, obtaining large amounts of in-domain text data with diverse lexical contexts is difficult, especially in low-resource scenarios. This paper proposes using text from a large generic-domain source and applying a domain filtering method to choose the relevant text data. This method involves two filtering steps: 1) selecting text based on its semantic similarity to the available in-domain text and 2) diversifying the vocabulary of the selected text using a greedy-search algorithm. Experimental results show that our proposed method outperforms the conventional text augmentation approach, with the relative reduction of word-error-rate ranging from 6% to 25% on the LibriSpeech dataset and 15% on a low-resource Vietnamese dataset.**

Index Terms: speech recognition, domain filtering, semantic similarity maximization, vocabulary coverage maximization

I. Introduction

The conventional automatic speech recognition (ASR) approach involves using separately trained acoustic, language, and pronunciation models. In contrast, the more recent end-to-end (E2E) ASR system integrates all these components into a single neural network that directly predicts text sequences from input speech waveforms [1]. Despite its impressive performance, the end-to-end (E2E) ASR system has some limitations. It necessitates a significant amount of training data to achieve optimal results and is extremely sensitive to out-of-vocabulary (OOV) words, which can negatively impact its performance [2]. This presents a significant challenge when developing a new E2E-ASR model for many real-world applications with limited in-domain training data, including low-resource language ASR.

Several methods have been proposed to tackle this low-resource problem for E2E-ASR, including unsupervised-pretraining [3]–[5], transfer learning [6]–[8], and data augmentation [9]–[13]. Of all the approaches mentioned, this paper focuses on data augmentation as it is easy to implement and has demonstrated outstanding performance. Recently, a deep-neural-network-based text-to-speech (TTS) model has been shown to generate speech with comparable quality to human speech. Therefore, many recent studies have tried to use synthetic data generated from a TTS model [14]–[21] to train an E2E-ASR model.

Most prior studies rely on transcribing the in-domain text data or generating simulated text with predefined templates to synthesize speech (cf. Section II). However, text data obtained with these methods might not have enough lexical diversity as seen in the real text, yielding the OOV issue. One possible solution is to use text data from a generic-domain source, which can diversify lexical context with large-scale data. However, using a generic-domain source is challenging since it results in a domain mismatch and can negatively impact the ASR performance.

To avoid this problem, this paper focuses on a domain filtering technique that can remove irrelevant sentences from the target domain and increase the performance of ASR. To achieve the best efficiency with augmented synthetic speech, we raise the following two desired factors for the text selection:

- Semantic similarity: the augmentation text should be similar to the in-domain text data (i.e., refers to the same topics, same writing styles, etc.).
- Vocabulary coverage: The vocabulary of the augmentation data should be diverse to minimize the OOV.

To satisfy these factors, this study proposes to use a domain filtering method based on semantic similarity maximization using the Sentence-BERT language model [22]. The proposed method also involves diversifying the vocabulary during data selection by using a greedy search algorithm to maximize the vocabulary size iteratively. We referred to this step as vocabulary coverage maximization.

The rest of this paper is organized as follows. In Section II, we introduce the related works to this study. In Section III, we introduced the proposed text selection algorithm based on semantic similarity and vocabulary maximization. In Section IV, we demonstrate the effectiveness of the proposed method on a common ASR framework in low-resource settings. Finally, we conclude our paper in Section V.

II. Related Work

Recently, neural speech synthesis models have gained significant advancements and shown to generate synthetic speech that achieves a naturalness level comparable to human speech [23], [24]. The advantage of synthetic speech is that it can be easily generated and manipulated to create diverse training data. In a study by Mimura, et al. [25], synthetic

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

speech generated by the Tacotron-2 sequence-to-sequence TTS model is used to alleviate the OOV problem of acoustic-to-word (ASW) E2E-ASR. The results suggest that adding synthetic speech, even from a single female speaker, can significantly improve ASR performance compared to using only real speech. In [16]–[20], acoustic variations in synthetic speech, such as speaker characteristics and speaking styles are introduced to further improve the robustness of ASR models across different conditions. This paper builds upon these recent advancements but places more emphasis on linguistic variations.

Conventionally, the text data needed to generate synthetic speech is collected from the transcription of the in-domain data [15] or generated using a language model. For example, it was proposed to utilize a MaxEnt language model [26] to generate new utterances, which increase the lexical diversity of the text data [16]. Unlike our approach of selecting real text from a generic domain, this method requires retaining generated sequences with a length of fewer than 20 words and a perplexity score under 500 to avoid utilizing unrealistic utterances. SynthASR was also proposed to utilize synthetic speech to improve ASR performance in specific low-resource domains [21]. The authors randomly combine predefined text utterance templates with common medication names to augment the text data. Then, for each generated sentence, synthetic speech with different voice profiles is generated. In study [27], the authors generate text based on data augmentation for natural language understanding using semantic frames [28]. The text augmentation comprises four steps: delexicalization, diversity classification, text generation using a sequence-to-sequence model, and surface realization. The result from the text augmentation process is a paraphrased version of the source sentence. Unlike these techniques that rely on task-specific knowledge, such as medication names or semantic frame information, our proposed methods are entirely data-driven based on a Sentence-BERT language model.

III. Proposed Method

This paper proposes to maximize the effectiveness of text selection by choosing text that resembles the target data domain (for example, audiobooks or conversations). To minimize errors resulting from OOV words, it is important to ensure diversity in the vocabulary, while maintaining not to cause the domain mismatch between real and synthetic data. Therefore, we propose a domain filtering method to maximize both the semantic similarity and the vocabulary coverage of the synthetic speech. Figure 1 shows the overview of the proposed domain filtering method for synthetic speech generation. The proposed method is two-staged: semantic similarity maximization (SSM) and vocabulary coverage maximization (VCM), each of which is described in the following subsections.

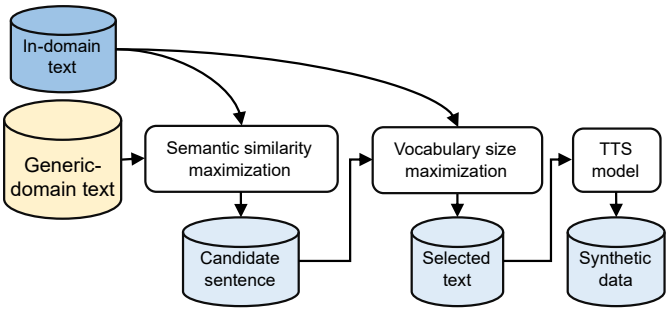


Fig. 1: Overview of the proposed text domain filtering method for synthetic speech generation.

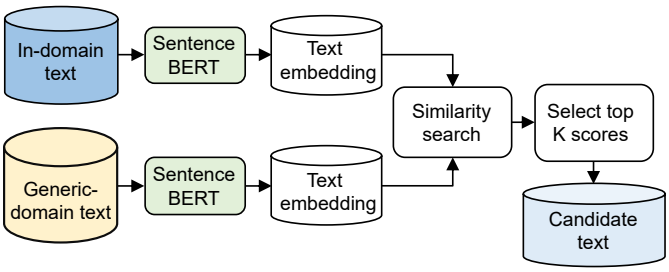


Fig. 2: Flow diagram of semantic similarity maximization.

A. Semantic similarity maximization

To exploit the training E2E-ASR advantages, the text data used for generating synthetic speech should ideally belong to the same domain as the in-domain text. Taking a virtual assistant ASR as an example, text from a casual conversation dialog is preferable to those from news articles.

The flow of the proposed text selection method is depicted in 2.

We first select relevant sentences that have high similarity to the in-domain text from the generic-domain text. To evaluate the semantic similarity, we use sentence embeddings extracted using a pre-trained Sentence-BERT model [22]. Sentence-BERT, which is a variant of BERT designed for processing sentences, has shown effectiveness in various natural language processing tasks, including semantic similarity, question-answering, and text classification. To assess the semantic similarity between the generic-domain and in-domain text, we initially extract embeddings for each sentence in both texts. Since the amount of generic-domain sentences is usually very large, we use the fast similarity search and clustering technique to reduce the search time [29]. For every sentence in the in-domain text, we search for 3 to 10 similar sentences in the generic-domain corpus based on the cosine similarity between embeddings. Finally, we eliminate duplicate sentences to obtain the final candidate sentence set.

B. Vocabulary coverage maximization

There is a trade-off between semantic similarity and vocabulary coverage. Relying only on semantic similarity to select text data often leads to limited vocabulary coverage

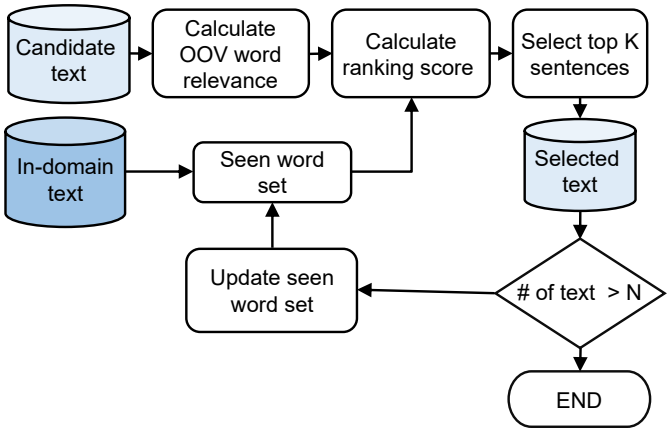


Fig. 3: Flow diagram of vocabulary coverage maximization.

because almost the same sentences may be selected. Selecting a larger set of sentences can increase vocabulary coverage, but an excessive amount of synthetic speech may degrade ASR performance due to the acoustic mismatch between real and synthetic speech data [16]. To overcome this issue, we propose a greedy search algorithm that maximizes the vocabulary coverage of the selected data while reducing the amount of text. Given a set of sentences that are selected to maximize semantic similarity using the method described in Section III-A, the second selection stage chooses a smaller set of sentences with maximum vocabulary coverage iteratively.

The process flow of the greedy-search algorithm is illustrated in Figure 3. To find which sentences should be included, we score each sentence in the set of candidate sentences based on how the sentence can increase the coverage. Let $\mathcal{D}$ be a set of words that have already appeared in either the in-domain text or in the text that has already been selected. The score of the i-th sentence is defined as

$$s_i = \frac{r_i}{|\mathcal{W}_i \cap \mathcal{D}|}, \tag{1}$$

where $\mathcal{W}_i$ is a set of words in the i-th sentence; thus, the denominator means the number of seen words in the sentence. r_i is the relevance of the unseen words to the target domain defined as

$$r_i = \sum_{u \in \mathcal{D} \setminus \mathcal{W}_i} r_i^{(u)}, \tag{2}$$

$$r_i^{(u)} = \max_{\substack{\mathcal{D}' \subset \mathcal{D} \\ |\mathcal{D}'|=K}} \frac{1}{K} \sum_{w \in \mathcal{D}'} \mathrm{sim}\left(w, u\right), \tag{3}$$

where $\mathrm{sim}\left(\cdot, \cdot\right)$ is a semantic similarity between two words. The similarity is calculated as a cosine similarity between two embeddings extracted using the GloVe model [30]. Eq. 3 intends to calculate the relevance score of the unseen word u on the basis of the top-K similar words in $\mathcal{D}$. The value of K is set to 20 in this paper.

With the calculated ranks, we select the sentence with the highest score and remove it from the set of candidate sentences. The word set $\mathcal{D}$ is also updated using the selected sentence. The steps above are repeatedly conducted until the predetermined number of sentences N is selected in total.

IV. Experiments

A. End-to-end Automatic Speech Recognition

For the E2E-ASR model, we used a Conformer-based connectionist temporal classification (CTC) / attention architecture [31]. The model consists of an encoder with 12 Conformer blocks and a decoder with six Transformer blocks. To train the model, we employed the ESPnet toolkit [32].

B. Experiment on LibriSpeech

We used the LibriSpeech [33] dataset as the in-domain data. LibriSpeech is derived from the larger LibriVox project, which contains free public-domain audiobooks. The dataset contains about 1,000 hours of speech from approximately 2,500 speakers.

Generic-domain text data was derived from the OSCAR corpus [34], which consists of text crawled from web pages. This text data covers different types of genres, reflecting the practical situation where only general domain text data is available. To reduce the computational amount, we used the subset of the corpus that includes 2 million sentences. As the OSCAR corpus text data is not in spoken form, we utilized the NeMo toolkit [35] for text normalization.

We adopted the Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) model [24] as our synthetic speech generation model. VITS is an end-to-end TTS model that generates raw speech waveform directly conditioned on the input text and speaker identity. We used the VITS model pretrained on speech data of 108 speakers of the VCTK dataset from the ESPNet toolkit [32]. For each sentence in the selected text, we randomly sampled a speaker embedding from the list of 108 speaker embeddings to synthesize speech.

According to [15], the amount of real speech should be at least 20% in the training data to avoid degrading ASR performance for the LibriSpeech dataset. Therefore, we empirically selected 30 hours of real speech from the train-clean-100 subset of LibriSpeech and kept the amount of synthetic speech consistent at 70 hours for all selection methods. The synthetic speech consists of 108 speakers from the VCTK dataset. Combining the real and synthetic data, we obtained a total of 100 hours of training data. We also included the results of using all 100 hours of clean speech to demonstrate the top-line performance of the E2E-ASR model. The experiment's specifications for synthetic speech are summarized below:

- In-domain: we directly use the transcription from the train-clean-360 subset of LibriSpeech. This provides the upper-bound performance of the ASR when in-domain text data is accessible.
- Synthetic text: we use the transcription of training data and generate synthetic text by text augmentation method

TABLE I: Results on LibriSpeech dataset. %OOV is the ratio between the number of OOV words and the total word count in each evaluation set.

	Speech data (hours)			%OOV				WER (%)			
Text for TTS	Real	Synthetic	# of vocab.	dev-clean	dev-other	test-clean	test-other	dev-clean	dev-other	test-clean	test-other
N/A	30	0	19,995	4.7	4.3	4.5	4.8	13.9	27.9	14.3	29.0
N/A	100	0	32,550	**2.5**	**2.5**	**2.4**	**2.8**	**6.4**	**17.5**	**6.7**	**17.6**
In-domain	30	70	38,071	2.1	2.1	1.9	2.3	8.3	20.3	8.7	20.5
Synthetic text	30	70	25,504	4.0	3.6	3.7	4.0	11.6	23.5	11.8	23.9
Random	30	70	54,249	2.7	2.7	2.5	2.8	9.4	22.6	9.6	22.6
SSM (ours)	30	70	47,846	2.3	2.4	2.2	2.6	**8.7**	**21.1**	**8.8**	**21.4**
SSM+VCM (ours)	30	70	**71,901**	**1.5**	**1.8**	**1.6**	**1.9**	8.8	22.0	9.0	22.2

using the nlpaug toolkit[1]. We used the contextual word embeddings substitution augmenter using BERT.

- Random: we randomly select sentences from the generic-domain text and synthesize 70 hours of speech.
- SSM: we selected text data from generic-domain sources using only semantic similarity maximization. After the selection, we obtained 22K sentences for generating synthetic speech.
- SSM+VCM: we selected approximately 136K sentences from a generic-domain source using SSM, and then further filtered to synthesize the necessary amount of data using the vocabulary maximization method.

Table I shows the experimental results. We observed that all the methods using synthetic data successfully improved the word-error-rate (WER). In terms of the text used for synthesizing speech, using in-domain text (i.e., transcriptions of the train-clean-360 subset) yields the lowest WER, indicating the importance of domain-specific augmented text.

Using synthetic text to keep the domain similarity between real and synthetic data outperforms the randomly selected text by approximately 7% to 9%. This result confirms the importance of selecting semantically similar text to that of the target domain. However, compared with 100 hours of clean speech, the WER is relatively 20% higher due to the mismatch of the acoustic characteristics of real and synthetic speech. Using VCM to further reduce the number of OOV words resulted in slightly higher WERs although the number of OOV words is significantly reduced. This might be because the ASR model is trained on sub-words instead of words, which enables the ASR model to recognize OOV words.

Our proposed method also exhibits better performance compared to the text augmentation method, with improvements ranging from 6% to 30% in terms of relative WER reduction.

C. Experiment with low-resource language

We conducted an experiment using the Vivos dataset [36], which includes 15 hours of Vietnamese speech recordings from 50 native speakers. The dataset is relatively small for training any E2E-ASR model, with a training set of 11,660 utterances and a testing set of 760 utterances. The recordings consist of excerpts mainly from news articles and conversational dialogues. We randomly selected 10% of the training utterances for validation. For generic-domain text sources, we used the Vietnamese text corpus collected from Wikipedia[2]. After performing text cleaning, we obtained approximately 1.2M unique sentences. We used the pretrained *paraphrase-multilingual-mpnet-base-v2*[3] Sentence-BERT model to evaluate the semantic similarity. To generate synthetic speech, we adopted the multi-speaker TTS model using BiLSTM acoustic model and HiFi-GAN vocoder [37] pretrained on Vivos dataset[4]. We conducted experiments with three different amounts of synthetic speech: 15 hours, 30 hours, and 60 hours. As the multi-speaker TTS model was trained from the Vivos dataset, the final synthetic dataset consists of 50 speakers from the Vivos dataset.

We conducted experiments following similar settings in Section IV-B. The specifications are summarized below:

- Synthetic text: synthetic text generated using Vietnamese text augmentation toolkit[5] with the substitution rate of 0.9.
- Random: synthetic speech is generated from random sentences collected from a generic-domain source.
- SSM: synthetic speech is generated from approximately 11K sentences using semantic similarity maximization from a generic-domain source.
- SSM+VCM: 55K sentences are collected from a generic-domain source using semantic similarity, then we further selected using vocabulary maximization to synthesize the required amount of speech.

We used the same ASR model described in Section IV-A and an LSTM-based language model (LM) for rescoring. The E2E-ASR model was trained for 120 epochs with the learning rate gradually increasing to 1×10^{-4} in the first 15K steps and then exponentially decaying. We used the raw speech waveform as an acoustic feature. We applied speed perturbation at ratios 0.9, 1.0, and 1.1 [12] and SpecAugment

[1]https://github.com/makcedward/nlpaug

[2]https://github.com/NTT123/viwik18

[3]https://www.sbert.net/

[4]https://github.com/NTT123/vietTTS

[5]https://github.com/longcld/vitextaug

TABLE II: Results on Vivos dataset. %OOV is the ratio between the number of OOV words and the total word count in each evaluation set. The OOV error rate is measured as the ratio between correctly recognized OOV words (of real speech data) and total OOV words in the evaluation set.

						WER ↓ (%)				OOV error rate ↓ (%)			
	Speech data (hours)			%OOV↓		w/o LM		w/ LM		w/o LM		w/ LM	
Text for TTS	Real	Synthetic	# of vocab.	dev	test	dev	test	dev	test	dev	test	dev	test
N/A	15	0	4,720	1.03	1.27	14.8	18.8	14.7	18.5	100.0	100.0	100.0	100.0
Synthetic text	15	15	5,139	0.85	1.00	10.7	13.0	10.9	12.8	100.0	98.8	100.0	98.8
Random	15	15	5,225	0.70	0.91	9.6	12.1	9.6	12.1	88.9	98.8	88.9	98.8
SSM	15	15	5,398	0.68	0.83	9.4	12.3	9.2	11.6	100.0	96.5	100.0	96.5
SSM+VCM	15	15	**6,030**	**0.40**	**0.47**	**9.2**	**11.4**	**9.0**	**10.9**	88.9	**90.6**	88.9	**89.4**
Random	15	30	5,515	0.59	0.74	7.9	10.0	7.7	9.6	88.9	96.5	88.9	97.6
SSM	15	30	5,728	0.52	0.61	7.6	10.0	7.4	9.3	88.9	95.3	88.9	94.1
SSM+VCM	15	30	**6,155**	**0.39**	**0.53**	**7.5**	**9.1**	**7.2**	**8.8**	**77.8**	**91.8**	**77.8**	**95.3**
Random	15	60	5,789	0.48	0.61	6.5	8.5	6.5	8.1	88.9	94.1	88.9	94.1
SSM	15	60	6,060	0.36	0.48	**6.1**	8.1	**5.9**	**7.5**	**55.6**	**78.8**	**55.6**	**78.8**
SSM+VCM	15	60	**6,711**	**0.25**	**0.36**	6.5	**7.9**	6.4	**7.5**	66.7	83.5	66.7	83.5

[13] for the data augmentation. Given that Vietnamese is a monosyllabic language comprising approximately 8,000 prevalent syllables, we train ASR model directly on word units instead of sub-word units, as detailed in Section IV-B. In our preliminary experiment, the model trained on word units demonstrates superior performance compared to the one trained on sub-word units. During inference, we averaged the weight of the top 10 models based on their validation accuracy. The search beam width was set to 20, while the CTC weight and LM weight were set to 0.3 and 0.1, respectively.

Table II illustrates the outcomes achieved by the proposed method in terms of ASR performance. When utilizing only 15 hours of synthetic speech, the proposed method exhibited a relative improvement of 15% and 9% compared to the synthetic text and random selection methods, respectively, on the test set with LM. Notably, a consistent reduction in WER was observed when augmenting the amount of synthetic speech to 30 and 60 hours. In all cases, the proposed data selection method outperformed the random selection approach. We also show the OOV error rate, which is measured as the ratio of correctly recognized OOV words (of the real speech data) to the total number of OOV words. The results in Table II indicate that while the baseline method without data augmentation failed to recognize any OOV word, our proposed method can reduce the OOV error rate to approximately 55.6% and 78.8%. Different from the experiment in Section IV-B, we have observed the efficacy of vocabulary expansion within the contexts of 15 hours and 30 hours of synthetic data configurations. This observation arises from the fact that OOV words cannot be accurately recognized, given that the ASR model is trained on word units rather than sub-word units. Consequently, the sole means of mitigating errors associated with OOV words is by increasing the vocabulary coverage. When the volume of synthetic speech was increased to 60 hours, the impact of vocabulary maximization on WER became less significant. Comparing the WER values between SSM and SSM+VCM, the difference was marginal (7.9% vs. 8.1% without LM, 7.5% vs. 7.5% with LM). This behavior was expected since, with the addition of more synthetic data, the most commonly occurring words were already included using the SSM alone. Consequently, the vocabulary maximization step did not further enhance the ASR performance.

V. Conclusions

In this paper, we proposed a data augmentation method using synthetic speech with text domain filtering to identify relevant text from generic-domain sources. The experiment results showed that combining semantic similarity search and vocabulary maximization gave better ASR performance than random selection and text augmentation methods. For future work, we aim to expand our data selection method into the acoustical domain to further improve the performance of E2E-ASR models in low-resource settings.

References

[1] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] C. Peyser, S. Mavandadi, T. N. Sainath, J. Apfel, R. Pang, and S. Kumar, "Improving tail performance of a deliberation E2E ASR model using a large text corpus," in *Proc. Interspeech*, 2020, pp. 4921–4925.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[6] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *Proc. ICASSP*, 2018, pp. 4909–4913.

[7] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Adversarial multilingual training for low-resource speech recognition," in *Proc. ICASSP*, 2018, pp. 4899–4903.

[8] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *Proc. SLT*, 2018, pp. 521–527.

[9] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *ASRU*, 2013, pp. 309–314.

[10] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *Proc. Interspeech 2020*, 2014, pp. 810–814.

[11] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[14] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, "Back-translation-style data augmentation for end-to-end asr," in *Proc. SLT*, 2018, pp. 426–433.

[15] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition," in *Proc. ICASSP*, 2019, pp. 6161–6165.

[16] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *Proc. ASRU*, 2019, pp. 996–1002.

[17] G. Wang, A. Rosenberg, Z. Chen, Y. Zhang, B. Ramabhadran, Y. Wu, and P. Moreno, "Improving speech recognition using consistent predictions on synthesized speech," in *Proc. ICASSP*. IEEE, 2020, pp. 7029–7033.

[18] C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *Proc. ICASSP*, 2020, pp. 7719–7723.

[19] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *Proc. ICASSP*, 2020, pp. 7069–7073.

[20] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end ASR systems," in *Proc. ICASSP*, 2021, pp. 5674–5678.

[21] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: Unlocking synthetic data for speech recognition," in *Proc. Interspeech*, 2021, pp. 896–900.

[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. EMNLP*, 2019, pp. 3982—-3992.

[23] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[24] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[25] M. Mimura, S. Ueno, H. Inaguma, S. Sakai, and T. Kawahara, "Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition," in *Proc. SLT*, 2018, pp. 477–484.

[26] F. Biadsy, M. Ghodsi, and D. Caseiro, "Effectively building tera scale MaxEnt language models incorporating non-linguistic signals," in *Proc. Interspeech*, 2017, pp. 2710–2714.

[27] R. Zevallos, N. Bel, G. Cámbara, M. Farrús, and J. Luque, "Data augmentation for low-resource quechua ASR improvement," in *Proc. Interspeech*, 2022, pp. 3518–3522.

[28] Y. Hou, Y. Liu, W. Che, and T. Liu, "Sequence-to-sequence data augmentation for dialogue language understanding," in *Proc. ICCL*, 2018, pp. 1234–1245.

[29] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[30] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[31] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[32] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[34] P. J. Ortiz Su'arez, B. Sagot, and L. Romary, "Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures," in *Proc. Workshop on Challenges in the Management of Large Corpora*, 2019, pp. 9–16.

[35] Y. Zhang, E. Bakhturina, K. Gorman, and B. Ginsburg, "NeMo inverse text normalization: From development to production," in *Proc. Interspeech*, 2021, pp. 4468–4472.

[36] H.-T. Luong and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese speech recognition system," in *Proc. Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies*, 2016, pp. 51–55.

[37] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

Multi-Self-Supervised Learning Model-Based Throat Microphone Speech Recognition

Kohta Masuda*, Jun Ogata†, Masafumi Nishida* and Masafumi Nishimura*‡
*Shizuoka University, Japan
†National Institute of Advanced Industrial Science and Technology, Japan
‡Aichi Sangyo University, Japan

Abstract — **Throat microphones (TMs) can record sounds and mitigate the effects of external noise. Ongoing works seek to apply TMs to speech recognition in high-noise environments. However, building a TM-based speech recognition model is difficult in the absence of a large-scale TM speech database. We previously proposed a speech recognition method using latent representations obtained from pretrained self-supervised learning (SSL) models for each speech recognition task and subsequent feature mapping. However, SSL models are built for speech recognition tasks and are not appropriate for feature mapping tasks. Therefore, in this work, fine-tuning is applied to the feature extraction part of the SSL model so that the SSL model can derive latent representations useful for the feature mapping task. In the training process of the TM speech recognition model, we first perform fine-tuning using a large amount of pseudo-TM speech data generated by the feature mapping model and then perform a second fine-tuning using a small amount of *real*-TM speech data. In the second fine-tuning, the parameters of the transformer part of the SSL model are fixed in order to retain the linguistic information. The character error rate was reduced by 21.38% (from 8.14% to 6.35%) compared to the conventional method.**

Introduction

Throat microphones (TMs) detect vibrations directly from the skin covering the throat. They record sounds with less interference from external noise compared with ordinary microphones. Several studies on speech recognition have been conducted using TMs in high-noise environments. However, TMs record speech with less high-frequency bands than speech captured by close-talk microphones (CM). Thus, detection accuracy is reduced when TMs are utilized in common speech recognition systems. Although a speech recognition system that directly recognizes TM speech is currently under study, the development of this system is difficult without a large-scale database of speech recorded using TM. Several approaches have been proposed to solve this problem, including knowledge distillation using many pseudo-TM speech features. These features are created using a Deep Neural Network to map CM speech features to TM speech features [1] and eliminate mismatches through domain adaptation [2]. However, the training process in these methods is complex and costly. Therefore, we focused on pretrained self-supervised learning (SSL) models. We constructed a speech recognition model for TM speech by employing a fine-tuning method using pseudo-TM speech generated by feature mapping [3]. This method can be easily learned and yields high recognition accuracy from a small amount of TM speech data. We also proposed using pretrained SSL models in feature mapping tasks to mitigate the shortage of training data for feature mapping models [4]. This method improved the feature mapping model and the recognition accuracy of the TM speech recognition model. However, the SSL model opposes the framework for speech recognition tasks; fine-tuning only the transformer part of the SSL model does not yield latent representations suitable for feature mapping tasks. Hung et al. proposed a method of adapting the SSL model to speech enhancement tasks, wherein the feature extractor part of the SSL model is also simultaneously fine-tuned [5]. Thus, we fine-tuned the feature extractor part of the SSL model to derive latent representations benefiting feature mapping. We trained the speech recognition model by first fine-tuning the model based on a large amount of generated pseudo-TM speech and, subsequently, fine-tuning the model using a small amount of TM speech. The second fine-tuning improved recognition accuracy by apprehending the acoustic features of TM speech. However, the second fine-tuning decreased the linguistic information attained from the large-scale speech data. We prevented the reduction of linguistic information by fixing the weight of the transformer part of the SSL model during the second fine-tuning.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

PROPOSED METHOD

A. *Feature Mapping Based on SSL Model*

We use an SSL model for the feature mapping method. Huang et al. [6] demonstrated that the use of an SSL model in speech enhancement tasks significantly improves recognition accuracy compared to conventional supervised learning approaches. In addition, Hung et al. [5] proposed adapting the SSL model for speech enhancement tasks by simultaneously training the feature extractor part of the SSL model. Inspired by this approach, we apply the SSL model to a feature mapping task that converts CM speech into TM acoustic features. Figure 1 shows the method of training in the feature mapping model. Parallel speech data was recorded simultaneously by a CM and a TM. The feature mapping model is time-frequency (T-F) mask-based and comprises a pretrained SSL model and long short-term memory networks (LSTM). The SSL model inputs raw CM speech and outputs a latent representation of speech. The LSTM estimates a short-time Fourier transform (STFT) mask of the pseudo-TM speech-based on the latent representation from the SSL model. Previously [4], we employed a training method that minimized losses between the STFT mask and the previously defined ideal phase-sensitive mask [7]. In this work, we estimate the STFT features of pseudo-TM speech by multiplying the estimated pseudo-TM speech STFT mask by the CM speech STFT features. Then, the parameters of the SSL model and the LSTM are updated by minimizing the mean square error (MSE) between the estimated pseudo-TM speech STFT features and the TM speech STFT features. More end-to-end learning can be performed using his method compared with conventional methods. This method enables us to perform end-to-end training of the feature mapping model.

The LSTM for the SSL model consists of three layers, 1,024 memory cells, and wav2vec 2.0 [8]. The pretrained wav2vec 2.0 model includes the feature extractor, the quantization, and the transformer. The weighted sum of the outputs from all layers of the transformer is utilized as the latent representation of speech rather than merely the output of the last layer of the transformer. The output of the last transformer layer is suitable for speech recognition tasks, while the outputs of layers close to the feature extractor part, such as the first layer, are suitable for tasks like speech enhancement and speaker separation. The wav2vec 2.0 model is pretrained with 56,000 h of speech data in 53 languages published by Conneau [9]. Pretraining models in various languages deliver better universal representations across multiple languages.

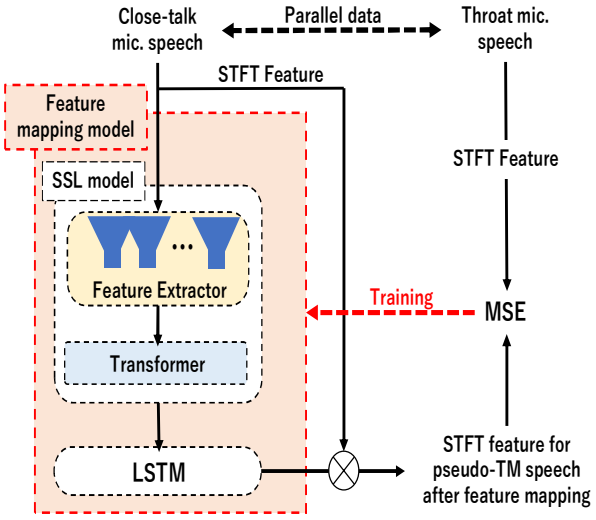


Fig. 1 The training procedure of the feature mapping model

B. *Data Augmentation and Fine-tuning of TM Speech Recognition Model*

At this stage, the acoustic features of TM speech are learned by fine-tuning the pretrained SSL model for speech recognition. The data is extended by generating pseudo-TM speech using the feature mapping model to compensate for the lack of training data for TM speech. Figure 2 shows the training process of the TM speech recognition model. Large-scale CM speech is input into the feature mapping model to obtain the STFT features of pseudo-TM speech, which are subsequently converted to a time domain using iSTFT to generate pseudo-TM speech. The pseudo-TM speech is used to fine-tune the pretrained SSL model. We use wav2vec 2.0 pretrained from 53 languages and 56,000 h of speech data as the SSL model for the speech recognition task and the feature mapping task. The fine-tuning method entails adding a linear layer to the final output layer of the SSL model and minimizing the connectionist temporal classification loss through correct labels. Training the acoustic features of TM speech by fine-tuning with only pseudo-TM speech is insufficient. Therefore, the speech recognition model is trained with two fine-tuning stages. The first fine-tuning for the SSL model is performed using pseudo-TM speech generated by feature mapping. The second fine-tuning is conducted using a small amount of TM speech. One problem with this method, however, is that the final fine-tuning is performed using a small amount of speech data, and therefore the linguistic information necessary for speech recognition cannot be extracted well. In other words, the first fine-tuning can be expected to acquire sufficient

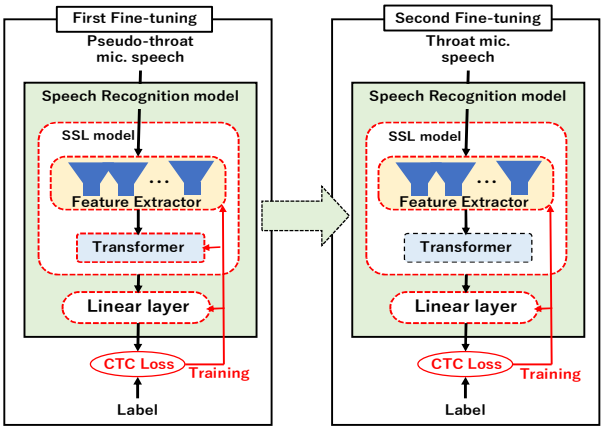


Fig. 2 Training process of throat microphone speech recognition model.

linguistic information from a large amount of data, but this information is replaced by poor ones in the second (final) fine-tuning. For this problem, we fix (freeze) the parameters of the transformer part trained in the first fine-tuning, and then perform the second fine-tuning with the fixed parameters. This method prevents the updating of linguistic information during the second fine-tuning. We also construct a speech recognition model using pseudo-TM speech generated via a feature mapping model without tuning the parameters of the feature extractor of the SSL model and compare this model with conventional methods. Additionally, we construct a speech recognition model without fixing the parameters of the transformer part of the SSL model to determine the influence of the extracted linguistic information. To determine if the linguistic problem could be prevented when different SSL models were used, the experiments were also carried out using HuBERT [10] and WavLM [11] as different SSL models.

EXPERIMENT

A. Experimental Conditions

Phonemically balanced, read-aloud parallel speech data (about 10 h) encompassing both CM and TM speech were used. The data were collected from ten male speakers and used as training data for the feature mapping model and fine-tuning the speech recognition model. We also utilized recordings of a newspaper-read speech (around 50 min) obtained from ten male speakers using TMs to evaluate the speech recognition model. All data were recorded in a quiet environment, and the speakers for the evaluation data were different from the speakers for the training data. Speech from the Corpus of Spontaneous Japanese (230 h), an existing large-scale Japanese dataset recorded through a CM, was also used. Pseudo-TM speech was created by applying feature mapping to the Corpus of spontaneous Japanese.

B. Experimental Results

The recognition experiment was accomplished by changing the dataset used to fine-tune the speech recognition model. The character error rate (CER) was used as an evaluation index. Table 1 displays the CERs for each model.

Table. 1 The character error rate of each throat microphones speech recognition model. Evaluation data are newspaper-read TM speech of 10 male speakers.

	SSL Model	Fine-tuning data	Feature mapping model	Speech Recognition model	CER[%]
1	Wav2vec	Thorat	–	Fine-tuning of all parts	25.86
2		Close talk			24.02
3		Close talk^{1st}→TM^{2nd}			14.38
4		Pseudo throat^{1st}→TM^{2nd}	Fine-tuning of only the Transformer part	Fine-tuning of all parts	8.14
5			Fine-tuning of all parts	Fine-tuning of all parts	7.5
6				Fine-tuning of only the Feature extractor part	6.35
7	HuBERT	Pseudo throat^{1st}→TM^{2nd}	Fine-tuning of all parts	Fine-tuning of all parts	6.94
8				Fine-tuning of only the Feature extractor part	4.8
9	WavLM	Pseudo throat^{1st}→TM^{2nd}	Fine-tuning of all parts	Fine-tuning of all parts	7.23
10				Fine-tuning of only the Feature extractor part	5.32

Models one–three denote the recognition results when the speech recognition model was fine-tuned without using pseudo-TM speech generated by feature mapping. We used existing speech data such as large-scale CM speech and a small amount of TM speech. In model one, fine-tuning was performed using only 10 h of TM speech. In model two, fine-tuning was performed using 230 h of CM speech. In model three, the first fine-tuning was accomplished using 230 h of CM speech, and the second fine-tuning was conducted using 10 h of TM speech. The recognition results for models one and two were 24%–25%, indicating that the small amount of TM speech was inadequate for training the speech recognition mode. The acoustic features of TM speech could not be recognized by training with only CM speech. The recognition results obtained from model three were significantly better than the results from models one and two. Thus, the efficacy of the advanced acquisition of general acoustic features and linguistic information by fine-tuning the model using large-scale CM speech before learning the acoustic features of TM speech by fine-tuning using TM speech was demonstrated. However, the

recognition accuracy was inadequate for fine-tuning the model with these datasets.

Models four–six denote the recognition results when the speech recognition model was fine-tuned using pseudo-TM speech generated by the feature mapping model. Model four represents a conventional method that only updates the parameters of the transformer part of the SSL model when fine-tuning the feature mapping model, to minimize the loss from the STFT masks. The recognition results from model four surpassed the recognition results from models one–three. Therefore, the use of pseudo-TM speech for fine-tuning eliminates the lack of training data for the speech recognition model, indicating that more acoustic features of TM speech were learned.

In models five–six, the parameters of all modules of the SSL model were updated during fine-tuning of the feature mapping to minimize the loss from the STFT features. Model five was constructed by updating the parameters of all modules of the SSL model during the second fine-tuning of the speech recognition model, as in model four. Model six was constructed by updating only the parameters of the feature extractor part of the SSL model during the second fine-tuning of the speech recognition model. The recognition results for models five and six were superior to the recognition results for model four, indicating that the SSL model can derive useful latent representations for feature mapping tasks by updating the parameters of the feature extractor part of the SSL model when fine-tuning the feature mapping. It was also found that minimizing STFT features instead of STFT masks when computing losses increases end-to-end learning. The results from model six were better than the results from model five. Hence, the linguistic information learned in the first fine-tuning was not reduced by fixing the parameters of the transformer part of the SSL model in the second fine-tuning of the speech recognition model and updating only the parameters of the feature extractor part; the recognition accuracy of the speech recognition model was further improved by this method.

In models seven–eight, HuBERT was used for the SSL model for speech recognition. Model seven was constructed by updating the parameters of all modules of the SSL model during the second fine-tuning of the speech recognition model as in model five. Model eight was constructed by updating only the parameters of the feature extractor part of the SSL model during the second fine-tuning of the speech recognition model as in model six. Models nine–ten used WavLM as the SSL model for the speech recognition model. Model nine was constructed by updating the parameters of all modules of the SSL model during the second fine-tuning of the speech recognition model as in model seven. Model ten was constructed by updating only the parameters of the feature extractor part of the SSL model during the second fine-tuning of the speech recognition model as in model eight. The recognition accuracy of fixing the transformer parameters in the second fine-tuning was good regardless of which SSL model was used, indicating that fixing the transformer part in fine-tuning is effective in preventing the reduction in linguistic information. Additionally, these results showed that HuBERT provides the best recognition accuracy as an SSL model.

CONCLUSIONS

In this study, in order to further improve the recognition performance of a speech recognition model suitable for TM speech, an approach was taken for both the feature mapping task and the speech recognition task. In the feature mapping task, by updating the parameters of the feature extractor part in the fine-tuning of the SSL model, the SSL model was able to derive a latent representation suitable for the feature mapping task and to learn more acoustic features of TM speech from pseudo-TM speech. It was shown to learn more features. In the speech recognition task, by fixing the parameters of the transformer part in the second fine-tuning of the SSL model, linguistic information obtained by fine-tuning using the large-scale pseudo-TM speech is ficed, and only the acoustic features of the TM speech are learned, which is more effective. We confirmed that by fixfing the parameters of the transformer part in fine-tuning, the recognition rete of the TM speech recognition model is further improved by learning only the acoutic features of the TM speech.

ACKNOWLEDGMENT

These results were obtained with a consignment project (JPNP 20006) by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

[1] T. Suzuki, J. Ogata, T. Tsunakawa, M. Nishida, M. Nishimura, “Knowledge Distillation for Throat Microphone Speech Recognition,” INTERSPEECH, pp. 461-465, 2019.

[2] M. A. T. Turan, E. Erzin, “Improving Phoneme Recognition of Throat Microphone Speech Recordings Using Transfer Learning,” Speech Communication, pp. 25-32, 2021.

[3] K. Masuda, J. Ogata, M. Nishida, M. Nishimura, “Throat Microphone Speech Recognition Using Wav2Vec 2.0 and Feature Mapping,” 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), pp. 395-397, 2022.

[4] K. Masuda, J. Ogata, M. Nishida, M. Nishimura, “Improved Throat Microphones Speech Recognition Using a Self-Supervised Learning Model,” 2023 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), 2023.

[5] K. H. Hung, S. W. Fu, H. H. Tseng, H. T. Chiang, Y. Tsao, C. W. Lin, “Boosting Self-Supervised Embeddings for Speech Enhancement,” arXiv preprint arXiv:2204.0339, 2022.

[6] Z. Huang, S. Watanabe, S. Yang, P. Garcia, S. Khudanpur, “Investigating Self-Supervised Learning for Speech Enhancement and Separation,” ICASP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6837-6841, 2022.

[7] M. Kolbaek, D. Yu, Z. H. Tan, J. Jensen, “Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1901-1913, 2017.

[8] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, “Wav2vec2.0: A Framework for Self-Supervised Learning of Speech Representations,” Advances in neural information processing systems, pp.12449-12460, 2020.

[9] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, “Unsupervised Cross-lingual Representation Learning for Speech Recognition,” arXiv preprint arXiv:2006.13979, 2020.

[10] W. N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 3451-3460, 2021.

[11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” IEEE Journal of Selected Topics in Signal Processing, pp. 1505-1518, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

ASR Model Adaptation for Rare Words Using Synthetic Data Generated by Multiple Text-To-Speech Systems

Kwok Chin Yuen*, Li Haoyang* and Chng Eng Siong
Nangyang Technological University, Singapore
E-mail: kwok002@e.ntu.edu.sg, li0078ng@e.ntu.edu.sg, aseschng@ntu.edu.sg

***Abstract*—Automatic speech recognition (ASR) for rare words is difficult as there are little relevant text-audio data pairs to train an ASR model. To obtain more text-audio pairs, text-only data are fed to Text-To-Speech (TTS) systems to generate synthetic audio. Previous works use a single TTS system conditioned on multiple speakers to produce different speaker voices to improve the output data's speaker diversity, and they show that training an ASR model on the more diverse data can avoid overfitting and improve the model's robustness. As an alternative way to improve the diversity, we study the speaker embedding distribution of audios synthesized by different TTS systems and found that the audios synthesized by different TTS systems have different speaker distributions even when they are conditioned on the same speaker. Inspired by this, this paper proposes to condition multiple TTS systems repeatedly on a single speaker to synthesize more diverse speaker data, so ASR models can be trained more robustly. When we apply our method to a rare word dataset partitioned from National Speech Corpus SG, which contains mostly road names and addresses in its text transcripts, experiments show that a pretrained ASR model adapted to our multi-TTS-same-SPK data gives relatively 9.8% lower word error rate (WER) compared to the ASR models adapted to same-TTS-multi-SPK data of the same data size, and our overall adaptation improves the model's WER from 57.6% to 16.5% without using any real audio as training data.**

I. Introduction

Rare words pose an ongoing challenge in the development of high-quality automatic speech recognition (ASR) systems [1]. These rare words, often associated with named entities like names and locations, play a vital role in preserving the intended meaning within the decoded transcript. Due to their limited occurrence in the audio-text pairs forming the training set of an ASR system, correctly predicting these "tail" words becomes inherently difficult.

One way to increase the number of text-audio training data pairs containing these rare words is to generate synthetic data using a Text-To-Speech (TTS) system. This method requires only text data as inputs to a TTS system, and does not require any change in the parameter or architecture of ASR models. Although there are many other methods like language model integration [2] and speech augmentation methods [3] [4] that can achieve a similar goal, [5] shows that the performance improvements are mostly independent.

As improving the speaker diversity of the synthetic training data can improve an ASR model's robustness [6] [7], previous works use a single TTS system conditioned on multiple speakers to produce different speaker voices. Alternatively, we found that using multiple TTS systems conditioned on a single speaker can also produce slightly different speaker voices. This is in line with several studies [8] [9], which shows that there can be noticeable differences in the real and TTS synthesized speaker voices. An analysis of the speaker embedding distribution of audios synthesized by conditioning different TTS systems on the same speaker is shown in Section V-C.

Inspired by this, we propose to synthesize data by repeatedly conditioning multiple TTS systems on the same speaker to synthesize more diverse speaker voices. Compared to methods that use a single TTS system to improve the speaker diversity by either conditioning the TTS system on multiple speakers or by tuning its parameters to control the speaker characteristics [10] of the synthesized audio, our method is less restricted by the trade-off [11] between the data's diversity and naturalness, i.e. improving the diversity of the synthesized audios costs less degradation in their quality. Also, our method can be applied to single-speaker TTS as the diversity can be compensated by the use of multiple TTS.

Our contributions are fourfold: 1) We explore the use of multiple TTS systems to synthesize training data to perform ASR adaptation. 2) We show that to generate diverse synthetic data, using multiple TTS systems conditioned on the same speaker is more effective than using a single TTS system conditioned on multiple speakers. 3) We show that freezing part of an ASR model when adapting to synthetic data gives better result than adapting the whole ASR model. 4) We provide baseline results on a high quality, publicly available rare word dataset, which consists of 13K utterances containing road names and addresses from National Corpus SG[1] [12]. The dataset is also released online[2].

*These authors contributed equally.

[1]For more information on National Corpus SG, visit: https://www.imda.gov.sg/how-we-can-help/national-speech-corpus

[2]Rare word dataset: https://github.com/cyaaronk/IMDA2RareWordDataset

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Related Works

A. Rare Word Recognition

Several directions have been explored for improving rare word recognition. The Contextual Listen, Attend, and Spell (CLAS) model [13] represents an end-to-end ASR system that builds upon the foundational Listen, Attend and Spell (LAS) model initially proposed by Chan et al. [14]. A notable distinguishing characteristic of the CLAS model is its incorporation of contextual information in an entirely end-to-end fashion, setting it apart from the LAS model. In [15], phonetic fuzzing of proper nouns are used to find hard negative examples for training CLAS. These methods requires real text-audio pairs during training.

Several techniques requires text-only data to bias ASR model outputs to rare words. They include shallow fusion [16]–[18], cold fusion [19], and deep fusion [20] of an external language model. Other techniques focus on modifying the training data by either data augmentation [3] [4] or TTS data generation [21] [22] [23] [24]. Our work aims to synthesize training data of diverse speaker voices for rare word ASR.

B. Diverse TTS Synthetic Data Generation

Numerous efforts are made to improve a TTS systems ability to synthesize data with diverse speaker voices. [6] uses a Tacotron 2 model through conditioning the network on multiple speaker embeddings. [5] investigates the use of different types of speaker embeddings for conditioning the TTS. [7] increases the speaker diversity through conditioning a Tacotron 2 model on randomly chosen virtual speakers sampled from a latent space.

All the above mentioned works, however, focuses on using just a single TTS system to improve the speaker diversity of the synthetic data. In our work, we show that an ASR model's robustness can be further improved by increasing the speaker diversity of the synthetic training data using multiple TTS systems. To the best of our knowledge, we are the first to leverage multiple TTS systems to improve the speaker diversity of the synthetic data for ASR adaptation.

III. Methodology

We now introduce our method to improve the speaker diversity of TTS synthetic speech. Specifically given a set of N transcripts $\{u_1, \ldots, u_N\}$, we first feed each transcript u_n to a TTS system M_1 and condition M_1 on a randomly sampled speaker s_i, where $i \in \{1, \ldots, K\}$ and K is the number of speakers. We then synthesize another audio for u_n by conditioning a different TTS system M_2 on the same speaker s_i. We call our approach multi-TTS-same-SPK as we use multiple TTS conditioned on the same speaker to generate audio for each transcript. In contrary, we call the approach used by previous works to synthesize audio as same-TTS-multi-SPK, as they use the same TTS conditioned on multiple different speakers for each transcript.

A. Multiple TTS Systems

We use two state-of-the-art TTS systems of different architecture to generate more diverse data. The first TTS system uses TransformerTTS [25] as the acoustic model and HiFi-GAN [26] as the vocoder. The second TTS system is VITS [27].

1) TransformerTTS and HiFi-GAN: TransformerTTS is an autoregressive TTS that utilizes transformers with multi-head attentions in an encoder-attention-decoder based architecture. It is the acoustic model in the TTS system which generates mel-spectrogram frames sequentially from textual input. The generated mel-spectrogram is then passed into HiFi-GAN, a GAN based vocoder to generate high-fidelity audio waveform by upsampling the input through a series of transpose convolutions with a Multi-Receptive Field Fusion module following each transpose convolution to process patterns of different lengths in parallel.

2) VITS: VITS is a fully end-to-end TTS that utilizes variational inference augmented with normalizing flows and an adversarial training phase to capture high-fidelity speech from textual input. VITS composes of a prior encoder for both text encoding and normalizing flow, a posterior encoder for spectrogram encoding, a duration predictor, a HiFi-GAN generator for decoding and a multi-period discriminator from HiFi-GAN.

B. Speaker Conditioning

To enable multi-speakers capability for the TTS systems, speaker embedding vectors that capture speaker style information are added into TransformerTTS and VITS so that the TTS systems can use the information to generate audios that sound similar to the target speaker.

Our TransformerTTS is made multi-speakers by first extracting Xvector [28] from audio recordings and normalize it. Then, a linear transformation is applied and the result is added to the output of the TransformerTTS encoder. The combination of Xvector with TTS has proven to be effective in creating multi-speakers TTS in [9].

To enable multi-speakers capability for VITS, an external speaker encoder based on GE2E [29] is used to extract speaker embedding vectors from audio recordings. The speaker embedding vectors are then added to the residual blocks within the posterior encoder, the normalizing flow within the prior encoder, the input latent space of the duration predictor and the decoder. Similar to Xvector, the GE2E speaker encoder has shown promising results in creating multi-speakers TTS in [30].

IV. Experimental Setup

A. TTS synthetic data generation

All the TTS models we use are trained on the same VCTK corpus [31], a popular benchmark dataset for TTS containing about 44 hours of speech data from 109 native English speakers with various accents. The ESPnet2-TTS toolkit [32] is used for all TransformerTTS and HiFi-GAN experiments, while the

TABLE I: Percentage of rare words (road names and addresses) in the train and test set's text transcript

Dataset	Total # of words	Total # of rare words	Total # of rare words / total # of words (%)	Total # of overlapped rare words in train and test set
Real train set	115119	53549	46.52	5466
Test set	11015	5825	52.88	-

TABLE II: Example sentences in the train and test set

Real train set	Test set
can i take the m r t to tiong poh avenue	chew kheng chuan hussein aljunied and yong melvin yik chye
weld road singapore centre for chinese language and telok blangah hill	bukit timah saddle club the concourse and bury road
farrer park road jalan selanting and sofitel singapore resort and spa	martin road bukit timah saddle club and sundridge park road
jalan mariam jalan shaer and holland road west	i am meeting audrey at spring street first
berrima road auckland road west and seraya road	jalan pinang seletar aerospace and south buona vista road
universal studios singapore south east asia hotel and bedford road	jalan binjai seletar airport and dalkeith road
faith bible dunman road and institute of high performance computing	liane road empress road and marsiling crescent
i am waiting for shasta to come back from kee sun avenue	upper bedok road rosas hall and jalan sukachita
gek poh shopping centre fernhill road and peiying primary school	kampong bahru prome road and north buona vista drive
regent street aljunied avenue two and sky suites	tai seng christian church seletar close and harrison road

TABLE III: Dataset overview

Dataset	Dataset Type	Total # of words	Total # of sentences	Total # of speakers	Total audio duration (h)
Real train set	real	115119	13000	186	17.4
VITS-SPKSET1	fake	115119	13000	108	11.6
VITS-SPKSET2	fake	115119	13000	108	11.6
TRANS-SPKSET1	fake	115119	13000	108	10.1
TRANS-SPKSET2	fake	115119	13000	108	10.2
Test set	real	11015	1176	76	1.63

Coqui-ai TTS toolkit [33] is used for all VITS experiments. All models uses 80-dimensional mel-spectrograms. Also, all input text are converted to phonemes before feeding into the models. As the TransformerTTS and VITS TTS systems originally generates waveforms at a sampling rate of 22050 Hz and 24000 Hz respectively, and the ASR model is trained to receive only 16000 Hz audio as input, down sampling of the TTS synthesized audio is required before they are fed into the ASR model. The process may cause the loss of high frequency acoustic information in the audio. However, we hypothesize that such information loss has little effect for ASR adaptation as a sampling rate of 16000 Hz is sufficient to cover most of the human voice frequency range [34] and keep most of the needed information for ASR. Future works are needed to study the effect of down sampling on ASR adaptation performance.

The TransformerTTS model and HiFi-GAN model we use is pretrained and released by ESPnet2-TTS [32]. The FFT, window, and hop size for both models were set to 2048, 1200, and 300, respectively. The TransformerTTS model uses 6 layers of transformer blocks in both its encoder and decoder with 8 heads for multi-head attention. The decoder prenet has 2 linear layers with 256 hidden units, and the postnet has 5 CNN layers with channel and filter size 256 and 5 respectively. The model was trained using the Adam optimizer [35] with the addition of a guided attention loss to help learning diagonal attentions [36]. The HiFi-GAN generator uses hidden dimension size of 512, kernel sizes of [10, 10, 8, 6] for the transposed convolutions, kernel sizes of [3, 7, 11] and dilation rates of [[1, 1], [3, 1], [5, 1]] × 3] for the MRF modules. The HiFi-GAN model was trained using the Adam optimizer with a learning rate of 0.0002.

The VITS model we use is pretrained and released by Coqui-ai TTS. The model's configurations follow the original work [27], with the FFT, window, and hop size set to 2048, 1024, and 256, respectively. The model was trained with the AdamW optimizer [37] with $\beta 1$ = 0.8, $\beta 2$ = 0.99, weight decay γ= 0.01 and learning rate of 0.001. A windowed generator training is used with a window size of 32 to improve efficiency and lowering memory consumption during training.

We subset 10 hours of real audio-text training data pairs containing road names and addresses from National Corpus SG [12] as our real train set. They are recordings of people asking for directions and consists of 13K unique utterances. We also partition 1.63h of parallel data as the test set. Table I shows the details of the rare word (road names and addresses) coverage for the train and test set. Table II shows some example sentences in the train and test set.

Given the 13K text-only transcripts we obtained from the

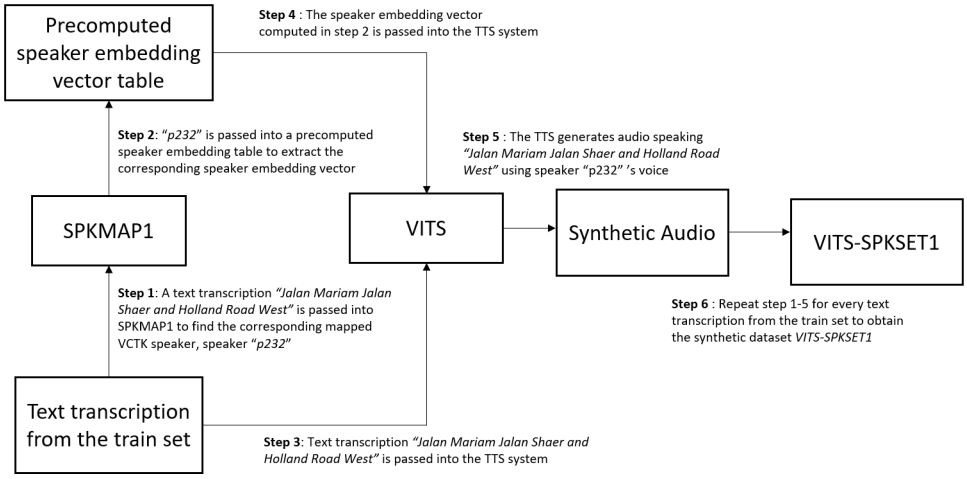


Fig. 1: Audio generation pipeline to obtain synthetic audio dataset *VITS-SPKSET1*. A similar pipline is also used to obtain synthetic audio dataset *VITS-SPKSET2*, *TRANS-SPKSET1* and *TRANS-SPKSET2*.

real train set, we first create a mapping named SPKMAP1 to map the text transcription of every utterance in the train set to a randomly sampled speaker in the VCTK corpus. We then obtain another mapping SPKMAP2 similarly, such that the mapped speaker for every text in SPKMAP2 is different from the mapped speaker in SPKMAP1. After creating SPKMAP1 and SPKMAP2, we generate 4 sets of synthetic audio data sets by varying the 2 speaker mappings SPKMAP1 and SPKMAP2 and varying the 2 TTS systems VITS and TransformerTTS systems used and name them VITS-SPKSET1, VITS-SPKSET2, TRANS-SPKSET1 and TRANS-SPKSET2.

Figure 1 shows the step-by-step procedure to obtain VITS-SPKSET1. In step 1, we first use SPKMAP1 to obtain the VCTK speaker id that is mapped to a text transcription in the train set. In step 2, we extract the speaker embedding vector that correspond to the VCTK speaker id obtained in step 1. In step 3, 4 and 5, we feed the text transcription (used in step 1) and the speaker embedding vector (extracted in step 2) into our VITS model to generate an synthetic audio file. We repeat step 1-5 for all text transcriptions in the train set, and the synthetic audio files obtained formed VITS-SPKSET1. To obtain VITS-SPKSET2, we simply replace SPKMAP1 with SPKMAP2, and repeat the pipeline just described. To obtain TRANS-SPKSET1 and TRANS-SPKSET2, we repeat the same steps using our TransformerTTS+HiFi-GAN TTS system in replacement of our VITS TTS system. Table III summarizes the basic information of all our datasets.

B. ASR adaptation

ESPnet [38] is used for all ASR experiments. Our pretrained ASR model is a SOTA Conformer model [39] trained on 960 hours of audio from the LibriSpeech dataset [40], which is a dataset containing audio recordings of English-language audiobooks recorded in 16kHz.

Our model follows an attention-encoder-decoder architecture [41], which consists of a Conformer encoder and an attention autoregressive (AR) decoder. Log Mel filterbank features [42] are extracted with a 25ms and 32ms window size and a 10ms and 16ms stride, model feature hidden dimension is 256 and 196, encoder Feed Forward Network (FFN) intermediate dimension is 1024 and 256 respectively. All our models use 80 dim. log Mel features, 4 attention heads, 6 decoder layers and 12 encoder layers. Swish activation [43] is used in FFN and the convolution modules [44] in the encoder has a convolution kernel of size 31. Decoder FFN intermediate dimension is 2048 and dropout rate is 0.1. Below the encoder is a convolutional subsampling module that receives log Mel feature inputs, and consists of a ReLU [45], two 2D convolution layers and a linear layer. A 2×2 stride and a 3×3 kernel is used in each convolution layer, effectively subsampling the input log Mel features by a factor of 4, and the convolution layer's channel size is the same as the model feature hidden dimension. We use joint CTC/attention decoding with CTC weight 0.3 [46] and beam size 20 during inference beam search [47].

We adapt the pretrained model to VITS-SPKSET1 (same-TTS-same-SPK) and TRANS-SPKSET1 (same-TTS-same-SPK) separately, to obtain 2 baseline ASR models. We then adapt the pretrained model to VITS-SPKSET1 + VITS-SPKSET2 (same-TTS-multi-SPK), and TRANS-SPKSET1 + TRANS-SPKSET2 (same-TTS-multi-SPK) respectively to form our second set of baselines. For our method, we adapt the pretrained model to VITS-SPKSET1 + TRANS-SPKSET1 (multi-TTS-same-SPK). Lastly we adapt the pretrained model to the real train set (ground-truth dataset) to find the performance upperbound.

During pretraining, we apply SpecAug [3] and set $m_F = 2$, $F = 27$, $m_T = 5$ and $p = 0.05$. Our dataset is augmented by speed perturbation with scaling factors of 0.9, 1.0, and 1.1. A joint CTC-attention loss function is employed with a CTC loss weight of 0.3. A label smoothing weight [48] of 0.1 is used in the attention loss, and the Adam optimizer [35] is used with a learning rate of 0.0025. For the learning

rate scheduling scheme, we follow ESPnet[38] implementation which has a warm-up and a decay stage. The warm-up stage has 25000 steps and the model is trained for 50 epochs. The final model is selected by averaging 10 checkpoints with the highest validation accuracy.

We also use the automatic mixed-precision (AMP) training implemented in PyTorch. During adaptation, we freeze the first 11 encoder layers and adapt only the last encoder layer, the CTC decoder and the attention decoder. We also reduce the learning rate to 0.000025 and the warmup steps to 1K, and train for 400 epochs.

V. Results and Discussion

Table IV shows that our model adapted to multi-TTS-same-SPK synthesized data outperforms all 4 baseline models adapted to same-TTS-same-SPK or same-TTS-multi-SPK synthesized data. Our model improves relative WER by 9.8% over the best baseline TRANS-SPKSET1+TRANS-SPKSET2, and improves relative WER by 71.4% over an unadapted model, showing the effectiveness of our adaptation method. However, there are still gaps between models adapted to real and synthesized data. Also, if we compare models adapted to single TTS data only, TransformerTTS performs better than VITS in all cases.

A. Discussion on why the multi-TTS-same-SPK approach outperform the same-TTS-multi-SPK approach in the experiments

As TTS synthesized data may contain artifacts that does not exist in the real data, we hypothesis that one factor contributing to the speaker diversity is the unique artifact pattern that exist in each TTS system. We further hypothesize that by adapting the ASR model to multiple TTS output distributions, it is less likely for the ASR model to over-fit to a specific TTS output distribution and therefore retaining the ASR model's capacity on processing real data.

B. Freeze model weights

Previous studies show that the top [49] or bottom [50] layers can be adapted to improve an ASR model performance in new domain data. Table V shows the result of adapting only the specific parts of the pretrained ASR model on VITS-SPKSET1. Compared to the unadapted results, adapting the last encoder layer improves relative WER by 31.6% and gives better result than adapting the first encoder layer. On the other hand, adapting both the CTC and attention-AR decoders improves relative WER by 52.3% and is the best adaptation result if the encoder is not adapted. Based on the results, we adapt the last encoder layer, the CTC and attention-AR decoders together and acheives the best WER of 21.1%, which improves relative WER 63.4% over the unadapted results, and improves relative WER by 8.3% over the results of adapting the whole ASR model.

C. Analysis on whether increasing the number of TTS systems can improve speaker diversity

Previous works have shown that conditioning a multi-speaker TTS system on a target speaker may produce synthesized speaker voice that sounds slightly different from the target speaker's ground-truth audio [8] [9]. We therefore hypothesize that a gap in speaker similarity may also exist between synthetic audios generated by different TTS systems conditioned on the same target speaker.

To investigate this hypothesis, we extract X-vector speaker embeddings from the ground-truth and synthesized audios and visualize the results using UMAP [51] in Figure 2. The structure of Figure 2 is as follows: It contains 4 sub-figures where each sub-figure shows the speaker embeddings of a target speaker. Within each sub-figure, there are 3 groups of speaker embeddings which corresponds to the ground-truth audio, the VITS TTS synthesized audio and the TransformerTTS synthesized audio respectively. Within each of the 3 groups, the speaker embeddings are extracted from audios speaking different transcripts.

From the results, we see that distinct clusters are formed from the ground-truth and synthetic audios. This shows that the speaker embeddings extracted from synthesized audios containing different transcripts are relatively close to each other if the audios are produced by the same TTS system conditioned on the same speaker, but the speaker embeddings are relatively far from each other if the audios are produced by different TTS systems conditioned on the same speaker.

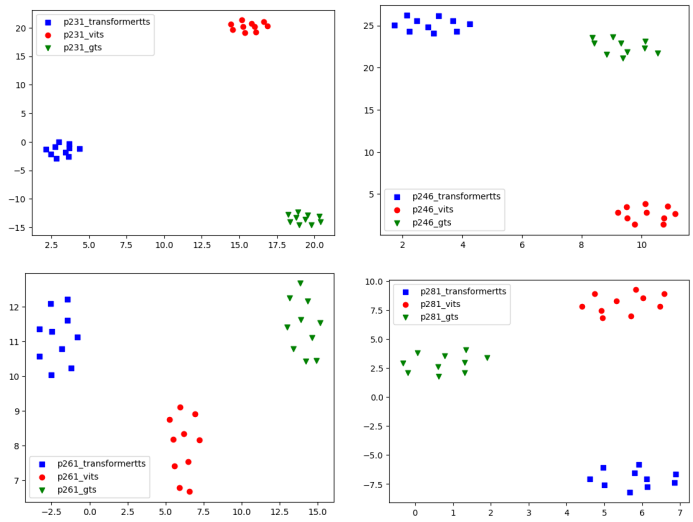


Fig. 2: UMAP plots of speaker embeddings extracted from ground-truth audios (gts), VITS TTS synthesized audios (vits) and TransformerTTS synthesized audios (transformertts). Best viewed in color.

TABLE IV: Effects of adapting a pretrained model on different combination of synthesized data. WER, substituion (sub.), insertion (ins.) and deletion (del.) errors are reported.

Data Combination	Data Type	Total # of sentences	WER (%)	Error (%) sub.	ins.	del.
None	n/a	0	57.6	44.0	5.5	8.2
Real train set	Real	13000	10.3	8.2	0.8	1.3
Our baselines						
VITS-SPKSET1	Synthesized	13000	21.1	16.2	1.1	3.8
TRANS-SPKSET1	Synthesized	13000	22.0	16.6	1.3	4.1
VITS-SPKSET1+VITS-SPKSET2	Synthesized	26000	19.2	14.9	0.9	3.3
TRANS-SPKSET1+TRANS-SPKSET2	Synthesized	26000	18.3	13.4	1.2	3.7
Our works						
VITS-SPKSET1+TRANS-SPKSET1	Synthesized	26000	**16.5**	**12.6**	1.0	**3.0**

TABLE V: Effects of adapting only specific parts of the pretrained ASR model on VITS-SPKSET1. WER, substituion (sub.), insertion (ins.) and deletion (del.) errors are reported on Aishell-1 dataset.

Encoder parts adapted	Decoders adapted	WER (%)	Error (%) sub.	ins.	del.
Unadapted					
-	-	57.6	44.0	5.5	8.2
Adapt encoder					
layer 1	-	62.5	47.3	3.7	11.6
layer 12	-	39.4	31.1	2.1	6.2
Adapt decoder					
-	CTC	62.5	46.8	4.1	11.6
-	attention-AR	43.1	32.3	2.8	8.0
-	CTC + attention-AR	27.5	20.8	1.9	4.9
Adapt encoder and decoder					
all	all	23.0	18.1	0.9	4.1
layer 12	attention-AR	21.7	16.8	1.2	3.8
layer 12	CTC + attention-AR	**21.1**	**16.2**	**1.1**	**3.8**

VI. Conclusion

In this work, we explore using multiple TTS systems to generate synthetic data for rare word ASR model adaptation, and show that it can improve the generated data's speaker diversity and improve ASR model adaptation results by up to relative 9.8% WER over same-TTS-multi-speaker systems. We also show a training strategy to freeze part of an ASR model when adapting to synthetic data, and it gives better result than adapting the whole ASR model. For future work, we will increase the number of TTS systems for ASR adaption and study the contributing factor of each TTS system.

References

[1] C. Peyser, S. Mavandadi, T. N. Sainath, J. Apfel, R. Pang, and S. Kumar, "Improving tail performance of a deliberation e2e asr model using a large text corpus," *arXiv preprint arXiv:2008.10491*, 2020.

[2] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*, IEEE, 2018, pp. 369–375.

[3] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[4] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages.," in *Interspeech*, 2014, pp. 1420–1424.

[5] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7069–7073.

[6] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6161–6165.

[7] C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7719–7723.

[8] R. Skerry-Ryan, E. Battenberg, Y. Xiao, *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*, PMLR, 2018, pp. 4693–4702.

[9] E. Cooper, C.-I. Lai, Y. Yasuda, *et al.*, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6184–6188.

[10] J. Melechovsky, A. Mehrish, D. Herremans, and B. Sisman, "Learning accent representation with multi-

level vae towards controllable speech synthesis," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 928–935.

[11] G. Sun, Y. Zhang, R. J. Weiss, *et al.*, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6699–6703.

[12] J. X. Koh, A. Mislan, K. Khoo, *et al.*, "Building the singapore english national speech corpus," *Malay*, vol. 20, no. 25.0, pp. 19–3, 2019.

[13] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: End-to-end contextual speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*, IEEE, 2018, pp. 418–425.

[14] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[15] U. Alon, G. Pundak, and T. N. Sainath, "Contextual speech recognition with difficult negative training examples," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6440–6444.

[16] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end speech recognition on voice search," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4764–4768.

[17] T. Hori, S. Watanabe, and J. R. Hershey, "Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2017, pp. 287–293.

[18] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 4945–4949.

[19] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[20] C. Gulcehre, O. Firat, K. Xu, *et al.*, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[21] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5674–5678.

[22] M. Mimura, S. Ueno, H. Inaguma, S. Sakai, and T. Kawahara, "Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 477–484.

[23] S. Xue, J. Tang, and Y. Liu, "Improving speech recognition with augmented synthesized data and conditional model training," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2022, pp. 443–447.

[24] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for asr using tts via a discrete representation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 68–75.

[25] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 6706–6713.

[26] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[27] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5530–5540.

[28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.

[29] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4879–4883.

[30] Y. Jia, Y. Zhang, R. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[31] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[32] T. Hayashi, R. Yamamoto, T. Yoshimura, *et al.*, "Espnet2-tts: Extending the edge of tts research," *arXiv preprint arXiv:2110.07840*, 2021.

[33] G. Eren and The Coqui TTS Team, *Coqui TTS*, version 1.4, Jan. 2021. DOI: 10.5281/zenodo.6334862. [Online]. Available: https://github.com/coqui-ai/TTS.

[34] Y. Iravantchi, M. Goel, and C. Harrison, "Digital ventriloquism: Giving voice to everyday objects," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–10.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4784–4788.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[38] S. Watanabe, T. Hori, S. Karita, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[39] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[41] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[42] A. Meghanani, C. Anoop, and A. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 670–677.

[43] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[44] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[45] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[46] T. Hori, S. Watanabe, and J. R. Hershey, "Joint ctc/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 518–529.

[47] I. Williams, A. Kannan, P. S. Aleksic, D. Rybach, and T. N. Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search.," in *Interspeech*, 2018, pp. 2227–2231.

[48] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[49] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2012, pp. 366–369.

[50] Y. Lee, A. S. Chen, F. Tajwar, *et al.*, "Surgical fine-tuning improves adaptation to distribution shifts," *arXiv preprint arXiv:2210.11466*, 2022.

[51] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Streaming End-to-End ASR Using CTC Decoder and DRA for Linguistic Information Substitution

Tatsunari TAKAGI,* Atsunori OGAWA†, Norihide KITAOKA* and Yukoh WAKABAYASHI*
* Toyohashi University of Technology, Japan
† Nippon Telegraph and Telephone Corporation, Japan

***Abstract*—This study proposes a method of streaming automatic speech recognition (ASR) for Japanese speech using domain adaptation. The development of ASR technology has led to its employment in various fields, however several issues, such as the decline recognition accuracy which occurs due to domain mismatches between training and test data. In particular, the accuracy of online speech recognition, also known as streaming ASR, which is becoming popular in real-world applications, also suffers from this issue. To realize domain-adapted streaming ASR, in this study we utilize a method of domain adaptation known as the Density Ratio Adaptation (DRA), which uses Language Models (LMs) trained with extensive text data. Our proposed method employs DRA for streaming ASR of Japanese speech, in a model using a CTC decoder for recognition. To facilitate streaming ASR processing, the CTC decoder successively replaces linguistic information on a frame-by-frame basis, obtaining recognition results through greedy searches. Considering CTC's assumption of conditional independence, we used unigram LMs for the subtraction and addition of language information when using DRA, during which selected linguistic information is replaced. Experimental results demonstrate that the proposed method enhances speech recognition accuracy when there is data domain mismatch.**

I. Introduction

Improvements in the performance of Automatic Speech Recognition (ASR) models in recent years have led to the widespread use of this technology across various fields. End-to-End ASR models , in particular, have gained considerable attention, and are being utilized in smart speakers, transcript generators and voice dialogue systems. Models employing bidirectional Long Short-Term Memory (LSTM) [1] and Self-Attention [2] can perform ASR while taking into account the entire input sequence, which is known as "batch ASR", yielding highly accurate recognition results. This technique is useful for ASR tasks that do not demand rapid responses, such as the generation of meeting minutes, although it requires waiting until the end of speech segment to output ASR results, making immediate recognition impossible. For this reason, batch ASR models are inadequate for tasks requiring real-time responses, such as verbal dialogue, voice assistants, and subtitling. Therefore, a technique is needed that can process speech feature input, then output the results in a forward direction in real time, which is called "streaming ASR". Chunk-based and frame-based [3] methods for streaming ASR have both been proposed. Chunk-based methods sequentially store input features in fixed-length chunks, then perform ASR when the chunks are filled. Frame-based methods recognize speech step by step by improving the decoding algorithm. Streaming ASR models such as Connectionist Temporal Classification (CTC) models [4] and RNN Transducer (RNN-T) models [5], using a Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN), have been proposed. One drawback of these methods, however, is lower recognition accuracy compared to batch ASR. This decrease in recognition accuracy is particularly evident when the domains of the data used to train the ASR model differ from the data to be recognized.

Various solutions have been proposed to tackle the issue of loss of recognition accuracy due to the domain mismatch, such as fine-tuning and domain adaptation methods using Language Models (LMs) [6], [7]. Fine-tuning requires both speech data and its transcription, from the domain to be recognized, which poses the challenge of time-consuming data collection. Thus, domain adaptation through fine-tuning, proves difficult when it is difficult to gather speech data and its transcription data. Conversely, domain adaptation methods using LMs have an advantage, as they use LMs trained solely on extensive text data, and thus require less time and effort to deploy, compared to acquiring speech and transcription data. The most common domain adaptation approach is Shallow Fusion, which combines the acoustic information from the ASR model and the linguistic information from the LM by adding their output probabilities in log-space with weighted coefficients [8]. Techniques for integrating LMs, such as Deep Fusion [9] and Cold Fusion [10], have also been proposed, but require retraining of the ASR model. Shallow Fusion, which does not require retraining of the ASR model, is commonly adopted due to its advantage of being able to easily integrate various types of LMs [11]–[13].

However, since the output of the ASR model already contains language information, overlapping of language information can occur when integrating an LM using Shallow Fusion. The Density Ratio Approach (DRA) [14], an extension of Shallow Fusion, has been proposed to address this issue. By using an LM trained with data from the same domain used for training the ASR model, and then subtracting the implicit language information contained within the ASR model, more precise domain adaptation has been demonstrated. It has been

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

shown that DRA is effective when using the RNN-T model for English, Spanish, and Italian speech data [15], and with the Hybrid CTC/Attention model and Joint CTC Attention Transformer model for Japanese data [16]. However, all of the methods using DRA incorporate decoding with a beam search [17], obtaining the final recognition results by reranking multiple candidates through substitution of the linguistic information using an LM. Although methods using beam searches can attain high recognition accuracy, they are slow to output the final recognition results, thus they cannot provide recognition results without a delay. In contrast, decoding using a greedy search allows sequential output of the recognition results at high speed, albeit at a lower level of recognition accuracy than that achieved by a beam search. DRA has been previously considered in the context of streaming ASR with RNN-T models [14], [15], however its application in streaming ASR models using CTC has yet to be explored. Although the recognition accuracy of CTC is lower than that of RNN-T, CTC offers advantages such as easier ASR model training, lower computational complexity, and smaller latency. Consequently, improving the recognition accuracy of streaming ASR models that utilize CTC is a significant issue yet to be resolved.

In this study, we refine the CTC decoder of a streaming ASR model that uses greedy searches, perform DRA on a per-time-frame basis, and demonstrate that domain adaptation is effective for enhancing accuracy, which is demonstrated through a performance evaluation using Character Error Rates (CERs).

II. Related work

A. Connectionist Temporal Classification (CTC)

CTC is a method specifically for labeling time-series data in ASR. To assign labels to time-series data, the probability of the occurrence of these labels in each frame is computed for a predefined set of labels. These labels include sets of words, phonemes and characters, as well as a `<blank>` symbol, which is neither a phoneme nor a character. The sequence of labels that maximizes the occurrence probability at each frame is determined, and a string is then obtained by eliminating repeated characters and `<blank>`. The loss function of CTC is expressed by the following formula:

$$\mathcal{L}_{\mathrm{CTC}} = -\log P(y|x), \tag{1}$$

Here, $P(y|x)$ represents the conditional probability of the correct that label sequence y is the correct label for input sequence x. This probability is calculated as the sum of the probabilities for all possible paths π. For instance, if the correct character string is $y = \texttt{cat}$, examples of the set of π of length 8 include:

$$\begin{aligned}\pi \;=\; & \{\texttt{c},_,\texttt{a},\texttt{a},\texttt{a},_,\texttt{t},_\},\{\texttt{c},\texttt{c},\texttt{a},\texttt{a},_,_,_,\texttt{t}\},\\ & \{_,_,_,\texttt{c},\texttt{a},\texttt{t},_,_\},\{_,_,\texttt{c},_,\texttt{a},_,\texttt{t},\texttt{t}\}.\end{aligned}$$

The sum of probabilities for all π that can lead to $y = \texttt{cat}$ becomes $P(y|x)$. Therefore, $P(y|x)$ can be represented as:

$$P(y|x) = \sum_{\pi \in B^{-1}(y)} P(\pi|x), \tag{2}$$

where B is a mapping from a path to a label sequence, and $B(\pi)$ is the label sequence corresponding to path π. $B^{-1}(y)$ is the inverse mapping of y, i.e., the set of paths corresponding to y. Because there are many such paths, it is difficult to calculate $P(y|x)$ directly. Therefore, in CTC, the forward-backward algorithm of Hidden Markov Models (HMMs) is employed to efficiently compute $P(y|x)$.

CTC calculates the probability of label appearance for the output of the ASR model at each frame, enabling streaming ASR by performing post-processing on the output for each frame.

B. Shallow Fusion

Shallow Fusion is a method that combines the outputs of an LM and an ASR model by taking a weighted sum in the log-space. This technique combines the acoustic information held by the ASR model and the linguistic information held by the LM. Shallow Fusion is a standard method for domain adaptation using an LM, and is formulated by the following equation:

$$y = \arg\max_{y} \{\log P_{\mathrm{ASR}}(y|x) + \lambda \log P_{\mathrm{LM}}(y)\}, \tag{3}$$

where $P_{\mathrm{ASR}}(y|x)$ is the probability of the ASR model outputting string y when given acoustic feature x, and $P_{\mathrm{LM}}(y)$ is the output probability of the LM trained with target domain recognition data. The advantage of Shallow Fusion is that it allows the training of the ASR model and an external LM with data from different domains.

C. Density Ratio Approach (DRA)

DRA, which is an extension of Shallow Fusion, is formulated by the following equation:

$$\begin{aligned}y = \arg\max_{y} \{&\log P_{\mathrm{ASR}}(y|x) - \lambda_{\mathrm{sub}} \log P_{\mathrm{LM,src}}(y)\\ &+ \lambda_{\mathrm{add}} \log P_{\mathrm{LM,tgt}}(y)\},\end{aligned} \tag{4}$$

Here, $P_{\mathrm{ASR}}(y|x)$ and $P_{\mathrm{LM,tgt}}(y)$ are the output probabilities of the ASR model and the LM trained with target domain data, respectively, similar to Shallow Fusion. $P_{\mathrm{LM,tgt}}(y)$ is called the "additive LM". $P_{\mathrm{LM,src}}(y)$ is the output probability of the LM trained using the same transcribed text used to train the ASR model, and is referred to as the "subtractive LM". By using this subtractive LM, the implicit language information included in the output of the ASR model is subtracted, and the language information of the target domain is added. This allows for the replacement of source domain language information with target domain information, enabling more accurate domain adaptation.

Algorithm 1 Decoding algorithm for applying DRA in a streaming CTC decoder. <none> is the symbol for nothing, which is removed during decoding. pack is a function that removes <none> from each element of the input and combines them into a string.

Input: $\mathbf{y}_t \in \mathbb{R}^{N\times 1}$, $\mathrm{LM}_{\mathrm{src}}$, $\mathrm{LM}_{\mathrm{tgt}}$,
Output: S_t

$\mathrm{idx}_t \leftarrow \mathrm{DRA}(\mathbf{y}_t, \mathrm{idx}_{t-1}, \mathrm{LM}_{\mathrm{src}}, \mathrm{LM}_{\mathrm{tgt}}, \mathrm{pack}(S_{0:t-1}))$
if $\mathrm{idx}_t \neq \mathrm{idx}_{t-1}$ **and** $\mathcal{D}(\mathrm{idx}_{t-1}) \neq$ <blank> **then**
  $S_{t-1} \leftarrow \mathcal{D}(\mathrm{idx}_{t-1})$
else
  $S_{t-1} \leftarrow$ <none>
end if
function $\mathrm{DRA}(\mathbf{y}_t, \mathrm{idx}_{t-1}, \mathrm{LM}_{\mathrm{src}}, \mathrm{LM}_{\mathrm{tgt}}, \mathrm{pack}(S_{0:t-1}))$
  for $i \leftarrow 0$ to N **do**
    if $i \neq \mathrm{idx}_{t-1}$ **then**
      $\mathbf{y}_t[i] \leftarrow \mathbf{y}_t[i] - \lambda_{\mathrm{sub}} \cdot \mathrm{LM}_{\mathrm{src}}[\mathcal{D}(i)] + \lambda_{\mathrm{add}} \cdot \mathrm{LM}_{\mathrm{tgt}}[\mathcal{D}(i), \mathrm{pack}(S_{0:t-1})]$
    end if
  end for
  return $\arg\max_i(\mathbf{y}_t[i])$
end function

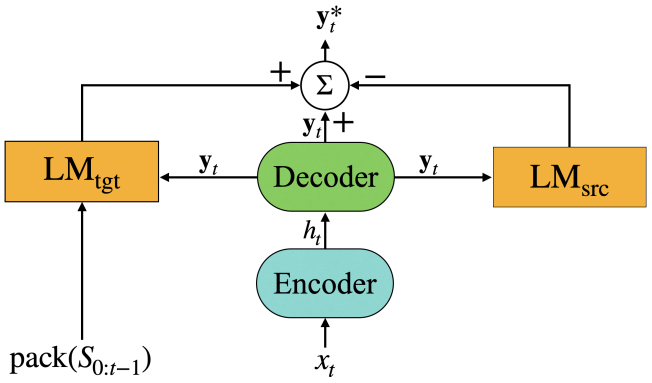


Fig. 1. Extension of DRA to apply to frame-by-frame output, where $\mathrm{LM}_{\mathrm{tgt}}$ and $\mathrm{LM}_{\mathrm{src}}$ are LMs trained with the target and source domain data, respectively. x_t is an acoustic feature, h_t is its embedding, $\mathbf{y}_t$ is a vector consisting of the output probabilities of the ASR model's decoder at the t-th frame.

III. Proposed Method

In this study, we apply DRA to streaming ASR in an end-to-end ASR model with a unidirectional LSTM encoder and a CTC decoder, using a greedy decoding algorithm[1], and a unigram LM. DRA replaces the implicit language information contained in text data by subtracting it via an LM. Ideally, the subtractive LM would be a suitable LM derived from the data used to train the ASR model, as with Hybrid Autoregressive Transducer (HAT) [18], but it is not feasible to identify the implicit language information from the training of the ASR model when using CTC.

CTC assumes conditional independence, i.e., that the output of the ASR model does not depend on the preceding or following context. We hypothesized that a unigram LM, which does not contain preceding or following context information, would be suitable as the subtractive LM used to apply DRA to the output of each frame.

In the frame-by-frame output of the CTC module, consecutive characters are eventually combined into a single character. Therefore, the timing of changes in language information occurs in two instances; when transitioning from <blank> to character, and when transitioning from one character to another. Thus, when we apply DRA to a frame defined as t, DRA is only applied to labels other than the character in the $(t-1)$-th frame. Here, we assume no LM processing for transitions to <blank>.

Figure 1 shows the extended form of DRA that was used on a frame-by-frame basis. The output of the CTC decoder on a frame-by-frame basis is typically the output probability of each label, and the recognition result is obtained by performing best path decoding, i.e., selecting the solution with the maximum likelihood value. However, in our proposed method, DRA is applied to each label in the frame-by-frame output before performing best path decoding. DRA as used in the proposed method is expressed by the following equation:

$$\begin{aligned}\mathrm{idx}_t = \arg\max_i \{&\log P_{\mathrm{ASR}}(\mathcal{D}(i)|x_t) \\ &- \lambda_{\mathrm{sub}} \log P_{\mathrm{uniLM,src}}(\mathcal{D}(i)) \\ &+ \lambda_{\mathrm{add}} \log P_{\mathrm{LM,tgt}}(\mathcal{D}(i)|\mathrm{pack}(S_{0:t-1}))\}, \end{aligned} \quad (5)$$

where $P_{\mathrm{ASR}}(\mathcal{D}(i)|x_t)$ is the label probability of the ASR model at the t-th frame, and where $P_{\mathrm{uniLM,src}}(\mathcal{D}(i))$ is the label probability of the unigram LM at the t-th frame. $P_{\mathrm{LM,tgt}}(\mathcal{D}(i)|\mathrm{pack}(S_{0:t-1}))$ is the label probability of the LM that is added to the output of the ASR model up to the $(t-1)$-th frame, excluding <blank> and continuous labels. $\mathcal{D}$ is a dictionary containing a character set and <blank> for ASR, i.e., $\mathcal{D}(i)$ indicates the i-th character, and $\mathrm{pack}(S_{0:t-1})$ is a determined string that does not change depending on the output of the ASR model after the t-th frame. Therefore, by outputting the string $\mathrm{pack}(S_{0:t-1})$ we can achieve streaming ASR. The pseudo-code for this approach is shown in Algorithm 1.

IV. Experiments

A. Experimental conditions

In our experiment, we used NVIDIA/NeMo [19], an audio processing toolkit, for training, inference and evaluation of the ASR and LMs. The data used in the experiment was taken from the Academic Presentation Speech (APS) and Spontaneous Presentation Speech (SPS) datasets of the Corpus of Spontaneous Japanese (CSJ). The APS and SPS datasets were each divided into "train", "dev" and "test" sets, which were

[1]In this paper, we propose a DRA algorithm which uses the greedy decoding algorithm. Note that our algorithm can be easily applied to beam search-based recognition by retaining multiple histories.

TABLE I
DETAILS OF APS AND SPS DATASETS

Dataset	Data Split	# of Speakers	Duration (h)
APS	train	889	253
	dev	49	10
	test	25	5
SPS	train	1555	303
	dev	80	13
	test	40	6

TABLE II
PERPLEXITY IN THE TEST SETS OF THE LANGUAGE MODELS USED FOR ADDITION AND SUBTRACTION

LM	Training set	Test set	
		APS	SPS
unigram	APS	184.1	199.9
unigram	SPS	232.2	162.5

used for training, development and cross-domain evaluation. Details of APS and SPS datasets are shown in Table I.

We used the LSTM-CTC ASR model and model architecture proposed in [20]. The encoder consisted of a 4-layer unidirectional LSTM, while the decoder used a single-layer conv1d, which applies 1D convolution over a multi-plane input signal. The input to the ASR model was an 80-dimensional log-mel spectrogram, and the per-frame output dimensionality of the ASR model was 3219, including the character set and <blank>. The learning rate was set at 1.0, AdamW [21] was used as the optimization algorithm, batch size was 128, and the maximum number of epochs was 50. The model used was selected based on the highest recognition accuracy evaluated using the dev dataset after each epoch.

We employed a unigram LM for both the subtractive and additive LMs. The unigram LM was trained using the training sets of both the APS and SPS datasets. We evaluated the performance of three LM integration methods (Baseline, Shallow Fusion and DRA) with different training and test sets. When using the baseline method, no addition or subtraction was performed by the LM, and the output of the ASR model was used "as is". The Shallow Fusion test approximated the ASR and additive unigram LM outputs through a weighted sum in the logarithmic domain. The DRA test subtracted the output of the subtractive unigram LM from the ASR model output, and added the output of the additive unigram LM, that the linguistic information contained in the two corpora differs. The addition and subtraction weights of the LM were tuned using the "dev" set. The evaluation metric used was the speech recognition CER.

In addition, we conducted experiments to confirm the accuracy of the timing of DRA application when using our proposed method, in which DRA is applied only to labels at the t-th frame, excluding the output label at the $(t-1)$-th frame. For our comparison experiment, DRA was also applied to all labels at the t-th frame, regardless of the output label at the $(t-1)$-th frame. The CERs when using each timing method were then compared.

TABLE III
RESULTS OF LM FUSION EXPERIMENT

Training set	Test set	Method	CER (%)
APS	SPS	Baseline	20.0
		Shallow Fusion	20.2
		DRA	**19.5**
SPS	APS	Baseline	23.2
		Shallow Fusion	23.0
		DRA	**22.2**

B. Experimental results

The results of the experiment are shown in Table III. We can see that the proposed DRA method achieved the lowest CER for both the APS and SPS test data. This suggests that implicit linguistic information can be replaced using a unigram LM to provide the language information to be subtracted from or added to, the per-frame output of CTC. We can also see that no improvement in CER from the baseline method was observed when LM integration was performed using the Shallow Fusion method, in contrast to the CER achieved when using DRA, which was slightly lower than the baseline method's CER. This can be attributed to the overlap of linguistic information contained in the output of the ASR model, and that contained in the output of the LM, when using Shallow Fusion. These results demonstrate the effectiveness of DRA for LM integration in streaming ASR using CTC.

The results of the experiment comparing the timing of DRA application showed that the proposed method achieved a higher ASR accuracy, with an improvement in CER of approximately 0.2%, compared to when DRA was applied without taking into consideration the output label at the $(t-1)$-th frame. We believe this is due to the fact that the proposed method was able to replace the linguistic information with the appropriate timing, by applying DRA at two points during the CTC ASR model's streaming output; when it changes from <blank> to a character and when it changes from one character to another.

V. CONCLUSIONS

In this study, we experimentally verified the effectiveness of domain adaptation through DRA in streaming ASR using a CTC decoder by demonstrating that the proposed method enhances speech recognition accuracy in cross-domain settings. We confirmed the efficacy of using a unigram LM for subtraction in DRA, and designed an algorithm for streaming ASR using DRA. Future work includes the choice of LMs for the addition of language information, measuring latency and real-time factor, and validating the results by using various dataset combinations because we used only two datasets.

ACKNOWLEDGMENT

This study was supported by JSPS KAKENHI Grants JP22H04863 and 23H00493.

References

[1] M. Schuster and K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. DOI: 10.1109/78.650093.

[2] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[3] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing Real-Time Streaming Transformer Transducer for Speech Recognition on Large-Scale Dataset," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5904–5908. DOI: 10.1109/ICASSP39728.2021.9413535.

[4] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: ACM, 2006, pp. 369–376, ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143891. [Online]. Available: http://dx.doi.org/10.1145/1143844.1143891.

[5] A. Graves, *Sequence Transduction with Recurrent Neural Networks*, 2012. arXiv: 1211.3711 [cs.NE].

[6] D. Wang and T. F. Zheng, "Transfer Learning for Speech and Language Processing," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 1225–1237. DOI: 10.1109/APSIPA.2015.7415532.

[7] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer Learning of Language-independent End-to-end ASR with Language Model Fusion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6096–6100. DOI: 10.1109/ICASSP.2019.8682918.

[8] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An Analysis of Incorporating an External Language Model into a Sequence-to-Sequence Model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5828. DOI: 10.1109/ICASSP.2018.8462682.

[9] C. Gulcehre, O. Firat, K. Xu, *et al.*, *On Using Monolingual Corpora in Neural Machine Translation*, 2015. arXiv: 1503.03535 [cs.CL].

[10] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold Fusion: Training Seq2Seq Models Together with Language Models," in *Proc. Interspeech 2018*, 2018, pp. 387–391. DOI: 10.21437/Interspeech.2018-1392. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1392.

[11] C. Peyser, S. Mavandadi, T. N. Sainath, J. Apfel, R. Pang, and S. Kumar, "Improving Tail Performance of a Deliberation E2E ASR Model Using a Large Text Corpus," in *Proc. Interspeech 2020*, 2020, pp. 4921–4925. DOI: 10.21437/Interspeech.2020-1465.

[12] T. Lee, M.-J. Lee, T. G. Kang, *et al.*, "Adaptable Multi-Domain Language Model for Transformer ASR," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7358–7362. DOI: 10.1109/ICASSP39728.2021.9413475.

[13] K. Hu, T. N. Sainath, B. Li, *et al.*, "Massively Multilingual Shallow Fusion with Large Language Models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094796.

[14] E. McDermott, H. Sak, and E. Variani, "A Density Ratio Approach to Language Model Fusion in End-to-End Automatic Speech Recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 434–441. DOI: 10.1109/ASRU46091.2019.9003790.

[15] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing RNN Transducer Technology for Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5654–5658. DOI: 10.1109/ICASSP39728.2021.9414716.

[16] D. Mori, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "Advanced Language Model Fusion Method for Encoder-decoder Model in Japanese Speech Recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 503–510.

[17] K. Hwang and W. Sung, "Character-level Incremental Speech Recognition with Recurrent Neural Networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5335–5339. DOI: 10.1109/ICASSP.2016.7472696.

[18] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid Autoregressive Transducer (HAT)," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6139–6143. DOI: 10.1109/ICASSP40776.2020.9053600.

[19] O. Kuchaiev, J. Li, H. Nguyen, *et al.*, *NeMo: A Toolkit for Building AI Applications Using Neural Modules*, 2019. arXiv: 1909.09577 [cs.LG].

[20] I. McGraw, R. Prabhavalkar, R. Alvarez, *et al.*, "Personalized Speech Recognition on Mobile Devices," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5955–5959. DOI: 10.1109/ICASSP.2016.7472820.

[21] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, 2019. arXiv: 1711.05101 [cs.LG].

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Biometric Signature Scheme with Template Protection and Authenticated Sample Recoverability

Wataru Nakamura* and Kenta Takahashi†
* Hitachi, Ltd., Japan
E-mail: wataru.nakamura.va@hitachi.com
† Hitachi, Ltd., Japan
E-mail: kenta.takahashi.bw@hitachi.com

***Abstract*—Biometric Template Protection (BTP) is an approach to protect biometric data in online biometric authentication. A hopeful scheme for BTP is *Biometric Signatures (BSs)*, which generates a digital signature for a given message using fuzzy biometric data as a private key, utilizing cryptographic tools such as fuzzy extractors. By advancing BTP, we introduce *Authenticated Sample Recoverability (ASR)*, a property that raw biometric samples such as images and videos of biometric traits having captured during past procedures can be recovered from the stored data and a biometric feature generated from the genuine user's sample captured again, while it is difficult to recover the past samples or features only from the stored data. This property is desired in various use cases. For example, in forensics, it is required to post-verify whether the genuine user used the system or fake sample was input during a past specific procedure, because the recovered sample is a strong evidence that the past sample was real or fake. Other examples include investigation of a cause when authentication accuracy in operation is low, and biometric reference update with multi-sample fusion using samples having captured in past procedures. To realize a template protected online biometric authentication system with ASR, we propose *Sample Recoverable Biometric Signatures (SRBSs)*. We present a generic construction of a SRBS scheme with EUF-CMA security, and give a specific construction in the random oracle model.**

I. Introduction

A. Background

As the spread of online services such as banking, shopping, file storage, and social media, online user authentication technology is becoming more and more important. Authentication procedures are performed with secret information, so it is crucial to prevent the secret information leakage. In addition, it is required to be secure against online attacks such as a replay attack and a real-time phishing. User convenience is also important because inconvenience hinders widespread use.

Widely used authentication methods are based on passwords, but it is difficult for password-based schemes to achieve both security and convenience. If users set a password to a simple one, it becomes more easy to guess. On the other hand, asking the user for a complex password reduces convenience.

Another method is based on a secret key stored on devices such as user's computer, IC cards, and hardware tokens. In particular, digital signature-based schemes such as FIDO[1] can protect online attacks such as a replay attack and a real-time phishing, and makes it difficult to guess the secret key from the verification key (or the public key) stored in the server and signatures transmitted during authentication procedures. However, users have to bring a device storing the secret key to authenticate themselves, which reduces convenience. Furthermore, if users lose the device, it leads to the risk of unauthorized use of the secret key[2].

A hopeful alternative is online biometric authentication, beginning to be used in commercial and governmental services [2][3]. In online biometric authentication systems, users can authenticate themselves using their own biometric traits such as face, fingerprint, and iris without remembering complex passwords or bringing a device for storing the secret key. A simple scheme might be one in which the server stores biometric data captured during an enrollment procedure, and compares it to biometric data captured during an authentication procedure. However, this scheme leads to the risk of biometric data leakage from the data stored in the server[3]. The leakage causes privacy invasion and increases impersonation risk. Furthermore, unlike passwords, it is difficult for users to change their biometric data if compromised. Also, there is a risk of online attacks for the simple scheme.

To protect biometric information from leakage, *Biometric Template Protection (BTP)* [5][6] has been studied and standardized. In BTP, biometric data is transformed into a biometric reference so that it is difficult to recover the original biometric data from the biometric reference. In particular, BTP which protects online attacks can be realized by schemes such as *Fuzzy Extractors (FEs)* [7][8] and *Fuzzy Signatures (FSs)* [9][10]. FEs generate a pair (r, p) of a secret key r and a helper string p from a fuzzy biometric feature y. The secret key r can be reproduced from the helper string p and a feature y' sufficiently close to y, and it is difficult to guess y or r from the helper string p alone. By using the secret key r for generating a digital signature, secure online authentication can

[1]FIDO is a trademark of FIDO Alliance.

[2]In order to lower the risk of unauthorized use, the device can perform additional authentication locally for access to the secret key. However, it is difficult for password-based schemes to achieve both security and convenience. Using biometric authentication locally is less convenient as compared to online biometric authentication because the former requires users to bring their devices after all.

[3]For example, due to a vulnerability in the Aadhaar system, anybody had access to biometric data of more than 1 billion Indian citizens [4].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

be realized. FSs generate a digital signature using biometric data as a private key without a helper data. We call schemes which generate a digital signature for a given message using biometric data as a private key a *Biometric Signature (BS)*. BSs include FE-based digital signatures and FSs.

B. Motivation for Authenticated Sample Recoverability (ASR)

BTP makes it difficult for biometric data to be recovered from the stored biometric reference. By advancing this, we introduce *Authenticated Sample Recoverability (ASR)*, a property of achieving the following two: (a) From the stored data and a biometric feature generated from the genuine user's sample captured again, raw biometric samples such as images and videos of biometric traits having captured during past procedures can be recovered[4]. (b) It is difficult to recover the past samples or features only from the stored data. ASR is desired in the following cases, for example.

(Case 1) *Post-verification of past procedures in forensics.* Consider a situation in which a suspect of a crime wants to prove an alibi by showing that he/she used the authentication system at a specific time and place. The fact that the authentication was successful might be insufficient for the proof when he/she is claimed that another person might have input a fake biometric sample imitating that of the subject. Conversely, there is a situation when a user wants to prove that he/she did not use the system at a specific time and place, e.g., when an attacker spoofed the system by a presentation attack [11]. In these situations, if samples during past procedures can be recovered with the help of the genuine user's feature, it is a strong evidence that the user did or did not use the system at a specific time and place. In other words, this property ensures the accountability of the procedure.

(Case 2) *Investigation of a cause of low authentication accuracy.* Consider a case in which a user is relatively frequently rejected from the system (e.g., the false rejection rate of the system is 0.1% by specification but the user is rejected with the probability 10%). If samples during past procedures can be recovered with the help of the genuine user's feature, they can be a powerful clue for investigating the cause in more detail. One can check, for example, whether the past samples were captured under appropriate lighting conditions and whether the biometric traits are presented in an appropriate posture.

(Case 3) *Biometric reference update.* When a biometric authentication system upgrades algorithms such as feature extraction, it has to update enrolled biometric references to reflect the upgrade. It is known as *multi-sample fusion* that authentication accuracy can be improved by the system generating a biometric reference from multiple samples (e.g., [12][13][14]). If samples during past procedures can be recovered during the update procedure, they can be used for multi-sample fusion.

C. Limitation of Previous Work on ASR

Online biometric authentication systems with authenticated recoverability can be realized to some extent by BEs based on Dodis et al.'s FE [7]. Biometric authentication systems based on their FE[5] can recover the feature y obtained in the enrollment procedure from the helper string and a feature y' sufficiently close to y. Also, biometric features generated in authentication procedures can also be recovered with a sufficiently close feature if the system runs the generation algorithm additionally and stores the generated helper string during each authentication procedure[6]. In this way, BEs based on their FE can realize a property that biometric *features* having generated during past procedures can be recovered from the stored data and a feature generated from the genuine user's sample captured again, while it is difficult to recover the past samples or features only from the stored data, which we call *Authenticated Feature Recoverability (AFR)*.

However, for (Case 1)–(Case 3) described above, ASR is needed rather than AFR. In BEs with AFR, the recoverable features are elements of a metric space such as a Hamming space, generated from samples by image processing and pattern recognition techniques. On the other hand, for (Case 1)–(Case 3), raw biometric samples such as images and videos of biometric traits is needed. The samples are a stronger evidence in (Case 1), a more helpful clue in (Case 2), and required to generate biometric references in (Case 3).

D. Our Contribution

Our goal is to construct a template protected online biometric authentication system with ASR. For this goal, we propose *Sample Recoverable Biometric Signatures (SRBSs)*. Our technical contributions are threefold.

(i) *Formal definition for SRBSs* (Sect. III). We formalize a biometric data setting which specifies how biometric data are generated. We then give a formal definition of SRBSs associated with the biometric data setting.

(ii) *Generic construction* (Sect. IV). We propose a generic construction of a SRBS scheme with EUF-CMA security based on a FE, a digital signature scheme, and a symmetric encryption scheme satisfying a specific property.

(iii) *Instantiation* (Sect. V). We present a specific construction of a SRBS scheme with EUF-CMA security in the random oracle model.

By using SRBSs with EUF-CMA security, we can construct online biometric authentication systems with ASR (Sect. VI). As described above, the systems ensure the accountability of past procedures, enable more detail investigation of the cause

[4]Strictly, our definition of recoverability is the property that past samples can be recovered from the stored data and a feature sufficiently close to ones generated from the past samples. We use the expression "the genuine user" because features generated from samples of the genuine user should be close with high probability.

[5]In their FE, the helper string includes a *secure sketch* $\mathrm{SS}(y)$ generated from a feature y so that y can be recovered from $\mathrm{SS}(y)$ and a feature y' sufficiently close to y.

[6]Strictly, when helper strings are generated multiple times, *reusable* FEs [8] have to be used for security.

if authentication accuracy in operation is low, and can apply multi-sample fusion for biometric reference update using past samples.

II. Preliminaries

In this section, we review basic notations and definitions of primitives.

A. Basic Notation

Let $\mathbb{N}$ denote the set of all positive integers, and let $\mathbb{R}$ and $\mathbb{R}_{>0}$ denote the sets of real numbers and positive real numbers, respectively. Throughout the paper, k denotes the security parameter. A function $f : \mathbb{N} \to [0,1]$ is said to be *negligible* if for all polynomials p and all sufficiently large k it holds $f(k) < 1/p(k)$. For random variables X and Y, the *min-entropy* $\mathbf{H}_\infty(X)$ of X is defined by $\mathbf{H}_\infty(X) := -\log_2(\max_x \Pr[X = x])$, and the average min-entropy $\tilde{\mathbf{H}}_\infty(X \mid Y)$ of X given Y is defined by $\tilde{\mathbf{H}}_\infty(X \mid Y) := -\log_2 \mathbf{E}_{y \leftarrow \mathrm{Y}}\big[\max_x \Pr[X = x \mid Y = y]\big]$, where Y denotes the probability distribution of Y. The *statistical distance* $\mathbf{SD}(X, X')$ between random variables X and X' is defined by $\mathbf{SD}(X, X') := (1/2)\sum_z |\Pr[X = x] - \Pr[X' = x]|$. For $\ell \in \mathbb{N}$, U_ℓ denotes a random variable uniformly distributed on $\{0,1\}^\ell$. For a positive function f, we write $f(k) = \Theta(k)$ if there exist c_1, c_2, and k_0 such that for any k larger than k_0 it holds that $c_1 < f(k)/k < c_2$.

B. Fuzzy Extractors

We review the definition of a fuzzy extractor introduced by Dodis et al. [7]. A $(\mathcal{Y}, \mu, \ell, t, \epsilon)$-fuzzy extractor Σ_{FE} is defined by the following algorithms $(\mathrm{Gen}_{\mathrm{FE}}, \mathrm{Rep}_{\mathrm{FE}})$.

- $\mathrm{Gen}_{\mathrm{FE}}(y) \to (r, p)$: The generation algorithm takes an element $y \in \mathcal{Y}$ as an input, and outputs an extracted string $r \in \{0,1\}^\ell$ and a public string $p \in \{0,1\}^*$. We require that for any random variable Y on $\mathcal{Y}$ of min-entropy μ, if $(R, P) \leftarrow \mathrm{Gen}_{\mathrm{FE}}(Y)$, then $\mathbf{SD}((R, P), (U_\ell, P)) \le \epsilon$.
- $\mathrm{Rep}_{\mathrm{FE}}(y', p) \to r'$: The reproduction algorithm $\mathrm{Rep}_{\mathrm{FE}}$ takes an element $y' \in \mathcal{Y}$ and a bit string $p \in \{0,1\}^*$ as inputs, and outputs a string r'. We require that for any $y, y' \in \mathcal{Y}$ satisfying $\mathtt{dist}(y, y') \le t$ and (r, p) generated by $(r, p) \leftarrow \mathrm{Gen}_{\mathrm{FE}}(y)$, it holds that $\mathrm{Rep}_{\mathrm{FE}}(y', p) = r$.

C. Digital Signature Schemes

A digital signature scheme Σ_{SIG} is defined by the following algorithms $(\mathrm{Gen}_{\mathrm{SIG}}, \mathrm{Sign}_{\mathrm{SIG}}, \mathrm{Ver}_{\mathrm{SIG}}, \mathrm{KeyVer}_{\mathrm{SIG}})$.

- $\mathrm{Gen}_{\mathrm{SIG}}(1^k) \to (sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}})$: The generation algorithm takes the security parameter 1^k as an input, and outputs a signing key sk_{SIG} and a verification key vk_{SIG}.
- $\mathrm{Sign}_{\mathrm{SIG}}(sk_{\mathrm{SIG}}, m) \to \sigma_{\mathrm{SIG}}$: The signing algorithm takes a signing key sk_{SIG} and a message $m \in \mathcal{M}$ as inputs, and outputs a signature σ_{SIG}.
- $\mathrm{Ver}_{\mathrm{SIG}}(vk_{\mathrm{SIG}}, m, \sigma_{\mathrm{SIG}}) \to result$: The verification algorithm takes a verification key vk_{SIG}, a message $m \in \mathcal{M}$, and a signature σ_{SIG} as inputs, and outputs the verification result $result \in \{\top, \bot\}$. We require that for any $k \in \mathbb{N}$, message $m \in \mathcal{M}$, and $(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}, \sigma_{\mathrm{SIG}})$ generated by $(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k)$ and $\sigma_{\mathrm{SIG}} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(sk_{\mathrm{SIG}}, m)$, it holds that $\mathrm{Ver}_{\mathrm{SIG}}(vk_{\mathrm{SIG}}, m, \sigma_{\mathrm{SIG}}) = \top$.
- $\mathrm{KeyVer}_{\mathrm{SIG}}(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \to result$: The key verification algorithm[7] takes a key pair $(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}})$ as an input, and outputs the verification result $result \in \{\top, \bot\}$. We require that for any $(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}})$ generated by $(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k)$ and sk'_{SIG}, if $sk_{\mathrm{SIG}} = sk'_{\mathrm{SIG}}$ then $\mathrm{KeyVer}_{\mathrm{SIG}}(sk'_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) = \top$ holds, and otherwise $\mathrm{KeyVer}_{\mathrm{SIG}}(sk'_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) = \bot$ holds.

EUF-CMA security for a digital signature scheme Σ_{SIG} is defined as follows. Consider the following EUF-CMA experiment $\mathrm{Exp}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SIG}},\mathcal{A}}(k)$ for $\Sigma_{\mathrm{SIG}} = (\mathrm{Gen}_{\mathrm{SIG}}, \mathrm{Sign}_{\mathrm{SIG}}, \mathrm{Ver}_{\mathrm{SIG}}, \mathrm{KeyVer}_{\mathrm{SIG}})$ and an adversary $\mathcal{A}$:

$$\begin{aligned}\mathrm{Exp}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SIG}},\mathcal{A}}(k) : [\ & (sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k);\ \mathcal{Q} := \emptyset;\\ & (m', \sigma'_{\mathrm{SIG}}) \leftarrow \mathcal{A}^{\mathcal{O}_{\mathrm{Sign}_{\mathrm{SIG}}}(\cdot)}(vk_{\mathrm{SIG}});\\ & \text{if } m' \notin \mathcal{Q} \wedge \mathrm{Ver}(m', vk_{\mathrm{SIG}}, \sigma'_{\mathrm{SIG}}) = \top\\ & \text{then return } 1 \text{ else return } 0\ ],\end{aligned}$$

where $\mathcal{O}_{\mathrm{Sign}_{\mathrm{SIG}}}$ is a signing oracle that takes a message $m \in \mathcal{M}$ as an iuput and operates as follows: $[\ \mathcal{Q} := \mathcal{Q} \cup \{m\};\ \sigma_{\mathrm{SIG}} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(sk_{\mathrm{SIG}}, m);$ return $\sigma_{\mathrm{SIG}}\ ]$. We say that a digital signature scheme Σ_{SIG} is *EUF-CMA secure* if for any PPTA adversary $\mathcal{A}$, $\mathrm{Adv}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SIG}},\mathcal{A}}(k) := \Pr[\mathrm{Exp}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SIG}},\mathcal{A}}(k) = 1]$ is negligible.

D. Symmetric Encryption Schemes

A symmetric encryption scheme Σ_{SE} is defined by the following algorithms $(\mathrm{Gen}_{\mathrm{SE}}, \mathrm{Enc}_{\mathrm{SE}}, \mathrm{Dec}_{\mathrm{SE}})$.

- $\mathrm{Gen}_{\mathrm{SE}}(1^k) \to sk_{\mathrm{SE}}$: The generation algorithm takes the security parameter 1^k as an input, and outputs a symmetric key sk_{SE}.
- $\mathrm{Enc}_{\mathrm{SE}}(sk_{\mathrm{SE}}, m) \to c$: The encryption algorithm takes a symmetric key sk_{SE} and a message $m \in \mathcal{M}$ as inputs, and outputs a ciphertext c.
- $\mathrm{Dec}_{\mathrm{SE}}(sk_{\mathrm{SE}}, c) \to m'$ or $\bot$: The decryption algorithm takes a symmetric key sk_{SE} and a string c as inputs, and outputs a decrypted message $m' \in \mathcal{M}$ or $\bot$. We require that for any $m \in \mathcal{M}$, $k \in \mathbb{N}$, and (sk_{SE}, c) generated by $sk_{\mathrm{SE}} \leftarrow \mathrm{Gen}_{\mathrm{SE}}(1^k)$ and $c \leftarrow \mathrm{Enc}(sk_{\mathrm{SE}}, m)$, it holds that $\mathrm{Dec}(sk_{,\mathrm{SE}}, c) = m$.

In Section IV, we define a specific security property of Σ_{SE} in order to use it as a building block of SRBSs. The property can be considered as a kind of *Key Dependent Message (KDM) security* [15], which is on encryption of messages depending on the secret key[8]. Specifically, KDM security in the single key setting[9] is defined as follows. Consider the

[7] $\mathrm{KeyVer}_{\mathrm{SIG}}$ is not included in the syntax of general digital signature schemes, but we include it because we use it for constructing a SRBS scheme. Most digital signature schemes can construct $\mathrm{KeyVer}_{\mathrm{SIG}}$ satisfying the requirement.

[8] When $\mathcal{F}$ contains all constant functions, $\mathcal{F}$-KDM security implies IND-CPA [16].

[9] In general, KDM security is considered in the multiple key setting. However, the property we define in Section IV is related to KDM security in the single key setting, so we explain the setting here.

following KDM experiment $\mathrm{Exp}^{\mathrm{KDM}}_{\Sigma_{\mathrm{SE}},\mathcal{F},\mathcal{A}}(k)$ for a symmetric encryption scheme $\Sigma_{\mathrm{SE}} = (\mathrm{Gen}_{\mathrm{SE}}, \mathrm{Enc}_{\mathrm{SE}}, \mathrm{Dec}_{\mathrm{SE}})$, a class $\mathcal{F}$ of functions, and an adversary $\mathcal{A}$:

$$\mathrm{Exp}^{\mathrm{KDM}}_{\Sigma_{\mathrm{SE}},\mathcal{F},\mathcal{A}}(k) : [\ b \leftarrow \{0,1\};\ sk_{\mathrm{SE}} \leftarrow \mathrm{Gen}_{\mathrm{SE}}(1^k);\ b' \leftarrow \mathcal{A}^{\tilde{O}_{\mathrm{Enc}_{\mathrm{SE}}}(\cdot)}();\ \text{return } b'\ ],$$

where $\tilde{O}_{\mathrm{Enc}_{\mathrm{SE}}}$ is a encryption oracle that takes a function $f \in \mathcal{F}$ as an input and operates as follows: $[\ m_1 := f(sk_{\mathrm{SE}});\ m_0 := 0^{|f(sk_{\mathrm{SE}})|};\ c \leftarrow \mathrm{Enc}_{\mathrm{SE}}(sk_{\mathrm{SE}}, c_b);\ \text{return } c\ ]$. We say that a symmetric encryption scheme Σ_{SE} is $\mathcal{F}$-KDM secure if for any PPTA adversary $\mathcal{A}$,

$$\mathrm{Adv}^{\mathrm{KDM}}_{\Sigma_{\mathrm{SE}},\mathcal{F},\mathcal{A}}(k) := \big|\Pr[\mathrm{Exp}^{\mathrm{KDM}}_{\Sigma_{\mathrm{SE}},\mathcal{F},\mathcal{A}}(k) = 1 \mid b = 1] - \Pr[\mathrm{Exp}^{\mathrm{KDM}}_{\Sigma_{\mathrm{SE}},\mathcal{F},\mathcal{A}}(k) = 1 \mid b = 0]\big|$$

is negligible.

III. Definitions for SRBSs

In this section we define how biometric data are generated, and the syntax and requirements of SRBSs.

A. Biometric Data Setting

In BSs, a raw biometric sample is an input of key generation and signing algorithms. In SRBSs, a sample is also required for recovery of samples. We define a setting of samples for these algorithms. Let $\mathcal{X}$ be the sample space. Let X be the distribution of a sample for key generation. We assume that when a sample $x_{\mathrm{G}} \leftarrow \mathrm{X}$ is captured from a user for key generation, a sample from the user for signing and recovery is generated according to the probability distribution denoted by $\Delta(x_{\mathrm{G}})$. In general, biometric authentication systems calculate closeness of features generated from samples. To formalize this, let $\mathcal{Y}$ denote the feature space, and let Feat be the feature generation algorithm. A feature y is generated from a sample x by $y \leftarrow \mathrm{Feat}(x)$. Features generated from samples of the genuine user should be close with high probability. This can be formalized with a threshold $t \in \mathbb{R}_{>0}$ and an error parameter $\alpha \in [0,1]$ as follows: $\Pr\{x_{\mathrm{G}} \leftarrow \mathrm{X};\ x' \leftarrow \Delta(x_{\mathrm{G}}) : \mathtt{dist}(\mathrm{Feat}(x_{\mathrm{G}}), \mathrm{Feat}(x')) > t\} \le \alpha$. We call $\mathcal{B} = (\mathcal{X}, \mathrm{X}, \Delta, \mathcal{Y}, \mathrm{Feat}, \mathtt{dist}, t, \alpha)$ satisfying the above conditions a *biometric data setting*.

Remark 1. If the system adopt multi-sample fusion, multiple samples may be required for each algorithm. We can treat this multi-sample setting by modifying the definition of the feature generation algorithm. Specifically, we can treat the case in which a_{G}, a_{S}, and a_{R} samples ($a_{\mathrm{G}}, a_{\mathrm{S}}, a_{\mathrm{R}} \in \mathbb{N}$) are required for key generation, signing, and recovery, respectively, by defining feature extraction algorithms[10] $\mathrm{Feat}_{\mathrm{G}}, \mathrm{Feat}_{\mathrm{S}}, \mathrm{Feat}_{\mathrm{R}}$ such that from samples $x_{\mathrm{G}} \in \mathcal{X}^{a_{\mathrm{G}}}, x_{\mathrm{S}} \in \mathcal{X}^{a_{\mathrm{S}}}, x_{\mathrm{R}} \in \mathcal{X}^{a_{\mathrm{R}}}$ features $y_{\mathrm{G}}, y_{\mathrm{S}}, y_{\mathrm{R}} \in \mathcal{Y}$ are generated by $y_{\mathrm{G}} \leftarrow \mathrm{Feat}_{\mathrm{G}}(x_{\mathrm{G}}), y_{\mathrm{S}} \leftarrow \mathrm{Feat}_{\mathrm{S}}(x_{\mathrm{S}}), y_{\mathrm{R}} \leftarrow \mathrm{Feat}_{\mathrm{R}}(x_{\mathrm{R}})$. Hereafter, we use the single-sample setting because discussions in the multi-sample setting can be done essentially in the same way.

[10] An example of generating a feature y from multiple samples $(x_1, \ldots, x_a)$ is to select the most appropriate one by a specific rule. Another example when $\mathcal{Y} = \{0,1\}^n$ is as follows. For each sample x_i, the algorithm extracts a feature y_i by $y_i = \mathrm{Feat}(x_i)$. Then, for each $j \in [n]$ it determines the j-th bit $y[j]$ of y by 1 if the j-th bit is 1 for more than half of $(y_i)_{i\in[a]}$, and determines $y[i]$ by 0 otherwise. Also, an example when $\mathcal{Y} = \mathbb{R}^n$ is to determine y by the average of $(y_i)_{i\in[a]}$.

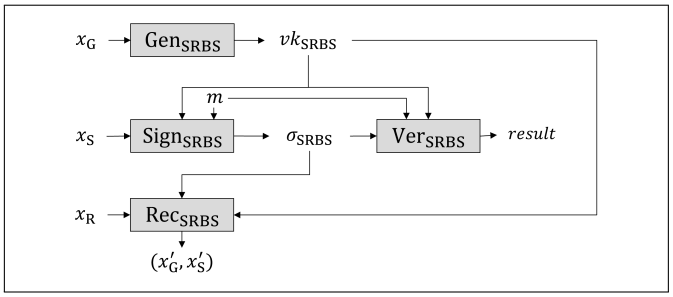


Fig. 1. Data Flow of a SRBS Scheme

B. SRBS Schemes

A SRBS scheme Σ_{SRBS} for a biometric data setting $\mathcal{B}$ is defined by the following algorithms $(\mathrm{Gen}_{\mathrm{SRBS}}, \mathrm{Sign}_{\mathrm{SRBS}}, \mathrm{Ver}_{\mathrm{SRBS}}, \mathrm{Rec}_{\mathrm{SRBS}})$. The data flow is shown in Fig. 1.

- $\mathrm{Gen}_{\mathrm{SRBS}}(1^k, x_{\mathrm{G}}) \to vk_{\mathrm{SRBS}}$: The key generation algorithm takes a sample $x_{\mathrm{G}} \in \mathcal{X}$ as an input, and outputs a verification key vk_{SRBS}.
- $\mathrm{Sign}_{\mathrm{SRBS}}(x_{\mathrm{S}}, vk_{\mathrm{SRBS}}, m) \to \sigma_{\mathrm{SRBS}}$: The signing algorithm takes a sample $x_{\mathrm{S}} \in \mathcal{X}$, a verification key vk_{SRBS},[11] and a message $m \in \mathcal{M}$ as inputs, and outputs a signature σ_{SRBS} The signature σ_{SRBS} may be $\perp$.
- $\mathrm{Ver}_{\mathrm{SRBS}}(vk_{\mathrm{SRBS}}, m, \sigma_{\mathrm{SRBS}}) \to result$: algorithm takes a verification key vk_{SRBS}, a message $m \in \mathcal{M}$, a signature σ_{SRBS} as inputs, and outputs the verification result $result \in \{\top, \perp\}$.
- $\mathrm{Rec}_{\mathrm{SRBS}}(x_{\mathrm{R}}, vk_{\mathrm{SRBS}}, \sigma_{\mathrm{SRBS}}) \to (x'_{\mathrm{G}}, x'_{\mathrm{S}})$ or $\perp$: algorithm takes a sample $x_{\mathrm{R}} \in \mathcal{X}$, a verification key vk_{SRBS}, and a signature σ_{SRBS} as inputs, and outputs recovered samples $(x'_{\mathrm{G}}, x'_{\mathrm{S}}) \in \mathcal{X} \times \mathcal{X}$ or $\perp$.

We require correctness defined as follows: for any $x_{\mathrm{G}}, x_{\mathrm{S}}, x_{\mathrm{R}} \in \mathcal{X}$ satisfying $\mathtt{dist}(\mathrm{Feat}(x_{\mathrm{G}}), \mathrm{Feat}(x_{\mathrm{S}})) \le t$ and $\mathtt{dist}(\mathrm{Feat}(x_{\mathrm{G}}), \mathrm{Feat}(x_{\mathrm{R}})) \le t$, $m \in \mathcal{M}$, and $(vk_{\mathrm{SRBS}}, \sigma_{\mathrm{SRBS}})$ generated by $vk_{\mathrm{SRBS}} \leftarrow \mathrm{Gen}_{\mathrm{SRBS}}(1^k, x_{\mathrm{G}})$ and $\sigma_{\mathrm{SRBS}} \leftarrow \mathrm{Sign}_{\mathrm{SRBS}}(x_{\mathrm{S}}, vk_{\mathrm{SRBS}}, m)$, it holds that $\mathrm{Ver}_{\mathrm{SRBS}}(vk_{\mathrm{SRBS}}, m, \sigma_{\mathrm{SRBS}}) = \top$ and $\mathrm{Rec}_{\mathrm{SRBS}}(x_{\mathrm{R}}, vk_{\mathrm{SRBS}}, \sigma_{\mathrm{SRBS}}) = (x_{\mathrm{G}}, x_{\mathrm{S}})$.

Next, we define EUF-CMA security for Σ_{SRBS}. Consider the following EUF-CMA experiment $\mathrm{Exp}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SRBS}},\mathcal{A}}(k)$ for a SRBS scheme $\Sigma_{\mathrm{SRBS}} = (\mathrm{Gen}_{\mathrm{SRBS}}, \mathrm{Sign}_{\mathrm{SRBS}}, \mathrm{Ver}_{\mathrm{SRBS}},$

[11] As Section I, some biometric schemes called FSs generate a signature without using the verification key. However, it is natural in applications such as online biometric authentication systems that the device for signing is connected to the network and can access the verification key. Therefore, in our definition, the inputs of the signing algorithm includes the verification key.

$\mathrm{Rec}_{\mathrm{SRBS}}$) and an adversary $\mathcal{A}$:

$$\begin{aligned}&\mathrm{Exp}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SRBS}},\mathcal{A}}(k):\\&[\ x_{\mathrm{G}} \leftarrow \mathrm{X};\ vk_{\mathrm{SRBS}} \leftarrow \mathrm{Gen}_{\mathrm{SRBS}}(1^k, x_{\mathrm{G}});\\&\ \ \mathcal{Q} := \emptyset;\ (m', \sigma'_{\mathrm{SRBS}}) \leftarrow \mathcal{A}^{\mathcal{O}_{\mathrm{Sign}_{\mathrm{SRBS}}}(\cdot)}(vk_{\mathrm{SRBS}});\\&\ \ \text{if } m' \notin \mathcal{Q} \wedge \mathrm{Ver}(vk_{\mathrm{SRBS}}, m', \sigma'_{\mathrm{SRBS}}) = \top\\&\ \ \text{then return } 1 \text{ else return } 0\ ],\end{aligned}$$

where $\mathcal{O}_{\mathrm{Sign}_{\mathrm{SRBS}}}$ is a signing oracle that takes a message $m \in \mathcal{M}$ as an input and operates as follows: [$\mathcal{Q} := \mathcal{Q} \cup \{m\}$; $x_{\mathrm{S}} \leftarrow \Delta(x_{\mathrm{G}})$; $\sigma_{\mathrm{SRBS}} \leftarrow \mathrm{Sign}_{\mathrm{SRBS}}(x_{\mathrm{S}}, vk_{\mathrm{SRBS}}, m)$; return σ_{SRBS}].

We say that a SRBS scheme Σ_{SRBS} is EUF-CMA secure if for any PPTA adversary $\mathcal{A}$, $\mathrm{Adv}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SRBS}},\mathcal{A}}(k) := \Pr[\mathrm{Exp}^{\text{EUF-CMA}}_{\Sigma_{\mathrm{SRBS}},\mathcal{A}}(k) = 1]$ is negligible.

Remark 2. SRBS schemes with EUF-CMA security satisfy ASR by storing vk_{SRBS} and σ_{SRBS}. Indeed, by correctness, a sample x_{G} having captured for $\mathrm{Gen}_{\mathrm{SRBS}}$ and x_{S} for $\mathrm{Sign}_{\mathrm{SRBS}}$ satisfying $\mathtt{dist}(\mathrm{Feat}(x_{\mathrm{G}}), \mathrm{Feat}(x_{\mathrm{S}})) \le t$ can be recovered from vk_{SRBS}, σ_{SRBS}, and a feature y satisfying $\mathtt{dist}(\mathrm{Feat}(x_{\mathrm{G}}), y) \le t$. Also, by EUF-CMA security, it is difficult to recover y satisfying $\mathtt{dist}(\mathrm{Feat}(x_{\mathrm{G}}), y) \le t$ from the stored vk_{SRBS} and σ_{SRBS}.

IV. Generic Construction

In this section, we give a generic construction of a SRBS scheme $\Sigma_{\mathrm{SRBS}} = (\mathrm{Gen}_{\mathrm{SRBS}}, \mathrm{Sign}_{\mathrm{SRBS}}, \mathrm{Ver}_{\mathrm{SRBS}}, \mathrm{Rec}_{\mathrm{SRBS}})$ for a biometric data setting $\mathcal{B} = (\mathcal{X}, \mathrm{X}, \Delta, \mathcal{Y}, \mathrm{Feat}, \mathtt{dist}, t, \alpha)$. For construction, we use a $(\mathcal{Y}, \mu, \ell, t, \epsilon)$-fuzzy extractor $\Sigma_{\mathrm{FE}} = (\mathrm{Gen}_{\mathrm{FE}}, \mathrm{Rep}_{\mathrm{FE}})$, a digital signature scheme $\Sigma_{\mathrm{SIG}} = (\mathrm{Gen}_{\mathrm{SIG}}, \mathrm{Sign}_{\mathrm{SIG}}, \mathrm{Ver}_{\mathrm{SIG}}, \mathrm{KeyVer}_{\mathrm{SIG}})$, and a symmetric encryption scheme $\Sigma_{\mathrm{SE}} = (\mathrm{Gen}_{\mathrm{SE}}, \mathrm{Enc}_{\mathrm{SE}}, \mathrm{Dec}_{\mathrm{SE}})$ as building blocks.

In order to describe requirements of the primitives, we define the *Biometric Encryption (BE) security* as follows. We assume that all samples $x \in \mathcal{X}$ are represented by the same bit-length $|x|$. Consider the following BE experiment $\mathrm{Exp}^{\text{BE}}_{\mathcal{B},\Sigma_{\mathrm{FE}},\Sigma_{\mathrm{SE}},\mathcal{A}}(k)$ for $\mathcal{B}, \Sigma_{\mathrm{FE}}, \Sigma_{\mathrm{SE}}$, and an adversary $\mathcal{A}$:

$$\begin{aligned}&\mathrm{Exp}^{\text{BE}}_{\mathcal{B},\Sigma_{\mathrm{FE}},\Sigma_{\mathrm{SE}},\mathcal{A}}(k):\\&\ \ [\ b \leftarrow \{0,1\};\ x_{\mathrm{G}} \leftarrow \mathrm{X};\ y_{\mathrm{G}} \leftarrow \mathrm{Feat}(x_{\mathrm{G}});\\&\ \ \ \ (r, p) \leftarrow \mathrm{Gen}_{\mathrm{FE}}(y_{\mathrm{G}});\\&\ \ \ \ c_1 \leftarrow \mathrm{Enc}_{\mathrm{FE}}(r, x_{\mathrm{G}});\ c_0 \leftarrow \mathrm{Enc}_{\mathrm{FE}}(r, 0^{|x_{\mathrm{G}}|});\\&\ \ \ \ b' \leftarrow \mathcal{A}^{\tilde{\mathcal{O}}_{\mathrm{Enc}_{\mathrm{SE}}}(\cdot), \tilde{\mathcal{O}}'_{\mathrm{Enc}_{\mathrm{SE}}}(\cdot)}(p, c_b);\ \text{return } b'\ ],\end{aligned}$$

where $\tilde{\mathcal{O}}_{\mathrm{Enc}_{\mathrm{SE}}}$ and $\tilde{\mathcal{O}}'_{\mathrm{Enc}_{\mathrm{SE}}}$ are defined as follows.

- The oracle $\tilde{\mathcal{O}}_{\mathrm{Enc}_{\mathrm{SE}}}$ takes a message m as an input, and operates as follows: [$m_1 := m$; $m_0 := 0^{|m|}$; $c \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, m_b)$; return c].
- The oracle $\tilde{\mathcal{O}}'_{\mathrm{Enc}_{\mathrm{SE}}}$ takes no input, and operates as follows: [$x_{\mathrm{S}} \leftarrow \Delta_{\mathrm{S}}(x_{\mathrm{G}})$; $y_{\mathrm{S}} \leftarrow \mathrm{Feat}(x_{\mathrm{S}})$; $r' \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{S}}, p)$; if $r' \neq r$ then return $\perp$; $x_1 := x_{\mathrm{S}}$; $x_0 := 0^{|x_{\mathrm{S}}|}$; $c \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, x_b)$; return c].

We say that a symmetric encryption scheme Σ_{SE} is *BE secure* with respect to a biometric data setting $\mathcal{B}$ and a fuzzy extractor scheme Σ_{FE} if for any PPTA adversary $\mathcal{A}$,

$$\begin{aligned}\mathrm{Adv}^{\text{BE}}_{\mathcal{B},\Sigma_{\mathrm{FE}},\Sigma_{\mathrm{SE}},\mathcal{A}}(k) := \big|\ &\Pr[\mathrm{Exp}^{\text{BE}}_{\mathcal{B},\Sigma_{\mathrm{FE}},\Sigma_{\mathrm{SE}},\mathcal{A}}(k) = 1 \mid b = 1]\\&- \Pr[\mathrm{Exp}^{\text{BE}}_{\mathcal{B},\Sigma_{\mathrm{FE}},\Sigma_{\mathrm{SE}},\mathcal{A}}(k) = 1 \mid b = 0]\big|\end{aligned}$$

is negligible.

We require the following for the building blocks:

1) The $(\mathcal{Y}, \mu, \ell, t, \epsilon)$-fuzzy extractor Σ_{FE} satisfies $\ell = \Theta(k)$ and $\mu = \mathbf{H}_\infty(Y_{\mathrm{G}})$, where X_{G} denotes a random variable distributed according to X, and Y_{G} denotes a random variable defined by $Y_{\mathrm{G}} := \mathrm{Feat}(X_{\mathrm{G}})$.
2) The digital signature scheme Σ_{SIG} is EUF-CMA secure.
3) The symmetric encryption scheme Σ_{SE}
   a) is BE secure with respect to $\mathcal{B}$ and Σ_{FE},
   b) has the key space $\{0,1\}^\ell$, and
   c) satisfies that for any distinct $sk_{\mathrm{SE}}, sk'_{\mathrm{SE}} \in \{0,1\}^\ell$ and any m, $\Pr[c \leftarrow \mathrm{Enc}_{\mathrm{SE}}(sk_{\mathrm{SE}}, m);\ m' \leftarrow \mathrm{Dec}_{\mathrm{SE}}(sk'_{\mathrm{SE}}, c) : m' = m]$ is negligible.

Also, we assume that all signing keys sk_{SIG} generated by $(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k)$ are represented by the same bit-length $|sk_{\mathrm{SIG}}|$.

Construction: Let k' be the security parameter for Σ_{SE} when the key space is $\{0,1\}^\ell$. We note that $k' = \Theta(k)$ follows from $\ell = \Theta(k)$. The construction is provided in Fig. 2. For any algorithm, if it fails a to operate "parse", then it returns $\perp$.

Theorem 1. $\Sigma_{\mathrm{SRBS}} = (\mathrm{Gen}_{\mathrm{SRBS}}, \mathrm{Sign}_{\mathrm{SRBS}}, \mathrm{Ver}_{\mathrm{SRBS}}, \mathrm{Rec}_{\mathrm{SRBS}})$ defined by the above construction is a SRBS scheme with EUF-CMA security.

Proof: It is easy to show correctness. We show EUF-CMA security. Let $\mathcal{A}$ be any PPTA adversary against Σ_{SRBS} that makes at most q signing queries. Let E_i be the event that $\mathcal{A}$ wins in Game_i.

$\underline{\mathrm{Game}_1}$: We define Game_1 as the actual EUF-CMA game for Σ_{SRBS}. In this game, the verification key vk_{SRBS} is generated as follows:

$$\begin{aligned}&[\ x_{\mathrm{G}} \leftarrow \mathrm{X};\ y_{\mathrm{G}} \leftarrow \mathrm{Feat}(x_{\mathrm{G}});\ (r, p) \leftarrow \mathrm{Gen}_{\mathrm{FE}}(y_{\mathrm{G}});\\&\ \ (sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k);\\&\ \ c_{\mathrm{G}} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, x_{\mathrm{G}});\ z \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, sk_{\mathrm{SIG}});\\&\ \ vk_{\mathrm{SRBS}} := (vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)\ ].\end{aligned}$$

Furthermore, for $i \in [q]$, when $\mathcal{A}$ makes the i-th signing query on a message m_i, the challenger generates a signature $\sigma_{\mathrm{SRBS},i}$ as follows:

$$\begin{aligned}&[\ y_{\mathrm{S},i} \leftarrow \mathrm{Feat}(x_{\mathrm{S}});\ \text{parse } vk_{\mathrm{SRBS}} \text{ as } (vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z);\\&\ \ r'_i \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{S},i}, p);\ sk'_{\mathrm{SIG},i} \leftarrow \mathrm{Dec}_{\mathrm{SE}}(r'_i, z);\\&\ \ \text{if } \mathrm{KeyVer}_{\mathrm{SIG}}(sk'_{\mathrm{SIG},i}, vk_{\mathrm{SIG}}) = \perp \text{ then return } \perp;\\&\ \ \sigma_{\mathrm{SIG},i} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(sk'_{\mathrm{SIG},i}, m_i);\ c_{\mathrm{S},i} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r'_i, x_{\mathrm{S},i});\\&\ \ \sigma_{\mathrm{SRBS},i} := (\sigma_{\mathrm{SIG},i}, c_{\mathrm{S},i})\ ].\end{aligned}$$

$\underline{\mathrm{Game}_2}$: In this game, the challenger generates $\sigma_{\mathrm{SRBS},i}$ as

$\mathrm{Gen}_{\mathrm{SRBS}}(1^k, x_{\mathrm{G}})$:
$y_{\mathrm{G}} \leftarrow \mathrm{Feat}(x_{\mathrm{G}})$;
$(r, p) \leftarrow \mathrm{Gen}_{\mathrm{FE}}(y_{\mathrm{G}})$;
$(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k)$;
$c_{\mathrm{G}} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, x_{\mathrm{G}})$;
$z \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, sk_{\mathrm{SIG}})$;
$vk_{\mathrm{SRBS}} := (vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
return vk_{SRBS}.

$\mathrm{Sign}_{\mathrm{SRBS}}(x_{\mathrm{S}}, vk_{\mathrm{SRBS}}, m)$:
$y_{\mathrm{S}} \leftarrow \mathrm{Feat}(x_{\mathrm{S}})$;
parse vk_{SRBS} as $(vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
$r' \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{S}}, p)$;
$sk'_{\mathrm{SIG}} \leftarrow \mathrm{Dec}_{\mathrm{SE}}(r', z)$;
if $\mathrm{KeyVer}_{\mathrm{SIG}}(sk'_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) = \perp$ then return $\perp$;
$\sigma_{\mathrm{SIG}} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(sk'_{\mathrm{SIG}}, m)$;
$c_{\mathrm{S}} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r', x_{\mathrm{S}})$;
$\sigma_{\mathrm{SRBS}} := (\sigma_{\mathrm{SIG}}, c_{\mathrm{S}})$;
return σ_{SRBS}.

$\mathrm{Ver}_{\mathrm{SRBS}}(vk_{\mathrm{SRBS}}, m, \sigma_{\mathrm{SRBS}})$:
if $\sigma_{\mathrm{SRBS}} = \perp$ then return $\perp$;
parse vk_{SRBS} as $(vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
parse σ_{SRBS} as $(\sigma_{\mathrm{SIG}}, c_{\mathrm{S}})$;
result $\leftarrow \mathrm{Ver}_{\mathrm{SIG}}(vk_{\mathrm{SIG}}, m, \sigma_{\mathrm{SIG}})$;
return *result*.

$\mathrm{Rec}_{\mathrm{SRBS}}(x_{\mathrm{R}}, vk_{\mathrm{SRBS}}, \sigma_{\mathrm{SRBS}})$:
$y_{\mathrm{R}} \leftarrow \mathrm{Feat}(x_{\mathrm{R}})$;
parse vk_{SRBS} as $(vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
parse σ_{SRBS} as $(\sigma_{\mathrm{SIG}}, c_{\mathrm{S}})$;
$r'' \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{R}}, p)$;
$x'_{\mathrm{G}} \leftarrow \mathrm{Dec}_{\mathrm{SE}}(r'', c_{\mathrm{G}})$;
$x'_{\mathrm{S}} \leftarrow \mathrm{Dec}_{\mathrm{SE}}(r'', c_{\mathrm{S}})$;
return $(x'_{\mathrm{G}}, x'_{\mathrm{S}})$.

Fig. 2. Generic construction of a SRBS scheme.

follows (where the difference is underlined):

[$y_{\mathrm{S},i} \leftarrow \mathrm{Feat}(x_{\mathrm{S}})$; parse vk_{SRBS} as $(vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
$r'_i \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{S},i}, p)$; $sk'_{\mathrm{SIG},i} \leftarrow \mathrm{Dec}_{\mathrm{SE}}(r'_i, z)$;
if $\underline{sk'_{\mathrm{SIG},i} \neq sk_{\mathrm{SIG}}}$ then return $\perp$;
$\sigma_{\mathrm{SIG},i} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(\underline{sk_{\mathrm{SIG}}}, m_i)$; $c_{\mathrm{S},i} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r'_i, x_{\mathrm{S},i})$;
$\sigma_{\mathrm{SRBS},i} := (\sigma_{\mathrm{SIG},i}, c_{\mathrm{S},i})$].

Because $\mathrm{KeyVer}_{\mathrm{SIG}}(sk'_{\mathrm{SIG},i}, vk_{\mathrm{SIG}}) = \perp$ is equivalent to $sk'_{\mathrm{SIG},i} \neq sk_{\mathrm{SIG}}$, $\Pr[\mathrm{E}_1] = \Pr[\mathrm{E}_2]$ holds.

$\underline{\mathrm{Game}_3}$: In this game, the challenger generates $\sigma_{\mathrm{SRBS},i}$ as follows:

[$y_{\mathrm{S},i} \leftarrow \mathrm{Feat}(x_{\mathrm{S}})$; parse vk_{SRBS} as $(vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
$r'_i \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{S},i}, p)$; $sk'_{\mathrm{SIG},i} \leftarrow \mathrm{Dec}_{\mathrm{SE}}(r'_i, z)$;
if $\underline{r'_i \neq r}$ then return $\perp$;
$\sigma_{\mathrm{SIG},i} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(sk_{\mathrm{SIG}}, m_i)$; $c_{\mathrm{S},i} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(\underline{r}, x_{\mathrm{S},i})$;
$\sigma_{\mathrm{SRBS},i} := (\sigma_{\mathrm{SIG},i}, c_{\mathrm{S},i})$].

Because $\Pr[(sk'_{\mathrm{SIG},i} = sk_{\mathrm{SIG}}) \wedge (r'_i \neq r)]$ is negligible, $\left|\Pr[\mathrm{E}_2] - \Pr[\mathrm{E}_3]\right|$ is negligible.

$\underline{\mathrm{Game}_4}$: In this game, the verification key vk_{SRBS} is generated as follows:

[$x_{\mathrm{G}} \leftarrow \mathrm{X}$; $y_{\mathrm{G}} \leftarrow \mathrm{Feat}(x_{\mathrm{G}})$; $(r, p) \leftarrow \mathrm{Gen}_{\mathrm{FE}}(y_{\mathrm{G}})$;
$(sk_{\mathrm{SIG}}, vk_{\mathrm{SIG}}) \leftarrow \mathrm{Gen}_{\mathrm{SIG}}(1^k)$;
$c_{\mathrm{G}} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, \underline{0^{|x_{\mathrm{G}}|}})$; $z \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, \underline{0^{|sk_{\mathrm{SIG}}|}})$;
$vk_{\mathrm{SRBS}} := (vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$].

Furthermore, the challenger generates $\sigma_{\mathrm{SRBS},i}$ as follows:

[$y_{\mathrm{S},i} \leftarrow \mathrm{Feat}(x_{\mathrm{S}})$; parse vk_{SRBS} as $(vk_{\mathrm{SIG}}, c_{\mathrm{G}}, p, z)$;
$r'_i \leftarrow \mathrm{Rep}_{\mathrm{FE}}(y_{\mathrm{S},i}, p)$; $__$;
if $r'_i \neq r$ then return $\perp$;
$\sigma_{\mathrm{SIG},i} \leftarrow \mathrm{Sign}_{\mathrm{SIG}}(sk_{\mathrm{SIG}}, m_i)$; $c_{\mathrm{S},i} \leftarrow \mathrm{Enc}_{\mathrm{SE}}(r, \underline{0^{|x_{\mathrm{S},i}|}})$;
$\sigma_{\mathrm{SRBS},i} := (\sigma_{\mathrm{SIG},i}, c_{\mathrm{S},i})$].

Because Σ_{SE} is BE secure and the key space is $\{0,1\}^\ell$ with $\ell = \Theta(k)$, $|\Pr[\mathrm{E}_3] - \Pr[\mathrm{E}_4]|$ is negligible. In addition, in order for $\mathcal{A}$ to win Game_4, $\mathcal{A}$ has to find a pair $(m', \sigma'_{\mathrm{SIG}})$ satisfying $m' \notin \{m_i\}_{i\in[q]}$ and $\mathrm{Ver}_{\mathrm{SIG}}(m', \sigma'_{\mathrm{SIG}})$. For finding such a $(m', \sigma'_{\mathrm{SIG}})$, the meaningful information contained in $(vk_{\mathrm{SRBS}}, \{(m_i, \sigma_{\mathrm{SRBS},i})\}_{i\in[q]})$ is $(vk_{\mathrm{SIG}}, \{(m_i, \sigma_{\mathrm{SIG},i}\}_{i\in[q]})$. Hence, it follows from EUF-CMA security of Σ_{SIG} and $k' = \Theta(k)$ that $\Pr[\mathrm{E}_4]$ is negligible.

Therefore, the advantage $\Pr[\mathrm{E}_1]$ for the original EUF-CMA game, upper-bounded as $\Pr[\mathrm{E}_1] \leq \sum_{i=1}^{3} |\Pr[\mathrm{E}_i] - \Pr[\mathrm{E}_{i+1}]| + \Pr[\mathrm{E}_4]$, is negligible. ■

V. Instantiation

In this section, we show that we can construct a SRBS scheme with EUF-CMA security in the random oracle model by giving an example of the building blocks $\Sigma_{\mathrm{FE}}, \Sigma_{\mathrm{SIG}}, \Sigma_{\mathrm{SE}}$ which satisfies the requirements in the generic construction.

We consider Σ_{SRBS} constructed by Section IV with the following building blocks. We use an $(\mathcal{Y}, \mu, \ell, t, \epsilon)$-fuzzy extractor Σ_{FE} such that $\ell = \Theta(k)$, $\mu = \mathbf{H}_\infty(Y_{\mathrm{G}})$, and ϵ is negligible. Such a scheme can be constructed if $\mathbf{H}_\infty(Y_{\mathrm{G}}) = \Theta(k)$ holds[12]. For example, Dodis et al. [7] showed that

[12]Considering keeping FNMR small, $\tilde{\mathbf{H}}_\infty(Y_{\mathrm{G}} \mid \mathrm{SS}(Y_{\mathrm{G}})) = \Theta(k)$ is required instead of $\mu = \mathbf{H}_\infty(Y_{\mathrm{G}})$. For example, in the case of a $[n, k, 2t+1]$ binary error-correcting code, $\mathbf{H}_\infty(Y_{\mathrm{G}}) + k - n = \Theta(k)$ is required.

for any $[n, k, 2t+1]$ binary error-correcting code and any μ, ϵ, one can construct an $(\mathcal{Y}, \mu, \ell, t, \epsilon)$-fuzzy extractor with $\ell = \mathbf{H}_\infty(Y_{\rm G}) + k - n - 2\log(1/\epsilon)$. By setting $\epsilon = 2^{(\log k)^2}$, it holds that ϵ is negligible and $\ell = \Theta(k)$. Also, as $\Sigma_{\rm SIG}$, we can use any digital signature scheme with EUF-CMA security, e.g., schemes in Ref. [17][18]. Furthermore, as $\Sigma_{\rm SE}$, we use a symmetric encryption scheme by Black et al. [15][13] with the key space $\{0,1\}^\ell$. It can be described as follows. Let $\mathcal{H}$ be the random oracle, and let $\mathcal{H}_n(x)$ be the first n bits of $\mathcal{H}(x)$. Then, $\Sigma_{\rm SE} = ({\rm Gen}_{\rm SE}, {\rm Enc}_{\rm SE}, {\rm Dec}_{\rm SE})$ is defined as follows:

- ${\rm Gen}_{\rm SE}(1^\ell) : sk_{\rm SE} \leftarrow \{0,1\}^\ell$; return $sk_{\rm SE}$.
- ${\rm Enc}_{\rm SE}(sk_{\rm SE}, m) : u \leftarrow \{0,1\}^\ell$; $c \leftarrow (u, \mathcal{H}_{|m|}(sk_{\rm SE} \parallel u) \oplus m)$; return c.
- ${\rm Dec}_{\rm SE}(sk_{\rm SE}, c)$: parse c as (u, b); $m \leftarrow \mathcal{H}_{|b|}(sk_{\rm SE} \| u) \oplus b$; return m.

Theorem 2. $\Sigma_{\rm SRBS}$ defined by the above construction is a SRBS scheme with EUF-CMA security in the random oracle model under the assumption necessary for EUF-CMA security of $\Sigma_{\rm SIG}$.

Proof: It is sufficient to show that the requirements 3(a) and 3(c) in Section IV because the other requirements are satisfied by the definition of the construction. 3(c) holds because ${\rm Dec}(sk', {\rm Enc}(sk_{\rm SE}, m)) = m$ implies $\mathcal{H}_{|m|}(sk_{\rm SE} \| u) = \mathcal{H}_{|m|}(sk'_{\rm SE} \parallel u)$ for $u \leftarrow \{0,1\}^\ell$, the probability of which is negligible.

We show 3(a), i.e., the BE security. In the BE experiment, consider replacing r with $\tilde{r}$, where $\tilde{r}$ satisfies $(r, p) = (\tilde{r}, p)$ with probability not less than $1 - \epsilon$, and $\tilde{r}$ is uniformly distributed and statistically independent of p. We can set $\tilde{r}$ in this way because of the requirement of $\Sigma_{\rm FE}$. The advantage difference is ϵ, which is negligible. Furthermore, in the modified BE experiment, the attacker cannot identify the ciphertexts $({\rm Enc}_{\rm SE}(\tilde{r}, x_{\rm G}), {\rm Enc}_{\rm SE}(\tilde{r}, x_{\rm S}), {\rm Enc}_{\rm SE}(\tilde{r}, sk_{\rm SIG}))$ and the fake ones, because both ciphertexts are uniformly distributed. Therefore, the BE security holds, which concludes the proof. ■

Remark 3. Theorem 2 holds in the random oracle model because $\Sigma_{\rm SE}$ relies on a random oracle. On the other hand, $\mathcal{F}$-KDM security without resorting to the use of random oracles is studied [19][20][21], which aims at enlarging the function class $\mathcal{F}$ or deriving a bound on $\mathcal{F}$. It is future work to construct a SRBS scheme which satisfies EUF-CMA security without random oracles.

VI. Application to Online Biometric Authentication

In this section, we show that by using SRBS schemes with EUF-CMA security we can construct an online biometric authentication system with ASR.

We consider an online biometric authentication system which consists of client devices and an authentication server, and assume that the server follows the protocol but data that appears in the server may leak. Also, we assume that the risk of leaking data that appears temporarily in client devices during procedures for legitimate users are sufficiently low. Under these assumptions, by using an SRBS scheme $\Sigma_{\rm SRBS} = ({\rm Gen}_{\rm SRBS}, {\rm Sign}_{\rm SRBS}, {\rm Ver}_{\rm SRBS}, {\rm Rec}_{\rm SRBS})$ with EUF-CMA security, we can construct an online biometric authentication system with ASR as follows.

During the enrollment procedure, An client device creates a user ID, captures a sample $x_{\rm G} \leftarrow {\rm X}$, and generates $vk_{\rm SRBS}$ by $vk_{\rm SRBS} \leftarrow {\rm Gen}_{\rm SRBS}(1^k, x_{\rm G})$. Then, the user ID and $vk_{\rm SRBS}$ are sent to and stored in the server.

During the authentication procedure, an client device gets the user ID, and sends it to the server. The server sends a fresh challenge m and $vk_{\rm SRBS}$ corresponding to the user ID to the client device. Then, the client device captures a sample $x_{\rm S} \leftarrow \Delta(x_{\rm G})$ from the user, generates $\sigma_{\rm SRBS}$ by $\sigma_{\rm SRBS} \leftarrow {\rm Sign}_{\rm SRBS}(x_{\rm S}, vk_{\rm SRBS}, m)$, and sends $\sigma_{\rm SRBS}$ to the server. The server determines the authentication result *result* by *result* $\leftarrow {\rm Ver}_{\rm SRBS}(vk_{\rm SRBS}, m, \sigma_{\rm SRBS})$. If *result* $= \top$, then the server stores $\sigma_{\rm SRBS}$.

During the recovery procedure, an client device gets the user ID, and sends it to the server. The server sends $vk_{\rm SRBS}$ and $\sigma_{\rm SRBS}$ corresponding to the user ID to the client device. Then, the client device captures a sample $x_{\rm R} \leftarrow \Delta(x_{\rm G})$ from the user, and recovers samples $(x'_{\rm G}, x'_{\rm S})$ by $(x'_{\rm G}, x'_{\rm S}) \leftarrow {\rm Rec}_{\rm SRBS}(x_{\rm R}, vk_{\rm SRBS}, \sigma_{\rm SRBS})$.

The above system satisfies ASR as follows. If the captured samples $(x_{\rm G}, x_{\rm S}, x_{\rm R})$ satisfy $\mathtt{dist}({\rm Feat}(x_{\rm G}), {\rm Feat}(x_{\rm S})) \le t$ and $\mathtt{dist}({\rm Feat}(x_{\rm G}), {\rm Feat}(x_{\rm R})) \le t$, then $(x_{\rm G}, x_{\rm S})$ can be correctly recovered by the recovery procedure. Also, EUF-CMA security of a SRBS scheme implies that it is difficult to recover the past samples or features only from the stored data $(vk_{\rm SRBS}, \sigma_{\rm SRBS})$, because the attacker who can recover the feature from $(vk_{\rm SRBS}, \sigma_{\rm SRBS})$ can forge a signature.

We further discuss biometric reference update. If the system executes the update procedure during an authentication procedure, the feature generated for authentication can be used for the recovery, so the user does not have to input additional biometric samples for the update, which keeps user convenience[14]. For example, consider the case in which the system requires $n_{\rm G}$ and $n_{\rm S}$ samples[15] for biometric reference generation and authentication, respectively, and assume that $n_{\rm G} > n_{\rm S}$ holds[16]. Then, when the system updates a biomeric reference of a user for whom the system performed q successful authentication procedures, the system can use $n_{\rm G}$ samples during the enrollment procedure, $n_{\rm S}$ samples during each of the past q authentication procedures, and $n_{\rm S}$ samples during the authentication procedure with biomeric reference update.

[13]The scheme in Ref. [15] is $\mathcal{F}^{\rm all}$-KDM secure in the random oracle model, where $\mathcal{F}^{\rm all}$ consists of all functions such that the bit-length $|f(sk_{\rm SE})|$ is independent of $sk_{\rm SE}$ and the random oracle accessed in f.

[14]An alternative for applying multi-sample fusion to generating a new biometric reference might be to capture multiple samples during the update procedure, but it is inconvenient for users.

[15]Examples of a fusion scheme is described in Remark 1 in Section III-A.

[16]The authentication procedures are performed more than the enrollment ones, so it is convenient for users if they are required to input less samples during the authentication.

Therefore, in order to generate a new biometric reference, the system can use n_{G} samples chosen from the $(n_{\mathrm{G}}+(q+1)n_{\mathrm{S}})$ samples. This is a merit of the proposed system, because systems without ASR have only n_{S} samples during the authentication procedure with update, which is insufficient for generating a new biometric reference when n_{G} samples are required.

We note that in the above update scheme a new biometric reference vk_{SRBS} is enrolled to the server. In order to make the system secure even when both of an old and a new vk_{SRBS} may leak, it suffices to use a reusable FE [8], which has the property that the randomness r does not leak even if multiple helper strings leak.

REFERENCES

[1] FIDO Alliance, *How FIDO works*, https://fidoalliance.org/how-fido-works/(Accessed: 2022/03/21).

[2] G. of India, *What is aadhaar*, https://uidai.gov.in/my-aadhaar/about-your-aadhaar.html(Accessed: 2022/03/21), 2019.

[3] T. Kawakami and Y. Hinata, *Pay with your face: 100m chinese switch from smartphones*, https://asia.nikkei.com/Business/China-tech/Pay-with-your-face-100m-Chinese-switch-from-smartphones(Accessed: 2022/03/21), 2019.

[4] V. Doshi, *A security breach in india has left a billion people at risk of identity theft*, https://www.washingtonpost.com/news/worldviews/wp/2018/01/04/a-security-breach-in-india-has-left-a-billion-people-at-risk-of-identity-theft/(Accessed: 2022/03/21), 2018.

[5] ISO, *ISO/IEC 24745:2011 information technology — security techniques — biometric information protection*, 2011.

[6] ——, *ISO/IEC 30136:2018 information technology — performance testing of biometric template protection schemes*, 2018.

[7] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," *SIAM Journal on Computing*, vol. 38, no. 1, pp. 97–139, 2008.

[8] X. Boyen, "Reusable cryptographic fuzzy extractors," in *Proceedings of the 11th ACM Conference on Computer and Communications Security (CCS2004)*, 2004, pp. 82–91.

[9] K. Takahashi, T. Matsuda, T. Murakami, G. Hanaoka, and M. Nishigaki, "Signature schemes with a fuzzy private key," *International Journal of Information Security*, vol. 18, no. 5, pp. 581–617, 2019.

[10] S. Katsumata, T. Matsuda, W. Nakamura, K. Ohara, and K. Takahashi, "Revisiting fuzzy signatures: Towards a more risk-free cryptographic authentication system based on biometrics," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS2021)*, 2021, pp. 2046–2065.

[11] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019, vol. 2.

[12] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2d+ 3d face biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 619–624, 2005.

[13] C. Rathgeb, T. Schlett, N. Buchmann, H. Baier, and C. Busch, "Multi-sample compression of iris images using high efficiency video coding," in *2018 International Conference on Biometrics (ICB2018)*, IEEE, 2018, pp. 291–296.

[14] U. Uludag, A. Ross, and A. Jain, "Biometric template selection and update: A case study in fingerprints," *Pattern Recognition*, vol. 37, no. 7, pp. 1533–1542, 2004.

[15] J. Black, P. Rogaway, and T. Shrimpton, "Encryption-scheme security in the presence of key-dependent messages," in *International Workshop on Selected Areas in Cryptography (SAC2002)*, Springer, 2002, pp. 62–75.

[16] T. Malkin, I. Teranishi, and M. Yung, "Key dependent message security: Recent results and applications," in *Proceedings of the First ACM Conference on Data and Application Security and Privacy (CODASPY2011)*, 2011, pp. 3–12.

[17] C.-P. Schnorr, "Efficient signature generation by smart cards," *Journal of Cryptology*, vol. 4, no. 3, pp. 161–174, 1991.

[18] D. Boneh and X. Boyen, "Short signatures without random oracles," in *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT2004)*, Springer, 2004, pp. 56–73.

[19] D. Boneh, S. Halevi, M. Hamburg, and R. Ostrovsky, "Circular-secure encryption from decision diffie-hellman," in *Annual International Cryptology Conference (CRYPTO2008)*, Springer, 2008, pp. 108–125.

[20] I. Haitner and T. Holenstein, "On the (im) possibility of key dependent encryption," in *Theory of Cryptography Conference (TCC2009)*, Springer, 2009, pp. 202–219.

[21] B. Barak, I. Haitner, D. Hofheinz, and Y. Ishai, "Bounded key-dependent message security," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT2010)*, Springer, 2010, pp. 423–444.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

IPFed: Identity protected federated learning for user authentication

Yosuke Kaga*, Yusei Suzuki*, Kenta Takahashi*
* Hitachi, Ltd., Japan

***Abstract*—With the development of laws and regulations related to privacy preservation, it has become difficult to collect personal data to perform machine learning. In this context, federated learning, which is distributed learning without sharing personal data, has been proposed. In this paper, we focus on federated learning for user authentication. We show that it is difficult to achieve both privacy preservation and high accuracy with existing methods. To address these challenges, we propose IPFed which is privacy-preserving federated learning using random projection for class embedding. Furthermore, we prove that IPFed is capable of learning equivalent to the state-of-the-art method. Experiments on face image datasets show that IPFed can protect the privacy of personal data while maintaining the accuracy of the state-of-the-art method.**

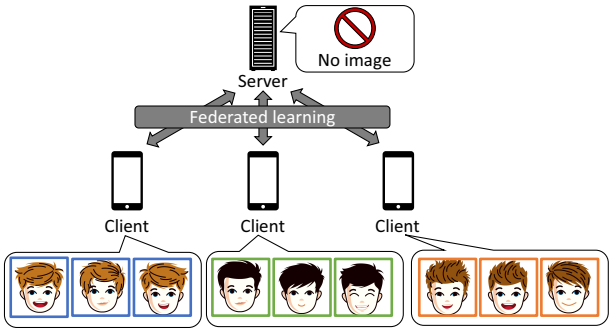


Fig. 1. Federated learning for user authentication.

I. Introduction

User authentication, such as face recognition, has recently achieved dramatic improvements in accuracy through the application of deep learning, and one of the reasons for this is that large numbers of images have been collected through web crawling and used as training data [1]. However, in recent years, GDPR [2] and other privacy-related regulations have made it more difficult to collect personal data. In order to continue to improve the accuracy of user authentication, Privacy-Preserving Machine Learning (PPML) [3] is needed, which performs machine learning without directly collecting personal data. One of the most popular methods in PPML is federated learning [4], which can perform privacy-preserving distributed learning using personal data on the client, and it has attracted attention. Federated learning has been applied to user authentication [5]–[8]. However, these methods face the challenges of privacy of training data [5], [6] and model accuracy [7], [8]. We propose Identity Protected Federated Learning (IPFed) as a solution to this problem. The main contributions of our work are as follows.

- We develop IPFed, which is a method to perform federated learning for user authentication while preserving the privacy of personal data using random projection for class embeddings.
- We prove mathematically that IPFed can perform equivalent learning to the state-of-the-art method while preserving the privacy of the training data.
- Experiments on face image datasets show that the proposed method can learn models with equivalent accuracy to the state-of-the-art method.

II. Related works

A. Federated learning

Federated learning is a method of machine learning in which training data is not shared but distributed across multiple devices. A typical federated learning algorithm is federated averaging (FedAvg) [4]. In this method, machine learning is performed while the personal data is kept in the clients, thus the privacy of training data can be preserved.

B. Federated learning for user authentication

Federated learning for user authentication is a method of learning a discriminator based on personal data such as fingerprints, faces, veins, etc., to determine a person's identity. An overview of federated learning for user authentication is shown in Fig.1. Since a client is often occupied by a single person in user authentication, it is natural for a single client to contain the personal data of only one user. Under this assumption, it is difficult to train a model that can discriminate between others because it is not possible for the client to refer to the personal data of others.

Two main approaches have been proposed to solve this problem. The first one is [5], [6], a method where the learning server performs the learning to increase the distance between the embeddings of the others, and the second is [7], [8], a method where the clients randomly assign representative embeddings for each class (called class embedding). In the following, we will give an overview of each method.

FedAwS [5], FedFace [6]: Federated Averaging with Spreadout (FedAwS) [5] is the algorithm to realize federated learning for user authentication by learning to increase the distance between each other's embeddings on the learning

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

server side. Furthermore, Aggarwal et.al. introduce FedFace, which applies FedAwS to face recognition[6].

Here, we introduce the algorithm of FedFace. The learning server contains a pre-trained model parameter θ_1, and the i^{th} client contains a local dataset S^i and a class embedding w_1^i. In federated learning, the goal is to improve accuracy by updating the model parameters θ_1 without sharing the local dataset S^i with other clients other than the learning server and the i^{th} client. In the t^{th} round, the i^{th} client receives the model parameter θ_t from the learning server. The model parameter θ_t and the class embedding w_t^i are updated by optimizing the positive loss function l_{pos} using the local data S^i.

$$l_{pos}(f_{\theta_t}(x), i) = \max\left(0, m - (w_t^i)^T f_{\theta_t}(x)\right)^2, \qquad (1)$$

where m denotes the margin parameter, $w_t^i \in \mathbb{R}^d$ represents the class embedding for the i^{th} client and $x \in S^i$ is training data, $f_{\theta_i} : X \to \mathbb{R}^d$ is a face feature extractor whose parameters are θ_i.

The updated model parameters θ_t^i and the class embedding w_t^i are sent to the learning server. The learning server aggregates the parameters θ_t^i by taking a weighted average and generates new model parameter $theta_{t+1}$. Then, the learning server obtains the class embedding matrix $W_t = [w_t^1, w_t^2, \ldots, w_t^C]^T$ where C is the number of clients.

Finally, the learning server performs spreadout regularization that minimizes the following equation:

$$reg_{sp}(W_t) = \sum_{c\in[C]} \sum_{\hat{c}\neq c} \left\{\max\left(0, v - d(w_t^c, w_t^{\hat{c}})\right)\right\}^2, \qquad (2)$$

where v represents the margin parameter and $d(\cdot,\cdot)$ is the function to calculate Euclidean distance. In this way, FedFace optimizes the model parameters θ_t and the class embedding matrix W_t and obtains the model with improved accuracy.

Hosseini at.el. [7], [8]: Next, we outline the second approach [7], [8]. In this method, class embeddings are randomly initialized and frozen on the client side, and are not shared with the learning server in order to preserve the privacy of the training data. The clients use the fixed class embedding to train the model, and the learning server aggregates the training results and updates the model parameter.

C. Challenging of related works

Next, we will describe the challenges of federated learning for user authentication. The first approach [5], [6] discloses the class embeddings w_t^i that contain information about the subject identity of the training data to the learning server. These class embeddings are representative embeddings for each subject. By using the class embeddings, the learning server could perform reconstruction of the original training data based on the model inversion attack [9]. Therefore, user privacy is not protected by this approach, since an attacker with access to the data held by the learning server can infer certain personal data, and the class embedding w_t^i should be kept secret from the learning server. On the other hand, in the second approach [7], [8], the class embeddings are assigned independently and randomly for each client, so it does not take into account the similarity of the class e.g. race or gender. Therefore, the performance of the model obtained in the second approach is inferior to the first approach that optimizes the class embedding.

From the above discussion, as far as we know, there is no method that can perform federated learning for user authentication with both high accuracy and privacy preservation of training data. Satisfying these two requirements is an important problem to be solved in the field of federated learning for user authentication.

Algorithm 1: Training algorithm of IPFed.

Input : A pre-trained face feature extractor f_{θ_1} at the learning server and the local dataset S^i at the i^{th} client.

i The i^{th} client initializes the class embedding w_t^i by average embedding obtained by inputting S^i into f_{θ_1}.;

for $t = 1, \ldots, T$ **do**

ii The learning server communicates the current global parameters θ_t to the i^{th} client;

iii The parameter server generates random transformation parameter r^t and send it to all clients;

for *each of the clients* $i = 1, 2, \ldots, C$ **do**

iv The client updates the model parameters based on the local data $x_j^i \in S^i$
$(\theta_t^i, \tilde{w}_t^i) \leftarrow (\theta_t, w_t^i) - \eta\nabla_{(\theta_t^i, w_t^i)} L_{pos}(S^i)$
where $L_{pos}(S^i) = \frac{1}{n_i}\sum_{j=1}^{n_i} l_{pos}(f_{\theta_t^i}(x_j^i), i)$;

v The i^{th} client perform random projection of class embedding $\bar{w}_t^i \leftarrow r_t \tilde{w}_t^i$;

vi The i^{th} client sends the updated parameters $(\theta_t^i, \bar{w}_t^i)$ back to the learning server

end

vii The learning server aggregates the updated parameters from all the clients
$\theta_{t+1} \leftarrow \frac{1}{n}\sum_{i\in[C]} n_i \cdot \theta_t^i$
where n represents the total number of samples on all the clients;

viii Finally, the learning server employs the spreadout regularizer to separate the class embeddings and obtain
$\hat{w}_t^i \in \hat{W}_t$
$\bar{W}_t = [\bar{w}_t^1, \ldots, \bar{w}_t^C]$
$\hat{W}_t \leftarrow \bar{W}_t - \lambda\nabla_{\bar{W}_t} reg_{sp}(\bar{W}_t)$;

ix The learning server sends $\hat{w}_t^i$ to the i^{th} client where $\hat{W}_t = [\hat{w}_t^i]$;

x The i^{th} client decodes class embedding and stores it:
$w_{t+1}^i \leftarrow (r_t)^{-1}\hat{w}_t^i$

end

xi **Output:** Output θ_{T+1}

III. IPFED

We propose a new method to solve the privacy and accuracy problems described in Sec.II. We call our proposed method Identity Protected Federated Learning (IPFed). The overview of IPFed and the training algorithm of IPFed are shown in Fig.2 and Algorithm 1, respectively.

IPFed consists of three entities: clients, a learning server, and a parameter server. The parameter server is managed separately from the learning server, and we assume that it is difficult to

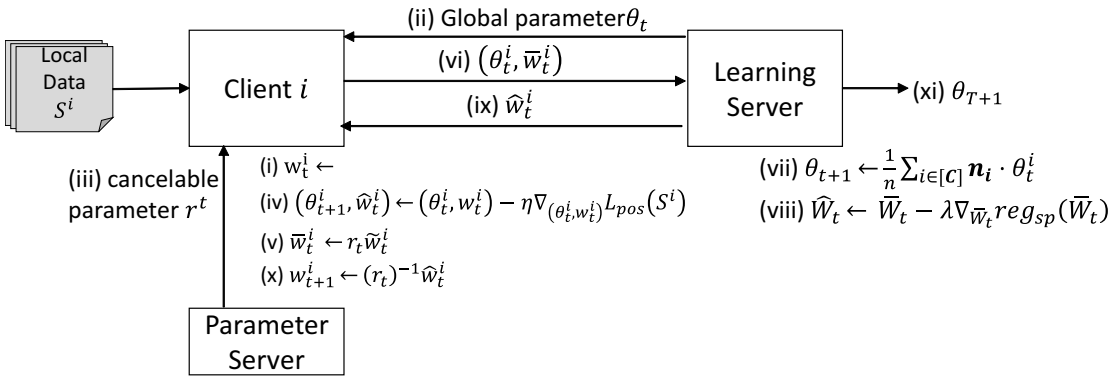


Fig. 2. The overview of IPFed.

steal data in the parameter server and the learning server at the same time. In the following, we explain the steps of IPFed. In the IPFed, a class embedding w_i^t is initialized and stored on each client (step (i)). The learning server sends the global parameter θ_t to each client (step (ii)). This parameter server randomly generates a transformation parameter $r^t \in \mathbb{R}^{d\times d}$ at each round of the federated learning and sends it to each client (step (iii)), where d is the dimension of the embeddings. Each client i updates the global parameter θ_t^i and the class embedding $\hat{w}_{t+1}^i$ (step (iv). In addition, the client calculates the inner product of the class embedding $\tilde{w}_t^i$ and the transformation parameter r^t (step (v)), and sends it to the learning server (step (vi)). The learning server aggregates the global parameters θ_t^i to θ_{t+1} (step (vii)). The learning server optimizes the received class embeddings $\bar{W}_t$ with spreadout similar to FedFace (step (viii)), and sends the result back to each client (step (ix)). The client transforms the received class embedding $\hat{w}_t^i$ by the inverse matrix of the transformation parameter $(r^t)^{-1}$ (step (x)), and keeps it as its own class embedding. After T times rounds, the trained global parameter θ_{T+1} is output.

In IPFed, the class embedding is multiplied by a random transformation parameter which is secret to the learning server. Furthermore, the updated class embedding is returned to the original feature space using the inverse matrix of the transformation parameters. This makes it possible to perform the optimization while keeping the class embeddings secret from any server. In the following, we show in Section III-A that our method can perform the same optimization as FedFace even when the class embedding is kept secret, and we also show in Section III-B that it is difficult for an attacker on any entity to obtain the user's personal data.

A. Derivation of IPFed

In this section, we analyze how the introduction of transformation parameters changes the spreadout process on the server side, and derive the conditions for learning equivalent to FedFace even with class embedding tramsformation. First, the loss function of spreadout reg_{sp} in FedFace is formulated in Eq.(2)[6]. Then, the update equation using this loss function is as follows.

$$\begin{aligned}\hat{w}_t^c &= \tilde{w}_t^c - \lambda\tfrac{\partial}{\partial\tilde{w}_t^c} reg_{sp}(\tilde{W}_t)\\ &= \tilde{w}_t^c - \lambda\tfrac{\partial}{\partial\tilde{w}_t^c}\left\{\textstyle\sum_{c\in[C]}\sum_{\hat{c}\neq c}\left(\max\left\{0, v - d(\tilde{w}_t^c, \tilde{w}_t^{\hat{c}})\right\}\right)^2\right\}\\ &= \tilde{w}_t^c - \lambda\tfrac{\partial}{\partial\tilde{w}_t^c}\left\{\textstyle\sum_{\hat{c}\in[C]} 2\left(\max\left\{0, v - d(\tilde{w}_t^c, \tilde{w}_t^{\hat{c}})\right\}\right)^2\right\}\\ &= \tilde{w}_t^c - \lambda\textstyle\sum_{\hat{c}\in[C]} 4(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})\cdot\min\left\{0, 1 - \frac{v}{d(\tilde{w}_t^c, \tilde{w}_t^{\hat{c}})}\right\}\end{aligned} \quad (3)$$

On the other hand, in Eq.(3) of the IPFed, the inner product of the transformation parameter and the class embedding is input to spreadout, so if we replace $\tilde{w}_t^c$ with $r_t\tilde{w}_t^c$, we get the following equation.

$$\begin{aligned}\hat{w}_t^c = &\quad r_t\tilde{w}_t^c - \lambda\sum_{\hat{c}\in[C]} 4(r_t\tilde{w}_t^c - r_t\tilde{w}_t^{\hat{c}})\cdot\min\{0, 1 - \tfrac{v}{d(r_t\tilde{w}_t^c, r_t\tilde{w}_t^{\hat{c}})}\}\\ \hat{w}_t^c = &\quad r_t\left\{\tilde{w}_t^c - \lambda\sum_{\hat{c}\in[C]} 4(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})\ \min\{0, 1 - \tfrac{v}{||r_t(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})||^2}\}\right\}\end{aligned} \quad (4)$$

In addition, the learning server sends the obtained class embedding to the client, and the client transforms it into the original feature space by taking the inner product with the inverse matrix of the transformation parameters $(r_t)^{-1}$. After adding this process, Eq.(4) can be expressed as follows.

$$(r_t)^{-1}r_t\left\{\tilde{w}_t^c - \lambda\sum_{\hat{c}\in[C]} 4(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})\cdot\min\left\{0, 1 - \frac{v}{||r_t(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})||^2}\right\}\right\} \quad (5)$$

Here, if the update equation of FedFace (Eq.(3)) and the update equation of IPFed (Eq.(5)) are equivalent, then IPFed can perform equivalent learning to FedFace even if the class embeddings are kept secret. The condition that the update equations of IPFed and FedFace are equivalent is expressed by the following equation.

$$(r_t)^{-1}r_t = I \quad (6)$$

$$||r_t(\tilde{w}_t^c - \tilde{w}_t^c)||^2 = ||(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})||^2 \quad (7)$$

where I is an identity matrix. Eq.(6) is valid if r_t is a regular matrix. Transforming Eq.(7), we obtain

$$\begin{aligned}(r_t(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}}))^T(r_t(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})) &= (\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})^T(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})\\ (\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})^T(r_t)^T r_t(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}}) &= (\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})^T(\tilde{w}_t^c - \tilde{w}_t^{\hat{c}})\\ (r_t)^T r_t &= I\end{aligned} \quad (8)$$

Therefore, Eq.(8) is valid if r_t is an orthonormal matrix. From the above, IPFed is equivalent to FedFace in terms of spreadout if the transformation parameter r_t is an orthonormal matrix.

B. Privacy analysis of IPFed

In this section, we discuss how privacy protection is achieved by our IPFed. We assume that an attacker against IPFed can obtain the data stored or communicated on any one of the three entities. We also assume a semi-honest model in which each entity follows correct protocols. We also assume that the goal of the attack is to obtain personal data of a specific individual.

Attacker against a client: From c^{th}client, the attacker can obtain the personal data S^c, the global parameter θ_t, the transformation parameter r_t, and the class embedding w_t^c,$\tilde{w}_t^c$. The attacker's goal is to obtain the personal data S^i (but $i \neq c$) stored on other clients from these data . S^c and r_t are completely independent of the personal data S^i ($i \neq c$) and do not contribute to the inference of the personal data. The attack to estimate training data from the global parameter θ_t is called Model Inversion Attack (MIA). However, in order to recover the training data using MIA, auxiliary information such as class labels and class embeddings are required [9]. Since they are kept secret in the IPFed, it is difficult to perform MIA. In addition, the class embedding w_{t+1}^c is obtained by updating w_t^c according to the distance from the class embeddings of other classes, and some information about other class embeddings may be leaked from the class embedding. However, in federated learning, [10] suggests that it is difficult enough to recover the original data from the result of aggregating many gradients, and it is difficult to estimate other class embeddings from one class embedding.

Attacker against the parameter server: From the parameter server, the attacker can obtain the transformation parameter r_t. Since the transformation parameter is randomly generated data independent of the personal data, it is difficult to obtain any personal data.

Attacker against the learning server: From the learning server, the attacker can obtain the global parameters θ_t,θ_{t+1}, the updated global parameters θ_{t+1}^i, the transformed class embeddings $\bar{w}_t^i$ can be obtained. The attacker's goal is to obtain specific personal data S^i from these data. Estimation of personal data from the global parameters θ^t and θ_{t+1} is possible by gradient inversion attack, but it is extremely difficult on practical settings [10]. Moreover, it is known that an attack to estimate training data from the updated global parameter θ_t^i can be made difficult by using secure aggregation [11] to calculate θ_{t+1} while keeping θ_t^i secret. By using these methods, user privacy can be protected under our assumption. On the other hand, random projection is used for template protection, called cancelable biometrics [12]. This random projection is known to make it difficult to estimate the original data from the transformed data when the transformation parameter is unknown. In general, it is known that the original data can be estimated when multiple random projections are available for the same data, but in federated learning for user authentication, the class embedding is updated and different each round, making the attack difficult. Furthermore, by setting up a shuffle server between the learning server and the client and hiding the client's ID from the learning server, it is possible to make it difficult to assign class embeddings in different rounds. This shuffle server can increase security. From the above discussion, it is difficult for the learning server to infer the class embedding $\tilde{w}_t^i$ from $\bar{w}_t^i$.

In conclusion, we have confirmed that IPFed can strongly protect the privacy of training data under the assumptions defined in this paper.

C. Efficiency analysis of IPFed

In this section, we discuss the efficiency of IPFed. In Fig.2, steps (4), (7), (11) are newly introduced steps in IPFed. Step (4) is related to communication efficiency. A transformation parameter $r^t \in \mathbb{R}^{d \times d}$ is very small data compared to the global parameter θ_t, and its transmission has a negligible effect on the learning efficiency. Furthermore, steps (7), (11) are related to computational efficiency on the client. In these steps, the inner product of the matrix is computed only once, thus it has a negligible impact on learning efficiency compared to updating the global parameter in step (6).

Note that the parameter server is a newly introduced and its operating costs are newly incurred. However, since the role of the parameter server is only to generate and send the transformation parameter, there is no problem even if the server has very little computing power.

From the above discussion, it can be said that the efficiency of IPFed is comparable to that of the conventional method [6]. However, quantitative evaluation of the efficiency of IPFed is a future work.

IV. Experiments

In this chapter, we show the effectiveness of the proposed method through experiments on face image datasets.

A. Setting

Datasets: We follow the setting in [6] and used CASIA-WebFace [14] for training. CASIA-WebFace consists of 494,414 images of 10,575 subjects. We randomly select 9,000 subjects for pre-training and 1,000 subjects for federated learning. To evaluate the performance of face verification, we use three datasets: LFW [15], IJB-A [16] and IJB-C [17].

Implementation: We use CosFace [18] for the face feature extractor. Only CosFace was used as the face feature extractor according to [6], but evaluation using more recent multiple the face feature extractors is a subject for future work. We use the scale parameter of 30 and the margin parameter of 10 for the CosFace loss function. The margin parameters are $m = 0.9$ and $v = 0.7$. The parameter λ = 25 in Eq.(3). We perform federated learning with a communication round of 200 and a learning rate of 0.1.

B. Evaluation

First, the face verification performance of each method for 1000 clients is shown in the Table.I. The comparison methods are shown below:

TABLE I
FACE VERIFICATION PERFORMANCE ON STANDARD FACE RECOGNITION BENCHMARKS LFW, IJB-A AND IJB-C.

Method	Training data	Class embedding	LFW Accuracy (%)	IJB-A TAR @ FAR=0.1%	IJB-C TAR @ FAR=0.1%
Baseline	Centrally aggregated	-	99.15%	74.21%	80.60%
Fine-tuning	Centrally aggregated	Not protected	99.30%	77.10%	82.71%
FCE	Distributed	Protected	99.17%	72.71%	80.06%
FedFace	Distributed	Not protected	99.22%	74.76%	79.77%
IPFed	Distributed	Protected	99.22%	75.70%	80.73%

TABLE II
IJB-A TAR @FAR=0.1% ON NUM. OF SUBJECTS.

Method	Num. of subjects			
	50	100	500	1000
Fine-tuning	75.15%	75.15%	75.76%	77.10%
FCE	72.54%	72.21%	74.38%	72.71%
IPFed	73.60%	72.82%	76.65%	75.70%

Baseline A typical CosFace model, pre-trained on 9000 subjects.

Fine-tuning A fine-tuned model of Baseline using CosFace on 1000 subjects.

FedFace A model trained according to [6].

IPFed A model trained based on the proposed random projection approach.

Fixed class embedding (FCE) A model trained by fixing the class embedding to the initial data. In FCE, the class embedding does not need to be shared with the server, and secure federated learning can be performed.

As shown in Table.I, IPFed achieves the same level of accuracy as FedFace. This means that the random projection based spreadout in IPFed is equivalent to the spreadout in FedFace. On the other hand, FCE is less accurate than IPFed. This is due to the lack of class embedding optimization, which means that the spreadout in IPFed contributes to the accuracy improvement.

Furthermore, the accuracy against the number of subjects used for federated learning is shown in TableII. The IPFed achieves higher accuracy than the FCE for any of the number of the subjects. This indicates that sharing and optimizing class embedding improve the accuracy. However, while Fine-tuning has a monotonic increase in accuracy for the number of the subjects, IPFed does not. This may be due to the fact that the hyperparameters in learning are not optimal for the number of the subjects, so automatic adjustment of hyperparameters is a future task.

In addition, in our experiments, only accuracy was evaluated, and attack defense performance was only theoretically evaluated. This is also a topic for future work.

V. CONCLUSION

In this paper, we focused on the problem of personal data leakage from class embedding in federated learning for user authentication, and proposed IPFed, which performs federated learning while protecting class embedding. We proved that IPFed, which is based on random projection for class embedding, can perform learning equivalent to the state-of-the-art method. We evaluate the proposed method on face image datasets and confirm that the accuracy of IPFed is equivalent to that of the state-of-the-art method. IPFed can improve the model for user authentication while preserving the privacy of the training data.

REFERENCES

[1] I. D. Raji and G. Fried, *About face: A survey of facial recognition evaluation*, 2021. arXiv: 2102.00813 [cs.CV].

[2] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, p. 3 152 676, 2017.

[3] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security Privacy*, vol. 17, no. 2, pp. 49–58, 2019. DOI: 10.1109/MSEC.2018.2888775.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[5] F. Yu, A. S. Rawat, A. Menon, and S. Kumar, "Federated learning with only positive labels," in *International Conference on Machine Learning*, PMLR, 2020, pp. 10 946–10 956.

[6] D. Aggarwal, J. Zhou, and A. K. Jain, "Fedface: Collaborative learning of face recognition model," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–8. DOI: 10.1109/IJCB52358.2021.9484386.

[7] H. Hosseini, S. Yun, H. Park, C. Louizos, J. Soriaga, and M. Welling, *Federated learning of user authentication models*, 2020. arXiv: 2007.04618 [cs.LG].

[8] H. Hosseini, H. Park, S. Yun, C. Louizos, J. Soriaga, and M. Welling, "Federated learning of user verification models without sharing embeddings," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 4328–4336. [Online]. Available: https://proceedings.mlr.press/v139/hosseini21a.html.

[9] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19, London, United Kingdom: Association for Computing Machinery, 2019, pp. 225–240, ISBN: 9781450367479. DOI: 10.1145/3319535.3354261. [Online]. Available: https://doi.org/10.1145/3319535.3354261.

[10] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[11] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[12] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001. DOI: 10.1147/sj.403.0614.

[13] H. Al-Assam, H. Sellahewa, and S. Jassim, "A lightweight approach for biometric template protection," in *Mobile Multimedia/Image Processing, Security, and Applications 2009*, SPIE, vol. 7351, 2009, pp. 194–205.

[14] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[15] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[16] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939.

[17] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 International Conference on Biometrics (ICB)*, IEEE, 2018, pp. 158–165.

[18] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

Privacy-Preserving Image Transformation Method for Person Detection and Re-ID

Yumo Ouchi, Hidetsugu Uchida and Narishige Abe
FUJITSU LIMITED, Japan

Abstract — **In recent years, the number of surveillance cameras is on the rise, and the limitation on image recognition including person re-identification (Re-ID) is lifted to ensure a safer and more convenient world. However, this has led to increasing privacy concerns. To address this issue, numerous privacy-preserving technologies have been proposed but they focused primarily on detection or classification of people. In this study, we propose a novel image transformation method to preserve privacy in detecting and re-identifying people. Privacy-preserving images visually protect images of people, but a machine can detect and classify them without any additional information. Our results show that the privacy-preserving images, transformed by our method, maintained the accuracy of person detection and Re-ID compared to that of the images with gaussian noise, blur, and pixelation.**

I. INTRODUCTION

With the evolving Artificial Intelligence technology, machine-learning techniques such as face recognition are being explored in the field of biometrics. Recently, the use of security camera images has advanced to ensure a safe and more convenient world. Moreover, the accuracy of the person re-identification (Re-ID)—which is the core technology in the technical field—has substantially improved. The Re-ID is a technique through which the person's identity is matched in video sequences captured by different cameras. Utilization of Re-ID for security camera images installed in the city can ensure surveillance (such as tracking suspicious persons and searching for lost and wandering persons), and marketing (such as analyzing customer traffic lines in stores).

However, psychological anxiety and discomfort are undeniable concerns owing to the possibility of privacy leakage due to continuous monitoring of the user. This has prompted the creation of many privacy-preserving technologies [1] [2] [3] [4]. Research in nonvisibility technology, which involves direct processing the data acquired from cameras and sensors obscuring the person's identity, is actively ongoing. The Re-ID system can be used in many situations if the psychological anxiety of the users can be minimized. There are two approaches to the nonvisibility of a person that involves 1) use of hardware instead of a camera and 2) use of software-based conversion techniques of general camera images.

In the hardware-based approach, Ohno et al. [5] proposed a method that tracks people using 3D point cloud information captured by multiple LiDAR sensors. Cheng et al. [6] also proposed a less precise and cheaper method than LiDAR using Radar to Re-ID a person using 4D Point Cloud information with 3D point cloud and Doppler velocity. Fan et al. [7] demonstrated a method for re-identifying signals using radio signals such as Wi-Fi. However, LiDAR and Radar are not commonly used in general settings, and their applications are limited. In addition, they are costly to install.

In the software-based approaches, RBG images captured by general cameras were made nonvisible by image processing. This could entail blurring or adding noise to the detected person or the face region [8]; however, the accuracy of the Re-ID is poor, and this approach is not practical. This prompted the creation of methods that ensure nonvisibility while suppressing deterioration in accuracy of Re-ID. For example, Ahmad et al. [9] proposed a Re-ID method using different information (Event Frame) from the image several frames before. Moreover, adding a patch image to a part of the face and using face anonymization to convert a detected face region into a face image of another person [10] have been elucidated. In addition, Kiya et al. [11] proposed a method in which the image data encrypted with a secret key is trained using vision transformer [12] and the images were encrypted with the same secret key during inference to identify images. Dou et al. [13] proposed a method that changes the appearance while maintaining the accuracy of Re-ID: using De-Identification technology with variational autoencoder [14], while utilizing the existing feature extraction model.

However, in the aforementioned techniques, the necessity to retrain the person detection model or the Re-ID feature

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

extraction model or focus on any one of the "person detection" and the "Re-ID feature extraction" cannot be overemphasized. Therefore, in this study, we propose a technology that ensures the nonvisibility of a person in a video image acquired from a camera while maintaining the accuracy of detecting and re-identifying the person without retraining the models. The following remarks were obtained in our study:

- We aimed at exploring nonvisibility technology for person detection and Re-ID
- We evaluated the usefulness of the proposed method by comparing with the Re-ID accuracy in the plain image
- We confirmed that the proposed method achieves effective nonvisibility by comparing the results with the Re-ID accuracy of pseudo-nonvisibility images with gaussian noise, blur, and pixelation

Section II describes the proposed method, Section III describes experiments and their results, and Section IV provides conclusions.

II. PROPOSED METHOD

The proposed method aims is to visually exclude the individuality of a person in a video frame acquired from a camera while suppressing the accuracy degradation of in detecting and re-identifying the person. Ito et al. [15] focused only on the image classification task. We extended this training framework and defined loss functions that suppress accuracy degradation in person detection and Re-ID feature extraction.

The model for transforming a plain frame image into a nonvisibility image is trained using pretrained person detection and Re-ID feature extraction models. First, plain frame images are fed to the transformation network to obtain transformed images. We have used U-net [16] as the image transformation network. The plain frame images and the transformed frame images are cropped as the bounding box area of the ground truth and input to the image similarity loss function. The image similarity loss function $L_{similarity}$, is L1 loss, which is calculated as follows:

$$L_{similarity}(B, \hat{B}) = \frac{1}{n}\sum_{i=1}^{n} l_i \tag{1}$$

$$l_i = |b_i - \hat{b}_i| \tag{2}$$

Here, let $B = \{b_1, b_2, \dots, b_n\}$ a cropped plain image set, $\hat{B} = \{\hat{b}_1, \hat{b}_2, \dots, \hat{b}_n\}$ is a cropped transformed image set, and n is the batch size.

Next, the detection result is obtained using the transformed frame image and the pretrained person detection model. In the experiment, among object classes detected using Faster RCNN [17], only person detection results are used for training. The person detection result and ground truth labels are input to the person detection network loss function, and we obtain a loss L_{object} from the person detection network loss function. L_{object} is the loss defined by Faster RCNN.

Thereafter, the Re-ID feature of the cropped transformed images is obtained using the cropped transformed image and the pretrained Re-ID feature extraction model. At the same time, the Re-ID feature amount is also obtained from the cropped plain image. In our experiment, the method proposed by Zheng et al. [18] is used as a Re-ID feature extraction model. The Re-ID features of the cropped transformed images and the Re-ID features of the cropped plain images are input into the Re-ID Network Loss function. The loss, L_{reid} is obtained from the Re-ID Network Loss function. L_{reid} is the cosine similarity of the features and is calculated as follows:

$$L_{reid}(F, \hat{F}) = 1 - \frac{1}{n}\sum_{i=1}^{n} s_i \tag{3}$$

$$s_i = \frac{f_i \cdot \hat{f}_i}{\max\left(\|f_i\|_2 \cdot \|\hat{f}_i\|_2, \in\right)} \tag{4}$$

Here, $F = \{f_1, f_2, \dots, f_n\}$ is the Re-ID feature set of the cropped plain image, $\hat{F} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_n\}$ is the Re-ID feature set of the cropped transformed images and n is the batch size. Using $L_{similarity}$, L_{object}, and L_{reid}, the image transformation network minimizes the loss function, L_{trans}. L_{trans} is defined as:

$$L_{trans} = L_{object} + L_{reid} - \alpha \cdot L_{similarity} \tag{5}$$

Here, α is adjustment parameter. However, the number of epochs in the first half of the training is defined as follows:

$$L_{trans} = L_{object} + L_{reid} + L_{similarity} \tag{6}$$

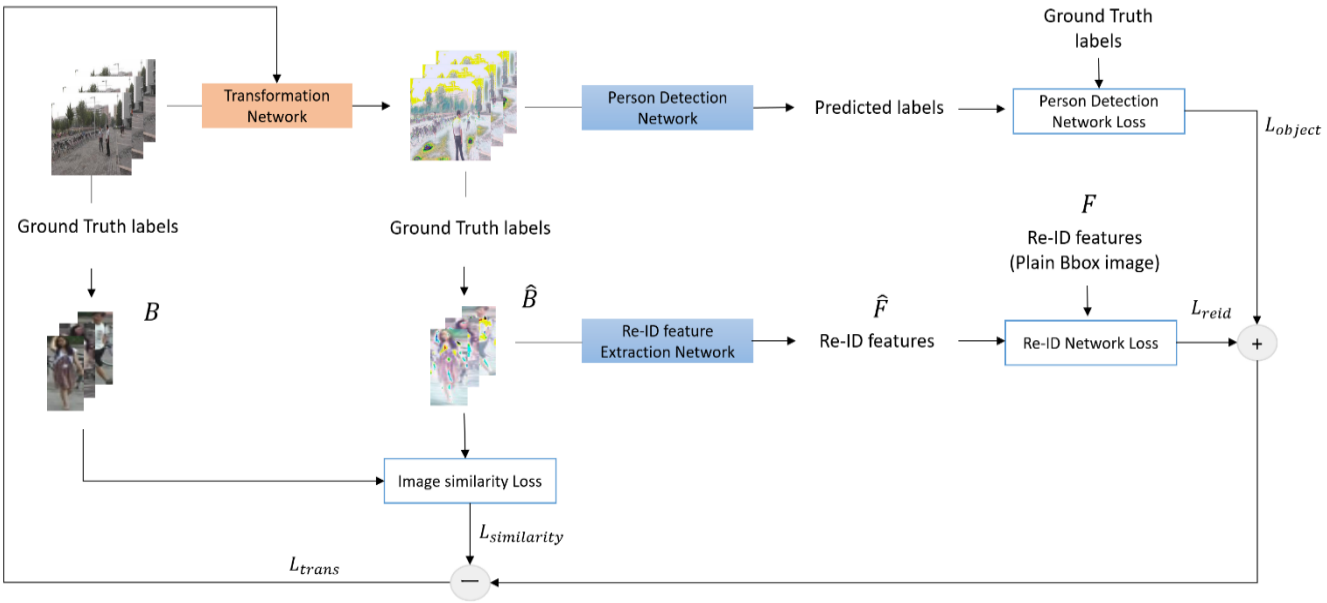


Figure 1: Proposed training process of transformation network

This ensures that the image transformation network generates an image close to the plain image.

III. EXPERIMENT

The degree of nonvisibility is evaluated in terms of the person detection accuracy, Re-ID accuracy from Re-ID feature extraction, and similarity with the plain image.

A. Experimental method

Faster RCNN—used as a person detection model—is a pretrained model provided by PyTorch[1]. The Re-ID feature extraction model is pretrained on the Market-1501 dataset [19] using the method proposed by Zheng et al. [18]. Using the two pretrained models, U-net was trained according to the training framework shown in Figure 1. In this training, we have used 5,703 images of Person Re-identification in the Wild (PRW) Dataset [20]: 4,563 images used as training data and 1,140 images as verification data. The activation function was Adam, the learning rate was 0.001, the batch size was 4, the number of epochs was 20, and the model of the epoch with the lowest loss was used for the evaluation. In the first half of training was conducted using the loss of equation (6) from one epochs to seven epochs. In the last half of training was conducted using the loss of equation (5) from eight epochs to twenty epochs, where the loss was sufficiently reduced. The adjustment parameter of equation (5) is $\alpha = 0.1$.

The evaluation data comprised 1,000 images extracted from the PRW Dataset. We used the average of Intersection over Union as the evaluation method of the person detection. We used the detection result classified as evaluation data of person detection. Moreover, we used Rank-1 accuracy and mean Average Precision (mAP) as evaluation method for Re-ID. The bounding box regions of the ground truth were cropped from the frame images and used for evaluation of Re-ID.

B. Privacy-Preserving Image for person detection

First, we conducted an ablation study to confirm the learning process shown in Figure 1. In addition, we created privacy-preserving image transformation model to detect persons only. The learning process is described in Section AThe loss function shown in Eq. (5) was used as $L_{reid} = 0$ between 8 epochs and 20 epochs. The performance of the transformed image is shown in Table 1. "Re-ID" in Table 1 indicates its matching accuracy using Re-ID feature extraction. From Table 1, the privacy-preserving image transformation model to solely detect persons maintained its accuracy. However, the Rank-1 accuracy of Re-ID decreased by 27% and mAP decreased by 37%. Therefore, the learning process of Figure 1 functioned as intended.

[1] https://pytorch.org/vision/main/models/faster_rcnn.html

C. *Privacy-Preserving Image for person detection and Re-ID*

Second, we created a privacy-preserving image transformation model for person detection and Re-ID. The learning method is described in Section A. The resulting transformed image is shown in Figure 2. Table 1 demonstrates that the person detection accuracy was equivalent to its performance using plain images. However, although the Rank-1 accuracy and mAP of Re-ID reduced by 14%, the accuracy of Re-ID was higher than that of privacy-protected images that solely detected person in Section B. This shows that the model learned using loss for person detection and Re-ID.

D. *Comparison of accuracy with image processing method*

Third we compared the accuracy in three image processing methods: gaussian noise, blur, and pixilation, to verify whether transformed images were as intended such that the person detection accuracy and Re-ID accuracy were not lowered. To evaluate person detection, we first evaluated transformed images with Peak signal-to-noise ratio (PSNR)—an image quality evaluation index and a higher PSNR value provides a higher image quality—and added gaussian noise to plain images to have the same PSNR value. The PSNR of blur and pixilation could not be reduced to the same level, and image processing was performed to such an extent that the face of the person in the image could not be identified. So, images of blur and pixelation were of higher image quality than those generated by the proposed method. Next, to evaluate Re-ID, we evaluated the cropped transformed images with PSNR, and added gaussian noise to plain cropped images to have the same PSNR value. Since the PSNR for blur and pixelation could not be reduced to the same level, bounding box image processing was performed such that the face of the person in the image could not be identified. Examples of a gaussian noise, blur, pixelation images are shown in Figure 2. Table 1 shows the results of the accuracy of person detection and Re-ID for the three methods. However, the PSNR values for blur and pixelation were not shown, since they could not generate an image based on PSNR.

From Table 1, we found that the transformed images generated by the proposed method were more accurate than that obtained from the three methods in person detection and Re-ID. Thus, the proposed method caused deterioration in image quality, maintained the accuracy of the Re-ID processing, and ensured nonvisibility.

IV. CONCLUSION

In this paper, we propose the novel image transformation method. Pretrained person detection model and Re-ID feature extraction model can be used to develop the proposed image transformation network. In the experiment, the Re-ID accuracy of the generated nonvisibility images were evaluated using the pretrained models. Our experimental results showed that the accuracy of person detection was maintained as that obtained using plain images, though the Re-ID feature extraction accuracy decreased by 14%. In addition, the nonvisibility images of the proposed method exhibited better accuracy compared to that of the images with gaussian noise, blur, and pixelation.

In the future, we will focus on improving the feature extraction accuracy and evaluating the nonvisibility images using more effective nonvisibility technology.

REFERENCES

[1] Jawad Yousef Ibrahim Alzamily, Syaiba Balqish Ariffin, Samy S. Abu-Naser. “Classification of Encrypted Images Using Deep Learning –ResNet50.” Journal of Theoretical and Applied Information Technology, 100, no. 20, 2022.

[2] T. Heitzinger and M. Kampel, "IPT: A Dataset for Identity Preserved Tracking in Closed Domains," 2020 25th International Conference on Pattern Recognition (ICPR), Milan,

Table 1: Person detection and re-identification (Re-ID) accuracy and Peak signal-to-noise ratio (PSNR)

Method	Person detection	Re-ID		Similarity	
	IoU	Rank-1	mAP	PSNR of images	PSNR of cropped images
Plain image	0.7893	0.9187	0.8383	-	-
Gaussian noise	0.7288	0.6049	0.4249	9.5010	8.5680
Blur	0.7121	0.6794	0.4873	-	-
Pixelation	0.0001	0.7449	0.6013	-	-
Proposed (for Person detection)	0.7789	0.6478	0.4676	9.6808	8.2936
Proposed (for Person detection + Re-ID)	0.7814	0.7855	0.6126	9.4529	8.5059



Figure 2: Example of frame images transformed by proposed method.

Italy, 2021, pp. 8228-8234, doi: 10.1109/ICPR48806.2021.9412979.

[3] C. Zhang, L. Zhu, S. Zhang and W. Yu, “TDHPPIR: An Efficient Deep Hashing Based Privacy-Preserving Image Retrieval Method,” In Proceedings of Neurocomputing 406, pp.386-398, 2020, doi: 10.1016/j.neucom.2019.11.119

[4] H. B. S. S, V. Agarwal, S. Ghosh, G. Ramena, S. Kumar and B. R. K. Raja, "PrivPAS: A real time Privacy-Preserving AI System and applied ethics," 2022 IEEE 16th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2022, pp. 9-16, doi: 10.1109/ICSC52841.2022.00010.

[5] M. Ohno, R. Ukyo, T. Amano, H. Rizk and H. Yamaguchi, "Privacy-preserving Pedestrian Tracking using Distributed 3D LiDARs," 2023 IEEE International Conference on Pervasive Computing and Communications (PerCom), Atlanta, GA, USA, 2023, pp. 43-52, doi: 10.1109/PERCOM56429.2023.10099061.

[6] Y. Cheng and Y. Liu, "Person Reidentification Based on Automotive Radar Point Clouds," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022, Art no. 5101913, doi: 10.1109/TGRS.2021.3073664.

[7] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan and D. Katabi, "Learning Longterm Representations for Person Re-Identification Using Radio Signals," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10696-10706, doi: 10.1109/CVPR42600.2020.01071.

[8] G. Megala, R. Venkatesan and T. Vigneshwaran, "Video Face Anonymization for Preserving Privacy," 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICDSAAI55433.2022.10028937.

[9] S. Ahmad, G. Scarpellini, P. Morerio and A. D. Bue, "Event-driven Re-Id: A New Benchmark and Method Towards

Privacy-Preserving Person Re-Identification," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2022, pp. 459-468, doi: 10.1109/WACVW54805.2022.00052.

[10] H. Lee, M. U. Kim, Y. Kim, H. Lyu, and H. J. Yang, "Development of a Privacy-Preserving UAV System With Deep Learning-Based Face Anonymization," in IEEE Access, vol. 9, pp. 132652-132662, 2021, doi: 10.1109/ACCESS.2021.3113186.

[11] H. Kiya, R. Iijima, A. Maungmaung, Y. Kinoshita, “Image and Model Transformation with Secret Key for Vision Transformer,” in IEICE Trans. Inf. & Sys., vol. E106.D, no.1, pp.2--11, January 2023, doi: 10.1587/transinf.2022MUI0001.

[12] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2021.

[13] S. Dou, X. Jiang, Q. Zhao, D. Li, C. Zhao, “Towards Privacy-Preserving Person Re-identification via Person Identify Shift,” arXiv:2207.07311, 2022.

[14] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in The 2nd International Conference on Learning Representations (ICLR), 2014.

[15] H. Ito, Y. Kinoshita, and H. Kiya, “Image to Perturbation: An Image Transformation Network for Generating Visually Protected Images for Privacy-Preserving Deep Neural Networks,” in IEEE Access, vol. 9, pp. 64629-64638, 2021, doi: 10.1109/ACCESS.2021.3074968.

[16] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[18] Z. Zheng, L. Zheng and Y. Yang, “A discriminatively learned CNN embedding for person reidentification,” in ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol.14, pp. 13,2018

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1116-1124, doi: 10.1109/ICCV.2015.133.

[20] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang and Q. Tian, "Person Re-identification in the Wild," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3346-3355, doi: 10.1109/CVPR.2017.357.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Eye Biometrics Combined with Periocular and Iris Recognition Using CNN

Taito Tonosaki*, Shokei Kawakami*, Koichi Ito*, Takafumi Aoki*,
Yoshiko Yasumura†, Masakazu Fujio†, Yosuke Kaga†, and Kenta Takahashi†
* Graduate School of Information Sciences, Tohoku University, Japan.
E-mail: ito@aoki.ecei.tohoku.ac.jp
† Hitachi Ltd., Japan.

***Abstract*—In this paper, we propose an eye biometric recognition method, which is combined with periocular and iris recognition. The proposed method extracts features using convolutional neural networks (CNNs) dedicated to periocular and iris recognition, respectively. The matching score is calculated by weighting scores from periocular and iris recognition. Through a set of experiments using public datasets, we demonstrate the effectiveness of the proposed method compared with conventional methods.**

I. Introduction

Accurate and reliable person authentication is essential for immigration control and electronic payment systems, which require high security. One of such personal authentication techniques is biometrics, which identifies individuals using their physical and behavioral characteristics [1]. Biometric characteristics include face, iris, fingerprints, gait, signature, and so on, which have different characteristics from each other. A fingerprint, for example, is highly reliable due to its high discrimination and permanence, while acceptability is low due to the contact with the sensor when a fingerprint is captured. A face is highly collectable and convenient since it can be captured using a common RGB camera, while it is unreliable since its features change with aging, makeup, facial expressions, and so on. Thus, biometric recognition using a single biometric characteristic has both high reliability but low convenience and high convenience but low reliability, and it is difficult to achieve both reliability and convenience in biometric recognition using a single biometric characteristic. In this paper, we focus on the region around the eyes, which has multiple biometric characteristics.

There are two types of biometric recognition using the region around the eye: iris and periocular. Iris recognition uses an iris, which is a ring-shaped region between the pupil and the sclera [2]. The iris has a fine pattern called the iris pattern, which is highly discriminative even in the eyes of two identical twins or the left and right eyes of the same person. Furthermore, the iris is invariant throughout life, and it is difficult to change the pattern through surgery and so on. As a result of these advantages of the iris, iris recognition is one of the most reliable biometric recognition methods, and is used in situations where high security is required. On the other hand, iris recognition is not convenient since it requires the user to look into the camera to capture the iris image. Periocular recognition uses features around the eyes such as eyebrows, eye shape, eyelid wrinkles, and skin texture [3]. Unlike iris recognition, periocular recognition can be performed without facing the camera, making it convenient and suitable for situations that require high throughput, such as walk-through authentication. On the other hand, periocular features are not very discriminative since they may be similar even among different people, such as identical twins, and they change with age, makeup, facial expressions, etc. Therefore, periocular recognition is not very reliable compared with iris recognition.

In this paper, we propose an eye biometrics system that combines reliable iris recognition and convenient periocular recognition. The proposed method extracts iris and periocular features using SE-ResNet-18, which is a combination of ResNet [4] and the attentional mechanism used in Squeeze-and-Excitation Networks [5]. For iris recognition, we basically adopt our proposed method [6]. The final matching score is calculated by fusing the matching scores calculated from the iris and periocular features. We improve the recognition accuracy by applying data augmentation that reproduces the variation of the iris and periocular images. Through experiments using the public datasets CASIA-Iris-Thousand, CASIA-Iris-Distance, and CASIA-Iris-Mobile-V1.0 [1] consisting of periocular images, we demonstrate the effectiveness of the proposed method compared with conventional methods.

II. Related Work

This section presents an overview of related work on iris recognition, periocular recognition, and their combination.

A. Iris Recognition

The most well-known iris recognition is iriscode [7], [8] proposed by Daugman, and many derivatives have been proposed. Among them, OSIRIS [9], which is an open-source iris recognition software modified from iriscode, is often used as a baseline for performance evaluation. OSIRIS first segments the iris region from a high-resolution periocular image and detects the iris contour from a low-resolution periocular image. Using

[1] http://biometrics.idealtest.org

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

both results, the parameters of the inner and outer bounding circles of the iris are obtained. Next, the iris region is normalized to a rectangle by polar coordinate mapping, and iriscodes are generated using a 2-D Gabor filter. Finally, the Hamming distance between iriscodes is used as the matching score. The accuracy of iris segmentation is significantly degraded by narrowing the eyes or wearing glasses, and iriscode cannot be generated accurately when the image is blurred. UniNet, which introduces deep learning into Iriscode, has been proposed [10]. UniNet is based on Fully-Convolutional Network (FCN) [11], which can preserve local features by keeping positional relationships between inputs and outputs, since iris consists of random patterns and has high discriminative capability in local features. UniNet consists of two types of FCNs: FeatNet and MaskNet. FeatNet plays the role of feature extraction by Gabor filter in iriscode, and MaskNet generates iris masks from normalized iris images.

B. Periocular Recognition

AttNet [12] has been proposed for periocular recognition using deep learning, which focuses on the high discriminability of the region between the eyes and eyebrows in periocular recognition, and applies an attention to this region. First, a mask representing the eye and brow regions in the periocular image is generated using an improved version of FCN called FCN-Peri. Next, the mask is multiplied by the intermediate features of AttNet, which extracts features from the periocular image, and the highly discriminative feature are extracted by applying attentions to these regions.

C. Periocular and Iris Recognition

A method combining UniNet [10] and AttNet [12] has been proposed to combine periocular and iris recognition [13]. Matching scores are calculated by inputting the iris features obtained by UniNet and the periocular features obtained by AttNet into a dynamic fusion block consisting of four FC layers. The recognition accuracy is significantly degraded when there is occlusion in the iris since iris segmentation is performed using RTV-L1 [14] which is based on handcrafted features.

III. Eye Biometrics Using CNN

This section describes the proposed method, which is a multi-biometric recognition method combining iris and periocular recognition. Fig. 1 shows an overview of the proposed method. We segment and localize the iris region in the periocular image to perform iris recognition. Three masks (iris mask, outer iris circle mask, and pupil circle mask) are generated from the periocular image using U-Net [15], and the ellipse parameters of the outer iris circle and the pupil circle are estimated. Based on the estimated ellipses, the iris region is localized, the iris is divided into four sub-regions, and each sub-region is normalized to a rectangular region. CNN-based feature extractors are used to extract features from the normalized iris image and the periocular image, respectively. The proposed method uses SE-ResNet-18 [4], [5] as feature extractors. In training the feature extractors, we introduce data augmentation that produces images similar to the variations in the iris and periocular images. When training the iris feature extractor, the training data is selected based on the ratio of non-iris regions in the iris image to reduce the influence of non-iris regions such as eyelids and eyelashes that are included in the iris image. The final matching score is calculated by integrating the matching score calculated from the iris and periocular regions. We describe the details of the iris detection, feature extraction, data expansion, and matching processes in the following.

A. Iris Detection

This section describes iris detection from periocular images. A segmentation network based on U-Net [15] is used to generate three binary masks of the iris. The proposed method uses U-Net whose encoder is replaced by SE-ResNet-18 [4], [5]. SE-ResNet-18 is a CNN model that introduces a Squeeze-and-Excitation block (SE block), which is an attention mechanism, into ResNet, which is a CNN model used in image recognition. In SE block, feature maps output from the convolutional layer are averaged in the channel direction and aggregated into a vector, and then passed through two fully-connected layers and activation functions (ReLU and Sigmoid) to convert it into a vector that represents the weights for each channel. The weight vectors are then multiplied by the feature map to obtain a refined feature map. U-Net outputs an iris mask representing the iris segmentation result, an outer mask representing the outer circle of the iris, and an inner mask representing the pupil segmentation result. The iris mask corresponds to the iris region detected at the pixel level, excluding occlusions such as eyelids, while the outer and inner masks represent the boundary circle between the iris and sclera, and the boundary circle between the iris and pupil, respectively. The outer and inner masks are not exact circles since a part of the circle may be lacking due to the occlusion. By applying ellipse fitting to both masks, we can estimate the center coordinates, the lengths of the major and minor axes, and the rotation angle of the outer and inner circles. The upper and lower parts of the iris are often occluded by the eyelids and eyelashes, making it more difficult to acquire features than the right and left sides of the iris. Therefore, the iris region is divided into four sub-regions: Upper (U), Bottom (B), Left (L), and Right (R), which are used for training and recognition. This approach effectively suppresses the effect of occlusion on the recognition accuracy. Each iris region is normalized to a rectangle by polar coordinate mapping. The normalized iris image has 512×64 pixels, and each sub-region has 128×64 pixels.

The loss function for training U-Net is defined by

$$L_{seg} = L_{iris} + \lambda_1 L_{outer} + \lambda_2 L_{inner}, \tag{1}$$

where L_{iris}, L_{outer}, L_{inner} indicate the dice loss [16] computed between the estimated and ground-truth values of the iris mask, inner mask, and outer mask, respectively, and λ_1 and λ_2 indicate the parameters for adjusting the weight of each loss.

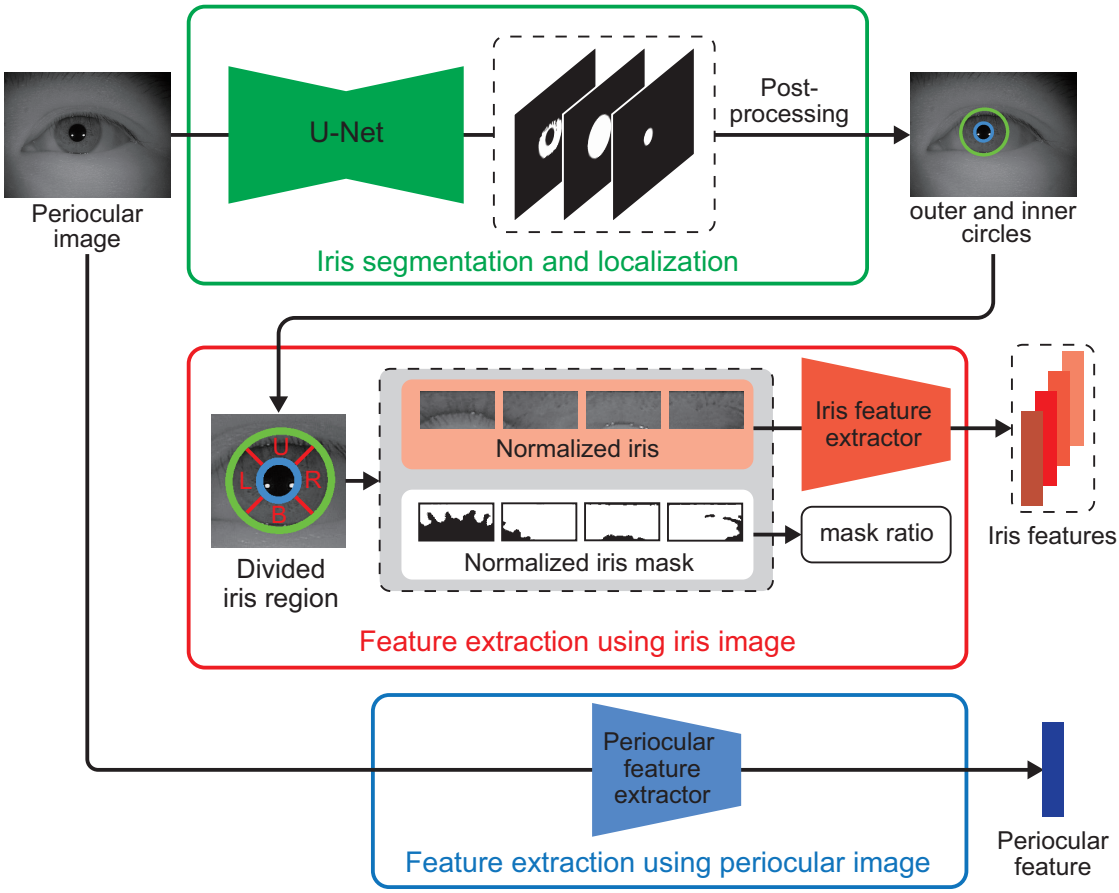


Fig. 1. Overview of the proposed method that combines periocular recognition and iris recognition.

The dice loss L_D is defined by

$$L_D = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2}, \tag{2}$$

where $p_i \in P$ indicates the pixel value of the estimated mask, $g_i \in G$ indicates the pixel value of the ground-truth mask, and N is the number of pixels in the mask. Although the number of pixels in the iris region and the number of pixels in the non-iris region are significantly different, the use of Dice loss allows for efficient training even when there is class bias.

B. Feature Extraction

We describe CNNs for feature extraction from iris and periocular images. In the proposed method, SE-ResNet-18 shown in Fig. 2 is used as feature extractors for iris and periocular images. The feature extractors are independent for the iris and periocular images as shown in Fig. 1. Their network architecture is the same, they do not share the weight parameters, and they are trained separately. The feature extractors extract 512-dimensional feature vectors from the iris and periocular images, respectively. The feature extractors are trained by inputting feature vectors to the classifier, which consists of a fully-connected layer and a softmax layer, and performing class classification. The cross entropy loss L_{CE} is used as the loss function. The feature extractor is trained to extract features that take into account the small differences in iris patterns by classifying the U, B, L, and R sub-regions as independent classes, even if the images are of the same person. Occlusions such as eyelids and eyeglasses may occupy the majority of the sub-regions obtained by iris segmentation. In particular, the U sub-region is often hidden by the eyelids and eyelashes. It is difficult to obtain features that are effective in identifying individuals from sub-regions that are mainly occlusions. Such low-quality training data degrades the CNN performance, and thus we have to select effective training data. To select sub-regions to be used for training, we use the mask ratio R defined by

$$R = \frac{A_{mask}}{A_{iris} + A_{mask}}, \tag{3}$$

where A_{iris} indicates an area of iris region, and A_{mask} indicates an area of non-iris region. The sub-regions with R higher than a threshold are excluded from the training data.

C. Data Augmentation

We introduce data augmentation to the training data to add variability to the periocular and iris images to improve the accuracy of the feature extractor. For the periocular image,

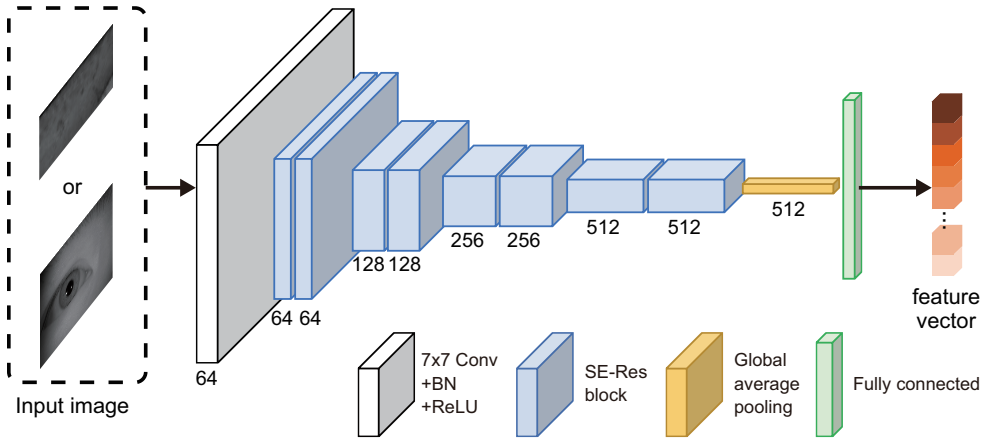


Fig. 2. Network architecture of the iris and periocular feature extractors, where the numbers under the blocks represent the number of output channels in each block.

we introduce the data augmentation of random crop, color jitter, random erasing, and random perspective. Random crop reproduces shifts in eye position by cropping random regions of the image. Color jitter reproduces changes in the ambient brightness of an image by randomly changing the brightness of the image. Random erasing replaces some pixel values in the image with zeros to reproduce the loss of periocular pattern due to glasses and others. Random perspective reproduces the orientation of a face by applying random deformations to the image. For iris images, we introduce the data augmentation of random erasing [17] and random horizontal shift. Random erasing is similar to data augmentation for periocular images. Random horizontal shift reproduces the rotation of the iris before rectangle normalization by randomly circularly shifting the image in the horizontal direction.

D. Matching

In the proposed method, the scores are calculated based on the features extracted from the iris and the periocular images, respectively, and the two matching scores are fused. Let us consider the image pair (α, β) to be verified. The iris features $v_U^\alpha, \ldots, v_R^\alpha$ and $v_U^\beta, \ldots, v_R^\beta$ are extracted from the single-regions, and the cosine similarity $S_U, \ldots, S_R$ are calculated for each sub-region. The iris similarity S_{iris} obtained by fusing $S_U, \ldots, S_R$ is defined by

$$S_{iris} = \frac{w_U S_U + w_B S_B + w_L S_L + w_R S_R}{w_U + w_B + w_L + w_R}, \tag{4}$$

where $w_U, \ldots, w_R$ indicate the weight according to the mask ratio for each sub-region and are defined by

$$w_X = \min(1 - R_X^\alpha, 1 - R_X^\beta), \tag{5}$$

where $X \in \{U, B, L, R\}$, and R_X^α and R_X^β indicate the mask ratio for region X in each image. The iris similarity S_{iris} is a weighted average of the similarity of each subregion based on the mask ratio, resulting in a score that gives more weight to subregions with less occlusion. The periocular features v_{peri}^α and v_{peri}^β are extracted from the periocular images, and the periocular similarity S_{peri} is calculated as the cosine similarity. The matching score S of the proposed method is defined by

$$S = (1 - \lambda) S_{iris} + \lambda S_{peri}, \tag{6}$$

where λ is a parameter within the range of 0 to 1.

IV. Experiments and Discussion

This section demonstrates the effectiveness of the proposed method through experiments using public datasets of periocular images.

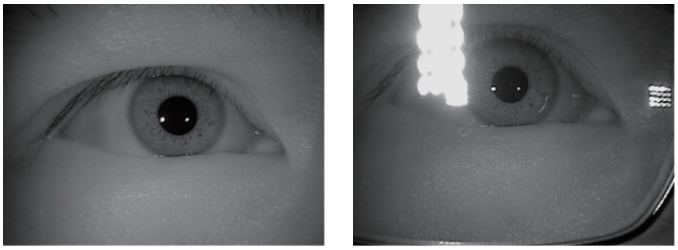
A. Dataset

In this experiment, we use three datasets provided by CASIA[2]: the CASIA-Iris-Thousand database, the CASIA-Iris-Distance database, and the CASIA-Iris-Mobile-V1.0 database. The details of the databases are described below.

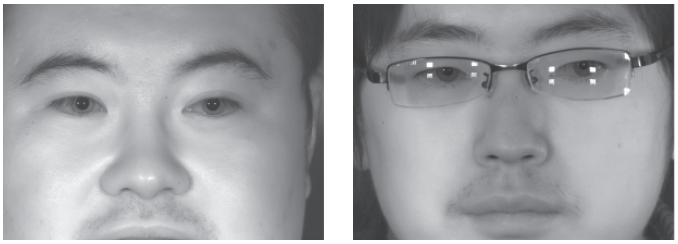
1) CASIA-Iris-Thousand (Thousand): Thousand consists of images taken around the eyes of subjects using a near-infrared camera, and contains a total of 20,000 images taking 10 images for each of the left and right eyes of 1,000 subjects. Fig. 3 (a) shows an example of images in Thousand. Most of the images show the iris clearly, although some images show the iris partially occluded or blurred by the reflections of the glasses. According to the evaluation protocol in [9], 13,980 images from subjects S5301 to S5999 were used for training and 6,020 images from subjects S5000 to S5300 were used for evaluation.

2) CASIA-Iris-Distance (Distance): Distance consists of images taken of the upper part of faces of the subjects with a near-infrared camera from a distance of 3m, and contains a total of 2,567 images taking from 142 faces. Fig. 3 (b) shows an example of images in Distance. Most of the images are blurred and do not show the iris clearly, making iris recognition particularly difficult. We detect the eyes using Mediapipe FaceMesh[3] and crop images around the eyes in the

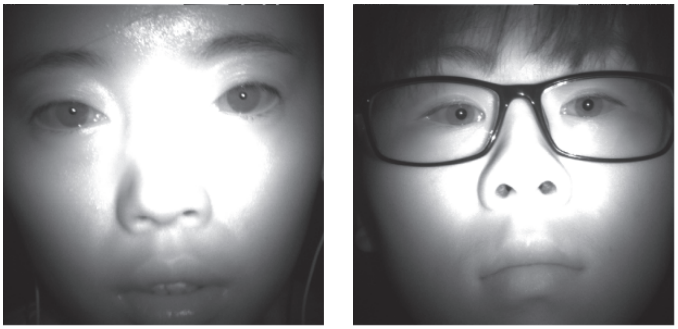
[2]http://biometrics.idealtest.org
[3]https://google.github.io/mediapipe



(a) CASIA-Iris-Thousand



(b) CASIA-Iris-Distance



(c) CASIA-Iris-M1-S3

Fig. 3. Examples of images in each dataset: (a) CASIA-Iris-Thousand (Thousand), (b) CASIA-Iris-Distance (Distance), and (c) CASIA-Iris-M1-S3 (M1-S3).

range of 640×480 pixels. According to the evaluation protocol in [13], 2,567 images of the right eye are used for training and 2,567 images of the left eye are used for evaluation.

3) CASIA-Iris-Mobile-V1.0 (M1-S3): The CASIA-Iris-Mobile-V1.0 database consists of three subsets. In this experiment, we use subset M1-S3, which consists of images taken from faces using cell phones with near-infrared cameras. M1-S3 contains a total of 3,600 images of 10 images each of 360 subjects. Fig. 3 (d) shows an example of the images in M1-S3. Some of the images clearly show the iris, while most of them are blurred. Similar to Distance, we detect the eyes using Mediapipe FaceMesh and crop images around the eyes in the range of 640×480 pixels. According to the evaluation protocol in [18], we use 3,600 images from subject M30001 to M30191 for training and 3,600 images from subject M30192 to M30380 for evaluation.

B. Evaluation Metrics

We describe the evaluation metrics used in this experiment. False Rejection Rate (FRR) is calculated from the matching score of genuine pairs and False Acceptance Rate (FAR) is calculated from the matching score of impostor pairs. We use Equal Error Rate (EER), which is equal to FRR and FAR. In practical authentication systems that require high security, such as immigration control systems, it is necessary to operate the system so that the FAR is low. Therefore, we use FRR@FAR0.01%, which is FRR when the threshold is set so that FAR is 0.01%. The smaller the value of EER and FRR@FAR0.01%, the higher the recognition accuracy. Note that the left and right eyes of the same person are considered as others, and the periocular image of one eye is used to determine the identity of the person in the experiments.

C. Training Setting

This section describes the training settings of the network used for iris detection and feature extraction. U-Net for iris detection is trained under the following conditions. The hyperparameters λ_1 and λ_2 of the loss function L are set to 1. The initial value of the learning rate is set to 0.0001, and the learning rate is multiplied by 0.1 if the loss to the validation data does not improve for 5 consecutive epochs. If the loss relative to the validation data does not improve for 30 consecutive epochs, training is terminated and the test data is evaluated using the weights from the epoch with the lowest loss. The batch size is set to 8 and Adam [19] is used as optimizer. Feature extractors for the iris and periocular images are trained under the following conditions. The initial value of the learning rate is set to 0.0001. If EER for the test data does not improve for 10 consecutive epochs, the learning rate is multiplied by 0.1. If EER does not improve for 30 consecutive epochs, training is terminated and evaluation is performed using EER at the minimum. All data augmentation methods are applied with a probability of 50%. The random horizontal shift is a circular shift within a range of 10 pixels on the left and right sides, respectively. Similar to traiing U-Net, the batch size is set to 8, and Adam is used as optimizer.

D. Combination of Periocular and Iris Recognition

We evaluate the effectiveness of the combination of periocular and iris recognition in the proposed method. In the proposed method, the matching scores calculated from iris and periocular are combined and used as the final matching score. We empirically explored the optimal value of λ by a grid search varying λ in Eq.(6) at intervals of 0.1 in the range [0.0, 1.0]. The closer λ is to 0, the stronger the matching score is for iris recognition, and the closer λ is to 1, the stronger the matching score is for periocular recognition. For Thousand, λ set to 0.4 resulted in the highest accuracy. The performance is improved by focusing on iris recognition, since the iris pattern is often clearly visible despite the occlusion caused by glasses and others. For Distance, λ set to 0.6 resulted in the highest accuracy. The performance is improved by focusing on periocular recognition since the iris pattern

TABLE I
COMPARISON OF PERIOCULAR RECOGNITION, IRIS RECOGNITION, AND OPTIMAL FUSION OF IRIS AND PERIOCULAR RECOGNITION, WHERE THE UNIT OF EACH VALUE IS %.

Modality	Thousand		Distance	
	EER	FRR @FAR0.01%	EER	FRR @FAR0.01%
Periocular	0.858	10.089	1.586	16.341
Iris	0.879	9.011	6.846	44.484
Periocular & Iris	**0.366**	**2.089**	**1.128**	**11.167**

TABLE II
EXPERIMENTAL RESULTS FOR DISTANCE, WHERE THE UNIT OF EACH VALUE IS %.

Method	Modality	EER	FRR@FAR0.01%
AttNet [12]	Periocular	14.27	38.4
OSIRIS [9]	Iris	17.13	100.0
UniNet [10]	Iris	5.54	24.7
UniNet+AttNet [13]	Periocular & Iris	2.29	13.7
Proposed	Periocular & Iris	**1.13**	**11.2**

is rarely captured clearly due to the distance between the camera and the subjects. Table I shows the comparison of iris recognition, periocular recognition, and optimal fusion of iris and periocular recognition. In both datasets, the combination of iris and periocular recognition significantly improves the accuracy.

E. Comparison with Conventional Methods

We compare the recognition performance of the proposed method with that of conventional methods using Distance and M1-S3. We chose Distance since many conventional methods have been evaluated. In Distance, we used the following conventional methods: AttNet [12] for periocular recognition, OSIRIS [9] for iris recognition, UniNet [10] for iris recognition, and UniNet+AttNet [10] for a combination of periocular and iris recognition. Table II shows the experimental results for Distance. The combination of periocular and iris recognition is more accurate than iris recognition alone. The proposed method has higher recognition accuracy than UniNet+AttNet [13], which combines periocular and iris recognition. In M1-S3, we evaluate the recognition accuracy when using one eye and both eyes. Note that the matching score is calculated by the average of the matching scores of the left and right eyes. Table III shows the experimental result for M1-S3. The proposed method has higher recognition accuracy in all cases than the method using Maxout-based CNN [18]. The proposed method can identify individuals with high accuracy even for low quality images such as those captured by mobile phone cameras.

V. CONCLUSION

We proposed an eye biometrics system that combines reliable iris recognition and convenient periocular recognition. The proposed method extracts iris and periocular features using SE-ResNet-18 and obtains the final matching score by fusing the matching scores calculated from the iris and periocular features. Through experiments using the public periocular image datasets: Thousand, Distance, and M1-S3, we demonstrated the effectiveness of the proposed method compared with conventional methods.

TABLE III
EXPERIMENTAL RESULTS FOR M1-S3, WHERE THE UNIT OF EACH VALUE IS %.

Method	Modality	EER	FRR@FAR0.01%
Maxout [18]	Periocular (single)	1.89	—
	Iris (single)	1.88	—
	Periocular & Iris (both)	0.60	—
Proposed	Periocular (single)	1.25	13.19
	Iris (single)	1.47	10.55
	Periocular & Iris (single)	0.66	4.74
	Periocular & Iris (both)	**0.57**	**3.79**

REFERENCES

[1] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer, 2008.

[2] K. W. Bowyer and M. J. Burge, *Handbook of Iris Recognition*. Springer, 2016.

[3] F. A. Fernandez and J. Bigun, "Survey on periocular biometrics research," *Pattern Recognition Letters*, vol. 82, no. 2, pp. 92–105, Oct. 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 770–778, Jun. 2016.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 7132–7141, Jun. 2018.

[6] S. Kawakami, H. Kawai, K. Ito, *et al.*, "A simple and accurate CNN for iris recognition," *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, pp. 1563–1568, Nov. 2022.

[7] J. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1148–1161, Nov. 1993.

[8] J. Daugman, "How iris recognition works," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 14, pp. 21–30, 1 Jan. 2004.

[9] N. Othman, B. Dorizzi, and S. Garcia-Salicetti, "OSIRIS: An open source iris recognition software," *Pattern Recognition Letters*, vol. 82, no. 2, pp. 124–131, Oct. 2016.

[10] Z. Zhao and A. Kumar, "Towards more accurate iris recognition using deeply learned spatially corresponding features," *Proc. Int'l Conf. Computer Vision*, pp. 3829–3838, Oct. 2017.

[11] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Patt. Anal. and Machine Intel.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[12] Z. Zhao and A. Kumar, "Improving periocular recognition by explicit attention to critical regions in deep neural network," *IEEE Trans. Information Forensics and Security*, vol. 13, no. 12, pp. 2937–2952, Dec. 2018.

[13] K. Wang and A. Kumar, "Periocular-assisted multi-feature collaboration for dynamic iris recognition," *IEEE Trans. Information Forensics and Security*, pp. 866–879, Sep. 2021.

[14] Z. Zhao and A. Kumar, "An accurate iris segmentation framework under relaxed imaging constraints using total variation model," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 3828–3836, Dec. 2015.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pp. 234–241, Oct. 2015.

[16] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proc. Int'l Conf. 3D Vision*, pp. 565–571, Jun. 2016.

[17] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *Proc. AAAI Conf. Artificial Intelligence*, pp. 13 001–13 008, Feb. 2020.

[18] Q. Zhang, H. Li, and Z. Sun, "Deep feature fusion for iris and periocular biometics on mobile devices," *IEEE Trans. Information Forensics and Security*, vol. 13, no. 11, pp. 2897–2912, Nov. 2018.

[19] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," *Proc. Int'l Conf. Learning Representations*, pp. 1–15, May 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Development of a Robust Ear Recognition Algorithm using Planar Approximation

Takahiko Arakawa*, Yuya Sato*, Hitoshi Sakano†, Tetsushi Ohki*‡,
* Shizuoka University, Shizuoka, Japan
E-mail: {arakawa,sato}@sec.inf.shizuoka.ac.jp, ohki@inf.shizuoka.ac.jp
† Shimane University, Shimane, Japan
E-mail: sakano@cis.shimane-u.ac.jp
‡ RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

***Abstract*—In recent years, multimodal authentication combining multiple authentication methods such as face recognition, iris recognition, and ear recognition has attracted attention. Ear recognition can achieve nearly 100% recognition accuracy in ideal conditions. However, it suffers from a significant decrease in recognition accuracy in the presence of occlusion, pose variations, and lighting variations. In this study, we focused on the influence of variations in ear poses on recognition accuracy. To address this issue, we employed a method of data augmentation for the training set, incorporating a variety of ear poses using plane approximation techniques. The results confirmed a noticeable improvement in recognition accuracy due to the inclusion of multiple poses.**

I. INTRODUCTION

Biometric authentication is a technology that uses biometric information to identify individuals. Compared to traditional authentication methods such as knowledge or possession-based authentication methods, biometric authentication is a convenient method that eliminates the need for memorization and the risk of loss, as it leverages individual characteristics [1]. In addition, biometric information has unique and immutable characteristics that reduce the risks associated with impersonation and theft.

However, there are challenges to its widespread adoption, such as the need for specialized sensors in certain cases, such as fingerprint recognition, or user resistance due to sensor contact. In light of these factors, facial recognition, a non-contact method that can be performed using simple sensors such as cell phone cameras, is gaining popularity. Facial recognition is primarily designed to recognize frontal faces or a significant portion of the face in input images. However, in real-world scenarios, it is not always guaranteed that a frontal face will be captured. For example, surveillance camera footage used in criminal investigations often captures only profiles, which significantly reduces the authentication accuracy [2]. To address this problem, an authentication method known as ear recognition has been proposed to improve authentication accuracy in such scenarios [3][4][5]. Ear recognition offers advantages over facial recognition because ear features exhibit minimal changes due to age, expression, or health conditions. In addition, ear images are expected to have fewer privacy concerns than facial images.

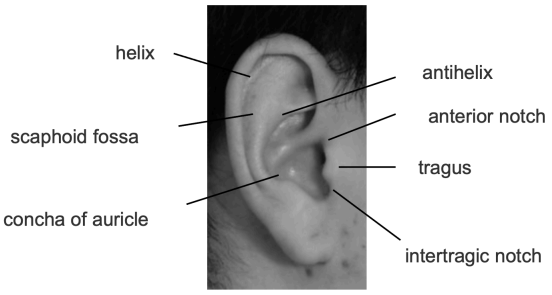


Fig. 1: The different parts of the ear

In this paper, we aim to construct an ear recognition system that achieves high accuracy in a wide range of capturing situations, including criminal investigations, by investigating methods to improve the robustness to capturing angles, which is a major challenge in ear recognition.

As shown in Fig. 1, the shape of the ear can be divided into several regions, including the tragus, concha of auricle, anterior notch, helix, scaphoid fossa, and intertragic notch. These regions exhibit variations in shape, appearance, and relative positions among individuals, leading to strong individuality.

In fact, studies conducted by Abaza et al. [6] and Emeršič et al. [7] have achieved near 100% identification rates in controlled experiments with controlled factors such as pose and lighting in ear recognition. Even in cases where suspects cover their faces with hats, sunglasses, masks, etc., and only the ears are observable, it is possible to identify individuals if the ears are clearly observed [8]. Furthermore, ear recognition can complement other biometric authentication methods, even when the face is visible, and it is being used in multimodal authentication systems [9].

On the other hand, ear recognition faces challenges related to occlusion, pose, lighting, capture devices, and variations in resolution, which can lead to a decrease in authentication accuracy. According to the research by Emeršič et al. [7], the recognition rates for ear images that were not controlled for occlusion, pose, lighting, capture devices, and resolution

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

were significantly lower than ear images captured in controlled environments suitable for existing ear recognition methods.

To address with these variations, there are two approaches: improving feature extraction methods and building robust recognition systems that can handle such variations, and training models with data that includes different variations to deal with them. In particular, in face recognition, an approach that constructs a 3D model of a person using some method and generates training data that includes various pose and lighting variations is used to build robust authentication systems [10]. The situation in ear recognition is considered similar to face recognition, and it is believed that using a 3D model-based approach can be effective in ear recognition as well.

On the other hand, in applications such as criminal investigations, there is a need for non-cooperative environments where it is difficult to perform 3D scanning of the subjects. Therefore, it is necessary to consider the generation of 3D models from captured 2D images. 1) Due to the limited number of registered images, only a few images are available for 3D modeling. 2) Ear images lack significant brightness gradients, and the features required for correspondence point extraction do not work effectively. These conditions make 3D reconstruction difficult in practice. Therefore, in this study, we propose a method based on *planar approximation*, which considers the non-planar ear as an object with additional convexity and concavity, aiming to approximate the ear to a planar object with texture applied to its surface. This method simplifies the process of data augmentation by using planar approximation techniques to manipulate the captured 2D ear images, rather than having to capture additional images or create 3D models. This can improve the effectiveness of ear recognition systems under practical conditions.

In Section II, we discuss related studies focusing on angle variations. Section III, we present our proposed method. Section IV, we describe the detail of evaluation experiments. Section V, we show the evaluation results and discuss their implications. Finally, Section VI, we address future challenges and provide a summary of our findings.

II. RELATEDWORK

In the field of ear recognition, some studies have been conducted using both three-dimensional models and two-dimensional models to address angle variations.

A. 3D model-based methods

One example of a study utilizing a three-dimensional model is the research conducted by Mursalin et al. [11]. They employed a 3D model of the ear to extract shape features such as curves, contours, and concavities. The registered ear model is then compared with the input ear model to evaluate the authentication accuracy. This database consists of images of the ears of selected volunteers taken from a variety of angles, including frontal, lateral, and oblique, under varying distance and lighting conditions. Three-dimensional models of the ears were created using specialized imaging equipment. However, these images were not taken under non-cooperative conditions, a practical scenario often encountered in ear recognition.

B. 2D models-based method

One example of a study utilizing a two-dimensional model is the research conducted by Watabe et al. [12]. They used Gabor filters to extract features from collected ear images in a database that produced angle variations. Their method assuming that the given database had known angle variations and extracting features through the transformation using Gabor filters. However, in practical scenarios, it is anticipated that estimating angles itself could be challenging, especially in non-cooperative environments. Additionally, the HOIP face image database used in this study, which included angle variations of the ear, has been discontinued. Regarding these prior researches, we propose a ear recognition method that takes into account angle variations in ear images captured in non-cooperative environments. Our approach involves approximating the ear to a planar surface to simulate pseudo pose variations. We experimentally show the effectiveness of this method.

III. METHOD

A. Planar approximation

As shown in Fig. 1, the human ear exhibits various concavities and convexities. This characteristic gives rise to regions that are impossible to capture (referred to as self-occlusion). The patterns of self-occlusion are influenced by the depth of concavities and the angle variations during image acquisition, leading to a decrease in authentication accuracy. In this paper, we propose a method for generating pseudo-rotated ear images that take into account various shooting angles and self-occlusion in non-cooperative environments. By utilizing these images for classifier training, we achieve a high-accuracy ear recognition system. We employ a planar approximation technique regarding the ear, which is not strictly flat, as an object with a textured surface, thereby approximating it to a planar object. Here, Fig. 2 illustrates an image of the ear captured from the side view. On the other hand, Fig. 3 depicts the same ear after being pressed against a transparent acrylic plate to eliminate the influence of concavities. It can be observed that the ear shape is maintained with similarity between Fig. 2 and Fig. 3

Indeed, this indicates that the concavities and convexities present on the surface of the ear are relatively small compared to the overall size of the ear. In this case, by using a planar approximation image derived from the three-dimensional shape, it becomes feasible to create pseudo-rotated ear images that account for the rotation of the ear.

B. Pseudo-rotation image generation method

The overview of the method for generating pseudo-rotated ear images is depicted in Fig. 4. In this approach, initially, an ear image captured from the lateral perspective is projected onto a virtual planar. Subsequently, the planar is rotated in a three-dimensional space to produce an image resembling



Fig. 2: A normal ear



Fig. 3: A ear pressed into a transparent acrylic plate

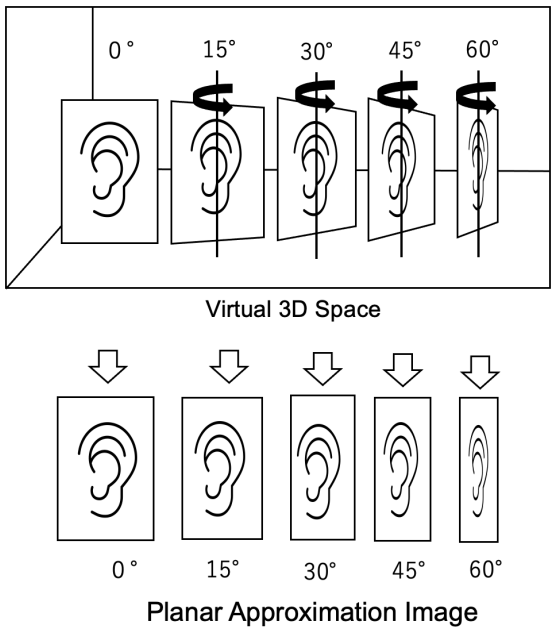


Fig. 4: Generation of Psudo spin ear image

the frontal view. This process enables the creation of pseudo-rotated ear images within the applicable range of the planar approximation.

In the proposed method, we assume a model that rotates an ear image on the face of a cylinder with radius r, using an axis that passes vertically through the cylinder. Here, the distance r between the rotation axis and the face determines the image generated after the rotation. In this method, we set r as zero.

IV. Experiment

We experimentally investigated the effectiveness of the proposed method. In the following sections, we will describe the database used, the detailed settings of the ear recognition system, and the procedure for generating pseudo-rotated images.

A. Environment

We have conducted the following experiment on a system running Windows 10, with an Intel(R) Core(TM) i5-8350U CPU @ 1.70GHz 1.90 GHz processor and 8.00 GB of installed RAM.

B. Database

Based on the best of our knowledge, there is no explicitly controlled database specifically for ear rotation. The AWE database (Annotated web ear database extended) is available and includes ear images with uncontrolled rotations. Emersic et al. constructed this unconstrained ear image database from images of celebrities available online[7]. The AWE database comprises 1,000 images, with 10 samples per individual for 100 individuals, simulating realistic scenarios with various disturbances that could occur in ear recognition systems. However, the database contains a mix of diverse disturbances, making it difficult to determine the specific disturbances that negatively affect the authentication accuracy. Nevertheless, the images in this database include intra-image rotation, yaw rotation, and pitch rotation of the ears. If the proposed data augmentation method demonstrates an improvement in authentication accuracy, it would serve as evidence of the effectiveness of the proposed approach.

C. Ear recognition system

The ear recognition system consists of preprocessing, shape descriptor, and classifier components.

1) Preprocessing: In this study, the objective is to investigate the impact of angle variations. Therefore, the normalization process, which corrects for angle variations, was not performed during the preprocessing stage. Instead, the image sizes varied during the database creation and were standardized to 128 $\times$ 128px.

2) Shape descriptor: Histogram of Oriented Gradients (HOG) was used, which is a well-known and effective shape descriptor for ear recognition[13].

3) Classifier: As for the classifier, we employed a 1-Nearest Neighbor (1-NN) classifier based on Euclidean distance. The 1-NN classifier is advantageous in that it can operate effectively even with a small number of training samples and allows for easy reproducing.

D. Protocol

1) Training: For the training phase, 9 samples per person were randomly selected from the AWE database, resulting in a training dataset of 100 individuals with 9 samples each. To introduce angle variations, the proposed method described in Section III was used for data augmentation. This method generated pseudo-rotated images at 0 degrees, 15 degrees, 30 degrees, 45 degrees, and 60 degrees from the frontal view for each sample. Consequently, the training dataset consisted of 4,500 images (100 individuals $\times$ 9 samples $\times$ 5 angles).

TABLE I: Rank-1 recognition rate: Compare the probability of the label determined to be the closest in the 1-NN classifier being the true label between the proposed method and the baseline.

	Proposed	Baseline
Rank-1 recognition rate	40.1%	35.5%

2) Testing: For the testing phase, the remaining data from the AWE database ware used. One sample per person, without pseudo-rotation, was selected, resulting in an evaluation dataset of 100 images(100 individuals $times$ 1 samples $times$ 1 angles). To demonstrate the effectiveness of the proposed method, the performance was compared to a baseline approach, which used a 1-NN classifier without the inclusion of pseudo-rotated images in the training data. The recognition rates of the proposed method and the baseline approach were compared. To ensure the generalization of the authentication results, 10-fold cross-validation was performed.

E. Metric

The evaluation metric used was the Rank-1 recognition rate, which represents the probability that the label output by the 1-NN classifier matches the true label.

V. Results and discussions

A. Qualitative evaluation with actual captured ear

Fig.5 shows (a) images of the human ear captured at various rotation angles, and (b) images generated by planar approximation based on a 0-degree actual image. Each column represents images rotated by 0 degrees, 15 degrees, 30 degrees, 45 degrees, and 60 degrees, respectively. Fig.5 shows differences due to the shadowing of the helix and tragus. However, the shape changes are not significant and the influence of self-occlusion appears to be small. Therefore, particularly in the case of actual ear image datasets with little rotation variation, it is expected that the use of plane approximation image data as training data will contribute to improved accuracy.

B. Rank-1 recognition rate

Table I shows the Rank-1 recognition rates described in IV-E. According to these results, the baseline method achieved a 1st rank authentication accuracy of 72.7%, while the proposed method achieved an accuracy of 75.2%. This demonstrates the effectiveness of the proposed approach.

In the following discussion, we will examine to what extent the improvement in accuracy is due to the planar approximation.

C. Limitations

1) Other rotation angles: In this study, we applied a planar approximation focusing solely on lateral rotation. This consideration is based on the premise that surveillance camera footage commonly used in criminal investigations is likely to capture a significant number of ear images that include lateral rotation. On the other hand, the AWE database used in this study includes not only lateral rotation but also rotations in the Roll and Pitch directions. Furthermore, actual ear images taken in practical situations are expected to contain even more diverse rotations in the yaw, roll and pitch directions than those in the AWE database used in this study.

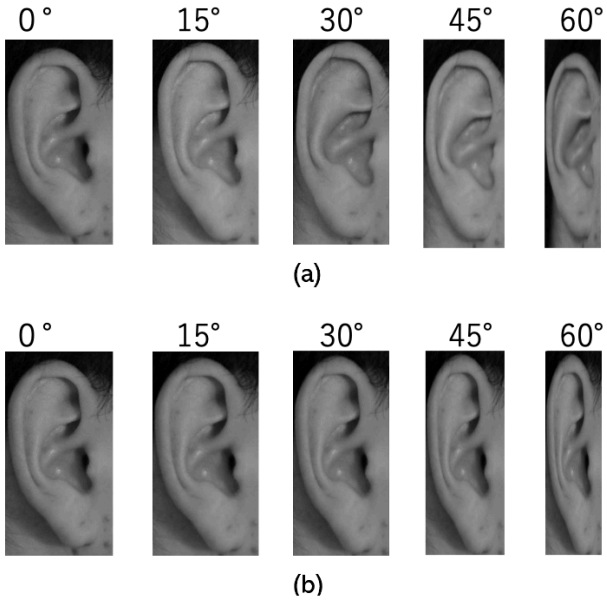


Fig. 5: Comparison with actual ear

(a) represents actual images of a person's ear captured by physically rotating it. (b) shows images of the person's ear captured at 0 degrees and then simulating angle variations using a planar approximation. Each column, from left to right, represents images rotated at angles of 0 degrees, 15 degrees, 30 degrees, 45 degrees, and 60 degrees, respectively.

In our experiment, as shown in I, we were able to achieve a 4.6% of improvement in accuracy by increasing the variations in the yaw direction, demonstrating the effectiveness of data augmentation using the proposed method. However, considering the roll and pitch directions in the plane approximation method may lead to further improvements in accuracy. These are issues that need to be addressed in future research.

2) Occurrence of self-occlusion: The proposed method assumes that the influence of the depth of the ear's concavity on the images, as shown in Fig. 2 and Fig. 3, is below a certain threshold, making self-occlusion less likely to occur and enabling the application of planar approximation. Therefore, it is expected that subjects with extremely large concavities in their ears or difficulties in accommodating certain levels of angle variations may pose challenges for the method. However, the optimal determination method for handling such cases has not been discussed in this paper, and further investigations are left as future work.

3) Other factors affecting accuracy: In addition to the angle variations addressed in this study, there are various factors in real-world environments that can reduce the accuracy of ear recognition, including occlusion, lighting conditions,

acquisition devices, and variations in resolution. To address these elements, it is necessary to develop more robust feature extraction systems and train them on data that encompasses various variations. By addressing these challenges, it is believed that the development of more practical ear recognition technology can be achieved.

VI. Conclusion and Future Work

In this paper, we proposed a novel data augmentation method for ear recognition using planar approximation and conducted authentication experiments based on the generation of rotated ear images. The results demonstrated the effectiveness of the proposed method. In future work, we plan to evaluate other types of rotations such as roll rotation, pitch rotation, in addition to the yaw rotation. We also aim to create an ear image database with controlled pose variations and further experimentally verify the effectiveness of the proposed method. Furthermore, we would like to explore the application of planar approximation to address lighting variations. Overall, the findings will open up new areas of research for developing more robust and accurate ear recognition systems.

ACKNOWLEDGEMENTS

This study was supported in part by JSPS KAKENHI JP23H00463 and JP23H03395.

References

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[3] M. Choraś, "Perspective methods of human identification: Ear biometrics," *Opto-electronics review*, vol. 16, no. 1, pp. 85–96, 2008.

[4] Ž. Emeršič, D. Štepec, V. Štruc, *et al.*, "The unconstrained ear recognition challenge," in *2017 IEEE international joint conference on biometrics (IJCB)*, IEEE, 2017, pp. 715–724.

[5] Ž. Emeršič, A. K. SV, B. Harish, *et al.*, "The unconstrained ear recognition challenge 2019," in *2019 International Conference on Biometrics (ICB)*, IEEE, 2019, pp. 1–15.

[6] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM computing surveys (CSUR)*, vol. 45, no. 2, pp. 1–35, 2013.

[7] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[8] I. Alberink and A. Ruifrok, "Performance of the fearid earprint identification system," *Forensic science international*, vol. 166, no. 2-3, pp. 145–154, 2007.

[9] Y. Ma, Z. Huang, X. Wang, and K. Huang, "An overview of multimodal biometrics using the face and ear," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–17, 2020.

[10] D. Smeets, P. Claes, D. Vandermeulen, and J. G. Clement, "Objective 3d face recognition: Evolution, approaches and challenges," *Forensic science international*, vol. 201, no. 1-3, pp. 125–132, 2010.

[11] M. Mursalin, M. Ahmed, and P. Haskell-Dowland, "Biometric security: A novel ear recognition approach using a 3d morphable ear model," *Sensors*, vol. 22, no. 22, p. 8988, 2022.

[12] D. Watabe, T. Minamidani, H. Sai, T. Maeda, T. Yamazaki, and J. Cao, "Estimating a shooting angle in ear recognition," in *Computer Information Systems and Industrial Management: 14th IFIP TC 8 International Conference, CISIM 2015, Warsaw, Poland, September 24-26, 2015, Proceedings 14*, Springer, 2015, pp. 559–568.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Ieee, vol. 1, 2005, pp. 886–893.

Word encoding for word-looking DGA-based Botnet classification

Sea Ran Cleon Liew* and Ngai Fong Law
* University of Waterloo, Canada
E-mail: srcliew@uwaterloo.ca
The Hong Kong Polytechnic University, Hong Kong
E-mail: ennflaw@polyu.edu.hk

Abstract — **There are two main types of domain name-generating algorithms (DGAs) – random-looking and word-looking. While existing methods can effectively distinguish between the two types of DGAs with high accuracy, classifying different types of word-looking DGAs has proven to be challenging, as they are often mistaken for legitimate domains. To address this issue, previous methods used character encoding with long short-term memory networks (LSTM) or convolutional neural networks (CNN) to model the character distribution of different word-looking DGAs. Since most word-looking DGAs are constructed using various dictionaries, we propose using word encoding instead of character encoding. Word encoding can provide a better characterization as it is based on the usage of different words in the dictionaries and their associations. Experimental results show that the classification accuracy for word-based DGAs increases by more than 7% (from 87% to 94%) using word encoding as compared to character encoding.**

I. INTRODUCTION

Domain name generation algorithms (DGAs) are techniques to generate a large number of domain names to facilitate the establishment of covert communication between bots and the command and control (C&C) server. By constantly changing domain names in communication, it makes the tracking and identification of malicious network traffic difficult. In this way, DGAs help to prevent the takedown of the C&C servers [1].

There are two main types of DGA: random-looking and word-looking [2-3]. Random-looking DGAs are produced by concatenating random strings of characters. In contrast, the word-looking DGAs are generated from concatenating words taken from a dictionary. Various methods have been proposed for detecting and classifying DGAs. Detection means to determine if the domain names are legitimate or generated by the DGA-based botnets, while classification identifies which method has been used to generate the DGA. Hence detection is a binary problem while classification is a multi-class problem. As reported in [4], the accuracy of the binary detection and multi-class classifications are over 90% and around 70% respectively using machine learning-based (ML) algorithms. Thus, it is much more difficult to classify the DGAs than to detect these domains from legitimate domains [5-7].

Various methods, including ML and deep learning-based (DL) approaches, have been proposed for detecting and classifying these domain names. In ML approaches, features about character distributions in the domain names have been used. Examples include vowels and consonant ratios or the numerals and English characters ratios [8]. Most of the extracted features were primarily targeted at random-looking domains and are effective in characterizing the randomness in the domain names. It, however, resulted in inferior performance in characterizing word-looking domain names [6]. In contrast, existing DL methods explore character encoding to characterize the character distribution and association within the domain names. Thus, DL methods are more effective in characterizing word-looking domain names than ML methods. Despite that, word-looking DGAs are produced by using different dictionaries. Instead of using character encoding, word-level encoding can better characterize the nature of the words used in the dictionary and their associations. We thus explore the use of word encoding and DL techniques for word-looking DGA classification.

This paper is organized as follows. First, a comparative study of existing methods is given in Section II. Then, our proposed method is described in Section III. Experimental results are given in Section IV. Finally, we conclude our work in Section V.

II. STUDY OF EXISTING METHODS

Domain name generation algorithms (DGAs) detection and classification are important problems in network security, as

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

these algorithms generate a large number of domain names to facilitate covert communication between bots and command and control servers. Table 1 illustrates the differences between the DGA detection and the classification problems. In detection, we only target to distinguish the legitimate and the DGAs. However, in classification, we want to know which algorithm has been used to generate the DGAs too.

A way to detect DGAs is to check if the domain names can be resolved into IP addresses successfully. This is based on the fact that only a small number of the generated domain names are registered. If it cannot be resolved, an error signal called NXDomain response will be received. Thus, one approach to DGA detection is to analyze network traffic characteristics and detect the error signal returned through network communication. However, the network traffic analysis can only detect DGAs, but not classify them. Classifying DGAs requires analyzing the patterns of the algorithmically generated domain names. To address this issue, both ML and DL methods have been used for DGA classification [5-12].

Feature extraction is a crucial step in ML methods. Features representing expert knowledge are extracted from the domain names to characterize the nature of domain names. Common types of features include statistical features, information theory features, and lexicographic features. For example, features like vowel-consonant ratio, TF-IDF features, n-gram distributions, the longest consecutive consonant/number/vowel sequences, pronounceability score, and entropy [9, 13-14] have been used to characterize the nature of the algorithmically generated domain names. While these features are somewhat effective in characterizing random-looking DGAs, accurately classifying word-looking DGAs remains a significant challenge. As reported in [4], the average F1 scores for classifying 11 word-looking DGAs are 0.54, 0.57, and 0.68 for kNN, decision tree, and random forest respectively. This suggests that the extracted features are not sufficient for characterizing different DGA classes. This is consistent with findings from other authors [6]. One main reason for this is that words in word-looking DGAs of different dictionaries are indeed valid English words. Thus, extracted features across different dictionaries would likely be similar.

DL has also been used for DGA classification. Unlike ML approaches, DL does not require feature extraction. Instead, the domain name is treated as a string of characters. With the use of a sufficient number of examples, a learning model is trained to distinguish and characterize the DGAs. As domain names are made up of characters, character encoding is commonly used to convert the domain names to numerical sequences. This process involves breaking down the domain name into individual characters and representing these characters as numerical values.

Convolutional neural networks (CNN) and long short-term memory networks (LSTM) are popular DL models [15-17]. LSTM is often used for acquiring patterns in long sequences [7, 10-11, 18]. CNN, on the other hand, uses a filter kernel with varied sizes to characterize sequential relationships [19-20]. Previous studies have shown that combining CNN and LSTM can improve detection and classification performance [6, 21]. According to a study in [5], the average F1 scores for classifying eleven word-looking DGAs are 0.87 and 0.90 for character-based Bi-LSTM and CNN-Bi-LSTM respectively. Comparing the performance of ML and DL approaches, we can see that the DL approach has shown a significant improvement in accurately classifying word-looking DGAs. It is promising to investigate whether the deep learning techniques can be improved further to classify the word-looking DGAs.

Domain name	Detection (Binary)		Classification (multi-class)	
	Class	Label	Class	Label
Polyu	Legit	0	legit	0
Pointreply	DGA	1	suppobox2	1
Duringsuppose	DGA	1	pizd	2
Somewhatlemon	DGA	1	nymaim	3

Table 1: DGA detection and classification problems.

III. PROPOSED METHODS

In existing character-based DL models for DGA classification, characteristics among the connecting characters are considered. However, in word-looking DGAs, words from different dictionaries are concatenated. Table 2 shows details regarding the dictionary used in constructing word-looking DGAs. As all dictionaries contain valid English words, character-based encoding may not be effective in distinguishing the English words in different dictionaries.

To better capture the linguistic and semantic structure [22, 23] of the word-looking DGAs, a word-based encoding method can be used. In this section, we discuss a DL method for characterizing the word relationships in word-looking DGAs. In particular, we consider both CNN and LSTM models. These models can learn to identify patterns and relationships among words in a way that is not possible with character-based encoding. In this way, the word-looking DGA classification can be done more accurately.

To extract the relevant parts of domain names from URLs, we followed the procedures outlined in [20]. If the URL contains a second-level domain name, we extract the second-level part. If the URL contains a third-level domain name, we check if the second-level domain name is from a popular dynamic domain name service such as "no-ip.com" or "ddns.net". If it is, we extract the third-level domain part. If the second-level domain name is not from a popular dynamic domain name service, we extract the longer string consisting of the second-level and the third-level domain names. This procedure helps extract the most relevant parts of the domain name while avoiding any unnecessary information.

DGAs	Dictionaries
gozi	gpl: 4379 words, luther: 1537 words nasa: 558 words, rfc: 2460 words
matsnu	Use two dictionaries: verb dictionary (878 words) and noun dictionary (1008 words)
nymaim	Use two dictionaries: the first has 2450 words and the second has 4387 words
pizd	The dictionary contains 384 words
rovnix	Uses the US Declaration of Independence as the dictionary
suppobox	The dictionary in all versions contains 384 words

Table 2: Dictionaries used in different DGAs

The next step in processing the domain names is to decompose them into words, which can help capture the semantic meaning of the domains. For example, the domains "preparetwenty" and "coveredpublicandfrom" can be decomposed into {"prepare", "twenty"} and {"covered", "public", "and", "from"} respectively. These words are valid English words that carry meaning and can provide important information for DGA classification. The subsequent DL model will be trained to characterize the association among the connecting words. Both CNN and Bi-LSTM will be used for DGA classifications. CNNs are effective at learning local features and patterns in the data [15-17], while Bi-LSTM can capture sequence dependencies and relationships among the words [22-23].

The proposed CNN and Bi-LSTM structure for DGA classification consists of an embedding layer with an M number of units, which maps each word to a vector representation. For CNN, the convolutional layer has 128 filters and a kernel size of 2, which applies the filter to the word embeddings to extract local features. The Bi-LSTM layer consists of 200 units to capture long-term dependencies and relationships among the words in the domain names. The output of the convolutional layer or the Bi-LSTM layer is then passed through three fully connected layers with 256, 128, and 128 units respectively, before being fed into the output layer with L units, corresponding to the L number of classes in the DGA classification.

IV. Experimental results

UMUDGA is a public dataset designed for detecting and profiling algorithmically generated domain names in DGA-based botnet detection [4]. It contains over 30 million domain names from 50 DGA classes. In addition to the domain names, UMUDGA contains also basic statistical features, n-gram features, and NLP (natural language processing) features. In our study, we focus on the word-looking DGAs classification. To train and evaluate our proposed models, the dataset contains 10,000 legitimate domain names and 10,000 algorithmically generated domains for each of the word-looking DGA classes. This dataset allows us to test the effectiveness of our models in classifying different types of word-looking DGAs.

To evaluate and compare the performance, the precision, recall, and F1 scores are used. Precision is defined as the ratio between TP and (TP+FP) where TP is the number of samples that are in class A and are identified as A, and FP is the number of samples that are not in class A but are identified as A. Recall is the ratio between TP and (TP+FN) where FN is the number of samples that are in class A but is identified as not in A. In a perfect classification, both precision and recall should be equal to 1. In practice, increasing precision may decrease recall. The F1 score is the weighted average of precision and recall.

Table 3 shows the average F1 scores of the proposed CNN and Bi-LSTM models. We can see that both models achieved an average F1 score of 0.94 for DGA classifications. We compared our proposed models with both ML models and other character-based DL models. For DL models such as CNN (character), LSTM.MI [24], and Bi-LSTM [5], their average F1 scores are smaller than 0.88. Hence, our proposed models can achieve an increase of at least a 7.7% improvement in the F1 score. The improvement was much larger as compared to the ML models such as RF, DT, and kNN. Hence word encoding is beneficial for word-looking DGAs classification.

Table 4 and Table 5 show respectively the percentage improvement in the F1 scores of the proposed CNN and Bi-LSTM models against different ML and DL models for different DGAs. It clearly shows that deep learning models can always outperform machine learning models. Compared to

other character-based DL models, our proposed word encoding CNN and Bi-LSTM improve the F1 scores for all classes.

Besides, F1 scores, precision, and recall are also considered. Fig. 1 and Fig. 2 show respectively the boxplot of the precision and recall achieved by different models. Consistent with the F1 scores results, DL models have better performance than the ML models. In addition, word encoding methods perform better than character encoding methods.

	F1 score	
	Legitimate	DGAs
Proposed (CNN)	0.805	0.936
Proposed (Bi-LSTM)	0.773	0.944
CNN (character)	0.711	0.854
LSTM.MI [24, 5]	0.711	0.866
Bi-LSTM [5]	0.721	0.874
Random Forest (RF) [4]	0.662	0.682
Decision Tree (DT) [4]	0.582	0.572
kNN [4]	0.484	0.540

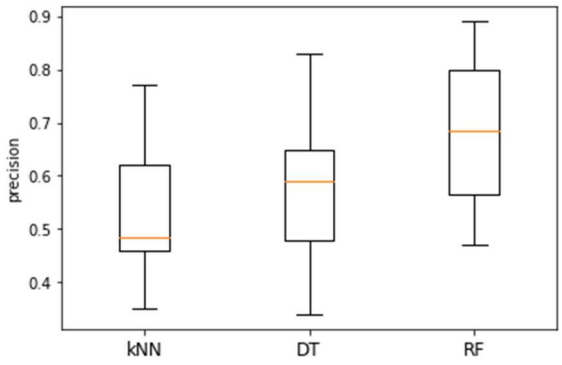
Table 3: The average F1 scores for the proposed CNN and Bi-LSTM models in comparison with other character-based deep learning models and machine learning models.

	kNN [4]	DT [4]	RF [4]	CNN (char)	LSTM .MI [24, 5]	Bi-LSTM [5]
legit	67	38	21	13	13	11
goziGpl	89	41	9.2	30	19	12
goziLuther	83	94	42	6.7	5.8	2.4
goziNasa	143	150	88	18	15	7.4
goziRfc	183	158	98	23	21	14
matsnu	20	20	7.9	6.7	6.1	8.0
nymaim	130	64	29	14	11	12
pizd	87	98	65	4.2	4.4	7.7
rovnix	92	68	45	12	9.6	6.8
suppobox1	102	106	70	2.1	3.0	6.5
suppobox2	24	24	18	2.1	3.1	4.2
suppobox3	23	20	11	1.0	0.6	0.7
Average	87	73	42	11	9.2	7.7

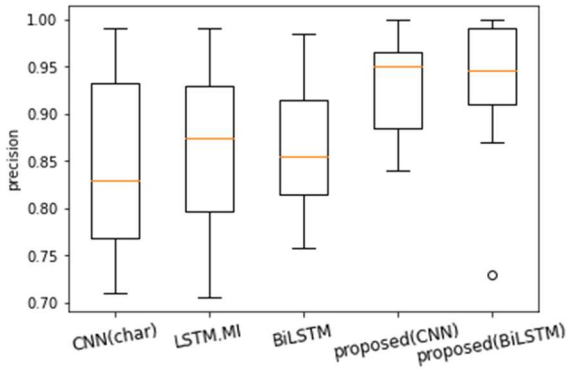
Table 4: The percentage improvement in the F1 score of the proposed CNN model against different ML and DL models for both legitimate and DGA classes.

	kNN [4]	DT [4]	RF [4]	CNN (char)	LSTM .MI [24, 5]	Bi-LSTM [5]
legit	60	33	17	8.5	8.3	6.8
goziGpl	93	44	12	33	21	15
goziLuther	83	94	42	6.7	5.8	2.4
goziNasa	158	156	92	21	17	9.8
goziRfc	186	160	100	25	23	15
matsnu	20	20	7.9	6.7	6.1	8.0
nymaim	133	65	30	15	12	13
pizd	87	98	65	4.2	4.4	7.7
rovnix	92	68	45	12	9.6	6.8
suppobox1	106	111	74	4.2	5.1	8.7
suppobox2	24	24	18	2.1	3.1	4.2
suppobox3	23	21	11	1.0	0.6	0.7
average	88	74	43	12	9.7	8.2

Table 5: The percentage improvement in the F1 score of the proposed Bi-LSTM model against different ML and DL models for both legitimate and DGA classes.

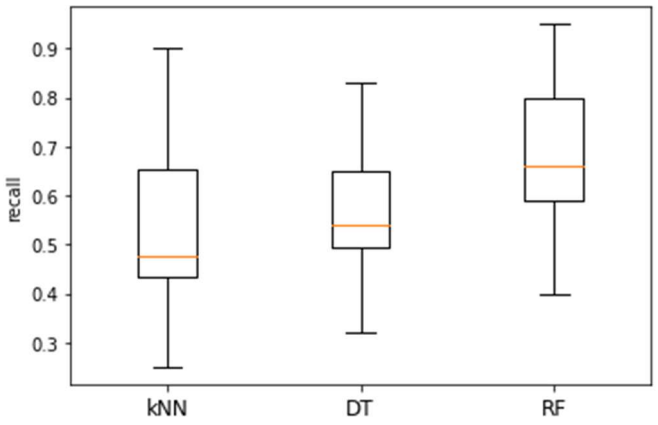


(a)

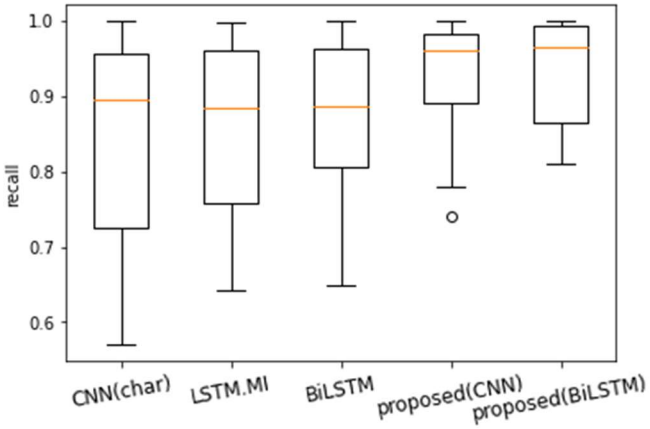


(b)

Fig 1: Boxplots of the precision achieved by (a) ML models such as kNN, DT, and RF, and (b) DL models such as character encoding CNN, LSTM.MI, Bi-LSTM, the proposed CNN model, and the proposed Bi-LSTM model.



(a)



(b)

Fig 2: Boxplots of the recall achieved by (a) ML models such as kNN, DT, and RF, and (b) DL models such as character encoding CNN, LSTM.MI, Bi-LSTM, the proposed CNN model, and the proposed Bi-LSTM model.

V. CONCLUSIONS

Existing state-of-the-art DGA classifications are ineffective at classifying word-looking DGAs. It is because word-looking DGAs are constructed by concatenating words from dictionaries. Features used in machine learning models are not designed to identify such patterns. The character encoding adopted by existing deep learning models achieves a better performance. To further improve the performance, the word encoding technique is applied to study the association of different words in the domain names. Experimental results show that the proposed word encoding-based CNN and Bi-LSTM models achieve better classification performance than both the machine learning models and character encoding-based deep learning models.

REFERENCES

[1] G. Vormayr, T. Zseby, and J. Fabini, "Botnet Communication Patterns", IEEE Communications Surveys & Tutorials, 19, 2768-2796, 2017.

[2] Z. Wang, and Y. Guo, "Neural networks based domain name generation", Journal of Information Security and Applications, 61, 102948, 2021.

[3] T. Wang, L. Chen, and Y. Genc, "A dictionary-based method for detecting machine-generated domains", Information Security Journal: A Global Perspective, Vol. 30, Issue 4, 2020.

[4] M. Zago, M. G. Pérez, and G. M. Pérez, "UMUDGA: A dataset for profiling DGA-based botnet", Computers & Security, 92, 101719, 2020.

[5] A. Cucchiarelli, C. Morbidoni, L. Spalazzi, and M. Baldi, "Algorithmically generated malicious domain names detection based on n-grams features", Expert Systems with Applications, 170, 114551, 2021.

[6] F. Ren, Z. Jiang, X. Wang, and J. Liu, "A DGA domain names detection modeling method based on integrating an attention mechanism and deep neural network", Cybersecurity, 3, 1-13, 2020.

[7] P. Vij, S.D. Nikam, and A. Bhatia, "Detection of Algorithmically Generated Domain Names using LSTM", International Conference on COMmunication Systems & NETworkS (COMSNETS), 1-6, 2020.

[8] H.P. Vranken, and H. Alizadeh, "Detection of DGA-Generated Domain Names with TF-IDF", Electronics, 11, 414, 2022.

[9] A.O. Almashhadani, M. Kaiiali, D. Carlin, and S. Sezer, "MaldomDetector: A system for detecting algorithmically generated domain names with machine learning", Computers & Security, 93, 101787, 2020.

[10] Y. Qiao, B. Zhang, W. Zhang, A.K. Sangaiah, and H. Wu, "DGA Domain Name Classification Method Based on Long Short-Term Memory with Attention Mechanism", Applied Sciences, 9, 4205, 2019.

[11] J.P. Selvi, R.J. Rodríguez, and E. Soria-Olivas, "Toward Optimal LSTM Neural Networks for Detecting Algorithmically Generated Domain Names", IEEE Access, 9, 126446-126456, 2021.

[12] Z. Wang, Y. Guo, and D. Montgomery, "Machine learning-based algorithmically generated domain detection", Computers and Electrical Engineering, 100, 107841, 2022.

[13] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon, "From Throw-Away Traffic to

Bots: Detecting the Rise of DGA-Based Malware", USENIX Security Symposium, 24, 2012.

[14] L. Bilge, S. Şen, D. Balzarotti, E. Kirda, and C. Krügel, "Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains", ACM Transactions on Information Systems Security, 16, 14, 2014.

[15] Y. Liu, Z. Zhou, Y. Yang, N.F. Law and A.A. Bharath, "Efficient Source Camera Identification with Diversity-enhanced Patch Selection and Deep Residual Prediction", Sensors, 21(14), 4701, 2021.

[16] A. Zhang, Z. Zou, Y. Liu, Y. Chen, Y. Yang, Y. Chen, N.F. Law, "Automated Detection of Circulating Tumor Cells using Faster Region Convolution Neural Network", Journal of Medical Imaging and Health Informatics, 9(1), 167-174, 2019.

[17] M. Irshad, N.F. Law, K.H. Loo and S. Haider, "IMGCAT: an Approach to Dismantle the Anonymity of a Source Camera using Correlative Features and an Integrated 1D Convolutional Neural Network", Array, 18, 100279, 2023.

[18] J. Woodbridge, H. Anderson, A. Ahuja, and D. Grant, "Predicting Domain Generation Algorithms with Long Short-Term Memory Networks", ArXiv, abs/1611.00791, 2016.

[19] Z. Feng, C. Shuo, and W. Xiaochuan, "Classification for DGA-Based Malicious Domain Names with Deep Learning Architectures", International Journal of Intelligent Information Systems, Vol. 6, Issue 6, 67-71, 2017.

[20] B. Yu, D.L. Gray, J. Pan, M.D. Cock and A.C. Nascimento, "Inline DGA Detection with Deep Networks", 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 683-692, 2017.

[21] H. Mac, D. Tran, V. Tong, L.G. Nguyen, and H.A. Tran, "DGA Botnet Detection Using Supervised Learning Methods", Proceedings of the Eighth International Symposium on Information and Communication Technology, 211-218, 2017.

[22] S.R.C. Liew, and N.F. Law, "BEAM – an Algorithm for Detecting Phishing Link", Proceedings of 2022 APSIPA Annual Summit and Conference, 598-604, 2022.

[23] S.R.C. Liew and N.F. Law, "Use of Subword Tokenization for Domain Generation Algorithm Classification", Cybersecurity (accepted), 2023, doi: 10.1186/s42400-023-00183-8.

[24] D. Tran, H. Mac, V. Tong, H.A. Tran, and L.G. Nguyen, "A LSTM based Framework for Handling Multiclass Imbalance in DGA Botnet Detection", Neurocomputing, 275: 2401, 2018.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection

Haowei Cheng, Candy Olivia Mawalim, Kai Li, Lijun Wang, and Masashi Unoki
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: {haowei.cheng, candylim, kai_li, lijun.wang, unoki}@jaist.ac.jp

***Abstract*—Deep-fake speech detection aims to develop effective techniques for identifying fake speech generated using advanced deep-learning methods. It can reduce the negative impact of malicious production or dissemination of fake speech in real-life scenarios. Although humans can relatively easy to distinguish between genuine and fake speech due to human auditory mechanisms, it is difficult for machines to distinguish them correctly. One major reason for this challenge is that machines struggle to effectively separate speech content from human vocal system information. Common features used in speech processing face difficulties in handling this issue, hindering the neural network from learning the discriminative differences between genuine and fake speech. To address this issue, we investigated spectro-temporal modulation representations in genuine and fake speech, which simulate the human auditory perception process. Next, the spectro-temporal modulation was fit to a light convolutional neural network bidirectional long short-term memory for classification. We conducted experiments on the benchmark datasets of the Automatic Speaker Verification and Spoofing Countermeasures Challenge 2019 (ASVspoof2019) and the Audio Deep synthesis Detection Challenge 2023 (ADD2023), achieving an equal-error rate of 8.33% and 42.10%, respectively. The results showed that spectro-temporal modulation representations could distinguish the genuine and deep-fake speech and have adequate performance in both datasets.**

I. Introduction

The development of deep learning technology is rapidly advancing. It offers numerous conveniences to our daily lives, such as the production of audiobooks[1], the creation of intelligent speech robots [2], and even aiding those who have lost their voices due to throat disease or other medical conditions. Speech is considered the most essential and natural way for humans to convey both linguistic and non-linguistic information. However, the emergence of malicious fake speech generated by deep learning poses significant threats to societal stability and individual property security. This form of fake speech is commonly known as "deep-fake speech". It deceives both human listeners and automatic speaker verification systems. Furthermore, it carries the risk of spreading false and harmful information, including distorting politicians' statements [3]. Therefore, it is crucial to develop a reliable method for effectively detecting deep-fake speech.

Several challenges have been organized to advance the field of deep-fake speech detection. One of the most renowned global challenges is the automatic speaker verification and spoofing countermeasures challenge (ASVspoof) [4]. Its primary objective is to promote the development of robust countermeasures against such spoofing attempts. Another notable recent challenge is the audio deep synthesis detection (ADD) challenge [5]. It specifically focuses on detecting deep-fake audio in realistic scenarios. This challenge aims to address the unique challenges posed by deep-fake situations encountered in real-life settings.

Although many challenges and methods are proposed in deep-fake speech detection tasks, it is difficult for machines to accurately distinguish them. The main reason for this difficulty lies in the inherent difference between genuine speech and fake speech. Genuine speech not only contains speech information but also reflects human vocal system activity, including characteristics like glottal vibration. On the other hand, fake speech generated by machines lacks these human-like characteristics. This disparity makes it challenging for machines to accurately distinguish between the two types of speech. Common features struggle to effectively capture the unique patterns associated with genuine speech information and human vocal system activity, resulting in insufficient discriminative information for neural network training. Therefore, achieving effective detection of deep-fake speech can be accomplished by successfully separating these components.

To address this issue, we were inspired by the human auditory mechanism. The human auditory cortex possesses dynamic and adaptive properties that enable us to effectively differentiate between speech from humans and machines [6]. Building upon this understanding, a study [7] has revealed that neurons in the auditory cortex can decompose spectrograms into spectro-temporal modulation (STM) content. This finding has led to the development of the STM, which is a multi-scale representation for speech analysis and has been shown to explain various psychoacoustic phenomena [8]. Another study [9] proposed an STM-based method for audio classification inspired by human auditory mechanisms. By utilizing an auditory model to capture relevant features, these methods have demonstrated their effectiveness in classification. Therefore, it is reasonable to consider that STM representation has the potential to discriminate deep-fake speech and enhance detection accuracy.

This paper focuses on exploring the cues and effectiveness of the STM representation which is based on the human auditory mechanisms for detecting deep-fake speech. We conducted an investigation into the role of various feature expressions and

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

implemented STM representations. To enhance the ability of neural networks to capture feature information within STM representations. We combined STM with a light convolutional neural network and bidirectional long short-term memory (LCNN-BiLSTM), which is a model widely used in classification tasks. We performed experiments on two datasets: the Automatic Speaker Verification and Spoofing Countermeasures Challenge 2019 (ASVspoof2019) and the Audio Deep synthesis Detection Challenge 2023 (ADD2023). The results clearly demonstrated that the STM representations were highly effective in distinguishing between genuine and fake speech, achieving adequate performance on both datasets.

II. Related work

A. ASVspoof challenge

To promote research and development in combating voice spoofing attacks, the ASVspoof challenge was first held in 2015 and has since become a significant platform in the field. The competition provides datasets containing various types of voice spoofing attacks, such as speech synthesis, playback, and voice conversion. The objective of the ASVspoof challenge is to evaluate and improve the robustness of ASV systems, enabling them to reliably distinguish between genuine and fake speech. Through the competition, researchers collaborate and propose innovative methods to address the challenges posed by voice spoofing attacks.

B. ADD challenge

However, there has been a recognition that many real-life scenarios have not been adequately covered in ASVspoof challenge. To address this gap, the Audio Deep synthesis Detection challenge (ADD) was motivated. The ADD challenge encompasses three tracks: low-quality fake audio detection, partially fake audio detection, and the fake audio game. These tracks aim to provide a more comprehensive evaluation of ASV systems' ability to detect various types of voice spoofing attacks, including both low-quality and partially manipulated audio samples. The inclusion of these tracks in the ADD challenge aims to advance research in the field and encourage the creation of more robust and efficient solutions to counter voice spoofing attacks.

C. Common methods and limitations

Many common methods have been proposed for detecting deep-fake speech, typically involving front-end feature extraction and back-end classification [10–16]. Baseline models in the two challenges applied linear frequency cepstral coefficients (LFCC) as feature and Gaussian mixture models (GMM) for classification [17]. Meanwhile, many features were explored for deep-fake speech detection, including Mel-frequency cepstral coefficients (MFCC) [18, 19], constant-Q cepstral coefficients (CQCC) [20, 21], Gammatone cepstral coefficients (GTCC) [22], etc. For classification, a variety of models are commonly employed, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), etc.

Despite many features are employed in deep-fake speech detection tasks, it is difficult for machines to accurately distinguish them. On the other hand, humans can distinguish genuine and fake speech through our sense of hearing. Taking inspiration from the human auditory mechanism, our approach involves exploring feature representations that not only capture the speech content but also the subtle cues associated with human vocal system activity.

III. Spectro-Temporal modulation representation

A. Definition

Temporal modulation refers to the changes in modulations over time in the spectrogram, while spectral modulation represents variations along the frequency axis. The concept of STM combines both temporal and spectral modulations simultaneously. In the field of auditory psychophysics and neuroscience, the auditory model is divided into two essential stages. One involves transforming the acoustic signal into an internal neural representation called an auditory spectrogram. Another analyzes this spectrogram to estimate the spectral and temporal modulation content using specialized filters that respond to specific modulations[23, 24]. This stage aims to separate different cues and characteristics associated with distinct auditory percepts [25–27]. It can be compared to the adaptive and masking properties of the human auditory system, where important information can still be perceived even in a noisy environment. Therefore, incorporating STM analysis provides a more comprehensive understanding of human perception. By examining STM, we can potentially uncover meaningful characteristics in the speech signal that can aid in the detection of deep-fake speech.

To obtain the STM representation from the original signal, a series of steps is followed. Initially, the input signal undergoes decomposition into frequency components using filterbanks. This process separates the speech signal into different frequency bands. Subsequently, squaring and low-pass filtering are used for computing power envelope from the output of the filterbank. In the final step, a two-dimensional spectro-temporal analysis is conducted on this power envelope to derive the STM spectrogram. The STM spectrogram thus provides a representation of the dynamic variations present in the speech signal across different spectral and temporal scales.

B. Investigation the role of feature expressions

In order to investigate the role of feature expressions, three filterbanks were employed to implement the STM independently. The Mel filterbank (Mel FB) and Gammatone filterbank (ERB FB) are both widely utilized filterbanks in the realm of speech signal processing [28]. Constant bandwidth filterbank (CBW FB) maintains a constant bandwidth across the frequency range.

1) Mel filterbank: The Mel FB is based on the Mel scale, which is derived from psychoacoustic experiments and provides a nonlinear mapping of frequency. Triangular-shaped filters are employed by the Mel FB, with each filter centered at a specific Mel frequency and possessing a bandwidth determined

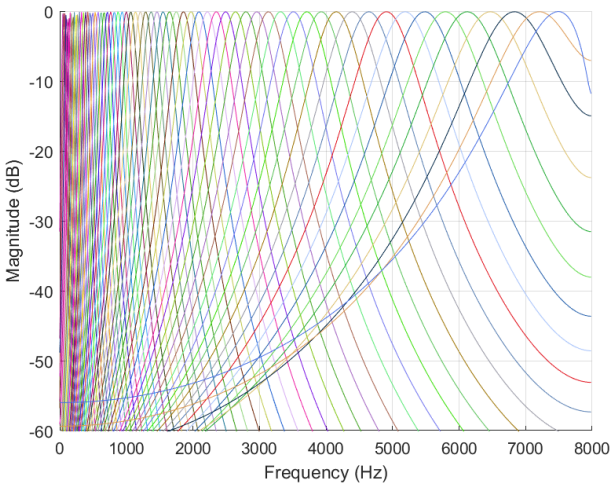


Fig. 1: Frequency response of ERB FB

by the adjacent filters. This triangular shape enables the Mel FB to approximate the perceptual frequency resolution of the human auditory system [29]. The formula to convert a linear frequency (f) to the Mel scale (m) is as follows:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{1}$$

2) Gammatone filterbank: The ERB FB is designed to accurately model the response characteristics of the cochlea in the human auditory system [30, 31]. It utilizes filters based on the Gammatone function, derived from a combination of complex exponential functions and low-pass filters. These filters effectively capture the shape of cochlear filters and the frequency selectivity of the auditory system. As a result, the filterbank enhances the representation of low-frequency components with narrow bandwidths and reduces the presence of high-frequency components with wider bandwidths, as shown in Figure 1. The integration of the ERB scale further enhances the accuracy by approximating the frequency resolution of the human auditory system. This integration allows the ERB FB to better capture the spectral characteristics of auditory signals and align with human auditory perception [32].

In the ERB FB, the center frequencies are based on the specified upper and lower frequency limits and the number of channels. These center frequencies are proportional to the corresponding bandwidths of the filters [33]. The output obtained from the ERB FB is as follows:

$$g_k(t) = \mathrm{A}t^{(n-1)} \exp(-2\pi b_f \mathrm{ERB}(f_k)t) \cos(2\pi f_k t), \tag{2}$$

where $At^{(n-1)} \exp(-2\pi b_f \mathrm{ERB}(f_k t))$ is the amplitude term represented by the Gamma distribution, A, n, and b_f are the amplitude, filter order, and bandwidth of the filter respectively. We apply the fourth order Gammatone. The formula to convert a linear frequency (f) to the ERB scale is as follows:

$$\mathrm{ERB} = 24.7(4.37 f_k + 1), \tag{3}$$

where f_k is the k-th center frequency (in kHz) of filterbank.

3) Constant bandwidth filterbank: Unlike Mel FB and ERB FB, which have variable bandwidths. CBW FB has a fixed bandwidth for each filter, regardless of their center frequency. It is constructed by employing a consistent bandwidth parameter, whereby the center frequencies and channel count are computed using the provided lower and upper frequency limits.

To analyze the differences between genuine and fake speech signals as illustrated in spectrograms. Figure 2 provides visualizations of the spectrogram representations.

C. Procedure of STM analysis

In the front-end input, the speech signal $s(t)$ is first filtered by a bank of filters. The output of the k-th channel is given by

$$y_k(t) = g_k(t) * \mathrm{s}(t), \tag{4}$$

where * represents the convolution, $g_k(t)$ is the impulse response of the k-th channel of filterbank. The power envelope is extracted by the Hilbert transform and squared. LPF represents a low-pass filter at cut-off frequency in 64Hz.

$$e_k^2(t) = \mathrm{LPF}\left[|\mathrm{y}_k(t) + j\mathrm{Hilbert}(y_k(t))|^2\right], \tag{5}$$

Finally, STM representation can be obtained by applying a two-dimensional Fourier transform to the squared envelope $e_k^2(t)$, as shown in Eq. (6). It is important to note that the result of the two-dimensional Fourier transform is typically a matrix comprising complex numbers, where each element consists of both real and imaginary parts. To obtain the STM representation utilized in this study, the absolute value of the result is taken. The STM representations of genuine and fake speech signals are shown in Figure 3.

$$\mathrm{STM} = \mathrm{2DFFT}(\log e_k^2(t)). \tag{6}$$

where 2DFFT represents a two-dimensional fast Fourier transform.

IV. Proposed Method

In this section, we present a deep-fake speech detection model that utilizes an LCNN and BiLSTM, as shown in Figure 4. LCNN is a convolutional neural network variant that is purposefully developed to strike a balance between computational complexity and performance. The advantage of LCNN is the implementation of a max feature map activation strategy, which involves using a max-out activation function. This characteristic enables faster training and inference times while minimizing the impact on overall performance. In order to extract useful information from STM, deep learning models such as BiLSTM have been widely used. It can effectively model the temporal dependencies in the STM, which are critical for deep-fake speech detection tasks. Specifically, BiLSTM has a "memory" mechanism that allows it to keep track of past information and use it to inform the current prediction. During the task using BiLSTM, the speech feature sequences are individually fed into the hidden layers of both the forward LSTM (LSTM_F) and the backward LSTM (LSTM_B). This process generates two feature vectors that encapsulate the forward and

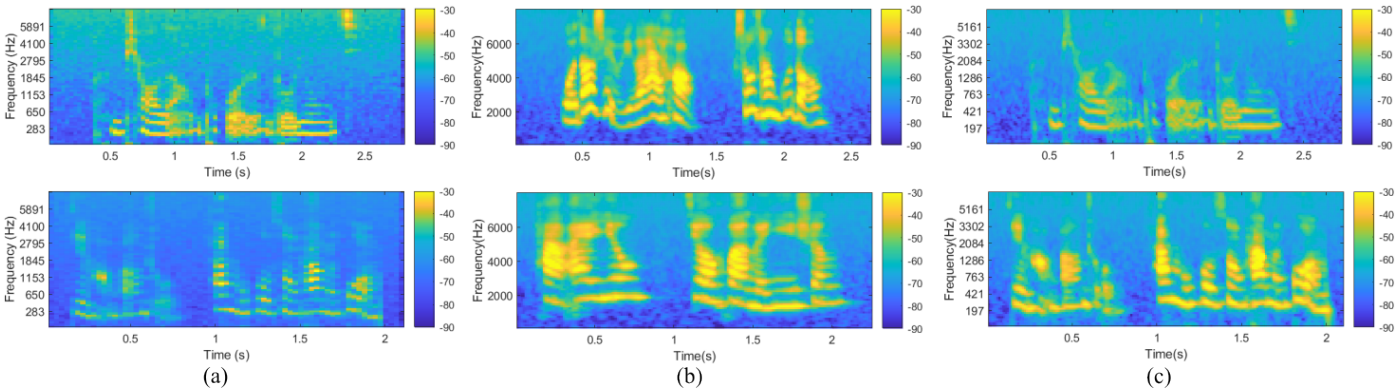


Fig. 2: Spectrograms of genuine (above) and fake (below) speech signals: (a) Mel FB, (b) CBW FB and (c) ERB FB.

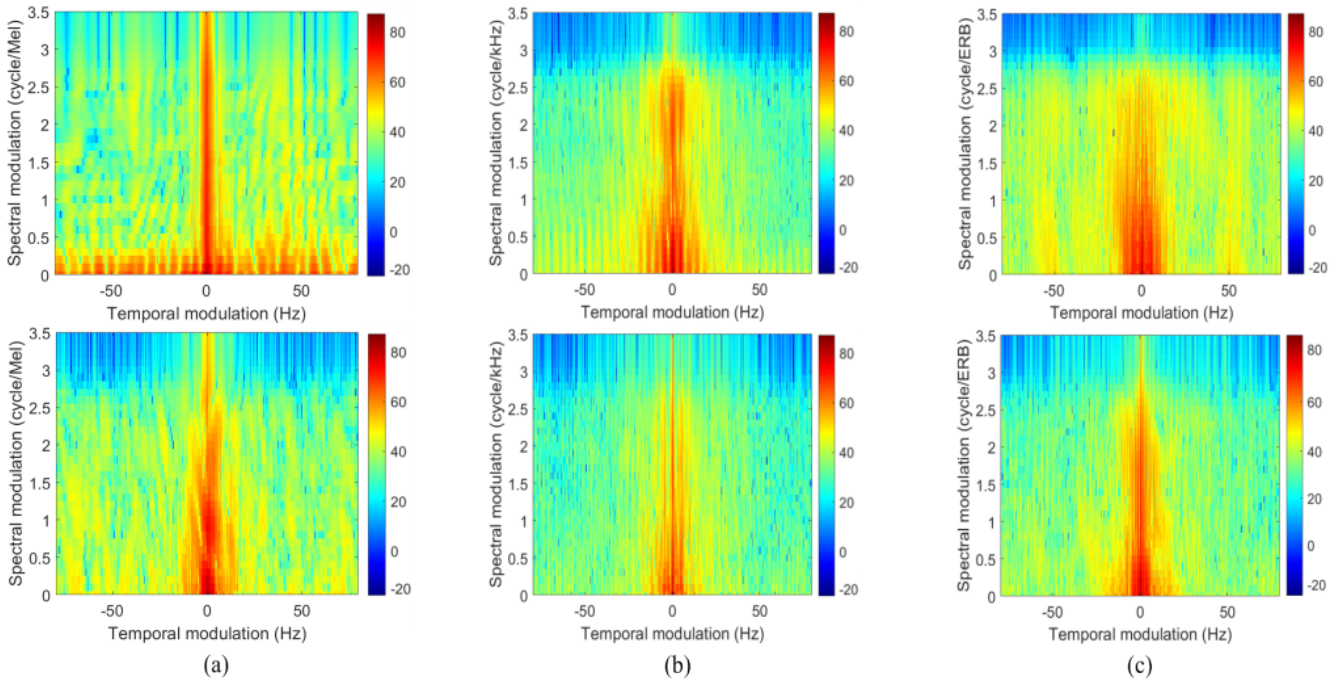


Fig. 3: STMs of genuine (above) and fake (below) speech signals: (a) Mel FB, (b) CBW FB, and (c) ERB FB.

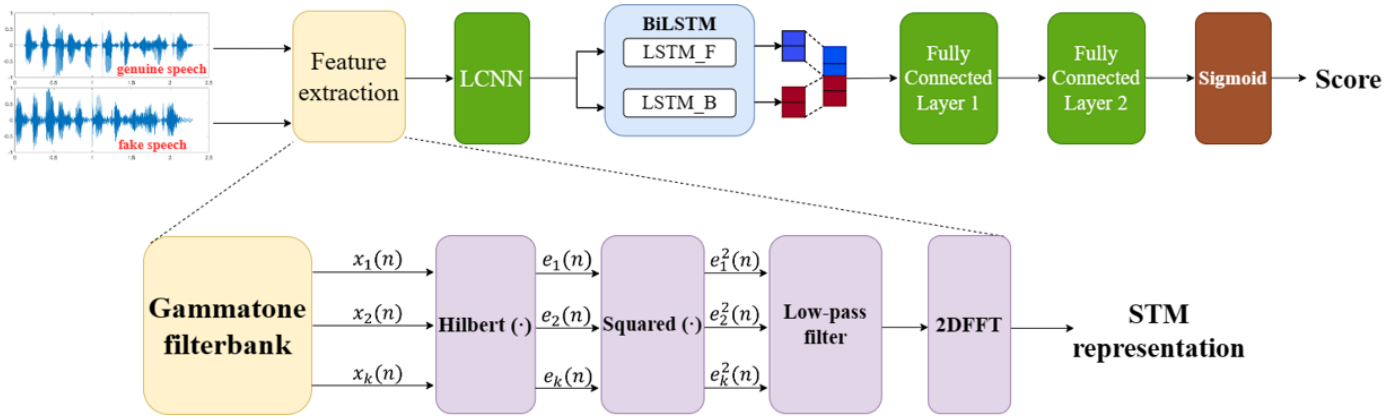


Fig. 4: Block diagram of the proposed method

TABLE I: Statistic for datasets of the ADD2023. (Durations with three values denoted with minimum/average/maximum)

Dataset	Number of utterances			Duration (sec)
	Genuine	Fake	Total	
Training	3,012	24,072	27,084	0.86/3.15/60.01
Development	2,307	26,017	28,324	0.86/3.16/60.01
Evaluation	-	-	111,977	0.35/5.51/217.49

TABLE II: statistic for datasets of the ASVspoof2019. (Durations with three values denoted with minimum/average/maximum)

Dataset	Number of utterances			Duration (sec)
	Genuine	Fake	Total	
Training	2,580	24,072	26,625	0.65/3.42/13.19
Development	2,548	22,296	24,844	0.69/3.49/16.51
Evaluation	7,355	63,882	71,237	0.47/3.14/16.55

backward information of the speech. Subsequently, the output vectors from these two layers are combined, forming a merged vector that is passed through two fully connected layers. Finally, the classification is performed by applying a sigmoid activation function to compute the score. This is particularly useful for distinguishing between genuine and fake speech, as speech signals often contain long-term dependencies. The dimensions of the BiLSTM layers are set to match the output dimensions of the LCNN. To optimize the model parameters, a binary cross entropy (BCE) objective function is utilized. The BCE objective function is defined as follows:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (7)$$

where N represents the total number of samples in the dataset, y_i and $\hat{y}_i$ denote the ground truth of the i-th training sample and its corresponding output probability from the model.

V. EXPERIMENT

A. Database and Metrics

In this study, two datasets were employed. The Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof2019) pioneered the comprehensive treatment of all major attack types, including text-to-speech, voice conversion, and replay spoofing attacks, effectively covering real-world voice spoofing scenarios [34]. Another dataset used is the Audio Deep synthesis Detection challenge (ADD2023) [35].

The ADD2023 dataset (as shown in Table I) consists of Mandarin speeches with neutral emotions. The training and development sets have high signal-to-noise ratios (SNR), while the evaluation set has low SNR with various real-world background noises. The evaluation set lacks publicly accessible labels provided by the organizers, therefore the final scores are required to be submitted to the ADD2023 challenge's website CODALAB for online evaluation.

The ASVspoof2019 dataset is divided into three subsets: the training set, development set, and evaluation set, as presented in Table II.

Notably, the evaluation dataset in ASVspoof2019 includes provided labels, facilitating local evaluation without the need to submit results for external assessment. The performance of the proposed method was evaluated using the equal error rate (EER). By comparing the results obtained from the two datasets, the objective was to prove the generalization ability and reliability of the proposed methods. This was achieved by utilizing a diverse and challenging collection of samples provided by the two datasets.

B. Experiment setup

The Mel FB, CBW FB, and ERB FB were implemented with consistent parameters. The frequency range was set from 50 Hz to 8000 Hz, utilizing 64 channels. This frequency range was chosen due to the typical perception of speech signals by the human auditory system. The lower limit of 50 Hz captures the fundamental frequency component of speech, while the upper limit of 8000 Hz includes high-frequency resonances and harmonics.

Setting the channel number to 64 aims to enhance spectral information and improve sound resolution. With 64 channels, a finer frequency division is achieved, enabling precise capture of speech characteristics across various frequency ranges. The increased channel count provides more frequency detail, leading to a more accurate representation of spectral features and capturing a wider range of speech features. However, it's important to consider the computational and memory costs associated with higher channel numbers, as they can impact real-time performance and computational efficiency. Thus, a careful balance was struck by selecting 64 channels. To accommodate the high-resolution STM representations, the TM domain underwent resampling at a rate of 1000 Hz, resulting in an STM representation size of [64, 1000]. Then the LCNN-BiLSTM model is trained using the labels, the batch size of all data is 64, and the epoch number is 30. An Adam optimizer with a learning rate of 0.0001 was used. Validation was performed using the development dataset, and the model achieving the lowest EER score was considered the best-performing one.

C. Comparison with different classic features

In addition to our proposed method, we conducted a comparative analysis by re-implementing three well-known features: MFCC, LFCC, and GTCC. This allowed us to assess the performance of our approach against these established methods [18–21].

Figure 5 illustrates the calculation diagram for the classic features. In the feature extraction stage, the input signal undergoes initial pre-processing, including windowing with a window length of 25 ms and a step length of 10 ms. The window type used is Hamming. Subsequently, a fast Fourier

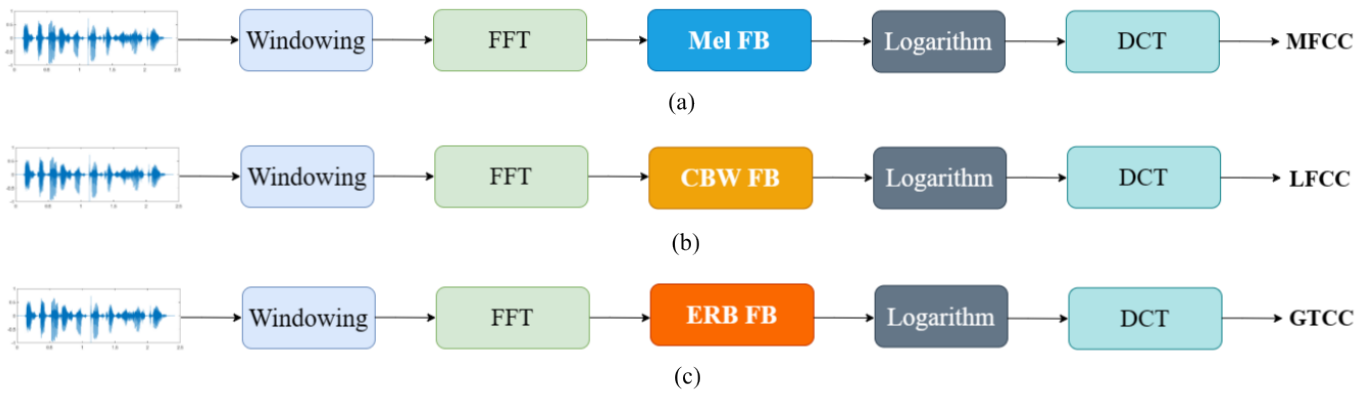


Fig. 5: Flow process diagram of classic features: (a) MFCC, (b) LFCC, and (c) GTCC.

transform (FFT) is applied to the windowed signal with 512 points. This yields the Mel spectrum, Linear spectrum, and Gammatone spectrum after passing through the respective Mel FB, CBW FB, and ERB FB. Finally, the resulting spectrum is subjected to logarithmic transformation and discrete Cosine transform (DCT) to obtain the MFCC, LFCC, and GTCC. The back-end classifier used here is LCNN-BiLSTM, which shares the same architecture as the STM experiment.

VI. Results and Discussion

To identify effective features for deep-fake speech detection, we analyze the STM representation among genuine and fake speech. Subsequently, we applied STM and common features to LCNN-BiLSTM model and performed comparative experiments to evaluate the performance of the proposed method.

We conducted experiments using the ASVspoof2019 dataset (Table III) and the ADD2023 dataset (Table IV). The baseline model in ASVspoof2019 utilized LFCC and GMM, achieving an EER of 18.89%. In comparison, our STM-based approach with the ERB FB achieved an EER of 8.33%, representing a significant improvement of 10.56%. Notably, the STM (ERB FB) outperformed both STM (Mel FB) and STM (CBW FB) in terms of performance. These results emphasize the effectiveness of integrating STM and the advantages of using the ERB FB for deep-fake speech detection. The improved performance can be attributed to the STM's ability to capture fine-grained temporal and spectral details, facilitating more precise discrimination between genuine and fake speech samples.

In the evaluation of the ADD2023 dataset, we compared the performance of our proposed method with the baseline published by the organizer, which utilized LFCC-LCNN and achieved an EER of 70.37%. In our experiments, we first applied Mel FB, CBW FB, and ERB FB as input features to the classifier. The results showed EERs of 77.61%, 83.37%, and 73.34%, respectively. Then we conducted further comparison experiments and found that the classic features outperformed the filterbanks. Meanwhile, the STM representations performed better than the classic features. Specifically, STM based on ERB FB achieved an EER of 42.10%, outperforming not only the baseline but also common features like MFCC (53.36%) and GTCC (63.69%). These results indicate that STMs provide critical information to detect genuine and fake speech.

While our method has demonstrated successful deep-fake speech detection and achieved superior results, it is important to discuss its underlying principles and address the remaining issues. The STM can be considered as a cepstrum, which can capture the subtle cues of human vocal system activity information and speech information in a two-dimensional representation. In the case of genuine speech, the vocal system activity information tends to concentrate near the origin of the STM representation, while the speech information spreads around it. On the other hand, the STM representation of fake speech lacks this characteristic pattern observed in genuine speech. Natural speech produced by humans exhibits a more regular pattern in the STM representation due to the commonalities in the human vocal system. However, machine-generated speech lacks this consistent pattern across individuals, resulting in a less regular waveform in speech signals. Leveraging this feature, we can effectively distinguish between fake and genuine speech.

Therefore, we investigated the role of different feature expressions and implemented STMs. Then combined STM representations with an LCNN-BiLSTM model, and experiments on the datasets of the ASVspoof2019 and the ADD2023. The results showed that STM can effectively distinguish genuine and fake speech with good performance both in two datasets.

VII. Conclusion

This study showed the effectiveness of utilizing STM based on the ERB FB, inspired by the human auditory mechanism for deep-fake speech detection. First, we investigated the role of different feature expressions and implemented STM representations. We then developed a classifier by combining LCNN with BiLSTM and conducted experiments. Additionally, we compared the performance of three classical features with the STM method. The proposed method was evaluated

TABLE III: Comparative results using the ASVspoof2019 dataset

Methods	Equal Error Rate (%)	
	Development set	Evaluation set
STM (Mel FB)	0.04	9.79
STM (CBW FB)	0.09	13.46
STM (ERB FB)	**0.02**	**8.33**

TABLE IV: Comparative results using the ADD2023 dataset

Method	Equal Error Rate (%)	
	Development set	Evaluation set
Mel FB	0.26	77.61
CBW FB	0.31	83.37
ERB FB	0.23	73.34
MFCC	0.14	53.36
LFCC	0.19	66.52
GTCC	0.21	63.69
STM (Mel FB)	0.14	47.65
STM (CBW FB)	0.26	55.55
STM (ERB FB)	**0.09**	**42.10**

on the ASVspoof2019 and ADD2023 datasets. Remarkably, our method exhibited significant results for ASVspoof2019 and ADD2023, achieving performance of 8.33% and 42.10% respectively. These results demonstrate the proposed method could effectively detect deep-fake speech. While the focus of this study was on evaluating the performance against baseline models, conducting comparisons with other existing deep-fake speech detection systems could provide a more comprehensive assessment of the proposed method's effectiveness. Future work will involve further investigation of the specific physics-based acoustic features that can be accurately captured and represented by the STM representation. Moreover, it is essential to expand the evaluation to encompass other state-of-the-art methods and to assess the system's performance across a diverse range of datasets.

ACKNOWLEDGEMENT

This work was supported by the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233) and Grant-in-Aid for Transformative Research Areas (A) (23H04344).

REFERENCES

[1] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: An overview," in *Proc. Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, 2021, pp. 557–566.

[2] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, 2010, pp. 537–541.

[3] R. Naika, "An overview of automatic speaker verification system," in *Proc. Intelligent Computing and Information and Communication, ICICC 2017*, pp. 603–610.

[4] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pp. 9216–9220.

[6] B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Adaptation of the human auditory cortex to changing background noise," *Nature communications*, vol. 10, no. 1, p. 2509, 2019.

[7] S. Shamma, "Encoding sound timbre in the auditory system," *IETE Journal of research*, vol. 49, no. 2-3, pp. 145–156, 2003.

[8] R. P. Carlyon and S. Shamma, "An account of monaural phase sensitivity," *Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 333–348, 2003.

[9] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, vol. 1, pp. I–601.

[10] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting deepfake voice using explainable deep learning techniques," *Applied Sciences*, vol. 12, no. 8, p. 3926, 2022.

[11] P. A. Ziabary and H. Veisi, "A countermeasure based on cqt spectrogram for deepfake speech detection," in *Proc. International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2021, pp. 1–5.

[12] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pp. 9241–9245.

[13] B. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[14] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, pp. 6359–6363.

[15] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[16] Z. Jiang, H. Zhu, L. Peng, W. Ding, and Y. Ren, "Self-supervised spoofing audio detection scheme." in *Proc. Interspeech*, 2020, pp. 4223–4227.

[17] Y. Ma, Z. Ren, and S. Xu, "Rw-resnet: A novel speech anti-spoofing model using raw waveform," *arXiv preprint arXiv:2108.05684*, 2021.

[18] R. K. Bhukya and A. Raj, "Automatic speaker verification spoof detection and countermeasures using gaussian mixture model," in *Proc. IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2022, pp. 1–6.

[19] R. Li, M. Zhao, Z. Li, L. Li, and Q. Hong, "Anti-spoofing speaker verification system with multi-feature integration and multi-task learning." in *Proc. Interspeech*, 2019, pp. 1048–1052.

[20] M. Sahidullah, H. Delgado, M. Todisco, A. Nautsch, X. Wang, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, "Introduction to voice presentation attack detection and recent advances," *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pp. 339–385, 2023.

[21] S. Ravindran and K. Geetha, "An overview of spoof detection in asv systems," *ECS Transactions*, vol. 107, no. 1, p. 1963, 2022.

[22] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, "Voice spoofing detector: A unified anti-spoofing framework," *Expert Systems with Applications*, vol. 198, p. 116770, 2022.

[23] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE transactions on speech and audio processing*, vol. 3, no. 5, pp. 382–395, 1995.

[24] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[25] F. Samson, L. Mottron, B. Jemel, P. Belin, and V. Ciocca, "Can spectro-temporal complexity explain the autistic pattern of performance on auditory tasks?" *Journal of autism and developmental disorders*, vol. 36, pp. 65–76, 2006.

[26] S. M. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature neuroscience*, vol. 8, no. 10, pp. 1371–1379, 2005.

[27] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[28] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016.

[29] R. S. Chavan and G. S. Sable, "An overview of speech recognition using hmm," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 6, pp. 233–238, 2013.

[30] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Replay attack detection in automatic speaker verification using gammatone cepstral coefficients and resnet-based model," *Journal of Signal Processing*, vol. 26, no. 6, pp. 171–175, 2022.

[31] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Proc. IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[32] T. Irino and M. Unoki, "A time-varying, analysis/synthesis auditory filterbank using the gammachirp," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1998)*, vol. 6, pp. 3653–3656.

[33] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[34] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[35] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Flexible Evidence Model to Reduce Uncertainty Mismatch Between Speech Enhancement and ASR Based on Encoder-Decoder Architecture

Ryu Takeda*, Yui Sudo† and Kazunori Komatani*
* SANKEN, Osaka University, Japan
E-mail: {rtakeda, komatani}@sanken.osaka-u.ac.jp Tel: +81-6-6879-8416
† Honda Research Institute Japan, Co., Ltd., Japan
E-mail: yui.sudo@jp.honda-ri.com

***Abstract*—Missing-data automatic speech recognition (ASR) can use the uncertainty of speech enhancement (SE) results to improve ASR performance without joint-training. Such uncertainty is represented by a probabilistic evidence model, whose design affects the performance. Previous evidence models have utilized a specific distribution regardless of the ASR model, and do not consider their robustness, i.e., the ability to recognize both clean and noisy speech (robust ASR). The SE uncertainty depends on the degree of speech cleanness, which causes an uncertainty mismatch between SE and a robust ASR model, resulting in recognition failures. We thus propose a flexible evidence model to reduce the mismatch. Our model features several parameters to control its distribution shape and can reweight the SE uncertainty via an encoder-decoder architecture by automatically selecting parameters in accordance with the ASR scores. Experiments showed that the character error rate improved by more than 3 points over the best baseline even with a robust ASR model under a -5-dB SNR condition.**

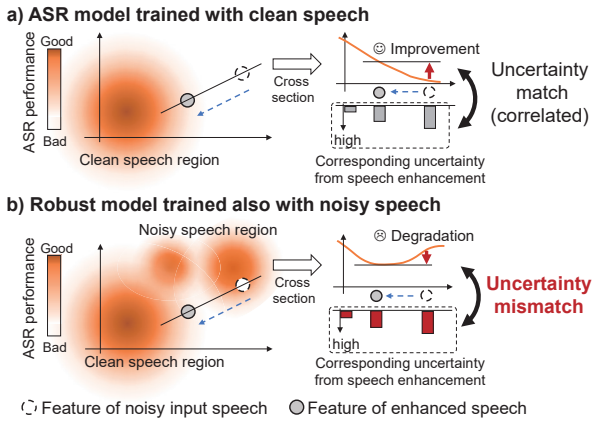


Fig. 1. Problem: uncertainty mismatch due to ASR models. Contour maps of ASR model space show ASR performance vs. feature space.

I. Introduction

Technologies for the integration of automatic speech recognition (ASR) and frontend signal processing (such as speech enhancement (SE)) have attracted more interest as their performance improves. Developers can easily apply state-of-the-art ASR or SE based on deep neural networks (DNNs) with various programming libraries [1], frameworks, and pretrained models [2], [3]. On the other hand, the training cost, including that required for the data-set and computer resources, increases as each DNN model becomes larger. This has made the joint-training or fine-tuning of the models for different applications and acoustic environments increasingly difficult for developers. For example, when we develop a robust ASR under noisy environments (low SNR conditions), ASR and SE are usually combined to reduce the influence of noise signals in advance. However, the training conditions of these two large models are usually different, which forces us to joint-train them in order to improve the performance. We believe that a *joint-training-free* interface between ASR and SE would make each model more flexible and portable for real applications.

Missing-data ASR (MD-ASR) is a promising approach that does not necessarily require *re-training* of models. This approach utilizes the *uncertainty* of speech features in the decoding process [4]–[6], which can propagate statistical information from SE to MD-ASR [7]–[10]. A probabilistic model of such uncertainty is called an *evidence* model [6], [11], [12] and is represented as a probability density function (pdf), which reduces misclassification risks by evaluating the *expectation* of target values such as the classification score or hidden vectors used in ASR. The formulation of the evidence model and MD-ASR usually depends on the base ASR structure, such as a hidden Markov model (HMM) with a Gaussian mixture model (GMM) [6], [11], DNN-HMM [13]–[16], or end-to-end ASR [17] based on encoder-decoder architectures [18]–[21]. Note that recent research [17] has extended the conventional frame-wise evidence models (e.g., [11]) to utterance-wise ones, which enables us to apply missing data techniques to the end-to-end ASR based on encoder-decoder model. The evidence model also utilizes the uncertainty pdf of enhanced speech to reduce the influence of noise or distortions, which also improves the ASR performance. Here, the uncertainty of enhanced speech usually represents how far its characteristics are from those of clean speech.

The problem with previous MD approaches and evidence models is that they ignore the influence of the ASR model, which sometimes causes an uncertainty mismatch between the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

enhanced speech and the ASR model. Figure 1 illustrates this problem. The uncertainty of enhanced speech by SE is usually lower than that of an observed noisy speech signal because SE tries to recover a clean speech signal. An ASR model trained on clean speech can recognize such enhanced speech accurately because its features are close to those of clean speech, which means that the uncertainties of SE and the ASR model match (Fig. 1 a)). However, such a relationship does not always hold, and the uncertainties of the two speech signals are reversed for robust ASR models (Fig. 1 b)). ASR models trained on clean and noisy speech can often recognize observed noisy speech better than enhanced speech, because speech data enhanced in a method-specific manner are not usually included in the ASR training set. *As the actual behavior of ASR models (e.g., pretrained models) is usually unknown, this mismatch should be solved flexibly in accordance with the given ASR model*.

Hence, we propose a flexible evidence model to reduce the uncertainty mismatch between SE and the ASR model based on encoder-decoder architechture. Our evidence model can control the important region of the ASR model's feature space with several parameters by referring to the features of enhanced speech and noisy speech. Effective parameters for a certain ASR model are selected automatically in accordance with the ASR recognition scores in an *offline* unsupervised manner. A beta distribution is used to represent the evidence model, which is a natural extension of the uniform model proposed in [17]. The posterior pdf from a neural SE based on a variational auto encoder (VAE) is also modified by this evidence model to further improve the ASR performance. Additionally, we eliminate the stochastic sampling procedure of Monte-Carlo implementation by applying the unscented transform and resampling technique used in filtering methods [22], [23]. Experiments were conducted with three types of ASR model trained on (1) clean speech, (2) clean, reverberant speech, and (3) clean, reverberant, noisy speech.

Our key contributions are as follows:

- We propose a flexible evidence model based on beta distribution and unsupervised model selection to reduce the model mismatch between SE and a given end-to-end ASR model, especially under low SNR conditions.
- We propose a Monte Carlo-based integration of SE and ASR featuring on encoder-decoder architecture to draw utterance-wise samples.
- We present a comprehensive investigation of the ASR performance in terms of SNRs, training data for ASR model, evidence models, integration frameworks with SE, and Whisper ASR models without fine-tuning.

II. Preliminary

We first present an overview of the existing end-to-end ASR, missing-data framework and evidence models. Note that conventional missing-data ASRs (e.g., GMM-HMM or DNN-HMM) cannot be applied to recent end-to-end ASRs because the assumption of independence among frames does not match these newer architectures. In this paper, we basically use almost the same notations as those in [11], [17].

A. Encoder-Decoder-Based Neural ASR

End-to-end ASR directly estimates a character sequence $c_{1:L} = [c_1, ..., c_L]$ with length L from a sequence of speech feature vectors $\mathbf{s}_{1:T} = [\mathbf{s}_1, ..., \mathbf{s}_T]$ with length T. The feature vector at frame t, $\mathbf{s}_t \in \mathbb{R}^D$, is a D-dimensional vector. The mapping from $\mathbf{s}_{1:T}$ to $c_{1:L}$, i.e., the posterior probability $p(c_{1:L}|\mathbf{s}_{1:T})$, is modeled by DNNs and obtained by supervised training.

The recent CTC/attention architecture [21] for end-to-end ASR comprises a shared encoder, CTC, and attention decoder networks. The shared encoder network converts the speech features $\mathbf{s}_{1:T}$ into a T'-length sequence of encoded hidden vectors $\mathbf{h}_{1:T'} = [\mathbf{h}_1, ..., \mathbf{h}_{T'}], \mathbf{h}_t \in \mathbb{R}^{D'} (t = 1, .., T')$. The CTC and attention decoder networks represent the posterior probabilities $p_{ctc}(c_{1:L}|\mathbf{h}_{1:T'})$ and $p_{att}(c_{1:L}|\mathbf{h}_{1:T'})$ in different schemes.

The score function J_{asr} in the decoding process is evaluated by a weighted sum with weights $w_i \in [0, 1], (i = 1, 2, 3)$:

$$\mathbf{h}_{1:T'} = \text{Encoder}(\mathbf{s}_{1:T}) := \mathbf{h}_{1:T'}(\mathbf{s}_{1:T}), \tag{1}$$

$$J_{asr}(c_{1:L}|\mathbf{h}_{1:T'}) = w_1 \log p_{ctc}(c_{1:L}|\mathbf{h}_{1:T'}) \tag{2}$$

$$+w_2 \log p_{att}(c_{1:L}|\mathbf{h}_{1:T'}) + w_3 \log p_{lm}(c_{1:L}), \tag{3}$$

where $p_{lm}(c_{1:L})$ denotes a language model. A character sequences that maximize this score function is searched for. Each of the functions ($\mathbf{h}_{1:T'}(\cdot)$, p_{ctc}, p_{att}, and p_{lm}) can be implemented with a recent NN model such as a Transformer.

B. Missing-Data Framework for End-to-End ASR

The *missing data/feature* in ASR means corrupted or distorted speech features caused by noise signals or signal processing methods, such as speech enhancement (SE). Missing-data ASR can recognize such distorted speech features $\hat{\mathbf{s}}_{1:T}$ by utilizing their uncertainty. In our SE scenario, an SE module estimates the original but distorted speech features $\hat{\mathbf{s}}_{1:T}$ from a single-channel, noisy observed spectrum $\mathbf{x}_{1:T} = [\mathbf{x}_1, ..., \mathbf{x}_T], \mathbf{x}_t \in \mathbb{C}^F (t = 1, ..., T)$ in the short-time Fourier transform (STFT) domain. The features extracted from the noisy spectrum $\mathbf{x}_{1:T}$ are represented as $\mathbf{u}_{1:T} = [\mathbf{u}_1, ..., \mathbf{u}_T]$, $\mathbf{u}_t \in \mathbb{R}^D (t = 1, ..., T)$. Hereafter, we assume 1) the same frame index t and length L for sequential variables in both the STFT and feature domains and 2) the mel-filterbank speech features for ASR.

The missing-data approach generally smooths a target function or target variables with an evidence model (pdf) that represents the uncertainty of input features [11]. *Small values of the evidence pdf represent large uncertainty.* The smoothing is implemented by evaluating the expectation of the target function or variables with the evidence pdf ε and a parameter set $\mathbf{\Theta}$. Note that the enhanced features $\hat{\mathbf{s}}_{1:T}$ and observed spectrum $\mathbf{x}_{1:T}$ can be included in $\mathbf{\Theta}$, i.e, $\{\hat{\mathbf{s}}_{1:T}, \mathbf{x}_{1:T}\} \in \mathbf{\Theta}$. For example, if the target function is a likelihood function, e.g., acoustic model $p(\mathbf{s}_{1:T}|c_{1:L})$ of ASR based on probabilistic

generative models, the smoothing is implemented by evaluating the expectation of the function with the evidence pdf ε.

$$\mathbb{E}[p(\mathbf{s}_{1:T}|c_{1:L})] := \int p(\mathbf{s}_{1:T}|c_{1:L})\varepsilon(\mathbf{s}_{1:T};\boldsymbol{\Theta})\mathrm{d}\mathbf{s}_{1:T} \quad (4)$$

The assumption or decomposition of (conditional) independence among frames (e.g., HMM) is usually applied to the likelihood and/or the evidence model to simplify the integral over a high dimensional variable $\mathbf{s}_{1:T}$, i.e., to evaluate each pdf of $\mathbf{s}_t$. Here, non-MD-ASR (i.e., conventional ASR) assumes a Dirac delta function for the evidence model: $\varepsilon(\mathbf{s}_{1:T};\boldsymbol{\Theta}) = \delta(\mathbf{s}_{1:T} - \hat{\mathbf{s}}_{1:T})$.

A missing-data approach for the end-to-end ASR based on an encoder-decoder model was recently proposed in [17], where experiments showed that the encoded vectors $\mathbf{h}_{1:T'}$ were effective as target variables. Frame-independent assumptions cannot be applied to the evaluation of the expectation calculation because end-to-end ASRs usually consider the correlation of features among frames by highly nonlinear networks such as self-attention networks. Therefore, the expectation is directly approximated as follows by a Monte-Carlo method:

$$\mathbb{E}[\mathbf{h}_{1:T'}(\mathbf{s}_{1:T})|\mathbf{x}_{1:T}] := \int \mathbf{h}_{1:T'}(\mathbf{s}_{1:T})\varepsilon(\mathbf{s}_{1:T};\boldsymbol{\Theta})\mathrm{d}\mathbf{s}_{1:T}, \quad (5)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\mathbf{h}_{1:T'}(\mathbf{s}_{1:T}^{(n)}),\ \mathbf{s}_{1:T}^{(n)} \sim \varepsilon(\mathbf{s}_{1:T};\boldsymbol{\Theta}), \quad (6)$$

where $\{\mathbf{s}_{1:T}^{(n)}\}_{n=1}^{N}$ are N samples drawn from $\varepsilon(\mathbf{s}_{1:T};\boldsymbol{\Theta})$ and the encoded vectors are then smoothed by the samples.

C. Utterance-wise Uniform Evidence Model

The MD-framework for end-to-end ASRs requires a process of drawing samples from the *utterance-wise* evidence model $\varepsilon(\mathbf{s}_{1:T};\boldsymbol{\Theta})$, of which the dimensions are usually high ($D \times T$ dimensions). Since the sampling from such a large dimensional function is almost impossible, a latent space model is introduced to the utterance-wise evidence to reduce the complexity of the sampling [17]. The utterance-wise evidence model is usually decomposed into the following pdf form to keep the continuity of features among frames ($\boldsymbol{\Theta}$ is omitted here):

$$\varepsilon(\mathbf{s}_{1:T}) = \int p(\mathbf{s}_{1:T}|\eta)p(\eta)\mathrm{d}\eta = \int \Big[\prod_{t,d} p(s_{t,d}|\eta)\Big]p(\eta)\mathrm{d}\eta, \quad (7)$$

where η is a latent variable and the conditional independence among the t and d of $p(\mathbf{s}_{1:T}|\eta)$ is assumed. Note that we cannot adopt the frame-independent model $\varepsilon(\mathbf{s}_{1:T}) = \prod_t p(\mathbf{s}_t)$ for the utterance-wise evidence model, since the samples drawn from this pdf become *noisy features* due to the frame independence and the ASR performance is thus seriously degraded.

The basic evidence model is an utterance-wise uniform pdf [17] extended from the frame-wise uniform pdf [11]. This pdf utilizes only the enhanced features $\hat{\mathbf{s}}_{1:T}$ and the observed features $\mathbf{u}_{1:T}$, and the latent variable η controls the interpolation between $\hat{\mathbf{s}}_{1:T}$ and $\mathbf{u}_{1:T}$:

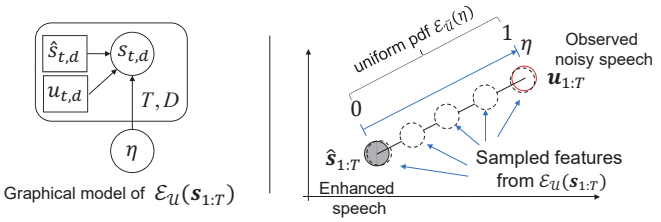


Fig. 2. Graphical model of utterance-wise uniform pdf (left) and samples drawn from it in feature space (right).

$$\varepsilon_{\mathcal{U}}(\mathbf{s}_{1:T};\hat{\mathbf{s}}_{1:T},\mathbf{u}_{1:T}) = \int \Big[\prod_{t,d}\varepsilon_{\bar{\delta}}(s_{t,d};(1-\eta)\hat{s}_{t,d}+\eta u_{t,d})\Big]\varepsilon_{\bar{\mathcal{U}}}(\eta;0,1)\mathrm{d}\eta, \quad (8)$$

$$\varepsilon_{\bar{\delta}}(\eta;\eta_m) = \delta(\eta-\eta_m),\ \varepsilon_{\bar{\mathcal{U}}}(\eta;\eta_l,\eta_u) = \frac{\mathbb{1}(\eta)}{(\eta_u-\eta_l)}, \quad (9)$$

where $s_{t,d}$, $\hat{s}_{t,d}$, and $u_{t,d}$ are the values of the d-th dimensions of $\mathbf{s}_t$, $\hat{\mathbf{s}}_t$, and $\mathbf{u}_t$, respectively. $\varepsilon_{\bar{\delta}}$ and $\varepsilon_{\bar{\mathcal{U}}}$ correspond to a scalar delta pdf and a scalar uniform pdf. η_u, η_l with $\eta_u > \eta_l$ specify the distribution boundaries, and $\mathbb{1}(\eta)$ is an indicator function that is equal to 1 if $\eta \in [\eta_l, \eta_u]$, and 0 otherwise. We can easily draw *frame-continuous* samples from this evidence model by applying ancestral sampling: specifically, we draw η from the uniform distribution and determine $s_{t,d}$ according to the delta function. The samples of η can be deterministically selected at even intervals in $[0, 1]$ with an arbitrary resolution. Figure 2 shows the graphical model of $\varepsilon_{\mathcal{U}}(\mathbf{s}_{1:T})$ and the samples drawn from it.

The posterior probability $p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T})$ for SE can be integrated with the uniform pdf if additional SE statistics are available. The mixture pdf of $\varepsilon_{\mathcal{U}}(\mathbf{s}_{1:T})$ and the posterior is defined with $\boldsymbol{\Theta} = \{\mathbf{x}_{1:T}, \mathbf{u}_{1:T}, \hat{\mathbf{s}}_{1:T}, \pi_1, \pi_2\}$ as follows:

$$\varepsilon_m(\mathbf{s}_{1:T};\boldsymbol{\Theta}) = \pi_1 p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}) + \pi_2\varepsilon_{\mathcal{U}}(\mathbf{s}_{1:T};\hat{\mathbf{s}}_{1:T},\mathbf{u}_{1:T}), \quad (10)$$

where π_1 and π_2 are mixture weights satisfying $\pi_1 + \pi_2 = 1$. Independence among frames is usually assumed in the posterior of SE, such that $p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}) = \prod_t p(\mathbf{s}_t|\mathbf{x}_{1:T})$. However, *this assumption causes the generation of discontinuous features among frames when samples are drawn*, which results in performance degradation of ASR [17]. Therefore, such a posterior of SE cannot be used by itself as the evidence pdf, and the uniform evidence model is necessary as a probabilistic floor.

III. Proposed Method

We first explain the concept behind our new evidence model design by considering the relationship among the feature space of ASR and the uncertainty of SE in terms of ASR performance. We then explain our new evidence model and the unsupervised model selection method. Finally, we present the Bayesian integration of SE and MD-ASR based on a latent space model of SE.

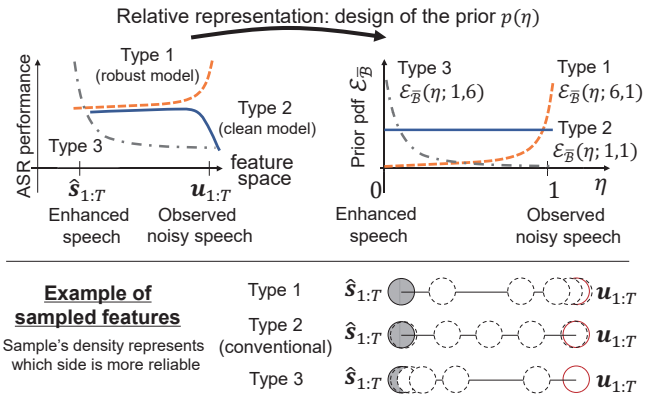


Fig. 3. Performance-feature curve (left) and its projection to our evidence model (right). The difference of drawn samples from each pdf type is also shown in the bottom.

A. Flexible Evidence Model for ASR Models

The structure of the ASR model's feature space is important for designing a new, flexible evidence model. For the ASR performance curve, we focus on a line segment defined by two end points: the enhanced features $\hat{\mathbf{s}}_{1:T}$ and the observed features $\mathbf{u}_{1:T}$, as shown on the upper left in Fig. 3. We assume that the curve can be roughly categorized into three types characterized by a peak around $\mathbf{u}_{1:T}$ (type 1), a dominant "good," flat region (type 2), or a peak around $\hat{\mathbf{s}}_{1:T}$ (type 3). An ASR model trained with clean speech corresponds to type 2, and a robust ASR model corresponds to type 1. The uniform pdf described above can only capture the curve of type 2 and its corresponding uncertainty. The pdf of type 1 is expected to capture the uncertainty-reverse phenomena in the case of robust ASR models.

Hence, we propose a new utterance-wise evidence model based on a beta distribution to represent the performance-feature curve in a relative coordinate space, as shown on the upper right in Fig. 3. This model can represent the previous uniform pdf as a special case. Our evidence model's structure is similar to the utterance-wise uniform pdf given by Eq. (8), but the prior distribution follows a beta distribution as follows:

$$\varepsilon_{\mathcal{B}}(\mathbf{s}_{1:T}; \hat{\mathbf{s}}_{1:T}, \mathbf{u}_{1:T}, \alpha, \beta) = \int \Big[\prod_{t,d} \varepsilon_{\delta}(s_{t,d}; (1-\eta)\hat{s}_{t,d} + \eta u_{t,d})\Big] \varepsilon_{\bar{\mathcal{B}}}(\eta; \alpha, \beta) \mathrm{d}\eta, \quad (11)$$

$$\varepsilon_{\bar{\mathcal{B}}}(\eta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \eta^{\alpha-1}(1-\eta)^{\beta-1} \ (\eta \in [0,1]), \quad (12)$$

where α and β are the beta distribution's shape parameters. For example, a curve of type 1 can be obtained by setting $\alpha = 6$ and $\beta = 1$. For $\alpha = \beta = 1$, the distribution is equivalent to the uniform pdf. We can also deterministically select samples of η at even intervals in $[0, 1]$ from the cumulative density function by applying an inverse transform method, since $\varepsilon_{\mathcal{B}}$ is a scalar density function. The example of drawn samples from each parameter setting of the beta distribution is illustrated at the bottom of Fig. 3. The density of samples represents which side, i.e., $\hat{\mathbf{s}}_{1:T}$ or $\mathbf{u}_{1:T}$, is more reliable.

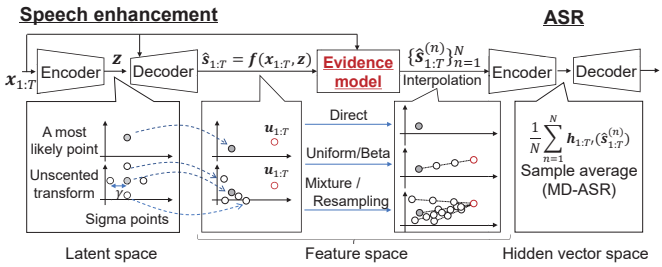


Fig. 4. Overview of deterministic sampling process and MD-ASR.

B. Unsupervised Selection of Appropriate Evidence Model

The flexible model requires us to select proper parameters (e.g., α and β) for the given ASR model and the target speech set $\mathcal{D} = \{\mathbf{x}_{1:T}^{(i)}\}_{i=1}^{N_d}$ (usually a test set). As mentioned above, the patterns of a typical parameter set or evidence model are usually limited; thus, we simply choose the best one from N_m preset models or parameters, $\mathcal{M} = \{\mathcal{M}_j\}_{j=1}^{N_m}$, by maximum likelihood estimation. Therefore, we need no back-propagation process of NNs and no transcriptions of speech data.

The criterion is the following length-normalized ASR score:

$$J(\mathcal{D}|\mathcal{M}_j) = \sum_{i=1}^{N_d} \frac{1}{L_i} \max_{\mathbf{c}_{1:L_i}} J_{asr}(\mathbf{c}_{1:L_i} | \mathbb{E}[\mathbf{h}_{1:T'} | \mathcal{M}_j, \mathbf{x}_{1:T}^{(i)}]). \quad (13)$$

Here, L_i is the length of the i-th recognized characters, and $\mathbb{E}[\mathbf{h}_{1:T'}|\mathcal{M}_j, \mathbf{x}_{1:T}^{(i)}]$ represents the smoothed encoded vector given the j-th parameters/model $\mathcal{M}_j$ and the i-th observed spectrum $\mathbf{x}_{1:T}^{(i)}$.

C. Bayesian Integration with Speech Enhancement

We also propose a Bayesian integration of the evidence model with a new posterior probability modeling of SE, since the posterior is also affected by the given ASR model. This means that mixture integration by Eq. (10) may not perform well for robust ASR models. Because our new posterior model does not assume the frame independence, we can draw samples directly from the pdf of SE without mixture models. Figure 4 presents an overview of the whole model and the sampling process.

An exactly integrated evidence model is derived by treating the enhanced features $\hat{\mathbf{s}}_{1:T}$ as random variables (the parameter set $\boldsymbol{\Theta}$ is omitted here):

$$\varepsilon_f(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}) = \int \varepsilon(\mathbf{s}_{1:T}|\hat{\mathbf{s}}_{1:T}, \mathbf{x}_{1:T}) p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T}) \mathrm{d}\hat{\mathbf{s}}_{1:T}, \quad (14)$$

where $p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T})$ corresponds to the posterior pdf of SE, and $\varepsilon(\mathbf{s}_{1:T}|\hat{\mathbf{s}}_{1:T}, \mathbf{x}_{1:T})$ corresponds to, for example, $\varepsilon_{\mathcal{U}}$ or $\varepsilon_{\mathcal{B}}$. These two evidence models have $\hat{\mathbf{s}}_{1:T}$ and $\mathbf{u}_{1:T}$ (calculated from $\mathbf{x}_{1:T}$) as parameters. Obviously, the posterior pdf is re-weighted by the primary evidence models. Ancestral sampling method is applied to draw samples from $\varepsilon_f(\mathbf{s}_{1:T}|\mathbf{x}_{1:T})$: we first draw $\hat{\mathbf{s}}_{1:T}^*$ from $p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T})$ and then draw $\mathbf{s}_{1:T}^*$ from $\varepsilon(\mathbf{s}_{1:T}|\hat{\mathbf{s}}_{1:T}^*, \mathbf{x}_{1:T})$.

The sampling from the posterior pdf of SE is simplified by assuming the following new latent space model for the SE process:

$$p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T}) = \int p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T}, \mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}, \quad (15)$$

$$\hat{\mathbf{s}}_{1:T} = \mathbf{f}(\mathbf{x}_{1:T}, \mathbf{z}) + \mathbf{n}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (16)$$

Here, $\mathbf{z} \in \mathbb{R}^K$ is a latent vector with dimension K, and $\mathbf{n}$ is a 0-mean noise random variable with *a small variance whose fluctuation is almost negligible*. The function $\mathbf{f}$ represents a deterministic process including SE and feature extraction from $\mathbf{x}_{1:T}$. When the dimension K is small, the sampling of $\mathbf{z}$ is easy because of the Gaussian assumption. The variational posterior pdf $q(\mathbf{z}|\mathbf{x}_{1:T})$ is utilized in the actual sampling phase instead of the prior, as explained in the next section.

We can eliminate the stochastic sampling process by applying the unscented transform (UT) [22] to the latent variable $\mathbf{z}$. The posterior pdf $p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T})$ can be approximated efficiently by a limited number of samples given by sigma points $\{\mathbf{z}^{(k)}\}_{k=1}^{2K+1}$ and their weights $\{W^{(k)}\}_{k=1}^{2K+1}$ under the assumption of $\mathbf{n}$:

$$p(\hat{\mathbf{s}}_{1:T}|\mathbf{x}_{1:T}) \approx \sum_{k=1}^{2K+1} \delta(\hat{\mathbf{s}}_{1:T} - \mathbf{f}(\mathbf{x}_{1:T}, \mathbf{z}^{(k)}))W^{(k)}. \quad (17)$$

The weight definition with a parameter κ follows [22], while the next section explains the sigma points. The deterministic sampling from $\varepsilon(\mathbf{s}_{1:T}|\hat{\mathbf{s}}_{1:T}, \mathbf{x}_{1:T})$ given $\{\mathbf{f}(\mathbf{x}_{1:T}, \mathbf{z}^{(k)})\}_{k=1}^{2K+1}$ and $\{W^{(k)}\}_{k=1}^{2K+1}$ is achieved by systematic resampling [24] without initial drawing.

D. Construction of Speech Enhancement Model

A framework similar to a VAE [25] is applied to learn the functions and parameters in Eqs. (15) and (16) from training data. Once a variational posterior probability q and the function $\mathbf{f}$ are specified by DNNs, training can be implemented by maximizing a lower bound of the log-likelihood given the observed noisy spectrum $\mathbf{x}_{1:T}$, the reference spectrum $\mathbf{y}_{1:T}$ and its features $\mathbf{s}_{1:T}$. We only describe key points here because the SE model itself and the improvement of the SE architecture are outside the scope of the present study.

The function $\mathbf{f}$ (called an decoder) comprises a spectrum mask process and feature extraction process $\mathbf{g}$. Mask-based SE estimates the reference spectrum, $\mathbf{y}_t$, by multiplying soft masks by the observed spectrum $\mathbf{x}_t$ [26], i.e., $\mathbf{y}_t \approx \mathbf{M} \circ \mathbf{x}_t$, where $\circ$ represents element-wise multiplication. The soft-mask vector $\mathbf{M}$ is a function of $\mathbf{x}_{1:T}$ or its subset modeled by DNNs, and each element of $\mathbf{M}$ is in $[0, 1]$.

$$\bar{\mathbf{y}}_t = \log(\mathbf{M}(\tilde{\mathbf{x}}_t, \mathbf{z}) \circ |\mathbf{x}_t|) + \mathbf{n}_t^y, \ \mathbf{n}_t^y \sim \mathcal{L}(\mathbf{0}, \epsilon\mathbf{I}), \quad (18)$$

$$\hat{\mathbf{s}}_t = \mathbf{g}(\bar{\mathbf{y}}_t) + \mathbf{n}_t^s, \quad \mathbf{n}_t^s \sim \mathcal{L}(\mathbf{0}, \epsilon\mathbf{I}), \quad (19)$$

where $\log(\cdot)$ and $|\cdot|$ represent element-wise logarithm and absolute functions, and $\mathcal{L}$ represents a 0-mean Laplace distribution. $\tilde{\mathbf{x}}_t$ is a concatenated vector of $\mathbf{x}_t$ from frame $(t-k)$ to $(t+k)$. ϵ represents a small variance for regularization, which means that $\mathbf{n}_t^y$ and $\mathbf{n}_t^s$ are small process noises. These processes are similar to those in [17] except for $\mathbf{z}$.

The variational posterior $q(\mathbf{z}|\mathbf{x}_{1:T}) = \mathcal{N}(\mathbf{m}, \mathbf{\Lambda})$ (called an encoder) is approximated by means of a VAE framework and used for sampling $\mathbf{z}$. The mean vector $\mathbf{m}$ and the diagonal covariance matrix $\mathbf{\Lambda}$ of q are inferred by DNNs from $\mathbf{x}_{1:T}$ similar to those used for the mask function $\mathbf{M}$, with a sliding window processing, networks to estimate the frame-wise mean and variance, and average pooling of the outputs. This inference networks may not be optimal, but it is enough to prove our concept of SE and MD-ASR integration. Specific configurations are described in the experiments.

Sigma points besides a mean are selected by adding a scaling parameter γ to the i-th element of $\mathbf{m}$, i.e., $m_i \pm \gamma$. The variances estimated by the DNN are empirically ignored for sigma points because they often take small values. In other words, the drawn samples become almost the same with the mean vector $\mathbf{m}$, which produces no effect for smoothing in MD-ASR.

IV. Experiment

We prepared several ASR models trained with different training data to evaluate our evidence model and determine *whether our unsupervised model selection can choose the appropriate one for each ASR model*. The detailed results using a robust ASR model are also discussed in the latter part of the results because of their high ASR performances. Note that our experimental conditions for the ASR models are open in terms of speaker, reverberation and noise sources. We also believe the language spoken in the data-set (e.g., English or Japanese) does not seriously affect the noise robustness of ASR and SE.

A. Experimental Setup

1) Data set for ASR and SE: We conducted experiments with speech data from the Corpus of Spontaneous Japanese (CSJ) [27]. The training set contained about 660 hours of speech (full set) for ASR and 190 hours (3–16 sec. of data in the core set) for SE. The official CSJ evaluation sets (eval{1, 2, 3}; 5 hours total) were used as test sets for MD-ASR. Each piece of speech data was segmented into utterances. The sampling rate for all data was set to 16 kHz. The WHAM! corpus [28] was utilized as a general background noise source for ASR training. The ProSoundEffects (PSE) corpus[1] (also used in [29]) was used as an application-specific, non-speech source for both the SE model's training set and the ASR test sets. The PSE corpus includes environmental sounds, animal noises, music signals, and so on. A validation set was also made from other data that were not included in either the training or test sets.

Three kinds of ASR model were trained on clean speech, reverberant speech, and reverberant speech with background noise, thus yielding (a) a clean speech model (clean), (b) a clean, reverberant speech model (clean+rev.), and (c) a clean, reverberant, noisy speech model (clean+rev.+bgn.). The reverberant speech signals were convolved with real impulse responses measured from 576 positions in a room (RT_{20} 640

[1]Pro Sound Effects Library. http://www.prosoundeffects.com

TABLE I

CHARACTER ERROR RATE (CER: %) FOR EACH SNR CONDITION WITH THREE ASR MODELS TO EVALUATE OUR UNSUPERVISED MODEL SELECTION. CER IS THE TOTAL OF EVAL1, EVAL2, AND EVAL3 TEST SETS. "UT" DENOTES THE UNSCENTED TRANSFORM.

	Conditions	ASR model (a): clean					ASR model (b): clean+rev.					ASR model (c): clean+rev.+bgn.					Total
		Test set SNR (dB)				Avg.	Test set SNR (dB)				Avg.	Test set SNR (dB)				Avg.	
		-5	0	5	10		-5	0	5	10		-5	0	5	10		
Baseline	No proc.	75.3	64.9	48.9	33.1	55.6	65.1	46.9	25.2	13.0	37.5	38.9	20.6	11.4	7.6	19.6	37.6
	Only SE	62.3	49.0	35.0	24.7	42.7	58.2	43.4	28.6	18.8	37.2	55.1	39.6	24.9	16.1	33.9	38.0
Proposed (isolated)	+ $\varepsilon_{\mathcal{U}}$: Uniform - Eq. (8) by [17]	55.9	41.5	27.9	18.6	36.0	50.6	34.3	19.9	11.9	29.2	42.4	25.8	14.7	9.3	23.0	29.4
	+ $p(\hat{\mathbf{s}}_{1:T}\|\mathbf{x}_{1:T})$: UT - Eq. (17)	58.4	44.6	29.7	19.2	38.0	54.2	39.3	23.8	14.4	32.9	51.3	35.0	20.0	12.1	29.6	33.5
	+ $\varepsilon_{\mathcal{B}}(6,1)$: Beta - Eq. (11)	67.0	53.6	37.0	24.6	45.6	56.5	37.9	20.3	11.2	31.5	35.3	19.3	11.0	**7.5**	18.3	31.8
	+ ε_m: UT and Uniform - Eq. (10)	**55.6**	**41.1**	**27.4**	**18.2**	**35.6**	**50.3**	34.3	19.9	12.0	29.1	43.0	26.5	15.1	9.5	23.5	29.4
	+ ε_m: UT and $\varepsilon_{\mathcal{B}}(6,1)$ - Eq. (10)	63.5	49.2	33.2	21.9	42.0	52.8	35.2	19.2	10.9	29.5	**34.8**	19.3	11.1	**7.5**	18.2	29.9
	+ ε_f: UT and Uniform - Eq. (14)	57.6	43.2	28.2	**18.2**	36.8	50.6	34.3	19.0	11.2	28.8	39.6	23.9	12.6	8.3	20.9	28.8
	+ ε_f: UT and $\varepsilon_{\mathcal{B}}(6,1)$ - Eq. (14)	68.9	56.3	39.6	26.0	47.7	58.1	39.8	21.5	11.7	32.8	35.1	**19.0**	**10.9**	**7.5**	**18.1**	32.9
Proposed	Model selection: Eq. (13)	56.5	41.2	**27.4**	18.3	35.9	50.4	**33.9**	**18.8**	**10.8**	**28.5**	35.2	19.2	11.1	**7.5**	18.2	**27.7**

ms) to simulate a real acoustic environment. The mixture SNRs of the background noise from WHAM! were randomly set to $0, 5, 10, 15, 20$ dB. Note that the speech data used for model (c) comprised $660 \times 3 = 1980$ hours in total.

The training data for the SE model were a mixture of reverberant speech and non-speech signals. Specifically, each speech signal was combined with a non-speech signal selected randomly from the PSE data. The mixture SNRs here were randomly set to $-5, 0, 5, 10, 15$ dB. The reference spectrum was the first arriving frame in the reverberant speech.

The test set data also were a mixture of reverberant speech and raw non-speech signals with different SNRs to investigate the robustness of our MD-ASR. We utilized a different impulse response recorded 1.5 m from a microphone in a different room. The speech source signals were from the CSJ test set, while the non-speech signals were from PSE data that were not used for the SE model's training set. The SNRs were $-5, 0, 5, 10$ dB.

2) DNN and MD-ASR Configurations: The ASR model was ESPnet [2] with a Transformer encoder-decoder acoustic model, which we selected because it performed the best for noisy data compared with a Conformer model [30]. The STFT parameters were a 512-point Hanning window and 128-point shift. An 80-dimensional mel-filterbank was used for the speech features, and global mean-and-variance normalization was applied. Speech-rate perturbation and *SpecAug* techniques were also applied to make the model more robust. See ESPnet's CSJ recipe and configuration files for the details of the networks, RNN language model, and hyper parameters.

The SE model was built from scratch with the PyTorch library and its pre-defined classes [1]. The STFT parameters were the same as those for ESPnet. The feature extraction networks comprised 80-dimensional complex filterbank networks, absolute and power activation functions, concatenation of 32 or 4 frames before and after the center frame, and positional encoding. A 6 and 2-layer transformer-encoder with eight heads and 512 hidden nodes was utilized as the main network to estimate the mask $\mathbf{M}$ and latent variable $\mathbf{z}$, respectively, and they were followed by a fully connected layer. The nonlinear function of the mask's head was a sigmoid function, and that for the frame-wise variance for $\mathbf{z}$ before average pooling was an absolute value function. Gradient clipping [31] and Adam [32] were applied. All parameters were updated for each utterance, i.e., the mini-batch size was 1. The learning rate was 1.0×10^{-4}. The parameter set at 50 epochs was used for the experiments.

TABLE II

CHARACTER ERROR RATE (CER: %) FOR EACH SNR CONDITION WITH ESPNET CONFORMER AND WHISPER MODEL

	Configuration	Test set SNR (dB)				Avg.
		-5	0	5	10	
ESPnet	Conformer (clean)	87.4	80.5	69.7	54.2	72.9
	Conformer (clean + rev.)	73.8	59.2	37.1	18.2	47.1
Whisper	large-v2	71.6	55.3	39.7	33.9	50.1
	large-v2 (text-normalized)	61.1	45.4	33.0	26.4	41.5

The decoding parameters of ASR were almost the same as those of the ESPnet CSJ recipe. The beam size was 20, and the respective weights for the CTC, attention, and language model were 0.5, 0.5, and 0.3. The number of samples N used for the isolated evidence models (e.g., uniform and beta) was 8. The sampling parameters in other cases were $N = 40$, $\pi_1 = 0.125$, and $\kappa = 0.5$. The parameters (α, β) of $\mathcal{E}_{\mathcal{B}}$ and γ were tested on the combinations of $\{(1,1),(2,1),(4,1),(6,1),(8,1)\}$ and $\{4, 8\}$, respectively. Our model selection method automatically chooses the best result from these parameter sets for each eval set and SNR.

B. Results among Different ASR Models for Evaluation of Unsupervised Model Selection

Table I lists the character error rate (CER) for each method with different ASR models and SNR conditions. A smaller CER means a better ASR performance, and the total number of characters over eval{1,2,3} was 115,745. The "Avg." columns indicate the averaged performance over all SNRs with each ASR model, while the "Total" column indicates the averaged performance over all ASR models. "No proc." refers to direct recognition of the noisy speech signal. The baselines were the methods with only SE and SE with the uniform evidence model. "ASR model (c)" is considered as a *robust* ASR model.

We found that SE seriously degraded the performance by 14 points compared to "No proc." in the case of the robust ASR model, which was due to the uncertainty mismatch between SE and ASR. The uniform evidence model improved the CER

TABLE III
DETAILED CHARACTER ERROR RATE (CER: %) OF EVAL1, EVAL2, AND EVAL3 TEST SETS FOR EACH SNR CONDITION WITH ROBUST ASR MODEL (CLEAN+REV.+ BGN.). "UT" DENOTES THE UNSCENTED TRANSFORM.

	Conditions	eval1					eval2					eval3					Total
		Test set SNR (dB)				Avg.	Test set SNR (dB)				Avg.	Test set SNR (dB)				Avg.	
		-5	0	5	10		-5	0	5	10		-5	0	5	10		
	No proc.	38.8	21.0	12.3	8.5	20.2	36.9	19.4	10.5	6.8	18.4	42.2	21.9	11.4	7.3	20.7	19.6
Baseline	Only SE	57.1	43.8	30.1	20.5	37.9	56.4	41.1	26.4	17.0	35.2	62.8	46.1	29.7	18.8	39.3	33.9
	+ $\varepsilon_{\mathcal{U}}$: Uniform - Eq. (8) by [17]	49.8	34.9	21.2	13.3	29.8	48.5	32.3	18.1	10.5	27.4	54.9	36.6	20.7	12.1	31.1	23.0
Proposed (isolated)	+ $p(\hat{\mathbf{s}}_{1:T}\|\mathbf{x}_{1:T})$: UT - Eq. (17)	51.0	36.2	21.7	13.7	30.6	49.5	33.0	18.2	10.7	27.9	54.6	36.1	20.0	11.7	30.6	29.6
	+ $\varepsilon_{\mathcal{B}}(6,1)$: Beta - Eq. (11)	35.6	20.0	**12.0**	**8.3**	19.0	32.9	17.8	10.1	**6.7**	16.9	38.6	20.5	11.0	7.3	19.3	18.3
	+ ε_m: UT and Uniform - Eq. (10)	43.4	27.9	16.4	10.7	24.6	41.0	24.5	13.7	8.5	21.9	45.6	27.5	15.2	9.2	24.3	23.5
	+ ε_m: UT and $\varepsilon_{\mathcal{B}}(6,1)$ - Eq. (10)	35.4	20.0	12.1	8.4	19.0	**32.5**	17.8	10.2	6.8	16.8	**37.4**	20.4	11.0	7.3	**19.0**	18.2
	+ ε_f: UT and Uniform - Eq. (14)	39.8	24.1	14.0	9.2	21.8	37.9	21.2	11.5	7.6	19.5	42.1	24.2	12.3	7.9	21.6	20.9
	+ ε_f: UT and $\varepsilon_{\mathcal{B}}(6,1)$ - Eq. (14)	**35.2**	**19.5**	**12.0**	**8.3**	**18.8**	32.9	**17.7**	**9.9**	6.8	**16.8**	38.4	**20.2**	**10.9**	**7.2**	19.2	**18.1**
Proposed	Model selection: Eq. (13)	35.6	19.8	12.1	**8.3**	18.9	32.7	17.9	10.2	**6.7**	16.9	38.4	**20.2**	**10.9**	7.3	19.2	18.2

but did not outperform "No proc." The same phenomena were found at 5 and 10 dB with the ASR model (clean+rev.). Thus, the simple connection of speech enhancement and ASR does not necessarily impove CERs.

As for isolated evaluation of the *fixed* evidence model, our "Beta" evidence model outperformed "No proc." by 1.3 points in "Avg." with the robust ASR model. Our evidence model design was matched for the robust ASR model. The sampling from speech enhancement (only UT by Eq. (17)) also improved the CER by 4.5 points compared to those of only SE in the "Avg." condition. This result indicated the effectiveness of speech enhancement based on encoder-decoder architechture for drawing utterance-wise samples . The Bayesian integration with SE (UT and beta) improved the CER by about 0.2 points more in the "Avg." condition. However, these settings did not work well with the ASR model trained by clean speech. In that case, the uniform evidence model and integration with SE outperformed the "Only SE" method in any SNRs.

On the other hand, our model selection could choose a better evidence model according to each eval set and SNR condition, and its total CER, 27.7, was the best. This result indicated that the ASR score from DNNs was reliable to some extent for model selection. However, some CERs deteriorated under certain conditions, and a more accurate criterion for model selection is necessary.

The CER (Avg.) of "No proc." for ESPnet Conformer and Whisper [33] are shown in Table II as references. The Conformer models were trained with clean and clean+rev. data sets, and fine-tuning was not applied to the Whisper "large-v2" model. Because the transcription rules used in CSJ and Whisper are different, normalization rules designed by manual (not perfect) were applied to the recognition results from the Whisper model. For example, we removed punctuation, symbols and abnormal recognition results, and we also unified representations of numbers and some orthographic variants. The CERs after the text normalization is represented as "large-v2 (text-normalized)." CERs of the Conformer models were obviously worse than those of Transformer in Table I (by over 9 points), which means the Conformer models are not robust against unknown noise sources that were not included in training data. CERs of Whisper models were also worse than those of Transformer trained with the "clean+rev.+bgn." dataset. This is partly because of a language model mismatch between the raw Whisper model and the CSJ dataset, e.g., orthographic variants and handling of fillers and hesitations.

C. Detailed Results using Robust ASR Model for Evaluation of Proposed Evidence Models

Table III lists the detailed CER for each method with the robust ASR model (clean+reverb.+bgn.) and SNR conditions to clarify the performance among different datasets and the effectiveness of our evidence models. This ASR model corresponds to "ASR model (c)" in Table I. The total number of characters of eval1, eval2, and eval3 was 43,897, 43,623, and 28,225, respectively. The "Avg." columns indicate the averaged performance over all SNRs, while the "Total" column indicates the averaged performance over eval sets. "No proc." refers to direct recognition of the noisy speech signal. The baselines were the methods with only SE and SE with the uniform evidence model. Since the 8.5, 6.8, and 7.3 CERs of "No proc." under the 10-dB condition were close to that of clean speech, the improvement under low SNR conditions is important in this experiment.

The overall tendency in terms of performance among different test sets (eval1, eval2, and eval3) was similar to the results in Table I. As for the isolated evaluation of the evidence model, our "Beta" evidence model outperformed "No proc." by 1.3 points in total and by over 3 points under the -5-dB condition. Since the CERs of only SE were worse than those of "No proc.", *our evidence model can utilize the poor results of SE*. The Bayesian integration with SE (UT and Beta) improved the CER by about 0.2 points more for some test sets. Although the unsupervised model selection also performed well, we can expect the selection accuracy to be improved in the future.

We found that our evidence models based on beta distribution could perform even with the robust ASR model, especially under noisy low SNR conditions. Therefore, the improvement under low SNR conditions (e.g., -10, -5, and 0 dB) will be the focus of future noise-robust ASR and MD-ASR research. The improvement without fine-tuning is challenging, so our findings here provide a useful setting for the era of integration

among huge neural model, such as ASR, SE, language models and dialogue models.

V. Conclusion

In this paper, we proposed a flexible evidence model to reduce the uncertainty mismatch between SE and ASR model. Our evidence model filters out the uncertainty of SE, and effective parameters for the model are selected in accordance with ASR scores in an unsupervised manner. Experiments showed that the CER improved by more than 3 points over the best baseline even with a robust ASR model under a -5-dB SNR condition.

Our future work will focus on evaluations using other pre-trained ASR models, real-time processing and the investigation of precise evidence models. The integration of MD-ASR with microphone array processing is another challenge, which will require a different design for the evidence model because the amount of available information (e.g., the directional patterns of arriving sound signals) increases.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number JP23H03457.

References

[1] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. of Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[2] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. of Interspeech*, 2018, pp. 2207–2211.

[3] C. Li, J. Shi, W. Zhang, *et al.*, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. of IEEE Spoken Language Technology Workshop*, 2021, pp. 785–792.

[4] A. C. Morris, M. P. Cooke, and P. D. Green, "Some solution to the missing feature problem in data classification, with application to noise robust ASR," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 1998, pp. 737–740.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[6] A. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: Soft data modelling for noise robust ASR," in *Proc. of Workshop Innovation Speech Process*, 2001.

[7] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. of Eurospeech*, 1999.

[8] S. Yamamoto, J.-M. Valin, K. Nakadai, *et al.*, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proc. of IEEE International Conference on Robotics and Automation*, 2005, pp. 1477–1482.

[9] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 82–85.

[10] D. Kolossa, H. Sawada, R. F. Astudillo, R. Orglmeister, and S. Makino, "Recognition of convolutive speech mixtures by missing feature techniques for ICA," in *Proc. of Fortieth Asilomar Conference on Signals, Systems and Computers*, 2006, pp. 1397–1401.

[11] M. Kuhne, R. Togneri, and S. Nordholm, "A new evidence model for missing data speech recognition with applications in reverberant multi-source environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 372–384, 2011.

[12] M. Kuhne, "Handling derivative filterbank features in bounded-marginalization-based missing data automatic speech recognition," in *Proc. of Interspeech*, 2015, pp. 3566–3570.

[13] A. H. Abdelaziz, S. Watanabe, J. R. Hershey, E. Vincent, and D. Kolossa, "Uncertainty propagation through deep neural networks," in *Proc. of Interspeech*, 2015, pp. 3561–3565.

[14] Y. Tachioka and S. Watanabe, "Uncertainty training and decoding methods of deep neural networks based on stochastic representation of enhanced features," in *Proc. of Interspeech*, 2015, pp. 3541–3545.

[15] K. Nathwani, J. A. Morales-Cordovilla, S. Sivasankaran, *et al.*, "An extended experimental investigation of DNN uncertainty propagation for noise robust ASR," in *Proc. of Hands-free Speech Communications and Microphone Arrays*, 2017, pp. 26–30.

[16] J. Novoa, J. Fredes, V. Poblete, and N. B. Yoma, "Uncertainty weighting and propagation in DNN-HMM-based speech recognition," *Computer Speech & Language*, vol. 47, pp. 30–46, 2018.

[17] R. Takeda, Y. Sudo, N. Kazuhiro, and K. Komatani, "Empirical sampling from latent utterance-wise evidence model for missing data asr based on neural encoder-decoder model," in *Proc. of Interspeech*, 2022, pp. 3789–3793.

[18] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *Proc. of NIPS Workshop on Deep Learning*, 2014.

[19] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. of Advances in Neural Information Processing Systems*, vol. 8, 2015.

[20] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 4960–4964.

[21] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[22] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," in *Proc. of Signal Processing, Sensor Fusion, and Target Recognition VI*, vol. 3068, 1997, pp. 182–193.

[23] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan, "The unscented particle filter," in *Proc. of International Conference on Neural Information Processing Systems*, 2000, pp. 563–569.

[24] R. Douc, O. Cappé, and bibinitperiod Moulines, "Comparison of resampling schemes for particle filtering," *Proc. of International Symposium on Image and Signal Processing and Analysis*, pp. 64–69, 2005.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. of International Conference on Learning Representations*, 2014.

[26] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[27] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[28] G. Wichern, J. Antognini, M. Flynn, *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. of Interspeech*, 2019, pp. 1368–1372.

[29] I. Kavalerov, S. Wisdom, H. Erdogan, *et al.*, "Universal sound separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 175–179.

[30] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, *Conformer: Convolution-augmented transformer for speech recognition*, 2020. arXiv: 2005.08100.

[31] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. of International Conference on Machine Learning - Volume 28*, 2013, pp. III-1310–III-1318.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of International Conference on Learning Representations*, 2015.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Investigating the Effectiveness of Speaker Embeddings for Shout Intensity Prediction

Takahiro Fukumori*, Taito Ishida*, and Yoichi Yamashita*
*Ritsumeikan University, Japan
E-mail: {fukumori@fc, is0394ie@ed, yyama@is}.ritsumei.ac.jp

***Abstract*—The automatic detection of shouted speeches has attracted much research attention as a core technology of audio surveillance systems. A common strategy in the past literature has been to train a binary classifier using labels of shouted or normal speeches. Although it is known that the acoustic properties of shouted speech usually differ among speakers, especially male and female groups, the conventional methods did not pay attention to encoding such personal and gender-related style information. There are recent findings that speaker embeddings, which are produced by a model trained for speaker identification, can improve other tasks such as speech emotion classification. Thus, this paper investigates the effectiveness of such speaker embeddings for a shouted speech detection problem. Specifically, we verify whether x-vector embeddings can work as effective auxiliary features for the target problem compared with a simple gender label of an input speech whose ground truth is generally unavailable in real situations. Our experiments on predicting the shout intensity beyond the traditional binary classification demonstrated the x-vector embeddings achieved performance improvement over the single use of speech features.**

I. Introduction

The maintenance of a safe living environment within society has increasingly become a priority, in which the early detection of suspicious or dangerous situations is essential for efficient rescue operations. Shouted speech detection is the fundamental task involved in many audio surveillance systems, and several machine learning approaches have been presented based on labeled datasets [1]–[9]. Most conventional studies treated a shouted speech detection problem as a binary classification of "normal" and "shouted." Pohjalainen et al. [1] and Nandwana et al. [2] used Gaussian mixture models, in which normal and shouted speeches were separately modeled using acoustic features. Similarly, Huang et al. [3] and Sharma et al. [8] employed support vector machines to find a decision boundary of shouted and normal speeches. More recent studies often employ deep learning-based approaches to improve classification performance [4]–[6]. For example, Baghel et al. [6] trained convolutional neural networks (CNNs) to capture local features that were effective for classification. Saeed et al. [4] introduced long short-term memory to a classifier to capture temporal structures. Our previous study [5] presented an architecture combining the CNNs with gated recurrent units (GRUs), showing state-of-the-art classification performance.

Each of the above conventional studies trained a classification model from scratch using a dataset of speech labeled either shouted or normal, without considering speaker attributes. However, conventional studies analyzing the acoustic properties of speech have shown personal variations in shouting [10]. Moreover, it is well known that the fundamental frequency of speech is affected by the speaker's gender [11]. Our recent study also revealed that the average intensity of shouted speeches differed among speakers, especially between male and female groups [12]. Unfortunately, there is currently no method that considers these variations in shouted speech detection.

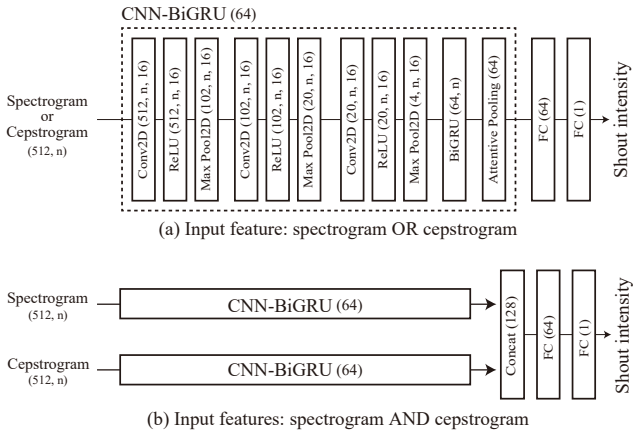
To extract such personal and gender-related speech features, we propose using an additional training resource from speaker identification research, a crucial technique in speech signal processing. Recent studies have demonstrated the applicability of models that were pre-trained for speaker identification to other related tasks, such as speech emotion classification [13]–[16]. For instance, in [14], a pre-trained speaker identification model converted variable-length speech segments to a fixed-length vector (namely, speaker embeddings), which complemented speech style information that other speech features lacked, enhancing the overall performance of the target speech classification task. Inspired by these pioneering findings, this paper aims to investigate whether the embeddings from the speaker identification model can also improve the performance of shouted speech detection.

Using existing neural architectures for shouted speech detection, this paper compares x-vector embeddings [17], a vector representation produced by a pre-trained speaker identification model, with gender labels. Since the gender label is generally unavailable in real situations, this comparison intends to confirm if x-vector embeddings can encode more effective, speaker-specific information than ground-truth gender labels under the target problem. We conducted experiments to predict shout intensity beyond the traditional binary classification using the dataset of Japanese speakers' speeches [12]. The results showed that the additional use of speaker embeddings effectively improved the prediction performance compared with the single use of speech features.

II. Methodology of shout intensity prediction

This section describes three elements of our shout intensity prediction method: speech features, speaker-specific features, and neural network architectures that use multiple features for prediction.

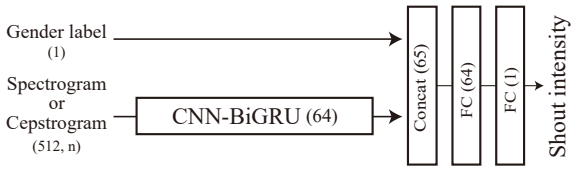
979-8-3503-0067-3/23/$31.00 ©2023 IEEE



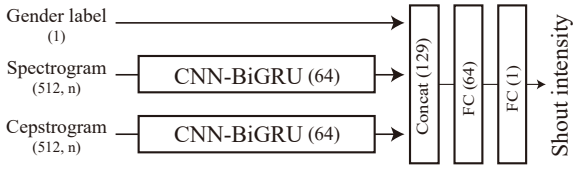
(a) Input feature: spectrogram OR cepstrogram

(b) Input features: spectrogram AND cepstrogram

Fig. 1. Conventional network architectures used in our previous work [12]. The size of each layer's output is represented in parentheses. In addition, n indicates the total number of frames.



(a) Input features: gender label + spectrogram OR cepstrogram



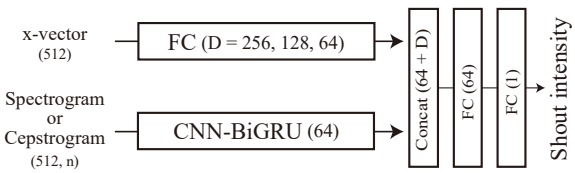
(b) Input features: gender label + spectrogram AND cepstrogram

Fig. 2. Network architectures using a gender label as an auxiliary feature. The size of each layer's output is represented in parentheses. Additionally, n indicates the total number of frames.

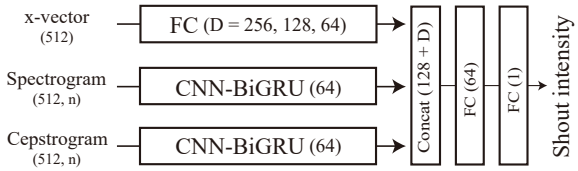
A. Speech features

For a given audio segment, we first partition it into successive n frames using a Hamming window with a length of 1,024 points (i.e., 64 ms) and a hop length of 512 points (i.e., 32 ms). Then, we obtained two types of speech features, the spectrogram and the cepstrogram, similar to [12]:

- **Spectrogram**: A spectrogram represents the temporal variation of a spectrum. Specifically, by applying the short-term Fourier transform to a speech signal, we obtain a 512-dimensional vector representing the power spectrum for each frame. Concatenating these vectors for n frames results in a spectrogram vector of size $512 \times n$.
- **Cepstrogram**: Applying the inverse discrete Fourier transform to the log power spectrum yields the cepstrum, and concatenating the cepstra of multiple frames yields the cepstrogram. The cepstrogram represents the temporal variation in the vocal tract and vocal cords. We set the dimensionality of each cepstrum to be the same as that of the spectrogram, that is 512. This results in a cepstrogram vector of size $512 \times n$ dimensions.



(a) Input features: x-vector + spectrogram OR cepstrogram



(b) Input features: x-vector + spectrogram AND cepstrogram

Fig. 3. Proposed network architectures using x-vectors as auxiliary features with fully connected layers. The size of each layer's output is represented in parentheses. Moreover, n indicates the total number of frames.

B. Speaker-specific features

In this paper, we compare the following two types of speaker-specific features as auxiliary information for the speech features.

- **Gender label**: We employ a binary labeling scheme to represent a gender, where females are assigned the label "0" and males are assigned the label "1". In real situations, we should automatically predict these labels using a certain classification model for a given speech input. However, our experiments use ground-truth data to examine how gender information can guide shout intensity prediction.
- **X-vector embeddings**: The x-vector is a state-of-the-art speaker identification method based on a time delay neural network trained in a supervised manner. Once the training is completed, it is used for obtaining speaker embeddings from variable-length utterances. Since our shout-intensity-prediction experiments are based on a dataset of Japanese speakers' speeches, the x-vector model should also be constructed using a speaker-identified dataset of Japanese speeches. Thus, we use a Japanese model provided in [18], which was pre-trained using a Japanese corpus, JTubeSpeech [19]. For input speeches, the pre-trained model produces 512-dimensional x-vector embeddings.

Most conventional methods used speech features only, while our study additionaly uses speaker-specific features (either gender label or x-vector embeddings) to predict the shout intensity.

TABLE I
EXPERIMENTAL CONDITIONS.

Shouted speech	2,500 samples - 50 sentences × 50 speakers - Training: 2,000 samples - Testing: 1,000 samples
Sampling frequency	16 kHz
Quantization	16 bit
Number of epochs	100
Batch size	8
Learning rate	0.0001
Optimizer	Adam
Loss function	Root mean squared error (RMSE)

C. Network architecture

As shown in figures 1 to 3, we prepare different neural network architectures for each combination use of the features. Figure 1 depicts base architectures presented by our recent work [12], which uses only speech features (i.e., the spectrogram and the cepstrogram). Figure 1(a) corresponds to a model that uses only a single feature. Specifically, the speech features were fed into the convolutional neural network and bidirectional gated recurrent unit (CNN-BiGRU) block. The CNN-BiGRU block consists of three convolutional layers, two max pooling layers, one BiGRU layer, and one attentive pooling layer. When using both a spectrogram and a cepstrogram, they are fed into separate CNN-BiGRU blocks, as in Figure 1(b).

Figure 2 shows the architectures when using not only speech features but also a gender label, which has two cases depending on the number of speech features used. Since the gender label is one-dimensional, the feature is directly concatenated to the outputs of CNN-BiGRU blocks.

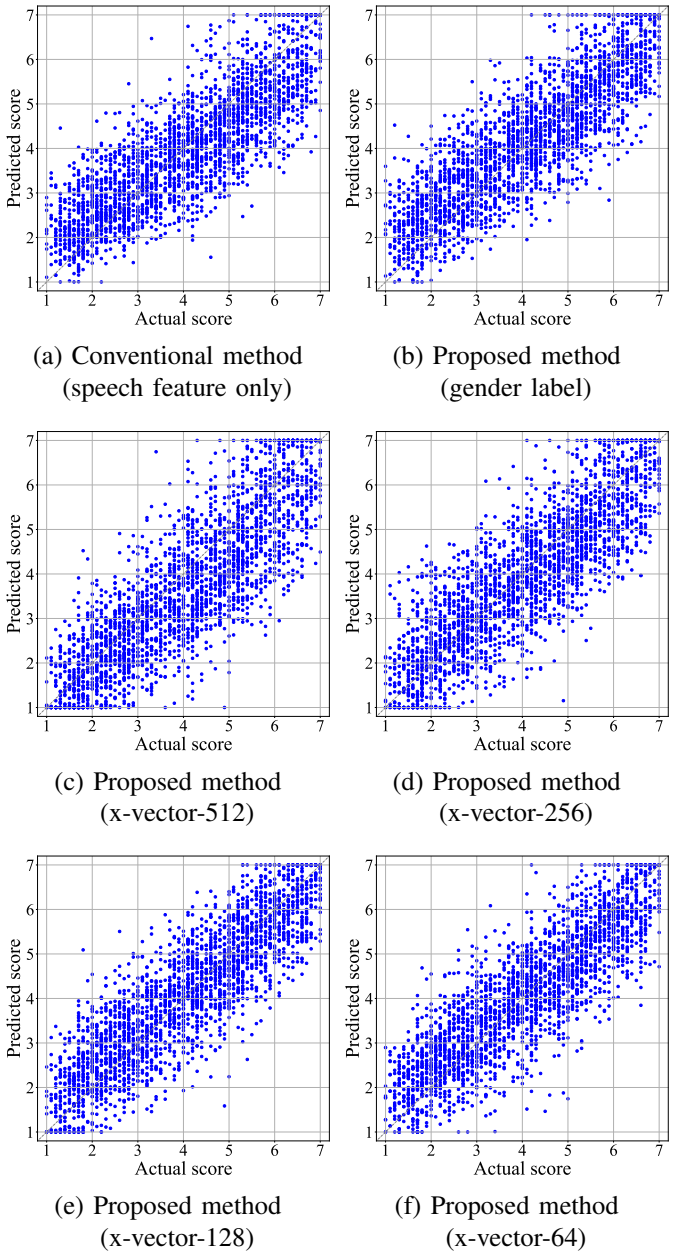
Figure 3 shows the architectures when using x-vector embeddings as auxiliary features. There are also two cases depending on the number of speech features used. The embeddings originally have 512 dimensions, and the dimension reduction can be achieved with a fully connected (FC) layer. The choices used in this paper are 256, 128, or 64 dimensions, in addition to 512 dimensions (i.e., no dimension reduction).

In all networks, two FC layers are placed at the end. When multiple speech features are used as shown in figures 1(b), 2(a)(b), and 3(a)(b), the outputs from the CNN-BiGRU and speaker information are concatenated using the concatenation block before being inputted to the FC layers. If only one feature is used, as described in figure 1(a), the output from the CNN-BiGRU block is directly fed into the FC layers.

III. EXPERIMENTS

A. Experimental settings

To investigate the effectiveness of the speaker-specific features described in Section II.B, we conducted an experiment using the Ritsumeikan shout corpus (RISC) [12], comprising Japanese speakers' shouted speeches. Each shouted speech in the RISC is assigned a numerical value corresponding to shout intensity, ranging from one to seven. The speeches having values of one are nearly equal to normal speeches. There are a total of 2,500 utterances (50 speakers × 50 sentences), collected from 50 speakers (21 females and 29 males).



(a) Conventional method (speech feature only)
(b) Proposed method (gender label)
(c) Proposed method (x-vector-512)
(d) Proposed method (x-vector-256)
(e) Proposed method (x-vector-128)
(f) Proposed method (x-vector-64)

Fig. 4. Scatter plot (Speech features: Spectorgram + Cepstrogram).

Table I presents the detailed experimental settings. We first normalized all speech data so that the maximum amplitude was one. The normalized data was then downsampled to a sampling frequency of 16,000 Hz. Subsequently, the 2,500 utterances were divided into 2,000 training utterances and 500 validation/test utterances, followed by a five-fold cross-validation. When extracting the speech features described in Section II.A from all speech samples to match the maximum frame length within each mini-batch, we performed zero-padding for each mini-batch. Regarding the learning conditions

TABLE II
RMSEs BETWEEN ACTUAL AND PREDICTED SCORES. (↓)

Speech features	Speech features only (Conventional)	Augmented with speaker-specific features (Proposed)				
		Gender label	x-vector-512	x-vector-256	x-vector-128	x-vector-64
Spectrogram	1.12	0.99	0.98	0.81	**0.79**	0.81
Cepstrogram	0.85	**0.79**	0.97	0.81	0.82	**0.79**
Spectorgram + Cepstrogram	0.87	0.82	0.92	0.82	**0.79**	0.80

TABLE III
CORRELATION COEFFICIENTS BETWEEN ACTUAL AND PREDICTED SCORES. (↑)

Speech features	Speech features only (Conventional)	Augmented with speaker-specific features (Proposed)				
		Gender label	x-vector-512	x-vector-256	x-vector-128	x-vector-64
Spectrogram	0.67	0.77	0.84	0.85	**0.86**	0.85
Cepstrogram	0.82	0.85	0.83	0.85	0.84	0.85
Spectorgram + Cepstrogram	0.82	0.84	0.84	**0.86**	**0.86**	0.85

of the network, we set the number of epochs to 100, the batch size to 8, and employed Adam [20] as the optimization method with a learning rate of 0.0001.

B. Results

We used the root mean square error (RMSE) and the correlation coefficient between the true and predicted intensity values as evaluation metrics. Tables II and III present the experimental results of intensity prediction for each combined use of features. The numbers following "x-vector-" in the table indicate the dimensionality after FC layer-based dimension reduction. Note that "x-vector-512" refers to the method that directly inputs the extracted x-vectors into the concatenation block without passing through the FC layer. Figure 4 shows scatter plots of the true and predicted intensity scores obtained using the "Spectrogram + Cepstrogram" feature combination, which achieved the higher prediction accuracy among all methods. We can see that the results of x-vector-128 displayed a small variance of the distribution compared with other models.

Comparing the overall conventional and proposed methods, the proposed method, which utilizes speaker information such as a gender label and an x-vector, demonstrated superior performance. The best scores for the RMSE and the correlation coefficient were 0.79 and 0.86, respectively, outperforming the conventional methods. When focusing on methods that utilize x-vectors, it was found that reducing the dimensionality through the FC layer, instead of directly using the 512-dimensional speaker embedding vectors, contributed to performance improvement.

The analysis based on each feature revealed that the use of both a spectrogram and a 128-dimensional x-vector resulted in improvements of 0.33 in RMSE and 0.19 in the correlation coefficient compared to the conventional method that only employed a spectrogram. Similarly, incorporating either a gender label or a 64-dimensional x-vector into a cepstrogram led to an RMSE improvement of 0.06 over the conventional method that exclusively utilized a cepstrogram. Furthermore, the combination of spectrogram, cepstrogram, and a 128-dimensional x-vector yielded improvements of 0.08 in RMSE and 0.04 in the correlation coefficient compared to not using an x-vector. Overall, the method that utilized x-vectors as a representation of the speaker demonstrated superior performance. These findings suggest that the utilization of information-rich x-vectors contributes to enhancing the performance of intensity prediction. Thus, the effectiveness of directly inputting speaker representation information into the prediction model has been confirmed by these results.

IV. CONCLUSION

This paper presented a shout intensity prediction method that augments speech features using x-vector embeddings as speaker-specific features. The results of the experiments demonstrated that incorporating x-vector embeddings or gender labels significantly improved the prediction performance. This indicates the effectiveness of leveraging a pre-trained model from the speaker identification task. In future work, we will conduct more large-scale experiments and further investigate how to characterize speaker styles in shouted speech detection.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP21K14381.

REFERENCES

[1] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4968–4971.

[2] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 161–165.

[3] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," *IEEE Conference on Industrial Electronics and Applications*, pp. 2115–2120, 2010.

[4] F. S. Saeed, A. A. Bashit, V. Viswanathan, and D. Valles, "An initial machine learning-based victim's scream detection analysis for burning sites," *Applied Sciences*, vol. 11, no. 18, 2021.

[5] T. Fukumori, "Deep spectral-cepstral fusion for shouted and normal speech classification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 4174–4178.

[6] S. Baghel, M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, "Shouted and normal speech classification using 1D CNN," *Pattern Recognition and Machine Intelligence*, pp. 472–480, 2019.

[7] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[8] A. Sharma and S. Kaul, "Two-stage supervised learning-based method to detect screams and cries in urban environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 290–299, 2016.

[9] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, "Audio-based event detection at different SNR settings using two-dimensional spectrogram magnitude representations," *Electronics*, vol. 9, no. 10, 2020.

[10] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech," in *The Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Aug. 2013.

[11] S. Baghel, S. R. M. Prasanna, and P. Guha, "Exploration of excitation source information for shouted and normal speech classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 1250–1261, Feb. 2020.

[12] T. Fukumori, T. Ishida, and Y. Yamashita, *RISC: A corpus for shout type classification and shout intensity prediction*, 2023. arXiv: 2306.04143.

[13] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale CNN and attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3020–3024.

[14] O. C. Phukan, A. B. Buduru, and R. Sharma, "Transforming the embeddings: A lightweight technique for speech emotion recognition tasks," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[15] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173.

[16] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "CopyPaste: An augmentation method for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6324–6328.

[17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[18] *X-vector extractor for Japanese speech*, https://github.com/sarulab-speech/xvector_jtubespeech, (Accessed on 06/01/2023).

[19] *JTubeSpeech: Corpus of Japanese speech collected from youtube*, https://github.com/sarulab-speech/jtubespeech, (Accessed on 06/01/2023).

[20] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Is the Ideal Ratio Mask Really the Best? — Exploring the Best Extraction Performance and Optimal Mask of Mask-based Beamformers

Atsuo Hiroe*, Katsutoshi Itoyama*†, and Kazuhiro Nakadai*
* Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan
† Honda Research Institute Japan Co., Ltd., Saitama, Japan
E-mail: {hiroe,itoyama,nakadai}@ra.sc.e.titech.ac.jp

***Abstract*—This study investigates mask-based beamformers (BFs), which estimate filters to extract target speech using time-frequency masks. Although several BF methods have been proposed, the following aspects are yet to be comprehensively investigated. 1) Which BF can provide the best extraction performance in terms of the closeness of the BF output to the target speech? 2) Is the optimal mask for the best performance common for all BFs? 3) Is the ideal ratio mask (IRM) identical to the optimal mask? Accordingly, we investigate these issues considering four mask-based BFs: the maximum signal-to-noise ratio BF, two variants of this, and the multichannel Wiener filter (MWF) BF. To obtain the optimal mask corresponding to the peak performance for each BF, we employ an approach that minimizes the mean square error between the BF output and target speech for each utterance. Via the experiments with the CHiME-3 dataset, we verify that the four BFs have the same peak performance as the upper bound provided by the ideal MWF BF, whereas the optimal mask depends on the adopted BF and differs from the IRM. These observations differ from the conventional idea that the optimal mask is common for all BFs and that peak performance differs for each BF. Hence, this study contributes to the design of mask-based BFs.**

I. Introduction

Target speech extraction is effective in improving speech intelligibility in telecommunication systems and the performance of automatic speech recognition systems [1]. Therefore, beamformers (BFs) are utilized to avoid nonlinear distortions such as musical noises and spectral distortions [2], [3]. In the last decade, combined frameworks comprising BFs and deep neural networks (DNNs), referred to as mask-based BFs, have been proposed [4]–[6]. In these frameworks, DNNs generate one or two time-frequency (TF) masks corresponding to the target, interferences, or both to inform the BF of the sound to be enhanced or suppressed. Afterward, the BF estimates a filter for extracting the target using masks. For filter estimation, the following BF methods are adopted: 1) maximum signal-to-noise ratio (max-SNR) or generalized eigenvalue (GEV) BF [4], [5], [7], 2) minimum variance distortionless response (MVDR) BF [5], [6], [8], and 3) multichannel Wiener filter (MWF) BF [9]–[11].

Several mask types have been examined to achieve improved extraction performance. Initially, ideal binary masks (IBMs) were employed to train DNNs for mask generation [4], [5]. Subsequently, ideal ratio masks (IRMs) have been utilized [6], [11], [12].

Overviewing these studies, we concluded that the following aspects are yet to be comprehensively investigated:

1) Which BF can achieve the best performance if an optimal mask is provided for each BF method?
2) Is an optimal mask common for all BF methods?
3) Are conventional ideal masks such as IRMs identical to an optimal mask?

Regarding these aspects, several studies such as [5], [13], and [14] considered that the mask optimal for the single-channel TF masking [12], [15] should commonly be optimal for all BF methods. However, this assumption was not verified in these studies. Moreover, they compared multiple BF methods that employed the same mask in terms of extraction performance. However, these results do not answer the first aspect unless the optimal mask is common.

The remainder of this paper is organized as follows. Sections II and III explain the related work and BF methods presented in this study, respectively. Section IV examines the method for obtaining the optimal mask for each BF method. Section V verifies the aforementioned points experimentally while Section VI discusses the experimental results. Finally, Section VII concludes the study.

II. Related work

First, we overview the history of mask-based BFs. The max-SNR, MVDR, and MWF BFs employ the statistics of the target, interferences, or both. The statistics are called target (or speech) and interference (or noise) covariance matrices. Given that the accuracy of both statistics affects the extraction performance, estimating them accurately is a fundamental issue [16], [17]. In [4], [5], both statistics were computed using two binary masks, each of which represents periods when only the target or interferences are present, and the masks were generated using a properly trained DNN. This idea was first adopted for the max-SNR and MVDR BFs [4]–[6], then applied to the MWF BF [18].

Second, we mention the studies that compare mask-based BFs. At least three methods have been adopted as aforementioned, and several studies have compared two or three of them,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

as presented in Table I. However, the best-performing method depends on experimental setups. Moreover, it is reported in [18] that the difference between the max-SNR and MWF BFs is marginal and the performance depends on the number of microphones used.

Third, we mention the mask types considered to be optimal for the mask-based BFs. Initially, the IBMs were considered to be optimal [4], [5]; then, the IRMs were considered optimal [6], [12], [20]–[22]. These ideas were based on the findings in the single-channel TF masking [12], [15]. However, these studies did not investigate whether these findings really apply to the mask-based BFs or the optimal mask is common for any BF method.

III. Framework of Mask-Based BFs

In this study, we define the best extraction performance as the BF output closest to the target in the TF domain, considering that a significant goal of BFs is to extract (or estimate) the target. Moreover, similar to conventional TF masks, we have the constraint that mask values are nonnegative and real-valued.

Fig. 1 illustrates our idea. The vertical and horizontal axes indicate the closeness of the BF output to the target and the variation of the mask values, respectively. Although the mask values vary multidimensionally, this figure conceptually represents the variation as a single axis. In the mask-based BFs, the extraction performance should depend on the variation and exhibit the peak on a particular mask. We refer to this mask as the optimal one. Considering that multiple BF methods and mask types are employed, we can rephrase the questions mentioned in Section I as follows:

- **Issue 1:** Which BF method has the highest peak, or are all peaks of the same height?
- **Issue 2:** Is the optimal mask common for all BF methods, or dependent on each one?
- **Issue 3:** Can the mask considered to be ideal achieve peak performance? More particularly, is the IRM optimal?

We assumed that the studies presented in Table I considered these issues as in Fig. 2. This indicates that the mask mentioned in each study is optimal for multiple BF methods such as BFs 1 and 2, whereas each BF method demonstrates a different height of the peak. However, this idea is yet to be verified.

Considering that this study is the first to examine these issues, we focus on the following methods for simplicity:

TABLE I
Studies comparing multiple BF methods. Methods in bold type represent those that performed the best in each study. In [18], the performance depends on the number of microphones. (SNR: Signal-to-noise ratio, PESQ: Perceptual evaluation of speech quality, WER: Word error rate)

	Metric	Methods compared
Heymann+16 [5]	SNR	**Max-SNR**, MVDR
Boeddeker+18[13]	SNR,PESQ	**Max-SNR**, MVDR
Wang+18 [14]	WER	Max-SNR, MVDR, **MWF**
Heymann+18 [18]	WER	Max-SNR, MWF
Shimada+18 [19]	WER	MVDR, **MWF**

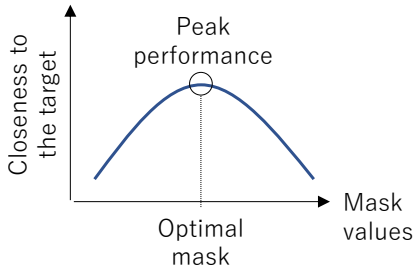


Fig. 1. Conceptual plot of the relationship between the closeness of the BF output to the target and mask values

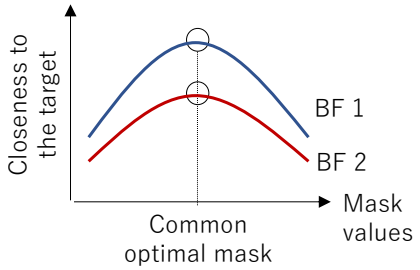


Fig. 2. Conventional ideas on the peak performance and optimal mask for multiple BF methods

- Mask-based MWF BF, which employs a single mask
- Mask-based max-SNR BF, which utilizes two masks, and its variants that employ a single mask

We do not include the MVDR BF in this study, given that the method has another issue regarding the accuracy of the estimation of the steering vector [23].

A. Signal models

In this study, all signals are in the TF domain. For simplicity, the frequency index is omitted, whereas the frame index t is adopted. Let $\boldsymbol{x}(t) = [x_1(t), \ldots, x_N(t)]^{\mathrm{T}}$ be an observation vector obtained with N microphones. The observation $\boldsymbol{x}(t)$ can be expressed as the following summation:

$$\boldsymbol{x}(t) = \boldsymbol{s}(t) + \boldsymbol{n}(t), \tag{1}$$

where $\boldsymbol{s}(t) = [s_1(1), \ldots, s_N(t)]^{\mathrm{T}}$ denotes the components arriving from the target and $\boldsymbol{n}(t) = [n_1(1), \ldots, n_N(t)]^{\mathrm{T}}$ represents the residuals called interferences. Using the observation $\boldsymbol{x}(t)$ and BF filter $\boldsymbol{w}$, the BF output $y(t)$ is expressed as

$$y(t) = \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x}(t). \tag{2}$$

We use the following three covariance matrices:

$$\boldsymbol{\Phi}_x = \left\langle \boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{H}} \right\rangle_t, \tag{3}$$

$$\hat{\boldsymbol{\Phi}}_s = \left\langle m_s(t)\boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{H}} \right\rangle_t, \tag{4}$$

$$\hat{\boldsymbol{\Phi}}_n = \left\langle m_n(t)\boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{H}} \right\rangle_t, \tag{5}$$

where $m_s(t)$ and $m_n(t)$ denote TF masks for the target and interferences, respectively, and $\langle \cdot \rangle_t$ computes the average over t. Each mask comprises nonnegative real values. We

refer to these matrices as observation, target, and interference covariance matrices, respectively. Unlike $\boldsymbol{\Phi}_x$, both $\hat{\boldsymbol{\Phi}}_s$ and $\hat{\boldsymbol{\Phi}}_n$ are estimated matrices computed from the masks and observations without using $\boldsymbol{s}(t)$ and $\boldsymbol{n}(t)$.

Moreover, consider $\mathrm{GEV}_{\max}(\boldsymbol{A}, \boldsymbol{B})$ and $\mathrm{GEV}_{\min}(\boldsymbol{A}, \boldsymbol{B})$ to be the eigenvectors corresponding to the maximum and minimum eigenvalues in the following GEV problem, respectively:

$$\boldsymbol{A}\boldsymbol{w} = \lambda \boldsymbol{B}\boldsymbol{w}. \tag{6}$$

B. Mask-based and ideal MWF BFs

The MWF BF is formulated as a problem of minimizing the mean square error (MSE) between the BF output and the corresponding reference $p(t)$ [14]:

$$\begin{aligned} \boldsymbol{w} &= \arg\min_{\boldsymbol{w}} \left\langle \left| p(t) - \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x}(t) \right|^2 \right\rangle_t & (7) \\ &= \boldsymbol{\Phi}_x^{-1} \left\langle \boldsymbol{x}(t) \overline{p(t)} \right\rangle_t, & (8) \end{aligned}$$

where $\overline{p(t)}$ denotes the conjugate of $p(t)$. The mask-based MWF BF employs the masked observation as the reference, and thus, corresponds to the case $p(t) = m_s(t) x_k(t)$ as

$$\begin{aligned} \boldsymbol{w}_{\mathrm{mwf}} &= \arg\min_{\boldsymbol{w}} \left\langle \left| m_s(t) x_k(t) - \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x}(t) \right|^2 \right\rangle_t & (9) \\ &= \boldsymbol{\Phi}_x^{-1} \left\langle m_s(t) \boldsymbol{x}(t) \overline{x_k(t)} \right\rangle_t, & (10) \end{aligned}$$

where $x_k(t)$ denotes the observation obtained with the kth microphone.

Another significant variant is the ideal MWF BF, which provides the upper-bound extraction performance for all BFs [24]. When $\boldsymbol{s}(t)$ in (1) is known, the upper-bound extraction performance can be achieved using the MWF with $p(t) = \boldsymbol{s}_k(t)$. The corresponding filter is obtained as

$$\begin{aligned} \boldsymbol{w}_{\mathrm{ideal}} &= \arg\min_{\boldsymbol{w}} \left\langle \left| s_k(t) - \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x}(t) \right|^2 \right\rangle_t & (11) \\ &= \boldsymbol{\Phi}_x^{-1} \left\langle \boldsymbol{x}(t) \overline{s_k(t)} \right\rangle_t. & (12) \end{aligned}$$

Note that the ideal MWF BF is not a particular case of the mask-based one. This is because a mask value $m_s(t)$ constrained to be real-valued and nonnegative cannot render (9) equivalent to (11), whereas $m_s(t)$ that can take any complex values can. Therefore, it is not evident whether the mask-based MWF BF can achieve the same extraction performance as the ideal alternative.

C. Max-SNR BF and its variants

The mask-based max-SNR BF is formulated as the following maximization problem [4], [5]:

$$\begin{aligned} \boldsymbol{w}_{\mathrm{snr}} &= \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}} \hat{\boldsymbol{\Phi}}_s \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}} \hat{\boldsymbol{\Phi}}_n \boldsymbol{w}} & (13) \\ &= \mathrm{GEV}_{\max}\left(\hat{\boldsymbol{\Phi}}_s, \hat{\boldsymbol{\Phi}}_n \right). & (14) \end{aligned}$$

Although this method originally utilizes two masks, we can derive two different variants that adopt a single mask by assuming the following relationship:

$$\hat{\boldsymbol{\Phi}}_s + \hat{\boldsymbol{\Phi}}_n = \boldsymbol{\Phi}_x. \tag{15}$$

One variant utilizes a mask for the interferences [17], [25]:

$$\begin{aligned} \boldsymbol{w}_{\mathrm{nor}} &= \arg\min_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}} \hat{\boldsymbol{\Phi}}_n \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}} \boldsymbol{\Phi}_x \boldsymbol{w}} & (16) \\ &= \mathrm{GEV}_{\min}\left(\hat{\boldsymbol{\Phi}}_n, \boldsymbol{\Phi}_x \right). & (17) \end{aligned}$$

We refer to this as the minimum noise-to-observation ratio (min-NOR) BF. Similarly, we can derive the other, which employs a mask for the target and is expressed as

$$\begin{aligned} \boldsymbol{w}_{\mathrm{sor}} &= \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}} \hat{\boldsymbol{\Phi}}_s \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}} \boldsymbol{\Phi}_x \boldsymbol{w}} & (18) \\ &= \mathrm{GEV}_{\max}\left(\hat{\boldsymbol{\Phi}}_s, \boldsymbol{\Phi}_x \right). & (19) \end{aligned}$$

We refer to this as the maximum signal-to-observation ratio (max-SOR) BF.

Furthermore, even without assuming (15), the min-NOR and max-SOR BFs can be equivalent to each other in terms of the filter estimation if one mask is computed from the other using (20).

$$m_s(t) + m_n(t) = \alpha, \tag{20}$$

where α is a nonnegative value, such that all the mask values are nonnegative. For example, assigninig $m_s(t) = \alpha - m_n(t)$ to (18) and (4) results in the problem represented by (16) if $m_n(t)$ is nonnegative for all t.

For the max-SNR BF and its variants, the scales of the filter and BF outputs are uncertain, unlike the MWF BF. Hence, a post-process for determining the proper scales is required [16], [17]. Considering that the formulation of these BFs differs from that of the MWF BF, it is not evident whether these BFs can achieve the same extraction performance as the ideal MWF BF.

IV. Obtaining the optimal mask

To explore the peak performance for each BF method, this study employs a bottom-up approach that obtains the optimal mask for each mixture of the target and interferences, instead of a priori deciding whether a particular mask type is optimal.

Let $\mathcal{M}$ be a set of mask values adopted in a BF. This set comprises $m_s(t)$, $m_n(t)$, or both for all t, depending on the BF method employed. We can formulate $\mathcal{M}$ as the solution to the problem of minimizing the following MSE:

$$\mathcal{M} = \arg\min_{\mathcal{M}} \left\langle \left| s_k(t) - y(t) \right|^2 \right\rangle_t, \tag{21}$$

where $s_k(t)$ and $y(t)$ denote the target included in the observation of the kth microphone and BF output, respectively. To render the mask values nonnegative and to avoid both diverging the mask values and converging them to zero, we impose the following constraints on (21):

$$\begin{aligned} m(t) &\geq 0 \quad \text{for all } t, & (22) \\ \left\langle m(t)^2 \right\rangle_t &= 1, & (23) \end{aligned}$$

where $m(t)$ denotes $m_s(t)$ or $m_n(t)$. In (21), $y(t)$ is computed as follows. First, the BF filter $\boldsymbol{w}(t)$ is computed depending on the adopted BF method, then the $y(t)$ is computed using (2).

To determine the ideal scale of $y(t)$ independent of the BF method, we apply the following post-filtering process, referred to as the *ideal scaling*:

$$\gamma = \frac{\left\langle s_k(t)\overline{y(t)}\right\rangle_t}{\langle |y(t)|^2\rangle_t}, \tag{24}$$

$$y(t) \leftarrow \gamma y(t). \tag{25}$$

Owing to the ideal scaling, the constraint represented as (23) does not affect the scale of $y(t)$.

Because $\mathcal{M}$, $\boldsymbol{w}(t)$, and $y(t)$ depend on each other, $\mathcal{M}$ cannot be obtained explicitly. In contrast, we employ the iterative algorithm based on gradient descent. Given that the mask estimation process adopts no explicit association between the masks and sources, the obtained masks such as $m_s(t)$ and $m_n(t)$ do not necessarily correspond to the sources such as the target and interferences.

Equations (21), (22), and (23) might seem to solve the same problem as the ideal MWF BF, which is represented as (11), based on the iterative algorithm and consequently achieve the same extraction performance as the ideal MWF BF. However, such a perspective is incorrect, as mentioned in Sections III-B and III-C. Rather, the objective of this study is to examine to what extent the output of the mask-based BF can approach that of the ideal MWF BF if the mask best fits (21).

V. Experiments

To clarify Issues 1–3 mentioned in Section III, we conducted the following experiments:

1) Exploring the peak performance for each BF method
2) Verifying whether the optimal mask is common for all the BF methods
3) Examining whether the IRM can achieve the peak performance

The following subsections first mention the dataset and system for these experiments, and then demonstrate the experimental results in order.

A. Dataset and experimental system

We employed the CHiME-3 simulated test set [26], which comprises both 330 utterances from four speakers and four background (BG) noises. In this dataset, sound data were recorded at 16 kHz with six microphones attached to a tablet device. We generated the TF domain signals using the short-time Fourier transform with window and shift lengths of 1024 and 256, respectively.

Fig. 3 illustrates the experimental system. The modules labeled *Absolute value*, *Normalization*, and *Ideal scaling* correspond to (22), (23), and (25), respectively. The observation data were generated by mixing clean speech and BG noise. To represent multiple scenarios in different SNRs, three multipliers such as 1, 2, and 4 were applied to the BG noise. We refer to these values as *BG multipliers*. The BF output $y(t)$ was generated as explained in Section IV; one or two masks were employed and the BF filter was estimated depending on the BF method, such as (10), (14), (17), and (19). Given that

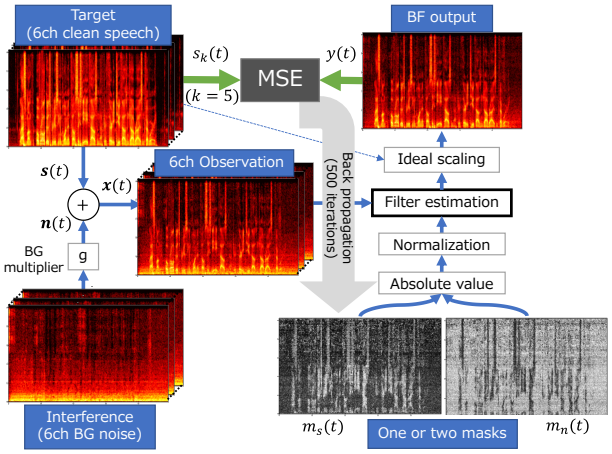


Fig. 3. Configuration of the system obtaining or employing the optimal mask(s) for each utterance

microphone #5 is the closest to the speaker, we set k to 5 in (21) and (24). The backpropagation based on the gradient descent algorithm was utilized only in the first experiment to obtain the optimal mask for each utterance. The estimation of the BF filter and output was implemented in PyTorch [27], which supports the backpropagation of matrix operations in the complex number domain.

Table II presents average SNR scores of the observation of microphone #5 for each BG multiplier.

For the evaluation score, we adopted the source-to-distortion ratio (SDR) [28], which is provided as a performance score in this dataset, considering that this score basically represents the closeness of the BF output to the target.

B. Experiment 1: Exploring the peak performance for each BF method

First, to explore the peak performance of each BF method, we obtained the optimal mask for each utterance, which is the solution to (21) under the constraints represented as (22) and (23), using the backpropagation of 500 iterations. Table III presents SDR scores of the mask-based BFs, including those of the ideal MWF BF using $k = 5$ in (12). All the mask-based BFs practically demonstrated identical scores comparable to the ideal MWF BF. These are remarkable results, given that they were not theoretically evident, as mentioned in Sections III-B and III-C.

C. Experiment 2: Verifying whether the optimal mask is common for all the BF methods

Furthermore, to verify whether the optimal mask is common for all the BF methods, we applied the optimal mask obtained

TABLE II
SNR scores [dB] of the observation of microphone #5

BG multiplier		
×1	×2	×4
7.540	1.536	-4.421

in a BF method to another one. If the same score is achieved after applying the mask, we can claim that the optimal mask is the same between these two BF methods. Although 12 permutations are possible from four methods, we examined six considered to be nontrivial. In the cases that $m_s(t)$ was applied to a BF method using $m_n(t)$, and vice versa, we converted the mask values using the following rules to satisfy (20):

$$m_s(t) = \max_t \{m_n(t)\} - m_n(t), \quad (26)$$

$$m_n(t) = \max_t \{m_s(t)\} - m_s(t), \quad (27)$$

where $\max_t \{\cdot\}$ denotes the maximum value over t.

The obtained results are presented in Table IV. The first and second rows indicate that the optimal masks obtained in the max-SNR BF were applied to the max-SOR and min-NOR BFs, respectively. These two rows present lower scores than Experiment 1. This suggests that the optimal mask obtained in the max-SNR BF is not necessarily optimal for the max-SOR or min-NOR BFs.

The third row indicates that the optimal mask $m_s(t)$ obtained in the max-SOR BF was applied to the min-NOR BF using (27), and the fourth row indicates that the mask $m_n(t)$ obtained in the min-NOR BF was applied to the max-SOR BF using (26). The two rows practically demonstrated the same scores as the max-SOR and min-NOR BFs in Table III. This suggests that the optimal mask can be converted between the max-SOR and min-NOR BFs, despite the fact that the two masks are not the same.

The fifth and sixth rows indicate that the optimal mask $m_s(t)$ obtained in a method was applied to the other for the max-SOR and MWF BFs. These two rows demonstrated considerably lower scores than the results of the max-SOR and MWF BFs in Table III. These results suggest that the optimal mask is not common between these two BF methods.

TABLE III
PEAK PERFORMANCE FOR EACH METHOD AND SCENARIO IN SDR [dB] COMPARED WITH THAT OF THE IDEAL MWF.

BF Method	Mask(s) used	BG multiplier ×1	×2	×4
Max-SNR	m_s, m_n	19.430	**14.276**	**9.451**
Max-SOR	m_s	19.426	**14.276**	**9.451**
Min-NOR	m_n	19.434	**14.276**	**9.451**
MWF	m_s	19.438	14.268	9.430
Ideal MWF		**19.441**	**14.276**	**9.451**

TABLE IV
EXTRACTION PERFORMANCE IN SDR [dB] AFTER APPLYING THE OPTIMAL MASK TO ANOTHER BF METHOD

BF for mask estimation	Applied to	Mask used	BG multiplier ×1	×2	×4
Max-SNR	Max-SOR	m_s	17.084	13.493	9.165
	Min-NOR	m_n	18.399	14.032	9.381
Max-SOR	Min-NOR	m_s & (27)	19.426	**14.276**	**9.451**
Min-NOR	Max-SOR	m_n & (26)	**19.434**	14.274	**9.451**
Max-SOR	MWF	m_s	15.465	10.259	5.337
MWF	Max-SOR	m_s	14.230	12.092	8.458

D. Experiment 3: Examining whether the IRM can achieve the peak performance

The third experiment is for examining if the IRM can achieve the peak performance for each BF method. Considering that the term IRM was ambiguously employed in the related studies [6], [12], [20]–[22], we define several masks that utilize the ratios of the target and interferences based on [12].

The IRMs for the target and interferences are defined as

$$m_s(t) = \left(\frac{|s_k(t)|^2}{|s_k(t)|^2 + |n_k(t)|^2} \right)^{\beta}, \quad (28)$$

$$m_n(t) = \left(\frac{|n_k(t)|^2}{|s_k(t)|^2 + |n_k(t)|^2} \right)^{\beta}, \quad (29)$$

where $k = 5$ and $\beta = 1$ or 0.5. As another type of ratio mask, we employed the spectral magnitude masks (SMMs) defined as

$$m_s(t) = \frac{|s_k(t)|}{|s_k(t) + n_k(t)|}, \quad (30)$$

$$m_n(t) = \frac{|n_k(t)|}{|s_k(t) + n_k(t)|}. \quad (31)$$

In the single-channel TF masking, these masks are ideal in terms of the magnitudes of the target and interferences [12].

The obtained results are presented in Table V. In this table, the combination of the max-SOR and IRM with $\beta = 0.5$ demonstrated the best score for all the scenarios. However, these scores were lower than that in Experiment 1. The results suggest that the IRM does not achieve the peak performance for any BF method examined in this study; hence, it differs from the optimal mask for each method.

In addition, Table V indicates that the max-SNR, max-SOR, and min-NOR BFs present the same scores for the IRM with $\beta = 1$. This is because this mask type evidently satisfies both (15) and (20). Hence, these three BFs are equivalent in this case, as mentioned in Section III-C.

TABLE V
EXTRACTION PERFORMANCE IN SDR [dB] WHEN THE IRM ($\beta = 1, 0.5$) AND SMM WERE EMPLOYED FOR EACH BF METHOD.

BF method	Type of ideal mask	Mask(s) used	BG multiplier ×1	×2	×4
Max-SNR	IRM ($\beta = 1$)	m_s, m_n	18.642	13.889	9.226
	IRM ($\beta = 0.5$)		18.313	13.790	9.203
	SMM		17.973	13.529	9.060
Max-SOR	IRM ($\beta = 1$)	m_s	18.642	13.889	9.226
	IRM ($\beta = 0.5$)		**18.725**	**13.948**	**9.275**
	SMM		13.640	11.447	8.147
Min-NOR	IRM ($\beta = 1$)	m_n	18.642	13.889	9.226
	IRM ($\beta = 0.5$)		18.267	13.747	9.160
	SMM		17.372	12.735	8.103
MWF	IRM ($\beta = 1$)	m_s	17.316	12.788	8.375
	IRM ($\beta = 0.5$)		16.228	11.799	7.477
	SMM		17.109	12.316	7.734

VI. Discussion

In this section, we discuss the experimental results for each issue mentioned in Section III.

A. Discussion on Issue 1

The results of Experiment 1 provide the answer to Issue 1. The peak performance of each BF method is practically identical and comparable to the upper bound given by the ideal MWF BF. These findings differ from the conventional idea presented in Fig. 2 and may deprive the meaning of the discussion on the BF that can achieve the best extraction performance. However, it is an open question whether these findings are applicable to any BF method and dataset. Therefore, further investigation is required.

B. Discussion on Issue 2

The results of Experiment 2 provide the answer to Issue 2. Although the optimal mask is not common, it depends on the BF method adopted. This finding differs from the conventional concept presented in Fig. 2, similar to the discussion on Issue 1. However, we consider this to be natural because the optimal mask is formulated as the solution to a different problem for each BF method, as explained in Section IV.

In addition, Table IV indicates several remarkable points. Although both the max-SOR and min-NOR BFs are derived from the max-SNR BF, as mentioned in Section III-C, the top two rows suggest that $m_s(t)$ and $m_n(t)$ are optimal for the max-SNR BF, while not for the max-SOR or min-NOR BFs. We consider that the reason for this is that the optimal masks obtained in the max-SNR BF do not satisfy (15) or (20). Hence, the max-SNR BF is not equivalent to the other two BFs in this case. In contrast, the third and fourth rows in this table suggest that the max-SOR and min-NOR BFs are equivalent if both $m_s(t)$ and $m_n(t)$ satisfy (20). Moreover, the bottom two rows exhibited lower scores than the other rows. Although $m_s(t)$ has commonly been interpreted as the mask for the target, the facts suggest that the optimal mask $m_s(t)$ for the max-SOR BF differs significantly from that of the MWF BF and vice versa.

From the discussion on Issues 1 and 2, we have obtained a novel idea for the peak performance and optimal mask among multiple BF methods, as illustrated in Fig. 4. This indicates that the optimal mask depends on the BF method adopted, while the peak performance is the same among multiple BF methods and comparable to the upper bound achieved by the ideal MWF.

C. Discussion on Issue 3

The results of Experiment 3 provide the answer to Issue 3. Although these masks have been considered to be optimal for any BF method, Table V suggests that the IRM and SMM do not achieve the upper-bound performance and differ from the optimal mask for any BF method. These findings can explain why the studies presented in Table I mentioned a different BF method as the best. Table V indicates that the BF method that appears the best depends on the mask type employed. For example, focusing on the IRM with $\beta = 0.5$ in this table, we can determine that the max-SOR BF performs the best. However, this result only suggests that this mask type is the closest to the optimal mask for the max-SOR in this dataset.

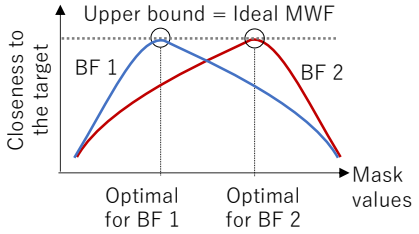


Fig. 4. Obtained concept of the peak performance and optimal mask among multiple BF methods

Furthermore, the fact that the IRM is not optimal imposes another issue on us. This is how the optimal mask is interpreted and represented as a formula. However, this is also an open question.

VII. Conclusions

In this study, we investigated mask-based BFs such as the max-SNR, max-SOR, min-NOR, and MWF BFs. To explore the peak performance for each BF method, we obtained the optimal mask for each utterance by minimizing the MSE between the BF output and target. We experimentally verified that these four methods have the same peak performance as the upper bound provided by the ideal MWF BF. Via additional experiments that applied the optimal mask across BF methods, we determined that the optimal mask differed for the BF method used. However, the mask values can be converted between the max-SOR and min-NOR BFs. These findings differed from the conventional idea that the optimal mask would be common and the peak performance would depend on the BF method. We verified that the IRM did not achieve the peak performance for these four BFs. Hence, this mask type was not optimal. We expect that these findings will contribute to the improvement of mask-based BFs.

Given that these findings are currently experimental, in the future, we would attempt to establish their theoretical background and investigate whether these apply to other BF methods and datasets.

The experimental system has been shared in https://github.com/hreshare/optimal_beamformers/.

References

[1] S. J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, pp. 1571–1575, 2018.

[2] M. Mizumachi and M. Origuchi, "Advanced delay-and-sum beamformer with deep neural network," *22nd International Congress on Acoustics (ICA)*, 2016.

[3] M. Mizumachi, "Neural network-based broadband beamformer with less distortion," *Proceedings of International Congress on Acoustics (ICA 2019)*, p. 2760, Sep. 2019.

[4] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3RD CHiME challenge," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, no. June 2016, pp. 444–451, 2016.

[5] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[6] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 1981–1985, 2016.

[7] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, International Speech Communication Association, 2019, pp. 1253–1257.

[8] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

[9] S. Stenzel, T. C. Lawin-Ore, J. Freudenberger, and S. Doclo, "A multichannel wiener filter with partial equalization for distributed microphones," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, unknown, Oct. 2013, pp. 1–4.

[10] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[11] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 66–70.

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," en, *IEEE/ACM Trans Audio Speech Lang Process*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[13] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural Mask-Based beamforming for Far-Field speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6697–6701.

[14] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments," *Comput. Speech Lang.*, vol. 49, pp. 37–51, May 2018.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," en, *IEEE/ACM Trans Audio Speech Lang Process*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[16] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 1, 2007, pp. I–41–I–44.

[17] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

[18] J. Heymann, M. Bacchiani, and T. N. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6722–6726.

[19] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5734–5738.

[20] Y.-H. Tu, J. Du, L. Sun, F. Ma, and C.-H. Lee, "On design of robust deep models for CHiME-4 multi-channel speech recognition with multiple configurations of array microphones," in *Interspeech 2017*, ISCA: ISCA, Aug. 2017.

[21] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on Teacher–Student deep learning using improved speech presence probability for Noise-Robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, Dec. 2019.

[22] S. Chakrabarty and E. A. P. Habets, "Time–Frequency masking based online Multi-Channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 4, pp. 787–799, Aug. 2019.

[23] Y. H. Tu, J. Du, L. Sun, and C. H. Lee, "LSTM-based iterative mask estimation and post-processing for multi-channel speech enhancement," *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, vol. 2018-Febru, no. December, pp. 488–491, 2018.

[24] J. Malek, Z. Koldovský, and M. Bohac, "Block-online multi-channel speech enhancement using deep neural

network-supported relative transfer function estimates," *IET Signal Proc.*, vol. 14, no. 3, pp. 124–133, May 2020.

[25] A. Hiroe, "Similarity-and-Independence-Aware beamformer with iterative casting and boost start for target source extraction using reference," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 1–20, 2022.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, 2017.

[27] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An imperative style, High-Performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Language modeling for spontaneous speech recognition based on disfluency labeling and generation of disfluent text

Koharu Horii*[1], Kengo Ohta†, Ryota Nishimura‡, Atsunori Ogawa§, Norihide Kitaoka*
* Toyohashi University of Technology, † National Institute of Technology, Anan College, ‡ Tokushima University,
§ NTT Corporation, Japan

***Abstract*—Disfluencies in spontaneous speech such as fillers and hesitations are major causes of automatic speech recognition (ASR) errors. In our previous work, we proposed a "disfluency labeling" method allows an end-to-end (E2E) ASR model to recognize such disfluencies as symbolized recognition targets. This study proposes a method of further improving spontaneous speech recognition by integrating a language model (LM) that can predict disfluencies as pre-defined symbols. First, we fine-tuned a pre-trained Bidirectional Encoder Representations from Transformers (BERT) to predict word boundaries where disfluency symbols ("#" for a filler and "@" for a hesitation) should be inserted within a corpus of normal written text. We then used it to generated a large corpus of disfluent text. Finally, we trained an LM to predict the disfluency symbols with the BERT-generated corpus, and integrated the LM into an E2E ASR model using Shallow Fusion. Our experimental results show that the LM trained using the generated disfluent text corpus improved disfluency prediction and, as a result, spontaneous speech recognition performance.**

I. Introduction

Automatic speech recognition (ASR) is a technology that has become well-known to the public through applications such as voice search services on smartphones and smart speakers, and through instant speech transcription. Thanks to significant progress in neural network (NN) modeling, the performance of end-to-end (E2E) ASR models, which are fully neural models that directly map input speech to the corresponding token sequence, has greatly improved. However, the recognition performance of these models when processing spontaneous speech remains unsatisfactory level. This is because spontaneous speech usually contains fillers such as "uh" and "um", and hesitations such as reparandum, repetition, or stuttering, and these disfluent acoustic phenomena are major causes of ASR errors [1]. In addition, since these disfluencies are generally unnecessary for downstream natural language processing (NLP) tasks, they should be removed from ASR results [2]. Several approaches have been proposed to alleviate this problem [3]–[7]. For example, in [6], the authors identified disfluencies based on fundamental frequency, and removed them before recognition.

Studies have also been conducted using E2E ASR approaches to cope with the disfluency problem. Lou and Johnson [8] developed a speech recognizer that generates fluent transcriptions directly from disfluent speech by training the ASR model with speech transcriptions which have had the disfluencies removed. Mendelev *et al.* [9] and we [10] replaced hesitations with hesitation labels, then trained an ASR with text containing hesitation labels and other regular tokens, in order to recognize hesitations. To target both fillers and hesitations, our previous study [11] also proposed a method called *disfluency labeling*, which replaces fillers and hesitations within transcriptions of spontaneous speech with symbols ("#" for fillers and "@" for hesitations). They then trained an ASR model using a corpus containing these disfluency symbols. Although spontaneous speech can include a theoretically unlimited variety of disfluencies, during decoding the ASR model treats all of these disfluencies as one of the two types of disfluencies represented by symbols, and thus the ASR robustly recognizes (detects) them, as it does with other tokens. Finally, by removing the two disfluency symbols from the ASR results we can obtain a transcription without unneeded information, i.e., one suitable for downstream tasks.

In [11], investigators performed inference using an ASR model without using a language model (LM). However, many studies have reported that using a method called *Shallow Fusion* to integrate an external LM into the main ASR model during inference is effective for improving recognition performance [12]–[14]. Shallow Fusion performs log-linear interpolation of the scores obtained by an ASR model and by a separately trained LM at each step of a beam search. By using an external LM and Shallow Fusion, it is possible to consider a larger amount of contextual information during inference.

In this study, we aim to improve the performance of the ASR model through disfluency labeling, and by using an external LM and Shallow Fusion. However, normal written text corpora do not contain fillers and hesitations, and thus LMs trained with these corpora are unsuitable for use with ASR models that employ disfluency labeling. Furthermore, manually building a large corpus of disfluent text would involve considerable costs in time and resources. To overcome this problem, we generated disfluent text by inserting disfluency symbols into written text using Bidirectional Encoder Representations from Transformers (BERT) [15]. The generated disfluent text is then

[1]Current affiliation is NTT Corporation.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

used to train an LM suitable for ASR of spontaneous speech. The BERT model can perform fill-in-the-blank tasks, in which, given a sentence with blanks, the model will complete the sentence by filling the blanks with suitable words. We treat written texts as disfluent texts, in which disfluency symbols are omitted. Then, using the BERT model with the fill-in-the-blank framework, we insert disfluencies to the written texts to generate disfluent texts. We then integrating it with the ASR through Shallow Fusion, we could improve the accuracy of the ASR system when it encounters spontaneous speech. Even though our method is simple, we experimentally show that it is very effective.

II. Related work

A. Insertion of disfluent acoustic events into text

Researchers have tried inserting disfluent events into text to increase the naturalness of text-to-speech (TTS) synthesis. In [16], a subject questionnaire was used to evaluate manual filler insertion during TTS. Then fillers were automatically inserted using various models. Recurrent neural network (RNN) and 4-gram LMs accurately inserted the fillers, and many of the subjects indicated that the inserted fillers were natural. In [17], high filler insertion accuracy was achieved by combining an LM and a decision tree using binary classification to determine whether or not to insert fillers between word boundaries. In [18], a Planner-Generator based disfluency generation model was created to augment the data used to train the disfluency detector, improving detection accuracy.

Other studies on phrase boundary prediction in text have inserted acoustic events (such as phrase boundaries) into the text instead of disfluencies, to improve speech articulation during TTS and readability of the ASR results. Models that predict time series data, such as RNNs [19], [20] and bidirectional long short-term memory based RNNs (BLSTMs) [21], [22], have been successful in this task. In [23], a self-attention-based model was used for punctuation prediction. Experiments on predicting the position and type of punctuation to be inserted were conducted using English IWSLT2011 datasets. Their results showed that the self-attention based model outperformed a state-of-the-art BLSTM-CRF (conditional random field) model. In this study, we employed BERT [15], a Transformer encoder model that can consider bidirectional context when inserting fillers and hesitations into written text, to generate the disfluent text needed to train an LM for spontaneous speech recognition.

B. Disfluency labeling

Our previous study [11] proposed a disfluency labeling method to improve the accuracy of ASR for spontaneous speech. Since the sentence that the speaker intended to say does not contain disfluencies, they need to be removed from the recognition results in order to interpret the target sentence. The box below shows an example of their disfluency labeling scheme:

Before	:	**Well...** the presentation is **all, uh,** almost over.
After	:	**#** the presentation is **@ #** almost over.
Target	:	The presentation is almost over.

This method replaces fillers and hesitations with the disfluency symbols "#" and "@", respectively. This approach makes the disfluency detection task easier for the ASR model as multiple types of fillers and hesitations can be represented with a single symbol, respectively, since the exact contents of these disfluencies do not need to be considered. Thus, a single symbol can be used to represent an entire class of disfluent phenomena. Fillers and hesitations have different features, so by using two different symbols they can be treated separately.

C. Shallow fusion

In this study, we used the Shallow Fusion method to integrate the external LM, trained with the generated disfluent text, with the ASR model. Shallow Fusion integrates the external LM only at the time of inference, performing log-linear interpolation using the LM at each step of the beam search [14]. Given acoustic feature vector sequence x, an E2E ASR model using an external LM through Shallow Fusion can predict the most probable output token sequence y^* as follows:

$$y^* = \underset{y}{\operatorname{argmax}}\{\log p(y|x) + \lambda \log p_{\mathrm{LM}}(y)\}, \tag{1}$$

where y is a possible output token sequence, $\log p(y|x)$ is the score of y given x obtained by the E2E ASR model, $\log p_{\mathrm{LM}}(y)$ is the score of y obtained by the external LM, and λ is an LM weight.

This method is very cost-effective because integration is performed only during inference. Applying the LM using Shallow Fusion makes it possible to consider a wider range of contextual expressions.

D. BERT

BERT is a Transformer encoder model that solves NLP tasks by considering context bidirectionally. Many open-source BERT models, pre-trained with large amounts of text data, are available. By fine-tuning such a pre-trained model using a small amount of data for the target task (such as sentence classification [24], question answering [15], or sentence tagging [25]), task-adapted models can easily be obtained, and their effectiveness has been experimentally confirmed.

In this study, we fine-tune our pre-trained BERT with a masked language modeling (MLM) task, using transcripts of labeled spontaneous speech. MLM is a self-supervised learning task in which the BERT automatically masks the sentences and then predicts the masked tokens. Tokens in the input sentences are randomly masked, then their content is predicted, allowing deep, bidirectional representation learning [15]. We hypothesized that taking additional context into consideration would allow us to more accurately insert the disfluency symbols.

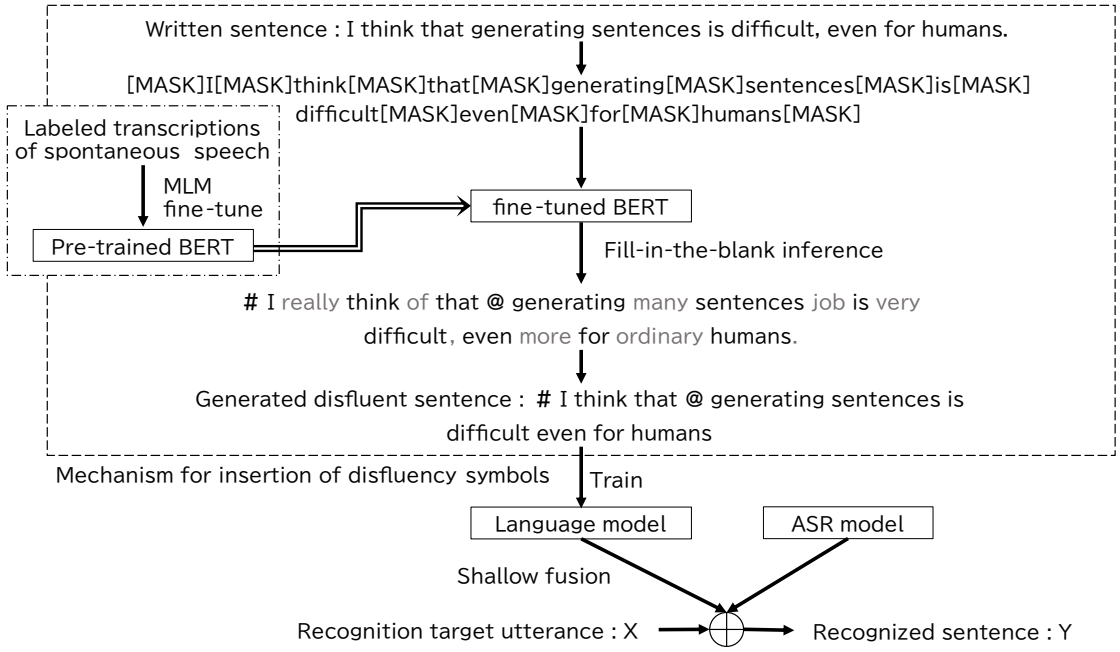


Fig. 1. Overview of disfluent sentence generation.

III. Disfluent sentence generation

To achieve further improvement in spontaneous speech ASR, we can train the LM using a large text dataset and then apply it during inference, however language models trained with written text are unsuitable for this task. Therefore, in this study we generated disfluent text by inserting disfluency symbols into written text using BERT [15]. The pre-trained BERT model was fine-tuned using the same spontaneous speech transcriptions used to train the ASR model, which were labeled in the manner described in Section II-B. Disfluency symbols were then inserted into the input sentences, and the generated text was then used to perform language modeling suitable for spontaneous speech. An overview of this approach is shown in Figure 1.

Fine-tuning was performed through the MLM task described above in II-D , during which 15% of the tokens in each sentence were randomly selected. Then 80% of the selected tokens were then replaced with "[MASK]" tokens, 10% were replaced with other random tokens, and the remaining 10% of the tokens were left unchanged [15]. This fine-tuning allows BERT to insert symbols into sentences based on the context on either side of a word. The BERT model can then make the inferences necessary to convert written text into disfluent text.

Suppose there is a written sentence such as, "I think that generating sentences is difficult, even for humans." First, "[MASK]" is inserted between each word and at the beginning and end of the sentence, as a preprocessing step to allow the insertion of symbols. Next, the masked sentence is entered into the fine-tuned BERT model, and the model performs inference for each masked word in the sentence, as to whether it is a filler (#), a hesitation (@), or some other word. An example is shown in Fig. 1, where inference is performed for all of the 11 masked words in the sentence. If a word inserted during fill-in-the-blank inference is not a symbol, it is discarded. In the example in Fig. 1, the words "really", "of", "many", etc. are all discarded. This discarding allows only disfluency symbols to be inserted into the sentences. Predicting all the masked elements using words, and then discarding them, prevents disfluency symbols from being inserted in positions where they are inappropriate (i.e., in a location where a word other than a disfluency symbol is predicted). Thus, disfluency symbols can only be inserted in positions where they are likely to be correct. This sequence of events generates disfluent text from written text. By applying this method to create a disfluent speech training corpus, the LM can be adapted to recognize disfluent spontaneous speech

The higher the BERT score (log-likelihood), the more likely the "[MASK]" element is a valid word. When generating disfluent text, a disfluency symbol is inserted even when the probability of a disfluency is low, but still higher than other words, regardless of its likelihood score. However, since the accuracy of symbol insertion is uncertain, a threshold p can be set. Symbols and words with likelihoods below p can be discarded to reduce the number of faulty disfluency symbol insertions, allowing more accurate generation of disfluent sentences.

IV. Evaluation experiments

A. Corpus

In order to evaluate the proposed method, we conducted experiments using the Corpus of Spontaneous Japanese (CSJ) [26], a dataset of spontaneous Japanese speech, as well as the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [27], a dataset of written Japanese text. The CSJ contains 700 hours of spontaneous speech and approximately 7 million morphemes. It consists of monologues, such as academic

TABLE I
TRAINING DATA FOR EACH LANGUAGE MODEL.

LM ID	Training data	Symbols inserted by	Sentences	Characters	Fillers (#)	Hesitations (@)
1	original BCCWJ	-	4,220,470	147,654,449	0	0
2	CSJ disfluency removed	-	398,578	11,021,130	0	0
3	Training data for 1 + 2	-	4,619,048	158,675,579	0	0
4	CSJ with disfluency labeled [11]	Actual disfluencies	413,408	11,553,641	441,729	90,782
5	BCCWJ with # inserted at BOS	Automatically	4,220,470	151,874,919	4,220,470	0
6	BCCWJ with disfluency symbols randomly inserted	Randomly	4,220,470	149,519,873	1,534,087	331,337
7	BCCWJ with disfluency symbols inserted by BERT	BERT-based (proposed)	4,220,470	211,025,390	28,076,932	1,530,249
8	BCCWJ with disfluency symbols inserted by BERT ($p = -10$)	BERT-based w/p limit (proposed)	4,220,470	185,143,059	3,443,116	281,734
9	Training data for 4 + 8	Actual and BERT-based w/p limit (proposed)	4,633,878	196,696,700	3,884,845	372,516

presentation speech (APS) and simulated public speech (SPS), as well as dialogues [26]. We used 413,377 CSJ utterances to train the ASR model and fine-tune the BERT. 4,000 utterances were used as development data to validate the LM and ASR model during training. CSJ dialogue speech, which included 23,263 utterances, was used as the recognition target. The dialogs consisted of interviews with the presenters of the APS and SPS monologues, task-oriented dialogues, and free dialogues. The BCCWJ corpus, which is currently the most extensive corpus of written Japanese, was used to train the LM. It contains 100 million words and 4,557,815 sentences, covering a wide range of documents, including books in general, magazines, newspapers, governmental white papers, internet bulletin boards, a blog, etc. [27].

B. Experimental settings

We used 'cl-tohoku/bert-base[2]', which was pre-trained with Japanese Wikipedia, as our pre-trained BERT model. Fine-tuning was then performed with the MLM task, using a transcription of the CSJ training data. The loss was converged in 3 epochs. Word-wise segmentation of the sentences was performed using MeCab with IPAdic 1.0.0[3] as the dictionary, and "[MASK]" tokens were inserted between the words. MeCab is a morphological analyzer for Japanese (a language in which written sentences are not segmented into separate words, i.e., there are no spaces between words). IPAdic is a dictionary used by Japanese morphological analyzers. To conserve memory, sentences longer than 512 words were excluded during mask insertion, since the maximum input length for the BERT is 512 words. We used the *BertForMaskedLM* class from Huggingface transformers[4] for inference. We conducted our ASR experiments using ESPnet2 (version: v.0.9.9). Processing in the BERT was word-based, while that in ESPnet was character-based. We used the joint CTC-attention Transformer ASR model trained with disfluency-labeled transcriptions, as in [11].

Nine identical LSTM language models were trained differently, using the data shown in Table I. LMs 1, 2, and 3 are the baseline LMs. LM1 was trained with original BCCWJ text, so this LM cannot predict sentences containing disfluency symbols, since there are no disfluencies in the BCCWJ text. LM2 was trained by using "CSJ disfluencies removed", which is close to a written style data set. LM2 does not predict disfluent sentences too. LM3 was trained by using a data set that combines those for training LM1 and LM2. LM4 was trained with transcriptions of the CSJ utterances with disfluency labeling. Since the CSJ contains transcriptions of spontaneous speech, including all of the fillers and hesitations actually uttered by the speakers, the training data used for LM4 always had the disfluency symbols inserted in the correct positions. LMs 5 and 6 were trained with BCCWJ text containing disfluency symbols that were automatically inserted without using BERT, for comparison with the proposed model. LM5 was obtained by inserting the filler symbol "#" at the beginning of sentences (BOS) as a pseudo rule-based method, based on the tendency of fillers to appear before utterances [28], [29]. LM6 was obtained by randomly inserting disfluency symbols into the masked BCCWJ text at the same rate as CSJ Eval1 (1.67% for "#," and 0.36% for "@").

LMs 7 and 8 are LMs trained on texts generated by our BERT-based disfluent sentence generation method, however LM8 was trained with text that had disfluency tags inserted using a p value of $p = -10$. The BERT score usually ranges from -20 to 0, and therefore, -10 is about the median value. LM9 was trained by combining the training data used to train LM8 with the 'ground truth' CSJ data used to train LM4, providing data augmentation for the accurately labeled CSJ data, and domain adaptation for the BCCWJ data with BERT-applied disfluency labels. Training of the language models was repeated for 40 epochs, and the model with the lowest validation loss for each training data category was selected. Test set perplexity (PPL) of each LM was also calculated.

The ESPnet2 default hyperparameter settings were used for ASR training and recognition. The CSJ character list was used as the token list (3,263 tokens), with the addition of disfluency symbols "#" and "@". Integration of the ASR model and the LMs during inference was performed using Shallow Fusion, with Shallow Fusion weights for all of the LMs set to 0.3 since this value brought the lowest error rate when using an LM in our preliminary experiments. We also compared the results of inference for each model with the results when the LM was not

[2]https://github.com/cl-tohoku/bert-japanese.
[3]https://taku910.github.io/mecab.
[4]https://github.com/huggingface/transformers.

TABLE II
ASR EVALUATION RESULTS.

ID	LM training corpus	Insertion method	PPL	CER	SER
0	-	-	-	10.3	32.8
1	BCCWJ	-	58.7	11.3	38.0
2	CSJ	-	49.6	11.4	38.9
3	BCCWJ & CSJ	-	46.2	11.3	38.7
4	CSJ	Actual locations	**15.1**	9.9	31.2
5	BCCWJ	# at BOS	63.3	10.0	31.8
6	BCCWJ	Random	27.8	10.1	32.6
7	BCCWJ	BERT	36.7	10.0	31.9
8	BCCWJ	BERT ($p = -10$)	25.6	9.9	31.3
9	BCCWJ & CSJ	BERT ($p = -10$)	16.0	**9.6**	**30.5**

used. The reference sentences were obtained by labeling them in the same manner as the training data. We used character error rate (CER) and sentence error rate (SER) to evaluate ASR performance. These calculations were performed after the symbols were removed from both the recognition results and the reference sentences. To measure the symbol detection rate during recognition, recall, precision, and F-measure were used.

C. Experimental results

Table II shows the results of our ASR experiments when using different LMs. Note that LM4, trained with only labeled CSJ data, had the best perplexity. This must be because the recognition target included interviews related to the contents of the APS and SPS monologues in the training data, therefore the distribution of the vocabulary is close to the target.

Based on the results of baseline LMs 1-3, using LMs trained on written sentences that do not contain disfluent symbols does not even overcome the accuracy of not using LMs. This suggests that the nature of disfluent spoken sentences is very different from that of written sentences, and that it is effective to use LMs trained on sentences in which disfluent symbols are inserted, rather than simply fusing LMs. Comparing the LMs trained with the labeled BCCWJ text, LM7, which used variations of our proposed method, performed better than both LM5 with labeling at BOS and LM6 with random labeling, suggesting the effectiveness of the proposed method. The accuracy of LM8, using parameter p to limit disfluency symbol insertions, was better than that of LM5 without parameter p, indicating that using p as a limit when performing symbol insertion, for greater certainty, is effective for improving ASR accuracy. In addition, although LMs 7 and 8 improved ASR accuracy in comparison to not using an LM (ID 0), and to low accuracy labeling (LMs 5 and 6), their performance was slightly lower than that of LM4, which was trained with the accurately labeled CSJ data. This may be because the symbols in LM4 represent actual disfluencies, whereas those inserted by LMs 7 and 8 were BERT-based, simulated symbols.

TABLE III
DISFLUENCY SYMBOL DETECTION RATE [%] DURING ASR.

LM ID	Symbol	Recall	Precision	F-measure
0	#	70.2	81.7	75.5
	@	**59.8**	69.7	**64.3**
9	#	**72.5**	**82.9**	**77.4**
	@	55.0	**71.5**	62.2

These results indicate that a certain degree of accuracy in labeling is required when training LMs to handle disfluencies. However, our proposed method is still effective, especially when manually labeled or accurately transcribed spontaneous speech are unavailable.

The best performing language model was LM9, which was trained with both the CSJ data used to train LM4 and the training data used to train LM8, which had the highest accuracy of the LMs trained with only BCCWJ text. ASR accuracy was improved when using LM9, without degrading the PPL of LM4, indicating that the proposed method's augmentation of the LM's knowledge was effective. Additionally, result of the LM3 baseline shows that simply combining BCCWJ and CSJ does not yield higher accuracy than the ASR without LMs. So, it is effective to generate disfluent text and combine this data for the language model, not simply train the LM with additional normal written text data.

Compared to the performance of the ASR without LM integration (ID0), the relative CER and SER were reduced by 6.8% and 7.0%, respectively, when using LM9, a steady improvement in accuracy. Considering the difficulty of performing ASR on spoken dialogue, and the good baseline accuracy achieved without LM integration, the proposed method provides satisfactory results.

Table III shows the disfluency symbol detection rate during ASR without LM integration (ID0) and with LM9 integration. While the detection rate for filler symbol (#) increased when the LM was integrated, the detection rate for hesitation symbol (@) decreased. This may be because low insertion accuracy for the hesitation disfluency symbol negatively affected hesitation detection during ASR.

V. CONCLUSION

In this study, a language model for spontaneous speech was created by training it with disfluent text data, generated by inserting disfluency symbols into a large amount of text. The LM was then integrated with an E2E ASR model using Shallow Fusion. Our evaluation experiment showed that ASR accuracy when processing disfluent speech was significantly improved when using the LMs trained with the proposed method. In addition, the importance of disfluency labeling was shown in the area of LM training.

In the future, we would like to develop a better method of generating training data, and to devise a new method of integrating the LM with the ASR. We would also like to try changing the LM network from an LSTM to a Transformer-based method, to further improve the accuracy of ASR for spontaneous speech.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP21K13641, JP22H04863, and JP23H00493.

REFERENCES

[1] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates," in *Proc. ACL-08: HLT*, 2008, pp. 380–388.

[2] M. Johnson and E. Charniak, "A TAG-based noisy-channel model of speech repairs," in *Proc. ACL-04*, 2004, pp. 33–39.

[3] R. Rose and G. Riccardi, "Modeling disfluency and background events in ASR for a natural language understanding task," in *Proc. ICASSP*, 1999, pp. 341–344.

[4] J. Wilpon, L. Rabiner, C.-H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[5] Y. Tsvetkov, Z. Sheikh, and F. Metze, "Identification and modeling of word fragments in spontaneous speech," in *Proc. ICASSP*, 2013, pp. 7624–7628.

[6] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Proc. ICASSP*, 1992, 521–524 vol.1.

[7] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[8] P. J. Lou and M. Johnson, "End-to-end speech recognition and disfluency removal," in *Proc. EMNLP*, Jan. 2020, pp. 2051–2061.

[9] V. Mendelev, T. Raissi, G. Camporese, and M. Giollo, "Improved robustness to disfluencies in RNN-Transducer based speech recognition," in *Proc. ICASSP*, 2021, pp. 6878–6882.

[10] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-end spontaneous speech recognition using hesitation labeling," in *Proc. APSIPA ASC*, 2021, pp. 1077–1081.

[11] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-end spontaneous speech recognition using disfluency labeling," in *Proc. Interspeech*, 2022, pp. 4108–4112.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[13] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," in *Proc. Interspeech*, 2017, pp. 523–527.

[14] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *Proc. ICASSP*, 2018, pp. 5824–5828.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[16] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King, "Investigating automatic & human filled pause insertion for speech synthesis," in *Proc. Interspeech*, 2014, pp. 51–55.

[17] J. Adell, A. Bonafonte, and D. Escudero, "Filled pauses in speech synthesis: Towards conversational speech," in *Proc. TSD*, 2007, pp. 358–365.

[18] J. Yang, D. Yang, and Z. Ma, "Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection," in *Proc. EMNLP*, 2020, pp. 1450–1460.

[19] A. Vadapalli and S. V. Gangashetty, "An investigation of recurrent neural network architectures using word embeddings for phrase break prediction," in *Proc. Interspeech*, 2016, pp. 2308–2312.

[20] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *Proc. Interspeech*, 2016, pp. 3047–3051.

[21] V. Klimkov, A. Nadolski, A. Moinet, *et al.*, "Phrase break prediction for long-form reading TTS: Exploiting text structure information," in *Proc. Interspeech*, 2017, pp. 1064–1068.

[22] Y. Zheng, J. Tao, Z. Wen, and Y. Li, "BLSTM-CRF based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end," in *Proc. Interspeech*, 2018, pp. 47–51.

[23] J. Yi and J. Tao, "Self-attention based model for punctuation prediction using word and speech embeddings," in *Proc. ICASSP*, 2019, pp. 7270–7274.

[24] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using BERT," in *Proc. EDBT/ICDT Workshops*, 2021.

[25] W. Liu, X. Fu, Y. Zhang, and W. Xiao, "Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter," in *Proc. ACL and IJCNLP*, 2021, pp. 5847–5858.

[26] K. Maekawa, "Corpus of Spontaneous Japanese : Its design and evaluation," *Proc. SSPR*, pp. 7–12, 2003.

[27] K. Maekawa, M. Yamazaki, T. Ogiso, *et al.*, "Balanced Corpus of Contemporary Written Japanese," *Language Resources and Evaluation*, vol. 48, pp. 345–371, 2014.

[28] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, no. 3, pp. 201–242, 1998.

[29] M. Watanabe, Y. Den, K. Hirose, and N. Minematsu, "Clause types and filed pauses in Japanese spontaneous monologues," in *Proc. Interspeech*, 2004, pp. 905–908.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Transformer-based Automatic Speech Recognition of Simultaneous Interpretation with Auxiliary Input of Source Language Text

Shuta Taniguchi* and Tsuneo Kato* and Akihiro Tamura* and Keiji Yasuda†
* Doshisha University, Japan
E-mail: tsukato@mail.doshisha.ac.jp
† MINDWORD, Inc., Japan
E-mail: ke-yasuda@mindword.jp

***Abstract*—Automatic speech recognition (ASR) of simultaneous interpretation is challenging due to disfluencies such as hesitations, filled pauses, interruptions, and self-repairs. Meanwhile, the simultaneous interpretation sometimes accompanies source language text. In particular, training materials for simultaneous interpreters have the source language text. The source language text can also be considered an ideal transcription of source language speech. Previously, we proposed a Transformer-based ASR with auxiliary input of source language text for transcribing simultaneous interpretation and verified the effectiveness of the model structure by using large-scale corpora of speech translation (ST). In this study, we fine-tuned and evaluated the proposed model using a corpus of simultaneous interpretation with transcription of the source language speech as the auxiliary input. We also added a filled pause model to the ASR and compared the effect on the reduction of a word error rate (WER) with the auxiliary input. The experimental results showed significant WER reductions by both the auxiliary input and the filled pause model, though the reduction made by the filled pause model was greater than that made by the auxiliary input.**

I. Introduction

Trainees of simultaneous interpreters take various interpretation tests in the course of the training program. Trainers score the quality of the trainees' interpretation with its transcribed text so that their subjective impression of the speech does not affect the score. Introduction of today's fast-evolving ASR is anticipated to transcribe the interpreters' speech automatically. However, the ASR of simultaneous interpretation is a challenging task due to disfluencies such as hesitations, filled pauses, interruptions, and self-repairs. Meanwhile, the ASR for the training programs of simultaneous interpreters has an advantage of having access to source language text because the test materials are prepared in advance. We aim to improve the accuracy of the ASR by using the source language text as an auxiliary input. We can consider the source language text a perfect transcription of the source language speech. If an accurate ASR of the source language speech is available, the proposed method can also be applied to general situations of simultaneous interpretation.

We have proposed a Transformer-based ASR model with auxiliary input of source language text toward transcribing simultaneous interpretation [1]. Since there is no corpus of simultaneous interpretation with source language text large enough to train the model, we simulated simultaneous interpretation data by changing the inputs and outputs of a large-scale corpus for developing end-to-end speech translation (ST), namely MuST-C corpus [2] for the initial evaluation of the model. The proposed model was trained and evaluated by using MuST-C corpus, with optimization of the model structure, specifically arrangement of Transformer blocks that have a source-target attention from the decoder and the audio encoder to the source language text encoder. The simulation showed that the proposed model significantly reduced WERs for four ST datasets.

In this study, we fine-tune and evaluate the proposed model using a simultaneous interpretation corpus. Since the speech of simultaneous interpretation contains many filled pauses, we added a filled pause model to the ASR [3]. We compared the effects on reduction of the WERs between the auxiliary input of the source language text and the filled pause model.

II. Related Studies

The proposed model is considered a new type of multimodal ASR. Having access to multiple modalities makes inference robust and provides complementary information [4]. In the field of ASR, audiovisual speech recognition (AV-ASR), which incorporates visual information from facial videos into ASR, has long been studied, and many studies have shown that the visual information contributes to noise robustness [5], [6]. AV-ASR now renews attention on how to incorporate two modalities effectively in an end-to-end manner [7], [8].

Textual input is more commonly used in combination with visual tasks such as visual question answering [9], [10] than audio tasks. Joint learning of visual and textual information with Transformer [11] flexibly captures the alignment between elements of an input text and regions in an image input with its attention mechanism, and it is even sensitive to syntactic relations such as associations between verbs and image regions that correspond to their arguments. Transformer-based vision-and-language models are extended to a unified Transformer encoder-decoder model that handles multiple tasks and domains in a single model [12].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In recent years, ASR with textual input has been studied in the context of low-resource settings [13]–[15]. To augment low-resource transcribed speech with much larger textual resources, multi-modal data augmentation [16] and pseudo-speech data augmentation architectures were proposed [14]. Pre-training an attention-based encoder-decoder ASR with the architectures showed significant improvement over a text-based augmentation without pre-training. Speech in low-resource languages sometimes accompanies translated speech in a high-resource language. Assuming an accurate transcription of the translated speech is available, ASR of the low-resource language with auxiliary input of the translation has been investigated. The auxiliary input of translation was studied in ASR on the basis of an encoder-decoder recurrent neural network with attention mechanisms [13], and that based on two Transformer encoders for audio and translation inputs and a decoder [15]. The Transformer-based model [15] has a similar architecture to our proposed model in a different context.

III. Transformer-based Multimodal ASR with Auxiliary Input of Source Language Text [1]

A. Model Structure

The proposed model consists of two Transformer encoders for inputting audio in the target language and source language text and one Transformer decoder for outputting recognized words. Figure 1 illustrates the model architecture. The blocks aligned in three vertical lines represent, from left to right, a source language text encoder, an acoustic encoder, and a decoder. The colors shading the blocks represent the types of Transformer blocks. The light gray blocks have a multi-head self-attention and a feed-forward network followed by a residual connection, layer normalization, and dropout with no source-target attention. The orange and green blocks have one and two multi-head source-target attentions, respectively, in between the multi-head self-attention and the feed-forward network. All Transformer blocks in the decoder have a source-target attention to the output of the audio encoder. Additionally, one Transformer block (the green one) in the decoder and one Transformer block (the orange one) in the audio encoder have a source-target attention to the output of the source language text encoder. The optimal arrangement of the two Transformer blocks was experimentally obtained through simulation using the ST corpus. The detailed process in the encoders and the decoder is described in [1].

B. Pre-training and Fine-tuning

The model is trained with two inputs of audio and source language text and an output of a correct transcription in an end-to-end manner. The training is conducted in teacher forcing, minimizing the cross-entropy error.

Since there is no corpus of simultaneous interpretation with source language text large enough to train the model, a large-scale ST corpus is used for pre-training instead. Then, a smaller corpus of simultaneous interpretation is used for fine-tuning the model.

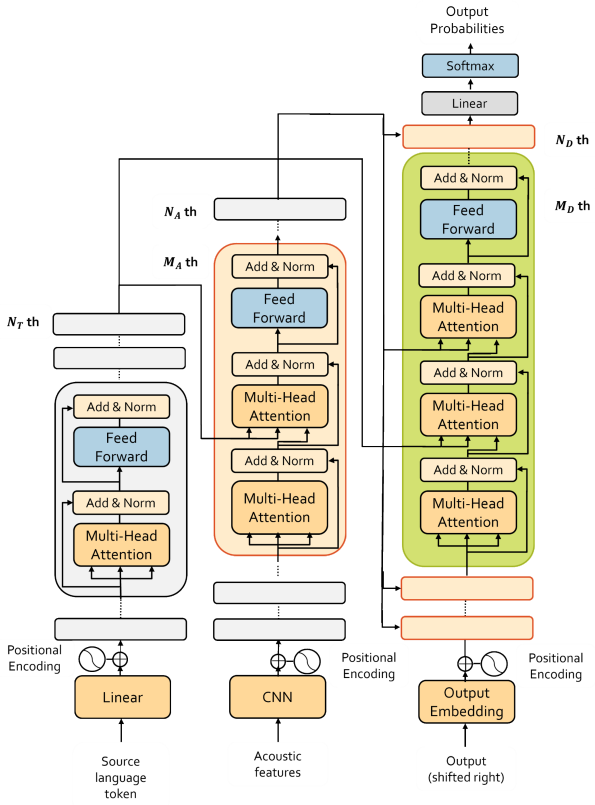


Fig. 1. Model structure.

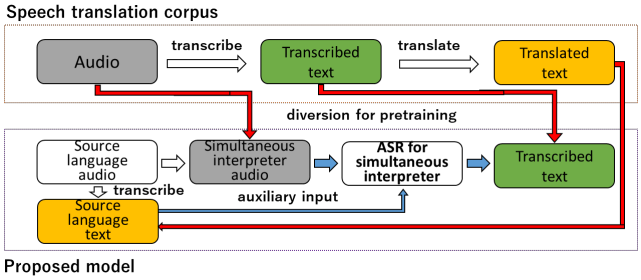


Fig. 2. Correspondence between components of speech translation (ST) corpus and input/output data of for training proposed model. Source language text (yellow block in bottom left) is substituted by translated text of ST corpus.

Figure 2 shows the correspondence between the components of the ST corpus and the input/output data for training the proposed model. The upper part shows three parts consisting of the ST corpus: audio, transcribed text, and translated text. The lower part shows the input/output data for training the proposed model: audio of simultaneous interpretation, transcribed text, and source language text. While the audio and the transcribed text are used as an input and an output of the model, the source language text (the yellow block in the bottom left) that does not exist is substituted by the translated text in the ST corpus.

The pre-trained model is further fine-tuned with a simultaneous interpretation corpus. The simultaneous interpretation

TABLE I
EXAMPLE OF PROCESSING FILLED PAUSES IN TRANSCRIPTION. WORDS WITH TAG OF FILLED PAUSE (IN RED FONTS) ARE SUBSTITUTED WITH LABEL '#' FOR TRAINING. RECOGNIZED LABELS '#' ARE REMOVED IN EVALUATION.

Stage	Transcription
Original	so uh depending on the uh depending on what uh this at uh modifies uh there are different ways of uh uh modification when understanding
Training	so # depending on the # depending on what # this at # modifies # there are different ways of # # modification when understanding
Evaluation	so depending on the depending on what this at modifies there are different ways of modification when understanding

corpus consists of source language speech, its simultaneous interpretation, and transcriptions of the two speech data.

C. Modeling Filled Pauses

Simultaneous interpretation contains many filled pauses and disfluencies. We extracted filled pauses in the transcription of the simultaneous interpretation corpus and added a filled pause model. Specifically, we introduced a filled pause model proposed in [3]. We substituted words tagged "filled pause" with a symbol '#' and trained the filled pause model.

IV. EXPERIMENTS

We conducted a comparative evaluation of the baseline model without the auxiliary input and the proposed model with the auxiliary input in two steps. First, we evaluated the two models with simulated data using the ST corpus and then evaluated them with the simultaneous interpretation corpus. We tested the filled pause model in the second step. We compared the effects of the auxiliary input and the filled pause model on the reduction of the WER.

A. Datasets

1) MuST-C Multilingual Speech Translation Corpus [2]: MuST-C is a large-scale speech corpus with translation in multiple languages for developing end-to-end ST models. The data consist of audio of English TED talks, transcribed text (i.e., subtitles), and manually translated text in multiple languages (i.e., translated subtitles). More than 500 hours of English TED talks were transcribed and translated into 15 languages. The dataset of the translated subtitles is maintained for each language, where 430 hours were transcribed per language on average and 237 hours at least. We used English speech as audio input and Dutch (MuST-C En-Nl), German (MuST-C En-De), and Japanese (MuST-C Ver.2 En-Ja) translations as the source language text. Must-C En-Nl, En-De, and En-Ja had 434.2, 400.0, and 540.9 hours of speech, respectively, as training sets, and they all had 4.1 hours of speech as a common test set.

2) CIAIR Simultaneous Interpretation Database [17]: CIAIR simultaneous interpretation database is a collection of simultaneous interpretations of monologue and dialogue speech between Japanese and English. The data consist of stereo audio of source language speech and its simultaneous interpretation in L and R channels, respectively, and transcriptions of both the source language and interpretation speech. We used monologue data consisting of lecture speech in Japanese and simultaneous interpretation in English. The Japanese-English monologue data contain 84 lectures that cover topics such as politics, history, culture, natural language processing, and spoken dialogue systems each lasting 10-30 minutes.

TABLE II
WERS OF BASELINE AND PROPOSED MODELS FOR THREE ST DATASETS: MUST-C EN-NL, MUST-C EN-DE, AND MUST-C VER.2 EN-JA.

	MuST-C		
	En-Nl	En-De	En-Ja
Baseline	14.4%	14.6%	11.4%
Proposed	9.5%	10.6%	10.4%

TABLE III
WERS OF BASELINE AND PROPOSED MODELS FOR CIAIR SIMULTANEOUS INTERPRETATION CORPUS. LEFT FIGURES ARE WERS INCLUDING FILLED PAUSES, WHEREAS RIGHT FIGURES IN PARENTHESES ARE WERS EXCLUDING FILLED PAUSES.

	W/o filled pause model	With filled pause model
Baseline	21.4% (19.7%)	18.5% (18.1%)
Proposed	20.5% (18.7%)	17.8% (17.3%)

The simultaneous interpretation data were processed in a sentence unit. Since the data were transcribed with time stamps and sentence boundary tags, the English speech data could be segmented into sentences. Then, the Japanese sentences corresponding to the segments of English speech were manually aligned after pre-processing of automatic alignment by Vecalign [18]. When a corresponding Japanese sentence was not found, the segment of English speech was removed from the dataset. As a result, a total of 6,019 utterances with 120,757 words, which amounted to 21 hours of simultaneous interpretation, were prepared with the source language text. We partitioned the sentence-segmented data into ten sets for cross-validation. In the data partitioning, we assigned all utterances from one interpreter into a specific set so that the cross-validation became an interpreter-open evaluation.

The transcription has various tags, such as a filled pause, repair, unclear voice, and general disfluency. We substituted tokens tagged "filled pause" with a symbol '#' for training the filled pause model. The total counts of '#' in the 6,019 utterances were 7,775, which occupied 6.4% of all tokens.

B. Experimental Setup

In the first step, we trained a baseline model and three types of the proposed model with the three language datasets of the MuST-C corpus. In the second step, we fine-tuned the baseline model and the proposed model with an auxiliary input of Japanese text, and evaluated them in cross-validation

TABLE IV

EXAMPLES OF RECOGNIZED TEXTS IMPROVED BY PROPOSED MODEL FOR CIAIR SIMULTANEOUS INTERPRETATION DATABASE. WORDS CORRECTED BY PROPOSED MODEL AND CORRESPONDING WORDS IN SOURCE LANGUAGE TEXT ARE UNDERLINED.

	A	B	C
Reference	but unfortunately north korea hasn't participated yet	the second is the bilingual dictionary is used	we have the incremental parsing and output control
Source language text input	ただ残念ながら北朝鮮は参加しておりません	もう 1 つは対訳辞書を用いる方法	漸進的構文解析とその出力を制御する部分
Recognized text by baseline model	but unfortunately NO SQUARE IT hasn't participated yet	the second is the BILATERAL dictionary is used	we have the incremental POLICY and output control
Recognized text by proposed model	but unfortunately north korea hasn't participated yet	the second is the bilingual dictionary is used	we have the incremental parsing and output control

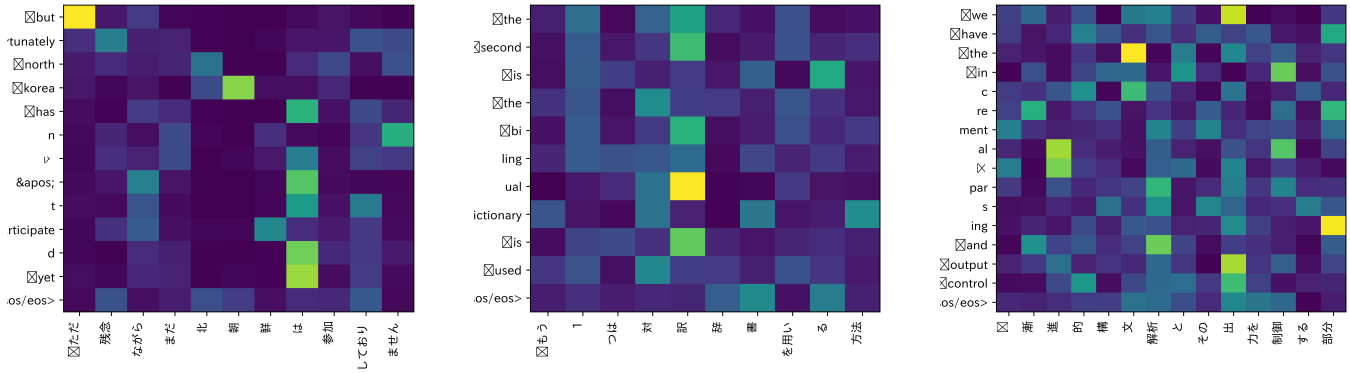


Fig. 3. Attention weights from decoder to source language text encoder in proposed model for samples listed in Table IV. Horizontal and vertical axes list source language tokens and recognized tokens, respectively. Yellow cells represent higher attention weights close to 1, while navy ones represent lower weights close to 0. Corresponding source language tokens or tokens close to those are attended to for corrected output tokens with proposed model.

with the CIAIR database. The two models were pre-trained with the MuST-C Ver.2 En-Ja dataset and fine-tuned with the CIAIR Japanese-English monologue dataset. This experiment examined the filled pause model, too.

We set model parameters as follows. The numbers of Transformer blocks in the source language text encoder, the audio encoder, and the decoder were 6, 12, and 6, respectively. The eleventh block in the audio encoder and the fourth block in the decoder had a source-target attention to the output of the source language text encoder, respectively. The dimension d_M of the embedding vectors in the source language text encoder, the audio encoder, and the decoder was 256. The number of heads in the multi-head attention was 4, and the dimension d_F of the vector in the feed-forward network in each Transformer encoder/decoder block was 2,048. The input acoustic features were 80-dimensional log Mel filterbank outputs sampled at 10 ms intervals and their temporal resolution was reduced to one-fourth by using a 2D-convolutional network.

The baseline model had the same model structure and parameters except that it had no source language text encoder.

English transcription and Japanese source language text were tokenized into 5,000 and 8,000 subwords, respectively, using SentencePiece unigram model [19]. The models were trained with a batch size of 64. The learning rate was set to 2.0e-4 with 20k warm-up steps. The training was conducted with early stopping with patience of five epochs at a maximum of 70 epochs. SpecAugment [20] was applied with parameters (m_F,m_T,f,t)=(2, 2, 15, 20). We used Espnet2 toolkit [21].

C. Experimental Results with MuST-C Corpus

Table II shows the WERs of the baseline and proposed model for three source language texts in the evaluation using the MuST-C corpus. The WERs were significantly reduced by the proposed model for all three source language texts. The reductions were greater for Dutch and German, which are from the same language family, than for Japanese, which is from a different language family.

D. Experimental Results with CIAIR Corpus

Table III shows the WERs of the baseline and proposed models with an option of the filled pause model for the CIAIR corpus. Percentage figures on the left and the right (in parentheses) are WERs, including and excluding filled pauses in the calculation, respectively. We will discuss the results on the basis of the WERs excluding the filled pauses here. To begin with, the baseline WER for CIAIR corpus was nearly double that for MuST-C corpus (11.4% vs. 19.7%). The auxiliary input of the source language text reduced the WER of the baseline by 1.0% (18.7%), whereas the filled pause model reduced it by 1.6% (18.1%). The auxiliary input reduced the WER of the baseline with the filled pause model by 0.8% (18.1% vs. 17.3%), whereas the filled pause model reduced the WER of the proposed model by 1.4% (18.7% vs. 17.3%). We conducted an ANOVA test on the WERs that were the mean of 84 interpreters' WERs with the model structure

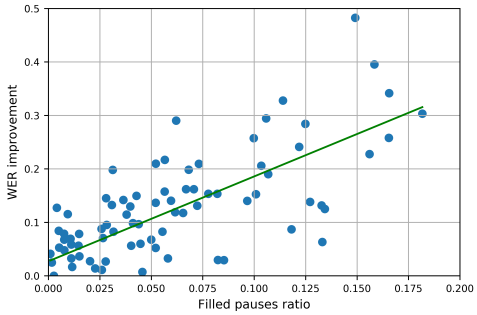


Fig. 4. Scatter plot of 84 interpreters' filled pause ratio and WER improvement for CIAIR corpus. Dot represents 10-30 minute long session. Straight green line represents regression line, and Pearson's correlation coefficient is 0.726.

and the filled pause model being within-subject factors. There was no significant interaction of the model structure and the filled pause model ($F_{(1,83)} = 1.5, p = 0.232$). There was a significant difference between the baseline and proposed models ($F_{(1,83)} = 15.6, p < 0.001$). There was a significant difference between with and without the filled pause model ($F_{(1,83)} = 102.6, p < 0.001$).

Table IV and Figure 3 show examples of recognized texts improved by the proposed model and the attention weights of a source-target attention from the decoder to the source language text encoder, respectively. The examples show that mis-recognitions of nouns and adjectives to similar-sounding words were corrected by the auxiliary input of the source language text. The attention maps show that the corresponding source language tokens or tokens close to those were attended from the corrected output tokens.

Figure 4 shows a scatter plot of the filled pause ratio and WER improvement by introducing the filled pause model into the proposed model for 84 sessions of simultaneous interpretation in CIAIR corpus. Blue dots represent 10-30 minute long sessions played by simultaneous interpreters. The green straight line represents a regression line. The ratio of filled pauses and the improvement of WER had a positive correlation. Pearson's correlation coefficient was 0.726.

V. Conclusions

We evaluated the Transformer-based end-to-end ASR that has an auxiliary input of source language text using a corpus of Japanese-English simultaneous interpretation. We also modeled filled pauses contained in the simultaneous interpretation and compared the effect on the reduction of WER to that made by the auxiliary input. The experimental results showed that both the auxiliary input of source language text and the filled pause model significantly reduced WERs though the reduction made by the filled pause model was greater than that made by the auxiliary input.

References

[1] Shuta Taniguchi, Tsuneo Kato, Akihiro Tamura, and Keiji Yasuda. Transformer-based automatic speech recognition with auxiliary input of source language text toward transcribing simultaneous interpretation. In *Proc. Interspeech 2022*, pp. 2813–2817, 2022.

[2] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, Vol. 66, pp. 1–14, 2021.

[3] Koharu Horii, Meiko Fukuda, Kengo Ohta, Ryota Nishimura, Atsunori Ogawa, and Norihide Kitaoka. End-to-End Spontaneous Speech Recognition Using Disfluency Labeling. In *Proc. Interspeech 2022*, pp. 4108–4112, 2022.

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 41, No. 2, pp. 423–443, feb 2019.

[5] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew Senior. Recent advances in the automatic recognition of audio-visual speech. In *Proc. IEEE*, pp. 1306–13261, 2003.

[6] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal Deep Learning. In *Proc. ICML 2011*, pp. 689–696, 2011.

[7] Hui Wang, Fei Gao, Yue Zhao, and Licheng Wu. WaveNet with cross-attention for audiovisual speech recognition. *IEEE Access*, Vol. 8, pp. 169160–169168, 2020.

[8] Fei Tao and Carlos Busso. End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia*, Vol. 23, pp. 1–11, 2020.

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32*, pp. 13–23, 2019.

[10] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, pp. 1–14, 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

[12] Ronghang Hu and Amanpreet Singh. UniT: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1439–1449, 2021.

[13] Antonios Anastasopoulos and David Chiang. Leveraging translation for speech transcription in low-resource settings. In *Proc Interspeech 2018*, pp. 1279–1283, 2018.

[14] Matthew Wiesner, Adithya Renduchintala, Shinji Watanabe, Chunxi Liu, Najim Dehak, and Sanjeev Khudanpur. Pretraining by Backtranslation for End-to-end ASR in Low-Resource Settings. In *Proc Interspeech 2019*, pp. 4375–4379, 2019.

[15] Kak Soky, Sheng Li, Masato Mimura, Chenhui Chu, and Tatsuya Kawahara. Leveraging Simultaneous Translation for Enhancing Transcription of Low-resource Language via Cross Attention Mechanism. In *Proc Interspeech 2022*, pp. 1362–1366, 2022.

[16] Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe. Multi-Modal Data Augmentation for End-to-End ASR. In *Proc Interspeech 2018*, pp. 2394–2398, 2018.

[17] Hitomi Tohyama, Shigeki Matsubara, Koichiro Ryu, N Kawaguch, and Yasuyoshi Inagaki. Ciair simultaneous interpretation corpus. In *Proc. Oriental COCOSDA*, 2004.

[18] Brian Thompson and Philipp Koehn. Vecalign: Improved sentence alignment in linear time and space. In *Proc. EMNLP 2019*, pp. 1342–1348, 2019.

[19] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. EMNLP 2018*, pp. 66–71, 2018.

[20] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[21] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech 2018*, pp. 2207–2211, 2018.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

An Analysis of Personalized Speech Recognition System Development for the Deaf and Hard-of-Hearing

Lester Phillip Violeta and Tomoki Toda
Nagoya University, Japan
E-mail: violeta.lesterphillip@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Deaf or hard-of-hearing (DHH) speakers typically have atypical speech caused by deafness. With the growing support of speech-based devices and software applications, more work needs to be done to make these devices inclusive to everyone. To do so, we analyze the use of openly-available automatic speech recognition (ASR) tools with a DHH Japanese speaker dataset. As these out-of-the-box ASR models typically do not perform well on DHH speech, we provide a thorough analysis of creating personalized ASR systems. We collected a large DHH speaker dataset of four speakers totaling around 28.05 hours and thoroughly analyzed the performance of different training frameworks by varying the training data sizes. Our findings show that 1000 utterances (or 1-2 hours) from a target speaker can already significantly improve the model performance with minimal amount of work needed, thus we recommend researchers to collect at least 1000 utterances to make an efficient personalized ASR system. In cases where 1000 utterances is difficult to collect, we also discover significant improvements in using previously proposed data augmentation techniques such as intermediate fine-tuning when only 200 utterances are available.**

I. Introduction

Deaf or hard of hearing (DHH) people typically have their development of speech stunted due to hearing impairedness. This results in difficult-to-understand speech [1], commonly referred to as atypical speech, which makes daily conversations a tedious task for them. As communication is one of the essential aspects of being human, there is a need to develop more systems to support DHH speakers and provide them with more communication alternatives.

Recent advancements in artificial intelligence (AI) has given way for computers to assist humans in their everyday life. One particular field with substantial improvements in the past decade is automatic speech recognition (ASR) [2], [3], [4], the conversion of speech into its corresponding text labels. Owing to neural network-based approaches and the availability of large-scale datasets, several systems have been proposed that could transcribe text at an astounding accuracy rate. Smart devices, in particular voice assistants, have used this technology as a means of communication between users' speech and a control system to execute the said command, enabling several conveniences such as controlling appliances around the house. Thus, with the growing popularity of ASR, several devices and software applications now support using speech to navigate through the device or an application. For DHH users, using ASR to transcribe their spoken speech into text becomes available as an alternative communication option. With the huge benefits and conveniences that ASR-based smart devices bring, improving and developing personalized ASR systems for DHH speakers can allow them a communication alternative and be able to use these devices as an assistant in everyday life.

Although ASR systems have been successful with typical speech, it has struggled with transcribing DHH speech due to the difference and high variability in the DHH speech patterns compared to typical speech. The deaf speaker usually cannot control the high-frequency components of speech [5] due to their hearing conditions, making how other humans and ASR models perceive their speech entirely different from typical speakers. Consequently, this means that although ASR models are usually trained on large-scale typical speech (usually a minimum of 1000 hours), the success of these ASR models is also limited to just typical speech. Replicating the same success on DHH speech would require training the ASR model on a large-scale DHH speaker dataset; however, recording at a large scale, due to strict ethics regulations and the effort it takes to setup recordings, makes this a difficult task to resolve. Thus, speakers with atypical speech have found these devices rather difficult to use [6], [7]. However, as mentioned, DHH users would be the ones who could benefit the most from ASR-based devices.

In this research, we provide an analysis of how to create personalized ASR systems for DHH speakers. Our goal is to provide a thorough analysis such that researchers would be able to understand the training size/performance tradeoff when collecting speech data, and how to make use of methodologies used in the literature to properly develop personalized ASR systems. Our previous work has shown the effectiveness of large-scale pretraining and using imperfect synthetic data to improve recognition accuracy in electrolaryngeal speech recognition [8]; however, there is still some doubt on whether the previously proposed method works on different types of atypical speech or on larger dataset sizes. Thus, this extension of the analysis provides information on the effects of varying the training data sizes and how it affects the performance of each ASR method. Moreover, compared to other atypical datasets used in the literature, we collected a large dataset of

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

four DHH speakers with 28.05 hours in total, allowing us a thorough examination of the dataset.

Our main findings are as follows:

- We perform a thorough analysis with publicly available ASR resources on four different DHH Japanese speakers to make a personalized ASR system. We train personalized models and analyze different scenarios with different training data sizes and data augmentation techniques to give researchers an idea of what requirements are needed to train high-performing personalized ASR systems.
- Through our findings, we confirm that our previously proposed speech synthesis-based data augmentations indeed improve ASR performance in low-resourced scenarios (200 utterances) even on DHH speech datasets. However, using at least 1000 utterances of real recordings to fine-tune pretrained models already shows huge performance boosts in ASR performance, and thus recommend researchers collect around this number of utterances as it is the best cost/performance tradeoff point.
- Finally, we emphasize the effectiveness of creating personalized ASR systems, as when conducting data augmentation, using the synthesized speech of the target speaker was always better than using real recordings from other speakers.

II. Related Literature

Since the number of devices and software applications that support the use of speech as an input grows more and more each day, using ASR as a communication alternative is becoming more available as an option. However, research such as [6], [7] has investigated that for DHH and atypical speech, currently productionized or open-sourced ASR models have up to 60% increase in word error rates (WER) compared to typical speech. Thus, although ASR-based technology has been gaining popularity and more support on devices, current models are still far from being inclusive to all people, and more investigations need to be done to resolve this issue.

Although not focused on DHH speech, several research works have investigated improving ASR performance for atypical speakers. The most common method to resolve this is by pretraining the model with large-scale typical speech data, and subsequently fine-tuning on the target low-resourced atypical speech data [9], [10], [11], [12]. Through large-scale pretraining, the model can generalize to the target better by using these as the initialization model parameters, rather than training the model from randomized initialization model parameters. The larger the pretraining dataset used, the more generalized the model could be and thus more robust to different types of speech during fine-tuning.

Research such as [11], [13] has shown that personalized ASR models will always outperform ASR models trained on other speakers' data. Although this work has shown that simply fine-tuning a pretrained model with target speakers can massively improve ASR performance, our previous research in [8] shows that naively fine-tuning the pretrained model on the target speaker is not very effective in a low-resourced scenario, such as when there are only 200 utterances available for each speaker. To further improve performance, we proposed using an intermediate fine-tuning approach using imperfect synthetic speech by using text-to-speech (TTS) [8]. The main point made here was that even low-quality synthetic speech of the target speaker (in this case, the electrolaryngeal speaker) could be utilized to improve ASR performance, contrary to the common intuition that data augmentations should always be of high quality. This has opened doors in improving low-resourced ASR, by requiring looser requirements to train ASR models of a target atypical speaker. However, more work has yet to be done to verify the effectiveness of this method on other types of atypical speakers (such as DHH) or if it would still work well in scenarios with a larger training data size as the experiments were only conducted using a small-scale electrolaryngeal speaker dataset.

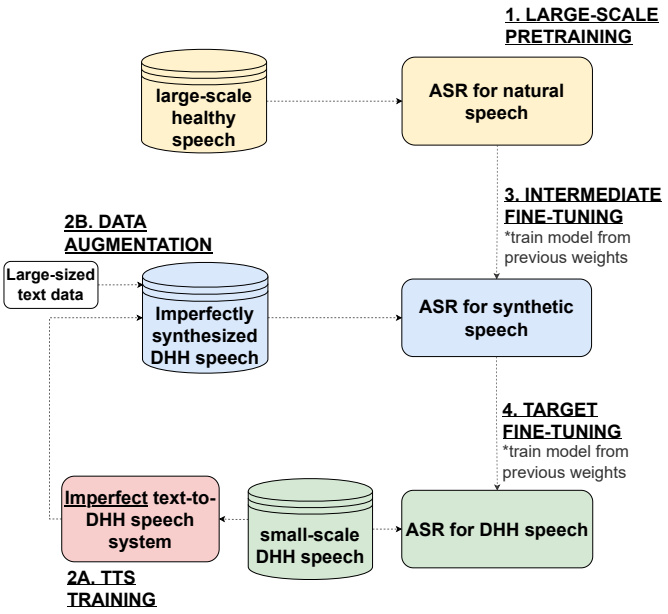


Fig. 1. An overview of the ASR training process. We follow the setup described in [8] and expand on the investigations of this work on deaf and hard-of-hearing speakers.

III. Methodology

A. Development of ASR model

We expand on our previous work shown in [8], to verify the effectiveness of the intermediate fine-tuning task on other atypical speech datasets and with different training data sizes. Following this experiment, we primarily used setups based on pretraining on a large-scale typical speech dataset and subsequently fine-tuning the model on each target speaker. We then perform a thorough analysis of their behavior to understand how to develop efficient ASR models. To generate synthetic speech used for data augmentation, we used the same TTS setup used in that experiment. The TTS setup involves using a pretrained model and fine-tuning it on each of the

TABLE I
AN OVERVIEW OF THE DHH SPEAKER DATASET COLLECTED.

Speaker ID	Gender	No. of utterances	Total length (Hours)
S1	Male	5000	11.41
S2	Male	4982	10.69
S3	Female	902	1.64
S4	Female	2493	4.31
Total	-	13377	28.05

speakers to create a larger training data. Then, text from a separate speech dataset is used to generate the larger speech dataset of each speaker. An overview of the ASR model training procedure is shown in Figure 1.

B. DHH dataset overview

The dataset we used was collected from four different DHH Japanese speakers in the comfort of their homes. To record the dataset, we asked them to utter the phrases from the Japanese speech corpus of Saruwatari Lab of the University of Tokyo (JSUT) dataset [14], a dataset spoken by a native female Japanese speaker. The dataset contains 7696 utterances and is commonly used for speech synthesis tasks. However, speakers were allowed to record only a subset of the dataset or skip some utterances if they felt uncomfortable saying some of the phrases. All datasets were recorded at a 44.1kHz sampling rate. A detailed look at the dataset can be found in Table I. Compared to commonly used atypical speech datasets such as UASpeech [15] and TORGO [16], our dataset is larger in total and contains a large number of utterances per speaker, making personalized ASR more possible. Moreover, compared to UASpeech [15] which only contains word-level utterances, our dataset contains sentence-level utterances, making it a more realistic scenario for certain personalized ASR scenarios. With a larger dataset, we aim to analyze the behavior of the network when varying the training data sizes.

IV. EXPERIMENTS

A. Experimental conditions

To conduct the experiments, we used ESPnet [17], [18], an open-sourced speech processing toolkit. For the ASR model architecture, we used a Conformer [19] encoder and a Transformer [20] decoder due to their proven strength in ASR tasks [21]. As for the TTS models, we used a Transformer [20] encoder and decoder. Since the TTS models only convert the text to a mel-spectrogram, we used a pretrained model based on HiFiGAN [22] as the vocoder to convert mel-spectrograms to waveforms. The ASR and TTS pretrained models are made openly-available by ESPNet, while we also used a pretrained HiFiGAN model[1] on the JSUT [14] dataset. To pretrain the ASR model, we used the large-scale LaboroTV [23] dataset, which contains around 2000 hours of Japanese speech collected from various TV recordings. On the other hand, to pretrain the TTS model, we used the entire JSUT [14] dataset.

To vary the training data sizes, we subset the datasets into different groups depending on the number of utterances available. More details on the number of utterances used are shown in Fig. 1 as they differ for each speaker due to the total dataset size. We evaluated the trained model using a test set with 250 utterances for speakers S1, S2, and S4, and 50 utterances for speaker S3. Although the TTS models were trained and synthesized the utterances in 24 kHz, the synthesized audio would be downsampled to 16 kHz when used to train the ASR model. Thus, all audio used in the ASR models were in a 16 kHz sampling rate.

B. Data augmentation setups

We describe the different setups used in our experiment. We follow the intermediate fine-tuning (IF) setup detailed in [8] and extend upon their findings.

- **Direct**: Directly using the out-of-the-box pretrained model and evaluating the performance without performing any fine-tuning.
- **Direct - speaker-independent**: Fine-tuning the dataset with the other DHH speakers' audio data. This means that we do not use the target DHH speaker's audio data to fine-tune the model.
- **Naive fine-tuning**: Simply fine-tuning the pretrained model on the target speaker without any extra procedures.
- **IF - TTS**: Based on the intermediate fine-tuning setup in [8]. We generate a larger dataset of the target DHH speaker using TTS, and use these during intermediate fine-tuning.
- **IF - TTS with text-swapping**: Similar to the TTS setup, but randomly swapping the text labels within the dataset. Note that the text labels are only swapped during the intermediate fine-tuning stage. We further investigate this setup due to its reported success in [8] but non-intuitive setup.
- **IF - speaker-independent**: Instead of using synthetic speech, we use the real speech recordings from other DHH speakers during intermediate fine-tuning. This allows us to assess if the improvements in intermediate fine-tuning come from the speaker-dependent or speaker-independent characteristics of the data.

C. Evaluation metric

To evaluate the ASR models, we use the character error rate (CER), as seen in Equation 1, a commonly used metric for Japanese ASR. Given the number of characters (N), and the number of wrongly predicted characters, which are categorized into three types, insertions (I), deletions (D), and substitutions (S), the CER can be calculated as the sum of the insertions, deletions, and substitutions divided by the number of characters, multiplied by 100. Thus, the lower the CER value, the better the performance of the ASR model.

$$\mathrm{CER} = \frac{\mathrm{I}+\mathrm{D}+\mathrm{S}}{\mathrm{N}} \,\mathrm{x}\, 100 \tag{1}$$

[1] https://github.com/kan-bayashi/ParallelWaveGAN

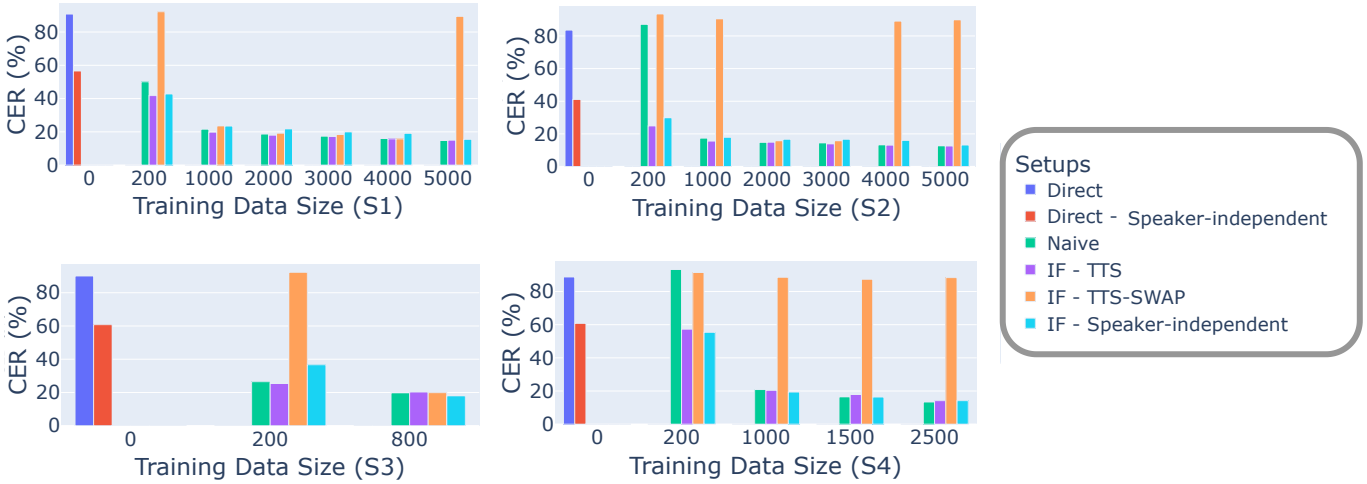


Fig. 2. Character error rate (CER) of the different developed ASR systems for all speakers (S1, S2, S3, and S4). The x-axis indicates the amount of data used in the training and development sets.

V. Results and Analysis

A. ASR model performance using out-of-the-box pretrained models

We present the summary of our results in Figure 2. As displayed, in all cases where fine-tuning is not used, the ASR model always performed above 80% CER as shown by the results of using the Direct setup. Thus, despite the ASR model being pretrained on a large-scale dataset, it still has difficulties adapting to the DHH speech, showing the difficulties faced by deaf speakers using out-of-the-box speech recognition devices. Aside from this, fine-tuning on speech data of other DHH speakers improves performance, but is still not effective enough as shown in the Direct - speaker-independent setup, proving the need for personalized ASR models due to the high variance in atypical speech. On the other hand, when fine-tuning the model on the available utterances (as shown by the Naive setup), the model's performance almost always improves, where the only exception is when using 200 utterances, similar to the findings in [8]. However, using a larger dataset of the target speaker should be sufficient enough for the model to generalize. In all speakers, we find improvements of between 5% to 30% CER by increasing the number of utterances to 1000 from 200. Thus, we recommend researchers collect at least 1000 utterances as the ASR models can be already be effectively fine-tuned using this amount of data.

B. Extending investigations on intermediate fine-tuning

We investigate whether our previously proposed data augmentation techniques for improving electrolaryngeal speech recognition also perform well on DHH speech. Using data augmentation techniques such as TTS to create a larger dataset of the target speaker and using intermediate fine-tuning, we see that this fine-tuning method can significantly improve performance in this low-resourced scenario, verifying that our

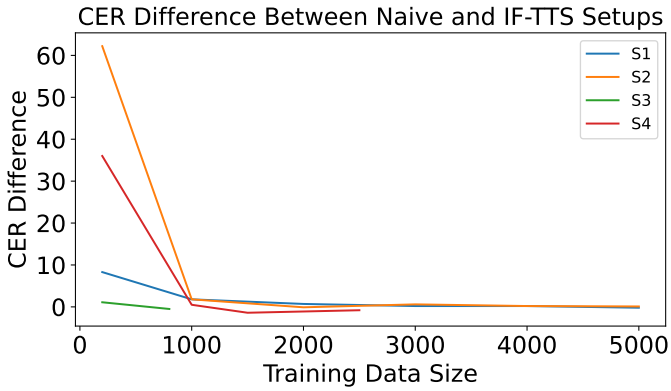


Fig. 3. Difference of the character error rate (CER) of the Naive and IF-TTS setups across different training data sizes.

method proposed in [8] can also be transferred to DHH speech data. However, we observed that using intermediate fine-tuning became less effective as the training data size increased. Amongst the different setups, we find that the IF-TTS setup was the most stable and showed the most improvements in the low-resourced scenarios. Although the IF-TTS text-swapping setup was reported to be successful in [8], we do not see the same trend in our experiments as we observe its instability. Specifically, in cases where the training data is too low or too high, we observed a collapse and the ASR model fails. One hypothesis on why this is the case is because of the difference in speech features between electrolaryngeal speech and DHH speech. For electrolaryngeal speech, the speaker identity is disrupted, but the intelligibility is still there, as several phonemes can still be pronounced almost the same as a typical speaker. On the other hand, DHH speech contains the speaker identity but has unintelligible speech. Thus, since the intelligibility information between the pretraining data

(LaboroTV) and the target data (electrolaryngeal speech) could be transferred and shared, the intermediate fine-tuning worked better on the electrolaryngeal dataset. On the other hand, in the case of our experiments, since the intelligibility information between the pretraining data (LaboroTV) and the target data (DHH speech) vastly differed, the information was not effectively shared. ASR models typically focus on the intelligibility information and disregard speaker identity information, which also explains why the text-swapping was not as effective in the case of DHH speakers. Finally, we observe that using data augmentations of the target speaker, even when the quality is low, was still always better than using real recordings of other DHH speakers as shown by the IF - speaker-independent setup results.

As a final observation, although we verified the validity of the intermediate fine-tuning in the low-resource scenario, we still found that it cannot beat the performance when using real recordings of the target speaker as the training data. A detailed look at the comparison of the Naive and IF-TTS setups is shown in Fig 3. As seen in the figure, using imperfect synthetic speech only became successful in the 200 utterance setup as there was a larger CER difference between using the two setups; however, as more real data was collected and used during training, the effects of using imperfect synthetic speech became close to none. Thus, although we acknowledge the difficulty of obtaining training data from the target speaker, we urge researchers to collect at least 1000 utterances to easily develop efficient personalized ASR systems, and such that they would not need to do any additional procedures such as intermediate fine-tuning and speech synthesis to improve the ASR models. On the other hand, in cases where collecting data is difficult, researchers can still opt to use the intermediate fine-tuning procedure as an alternative, as it has been verified to work even on other types of atypical speech.

VI. Conclusion

We presented a thorough analysis of developing personalized ASR systems for DHH Japanese speakers. We provided a thorough extension of our previous work to verify its effectiveness by analyzing its behavior when varying the number of utterances in the training data and with a different atypical speech dataset. We collected a large-scale DHH dataset, totaling around 28.05 hours of four speakers. Our findings show the following observations: 1) the most efficient setup is to collect 1000 utterances from the target speaker, as we see the most performance boost compared to using just 200 utterances, 2) in cases where 1000 utterances are challenging to collect, data augmentation techniques such as using intermediate fine-tuning with synthetic speech is still an available option to improve performance, and finally, 3) we emphasized the importance of using personalized models, as using real recordings from the target speaker as training data greatly outperforms using synthetic speech or recordings from other speakers. Future work would be to improve the speech intelligibility of atypical speakers using voice conversion techniques such that they can easily use out-of-the-box ASR models.

Acknowledgment

This work was partly supported by AMED under Grant Number JP21dk0310114, Japan, and a project, JPNP20006, commissioned by NEDO.

References

[1] M. J. Osberger and N. S. McGarr, "Speech production characteristics of the hearing impaired," ser. Speech and Language, N. J. LASS, Ed., vol. 8. Elsevier, 1982, pp. 221–283.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[3] H. A. Bourlard and N. Morgan, *Connectionist speech recognition*. Kluwer Academic Publishers, 1994.

[4] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[5] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 18 866–9, 01 2007.

[6] A. T. Glasser, K. R. Kushalnagar, and R. S. Kushalnagar, "Feasibility of using automatic speech recognition with voices of deaf and hard-of-hearing individuals," in *Proc. ASSETS*. New York, NY, USA: Association for Computing Machinery, 2017, p. 373–374.

[7] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems," in *Proc. Interspeech*, 2018, pp. 466–470.

[8] L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Intermediate Fine-Tuning Using Imperfect Synthetic Speech for Improving Electrolaryngeal Speech Recognition," in *Proc. ICASSP*. IEEE, 2023.

[9] R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, J. R. Green, and K. Tomanek, "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia," in *Proc. Interspeech*, 2021, pp. 4833–4837.

[10] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, and Y. Matias, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Proc. Interspeech*, 2019, pp. 784–788.

[11] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, and K. Tomanek, "Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases," in *Proc. Interspeech*, 2021, pp. 4778–4782.

[12] L. P. Violeta, W. C. Huang, and T. Toda, "Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition," in *Proc. Interspeech*, 2022, pp. 41–45.

[13] J. Tobin and K. Tomanek, "Personalized automatic speech recognition trained on small disordered speech datasets," in *Proc. ICASSP*, 2022, pp. 6637–6641.

[14] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017. [Online]. Available: https://arxiv.org/abs/1711.00354

[15] H. Kim, M. H. Johnson, J. Gunderson, A. Perlman, T. Huang, K. Watkin, S. Frame, H. V. Sharma, and X. Zhou, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[16] F. Rudzicz, A. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 1–19, 01 2010.

[17] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[18] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo, S. Karita, C. Li, J. Shi, A. Subramanian, and W. Zhang, "The 2020 espnet update: New features, broadened applications, performance improvements, and future plans," 06 2021, pp. 1–6.

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on espnet toolkit boosted by conformer," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5874–5878.

[22] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.

[23] S. Ando and H. Fujihara, "Construction of a Large-Scale Japanese ASR Corpus on TV Recordings," in *Proc. ICASSP*, 2021, pp. 6948–6952.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Classification of Vocal Cord Disorders: Comparison Across Voice Datasets, Speech Tasks, and Machine Learning Methods

Ching-Chieh Chen*, Wei-Cheng Hsu†, Tzu-Han Lin*, Kuan-Dar Chen*, Yung-An Tsou† and Yi-Wen Liu*
* Dept. Electrical Engineering, National Tsing Hua University, Taiwan
† Dept. Otorhinolaryngology, Head and Neck Surgery, China Medical University Hospital, Taiwan

***Abstract*—Dysphonia may originate from various vocal cord disorders (VCDs), and determination of its etiology usually requires image-based examination. However, such examination is invasive and not available in rural areas. Therefore, we aim to evaluate whether the voice of VCD patients may provide cues that could be utilized for screening purposes. Presently, a voice dataset containing entries of vocal fold atrophy, paralysis, benign organic lesions, and laryngeal cancer was prepared and support vector machine and neural network models were trained for VCD classification. Features were first extracted from /a/, and performance on the present dataset was compared against the Saarbruecken voice database. Next, the utterances of counting from 1 to 5 were processed, and "4", pronounced as /si/ in Mandarin, was found most suitable for classifying non-cancer VCDs. Finally, we demonstrated that inclusion of the subjective GRBAS scale consistently raised the classification accuracy.**

I. Introduction

Dysphonia, or having troubles with voice when one attempts to talk, may be a consequence of various vocal cord disorders (VCDs), including laryngeal cancer. Since dysphonia is common, the underlying VCDs are often overlooked and could therefore miss the best time for treatment. Accurate clinical diagnosis of VCDs requires video stroboscopic examination. However, the examination may not be available in rural or other medically under-staffed areas. To help clinicians identify the etiology of VCDs in a timely fashion and promote the welfare of residents in these areas, it is desirable to deploy a voice disorder screening tool that is easy to use by patients remotely.

Recordings of the patients' own voice may be a good starting point. Thanks to advances in speech and voice processing by machine-learning algorithms, one can envision that different VCDs possibly exhibit distinct spectrotemporal cues to be discovered and utilized by machines automatically. Indeed, pioneering work was conducted by Al-Nasheri *et al.* to detect and classify VCDs using feature vectors extracted from patients' sustained vowel /a/; the feature vectors could be derived from auto-correlation and entropy [1], or simply from the *Multidimensional Voice Program* (MDVP) [2]. In these studies, the efficacy of support vector machine (SVM) models were evaluated across German, English, and Arabic databases. More recently, Fang *et al.* thoroughly compared the performance of SVM, Gaussian mixture models, and deep neural networks (DNNs) for pathological voice detection [3], while others transformed voice data into images and applied convolutional neural networks (CNNs) to assess the health of the vocal cord [4]–[7].

The above studies commonly used the Saarbruecken Voice Database (SVD) [8] and the Massachusetts Eye and Ear Infirmary (MEEI) [9] database to validate performance of the models, but several limitations were noticed. According to Al-Nasheri *et al*. [1], [2], the performance that could be achieved by any model depended on (1) variation in the severity of pathological voice in a database, (2) differences in the recording environment and recording rules between databases, (3) mismatch in the demographic distributions between different types of pathology within the same database, and (4) imbalance between the number of samples for different disorders within a database. In addition, differences in the ethnic and linguistic background of patients might also be a factor to consider for VCD detection and classification purposes.

Though these challenges exist, nevertheless, accuracy of $> 90\%$ on the binary classification of pathological against normal voice has been achieved in previous studies (e.g., [3]). Therefore, in this study, we chose to focus on discriminating VCDs that might differ in subtle ways. Based on prevalence of the disorder and the importance of early detection to minimize the need for surgical intervention, we selected four categories of VCDs: vocal cord atrophy, vocal cord paralysis, benign organic lesions, and cancer lesions. We endeavored to establish a dataset that is well-balanced and age-matched between different disorders. The dataset contains voice recordings of the sustained /a/, counting from 1 to 10 (in Mandarin), and reading a short text. In terms of machine learning methods, we evaluated SVM and various pre-trained CNNs. The rest of this paper is organized as follows. Sec. II describes the present dataset and its counter-part that was sampled from the Saarbruecken database. Sec. III describes machine learning methods for VCD classification. Sec. IV presents the performance of models in terms of their accuracy on various classification tasks. Finally, discussion and conclusions are given in Sec. V and Sec. VI, respectively.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Dataset preparation

As mentioned in Sec. 1, the following four categories were chosen as the targets for classification: vocal cord atrophy, vocal cord paralysis, benign organic lesions (BOL), and cancer lesions. Laryngitis, though common, was not included because of its relatively short time from symptom onset to recovery.

Each of the four categories may contain several subtypes of disorders – the atrophy class (atrophy and sulcus), the paralysis class (palsy), the benign organic lesions (polyps, nodules, cysts,and edema) and the cancer lesions (leukoplakia and laryngeal cancer). To investigate the feasibility of machine classification, we trained SVM models and various CNNs and tested them on two datasets (namely, Saarbruecken and *the present*) that contain sufficient samples of the VCDs listed above. The datasets are introduced next.

A. The Saarbruecken voice database

The Saarbruecken voice database [8] is a German vocal cord disease database which contains more than 2,000 records. Among the available records, 1,356 records are pathological voice, and 687 records are healthy voice. The collection includes vowels /a/, /i/, /u/ with 4 types of intonation (low, normal, high, low-high-low), along with a spoken sentence "Guten Morgen, wie geht es Ihnen?" To facilitate the comparison across datasets, we selected the voice files of normal intonation sustained vowel /a/. The phonation time in the Saarbruecken dataset ranges between 1-4 seconds, and the sampling rate of the recording is 50 kHz with 16 bit resolution. The original database includes 71 disease types. After excluding the data with multiple diseases, the number of records in many categories became fewer than 10. We selected atrophy, palsy, polyp, and cancer for comparison purposes, and there were 35, 198, 36, and 20 records, respectively.

B. The present dataset

The voice data used in this research were retrospectively collected from a larger database that was established by the Voice and Swallowing Center of China Medical University Hospital (CMUH) in Taichung, Taiwan since 2017. Voice and electroglottography (EGG) signals were recorded by speech therapists for patients with laryngeal lesions who needed a diagnosis. The electroglottography instrument Model 6103 (KayPENTAX, Lincoln Park, NJ) was utilized, and an external microphone was connected to the EGG hardware interface (Computerized Speech Lab, model 4500). The signals were synchronously sent to the an analysis software (Real-Time EGG Analysis software, model 5138). The software computed the patients' MDVP, which helped the doctor to assess the patients' condition. In total, about 2300 entries were recorded.

The larger database contains a variety of VCDs; however, some disorders are quite rare, and in some cases the vocal cords were indirectly affected by lesions in other body parts of the patient. Some patients had concurrent VCDs, such as left vocal cord paralysis with right vocal cord atrophy. To avoid complications, we excluded the cases of multiple disorders in this study. The original database also included some data from the same patients before and after treatment, and we selected the ones before treatment. Thus, a total of 459 data entries would be studied in the present research, including 138 of atrophy (65M, 73F), 118 of palsy (56M, 62F), 142 of polyp and other BOL (31M, 111F), and 40 of cancer (36M, 4F). Statistics for their ages at the time of voice recording were as follows: atrophy = 58 ± 13, palsy = 59 ± 14, BOL = 54 ± 12, and cancer = 61 ± 13 years. This research was approved by the IRB of CMUH.

The database contains voice samples for the following speech tasks: (1) sustaining the vowel /a/, (2) counting from 1 to 10 in Mandarin, and (3) reciting a fixed passage from an article in Mandarin which contains 9 phrases and 71 Chinese characters. In this paper, only the /a/ task and the counting from 1 to 5 were analyzed. Table I summarizes the total number of data entries by different categories and speech tasks, as compared against those of the Saarbruecken dataset. For the present dataset, the counting task contributes to slightly fewer entries than /a/ because some patients had difficulties completing the task.

TABLE I
TOTAL NUMBER OF DATA ENTRIES IN THIS STUDY. THE SYMBOLS $/a/_1$ AND $/a/_2$ STAND FOR THE /a/ VOICE IN OUR PRESENT DATASET AND THE SAARBRUECKEN DATASET, RESPECTIVELY; "1-5" STANDS FOR THE MANDARIN COUNTING FROM ONE TO FIVE IN OUR PRESENT DATASET.

Disorder Type	$/a/_1$	"1-5"	$/a/_2$
Atrophy	138	135	35
Paralysis	118	106	198
Benign organic lesions	142	138	36
Cancer	40	34	20

To enhance the data, we also rated the voice quality of 459 patients in the GRBAS scale (*grade, roughness, breathiness, asthenia*, and *strain*) [10]. The GRBAS scale is commonly used in laryngology for voice quality evaluation. A score of 0, 1, 2, or 3 was given for each of the five scales to represent severity (normal = 0, and the most severe = 3). Existing research suggested that these five indicators can be useful for specifying VCDs [10]. In our study, two different versions of the GRBAS scale were collected. The first version was rated by the first author of this paper while focusing on the voice quality of the sustained vowel /a/, and the second one was rated by a speech therapist based on the overall voice quality in performing the sustained /a/ task, the counting task, and the reciting task.

III. Vocal cord disorder classification

In this section, we disclose details in the signal processing and machine learning methods that were used in this research.

A. Signal processing and feature extraction

First, the voice recordings in both the present and the Saarbruecken datasets were down-sampled to 16 kHz. A 512-point (or 32 ms, equivalently) Hann window was applied to calculate the short-time Fourier transform with 50% overlap. Then, the magnitude spectrogram was subject to triangular

filtering to compute 128 Mel energies. For training SVM models, we derived 20 Mel-frequency cepstral coefficients (MFCC); for training neural networks, 12-bins chroma vectors [11], which are typically adopted for music analysis and recognition purposes, were also extracted using *librosa* [12]. We also defined several aggregated features by stacking Mel-spectrogram, MFCC, and the chroma vector vertically, which was shown to improve classification performance empirically in existing research [13].

B. Support vector machine (SVM) configurations

The length of each person's phonation of /a/ is different. To reduce the dimension for training SVM, we computed the statistical functionals (mean, variance, median, and first and third quartile) of MFCC along the time axis as the features. After concatenation, the dimension of the feature space became $5 \times 20 = 100$.

Among various commonly used kernel functions, we selected the linear kernel and the radial basis function (RBF) for the experiments. In its simplest form, SVM was developed for binary classification. To classify data into three or more categories, multiple binary SVM models should be first constructed. In this research, we adopted the *one vs. rest* [14] strategy, which predicts data labels by considering the distance of a datapoint to the separation hyperplane in every binary model.

The penalty parameter C, the kernel type, and the RBF shape parameter γ were jointly determined by grid search within a reasonable range (Table II). In Sec. IV, results achieved by the best hyperparameter combination in terms of classification accuracy will be reported.

TABLE II
THE RANGE OF GRID SEARCH FOR SVM PARAMETERS

Parameter	Values
Kernel type	'linear','rbf'
C	0.8, 1, 1.2, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10
γ	0.1, 0.01, 0.001, 0.0001

C. Neural network configurations

We tested various off-the-shelf ImageNet pre-trained convolutional neural networks from Tensorflow-Keras framework, including ResNet50 [15], DenseNet121 [16], InceptionV3 [17], Xception [18], MobileNet [19] and EfficientNetV2M [20]. Each network model was provided with various input features and performance of all feature-model combinations was evaluated. To modify and fine-tune the models, we loaded the models with ImageNet pre-trained weights and set all hyperparameters trainable. The input layer was modified to fit the shape of each feature, and softmax activation was utilized in the output layer. Each model was fine-tuned by 20 epochs with a batch size of 16, using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} after every epoch. All the training and validation were conducted on a single NVIDIA RTX-3060-12G GPU and AMD 5600x CPU with 16GB of RAM.

D. Cross validation settings

To evaluate the SVM and CNNs, we applied 5-fold cross validation to avoid overfitting. Note that the outcome depended on how data were assigned into five folds; therefore, we performed 10 rounds of random assignment, conducted all the experiments for each round, and will report the mean and standard deviation of accuracy across the 10 rounds in Sec. IV.

IV. RESULTS

In this section, we report the accuracy of vocal cord disorder classification.

A. Performance on different datasets

Table III shows the SVM classification accuracy for different tasks and compares the results across the two datasets. In the first two rows of the table, the "cancer/non-cancer" task refers to binary classification of cancer vs. the three other categories, and "3-class" refers to classifying the three benign cateogories (atrophy, paralysis, and BOL).

Generally, we observe that the rates of accurate classification on the Saarbruecken dataset exceed that of the present dataset. However, this is probably due to the fact that the former is less balanced than the latter in terms distribution of VCDs. For the present dataset, the accuracy of classifying the three types of benign disorders is merely 46.12% on average. Based on physicians' subjectively experiences, this task is also quite difficult for human ears. Therefore, it is not surprising that Table III did not clearly indicate the potential of using MFCCs to discriminate the three benign disorders.

Next, we examine if there was a clear difference between any two categories of the benign disorders. It is noteworthy that the average classification accuracy between BOL and atrophy is only 52.32%. From a pathological point of view, vocal cord atrophy can lead to incomplete closure, and the BOLs may also cause gaps in the closure of the vocal cords. When the gaps caused by the two diseases are similar, the voices may be difficult to distinguish. In contrast, patients with palsy generally had more severe voice abnormalities in the present dataset. Consequently, a higher accuracy could be achieved between paralysis/atrophy or paralysis/BOL than between BOL/atrophy. The same trend is also observed with the Saarbruecken dataset.

In the last column of Table III, we report the results of mixing the two datasets and training a single model to perform classification. The rates of accurate classification by mixed modeling consistently fall between what were obtained on individual datasets.

B. Vocal cord disorder classification by neural networks

In this part of the experiments, we tested the neural networks with the 3-class task. The rates of accurate classification for all feature-model combinations are listed in Table IV. Because the sizes of MFCC and the aggregated MFCC+Chroma feature

TABLE III
PERFORMANCE OF DISORDER CLASSIFICATION (%) BY THE SUPPORT VECTOR MACHINE MODEL

Classification task	/a/$_1$	/a/$_2$	Mixed
Cancer/non-cancer	75.14±4.96	85.00±6.44	78.07±4.10
3-class	46.12±7.17	67.22±7.59	56.57±3.02
Atrophy/Paralysis	66.15±3.77	82.77±3.86	76.94±3.82
Paralysis/BOL	65.58±7.37	79.79±4.78	70.20±5.10
BOL/Atrophy	52.32±5.59	61.33±4.00	53.94±6.73

were below the minimal requirement of InceptionV3, the corresponding evaluations are absent. Otherwise, it appears that the aggregated feature Mel+Chroma yielded better results than other features on most of the models, while Xception produced the highest accuracy among all the models regardless of the features being used. When used jointly, the combination of the Mel+Chroma feature and Xception (with 49.52% accuracy) outperformed the baseline SVM (46.12%, Table III) by approximately 3.4%.

C. Comparing the sustained /a/ and the counting task

The Mandarin numbers 1 to 10 have the unique property in that every number has exactly one syllable. This simple fact provides a research opportunity to investigate whether utterances of the numbers could contain useful cues for different vocal disorder classification tasks. In this research, we first performed forced alignment to the counting task voice data using AISHELL [21] on the Kaldi platform [22]. The utterances were thus parsed automatically, so that we could subsequently apply the same feature extraction and SVM training settings that were used for sustained /a/ data to the voice recordings in the counting task. Presently, we focus on "1" (yi1), "2" (er4), "3" (san1), "4" (si4), and "5" (wu3). Here, the pronunciation of each character is transcribed in Pinyin with the tone number. Note that "1" and "5" happen to contain sustained vowels /i/ and /u/, respectively. Therefore, we expected that they might provide information that is complementary to /a/ for the purpose of voice disorder classification.

Performance of the SVM model is summarized in Table V. From the first row, it appears that /a/ and "1" were most effective for cancer detection. Quite peculiarly, the model trained on utterances of "4" achieved the highest accuracy in many of the classification tasks. For classifying Paralysis vs. BOL in particular, the performance using "4" was approximately 17% higher than using /a/. Implications of these observations will be further discussed in Sec. V.

D. The potential value of adding GRBAS

We combined GRBAS and MFCC and conducted another set of experiments. When combined with MFCC functionals, the feature space extended by 5 dimensions because G, R, B, A, and S each provided a numeric score. SVM models were trained in this extended space. In Table VI, the label $+$GRBAS^1 refers to the results of adding the GRBAS scale that was evaluated on the sustained vowel /a/, while $+$GRBAS^2 refers to the results of adding the overall speech sound quality evaluation. Apparently, adding the subjective GRBAS scores improved the performance in all classification tasks. The improvement appears greatest in cancer detection (from approximately 75% to 83%) and in Atrophy vs. BOL (from approximately 52% to 63%). Implications will be discussed in Sec. V.

V. DISCUSSION

A noteworthy finding from Table IV would be the performance boost produced by adding the chroma vectors. Since chroma is a pitch-related feature, its usefulness might imply that benign VCDs had tendencies to alter the pitch of patients' voice differently while they delivered the sustained /a/. If so, these tendencies probably were not captured explicitly by Mel energies or MFCC. Therefore, the chroma feature might have provided information that was complementary to Mel-based features so as to raise the classification accuracy in our experiments.

The most surprising finding in this research is probably the efficacy of "4" for classification between paralysis, atrophy, and BOLs (Table V). While the Mandarin "4" is transcribed as *si4* in Pinyin, the vowel is not the front vowel /i/; it has been either considered as the central vowel /ɨ/ or a denti-alveolar approximant /ɹ/ which follows /s/ [23]. The promising findings with Mandarin "4" suggest that it might be beneficial to explore continuous speech data for VCD classification purposes [24]. As long as the short text is phonetically balanced, rich information in all the phonemes could be analyzed either separately or sequentially to assess vocal cord conditions.

Finally, GRBAS required subjective evaluation in the present research and this may seem to defy the purpose of serving medically under-staffed areas. Nevertheless, the GRBAS scoring process could also be automated [25]. Since results in Table VI clearly indicate the benefit of including GRBAS, in the future, it would be worthwhile to develop a separate GRBAS prediction module and integrate it with the current systems. Until then, however, the results listed in Table VI should be viewed as a performance upperbound that is not yet achieved by the present methods.

VI. CONCLUSIONS

This research focused on distinguishing the voice of vocal cord atrophy, paralysis, BOLs, and cancers. The first three categories are benign but atrophy and BOLs are difficult to differentiate automatically by voice. Automatic classification of cancer/non-cancer voice, nevertheless, produced promising results especially if GRBAS could be jointly considered. The fact that a model that was developed based on /sɨ/ outperformed /a/ in many tasks may point toward further research directions as to process continuous speech data for voice-based classification of VCDs.

VII. ACKNOWLEDGEMENTS

We would like to thank the National Science and Technology Council of Taiwan for supporting this research under grant No. 109-2221-E-007-094-MY3.

TABLE IV
THE ACCURACY (%) OF PRE-TRAINED NEURAL NETWORKS ON THE "3-CLASS" TASK

Model	**Feature(s)** Mel	MFCC	Mel+Chroma	Mel+MFCC	MFCC+Chroma
ResNet50	36.59±5.92	36.92±2.93	36.63±4.11	39.57±7.56	**40.67±7.34**
DenseNet121	42.56±5.71	38.56±4.58	**42.69±4.96**	41.13±4.43	37.76±5.67
InceptionV3	47.79±5.33	-	**47.96±3.61**	46.06±4.26	-
Xception	48.38±6.15	44.88±4.55	**49.52±5.19**	48.13±4.04	48.46±5.13
MobileNet	45.98±5.34	40.46±4.20	**47.62±5.15**	44.13±4.99	37.31±4.73
EfficientNetV2M	41.58±5.79	39.95±2.91	**43.68±1.58**	43.16±1.05	41.05±3.16

TABLE V
THE ACCURACY (%) OF USING SUSTAINED /a/ VS THE COUNTING TASK

Classification task	/a/$_1$	"1" (*yi1*)	"5" (*wu3*)	"2" (*er4*)	"3" (*san1*)	"4" (*si4*)
Cancer/non-cancer	**75.14±4.96**	75.00±3.42	71.88±7.53	74.38±7.88	73.12±4.88	72.81±5.60
3-class	46.12±7.17	46.32±3.71	49.21±5.46	46.05±3.86	46.32±2.02	**50.66±5.04**
Atrophy/Paralysis	66.15±3.77	59.80±7.75	67.76±8.84	64.29±4.21	64.29±4.59	**69.18±6.61**
Paralysis/BOL	65.58±7.37	70.82±5.40	76.12±4.19	71.02±6.11	71.22±7.82	**82.65±3.79**
BOL/Atrophy	52.32±5.59	**58.00±4.34**	50.00±5.09	53.45±5.03	52.73±4.53	57.64±4.95

TABLE VI
THE ACCURACY (%) OF MFCC AND MFCC+GRBAS

Classification task	/a/$_1$ only	+GRBAS[1]	+GRBAS[2]
Cancer/non-cancer	75.14±4.96	**83.14±5.64**	77.71±4.92
3-class	46.12±7.17	**51.38±3.08**	49.38±3.76
Atrophy/Paralysis	66.15±3.77	67.69±5.76	**71.92±6.09**
Paralysis/BOL	65.58±7.37	66.15±5.10	**69.23±6.49**
BOL/Atrophy	52.32±5.59	**62.86±4.91**	56.79±5.80

[1] GRBAS based on the sustained /a/.
[2] GRBAS based on each patient's overall quality.

REFERENCES

[1] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, *et al.*, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2017.

[2] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, *et al.*, "An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification," *J. Voice*, vol. 31, no. 1, 113.e9–113.e18, 2017.

[3] S.-H. Fang, Y. Tsao, M.-J. Hsiao, *et al.*, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[4] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41 034–41 041, 2018.

[5] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, *et al.*, "Voice pathology detection and classification using convolutional neural network model," *Applied Sciences*, vol. 10, no. 11, p. 3723, 2020.

[6] H.-C. Hu, S.-Y. Chang, C.-H. Wang, *et al.*, "Deep learning application for vocal fold disease prediction through voice recognition: Preliminary development study," *J. Medical Internet Research*, vol. 23, no. 6, e25247, 2021.

[7] E. C. Compton, T. Cruz, M. Andreassen, *et al.*, "Developing an artificial intelligence tool to predict vocal cord pathology in primary care settings," *The Laryngoscope*, 2022. DOI: https://doi.org/10.1002/lary.30432. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lary.30432.

[8] B. Woldert-Jokisz, *Saarbruecken voice database*. Saarbrücken, Germany: Institut für Phonetik, Universität des Saarlandes, 2007.

[9] Massachusette Eye Ear Infirmary, *Disordered Voice Database (Ver. 1.03)*. Lincoln Park, NJ, USA: Kay Elemetrics Corp, 1994.

[10] K. Omori, "Diagnosis of voice disorders," *Jap. Med. Assoc. J.*, vol. 54, no. 4, pp. 248–253, 2011.

[11] M. Müller and S. Ewert, "Chroma Toolbox: Matlab implementations for extracting variants of chroma-based audio features," in *Proc. 12th Int. Soc. Music Information Retrieval Conf.*, (Miami, United States), Miami, United States: ISMIR, Oct. 2011, pp. 215–220.

[12] B. McFee and et al., *Librosa/librosa: 0.10.0*, version 0.10.0, Feb. 2023. DOI: 10.5281/zenodo.7657336. [Online]. Available: https://doi.org/10.5281/zenodo.7657336.

[13] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," *Applied Acoustics*, vol. 158, p. 107 050, 2020, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2019.107050.

[14] J. Milgram, M. Cheriet, and R. Sabourin, ""One against one" or "one against all": Which one is better for handwriting recognition with SVMs?" In *10th Int. work-*

shop on Frontiers in Handwriting Recognition, Suvisoft, 2006.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[16] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CoRR*, vol. abs/1610.02357, 2016.

[19] A. G. Howard, M. Zhu, B. Chen, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[20] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *CoRR*, vol. abs/2104.00298, 2021.

[21] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *20th Conf. Oriental Chapter of the Int. Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5. DOI: 10.1109/ICSDA.2017.8384449.

[22] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[23] W.-S. Lee and E. Zee, "Standard Chinese (Beijing)," *J. Int. Phonetic Assoc.*, vol. 33, no. 1, pp. 109–112, 2003.

[24] Z. Liu, M. Huckvale, and J. McGlashan, "Automated voice pathology discrimination from continuous speech benefits from analysis by phonetic context," in *Proc. Interspeech*, 2022, pp. 2158–2162.

[25] S. Hidaka, Y. Lee, M. Nakanishi, K. Wakamiya, T. Nakagawa, and T. Kaburagi, "Automatic GRBAS scoring of pathological voices using deep learning and a small set of labeled voice data," *Journal of Voice*, 2022, ISSN: 0892-1997. DOI: https://doi.org/10.1016/j.jvoice.2022.10.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199722003472.

Detecting the Underwater Distance and Swimming Direction of Tilapia using YOLO

Yi-Zeng Hsieh*, Chia-Hsuan Wu and Cheng-Hou Chou
Department of Electrical Engineering, National Taiwan University of Science and Technology Taipei City, Taiwan

yzhsieh@mail.ntust.edu.tw

Abstract— **The article introduces a groundbreaking system designed to revolutionize underwater operations involving the recognition of tilapia fish by underwater unmanned vehicles (UUVs). This innovative technology aims to replace the traditional method of employing human divers, offering a safer and more cost-effective approach to underwater tasks. By harnessing the capabilities of YOLOv3, this system achieves real-time image recognition, enabling the determination of both the distance and direction of tilapia fish with remarkable speed and accuracy.**

Keywords—machine learning, deep learning, underwater technology

I. INTRODUCTION

According to statistics from the Fisheries Agency, the number of people employed in Taiwan's fishing industry in 2018 was 325,940, accounting for approximately 2.86% of the total employed population in Taiwan. The number of employees in Taiwan's fishing industry has been decreasing year by year, and there has already been a shortage of front-line fishing workers. In addition, with the global population increasing and per capita consumption of aquatic products on the rise, the fishing industry is facing serious challenges. Taiwan's total catch has only reached 1 million tons. Despite continuous technological advancements in fishing, production levels have not been able to increase, leading to the vicious cycle of overfishing in global fishing industries. Therefore, many countries have turned to aquaculture as a source of seafood, and Taiwan's aquaculture is largely based on manual methods. To address these problems, the academic and industrial sectors of various countries have begun to develop various intelligent aquaculture systems. For example, the Internet of Things (IoT) can be applied to environmental monitoring in aquaculture, with sensors collecting environmental data and transmitting it to servers to provide users with real-time information.

Automatic feeding can also be achieved through algorithms, reducing inter-fish competition and food waste. This not only lowers costs but also helps to improve water quality and avoid turbidity. With the rapid growth of hardware, high-performance GPUs can be used to execute complex calculations in real-time, driving the further development of machine learning. Therefore, some researchers have proposed using Recurrent Neural Networks (RNN) to identify abnormal behavior in fish. Currently, research on intelligent aquaculture mainly focuses on environmental monitoring and automatic feeding, with little related to underwater operations. In Taiwan, underwater operations are mostly carried out by divers, so this paper proposes using a Remotely Operated Vehicle (ROV) to replace divers for underwater work. The focus of this paper is on identifying the distance and swimming direction of the Tilapia fish based on YOLO recognition. Tilapia is an important cultured fish species in Taiwan, with a production of 63,329 tons in 2017, ranking first in Taiwan's aquaculture and covering an area of approximately 4,052 hectares, second only to sea bass.

Underwater operations have long relied on human divers to perform essential tasks such as fish monitoring, maintenance, and data collection. However, these operations come with inherent risks and significant costs associated with deploying divers and vessels. This article addresses these challenges by introducing a novel system that employs advanced computer vision technology to recognize tilapia fish, one of Taiwan's most important cultured fish species, and their precise locations. The primary objective is to develop a cost-effective and safe alternative to traditional underwater personnel and vessels.

The research highlights several key points regarding Taiwan's fishing industry and the challenges it faces. Here are five potential contributions that can be made based on the information provided:

1.Development of Intelligent Aquaculture Systems: One significant contribution can be the development and implementation of intelligent aquaculture systems in Taiwan. Drawing inspiration from global efforts, Taiwan can invest in research and technology to create advanced systems that incorporate IoT, algorithms for automatic feeding, and real-time data analysis. These systems can help optimize production, reduce costs, and promote sustainable aquaculture practices.

2.Sustainable Aquaculture Practices: Given the dwindling number of front-line fishing workers and concerns about overfishing, Taiwan can focus on sustainable aquaculture practices. By embracing technological advancements, Taiwan can contribute to global efforts in responsible seafood production, ensuring that its aquatic resources are utilized efficiently and with minimal environmental impact.

3. Application of Machine Learning in Aquaculture: Taiwan can make strides in the application of machine learning, such as Recurrent Neural Networks (RNN), to identify abnormal behavior in fish. By developing and deploying AI-powered systems for fish behavior analysis, Taiwan can enhance fish health monitoring and management, ultimately leading to improved production and resource conservation.

4. ROV-Based Underwater Operations: Introducing Remotely Operated Vehicles (ROVs) for underwater operations in Taiwan's aquaculture industry can be a significant advancement. ROVs can replace human divers for tasks such as fish monitoring and maintenance, making underwater operations safer, more efficient, and cost-effective. This transition can contribute to the safety and well-being of workers in the industry.

5. Tilapia Fish Farming Enhancement: Focusing on the specific case of Tilapia fish, Taiwan can lead research and innovation to optimize its production. By utilizing YOLO

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

recognition technology to identify the distance and swimming direction of Tilapia fish, the industry can improve farming techniques, reduce waste, and enhance the overall quality of Tilapia production. This approach can serve as a model for other fish species in Taiwan's aquaculture sector.

In summary, Taiwan has the opportunity to play a pivotal role in the development of intelligent and sustainable aquaculture practices, benefiting both its economy and the global seafood industry. By embracing cutting-edge technology and innovative approaches, Taiwan can contribute to the long-term health and viability of its fishing and aquaculture sectors.

II. Related works

With the advent of the deep learning era, deep image analysis has become a popular topic of discussion. Since 1949, the first step towards machine learning was opened by the mechanism of neural psychology-based learning, but machine learning algorithms require a large number of iterations to produce training results, require a lot of computing power, and have a high dependence on hardware devices. In the early stages of machine learning development, the efficiency was relatively high in cost.

However, with the rapid growth of computer hardware in recent years, the computing power of computers has significantly improved, which has led to the flourishing development of neural network technology and has been once again promoted by many scholars. The architecture of computer vision systems has become increasingly sophisticated, and with the improvement of deep perception capabilities, images can further enhance recognition accuracy, making it a key technology for deep image analysis in recent years. With the improvement of hardware, many scholars have applied deep learning to an important issue in computer vision, object recognition.

[1] Using the Selective Search algorithm, the input image generates 2,000 to 3,000 candidate regions, then extracts the features of the candidate regions through Convolutional neural network (CNN), inputs the features into SVM (Support Vector Machine) for classification, and finally adjusts the bounding box through a linear regression model. [2] Improved R-CNN, in which traditional R-CNN extracts features for each candidate region, while fast R-CNN produces a feature map after passing through Convolution layer and Max Pooling. Then, it extracts the feature vector from the feature map through RoI Pooling, and finally sends the feature vector into the fully connected layer, and is divided into two outputs, one for classification using softmax and the other for adjusting the bounding box using a regression model. Both R-CNN and fast R-CNN need to generate candidate regions through the Selective Search algorithm, while [3] proposes faster R-CNN, which does not need to generate candidate regions in advance. Instead, the image is input and passed through the Convolution layer to generate a feature map. Then, the feature map is input into RPN (Region Proposal Network), whose output is the bounding box and the probability of the bounding box containing an object, and then through RoI Pooling similar to fast R-CNN, each region is classified, and the precise bounding box is found.

The key research problem highlighted in these sentences revolves around the efficiency and accuracy of object detection algorithms, particularly in the context of computer vision and image processing. Specifically, the sentences discuss various approaches to object detection, and the problem can be summarized as follows:

Improving Object Detection Efficiency and Accuracy: The central research problem is to enhance the efficiency and accuracy of object detection in images. Traditional methods like Selective Search generate a large number of candidate regions, which can be computationally expensive and may lead to inefficiencies. The challenge is to develop methods that can accurately detect objects while minimizing the number of candidate regions generated, thus improving computational efficiency.

The evolution of object detection algorithms, from R-CNN to fast R-CNN and finally to faster R-CNN, highlights the ongoing effort to address this problem. The goal is to extract meaningful features from images and efficiently classify and locate objects within them, ultimately reducing the reliance on time-consuming candidate region generation processes like Selective Search. Faster R-CNN, in particular, represents an attempt to eliminate the need for pre-generated candidate regions altogether, which is a significant step towards more efficient and accurate object detection.

III. Method

This article develops a system that recognizes the distance and direction of tilapia fish captured by underwater unmanned vehicles (UUVs) to replace divers for underwater operations, thus avoiding danger and reducing the cost of employing underwater personnel and vessels. YOLOv3 is

used for the development of the model, and its fast characteristics enable real-time recognition of images captured by the UUV and the determination of the distance and direction of the tilapia fish. The main research methods are divided into: (1) UUV device body; (2) data set collection; and (3) tilapia fish direction and distance model training.

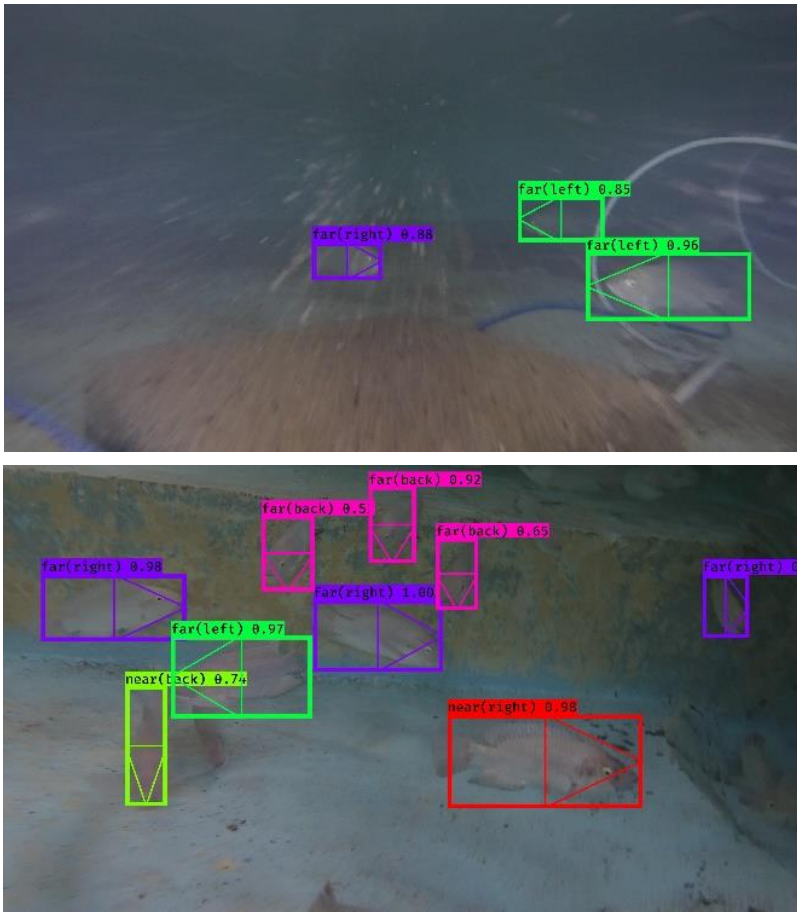


Fig.1 Experiment results

- UUV Device Body Development: The first phase involves the creation and optimization of the UUV device itself. This encompasses the selection of appropriate hardware and software components, ensuring that the UUV is equipped to capture high-quality underwater images and transmit data in real-time.
- Data Set Collection: To train the recognition model effectively, a comprehensive dataset of underwater images containing tilapia fish in various orientations and distances is collected. This dataset forms the foundation for the system's machine learning capabilities, allowing it to recognize and differentiate between different fish positions and distances.
- Tilapia Fish Direction and Distance Model Training: The core of the system's functionality lies in its ability to accurately determine the direction and distance of tilapia fish. This is achieved through the training of a sophisticated machine learning model, powered by YOLOv3. YOLOv3's speed and accuracy make it an ideal choice for real-time image recognition. The model is trained to identify tilapia fish within the collected dataset and provide precise information about their locations relative to the UUV.

A. UUV Device Body

To be able to move quickly and flexibly in the water, a flexible ROV, the Power Ray from Power Vision, was chosen as it has excellent maneuverability and low power consumption, with a working time of up to 4 hours underwater. It can completely replace underwater workers for long-term underwater work, making it the most suitable UUV for this system.

B. Data Set Collection

The dataset used in this article is composed of tilapia fish images captured by the ROV, in the breeding grounds of tilapia fish at the Gongliao Aquatic Research and Conservation Center shown in Fig.1. The video captured by the ROV was cut into photos, and some of the photos were either blurry or without fish, making them unsuitable. Therefore, there are about 5000 tilapia fish images, and labeling was used to mark the distance and direction of the fish in the images.

C. Tilapia Fish Direction and Distance Model Training

To be able to recognize the direction and distance of tilapia fish in real time, YOLOv3 was chosen as the prediction algorithm, which has top performance in terms of accuracy and time. This deep learning network can be divided into two parts: Darknet-53, which is the main architecture for feature extraction, and the YOLO layer architecture, which allows YOLOv3 to predict at multiple scales.

1) Darknet-53

This is the first to seventy-fifth layer of the entire YOLOv3 architecture, with a total of 53 convolution layers and the rest are residual layers shown in Fig.2. The network is mainly composed of a series of 11 and 33 convolutional layers, and its performance is better than resnet-101 and its efficiency is 1.5 times that of resnet-101. Compared with resnet-152, the effect is not much different, but the efficiency is twice as much.

2) YOLO Layer Architecture

It is divided into three scales: small YOLO layer (13*13), medium YOLO layer (26*26), and large YOLO layer (52*52), which produce different feature maps at each scale. Then, the feature maps of different scales are merged with filters of different sizes (3*3 and 1*1), and the final output feature map has a depth of 39 shown in Fig.3. [4] Mentions that the use of anchor boxes can effectively help with object detection, and they used K-Means to obtain 9 sets of anchor boxes. However, these 9 sets of anchor boxes such as (137,163), (192,294), (250,118), (329,148), (342,355), (437,200), (568,592), (594,258), (905,363), are not suitable for our own dataset, so we used K-Means for our dataset to obtain anchor boxes that are more suitable for our purposes.

IV. Experiment Result

At the beginning of the experiment, the video or photo of tilapia captured by the ROV can be inputted into the YOLO model that has completed training, and the final output judgment results can be obtained as shown in the figure 1. It can be seen from the figure 1 that the recognition ability of the model trained in this article is good and the accuracy is high when the object being recognized is not obstructed. However, if the object is obstructed, the model will not be able to recognize its swimming direction smoothly. In terms of recognition of distance, there is a problem of simultaneously recognizing both far and close objects in the middle distance. In order to quickly and intuitively determine the swimming direction, we also added arrows pointing to the direction of movement in the recognition results. Applying this system to aquaculture systems can reduce the cost of underwater personnel and avoid dangerous situations. With this system, tilapia swimming direction can be quickly determined from shore.

The developed system represents a significant advancement in underwater operations and aquaculture management. By replacing human divers with UUVs equipped with YOLOv3-based recognition capabilities, the system enhances safety by eliminating the risks associated with underwater personnel. Additionally, it offers substantial cost savings by reducing the need for expensive vessels and diver employment.

V. Conclusion:

This paper introduces an innovative system that leverages YOLOv3 technology to revolutionize underwater operations, particularly in the context of tilapia fish recognition. The combination of advanced hardware, comprehensive dataset collection, and machine learning model training yields a solution that significantly enhances safety and reduces costs in underwater tasks. The potential applications extend beyond tilapia fish recognition, offering a versatile approach to various underwater operations and aquatic resource management challenges. This technology represents a pivotal step towards the future of efficient and cost-effective underwater operations.

Acknowledgment

This paper thanks Prof. Fan-Hua Nan, National Taiwan Ocean University for aquaculture supporting and the Tanhou company, Penghu, Taiwan, to provide the experimental field. This study was partly supported by National Science and Technology Council (NSTC), Taiwan, under NSTC 111-2221-E-011 -162 -MY3 and 111-2221-E-011 -163 -MY3.

References

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. 2014 IEEE Conference on Computer Vision and Pattern Recognition.

[2] Ross Girshick. “Fast R-CNN”. 2015 IEEE International Conference on Computer Vision (ICCV).

[3] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6): 1137-1149.

[4] Joseph Redmon, Ali Farhadi. "Yolov3: An incremental improvement." arXiv 2018, arXiv:1804.02767

Application of Deep Learning Techniques for Thermal Imagery Analysis in Abnormal Identification of Floor Tiles in Heritage Environments

Chen-Xin-Yu [1], Wu-Pei-Chen [1], Chin-Yen-Ju [1], Tsung-Yi Chen [2], Kuo-Chen Li [3], Chiung-An Chen [4], Mei-Ling Chan[5],and Shih-Lun Chen [1, *]

* Department of Electronic Engineering, Chung Yuan Christian University, Chung Li City 32023, Taiwan;
E-mail: s10926107@cycu.edu.tw; s10926111@cycu.edu.tw; s10926204@cycu.edu.tw; chischen@cycu.edu.tw
Tel: +886-3-265-4644

Department of Electronic Engineering, Feng Chia University, Taichung City 407102 , Taiwan;
E-mail: tsungychen@fcu.edu.tw Tel: +886-4-2451-7250#4910

Department of Information Management, Chung Yuan Christian University, Chung Li City 32023, Taiwan;
E-mail: kuochen@cycu.edu.tw; Tel: +886-3-265- 2050

Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 243303, Taiwan
E-mail: joannechen@mail.mcut.edu.tw; Tel: 886-2-2908-9899#4813

School of Physical Educational College, Jiaying University, Meizhou City 514000, Guangdong Province, China;
E-mail: Lynn202207017@jyu.edu.cn

Abstract — **Thermal imaging has gained significant use in recent years, particularly during the epidemic, including its application in architecture for damage detection on archaeological monuments through high-temperature analysis. The non-invasive nature of thermal imaging, along with its ability to visualize temperature levels, allows for problem identification while preserving the building's structure. The integration of artificial intelligence further enhances its potential applications. This study aims to propose an automated inspection system using a convolutional neural network (CNN) for analyzing materials in high-temperature floor blocks. Data for this study are provided by the Bureau of Cultural Heritage, Ministry of Culture (BOCH), Republic of China (Taiwan). In addition, a professor with more than seven years of experience in monument maintenance is responsible for assisting in the identification of overheated data and collecting thermal images of the floor. Analyzing materials at elevated temperatures for restoration purposes offers several benefits: efficient identification of problematic materials, identification of materials suitable for specific environmental conditions through high-temperature analysis, and determining the sequence of temperature variations among different materials to aid in restoration planning. The identification accuracy rate of this study is as high as 92.1%. Compared with the speed of 3600s for professionals to identify 100, this research only needs 3s. The efficiency is increased by about 99.91, which is an amazing improvement. These functions increase the practicality of restoration efforts, improve restoration quality and efficiency, and contribute to academic research on ancient monument preservation.**

I. INTRODUCTION

Thermal imaging has emerged as a crucial tool in various fields, including architecture and archaeological research. By capturing the thermal radiation energy emitted by objects, thermal images provide valuable insights into temperature distributions and thermal properties. These images convert thermal energy into two-dimensional temperature data, with colors representing different temperature levels. The non-invasive nature of thermal imaging makes it particularly suitable for architectural inspections, allowing for the identification of potential issues while preserving the integrity of the structure. Similarly, in archaeological research, thermal imaging aids in locating damage and understanding the thermal characteristics of ancient monuments.

In a related study highlighted in [3], a similar approach was employed to assess and repair bridge concrete using thermal images. Abnormal temperature distributions detected in the thermal images were indicative of damaged areas, leading to successful bridge repair. With an accuracy of 80% as reported in the literature, the technique demonstrated the potential of abnormal temperature distribution analysis in thermal images for guiding repairs. Applying a similar approach to identify potential damage and mark out vulnerable areas in ancient site sheet metal can greatly contribute to restoration efforts by

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

providing valuable insights into the condition of the site. To specifically address the analysis of floor materials, the project focuses on processing the thermal images to remove non-objective objects that could interfere with an accurate analysis. Drawing inspiration from the powerful CNN technique mentioned in [4], which has proven successful in various applications such as medicine and face recognition, the project leverages CNN's capabilities in image recognition and target identification through feature extraction. The CNN architecture allows for the extraction of relevant features from thermal images, enabling the identification and classification of materials based on their thermal characteristics.

However, retraining the CNN model from scratch can be time-consuming. To enhance execution efficiency, the project adopts a strategy known as migration learning, as outlined in [5]. This approach involves training the AlexNet CNN model, a widely used architecture, by adapting the pre-trained model to the specific requirements of the project. By modifying specific link layers, adjusting the number of recognition types, and providing extensive training and validation datasets, the project achieves optimal performance while reducing the training time and computational resources required. The results reported in the literature further validate the effectiveness of migration learning, with recognition accuracy reaching an impressive 99% and demonstrating the versatility of the approach in creating different models with varying training sets and parameter configurations. In this research, the trained AlexNet CNN model is employed to identify materials on demand. By accurately identifying and quantifying the percentage of vulnerable materials, maintenance methods can be appropriately planned and executed, ensuring the longevity and preservation of the architectural elements. Furthermore, the analysis of suitable materials based on their thermal characteristics adds valuable insights to the restoration process, enabling informed decisions regarding material selection and facilitating the restoration of ancient monuments to their original splendor.

In summary, thermal imaging offers unparalleled visibility and non-invasiveness, making it an invaluable tool for architectural inspections and archaeological research. By adopting a similar approach to the assessment of bridge concrete, the project aims to explore the potential damage in ancient site sheet metal using thermal imaging techniques. Leveraging the power of CNN and migration learning, the project seeks to develop an automated system capable of accurately identifying and quantifying vulnerable materials. This information will assist in the planning of restoration methods and the analysis of suitable materials, ultimately enhancing the quality and efficiency of restoration efforts.

II. Methods

This research aims to leverage artificial intelligence to rapidly detect high-temperature areas in thermal images of historic sites, enabling professionals to promptly assess the condition of the sites and identify areas in need of repair. Furthermore, this technology can serve as a preventive measure by identifying excessively smooth areas that may pose safety risks. The research encompasses three main components: image processing, image segmentation, and CNN recognition. The flow chart illustrating the research process is depicted in Fig. 1.

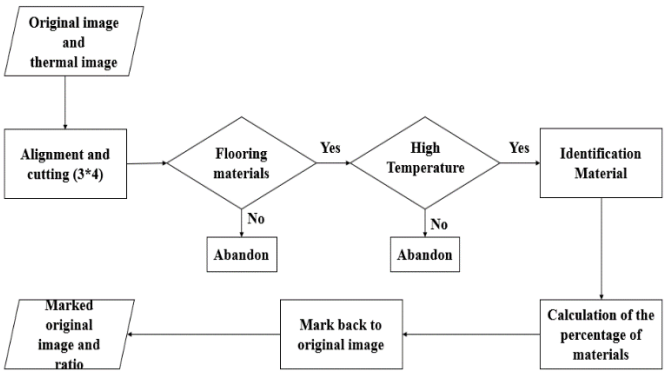


Fig. 1 This research process.

A. Image proofreading

To ensure a seamless alignment between the vulnerable regions identified in the thermal image and the corresponding areas in the original image, it is crucial to crop both images to the same dimensions. This process involves centering the images and removing any excess portions, as illustrated in Fig. 2 and Fig. 3.

In the first step, shown in Fig. 2, the centers of the thermal image and the original image are aligned. This alignment establishes a common reference point for subsequent analysis and marking of the vulnerable areas. By centering the images, we ensure accurate mapping and facilitate the integration of the thermal image's findings with the original image. Once the centers are aligned, the excess parts of both images are cropped, as depicted in Fig. 3. This cropping procedure ensures that both images share the same size and removes any unnecessary areas. The resulting images can then be seamlessly combined, enabling a smooth and accurate representation of the vulnerable regions within the context of the original image.

By implementing this cropping technique, the research aims to enhance the visual coherence and precision in marking the vulnerable parts of the thermal image onto the corresponding areas in the original image, aiding in the assessment and evaluation of the site's condition.

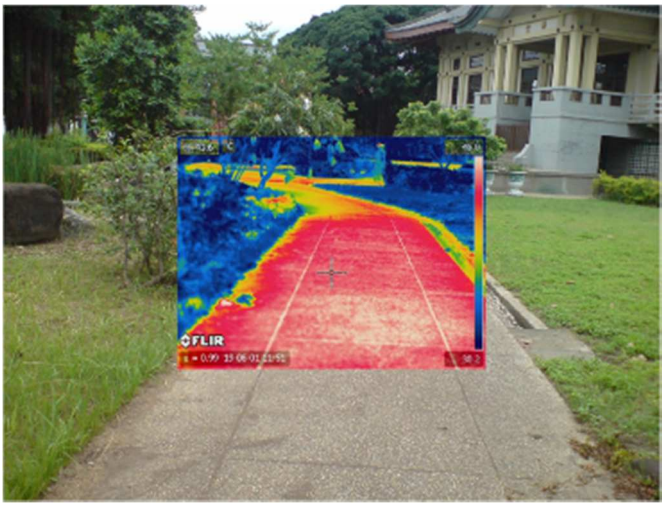


Fig. 2 Center Point Alignment



Fig. 3 Cutting results.

B. Image cutting

Since the floor block street is located in the lower half of the image, this study decided to divide the raw and thermal images into 12 smaller images to form a grid of 3 rows and 4 columns, which can effectively help to identify the high temperature status of the floor area later.

In addition to this by splitting the images into smaller ones, a more detailed and comprehensive dataset can be generated to train the model and improve the drawback of the study itself which has not many sources of images. And the increased amount of data enables more robust and accurate thermal pattern analysis and helps to identify weak regions. Fig. 4 and Fig. 5 visually show the 12 small images resulting from the segmentation of the original and thermal images. Each image represents a different region of interest in the whole image, enabling a more refined analysis.

The aim of the total study is to improve the accuracy of the analysis by utilizing the enhanced dataset obtained from the segmented images. This will improve the ability to identify and assess vulnerable areas within historic sites based on thermal imaging data.



Fig. 4 The original picture is cut into 3*4 sheets.

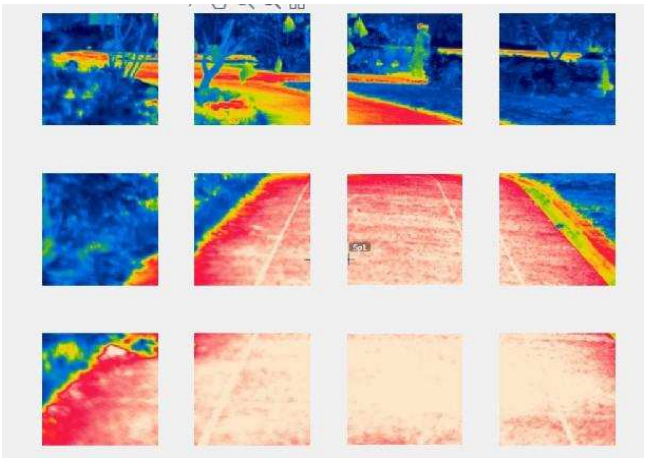


Fig. 5 Thermal image cut into 3*4 sheets.

C. Defining model recognition objects and training sets.

The CNNs for material identification are designed based on the functional requirements, specifically targeting four materials: standard brick, asphalt, red brick, and concrete. RGB images of these materials are classified and organized into a training set.

Fig. 6 visually presents the arrangement of the training set, where the identified materials are categorized and stored for training the CNNs. This process enables the networks to learn the distinctive features and characteristics of each material, facilitating accurate material identification in subsequent analysis. By tailoring the CNNs to address these specific materials and leveraging a well-structured training set, this research aims to enhance the precision and reliability of material recognition. This approach contributes to the

advancement of robust techniques for material identification and analysis within the context of historic sites, providing valuable insights for conservation and restoration efforts.



standard brick | asphalt | red brick | concrete

Fig.6 Four types of identification materials.

D. CNN Object Recognition Training

The training process for the AlexNet model is conducted using MATLAB's integrated deep learning network program. The material-aware floor model is represented by four distinct types, which are inputted into the program's Data section after completion. To augment the dataset and enhance training accuracy, various techniques are employed, such as rotating the images within the training set and adjusting the parameters in the training block. Through several iterations and experimentation, the optimal training parameters have been determined, as outlined in Table I and Table II. These parameter adjustments play a crucial role in fine-tuning the training process and improving the overall accuracy of the model. By systematically refining the training parameters, the research endeavors to optimize the performance of the AlexNet model and ensure its effectiveness in material recognition and analysis. The utilization of MATLAB's deep learning network program and the systematic parameter adjustments contribute to the robustness and reliability of the training process, enhancing the model's ability to accurately identify materials in the context of the floor analysis.

Table I. Rotation parameters of x-axis

	Min	Max
Random rotation	-90	90
Random rescaling	1	2

Table II. Training Parameters

InitialLearnRate	0.0001
ValidationFrequency	4
MaxEpochs	8
MiniBatchSize	10

E. Mark back to the original image

(i). Calculation method and encoding

The corrected image will be saved as 640*480 pixels and then each small piece of the image will be cut equally into 160*160 pixels. The method is described as follows, take the coordinates x=0 and y=0 of the upper left corner of the image as the starting point, increase x to the right, increase y to the bottom, divide the length (x direction) into four equal parts, divide the width (y direction) into three equal parts, and the cutting points in the x direction are x=160, x=320, and x=480, and the cutting points in the y direction are y=160, and y=320. In order to find the cutting images that are the same symptom as that in the original image, it is necessary to number the cutting images. of the cut pictures, the cut pictures need to be numbered. Based on the number of images entered, the cuts are named alphabetically, and then cut into 3*4 images, numbered from top left to bottom right as 1 to 12. Finally, the code can be used to retrieve the location from the thermal image that corresponds to the original image.

(ii). Matching the thermal image to the original map

By applying the established coding methodology, the blocks identified in the thermal image are accurately matched back to their corresponding locations on the original map. The matching process enables the precise localization of high-temperature symptomatic areas. These identified blocks are visually highlighted by encircling them with bounding boxes, providing a clear visual representation of the regions exhibiting elevated temperature symptoms. This visual marking aids in distinguishing and analyzing the areas that require attention and potential restoration interventions. The systematic matching of the thermal image to the original map, along with the prominent highlighting of selected blocks, facilitates effective analysis and assessment of high-temperature phenomena within the floor structure.

F. Calculated Material Share

In this study, the material composition analysis involves calculating the percentage of each material within the high-temperature area. This calculation is performed by considering the total number of sheets identified as high temperature as the denominator. The numerator represents the number of sheets of each material detected within the high temperature area using the material Convolutional Neural Network (CNN). To obtain the percentage of each material, the numerator is divided by the denominator and then multiplied by 100, as shown in

$$\frac{(Total\ number\ of\ sheets\ of\ material)}{(Total\ number\ of\ sheets\ of\ high\ temperature\ material)} * 100\%. \quad (1)$$

This approach allows for a quantitative analysis of the distribution and prevalence of different materials within the regions exhibiting elevated temperatures. By calculating the percentage of each material, valuable insights can be derived regarding the material makeup and its relative contribution to the high-temperature areas. Such information is crucial for understanding the composition of the structure, identifying potential material-related issues, and facilitating appropriate maintenance and preservation strategies.

III. Result

A. CNN Accuracy

For the training phase of the material identification model, a comprehensive dataset comprising 1531 material images was employed. This dataset was divided into a training set, consisting of 1072 images, and a validation set, consisting of 459 images. The distribution of training sessions for each item is outlined in Table III, highlighting the number of iterations dedicated to each material during the training process.

In order to assess the performance of the trained models, confusion matrix analysis was used to measure the final accuracy, the confusion matrix will provide a detailed breakdown of the predictive accuracy of the models, enabling a comprehensive assessment of the effectiveness of the various models applied in this study. The four functions that make up the confusion matrix are TP (correctly predicted successful samples), TN (correctly predicted successful negative samples), FP (incorrectly predicted negative samples), and FN (incorrectly predicted negative samples), and the accuracy rate is calculated from these four elements, as shown in

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \quad (2)$$

As shown in Tables IV and V, the accuracy of the model in recognizing various categories calculated by (2) reaches more than 89%, where the overall accuracy of the model as a whole, also calculated by (2), reaches an impressive 92.1%.The confusion matrix provides a detailed breakdown of the model's predictive accuracy, enabling a comprehensive assessment of its effectiveness in correctly identifying different materials within the thermal images. The high accuracy rate obtained by the model demonstrates its robustness and reliability in accurately recognizing and classifying materials based on their thermal signatures. This result signifies the effectiveness of the training process and the successful adaptation of the Convolutional Neural Network model to the specific requirements of material identification in the context of this study.

Table III. Number of training sheets by material.

	standard brick	asphalt	concrete	red brick
Training Data	258	140	324	350
Validation Data	110	70	139	150
Total	368	200	463	500

Table IV. Material CNN Confusion matrix

		Target Class				
	Category	Standard Brick	Asphalt	Concrete	Red Brick	Accuracy
Output Class	Standard Brick	333	1	12	22	90.4%
	Asphalt	1	180	19	0	90%
	Concrete	6	11	443	3	95.7%
	Red Brick	25	1	20	454	90.8%
	Accuracy	91.2%	93.3 %	89.7%	94.8 %	92.1%

Table V Material CNN Accuracy

	Accuracy	Standard Brick	Asphalt	Concrete	Red Brick
Alexnet	92.1%	91.2%	93.3%	89.7%	94.8%

B. Research Results

The output of the analysis produces two images: the first image overlays boxes on the original image to highlight areas exhibiting high temperature symptoms, while the second image presents the corresponding thermal image. Fig.7. showcases the corresponding thermal image obtained from the analysis.

The analysis also provides information regarding the distribution of materials within the high temperature areas. The upper column of the output image displays the percentage of each material present in the identified high temperature region.

Notably, the results depicted in Fig.7 indicate that red brick displays the highest vulnerability in the given environment. Following red brick, standard brick exhibits a lesser degree of vulnerability, and finally, the concrete material demonstrates the least susceptibility to high temperature effects. By identifying the most vulnerable materials, such as red brick, appropriate maintenance and preservation strategies can be developed to mitigate potential risks and ensure the longevity of historic sites. The findings the significance of utilizing thermal imaging and material identification techniques to prioritize areas in need of attention and implement targeted interventions to safeguard vulnerable materials.

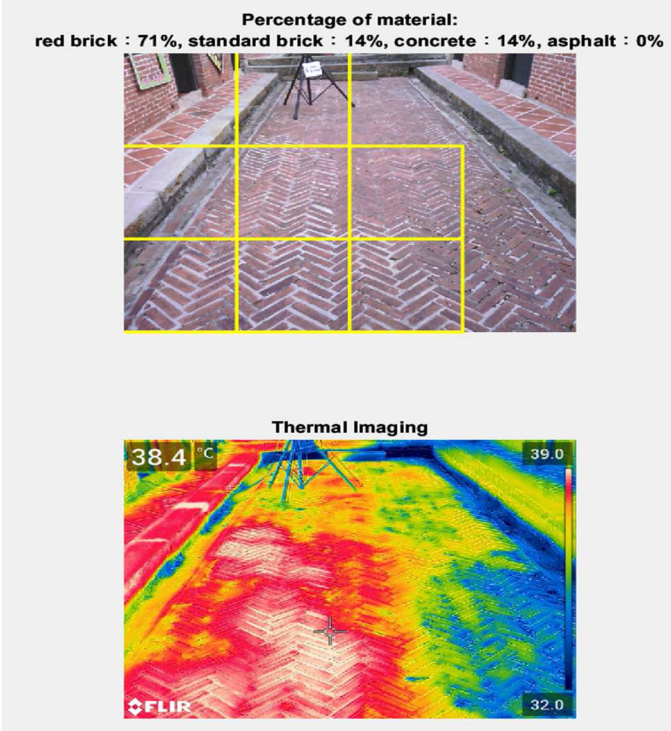


Fig.7 The actual use of the results of this study

In addition, the results of this study were compared with the results of human identification in terms of time and accuracy in recognizing 100 images, with the assistance of a professor with more than seven years of experience in monument maintenance. The results revealed that while manual recognition achieved a slightly higher accuracy rate of approximately 3%, the proposed method demonstrated a remarkable improvement in speed, requiring only 3 seconds to recognize the entire set of images. This represents an impressive enhancement of approximately 99.16%. The experimental findings indicate that the proposed approach significantly contributes to time-saving and enhances overall work efficiency, as illustrated in Table VI.

Table VI Material CNN Accuracy

	Manual Judgment	**This research**
Accuracy	95.8%	92.1%
Elapsed time(s)	3600	3

IV. CONCLUSIONS

This research has highlighted the impact of rising temperatures on historical sites and emphasized the importance of proactive measures to mitigate damage. The study focused on detecting high-temperature areas within floor materials, specifically red bricks, concrete floor tiles, asphalt, and square bricks. The research also underscored the significance of efficient material identification and analysis for selecting suitable materials in varying environmental conditions, aiding restoration efforts. Additionally, it is worth noting that the research achieved a remarkable recognition accuracy rate of 92.1%. This high level of accuracy signifies significant success in accurately identifying and classifying materials based on their response to high temperatures. The achieved accuracy rate demonstrates the effectiveness of the developed methodology and the reliability of the material recognition system. It is worth noting that the amount of training data varies greatly, and it is easy to make mistakes in judgment. To enhance the research's comprehensiveness and reliability, future collaborations with the Cultural Affairs Bureau are planned to obtain additional samples.

Furthermore, this research extends its implications beyond historical sites. The methodology employed can be applied to other material types used in road construction, contributing to the development of resilient infrastructure. Expanding the research's scope has the potential to benefit broader contexts by addressing the challenges posed by increasing temperatures and promoting sustainable practices in construction and maintenance. Through further collaborations and research, these findings can be applied to preserve cultural heritage and foster the development of sustainable infrastructure.

ACKNOWLEDGMENT

The authors gratefully thank the Bureau of Cultural Heritage, Ministry of Culture of Taiwan for its funding and equipment support.

REFERENCES

[1] Kordatos, E. Z., Exarchos, D. A., Stavrakos, C., Moropoulou, A., & Matikas, T. E. (2013). Infrared

thermographic inspection of murals and characterization of degradation in historic monuments. *Construction and Building Materials*, *48*, 1261–1265. https://doi.org/10.1016/j.conbuildmat.2012.06.062

[2] ElMasry, G., ElGamal, R., Mandour, N., Gou, P., Al-Rejaie, S., Belin, E., & Rousseau, D. (2020). Emerging thermal imaging techniques for seed quality evaluation: Principles and applications. *Food Research International*, *131*, 109025. https://doi.org/10.1016/j.foodres.2020.109025

[3] 弘一川西, 詳悟林, 和明橋本, 勲氏家, & 邦釘全. (2020). 赤外線サーモグラフィ法における損傷自動判別技術. *Ai・データサイエンス論文集*, *1*(J1), 382–391. https://doi.org/10.11532/jsceiii.1.J1_382

[4] Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018). Convolutional Neural Network (CNN) for Image Detection and Recognition. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 278–282. https://doi.org/10.1109/ICSCCC.2018.8703316

[5] Almisreb, A. A., Jamil, N., & Din, N. M. (2018). Utilizing AlexNet Deep Transfer Learning for Ear Recognition. *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 1–5. https://doi.org/10.1109/INFRKM.2018.8464769

[6] *A Review of Convolutional Neural Networks | IEEE Conference Publication | IEEE Xplore*. (n.d.). Retrieved June 8, 2023, from https://ieeexplore.ieee.org/document/9077735

[7] Preda, A., & Scurtu, I. C. (2019). Thermal image building inspection for heat loss diagnosis. *Journal of Physics: Conference Series*, *1297*(1), 012004. https://doi.org/10.1088/1742-6596/1297/1/012004

Wavelet and Cutout in YOLO Architecture for Road Pothole Detection

Shao-Hua Lu [1], Jia-Teng Lu [2], Szu-Yin Lin [1], and Chih-Hsien Hsia [1,3*]

[1] Department of Computer Science and Information Engineering, National llan University, Taiwan

[2] Kang Chiao International School, Taiwan

[3] Department of Business Administration, Chaoyang University of Technology, Taiwan

[*]E-mail: hsiach@niu.edu.tw

***Abstract* — The problem of road potholes caused by vehicle pressure, plate movement, and thermal expansion and contraction must be overcome for road safety and comfortable driving. Many experts have proposed solutions to this problem, including image analysis, sensors that can detect vibrations caused by potholes, and other tactics. Deep learning (DL) has been utilized widely in several sectors in recent years, including object identification, which can be used to find potholes in the road. In order to preprocess the information and remove less important features, which will increase average precision (AP), this study suggests using the you only look once (YOLO) series of models in combination with data augmentation (DA) and discrete wavelet transform (DWT) features. According to experimental results, when combined with DWT and Cutout-based DA, trained using two different publicly available pothole datasets, and then objectively compared, our suggested method outperforms several other model designs in the task of pothole detection, with an improvement of 2.5% in AP. This illustrates the study's effective improvement in detection accuracy and real-time performance.**

I. INTRODUCTION

In today's society, people rely on transportation for convenience, and the traffic burden on roads is increasing. Due to the weight of moving cars and the influence of the outside climate, asphalt pavement on roadways is susceptible to developing potholes. However, the size and depth of potholes are important indications that affect drivers' safety. Severe cases might result in traffic accidents, while moderate ones can cause passengers' discomfort. When a pothole detection system is placed in a car, it can assist drivers in assessing road conditions. This enables them to slow down or change lanes ahead of time to avoid potholes, enhancing passenger safety and comfort. Drivers must concentrate on driving while on the road, thus this study will focus on detecting potholes on the road, which informs drivers about potholes ahead.

Many research projects have proposed pothole detection solutions in recent years. Kim *et al*. [1] classified three approaches to road pothole detection: vibration-based systems, 3D reconstruction, and computer vision (CV). The authors' first method, vibration-based systems, uses sensors [2] they designed to determine whether a vehicle has driven over a pothole. Its benefits include the fact that it does not require a big quantity of data storage and is cost-effective. However, there are three major disadvantages. For instance, because the sensors are attached to the vehicle, they can only detect potholes when the tires contact them, limiting the detection range. Second, the large vibrations that trigger the sensors can be generated by anything from potholes to road gaps or manhole covers, resulting in false positives. The second method relies on 3D reconstruction. According to [3], the authors employed a 3D laser scanning system to detect the state of road potholes. The device emits and absorbs reflections from an invisible light spectrum in order to reconstruct a 3D image known as a point cloud. In Fig. 1, the point cloud contains information such as distances on the *x*, *y*, and *z* axes, and the scanning speed is 5000 points per second with a 360° x 60° field of view. Drivers can use the point cloud data to detect road potholes through analysis and aggregation. Although this method gives excellent accuracy, high-precision laser scanning equipment can be quite expensive. The third method is to detect potholes using CV. For pothole identification, Rahman *et al*. [4] proposed utilizing the single shot multibox detector (SSD) [5], a deep learning (DL) technique. To achieve lightweight implementation, they replaced the VGG16 design in the front end with the MobileNet architecture [6], as shown in Fig. 2.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

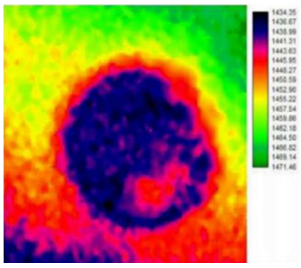


Figure 1. Pothole heatmap [3].

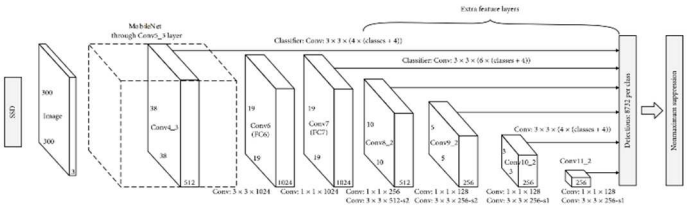


Figure 2. SSD MobileNet model architecture [5].

According to Bochkovskiy *et al*. [7] on you only look once (YOLO)v4, as shown in Fig. 3, the backbone portion is used for feature extraction. The neck portion is utilized to collect features from several scales, while the head portion is used for localization and categorization. Object identification algorithms can be classified as two-stage (such as R-CNN) or one-stage (such as YOLO and SSD). The primary distinction between the two is that two-stage algorithms select region proposals and then perform feature extraction and classification on those regions. One-stage approaches, on the other hand, perform both localization and classification at the same time. Girshick *et al.* [8] suggested a region-convolutional neural network (R-CNN) object detector. First, the selective search algorithm selects 2000 candidate regions in an image. Then, on each potential region, feature extraction is conducted, and support vector machines (SVM) are utilized for classification, along with bounding box (BBox), which refine the target's position. However, R-CNN requires feature extraction for each candidate region, resulting in a longer calculation time. To address this issue, Girshick *et al*. [9] established fast R-CNN, which performs feature extraction on the input picture only once to obtain a feature map, to improve on R-CNN. Then, using fully connected layers, the candidate regions are mapped onto the feature map. The regions of interest (ROIs) in fast R-CNN are the same as those in R-CNN. The authors used ROI pooling to convert each ROI's image into a fixed-scale feature map, and Softmax was used instead of SVM for classification. Ren *et al*. [10] later combined the region proposal network (RPN) with the fast R-CNN design, resulting in faster R-CNN.

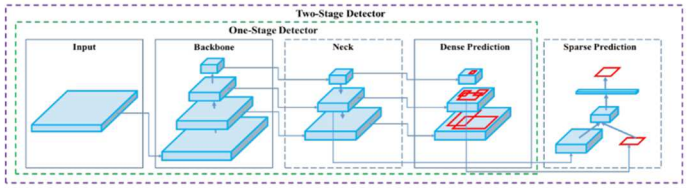


Figure 3. Object detection architecture [7].

YOLOv4 [7] promoted integrating the original backbone network with the cross stage partial network (CSPNet) [11] to form CSPDarknet53. YOLOv4 replaced the feature pyramid network (FPN) in the neck region with the path aggregation network (PAN). This combination improves network speed and accuracy while lowering hardware requirements. In contrast, YOLOv5 [13] added the focus structure to the backbone, which splits the incoming image. It employs two distinct CSPNets in the backbone and head. YOLOv5 does not require the explicit definition of anchor boxes; instead, the model learns them based on the training data. Its key advantage is that it offers four alternative sizes of training models for customers to pick from, letting them customize the network depth to their unique requirements. The smallest model, YOLOv5s, has only 27 MB of parameters, a 90% reduction in parameter count over YOLOv4. In [13], on the other hand, presented YOLOR, which describes human learning and comprehension in terms of the auditory, visual, and tactile senses, as well as learning from previous experiences. Visual and auditory information, which constitute explicit knowledge, can be grasped through regular learning. Conversely, the ability to draw conclusions and learn from the past depends on implicit knowledge or subconscious understanding. YOLOR combines explicit and implicit knowledge encoding, as seen in Fig. 4.

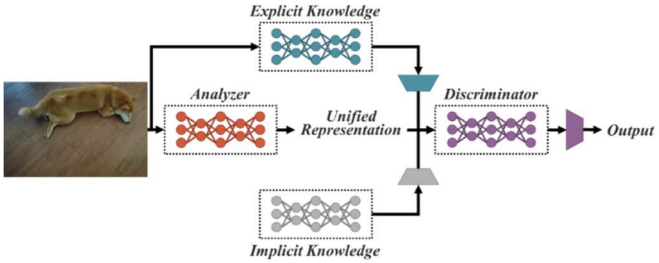


Figure 4. The integration of implicit and explicit knowledge processing [13].

SSD [5] network structures come in two sizes: SSD 300 and SSD 512. The primary distinction is the size of the input image (300x300 *vs*. 512x512), with both variants employing the VGG16 neural network for feature extraction. SSD is capable of handling multiple scales and sized feature maps. The feature maps are larger in the early shallow layers of the CNN, and pool with a stride of 2 is used in the succeeding convolutional layers to lower their size. For detection, both shallow and deep feature maps are used. It can be used to find tiny item targets

in the shallow levels of the network. When dealing with larger targets, the bounding box becomes too narrow to effectively encompass the target. As a result, in the network's deep layers, the feature maps are pooled, lowering their size while increasing the receptive field. As a result, the network can identify larger targets with more accuracy.

According to [7], YOLOv4 falls behind SSD in terms of real-time speed, although having a higher AP. YOLOv5, on the other hand, provides AP similar to YOLOv4 [14], but with improved real-time performance. Furthermore, YOLOR [15] achieves an AP of about 58% when trained on the publicly available MSCOCO dataset [16]. As a result, this study will examine the performance of YOLOv4, YOLOv5, and YOLOR as object detection algorithms in road pothole detection. The above-mentioned road pothole detection aims to address the issue of high cost. As a result, we decided to use a combination of CV and DL techniques. This research will focus on road pothole identification in this direction. Furthermore, the following issues must be addressed in this research: *1)* A real-time object detection algorithm, *2)* Image preprocessing prior to training to improve detection accuracy, and *3)* DA techniques to increase data diversity because the dataset may not cover all real-world circumstances.

II. Proposed Method

This work presents a YOLO-based object identification technique for training road potholes. The dataset is provided to DWT before training the model, and its results of applying alternative YOLO structures to the training model for road pothole identification are compared. The YOLO structure and DWT are investigated in order to improve average precision (AP). Figure 5 depicts the structure of the road pothole detecting model used in this study.

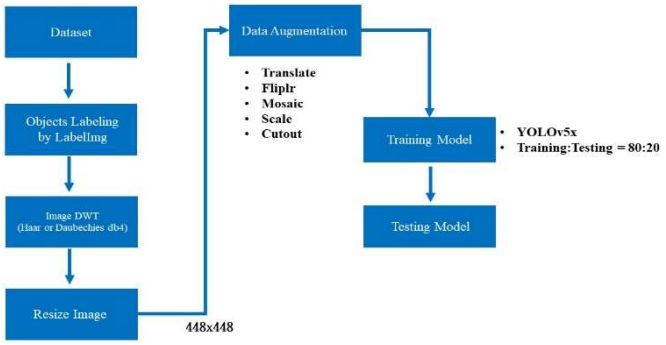


Figure 5. Road pothole detection model system.

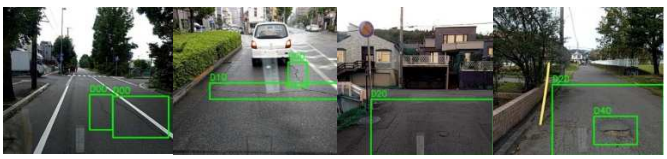
A. Data Preprocessing

We labeled the dataset using the open-source software application LabelImg [17], which allows us to save the labeled data in YOLO format to analyze the annotations. In this research, we used two public datasets: Rahman *et al*. [4] pothole dataset and Arya *et al*. [18] RDD2020 pothole dataset. The usage of these two datasets aims to highlight the model's generalization capacity. Regarding the variety of potholes, the dataset given in [4] is wide. The authors compiled it using smartphone photographs or Google searches. Figure 6 depicts the dataset distribution, which includes 532 photos for training and 133 images for validation. The RDD2020 dataset [18], on the other hand, was obtained by placing a smartphone inside a car to capture photographs from the inside looking out. Road damage is classified into four types: vertical cracks, alligator cracks, diagonal cracks, and potholes. The training set, however, is labeled, while the test set is not. As a result, we utilized only the training set in this study, which was further separated into training, validation, and test sets. In addition, we excluded all other characteristics from the dataset because our focus was on pothole detection. After organizing the dataset, the distribution is as follows: 2275 photos for training, 578 images for validation, and 218 images for testing. The training to testing data ratio is about 80:20, as shown in Fig. 7. The RDD2020 dataset excludes potholes with collected water under rainy weather. However, it contains more image examples than [6]. These two datasets' constraints complement each other, making them acceptable for comparing our model's training experiments.



(a) (b) (c) (d)

Figure 6. Pothole dataset [19]: (a) Pothole in shadow; (b) Pothole with accumulated water after rain; (c) Pothole with interference from various objects near the water; (d) Pothole exposed to sunlight.



(a) (b) (c) (d)

Figure 7. RDD2020 Pothole Dataset [18]: (a) D00 vertical cracks; (b) D10 diagonal cracks; (c) D20 alligator cracks; (d) D40 potholes.

B. Wavelet Transform

DWT is widely used in image analysis [20-21], where the transformation separates the original image into four sub-band frequency domains: low low-band (LL-band), low high-band (LH-band), high low-band (HL-band), and high high-band

(HH-band). When CNN extract features from an image, certain regions of the image can cause a drop in recall, resulting in prediction errors in the trained model. As a result, DWT is applied to the dataset images in this study, and multiple frequency domains are kept for training the model. We conduct tests to determine which frequency domain enhances the trained models' AP. Before training or testing the object detection algorithm, the dataset is processed using DWT preprocessing to eliminate unnecessary information from the images.

DWT provides a variety of mother wavelet functions, such as Haar, Daubechies, Symlets, Coiflets, PyWavelets, and so on. For this study, we use PyWavelets, which supports 38 Daubechies functions (db1 through db38). As a result, we use Daubechies db4 [22] and Haar as wavelet functions. Different wavelet functions are applied to the dataset for DWT under the same parameter settings for the object detection technique, and the models are trained. We examine the influence of wavelet transform as a data preparation strategy on road pothole detection by comparing the AP differences between datasets without wavelet and those processed with different mother wavelet functions.

C. Data Augmentation

Road potholes have variable shapes and are influenced by external influences, making it difficult to compile a comprehensive dataset that includes all sorts or variations of potholes. To improve the learning effectiveness of DL models, DA approaches can change the picture features in the dataset. The object identification methods employed in this study, such as YOLOv4, YOLOv5, YOLOR, and others [12,15,23], often incorporate DA by modifying parameters or disabling certain augmentation strategies. Image attributes that are effective for detecting road potholes include edges, colors, orientations, locations, textures, and more. However, capturing all of these differences in the obtained dataset is difficult. In this study, various strategies are used to enrich the gathered data [24-25] and increase the complexity of the dataset due to the irregular shape, angle-dependent changes, and influence of external influences on road potholes:

1) Scale: Image resizing while retaining original resolution. *2)* Translate: Perform horizontal or vertical image translations. *3)* Rotation: Rotate the images clockwise or counterclockwise around their axis. *4)* Mosaic: The process of creating a new image by merging four separate images. *5)* Cutout: Using a fixed-size rectangular mask to obstruct portions of an image. The first three strategies are standard DA methods that emphasize and improve the presence of useful information in a dataset. In a study [26], it was demonstrated that conducting translation or mirroring DA on training data resulted in lower error rates across multiple algorithm models tested on CIFAR-10 and CIFAR-100 datasets. This study recommends the usage of Cutout [27], a technique that randomly eliminates pixels from the original image and fills them with surrounding pixels. This strategy has been shown to be effective in lowering mistake rates. The Mosaic approach is based on Cutmix [28], as updated by the authors of the YOLOv4 [7]. Cutmix picks a section from one image at random and replaces it with the same region in another image, whereas the Mosaic approach takes four images at random from the dataset, performs region cropping, and merges them into a single image. When compared to Cutmix, using the Mosaic approach can result in improved model correctness.

D. System Design

We compare three object detection methods in this study: YOLOv4 [12], YOLOv5x [23], and YOLOR with version YOLOR-D6 [15]. We use these approaches to train models with the experimental data from [4]. For model training, the TensorFlow version of SSD MobileNetv2 is used. We compare the AP of YOLOv4, YOLOv5, and YOLOR with the YOLOR-D6 variation based on the above. Table 1 displays the model structures. After training the models with a 50% intersection over union (IoU) criterion (Table 2), the performance of all three object detection algorithms exceeds the strategy provided in [4]. Furthermore, the difference in AP between the three approaches is minor. To provide a more precise comparison, we train the models using the dataset from [18], as shown in Table 3, where YOLOv5x has higher AP values for both public datasets. As a result, we use YOLOv5x as the fundamental method for detecting road potholes in this study.

Table 1. Object detection model structure.

Models	YOLOv4	YOLOv5x	YOLOR-D6
Backbone	CSPDarknet53	CSPDarknet53	CSPDarknet53
Neck	PAN, SPP	PAN, SPPF	PAN, SPPCSP
Head	YOLOv3	YOLOv3	YOLOv4
Activation	Leaky, Linear, Mish	Leaky ReLU	Mish, Linear
Optimization	SGD	SGD	SGD
Image size	448x448	448x448	448x448
Batch size	8	8	8
Epochs	300	300	300

Table 2. Comparison of training results on dataset [4].

Detection Methods	AP50
SSD MobileNetv2	65.0%
YOLOv4	78.5%
YOLOv5x	**78.8%**

YOLOR-D6	78.7%

Table 3. Comparison of training results on dataset [18].

Detection Methods	AP50
YOLOv4	54.8%
YOLOv5x	**55.6%**
YOLOR-D6	53.0%

III. EXPERIMENTAL RESULTS

This part conducts experiments based on the research method given in Section II. The experimental findings suggest that YOLOv5x outperforms alternative object detection algorithms on two public datasets, as shown in Tables 2 and 3. The two datasets are merged and combined with DWT and DA to increase the model's generalization and AP.

A. Research Environment

To verify whether the proposed algorithm is effective when applied to two public datasets, the work utilized variant road environment data that were captured from public datasets. The experimental environment is established using an Intel Core i11-11400 CPU@2.6GHz, 32GB RAM, Microsoft Windows 10, Nvidia GeForce RTX3060 12GB for display card, SSD is 1TB, and Anaconda3-5.2.0 + Python3.6, Pytorch 1.10.1, CUDA 11.3 + CuDNN8.4.0 are chosen as the software development platform. The two datasets were combined based on the description in Section 2. Since the dataset proposed in [4] lacks a distinct test set, 67 photos with their associated annotation files were removed from the training set to produce a test set. The structure is as follows: The training set contained 465 photos, the validation set contained 133 images, and the test set contained 67 images. Lastly, the two datasets were combined, yielding the following distribution after organization: 2740 photos in the training set, 711 images in the validation set, and 285 images in the test set.

B. Research Analysis

This section discusses numerous aspects of the research analysis. Specifically, the comparison of several wavelet functions, frequency domains, and levels is shown, as well as the training results of mixed datasets employing DWT. In the experiment, Haar and Daubechies db4 wavelet functions were compared. The frequency domains taken into account include LL, HL, LH, and HH. In Table 4, the experimental results show that training the model with a first-order wavelet utilizing a two-dimensional technique gives higher detection rates than training with a second-order DWT. Furthermore, when the two DWT wavelet functions are compared, the Haar wavelet function performs better in terms of preserving the LL and LH bands in the frequency domain, resulting in a 1.2% improvement. This demonstrates that DWT can improve the learning effectiveness of the YOLOv5x model.

Table 4. Training results using a two-dimensional multi-level DWT.

Methods	Precision	Recall	AP50
YOLOv5x	68.5%	56.5%	60.3%
YOLOv5x + Haar-Based LL1、LH1	68.5%	56.2%	61.5%
YOLOv5x + Haar-Based LL1、HL1	64.8%	57.1%	59.8%
YOLOv5x + Haar-Based LL1、HH1	73.6%	53.1%	60.9%
YOLOv5x + Haar-Based LL2、LH2	69.8%	50.5%	57.4%
YOLOv5x + Haar-Based LL2、HL2	63.6%	50.9%	55.3%
YOLOv5x + Haar-Based LL2、HH2	69.8%	50.5%	57.4%
YOLOv5x + Daubechies-Based LL1、LH1	68.5%	54.9%	60.2%
YOLOv5x + Daubechies-Based LL1、HL1	62.7%	57.1%	59.4%
YOLOv5x + Daubechies-Based LL1、HH1	69.5%	54.8%	60.4%
YOLOv5x + Daubechies-Based LL2、LH2	68.1%	52.6%	58.4%
YOLOv5x + Daubechies-Based LL2、HL2	67.4%	50.3%	56.4%
YOLOv5x + Daubechies-Based LL2、HH2	67.0%	51.0%	56.7%

Table 5. Training results using a two-dimensional first-order DWT combined with DA.

Method	Precision	Recall	AP50
YOLOv5x + Cutout	68%	58.5%	62.4%
YOLOv5x + Haar-Based LL、LH + DA	71.6%	54.8%	62.8%
YOLOv5x + Haar-Based LL、HL + DA	69.7%	55.2%	60.1%
YOLOv5x + Haar-Based LL、HH + DA	63.6%	57.0%	59.6%
YOLOv5x + Daubechies-Based LL、LH + DA	70.1%	55.7%	61.7%
YOLOv5x + Daubechies-Based LL、HL + DA	70.8%	56.1%	61.9%
YOLOv5x + Daubechies-Based LL、HH + DA	68.3%	57.0%	61.2%

After conducting experiments and demonstrating that DWT processing of datasets improves the learning of the YOLOv5x model, it was observed that the model gets better learning outcomes when employing a two-dimensional first-order DWT rather than a second-order DWT. Based on the results of this experiment, the previously mentioned Cutout-based DA was implemented in YOLOv5x. Table 5 displays the outcomes produced when the two-dimensional first-order DWT was combined with Cutout. When Cutout is employed without the two-dimensional first-order DWT, the training results show a 2.1% improvement over when the Cutout is not used. The Haar wavelet function surpasses the Daubechies db4 wavelet function in retaining the LL and LH bands in the frequency domain when the two-dimensional first-order DWT and DA techniques are used. Combining Haar DWT and LH + Cutout yields a 0.4% improvement over not utilizing DWT and a 1.3% improvement over using only Haar DWT with LL and LH

bands. Based on these findings, combining DWT and DA improves the learning effectiveness of the YOLOv5x model.

IV. Conclusion

Various YOLO detection algorithms were used to detect road potholes in this study. Based on training results with two separate publicly available datasets, YOLOv5 revealed higher average precision (AP) in detection. This article is based on YOLOv5, which is a single-class classification job for detecting road potholes. It also includes the image pre-processing step, which enhances the focus structure. In addition to YOLO detection models, this study used the DA of Cutout approach and the Haar-based DWT to pre-process the dataset by maintaining the low-low and low-high bands. The experimental findings demonstrated that the proposed model in this study improved AP for road pothole identification. The suggested method has the highest AP, with a 2.5% increase over the model trained without data preprocessing, making it appropriate for road pothole identification tasks.

References

[1] T. Kim and S.-K. Ryu, "Review and analysis of pothole detection methods," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 7, no. 8, pp. 603-608, 2014.

[2] X. Yu and X. Yu, "Vibration-based system for pavement condition evaluation," *Applications of Advanced Technology in Transportation*, pp. 183-189, 2006.

[3] K. Chang, J. Chang, and J. Liu, "Detection of pavement distresses using 3D laser scanning technology," *Computing in Civil Engineering*, pp. 1-11, 2005.

[4] A. Rahman, R. Mustafa, and M. S. Hossain, "Real-time pothole detection and localization using convolutional neural network," *International Conference on Big Data, IoT, and Machine Learning*, pp. 579–592, 2022.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," *European Conference on Computer Vision*, pp. 21-37, 2016.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.

[7] A. Bochkovskiy, C. Y. Wang, and H. Y. Liao, "YOLOv4: optimal speed and accuracy of object detection," *arXiv preprint*, arXiv: 2004.10934.

[8] R. Grishick, J. Donahue, T. Darrel, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", *arXiv preprint*, arXiv: 1311.2524, 2014.

[9] R. Girshick, "Fast R-CNN", *arXiv*: 1504.08083, 2015.

[10] S. Ren, K. He, R. Grishick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks", *arXiv*: 1506.01497, 2017.

[11] C.-Y. Wang, H.-Y. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: a new backbone that can enhance learning capability of CNN," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1571-1580, 2020.

[12] Ultralytics, "YOLOv5", [Online] Available: https://github.com/ultralytics/yolov5.

[13] C.-Y. Wang, I.-H. Yeh, and H.-Y. Liao, "You only learn one representation: unified network for multiple tasks," *arXiv*: 2105.04206, 2021.

[14] YOLOv4 vs YOLOv5 evaluation, [Online] Available: https://pedin024.medium.com/%E5%88%9D%E6%8E%A2yolov5-71f13b4ba78d..

[15] YOLOR Github, [Online] Available: https://github.com/WongKinYiu/yolor.

[16] MS COCO dataset, [Online] Available: https://cocodataset.org/#home.

[17] LabelImg Github, [Online] Available: https://github.com/tzutalin/labelImg.

[18] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "RDD2020: An annotated image dataset for automatic road damage detection using deep learning," *Data in Brief*, vol. 36, no. 107133, pp. 1-7, 2022.

[19] R. Atikur, Annotated potholes image dataset. [Online] Available: https://www.kaggle.com/chitholian/annotated-potholes-dataset.

[20] C.-H. Hsia, J.-S. Chiang, and J.-M. Guo, "Memory-efficient hardware architecture of 2-D dual-mode lifting-based discrete wavelet transform," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 671-683, 2013.

[21] C.-H. Hsia, J.-M. Guo, and J.-S. Chiang, "Improved low-complexity algorithm for 2-D integer lifting-based discrete wavelet transform using symmetric mask-based scheme," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 8, pp. 1201-1208, 2009.

[22] T. Lan, C.-H. Hsia, P.-T. Lai, H.-W. Tseng, and C.-F. Yang, "Memory-efficient very large scale integration architecture of 2D algebraic-integer-based Daubechies discrete wavelet transform," *Sensors and Materials*, vol. 34, no. 9(4), pp. 3623-3636, 2022.

[23] WongKinYou, "YOLOv4 in PyTorch," [Online] Available: https://github.com/WongKinYiu/PyTorch_YOLOv4.

[24] C. Shorten, T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 60, pp. 1-48, 2019.

[25] Imgaug Github. [Online] Available: https: //github.com/aleju/imgaug.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261-2269, 2017.

[27] T DeVries and G W Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint* arXiv: 1708.04552, 2017.

[28] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo and J. Choe, "CutMix: regularization strategy to train strong classifiers with localizable features," *IEEE/CVF International Conference on Computer Vision*, pp. 6022-6031, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Improved One-class Learning for Voice Spoofing Detection

Lixiang Li* † §, Xiaopeng Xue§ †, Haipeng Peng§ †, Yeqing Ren§ † and Mengmeng Zhao§ † ‡
§ Information Security Center, State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
† National Engineering Laboratory for Disaster Backup and Recovery,
Beijing University of Posts and Telecommunications, Beijing, China
‡ Department of Information Science and Engineering, Zaozhuang University, Zaozhuang, China
* Corresponding author. Email: lixiang@bupt.edu.cn

***Abstract*—Most of the studies on one-class learning for voice spoofing detection have poor generalization ability for unknown attacks and lack flexibility for different tasks. To address these problems, we propose an improved one-class softmax (IOC-Softmax) by reformulating the scale factor of one-class softmax (OC-Softmax) into two parts such that it down-weights the loss assigned to bona fide samples. Our IOC-Softmax is a novel dynamically scaled OC-Softmax, which can select different scale factors for different tasks, so that it can prevent a large number of bona fide samples from overwhelming the anti-spoofing model during training. To verify the effectiveness of the proposed IOC-Softmax, we perform numerous experiments on the ASVspoof 2019 dataset. The results show that our IOC-Softmax achieves better performance when compared with OC-Softmax.**

I. Introduction

Automatic speaker verification (ASV) is a technique that verifies the identity of the speaker by the speech information [1], which has been applied successfully to various practical scenarios, such as remote authentication. However, spoofing attacks created by text-to-speech (TTS), voice conversion (VC) and replay bring big threats to the ASV system.

To enhance the anti-spoofing ability of the ASV system, voice spoofing detection methods are investigated to detect spoofing attacks. Four ASVspoof challenges [2]–[5] have been successfully organized, which provide datasets, evaluation metrics and baselines for research on anti-deception tasks. The common anti-spoofing research centers on features and classifiers. Well-designed acoustic features have achieved good performance, such as Linear Frequency Cepstral Coefficients (LFCC) [6], constant Q transform (CQT) [7], etc. Gaussian Mixture Model (GMM) [8]–[10] and deep neural network-based classifiers [11]–[14] are the mainstream back-end classifiers. Other efforts have investigated the data augmentation [15], feature fusion [16] and other training strategies to improve the detectability of the anti-spoofing model. Ge et al. [17] employed the LFCC as the front-end feature, and explored a differentiable architecture search method to detect spoofing attacks. Li et al. [18] introduced channel-wise gating into Res2Net to reduce the redundant information.

Although great progress has been achieved, the generalization ability of existing efforts to unknown attacks is not satisfactory. One possible reason is that most anti-spoofing efforts are considered to be the binary classification problem, which assumes that test data and training data have similar data distributions for the same class. However, in the actual application, with the rapid development of TTS and VC technologies, the data distribution of the training set is difficult to catch up with the expansion of test data.

To this end, a potential direction for improving generalization ability is one-class learning. Villalba et al. [19] presented a one-class SVM to fit the distribution of bona fide speech, but the feature space is still obtained by the form of binary classification. Zhang et al. [13] proposed the OC-Softmax for synthetic voice spoofing detection, which builds the feature space by employing two margins for genuine and spoofed speech. Lou et al. [20] used OC-Softmax to model the bona fide samples for replay attack detection, and designed the residual variability block to reduce the variance of bona fide embedding space.

Although OC-Softmax has improved the detection ability of the anti-spoofing model, it still has the following two problems: first, the generalization ability is still unsatisfactory, and second, the flexibility to different tasks is insufficient. We discover that the scale factor in OC-Softmax is the central cause, which causes genuine samples to overwhelm the training of the model. Specifically, when real speech and fake speech are classified correctly, the loss value contributed by genuine speech is much greater than that of spoofed speech. And the genuine speech is compact enough around the weight center, it is difficult to further reduce its loss. Therefore, in the process of training, the model can only further force the spoofing speech to converge in the opposite direction of the weight vector to reduce the total loss. However, the aggregation of spoofing speech will lead to poor generalization ability. To overcome the above problems, this paper proposes an improved OC-Softmax, named IOC-Softmax. Our method reformulates the scale factor of OC-Softmax into two parts, one tailored for genuine speech and the other tailored for spoofing speech. These two scale factors enhance the generalization ability of the anti-spoofing model and the flexibility of one-class learning by adjusting the loss values contributed by genuine samples

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

and spoofed samples.

The main contributions are as follows:

- We empirically investigate and analyze the effects of the scale factor.
- We propose an improved one-class softmax (IOC-Softmax) by reshaping the scale factor into two parts to improve generalization ability of anti-spoofing model and flexibility for different tasks.
- Comprehensive experiments are carried out on both the LA and PA scenarios of the ASVspoof 2019 dataset to demonstrate the advantage of the proposed IOC-Softmax.

II. Method

A. Preliminary: One-class learning loss function

The original one-class learning loss function [13] can be defined as

$$L_{OC} = \frac{1}{M}\sum_{i=1}^{M}\log(1+e^{s(m_{y_i}-\hat{w}_0\cdot\hat{x}_i)(-1)^{y_i}}), \qquad (1)$$

where M is batch size. s is the scale factor. $\hat{w}_0$ is a normalized weight vector, and it represents the optimization direction of real samples. $\hat{x}_i \in R^D$ is the normalized feature embedding, where D is the dimension of feature embedding. In this paper, we take D as 256. y_i is the label. m_0 and m_1 are two margins used to constrain the angle α_i between $\hat{w}_0$ and $\hat{x}_i$. The range of values for m_0 and m_1 is -1 to 1, and the value of m_0 is greater than that of m_1. When the input is the genuine speech ($y_i = 0$), m_0 is used to make α_i less than $\arccos(m_0)$, and a small $\arccos(m_0)$ can make the genuine samples more compact around $\hat{w}_0$. m_1 is employed to force α_i larger than $\arccos(m_1)$ when $y_i = 1$, and a relatively large $\arccos(m_1)$ can keep the spoofing samples away from the $\hat{w}_0$.

B. Improved one-class learning

In this subsection, we first delve into the mechanism of OC-Softmax. Table I elaborates four cases described in Fig. 1(a), Fig. 1(b) and Fig. 1(c), i.e., $B1$, $B2$, $B3$ and $B4$. $\hat{w}_0\cdot\hat{x}_i$ represents the similarity of center vector $\hat{w}_0$ and speech samples $\hat{x}_i$. Fig. 1(a) is the schematic of OC-Softmax, where $B1$, $B2$, $B3$ and $B4$ are four cases discussed in Fig. 1(b) and Fig. 1(c). Fig. 1(b) describes that the value of OC-Softmax ($m_0 = 0.9, m_1 = 0.2$) varies with the scale factor s under four cases. The corresponding loss functions of these cases are L_{OC-B1}, L_{OC-B2}, L_{OC-B3} and L_{OC-B4}. We rethink OC-Softmax based on these four cases. As shown in Fig. 1(b), compared with $L_{OC-B2,B3}$, the loss values of L_{OC-B1} and L_{OC-B4} are much smaller, so the genuine samples will converge to $B1$, and the spoofed samples will converge to $B4$. For $B1$ and $B4$, when the genuine samples and the spoofed samples share the same scale factor, the loss values contributed by these two types of samples are either almost zero or very different. The former makes the embedding space of both genuine speech and spoofed speech more compact, resulting in the poor generalization ability of the model. For the latter, taking the setting $s = 20$ in OC-Softmax as an example, the loss value contributed by the genuine sample is 0.127, while the loss value contributed by the spoofed sample is 3.775×10^{-11}, which is quite different, resulting in the genuine samples overwhelming the training of the model. Further, Fig. 1(c) illustrates the loss range of OC-Softmax when genuine samples and spoofing attacks are classified correctly, where $\gamma = s(m_{y_i} - \hat{w}_0\cdot\hat{x}_i)(-1)^{y_i}$. For genuine samples, the range of γ is $[-2, 0]$ and the corresponding range of loss value is $[0.127, 0.639]$. For spoofing attacks, the range of γ is $[-24, 0]$ and the corresponding range of loss value is $[3.775 \times 10^{-11}, 0.639]$. When the genuine speech converges to $B1$, its loss value reaches the minimum, that is, 0.127. To further reduce the total loss value, the spoofing speech can only be compressed to $B4$ during the training of the model. This leads to a compact spoofing embedding space and then results in poor generalization ability of the anti-spoofing model.

In addition, genuine speech and spoofing attacks share the same scale factor, which limits the flexibility to balance two classes. For example, replay attacks on the PA scenario are produced by different acoustic and replay configurations, and the training data and testing data still have similar data distribution in essence. Synthetic samples are produced by different TTS and VC technologies, which results in the different distribution of testing data and training data. Because of the above, it is difficult to use only one scale factor to achieve universality.

To solve these problems, we reformulate the formula of OC-Softmax, called IOC-Softmax. The definition of IOC-Softmax is given as follows:

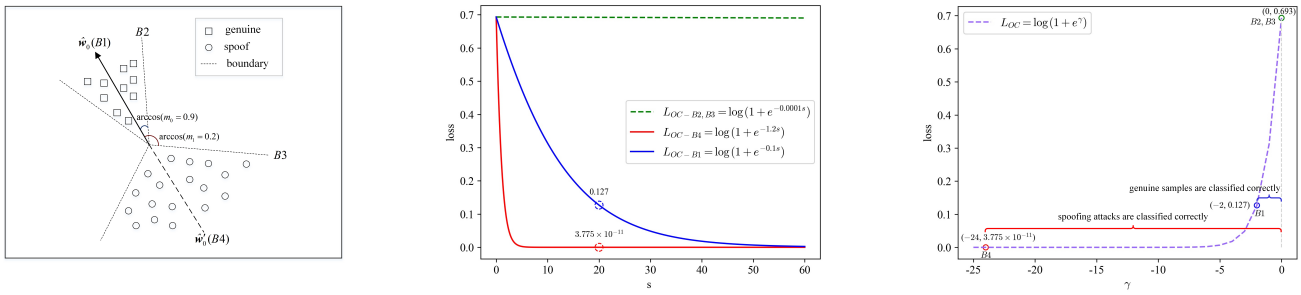
$$L_{IOC} = \frac{1}{M}\sum_{i=1}^{M}\log(1+e^{s_{y_i}(m_{y_i}-\hat{w}_0\cdot\hat{x}_i)(-1)^{y_i}}), \qquad (2)$$

where s_{y_i} is the scale factor, including two parts, i.e. s_0 tailored for genuine samples and s_1 customized for spoofing attacks. These factors control the loss values contributed by speech samples.

III. Experiments

A. Datasets

We perform all experiments on the ASVspoof 2019 dataset [21], which includes two subsets, i.e. LA scenario and PA scenario. Each scenario consists of three partitions, i.e. training set, development set and evaluation set. For brevity, we abbreviate these partitions to Train, Dev and Eval. Both LA and PA scenarios contain bona fide speech and spoofed speech. On the LA scenario, the spoofed speech is generated by representative TTS and VC algorithms. The spoofing attacks of the Trian set and the Dev set are both A01-A06 attacks, while the attack types of the Eval set are different from those of the Train and Dev sets. This results in the spoofing samples of the Eval set being unknown attacks to the model trained with the Train set. On the PA scenario, the spoofed samples are produced by 27 acoustic configurations and 9 replay configurations, and all sets share the same spoofing attack types, namely AA-CC. Table II describes this dataset in detail.



(a) (b) (c)

Fig. 1. The rethinking of OC-Softmax. (a) The schematic of OC-Softmax. (b) The loss value of OC-Softmax varies with scale factor under $B1$, $B2$, $B3$ and $B4$. (c) The loss range of OC-Softmax when genuine samples and spoofing attacks are correctly classified.

TABLE I
FOUR CASES SHOWN IN FIG. 1.

	Case	Description
$B1$	$\hat{w}_0 \cdot \hat{x}_i = 1$	genuine samples $\hat{x}_i$ converge to the weight vector $\hat{w}_0$.
$B2$	$\hat{w}_0 \cdot \hat{x}_i = 0.9001$	$\hat{w}_0 \cdot \hat{x}_i = 0.9$ is the critical condition for the correct classification of genuine samples.
$B3$	$\hat{w}_0 \cdot \hat{x}_i = 0.1999$	$\hat{w}_0 \cdot \hat{x}_i = 0.2$ is the critical condition for the correct classification of spoofed samples.
$B4$	$\hat{w}_0 \cdot \hat{x}_i = -1$	spoofed samples $\hat{x}_i$ are forced to the opposite direction of weight vector $\hat{w}_0$.

TABLE II
THE ASVSPOOF 2019 DATASET.

Subset		#Genuine	#Spoof	Attack Type
LA	Train	2580	22800	A01-A06
	Dev	2548	22296	A01-A06
	Eval	7355	63882	A07-A19
PA	Train	5400	48600	AA-CC
	Dev	5400	24300	AA-CC
	Eval	18090	116640	AA-CC

TABLE III
EER(%) AND T-DCF PERFORMANCE COMPARISON WITH OTHER LOSS FUNCTIONS ON THE LA SCENARIO.

Loss	Dev		Eval	
	EER	t-DCF	EER	t-DCF
Softmax	0.35	0.010	4.69	0.125
AM-Softmax	0.43	0.013	3.26	0.082
OC-Softmax	**0.20**	**0.006**	2.19	0.059
IOC-Softmax	0.23	0.007	**1.71**	**0.046**

TABLE IV
EER(%) AND T-DCF PERFORMANCE COMPARISON WITH OTHER LOSS FUNCTIONS ON THE PA SCENARIO.

Loss	Dev		Eval	
	EER	t-DCF	EER	t-DCF
Softmax	1.34	0.037	1.90	0.055
AM-Softmax	1.71	0.039	1.89	0.051
OC-Softmax	1.33	0.040	1.99	0.058
IOC-Softmax	**1.18**	**0.028**	**1.34**	**0.038**

B. Evaluation metrics and implementation details

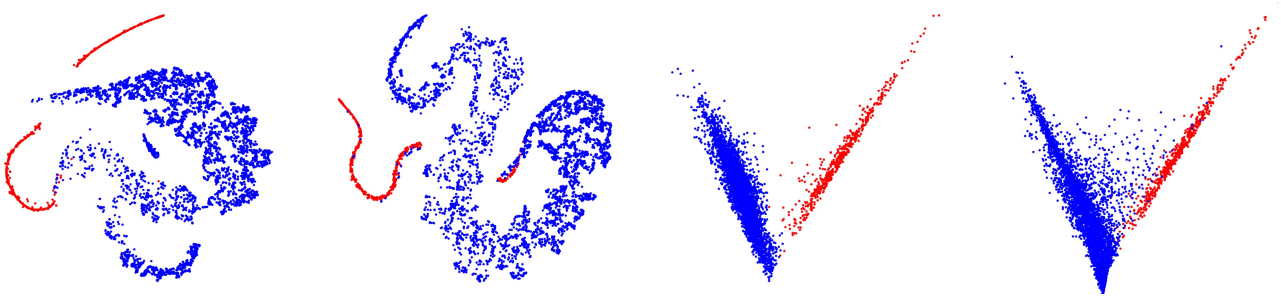
We use equal error rate (EER) and the tandem detection cost function (t-DCF) [22] to assess the proposed IOC-Softmax, the former measures the capability of the voice spoofing detection model and the latter appraises the reliability of the voice spoofing detection model to the ASV system. The lower the two indicators are, the stronger the system performance is.

For a fair comparison with OC-Softmax, we employ the network structure from [13]. The input of the network is extracted 60-dimensional LFCC features, and the 750 frames are set as the fixed length. The output of the network is 256-dimensional feature embedding. The feature embedding is then fed into IOC-Softmax to get the score that indicates the classification result. For AM-Softmax, s and m are set to 20 and 0.9, respectively. For OC-Softmax, s, m_0 and m_1 are 20, 0.9 and 0.2, respectively. And for IOC-Softmx, m_0 and m_1 are same as those of OC-Softmax, and s_0 and s_1 are 40 and 20 on the LA scenario, respectively. On the PA scenario, s_0 and s_1 are set to 60 and 40, respectively.

We train all the models using Pytorch on a single NVIDIA RTX 2080 Ti GPU. The Adam optimizer and Stochastic Gradient Descent (SGD) optimizer are employed to update the weights of the backbone and the loss function respectively, and the β_1 and β_2 in the Adam optimizer are 0.9 and 0.999 respectively. The epoch and batch size are set to 100 and 64 for all models. The initial learning rate is 0.0003, which decreases by 50% every 10 epochs. The model with a minimum EER on the Dev set was chosen as the final model for evaluation.

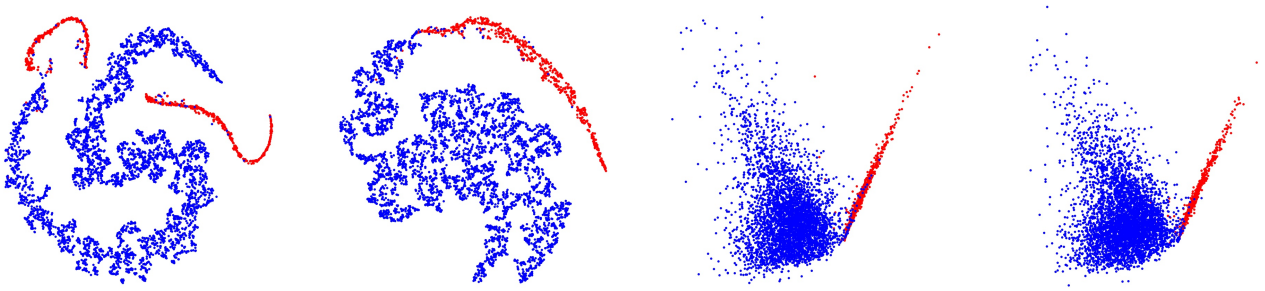
C. Results

1) Comparison with other loss functions: To validate the effectiveness of the proposed method, we conduct comparative tests with Softmax, AM-Softmax and OC-Softmax on the Dev and Eval sets of the ASVpoof 2019 dataset. Experimental re-



(a) (b) (c) (d)

Fig. 2. t-SNE and PCA visualizations on the LA scenario.



(a) (b) (c) (d)

Fig. 3. t-SNE and PCA visualizations on the PA scenario.

TABLE V
COMPARISON WITH SOFTMAX VARIANTS ON INDIVIDUAL ATTACKS OF THE LA SCENARIO.

Attack	Softmax	OC-Softmax	AM-Softmax	IOC-Softmax
A07	0.36*	0.12	0.22	0.12
A08	0.13	0.17*	0.06	0.04
A09	0.04	0.12*	0.01	0.02
A10	1.18*	1.14*	0.63	0.50
A11	0.34*	0.12	0.02	0.04
A12	0.43*	0.46*	0.52*	1.01
A13	0.69*	0.22	0.26	0.28
A14	4.98*	0.69	0.50*	0.52
A15	5.43*	1.40	0.69*	1.12
A16	0.22	0.32	0.50*	0.28
A17	23.48*	9.21*	13.47*	7.77
A18	0.20*	0.89*	4.23*	1.44
A19	1.34*	0.89	0.85	0.83

TABLE VI
COMPARISON WITH SOFTMAX VARIANTS ON INDIVIDUAL ATTACKS OF THE PA SCENARIO.

Attack	Softmax	OC-Softmax	AM-Softmax	IOC-Softmax
AA	2.68*	2.67*	2.62*	1.87
AB	1.37	1.48*	1.36	1.01
AC	1.38*	1.75*	1.52 *	0.88
BA	2.38 *	2.34*	2.31	1.83
BB	1.41*	1.32	1.35	1.00
BC	1.41*	1.82*	1.44*	0.86
CA	2.34*	2.54*	2.23*	1.70
CB	1.66*	1.50*	1.52*	1.10
CC	1.62*	1.89*	1.71*	0.96

sults in Tables III and IV show that the proposed IOC-Softmax outperforms other loss functions. Specifically, compared with OC-Softmax, our IOC-Softmax improves EER by 21.91% and t-DCF by 22.03% on the Eval set of the LA scenario. And on the Eval set of the PA scenario, our method achieves an EER improvement of 33.66% and a t-DCF improvement of 34.48% when compared with OC-Softmax.

2) Performance comparison on individual attacks: For four different loss functions, we calculate their detection capabilities for each attack on the Eval set, and the comparison results are presented in Tables V and VI. In order to evaluate performance differences more clearly, we use the method in [23] to conduct the statistical significance test. "*" indicates a statistically significant difference in EER performance of individual attacks between other loss functions and IOC-Softmax, with a confidence interval of 99%. Table V shows that our IOC-softmax outperforms other loss functions on individual attacks, especially on A17. As shown in Table VI, our IOC-softmax achieves significantly superior performance on individual attacks.

3) Visualizations: We perform feature embedding visualizations using Principal Component Analysis (PCA) [28] and t-distributed Stochastic Neighbor Embedding (t-SNE) [29]. In Fig. 2 and Fig. 3, (a) and (b) are the results of t-SNE, and (c) and (d) are visualization results using PCA. The (a) and (c) in both Fig. 2 and Fig. 3 are embedding visualizations on the Dev set, and the (b) and (d) are embedding visualizations on the Eval set. Red dots represent genuine speech and blue dots are spoofed speech. As can be seen from Fig. 2, on the LA scenario, the data distribution of spoofing attacks on the Dev is different from that of spoofing attacks on the Eval.

TABLE VII
PERFORMANCE COMPARISON WITH SOME OF EXISTING SINGLE SYSTEMS

Single System	LA		PA	
	EER	t-DCF	EER	t-DCF
Baseline CQCC+GMM [4]	9.57	0.237	11.04	0.245
Baseline LFCC+GMM [4]	8.09	0.212	13.54	0.302
CQMOC+LCNN [12]	8.59	0.26	1.32	0.036
CQMOC+GMM [12]	7.10	0.199	6.95	0.135
CQT-MMPS+LCNN [12]	5.99	0.176	**0.90**	0.024
CQT-MMPS+ResNet [12]	3.72	0.119	1.08	0.031
CQCC+NF-ResNeXt [24]	6.72	0.160	-	-
MFCC+NF-ResNeXt [24]	5.13	0.132	-	-
sm-ALTP+Asymmetric Bagging [25]	5.22	0.132	1.10	0.034
STFT+LCGRNN+GKDE-Triplet [26]	3.03	0.078	0.92	**0.020**
Spec+scDenseNet [27]	4.45	0.131	-	-
LFCC+scDenseNet [27]	1.89	0.047	-	-
ResNet+IOC-Softmax	**1.71**	**0.046**	**1.34**	**0.038**

However, on the PA scenario of Fig. 3, the data distributions of spoofing attacks on the Dev and Eval sets are basically the same. This indicates that it is necessary to customize scale factors for different scenarios. As can be seen in Fig. 2 and Fig. 3, our IOC-Softmax can separate genuine speech from spoofing speech very well, showing good detection capability.

4) Comparison with other single systems: In order to further verify the advantage of the proposed IOC-Softmax, we compare our method to advanced single systems on the ASVspoof 2019 Eval set, and Table VII describes EER(%) and t-DCF results. These results are from reports in the [4], [12], [24]–[27], and "-" indicates that the LA scenario or PA scenario is not investigated in the corresponding paper. As can be seen from Table VII, our method achieves comparable or better results than other advanced single systems.

IV. CONCLUSIONS

In this paper, we study and analyze the effect of scale factor on OC-Softmax, and find that setting the same scale factor for real speech and spoofing speech reduces the detectability of the model and limits its flexibility for different tasks. For this purpose, this paper presents an improved one-class learning loss, called IOC-Softmax. We reformulate the scale factor of OC-Softmax into two parts, one tailored for genuine speech, and the other customized for spoofing speech, to balance the impact of two class samples on the loss value. We perform comprehensive experiments on the ASVspoof 2019 dataset, and the experimental results show that the proposed IOC-Softmax is effective. Compared with OC-Softmax, our approach achieves better performance. In future work, we will apply the proposed IOC-Softmax to other fields.

ACKNOWLEDGMENT

This work is partly supported by the Key R&D Program of Shandong Province, China (Grant No. 2021CXGC010107), the National Natural Science Foundation of China (Grant Nos. 62032002, 61932005), and the 111 Project (Grant No. B21049).

REFERENCES

[1] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, Jun. 2006.

[2] Z. Wu, T. Kinnunen, N. W. D. Evans, *et al.*, "Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 2037–2041.

[3] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 2–6.

[4] M. Todisco, X. Wang, V. Vestman, *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 1008–1012.

[5] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection," 2021. arXiv: 2109.00537.

[6] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 2087–2091.

[7] J. E. Youngberg and S. F. Boll, "Constant-q signal analysis and synthesis," in *Proc. ICASSP 1978 – 3rd IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tulsa, Oklahoma, USA, Apr. 1978, pp. 375–378.

[8] M. Todisco, H. Delgado, and N. W. D. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for

automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, May 2017.

[9] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 2092–2096.

[10] J. Sánchez, I. Saratxaga, I. Hernáez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, Feb. 2015.

[11] J. Monteiro, M. J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, p. 101 096, Apr. 2020.

[12] J. Yang, H. Wang, R. K. Das, and Y. Qian, "Modified magnitude-phase spectrum information for spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1065–1078, Feb. 2021.

[13] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, Apr. 2021.

[14] H. Tak, J. Jung, J. Patino, M. Todisco, and N. W. D. Evans, "Graph attention networks for anti-spoofing," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 2356–2360.

[15] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, "A study on data augmentation in voice anti-spoofing," *Speech Communication*, vol. 141, pp. 56–67, May 2022.

[16] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 1078–1082.

[17] W. Ge, M. Panariello, J. Patino, M. Todisco, and N. W. D. Evans, "Partially-connected differentiable architecture search for deepfake and spoofing detection," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 4319–4323.

[18] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 4314–4318.

[19] J. A. V. López, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the asvspoof 2015 challenge," in *Proc. INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 2067–2071.

[20] Y. Lou, S. Pu, J. Zhou, X. Qi, Q. Dong, and H. Zhou, "A deep one-class learning method for replay attack detection," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 4765–4769.

[21] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101 114, May 2020.

[22] T. Kinnunen, K. Lee, H. Delgado, *et al.*, "T-dcf: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey 2018 – 11st The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, Jun. 2018, pp. 312–319.

[23] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Odyssey 2004 – 4th The Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004, pp. 237–244.

[24] Z. Zhang, X. Zhao, and X. Yi, "Improving robustness of speech anti-spoofing system using resnext with neighbor filters," in *Proc. ICME 2022 – 23rd IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, July 2022, pp. 1–6.

[25] M. Aljasem, A. Irtaza, H. Malik, *et al.*, "Secure automatic speaker verification (SASV) system through sm-altp features and asymmetric bagging," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3524–3537, May 2021.

[26] A. G. Alanıs, J. A. G. López, and A. M. Peinado, "A kernel density estimation based loss function and its application to asv-spoofing detection," *IEEE Access*, vol. 8, pp. 108 530–108 543, Jun. 2020.

[27] S. Cui, B. Huang, J. Huang, and X. Kang, "Synthetic speech detection based on local autoregression and variance statistics," *IEEE Signal Processing Letters*, vol. 29, pp. 1462–1466, 2022.

[28] H. Hotelling, "Analysis of a complex of statistical variables in principal components," *Journal of Educational Psychology*, vol. 24, no. 7, pp. 498–520, 1933.

[29] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

Lightweight CNN and Image Enhancement Using in Palm Vein Recognition

Ping-Han Chen[1], Yung-Sheng Hung[2], and Chih-Hsien Hsia[3,*]
[1] Internet of Things Laboratory, Chunghwa Telecom Laboratories, Taoyuan, Taiwan
[2] Kang Chiao International School, New Taipei, Taiwan
[3] Department of Computer Science and Information Engineering, National llan University, Yilan, Taiwan
[*]E-mail: hsiach@niu.edu.tw

***Abstract* — Information safety is becoming increasingly important in modern society, and biometric recognition technologies have become a key to achieve information security due to its universality, permanence, distinctiveness, and collectability. This research proposed a palm vein verification system based on CNN (convolution neural network). In preprocessing step, we first perform ROI localization and image contrast adjustment, and then proceed to feature extraction through lightweight convolutional neural network. Finally, we validated our method on the CASIA palm vein dataset. As a result, the system achieved an Equal Error Rate (EER) of 0.529%, which is superior compared to other renowned architectures, and it also possesses the minimal parameter count and the fastest execution speed.**

I. INTRODUCTION

As advances in modern society continue, the importance of information security has been raised to an unprecedented height. A system that is highly secured and convenient at the same time is in the society's need. Traditionally information safety has been protected by passwords, identity card, smart card, and etc., but these systems have risks like forgetting the passwords and losing the card. Moreover, traditional methods are also vulnerable to people with bad intentions who may attempt decoding or stealing, which is clearly a sign to its shortcoming in not meeting the society's need for a well-developed information security technology. Under such circumstances, research and studies have shown that biometric recognition is a great solution to the rising issue mainly because of its following four characteristics: Universality: feature that every human being possesses naturally, such as DNA, facial features, fingerprints, veins, iris, and etc., Permanence: Biological traits do not undergo significant changes over time; any variations are minor and exhibit a linear pattern of observation. Distinctiveness: Every individual's biological characteristics is different and identifiable. Not even the left and right palm veins of the same person nor the palm veins of a pair of twins would be the same [1]. Collectability: biological characteristics can be obtained quickly and conveniently through special instrument [6].

Apart from being efficient and accurate in verifying the user's identity, utilizing human biological characteristics also reduces the risks with forgetting passwords and losing an ID card, making it much safer, more reliable, and convenient compared to the traditional methods of identity cards and passwords.

In the scheme of human biological characteristics, there are generally two categories: external biological characteristics and internal biological characteristics. External biological characteristics are the qualities of the human body that can be directly recorded and further recognized, such as facial feature [2], fingerprints [3], palm prints [4], while internal biological characteristics are the qualities that reside inside the human body and requires more sophisticated machinery to detect or collect data, such as DNA, finger veins [5], palm veins. However, external biological characteristics can be easily stolen and used by unauthorized individuals [6], For instance, facial feature can be frequently captured in unknown circumstances, while fingerprints leave traces every time objects are touched, making them exploitable by malicious actors. Therefore, systems that are based on external biological characteristics are more susceptible to attacks, lowering the overall safety of the system [7].

Compared to external biological characteristics, palm vein and other internal biological characteristics are hidden within the human body, which makes them more difficult to obtain. Slight stains and injuries would also have little to no effect on the systems' functionalities [6], and vein is a relatively consistent characteristic that does not change significantly through time [8]. It is by these qualities that vein recognition has stood out amongst human biological characteristics, having

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the potential to satisfy all the needs for a well-developed biometric system.

Vein images are mainly composed of the vein networks underneath the human skin. Under visible light, it is almost impossible for the naked eye to recognize a complete vein network. Under near-infrared light, hemoglobin (Hb) in the human blood would absorb the spectrum within 750 nm to 1050 nm of wavelength while human muscle would reflect it [9]. The vein images only can be captured using a camera with near-infrared light, it's ensuring that this biological characteristics can only be obtained from a living human being [10].

Despite its benefits, palm vein recognition still has some issues. For instance, collecting palm vein images requires the use of near-infrared light, which makes it more difficult compared to other biological databases. This limitation results in a smaller dataset, which in turn increases the difficulty of training Convolution Neural Network (CNN) models. Furthermore, biometric recognition systems are commonly used in edge devices, which often have limitations in terms of network computation and system complexity.

In the field of biometric recognition, depending on the different application scenarios, there are primarily two modes: identification and verification. The identification mode aims to identify a user's identity from a large database, following a 1:N matching method. The objective of the verification mode is to verify whether the current biological characteristic matches the specific identity previously stored in the database, adhering to a 1:1 matching method. In this research, we will focus on the verification mode.

II. RELATED WORK

In recent years, there has been rising attention to vein-based recognition systems, which includes palm vein [7], finger vein [11]. Due to its convenience in extraction and complexity in feature textures compared to other vein features in the human body, palm vein and its related studies are the focuses of this research. Typically palm vein recognition system primarily captures images using near-infrared cameras. The captured images are then processed for region of interest (ROI) localization and vein feature extraction. Finally, compares the extracted features with the enrolled feature and outputs the recognition result. Currently, palm vein recognition algorithm field can be separated into two types:

A. Handcrafted Feature:

Handcrafted features are an early and common method in the field of vein recognition. This type of method mainly defines vein images and extracts features with human identification significance. It can be further categorized into these three methods:

a. Geometry-based methods [12]: This method mainly utilizes the linear structures presented on the palm vein images and extracts data of such segments, curves, and points, which are supposedly close to the shape of the vein texture. It highly on the processed image, which requires a ROI localization; however, features obtained from this method sometimes have blurs and only few feature points, so it is more difficult to overcome sample rotation, zoom, and displacement [13]. Among these methods, radon transform [10] and Gabor feature [13] have been commonly used in the palm vein recognition field. Despite this, a clear limitation to these methods is that their performance highly relies on the artificially selected features. Consequently, these handcrafted methods may face difficulties when handling different lightings, changes in hand gesture, blockings, and other complex obstacles.

b. Statistical-based methods [11]: These methods use statistics from data to perform feature recognition, including through local binary histograms and image invariant moments, and can be distinguished as global or local statistics. However, these methods are equivalently sensitive to rotation, zoom, and displacement.

c. Local invariant-based methods [14]: Inspired by scale-invariant feature transform (SIFT), these methods directly extract local invariant data from the palm vein image feature. Although they effectively contend feature displacement and rotation, they are susceptible to small changes in the surrounding lightings and the device lightings, resulting in lesser extracted invariant feature points.

B. CNN based:

CNN based method have demonstrated its strength in feature extraction and have been successfully applied to several computer visual tasks, such as image classification [15]. It has also been used in the field of vein recognition, as it is

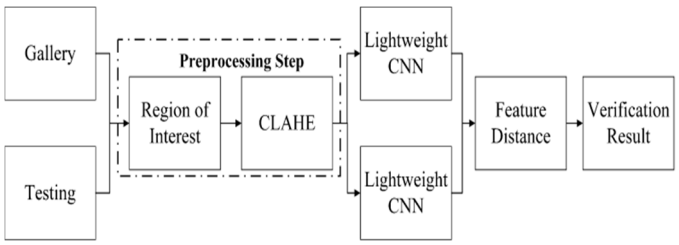


Fig. 1, System operation flowchart

applied to vein image quality assessment by Qin et al. [16] and also to vein pattern segmentation [17]. Additionally, Chen et al. [6] proposed a lightweight neural network for palm vein, using Gabor filter and Weight Selection Mechanism to enhance system safety, and Qin et al. [18] proposed a novel palm vein defensive framework, reducing the negative effects caused by low image qualities. Eladlani et al. [25] proposed a method that utilizes model merging, and ensemble network techniques to achieve commendable recognition results. These previous working examples have proved that CNN based algorithms have decent recognition ability in the vein recognition field, but they mostly overlooked the calculation inabilities of typical edge devices. Therefore, this research introduced lightweight system architecture, implementing CLAHE and lightweight CNN to lower the error rate of the system.

III. METHOD

A. System flowchart

Fig. 1 illustrates the workflow of the palm vein verification system proposal in this research. To reduce system load, the system mainly consists of the preprocessing stage and the feature extraction stage.

Initially, after obtaining the palm vein image, the system performs region of interest (ROI) localization enhances its contrast by applying Contrast Limited Adaptive Histogram Equalization (CLAHE). The adjusted image is then fed into a CNN for feature extraction. This compresses the original two-dimensional palm vein image into one-dimensional feature. During the verification step, the extracted features will calculate the feature distance with the enrolled features. If the distance is smaller than the threshold, the system will classify them as the same user.

B. Preprocessing – ROI

In the image captured through a near-infrared camera, background information is unnecessary to the system apart from the main portion that contains the palm veins. A great

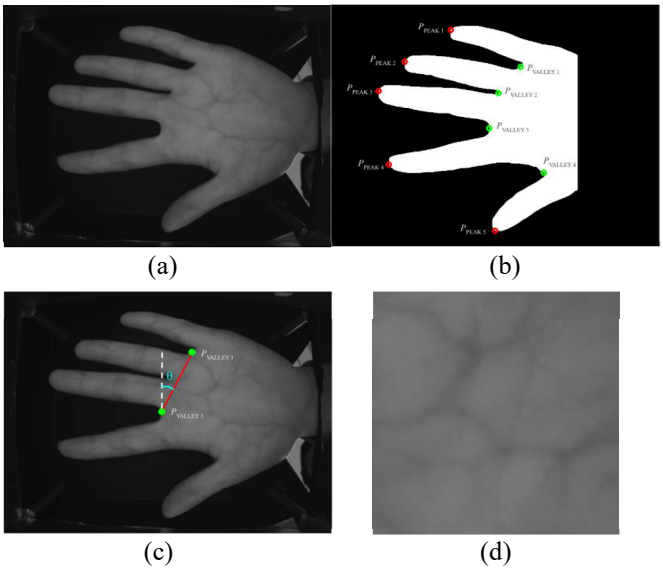


Fig. 2, Palm vein extraction and ROI process diagram (a) Original palm vein image, (b) Locating finger peaks and valleys, (c) Rotation angle θ, (d) Palm vein ROI image.

image localization method can effectively improve the system's effectiveness. The palm-vein capturing device on the current market can be separated into two types: contact and contactless devices. Even though contact-required devices can limit and reduce inconsistencies in hand placement, most systems employ contactless devices due to their superiority in terms of hygiene, allowance for unrestrained hand placement, and higher acceptance from users [13]. However, this inevitably causes more noise and varieties in size ratio, displacement, and rotation due to inconsistent camera angles and hand placements from different images that are captured at a different occasion. Consequently, several algorithms, such as ROI contouring, angle correction, scale correction, and other algorithms, are needed to ensure that the extracted palm vein data would not be heavily influenced by the aforementioned variables.

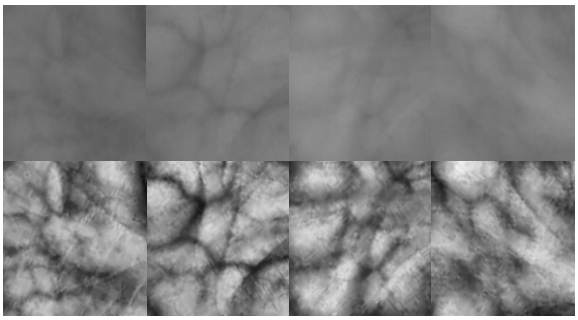


Fig. 3, CLAHE Sample Comparison. The first row shows the original image, the second row shows the image after applying CLAHE.

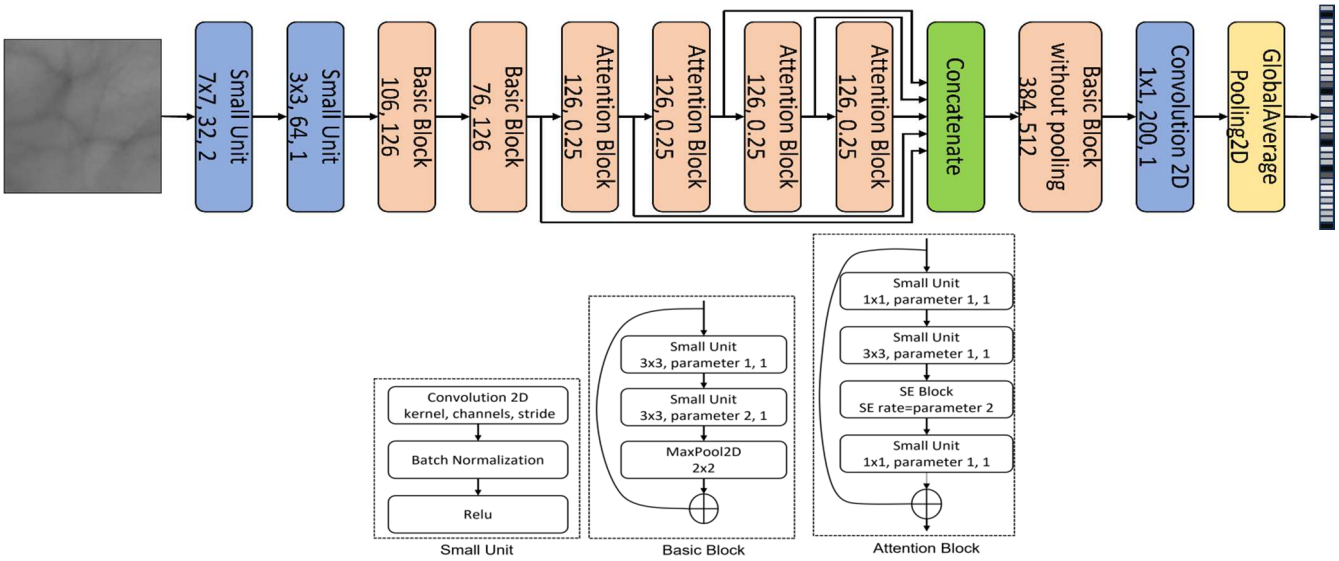


Fig. 4, Proposal architecture of the Lightweight CNN model.

The proposal ROI localization algorithm in this research, as shown in Fig. 2, involves filtering the foreground and background using OTSU's method [19]. Next, the radial distance function (RDF) is used to calculate the distance between the palm centroid and the palm contour, allowing for the localization of fingertip and finger valley positions. By identifying the finger valleys between the index finger and middle finger, as well as between the ring finger and little finger, the hand rotation angle is determined, and the ROI region size is defined. Finally, the palm vein information is extracted.

C. Preprocessing - contrast adjustment

Although the addition of an external near-infrared light source improves the contrast to some extent during palm vein extraction, the contrast in vein images still lack adequacy. After reference [18], we found that adding the CLAHE method would noticeably increase the contrast between veins and muscle tissues in palm vein images.

In CLAHE, there are two parameters, Clip-Limit and Tile-Size. The Clip-Limit primarily serves as a cap, redistributing excess histogram values evenly to lower values. This method effectively prevents the loss of image detail due to excessively enhanced contrast. The Tile-Size determines the number of sections into which the image will be divided for processing. More subdivisions can enhance the image's local contrast, balancing brightness and contrast within each block, but it can also introduce conspicuous block effects. Fig. 3 displays images of the veins before and after implementing the CLAHE algorithm. When compared to images processed without the use of CLAHE, it is observed that the image processed with CLAHE has an enhanced contrast between the veins and muscle tissues.

D. Lightweight Convolution Neural Network

To operate palm vein image verification, the processed image would go through a refined CNN structure that is adaptable to vein images to perform feature extraction. Most of the CNNs architectures are not specifically designed for palm vein verification and may face challenges such as inadequate recognition accuracy or excessively large network parameters when confronting palm vein images.

To address this problem, we came up with an improved version of the CNN structure, in which the main structure is illustrated in Fig. 4. This architecture is made up of multiple convolution layers with batch normalization layer (BN) and Relu after each of the convolution layers. Additionally, in order to reduce the number of network parameters, we employed separable convolutions [16] for all convolution layers except the first two convolutions. This choice enhances processing speed and reduces the overall number of model parameters. Residual connections [17] are also used on Basic Block and Attention Block. Finally, a Dropout layer is added before the output layer to mitigate overfitting.

In Loss function, we hired the Triplet loss function to enhance the network's ability to separate feature distance from different samples. The Triplet loss aims to minimize the feature distance between the same user's samples and maximize the distance between the different user's samples. In the training step, the calculated value is utilized to optimize the lightweight CNN.

IV. EXPERIMENT RESULT

A. Database

To assess our proposal method's performance, we chose to use the database Multi-Spectral Palmprint Image Database 1.0 from CASIA (Institute of Automation of the Chinese Academy of Sciences) as experimental dataset, which was collected through contactless devices and consists of the right and left hand images of a hundred people. This dataset was created with two sessions, with at least a month spanning in between each session, in order to simulate real-life circumstances. Each collection will add six different light spectra, including 460nm, 630nm, 700nm, 850nm, 940nm, and white. For each session, three images were captured for each type of spectrum. The entire dataset consists of 7200 images, 2400 of which used near-infrared spectra and are suitable to perform palm vein recognition. To simulate real-life situations, we separated the images collected from those two sessions into categories called gallery and test, each having 1200 images. Apart from serving as training samples for deep learning, gallery images also represent registered users and are used in the EER experiments.

B. Evaluation Metrics

In biometric recognition technologies, there are typically two indexes used to evaluate its performance: Accuracy and Equal Error Rate (EER). Each of these two indexes corresponds to two major fields of biometric recognition: Identification and Verification.

Identification is the process of comparing the input biometric characteristic with the registered biometric characteristic to identify the identity. In this scenario, Accuracy is an important index. It is comparably simpler and more straightforward, representing the ability of the system to distinguish and identify individual samples. In an ideal situation, a perfectly accurate system would have an Accuracy of 100%, meaning that it has classified every input correctly.

On the other hand, Verification's goal is to confirm whether the user is the identity of which they claimed to be, so we only have to compare the input data with the said identity's known data. EER is often used to evaluate the security of a system. EER is composed of the two parameters False Accept Rate (FAR) and False Reject Rate (FRR). During evaluation, by adjusting the threshold, FAR and FRR will change correspondingly; when their values are the same, this specific point represents the system's EER. A lower EER indicates higher recognition performance and better security of the system.

Table 1, The impact of adjusting Tile Size on the EER, with Clip-Limit of 1, 9 and 17.

Tile Size	*Clip-Limit (EER %)*		
	1	*9*	*17*
4	0.547	**0.529**	0.646
8	1.6	0.963	1.162
16	2.146	4.111	1.724
20	2.281	2.946	1.996
32	0.592	2.312	2.424
40	2.969	3.413	3.268
80	3.444	5.380	2.632
160	3.464	2.487	3.291

C. Contrast Adjustment

To evaluate the effectiveness of the CLAHE algorithm in adjusting image contrast, we compared the original image with the processed one. In addition, to find out their effects on the palm vein recognition system, we performed the following experiment based on the two parameters Clip-Limit and Tile Size of the CLAHE algorithm. Firstly, we hold Clip-Limit constant at 1, 9 and 17 while adjusting Tile Size. As shown in Table 1, there are variations in EER with the increase of Tile Size. In the end, we selected a Tile Size of 4 and a Clip-Limit of 9 as the final parameter values.

D. Lightweight Convolution Neural Network

In this research, we designed a lightweight convolution neural network specifically for palm vein verification. This architecture's intention is to improve the precision and efficiency of biometric verification and offer opportunities for future applications to edge devices. To thoroughly assess the architectures designed, we compared our design with various common convolutional neural networks.

As shown in Table 2, our network architecture performs excellently in terms of EER and has the lowest processing-time and parameter count among all the comparison subjects. This demonstrates that the proposal network architecture is the most applicable to be implemented in edge devices, because while promising a decent accuracy, it effectively lowers the processing time and parameters, which is a significant advantage as a system operating in an edge device that requires specific capabilities.

Table 2, Comparison among different network structures

Method	*Parameter*	*Processing -time*	*EER (%)*
Xception[20]	21.31M	0.0709s	0.882
VGG16[21]	14.85M	0.0673s	2.274
MobileNetV2[22]	2.27M	0.0287s	6.260
NASNetMobile[23]	4.52M	0.0680s	5.251
ResNet50 V2[24]	24.03M	0.0570s	4.293
Proposal	**1.12M**	**0.0285s**	**0.529**

V. CONCLUSION

This research proposes a lightweight palm vein verification system. During the image preprocessing stage, the Region of Interest is utilized to filter out background information, and CLAHE is applied to enhance the vein information. Lightweight CNN is then used for feature extraction. The extracted features are then compared with the features of enrolled users, and the feature distance is calculated. If the distance is below a threshold, it is recognized as the same user. The experiment was validated using the CASIA multispectral palm vein database, achieving an EER of 0.529%. The experimental results also demonstrate that our proposed network has an excellent verification rate and the potential for deployment on edge devices.

VI. REFERENCES

[1] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki and A. T. S. Ho, "Detection of Face Spoofing Using Visual Dynamics," *IEEE Tran. Inform. Fore. Secu.*, vol. 10, no. 4, pp. 762-777, 2015.

[2] R. He, J. Cao, L. Song, Z. Sun and T. Tan, "Adversarial Cross-Spectral Face Completion for NIR-VIS Face Recognition," *IEEE Tran. Patt. Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025-1037, 2020.

[3] S. Balakrishnan, Venkatesan K. and M. S. S. Hameed, "An Embarking User Friendly Palmprint Biometric Recognition System with Topnotch Security," *2021 5th Inter. Conf. Intell. Comput. and Cont. Sys. (ICICCS)*, Madurai, India, 2021, pp. 1028-1032.

[4] A.-S. Ungureanu, S. Salahuddin and P. Corcoran, "Toward Unconstrained Palmprint Recognition on Consumer Devices: A Literature Review," *IEEE Access*, vol. 8, pp. 86130-86148, 2020.

[5] C.-H. Hsia and C.-H. Liu, "New Hierarchical Finger-Vein Feature Extraction Method for iVehicles," *IEEE Sens. Jour.*, vol. 22, no. 13, pp. 13612-13621, 2022.

[6] Y.-Y. Chen, C.-H. Hsia and P.-H. Chen, "Contactless Multispectral Palm-Vein Recognition With Lightweight Convolutional Neural Network," *IEEE Access*, vol. 9, pp. 149796-149806, 2021.

[7] H. Qin, M. A. El-Yacoubi, Y. Li and C. Liu, "Multi-Scale and Multi-Direction GAN for CNN-Based Single Palm-Vein Identification," *IEEE Tran. Inform. Fore. Secu.*, vol. 16, pp. 2652-2666, 2021.

[8] F. Marattukalam, W. Abdulla, D. Cole and P. Gulati, "Deep Learning-Based Wrist Vascular Biometric Recognition," *Sensors*, vol. 6, 2023.

[9] M. Zhou, L. Lin, M. Wang, X. Li and G. Li, "Influence of water on noninvasive hemoglobin measurement by Dynamic Spectrum," *Analy. Method*, vol. 5, no.18, pp.4660-4665, 2013.

[10] J.-D. Wu and S.-H. Ye, "Driver identification using finger-vein patterns with radon transform and neural network," *Expert Sys. with Appl.*, vol. 36, no. 3, pp. 5793-5799, 2009.

[11] L. Mirmohamadsadeghi and A. Drygajlo, "Palm vein recognition with Local Binary Patterns and Local Derivative Patterns," *2011 Interna. Joint Conf. on Biometrics (IJCB)*, Washington, DC, 2011, pp. 1-6.

[12] Y. Zhou and A. Kumar, "Human Identification Using Palm-Vein Images," *IEEE Tran. Inform. Fore. Secu*, vol. 6, no. 4, pp. 1259-1274, 2011.

[13] W. Kang and Q. Wu, "Contactless Palm Vein Recognition Using a Mutual Foreground-Based Local Binary Pattern," *IEEE Tran. Inform. Fore. Secu*, vol. 9, no. 11, pp. 1974-1985, 2014.

[14] W. Kang, Y. Liu, Q. Wu and X. Yue, "Contact-free palm-vein recognition based on local invariant features," *PLoS One*, vol. 9, no. 5, p. e97548, 2014.

[15] Y. Li, S. Ruan, H. Qin, S. Deng and M. A. El-Yacoubi, "Transformer Based Defense GAN Against Palm-Vein Adversarial Attacks," *IEEE Tran. Inform. Fore. Secu*, vol. 18, pp. 1509-1523, 2023

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861*, 2017.

[17] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conf. on Comput. Vis. Patt. Recog. (CVPR*), Las Vegas, NV, USA, 2016, pp. 770-778.

[18] Y. Ayoub et al., "Diagnostic Superficial Vein Scanner," *2018 Interna. Conf. on Comput. and Appli. (ICCA)*, Beirut, Lebanon, 2018, pp. 321-325.

[19] N. Otsu, "A Threshold Selection Method from Gray-Level

Histograms," *IEEE Tran. Sys., Man, and Cybern.*, vol. 9, no. 1, pp. 62-66, 1979.

[20] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, I. Ullah and X. Zhang, "DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition," *Expert Sys. Appli.*, vol. 191, 2022.

[21] B. Maser and A. Uhl, "Using CNNs to Identify the Origin of Finger Vein Sample Images," *2021 IEEE Interna. Workshop on Biom. and Fore. (IWBF)*, Rome, Italy, 2021, pp. 1-6.

[22] C. Nadir, B. Attallah and Y. Brik, "Finger Vein Based CNN Algorithms for Human Recognition," *2022 Interna. Conf. Advan. Tech. Electron. Electr. Eng. (ICATEEE)*, M'sila, Algeria, 2022, pp. 1-6.

[23] Wei Jia, W. Xia, Y. Zhao, H. Min and Y.-X. Chen, "2D and 3D Palmprint and Palm Vein Recognition Based on Neural Architecture Search," *Interna. Jour. Automa. Comput.*, vol. 18, pp. 377-409, 2021.

[24] A. S. M. Htet and H. J. Lee, "Contactless Palm Vein Recognition Based on Attention-Gated Residual U-Net and ECA-ResNet," *Appli. Sci.*, vol. 13, no. 11, 2023.

[25] M. A. -E. Eladlani, S. Si Kaddour, L. Boubchir and B. Daachi, "On the Use of Convolutional Neural Networks for Palm Vein Recognition," *2022 IEEE Interna. Conf. on Big Data (Big Data)*, Osaka, Japan, 2022, pp. 3586-3592.

Breast Cancer Detection Auxiliary System

Leveraging Deep Learning and Mixed Reality

Szu-Yin Lin*, Ming-Chun Chien, Edwin Tiong Kwong Meng, Yu-Chien Wang, Yu-Yi Kuo, Che-Hsuan Lin
Computer Science and Information Engineering, National Ilan University, Taiwan
E-mail: szuyin@niu.edu.tw

Abstract — **According to the World Health Organization (WHO), in 2022, breast cancer is the most diagnosed cancer among women worldwide, irrespective of age. In Taiwan, it is the most prevalent cancer among women and has the fourth-highest mortality rate. Additionally, diagnosing breast cancer often takes a long time searching for symptoms. In this study, we propose utilizing artificial intelligence image analysis methods and mixed reality interfaces to track and compare breast cancer images. We employed a deep learning convolutional neural network to analyze and classify acquired images of benign or malignant cases. Thus, establishing breast cancer tracking and a diagnostic decision support system for physicians to reference would benefit the future medical field. By incorporating mixed reality technology, doctors can save time and reduce labor costs without being constrained by geographic limitations.**

I. INTRODUCTION

This study aims to integrate various technologies, including deep learning, image processing, and mixed reality, to develop a specialized breast cancer diagnostic system that helps address physicians' challenges in diagnosing this disease. Breast cancer is one of the most prevalent cancers among women worldwide, and early diagnosis and treatment are crucial for improving patients' survival rates and quality of life. However, the current diagnostic process for breast cancer often relies on subjective assessments, leading to reduced diagnostic accuracy. Therefore, there is a pressing need to develop an accurate, reliable, and visualized breast cancer diagnostic system to enhance diagnostic precision and alleviate the burden on medical staff.

This research aims to apply deep learning techniques to analyze and predict calcification-related features in breast cancer diagnosis. The Efficient Net B7 model enhances the accuracy of predicting benign and malignant breast cancer radiographs. Techniques such as histogram equalization and data augmentation, are used during the training phase. Furthermore, we have developed a case comparison and marking feature that facilitates the identification of discrepancies and abnormal areas between different reports. This provides physicians with more comprehensive diagnostic information. To ensure a seamless user experience, we utilize mixed reality devices as the user interface for the breast cancer diagnostic system. These devices offer portability and enable hands-free operation, empowering doctors to access data and diagnose without being limited to a specific location. The system operates seamlessly, with only a wireless network connection being required.

II. LITERATURE REVIEW

Breast cancer diagnosis is a critical area of medical research, with numerous studies dedicated to enhancing the accuracy and efficiency of diagnostic methods. This section reviews prior studies and relevant literature to gain insights into the current knowledge and approach to breast cancer diagnosis. In recent years, the application of artificial neural networks in medical image processing has garnered significant attention. For instance, Wessam et al. [1] developed a breast cancer diagnosis model based on convolutional neural networks, achieving remarkable accuracy. Furthermore, image processing techniques have found extensive use in breast cancer diagnosis. Petrick et al. [2] utilized image processing techniques to extract features from mammography images, improving diagnostic accuracy.

Despite the progress made in research thus far, there remain challenges and limitations in breast cancer diagnosis. The accuracy of existing models in predicting benign and malignant cases necessitates further improvement. Additionally, more research is required to enhance preprocessing techniques and image segmentation methods. Furthermore, applying adjunctive diagnostic systems in real clinical settings requires further investigation and validation.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

This study employs the CBIS-DDSM dataset [3] for experimentation and analysis. CBIS-DDSM is a publicly available medical image dataset comprising thousands of mammography images. This dataset provides valuable resources to validate the effectiveness of the developed adjunct diagnostic system in real-world scenarios. Integrating various technologies, such as deep learning, image processing, and mixed reality, would develop a specialized assistive diagnostic system for breast cancer that enhances the accuracy and efficiency of diagnosis.

III. METHODOLOGY

Fig. 1 shows the system architecture. The system in this study has two main functions: cancer prediction and X-ray image comparison. All image data used in this study are from the CBIS-DDSM dataset.

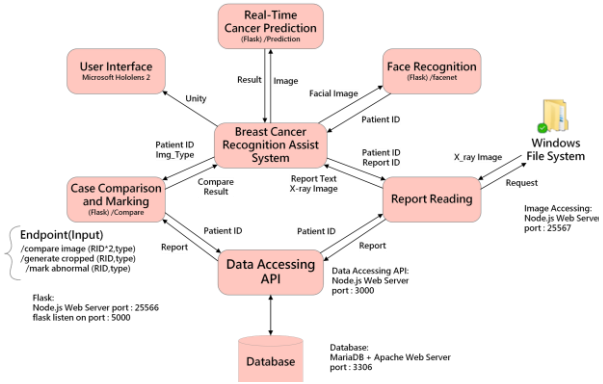
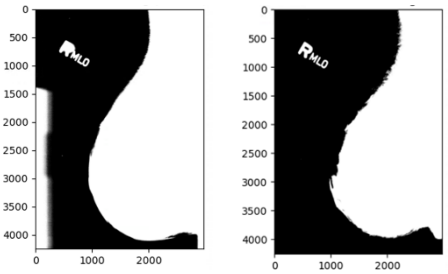


Fig. 1 System Architecture

A. Image Masking

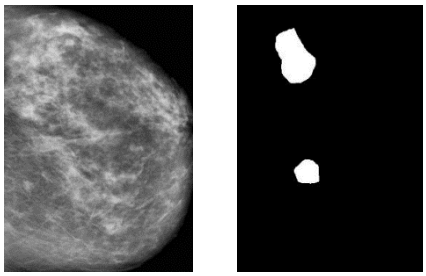
The Opening and Closing operations in morphology enable us to denoise [4], fill holes, and refine the edges of images. Opening and Closing require the assignment of a structural element and the specification of the number of iterations. Structural factors, such as square, oval, and more, can vary in size and shape. The number of iterations determines the strength and impact of the operation. The Opening operation removes small items, reduces noise, and smoothes out the edges, while the Closing operation fills holes and connects fragmented edges. In this study, we utilize OpenCV's function, cv2.morphologyEx(), to implement these operations. Image masking is a common technique for edge processing, image segmentation, and object detection. It significantly improves image quality and enhances readability.

We utilized OpenCV's cv2.morphologyEx() function to perform the Opening operation. This function requires specifying the structural element's shape, size, and number of iterations. Additionally, because the text in the corners of the X-ray image can affect subsequent processes, we used cv2.contourArea to locate and remove smaller contours representing the text. Subsequently, we employed cv2.drawContours to obtain the most significant contour, the image mask. This mask then creates an X-ray image that includes only the patient's breast. We can improve the quality and clarity of X-ray images in this study by employing these techniques and utilizing OpenCV's functions.



(a) (b)

Fig. 2 Morphology Operation (a) Opening (b) Closing

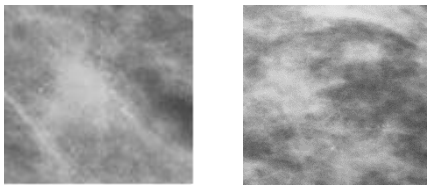


(a) (b)

Fig. 3 Image Data (a) Original Image (b) Corresponding ROI

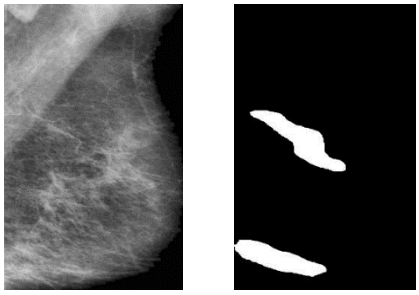
B. Cropped Image Generation

To facilitate physicians in observing the lesion area in detail and conducting subsequent benign and malignant classification, we use the contour of the ROI image to get the crop area. When the original image is appropriately positioned, the corresponding section is cropped and saved as a separate image to assist physicians in diagnosing. To accomplish this, we first employed the cv2.findContours function to identify the contours within the ROI. Then, using the cv2.boundingRect function, we obtained the boundary frame of the contour. With these coordinates, we could precisely crop only the abnormal portion of the X-ray image. The resulting output includes two cropped images: one from the original image and another from the corresponding ROI image. These cropped images aid in subsequent observation and diagnosis by medical professionals.



(a) (b)

Fig. 4 Cropped Image (a) cropped from the first contour (b) cropped from the second contour



(a) (b)

Fig. 5 Image Data (a) Original Image (b) Corresponding ROI

C. *Calculation of Contour Information of the Lesion Area*

We utilized the patient ROI in the dataset and employed OpenCV to calculate the lesion area information. This information was then added to the original image, allowing doctors to access additional reference information while observing the X-ray film, aiding in subsequent diagnosis. The following images display the original images and their corresponding ROIs to demonstrate the functionality. For each contour within the ROI, we drew a contour line and framed the boundary of the contour using the coordinates obtained from cv2.boundingRect on the original image. Subsequently, we utilized OpenCV's functions to calculate the contour length, width, area, and perimeter. Finally, we displayed this information on the original image. It is important to note that the values obtained from the drawing are in pixel units.

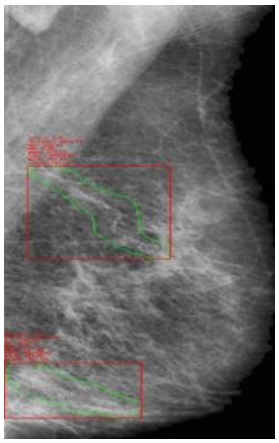


Fig. 6 Mammography image with the information marked.

D. *Comparison of Mammography Images*

To enable physicians to compare mammograms of the same patient at different times and identify any differences (lesion areas), we simulated the differences between the two images. We highlighted them in rectangular boxes in the second image. The following images display the original mammograms and our simulated difference maps. We calculated the differences between the two images using a metric subtraction approach, and the areas with variations were marked with red boxes. While this method works well with our simulated data, we have concerns about its effectiveness with real-world data. Therefore, we plan to develop a series of processes to deal with potential issues such as misalignment. For now, here are the results of the image comparison.

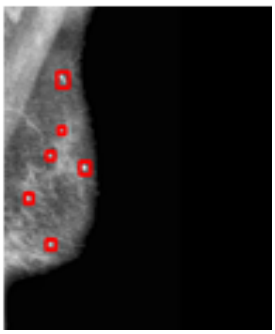


Fig. 7 Mammography image with differences marked.

IV. EXPERIMENTAL RESULTS

We have successfully developed a Breast Cancer Prediction model with a 70% accuracy rate for calcification data. Additionally, we have created a user interface for AR devices and a corresponding server for data access and computing service. We used the CBIS-DDSM dataset's crop images as our model training data. The data ratios for the training sets were as follows: Calc training set - 1002 benign cases and 544 malignant cases; Mass training set - 678 benign and 636 malignant cases. To solve the data imbalance problem, we employed downsampling, reducing benign data to match the number of malignant cases. To improve the image's visual quality to make it clearer and more distinctive for model training, we used histogram equalization to enhance the images [5]. Due to the limited number of samples available for training data relevant to medical applications, it becomes necessary to increase the data through data augmentation. This helps alleviate the issues caused by insufficient data. The data scaling process involves several steps. First, an image is taken as input. Then, the image is subjected to horizontal and vertical flipping, which creates mirror images. Next, the image is rotated 30 degrees clockwise while simultaneously flipping it. Subsequently, the image is rotated 60 degrees clockwise and flipped again. At this point, the resulting images from the first few steps are written or saved. At last, the data scaling process

concludes and produces the transformed images with various flips and rotations.

We chose six common CNN models used in medical imaging for prediction models: ResNet50, VGG16, DenseNet121, InceptionV3, EfficientNetB7, and MobileNet [6]. We took these models as a foundation and trained all of them using transfer learning techniques by incorporating customizable, fully connected layers. To quantify the models' performance and efficiency, we evaluated them using various metrics, including Precision, F1-Score, AUC, FPR, Accuracy, and Recall. These metrics helped us determine the most suitable models. Below are some of the training results:

Table 1 Calcified Histogram Equalization Model Training Data

Model	EfficientNetB7	ResNet50	MobileNet	VGG16	DenseNet121	InceptionV2
Sensitivity (%)	71.3564	75.9748	61.7605	68.6147	68.2539	59.8124
Precision (%)	70.2914	62.4552	65.0951	63.3155	63.4473	58.5038
F1-Score (%)	70.8199	68.5547	63.3839	65.8587	65.7629	59.1509
AUC	0.7721	0.7278	0.6840	0.7081	0.7071	0.6001
FRP	0.3015	0.4567	0.3311	0.3975	0.3932	0.4242
Accuracy	0.7092	0.6645	0.6616	0.6529	0.6500	0.5966
Recall	0.7135	0.7597	0.6176	0.6861	0.6825	0.5981

Table 2 Mass Histogram Equalization Model Training Data

Model	EfficientNetB7	ResNet50	MobileNet	VGG16	DenseNet121	InceptionV2
Sensitivity (%)	60.8311	61.3616	65.7382	64.3678	43.1476	65.2077
Precision (%)	74.9455	64.3486	61.2186	60.4149	62.2052	58.0480
F1-Score (%)	67.1547	68.5547	63.3839	62.3287	50.9527	61.4199
AUC	0.7768	0.6869	0.6824	0.6655	0.6353	0.6311
FRP	0.2033	0.3399	0.4164	0.4217	0.2621	0.4712
Accuracy	0.7285	0.6441	0.6273	0.6242	0.6109	0.5923
Recall	0.6083	0.6136	0.6573	0.6436	0.4314	0.652

The above results show that the best result of both is EfficientNetB7, which achieves 70% accuracy for quality and calcification data. Since this study focuses on proposing a complete diagnostic system, the current calcification accuracy rate of 70% is believed to be able to assist doctors in their diagnostic work to a certain extent.

V. Conclusion and Discussion

This study integrates technologies such as deep learning, image processing, and mixed reality to develop a breast cancer-specific assistive diagnostic system addressing the challenges doctors face in diagnosing breast cancer. Firstly, in terms of functionality, we employed artificial neural networks to predict breast cancer's benign or malignant nature from X-ray images. Through discussions with medical professionals, it became evident that existing research primarily focuses on cases involving the masses. Therefore, our system is primarily designed to identify calcifications. By utilizing the EfficientNet B7 model and employing techniques such as histogram equalization and data augmentation during training, we achieved an accuracy rate of 70%. Secondly, we have developed a case comparison and marking feature that enables the comparison of two reports, highlighting differences and abnormal regions. However, this feature requires the inclusion of region of interest (ROI) images in addition to the patient's breast radiographs. Since this study did not encompass the generation of ROI images, we hope future studies can fill this gap by developing accurate image segmentation models to provide reliable ROI data. Finally, we have successfully implemented the HoloLens 2 mixed reality device as a user interface for the Breast Cancer Assistive Diagnostic System. The mobility of the HoloLens 2 allows users to access data and make diagnoses regardless of their location, only requiring a wireless network connection.

Acknowledgment

This research was supported by the National Science and Technology Council, Taiwan, under Grant 109-2410-H-197-002-MY3 and 112-2410-H-197-002-MY2. Special thanks to Dr. Chin-Yau Chen, a breast surgeon at National Yang Ming Chiao Tung University Hospital, for discussing this with us during this research, allowing us to understand the needs of physicians and improve the system.

References

[1] Salama, W. M., & Aly, M. H. (2021). Deep learning in mammography images segmentation and classification: Automated CNN approach. *Alexandria Engineering Journal*, 60(5), 4701-4709.

[2] Petrick, N., Chan, H. P., Sahiner, B., & Helvie, M. A. (1999). Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Medical physics*, 26(8), 1642-1654.

[3] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1), 1-9.

[4] Dehghani, S., & Dezfooli, M. A. (2011). A method for improve preprocessing images mammography. *International Journal of Information and Education Technology*, 1(1), 90.

[5] Simi, V. R., Edla, D. R., Joseph, J., & Kuppili, V. (2020). Parameter-free fuzzy histogram equalisation with illumination preserving characteristics dedicated for contrast enhancement of magnetic resonance images. *Applied Soft Computing*, 93, 106364.

[6] Elkorany, A. S., & Elsharkawy, Z. F. (2023). Efficient breast cancer mammograms diagnosis using three deep neural networks and term variance. *Scientific Reports*, 13(1), 2663.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Efficient Reversible Data Hiding for 3D Mesh Models Based on Multi-LSB Substitution and Ring-prediction

Wanli Lyu* and Lulu Cheng*and Zhaoxia Yin† and Bin Luo*
* Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
E-mail:lwl@ahu.edu.cn; 2357610293@qq.com; luobin@ahu.edu.cn
† Shanghai Key Laboratory of Multidimensional Information Processing
East China Normal University, Shanghai, 200241, China
E-mail: zxyin@cee.ecnu.edu.cn

***Abstract*— Reversible data hiding (RDH) technology can hide secret data in multimedia to realize the purpose of secret communication, copyright and content protection. However, the current RDH methods for 3D models are challenging to balance embedding capacity and carrier distortion. This paper proposes an RDH method for 3D mesh models with high embedding capacity and low model distortion based on multiple least significant bit (multi-LSB) substitution and ring-prediction. The coordinates of the vertices are converted into specific format by preprocessing. The secret data are embedded in multi-LSB of converted coordinates by multi-LSB substitution. The cover 3D mesh model is reconstructed by correlation of 1-ring of neighborhood vertices. Experimental results show that the proposed method achieves higher embedding capacity and lower distortion than the state-of-the-art methods.**

***Index Terms*—3D mesh model, Reversible data hiding, Multiple least significant bits, Ring prediction**

I. Introduction

Reversible data hiding (RDH) technology can hide secret data in multimedia to realize secret communication, copyright, and content protection and recover the carrier after extracting secret data [1], [2]. This technology studies how to reversibly embed data into the carrier without introducing permanent distortion. The early RDH technology mainly uses the digital image as the carrier. RDH technology based on digital images has matured, and theoretical optimal solutions have been achieved. However, mature RDH technology of digital images cannot be directly applied to other carriers (such as video, audio, and 3D model). The 3D model is an emerging digital media after image, audio, and video, widely used in video game production, building structure display, and other fields. Research on RDH techniques for 3D models meets the urgent need for multimedia data security in complex and diverse social environments. According to different application scenarios, current RDH methods for 3D models can be divided into encryption domain and plaintext domain. RDH in encrypted domain (RDH-ED) method for 3D models [3]–[10] is proposed to protect carrier content from disclosure. This kind of method encrypts the carrier before embedding the data. Jiang et al. [3] first extended the RDH method on encrypted images to encrypted 3D mesh models. Yin et al. [5] proposed an RDH method for 3D mesh models based on multiple most significant bit (multi-MSB) prediction for the first time, which effectively improved the embedding capacity of 3D mesh models, and there are no corresponding attacks as there may be with digital images [11]. Embedding information in encrypted domain usually does not consider the deformation of the carrier caused by data embedding, and the average embedding rate of current reversible data hiding methods in encrypted domain of 3D models is as high as 33.15 bit per vertex (bpv) [10].

Different from the RDH-ED method for 3D models, the common method to protect privacy is encryption, but the special form of ciphertext encoding is messy and easy to attract the attention of attackers. RDH in plaintext domain pays more attention to the invisibility of the method. It is clear that covert storage requires both properties, namely, undetectability and reversibility. That is to say, this algorithm also focuses more on how to improve the embedding capacity while maintaining low distortion. The typical RDH method for images [12]–[26] can be divided into three categories: (1) histogram shifting [12], [19]–[23], (2) difference expansion [13], [14], [24]–[26], and (3) lossless compression [15]–[18]. Many ideas of RDH for multimedia are based on the classic method of RDH for digital images. In [27], DCT transformation is performed on the vertices cluster to embed the data by changing the high-frequency component coefficients. However, this method uses multiple vertices to embed 1-bit data, and the utilization rate of vertices is low. [28]–[30] uses compression algorithms to compress and encode 3D models and then embeds the secret data. In [31]–[36], secret data is embedded by modifying the difference between the predicted vertex coordinates and the actual vertex coordinates. However, controlling the cover model distortion is difficult when improving the embedding capacity by modifying the prediction error directly.

The performance of the RDH method for 3D models in

† Corresponding author
This research work is partly supported by National Natural Science Foundation of China (62172001,61872003,61860206004).

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

plaintext domain is evaluated by embedding capacity and carrier distortion. In most studies, researchers strike balance between these two indicators. To embed multi-bit data into vertices, the multi-MSB prediction method of [5] is applied to our method. Compared with [5], to control the distortion of the model, instead of embedding the data into the multi-MSB of the vertex coordinates, the proposed method embeds the secret data into the multiple least significant bit (multi-LSB) that has the least impact on the vertex through vertex transformation. Different from the existing work of RDH in plaintext domain for 3D models, this paper does not directly calculate the difference between the coordinates of adjacent vertices, but first normalized to a 32-bit binary form, which provides more embeddeable room for secret data embedding. We propose a new method to effectively improve the embedding capacity and reduce the introduced distortion. The main contributions of this paper are summarized as follows:

(1) The data embedding room is obtained by using the correlation of local regions of 3D mesh model, and multi-bit data can be embedded in each embeddable vertex by multi-LSB substitution to improve embedding capacity.

(2) The vertex coordinates are converted to a specific format, and the carrier distortion is controlled by modifying the position that has the least impact on the vertex to embed the secret data.

(3) The cover 3D mesh model is restored by predicting the modified vertices through the multi-LSB of the vertices of 1-ring neighborhood and realizing the separation of data extraction and model reconstruction.

The rest of this paper is organized as follows. Section II introduces the related works. An efficient RDH method for 3D mesh model is designed in Section III. The experimental results are discussed in Section IV, and finally we conclude this paper in Section V.

II. Related Works

To embed the secret data into the plaintext domain, it is necessary to better balance the distortion caused by the data embedding and the embedding capacity. Most works are based on difference expansion to hide secret data into unencrypted 3D mesh models. This kind of method can effectively embed secret data with low distortion by using the spatial correlation of adjacent vertices.

An RDH method for 3D mesh models based on prediction error expansion is proposed in [31]. The method predicts the position of the vertex by traversing the center of mass of the adjacent vertices, and the secret data is embedded in the difference between the predicted and actual position. In addition, too large prediction errors will not be enlarged to control the distortion introduced by selecting the threshold adaptively. This method extends the prediction error extension method to 3D mesh models for the first time, but its embedding capacity and model loss still have a lot of room for improvement. At the maximum embedding capacity, the SNR of the 3D mesh models tested by this method is all less than 35 dB. The three-dimensional model is spatially distributed, so the x, y, and z axis coordinates of each vertex can be used to embed secret data. To make full use of the high-dimensional characteristics of the 3D mesh model, Jiang et al [34] proposed to divide vertices according to odd-even property and obtained the prediction error by using the geometric similarity of adjacent vertices, combined every three neighboring prediction errors into a triplet of prediction errors, and constructed a three-dimensional prediction error histogram with lower entropy than the one-dimensional prediction error histogram by using the correlation between the prediction errors. According to the estimated optimal probability transfer matrix, the prediction error histogram is modified to embed the data to achieve the minimum distortion. This method can observe the obvious deformation of the model when the embedding capacity reaches 0.13bpv. In 2019, Zhang et al. [32] proposed to define the smoothness of the local area by using local average curvature, and priority is given to embedding data in smooth cells. The twice-layered strategy uses candidate vertices and reference vertices interchangeably to improve embedding capacity. This method effectively improves the embedding capacity, but the model still produces distortion visible to human eyes when the embedding capacity reaches 1.4 bpv. To further improve the embedding performance, Girdhar et al. [36] embedded data by selecting the average difference of appropriate vertex pairs, and logical mappings are introduced to randomly embed data into vertex coordinates. The secret data is distributed among the three coordinates of all vertices to control the distortion of the cover model, which effectively controls the distortion and greatly improves the embedding capacity. In 2022, Bhardwaj[30] proposed a more effective RDH method in the plaintext domain for 3D models, this method first uses hierarchical absolute moment block truncation coding compression model, and binary secret data is embedded in a particular block through interchanged operations. The model distortion is effectively controlled by embedding data according to the characteristics of each block, and the embedding capacity is higher than the existing RDH methods for 3D mesh model in plaintext domain.

III. Proposed Method

This paper proposes an RDH method for 3D mesh models based on multi-LSB substitution and ring-prediction. Figure. 1 is a sketch describing the proposed method. The proposed method is mainly divided into three parts. (1) The cover model M is first preprocessed by the sender. (2) The senders embed the secret data into the model through prediction error detection, generating data-embedded model M'. (3) The receiver with different permissions can extract secret data and reconstructed cover model $\widetilde{M}$, respectively.

A. Preprocessing

1) Vertex coordinates transformation: This paper studies the most popular triangular 3D meshes. The 3D mesh model is a carrier distributed in space, and the vertex coordinates of an uncompressed 3D mesh model are represented as valid 6-bit floating-point numbers. For example,

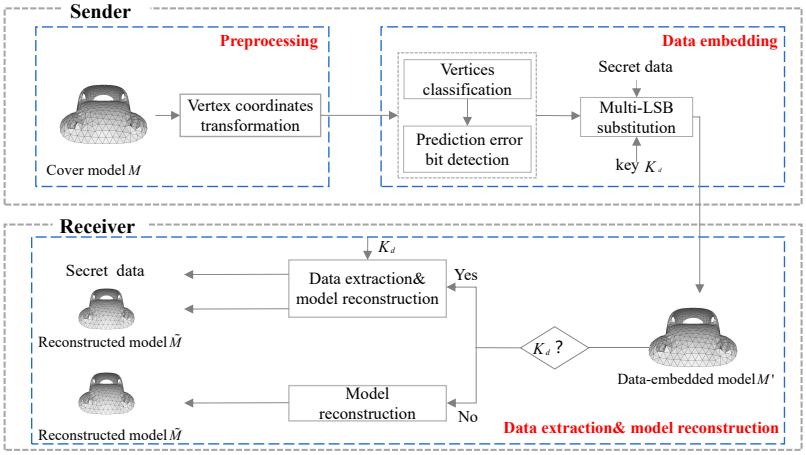


Fig. 1. Framework of the proposed method.

$v = \{0.295122, 0.338337, 0.103678\}$. Most applications of 3D models should be carried out at an acceptable level of accuracy, without the need to operate on every bit of floating-point [37]. So we do not directly modify vertex coordinates and then embed secret data. All vertices position coordinates are normalized to the accuracy of u bits and represented as integers between 0 and $2^u - 1$, where $u \in [1, 33]$. Vertex $v_{i,j}$ is normalized to $v'_{i,j}$,

$$v'_{i,j} = \lfloor v_{i,j} \times 10^u \rfloor, \quad j = x, y, z, \tag{1}$$

where i represents i_{th}th vertex, u is truncation coefficient.

To simplify data embedding operations, vertex coordinates are first converted into 32-bit binary stream form according to IEEE754 standards [38]. In modifying such a format the coordinates of the multi-LSB have little effect on the vertices according to the characteristics of the IEEE754 standard. Figure. 2 shows the transformed vertex coordinate and is calculated by

$$v'_{i,j} = (-1)^s \times \text{mant} \times 2^{e-127}, \tag{2}$$

where s represents the positive (0) or negative (1) of vertex coordinate, e is exponent (8 bits), and mant is mantissa (23 bits).

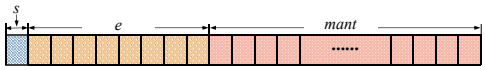


Fig. 2. IEEE754 format for vertex coordinates.

B. Data embedding

1) Vertices classification: The sender modifies part of the coordinates of the 3D mesh model to embed data. Since the neighborhood of a vertex contains multiple triangles, if the vertex is modified, the neighborhood vertex should remain unchanged to reconstruct the vertex coordinates reversibly. Therefore, vertices are divided into "embedded set" S_e for embedding data and "reference set" S_r for recovering modified vertices. Note that vertices in "reference set" are not modified during data embedding, and vertices are traversed in ascending order according to the topological structure of the triangle face. The triangle face list is denoted as $F = (f_1, f_2, \cdots, f_v)$, where $f_i = (v_{i,x}, v_{i,y}, v_{i,z})$ is any triangle face, $f_t = (v_{t,x}, v_{t,y}, v_{t,z})$ is next triangle face to be traversed. The initial states of S_e and S_r are both empty sets. If the vertices in f_t are neither in S_e nor in S_r, add $v_{t,x}$ to S_e and add $v_{t,y}$, $v_{t,z}$ to S_r. If there is a vertex $v_{t,j}$ is contained in S_e but not in S_r, add other vertices to S_r. If there are at least one vertex $v_{t,j}$ is not contained in S_e but contained in S_r, add the first vertex in f_t except $v_{t,j}$ to S_e, and add other vertices to the S_r. Figure. 3 shows the "Beetle" model and its magnified local area. The vertices marked red, blue, and yellow represent vertices in "embedded set", vertices in "reference set", and the vertices to be traversed.

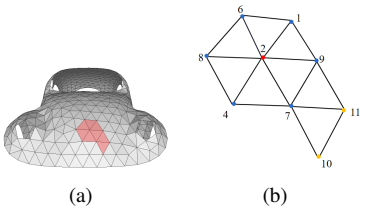


Fig. 3. Local vertex distribution of "Beetle" 3D mesh model.

2) Prediction error bit detection: The data encryption key K_d is used to generate a random bitstream to encrypt secret data by XOR operation to improve data security. Based on the characteristics of the IEEE754 standard, modifying the multi-LSB of the transformed vertex coordinates has the least impact on the vertex coordinates, so secret data is embedded in the multi-LSB of coordinates to control the distortion of the cover mesh model. The vertices in "embedded set" are predicted by multiple vertices in "reference set" of 1-ring of neighbors. The

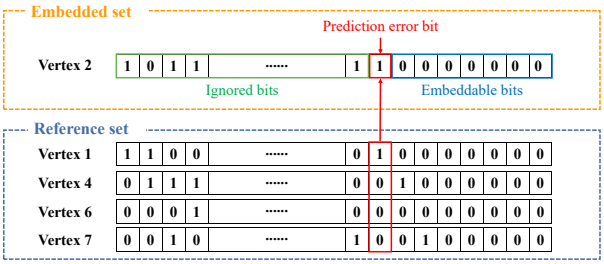


Fig. 4. Prediction error bit detection.

ring-prediction results can be obtained by

$$p_{i,j}^{k}=\begin{cases}1, & N_{1,k}\geq N_{0,k}\\ 0, & N_{0,k}>N_{1,k}\end{cases}\qquad k=1,2,\ldots,23, \tag{3}$$

where $p_{i,j}^{k}$ is the k_{th} LSB predicted value of vertex coordinate. $N_{1,k}$ denotes the number of '1' in the k_{th} LSB of vertices in S_r. $N_{0,k}$ denotes the number of '0' in the k_{th} LSB of vertices in S_r. If the number of '1' is greater than or equal to the number of '0' bit, then the predicted value the LSB of the current vertex coordinate in S_e is 1, otherwise, predicted value is 0. Compare ${v_{i,j}}'^{rLSB}$ and $p_{i,j}^{rLSB}$ sequentially from LSB to MSB until a certain bit is different,

$$\arg\max_{r}\ {v_{i,j}}'^{rLSB}=p_{i,j}^{rLSB},\quad r=1,2,\ldots,23, \tag{4}$$

where $v'_{i,j}$ is integer coordinate, ${v_{i,j}}'^{rLSB}$ and $p_{i,j}^{rLSB}$ are represent the r_{th}LSB of the 32-bit IEEE754 form coordinates and its predicted value $p_{i,j}$.

3) Multi-LSB substitution: Through the above operation, the embeddable length of each vertex can be obtained, and the multi-LSB that are correctly predicted are substituted by the secret data. For example, vertex 2 is the vertex in "embedded set", and vertices numbered 1, 4, 6, and 7 are the vertices of the 1-ring neighborhood of vertex 2. Figure. 4 describes the process of prediction error detection. The 8_{th} LSB of vertex 2 is inconsistent with the predicted result. The secret data is embedded in 7-LSB of the vertex coordinates of the "embedded set" by multi-LSB substitution. A 3D mesh model is a carrier in space, and each vertex is composed of x, y, and z coordinates. Assuming that r_1, r_2, and r_3 represent the embeddable lengths of the x-, y-, and z-axis coordinate, respectively, the embeddable length of the vertex is $r=\min\{r_1,r_2,r_3\}$.

C. Data extraction and model reconstruction

The sender classifies vertices into "embedded set" and "reference set" by ascending traversal according to the topological relation of neighborhood vertices. Similarly, the receiver also classifies the vertices into "embedded set" and "reference set" in the above way. As shown in Figure. 5, the receivers with different permissions implement data extraction and model recovery, respectively, in the following two cases:

Case 1: the receiver with the encryption key K_d can extract and decrypt secret data directly. The vertex coordinates in "embedded set" can be reconstructed using the ring-prediction rules of Eq.(3). Then the model $\widetilde{M}$ can be reconstructed by inverse coordinate transformation.

Case 2: legitimate receivers without the encryption key K_d can recover the cover model by predicting the modified vertices in "embedded set" using multi-LSB of the vertices in "reference set" of the 1-ring neighborhood. Then, the cover model can be reconstructed by inverse coordinate transformation. The secret data cannot be decrypted because there is no encryption key, so the original data cannot be extracted.

As the above analysis, both the secret data and the cover model at the receiver end are successfully retrieved in a separable manner.

IV. Experiment

The commonly-used princeton shape benchmark [39] which consists of 380 different 3D objects, and we selected four common 3D mesh objects: Cow, Beetle, Elephant, and Mushroom are adopted to test the validity of the proposed method. To demonstrate the performance of the proposed method, the corresponding analysis and discussions are given in this section in terms of analysis of embedding capacity, imperceptibility, choice of truncation coefficient u, and comparison with the state-of-the-art works.

A. Embedding capacity

In the Cartesian coordinate system, multi-dimensional embedding is realized by modifying the coordinate values of the x-, y-, and z-axis. To minimize the distortion of the cover model caused during data embedding, we choose the mantissa of vertices to embed the secret data; that is, the maximum embeddable length of each direction coordinate of each vertex is 23. The embeddable length of each vertex can be obtained by prediction error detection. In this paper, the embedding threshold l is set , $l\in[1,23]$, to mark the embeddable length of vertices. Vertices with embeddable length less than l are judged to be predictive errors and therefore not used for embedding data. The embedding rate can be calculated as

$$ER=\frac{3l\left(n_e-n_{error}\right)}{n}, \tag{5}$$

where n_e represents the number of embeddable vertices, n_{error} is the number of vertices with embeddable length less than l, and n is the total number of vertices of 3D mesh model.

B. Imperceptibility

The vertex coordinates are modified when data is embedded, making the difference between the data-embedded and the cover models. We measure the imperceptibility of the difference by subjective evaluation of human eye evaluation and objective evaluation of statistical analysis. The less easily the data embedding operation is perceived, the lower the distortion caused by the data embedding to the carrier. Experiments are conducted on four models. Figure. 6 shows the appearance of

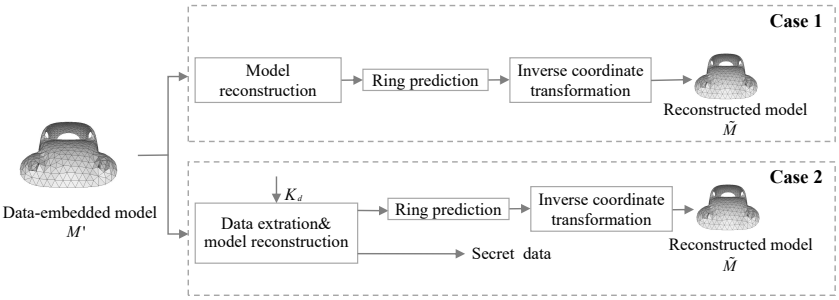


Fig. 5. Two cases of data extraction and model reconstruction for the receiver.

four 3D mesh models of each phase. The first row contains four original 3D mesh models for testing, and the second row contains the data-embedded 3D mesh models. 30 observers are invited to evaluate the difference among the data-embedded and the cover models. According to the feedback, 30 observers cannot distinguish the difference between the cover, data-embedded. Signal noise ratio (SNR), Hausdorff distance, the maximum moving distance, and average moving distance were introduced as objective evaluation indicators of difference. Statistical results under the maximum embedding rate are listed in Table I. SNR measures geometric distortion due to added noise and Hausdorff distance to measure the degree of deviation of 3D mesh models vertices after embedding secret data. The closer the SNR value is to ∞, the more similar the structure of the two models. The closer the Hausdorff distance is to 0, the smaller the difference between the two models. SNR is calculated as

$$\text{SNR} = 10\lg\frac{\sum_{i=1}^{N}\left[(v_{i,x}-\bar{v}_x)^2+(v_{i,y}-\bar{v}_y)^2+(v_{i,z}-\bar{v}_z)^2\right]}{\sum_{i=1}^{N}\left[(\hat{v}_{i,x}-v_{i,x})^2+(\hat{v}_{i,y}-v_{i,y})^2+(\hat{v}_{i,z}-v_{i,z})^2\right]}, \tag{6}$$

where $\bar{v}_x$, $\bar{v}_y$, $\bar{v}_z$ are the mean coordinates of the models; $v_{i,x}$, $v_{i,y}$, $v_{i,z}$ are the original coordinates of v_i; and $\hat{v}_{i,x}$, $\hat{v}_{i,y}$, $\hat{v}_{i,z}$ are the modified coordinates of v_i, and N is the total number of vertices. Hausdorff distance is the distance between two subsets of a metric space. It can be represented by

$$HD(W,V)=\max_{w\in W}\left\{\min_{v\in V}\{d(w,v)\}\right\}, \tag{7}$$

where W and V are any two sets in Euclidean space, and w and v are two arbitrary vertices in these two sets respective. $d(\cdot)$ is the distance between two vertices. In addition, the maximum moving distance $MaxD$ of the vertices in cover mesh model and the average moving distance $AvgD$ are calculated by

$$MaxD=\max\left(\sqrt{(v_{ix}-v'_{ix})^2+(v_{iy}-v'_{iy})^2+(v_{iz}-v'_{iz})^2}\right), \tag{8}$$

$$AvgD=\frac{1}{N}\sum_{i=0}^{N-1}\sqrt{(v_{ix}-v'_{ix})^2+(v_{iy}-v'_{iy})^2+(v_{iz}-v'_{iz})^2}, \tag{9}$$

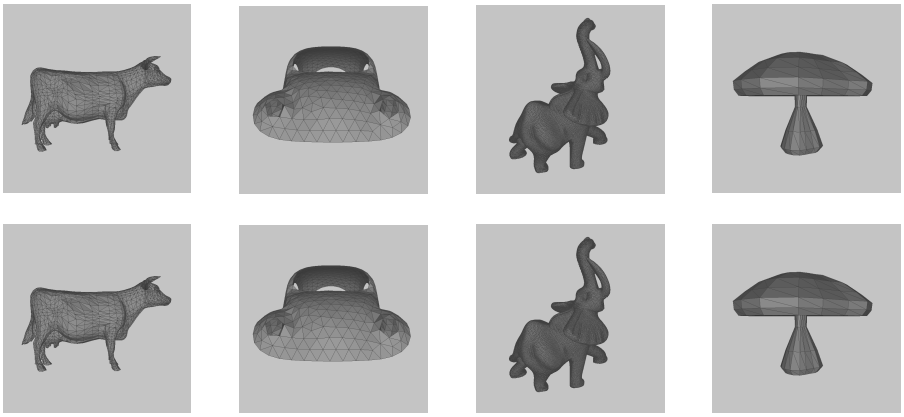
where v_{ix}, v_{iy}, and v_{iz} represent the coordinates of the cover 3D mesh model, v'_{ix}, v'_{iy}, and v'_{iz} are the coordinates of the data-embedded 3D mesh model, and N is the total number of vertices. As shown in Table I, the statistical analysis results show that the difference between the cover mesh model and the data-embedded mesh model is only a small order of magnitude. The above experimental results and analysis effectively demonstrate that our method introduces lower distortion.

TABLE I
STATISTICAL ANALYSIS OF DISTORTION INTRODUCED BY DATA EMBEDDING FOR TEST MODELS

Models	Embedding Rate	Hausdorff Distance	SNR	MaxD	AvgD
Cow	7.93	1.60e-05	100.39	3.18e-04	1.10e-07
Beetle	8.45	1.51e-05	95.07	1.82e-04	1.84e-07
Elephant	8.28	2.80e-05	93.60	10.50e-04	41.90e-07
Mushroom	9.19	1.50e-05	100.50	0.94e-04	4.16e-07

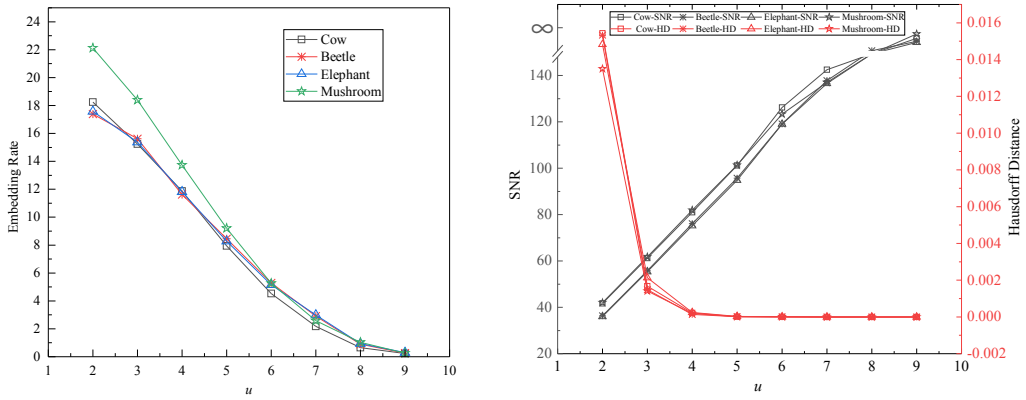
C. Choice of truncation coefficient u

In the proposed method, u is an important parameter that determines the precision of vertex coordinates. The embedding capacity and the quality of the data-embedded model are affected by u. In general, when u is 1, the decimal reservation precision is too low to apply to most application scenarios and has no practical significance. Therefore, the proposed method takes the common precision of u for paper. The correlation between u and embedding rate is explored, as shown in Figure. 7(a). With the change of u, the embedding rate also changes to different degrees. The spatial distribution of 3D mesh models with different shapes is disorganized. As u increases, the vertex coordinates become more disordered, so the correlation of the LSBs of vertex coordinates in the 1-ring neighborhood decreases accordingly. Figure. 7(b) show the Hausdorff distance and SNR of the cover and data-embedded mesh model on different u under maximum embedding rate on four test models. As u increases, the Hausdorff distance approaches 0, and SNR approaches ∞. As u increases, the embedding rate gradually decreases, and less distortion is introduced. To better balance the embedding capacity with the introduced distortion, the truncation coefficient u = 5 is selected. According to the different application scenarios, the sender can choose the appropriate truncation coefficient according to the actual requirements.



(a) (b) (c) (d)

Fig. 6. Visualization of four 3D mesh models in different stages ($u = 5$): The first row contains four original 3D mesh models for testing, and the second row contains the data-embedded 3D mesh models.



(a) (b)

Fig. 7. Influence of truncation coefficient u on embedding rate and imperceptibility.

D. Comparison with the state-of-the-art works

To better highlight the effectiveness of our method, the performance of the proposed method is compared with several similar methods. Table II shows the experimental results of our method and other methods on the dataset. Jiang et al. [34] constructed prediction error histogram with lower entropy and obtained the prediction error triplet. The coding optimization embedded data is constructed to achieve the optimal prediction error histogram modification, however, the method is embedding capacity less than 1.5 bpv. Zhang et al. [32], [35] proposed two RDH methods successively. The first emthod is a hybrid prediction method based on the uniformity factor, which produces a small prediction error. The embeddable space can be obtained by using overlapping segmentation and adaptive embedding in the multi-layer embedding process. The second method is an improved RDH based on prediction error expansion and ordering. However, both methods produce significant distortion when the embedding capacity exceeds 1.5 bpv. Girdhar et al. [36] embedded secret data using the difference between the coordinates of the vertex pairs on the selected coordinate axis. Their method introduces chaotic logical mapping to select vertex coordinates embedded in secret data and extract secret data. The secret data can be extracted using the logical map without restoring the model, thus realizing the separability of the method. The embedding rate of this method is affected by the number of adjacent embeddable vertices and threshold parameters. The embedding rate of their method still has much room to improve. In [30], hierarchical absolute moment block truncation coding is used to compress the mesh blocks and then embed secret data in the compressed domain. The method proposed by Bhardwaj [30] greatly improves the embedding rate of secret data in plaintext domain of the 3D model. The receiver can reconstruct the cover model using the designed embedding rules without extracting secret data. However, the method will generate auxiliary data in the process of data embedding. In our method, vertices are divided into two spatially adjacent sets. Data embedding modifies the coordinates of the vertices in "embedded set" and recovers the modified vertices by the vertices in "reference set" without extracting secret data. After the vertex coordinates in "embedded set" are converted into 32-bit IEEE754 format,

the embeddable length of the vertices in "embedded set" is obtained by comparing the consistency of the coordinates of the embeddable vertices with their reference vertices. Each vertex coordinate can embed multi-bit data to effectively improve each vertex's embedding capacity. These modified bits have minimal impact on the coordinates and effectively control the distortion of the cover model. Receivers with different permissions can extract secret data and reconstruct the cover model. So the separability of data extraction and model reconstruction of our method is satisfied. As shown in Table II, our method achieves higher embedding rate and better imperceptibility than other methods.

TABLE II
COMPARISON OF FEATURES WITH THE STATE-OF-THE-ART METHODS

Methods	Features				
	Embedding Rate	Hausdorff Distance	Data Error(%)	SNR(dB)	Separable
Jiang et al.[34]	1.29	-	0	46.78	No
Zhang et al.[32]	1.40	-	0	48.00	No
Zhang et al.[35]	1.50	-	0	40.00	No
Girdhar et al.(k=1)[36]	1.50	0.2610	0	98.74	Yes
Girdhar et al.(k=2)[36]	3.00	0.3360	0	97.50	Yes
Girdhar et al.(k=3)[36]	4.50	0.4829	0	95.08	Yes
Bhardwaj[30]	6.94	0.2358	0	96.40	Yes
Proposed method	**8.08**	**0.000017**	0	**99.81**	Yes

V. CONCLUSIONS

In the paper, an RDH method for 3D mesh model with higher embedding rate and lower distortion is proposed based on multi-LSB substitution and ring-prediction. The vertex coordinates are converted to a specific format, and the multi-LSB with the least impact on the vertices is selected to embed the data by bit substitution. Each embeddable vertex can embed multi-bit data without significant distortion to the cover model. 3D mesh model can be reconstructed without extracting secret data by exploring the correlation of vertices in the local region. Experimental results show the proposed method outperforms state-of-the-art methods in embedding rate and model distortion. However, the proposed method is not yet against malicious attacks. In follow-up research, it is urgent to enhance the robustness of the method against external attacks. We also consider using genetic algorithm to find the optimal vertex embeddable set to make full use of the vertex topological relationship of 3D model.

REFERENCES

[1] J. M. Barton, "Method and apparatus for embedding authentication information within digital data," *United States Patent, 5,646,997*, 1997.

[2] Y.-Q. Shi, X. Li, X. Zhang, H.-T. Wu, and B. Ma, "Reversible data hiding: Advances in the past two decades," *IEEE Access*, vol. 4, pp. 3210–3237, 2016.

[3] R. Jiang, H. Zhou, W. Zhang, and N. Yu, "Reversible data hiding in encrypted three-dimensional mesh models," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 55–67, 2017.

[4] N. Xu, J. Tang, B. Luo, and Z. Yin, "Separable reversible data hiding based on integer mapping and msb prediction for encrypted 3d mesh models," *Cognitive Computation*, vol. 14, pp. 1172–1181, 2022.

[5] Z. Yin, N. Xu, F. Wang, L. Cheng, and B. Luo, "Separable reversible data hiding based on integer mapping and multi-msb prediction for encrypted 3d mesh models," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Springer, 2021, pp. 336–348.

[6] M. Shah, W. Zhang, H. Hu, H. Zhou, and T. Mahmood, "Homomorphic encryption-based reversible data hiding for 3d mesh models," *Arabian Journal for Science and Engineering*, vol. 43, no. 12, pp. 8145–8157, 2018.

[7] B. J. Van Rensburg, P. Puteaux, W. Puech, and J.-P. Pedeboy, "Homomorphic two tier reversible data hiding in encrypted 3d objects," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 3068–3072.

[8] Y.-Y. Tsai, "Separable reversible data hiding for encrypted three-dimensional models based on spatial subdivision and space encoding," *IEEE Transactions on Multimedia*, vol. 23, pp. 2286–2296, 2020.

[9] W.-L. Lyu, L. Cheng, and Z. Yin, "High-capacity reversible data hiding in encrypted 3d mesh models based on multi-msb prediction," *Signal Processing*, vol. 201, p. 108 686, 2022.

[10] Y. Tang, L. Cheng, W. Lyu, and Z. Yin, "High capacity reversible data hiding for encrypted 3d mesh models based on topology," in *Digital Forensics and Watermarking: 21st International Workshop, IWDW 2022, Guilin, China, November 18-19, 2022, Revised Selected Papers*, Springer, 2023, pp. 205–218.

[11] I. C. Dragoi and D. Coltuc, "On the security of reversible data hiding in encrypted images by msb prediction," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 187–189, 2020.

[12] Z. Ni, Y.-Q. Shi, N. Ansari, and W. Su, "Reversible data hiding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 354–362, 2006.

[13] I. Caciula, H. G. Coanda, and D. Coltuc, "Multiple moduli prediction error expansion reversible data hiding," *Signal Processing: Image Communication*, vol. 71, pp. 120–127, 2019.

[14] G. Fan, Z. Pan, E. Gao, X. Gao, and X. Zhang, "Reversible data hiding method based on combining ipvo with bias-added rhombus predictor by multi-predictor mechanism," *Signal Processing*, vol. 180, p. 107 888, 2021.

[15] M. U. Celik, G. Sharma, A. M. Tekalp, and E. Saber, "Lossless generalized-lsb data embedding," *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 253–266, 2005.

[16] M. U. Celik, G. Sharma, and A. M. Tekalp, "Lossless watermarking for image authentication: A new frame-

work and an implementation," *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 1042–1049, 2006.

[17] X. Zhang, Z. Qian, G. Feng, and Y. Ren, "Efficient reversible data hiding in encrypted images," *Journal of Visual Communication and Image Representation*, vol. 25, no. 2, pp. 322–328, 2014.

[18] W. Zhang, X. Hu, X. Li, and N. Yu, "Recursive histogram modification: Establishing equivalency between reversible data hiding and lossless data compression," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2775–2785, 2013.

[19] X. Li, B. Li, B. Yang, and T. Zeng, "General framework to histogram-shifting-based reversible data hiding," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2181–2191, 2013.

[20] G. Gao, X. Wan, S. Yao, Z. Cui, C. Zhou, and X. Sun, "Reversible data hiding with contrast enhancement and tamper localization for medical images," *Information Sciences*, vol. 385, pp. 250–265, 2017.

[21] B. Ou and Y. Zhao, "High capacity reversible data hiding based on multiple histograms modification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2329–2342, 2019.

[22] X. Gao, Z. Pan, E. Gao, and G. Fan, "Reversible data hiding for high dynamic range images using two-dimensional prediction-error histogram of the second time prediction," *Signal Processing*, vol. 173, p. 107 579, 2020.

[23] T. Zhang, C. Yang, S. Weng, and T. Hou, "Adaptive multi-histogram reversible data hiding with contrast enhancement," *Journal of Visual Communication and Image Representation*, vol. 89, p. 103 637, 2022.

[24] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 890–896, 2003.

[25] X. Liao and C. Shu, "Reversible data hiding in encrypted images based on absolute mean difference of multiple neighboring pixels," *Journal of Visual Communication and Image Representation*, vol. 28, pp. 21–27, 2015.

[26] I. C. Dragoi, I. Caciula, and D. Coltuc, "Improved pairwise pixel-value-ordering for high-fidelity reversible data hiding," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 1668–1672.

[27] H. Luo, Z.-M. Lu, and J.-S. Pan, "A reversible data hiding scheme for 3d point cloud model," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, IEEE, 2006, pp. 863–867.

[28] Z. Sun, Z.-M. Lu, and Z. Li, "Reversible data hiding for 3d meshes in the pvq-compressed domain," in *2006 International Conference on Intelligent Information Hiding and Multimedia*, IEEE, 2006, pp. 593–596.

[29] Z.-M. Lu and Z. Li, "High capacity reversible data hiding for 3d meshes in the pvq domain," in *International Workshop on Digital Watermarking*, Springer, 2007, pp. 233–243.

[30] R. Bhardwaj, "Efficient separable reversible data hiding algorithm for compressed 3d mesh models," *Biomedical Signal Processing and Control*, vol. 73, p. 103 265, 2022.

[31] H.-T. Wu and J.-L. Dugelay, "Reversible watermarking of 3d mesh models by prediction-error expansion," in *2008 IEEE 10th Workshop on Multimedia Signal Processing*, IEEE, 2008, pp. 797–802.

[32] Q. Zhang, X. Song, T. Wen, and C. Fu, "Reversibility improved data hiding in 3d mesh models using prediction-error expansion and sorting," *Measurement*, vol. 135, pp. 738–746, 2019.

[33] Y.-H. Huang and Y.-Y. Tsai, "A reversible data hiding scheme for 3d polygonal models based on histogram shifting with high embedding capacity," *3D Research*, vol. 6, no. 2, pp. 1–12, 2015.

[34] R. Jiang, W. Zhang, D. Hou, H. Wang, and N. Yu, "Reversible data hiding for 3d mesh models with three-dimensional prediction-error histogram modification," *Multimedia Tools and Applications*, vol. 77, no. 5, pp. 5263–5280, 2018.

[35] Q. Zhang, X. Song, T. Wen, and C. Fu, "Reversible data hiding for 3d mesh models with hybrid prediction and multilayer strategy," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 29 713–29 729, 2019.

[36] A. Girdhar and V. Kumar, "A reversible and affine invariant 3d data hiding technique based on difference shifting and logistic map," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 12, pp. 4947–4961, 2019.

[37] M. Deering, "Geometry compression," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 1995, pp. 13–20.

[38] P. Markstein, "The new ieee-754 standard for floating point arithmetic," in *Dagstuhl Seminar Proceedings*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2008.

[39] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Proceedings Shape Modeling Applications, 2004.*, IEEE, 2004, pp. 167–178.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

MAEDefense: An Effective Masked AutoEncoder Defense against Adversarial Attacks

Wanli Lyu* and Mengjiang Wu*and Zhaoxia Yin† and Bin Luo*
* Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
E-mail:lwl@ahu.edu.cn; mengjiangwu@stu.ahu.edu.cn; luobin@ahu.edu.cn
† Shanghai Key Laboratory of Multidimensional Information Processing
East China Normal University, Shanghai, 200241, China
E-mail: zxyin@cee.ecnu.edu.cn

***Abstract*— Recent studies have demonstrated that deep neural networks (DNNs) are vulnerable to attacks when adversarial perturbations are added to the clean samples. Reconstructing clean samples under the premise of inputting adversarial perturbations is a challenging task. To address this issue, this paper proposes a Mask AutoEncoder Defense (MAEDefense) framework to counter adversarial attacks. Firstly, the adversarial sample is divided into two complementary masked images. Secondly, the two masked images with carefully crafted adversarial noise locations are reassigned to non-adversarial noise locations. Finally, the two reconstructed images are pixel-wise fused (weighted average) to obtain a "clean image". The proposed method requires no external training and is easy to implement. Experimental results show that the proposed method significantly defends against white-box attacks and black-box transferable attacks compared with state-of-the-art methods.**

I. Introduction

Due to their success in vision tasks such as image classification[1], target detection [2], target tracking [3], and image segmentation[4], deep neural networks (DNNs) have been widely used in critical areas such as autonomous driving [5]and face recognition[6]. However, recent studies have shown that deep neural networks are vulnerable to adversarial attacks and can be easily fooled [7]. These attacks are difficult for the human visual system to detect. Adversarial samples are generated by deliberately introducing imperceptible perturbations to clean samples, causing the model to produce false results with a high level of confidence. Furthermore, research has shown that adversarial samples can also exist in the physical world[6], posing significant security risks. Therefore, considering the security of deep neural networks has become a necessity rather than an option.

Numerous defense methods have been proposed to counter adversarial attacks. These defense mechanisms can be broadly categorized into two main types: robustness-based defense and input transformation-based defense. Robustness-based defense aims to improve the performance of deep convolutional neural network (DCNN) classifiers by retraining them and utilizing augmented datasets that include adversarial examples.[7]–[10]. These defense methods demonstrate a strong resistance against specific attacks that generate adversarial examples. However, these defense methods are typically time-consuming and computationally expensive when dealing with large-scale, high-dimensional image datasets (such as ImageNet[11]). Additionally, robust defense methods, when faced with unknown attacks, require retraining of the classifier, which can lead to a decrease in the accuracy of recognizing clean examples[7], [10]. Unlike robust defense methods, input transformation-based defense methods [12]–[15] are more practical. However, most existing methods heavily rely on external priors learned from training datasets, while overlooking the rich intrinsic priors inherent in the input data. For instance, methods like PixelDefend[16] and ComDefend [12]transform adversarial examples into "clean samples" and train models using adversarial examples as input. Similarly, DefenseGAN[17] utilizes generative adversarial networks (GANs) to mitigate the impact of adversarial perturbations. Defense methods that rely on learning external priors suffer from the issue of significantly reducing the defensive capabilities of the model when the counter-attack method changes or when the attack method is unknown.

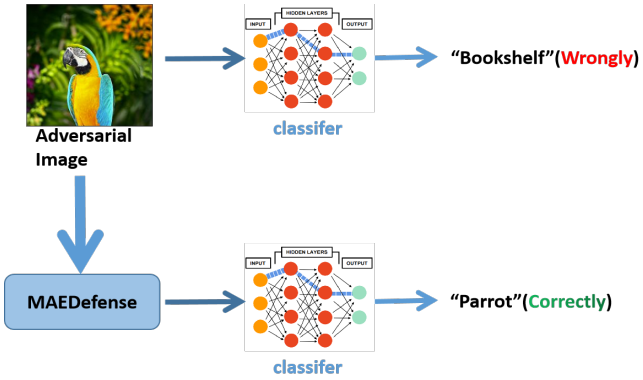


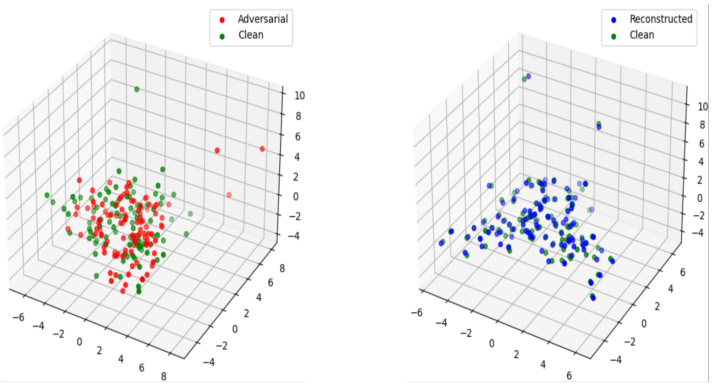
Fig. 1. An example of the proposed MAEDefense.

Adversarial perturbations are carefully designed noises, and disrupting their structure is a natural defense mechanism. A straightforward idea is to directly mask the pixels in order to eliminate the structure of the noise. While this method may eliminate the noise structure, it also damages the structural

† Corresponding author
This research work is partly supported by the National Natural Science Foundation of China No. 62172001 and Natural Science Research Project of Anhui Educational Committee Project number 2022AH050120.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

information of the original image, resulting in a decline in the neural network's performance. However, this naive idea presents an intriguing suggestion: By utilizing random masking, the meticulously designed structure can be disrupted, while an image reconstruction model can be employed to restore the overall structure of objects in the image. If the reconstruction successfully preserves the underlying global structure, the pixel masking and image reconstruction can effectively redistribute the carefully designed adversarial noise into non-adversarial structures. Recently, He et al[18] proposed a simple and efficient unsupervised training model called MAE. This model leverages the abundant redundancy in image data and employs an asymmetric codec to restore partially obscured information. Inspired by their work, we have incorporated the MAE model as the reconstruction component in our defense approach.



(a) (b)

Fig. 2. (a) The left image shows a 3D map illustrating adversarial image features (highlighted in red) and corresponding clean image features (highlighted in green). (b) On the right side, we depict the features of the corresponding defense image (highlighted in blue). This visual representation clearly demonstrates how the masked reconstruction model effectively redistributes the carefully crafted adversarial noise into non-adversarial structures.

This paper presents a model-agnostic defense (MAEDefense) that effectively restores adversarially perturbed images to their original class labels. As shown in Fig**??**, an image of a "parrot" with adversarial perturbation is initially misclassified as a "bookshelf," but it is correctly classified after undergoing the MAEDefense process. The defense is based on the MAE model, which randomly masks image pixel blocks and reconstructs the masked pixels. This paper explores the feature structures of clean samples, adversarial samples, and reconstructed samples. Fig2 displays 100 randomly selected features projected onto a 3D space using principal component analysis. This figure clearly demonstrates how the masking and reconstruction model reallocates carefully crafted adversarial noise into non-adversarial structures, thereby disrupting the structure of the adversarial noise. The main contributions of our paper are as follows:

- The paper proposes a simple and effective defense method called MAEDefense. The carefully crafted adversarial noise locations are reassigned to non-adversarial noise locations, which effectively breaks the inherent structure of the adversarial noise. By utilizing the class activation map visualization, the MAEDefense method successfully redirects the classifier's attention away from the noise region and towards the main structure region of the image.
- The prior information embedded in the image and Mask AutoEncoder model is utilized to reconstruct the adversarial examples, resulting in a significant reduction in computational time. Through extensive experimentation, it has been demonstrated that MAEDefense is an effective countermeasure against various attack strategies.

II. Relate Work

A. Adversarial Attacks

Existing adversarial attack methods can be broadly categorized into two types: black-box attacks and white-box attacks. In a white-box setting, the attacker has knowledge of the model's structure and specific parameters, enabling them to launch adversarial attacks against the victim model with ease. In black-box attacks, the attacker has no knowledge of the internal structure, training parameters, etc., of the target model. The attacker can generate a large number of adversarial samples by querying or training substitute models, and then use these examples to attack the victim model.

Several attack methods have been proposed to effectively work in the white-box environment. Goodfellow et al[7] proposed a single-step attack called the Fast Gradient Symbolic Method (FGSM). Unlike FGSM,BIM[6] perturbs small multi-step perturbations along increasing gradients in an iterative manner.The projected gradient descent(PGD)[8] attack is BIM with random starts.The C&W attack[19] is an efficient optimization method used to find the smallest perturbation.

In the "arms race" between attackers and defenders, classifier owners may employ defensive measures to protect the targeted classifier. The effectiveness of many attack methods is compromised, but black-box attacks, such as transfer-based attacks, are more powerful. Momentum FGSM (MIM) [20] introduces momentum into gradient accumulation at each iteration, enhancing both robustness and transferability. DIM[21] further improves MIM by applying random transformations to the input image. TIM utilizes a set of images to compute gradients, which is particularly useful for black-box models with defensive mechanisms. To reduce the computation of gradients, Dong et al[22] propose shifting the image within a small range and approximating the gradient by convolving the gradients of untransformed images with a kernel matrix. Wang et al. [23] propose a method based on variance tuning to enhance transferability (VMI).

B. Adversarial Defenses

The goal of adversarial defense is to protect victim classifiers from adversarial attacks. Among the existing defense methods, reconstruction-based defense methods are becoming increasingly popular. These methods aim to reconstruct a "clean sample" by disrupting the fixed structure of the adversarial perturbation.

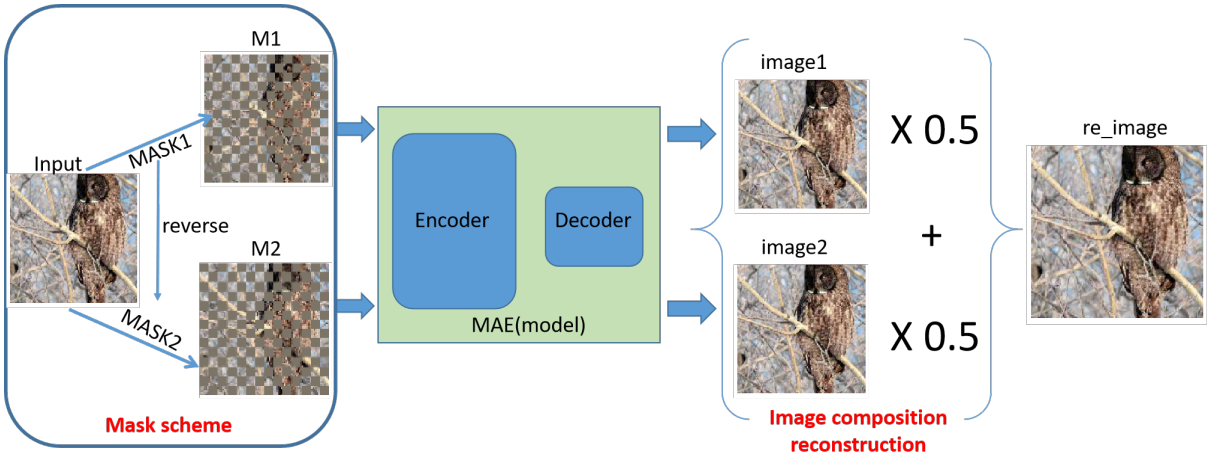


Fig. 3. Framework diagram of the MAEdefense.

Recently, a new approach called ComDefend[12] was proposed to remove adversarial noise from images while preserving their structural information. This approach is an end-to-end compression model. In addition, Mustafa et al[14] proposed a method that combines wavelet denoising and pre-trained super-resolution (SR) networks to reconstruct clean images. To address the issue of adversarial perturbations in the feature space, Naseer et al[15] developed a neural representation purifier (NRP). Such defenses require training on external datasets or pre-trained models, although they are effective in defending against certain attacks. Unlike previous methods, DIPdefend[13] exploits the internal priors of a target image to reconstruct clean images, achieving state-of-the-art performance in adversarial defense. However, DIPDefend can only operate on a single image and requires multiple iterations, which is not only time-consuming but also susceptible to overfitting.

C. Masked AutoEncoder

Masked Autoencoder[18] is a specialized neural network that is extensively used in image processing tasks. It is designed to reconstruct input data while simultaneously learning to mask or remove specific regions. This approach effectively removes noise, occlusions, and perturbations from images while preserving essential visual information. MAE consists of two main components: an encoder, which compresses input data (such as images) into a potential encoding, and a decoder, which utilizes this code to reconstruct the original input. During the training phase, MAE introduces masks that are applied to the input data, blocking certain portions. The network learns to reconstruct the original input data from these masked inputs, filling in missing or obscured regions with information from unmasked areas. This strategy allows MAE to focus on specific areas of interest, making it particularly valuable for tasks such as inpainting, denoising, and image completion. In summary, the Masked Autoencoder serves as a potent tool for image reconstruction and restoration. It effectively eliminates unwanted perturbations while preserving the visual quality of the original input data.

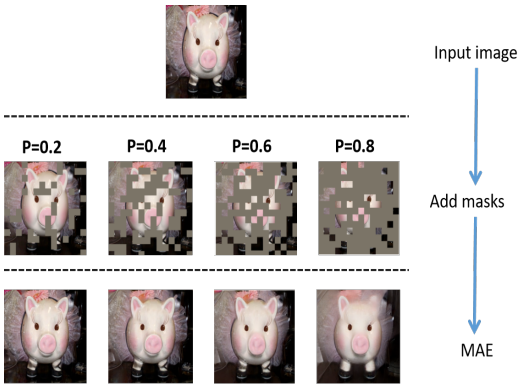


Fig. 4. An example of how a Masked AutoEncoder (MAE) affects input images. We apply various masks and display the reconstructed images using MAE

III. METHODOLOGY

A. Design Motivation

Typically, the ”clean images” obtained also contain white noise present in the environment, which is considered as part of the image. While modern deep neural networks perform at a human-level when dealing with unstructured small noise, the situation is different when it comes to carefully crafted structured adversarial noise, known as adversarial examples. This type of adversarial noise can easily disrupt the output of neural networks, leading to predictions that are inconsistent with human perception. To enhance the robustness of neural networks against adversarial noise, we need to carefully eliminate or reduce adversarial structures that are specifically crafted, while still preserving the intrinsic structure within the image to ensure accurate classification.

Through previous studies, we have learned that adversarial perturbations are highly unstable and can be eliminated using simple input transformation methods. Our approach involves initially neutralizing a significant portion of the structure

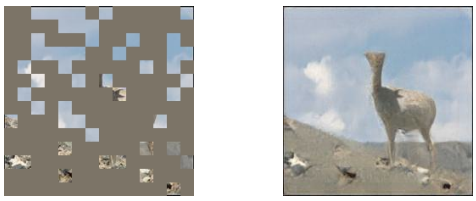
created by the adversarial noise through pixel masking (zeroing out blocks of pixels) in the image. However, this process also introduces non-adversarial noise to the image, which negatively affects the global structure of the image. Fortunately, we can use the MAE model to effectively recover the global structure of an image using information from the unmasked portion. the MAE model recovers the intrinsic structure of a clean image, such as the primary structure of a dog in a dog image.

Next, we present the reconstruction results based on the MAE model with different masking ratios, as shown in Fig4. It is worth noting that the global structure is maintained even when the image has a masking rate of 0.6. This indicates that the model is able to reconstruct the underlying global image structure from the masked adversarial image, despite the fact that a large amount of pixel block information has been removed. By applying a mask with a specific ratio, we can preserve the overall structure while reducing the impact of other potential adversarial structures.

B. MAEDefense as a defense mechanism

The analysis above demonstrates that the MAE reconstruction model is capable of restoring the visual information of an image. Even when some pixel block information is omitted, the remaining pixel information can still be utilized to reconstruct the global structure of the image. In this paper, we first reconstruct two adversarial images using the uniform mask and MAE model. Then, we fuse the two reconstructed images pixel by pixel to obtain the final clean image. The framework of the method is shown in Fig3. Firstly, the adversarial sample is divided into two complementary masked images. Secondly, the two masked images with carefully crafted adversarial noise locations are reassigned to non-adversarial noise locations. Finally, the two reconstructed images are pixel-wise fused (weighted average) to obtain a ”clean image”.

C. Mask scheme



(a) Random Mask



(b)Uniform Mask

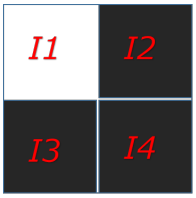
Fig. 5. Reconstructed images under different masking schemes. (a) random mask reconstructed at 75% ratio, and (b) uniform mask reconstructed at 75% ratio.

The MAE reconstructs the complete image after randomly masking 75% of the image in the current self-supervised learning task. Unlike the task of MAE itself, our goal is to restore the visual information of the image while disrupting the adversarial perturbation.

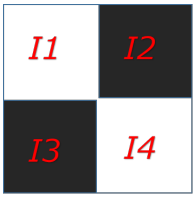
It has been found that when the pixels of an image are widely and randomly discarded, the MAE model is unable to predict the discarded pixel information based on the remaining pixel information. This leads to incomplete recovery of the global structure of the image. As shown in Fig5(a), features such as the antelope head and antelope horns could not be recovered, resulting in incorrect classification of the reconstructed images. This ultimately led to a decrease in the accuracy of image classification. This is because the visible elements within each local window are completely obscured, making it difficult to recover structural features within that window. To address this issue, a uniform masking scheme is proposed in this paper, as illustrated in Fig5(b). The specific operation is as follows: the image is divided into 49 pixel blocks, each measuring 32x32 in size. Again, the pixel blocks I are uniformly refined into four 16×16 pixel blocks, as shown in Fig 6(a). These four regions are named I1, I2, I3, and I4. We use the generated mask to discard all pixel information of regions I2, I3, and I4, and only keep the pixel information from the I1 region. Finally, we chose the mask scheme shown in Fig6(b) to balance the classification accuracy after reconstructing the adversarial and clean samples.

D. reconstruction scheme

When the perturbation is small and the attack intensity is low, reconstructing the recovery after discarding only a few pixels can provide effective defense. However, as the intensity of the attack increases, the structure against noise becomes more robust, and the defense performance of reconstructing clean samples by discarding only some blocks of pixels is greatly reduced. Therefore, the key to solving the problem lies in disrupting the structure of the antidisturbance by discarding more pixel information. In this paper, we propose a method to mitigate full-image adversarial perturbations using two masks. We reconstruct two similar images by MAE model, and then enhance the global structure of the reconstructed image by fusing the two images (weighted average) to achieve better defense performance.



(a)75% Ratio



(b)50% Ratio

Fig. 6. Masking schemes at different masking rates. (a) shows the uniform masking scheme with a 75% masking rate and (b) shows the uniform masking scheme with a 50% masking rate.

IV. EXPERIMENTS

A. Experimental Settings

Dataset: We randomly pick 1,000 clean images pertaining to the 1,000 categories from the ILSVRC 2012 validation set. **Attacks and Defense Methods:** To verify the effectiveness of the defense framework, five classical methods,FGSM[7],BIM[6],PGD[8],MIM[20] and C&W[19] are chosen to generate different types of adversarial samples for defense testing. In this attack setup, the model owner evaluates the effectiveness of the defense methods in protecting the victim model. The defense methods include three methods that require external prior information training,ComDefend[12],SR[14],NRP[15] and DIPDefend[13], which does not require any external training. These methods are evaluated and compared with the proposed method. Additionally, with the development of adversarial attack and defense techniques, more and more advanced attacks are proposed. Black-box attack methods, such as transferable attacks, are more robust, and many defense mechanisms fail under advanced transferable attacks, such as DIM[21],TIM[22],VMI and VNI[23],etc. We have also conducted experiments to evaluate the performance of our defenses against transferable attack methods.

B. Assessment under White-Box Attack

In this experiment, ResNet-50[25] is used as the victim model and achieves an accuracy of 94.7% for clean image samples without any defense mechanism. However, under classical adversarial attacks, the classification accuracy of the adversarial samples decreases rapidly, and it is difficult to disrupt the fixed structure of the adversarial perturbation. As a preprocessing method that does not require external training data, this method still achieves a high average classification accuracy (ACC) of 69.2% under various classical adversarial attacks, as shown in Table I. In comparison to previous defense methods, our proposed method not only achieves higher classification accuracy after reconstruction but also does not require any additional training. When compared to the advanced defense method DIPDefend, which does not require multiple iterations, our method is simpler and more efficient to implement. Our defense method achieves a significant improvement not only on the adversarial examples but also on the clean examples.

Although the proposed defenses achieve good performance against classical attack methods, many of them fail to withstand the emergence of various attack techniques. The most notable among these are black-box attacks, which are based on tranferrable attacks. These attacks generate more robust adversarial examples and perturbations that are more difficult to eliminate. We selected three classifier models ResNet50[25],VGG16[26] and DenseNet-121[27] as our victim models to generate adversarial examples under four robust transferable attack methods. We present the classification accuracy of the models before and after the adversarial samples are defended by our defense methods, as shown in Table II. Maintain more than 40% model classification accuracy in the face of advanced attack methods.Comparing different attack methods, we find that the performance of our defense method is degraded by image pre-processing based data enhancement attacks DIM and TIM attacks, while the gradient obfuscation-based adversarial attack methods have relatively little impact on our defense methods.

In addition to comparing defense capabilities, we also compared the image quality generated by different defense methods. In Table III, we quantitatively compared the ImageNet dataset using the SSIM[24] metric. We conducted a visual comparison of image quality among different defense methods under five different types of adversarial attacks. Our observations indicate that the proposed method generally outperforms the baseline methods in terms of visual quality. The superiority of the reconstructed images further confirms the effectiveness of our proposed strategy for mask reconstruction denoising in improving image quality.

C. Assessment under Black-box Attack

In this section, we assess different defense methods counter cross-model black-box attacks. In such attacks, the adversary does not have access to the parameters of the target classifier or defense model. Instead, a substitute model (referred to as the source model) with a distinct architecture is trained. The adversarial examples generated from the source model can deceive the target classifier. Black-box attacks can easily be transferred between different models.

Our method is compared to the DIPDefend approach, as it has demonstrated superiority over several previous defense methods in similar settings. To evaluate the robustness of our method against black-box attacks, we utilize the ResNet-50 model as the substitute model to generate adversarial samples using various attack methods. The Inception-v3 and VGG16 models have been chosen as our target models. The overall accuracy results are presented in Table IV. We observed a significant decrease in accuracy for the target models when they were tested on the adversarial samples generated from the substitute model. This confirms the transferability of adversarial examples. However, by implementing various defense strategies, the precision of the target models significantly enhances across different attack scenarios. Comparatively, our method demonstrates superior performance in the majority of black-box attack scenarios.

D. Analysis of the Proposed Method

Why is the mask ratio 50%?We investigated the relationship between the classification accuracy of reconstructed images and the masking rate. Different masking rates were applied to both the original and adversarial samples, and then these images were reconstructed. The reconstructed images were inputted into the ResNet-50 network for classification. Fig8 illustrates the classification accuracy and the corresponding masking rates.From the figure, it can be observed that the recognition accuracy of reconstructed clean samples gradually decreases as the masking rate increases. However, for highly perturbed adversarial samples, the classification accuracy of

TABLE I
PERFORMANCE COMPARISONS WITH FIVE ADVERSARIAL DEFENSE METHODS ON THE IMAGENET DATASETS UNDER WHITEBOX ATTACKS.(FGSM,PGD,BIM AND MIM ATTACKS ADOPT THE ℓ_∞ NORM WITH ϵ = 2/4/8)

Defence method	clean	FGSM[7]	PGD[8]	BIM[6]	MIM[20]	C&W[19]	ACC
No defence	94.7	13.0/8.7/7.6	0.5/0.0/0.0	3.9/0.2/0.0	0.9/0.0/0.0	0	9.25
Comdefend[12]	81.1	34.2/19.7/12.7	42.2/18.4/2.8	30.2/11.7/3.8	23.9/4.8/0.6	77.8	20.43
SR[14]	73.3	64.0/56.6/43.7	68.1/63.4/49.3	65.6/59.1/53.3	63.7/53.5/38.2	70.7	58.75
NRP[15]	**86.7**	32.4/21.4/18.5	38.4/16.9/13.3	26.5/7.1/2.2	18.4/2.8/3.3	**82.1**	25.60
DIPdefend[13]	75.1	61.2/48.3/40.1	69.7/65.4/50.8	60.3/57.2/49.7	60.9/53.7/40.6	70.5	54.0
MAEdefense(Ours)	85.8	**74.3/64.6/54.7**	**75.5/72.5/64.3**	**75.4/71.7/66.3**	**74.5/65.9/45.7**	77.2	**69.2**

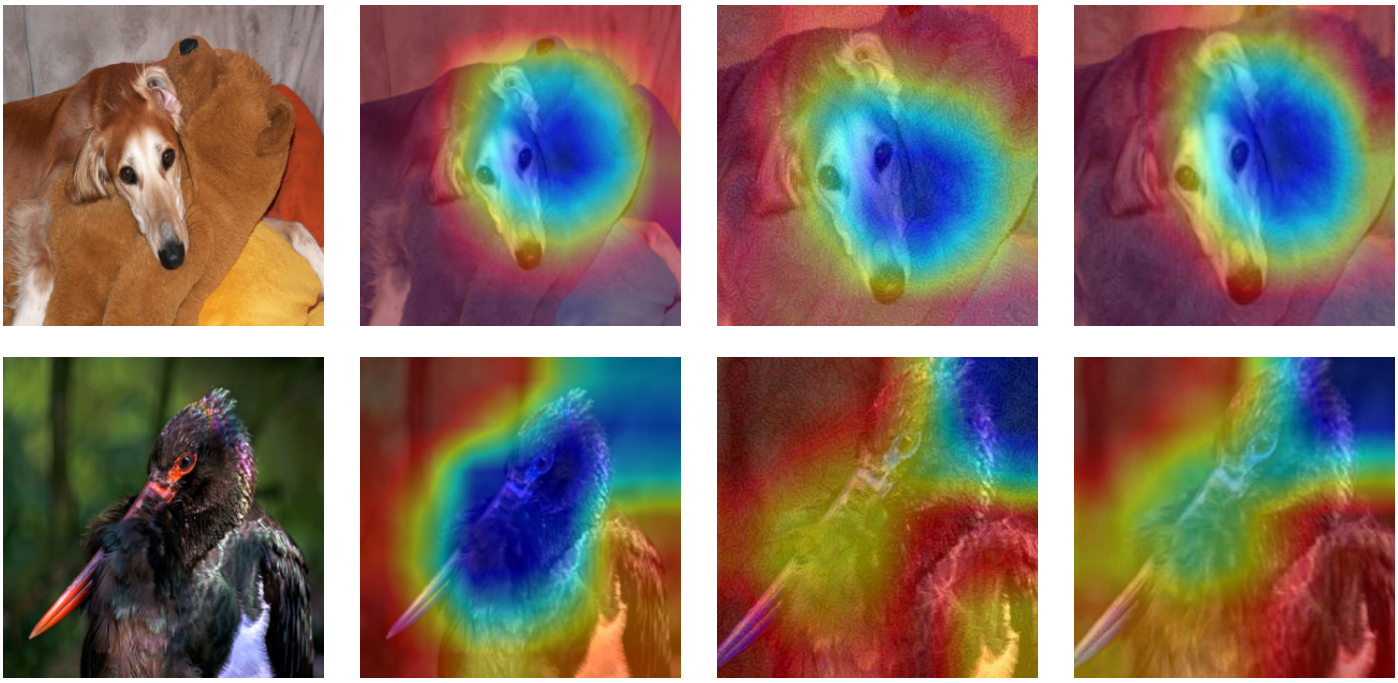


Fig. 7. Defense Visualization against MIM Attacks. The first column displays three clean images. The following three columns show the class activation maps for the clean, MIM-attacked, and reconstructed images.

TABLE II
CLASSIFICATION ACCURACY ON DIFFERENT CLASSIFIERS BEFORE AND AFTER RECONSTRUCTION OF ADVERSARIAL SAMPLES UNDER ROBUST ATTACK($\epsilon = 2$)

attacke method	ResNet-50	VGG16	DeseNet-121
DIM[21]	0.9	0.2	0.4
	58.5	44.9	52.6
TIM[22]	5.9	2.0	1.2
	64.4	42.8	55.9
VMI[23]	0.4	0.6	0.2
	70.2	49.7	61.8
VNI[23]	1.0	0.8	0.1
	69.6	49.8	60.8

the reconstructed samples is highest when the masking rate of 50%. This could be attributed to the fact that preserving a certain proportion of the image information enables the reconstructed samples to retain some similarity to the original samples, thereby enhancing the classification accuracy. However, as the masking rate increases further, more information is lost in the reconstructed images, leading to a decrease in classification accuracy. Therefore, in this experiment, a masking rate of 50% was found to be the threshold for achieving optimal classification accuracy.

In order to analyze the performance of the proposed method, we compare it with the direct reconstruction method using MAE[18], analyze the proposed method. As shown in Table V, the proposed method not only outperforms MAE in reconstructing clean samples but also surpasses MAE in countering adversarial attacks.The effectiveness of our proposed method is fully demonstrated by the analysis of experimental results. To validate the shorter runtime of our method, we compared its runtime with that of DIPDefend on a single image. The Table VI shows that our method has a significantly shorter runtime compared to DIPDefend. This is because DIPDefend requires multiple iterations, typically around 1000 iterations, to generate normal samples, whereas our proposed method does not require any iterative computation. Therefore, compared to other methods, our approach has an advantage in terms of defense accuracy and computational cost.

In the final part of this paper, we utilized Class Activation

TABLE III
COMPARE THE IMAGE QUALITY OF RECONSTRUCTED IMAGES USING VARIOUS ADVERSARIAL DEFENSE METHODS, USING THE SSIM (STRUCTURAL SIMILARITY INDEX) METRIC[24]

Defence method	clean	FGSM[7]	PGD[8]	MIM[20]	VMI[23]	C&W[19]
Comdefend[12]	0.862	0.855/0.832/0.764	0.857/0.847/0.815	0.858/0.853/0.843	0.857/0.843/0.793	0.861
SR[14]	0.885	0.880/0.861/0.762	0.882/0.874/0.842	0.882/0.879/0.869	0.881/0.871/0.813	0.884
NRP[15]	0.823	0.825/0.830/0.869	0.823/0.823/0.844	0.826/0.825/0.825	0.826/0.821/0.835	0.827
DIPdefend[13]	0.802	0.789/0.746/0.640	0.795/0.771/0.709	0.794/0.782/0.763	0.795/0.771/0.696	0.807
MAEdefense(Ours)	**0.915**	**0.903/0.898/0.839**	**0.925/0.891/0.845**	**0.905/0.896/0.889**	**0.901/0.892/0.851**	**0.926**

TABLE IV
EVALUATE THE TRANSFERABILITY OF MAEDEFENSE ACROSS DIFFERENT MODELS AND COMPARE IT WITH DIPDEFEND(IMAGENET, ATTACK STRENGTH ϵ = 8/16). THE SOURCE MODEL USED IS RESNET50, AND THE TARGET MODELS ARE INCEPTION-V3 AND VGG16.

Target Network	Defence method	FGSM[7]	PGD[8]	MIM[20]	DIM[21]	TIM[22]	VMI[23]
Inception-v3	Normal Inception-v3	67.4/55.9	74.3/67.2	60.6/38.4	31.7/11.0	61.9/40.6	44.3/40.9
	DIPdefend[13]	56.6/39.7	66.5/61.5	55.8/40.9	32.2/23.4	46.7/39.2	42.5/36.8
	MAE[18]	59.0/45.5	68.0/60.4	59.6/46.3	46.1/28.5	50.9/38.7	52.3/47.4
	MAEdefense(ours)	61.0/47.5	75.3/67.4	69.6/51.3	55.3/37.9	55.3/45.6	57.8/53.4
VGG16	Normal VGG16	50.2/30.4	51.3/43.0	36.2/20.4	15.2/6.4	39.6/16.9	25.1/22.1
	DIPdefend[13]	40.0/31.2	48.3/47.2	35.7/22.3	20.0/17.6	27.6/18.2	23.5/20.3
	MAE[18]	43.8/33.5	50.1/46.8	44.4/33.5	33.4/19.9	29.5/20.3	39.6/33.7
	MAEdefense(ours)	55.7/36.3	52.3/37.6	50.2/37.8	55.3/37.9	55.3/35.6	53/36.4

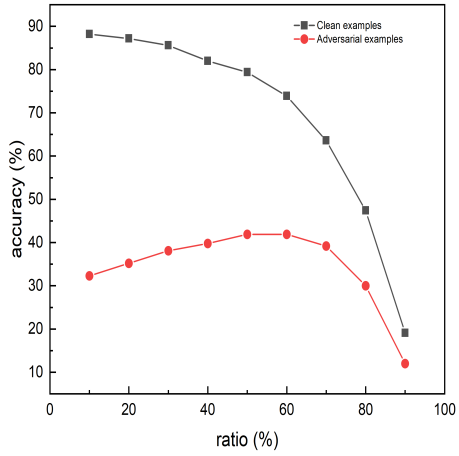


Fig. 8. Classification accuracy of reconstructed images with adversarial samples and clean samples at different mask rates on ResNet-50

TABLE V
PERFORMANCE COMPARISONS WITH MAE ON THE IMAGENET DATASETS UNDER WHITEBOX ATTACKS ($\epsilon = 2$)

Defence Method	clean	FGSM	PGD	BIM	MIM	C&W
MAE[18]	60.2	52.4	51.9	52.4	51.0	54.4
Ours	85.8	74.3	75.5	75.4	74.5	77.2

Mapping (CAM) to interpret the predictions made by the Convolutional Neural Network (CNN) model. CAM generates heatmaps by replacing the last fully connected layer with a Global Average Pooling (GAP) layer. By applying a class-weighted average to the feature maps, the Class Activation Map (CAM) algorithm highlights the regions in the image that are relevant to the predicted class label. The Fig7 illustrates the results of the CAM analysis for the top-1 predictions of the model on clean samples, adversarial samples, and recovered samples. It can be observed that the masked reconstruction disrupts most of the noise structure, resulting in CAM results that resemble those of clean images. The MAE-based masked reconstruction helps restore the model's attention to distinctive regions that correspond to the correct class label.

TABLE VI
COMPARE THE RUNTIME OF OUR PROPOSED IMAGE RECONSTRUCTION DEFENSE METHOD WITH OTHER EXISTING METHODS

Defence Method	Runtime (seconds)
DIPDefend	600.038
Ours	5.316

V. CONCLUSION

The paper presents a novel and simple defense method against adversarial examples. Unlike previous work, the scheme disrupts the fixation structure of adversarial perturbations by masking blocks of pixels and reconstructing the global structure of classified images using a Masked AutoEncoder (MAE). Extensive experiments on a public image dataset demonstrate that the proposed approach surpasses numerous existing defense methods in terms of defense performance and visual quality. In addition, our defense method is independent of the type of attack, does not require additional training, and is equally effective against both black-box and white-box attacks.

REFERENCES

[1] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[3] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3516–3527, 2019.

[4] C. Liu, L.-C. Chen, F. Schroff, *et al.*, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 82–92.

[5] M. Rahul, P. L. Uppunda, R. S. Vinayaka, B. Sumukh, and C. Gururaj, "Simulation of self-driving cars using deep learning," *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Applications*, pp. 379–396, 2022.

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, Chapman and Hall/CRC, 2018, pp. 99–112.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[9] C. Xiao and C. Zheng, "One man's trash is another man's treasure: Resisting adversarial examples by adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 412–421.

[10] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 501–509.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[12] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6084–6092.

[13] T. Dai, Y. Feng, B. Chen, J. Lu, and S.-T. Xia, "Deep image prior based defense against adversarial examples," *Pattern Recognition*, vol. 122, p. 108 249, 2022.

[14] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2020.

[15] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 262–271.

[16] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJUYGxbCW.

[17] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=BkJ3ibb0-.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[19] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, Ieee, 2017, pp. 39–57.

[20] Y. Dong, F. Liao, T. Pang, *et al.*, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[21] C. Xie, Z. Zhang, Y. Zhou, *et al.*, "Improving transferability of adversarial examples with input diversity," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

Preemptive Image Protection against Steganography

Yusheng Guo*, Nan Zhong *, Zhenxing Qian * †, Xinpeng Zhang * † and Hsunfang Cho ‡
* School of Computer Science Fudan University, Shanghai, China
E-mail: ysguo20@fudan.edu.cn, nzhong20@fudan.edu.cn
† Key Laboratory of Digital Culture Protection and Tourism Data Intelligent Computing, Shanghai, China
E-mail: zxqian@fudan.edu.cn, zhangxinpeng@fudan.edu.cn
‡ Shanghai Data Exchange Corporation, Shanghai, China
E-mail: alex-zhuo@chinadep.com

***Abstract*—Image steganography plays a vital role in the realm of information hiding, involving the discreet embedding of confidential data within an image. However, the misuse of this technology by malicious actors on social media platforms poses a significant threat to public security. Presently, the primary defense against image steganography is steganalysis technology, which effectively distinguishes steganographic images from regular ones by detecting subtle modifications. Nevertheless, there is considerable room for improvement in existing steganalysis techniques, especially when dealing with small payloads. In this paper, we propose a proactive strategy to enhance the detection of steganographic images with small payloads. Specifically, as the image is captured, we apply a meticulous-designed and imperceptible mask to it. This mask is aimed at increasing the success rate of detection. We have developed three convolutional neural networks (CNNs) to facilitate this proactive image processing: preemptive-CNN, steganography network (Stego-CNN), and steganalysis discriminator. During the training process, Stego-CNN remains unaltered, while preemptive-CNN and the discriminator are jointly trained to ensure that the applied mask enhances the detection accuracy. Experimental results demonstrate the effectiveness of our proposed approach in enhancing the detection accuracy, all without compromising the visual quality of the images.**

I. Introduction

Thousands of images are transmitted on social networks every day. With communication technology rapidly developing and photography equipment becoming increasingly portable, people can record and share images almost anytime and anywhere. However, while this brings great convenience to social communication, it also poses some security risks. Image steganography enables an image to carry secret information with minimal modification, typically less than 1 per pixel. Due to the minimal amplitude of modification per pixel, the human visual system finds it difficult to detect the difference between the original and steganography-altered images. Criminals can exploit steganography to embed criminal information in seemingly normal images, making it appear no different from regular image transmission. Therefore, studying the technology of resisting steganography holds great practical significance in maintaining social security and fighting crime.

Image steganography [1]–[5] is a technology that enables the embedding of secret information into an image using an imperceptible method. Steganography typically introduces only slight pixel value distortions to an image to maintain its invisibility. Moreover, image steganography prefers to modify the textured and complex regions of the image [6]–[9] to preserve statistical characteristics and avoid visible artifacts. This process is similar to conventional image processing techniques, such as JPEG compression. However, steganography introduces even subtler changes to the images compared to JPEG compression. Fig. 1 illustrates the influence of steganography and JPEG compression on an image. For the sake of simplicity, the original image is referred to as the cover, and the steganographic image is referred to as the stego in the subsequent content of this paper.

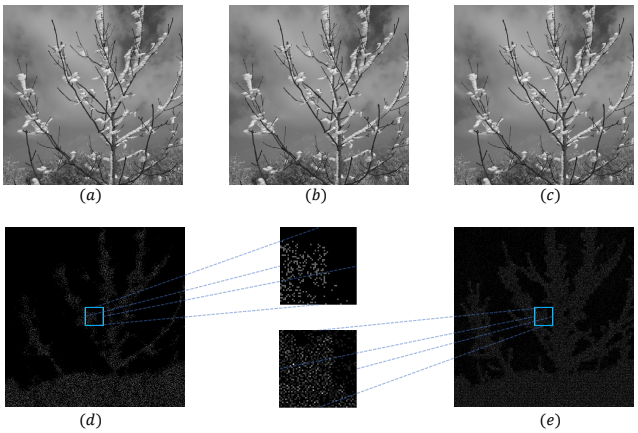


Fig. 1. (a): Original image. (b): Steganographic image, where the embedding algorithm is HILL with payload $= 0.2$. (c): Compressed image with the quality factor of 95. (d): The residual image between steganographic image and the original image (magnified 100 times). (e): The residual image between compressed image and the original image (magnified 30 times).

In order to mitigate the potential security risks linked to image steganography, a plethora of countermeasures have been put forth. Among these countermeasures, steganalysis technology emerges as the predominant and highly efficient approach. Steganalysis typically comprises three key steps: initially, computing the residuals of both cover images and stego images; next, extracting high-dimensional statistical features from these residuals; and ultimately, employing these features to identify stego images through the use of a discriminator.

Classical steganalysis primarily focuses on the disparity of hand-crafted statistical features during the feature extraction

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

stage. These features exhibit strong mathematical interpretability between cover and steganographic images [10]–[16]. The detection accuracy of steganalysis discriminators heavily relies on the dimensionality of the extracted features. Higher-dimensional features result in better detection performance. In primary steganalysis, discriminators often employ nonlinear Support Vector Machines (SVMs) to classify the extracted features [17]–[21]. However, the computational complexity of SVMs increases significantly with the feature dimension, which limits the timeliness of the discriminator. In current steganalysis practices, Fisher Linear Discriminant (FLD) ensemble classifiers are commonly employed to handle high-dimensional features [22]. Specifically, FLD serves as the fundamental classifier, and the final decision result is obtained through voting aggregation based on the decisions of multiple basic classifiers, such as the Spatial Rich Model (SRM) [23] and the selection-channel-aware rich model (maxSRM) [24].

On the other hand, all three modules of steganalysis can be simulated and implemented using convolutional neural networks (CNNs). The steganalysis technology based on CNN is gradually emerging [25]. The performance of CNN has even surpassed that of classical steganography since the advent of YeNet [26], which is the first CNN-based steganography that made a breakthrough in performance. Steganalysis Residual Network (SRNet) [27] overcomes the constraints of YeNet on the database and prior knowledge, making network training an end-to-end process. SiameseNet breaks the constraints of steganography on image size and can detect images of multiple sizes without degrading performance.

However, both traditional steganalysis methods and modern neural network-based steganalysis techniques have shown suboptimal performance when dealing with small steganographic payloads. In this research, we introduce an innovative CNN-based approach to combat image steganography. Our method involves superimposing an imperceptible mask on the image during its creation, which we call the ”masked cover.” Detecting steganography in the masked cover is more achievable compared to the original cover, even when the masked cover undergoes standard image processing before steganography. To implement our approach, we offer a discriminator to social network operators, empowering them to identify images containing concealed information. By identifying steganographic images, they can prevent them from being uploaded to social networks.

The main contributions of the proposed strategy are summarized as follows:

- We implemented an approximation to image steganography using convolutional neural networks.
- The proposed scheme enhances the complexity of image steganography while preserving the visual quality of the image..
- Our scheme improves the detection accuracy of stego images and is compatible with existing steganalysis techniques.

The remaining structure of this paper is as follows. The second section II is the related work of steganography and steganalysis. The implementation details of our scheme are in section III. The fourth section IV presents our experimental results. Finally, the conclusion of this study are presented in section V.

II. Related work

A. Steganography

Image steganography represents a fundamental method for covert communication. It involves concealing information within an ordinary, unaltered image, effectively shrouding the secrecy of transmission. The key differentiation between steganography and encrypted communication lies in the visibility of the communication process: the latter is observable, while the former adeptly disguises it. Historically, steganography can be broadly classified into three categories: naive steganography, adaptive steganography, and CNN-based steganography.

Naive steganography includes Least Significant Bit (LSB) substitution [24], Bit-Plane Complexity Segmentation (BPCS) [28], vector pixel randomly Plus or Minus 1 (PM1) [29], MMD [30] encoding and ZZW [31] encoding, among others. These algorithms, collectively known as non-adaptive steganography, solely prioritize the visual quality of stego without thoroughly considering image statistical features. Consequently, they are prone to detection by steganalysis.

Adaptive steganography is the prevailing steganography technique that assigns different distortion values to pixels in varying positions. It comprises two frameworks based on minimizing distortion and mathematical models. The minimization distortion steganography framework conceals secret information while minimizing overall distortion. Ever since Filler et al. introduced the convolution code-based Syndrome-Trellis Codes (STC) [32], which approximates the theoretical distortion limit, the focus of research in the minimization distortion steganography field has shifted towards defining better distortion value functions. Highly Undetectable steGO (HUGO) [33] defines the distortion value based on the deviation of features extracted from the cover and stego using the steganalysis algorithm SPAM [34]. A smaller deviation corresponds to a smaller distortion value. Wavelet Obtained Weights (WOW) [2] and UNIversal WAvelet Relative Distortion (UNIWIARD) [3] employ a set of multidirectional wavelet filters to predict the pixel residuals in the cover. Pixels that are difficult to predict in any direction are assigned a small distortion value. HIgh-pass, Low-pass, and Low-pass (HILL) [5] uses a high-pass filter to identify the hard-to-predict areas in the cover, while two low-pass filters concentrate pixels with low distortion costs. Steganography based on a mathematical model offers more theoretical support and interpretability compared to methods based on distortion value. Multivariate Gaussian (MG) [35] models the cover and stego using a multivariate Gaussian model, optimizing the KL-divergence of the probability distribution to achieve indistinguishability. Fridrich et al. [4] employ the Multivariate Generalized Gaussian model

(MVG), which incorporates optimization parameter measures to specify modifications to the carrier image, resulting in improved security compared to MG.

Currently, the rapid advancement of Convolutional Neural Networks (CNN) is forging a novel path within the realm of steganography. A series of CNN-based methodologies have been sequentially introduced to this field. Volkhonsiky et al. [36] have presented a steganographic approach rooted in Generative Adversarial Networks (GANs). In this paradigm, the generator handles the process of embedding confidential information, while the discriminator specializes in steganalysis. By leveraging GANs for continual optimization of the distortion function, Tang et al. have achieved a level of security performance equivalent to UNIWARD. In the HiDDeN framework [37], the encoder is entrusted with the task of generating an encoded image, referred to as stego, from a combination of secret information and a cover image. The resulting stego image is meticulously designed to be visually indistinguishable from the original cover image, while still allowing for the extraction of concealed information. Tang et al. [38] have introduced the ADV-EMB algorithm, which adapts the distortion function based on the gradient of the neural network, drawing inspiration from adversarial samples. Furthermore, UT-6HPF-GAN [39] is capable of crafting a modified probability graph to confound the discriminator by utilizing the cover image as input to the generator. This innovative approach showcases the ever-evolving landscape of steganography through CNN-driven techniques.

B. Steganalysis

Steganography is a countermeasure to steganalysis, which is used to detect whether an image has embedded secret information. Steganography requires changing the pixel values in the cover, which inevitably destroys the neighborhood correlation of pixels. Even if the correlation is difficult to be perceived by human eyes, steganalysis can distinguish cover from stego. Generally, steganalysis includes three modules: residual calculation, feature extraction, and discriminator. Obviously, the performance of steganalysis depends on the quality of extracted features and the discriminating ability of the discriminator, which also guides the direction of technical development.

The nonlinear Support Vector Machine (SVM) [40] is a commonly employed discriminator in naive steganalysis. However, since the introduction of SRM, the number of feature dimensions has surged to a staggering 34,671. Persisting with SVM as the discriminator in such circumstances would result in a substantial computational burden. Consequently, contemporary steganalysis approaches often opt for an ensemble discriminator based on Fisher's Linear Discriminant (FLD) to handle high-dimensional features. Fridrich et al. first introduced SRM [23], which leverages multiple submodels and high-pass filters to capture a wide range of steganographic features. This methodology strives to enhance the detection of potential damage to the neighborhood correlation of pixels. As an upgraded iteration of SRM, maxSRM [24] assigns varying statistical weights to the pixels in the cover image. It examines the maximum pixel modification probability in four positions to compute the co-occurrence matrix of the cover image.

The potent feature extraction and classification capabilities of Convolutional Neural Networks (CNNs) offer novel opportunities to enhance steganalysis detection accuracy. GNCNN [41] stands as the pioneering CNN-based scheme that can rival classical steganalysis in terms of performance. It comprises three integral components: a preprocessing layer (utilizing a non-random high-pass filter), a feature extractor (a convolution block), and a discriminator (a fully connected layer). YeNet [26] adopted an initial layer of thirty 5×5 SRM filters in their CNN architecture, achieving superior detection performance compared to previous classical approaches, marking a significant breakthrough in CNN-based steganalysis. Yedroudj et al. further refined the YeNet structure, introducing Yedroudj-net [42] to enhance detection performance. In pursuit of improved detection capabilities, Li et al. devised a broad structure featuring parallel subnets with several filter groups for preprocessing. This innovative structure goes by the name ReST-Net [43]. SRNet, introduced by Boroumand et al. [27], represents the first scheme that dispenses with manually designed filters for feature extraction. Instead, they initialized filters randomly and extracted ample noise residuals through multiple non-pooled convolutional blocks. Weike et al. expanded the horizons of steganalysis by leveraging the relationships between image sub-regions to distinguish whether an image contains steganographic content or not [44]. This approach broadens the applicability and versatility of steganalysis techniques.

III. METHOD

In this section, we will elucidate the intricate details of our innovative approach. As illustrated in Fig. 2, our proposed scheme is primarily composed of three essential core modules: the Preemptive-CNN, Stego-CNN, and Steganalysis Discriminator. Initially, the Preemptive-CNN generates the original cover and subsequently creates the masked cover. Following this, the masked cover is fed into the Stego-CNN, where it is utilized to embed confidential information and generate the stego image. Finally, both the stego image and the masked cover serve as inputs for the discriminator. To ensure that the first processed image can receive feedback from the subsequent discriminator, it is imperative that all three aforementioned modules be differentiable.

A. Preemptive-CNN

The purpose of the preemptive-CNN is to enhance the original covers prior to steganography, making it easier to detect steganography on the modified covers compared to the original ones. To accomplish this, an encoder-and-decoder neural network is employed to implement the preemptive-CNN. The input to the preemptive-CNN consists of the original covers, while the output is masked covers of the same size. Given that the covers are images used in a social network, the visual quality of the masked covers is crucial. Therefore, any arbitrary modification of pixels in the original covers is not allowed. In the proposed method, we impose restrictions

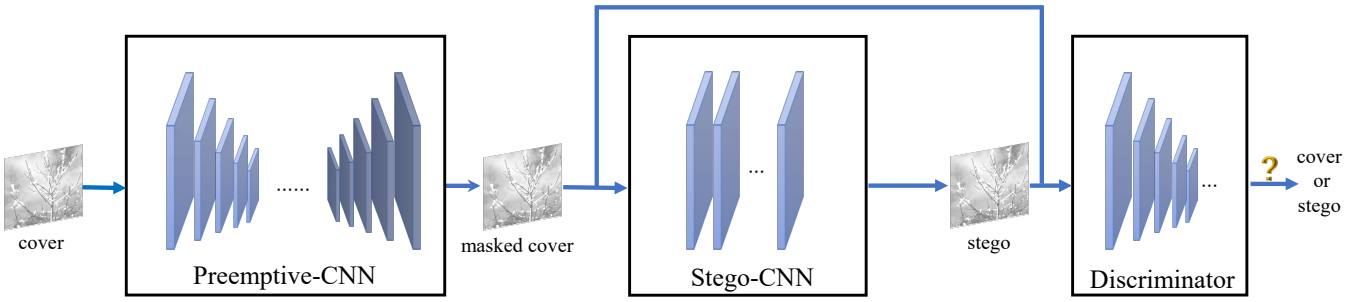


Fig. 2. Framework of proposed method

on the Structural SIMilarity (SSIM) [45] value of the original cover to ensure that the difference between the masked covers and the original covers is imperceptible to the human visual system.

Mathematically, this can be expressed as follows:

$$ssim(x_{ori}, Clip(f_P(x_{ori}))) > \mathcal{T}. \tag{1}$$

Where x_{ori} is an original cover, f_P is the preemptive-CNN, $Clip$ is used to force the value of pixels in the output in $[0, 255]$, and $\mathcal{T}$ is the constraint threshold. The SSIM is defined as follow:

$$ssim(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\mu_x^2 + \mu_y^2 + c_2)}. \tag{2}$$

where μ_x and σ_x^2 is the mean and variance of x, respectively. σ_{xy} is the covariance of x and y. $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$ where L is the fluctuation range of pixel values.

Considering the nature of the image data, it is important to note that the value of each pixel is an integer within the range of [0, 255]. Consequently, the value of each pixel in the masked cover should also be an integer. However, due to the characteristics of CNN, the generated masked cover, produced through encoding and decoding processes, often exist in a floating-point format. This poses a challenge, as the usage of the $Round()$ function for approximating the masked cover is non-differentiable, leading to the disappearance of gradients throughout the entire network.

To address the aforementioned issues, we propose a solution by employing the $Sigmoid()$ function to approximate the masked cover image. This approach allows us to maintain the integer representation of the pixels while ensuring the differentiability of the overall network. Mathematically:

$$\widetilde{X} = \lfloor X \rfloor + sigmoid\left(C \cdot (X - \lfloor X \rfloor)\right). \tag{3}$$

Where X is the output of preemptive-CNN, C is a constant and permitted to select in $[10, 50]$.

B. Stego-CNN

For reference [47], we design the Stego-CNN to simulate the embedding process of HILL [5], which is one of the most effective steganography. HILL is constituted by a highpass filter H and two low-pass filters L_1 and L_2, which are implemented by convolution in proposed method. The cost value ϱ of the cover X is defined as:

$$\varrho = \frac{1}{|X \otimes H| \otimes L_1} \otimes L_2. \tag{4}$$

Where the notation $\otimes$ stands for mirror-padded convolution. L_1 and L_2 are both average filter with sizes of 3×3 and 15×15 respectively, and the form H is as:

$$H = \begin{bmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{bmatrix}. \tag{5}$$

Employing STC [32] to embed secret information, the theoretical steganographic distortion can be noted as:

$$D = \sum_{i=1}^{M}\sum_{j=1}^{N}\left[p_+(i,j)\rho_+(i,j) + p_-(i,j)\rho_-(i,j)\right]. \tag{6}$$

Where $\rho_+(i,j)$ and $\rho_-(i,j)$ stand the cost values of modifying $x_{i,j}$ by $+1$ and -1, and $\rho_+(i,j) + \rho_-(i,j) = 2/3\varrho_{i,j}$. $p_+(i,j)$ and $p_-(i,j)$ are the corresponding modification probabilities, and further

$$p_+(i,j) = \frac{e^{-\alpha\rho_+(i,j)}}{1 + e^{-\alpha\rho_+(i,j)} + e^{-\alpha\rho_-(i,j)}}. \tag{7}$$

$$p_-(i,j) = \frac{e^{-\alpha\rho_+(i,j)}}{1 + e^{-\alpha\rho_+(i,j)} + e^{-\alpha\rho_-(i,j)}}. \tag{8}$$

Where α is the parameter to adjust the payload. The value of α for each cover is randomly selected between $[10, 150]$ to achieve various payload during training.

C. discriminator and loss function

The discriminator's inputs consist of two parts: masked covers and masked stegos. Its outputs determine whether the input images are covers or stegos. Masked covers refer to the original covers processed with a preemptive-CNN, while masked stegos are the stegos corresponding to the masked covers. Both the discriminator and the preemptive-CNN are updated synchronously during training, while the StegoCNN

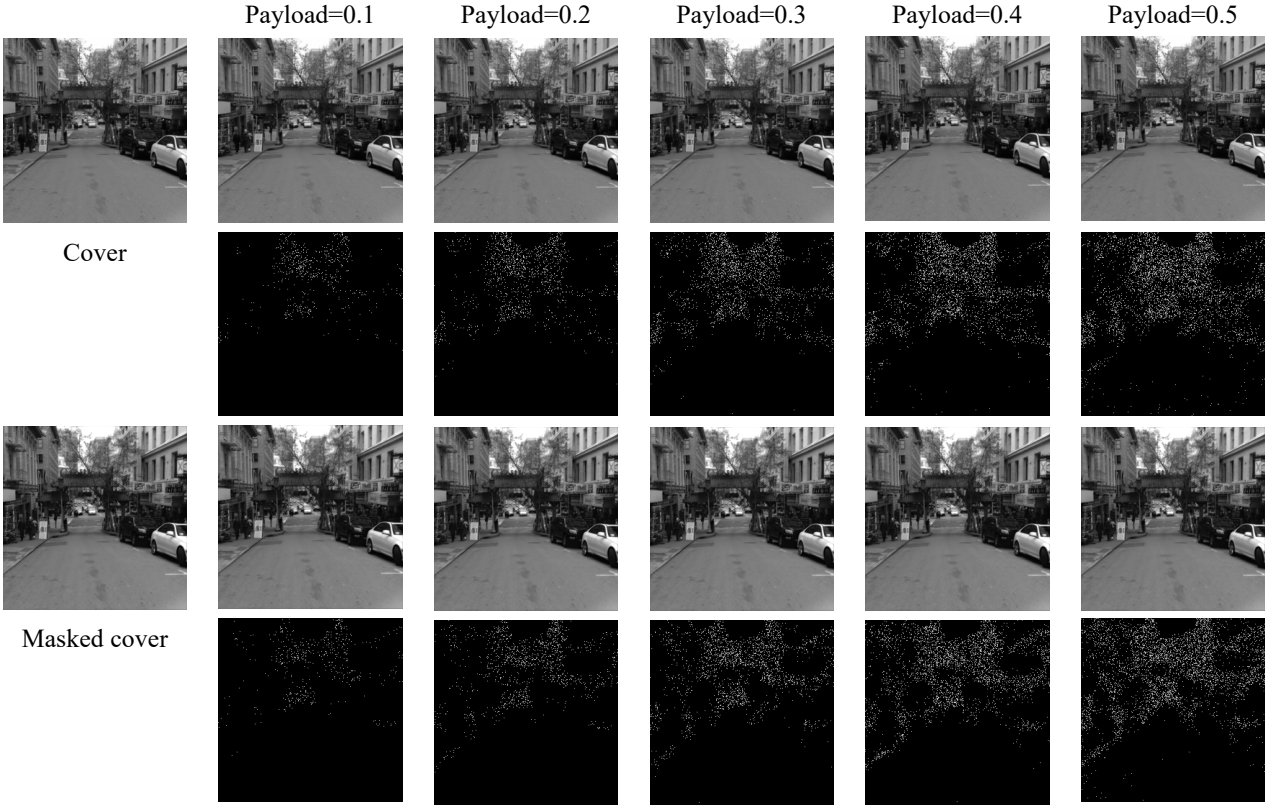


Fig. 3. Visual quality comparison of the original cover image, the masked cover image, and the stego images under various steganography payloads. The leftmost image of the first row displays the original cover, while the right side showcases the stego images corresponding to different payloads in sequence. The second row presents the residual image obtained by comparing the stego image with the original cover. The far left of the third row illustrates the masked cover, and the right side exhibits the stego images corresponding to different payloads in sequence. The fourth row showcases the residual image resulting from comparing the stego images with the masked cover.

remains unchanged. The loss function for the entire method is denoted as follow:

$$L = L_P + \lambda L_D. \tag{9}$$

Where the loss function of preemptive-CNN L_P is to ensure the visual similarity between the original covers and masked covers. We measure visual similarity with ssim as stated III-A.

$$L_P = ssim(X_{ori}, X_{masked}). \tag{10}$$

L_D denotes the loss function of discriminator and includes two components

$$L_D = L_{cls}. \tag{11}$$

Where L_{cls} is the cross-entropy loss to binary classification.

IV. Results

A. Setting

Unless otherwise specified, all experimental details and hyperparameters are set as follows. Select $\mathcal{T} = 0.95$ in equation (1), $C = 10$ in equation (3), $\alpha = 50$ in equations (7) and (8) To reduce the computation, we presume $\rho_+(i,j) = \rho_-(i,j) = 1/3\varrho_{i,j}$ in equation (6).

We select the Adam optimizer with learnning rate $lr = 1e-3$ to minimize L_P, and Adamax optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-8$ to minimize L_D. The balancing hyperparameter $\lambda = 1$ in equation (9).

The complete CNN model is implemented using PyTorch version 1.10.2, and the runtime results presented herein were achieved employing an Nvidia GeForce 2080Ti GPU. The batch size is configured to 6, necessitating approximately 10GB of GPU memory for training images of dimensions 256 × 256. We have conducted network training on diverse datasets, spanning 1000 epochs, a computational endeavor that consumes approximately 55 hours on our specific hardware configuration.

We selected 10,000 images from BOSSbase 1.01 as our original covers. These images were divided into three sets: the training set, the validation set, and the test set, with a ratio of 5:1:4 respectively. There was no overlap between these sets, and they were all resized to dimensions of 256 × 256. In all of our CNN experiments, we selected the model with the best performance on the validation set. Subsequently, we ran this model five times on the test set to calculate the mean and standard deviation of its performance.

TABLE I
COMPARISON OF THE DETECTION PERFORMANCE OF THE ORIGINAL COVER AND THE MASKED COVER UNDER DIFFERENT EMBEDDING PAYLOADS USING DIFFERENT STEGANALYSIS ALGORITHMS

Payload	Method	SRM [23]		SRMQ1 [23]		maxSRMd2 [24]		YeNet [26]		F-Net [46]	
		Original	Masked	Original	Masked	Original	Masked	Original	Masked	Original	Masked
0.1	HILL [5]	61.90	66.13	60.48	65.86	69.20	74.71	67.88	75.13	69.19	77.13
	WOW [2]	65.60	69.59	63.37	66.28	74.65	75.44	68.54	76.23	70.12	78.29
	UNIWARD [3]	64.85	67.63	64.38	66.58	68.35	72.87	68.14	76.18	70.44	78.57
	MVG [4]	62.47	66.58	61.65	66.17	65.33	72.91	68.39	76.54	70.30	78.33
0.2	HILL [5]	71.81	78.62	70.51	76.01	76.61	86.06	75.04	87.10	77.09	99.11
	WOW [2]	75.66	78.06	73.67	76.10	82.28	84.84	75.73	88.23	77.82	99.18
	UNIWARD [3]	75.19	80.36	74.15	78.42	77.48	83.55	76.16	88.47	76.28	99.24
	MVG [4]	72.41	79.75	70.73	77.58	73.80	80.44	76.28	88.62	78.18	99.20
0.3	HILL [5]	78.94	87.83	77.33	86.63	81.91	92.01	77.85	91.68	80.37	99.34
	WOW [2]	81.68	87.43	80.50	85.38	86.10	91.01	79.49	95.06	81.35	99.78
	UNIWARD [3]	81.89	89.89	81.44	88.19	83.66	90.47	78.63	95.55	82.14	99.85
	MVG [4]	78.59	88.92	77.84	87.14	80.23	89.17	77.15	95.18	80.67	99.89
0.4	HILL [5]	83.24	93.24	82.08	91.73	85.21	95.57	82.86	97.49	84.66	99.79
	WOW [2]	85.10	92.40	84.33	91.74	88.86	94.36	83.08	98.54	84.35	99.87
	UNIWARD [3]	86.53	94.42	85.95	93.64	87.39	94.70	83.78	98.58	84.64	99.94
	MVG [4]	82.84	93.46	82.52	92.38	83.93	93.93	83.84	98.16	85.27	100.00
0.5	HILL [5]	86.89	96.44	86.18	95.39	87.98	97.73	86.38	98.95	88.30	99.91
	WOW [2]	88.26	95.56	88.06	95.21	90.93	96.70	86.53	99.11	88.75	100.00
	UNIWARD [3]	89.48	96.74	89.19	96.43	90.26	97.28	86.93	99.36	89.49	100.00
	MVG [4]	86.58	96.23	86.04	95.49	86.96	96.68	87.18	99.45	88.94	100.00

B. Advantage of preemptive processing

Fig. 3 illustrates the visual distinctions among the original cover, the masked cover, and the stego images across different steganography payloads. It is evident that both the original cover and the masked cover undergo imperceptible modifications due to steganography, as perceived by the human visual system. It is important to note that the visual disparity between the original cover and the masked cover is exceedingly small, primarily because the SSIM threshold is set at 0.95. Consequently, it becomes arduous for the human visual system to discern any dissimilarities between them. Furthermore, in Fig. 3, the second and fourth rows reveal that the respective altered residuals of the masked cover exhibit a higher concentration compared to those of the original cover under the same payload. This characteristic contributes to facilitating the detection of the steganography within the mask cover.

TABLE I showcases a comprehensive comparison of detection performance between the original carrier and the mask carrier, both processed by a preemptive network, using various steganalysis algorithms across different embedding loads. The steganalysis techniques employed encompass SRM, SRMQ1, and maxSRMd2, which represent traditional, rich model-based methods. These experiments were executed within MATLAB R2022b. Each steganalysis experiment was repeated ten times, and the resulting average detection accuracy was deemed the ultimate detection outcome.

Additionally, YeNet and F-Net are steganalysis models based on convolutional neural networks. To implement these models, we utilized PyTorch 1.10.2, set the batch size to 32, selected ADAM as the optimizer, and established a learning rate of $1e-4$. Similar to the traditional steganalysis experiments, each trial for the neural network-based models was performed five times, and the average detection accuracy was used as the final detection result.

TABLE I illustrates the progressive increase in the detection success rate of various steganalysis algorithms as the steganographic payload rises. This phenomenon occurs due to the augmented steganography payload leading to a greater number of modified pixels within the overall image. Consequently, the detection of steganography becomes easier. Notably, different steganalysis algorithms exhibit varying levels of detection accuracy. Generally, CNN-based steganalysis outperforms traditional methods in terms of detection performance.

Regarding the original cover, the HILL algorithm demonstrates superior concealment compared to other algorithms. It proves to be the least easily detected under a range of detection algorithms. Moreover, the detection success rate for the masked cover consistently surpasses that of the original cover image. For traditional steganalysis algorithms, the preemptive strategy contributes significantly to performance improvement as the embedding load increases.

It is worth mentioning that the advantage of the HILL algorithm's invisibility becomes less pronounced, primarily because Stego-CNN only simulates the HILL algorithm during joint training. The proposed strategy greatly enhances CNN-based steganalysis algorithms since the joint training process incorporates a CNN architecture-based discriminator. When the load is set to 0.1, the preemptive strategy improves detection accuracy by an average of 8 percentage points. However, with a payload of 0.3, the preemptive strategy can boost accuracy by up to 20 percentage points. Moreover, when the load exceeds 0.3, the proposed method achieves a near 100% detection success rate, eliminating all stego images.

To summarize the discussion, we propose a preemptive strategy that significantly enhances the detection precision of steganographic analysis algorithms, regardless of whether the steganographic analysis algorithm is based on a sophisticated

model or CNN.

V. Conclusion

The current image steganalysis technology exhibits poor detection performance when the embedding payload is low. In light of this issue, this paper proposes a preemptive detection strategy. The strategy involves incorporating a carefully designed imperceptible mask into the original cover image, thereby facilitating the detection of steganography in the masked cover. The proposed method consists of three main modules: preemptive-CNN, Stego-CNN, and discriminator. Each module is implemented using a differentiable convolutional neural network. During the training phase, the Stego-CNN remains unchanged while the preemptive-CNN and discriminator are jointly trained. Experimental results demonstrate that the detection of steganography becomes easier when applied to the masked cover after processing the original cover using the preemptive-CNN. Moreover, the proposed method minimally impacts the visual quality of the image and remains compatible with existing steganalysis techniques.

Acknowledgment

This work was supported by Shanghai Data Exchange Corporation and the National Natural Science Foundation of China (U20B2051, U1936214, U22B2047).

References

[1] J. Fridrich, *Steganography in digital media: principles, algorithms, and applications*. Cambridge University Press, 2009.

[2] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *2012 IEEE International workshop on information forensics and security (WIFS)*, IEEE, 2012, pp. 234–239.

[3] V. Holub and J. Fridrich, "Digital image steganography using universal distortion," in *Proceedings of the first ACM workshop on Information hiding and multimedia security*, 2013, pp. 59–68.

[4] V. Sedighi, J. Fridrich, and R. Cogranne, "Content-adaptive pentary steganography using the multivariate generalized gaussian cover model," in *Media Watermarking, Security, and Forensics 2015*, SPIE, vol. 9409, 2015, pp. 144–156.

[5] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 4206–4210.

[6] N. Zhong, Z. Qian, Z. Wang, X. Zhang, and X. Li, "Batch steganography via generative network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 88–97, 2020.

[7] Y. Cao, Z. Zhou, C. Chakraborty, *et al.*, "Generative steganography based on long readable text generation," *IEEE Transactions on Computational Social Systems*, 2022.

[8] S. Dhawan, C. Chakraborty, J. Frnda, R. Gupta, A. K. Rana, and S. K. Pani, "Ssii: Secured and high-quality steganography using intelligent hybrid optimization algorithms for iot," *IEEE Access*, vol. 9, pp. 87 563–87 578, 2021.

[9] A. Kishor, C. Chakraborty, and W. Jeberson, "Intelligent healthcare data segregation using fog computing with internet of things and machine learning," *International Journal of Engineering Systems Modelling and Simulation*, vol. 12, no. 2-3, pp. 188–194, 2021.

[10] S. Lyu and H. Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *International Workshop on information hiding*, Springer, 2002, pp. 340–354.

[11] I. Avcibas, N. Memon, and B. Sankur, "Steganalysis using image quality metrics," *IEEE transactions on Image Processing*, vol. 12, no. 2, pp. 221–229, 2003.

[12] İ. Avcıbaş, M. Kharrazi, N. Memon, and B. Sankur, "Image steganalysis with binary similarity measures," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 17, pp. 1–9, 2005.

[13] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2011.

[14] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2014, pp. 48–53.

[15] W. Tang, H. Li, W. Luo, and J. Huang, "Adaptive steganalysis against wow embedding algorithm," in *Proceedings of the 2nd ACM workshop on Information hiding and multimedia security*, 2014, pp. 91–96.

[16] R. Cogranne, V. Sedighi, J. Fridrich, and T. Pevnỳ, "Is ensemble classifier needed for steganalysis in high-dimensional feature spaces?" In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2015, pp. 1–6.

[17] S. Lyu and H. Farid, "Steganalysis using color wavelet statistics and one-class support vector machines," in *Security, steganography, and watermarking of multimedia contents VI*, SPIE, vol. 5306, 2004, pp. 35–45.

[18] T. Pevny and J. Fridrich, "Merging markov and dct features for multi-class jpeg steganalysis," in *Security, steganography, and watermarking of multimedia contents IX*, SPIE, vol. 6505, 2007, pp. 28–40.

[19] Q. Liu, A. H. Sung, and M. Qiao, "Temporal derivative-based spectrum and mel-cepstrum audio steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 359–368, 2009.

[20] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[21] A. Nissar and A. H. Mir, "Classification of steganalysis techniques: A study," *Digital Signal Processing*, vol. 20, no. 6, pp. 1758–1770, 2010.

[22] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2011.

[23] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[24] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2014, pp. 48–53.

[25] M. Chaumont, "Deep learning in steganography and steganalysis," in *Digital Media Steganography*, Elsevier, 2020, pp. 321–349.

[26] J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2545–2557, 2017.

[27] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2018.

[28] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM systems journal*, vol. 35, no. 3.4, pp. 313–336, 1996.

[29] E. Kawaguchi and R. O. Eason, "Principles and applications of bpcs steganography," in *Multimedia systems and applications*, International Society for Optics and Photonics, vol. 3528, 1999, pp. 464–473.

[30] J. Fridrich and D. Soukal, "Matrix embedding for large payloads," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 3, pp. 390–395, 2006.

[31] W. Zhang, X. Zhang, and S. Wang, "Maximizing steganographic embedding efficiency by combining hamming codes and wet paper codes," in *International Workshop on Information Hiding*, Springer, 2008, pp. 60–71.

[32] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, 2011.

[33] T. Pevnỳ, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *International Workshop on Information Hiding*, Springer, 2010, pp. 161–177.

[34] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[35] J. Fridrich and J. Kodovskỳ, "Multivariate gaussian model for designing additive distortion for steganography," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 2949–2953.

[36] D. Volkhonskiy, I. Nazarov, and E. Burnaev, "Steganographic generative adversarial networks," in *Twelfth International Conference on Machine Vision (ICMV 2019)*, International Society for Optics and Photonics, vol. 11433, 2020, p. 114333M.

[37] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 657–672.

[38] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "Cnn-based adversarial embedding for image steganography," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2074–2087, 2019.

[39] J. Yang, D. Ruan, J. Huang, X. Kang, and Y.-Q. Shi, "An embedding cost learning framework using gan," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 839–851, 2019.

[40] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[41] Y. Qian, J. Dong, W. Wang, and T. Tan, "Learning and transferring representations for image steganalysis using convolutional neural network," in *2016 IEEE international conference on image processing (ICIP)*, IEEE, 2016, pp. 2752–2756.

[42] M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-net: An efficient cnn for spatial steganalysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 2092–2096.

[43] B. Li, W. Wei, A. Ferreira, and S. Tan, "Rest-net: Diverse activation modules and parallel subnets-based cnn for spatial image steganalysis," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 650–654, 2018.

[44] W. You, H. Zhang, and X. Zhao, "A siamese cnn for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 291–306, 2020.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] X. Deng, B. Chen, W. Luo, and D. Luo, "Fast and effective global covariance pooling network for image steganalysis," in *Proceedings of the ACM workshop on information hiding and multimedia security*, 2019, pp. 230–234.

[47] N. Zhong, Z. Qian, Z. Wang, and X. Zhang, "Steganography in stylized images," *Journal of Electronic Imaging*, vol. 28, no. 3, p. 033 005, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Zero-shot multi-speaker accent TTS with limited accent data

Mingyang Zhang*, Yi Zhou†, Zhizheng Wu*, Haizhou Li*†
* Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
E-mail: zhangmingyang@cuhk.edu.cn, wuzhizheng@cuhk.edu.cn, haizhouli@cuhk.edu.cn
† Department of Electrical and Computer Engineering, National University of Singapore
E-mail: yi.zhou@u.nus.edu

***Abstract*—In this paper, we present a multi-speaker accent speech synthesis framework. It can generate accented speech of unseen speakers using only a limited amount of accent training data. Without relying on the accent lexicon, the proposed network is able to learn the accent phoneme embedding via a simple model adaptation. In specific, a standard multi-speaker speech synthesis is first trained with native speech. Then, an additional neural network module is appended for adaptation to map the native speech to the accented speech. In the experiments, we have synthesized English speech with Singapore and Hindi accents. Both objective and subjective evaluation results successfully confirm that our proposed technique with phoneme mapping is effective to generate high-quality accent speech for unseen speakers.**

I. Introduction

With an increasing demand for personalized speech synthesis, text-to-speech (TTS) synthesis systems need to generate speech not only in the target speaker's voice, but also the target accent. Accent speech synthesis is a task seeking to generate high-quality accent speech in the target speaker's voice. Accent speech synthesis has several significant real-world applications that make it an important area of research and development. Accent TTS can be used to create language learning tools that provide learners with exposure to different accents and dialects. In the entertainment industry, accent TTS can be used to create more immersive experiences in video games and movies. Accent TTS can be employed in virtual assistants and customer service applications to provide users with a personalized and culturally relevant experience. In some cases, accent TTS can play a role in preserving and documenting endangered languages and dialects, contributing to linguistic diversity and heritage conservation.

Accent refers to the variations in pronunciation, intonation, and individual speech habits associated with specific geographical regions or linguistic communities. These variations can include differences in vowel and consonant sounds, rhythm, stress patterns, and prosody. Accents may also encompass dialectical differences and linguistic features unique to certain cultural or regional groups. In the accent TTS, we are primarily addressing the pronunciation and prosodic aspects of speech associated with different regions or linguistic backgrounds.

Most of the previous studies on accent speech synthesis are based on the traditional speech synthesis system. By employing FestVox [1], a voice generation toolkit, the researchers obtained paired speech data between the accent phoneme representations and accent speech features. Therefore, accent speech can be generated directly from the accent phoneme representation. This approach has been adopted for Indian accent languages [2]–[4]. Another attempt utilizes the grapheme-to-phoneme (G2P) to construct the accent phoneme space and generate an accent phoneme representation [5]. These methods have demonstrated their effectiveness in accent speech synthesis, where phoneme plays a key role to characterize the accent pronunciation. However, their success rely on the accent language lexicon, which is difficult to obtain from limited accent data.

Alternatively, some works imitated the application of speaker embedding in speaker adaptation[6]–[8], and implemented an accent encoder to learn a fixed dimension vector to represent the accent identity, which is used for accent speech generation [9], [10]. However, learning speaker embedding usually requires thousands of multi-speaker data [11]. Consequently, the accuracy of accent embedding may not be ideal. Moreover, prior foreign accent studies have reported that both phonemic and prosodic qualities essential for identifying an accent [12], [13]. It might not be enough to fully express one accent with a single vector. Besides, the accent ID has been augmented to the input features [14] to learn the tone variance. Yet, these methods didn't account for the phoneme information change between accents and also trained in a data-driven manner requiring large training data.

With recent advances in deep learning, the performance of neural TTS systems have achieved great success. Many end-to-end speech synthesis systems can generate human-like voice with high quality [15]–[18], where sequence-to-sequence architecture is adopted. Generally, sequence-to-sequence models include an encoder to convert the input text sequence into a latent linguistic representation, an attention mechanism [19] or length regulator [20] to learn the alignment between the linguistic representation and Mel-spectral features, and a decoder to convert the aligned linguistic representation into Mel-spectral features. Such a complex network usually owns a huge parameter number, and the model training requires a large amount of data to generalize well. In the personalized TTS scenario, it learns to synthesize a new voice with few speech

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

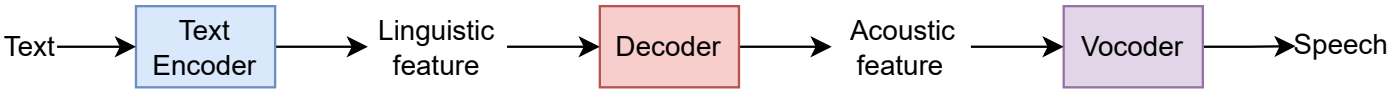


Fig. 1. The main modules in end-to-end TTS.

data from the target speaker [21]. Using limited data to train the large TTS model may result in under-fitting problem, thus leading to poor performance. Speaker adaptation successfully addresses the problem, which has been popularly adopted in personalized TTS [22], [23].

Inspired by the speaker adaptation technique, we propose an accent adaptation method for accent speech synthesis with limited data. The proposed method is able to learn the phoneme mapping between different accents. The TTS model leverages large-scale speech data by training a standard multi-speaker TTS. Then, a phoneme mapping neural network module is added for adaptation. In particular, only the phoneme mapping module and the attention will be updated during adaptation. The mapping module is carefully designed and the adaptation is dedicated for accent speech synthesis. By doing so, the mapping module is able to pick up the accent pronunciation whereas the decoder is still speaker-independent.

The rest of this paper is organized as: Section 2 introduces the proposed method; Section 3 presents the experiments and discusses the results; Section 4 concludes this work.

II. Related Work

A. End-to-end TTS

Neural TTS has experienced tremendous growth in recent years, leading to remarkable improvements in the quality of generated speech. End-to-end models (e.g., Tacotron 1/2 [15], [16], and FastSpeech 1/2 [18], [20]) drastically decrease the complexity of text analysis and allow direct input of characters or phonemes, while Mel-spectrograms simplify acoustic features. Fully end-to-end TTS systems, such as ClariNet[24] and EATS [25], are able to produce waveform directly from text, outperforming traditional concatenative and parametric synthesis in terms of both intelligibility and naturalness with minimal preprocessing and feature engineering.

A modern TTS system generally consists of three primary components, as shown in Figure 1: a text encoder module for translating text into linguistic features, a decoder module for learning alignment between linguistic and acoustic features and generating acoustic features from the output, and a neural vocoder for synthesizing waveform from the acoustic features. In this work, we focus on manipulating the linguistic aspect of TTS as it is closely related to accent-related information which will be explored further in the next section.

B. Identify of Foreign Accent

To understand what is meant by a "foreign accent," we must take into account the contact of two phonological systems, one being from a native language (L1) and the other from a non-native second language (L2). To identify a foreign accent we must be aware of the differences in cultural backgrounds, the environment of language acquisition, and any psychological and pragmatic factors that come into play. In other words, the imitation of a foreign accent involves the integration of multiple elements, so as to form an effective cognitive connection between the two different linguistic systems. But in the aspect of audio processing, both segmental and suprasegmental features are important when identifying a foreign accent, and there should be a reflection of both phonemic and prosodic aspects [13]. Therefore, when it comes to accent speech synthesis, the foremost components that need to be taken into consideration are phonemes and prosody. Phonemes are particularly essential, as the duration of each can be inferred from its representation.

The encoder part of a speech synthesis system plays an integral role in constructing the linguistic features of a synthesized speech. It generates phonemes and prosodic information, which then feed into the subsequent decode stages. In this work, we focus on transforming the linguistic features generated by the encoder to more accurately represent the desired accent. To do this, We take measures to modify the phoneme information and improve the accuracy of the prosodic characteristics. With this, we are able to constrain the generation of accent information that is better aligned to the desired outcome.

III. Methodology

Our proposed system is based on the Tacotraon2 framework [16] and the architecture of our system is illustrated in Fig. 2. The model consists of an encoder followed by a residual layer, an attention mechanism, and a decoder. The proposed accent speech synthesis approach is an accent adaption technique based on a pre-trained model.

A. TTS pre-training

First, we train a standard TTS model on a publicly accessible multi-speaker corpus. The pre-trained model has a similar architecture to Fig. 2, except that the residual layer after the encoder is absent. We follow the voice cloning technique to train the multi-speaker TTS system, as described in [8]. Instead of concatenating the speaker embedding to the encoder output, our TTS model concatenates it to the context vector that produced by the attention mechanism. In this way, the attention mechanism takes only the phoneme information as input, the alignment prediction between the phoneme representation and the Mel-spectral features should be speaker-independent.

We used a G2P module to convert the input text sequence into a standard phoneme sequence [26]. When trained with native speech data from multiple speakers, the encoder learns to transform the standard phoneme sequence to the average

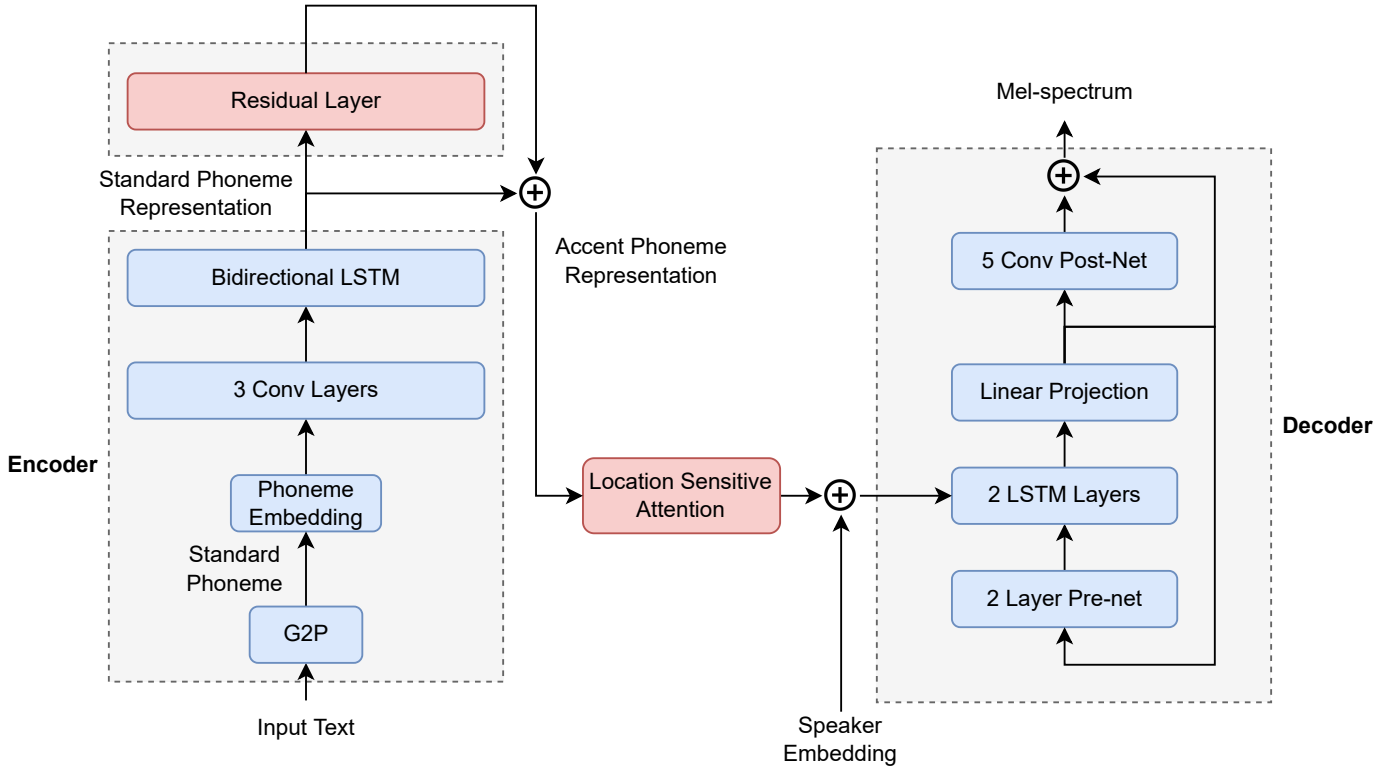


Fig. 2. Architecture of the proposed accent speech generation system

standard phoneme representation. Then the attention-based decoder will learn to convert the standard phoneme representation to the Mel-spectral features, conditioned on the speaker embedding.

The pre-trained TTS system is a well-trained zero-shot multi-speaker speech synthesis model that can generate high-quality speech with the voice of an unseen speaker.

B. Accent adaptation

According to the previous research on accent speech synthesis, the most critical step is to obtain accent-specific phoneme [2]–[5]. In the sequence-to-sequence speech synthesis system, the encoder module is primarily responsible for generating the phoneme information. If the encoder can generate an accent-specific phoneme representation from the standard phoneme sequence input, the decoder is able to generate accent sound speech accordingly by taking the accent information as input.

In many speaker adaptation works, the parameters of the entire model are updated during the adaptation stage. As a result, the model becomes speaker-dependent and can only produce the voice of one target speaker. To preserve the ability of the pre-trained model to generate multi-speaker voices during adaptation, we need to fix the decoder parameters.

It also can be risky to update all of the encoder parameters, especially with limited training data available. The distribution of phonemes in the adaptation training data is imbalanced. If all parameters of the encoder are updated, the encoder will be over-fitted to the sparse training data. This affects the accuracy of out of vocabulary phonemes generation. Therefore, we propose to use an additional module dedicated to the encoder to learn the mapping from the standard phoneme representation to the accent phoneme representation.

We have investigated the use of a standard convolutional or long-short term memory (LSTM) layer as the additional module in our work. However, the training process becomes challenging because the following attention module fails to learn how the phoneme representations and Mel-spectral features are aligned. Since the outputs of the newly added additional layer are completely new to the attention mechanism, it is difficult to infer alignment from such limited data. We propose to use the residual layer architecture to combine the output of the previous encoder output with the output of the additional layer. The residual layer does not discard the original information, but provides residual knowledge to learn the difference between the standard phoneme and the accented phoneme. In addition, this also helps in learning the phoneme variations. Since not all phonemes vary between accents, the residual layer only affects the residual information while maintaining the consistency of the remaining phonemes. Given that their functions are comparable, the residual layer is a 1D convolutional layer with ReLU activation that has the same architecture as the decoder Post-Net. The Post-Net is also provided as residual compensation for the generated Mel-spectral features.

The attention module is also essential to the accent learning.

TABLE I
HYPER-PARAMETERS AND MODEL ARCHITECTURE

Encoder	3-layer conv-5-512-RELU Bi-directional LSTM (512 units)
Residual Layer	1-layer conv-5-512-RELU
Accent Classifier	3-layer LSTM (786 units) Linear Projection (6 units)
Decoder	Pre-net: 2-layer fully connection (256 units) 2-layer LSTM (1,024 units) Linear Projection (80 units) Post-net: 4-layer conv-5-512-RELU + 1-layer conv-5-80-Linear

In addition to the phoneme variation, prosodic quantities are also necessary to form an accent [12], [13]. The attention mechanism provides the prosodic information by learning the duration from the accented phoneme representations. Therefore, only the additional residual layer and attention module are updated during adaptation, as illustrated in red color in Fig. 2. The encoder and decoder are frozen to preserve the ability to synthesize speech of different speakers. The training loss function in the adaptation stage is the same as the pre-trained model, i.e., the combination of the Mel-spectral feature loss and the stop token loss [16].

During inference, by conditioning on any speaker's speaker embedding, the model is capable of generate the target accent sound speech with any speaker's voice.

IV. EXPERIMENTS

A. Experimental Setup

To train the pre-trained TTS model, we use VCTK corpus [27]. It's a multi-speaker corpus that contains speech data uttered by 110 English speakers. For the accent speech data, we use two different accents Singapore and Hindi data to verify our proposed approach. And for each accent, we only use 300 utterances for training. Mel-spectral features are extracted using 12.5ms frame shaft with 50ms frame length. To convert the generate Mel-spectral features to waveform, we trained a ParallelWaveGAN neural vocoder [28]. The vocoder is trained with the same data of the pre-trained TTS model and the accent data is not used.

The details of model architecture are listed in table I. Where "conv-k-c-ReLU" denotes 1-D convolution with k filter size and c output channels with ReLU activation. For optimization, we use Adam optimizer with the learning rate of 0.001. The batch size is set to 64 in our experiments.

B. Objective Evaluation

To evaluate the generation accuracy of our proposed method, we calculated the objective evaluation metrics from two aspects: phoneme and pitch.

For the phoneme accuracy evaluation, we used Google Speech-to-Text API to convert the generated samples to text, then calculate the Word Error Rate (WER) to compare. The speech recognition model we were using is United States English model. We calculate the WER of both ground-truth speech and the generated speech. The results are shown in II. We can see from the results that our proposed system can generate speech with low distortion levels and high phoneme accuracy that is closed to the ground-truth for both accents.

TABLE II
WER RESULTS OF THE GROUND-TRUTH AND GENERATED ACCENT SPEECH

	Ground-truth	Generated
Singapore accent	20.09%	26.67%
Hindi accent	25.38%	29.63%

TABLE III
RMSE, CORRELATION AND DTW DISTANCE OF PITCH IN GENERATED AND TARGET SAMPLES

	RMSE	Corr	DTW
	Singapore		
Pre-trained	38.16	0.85	17.84
Proposed	**31.53**	**0.89**	**13.11**
	Hindi		
Pre-trained	32.68	0.87	13.95
Proposed	**28.64**	**0.91**	**11.67**

For the pitch accuracy evaluation, we compute the RMSE, correlation, and DTW distance of F0 between the model outputs and the ground truth speech. We compare the results between the pre-trained model and our proposed model. We use the pre-trained model to generate non-accent sound speech using the accent speaker's embedding. Evaluation results are shown in Table III. The F0 comparison demonstrates that our proposed method outperforms the pre-trained model in terms of the pitch accuracy.

C. Subjective Evaluation

We also conduct listening experiments to evaluate the similarity of accents and speakers, as well as speech quality, in addition to the objective evaluations. The objective evaluation only applies to the seen speaker because there is no ground-truth for the accented speech of unseen speakers. We evaluate the performance of unseen speaker generation in this section as well [1].

In the accent and speaker similarity test, we generate accented speech of both seen speaker and unseen speakers. Listener are required to compare the accent and speaker similarity between the generated speech and target speech, and select one of the four options for each sample: "Different, Sure", "Different, Not sure", "Same, Not sure" and "Same, Sure". The similarity test results are shown in Fig. 3 and Fig. 4. From both results, we can see that our proposed system can achieve high performance in terms of accent and speaker similarity for both seen and unseen speakers. The unseen speaker generation is slightly worse than the seen speaker generation, but still outperforms the pre-trained model.

To evaluate the overall speech quality of the proposed system, we conduct mean opinion score (MOS) test for each TTS system and the target. Listeners are required to listen to all samples one by one and rate the quality and naturalness of the

[1]Speech samples are at: https://arkhamimp.github.io/AccentTTS/

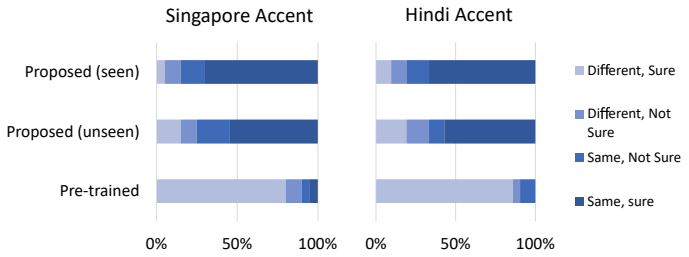


Fig. 3. Accent similarity test results between the generated speech and the target

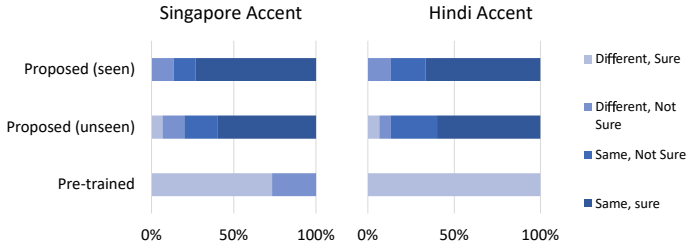


Fig. 4. Speaker similarity test results between the generated speech and the target

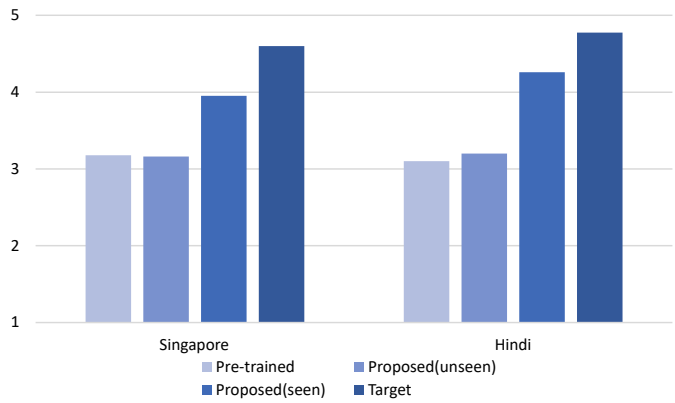


Fig. 5. MOS test results of different models and the ground-truth target

converted speech on a 5-point scale. The higher score indicates a better speech quality. Fig. 5 shows the MOS test results. Similar conclusions are also revealed by the MOS test that the unseen speaker generation achieves comparable performance with the pre-trained model, while the speech quality of the seen speaker generation is significantly improved.

V. Conclusion

This paper presents a multi-speaker accent speech synthesis system that capable of generate any speaker's voice with the target accent while only use a little amount of accent speech training data. Both the objective evaluation and subjective evaluation experiments demonstrate our proposed system outperform the pre-trained system in terms of both accent and speaker similarity and speech quality.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 62271432.

References

[1] G. K. Anumanchipalli, K. Prahallad, and A. W. Black, "Festvox: Tools for creation and analyses of large speech corpora," in *Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia*, 2011, p. 70.

[2] S. Kishore, R. Sangal, and M. Srinivas, "Building hindi and telugu voices using festvox," *Proceedings of ICON*, pp. 303–309, 2002.

[3] M. Waseem and C. Sujatha, "Speech synthesis system for indian accent using festvox," *International journal of Scientific Engineering and Technology Research, ISSN*, pp. 2319–8885, 2014.

[4] S. Kayte, M. Mundada, and D. C. Kayte, "Speech synthesis system for marathi accent using festvox," *International Journal of Computer Applications*, vol. 130, no. 6, pp. 38–42, 2015.

[5] B. Kolluru, V. Wan, J. Latorre, K. Yanagisawa, and M. J. Gales, "Generating multiple-accent pronunciations for tts using joint sequence model interpolation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *Advances in neural information processing systems*, vol. 31, 2018.

[7] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*, PMLR, 2018, pp. 3683–3691.

[8] Y. Jia, Y. Zhang, R. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[9] W. Li, B. Tang, X. Yin, *et al.*, "Improving accent conversion with reference encoder and end-to-end text-to-speech," *arXiv preprint arXiv:2005.09271*, 2020.

[10] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, vol. 72, p. 101 302, 2022.

[11] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 4879–4883.

[12] J. Vaissière and P. B. de Mareüil, "Identifying a language or an accent: From segments to prosody," in *Workshop MIDL 2004*, 2004, pp. 1–4.

[13] P. B. De Mareüil and B. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, no. 4, pp. 247–267, 2006.

[14] Z. Wang, W. Ge, X. Wang, *et al.*, "Accent and speaker disentanglement in many-to-many voice conversion," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2021, pp. 1–5.

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[16] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[17] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.

[18] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA.

[19] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[20] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," *arXiv preprint arXiv:1812.05253*, 2018.

[22] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[23] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4475–4479.

[24] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HklY120cYm.

[25] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=rsf1z-JSj87.

[26] J. Park Kyubyong & Kim, *G2pe*, https://github.com/Kyubyong/g2p, 2019.

[27] J. Yamagishi, C. Veaux, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[28] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203. DOI: 10.1109/ICASSP40776.2020.9053795.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Speech Enhancement with Multi-granularity Vector Quantization

Xiaoying Zhao[‡] and Qiushi Zhu[*] and Jie Zhang[*] and Yeping Zhou[†] and Peiqi Liu[†]
[‡] University of Science and Technology of China (USTC), Hefei, China
[*] NERC-SLIP, University of Science and Technology of China (USTC), Hefei, China
[†] China Merchants Bank
E-mail: {xyzhao1123, qszhu}@mail.ustc.edu.cn, jzhang6@ustc.edu.cn
ypzhou@cmbchina.com, liupeiqi@cmbchina.com

***Abstract*—Neural network based speech enhancement (SE) has developed rapidly in the last decade. Meanwhile, the self-supervised pre-trained model and vector quantization (VQ) has achieved excellent performance on many speech-related tasks, while they are less explored on SE. As it was shown that utilizing a VQ module to discretize noisy speech representation is beneficial for speech denoising, in this work we therefore study the impact of using VQ at different layers with different number of codebooks. Different VQ modules indeed enable to extract multiple-granularity speech features. Following an attention mechanism, the contextual features extracted by a pre-trained model are fused with the local features extracted by the encoder, such that both global and local information are preserved to reconstruct the enhanced speech. Experimental results on the Valentini dataset show that the proposed model can improve the SE performance, where the impact of choosing pre-trained models is also revealed.**

I. Introduction

Speech enhancement (SE) aims at mitigating background noises and reverberation contained in noisy speech recordings, which is widely configured as a front-end in, e.g., speech recognition [1], [2], assistive hearing [3], robust speaker recognition [4]. Traditional statistics-based SE methods usually depend on the stationarity assumption, which, however hardly holds in practice. Recently, the development of deep neural networks (DNNs) facilitates a new direction for performing SE, which even shows a superiority over conventional approaches, particularly in non-stationary scenarios. In general, DNN based SE models can be implemented in a generative [5], [6] or discriminative fashion [7], [8]. The generative models aim to learn the speech distribution, where noisy speech is fed into the model to generate the estimated speech signal subject to a consistency constraint on the distributions of enhanced and clean speech. The latter is based on directly minimizing the difference between enhanced and clean signals by training discriminative models, such that the impact of ambient noises on the spectrum can be removed. The discriminative model can be implemented in the frequency domain [9], [10] or time domain [7], [11], where the frequency- domain SE is achieved by estimating a mask matrix, while the time-domain counterpart is usually implemented following an encoder-decoder fashion.

Discrete representation learning has demonstrated an efficacy in many fields, especially for low-resource tasks, where vector quantization (VQ) is often used. For example, in [12], [13] VQ was used to learn discrete latent representations to generate high quality images. In [14], [15], VQ was used for speech synthesis to generate more natural speech by learning discrete representations. In [16], scalable VQ was applied to encode multi-scale features for speech coding. Considering that learning discrete representations can yield meaningful discrete units and diminish noise components, it was shown in [17] that VQ was also applicable to SE by reconstructing speech from discrete noisy representations. However, it was also shown that VQ with a single fixed codebook cannot adequately learn the clustering units at different granularities, which limits its applicability for SE.

In the speech domain, self-supervised pre-training models [18]–[22], e.g., wav2vec2.0 [18], HuBERT [20], data2vec [21], have been rapidly developed by exploiting large amounts of unlabeled data, which were shown to be helpful for various downstream tasks. The difference therein is that indeed wav2vec2.0, HUBERT and data2vec perform contrastive learning, masked prediction and regression tasks, respectively. The SE performance can also be boosted by using self-supervised pre-trained models. In [23], thirteen pre-trained models were applied to SE tasks, which are taken as feature extractors to estimate spectral masks to reconstruct the clean speech waveform. This implies that the high-level features extracted by self-supervised pre-trained models are applicable to SE, which can also be combined with traditional acoustic features [24]. These pre-trained features are all utilized in the estimation of time-frequency mask, but not validated in the U-Net based SE models.

In this paper, based on [17], which uses a single VQ module with a fixed codebook, we therefore explore how VQ in combination with pre-training models can be better leveraged for the SE task. Considering that different layers of the SE model learn different granularity of information, we first propose a multi-granularity VQ technique to model different levels of information by utilizing different VQ at different layers. Specifically, different VQ modules have different numbers of codebooks and learnable codevectors. Compared to [17], it is shown that the performance can be improved due to the use of multi-granularity VQ at the cost of a small increase in the model parameters. Secondly, to enable the model to

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

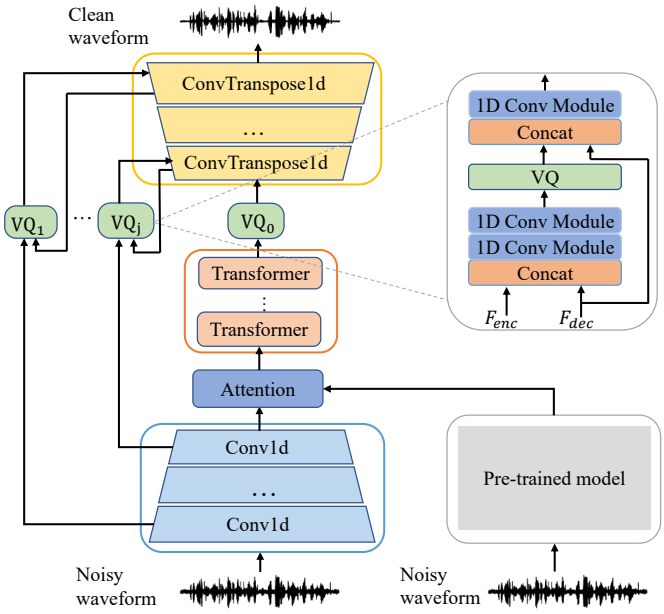


Fig. 1. An illustration of the proposed SE model with multi-granularity VQ modules and self-supervised pre-trained model. The feature fusion mechanism of the proposed multi-granularity VQ is shown at the upper right corner.

better make use of global and local information, we fuse the contextual features extracted by the pre-trained model with the local features extracted by the encoder of the SE model. Both the contextual and local features are input to the U-Net framework. This differs from the scheme in [23], [24], where only the features extracted by the pre-trained model are incorporated to estimate the time-frequency mask, as the considered features for the proposed method become much richer. In addition, it is shown that different pre-trained models have a different effect on the performance, where the data2vec model performs the best with nearly the same amount of parameters. Experimental results on the Valentini dataset demonstrate that utilizing the multi-granularity VQ in different layers is beneficial for SE. The rest of this paper is organized as follows. Section 2 presents the proposed SE method. Experimental setup and results are shown in Section 3 and Section 4, respectively. Finally, Section 5 concludes this work.

II. Method

To guide the reader, in this section we first briefly introduce the SE model in [17], which is the basis of the proposed method. The SE model follows the U-net structure, which contains convolution-based encoders and decoders and the Transformer encoder based bottleneck layer. This is similar to the DEMUCS network. The VQ module is used to discretize the representations output of the bottleneck layer, which is then fed to the decoder to reconstruct clean speech waveforms, and it is shown in Section 4 that both the VQ module and pre-trained models can be beneficial for the performance. However, we found that a single VQ cannot learn speech representations at different granularities. For more details, we refer to [17].

A. Model structure

The structure of our proposed model is shown in Fig. 1, which mainly contains encoder, bottleneck layer, multi-granularity VQ and decoder. The encoder and decoder contain D-layer one-dimensional (1D) convolution and D-layer 1D transposed convolution, respectively, and the bottleneck layer consists of N-layer Transformer encoder. A skip connection is utilized between the i-th encoder layer and the i-th decoder layer. The encoder adopts a convolutional module that can efficiently extract local features. To better utilize the pre-trained model to model long-time contextual information, we take the pre-trained model as a feature encoder. Local features and global features are fed into the attention network for fusion. The fused features are sent to the bottleneck layer, whose output is fed to the decoder after being discretized by quantizer VQ_0. In order to incorporate discrete units of different granularity, we utilize different vector quantizers. Specifically, the output of the i-th encoder layer and the output of the i-th decoder layer are fused into VQ_i to obtain the i-th discrete representation, which is then fused with the output of the i-th decoder layer.

B. Multi-granularity VQ

The proposed multi-granularity VQ mechanism is shown in the upper right corner of Figure. 1, which mainly contains the fusion VQ modules. We first concatenate the output of the i-th encoder layer and the output of the i-th decoder layer. The combined features are fed into the two-layer 1D convolution to obtain the fused features,which are input into the VQ module to obtain the discrete features. Then, the discrete features and the decoder output are concatenated and fed to the 1D convolution module to construct the final output. The VQ module is built on production quantization, where different number of codebooks and codewords for different layers are considered. Given G codebooks, each with V learnable d-dimensional codewords, we first map the input feature to $\boldsymbol{l} \in R^{G \times V}$ logits, and then select the discrete vectors by the Gumbel-softmax [25] operation in a differential way. The probability of choosing the v-th codeword from the g-th codebook is given by

$$\overline{p}_{g,v} = \frac{\exp(\bar{l}_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp(\bar{l}_{g,k} + n_k)/\tau}, \tag{1}$$

where $n_v = -\log(-\log(u))$ with u uniformly distributed over [0, 1] and τ is a non-negative temperature coefficient, In the forward stage, the code vector is selected by $argmax$ operation, and in the backward propagation stage the true gradient of the Gumbel-softmax output is utilized. As it is expected that more codewords can be used for quantization by maximizing the softmax distribution l, we consider the

diversity loss function L_d, which is given by

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \overline{p}_{g,v} \log \overline{p}_{g,v}. \quad (2)$$

C. The total loss function

For each VQ module, a codebook diversity loss function is utilized. The overall loss function consists of the SE loss function and weighted codebook diversity loss functions, given by

$$L_{\text{total}} = L_{\text{se}} + \sum_{i=0}^{D} \lambda_i L_{d_i}, \quad (3)$$

where L_{d_i} is the codebook diversity loss of the i-th VQ and λ_i is the weighting parameter, and L_{se} is the SE loss function, which is given by [7], [17]

$$L_{\text{se}} = \frac{1}{T} \left(||\boldsymbol{y} - \hat{\boldsymbol{y}}||_1 + \sum_{i=1}^{M} L_{\text{stft}}^{(i)}(\boldsymbol{y}, \hat{\boldsymbol{y}}) \right), \quad (4)$$

where

$$L_{\text{stft}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\text{sc}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + L_{\text{mag}}(\boldsymbol{y}, \hat{\boldsymbol{y}}), \quad (5)$$

$$L_{\text{sc}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{|||\text{STFT}(\boldsymbol{y})| - |\text{STFT}(\hat{\boldsymbol{y}})|||_F}{||\text{STFT}(\boldsymbol{y})||_F}, \quad (6)$$

$$L_{\text{mag}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T} || \log |\text{STFT}(\boldsymbol{y})| - \log |\text{STFT}(\hat{\boldsymbol{y}})|||_1, \quad (7)$$

where T denotes the speech length, $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are the clean and enhanced speech signals, respectively, M denotes the number of multi-scale STFT loss functions, $||\cdot||_F$ the Frobenius norm and $||\cdot||_1$ the L_1 norm. The multi-resolution loss $L_{\text{stft}}^{(i)}$ utilizes the STFT loss with the number of FFT bins ranging from {512, 1024, 2048}, hop size from {50, 120, 240} and window length from {240, 600, 1200}, respectively.

III. Experimental setup

A. Data description

To evaluate the effectiveness of the proposed method, we utilize the Valentini [32] dataset, which contains 28.4 hours of clean-noisy speech pairs consisting of 34647 mono audio samples for training and 824 mono audio samples for testing at a sampling rate of 48 kHz. The duration of the audio ranges from 1 second to 15 seconds (the average duration is around 3 seconds). These data are collected from 84 speakers with 10 different types of noises added to the clean speech at 4 signal-to-noise ratios (SNRs) (i.e., 0, 5, 10 and 15 dB) in the training set and 5 types of noise added to the speech with 4 SNRs (2.5, 7.5, 12.5 and 17.5 dB) in the test set, and all speakers are native English. The raw speech waveforms are downsampled to 16kHz and then applied the remix and bandmask techniques for data augmentation as in [7].

The performance of the proposed method is evaluated using objective metrics, including 1) perceptual evaluation of speech quality (PESQ) [33], 2) short-time objective intelligibility (STOI) [34], 3) mean opinion score (MOS) prediction of distortion of speech signal (SIG) [35], 4) MOS prediction of intrusiveness of background noise (BAK) [35], 5) MOS prediction of overall quality (OVRL) [35].

B. Model configuration

Our network is built on the basis of the DEMUCS model [7], which is modified by replacing the LSTM with Transformer and using an attention module as feature fusion module. Both the encoder and decoder have D = 5 layers and the dimension is 512. We utilize 2 Transformer encoders in the bottleneck layer, where each has a dimension of 512, the feedforward neural network has a dimension of 2048, and the self-attention module has 12 heads. The 1D convolution in the encoder uses a kernel size of 8 and a stride size of 2. For the pre-trained models, we use the publicly available wav2vec2.0 base[1], HuBERT base[2] and data2vec base[3] models in experiments. All three models are obtained by self-supervised pre-training using the same unlabeled data and have similar parameters. These model parameters are not used for training, and we only use them to extract features. For the attention fusion module, we adopt a cross-attention layer with a dimension of 512. For VQ_0 and $VQ_i (i = 1, 2, 3, 4, 5)$ modules, we adopt G = 2 codebooks with V = 320 and $G = 1$ codebook with V = {320, 640, 960 2560, 5120} learnable 128-dimensional vectors, respectively. λ_i in (3) is set to be 0.01 due to the range of the diversity loss, since the amplitude of the codebook diversity loss function L_d is approximately 100 times larger than that of the SE loss function L_{se}. We use the Adam optimizer to train the SE model for 1M iterations, where the batch size is 30 and the maximum learning rate is 2×10^{-4}. All models are trained on 4 Tesla-V100-32G GPUs.

IV. Experimental results

Comparison methods: We first measure objective evaluation metrics on the noisy test set as a reference. The time-domain comparison methods include SEGAN [26], SEGAN-D [28] and MetricGAN [30], which are based on the generative adversarial network (GAN). The time-frequency domain comparison approaches include MMSE-GAN [29] and Deep-MMSE [31], which estimates the noise spectral density under the minimum mean square error (MMSE) criterion. Wave U-Net [27], DEMUCS [7] and CleanUNet [8] are also compared, which depend on the U-net network in the time domain.

Table I shows the experimental results of different models on the Valentini dataset. It can be seen that the proposed method using pre-trained model and multi-granularity VQ can achieve a PESQ of 3.18 and an STOI of 96.7, respectively, which is the best among all comparison methods and is better than the method in [17], which utilizes a fixed codebook and adopts a pre-trained model to initialize the encoder. This implies that using multi-granularity VQ modules and a pre-trained model as the feature extractor is more effective than the single-fixed counterpart. In addition, the deployment of multiple

[1]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[2]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert
[3]https://github.com/facebookresearch/fairseq/tree/main/examples/data2vec

TABLE I
PERFORMANCE COMPARISON OF OBJECTIVE EVALUATION METRICS ON THE VALENTINI TEST SET.

Model	Domain	PESQ (WB)	STOI (%)	pred. CSIG	pred. CBAK	pred. COVL
Noisy	-	1.97	92.1	3.35	2.44	2.63
SEGAN [26]	waveform	2.16	-	3.48	2.94	2.80
Wave U-Net [27]	waveform	2.40	-	3.52	3.24	2.96
SEGAN-D [28]	waveform	2.39	-	3.46	3.11	3.50
MMSE-GAN [29]	time-frequency	2.53	93.0	3.80	3.12	3.14
MetricGAN [30]	waveform	2.86	-	3.99	3.18	3.42
DeepMMSE [31]	waveform	2.95	94.0	4.28	3.46	3.64
DEMUCS [7]	waveform	3.07	95.0	4.31	3.40	3.63
CleanUNet [8]	waveform	3.09	95.8	4.38	3.47	3.69
Zhao et al. [17]	waveform	3.13	96.1	4.46	3.56	3.82
Ours	waveform	3.18	96.7	4.53	3.64	3.85

TABLE II
THE IMPACT OF DIFFERENT MODULES ON THE PERFORMANCE.

Model	PESQ (WB)	STOI (%)	pred. CSIG	pred. CBAK	pred. COVL
Base	3.07	95.2	4.31	3.43	3.64
+ Multi-VQ	3.12	96.2	4.42	3.53	3.77
+Wav2vec2.0	3.14	96.3	4.48	3.58	3.83
+HuBERT	3.14	96.4	4.48	3.59	3.83
+Data2vec	3.18	96.7	4.53	3.64	3.85

TABLE III
THE IMPACT OF VQ OF DIFFERENT LAYERS.

VQ0	VQ1	VQ2	VQ3	VQ4	VQ5	PESQ (WB)	STOI (%)
✓	✓	✓	✓	✓	✓	3.12	96.2
✗	✓	✓	✓	✓	✓	3.09	95.6
✓	✗	✓	✓	✓	✓	3.12	96.0
✓	✓	✗	✓	✓	✓	3.11	96.0
✓	✓	✓	✗	✓	✓	3.10	95.9
✓	✓	✓	✓	✗	✓	3.10	95.8
✓	✓	✓	✓	✓	✗	3.10	95.8

codebooks only results in a small increment of the parameter amount. (The codebook has codevectors that are learnable, so the number of model parameters added can be calculated as $(320+640+960+2560+5120+320\times2)\times128/10^6 \approx 1.3M$.) In order to more clearly see the impact of each module, we present a comparison in Table II. The base model does not use VQ or pre-trained models, leading to a PESQ of 3.07 and an STOI of 95.2, which are similar to that of CleanUNet, because their model structures are rather similar. In case the multi-granularity VQ modules are included for the base model as shown in Figure 1, the SE performance in all metrics is clearly improved, showing the effectiveness of using multiple codebooks. By the inclusion of multiple VQs and using any pre-trained model (wav2vec2.0, HuBERT or data2vec) as the feature extractor (i.e., fusing the extracted contextual features with the local features extracted by the SE encoder), the performance can be further improved to some extent, while the features extracted from the data2vec model seem the best and the benefits from using wav2vec and HuBERT are almost the same.

Further, we experimentally analyze the impact of the VQ of different layers in Table III. It is clear that in case the VQ

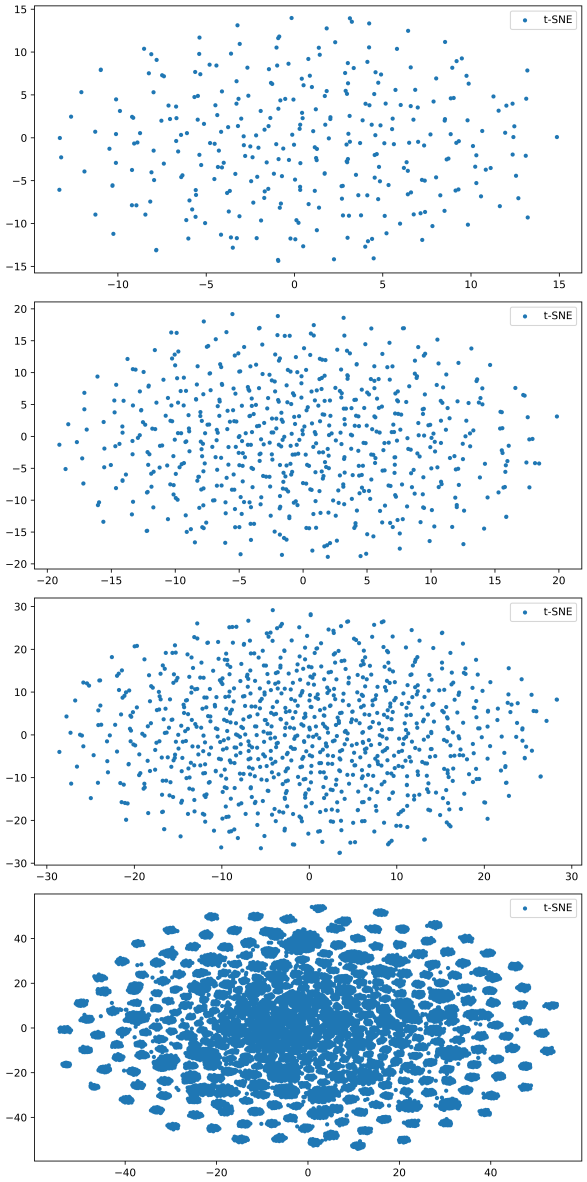


Fig. 2. Codebook visualization with different number of learnable vectors, where a single codebook with 320, 640 and 960 learnable vectors and two codebooks with 320 learnable vectors are included from top to bottom.

module is included for each layer, the best SE performance is achieved. The removal of the VQ at any layer leads to a performance decrease, meaning the necessity of the proposed multi-granularity VQ operations. It seems that the removal of VQ_0 has a more significant impact on the performance, indicating that the discretized contextual features are more beneficial for model denoising. VQ_1 has the smallest impact among all VQ modules, and a higher-layer VQ roughly affects the performance more. Due to the limited size of this dataset, these results might not reflect the analysis comprehensively, so in the future we will evaluate the proposed method in a larger-sized dataset.

Finally, we visualize the learned 128-dimensional codebooks using t-SNE, which are reduced to two dimensions and shown in Figure 2, where a single codebook with 320 (VQ_1), 640 (VQ_2) and 960 (VQ_3) learnable vectors and two codebooks with 320 learnable vectors (320×320 vectors in total, i.e., VQ_0) are included from top to bottom. It is clear that more learnable vectors used in the VQ codebook, more fine-grained speech embeddings can be extracted. From the perspective of t-SNE distributions, the speech embedding extracted by VQ_0 shows a clearer clustering property, which is useful for classification. On the other hand, the multiple embeddings are not spatially overlapped, resulting in the so-called multi-granularity speech features, they are somehow complementary to represent the speech domain. That is why the inclusion of each VQ module positively contributes to the performance in Table III.

V. Conclusion

In this paper, we investigated the application of multi-granularity VQ and pre-trained models to speech enhancement. It was shown that using different VQ modules at different layers can improve the SE performance compared to using a single-fixed VQ. The fusion of contextual features extracted by the pre-trained model and local features extracted by the encoder allows the model to better model global and local information. Compared with the self-supervised pre-trained wav2vec2.0 and HuBERT models that are dedicated to speech representation learning, data2vec seems more effective for SE as it can learn representations from general data (e.g., speech, images, video).

VI. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62101523), Hefei Municipal Natural Science Foundation (2022012), USTC Research Funds of the Double First-Class Initiative (YD2100002008).

References

[1] Z. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 796–806, 2016.

[2] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, and L.-R. Dai, "A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1927–1939, 2023.

[3] J. Zhang and G. Zhang, "A parametric unconstrained beamformer based binaural noise reduction for assistive hearing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 292–304, 2022.

[4] H. Taherian, Z. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1293–1302, 2020.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *ISCA Interspeech*, 2017, pp. 3642–3646.

[6] D. Michelsanti and Z. Tan, "Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification," in *ISCA Interspeech*, 2017, pp. 2008–2012.

[7] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *ISCA Interspeech*, 2020, pp. 3291–3295.

[8] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *ICASSP*, 2022, pp. 7867–7871.

[9] N. Shah, H. Patil, and M. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *Proc. of APSIPA ASC*, 2018, pp. 1246–1251.

[10] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *ISCA Interspeech*, 2020, pp. 2472–2476.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Proc. of NIPS*, vol. 32, 2019.

[13] S. Gu, D. Chen, J. Bao, *et al.*, "Vector quantized diffusion model for text-to-image synthesis," in *Proc. of CVPR*, 2022, pp. 10 696–10 706.

[14] C. Du, Y. Guo, X. Chen, and K. Yu, "VQTTS: High-Fidelity Text-to-Speech Synthesis with Self-Supervised VQ Acoustic Feature," in *ISCA Interspeech*, 2022, pp. 1596–1600.

[15] Y. Liu, R. Xue, L. He, X. Tan, and S. Zhao, "DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders," in *ISCA Interspeech*, 2022, pp. 1581–1585.

[16] X. Jiang, X. Peng, H. Xue, Y. Zhang, and Y. Lu, "Cross-Scale Vector Quantization for Scalable Neural Speech Coding," in *ISCA Interspeech*, 2022, pp. 4222–4226.

[17] X. Zhao, Q. Zhu, and J. Zhang, "Speech enhancement using self-supervised pre-trained model and vector quantization," in *Proc. of APSIPA ASC*, 2022, pp. 330–334.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. of NIPS*, vol. 33, pp. 12 449–12 460, 2020.

[19] Q. Zhu, J. Zhang, M. Wu, X. Fang, and L. Dai, "An improved wav2vec 2.0 pre-training approach using enhanced local dependency modeling for speech recognition.," in *ISCA Interspeech*, 2021, pp. 4334–4338.

[20] W. Hsu, B. Bolte, Y. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[21] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Int. Conf. Machine Learning (ICML)*, vol. 162, 2022, pp. 1298–1312.

[22] Q. Zhu, J. Zhang, Z. Zhang, M. Wu, X. Fang, and L. Dai, "A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition," in *ICASSP*, 2022, pp. 3174–3178.

[23] Z. Huang, S. Watanabe, S. Yang, P. García, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *ICASSP*, 2022, pp. 6837–6841.

[24] K. Hung, S. Fu, H. Tseng, H. Chiang, Y. Tsao, and C. Lin, "Boosting Self-Supervised Embeddings for Speech Enhancement," in *ISCA Interspeech*, 2022, pp. 186–190.

[25] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[26] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *ISCA Interspeech*, 2017, pp. 3642–3646.

[27] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[28] H. Phan, I. McLoughlin, L. Pham, *et al.*, "Improving gans for speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1700–1704, 2020.

[29] M. Soni, N. Shah, and H. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *ICASSP*, 2018, pp. 5039–5043.

[30] S. Fu, C. Liao, Y. Tsao, and S. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Int. Conf. Machine Learning (ICML)*, vol. 97, PMLR, 2019, pp. 2031–2041.

[31] Q. Zhang, A. Nicolson, M. Wang, K. Paliwal, and C. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1404–1415, 2020.

[32] C.Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[33] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[34] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[35] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-party Meetings

Mohan Shi*, Jie Zhang*, Zhihao Du†, Fan Yu†, Qian Chen†, Shiliang Zhang†, Li-Rong Dai*
* NERC-SLIP, University of Science and Technology of China (USTC), China
E-mail: smohan@mail.ustc.edu.cn, {jzhang6; lrdai}@ustc.edu.cn
† Speech Lab of DAMO Academy, Alibaba Group, China
E-mail: {neo.dzh; yf352572; tanqing.cq; sly.zsl}@alibaba-inc.com

***Abstract*—Speaker-attributed automatic speech recognition (SA-ASR) in multi-party meeting scenarios is one of the most valuable and challenging ASR tasks. It was shown that Single-Channel Frame-level Diarization with Serialized Output Training (SC-FD-SOT), Single-Channel Word-level Diarization with SOT (SC-WD-SOT), and joint training of Single-Channel Target-Speaker separation and ASR (SC-TS-ASR) can be exploited to partially solve this problem, where the latter achieves the best performance. In this paper, we propose three corresponding MultiChannel (MC) SA-ASR approaches, namely MC-FD-SOT, MC-WD-SOT, and MC-TS-ASR. For different tasks/models, different multichannel data fusion strategies are considered, including channel-level cross-channel attention for MC-FD-SOT, frame-level cross-channel attention for MC-WD-SOT and neural beamforming for MC-TS-ASR. Experimental results on the AliMeeting corpus reveal that our proposed multichannel SA-ASR models can consistently outperform the corresponding single-channel counterparts in terms of the speaker-dependent character error rate (SD-CER).**

I. Introduction

Multi-talker speech modeling in real meeting scenarios (also known as the cocktail party problem) is one of the most valuable and challenging problems in speech signal processing [1], [2], where speaker-attributed automatic speech recognition (SA-ASR) is expected to correctly identify "who spoke what" [3]–[5]. Therefore, SA-ASR not only needs to recognize the possible overlapping speech sounds of different speakers [6]–[8], but also needs to identify the corresponding speaker for each recognized utterance.

Serialized **O**utput **T**raining (**SOT**) [8] was proposed for multi-talker ASR in order to avoid the limit of the maximum number of speakers and duplicated hypotheses [6], [9], which introduces a special symbol to represent the speaker change with only one output layer. To obtain the speaker attribute transcription, three **S**ingle-**C**hannel (SC) SA-ASR approaches were proposed by modular combination [10], namely **S**ingle-**C**hannel **F**rame-level **D**iarization with **SOT** (**SC-FD-SOT**), **S**ingle-**C**hannel **W**ord-level **D**iarization with **SOT** (**SC-WD-SOT**) and **S**ingle-**C**hannel **T**arget-**S**peaker separation and **ASR** (**SC-TS-ASR**). In SC-FD-SOT, the frame-level speaker diarization and SOT-based ASR are combined to obtain the speaker-attributed transcription. However, due to the modular independence, such an alignment strategy may suffer from erroneous timestamps, which severely hinder the speaker assignment performance. In order to alleviate this problem, SC-WD-SOT was proposed, which uses an attention mechanism to diarize the hypotheses of SOT-based ASR at the token level to obtain the corresponding speaker of each character. In order to further get rid of the dependence on SOT-based ASR hypotheses, SC-TS-ASR was proposed to train the joint model of target speaker separation and ASR, which achieves the best SA-ASR performance.

Although these approaches can work well on real-meeting corpora, they all use single-channel audio after beamforming as the input of the model. However, in real meeting scenarios, we usually have multichannel signals recorded using microphone arrays, which can be used more efficiently inside the model to learn richer temporal and spatial information. In this work, we therefore propose three **M**ulti**C**hannel SA-ASR approaches accordingly, called **MC-FD-SOT**, **MC-WD-SOT** and **MC-TS-ASR**, respectively.

In literature, **F**rame-**L**evel **C**ross-**C**hannel **A**ttention (**FLCCA**) [11], [12] and **C**hannel-**L**evel **C**ross-**C**hannel **A**ttention (**CLCCA**) [13], [14] were proposed in the context of ASR, speech separation and speaker diarization, where the former attends more on the time frames and the latter more on channels. **M**ulti-**F**rame **C**ross-**C**hannel **A**ttention (**MFCCA**) [15] reaches a tradeoff in-between by integrating cross-channel information between adjacent frames. As the MFCCA-based SOT works best for the multi-talker ASR and outperforms the best system in M2MeT challenge [16]–[18], we adopt the MFCCA-based SOT to replace SOT in SC-FD-SOT and SC-WD-SOT, namely MC-SOT. In track 1 of M2MeT, as the CLCCA was applied to **T**arget-**S**peaker **V**oice **A**ctivity **D**etection (**TS-VAD**) [19], [20] to fuse multichannel speech information and achieved the best performance (i.e., MC-TS-VAD [14]), we propose to combine MC-TS-VAD with MC-SOT, i.e., MC-FD-SOT. As for word-level diarization, the text and speech representations based attention is required, which means the importance of inter-frame speech context, we therefore adopt FLCCA in word-level diarization to fuse multichannel information. Combined with multichannel

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

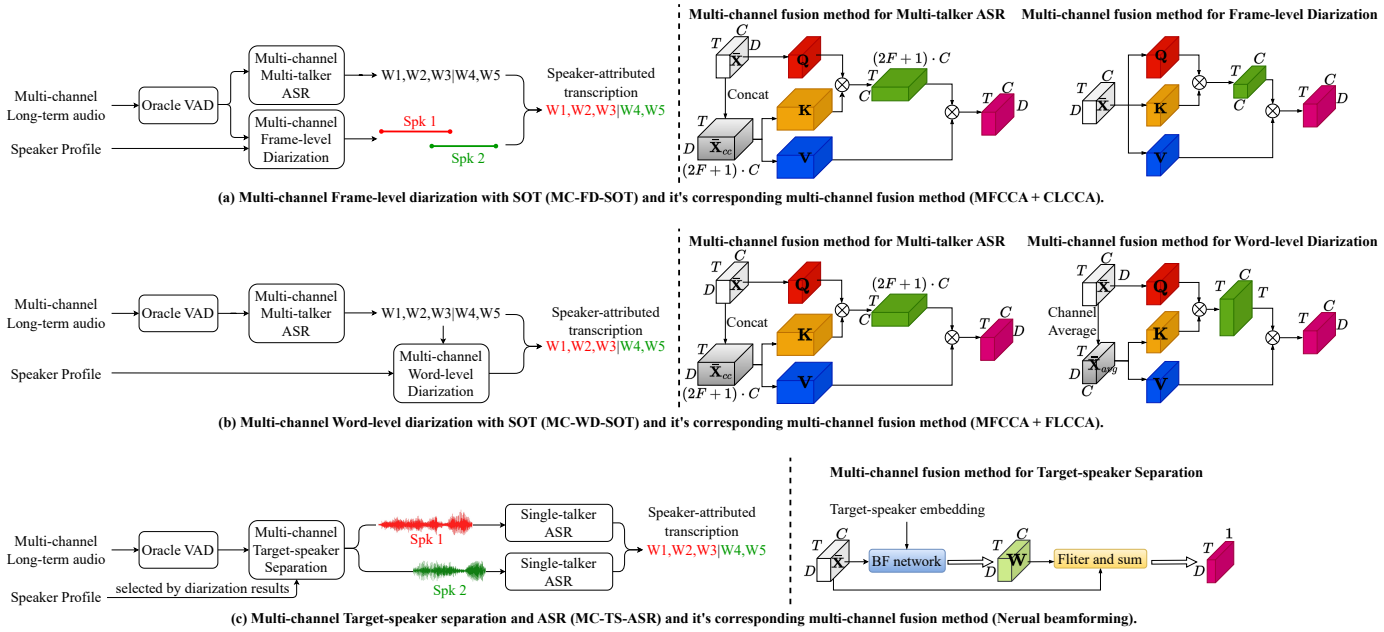


Fig. 1. The diagram of MC-FD-SOT, MC-WD-SOT and MC-TS-ASR and the corresponding simple diagram of multichannel fusion.

word-level diarization and MC-SOT, MC-WD-SOT is then obtained. Neural beamforming has been widely used for multichannel speech separation recently [21]–[24], where a trainable neural network is used to adapt the filter weights and the single-channel output speech spectrum is then given by the filter-and-sum operation. Considering the noise reduction capacity of neural beamforming techniques, we further propose the MC-TS-ASR method by jointly training the pre-trained neural beamforming based multichannel target speaker separation and ASR modules.

Experimental results on a real meeting scenario corpus, AliMeeting, show that the MC-FD-SOT approach achieves an average speaker-dependent character error rate (SD-CER) of 33.5%, which outperforms the SC-FD-SOT. The MC-WD-SOT approach achieves an average SD-CER of 30.7%, which is better than SC-WD-SOT. The MC-TS-ASR approach obtains the best performance with an average SD-CER of 28.3%, which is still better than SC-TS-ASR. The remainder of the paper is organized as follows. Section 2 presents the proposed MC-SA-ASR methods, including MC-FD-SOT, MC-WD-SOT, and MC-TS-ASR. Section 3 shows the experimental setup, followed by results in Section 4. Finally, Section 5 concludes this work.

II. MC-SA-ASR

A. MFCCA-based MC-SOT

The FLCCA attends more to the information between frames but simply averages the information between channels, while the CLCCA applies attention between channels but ignores the information between local frames. The MFCCA strategy can be regarded as a combination of FLCCA and CLCCA by attending to both the channel-level and frame-level information.

The i-th head of MFCCA is calculated as

$$\mathbf{Q}_i = \bar{\mathbf{X}}\mathbf{W}_i^q + (\mathbf{B}_i^q)^\top \in R^{T\times C\times D}, \quad (1)$$

$$\mathbf{K}_i = \bar{\mathbf{X}}_{cc}\mathbf{W}_i^k + (\mathbf{B}_i^k)^\top \in R^{T\times (2F+1)\cdot C\times D}, \quad (2)$$

$$\mathbf{V}_i = \bar{\mathbf{X}}_{cc}\mathbf{W}_i^v + (\mathbf{B}_i^v)^\top \in R^{T\times (2F+1)\cdot C\times D}, \quad (3)$$

$$\mathbf{H}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{D}}\right)\mathbf{V}_i \in R^{T\times C\times D}, \quad (4)$$

where T, C, D stand for time, channel and feature dimension, respectively, $(\cdot)^\top$ denotes transpose operation, F is the number of local context frames at each time step, $\bar{\mathbf{X}}_{cc} = [\bar{\mathbf{X}}_{cc}^0, \cdots, \bar{\mathbf{X}}_{cc}^T]$ with $\bar{\mathbf{X}}_{cc}^t, \forall t \in \{0, \cdots, T\}$ being the concatenation of the context frames at time step t, which is constructed as $\bar{\mathbf{X}}_{cc}^t = [\bar{\mathbf{X}}^{t-F}, \cdots, \bar{\mathbf{X}}^t, \cdots, \bar{\mathbf{X}}^{t+F}] \in R^{(2F+1)\cdot C\times D}$, $\mathbf{W}^*$ and $\mathbf{B}^*$ are learnable parameters.

It can be clearly seen that the MFCCA can combine both channel-level and frame-level information, and it was shown in [15] that the combination with the attention based encoder-decoder (AED) to configure an MFCCA-based multichannel ASR model is beneficial for performance with a small increase in the computational cost. Considering the issues of overlapping speech and an unknown number of speakers in real-world meeting scenarios, the MFCCA was combined with SOT to tackle the multichannel multi-speaker ASR task, which is called MC-SOT. It was shown that since the SOT does not require a defined number of speakers and enables to model the transcribed dependencies of different speakers and transcripts from different speakers are concatenated in line with the corresponding starting time with a separator $\langle$sc$\rangle$ in-between, the SOT can thus work better than permutation invariant training (PIT) [6].

B. MC-FD-SOT

As the MC-TS-VAD method performs the best in the M2MeT Challenge Speaker Diarization Track [14], which utilizes a CLCCA-based TS-VAD to fuse multichannel features (e.g., the diarization error rates (DERs) are 2.26% and 2.98% on AliMeeting eval and test sets, respectively). As shown in Fig. 1(a), here we combine MC-TS-VAD and MC-SOT by aligning their timestamps to obtain the speaker-attributed transcription.

The detailed procedure of MC-FD-SOT is the same as SC-FD-SOT [10]. First, we estimate the utterance number $\hat{N}$ by diarization output of MC-TS-VAD and oracle sentence segmentation, and the utterance number of MC-SOT output is defined as N. In case $\hat{N} = N$, no further effort is required; if $\hat{N} > N$, we select the top-N longest utterances from the output of MC-TS-VAD and discard the rest; otherwise we select the $\hat{N}$ utterances with the longest text length from the MC-SOT output and discard the others. Finally, we align the utterances between MC-TS-VAD and MC-SOT in a chronological order.

C. MC-WD-SOT

In SC-WD-SOT, three independent encoders are used to encode the multi-talker hypotheses, speech features and speaker embeddings. The encoded hypotheses $H = \{\mathrm{h}_l | l = 1, \ldots, L\}$ and encoded features $X = \{\mathrm{x}_t | t = 1, \ldots, T\}$ are input into an attention layer, where H is viewed as query, X as key and value, and the output is the word-level representation $R = \{\mathbf{r}_l | l = 1, \ldots, L\}$. Next, the context-independent (CI) score $S_{l,n}^{CI}$ is calculated using the dot product between the encoded speaker embeddings $V = \{\mathrm{v}_n | n = 1, \ldots, N\}$ and the aggregated representations R. Meanwhile, the self-attention based network (SAN) [25] is also applied to H and V to compute the context-dependent (CD) score $S_{l,n}^{CD}$. Finally, the CI and CD scores are concatenated and fed to a post-processing network to predict the corresponding speaker for each character.

In MC-WD-SOT, the FLCCA is used in prior to the speech encoder to fuse multichannel speech information. As shown in Fig. 1(b), it can better extract the frame-level speech context information while fusing the cross-channel information. FLCCA is followed by 2D-convolution to obtain the single-channel representation required by the speech encoder. The speaker attribute transcription can then be obtained by merging the multichannel word-level diarization results with the MC-SOT hypotheses.

D. MC-TS-ASR

The multichannel target speaker separation module uses the given target speaker embedding to separate the multichannel mixed speech into single-channel target speaker representations, and the transcription of each speaker is then obtained by joint modeling through the ASR module, e.g., see Fig. 1(c). In general, multichannel speech separation can be achieved by neural beamforming techniques in e.g., [21]–[24].

In this work, we adopt the embedding and beamforming network (EaBNet) [24] for multichannel target speaker separation, which is an effective neural beamforming approach. The EaBNet mainly consists of an embedding module (EM) and beamforming module (BM), where the former follows an encoder-TCN-decoder structure. The encoder gradually extracts features with multiple downsampling operations, and the decoder is a mirror version of the encoder that recovers the features in the original dimensions. Temporal convolutional networks (TCNs) [26] are the bottleneck connecting the encoder and decoder, which functions as a separator. Unlike traditional beamformers [27], the BM directly uses the mapping capability of the network to estimate the beamformer weights, which consists of several LSTM layers. Finally, the desired spectrum and speech are obtained using the filtered spectrum and filter-and-sum operation. In order to better extract the target speaker, feature-wise linear modulation (FILM) [28] is used to integrate the speaker embedding into the feature before each TCN layer.

III. Experimental Setup

A. Dataset and evaluation metrics

In this work, the AliMeeting corpus [16], [17] is used to evaluate the proposed MC-SA-ASR systems. The AliMeeting corpus contains 104.75 hours data for training (Train), 4 hours for evaluation (Eval) and 10 hours for testing (Test). Each subset contains several meeting sessions and each consists of a 15 to 30-minutes discussion by 2 to 4 participants. The dataset includes 8-channel far-field audio recorded by microphone arrays (*Ali-far*) and single-channel near-field audio recorded by participants' headsets (*Ali-near*). For MC-TS-ASR, we further simulate 50 hours mixed training data (called *Train-Ali-simu*) from *Train-Ali-near* to train the front-end module as well as the corresponding eval *Eval-Ali-simu* and *Test-Ali-simu* sets.

For SA-ASR, the speaker-dependent character error rate (SD-CER) [29] is used for performance evaluation. SD-CER is calculated by matching the ASR hypothesis to the corresponding speaker reference transcription, which is a rigorous evaluation metric for meeting scenarios. Besides, we adopt the scale-invariant SNR (SI-SNR) [30] to measure the speech separation performance.

B. Model configuration

For MC-SOT, we employ a transformer model with MFCCA-based Conformer [31] encoder to construct MC-SOT, which includes 11 encoder layers and 6 decoder layers with 4-head multi-head self-attention (MHSA), where the dimensions of MHSA and feed-forward network (FFN) layer are set to be 256 and 2048, respectively. The obtained CERs on Eval and Test sets of AliMeeting corpus are 17.3% and 18.4%, respectively. For MC-WD-SOT, similarly to SC-WD-SOT, a four-layer self-attention based encoder is employed to encode the recognized text with 8 attention heads and 256 hidden units in each layer. For the speech encoder, a four-layer feed-forward sequential memory network (FSMN) [32] is adopted with 256 memory blocks and a kernel size of 31. We adopt a one-layer

TABLE I
THE PERFORMANCE COMPARISON BETWEEN MC-SA-ASR AND SC-SA-ASR APPROACHES IN TERMS OF SD-CER (%).

Approach	MC Fusion	Eval	Test	Average
SC-FD-SOT [10]	CDDMA bf	41.0	41.2	41.2
SC-WD-SOT [10]	CDDMA bf	36.0	37.1	36.8
SC-TS-ASR [10]	CDDMA bf	32.5	35.1	34.4
MC-FD-SOT	MFCCA+CLCCA	32.8	33.7	33.5
MC-WD-SOT	MFCCA+FLCCA	30.7	30.7	30.7
MC-TS-ASR	Neural bf	**30.4**	**27.5**	**28.3**

TABLE II
THE PERFORMANCE OF THE PROPOSED MC-WD-SOT METHOD WITH DIFFERENT MULTICHANNEL FUSION STRATEGIES IN SD-CER (%).

Approach	MC Fusion	Eval	Test	Average
SC-WD-SOT	CDDMA bf	36.0	37.1	36.8
MC-WD-SOT	MFCCA+SC-WD	30.7	33.1	32.4
	MFCCA+FLCCA	30.7	30.7	30.7
	+ Oracle $\langle$sc$\rangle$	**30.1**	**30.5**	**30.4**

TABLE III
THE PERFORMANCE (IN SD-CER (%)) OF MC-TS-ASR IN TERMS OF THE DIFFERENT MINIMUM LENGTHS OF DIARIZATION UTTERANCES.

Set	Oracle	Speaker diarization				
		0s	0.3s	0.6s	0.9s	1.2s
Eval	33.2	30.4	30.1	28.6	**28.2**	28.5
Test	28.9	27.5	27.5	**27.3**	27.6	28.5

TABLE IV
THE PERFORMANCE COMPARISON OF SPEECH SEPARATION IN SI-SNR (DB) AND ASR IN SD-CER (%) IN DIFFERENT STAGES FOR THE PROPOSED MC-TS-ASR METHOD.

Stage	Strategy	SD-CER		SI-SNR (simu)	
		Eval	Test	Eval	Test
Fine-tune (simu)	Joint	73.2	73.1	17.24	16.95
Fine-tune (real)	ASR-only	46.2	43.8	17.24	16.95
	Joint	**30.4**	**27.5**	-1.62	-1.99

of FLCCA with 8 attention heads and 560 hidden units in prior to the speech encoder. FLCCA is followed by three-layer 2D-convolutions, where the number of kernels is $\{16, 8, 1\}$ and the kernel size is (5, 7). For MC-TS-ASR, to reduce the computational burden, a squeezed version of TCNs (S-TCNs) is used, where three S-TCNs are stacked and each consists of 6 squeezed temporal convolutional modules with a kernel size of 5 and dilation rate of $\{1, 2, 4, 8, 16, 32\}$. The feature dimension in S-TCNs is 256, and 2 uni-directional LSTM layers with 64 hidden nodes are used as the BM. We employ a Res2Net-based d-vector extraction network trained on the VoxCeleb corpus [33], [34] as the speaker embedding model. For a fair comparison, the conformer with 8 layer encoder is used as the back-end ASR model, so that the amount of model parameters is 44M as close as that of MFCCA-based MC-SOT (45M).

C. Training details

The 80-dimensional log Mel-filter bank feature (Fbank) is calculated as the input feature using the ESPnet toolkit [35]. We use 4950 commonly-used Chinese characters as the modeling units. The MFCCA-based MC-SOT model is trained using *Train-Ali-far* and *Train-Ali-near* (i.e., around 210 hours of speech data in total). For MC-TS-ASR model, we first pretrain the front-end model with *Train-Ali-simu* and the ASR model with *Train-Ali-near*, respectively, which are then finetuned jointly using *Train-Ali-simu*. To maintain the front-end performance, we still adopt signal-level separation loss for pre-training, which is linearly combined with the ASR loss, such that the total loss function can be given by

$$Loss = \lambda L^{\text{SEP}}(\hat{\boldsymbol{w}}, \boldsymbol{w}) + (1-\lambda) L^{\text{ASR}}(\hat{\boldsymbol{y}}, \boldsymbol{y}), \qquad (5)$$

where $\boldsymbol{w}$ and $\boldsymbol{y}$ are the true speech and transcription, and $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{y}}$ are corresponding hypothesizes, respectively. L^{SEP} and L^{ASR} are loss functions for separation and ASR, respectively. The negative SI-SNR is used as the separation loss, while joint attention-CTC loss is used for ASR, and λ is set to be 0.5. Finally, we use *Train-Ali-far* and *Train-Ali-simu* to jointly finetune the model with only the ASR loss being taken into account. The total amount of data used in the experiment is consistent with that in [10] to ensure a fair comparison. All models are trained with the Adam optimizer [36]. When pre-training the front-end model, the learning rate is initialized to be 1.5e-4, which will be halved if the loss does not decrease for consecutive two epochs. Besides, we use a learning rate of 1e-3 with 10000 warmup steps to jointly fine-tune the model.

IV. EXPERIMENTAL RESULTS

First of all, we show the performance of the proposed MC-SA-ASR methods in Table I in comparison with the SC-SA-ASR counterparts, which use the *Ali-far-bf* data generated by the CDDMA beamformer (bf) [37] for training and inference. In order to utilize multichannel information, both MFCCA and CLCCA are used in SOT and frame-level diarization module for MC-FD-SOT, which achieves an average relative SD-CER reduction of 18.8% in Eval and Test sets compared to SC-FD-SOT (from 41.2% to 33.5%). For MC-WD-SOT, MFCCA and FLCCA are used in SOT and word-level diarization module, an average relative SD-CER reduction of 16.6% is achieved compared to SC-WD-SOT (from 36.8% to 30.7%). It is clear that the proposed MC-TS-ASR method, which calculates the filter weights of a neural beamformer and then performs filter-and-sum to obtain the fused spectrum, results in the best performance among all MC-SA-ASR systems and an average relative SD-CER reduction of 17.7% compared to SC-TS-ASR (from 34.4% to 28.3%).

Second, we show the performance of the application of different multichannel information fusions to the proposed MC-WD-SOT method in Table II. To improve the performance, similarly to [10] we use the combination of ASR hypotheses and real transcriptions for training. It is clear that combining the single-channel word-level diarization (SD-WD) and MC-SOT results, the performance can be largely improved compared to SC-WD-SOT. Using the FLCCA before speech encoder to fuse the multichannel information can further improve the performance, which converges to the case when an oracle separator ($\langle$sc$\rangle$) is additionally exploited.

Since short utterances might cause interference and lead to too many insertion errors, we further show the impact of the minimum length of diarization utterances on the MC-TS-ASR performance in Table III. Compared with the original oracle speaker labels, the ones obtained from speaker diarization (MC-TS-VAD) obtain better results in most cases, due to the fact that there are some interfering utterances that are not recognized by the diarization model. A further improvement can be achieved by removing the very short diarization utterances, especially on the Eval set (e.g., from 30.4% to 28.2%).

Finally, we compare the performance of speech separation and ASR in different stages is shown in Table IV for the proposed MC-TS-ASR method. Specifically, after pre-training we first fine-tune the whole model jointly on *Train-Ali-simu*, such that the front-end separation module can adapt to the back-end ASR module. We then fine-tune the model using real data composed of *Train-Ali-far* and *Train-Ali-simu*. In this step, we use ASR-only and joint training strategies for comparison. The difference between them is whether to freeze the front-end module during training. It can be seen that the relative SD-CER of joint training decreases by 34.2% (from 46.2% to 30.4%) and 37.2% (from 43.8% to 27.5%) on the Eval and Test sets, respectively, compared to the ASR-only training. However, joint training also leads to a decrease in the SI-SNR. In other words, joint fine-tuning enforces the front-end module to be adaptive with the ASR module at a sacrifice in the separation performance. This implies that given a speaker embedding, the end-to-end ASR model directly recognizes the speaker's utterances rather than the separated speech in multi-talker scenarios. This also shows that a better front-end signal quality (e.g., in SI-SNR) does not necessarily mean a better back-end ASR performance in practice.

V. Conclusions

In this paper, we proposed three multichannel SA-ASR approaches in the context of multi-party meetings. The MC-FD-SOT corresponds to SC-FD-SOT, which introduces MFCCA and CLCCA into ASR and frame-level diarization. While the MFCCA is also used in the ASR module of MC-WD-SOT, merged with FLCCA-based word-level diarization. The MC-TS-ASR uses neural beamforming to fuse the multichannel information. It was shown that the proposed multichannel methods outperform the corresponding single-channel counterpart. The joint fine-tuning of MC-TS-ASR enables the front-end module to match the ASR module, but the separation performance drops, implying that a direct recognition of the target speaker's speech might be a more appropriate choice, particularly in multi-talker scenarios. To our knowledge, this work is the first attempt to the multichannel SA-ASR problem in real meeting scenarios. In the future, we will explore more effective multichannel multi-talker modeling methods in meeting scenarios and attempt to further take speaker localization (or direction-of-arrival estimation) into account, such that the considered SA-ASR can identify "who spoke what from where".

Acknowledgment

This work was done while the first author was an intern at Speech Lab of DAMO Academy, Alibaba Group, and supported by the National Natural Science Foundation of China (62101523), Hefei Municipal Natural Science Foundation (2022012) and USTC Research Funds of the Double First-Class Initiative (YD2100002008). (*Correspondence*: jzhang6@ustc.edu.cn)

References

[1] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *MLMI*, ser. Lecture Notes in Computer Science, vol. 4299, Springer, 2006, pp. 309–322.

[2] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *CLEAR*, ser. Lecture Notes in Computer Science, vol. 4625, Springer, 2007, pp. 373–389.

[3] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *INTERSPEECH*, ISCA, 2018, pp. 1561–1565.

[4] N. Ryant, P. Singh, V. Krishnamohan, *et al.*, "The third DIHARD diarization challenge," in *INTERSPEECH*, ISCA, 2021, pp. 3570–3574.

[5] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, *et al.*, "The AMI meeting corpus," in *ICMT*, Citeseer, vol. 88, 2005, p. 100.

[6] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *INTERSPEECH*, ISCA, 2017, pp. 2456–2460.

[7] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 184–196, 2018.

[8] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *INTERSPEECH*, ISCA, 2020, pp. 2797–2801.

[9] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," in *ASRU*, IEEE, 2019, pp. 237–244.

[10] F. Yu, Z. Du, S. Zhang, Y. Lin, and L. Xie, "A comparative study on speaker-attributed automatic speech recognition in multi-party meetings," in *INTERSPEECH*, ISCA, 2022, pp. 560–564.

[11] F. Chang, M. Radfar, A. Mouchtaris, B. King, and S. Kunzmann, "End-to-end multi-channel transformer for speech recognition," in *ICASSP*, IEEE, 2021, pp. 5884–5888.

[12] F. Chang, M. Radfar, A. Mouchtaris, and M. Omologo, "Multi-channel transformer transducer for speech recognition," in *INTERSPEECH*, ISCA, pp. 296–300.

[13] D. Wang, Z. Chen, and T. Yoshioka, "Neural speech separation using spatially distributed microphones," in *INTERSPEECH*, ISCA, 2020, pp. 339–343.

[14] W. Wang, X. Qin, and M. Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *ICASSP*, IEEE, 2022, pp. 9171–9175.

[15] F. Yu, S. Zhang, P. Guo, *et al.*, "Mfcca: Multi-frame cross-channel attention for multi-speaker asr in multi-party meeting scenario," in *SLT*, IEEE, 2022.

[16] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, *et al.*, "M2Met: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP*, IEEE, 2022, pp. 6167–6171.

[17] F. Yu, S. Zhang, P. Guo, Y. Fu, Z. Du, *et al.*, "Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge," in *ICASSP*, IEEE, 2022, pp. 9156–9160.

[18] S. Ye, P. Wang, S. Chen, X. Hu, and X. Xu, "The royalflush system of speech recognition for m2met challenge," in *ICASSP*, IEEE, 2022, pp. 9181–9185.

[19] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *INTERSPEECH*, ISCA, 2020, pp. 274–278.

[20] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker," in *Interspeech*, ISCA, 2021, pp. 3555–3559.

[21] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *ASRU*, IEEE, 2019, pp. 260–267.

[22] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP*, IEEE, 2020, pp. 6394–6398.

[23] X. Ren, X. Zhang, L. Chen, X. Zheng, C. Zhang, *et al.*, "A causal u-net based neural beamforming network for real-time multi-channel speech enhancement," in *INTERSPEECH*, ISCA, 2021, pp. 1832–1836.

[24] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP*, IEEE, 2022, pp. 6487–6491.

[25] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[26] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *CVPR*, IEEE, 2017, pp. 1003–1012.

[27] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[28] E. Perez, F. Strub, H. d. Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, AAAI Press, 2018, pp. 3942–3951.

[29] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, *et al.*, "AISHELL-4: an open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *INTERSPEECH*, ISCA, 2021, pp. 3665–3669.

[30] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" In *ICASSP*, IEEE, 2019, pp. 626–630.

[31] A. Gulati, J. Qin, C. Chiu, N. Parmar, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, ISCA, 2020, pp. 5036–5040.

[32] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *CoRR*, vol. abs/1512.08301, 2015.

[33] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*, ISCA, 2017, pp. 2616–2620.

[34] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, ISCA, 2018, pp. 1086–1090.

[35] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, *et al.*, "Espnet: End-to-end speech processing toolkit," in *INTERSPEECH*, ISCA, 2018, pp. 2207–2211.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[37] W. Huang and J. Feng, "Differential beamforming for uniform circular array with directional microphones," in *INTERSPEECH*, ISCA, 2020, pp. 71–75.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Hybrid Syllable and Character Representations for Mandarin ASR

Fengrun Zhang, Chengfei Li*, Shuhao Deng, Yaoping Wang, Jinfeng Bai
TAL Education Group, Beijing, China
E-mail: {zhangfengrun1, lichengfei, dengshuhao1, wangyaoping, baijinfeng1}@tal.com

***Abstract*—With the development of deep learning, End-to-End (E2E) automatic speech recognition (ASR) based on Connectionist Temporal Classification (CTC) and attention has achieved great success and become the most popular method. In speech recognition, the selection of modeling units is critical. Most of the time, the modeling units of Mandarin are Chinese characters. However, the phenomenon of homophones and polyphonic characters in Chinese is very common, which degrades ASR performance. Pinyin can be regarded as the syllables of Chinese characters, which can reflect the pronunciation information of Chinese characters. In E2E ASR, due to the sequence-to-sequence form, Chinese characters directly correspond to the acoustic features and lack intermediate-level representations. In this paper, we introduce pinyin with tones as an auxiliary modeling unit to compensate for the mismatch between Chinese characters and acoustic features. On the basis of the hybrid modeling of syllables and Chinese characters, we propose a multi-task ASR model based on syllables and characters, which introduces a syllable CTC decoder and an attention decoder from syllables to Chinese characters to the joint CTC-attention model. Furthermore, a method of syllable auxiliary attention-rescoring method is proposed. Compared with the character-based ASR model, our method achieves a relative 8.6%/9.4% character error rate (CER) drop on Aishell-1 by greedy-search/attention-rescoring.**

I. Introduction

Automatic speech recognition (ASR) is the technology that converts speech into text. Traditionally, the entire ASR process is divided into multiple modules that are independent, including acoustic model, language model, and pronunciation lexicon, and then a decoder, such as weighted finite-state transducers (WFST) [1], is designed to search for the best path in search space according to the output of each module.

With the development of deep learning, End-to-End (E2E) ASR has become the mainstream speech recognition method. Unlike traditional speech recognition approaches that are generally divided into multiple modules, E2E ASR is trained using paired speech and text transcriptions, directly modeling the mapping of acoustic features to characters. In E2E ASR process, Connectionist Temporal Classification (CTC) can well solve the problem that acoustic features and text sequences have different lengths and cannot be aligned [2], and the attention mechanism of transformer can better capture context information [3]. At present, the model based on joint CTC and attention structure has become the mainstream E2E ASR framework [4].

*Chengfei Li is the corresponding author.

Due to the sequence-to-sequence form of E2E ASR, acoustic features are usually mapped directly onto characters without intermediate phoneme representations. However, this mapping may have some bias. For example, among all Chinese characters, a part derives from pictographs, and a part is ideographic in origin, including compound ideographs, but the vast majority originated as phono-semantic compounds [5]. This means that many Chinese characters are not directly related to acoustic features. With the frames' conditional independence assumption of CTC, in practical applications, the prediction of CTC often gains other characters with the same pronunciation as the target.

E2E ASR based on auxiliary modeling of phonemes or syllables has raised the interest of some researchers. R. Sanabria et al. [6] used a hierarchical multi-task learning method for hybrid modeling of phonemes and characters, and K. Krishna et al. [7] further studied the impact of resources, number of layers on hierarchical multi-task learning based on phoneme and character modeling. H. Futami et al. [8] modeled phonemes and characters, then used a phone-conditioned masked language model (PC-MLM) to correct characters with low confidence in the predicted sequence. X. Wan et al. [9] and Y. Yang et al. [10] only used syllables as modeling units, respectively established the mapping from acoustic features to syllables and from syllables to characters, and then cascades the two as the acoustic model. J. Yuan et al. [11] studied the improvement of pinyin with tones and pinyin without tones on Mandarin ASR and achieved high syllable accuracy using wav2vec 2.0 [12]. The result shows that pinyin with tones is better for Mandarin ASR.

Aiming at the problems existing in character-based Mandarin ASR, this paper uses characters as the main modeling unit, pinyin with tones as an auxiliary syllable modeling unit. The contributions of this paper are summarized as follows:

- This paper proposes an acoustic model for simultaneously modeling syllables and characters with a special transformer decoder that uses syllables to predict characters in order to explicitly construct the relationship between syllables and characters.
- This paper proposes a rescoring method using a hybrid attention-rescoring of characters and syllables. Compared with the current best attention-rescoring method using characters, the introduction of syllable information helps to improve the accuracy of the inference process.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. Related Work

A. CTC-based ASR

CTC assumes conditional independence between different frames and tokens [2]. Given the acoustic features $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$ of a speech, and the corresponding label $\boldsymbol{Y} = (y_1, \ldots, y_i, \ldots, y_L)$, the number of time steps T and the sequence length L are usually different. In ASR, T is always larger than L.

The main idea of CTC is that the posterior probability of **Y** given **X** can be directly computed by multiplying the probabilities for each frame with each other. Due to the silent segment in the speech, and to facilitate the calculation of the loss function, a special label $< blank >$ is added to the dictionary, and $< blank >$ can be predicted as the output of the model.

Let $\boldsymbol{H} = \{h_1, \ldots, h_l, \ldots, h_{T'}\}$ be the output of the linear layer after encoder, to map **H** to a shorter sequence **Y**, $< blank >$ tokens in H are removed and adjacent repeated tokens are merged, which is defined as the mapping rule α. The loss function of CTC is defined as the equation 1:

$$\begin{aligned} \mathcal{L}_{CTC} &= -\log P(Y \mid X) = -\log \sum_{H \in \alpha^{-1}(Y)} P(H \mid X) \\ &= -\log \sum_{H \in \alpha^{-1}(Y)} \prod_{t=1}^{T'} P\left(h_t \mid X\right) \end{aligned} \tag{1}$$

Where α^{-1} is the inverse of the mapping rule α, representing the set of **H** of length T' that satisfy the mapping rule α.

B. Joint CTC-Attention based ASR

In recent years, transformer has been the most famous and widely used model among the attention-based ASR models[13], [14]. Subsequently, many excellent acoustic models have been proposed, such as Conformer [15], Efficient Conformer [16], Squeezeformer [17], but the whole follows the transformer's encoder-decoder architecture.

In the training process of joint CTC-Attention model, the CTC loss and Attention loss are combined in equation 2.

$$\mathcal{L}_{\text{combined}}(\mathbf{X}, \mathbf{Y}) = \lambda \mathcal{L}_{\text{CTC}}(\mathbf{X}, \mathbf{Y}) + (1-\lambda)\mathcal{L}_{\text{Attention}}(\mathbf{X}, \mathbf{Y}) \tag{2}$$

Where λ is a hyperparameter that balances the importance of CTC and Attention loss.

In the inference process, since both CTC and Attention decoder have the ability to map the encoder output to character sequences, many decoding methods have emerged [18], [19]. The current mainstream non-autoregressive methods are greedy search, prefix beam search and attention-rescoring. The first two only use the CTC decoder, while attention-rescoring combines the CTC and attention decoder so that gets better performance obtained among non-autoregressive methods.

The attention-rescoring method is a two-stage approach. The first stage obtains a set of complete hypotheses with CTC scores, then passes these hypotheses through the attention decoder to get the attention scores. The second stage rescores the complete hypotheses using the CTC and attention scores. Let $\hat{\Omega}$ be a set of hypotheses of **X** by CTC prefix beam search, and $y \in \hat{\Omega}$, the rescoring pass obtains the final result according to equation 3,

$$\text{Prediction} = \arg\max_{y \in \hat{\Omega}} \left\{ \lambda \alpha_{\text{ctc}}(y, X) + (1-\lambda)\alpha_{\text{att}}(y, X) \right\}, \tag{3}$$

where

$$\begin{cases} \alpha_{\text{ctc}}(y, X) = \log p_{\text{ctc}}(y \mid X) \\ \alpha_{\text{att}}(y, X) = \log p_{\text{att}}(y \mid X) \end{cases} \tag{4}$$

The inference speed of attention-rescoring is faster than the autoregressive method due to the non-autoregressive nature, while it has better performance than other non-autoregressive methods.

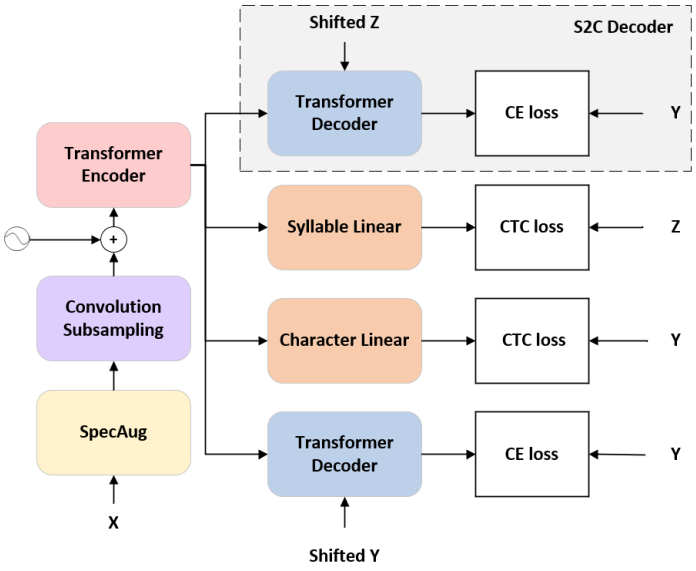


Fig. 1. Model architecture. X. Y and Z represent acoustic features, character sequences, and syllable sequences, respectively. In particular, the transformer decoder that inputs the shifted Z and outputs the probability of Y is called Syllable-to-Character (s2c) Decoder.

III. Proposed Method

In this section, two proposed methods based on syllables and characters will be introduced. One is an acoustic model with multi-task form loss functions that takes character accuracy as the main optimization goal. The other is an inference method that uses syllable information to improve character recognition performance.

A. Acoustic Model

Conventionally, the modeling units of the Hidden Markov Model (HMM) based ASR system are usually pronunciation representations. Common pronunciation representations include phonemes, initial and final consonants of syllables, syllables, etc. Introducing syllable modeling units into existing E2E acoustic models to compensate for the pronunciation information of characters is regarded as a central issue. Figure 1 shows the structure of the acoustic model.

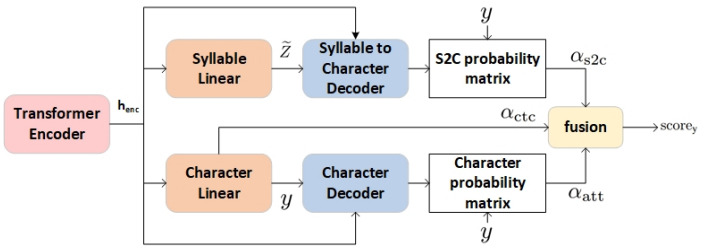


Fig. 2. The pipeline of the inference process. $\widetilde{Z}$ is obtained through greedy search, and $\hat{\Omega}$ is obtained through prefix beam search, while $y \in \hat{\Omega}$ is one of the hypotheses.

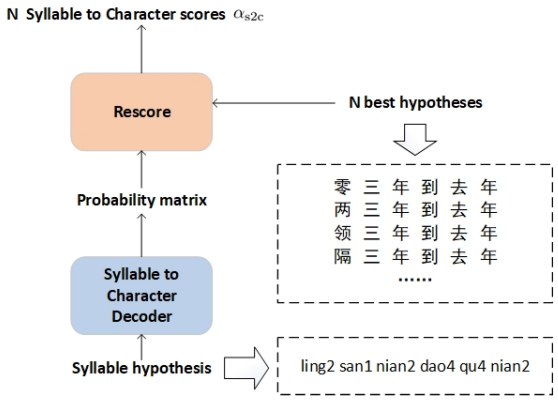


Fig. 3. The pipeline of the syllable auxiliary attention rescoring. The best syllable sequence is obtained by greedy search, and the N best character sequences are obtained by prefix beam search. The rescore process is to calculate the negative logarithm of the probability of the current character hypotheses in the probability matrix as α_{s2c}.

A hybrid modeling method of syllables and characters is adopted. For each training sample, in addition to the paired acoustic features $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$ and character sequence $\boldsymbol{Y} = (y_1, \ldots, y_i, \ldots, y_L)$, the syllable sequence $\boldsymbol{Z} = \{z_1, \ldots, z_i, \ldots, z_L\}$ is obtained corresponding to $\mathbf{Y}$. In Chinese, $\mathbf{Y}$ and $\mathbf{Z}$ have the same length L.

After the acoustic feature is fed into the encoder and obtains h_{enc}, the character CTC decoder maps h_{enc} to $\hat{Y}_{ctc}$, then calculates the CTC loss L_{char}^{ctc} with ground truth $\mathbf{Y}$. The character transformer decoder obtains $\hat{Y}_{att}$ by inputting shifted $\mathbf{Y}$, then calculates the cross entropy (CE) loss L_{char}^{att} with ground truth $\mathbf{Y}$. The part of the model mentioned above is the same as the current transformer-based acoustic model.

The success of unsupervised pre-training models such as wav2vec 2.0 [12], HuBERT [20], and WavLM [21] use self-supervised learning to learn speech representations, and have achieved performance close to supervised learning, which proves that pronunciation features are easy to distinguish and learn. S. Zhang et al. [22] used hybrid character and syllable modeling in the Listen, Attend and Spell (LAS) [13] to reduce substitution errors, especially those caused by homophones in Mandarin. Inspired by their work, we aim at the auxiliary modeling of syllables, adding a parallel syllable CTC decoder on the decoder side, which maps h_{enc} to syllable sequences $\hat{Z}_{ctc}$ and calculates L_{syl}^{ctc} with ground truth $\mathbf{Z}$. The above three decoders have a parallel structure. The design of multi-task loss functions enables each decoder to optimize its parameters, and jointly optimize the parameters of the encoder so that the encoder can learn representations that are applicable to both characters and syllables. Since syllables and characters have an explicit correspondence and the above decoders only establish the mapping from the acoustic features to the characters/syllables, we propose a syllable-to-character (s2c) decoder to further strengthen the mapping between syllables and characters. The process of syllables to characters is shown in equation 5, 6, 7.

$$\mathrm{h_{dec}} = \mathrm{Transformer}\left(\mathrm{h_{enc}}, \mathrm{Embedding}(\widetilde{\mathrm{Z}})\right) \tag{5}$$

$$\mathrm{Prob_{char}} = \mathrm{Softmax}\left(\mathrm{Linear}\left(\mathrm{h_{dec}}\right)\right) \tag{6}$$

$$\mathcal{L}_{s2c}^{att} = -\log P_{\mathrm{att}}\left(\mathrm{Y} \mid \mathrm{Prob_{char}}\right) \tag{7}$$

In the above equations, $\widetilde{Z}$ is obtained by shifting $\mathbf{Z}$ and adding an attention mask. Linear layer projects the feature dimension of $\mathrm{h_{dec}}$ to the dictionary size of character. Finally calculate the CE loss L_{s2c}^{att} of the probability matrix $\mathrm{Prob_{char}}$ and ground truth $\mathbf{Y}$.

The total loss function of the model can be described as equation 8, and each sub-loss function corresponds to a decoder.

$$\mathcal{L}_{\mathrm{combined}} = \lambda_1 \mathcal{L}_{char}^{ctc} + \lambda_2 \mathcal{L}_{char}^{att} + \lambda_3 \mathcal{L}_{syl}^{ctc} + \lambda_4 \mathcal{L}_{s2c}^{att} \tag{8}$$

Compared to conventional character-based methods, the additional $\mathcal{L}_{syl}^{ctc}$ and $\mathcal{L}_{s2c}^{att}$ are used for syllable modeling.

B. Syllable Auxiliary attention-rescoring

As is described in equation 3 and 4, conventional character-based attention-rescoring method uses the fusion of CTC score α_{ctc} and attention score α_{att} to rescore N hypotheses $y \in \hat{\Omega}_N$ from prefix beam search. CTC prefix beam search to get N hypotheses $\hat{\Omega}_N$ of $\mathbf{X}$, where N denotes the beam width. Then the attention decoder calculates the probability of $\hat{Y}_i \in \hat{\Omega}_N$ to get the attention score α_{att}, and weights with the CTC score α_{ctc} to regain the score of each hypothesis, and the hypothesis with the highest score is the prediction.

Since syllables only contain phonetic information, unlike the situation where multiple characters may correspond to the same pronunciation, a syllable can only correspond to one pronunciation and there is no syllable with the same pronunciation. In this paper, it can be seen from the experimental results in [11] that the mapping of acoustic features to syllables is simpler and easier to learn than characters, which means that on the same experimental configuration and dataset, syllables have a high probability of achieving higher accuracy than characters.

We assume that the prediction $\widetilde{Z}$ obtained from the syllable CTC decoder using greedy search is close enough to ground truth, the shifted $\widetilde{Z}$ is used as the input of the s2c decoder to get the score α_{s2c} of y, then combine it with the CTC score α_{ctc} and the score α_{att} of character attention above as the final

score of y. Figure 2 depicts this process. The final prediction is obtained by equation 9.

$$\text{Prediction} = \arg\max_{\bar{Y}_i \in \hat{\Omega}} \{\lambda_1 \alpha_{\text{ctc}} + \lambda_2 \alpha_{\text{att}} + \lambda_3 \alpha_{\text{s2c}}\} \quad (9)$$

where

$$\begin{cases} \alpha_{\text{ctc}}(y, X) = \log p_{\text{ctc}}(y \mid X) \\ \alpha_{\text{att}}(y, X) = \log p_{\text{att}}(y \mid X) \\ \alpha_{\text{s2c}}(y, X) = \log p_{\text{s2c}}(y \mid X, \widetilde{Z}) \end{cases} \quad (10)$$

For all $y \in \hat{\Omega}$, the $\widetilde{Z}$ of the s2c decoder input is the same, and α_{s2c} represents the possibility of character $\bar{Y}_i$ mapping to $\widetilde{Z}$.

In the real-world inference process, there are usually insertion and deletion errors, which means that the inferred result and the ground truth may have different lengths. We consider that insertion and deletion errors account for a very small fraction of substitution errors. In the process of syllable auxiliary rescoring, among the N best hypotheses of characters, the hypotheses whose length is different from the syllable hypothesis are discarded. The following experiments also proved that this selection is effective and does not increase the number of insertion and deletion errors.

IV. Experiments

A. Dataset

In this paper, we evaluate the proposed syllable auxiliary modeling method on the Aishell-1 [23], a commonly used dataset for evaluating Mandarin ASR. The Aishell-1 dataset contains recordings and corresponding character transcriptions. The syllable of the corresponding character is obtained through the open-source toolkit pypinyin[1]. Pypinyin has two forms when converting characters into syllables, one is pinyin with tones, and the other is pinyin without tones. Fig 4 shows a sample from the dataset. It has been proved in [11] that although the use of pinyin without tones can achieve higher accuracy of syllables, the improvement of the accuracy of characters is not as good as that of pinyin with tones. Therefore pinyin with tones is used as the modeling unit of syllables. Since multiple characters can correspond to the same syllable called homophones, the character dictionary consists of 4233 tokens and the syllable dictionary consists of 1183 tokens, including $< blank >$ and $< sos/eos >$ tokens.

B. Experimental Setup

In all experiments, the acoustic feature is 80-dimensional log Mel-filterbank (FBank) with window size of 25ms and window shift of 10ms. For data augmentation, 0.9, 1, 1.1 times speed perturbation and SpecAugment [24] are applied to the input. The acoustic model consists of a transformer encoder with L = 12, H = 512, A = 8, two parallel transformer decoders with each L = 6, H = 512, A = 8, and two parallel CTC linear layers, where L, H and A denotes the number of layers, hidden units, and attention heads respectively. Adam optimizer is used with Noam learning rate scheduling of learning rate at 0.002, warmup steps at 25, 000, gradient clipping at 5, and gradient accumulation at 1 [14]. During training, by experience, we set $\lambda_1 = 0.21, \lambda_2 = 0.4, \lambda_3 = 0.09, \lambda_4 = 0.3$ in equation 8. During inferring, we set $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.5$ in equation 9. When using prefix beam search, the beam width is set to 10.

[1]https://github.com/mozillazg/python-pinyin

```
Characters          : 在 长 达 两 年 的 时 间 里
Pinyin without tones: zai chang da liang nian de shi jian li
Pinyin with tones   : zai4 chang2 da2 liang3 nian2 de5 shi2 jian1 li3
```

Fig. 4. The character and syllable sequence of a sample in Aishell-1. Compared with pinyin without tones, the last number of pinyin with tones expresses tone information, and no number indicates a special soft tone.

V. Results and Analysis

A. Main Results

Table I shows the comparison of the proposed method with other approaches based on syllable and character modeling. All the methods only used the data of Aishell-1, without introducing external data. Different approaches are compared in terms of modeling units and Character Error Rate (CER). The character-only approach is to directly use an E2E method to establish the mapping of acoustic features to character sequences. A syllable-only approach is roughly a two-stage form. First, the syllable sequences are obtained through the acoustic features, and then the internal or external language model is used to map the syllable sequences to the character sequences. The approach of hybrid character and syllable modeling obtains the mapping of acoustic features to characters and syllables respectively, and then uses rescoring or error correction methods to obtain better character sequences through syllable sequences.

TABLE I
CER of different methods on Aishell-1 test set

Methods	Modeling units	CER (%)
Transformer [25]	Character	5.3
RNN-T [9]	Character	6.1
Syllable-level Transformer [10]	Syllable	5.2
Cascade RNN-T [9]	Character + Syllable	5.7
Decoupling Pinyin-T [11]	Syllable	7.1
Decoupling Pinyin+T [11]	Syllable	5.3
Proposed method	Character + Syllable	**4.8**

As can be seen from Table I, the proposed method achieves a competitive CER of 4.8%. Our model structure is similar to [25] and [10] which forms a transformer encoder-decoder architecture, but achieves better results. We consider that mapping from acoustic features to both character sequences and syllable sequences is important. Only mapping to character sequences will lose pronunciation information. Since the transformer encoder has the ability to capture contextual information, the character CTC decoder can use this information to better predict, and only mapping to syllables will lose the contextual information of the characters.

B. Syllable Accuracy

Table II shows the syllable error rate (SER) of part of the methods that counted the SER of syllables in Table I.

TABLE II
SER OF DIFFERENT METHODS ON AISHELL-1 TEST SET

Methods	SER (%)
Syllable-level Transformer [10]	3.2
Decoupling Pinyin-T [11]	2.7
Decoupling Pinyin+T [11]	3.6
Proposed method	3.0

In Table II, only Decoupling Pinyin-T [11] uses pinyin without tones, and pinyin with tones is used in all other methods. Among the results for pinyin with tones, the SER of our proposed method is the lowest. In the process of multi-task learning, although the CTC decoder of the syllable has only a weight of 0.09, it can still achieve a high accuracy rate. Besides, it can be seen that compared with the modeling method using only syllables, the modeling of hybrid characters and syllables can improve the accuracy of both syllables and characters. This is due to the fact that the modes of multi-task learning of syllables and characters are mutually reinforcing. Moreover, it can also be seen that the accuracy of syllables is significantly higher than characters, which means that the inference process of using syllables to assist characters' recognition is reasonable.

C. Analysis of Character Error Rate

Table III shows the error type of Transformer [25] with attention-rescoring and our proposed method with syllable auxiliary attention-rescoring.

From Table III it can be seen that in Mandarin speech recognition, substitution error is the main error type. The proposed auxiliary syllable modeling method significantly reduces the substitution error. We consider that by introducing the modeling of syllables, a more accurate mapping relationship between Chinese characters and acoustic features is established while preserving the semantic information of Chinese characters.

Discording the character hypotheses whose lengths are different from the syllable hypothesis in the rescoring process does not increase the number of insertion and deletion errors, which confirms our previous assumption. Compared to characters, syllables also have fewer insertion and deletion errors, and it is feasible to align the length of character hypotheses with the length of the syllable hypothesis.

TABLE III
ERROR OF DIFFERENT METHODS ON AISHELL-1 TEST SET

Methods	Transformer	Proposed method	Relative drop (%)
Insert	106	101	4.7
Delete	139	128	7.9
Substitute	5260	4804	8.7
total	5505	5033	8.6

D. Ablation Study

Ablation studies are conducted to demonstrate the effectiveness of each proposed method as shown in Table IV. The acoustic model shown in Figure 1 and the rescoring method shown in Figure 2 are used as benchmarks (A1). Auxiliary rescoring for syllables is first removed, and only character rescoring is used for inference (A2). No rescoring method is then used, only CTC greedy search for character is used during inference (A3). Then the inference method is not changed, and the s2c decoder is removed. (A4). Compared with the conventional transformer-based acoustic model for character modeling [14], the model only has an extra syllable CTC decoder. Finally, the syllable CTC decoder is removed, and the model is exactly the same as the implementation in [25]. From the results, it can be seen that introducing hybrid modeling of syllables and characters gains a relative drop of 6.9% from 5.8% to 5.4% in CER even without introducing complex inference methods. After adding s2c decoder and using syllable auxiliary rescoring, compared with the CTC greedy search and attention-rescoring in [25], CER obtioned by proposed method is relatively dropped by 17.2%/9.4%.

TABLE IV
ABLATION STUDY OF ALL COMPONENTS PROPOSED

ID	model	CER (%)
(A1)	Proposed method	4.8
(A2)	-Syllable rescoring	4.9
(A3)	-Character rescoring	5.3
(A4)	-s2c decoder	5.4
(A5)	-Syllable modeling	5.8

VI. CONCLUSIONS

In this paper, we propose a novel hybrid syllable and character modeling approach to better model the mapping of acoustic features to characters. The paradigm of multi-task learning is applied to our model. In particular, we propose a transformer-based decoder from syllables to characters. On this basis, we propose a method of syllable auxiliary rescoring. By introducing these methods, our method outperforms other methods based on syllable modeling on the test set of Aishell-1. We will explore pronunciation representations suitable for other languages in our future work.

VII. ACKNOWLEDGEMENT

This work was supported by National Key RD Program of China, under Grant No.2020AAA0104500.

REFERENCES

[1] M. Mohri, F. C. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, pp. 69–88, 2002.

[2] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the 23rd international conference on Machine learning*, 2006.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[4] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, 2016.

[5] Wikipedia contributors, *Chinese character classification — Wikipedia, the free encyclopedia*, [Online; accessed 15-May-2023], 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Chinese_character_classification&oldid=1137995585.

[6] R. Sanabria and F. Metze, "Hierarchical multitask learning with ctc," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 485–490. DOI: 10.1109/SLT.2018.8639530.

[7] K. Krishna, S. Toshniwal, and K. Livescu, *Hierarchical multitask learning for ctc-based speech recognition*, 2019. arXiv: 1807.06234 [cs.CL].

[8] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Non-autoregressive Error Correction for CTC-based ASR with Phone-conditioned Masked LM," in *Proc. Interspeech 2022*, 2022, pp. 3889–3893. DOI: 10.21437/Interspeech.2022-10062.

[9] X. Wang, Z. Yao, X. Shi, and L. Xie, "Cascade rnn-transducer: Syllable based streaming on-device mandarin speech recognition with a syllable-to-character converter," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 15–21. DOI: 10.1109/SLT48900.2021.9383506.

[10] Y. Yang, B. Du, and Y. Li, *Multi-level modeling units for end-to-end mandarin speech recognition*, 2022. arXiv: 2205.11998 [cs.CL].

[11] J. Yuan, X. Cai, D. Gao, R. Zheng, L. Huang, and K. W. Church, "Decoupling recognition and transcription in mandarin asr," *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1019–1025, 2021.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964. DOI: 10.1109/ICASSP.2016.7472621.

[14] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, 2018.

[15] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040. DOI: 10.21437/Interspeech.2020-3015.

[16] M. Burchi and V. Vielzeuf, "Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition," *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 8–15, 2021.

[17] S. Kim, A. Gholami, A. Shaw, *et al.*, "Squeezeformer: An efficient transformer for automatic speech recognition," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 9361–9373.

[18] Y. Higuchi, H. Inaguma, S. Watanabe, T. Ogawa, and T. Kobayashi, "Improved mask-ctc for non-autoregressive end-to-end asr," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8363–8367. DOI: 10.1109/ICASSP39728.2021.9414198.

[19] R. Fan, W. Chu, P. Chang, and A. Alwan, "A ctc alignment-based non-autoregressive transformer for end-to-end automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1436–1448, 2023. DOI: 10.1109/TASLP.2023.3263789.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A.-r. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[21] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[22] S. Zhang, M. Lei, and Z. Yan, "Investigation of transformer based spelling correction model for ctc-based end-to-end mandarin speech recognition," in *Interspeech*, 2019.

[23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017.

[24] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.

[25] Z. Yao, D. Wu, X. Wang, *et al.*, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Interspeech*, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Enhancing Whisper Model for Pronunciation Assessment with Multi-Adapters

Jing Li*, Rui Li*, Shen Guo*, Aishan Wumaier*†
* School of Information Science and Engineering, Xinjiang University, Urumqi, China
† Xinjiang Key Laboratory of Multi-lingual Information Technology, Urumqi, China
E-mail: hasan1479@xju.edu.cn

***Abstract*—Automatic pronunciation assessment is an important part of computer-aided pronunciation training. Due to the scarcity of non-native pronunciation assessment datasets and the fact that traditional speech assessment usually uses Goodness of Pronunciation(GOP) features, this may not provide enough information for word or sentence-level assessment. This paper aims to improve the performance of automatic pronunciation assessment from two aspects. First, to alleviate the problem of insufficient training data for pronunciation assessment, we use the weakly supervised learning model Whisper to build a pronunciation assessment model. With the Whisper encoder, Pearson correlation coefficient(PCC) performance is significantly improved compared to traditional acoustic features. Second, we propose a multi-adapters method that uses a multi-task loss to fine-tune the adapter while simultaneously learning phone, word, and sentence assessment tasks to boost sentence-level assessment task performance. In addition, through the experimental comparison of different scale models in the Whisper, the experimental results on the open-source dataset speechocean762 show that our proposed method achieves the best results in the medium.en model.**

Index Terms: pronunciation assessment, weakly supervised learning, Whisper, multi-adapters

I. Introduction

Automatic pronunciation assessment technology is a very important technology for computer-aided pronunciation training systems (CAPT). Compared with traditional manual assessment technology, automatic pronunciation assessment technology can not only provide accurate, timely, and comprehensive pronunciation feedback and pronunciation scores to second language learners but also has more objectivity[1]. However, it is a low-resource task due to the scarcity of high-quality pronunciation assessment datasets annotated by experts. Therefore, the improvement of model performance requires full use of data and effective modeling.

Traditional automatic pronunciation assessment technology usually uses Goodness of Pronunciation(GOP) to measure the pronunciation quality of non-native speakers[2], [3]. Specifically, it is pointed out in [4] that the log-posterior probability score has the highest correlation with human expert assessment scores. Subsequently, a posterior probabilistic variant, also known as GOP, is proposed for phone-level pronunciation quality scoring [5]. Several recent studies are based on deep neural network pronunciation assessment methods[6], [7]. A novel GOPT model that can assess different aspects of different pronunciation granularities at the same time was proposed[8]. In subsequent studies, a multi-view, multi-granularity, and multi-aspect(3M) pronunciation assessment model was proposed[9]. Furthermore, considering phones, words, and sentences contain grammatical structures, a Hierarchical Pronunciation Assessment with Multi-aspect Attention(HiPAMA) model is proposed[10]. The accuracy of pronunciation quality assessment for second language learners depends partly on manually designed features.

Since end-to-end (E2E) pronunciation assessment or mispronunciation detection models can avoid the mismatch problem between feature extraction and modeling in traditional methods, E2E models have become an important research direction in the field of automatic pronunciation assessment[11]–[14]. In [15], scoring spontaneous speech in a second language is achieved by combining temporally aggregated features with time-series features. In [16], it is proposed to use acoustic cues and vocabulary for encoding to score spontaneous spoken speech. In [17], a Siamese network is proposed to extract phone instance-level distance metric features to score the proficiency of non-native speakers' spontaneous speech. However, these methods still may require the design of more advanced features by knowledgeable people in the professional field. More importantly, it will be limited by data scarcity.

In this paper, we try to improve the performance of pronunciation assessment models from two perspectives. First, we experimented with using a Whisper encoder as the backbone of the pronunciation assessment model and explored the impact of using a single adapter on different sizes of Whisper models on pronunciation assessment performance. Then, inspired by the recent weakly supervised learning method Whisper's use of multi-task fine-tuning to achieve SOTA performance in speech recognition[18], we use this technique to improve pronunciation assessment performance. According to the alignment information, the multi-adapters are used to fine-tune the representations of different pronunciation granularities obtained by different layers of the Whisper encoder. Experimental results show that the performance of multi-adapters fine-tuning for multi-tasking is better.

In general, the main contributions of this paper include the following aspects: 1) To the best of our knowledge, we are the first to employ a weakly supervised pre-trained Whisper model on the pronunciation assessment task and demonstrate its effectiveness by fine-tuning only a single adapter. 2) In

Code at https://github.com/LijingDK/pronunciation-accessment-whisper

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

order to make use of multi-aspect information, we propose to use different aspects of adapters in different layers, and use multi-task loss to fine-tune multi-adapters, and the performance has been further improved.

II. RELATED WORKS

A. Whisper for WSL Representation

Weakly supervised learning methods in Whisper have been widely used in multilingual and multitask speech recognition and translation. It is worth noting that its performance in these domains has reached the same level as humans. However, current pre-trained models based on weak supervision mainly focus on ASR performance improvement[19]. Pronunciation assessment as a downstream task has not received enough attention. Therefore, in this study, we use Whisper to generate weakly supervised learning representations for our pronunciation assessment model. Unlike the Wav2vec 2.0 self-supervised pre-training learning model [20], Whisper uses 680k hours of weakly supervised data for training, and uses the encoder and decoder of the Transformer[21] architecture as the backbone.

B. Adapter for Fine-tuning

Pre-trained models usually contain a large number of parameters and require a lot of computing resources, making the optimization of the model more difficult. To solve this problem, research in recent years has paid more attention to how to introduce a small number of task-specific parameter-sharing techniques to improve the efficiency of pre-trained models. Adapter[22] is a novel fine-tuning method, which is usually injected between pre-trained network layers, and then fine-tuned on the adaptive corpus to adapt to specific tasks. Crucially, the Adapter can effectively reduce the parameter amount and size of the model, making the model have stronger generalization ability.

III. PROPOSED METHOD

Since Whisper is a weakly supervised model for multi-language and multi-task training, the multi-task pronunciation assessment task mode needs to be fine-tuned on the basis of Whisper. Our proposed model architecture is shown in Figure 1(c).

A. Single-Adapter

Adapter fine-tuning has been widely used in various tasks, but it is better at fine-tuning in the pre-training field. In addition, a lightweight Adapter fine-tuning method can achieve the same function as the full fine-tuning[23]. The mechanism of adapters is defined as:

$$F_A(Z) = f_{\theta^S}(f_{\theta^U}(\sigma(f_{\theta^D}(Z))) + Z) \tag{1}$$

where Z is the input of the adapter, $\sigma(\cdot)$ is an activation function, and the three linear layers $f_{\theta^D}, f_{\theta^U}, f_{\theta^S}$ represent downsampling, upsampling ,and scoring respectively.

Specifically, given a speech signal, we define $X = (x_t \in \mathbb{R}^d \mid t = 1, ..., T)$ as the input fbank features. We use the Whisper encoder architecture as the backbone model, denoted F_w^h, which maps fbank features to the speech representation Z^h. Here, h represents the number of layers of the Whisper encoder. We add a sentence adapter on top of the Whisper encoder, and during the entire training period, freeze the Whisper encoder and only fine-tune the sentence adapter. The whole procedure is as follows:

$$\begin{aligned} Z^h &= F_w^h(X) \\ \bar{Z} &= AvgPool(Z^h) \\ Score_{sent} &= F_A^{sent}(\bar{Z}) \end{aligned} \tag{2}$$

$$Loss_{sent} = \sum_{i=1}^{n}(Label_{sent}^i - Score_{sent}^i)^2 \tag{3}$$

where Z^h is the output of the top layer of the Whisper encoder, and $\bar{Z}$ is the mean vector after the global average pooling. n is the total number of samples.

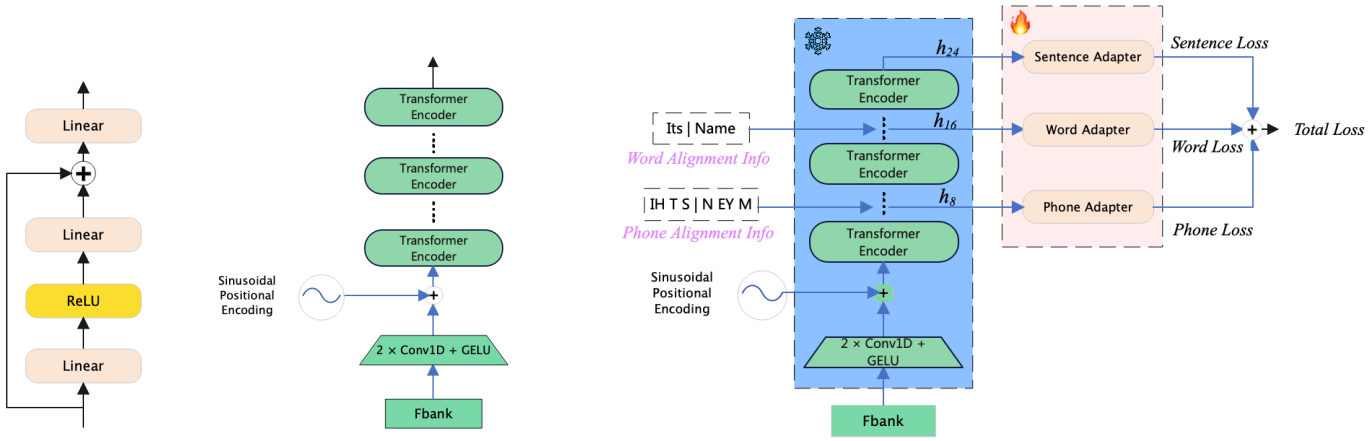
B. Multi-Adapters

Although weakly supervised learning can learn rich speech representations, how to apply general speech representations to downstream pronunciation assessment tasks is still a challenging problem. In the previous work, the research found that the phone information is encoded in the shallow layer of the pre-training model, and the content related information is encoded in the top layer[24], [25]. Inspired by this, we use the Whisper backbone to explore a multi-adapters architecture that is effective for pronunciation assessment tasks, as shown in the right architecture of Figure 1(c).

In the process of model training, the training samples of the pronunciation assessment data set are limited and the scoring tasks of phones, words and sentences are extremely related. As previous studies have shown, simultaneous training and sharing of network parameters across tasks can improve performance. Therefore, we take sentence assessment as the main task, phone and word assessment as auxiliary tasks, and then use multi-task learning (MTL) to combine the three. Specifically, the phone adapter, word adapter, and sentence adapter are respectively injected into the 8th, 16th, and 24th layers of the Whisper encoder. We use the kaldi based acoustic model[1] to obtain alignment information, and according to the alignment information, perform average pooling on the weakly supervised pre-training representations of the corresponding frames of each phone to obtain the acoustic representation of the phone, then input to the phone adapter to get the phone score. The entire process can be interpreted with the following expressions:

$$\begin{aligned} Z_t^8 &= F_w^8(X) \\ \bar{Z}_{phone_i} &= \frac{1}{m}\sum_{t=1}^{m} Z_{t \in phone_i}^8 \\ Score_{phone_i} &= F_A^{phone}(\bar{Z}_{phone_i}) \end{aligned} \tag{4}$$

[1]https://kaldi-asr.org/models/m13



(a) Adapter (b) Whisper encoder (c) Multi-Adapters

Fig. 1: Pronunciation assessment architecture based on multi-adapters Whisper model. Snowflakes represent frozen parameters, and flames represent trainable.

where m refers to the number of corresponding frames belonging to the same phone, which is obtained from the alignment information.

In the same way, the acoustic representation for a word is obtained by averaging the phone representations in the word. The typical part of the process is as follows:

$$\begin{aligned} Z_t^{16} &= F_w^{16}(X) \\ \bar{Z}_{word_i} &= \frac{1}{n}\sum_{t=1}^{n} Z_{t\in word_i}^{16} \\ Score_{word_i} &= F_A^{word}(\bar{Z}_{word_i}) \end{aligned} \tag{5}$$

where n refers to the corresponding phone belonging to the same word, which is obtained according to the alignment information and the pronunciation dictionary.

Finally, we fine-tune the network based on the MTL architecture, the loss is defined as shown in Eq. 6.

$$L_{total} = \alpha L_{phone} + \beta L_{word} + \gamma L_{sent} \tag{6}$$

where L_{phone}, L_{word} and L_{sent} are defined as Mean squared error (MSE) losses for phones, words, and sentences, respectively. α, β, and γ is a weight parameters to balance the loss between phone, word, and sentence scores.

IV. Experiments

A. Data Description

In this paper, we conduct experimental research on the speechocean762[2] dataset[26], an open-source pronunciation assessment corpus and also known as a benchmark for pronunciation assessment tasks. The dataset was scored by five experts on the phone with 0-2 points, and the pronunciation quality of words and sentences was scored from 0-10 points. In order to achieve the same scoring interval as the phone score, we scaled the word and sentence level scores to the 0-2 interval. Although the score range is narrowed, it does not affect the calculation of subsequent PCC values. The Train and Test sets are subject to the official release numbers, each containing 2,500 sentence pronunciations.

TABLE I: Architecture details of the Whisper model family.

Model	Layers	Width	Heads	Parameters
base.en	6	512	8	74M
small.en	12	768	12	244M
medium.en	24	1024	16	769M
large-v2	32	1280	20	1550M

B. Experiment Setup

All experiments in this paper are only trained on 2080TiGPU with 16GB of memory, batch size=20, epoch=25, learning rate=1e-5, $\alpha = \beta = \gamma$=1, because the labels in speechocean762 tend to have high scores, we use PCC to evaluate the model and expert scoring consistency.

C. Backbone Architecture

We use the ESPnet speech toolkit[27] to build a pronunciation assessment model based on the Whisper encoder. The input part of the model consists of two CNNs and an activation function GELU, where the stride of the second CNN is 2, and the 80-dimensional Fbank feature becomes 1024 dimensions after passing through the second CNN layer, and the number of frames will become the original half, the frame corresponding to each phone is halved. Then add the input module and the sinusoidal position code input to the transformer encoder module. Figure 1(b) shows the Whisper encoder architecture, Table I shows the architecture of the Whisper model family details.

D. Adapter Fine-tuning Architecture

For the fine-tuning stage of the pronunciation assessment model, the internal structure of the three Adapters used in the experiment is consistent. Each Adapter module consists of four parts, including three linear layers, an activation layer (ReLU), and a residual connection, as shown in Figure 1(a).

[2]https://www.openslr.org/101/

V. RESULTS

First, we show the performance comparison results using a single-adapter under different scale models of Whisper. Then, on this basis, the best scale model is selected and compared with different baselines. Finally, training with multiple losses further demonstrates the effectiveness of our proposed multi-adapters fine-tuning method.

From Table II we can find that larger scale models can achieve more significant results compared to using small scale models. However, large-v2 is 0.006 lower than the medium.en model. We guess that this may be because large-v2 is a multilingual training model and may not be good at fine-tuning pure English corpora. In the following experiments, we use the medium.en model to conduct multi-adapters experiments and conduct a comparative study with the baseline.

TABLE II: The PCC results for Whisper at different scales using a single sentence adapter

Model	Size	PCC
Whisper	base.en	0.66
	small.en	0.68
	medium.en	0.717
	large-v2	0.711

We used five baselines to compare with our proposed method. The two methods using GOP features as input are based on GOP[5] and GOPT model[8], one uses deep features as input[7], the other two baselines are a zero-short pronunciation assessment method proposed using the self-supervised learning(SSL) pre-training model[28] and an E2E multi-encoder method[11]. The experimental results are shown in Table III, we can observe that compared with the state-of-the-art GOPT baseline, the performance of single-adapter is about 0.023 lower than that of GOPT, but much higher than the value of about 0.077 of the most basic GOP method. Moreover, in the pre-training domain, it is about 0.117 higher than the latest benchmark model zero-shot. It is worth noting that the weakly supervised learning method achieves more significant performance improvements compared to the SSL method.

To further verify the performance of multi-adapters, we compared this method with the single-adapter and the baselines. It can be seen that multi-task training method is better than the result of single-task training, and the performance is significantly improved by the value of 0.003. This also demonstrates the effectiveness of our proposed method.

VI. DISCUSSION

Although our proposed weakly supervised pronunciation assessment method does not outperform the GOPT baseline, there may be several reasons for this result. First, the hyperparameters in our model may not have been chosen well, and a more complete hyperparameter search algorithm may need to be designed. Secondly, we did not use GOP features, were limited to unilateral research at the sentence level, and did not study the assessment of pronunciation quality at multiple granularities from multiple aspects like the GOPT method, which may have led to the low performance of our model, in future work, we will try to combine GOP features and weakly supervised pre-training representations to explore multi-aspected pronunciation assessment studies of phones, words, and sentences. Despite the limitations of the above factors, our proposed method still makes a certain contribution to the field. We use Whisper to learn a more general feature representation, which reduces the tendency of the model to overfit and helps the model to quickly adapt to various speech pronunciation assessment tasks. Moreover, the weakly supervised representation we use can capture not only phone information but also semantic information, so speech with high articulation quality can still achieve high scores even without artificially designing higher-level features.

TABLE III: Comparison of our method with different baselines on PCC.

Model	**PCC**
Regression based	
GoP[5]	0.64
DeepFeature[7]	0.72
GOPT[8]	0.74
End2End based	
Zero-shot[28]	0.60
Multi-encoder[11]	0.70
Proposed Methods	
Only Utterance Adapter	0.717
Phone + Word + Utterance Adapter	0.72

VII. CONCLUSIONS

In this paper, we use weakly supervised learning representations extracted from Whisper and fine-tune a single adapter to address the data insufficiency problem in the pronunciation assessment task. The use of the weakly supervised method achieves more significant performance improvements compared to E2E baselines, demonstrating the efficacy and importance of the weakly supervised method in pronunciation assessment tasks, and worthy of further research. In addition, we propose different aspects of using adapters at different layers of the Whisper encoder considering the correlation between different pronunciation granularities. We use multi-task loss to fine-tune the adapters, and the performance is further improved significantly. Experiments demonstrate the effectiveness of our proposed method on an open-source speech assessment dataset.

VIII. ACKNOWLEDGEMENTS

This work is supported by the unique fund project "Wireless broadcast channel content production application demonstration" guided by the central government and the key project "Research and development of key technologies for intelligent learning of national common language and characters" of the Education Department of Xinjiang Uygur Autonomous Region.

REFERENCES

[1] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: When technology meets pedagogy," *circulation journal official journal of the japanese circulation society*, 2002.

[2] H. Huang, H. Xu, Y. Hu, and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3165–3177, 2017.

[3] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for l2 pronunciation.," in *Interspeech*, 2020, pp. 3022–3026.

[4] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: A detailed performance study," *Speech Communication*, 2009.

[5] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[6] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities.," in *INTERSPEECH*, 2019, pp. 954–958.

[7] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment.," in *Interspeech*, 2021, pp. 4438–4442.

[8] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *Proc. ICASSP 2022*, IEEE, 2022, pp. 7262–7266.

[9] F.-A. Chao, T.-H. Lo, T.-I. Wu, Y.-T. Sung, and B. Chen, "3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment," in *Proc. APSIPA ASC 2022*, IEEE, 2022, pp. 575–582.

[10] H. Do, Y. Kim, and G. G. Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *Proc. ICASSP 2023*, IEEE, 2023, pp. 1–5.

[11] B. Lin and L. Wang, "Attention-based multi-encoder automatic pronunciation assessment," in *Proc. ICASSP 2021*, IEEE, 2021, pp. 7743–7747.

[12] Z. Zhang, Y. Wang, and J. Yang, "End-to-end mispronunciation detection with simulated error distance," *Proc. Interspeech 2022*, pp. 4327–4331, 2022.

[13] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *arXiv preprint arXiv:2204.03863*, 2022.

[14] T.-H. Lo, Y.-T. Sung, and B. Chen, "Improving end-to-end modeling for mispronunciation detection with effective augmentation mechanisms," in *Proc. APSIPA ASC 2021*, IEEE, 2021, pp. 1049–1055.

[15] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, *et al.*, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Proc. ASRU 2015*, IEEE, 2015, pp. 338–345.

[16] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *Proc. ICASSP 2018*, IEEE, 2018, pp. 6234–6238.

[17] K. Kyriakopoulos, K. Knill, and M. Gales, "A deep learning approach to assessing non-native pronunciation of english using phone distances," in *Interspeech 2018*, 2018.

[18] H. Yang, M. Zhang, S. Tao, M. Ma, and Y. Qin, "Chinese asr and ner improvement based on whisper fine-tuning," in *Proc. ICACT 2023*, IEEE, 2023, pp. 213–217.

[19] R. Tang, K. Kumar, G. Yang, *et al.*, "Speechnet: Weakly supervised, end-to-end speech recognition at industrial scale," *arXiv preprint arXiv:2211.11740*, 2022.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] S. A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," 2017.

[23] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Lightweight adapter tuning for multilingual speech translation," 2021.

[24] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[25] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU 2021*, IEEE, 2021, pp. 914–921.

[26] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[27] S. Watanabe, T. Hori, S. Karita, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[28] H. Liu, M. Shi, and Y. Wang, "Zero-shot automatic pronunciation assessment," *arXiv preprint arXiv:2305.19563*, 2023.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

DoA Estimation of Room Reflections Using NN-Based MUSIC Algorithm

Haowen Li*, Wen Zhang* Lijun Zhang*
* Center of Intelligent Acoustics and Immersive Communications,
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
E-mail: AndreaLee@mail.nwpu.edu.cn, wen.zhang@nwpu.edu.cn, zhanglj7385@nwpu.edu.cn

***Abstract*—The estimation of direction of arrival (DoA) for coherent sound sources, such as direct sound and early reflections in a room environment, is crucial. However, most traditional algorithms, like the multiple signal classification (MUSIC) algorithm, require observed signals to be non-coherent and can only be applied to narrowband signals. To tackle this issue, this paper introduces a neural network (NN)-based broadband MUSIC algorithm. By substituting the covariance matrix in the MUSIC algorithm with a pseudo-covariance matrix estimated through the network, this method can overcome the limitation of locating coherent broadband sources. Simulation results have shown that this proposed approach can accurately estimate the DoAs of direct sound and first-order reflections in both noisy and reverberant environments.**

I. Introduction

In a room sound field, the speech signals captured by a microphone array typically consist of direct sound, early reflections, and reverberation. For many tasks, such as in speech enhancement [1], immersive audio, and augmented reality (AR) [2], obtaining the DoA information for direct sound and early reflections is essential. However, accurately estimating the DoAs of direct sound and early reflections poses a challenge for traditional algorithms since they are coherent signals.

Traditional DoA estimation algorithms such as MUSIC and minimum variance distortionless response (MVDR) are unable to estimate the DoAs of coherent sound sources because of the rank-deficient covariance matrix [3],[4]. To address this limitation, additional processing is often required, such as spatial smoothing using the MUSIC algorithm. Spatial smoothing involves dividing the original array into multiple subarrays and reconstructing the received data covariance matrix using the covariance of the subarray matrices. This process partially restores the rank of the covariance matrix, thereby reducing coherence [5],[6]. However, this method has some drawbacks, including compromised accuracy and a limited ability to resolve coherent signals.

With the advancements in deep learning, numerous NN-based DoA estimation techniques have emerged for various scenarios and tasks. These methods often employ popular network architectures, including multi-layer perceptron (MLP) [7],[8],[9], convolutional neural network (CNN) [10],[11],[12], and convolutional recurrent neural network (CRNN) [13],[14]. However, these algorithms suffer from two key limitations. Firstly, due to the black-box nature of deep neural networks, the models lack interpretability, making it difficult to understand how the model works. Secondly, these algorithms heavily rely on large amounts of training data, which can be challenging and time-consuming to acquire.

As a solution to the limitations of NN-based DoA estimation methods, a recent trend has been to combine NNs with traditional algorithms. In these approaches, the NN replaces certain steps of the traditional algorithm. For instance, [15] proposed a method that uses the spatial pseudo-spectrum generated by the MUSIC algorithm as input to a CNN. The CNN directly outputs the posterior probability of the DoA. Similarly, [16] introduced a hybrid method that combines the traditional MUSIC algorithm with an NN-based algorithm to handle the processing of broadband coherent signals. This approach employs the coherent signal subspace method (CSSM) [17] and replaces the steps of computing the observed signal covariance matrix with a recurrent neural network (RNN). Nonetheless, accurately estimating the DoAs of coherent broadband sound sources remains an unresolved challenge.

Motivated by the successful combination of traditional algorithms and deep learning, this study aims to address the challenge of accurately estimating the DoAs of broadband coherent signals. Specifically, we focus on the direct sound and early reflections of speech signals in a reverberant room as typical examples of broadband coherent sound sources. Our proposed approach builds upon the framework of the broadband MUSIC algorithm using the independent subspace spatial matching (ISSM) technique [18],[19]. To achieve this, we first perform a short-time Fourier transform (STFT) on the observation signal. Next, we utilize the frequency bins of the STFT as input to the NN. The designed NN learns the pseudo-covariance matrix of both direct and reflected sources for all frequency bins, which can to some extent overcome the rank deficiency problem inherent in the original covariance matrix. By applying the ISSM-based MUSIC algorithm, we compute the MUSIC spatial spectrum for each frequency point and average them to obtain a final MUSIC spatial pseudo-spectrum. The peak of the final MUSIC spatial pseudo-spectrum corresponds to the estimated DoA. This enables us to estimate the DoAs of the direct sound and first-order reflections accurately. Compared to traditional algorithms, our proposed method demonstrates superior performance in precisely estimating the DoAs of direct sound and first-order reflections in a room sound field.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

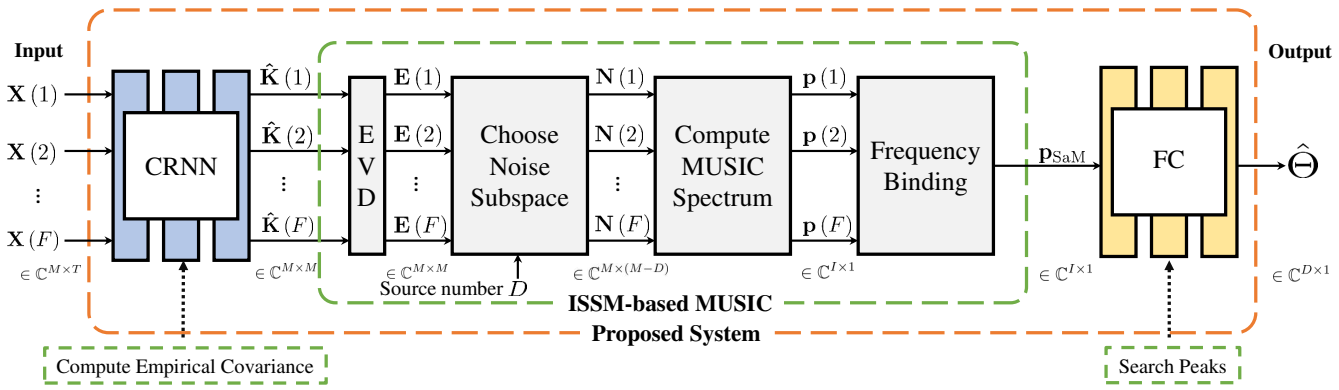


Fig. 1. Block diagram of the proposed NN-based DoA estimation algorithm, where the block in green dashed line displays the basic steps of the ISSM-based MUSIC algorithm, the block in orange dashed line displays the basic steps of the proposed approach, and the black dashed line indicates that these two steps are replaced by NNs.

The rest of the paper is structured as follows. Section II presents the signal model, along with an analysis of the challenges and underlying causes associated with the traditional MUSIC algorithm. Section III describes the overall framework and details of the proposed algorithm. Section IV validates the performance of the proposed algorithm through simulation experiments. Section V provides a discussion and summary of the entire paper.

II. Signal Model and Problem Formulation

Consider the scenario where there are D broadband signals captured by a microphone array with $M(M > D)$ elements, the observed signal of the mth sensor for $m = 1, 2, \ldots, M$ can be presented in the time-frequency (TF) domain through STFT transformation as

$$X_m(t,f) = \sum_{d=1}^{D} e^{-j2\pi f \tau_{md}} S_d(t,f) + V_m(t,f), \tag{1}$$

where $t = 1, 2, \ldots, T$ and $f = 1, 2, \ldots, F$ are the indexes of the time frames and frequency bins, respectively, T is the number of time frames and F is the number of frequency bins. S_d denotes the dth signal source, where d ranges from 1 to D, and D is the number of sources. The term $e^{-j2\pi f \tau_{md}}$ is the phase delay of the dth source at the mth array element relative to the reference element. This phase delay is used to construct the steering vectors, which describe the spatial characteristics of the array. V_m is the additive noise.

Equation (1) can be written in matrix form as

$$\mathbf{x}(t,f) = \mathbf{A}(f,\mathbf{\Theta})\mathbf{s}(t,f) + \mathbf{v}(t,f), \tag{2}$$

with

$$\mathbf{x}(t,f) = [X_1(t,f)\ X_2(t,f)\ \cdots\ X_M(t,f)]^T, \tag{3}$$

$$\mathbf{s}(t,f) = [S_1(t,f)\ S_2(t,f)\ \cdots\ S_D(t,f)]^T, \tag{4}$$

$$\mathbf{v}(t,f) = [V_1(t,f)\ V_2(t,f)\ \cdots\ V_M(t,f)]^T, \tag{5}$$

where $\mathbf{\Theta} = [\Theta_1\ \Theta_2\ \cdots\ \Theta_D]^T$ contains all DoA of D source signals, $\Theta_d = [\theta_d, \phi_d]$ contains the azimuth angle θ_d and the elevation angle ϕ_d of the dth signal. $\mathbf{A}(f,\mathbf{\Theta}) = [\mathbf{a}(\Theta_1)\ \mathbf{a}(\Theta_2)\cdots\ \mathbf{a}(\Theta_D)] \in \mathbb{C}^{M\times D}$, where $\mathbf{a}(\Theta_d) \in \mathbb{C}^{M\times 1}$ is the steering vectors of the dth signal, that is $[\mathbf{a}(\Theta_d)]_m = e^{-j2\pi f \tau_{md}}$. $\mathbf{x}(t,f) \in \mathbb{C}^{M\times 1}$, $\mathbf{s}(t,f) \in \mathbb{C}^{D\times 1}$, and $\mathbf{v}(t,f) \in \mathbb{C}^{M\times 1}$ are the signal vectors of the observed signal, source signal, and additive noise signal in the STFT domain, respectively.

The MUSIC algorithm is based on matrix eigenspace decomposition that relies on an accurate mathematical description of non-coherent narrowband sources. As the following processing is implemented for each frequency, the notation f is omitted for notional simplicity, so (2) becomes

$$\mathbf{x}(t) = \mathbf{A}(\mathbf{\Theta})\mathbf{s}(t) + \mathbf{v}(t). \tag{6}$$

The covariance matrix of $\mathbf{x}(t)$ can be calculated using $\mathbf{R}_x = \mathbb{E}\left[\mathbf{x}(t)\mathbf{x}^H(t)\right] \in \mathbb{C}^{M\times M}$. Here, $\mathbb{E}[\cdot]$ denotes the expectation operation using an empirical estimate $\hat{\mathbf{R}}_{\mathbf{x}} = \sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}^H(t)/T$, where T is the total number of samples. The symbol $(\cdot)^H$ represents the Hermitian transpose. We can obtain the eigenvectors and eigenvalues of $\mathbf{R}_x$ by performing the eigenvalue decomposition (EVD). The eigenvectors corresponding to the $M - D$ smallest eigenvalues form the noise subspace $\mathbf{N} \in \mathbb{C}^{M\times(M-D)}$. Then, the spatial spectrum of the MUSIC algorithm can be obtained as

$$\mathbf{p}_{\text{MUSIC}}(\mathbf{\Phi}) = \frac{1}{\mathbf{a}^H(\mathbf{\Phi})\mathbf{N}\mathbf{N}^H\mathbf{a}(\mathbf{\Phi})}, \tag{7}$$

where $\mathbf{a}(\mathbf{\Phi})$ is the steering vector for all possible looking directions $\mathbf{\Phi} = [\Phi_1\ \Phi_2\ \cdots\ \Phi_I]^T$ in space. The values corresponding to the peaks in the MUSIC spatial spectrum $\mathbf{p}_{\text{MUSIC}}(\mathbf{\Phi}) \in \mathbb{C}^{I\times 1}$ can be used to estimate DoAs of the sources.

Taking the microphone array receiving two incident signals $\mathbf{s}(t) = [S_1(t)\ S_2(t)]^T$ as an example, the covariance matrix can be obtained as

$$\begin{aligned}\mathbf{R}_x =& \mathbb{E}\left[\mathbf{x}(t)\mathbf{x}^H(t)\right] \\ =& \mathbf{A}(\mathbf{\Theta})\mathbf{R}_s\mathbf{A}^H(\mathbf{\Theta}) + \mathbb{E}\left[\mathbf{v}(t)\mathbf{v}^H(t)\right],\end{aligned} \tag{8}$$

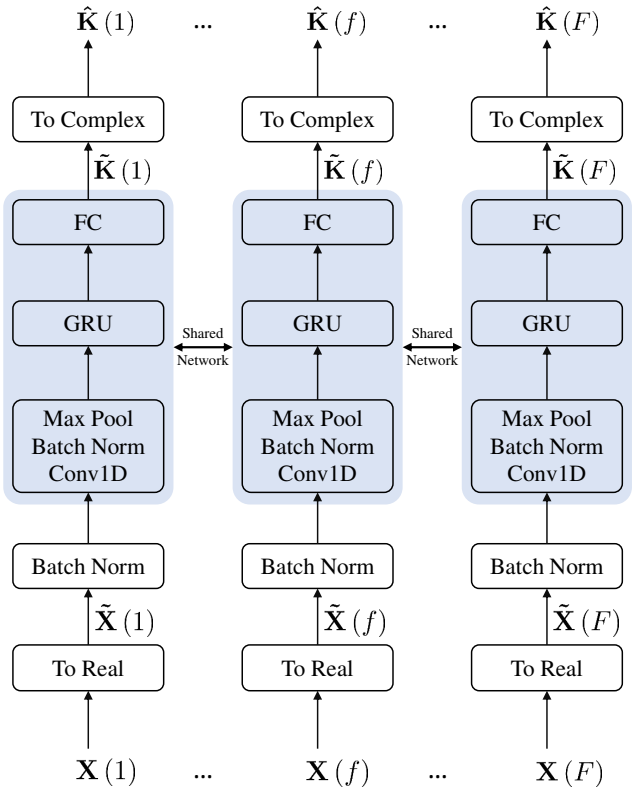


Fig. 2. Architecture of CRNN to obtain all pseudo-covariance matrix $\hat{\mathbf{K}} = \left[\hat{\mathbf{K}}(1)\ \hat{\mathbf{K}}(2)\ \cdots\ \hat{\mathbf{K}}(F)\right]$ corresponding to different frequency bins.

and the covariance matrix of the incoming signal is given by

$$\begin{aligned}\mathbf{R}_s =& \mathbb{E}\begin{bmatrix} S_1(t)S_1^H(t) & S_1(t)S_2^H(t) \\ S_2(t)S_1^H(t) & S_2(t)S_2^H(t) \end{bmatrix} \\ =& \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}^H\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},\end{aligned} \tag{9}$$

where σ_1^2 and σ_2^2 are the auto-correlation results of signals $S_1(t)$ and $S_2(t)$, respectively. When the two signals are coherent, meaning that they have a high degree of correlation with each other, $\rho_{12} = 1$, and $\mathbf{R}_s$ becomes rank deficiency. This leads to a situation where the eigen decomposition of $\mathbf{R}_s$ cannot accurately estimate the DoAs of coherent signals.

III. Proposed Approach

A. Architecture

To tackle the challenge of estimating the DoAs of broadband signals using the narrowband MUSIC algorithm, the ISSM-based deep MUSIC algorithm is proposed. As illustrated in Fig. 1, the proposed method is based on applying the narrowband MUSIC algorithm to each frequency bin to calculate the corresponding MUSIC spatial spectrum [18],[19]. Additionally, the spectra of all frequency bins are summed and averaged according to the equation

$$\mathbf{p}_{\text{MUSIC-ISSM}}(\mathbf{\Phi}) = \frac{1}{F}\sum_{f=1}^{F}\mathbf{p}_{\text{MUSIC}}(f, \mathbf{\Phi}), \tag{10}$$

where $\mathbf{p}_{\text{MUSIC}}(f, \mathbf{\Phi}) \in \mathbb{C}^{I\times 1}$ is the MUSIC spatial spectrum of the fth frequency bin, and F is the total number of frequency bins. The peaks of $\mathbf{p}_{\text{MUSIC-ISSM}}(\mathbf{\Phi})$ correspond to the estimated DoAs of the sources.

In Section II, it was highlighted that the limitation of the narrowband MUSIC algorithm in estimating the DoAs of coherent signals is attributed to the rank deficiency of the covariance matrix. This deficiency causes the energy of the incoming signal to disperse into the noise subspace, making it challenging to accurately compute the DoA. As a solution, this study proposes the use of a NN to replace the covariance matrix computation step.

The proposed algorithm's overall framework flow is illustrated in Fig. 1. It starts with the microphone array receiving broadband speech signals, which are then transformed into $\mathbf{X} \in \mathbb{C}^{M\times T\times F}$ using STFT. Each frequency bin $\mathbf{X}(f) \in \mathbb{C}^{M\times T}$ is then input into a CRNN composed of a convolutional module, gated recurrent unit (GRU), and uses a fully connected (FC) layer to perform linear transformation to obtain all pseudo-covariance matrix $\hat{\mathbf{K}} = \left[\hat{\mathbf{K}}(1)\ \hat{\mathbf{K}}(2)\ \cdots\ \hat{\mathbf{K}}(F)\right]$ corresponding to different frequency bins, where $\hat{\mathbf{K}}(f) \in \mathbb{C}^{M\times M}$. This network produces the pseudo-covariance matrix $\hat{\mathbf{K}}$ that replaces the conventional $\hat{\mathbf{R}}_{\mathbf{x}}$. Notably, the NN processes each frequency bin independently. Fig. 2 provides specific details of this NN architecture. It should be noted that before feeding into the NN, the real and imaginary parts of $\mathbf{X}(f)$ need to be concatenated to obtain the real-valued $\tilde{\mathbf{X}}(f) \in \mathbb{R}^{2M\times T}$. The NN outputs $\tilde{\mathbf{K}}(f) \in \mathbb{R}^{2M\times M}$, which should then be split into real and imaginary parts to obtain $\hat{\mathbf{K}}(f) \in \mathbb{C}^{M\times M}$.

Subsequently, the MUSIC spatial spectrum is computed for each frequency bin, resulting in $\mathbf{p}_{\text{all}} = [\mathbf{p}(1)\ \mathbf{p}(2)\ \cdots\ \mathbf{p}(F)] \in \mathbb{C}^{I\times F}$. These responses are then averaged using (10) to obtain $\mathbf{p}_{\text{SaM}} \in \mathbb{C}^{I\times 1}$. Because the gradient propagation of the NN affects the conventional search-peaks operation, it cannot be used to extract the estimated DoA values. To address this, FC layers are employed as a replacement, enabling the estimation of $\hat{\mathbf{\Theta}}$ for the direct sound and first-order reflection. Consequently, the entire DoA estimation algorithm is treated as a supervised regression task.

B. Training Loss

To address the issue of permutation inconsistency between the estimated and ground truth DoA orders, an improved Root Mean Square Period Error (RMSPE) based on permutation invariant training (PIT) [16] is integrated into the proposed system. This metric is defined as follows

$$\text{RMSPE}(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) = \min\left(\frac{1}{d}\left\|\text{G}(\mathbf{\Theta}, \hat{\mathbf{\Theta}})\right\|^2\right)^{\frac{1}{2}}, \tag{11}$$

$$\text{G}(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) = \min\left[|\mathbf{\Theta} - \text{P}(\hat{\mathbf{\Theta}})|, 2\pi - |\mathbf{\Theta} - \text{P}(\hat{\mathbf{\Theta}})|\right], \tag{12}$$

where $\mathbf{\Theta} = [\Theta_1\ \Theta_2\ \cdots\ \Theta_D]$ and $\hat{\mathbf{\Theta}} = [\hat{\Theta}_1\ \hat{\Theta}_2\ \cdots\ \hat{\Theta}_D]$ are the true and estimated DoAs of all D sources, respectively, and $\text{P}(\cdot)$ computes all possible permutations of the DoAs.

TABLE I
PARAMETER SETTINGS

Parameter	Value	Parameter	Value
Array elements M	9	Epoch	50
Room size	$(4-7, 3-5)$m	Learning rate init	10^{-3}
Wall absorption	$(0.2-0.8)$	Optimizer	*Adam*
Batch size	16	Train samples	3696
Test samples	1680	Validation samples	924

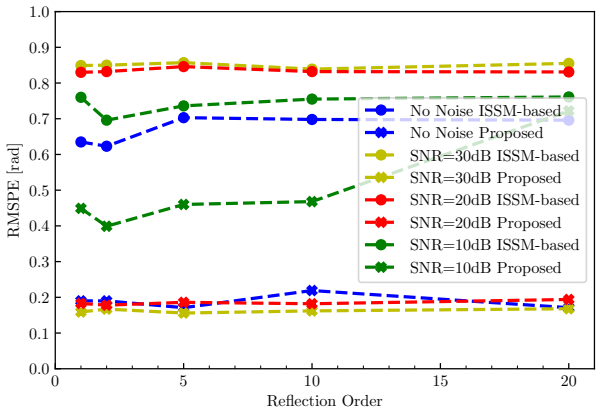


Fig. 3. DoA estimation error RMSPE [rad] under different reflection orders and SNRs.

TABLE II
DOA ESTIMATION ERROR RMSPE [RAD] USING THE PROPOSED METHOD FOR THE SITUATION WHERE THE ISSM-BASED MUSIC CANNOT OBTAIN FIVE ESTIMATED DOAS, I.E., THE CASE OF "UNDETECTED TEST SAMPLES".

(a) No Noise					
Reflection orders	1	2	5	10	20
Undetected test samples	21	81	161	158	204
Proposed	0.162	0.175	0.178	0.206	0.160
(b) SNR = 30 dB					
Reflection orders	1	2	5	10	20
Undetected test samples	120	121	88	91	110
Proposed	0.164	0.168	0.152	0.177	0.157
(c) SNR = 20 dB					
Reflection orders	1	2	5	10	20
Undetected test samples	46	38	31	31	29
Proposed	0.182	0.165	0.183	0.197	0.180
(d) SNR = 10 dB					
Reflection orders	1	2	5	10	20
Undetected test samples	1	150	161	1	0
Proposed	0.582	0.371	0.418	0.146	-

IV. EXPERIMENTS AND ANALYSES

A. Experimental Setup

To simulate the 2D room sound field, we generate the reverberated signals by convolving the audio signals with the room impulse responses (RIRs). We generate the RIRs using the *pyroomacoustics* library [20]. The speech signals used in the simulation are obtained from the TIMIT dataset [21]. We consider different reflection orders to achieve a more realistic representation of the sound field within the room.

The dataset we use is specifically designed to cover various room characteristics such as size, wall sound absorption coefficients, microphone position, and sound source position. To obtain all angles in a 2D room, a uniform circular array (UCA) with $M = 9$ and a radius $R = 0.05$m is used. The specific parameter settings can be found in Table I. To ensure diversity in our data, each speech sample in the TIMIT database is convolved with different RIR parameters. We generate RIR signals with random room sizes ranging from $4-7$m in width and $3-5$m in height. The absorption coefficients for each wall are also randomly set between $0.2-0.8$. Furthermore, the source position of each direct sound signal and the receiving position of the microphone array are randomly placed within the room, ensuring a separation of at least 1m for far-field signals.

For model training, validation, and testing, we obtain true DoA values using the image source method [22]. This method provides the DoA for both the direct sound and the first-order reflection for each sample in the dataset.

To evaluate the robustness of the model, we train and test it under various reflection orders and noise conditions. Specifically, we consider reflection orders of $1, 2, 5, 10$, and 20, representing different levels of sound reflections in the room. We also vary the signal-to-noise ratio (SNR) during training and testing. The SNR values we use are infinity, 30 dB, 20 dB, and 10 dB, which cover a range of noise levels from clean signals to increasingly noisy conditions.

To ensure a fair evaluation, we divide the generated dataset into training and testing sets based on the TIMIT dataset. One-tenth of the training set is further set aside as a validation set for monitoring the model's performance during training. The final dataset consists of 3696 samples in the training set, 924 samples in the validation set, and 1680 samples in the testing set. This division allows us to adequately train and evaluate the model's performance across different reflection orders and noise conditions.

During the training process, the training set is first transformed using the STFT before being passed through the model. We train the model for a total of 50 epochs using the *Adam* optimizer [23]. To optimize the learning process, we set an initial learning rate of 10^{-3}. This learning rate is adjusted throughout training based on the validation set's error rate. The learning rate decreases gradually, ensuring that the model converges effectively. We set a minimum learning rate of 10^{-4} to prevent the learning rate from decreasing too much.

For more detailed information on the model's parameters, please refer to Table I where all the parameter settings are listed. These settings ensure the model is trained effectively and allow for meaningful comparisons across different reflection orders and noise conditions.

TABLE III
COMPARISON OF THE ESTIMATED AND TRUE DOAS OF THE DIRECT SOUND AND FIRST-ORDER REFLECTION.

Wall	direct	left wall	right wall	up wall	under wall	RMSPE
(a) No Noise and Reflection Order = 20						
Ground truth	164.525°	169.250°	251.570°	5.826°	135.295°	-
ISSM-based MUSIC	164.000°	106.000°	252.000°	8.000°	140.000°	0.301 rad(17.3°)
Proposed-NN-MUSIC	158.094°	167.422°	235.426°	11.232°	121.914°	0.177 rad(10.1°)
(b) SNR = 30 dB and Reflection Order = 5						
ISSM-based MUSIC	164.000°	292.000°	252.000°	10.000°	52.000°	0.970 rad(55.6°)
Proposed-NN-MUSIC	163.830°	168.453°	236.283°	7.896°	127.621°	0.134 rad(7.7°)
(c) SNR = 20 dB and Reflection Order = 2						
ISSM-based MUSIC	294.000°	168.000°	218.000°	326.000°	78.000°	0.743 rad(42.6°)
Proposed-NN-MUSIC	167.903°	170.619°	237.380°	4.274°	125.797°	0.137 rad(7.9°)
(d) SNR = 10 dB and Reflection Order = 10						
ISSM-based MUSIC	202.000°	168.000°	254.000°	8.000°	144.000°	0.267 rad(15.3°)
Proposed-NN-MUSIC	173.826°	182.513°	240.449°	359.725°	119.676°	0.202 rad(11.6°)

To establish a comparison baseline for broadband coherent signal localization, we utilize the traditional ISSM-based MUSIC method. Since there is currently no existing baseline method specifically designed for this task, the ISSM-based MUSIC method serves as a suitable alternative.

To evaluate the performance of our proposed model, we use the RMSPE metric, as defined in equation (11). RMSPE provides a comprehensive measure of the model's accuracy in localizing broadband coherent signals. By comparing the RMSPE scores of our model with those of the ISSM-based MUSIC method, we can assess the effectiveness of our proposed approach.

B. Results and Discussion

During the testing of the traditional ISSM-based MUSIC algorithm, it is observed that not all samples are able to obtain the direct sound and first-order reflections in a 2D room, i.e., at least five peaks, resulting in the estimation of fewer than five $\hat{\theta}$ values. As a consequence, the RMSPE metric cannot be calculated for these un-detected samples. In order to ensure consistency in the test samples used by both methods, the entire test set is divided into two parts based on whether the traditional ISSM-based MUSIC algorithm can obtain five peaks (resulting in five estimated $\hat{\theta}$ values).

For the traditional ISSM-based MUSIC algorithm, we only calculate RMSPE on the samples where five peaks can be obtained. All tests are carried out under varying reflector orders and SNRs. In Fig. 3 and Table II, the abbreviation "Proposed" represents the Proposed-NN-MUSIC algorithm, while "ISSM-based" represents the traditional ISSM-based MUSIC algorithm.

Fig. 3 displays the RMSPE of both methods on the test samples where the ISSM-based MUSIC algorithm can extract five estimated DoAs. Table II provides the results for the situation where the ISSM-based MUSIC algorithm fails to obtain five estimated DoAs (referred to as "Undetected test samples") while the Proposed-NN-MUSIC algorithm still works effectively. The results as shown in Fig. 3 and Table II demonstrate that the Proposed-NN-MUSIC algorithm achieves significantly smaller RMSPE errors compared to the conventional ISSM-based MUSIC algorithm. This indicates that the proposed system effectively overcomes the limitations of the traditional method, successfully estimating all target DoA values in certain cases. Consequently, the Proposed-NN-MUSIC algorithm enhances the overall robustness and accuracy of the DoA estimation algorithm.

Furthermore, Fig. 3 clearly illustrates that the Proposed-NN-MUSIC algorithm consistently outperforms the ISSM-based MUSIC algorithm across all scenarios. As the SNR increases, the RMSPE tends to decrease and reach a stable level. Particularly under high SNR conditions, the RMSPE remains consistently low and exhibits minimal variation with respect to the reflector order, mostly below 0.2 rad. However, it is worth noting that under low SNR conditions, the RMSPE increases significantly, reaching up to 0.5 rad even when the reflector order is 1.

To gain a deeper understanding of the estimation by the two algorithms, Table III presents the true values, estimated values, and RMSPE for a set of direct sound DoA and its four first-order reflections for both algorithms. It is evident that the traditional ISSM-based MUSIC algorithm can accurately estimate three or four angles, but there is typically one angle with a large deviation or is unattainable, resulting in a high overall RMSPE for the algorithm. On the other hand, although the Proposed-NN-MUSIC algorithm may have slightly lower accuracy in estimating the DoA of the direct sound or one of its reflections in comparison to the traditional ISSM-based MUSIC algorithm, it avoids significant biases in estimating all angles. As a result, the Proposed-NN-MUSIC algorithm achieves a noteworthy reduction in its overall performance.

V. CONCLUSIONS

In this paper, we proposed a system for accurately estimating the DoA of coherent broadband sound sources in a room environment. Our system expands upon the traditional ISSM-based MUSIC algorithm by incorporating a neural network

that is specifically designed to learn the pseudo-covariance matrix for each coherent source. This novel design allows our Proposed-NN-MUSIC algorithm to overcome the limitations of the traditional ISSM-based MUSIC algorithm in accurately estimating the DoA of coherent broadband sound sources. By utilizing the neural network to process multiple frequency bins simultaneously, our algorithm achieves improved accuracy and can handle a wider range of sound sources. We validate the effectiveness of our Proposed-NN-MUSIC algorithm through experiments using speech signals as coherent broadband sound sources in various room environments. These experiments demonstrate that our algorithm achieves higher accuracy in DoA estimation compared to the traditional algorithm.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61831019 and 62271401. The author would like to thank Mr. Yichen Yang for the helpful discussions.

REFERENCES

[1] A. Xenaki, J. Bünsow Boldt, and M. Græsbøll Christensen, "Sound source localization and speech enhancement with sparse bayesian learning beamforming," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3912–3921, 2018.

[2] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2005.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[5] Q. Chen and R. Liu, "On the explanation of spatial smoothing in MUSIC algorithm for coherent sources," in *International Conference on Information Science and Technology*, pp. 699–702, IEEE, 2011.

[6] M. Wax and J. Sheinvald, "Direction finding of coherent signals via spatial smoothing for uniform circular arrays," *IEEE transactions on antennas and propagation*, vol. 42, no. 5, pp. 613–620, 1994.

[7] Y. Kim and H. Ling, "Direction of arrival estimation of humans with a small sensor array using an artificial neural network," *Progress In Electromagnetics Research B*, vol. 27, pp. 127–149, 2011.

[8] H. Tsuzuki, M. Kugler, S. Kuroyanagi, and A. Iwata, "An approach for sound source localization by complex-valued neural network," *IEICE TRANSACTIONS on Information and Systems*, vol. 96, no. 10, pp. 2257–2265, 2013.

[9] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2927–2932, IEEE, 2013.

[10] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, Audio Engineering Society, 2015.

[11] Y. Hao, A. Küçük, A. Ganguly, and I. M. Panahi, "Spectral flux-based convolutional neural network architecture for speech source localization and its real-time implementation," *IEEE Access*, vol. 8, pp. 197047–197058, 2020.

[12] G. Bologni, R. Heusdens, and J. Martinez, "Acoustic reflectors localization from stereo recordings using neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2021.

[13] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 241–245, IEEE, 2018.

[14] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," *arXiv preprint arXiv:2010.00140*, 2020.

[15] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, 2020.

[16] J. P. Merkofer, G. Revach, N. Shlezinger, T. Routtenberg, and R. J. van Sloun, "DA-MUSIC: Data-Driven DoA Estimation via Deep Augmented MUSIC Algorithm," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[17] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[18] G. Su and M. Morf, "The signal subspace approach for multiple wideband emitter location," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 6, pp. 1502–1522, 1983.

[19] S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for Sound Source Localization in Robotics," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007.

[20] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355, IEEE, 2018.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Hybrid Multi-Task Learning for End-To-End Multimodal Emotion Recognition

Junjie Chen*, Yongwei Li†, Ziping Zhao*, Xuefei Liu†, Zhengqi Wen†, Jianhua Tao‡,§
* College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
E-mail: 2110090009@stu.tjnu.edu.cn, ztianjin@126.com
† National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-mail: yongwei.li@nlpr.ia.ac.cn, liuxuefei@jixianyuan.com, zqwen@nlpr.ia.ac.cn
‡ Department of Automation, Tsinghua University, Beijing, China
§ Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
E-mail: jhtao@tsinghua.edu.cn

***Abstract*—Multimodal emotion recognition plays a pivotal role in the advancement of natural human-computer interaction systems. Recent studies have attempted to apply multi-task learning to emotion recognition. However, the multi-task shared feature extractor of traditional methods needs to integrate the feature representations of different tasks, which may lead to the feature extractor failing to focus on the learning of emotion representations. To address this problem, we propose a hybrid multi-task learning framework for end-to-end multimodal emotion recognition, in which the primary task is emotion classification, and the auxiliary tasks are emotion regression and gender classification. This framework consists of two networks specialized in gender and emotion recognition, where the latter transfers knowledge from the former through our proposed Deep Aggregation LSTM (DA-LSTM). The DA-LSTM could more precisely capture emotional information in discourse by aggregating emotion and gender feature extractors. Experimental results on a commonly used dataset IEMOCAP demonstrate the effectiveness of our proposed method.**

I. Introduction

Emotion recognition plays a pivotal role in the advancement of natural human-computer interaction systems, which could automatically detect emotions by analyzing information from multiple modalities such as speech, text, and facial expressions. Emotion recognition has diverse applications across various domains, including the prevention of traffic accidents by detecting driver fatigue and emotional distress [1], as well as aiding physicians in the diagnosis of mental health conditions such as depression and anxiety during psychological evaluations [2]. In human-computer interaction systems, speech and text are the most convenient and straightforward media for expressing emotions. Therefore, we focus on speech and text-based multimodal emotion recognition (MER) in this paper.

Traditional MER methods [3]–[5] commonly employ acoustic features such as MFCC and spectrograms, as well as word embeddings generated using pre-trained language models like Glove and ALBERT. Recent researches [6]–[8] have shown that extracting high-level features directly from raw utterances using pre-trained self-supervised models, particularly WavLM [9] and BERT [10], yields superior results compared to traditional MER approaches. Additionally, incorporating multi-head attention mechanisms to capture complementary information between modalities has also been shown to achieve remarkable performance in recent studies [11]–[13].

Furthermore, several studies have explored the incorporation of multi-task learning techniques to enhance emotion recognition performance. For instance, Mukherjee et al. [14] adopted a joint training approach to simultaneously perform emotion classification and emotion regression, to detect emotions expressed in tweets. Similarly, Li et al. [15] incorporated gender classification as an auxiliary task and shared the feature extractors with the emotion classification task, thereby facilitating the learning of emotion representations that capture gender-related characteristics. Despite the impressive achievements of these approaches, they suffer from a limitation. The multi-task shared feature extractor in traditional methods needs to balance and integrate the feature representations of different tasks, potentially resulting in the feature extractor neglecting the focused learning of emotion representation.

In this paper, we propose a hybrid multi-task learning framework for end-to-end multimodal emotion recognition, in which the primary task is emotion classification, and auxiliary tasks are emotion regression and gender classification. Considering that both emotion regression and emotion classification focus on emotional information, while gender classification focuses more on gender-related cues, we design the framework as a combination of two networks specialized in gender and emotion recognition, respectively. The first network is composed of a gender feature extractor and a gender classifier. The second network is composed of an emotion feature extractor, two cross-modal fusion modules, an emotion classifier, and three emotion regression modules that share representations with the emotion classifier. In addition, a Deep Aggregation LSTM (DA-LSTM) is designed to enable the emotion feature extractor to focus on the learning of emotion representation while transferring knowledge from the gender feature extractor.

We can summarize the main contributions of this paper as: 1) We propose a hybrid multi-task learning framework for end-to-end multimodal emotion recognition, which could comprehensively capture and analyze emotional information

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

in multimodal data; 2) We propose an LSTM-based module, referred to as Deep Aggregation LSTM (DA-LSTM), which enables the integration of both emotion and gender feature extractors for conditioning emotion recognition on gender; 3) Experimental results on the dataset IEMOCAP [16] demonstrate the effectiveness of our proposed method.

II. Method

In this section, we address the proposed hybrid multi-task learning framework, with emotion classification as the primary task while emotion regression and gender classification as auxiliary tasks. The proposed framework consists of feature extractors, cross-modal fusion modules, and output modules. These modules are introduced in order. After that, two different multi-task learning strategies with these modules are introduced, including our proposed hybrid multi-task learning and traditional multi-task learning.

A. Feature Extractors

In this subsection, we provide a detailed description of the gender feature extractor and the emotion feature extractor. Both extractors comprise a speech feature encoder and a text feature encoder, each serving a specific purpose in capturing gender-related and emotional cues, respectively.

1) Gender Feature Extractor: As shown in Fig. 1, the speech encoder consists of a pre-trained WavLM-base-plus model, two LSTM layers, and one LayerNorm layer. WavLM-base-plus is a variant of WavLM, which consists of a CNN encoder and 24 Transformer encoders. WavLM-base-plus takes the raw speech signal and outputs a frame-level embedding in the shape of N $\times$ 768 after a nonlinear transformation, where N and 768 denote the number of frames and the representation of one frame, respectively. Then, two LSTM layers with 256 cells process these embedding features to obtain high-level speech feature representations of shape N $\times$ 256. Finally, layer normalization is applied to these representations to obtain a more stable distribution. The text encoder consists of a pre-trained BERT model, two LSTM layers, and one LayerNorm layer. Sentences are segmented into words, mapped into word IDs based on the BERT lexicon, and fed into the pre-trained BERT model to obtain M $\times$ 768 word-level embeddings, with M and 768 denoting the number of words and word representations, respectively. Finally, these word-level embeddings are processed into high-level text feature representations of shape M $\times$ 256 using LSTM and LayerNorm with the same structure as the speech encoder.

2) Emotion Feature Extractor with DA-LSTM: As illustrated in Fig. 2, the emotion feature extractor is structurally identical to the gender feature extractor except that the LSTM is replaced by the DA-LSTM, which could transfer knowledge from the gender feature extractor to the emotion feature extractor. Thus, our proposed DA-LSTM is described in detail below.

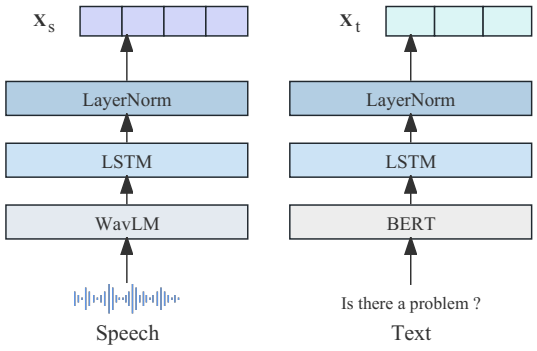


Fig. 1. A speech feature encoder and a text feature encoder.

After the k-1_{th} epoch of training, the parameters of two feature extractors with the same shape are obtained as follows:

$$\mathcal{P}^{em,k-1} = \{W_x^{em,k-1}, W_h^{em,k-1}, b^{em,k-1}\} \quad (1)$$

$$\mathcal{P}^{ge,k-1} = \{W_x^{ge,k-1}, W_h^{ge,k-1}, b^{ge,k-1}\} \quad (2)$$

where $\mathcal{P}^{em,k-1}$ comes from the emotion feature extractor, which extracts emotional information from utterances, and $\mathcal{P}^{ge,k-1}$ comes from the gender feature extractor, which extracts gender-related information from utterances. It is imaginable that $\mathcal{P}^{ge,k-1}$ could complement $\mathcal{P}^{em,k-1}$. Thus, it is valuable to aggregate them as the emotion feature extractor in the k_{th} epoch.

In this work, we propose a Deep Aggregation LSTM (DA-LSTM) as shown in Fig. 2. After the training of the k-1_{th} epoch, the $\mathcal{P}^{em,k-1}$ and $\mathcal{P}^{ge,k-1}$ are aggregated as follows:

$$W_x^{em,k} = \alpha * W_x^{em,k-1} + (1-\alpha) * W_x^{ge,k-1} \quad (3)$$

$$W_h^{em,k} = \alpha * W_h^{em,k-1} + (1-\alpha) * W_h^{ge,k-1} \quad (4)$$

$$b^{em,k} = \alpha * b^{em,k-1} + (1-\alpha) * b^{ge,k-1} \quad (5)$$

where $W_{(\cdot)}^{em,k-1}$ and $b^{em,k-1}$ are parameters in DA-LSTM of emotion feature extractor after the training of the k-1_{th} epoch, $W_{(\cdot)}^{ge,k-1}$ and $b^{ge,k-1}$ are parameters in LSTM of gender feature extractor after the training of the k-1_{th} epoch, α is the weight of the emotion classification task, and $W_{(\cdot)}^{em,k}$ and $b^{em,k}$ are set as the initial parameters of DA-LSTM in the k_{th} epoch.

B. Cross-Modal Fusion Modules

Since the high-level feature representations (HFR) of speech and text are in different dimensional spaces, we compress and map the HFRs of different modalities into the same shape before modal fusion to further extract emotional information and reduce the number of parameters. Taking speech modality as an example, the HFR with shape N $\times$ 256 is downscaled and input to a fully connected layer for compression, and then upscaled and fed to a LayerNorm layer to obtain the aligned HFR with shape D $\times$ 32. A similar process is applied for the HFR of text to obtain the aligned HFR with shape D $\times$ 32, and D is the minimum of M and N.

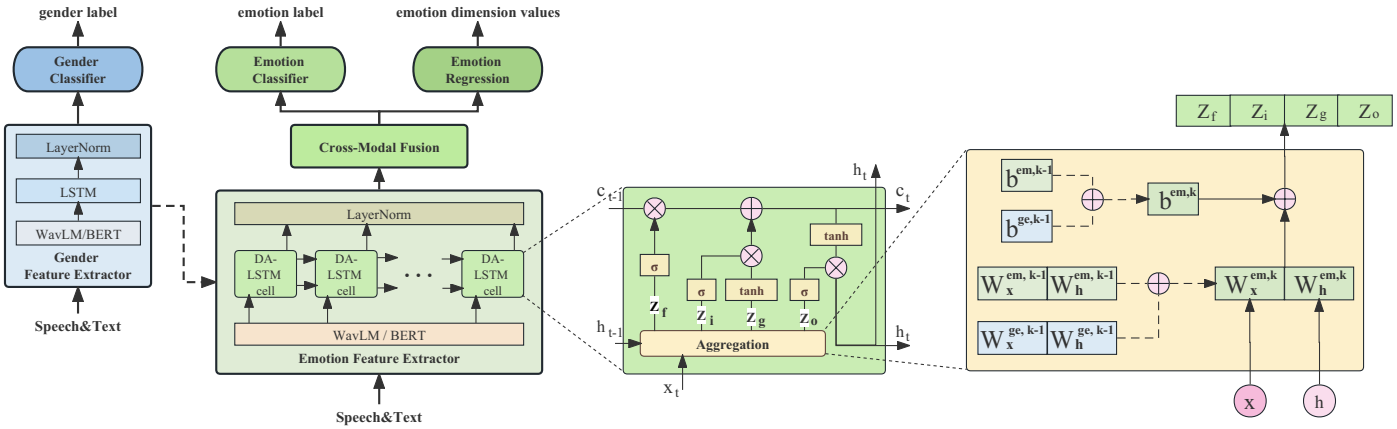


Fig. 2. Our proposed hybrid multi-task learning framework with Deep Aggregation LSTM (DA-LSTM) for end-to-end multimodal emotion recognition.

After aligning the HFRs of the two modalities, multi-head attention is employed to capture the interaction information between different modalities, allowing parallel processing of different attention heads to accelerate the inference process. Given a source sequence $X_s \in \mathbb{R}^{D\times 32}$ and a target sequence $X_t \in \mathbb{R}^{D\times 32}$, these sequences are employed with linear transformation and performed with attention h times to obtain h heads, and these heads are concatenated and employed with linear transformation. The process by which the source sequence X_s obtains complementary information from the target sequence X_t is denoted as Cross-Modal Multi-Head Attention $CM\text{-}MHA_{s\to t}$:

$$head_i = Attention(X_s W_i^Q, X_t W_i^K, X_t W_i^V) \quad (6)$$

$$CM\text{-}MHA_{s\to t} = Concat(head_1, ..., head_h)W^O \quad (7)$$

where W_i^Q, W_i^K, W_i^V and W^O are learnt parameters. Based on the $CM\text{-}MHA_{s\to t}$, the cross-modal representations of speech and text are obtained by treating the aligned HFRs of speech and text as source sequences respectively.

C. Output Modules

In this work, we employ five output modules, including an emotion classifier, three emotion regression modules (Valance, Activation, Dominance), and a gender classifier, all of which consist of a fully connected layer, a ReLU activation function, a dropout layer, and an output layer. In unimodal systems, the five output modules use only unimodal HFR as input. In multimodal systems, the emotion classifier and the emotion regression modules predict four emotion labels and three sentiment-dimensional values using a joint emotion representation that combines unimodal HFRs and cross-modal representations, respectively, and the gender classifier predicts two gender labels with a joint gender representation that combines HFRs of speech and text.

D. Multi-Task Learning Strategies

To demonstrate the difference between our proposed hybrid multi-task learning and traditional multi-task learning, these two strategies are described in detail in this subsection. Both of them consider three tasks, with emotion classification as the primary task while emotion regression and gender classification as auxiliary tasks. Let $\mathcal{L}_{ec}$ and $\mathcal{L}_{ge}$ denote the cross entropy losses of emotion and gender classification respectively, $\mathcal{L}_{er}$ the sum loss of emotion regression, $loss_{val}$, $loss_{act}$, $loss_{dom}$ the MSE losses of the valance, activation, dominance dimensional value prediction tasks respectively, and β the weight of the emotion regression task. The detailed description of the computation functions for these two strategies is as follows:

1) Hybrid Multi-Task Learning: As illustrated in Fig. 2, on the one hand, the knowledge of gender classification is transferred to emotion classification through our proposed DA-LSTM, thus improving the discrimination of the same emotion on different genders. On the other hand, since the goal of both emotion regression and emotion classification is to learn discriminative emotion representations, they can naturally share the emotion feature extractor and cross-modal fusion modules. The entire emotion recognition model is optimized by the following loss function $\mathcal{L}$:

$$\mathcal{L}_{er} = loss_{val} + loss_{act} + loss_{dom} \quad (8)$$

$$\mathcal{L} = \mathcal{L}_{ec} + \beta * \mathcal{L}_{er} \quad (9)$$

2) Traditional Multi-Task Learning: The output modules of all auxiliary tasks share the emotion feature extractor and cross-modal fusion modules with the emotion classifier. The entire model is optimized by the following function $\mathcal{L}$:

$$\mathcal{L}_{er} = loss_{val} + loss_{act} + loss_{dom} \quad (10)$$

$$\mathcal{L} = \alpha * (\mathcal{L}_{ec} + \beta * \mathcal{L}_{er}) + (1-\alpha) * \mathcal{L}_{ge} \quad (11)$$

where α is the weight of the weighted combination of emotion classification and emotion regression.

III. Experiment

A. Emotion Dataset

To evaluate the performance of our proposed method, we conducted experiments on the Interactive Emotion Dyadic Motion Capture (IEMOCAP) dataset, which is a public English

emotion corpus recorded by the University of Southern California. The IEMOCAP dataset contains 12 hours of audio-visual data with corresponding transcripts. The audio-visual data is guided by 10 professional actors in a dialogue scene with script and improvisation, and divided into five sessions. Each session contains some conversation between a female actor and a male actor. These conversations are artificially split into sentence-level utterances. Each utterance is marked by at least three markers for discrete emotion labels (neutral, happy, sad, angry, excited, etc.) and three dimension values (Valance, Activation, Dominance). To stay consistent with the prior studies [17]–[19], we select 5531 utterances with four emotions: neutral, happy (the emotion 'excited' is merged into 'happy'), sad, and angry. And five-fold cross-validation with a leave-one-session-out (LOSO) strategy is conducted for classification evaluation.

B. Experimental Setup

In this experiment, the maximum length of the speech utterance was adjusted to 3 seconds by truncating and zeroing. The same operation was used to set the maximum length of the text utterance to 40 tokens. The optimizer AdamW with a learning rate of $1e^{-5}$ and a weight decay of $1e^{-3}$ is employed to optimize the model. The rate of dropout is set to 0.5. Moreover, in order to obtain the best performance, the coefficients of multi-task learning α and β are set to 0.7 and 0.3 respectively.

IV. Results and Discussion

In this section, we evaluate the performance of the proposed method by weighted accuracy (WA) and unweighted accuracy (UA), which are commonly employed in the evaluation of emotion classification models. The weighted accuracy is the overall classification accuracy, and the unweighted accuracy is the average recall over the emotion categories. TABLE I shows the performance comparison of multi-task learning methods for unimodal and multimodal emotion recognition systems, and all results are averaged over five-fold cross-validation on the IEMOCAP dataset. TABLE I is composed of three parts: speech emotion recognition systems (Sys. 1-4), text emotion recognition systems (Sys. 5-8), and multimodal emotion recognition systems (Sys. 9-12). Each part incorporates four different multi-task learning strategies. The first strategy (Sys. 1,5,9) involves solely the primary task without any auxiliary tasks. The second strategy (Sys. 2,6,10) includes emotion regression as the only auxiliary task. The third strategy (Sys. 3,7,11) employs traditional multi-task learning. Lastly, the fourth strategy (Sys. 4,8,12) adopts our proposed hybrid multi-task learning. The performance comparisons of systems in TABLE I with different modalities or multi-task learning strategies are as follows:

A. Comparison of Different Modal Systems

The audio modal systems (Sys. 1-4) demonstrated a lower performance than text modal systems (Sys. 5-8), which may be attributed to the fact that audio utterances are more individual in terms of emotional expression than text utterances. The multimodal systems (Sys. 9-12) achieve at least 7.01% and 6.97% improvement in WA and UA respectively compared to the unimodal systems (Sys. 1-8), which shows that multimodal systems perform better than unimodal systems. These results are consistent with many previous studies [24]–[26].

TABLE I
PERFORMANCE COMPARISON OF MULTI-TASK LEARNING METHODS FOR UNIMODAL AND MULTIMODAL EMOTION RECOGNITION SYSTEMS. "A" AND "T" REFER TO AUDIO AND TEXT MODALITIES RESPECTIVELY. EC, ER, AND GC DENOTE EMOTION CLASSIFICATION, EMOTION REGRESSION, AND GENDER CLASSIFICATION. THE SUPERSCRIPT * INDICATES THAT DA-LSTM IS ADOPTED FOR THIS TASK.

Sys.	Modal	Task	WA(%)	UA(%)
1	A	EC	68.32	68.17
2	A	EC + ER	66.57	67.55
3	A	EC + ER + GC	67.23	67.17
4	A	EC + ER + GC*	68.72	69.21
5	T	EC	69.11	69.74
6	T	EC + ER	69.51	69.79
7	T	EC + ER + GC	68.80	69.81
8	T	EC + ER + GC*	69.72	69.91
9	A&T	EC	76.12	76.71
10	A&T	EC + ER	76.36	76.82
11	A&T	EC + ER + GC	76.89	77.63
12	A&T	EC + ER + GC*	**77.34**	**77.91**

TABLE II
PERFORMANCE COMPARISON BETWEEN OUR METHOD AND LATEST METHODS ON IEMOCAP. THE SYMBOL '-' INDICATES THAT THE VALUE IS NOT MENTIONED IN THE REFERENCE. BOLD INDICATES THE BEST PERFORMANCE.

Method	WA(%)	UA(%)
Zhang et al. [20]	56.87	59.47
Deschamps et al. [21]	62.00	63.00
Li et al. [22]	63.40	-
Ours (audio only)	68.72	69.21
Liu et al. [5]	72.39	70.08
Chen et al. [6]	74.30	75.30
Lin et al. [23]	76.40	76.90
Wang et al. [4]	76.80	77.10
Ours (audio&text)	**77.34**	**77.91**

B. Comparison of Different Multi-Task Learning Strategies

As shown in TABLE I, using more auxiliary tasks can usually increase the accuracy of emotion classification (except Sys. 2,7), which demonstrates that multi-task learning helps to more comprehensively parse the emotions contained in the discourse and that traditional multi-task learning could occasionally experience negative migration issues. Significantly, systems with hybrid multi-task learning (Sys. 4,8,12) achieved superior results in all modalities compared to those with traditional multi-task learning (Sys. 3,7,11). This shows that DA-LSTM can more effectively learn gender-specific emotional representation thus accurately identifying emotions. Finally, the multimodal system with hybrid multi-task learning achieves the best performance.

TABLE II shows the performance comparison between our proposed method and the latest methods, where the upper part

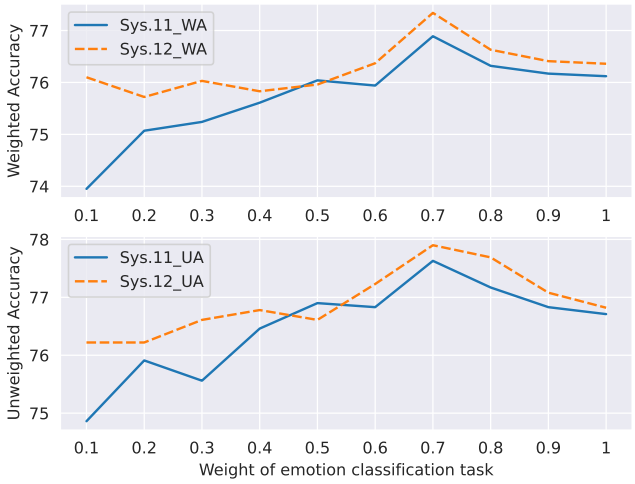


Fig. 3. Performance comparison between hybrid multi-task learning (the orange dotted line) and traditional multi-task learning (the blue line) with different weights of the emotion classification task.

uses only audio modality and the lower part uses multimodal data as input. As shown in the results, our hybrid multi-task learning model achieves improved performance compared to the recent studies employing either multi-task [20]–[23] or single-task [4]–[6] approaches. We incorporated multiple auxiliary tasks based on similar architecture as in [6] and achieved gains of 3.04% and 2.61% in WA and UA, respectively.

In addition, to comprehensively compare the performance of hybrid multi-task learning and traditional multi-task learning, we adjust the value of α based on the multimodal systems Sys. 11 and Sys. 12 in TABLE I. Fig. 3 shows the performance of Sys. 11 and Sys. 12 with different values of α which is the weight of emotion classification. As shown in Fig. 3, the performance of Sys. 11 and Sys. 12 under different weights of the primary task has roughly the same trend, with the latter having a flatter trend, which indicates that DA-LSTM can effectively mitigate the negative migration problem. Second, Sys. 12 shows better performance than Sys. 11 in most settings, and when both systems reach the best performance, Sys. 12 achieves 0.45% and 0.28% improvement in WA and UA, respectively, which demonstrates that DA-LSTM can improve the emotion recognition accuracy more effectively than sharing the feature extractors within traditional multi-task learning.

V. Conclusion

In this paper, we propose a hybrid multi-task learning framework for end-to-end multimodal emotion recognition, where the primary task is emotion classification and the auxiliary tasks are emotion regression and gender classification. In addition, we design a Deep Aggregation LSTM (DA-LSTM) to transfer knowledge from gender classification to the primary task. The experimental results show that: 1) hybrid multi-task learning performs better than traditional multi-task learning. 2) DA-LSTM can effectively transfer knowledge from the gender feature extractor to the emotion feature extractor.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62071330, No. 62201571, and No. U21B2010).

References

[1] W. Li, R. Tan, Y. Xing, G. Li, S. Li, G. Zeng, P. Wang, B. Zhang, X. Su, D. Pi *et al.*, "A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks," *Scientific Data*, vol. 9, no. 1, p. 481, 2022.

[2] R. Francese and P. Attanasio, "Supporting depression screening with multimodal emotion detection," in *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, 2021, pp. 1–8.

[3] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[4] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition." in *Interspeech*, 2021, pp. 4518–4522.

[5] P. Liu, K. Li, and H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," *arXiv preprint arXiv:2201.06309*, 2022.

[6] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[7] J. Li, S. Wang, Y. Chao, X. Liu, and H. Meng, "Context-aware multimodal fusion for emotion recognition," *Proc. Interspeech 2022*, pp. 2013–2017, 2022.

[8] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Context-dependent domain adversarial neural network for multimodal emotion recognition." in *INTERSPEECH*, 2020, pp. 394–398.

[12] D. Krishna and A. Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks." in *Interspeech*, 2020, pp. 4243–4247.

[13] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition." in *Interspeech*, 2020, pp. 374–378.

[14] R. Mukherjee, A. Naik, S. Poddar, S. Dasgupta, and N. Ganguly, "Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2303–2307.

[15] Y. Li, T. Zhao, T. Kawahara *et al.*, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning." in *Interspeech*, 2019, pp. 2803–2807.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[17] M. Hou, Z. Zhang, and G. Lu, "Multi-modal emotion recognition with self-guided modality calibration," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4688–4692.

[18] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6269–6273, 2020.

[19] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Interspeech*, 2019.

[20] H. Zhang, M. Mimura, T. Kawahara, and K. Ishizuka, "Selective multitask learning for speech emotion recognition using corpora of different styles," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7707–7711.

[21] T. Deschamps-Berger, L. Lamel, and L. Devillers, "End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[22] Y. Li, P. Bell, and C. Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.

[23] B. Lin and L. Wang, "Robust multi-modal speech emotion recognition with asr error adaptation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] S. Dutta and S. Ganapathy, "Multimodal transformer with learnable frontend and self attention for emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6917–6921.

[25] V. Rajan, A. Brutti, and A. Cavallaro, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4693–4697.

[26] D. Sun, Y. He, and J. Han, "Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and bert for multimodal emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

It's What You Say and How You Say It: Exploring Audio and Textual Features for Podcast Data

Neil Shah, Vivek Srivastava, Mohit Bhardwaj, Satej Kadlay,
Dharmeshkumar Agrawal, Savita Bhat, Niranjan Pedanekar
TCS Research
Pune, India
{neilkumar.shah, srivastava.vivek2, mohit.bhardwaj4, satej.kadlay, dharmesh.agrawal, savita.bhat, n.pedanekar}@tcs.com

***Abstract*—Podcasts are a relatively new media type, consisting of spoken documents or conversations with a wide range of topics, genres, and styles. It is of utmost importance to understand what makes a podcast more popular than others or which tags help to characterize a podcast. Here, we provide a comprehensive analysis of widely used traditional features from two modalities, i.e., audio and text. We explore multiple feature combinations considering podcast popularity prediction and multi-label tag assignment as proxy downstream tasks. Our analysis is focused on examining prevalent hypotheses and identifying new perspectives about the podcast content.**

I. Introduction

Podcasts have emerged as an exciting medium for entertainment, advertisement, news, and information dissemination. According to Nielsen [1], the total number of podcast titles is reaching 2 million with a steady increase in listeners across all demographics. With a significant increase in the listener base and podcast content, several open problems in handling and accessing this information have emerged. Jones et al.[2] highlight some of the unique challenges and future directions in the domain of podcast information access. They point out that the existing technologies addressing tasks such as *Search*, *Recommendation*, *Summarization*, and *User experience* are inadequate to handle the multi-genre, multi-style, and multi-format composition of podcasts. As this opens up an exciting landscape for future research, we believe that a comprehensive feature analysis of podcasts can serve as an important groundwork in tackling some of these problems.

Podcasts represent inherently heterogeneous data consisting of music and speech in different spoken and written styles and formats. With this work[1], we investigate the efficacy of individual traditional features and their combinations in understanding *what* the podcast contains and *how* the podcast is delivered. Traditionally various downstream applications use Textual and audio features such as tags, sentiment, emotions, Mel-frequency Cepstrum Coefficients (MFCC), and the multiband modulation features. We formulate an application-oriented framework to evaluate and understand the interplay among these features. We consider two separate applications, i.e., *Podcast Popularity Prediction*, and *Podcast Tag Assignment*. To summarize, we evaluate various hand-crafted audio and textual features described in Section III. In Section IV, we curate and enhance the dataset introduced in [3]. Experiments and in-depth analysis are detailed in Sections V and VI respectively. Our contribution in this work is a detailed feature analysis on the two aforementioned downstream tasks along with key insights focused on podcast data. We pose a baseline for the podcast tag assignment task. We also share the enhanced dataset[2] to facilitate future research in the podcast domain.

II. Related Work

Clifton et al. [4] introduced a large-scale *Spotify* podcast dataset (≈1,00,000 podcasts & ≈60,000 hours of speech) for search and summarization. Recent work in [5] further enriches this dataset with precomputed audio features based on voice and MFCCs. They demonstrate the effectiveness of these features in podcast segment categorization based on delivery. We use a similar feature set and even capture the speech modulation variation to understand podcast data but on a different dataset. In our study, we enhance the data compiled by Yang et al.[3]. They introduce an Adversarial Learning-based Podcast Representation (ALPR) that captures non-textual aspects of podcasts. They evaluate these representations in the context of podcast popularity prediction. On the same dataset, Joshi et al. [6] consider DistilBERT based embeddings with triplet loss and report state-of-the-art results for podcast popularity prediction. Due to the unavailability of the source code and unclear experimental setup, direct comparison with their score is not appropriate. Alternatively, we seek to understand the interplay between various hand-crafted features from podcast audio and its transcript.

Recently, Reddy et al. in [7] investigated the correlation between various syntax level linguistic features and podcast engagement. In contrast, we use content-specific textual features in our study. The efficacy of hand-crafted audio and textual features remains relatively unexplored in the podcast domain. Inspired by these works, we seek to understand

[1]Neil Shah, Vivek Srivastava, and Mohit Bhardwaj contributed equally in implementation, experimentation, and writing the manuscript. Satej Kadlay and Dharmeshkumar Agrawal worked on dataset collection and feature extraction. Savita Bhat and Niranjan Pedanekar contributed to the formulation and manuscript writing.

[2]https://github.com/mohit19014/Feature-Analysis-in-podcast-domain.git

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

how data from different modalities perform (separately or combined) in the context of podcast data.

III. Features

A podcast is essentially an audio experience with focused textual content. Accordingly, we consider both audio and text modalities in this study. Next, we present a brief overview of hand-crafted features and their characteristics in the context of a podcast.

A. Audio Features

We consider audio features that are appropriate to represent *how* a podcast is delivered. To that extent, we use short-time modulation features derived through multi-band filtering and energy operators and the other traditional speech processing features such as MFCC, voice features, and emotions to extract salient properties.

Modulation Features: The presence of amplitude and frequency modulations in speech formats improves the performance in voice activity detection, audio event detection, and audio summarization tasks [8], [9]. To extract these representations, we rely on the multiband filtering method proposed in [8] that tracks a filter that records Maximum average Teager Energy (MTE) and also computes their corresponding values of Mean Instantaneous Amplitude (MIA) and Mean Instantaneous Frequency (MIF). In this work, we extract 1198-dimensional modulation features using 40 Gabor filters, resample the podcast at $16kHz$, and use a window size and a window shift of $25ms$ and $10ms$, respectively.

MFCC: MFCC is based on the human auditory system [10]. It employs a non-linear scale to correlate with the human perception of the spectral contents of a sound. We consider 39-dimensional delta and double-delta MFCC representation using the similar parameters adopted for extracting modulation features.

Voice Features: It is possible to capture the speaker characteristics, their speaking style, and approximately identify the listener's interest towards a section of an audio [11]. We use PRAAT to extract 15-dimensional voice features on the entire podcast data [12]. The voice features include $F0$-median, $F0$-mean, $F0$-standard deviation, $F0$-minimum, $F0$-maximum, number of pulses, number of periods, their mean and standard deviation, number of unvoiced frames, number and degree of voice breaks, mean autocorrelation, mean noise-to-harmonic (NHR) and mean harmonic-to-noise (HNR) ratio.

Affective Audio Features: The emotional state in audio is one of the most important paralinguistic messages captured during human interactions. We use a Wav2Vec2 fine-tuned model on the IEMOCAP database to extract 4-dimensional audio emotions features (*neutral*, *happy*, *sad*, and *angry*) of an entire podcast [13], [14].

B. Textual Features

According to Nielsen[1], podcast engagement has steadily seen growth in the number of listeners owing to quality content. As an exercise to understand the characteristics of such content, we evaluate various textual features. We consider a set of textual features that can help understand *what* a podcast talks about.

Tags: It is observed that every podcast is assigned a set of tags representing its categorization such as *Arts*, *Society & Culture*, *Sports*, *Business*, etc. This assignment is user-defined and captures the scope of the podcast. We use an average 100-dimensional Glove embedding vector representing the assigned set of tags to the podcast.

Topics: Even though a set of tags can describe a podcast, topics give a detailed list of concepts covered in a podcast. We use the unsupervised topic detection algorithm Top2Vec [15] to generate a list of topics for a given podcast. We represent the topic words using an average 100-dimensional Glove embedding vector for a given podcast.

Affective Textual Features: Emotions play an important role in human cognition, including perception, attention, learning, and reasoning [16]. Affective content is more engaging to users than neutral content [17]. Accordingly, we consider two perspectives of affective features given below. leftmargin=*

1) **Sentiment**: Sentiment of a podcast refers to the inclination of the podcast content towards positive or negative polarity. We use the NLTK sentiment analyzer to create a 3-dimensional sentiment representation vector for each podcast with scores corresponding to *negative*, *neutral*, and *positive* sentiment that sum up to 1.
2) **Text Emotions**: We consider six basic human emotions, i.e., *happy*, *sad*, *anger*, *fear*, *surprise*, and *disgust*, assigned to sentences in podcast transcripts. We use a zero-shot sentence classification setup with task-aware sentence representations [18] to predict the emotion and the corresponding confidence for each sentence. We combine confidence score of the emotion detection model and sentence count for an emotion to calculate a posterior probability of each emotion. This results in a 6-dimensional emotion vector representing a given podcast.

Named Entities: Named entities like *person*, *organization*, *location*, and *others* can play a significant role in attracting audience attention. From the podcast title, we predict named entities and the corresponding confidence score using the TARS [18] zero-shot sentence classification setup. To obtain a four-dimensional named entities representation for a podcast, we use the similar formulation used in the text emotion feature representation.

Document Embeddings: Pre-trained language models are often used to extract contextual representations. Considering the longer length of podcast transcripts, we use Longformer embeddings (LF Embeddings) [19] to extract podcast content representations.

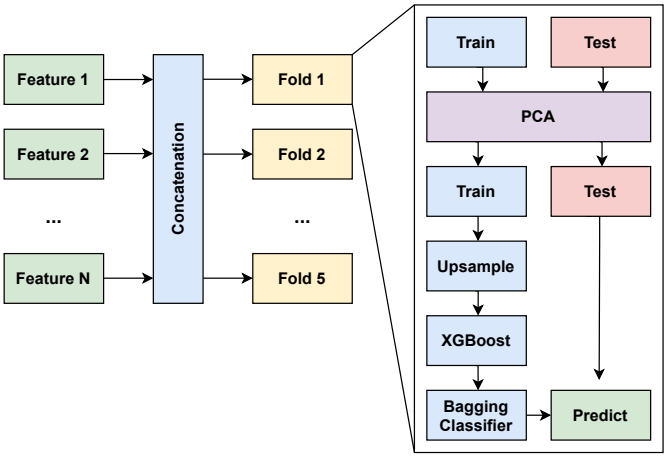


Fig. 1. Proposed architecture for podcast popularity prediction. For every fold, the classification setup involves PCA, upsampling, XGBoost, and a bagging classifier.

IV. Dataset

For the task of popularity prediction, Yang et al.[3] introduced a dataset of $6,511$ English language podcasts - 837 popular, and 5674 unpopular. For a podcast, we scrape its audio file and tags using the RSS feed of the podcast. It is possible to assign one or more tags to a podcast from 105 fine-grained tags. However, to increase the reach of a podcast, some creators label their podcast into multiple categories that don't especially fit their content [20]. To reduce noise in the tag set, we merge these 105 fine-grained tags into 19 coarse-grained tags as described by Apple podcasts[3], which results in a cleaner data with only 0.2% of the podcasts having three or more tags as opposed to 46.18% earlier. Since the transcripts had no punctuation, we punctuate the transcripts using a bidirectional RNN and attention-based punctuation restoration technique [21]. Further, we extract the audio and text features mentioned in Section III. Finally, we end up with a dataset of 3306 podcasts (444 popular, and 2862 unpopular) as either some audio links were broken or information was missing.

V. Experimental Setup

The podcast data is multimodal and heterogeneous, motivating us to experiment with combinations of diverse feature sets for both the downstream tasks. We conduct our experiments using a 20 core CPU with $64GB$ RAM.

A. Podcast Popularity Prediction

In this task, we seek to investigate the factors influencing the popularity of a podcast using a fine-tuned XGBoost and bagging classifier-based architecture. We perform the grid search on hyperparameter space to find the best parameter combination[4]. Fig. 1 shows our binary podcast popularity classification model. We use upsampling with SMOTE[22] and bagging to address the data imbalance. We report the experimental results using the macro-F1 score.

[3]https://podcasts.apple.com/us/genre/podcasts/id26

[4]*gamma* = 0.2, *maximum depth* = 14, *estimators* = 120, *reg_alpha* = 0.8, and *reg_lambda* = 1.2

TABLE I
Ablation analysis for Popularity prediction task.

Feature(s)	Macro-F1
All Text	0.545
(−) Tags	0.500
(−) Text Emotions	0.537
(−) Sentiment	0.539
(−) Topics	0.547
(−) Named Entities	0.549
All Audio	0.802
(−) MIF	0.603
(−) Voice	0.784
(−) MFCC	0.801
(−) MTE	0.801
(−) MIA	0.801
(−) Audio Emotions	0.802
All Audio + All Text	0.807
(−) MIF	0.603
(−) Voice	0.791
(−) Topics	0.798
(−) MTE	0.801
(−) MIA	0.801
(−) Sentiment	0.801
(−) MFCC	0.801
(−) Audio Emotions	0.802
(−) NER	0.802
(−) Tags	0.809
(−) Text Emotions	0.812

B. Podcast Tag Assignment

In the novel multi-label podcast tag assignment task, we seek to automatically assign appropriate coarse-grained tags to a podcast. We consider a unified two-layer perceptron network with dropout after the first layer[5]. We use binary cross-entropy loss function with *class_weights* to handle dataset imbalance. We calculate *class_weights* for each tag l as N/N_l, where N is the total number of podcasts in the training dataset, and N_l denotes the numbers of podcasts with tag l. Since the dataset is highly imbalanced, we use the weighted-F1 score (w-F1) as the primary evaluation metric. Similar to [23], we also evaluate the features on tag recall (TR) and unique tags learned (TL) over weighted-F1 scores.

VI. Results and Analysis

In this section, we report our results and analysis on the performance of features in different modalities for the two aforementioned downstream tasks.

A. Podcast Popularity Prediction

In Table I, we present the ablation study of text, audio, and multimodal features for popularity prediction. Using the textual modality, the tags and affective features (text emotions and sentiment) are the most significant in predicting the podcast popularity since their exclusion results in a drop of

[5]*no. of epochs* = 75, *learning rate* = 0.0001, *batch size* = 4, *optimizer* = Adam, *dropout=0.3*

TABLE II
ABLATION ANALYSIS FOR MULTI-LABEL TAG ASSIGNMENT TASK.

Feature(s)	w-F1	TR	TL
All Text	44.02	78.22	19
(−) Topics	35.31	76.34	19
(−) LF Embeddings	41.89	71.85	19
(−) Named Entities	43.37	79.44	19
(−) Sentiment	43.42	77.07	19
(−) Text Emotions	43.83	79.34	19
All Audio	18.79	80.40	18
(−) Voice	16.26	57.89	14
(−) MIA	17.55	51.20	15
(−) MTE	17.87	40.78	15
(−) Audio Emotions	18.68	71.84	15
(−) MFCC	19.30	63.84	17
(−) MIF	22.54	70.39	17
All Text + Audio	18.62	69.36	17
(−) LF Embeddings	17.57	70.23	16
(−) Voice	17.71	63.86	15
(−) MTE	17.90	53.89	15
(−) Audio Emotions	17.91	54.09	12
(−) Topics	18.36	71.92	16
(−) Sentiment	18.49	52.59	15
(−) Named Entities	18.60	59.59	17
(−) MIA	18.63	58.46	15
(−) Text Emotions	18.98	65.40	17
(−) MFCC	19.26	55.45	15
(−) MIF	31.14	55.47	16

8.25%, 1.46%, and 1.10% in the macro-f1 score, respectively. However, the topics and named entities negatively contribute to the popularity prediction results. The MIF and voice features contribute the most in popularity prediction amongst all audio features as with their exclusion, we report a drop of 24.81% and 2.24% in the macro-f1 score. Additionally, we observe that the features such as MFCC, MTE, MIA, and audio emotions do not show significant performance improvement in audio and multimodal experiments.

Since the combinations of all audio and text representations report the highest macro-F1 score of 0.807, we further experiment with different combinations of these features to identify the best performing multimodal feature combination. The top multimodal feature combination with the highest macro-F1 score is 0.820 for (modulation + voice + tags). Experiments with only textual dataset in [6] report the best performing system with a 0.820 macro-F1 score. In contrast, we use a much smaller dataset of 3306 podcasts compared to 6511 podcasts in [6] and report similar results for the best performing system. The multiband modulation features with their ability to model dynamic aspects of audio are the most significant ones in capturing the intricacies of a podcast for this task [8].

B. Podcast Tag Assignment

In Table II, we present our ablation study for multi-label tag assignment for text, audio, and multimodal frameworks. Topics emerge as the most prominent feature as their exclusion results in a 19.78% drop in the weighted-F1 score. The audio features as standalone representations do not perform well in tag assignment tasks. Even with all the audio features, the model fails to learn all the tags and perform relatively poor when compared with the text modality. Voice features dominate the audio modality as with their exclusion, the model fails to learn five tags altogether while also producing a significant drop of 13.46% and 27.99% in the weighted-F1 score and tag recall, respectively.

The model also does not benefit from multimodal setup. Similar to the audio modality, the MFCC, and MIF features fail to capture the tags appropriately. We note an increase of 67.23% in the weighted-F1 score with the removal of MIF features. In a multimodal framework, LF embeddings perform the best, followed by audio emotions, voice features, and sentiments.

C. Key Insights

In addition to the above findings, we report insights that help in understanding *whats* and *hows* of the podcast. We select all available 444 popular podcasts and select an equal subset of the unpopular podcast in 6-folds.

Young and energetic voice contribute to the podcast's popularity The mean pitch values in voice features are consistently higher in the popular podcast, as is high MIA values. The lower *p-values* in the statistically significant *t-test* metric of 0.003 and 0.014 for mean pitch and MIA between popular and unpopular distributions indicates that on average, a more young and energetic voice contributes to the podcast's popularity [24].

Speakers in popular podcasts tend to speak relatively fast The higher pulses and periods in data indicates that on average popular podcast tends to have high periodic speech variations and high speech rate, as also reported in [7]. We report the *p-value* of 0.042 for periodicity in the *t-test* metric between the popular and unpopular distributions.

Emotions from audio and text modalities capture different aspects of a podcast We observe a non-trivial dissimilarity between emotion identified using audio signal and textual content. The audio emotion features associate the music category with the most share for '*sad*' emotion as compared to other categories. Whereas text emotions portray a low share for '*sad*' emotion and instead associate music category having maximum share for '*happy*' emotion. Since audio and text capture different aspects of the content, both modalities are equally important to understand the podcast content better.

Podcasts categories tend to have preference for speaker's gender Though the voice features help us to extract the speaker's gender information, we rely on the pre-trained model [25] to gather the gender-specific segments. The gender share for every podcast varies depending on the podcast category. It is interesting to note that podcasts under '*Kids & Family*' category have significant '*Female*' voice (56%) whereas '*Sports*', '*Technology*', and '*News*' categories are dominated by speakers tagged as '*Male*' (82%, 79% & 77% respectively).

At times speaker's gender influences podcast popularity We investigate the distribution of speakers' gender among popular and unpopular podcasts. We observe that unpopular podcasts under categories such as '*Kids & Family*' have a high percentage of '*Male*' voice share (72%). However, popular podcasts under the same category have a high '*Female*' voice share (65%). On the other hand, unpopular podcasts within the '*Sports*' category have a high percentage of '*Female*' voice share (48%) as compared to the '*Male*' voice share (42%). Popular podcasts in '*Sports*' have 82% '*Male*' voice share as compared to 8% '*Female*' voice share. Even though we observe a slight association between gender and popularity across certain categories, its causality should be validated. We need to exercise caution in generalizing these associations because of uncertainty in underlying models used for identifying voice information and the lack of enough data.

Popular podcasts contain more musical content We use the pre-trained model [25] to extract music segments for every podcast. Popular podcasts have 64.81% more musical content than unpopular podcasts. Also, we observe that popular podcasts under the '*Music*' category on average have 46.05% of musical content as compared to 3.17% in unpopular podcasts.

Topics effectively captures *what* is covered in a podcast Topics provide a detailed list of concepts covered within the podcast. It is essential to understand *what* the podcast is about, which also helps to assign relevant tags. We observe that podcasts under the '*Technology*' category are associated with topics like apps, software, iPhone, and similarly, podcasts under the '*Leisure*' category are associated with topics like games, racing, and cards.

VII. Conclusion

Podcasts are spoken documents ranging across several genres, topics, and styles. With the rapid growth in popularity and global reach, there is a definite need to explore and investigate this new engagement medium and relevant research landscape. In this work, we use the podcast audio and its transcript to characterize the podcast data. We observe that the hand-crafted features capturing distinct qualities like speaker's style, emotion, and subject content coverage (i.e., voice, emotions, sentiment, and topics) are significant irrespective of the downstream task. Even though we report various insights relevant to podcast popularity, tag assignment, and content features, care must be taken while using and generalizing these results. We believe this analysis can benefit several other downstream tasks such as podcast summarization, retrieval, and recommendation.

References

[1] "Podcasting today: Insights for podcast advertisers," https://www.nielsen.com/us/en/insights/report/2021/podcasting-today-2/, accessed: 2022-03-06.

[2] R. Jones, H. Zamani, M. Schedl, C.-W. Chen, S. Reddy, A. Clifton, J. Karlgren, H. Hashemi, A. Pappu, Z. Nazari *et al.*, "Current challenges and future directions in podcast information access," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1554–1565.

[3] L. Yang, Y. Wang, D. Dunne, M. Sobolev, M. Naaman, and D. Estrin, "More than just words: Modeling non-textual characteristics of podcasts," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 276–284.

[4] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette *et al.*, "100,000 podcasts: A spoken english document corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5903–5917.

[5] A. Alexander, M. Mars, J. C. Tingey, H. Yu, C. Backhouse, S. Reddy, and J. Karlgren, "Audio features, precomputed for podcast retrieval and information access experiments," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2021, pp. 3–14.

[6] B. Joshi, S. Mittal, and A. Chetan, "Did you "read" the next episode? using textual cues for predicting podcast popularity," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 13–17.

[7] S. Reddy, M. Lazarova, Y. Yu, and R. Jones, "Modeling language usage and listener engagement in podcasts," *arXiv preprint arXiv:2106.06605*, 2021.

[8] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2024–2038, 2006.

[9] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1294–1298.

[10] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[11] J. Adell, A. Bonafonte, and D. Escudero, "Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech," *Procesamiento del lenguaje natural*, no. 35, pp. 277–283, 2005.

[12] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[14] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[15] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*, 2020.

[16] C. M. Tyng, H. U. Amin, M. N. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in psychology*, vol. 8, p. 1454, 2017.

[17] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu, "A three-level framework for affective content analysis and its case studies," *Multimedia tools and applications*, vol. 70, no. 2, pp. 757–779, 2014.

[18] K. Halder, A. Akbik, J. Krapac, and R. Vollgraf, "Task-aware representation of sentences for generic text classification," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3202–3213.

[19] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[20] M. Sharpe, "A review of metadata fields associated with podcast rss feeds," *arXiv preprint arXiv:2009.12298*, 2020.

[21] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration." in *Interspeech*, 2016, pp. 3047–3051.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[23] S. Kar, S. Maharjan, and T. Solorio, "Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2879–2891.

[24] C. A. Klofstad, S. Nowicki, and R. C. Anderson, "How voice pitch influences our choice of leaders: when candidates speak, their vocal

characteristics–as well as their words–influence voters' attitudes toward them," *American Scientist*, vol. 104, no. 5, pp. 282–288, 2016.
[25] "Python pip project for gender, silence and music identification from audio," https://pypi.org/project/inaSpeechSegmenter/, accessed: 2022-03-06.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Robust Finger Vein Recognition Based on Lightweight Attention Convolutional Neural Networks

Ming-Yi Wei[1], Yu-Chi Wang[2], Liang-Ying Ke[3] and Chih-Hsien Hsia[3,*]
[1]Department of Electrical Engineering, National Taiwan University of Science and Technology, Taiwan
E-mail: B10832007@mail.ntust.edu.tw
[2]Department of Mathematics, King's College London, United Kingdom
E-mail: k21098296@kcl.ac.uk
[3]Department of Computer Science and Information Engineering, National Ilan University, Taiwan
[*]E-mail: r1143001@ems.niu.edu.tw; hsiach@niu.edu.tw

***Abstract* — In the field of computer vision (CV), convolutional neural networks (CNN) have demonstrated superior capability in feature extraction. However, the translation during the process of capturing finger vein images can potentially degrade the accuracy rate of the model, making it challenging to apply CNN to real-time and high accuracy finger vein recognition. Additionally, while CNN offers high accuracy, they also require a high number of parameters. Previous studies also confirmed the absence of shift-invariant features in CNN. In the light of these issues, this research proposes a lightweight attention convolution neural network (LACNN) for applications in finger vein recognition. The model incorporates diverse branch block (DBB), adaptive polyphase sampling (APS), and coordinate attention mechanism (CAM). To evaluate the performance of the model under finger vein recognition, the finger-vein by university sains malaysia (FV-USM) public database is employed for analysis and comparison with recent research methods. According to the result, the new model achieves an accuracy recognition rate of 99.82% with a parameter count of only 1.23 million. Thus, the LACNN presented in this study surpasses the other recent finger vein recognition methods.**

I. INTRODUCTION

With rapid development of information technology, the significance of personal data has grown increasingly prominent. Utilizing biometric recognition, which includes the identification of individuals based on distinctive physiological or behavioral characteristics, has captured the attention of the public. Hsia *et al.* [1] identified the four fundamental characteristics of biometric recognition: *1)* Universality: the presence of biological characteristics that are possessed by every individual, such as fingerprints, iris patterns, and finger veins; *2)* Distinctiveness: the uniqueness of these biological features among different individuals. Even in the case of identical twins, their biometric characteristics are distinct and can be utilized for identity recognition; *3)* Permanence: an individual's biometric traits should remain constant and stable over time, without significant changes; *4)* Collectability: the feasibility of conveniently and rapidly measuring biometric features using sensors. Chen *et al.* [2-3] further categorized the acquisition regions of biometric traits into two types: *1)* external biological characteristics, including features such as fingerprints, iris patterns, and facial characteristics; and *2)* internal biological characteristics, such as finger veins and palm veins. However, the accuracy of external biological characteristics is more susceptible to environmental factors, leading to a decrease in accuracy compared to internal biological characteristics. Finger veins, on the other hand, represent a more stable internal biological characteristic. Therefore, veins are more suitable for applications related to identity recognition.

In recent years, Liu *et al.* [4] proposed a novel CNN architecture for finger vein recognition. However, the complexity of this model results in a parameter count of over 5 million, which may pose limitations when deploying the model in practical settings due to hardware constraints. Chai *et al.* [5] utilized convolution layers and global average pooling (GAP) to avoid an increase caused by fully connected layers. However, significantly increasing the channel dimension of the model can lead to an increase in the number of parameters, making the model less amenable to lightweight implementation. Chai *et al.* [6] uses continuous convolutional layers and ACE to improve model generalization. However, this model did not address the translation issue of finger vein features, potentially resulting in inconsistent recognition due to variations in user operations.

The objective of this study is to address the issue of excessive parameter count in deep learning networks applied to finger vein recognition, while simultaneously improving the model's ability to handle translation variations in finger vein features. To achieve this, we propose an lightweight attention convolution neural network (LACNN) structure and evaluate its performance using the finger-vein by university sains malaysia (FV-USM) database. Experimental results demonstrate that our proposed approach successfully reduces the model's parameter count while maintaining high accuracy in recognition results.

II. METHODOLOGY

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In this study, we propose a finger vein recognition system based on the LACNN model, which can be divided into training and testing stages, as illustrated in Fig. 1. During the training stage, the back propagation network (BPN) is employed to iteratively optimize the model's weight parameters. In the testing stage, the optimized lightweight model is used for finger vein recognition. The model takes an input image and predicts the identity based on the vein patterns. When the model's output probability is above 50%, the image is classified as the predicted identity. Otherwise, it is considered a prediction error.

A. Model Compression

The classical CNN structure often exhibits high accuracy in many recognitions but comes with a large number of parameters, making them less suitable for practical deployment. Structural re-parameterization is a method that uses functions to transform the parameters of a complex model structure into those of a simplified model structure. This training strategy allows the model to utilize abundant computational resources during the training stage for optimization. A series of transformation functions is applied to convert the complex model structure into an equivalent simplified model structure for inference. Ding *et al.* [7] proposed the diverse branch block for structural re-parameterization. By extracting feature maps of different scales using branches of complexities, the module enriches the model's diversity in the feature space. Therefore, we use DBB modules to replace all convolutional layers in the model structure for training. After completing the training, we will use the original model structure for inference.

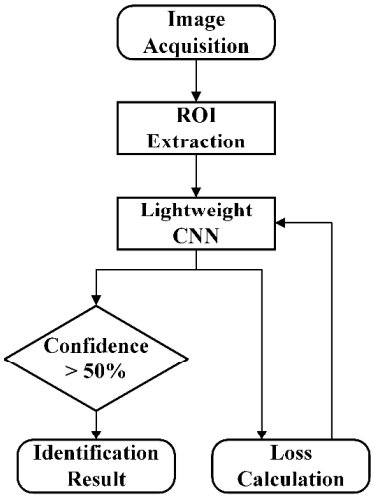


Figure 1. Flowchart of the proposed finger vein recognition

B. Attention Mechanism

To further enhance the ability of the model to extract effective features in finger vein recognition tasks, the proposed lightweight model structure incorporates the use of an attention mechanism. Hou *et al.* [8] introduced a novel attention mechanism named the coordinate attention mechanism (CAM). CAM compresses the spatial information of the input feature map using two separate average pooling layers in different directions. The resulting one-dimensional features from the two directions are then concatenated and fed into a convolutional layer to encode the channel-wise representation of the feature map. By using the same convolution weights for the two directions, the information between the vectors can be shared. It helps the model improve its ability to recognize vein features. In this study, we will use the CAM attention mechanism to enable the model to extract more discriminative features for recognition and enhance the accuracy of identity recognition for different individuals.

C. Shift-Invariant Features

CNN is generally believed to be able to extract features that are invariant to translation. However, recent studies have found that making subtle changes to the input image can greatly affect the output of the model. In the task of finger vein recognition, the captured images may undergo translation due to user manipulation. Therefore, image translation becomes one of the factors that have to be considered when designing finger vein recognition model structure. Chaman *et al.* [9] proposed adaptive polyphase sampling (APS) to address the issue of inconsistency features in neural network structures. When down sampling the input feature map, APS generates four potential candidate sampling results, which are referred to as polyphase components. At this point, we can compute the L_p normalization for each candidate sampling result and choose the one with the highest computed value as the optimal sampling result. In this study, we will use APS sampling to replace the original pooling layers, allowing the model to maintain consistency in the extracted features and achieve stable recognition results in finger vein recognition.

D. Margin-based Loss Functions

In recent years, various forms of margin-based Softmax loss function have been proposed with the aim of minimizing intra-class variations and maximizing inter-class differences to enhance the accuracy of model recognition. Deng *et al.* [10] applied L_2 normalization to the deep feature x and the weights W_{y_i} of the last fully connected layer in the model. However, normalizing both the deep feature and the weights with L_2 normalization can sometimes lead to less distinct feature differences in the model's output. Therefore, we employ elastic penalty margin loss (EPML) function [11] for finger vein recognition, in which the margin penalty is determined by

random values obtained from a Gaussian distribution. This approach provides the model with greater flexibility and a more adaptable space for learning class separability.

E. Proposed Lightweight Model Structure

Given the need to reduce model parameters and maintain high accuracy in finger vein recognition tasks, this study proposes an LACNN structure for one-to-many identity recognition. To address the challenges of gradient degradation caused by increasing model depth and decreasing gradient correlation between neurons, we propose an improved DBB residual module based on the ResNet residual module. Residual modules are effective in solving the problem of gradient vanishing. The DBB residual module consists of a DBB module, APS sampling, and CAM attention mechanism. The DBB module replaces all convolutional layers in the original residual module to increase the model's complexity and enhance the diversity of feature extraction. During the inference phase, the module's parameters are transformed using a function that maintains the optimal parameters while achieving lightweight representation. The CAM attention mechanism identifies important features in the input feature map to enhance the model's sensitivity to critical features, thereby improving overall accuracy. APS sampling effectively addresses the translation effects in the input feature images caused by user operations and improves the consistency of the model's feature extraction across different translations.

The LACNN structure, constructed using the proposed DBB residual module (DBB ResBlock), is shown in Fig. 2. The model structure is divided into two main parts: Input flow, and Output flow and described in the following:

1) Input flow

First, to meet the computational limitations of the model structure and avoid distortion caused by changes in image scale, we employ padding to expand the vein image to a square shape. Then, the image is resized to a side length of 112 before being inputted into the model for feature extraction. After using the DBB module to extract features, the ReLU activation function is applied to filter the features and enable the model to express nonlinear characteristics. Finally, the CAM attention mechanism and APS sampling are applied consecutively to focus on important feature regions in the feature maps and reduce their size. This reduction in feature map size helps reduce the computational load in subsequent stages of the model.

2) Output flow

In the final stage, we first utilize global average pooling to compress the spatial information of the feature maps, avoiding direct computations with fully connected layers to reduce computational complexity. Simultaneously, we obtain a feature vector of length 256, which is then input into fully connected layers to obtain two types of feature results: normalized features obtained by applying normalization to the weights and input features, and identification results without normalization. In particular, the normalized features are used to calculate the loss value in the EPML, which is utilized for model weight updates to optimize the model. On the other hand, the identification results can be directly used to examine the predicted identity class by the model and perform subsequent metric computations. It is worth noting that in the output flow, we place the Dropout layer before the output layer to improve generalization ability, thereby further enhancing recognition accuracy of the model.

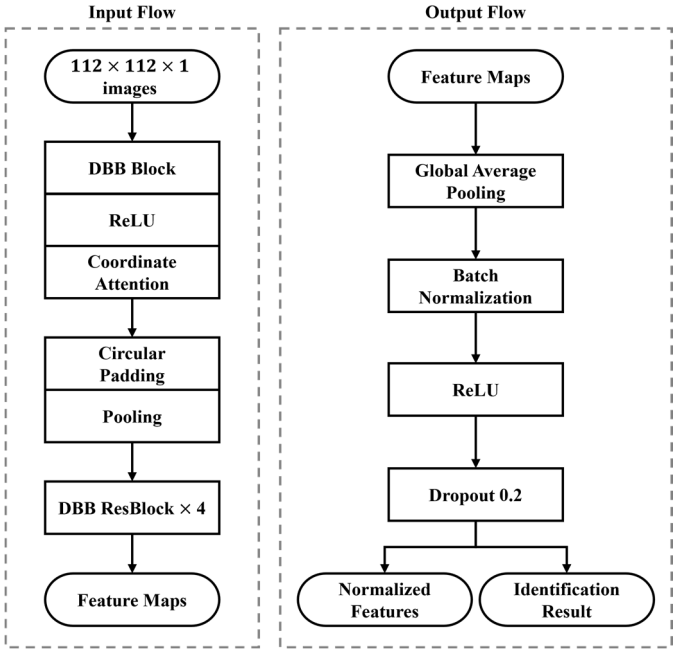


Figure 2. Proposed LACNN structure

III. EXPERIMENTS

A. Database

This study utilized the FV-USM public database [12] for implementing deep learning. The database contains 5,904 infrared finger vein images with a resolution of 640x480, collected from 123 subjects, resulting in a total of 492 classes in the database.

B. Experimental Environment and Model Parameter Setting

In this study, we utilized a custom designed LACNN structure for finger vein recognition. For model training, we employed the EPML and used the AdamW optimizer. In terms of hyperparameter settings, the image size was set to 112, the batch size was set to 32, the number of epochs was set to 500,

and the initial learning rate was set to 2e-4. In this study, we implemented all training processes using the Pytorch DL framework on Intel® Core TM i7-8700 CPU and Nvidia GTX 1060 GPU.

C. Experiment Results

We evaluated the overall performance of the proposed structure using the FV-SUM database. The evaluation metric used was the correct identification rate (CIR), which measures the security of the model. Based on the same database, we compared the performance with other lightweight research methods in terms of CIR and model parameters, as shown in Table 1. It can be observed that our proposed model structure achieves the best result in terms of CIR while having the lowest number of model parameters. Compared to the previous model with the minimum number of parameters, our model structure reduces the parameter count by approximately 53%. This reduction makes our method less contained by hardware limitations and easily applicable to handheld platforms or embedded systems.

Table 1. CIR and number of parameters of each method.

Methods	CIR (%)	Parameters (M)
Liu *et al.* (2021) [4]	98.58	5.85
Semi-PFVN (2022) [5]	94.67	3.35
LightFVN + ACE (2022) [6]	96.17	2.65
LFVRN_CE + ACE [13]	99.09	4.93
Ke *et al.* (2023) [14]	99.80	75.00
LACNN (2023)	**99.82**	**1.23**

IV. Conclusion

Finger vein recognition technology has undergone continuous development and has been proven to be an effective and reliable biometric identification method. Yet for high security applications requiring one-to-many identification, there are still many challenges that need to be addressed. In this study, we propose a novel finger vein recognition technique that enhances the model's diversity in feature learning by increasing the complexity of the model structure during the training stage. The proposed technique also enables the model to capture globally important features and exhibit robustness against translation, thus addressing the issue of feature instability in extraction. Experimental results demonstrate the proposed model achieves a high recognition rate of 99.82%. Furthermore, the performance of the proposed model surpasses that of the latest research methods.

V. References

[1] Chih-Hsien Hsia and Chin-Hua Liu, "New hierarchical finger-vein feature extraction method for iVehicles," *IEEE Sensors Journal*, vol. 22, no. 13, pp. 13612-13621, 2022.

[2] Yung-Yao Chen, Chih-Hsien Hsia, and Ping-Han Chen, "Contactless multispectral palm-vein recognition with lightweight convolutional neural network," *IEEE Access*, vol. 9, pp. 149796-149806, 2021.

[3] Yung-Yao Chen, Sin-Ye Jhong, Chih-Hsien Hsia, and Kai-Lung Hua, "Explainable AI: A multispectral palm-vein identification system with new augmentation features," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 111, pp. 1-21, 2021.

[4] Weiye Liu, Huimin Lu, Yupeng Li, Yifan Wang, and Yuanyuan Dang, "An improved finger vein recognition model with a residual attention mechanism," *Chinese Conference on Biometric Recognition*, pp. 231-239, 2021.

[5] Tingting Chai, Jiahui Li, Shitala Prasad, Qi Lu, and Zhaoxin Zhang, "Shape-driven lightweight CNN for finger-vein biometrics," *Journal of Information Security and Applications*, vol. 67, pp. 103211, 2022.

[6] Tingting Chai, Jiahui Li, Yanhong Wang, Guoying Sun, Changyong Guo, and Zhaoxin Zhang, "Vascular enhancement analysis in lightweight deep feature space," *Neural Processing Letters*, 2022.

[7] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding, "Diverse branch block: Building a convolution as an inception-like unit," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10886-10895, 2021.

[8] Qibin Hou, Daquan Zhou, and Jiashi Feng, "Coordinate attention for efficient mobile network design," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713-13722, 2021.

[9] Anadi Chaman and Ivan Dokmanic, "Truly shift-invariant convolutional neural networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3773-3783, 2021.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefansos Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690-4699, 2019.

[11] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1578-1587, 2022.

[12] Mohd Shahrimie Mohd Asaari, Shahrel A. Suandi, and Bakhtiar Affendi Rosdi, "Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3367-3382, 2014.

[13] Yiqi Zhong, Jiahui Li, Tingting Chai, Shitala Prasad, and Zhaoxin Zhang, "Different dimension issues in deep feature space for finger-vein recognition," *Chinese Conference on Biometric Recognition*, pp. 295-303, 2021.

[14] Liang-Ying Ke, Shu-Yen Yeh, and Chih-Hsien Hsia, "Attention-based deep learning model using adaptive margin loss for finger-vein recognition," *IEEE International Conference on Electronic Communications, Internet of Things and Big Data*, pp. 144-146, 2023.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Sound Field Estimation around a Rigid Sphere with Physics-informed Neural Network

Xingyu Chen*, Fei Ma*, Amy Bastine*, Prasanga Samarasinghe*, Huiyuan Sun†
* Audio and Acoustic Signal Processing Group, The Australian National University, Canberra, Australia
E-mail: {xingyu.chen1, fei.ma, amy.bastine, prasanga.samarasinghe}@anu.edu.au
† School of Electrical and Information Engineering, The University of Sydney, Australia
E-mail: {huiyuan.sun}@sydney.edu.au

***Abstract*—Accurate estimation of the sound field around a rigid sphere necessitates adequate sampling on the sphere, which may not always be possible. To overcome this challenge, this paper proposes a method for sound field estimation based on a physics-informed neural network. This approach integrates physical knowledge into the architecture and training process of the network. In contrast to other learning-based methods, the proposed method incorporates additional constraints derived from the Helmholtz equation and the zero radial velocity condition on the rigid sphere. Consequently, it can generate physically feasible estimations without requiring a large dataset. In contrast to the spherical harmonic-based method, the proposed approach has better fitting abilities and circumvents the ill condition caused by truncation. Simulation results demonstrate the effectiveness of the proposed method in achieving accurate sound field estimations from limited measurements, outperforming the spherical harmonic method and plane-wave decomposition method. Our code is available at github.com/xingyuaudio/pinn_estimation.**

I. Introduction

Sound field analysis is a key topic of research with a wide range of applications, including spatial audio processing [1], [2], augmented reality/virtual reality [3], and active noise control [4], [5]. To determine the distribution of acoustic pressure in a region, researchers often use microphone arrays. However, microphone arrays cause reflections and scattering, which potentially impacts the accuracy of the sound field estimation. Therefore, understanding the scattering field is essential to ensure the accuracy of the sound field estimation.

Rigid spherical microphone arrays are extensively employed in sound field analysis due to their theoretically derived scattering fields, enabling the estimation of sound fields in three-dimensional space surrounding the array [6]–[8]. Conventional methods, such as the spherical harmonic (SH) method [9], represent the sound field by a set of orthonormal modes. However, these methods have inherent limitations in practical applications. For instance, the limited number of microphones leads to a discrete representation through integration, resulting in truncation errors [10]. Plane-wave decomposition method assumes that the sound field only consists of plane waves [11]. It neglects the spatial variation within each wavefront, which may not fully capture the intricacies of the scattering fields.

Learning-based methods have been employed in scattering problems, and have demonstrated effectiveness in fitting nonlinear mappings [12], [13]. Among these methods, Convolutional Neural Networks (CNNs) have gained popularity due to their ability to promote parameter sharing and reduce the number of trainable parameters [14]. However, in the context of sound field estimation, convolutional and max-pooling layers introduce pixel spaces that can lead to discontinuities in sound pressure continuity. This discontinuity potentially undermines the physical interpretation of the data [15]–[17]. Additionally, the heavy reliance on training data often necessitates a large simulated dataset.

Recently, researchers introduced Physics-Informed Neural Networks (PINNs) as a method to incorporate physical knowledge into the network architecture and training [17]–[24]. This integration allows PINNs to achieve physically feasible estimations and require less training data. Fully connected Neural Networks (FNNs) are commonly employed as part of PINN due to their ability to approximate any continuous function, as guaranteed by the Universal Approximation Theorem [25].

In this paper, we propose a method for sound field estimation around a rigid sphere using PINNs. We employ a small FNN network structure and incorporate two constraints: the Helmholtz equation and the fact that sound pressure cannot generate radial motion at the sphere boundary. We use the FNN to model the mapping between Euclidean space positions and sound pressure and use these constraints as a loss function to train the network. Notably, our method relies solely on measured data from microphones, eliminating the need for setting up a grid and using simulated datasets. Simulation results demonstrate the effectiveness of our method, outperforming the spherical harmonic method and the plane-wave decomposition method.

II. problem formulation

We introduce two coordinate systems, as shown in Fig. 1, where $\boldsymbol{x} = (x, y, z)$ and $\boldsymbol{r} = (r, \theta, \phi)$ represent Cartesian and spherical coordinates, respectively. Our study focuses on a sound field $\mathcal{V} \subset \mathbb{R}^3$, which includes a rigid sphere located at the origin O with a radius of a. We aim to estimate the sound field $P(\boldsymbol{r}, \omega)$ at position $\boldsymbol{r}$ and angular frequency ω around the sphere $r \geq a$, by using measurements from Q microphones mounted on the rigid sphere.

For a homogeneous sound field (absence of sound sources) around a rigid sphere, theoretical solutions for the sound field have been established [6]. The spherical harmonic (SH) method decomposes $P(\boldsymbol{r}, \omega)$ into a series of orthonormal

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

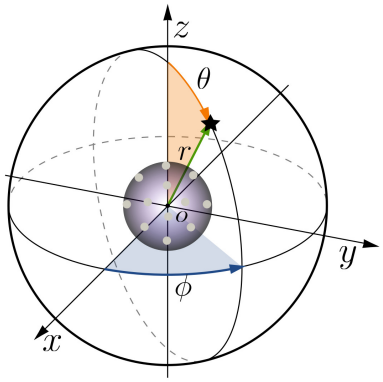


Fig. 1: System setup: A rigid sphere with microphones showing Cartesian and spherical coordinates. • represent the microphones, and ⋆ represents a reference point.

modes [6]

$$P(\boldsymbol{r},\omega)=\sum_{n=0}^{\infty}G_n(r,a,\omega)\sum_{m=-n}^{n}\mathcal{P}_n^m(a,\omega)Y_n^m(\theta,\phi), \quad (1)$$

where $G_n(r,a,\omega)$ is the radial propagator, $\mathcal{P}_n^m(a,\omega)$ are the spherical harmonic coefficients, and $Y_n^m(\theta,\phi)$ is the spherical harmonic of order n and degree m. $\mathcal{P}_n^m(a,\omega)$ are obtained by integrating $P(\mathbf{a},\omega)$ over the surface of the sphere

$$\mathcal{P}_n^m(a,\omega)=\iint P(\mathbf{a},\omega)Y_n^m(\theta,\phi)^*d\Omega, \quad (2)$$

where $\mathbf{a}=(a,\theta,\phi)$, $\Omega=(\theta,\phi)$, and $\iint(\cdot)d\Omega\equiv\int_0^{2\pi}\int_0^{\pi}(\cdot)\sin\theta d\theta d\phi$. $G_n(r,a,\omega)$ must satisfy the boundary condition on the rigid sphere, leading to

$$G_n(r,a,\omega)=j_n(\frac{\omega r}{c})-\frac{j_n'\left(\frac{\omega a}{c}\right)}{h_n^{(2)'}\left(\frac{\omega a}{c}\right)}h_n^{(2)}(\frac{\omega r}{c}), \quad (3)$$

where $j_n(\cdot)$ and $h_n^{(2)}(\cdot)$ are the spherical Bessel function of the first kind and the spherical Hankel function of the second kind, respectively, and $j_n'(\cdot)$ and $h_n^{(2)'}(\cdot)$ are corresponding derivatives [8], c is the speed of sound.

There are three issues with the implementation of the theoretical method. First, the infinite series in (1) needs to be truncated at limited order N [10]. Second, the pressure on the sphere is sampled at a limited (Q) number of positions. Hence, the integral in (2) needs to be discretely approximated. Third, in the low-frequency region where the distance r increases ($r>a$), the propagators $G_n(r,a,\omega)$ approximately follow a power law of the form $\frac{n+1}{2n+1}\left(\frac{r}{a}\right)^n$. These increasing values have significant implications as they indicate an ill-posed problem [8]. Consequently, small variations in $\mathcal{P}_n^m(a,\omega)$ resulting from truncation errors or noise will be magnified through the propagator.

To overcome these challenges and obtain more accurate estimations, we propose a neural network approach integrated with physics laws. This method aims to establish a regression relationship between coordinates and sound pressure while satisfying the governing Helmholtz equation and the boundary condition.

III. PINN METHOD

In this section, we present the PINN method, which consists of three steps: (1) constructing an FNN using microphone positions and measured sound pressure as training pairs, (2) incorporating physical knowledge as a constraint in the network, and (3) adjusting the weight of the loss function to facilitate convergence. The workflow of the PINN method is presented in Fig. 2.

FNN is a composition of multiple layers, and the nodes between adjacent layers are fully connected. A single layer of the network is denoted as

$$F(\boldsymbol{x})=\sigma\left(\boldsymbol{x}^{\intercal}\mathbf{w}+b\right), \quad (4)$$

where $\boldsymbol{x}\in\mathbb{R}^3$ is the inputs, which in this case are the Cartesian positions in the sound field $\mathcal{V}$, $\mathbf{w}$ represents the weights, b is the bias, and σ is the activation function. The FNN is a composition of multiple layers

$$\Phi(\boldsymbol{x},\omega;\psi)=\left(F_n\circ\ldots\circ F_2\circ F_1\right)(\boldsymbol{x}), \quad (5)$$

where ψ represents the set of all trainable parameters $\mathbf{w}$ and b. Hereafter, ω is omitted in $\Phi(\boldsymbol{x};\psi)$ when representing acoustic quantities for notional simplicity. ψ is learned by minimizing the mean squared error (MSE) loss $\mathcal{L}_{\text{data}}$ with respect to the training data

$$\psi^*=\arg\min_{\psi}\mathcal{L}_{\text{data}}(\boldsymbol{x};\psi), \quad (6)$$

where $\mathcal{L}_{\text{data}}$ is defined as

$$\mathcal{L}_{\text{data}}(\boldsymbol{x}_i;\psi)=\frac{1}{Q}\sum_{i=1}^{Q}\left(P(\boldsymbol{x}_i,\omega)-\Phi\left(\boldsymbol{x}_i;\psi\right)\right)^2. \quad (7)$$

It fits the regression relationship between the spatial coordinates of the microphones $\boldsymbol{x}$ and the corresponding sound pressure values $P(\boldsymbol{x},\omega)$. The network results $\Phi(\boldsymbol{x};\psi)$ represent the estimated sound pressure $\hat{P}(\boldsymbol{x},\omega)$.

The aforementioned approach represents a purely data-driven method. It needs a large dataset, which is not always possible. We incorporate domain knowledge to reduce the data requirements and generate physically meaningful estimations.

For sound field estimation around a rigid sphere, we can leverage two domain knowledge to improve the estimation accuracy. Firstly, sound fields are governed by the Helmholtz equation

$$\left(\triangle+(\frac{c}{\omega})^2\right)P(\boldsymbol{r},\omega)=0, \quad (8)$$

where $\triangle=\partial^2/\partial x^2+\partial^2/\partial y^2+\partial^2/\partial z^2$ denotes the Laplace operator. Therefore, we require the PINN estimation to satisfy (8), i.e.,

$$\left(\triangle+(\frac{c}{\omega})^2\right)\Phi(\boldsymbol{x};\psi)=0. \quad (9)$$

Secondly, radial velocity is zero on the rigid sphere:

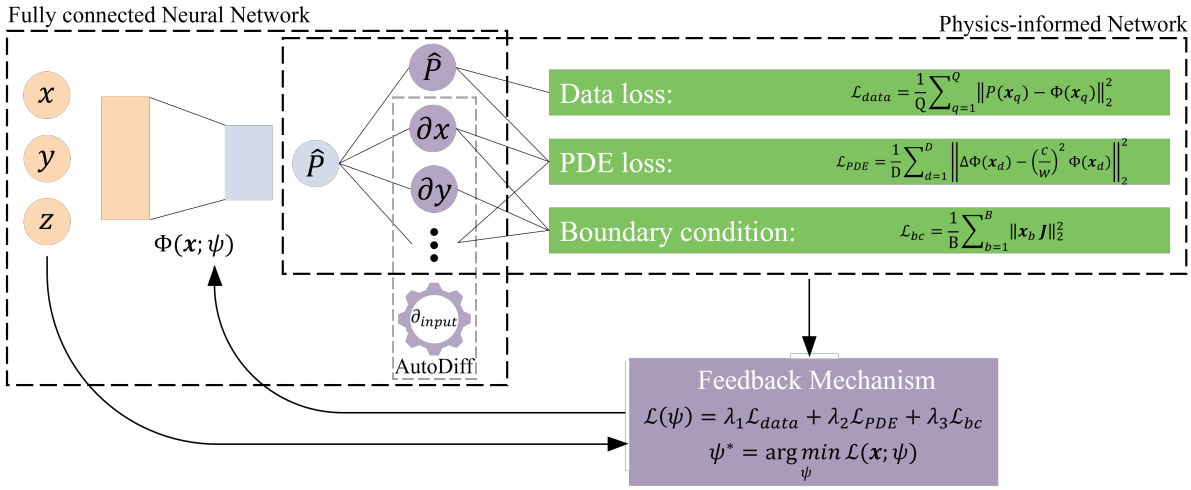


Fig. 2: The workflow includes three components: an FNN for mapping inputs to outputs with trainable parameters ψ, a Physics-Informed Network that incorporates governing Helmholtz equations and boundary conditions, and a feedback mechanism to minimize the loss function and update ψ.

$$\left.\frac{\partial P(\boldsymbol{r},\omega)}{\partial r}\right|_{r=a} = 0. \tag{10}$$

It represents the derivative of the sound pressure with respect to the radius of the sphere and can be written in terms of derivatives with respect to the Cartesian coordinates as

$$\frac{\partial P(\mathbf{r},\omega)}{\partial r} = \frac{\partial P(\mathbf{r},\omega)}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial P(\mathbf{r},\omega)}{\partial y}\frac{\partial y}{\partial r} + \frac{\partial P(\mathbf{r},\omega)}{\partial z}\frac{\partial z}{\partial r}. \tag{11}$$

Conduct the substitutions $\frac{\partial x}{\partial r} = \frac{x}{r}$, $\frac{\partial y}{\partial r} = \frac{y}{r}$, $\frac{\partial z}{\partial r} = \frac{z}{r}$, (10) can be expressed in a matrix form

$$\left.\frac{\partial P(\boldsymbol{r},\omega)}{\partial r}\right|_{r=a} = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} \frac{\partial P(\mathbf{r},\omega)}{\partial x} \\ \frac{\partial P(\mathbf{r},\omega)}{\partial y} \\ \frac{\partial P(\mathbf{r},\omega)}{\partial z} \end{bmatrix} = 0. \tag{12}$$

We denote $\boldsymbol{x_b}$ as a point on the rigid sphere with radius a and $\begin{bmatrix} \frac{\Phi(\boldsymbol{x};\psi)}{\partial x} & \frac{\Phi(\boldsymbol{x};\psi)}{\partial y} & \frac{\Phi(\boldsymbol{x};\psi)}{\partial z} \end{bmatrix}^{\mathsf{T}}$ as $\boldsymbol{J}$. Therefore, we require the PINN estimation to satisfy (12), i.e.,

$$\boldsymbol{x_b J} = 0. \tag{13}$$

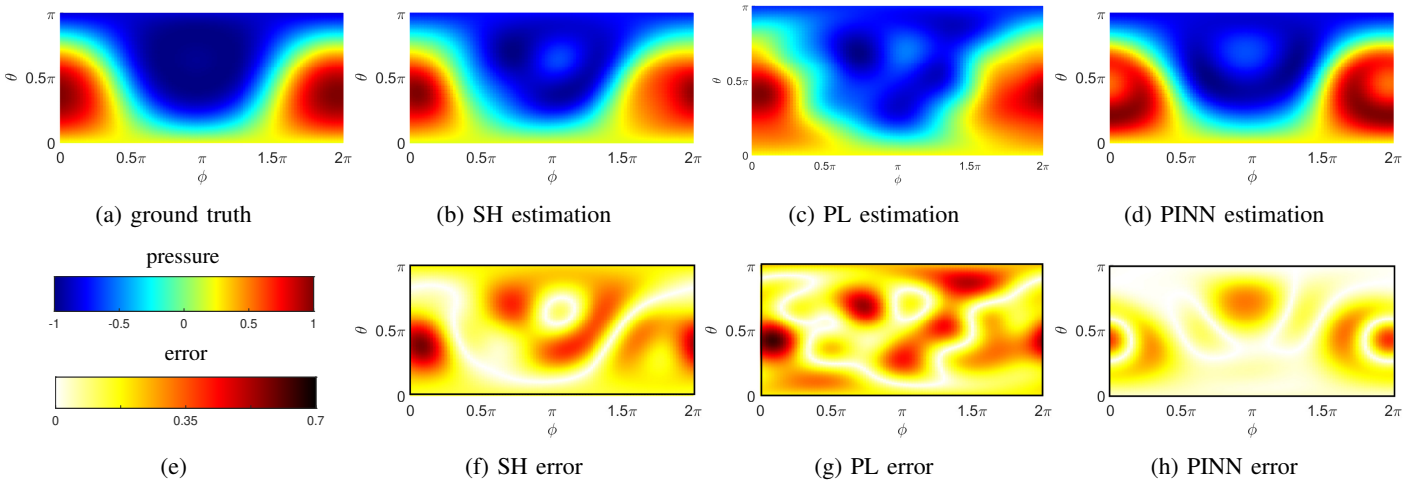
At present, we have constructed an FNN and acquired physical knowledge, which allows us to develop a PINN. Leveraging automatic differentiation within the neural network, we can efficiently compute the first-order and second-order derivatives of $\Phi(\boldsymbol{x};\psi)$ as required in (9) and (13). These derivatives are essential for updating the network parameters ψ using gradient descent.

We formulate the loss function to include three terms, each contributing to the overall training process

$$\begin{aligned} \mathcal{L}(\psi) = \lambda_1 \underbrace{\frac{1}{Q}\sum_{q=1}^{Q} \|P(\boldsymbol{x}_q) - \Phi(\boldsymbol{x}_q;\psi)\|_2^2}_{\mathcal{L}_{\text{data}}} \\ + \lambda_2 \underbrace{\frac{1}{D}\sum_{d=1}^{D} \left\|\triangle\Phi(\boldsymbol{x}_d;\psi) + (\frac{c}{\omega})^2\Phi(\boldsymbol{x}_d;\psi)\right\|_2^2}_{\mathcal{L}_{\text{PDE}}} \\ + \lambda_3 \underbrace{\frac{1}{B}\sum_{b=1}^{B} \|\boldsymbol{x}_b\boldsymbol{J}\|_2^2}_{\mathcal{L}_{\text{bc}}}, \end{aligned} \tag{14}$$

where $\|\cdot\|_2$ is the 2-norm, $\mathcal{L}_{\text{data}}$ represents the discrepancy between the PINN estimations $\Phi(\boldsymbol{x})$ and the actual measurements $P(\boldsymbol{x}_q)$ obtained from microphones. $\mathcal{L}_{\text{PDE}}$ ensures that the Helmholtz equation is satisfied for the positions $\boldsymbol{x}_d$ within the domain $\mathcal{V}$, and $\mathcal{L}_{\text{bc}}$ imposes the boundary condition of zero radial velocity on the rigid sphere. D are positions surrounding the sphere, and B are positions on the rigid sphere. To control the relative importance of these terms and ensure similar influences on the training process, we introduce weight factors λ_1, λ_2, and λ_3.

The weights in the loss function significantly influence the convergence and accuracy of the PINN. By balancing loss weights, a compromise can be achieved between accurately fitting the microphone measurements and adhering to the physical constraints. Since all three losses aim to approach zero, direct normalization is not feasible. $\mathcal{L}_{\text{data}}$ is influenced by data noise. To address the issue, we balance $\mathcal{L}_{\text{PDE}}$ and $\mathcal{L}_{\text{bc}}$ to the same order of magnitude during the initial stages of training. Specifically, We set λ_1 to 1, λ_2 to $(\frac{c}{\omega})^2$, and λ_3 to r.



(a) ground truth (b) SH estimation (c) PL estimation (d) PINN estimation

(e) (f) SH error (g) PL error (h) PINN error

Fig. 3: Sound field at f=1000f = 1000 Hz, r=0.072r=0.072 m (a) The ground truth, (b) SH estimation, (c) PL estimation, and (d) PINN estimation. (f), (g), (h) represent the corresponding errors, respectively, and (e) shows the color bar for reference.

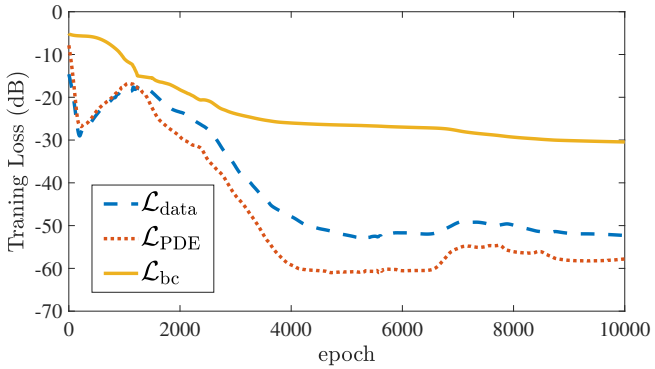


Fig. 4: PINN training loss: data loss, PDE loss, and boundary condition loss.

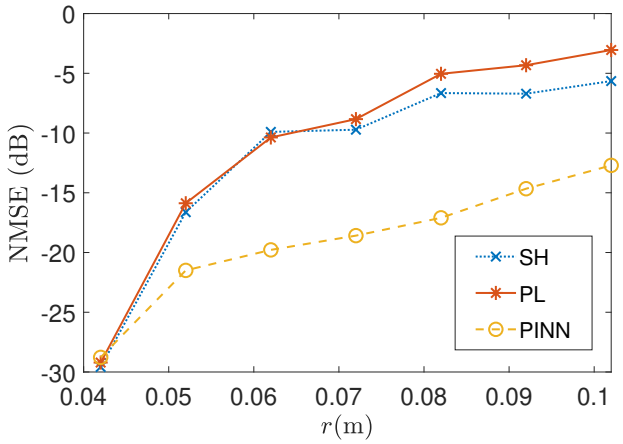


Fig. 5: The sound field estimation error NMSE at 1000 Hz as a function of radius.

IV. SIMULATION

We conduct simulations in a 3D free field to evaluate the proposed method regarding the sound field estimation surrounding a rigid sphere and compare the performance of three methods: the spherical harmonic method (SH) [9], the plane wave decomposition method (PL) [26], and the proposed method (PINN).

We uniformly mount 32 omnidirectional microphones on a rigid sphere with a radius of $a = 0.042$ m as shown in Fig. 1. We simulate the transfer functions between the sound source and the microphones on the rigid sphere according to [27]. Specifically, we place two point sources at coordinates $(2.5, 0.8, 0.0)$ m and $(-2.0, -0.6, 1.2)$ m, both with an amplitude of 1. The speed of sound is set to $c = 343$ m/s. In order to account for realistic conditions, we add white Gaussian noise to the observation signals, resulting in a signal-to-noise ratio (SNR) of 30 dB. The pressure values are then normalized within the range of $[-1, 1]$.

For the PINN method, we construct a 3-layer FNN with 4 nodes with the hyperbolic tangent (tanh) activation function. To train the PINN, we employ the Adam optimizer [28] with a learning rate of 10^{-5} for a total of 10,000 epochs. We set $\mathcal{L}_{\text{data}}$ based on the 32 microphone measurements, $\mathcal{L}_{\text{PDE}}$ based on 1000 randomly distributed points surrounding the sphere, and $\mathcal{L}_{\text{bc}}$ based on 500 uniformly distributed sampling points on the rigid sphere. As we train the network in Cartesian coordinates, we calculate the spherical coordinates $\boldsymbol{r}$ and then convert them to Cartesian coordinates $\boldsymbol{x}$.

In Fig. 4, we illustrate the training process. It demonstrates the simultaneous and nearly synchronized reduction of the training error $\mathcal{L}_{\text{PDE}}$ and $\mathcal{L}_{\text{bc}}$ in the initial stages of training for the first 2000 epochs. Subsequently, a gradual decline in $\mathcal{L}_{\text{data}}$ is observed. This observation suggests that the weight configurations of $\lambda_1 = 1$, $\lambda_2 = (\frac{c}{\omega})^2$, and $\lambda_3 = r$ achieve a good balance.

We define the pressure error between the true and estimated

sound fields as

$$\text{error} := \|P(\boldsymbol{x}_g, \omega) - \hat{P}(\boldsymbol{x}_g, \omega)\|, \tag{15}$$

where $P(\boldsymbol{x}_g, \omega)$ and $\hat{P}(\boldsymbol{x}_g, \omega)$ represent the true and estimated pressures, respectively.

Fig. 3 presents the ground truth, estimated sound pressure using these three methods, and error at a frequency of 1000 Hz and a radius of 0.072 m. It is observed that the SH method exhibits higher errors around $\theta = 0.5\pi$, the PL method performs the worst among the three due to its assumption of far-field conditions, which ignore the scattering information. the PINN method achieves the best results overall, although it still experiences relatively higher errors at $\theta = 0.5\pi$.

To evaluate the overall performance, we defined the normalized mean squared error (NMSE) as

$$\text{NMSE} := 10\log_{10} \frac{\sum_{g=1}^{10000} \|P(\boldsymbol{x}_g, \omega) - \hat{P}(\boldsymbol{x}_g, \omega)\|_2^2}{\sum_{g=1}^{10000} \|P(\boldsymbol{x}_g, \omega)\|_2^2}, \tag{16}$$

The estimation was performed at 10,000 evaluation positions uniformly sampled within a sphere with a specific radius and frequency.

Fig. 5 illustrates the relationship between the radius r of the estimation sphere and the NMSE at a frequency of 1000 Hz. When $r = a$, which corresponds to estimating the sound pressure on the rigid sphere, all three methods yield comparable results. However, as the radius increases, both the SH method and PL method show a significant surge in error. In contrast, the PINN method exhibits a more gradual increase in error as the radius increases. It has a lower NMSE of $7 - 10$ dB for $r > 0.05$ m than SH and PL methods. This indicates that the PINN method performs better compared to the conventional methods.

V. Conclusion

In this study, we propose a PINN method to estimate the sound field around a rigid sphere based on the measurement of a small number of microphones mounted on the rigid sphere. Our method employs an FNN to model the mapping between space positions and sound pressure. We incorporate the Helmholtz equation and the fact that sound pressure cannot generate radial motion at the sphere boundary as part of the loss function. This integration of physical knowledge into the network architecture and training process enhances the accuracy of estimation. A major advantage of our approach is that it relies solely on measured data, eliminating the need for grid setup and simulated datasets. Simulation results demonstrate that our method outperforms the spherical harmonic method and the plane-wave decomposition method.

It is hard to derive theoretical solutions for the scattering sound of complex geometries, but PINN can use physical and geometric constraints to fit the sound pressure distribution. It allows our method to potentially estimate the sound field of arbitrary geometry.

References

[1] J. Breebaart and C. Faller, *Spatial audio processing: MPEG surround and other applications*. John Wiley & Sons, 2007.

[2] F. Rumsey, *Spatial audio*. Taylor & Francis, 2012.

[3] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.

[4] J. Zhang, T. D. Abhayapala, W. Zhang, P. N. Samarasinghe, and S. Jiang, "Active noise control over space: A wave domain approach," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 26, no. 4, pp. 774–786, 2018.

[5] F. Ma, W. Zhang, and T. D. Abhayapala, "Active control of outgoing broadband noise fields in rooms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 529–539, 2020.

[6] E. G. Williams and J. A. Mann III, *Fourier acoustics: Sound radiation and nearfield acoustical holography*, 2000.

[7] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2004.

[8] E. G. Williams and K. Takashima, "Vector intensity reconstructions in a volume surrounding a rigid spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 773–783, 2010.

[9] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 2, 2002, pp. II–1949.

[10] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.

[11] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, 2004.

[12] Z. Wei and X. Chen, "Deep-learning schemes for full-wave nonlinear inverse scattering problems," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 1849–1860, 2018.

[13] X. Chen, Z. Wei, M. Li, and P. Rocca, "A review of deep learning approaches for inverse scattering problems (invited review)," *Progress In Electromagnetics Research*, vol. 167, pp. 67–81, 2020.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] F. Lluis, P. Martinez-Nuevo, M. Bo Moller, and S. Ewan Shepstone, "Sound field reconstruction in rooms: Inpainting meets super-resolution," *The Journal of the*

Acoustical Society of America, vol. 148, no. 2, pp. 649–659, 2020.

[16] M. S. Kristoffersen, M. B. Møller, P. Martınez-Nuevo, and J. Østergaard, "Deep sound field reconstruction in real rooms: Introducing the isobel sound field dataset," *arXiv preprint arXiv:2102.06455*, 2021.

[17] K. Shigemi, S. Koyama, T. Nakamura, and H. Saruwatari, "Physics-informed convolutional neural network with bicubic spline interpolation for sound field estimation," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[18] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10561*, 2017.

[19] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[20] R. Liu and P. Gerstoft, "Sd-pinn: Physics informed neural networks for spatially dependent pdes," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[21] M. Olivieri, M. Pezzoli, F. Antonacci, and A. Sarti, "A physics-informed neural network approach for nearfield acoustic holography," *Sensors*, vol. 21, no. 23, p. 7834, 2021.

[22] N. Borrel-Jensen, A. P. Engsig-Karup, and C.-H. Jeong, "Physics-informed neural networks for one-dimensional sound field predictions with parameterized sources and impedance boundaries," *JASA Express Letters*, vol. 1, no. 12, 2021.

[23] F. Ma, T. D. Abhayapala, P. N. Samarasinghe, and X. Chen, "Physics informed neural network for head-related transfer function upsampling," *arXiv preprint arXiv:2307.14650*, 2023.

[24] F. Ma, T. D. Abhayapala, and P. N. Samarasinghe, "Circumvent spherical bessel function nulls for open sphere microphone arrays with physics informed neural network," *arXiv preprint arXiv:2308.00242*, 2023.

[25] F. Scarselli and A. C. Tsoi, "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results," *Neural networks*, vol. 11, no. 1, pp. 15–37, 1998.

[26] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3094–3103, 2005.

[27] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

A Controlled Noise Reduction Wiener Filter Based on the Quadratic Eigenvalue Problem

Ningning Pan*, Jacob Benesty†, and Jingdong Chen‡
*Research Institute for Digital Economy and Interdisciplinary Sciences,
Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China
‡Center of Intelligent Acoustics and Immersive Communications,
Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
†INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada

***Abstract*—This paper deals with the problem of multichannel noise reduction and a controlled Wiener filter is developed in the frequency domain by minimizing the subband mean-squared error between the clean speech of interest and its estimate subject to a constraint on the residual noise level. Using the Lagrange multiplier, we then transform the constrained minimization problem into one of quadratic eigenvalue problem. Different forms of the controlled Wiener filter are then derived. Depending on how the quadratic eigenvalues are used, these filters can control the amount of noise reduction as well as the compromise between the amount of noise reduction and the level of speech distortion. Simulation results justify the properties of the controlled Wiener filter.**

I. Introduction

Noise reduction is the process of recovering a clean speech signal of interest from microphone observations corrupted by additive noise [2]–[10]. Depending on the number of microphones, those methods can be divided into two classes, i.e., single-channel [11], [12], [13], [14], [15], [16] and multichannel methods [18], [19], [20].

Typical multichannel noise reduction approaches include the optimal filtering [18], [19], [20], the subspace based methods [21], [22], the statistical approach [23], [24], and the data driven based machine learning approaches [25], [26]. Optimal filtering based algorithm is one of the most straightforward ways to achieve noise reduction. Representative algorithms include the classical Wiener filter, maximum SNR filter, minimum variance distortionless response (MVDR) filter, linearly constrained minimum variance (LCMV) filter, etc, which are designed based on different criteria [17]. The classical Wiener filter is derived by minimizing the MSE between the desired clean speech and its estimate. Though widely used, the Wiener filter lacks flexibility to control the amount of noise reduction and speech distortion. One way to deal with the flexibility issue is by using the so-called tradeoff Wiener filter filter, which controls the compromise between noise reduction and speech distortion by a tunable parameter. The tunable parameter can be either set in a heuristic manner [17], or converted into a masking threshold, which can then be determined by the psychoacoustically motivated algorithm [27] or other methods as in [28].

In this paper, we present a controlled Wiener filter in the frequency domain. This controlled Wiener filter is developed by minimizing the subband MSE between the desired clean speech and its estimate subject to a constraint on the level of the residual noise. Using the Lagrange multiplier, we then transform the constrained minimization problem into one of quadratic eigenvalue problem. The controlled Wiener filters are then derived. The experiment results justify the flexibility of the deduced filters.

II. Signal Model, Problem Formulation, and Performance Measures

We consider a microphone array consisting of M microphones. In a general way, the received signals at the frequency index f are expressed as [1], [2]

$$\begin{aligned} Y_m(f) &= G_m(f)S(f) + V_m(f) \\ &= X_m(f) + V_m(f),\ m = 1, 2, \ldots, M, \end{aligned} \tag{1}$$

where $Y_m(f)$ is the mth microphone signal, $S(f)$ is the unknown speech source, $G_m(f)$ is the acoustic channel transfer function from the position of $S(f)$ to the mth microphone, $X_m(f) = G_m(f)S(f)$ is the convolved speech signal, and $V_m(f)$ is the additive noise. All signals are considered to be zero mean. It is assumed that $X_i(f)$ and $V_j(f)$ are incoherent, i.e., $E[X_i(f)V_j^*(f)] = 0,\ \forall i, j = 1, 2, \ldots, M$, where $E[\cdot]$ denotes mathematical expectation and the superscript * is the complex-conjugate operator. By definition, the terms $X_m(f),\ m = 1, 2, \ldots, M$ are coherent while the terms $V_m(f),\ m = 1, 2, \ldots, M$ are either partially coherent or incoherent. Without loss of generality, let us choose Microphone 1 as the reference. Then, the objective of multichannel noise reduction in the frequency domain is to estimate the desired signal, $X_1(f)$, from the M observations $Y_m(f),\ m = 1, 2, \ldots, M$, in the best way we can.

In a vector form, (1) is expressed as

$$\begin{aligned} \mathbf{y}(f) &= [Y_1(f)\ Y_2(f)\ \cdots\ Y_M(f)]^T \\ &= \mathbf{x}(f) + \mathbf{v}(f), \end{aligned} \tag{2}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{x}(f)$ and $\mathbf{v}(f)$ are defined similarly to $\mathbf{y}(f)$. It follows immediately that the covariance matrix of $\mathbf{y}(f)$ is

$$\mathbf{\Phi}_{\mathbf{y}}(f) = E\left[\mathbf{y}(f)\mathbf{y}^H(f)\right] = \mathbf{\Phi}_{\mathbf{x}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f), \tag{3}$$

where the superscript ^{H} is the conjugate-transpose operator, and $\mathbf{\Phi}_{\mathbf{x}}(f)$ and $\mathbf{\Phi}_{\mathbf{v}}(f)$ are the covariance matrices of $\mathbf{x}(f)$ and $\mathbf{v}(f)$, respectively.

In the frequency domain, multichannel noise reduction is generally performed by applying a complex-valued filter, $\mathbf{h}(f)$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of length M, to the observation signal vector, $\mathbf{y}(f)$, i.e.,

$$Z(f) = \mathbf{h}^H(f)\mathbf{y}(f) = X_{\rm fd}(f) + V_{\rm rn}(f), \tag{4}$$

where the filter output, $Z(f)$, is the estimate of $X_1(f)$, $X_{\rm fd}(f) = \mathbf{h}^H(f)\mathbf{x}(f)$ is the filtered desired signal, and $V_{\rm rn}(f) = \mathbf{h}^H(f)\mathbf{v}(f)$ is the residual noise. The objective of multichannel noise reduction is then to find an optimal filter, $\mathbf{h}(f)$, in such a way that $Z(f)$ is as close as possible to the desired signal, $X_1(f)$. According to (4), the variance of $Z(f)$ is written as

$$\phi_Z(f) = \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{y}}(f)\mathbf{h}(f) = \phi_{X_{\rm fd}}(f) + \phi_{V_{\rm rn}}(f), \tag{5}$$

where $\phi_{X_{\rm fd}}(f) = \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{x}}(f)\mathbf{h}(f)$ and $\phi_{V_{\rm rn}}(f) = \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f)$ are the variances of $X_{\rm fd}$ and $V_{\rm rn}$ respectively.

Given the signal model in (1), the input signal-to-noise ratio (SNR) is defined as

$$\text{iSNR}(f) = \frac{\phi_{X_1}(f)}{\phi_{V_1}(f)}, \tag{6}$$

where $\phi_{X_1}(f)$ and $\phi_{V_1}(f)$ are the variances of $X_1(f)$ and $V_1(f)$, respectively. Now, with the signal model in (1) and the multichannel noise reduction model in (4), we can define the following narrowband performance measures [2]:

- the noise reduction factor:

$$\xi_{\rm n}\left[\mathbf{h}(f)\right] = \frac{\phi_{V_1}(f)}{\phi_{V_{\rm rn}}(f)}, \tag{7}$$

- the speech reduction factor:

$$\xi_{\rm s}\left[\mathbf{h}(f)\right] = \frac{\phi_{X_1}(f)}{\phi_{X_{\rm fd}}(f)}, \tag{8}$$

- the output SNR:

$$\text{oSNR}\left[\mathbf{h}(f)\right] = \frac{\phi_{X_{\rm fd}}(f)}{\phi_{V_{\rm rn}}(f)}, \tag{9}$$

- and the speech distortion index:

$$\upsilon_{\rm sd}\left[\mathbf{h}(f)\right] = \frac{E\left\{\left|X_1(f) - X_{\rm fd}(f)\right|^2\right\}}{\phi_{X_1}(f)}. \tag{10}$$

It can be verified that

$$\frac{\text{oSNR}\left[\mathbf{h}(f)\right]}{\text{iSNR}(f)} = \frac{\xi_{\rm n}\left[\mathbf{h}(f)\right]}{\xi_{\rm s}\left[\mathbf{h}(f)\right]}. \tag{11}$$

III. Multichannel Wiener Filter

We define the error signal between the estimated and the desired signals at frequency f as

$$\begin{aligned}\mathcal{E}(f) &= Z(f) - X_1(f)\\ &= \mathbf{h}^H(f)\mathbf{y}(f) - X_1(f).\end{aligned} \tag{12}$$

The (narrowband) mean-squared error (MSE) criterion is then

$$\begin{aligned}J\left[\mathbf{h}(f)\right] &= E\left[\left|\mathcal{E}(f)\right|^2\right]\\ &= \phi_{X_1}(f) + \mathbf{h}^H(f)\mathbf{\Phi}_{\mathbf{y}}(f)\mathbf{h}(f) - \mathbf{h}^H(f)\boldsymbol{\phi}_{\mathbf{x}}(f)\\ &\quad - \boldsymbol{\phi}_{\mathbf{x}}^H(f)\mathbf{h}(f),\end{aligned} \tag{13}$$

where $\boldsymbol{\phi}_{\mathbf{x}}(f) = \mathbf{\Phi}_{\mathbf{x}}(f)\mathbf{i}$, with $\mathbf{i}$ being the first column of the $M \times M$ identity matrix, $\mathbf{I}_M$.

The minimization of $J\left[\mathbf{h}(f)\right]$ leads to the well-known multichannel Wiener filter:

$$\begin{aligned}\mathbf{h}_{\rm W}(f) &= \mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\boldsymbol{\phi}_{\mathbf{x}}(f)\\ &= \mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{\Phi}_{\mathbf{x}}(f)\mathbf{i}\\ &= \left[\mathbf{I}_M - \mathbf{\Phi}_{\mathbf{y}}^{-1}(f)\mathbf{\Phi}_{\mathbf{v}}(f)\right]\mathbf{i}.\end{aligned} \tag{14}$$

It can be shown that [2], [16]

$$\text{oSNR}\left[\mathbf{h}_{\rm W}(f)\right] \geq \text{iSNR}(f) \tag{15}$$

and

$$\xi_{\rm n}\left[\mathbf{h}_{\rm W}(f)\right] \geq 1. \tag{16}$$

So, noise reduction is possible with the Wiener filter. However, there is no way to control the amount of noise reduction (or, equivalently, the amount of speech distortion).

IV. Controlled Multichannel Wiener Filter

In this section, we drop the dependence on f for the sake of readability. Let us assume that we want to find a filter, $\mathbf{h}$, whose noise reduction factor is equal to $\xi_0 > 1$, i.e., $\xi_{\rm n}(\mathbf{h}) = \xi_0$, and, at the same time, minimizes the MSE. The controlled Wiener filter is obtained from the criterion:

$$\min_{\mathbf{h}} J\left(\mathbf{h}\right) \quad \text{subject to} \quad \mathbf{h}^H\mathbf{\Phi}_{\mathbf{v}}\mathbf{h} = \xi_0^{-1}\phi_{V_1}. \tag{17}$$

From the above criterion, we can write the Lagrange function:

$$\begin{aligned}\mathcal{L}\left(\mathbf{h},\lambda\right) =& \phi_{X_1} + \mathbf{h}^H\mathbf{\Phi}_{\mathbf{y}}\mathbf{h} - \mathbf{h}^H\boldsymbol{\phi}_{\mathbf{x}} - \boldsymbol{\phi}_{\mathbf{x}}^H\mathbf{h}\\ &- \lambda\left(\mathbf{h}^H\mathbf{\Phi}_{\mathbf{v}}\mathbf{h} - \xi_0^{-1}\phi_{V_1}\right),\end{aligned} \tag{18}$$

where λ is the Lagrange multiplier. Differentiating $\mathcal{L}\left(\mathbf{h},\lambda\right)$ with respect to $\mathbf{h}$ and λ yields the following equations:

$$\mathbf{\Phi}_{\mathbf{y}}\mathbf{h} - \lambda\mathbf{\Phi}_{\mathbf{v}}\mathbf{h} = \boldsymbol{\phi}_{\mathbf{x}}, \tag{19}$$

$$\mathbf{h}^H\mathbf{\Phi}_{\mathbf{v}}\mathbf{h} = \xi_0^{-1}\phi_{V_1}. \tag{20}$$

It can be shown [3], [4] that the smallest solution λ of the two previous equations is needed to solve (17).

Assume that λ is not an eigenvalue of $\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{y}}$, so that the matrix $\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}$ is invertible, and let

$$\begin{aligned}\overline{\mathbf{h}} &= \left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)^{-1}\mathbf{\Phi}_{\mathbf{v}}\left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)^{-1}\boldsymbol{\phi}_{\mathbf{x}}\\ &= \left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)^{-1}\mathbf{\Phi}_{\mathbf{v}}\mathbf{h}.\end{aligned} \tag{21}$$

We can rewrite (19) and (20) in terms of $\overline{\mathbf{h}}$:

$$\left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)\mathbf{\Phi}_{\mathbf{v}}^{-1}\left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)\overline{\mathbf{h}} = \boldsymbol{\phi}_{\mathbf{x}}, \tag{22}$$

$$\begin{aligned}\overline{\mathbf{h}}^H\left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)\mathbf{\Phi}_{\mathbf{v}}^{-1}\left(\mathbf{\Phi}_{\mathbf{y}} - \lambda\mathbf{\Phi}_{\mathbf{v}}\right)\overline{\mathbf{h}} &= \xi_0^{-1}\phi_{V_1}\\ &= \overline{\mathbf{h}}^H\boldsymbol{\phi}_{\mathbf{x}}.\end{aligned} \tag{23}$$

From (23), we can check that

$$\frac{\boldsymbol{\phi}_{\mathbf{x}}^H\overline{\mathbf{h}}}{\xi_0^{-1}\phi_{V_1}} = 1. \tag{24}$$

Developing (22) and exploiting the previous result, we get

$$\begin{aligned}\left(\mathbf{\Phi}_{\mathbf{y}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{y}} - 2\lambda\mathbf{\Phi}_{\mathbf{y}} + \lambda^2\mathbf{\Phi}_{\mathbf{v}}\right)\overline{\mathbf{h}} &= \boldsymbol{\phi}_{\mathbf{x}}\\ &= \frac{\boldsymbol{\phi}_{\mathbf{x}}\boldsymbol{\phi}_{\mathbf{x}}^H}{\xi_0^{-1}\phi_{V_1}}\overline{\mathbf{h}}.\end{aligned} \tag{25}$$

As a consequence, (25) can be rewritten as

$$\left[\lambda^2\mathbf{\Phi_v} - 2\lambda\mathbf{\Phi_y} + \left(\mathbf{\Phi_y}\mathbf{\Phi_v}^{-1}\mathbf{\Phi_y} - \frac{\boldsymbol{\phi}_\mathbf{x}\boldsymbol{\phi}_\mathbf{x}^H}{\xi_0^{-1}\phi_{V_1}}\right)\right]\overline{\mathbf{h}} = \mathbf{0} \quad (26)$$

or, in a more standard form:

$$\left(\lambda^2\mathbf{A}_2 + \lambda\mathbf{A}_1 + \mathbf{A}_0\right)\overline{\mathbf{h}} = \mathbf{Q}(\lambda)\overline{\mathbf{h}} = \mathbf{0}, \quad (27)$$

where

$$\mathbf{A}_2 = \mathbf{\Phi_v},$$
$$\mathbf{A}_1 = -2\mathbf{\Phi_y},$$
$$\mathbf{A}_0 = \mathbf{\Phi_y}\mathbf{\Phi_v}^{-1}\mathbf{\Phi_y} - \frac{\boldsymbol{\phi}_\mathbf{x}\boldsymbol{\phi}_\mathbf{x}^H}{\xi_0^{-1}\phi_{V_1}},$$

and $\mathbf{Q}(\lambda)$ is called the λ-matrix, which is an $M \times M$ matrix polynomial of degree 2. In (27), we recognize the quadratic eigenvalue problem (QEP) [5], where $\overline{\mathbf{h}}$ is the right eigenvector corresponding to the eigenvalue λ. Therefore, the solution of (17) is $\mathbf{h} = \left(\mathbf{\Phi_y} - \lambda\mathbf{\Phi_v}\right)^{-1}\boldsymbol{\phi}_\mathbf{x}$, where λ is the smallest eigenvalue of (27). Depending on the value of λ, we have the following cases for $\mathbf{h}$:

- $\lambda = 0$ gives the classical Wiener filter,
- $\lambda = 1$ leads to $\mathbf{h} = \mathbf{i}$, an all-pass filter which achieves no noise reduction and adds no speech distortion as well,
- $\lambda < 0$ gives more noise reduction at an expense of more speech distortion than the classical Wiener filter,
- $0 < \lambda < 1$ gives less noise reduction and less speech distortion than the classical Wiener filter, and
- $\lambda > 1$ gives less noise reduction and more speech distortion, which is the worst case.

The QEP is a generalization of the standard eigenvalue problem (SEP), $\mathbf{Au} = \lambda\mathbf{u}$, and the generalized eigenvalue problem (GEP), $\mathbf{Au} = \lambda\mathbf{Bu}$. With a λ-matrix of size $M \times M$, the QEP has $2M$ eigenvalues (finite or infinite) with up to $2M$ right and $2M$ left eigenvectors, and if there are more than M eigenvectors they are not, of course, linearly independent [5]. In our context, since $\mathbf{A}_2 = \mathbf{\Phi_v}$ is nonsingular, the QEP has $2M$ finite eigenvalues. Since $\mathbf{A}_0, \mathbf{A}_1$, and $\mathbf{A}_2$ are Hermitian, the eigenvalues are real or come in pairs (λ, λ^*) [5].

One important aspect of the QEP is the linearization process, which is not unique. We say that a $2M \times 2M$ matrix $\mathbf{A} - \lambda\mathbf{B}$ is a linearization of the $M \times M$ λ-matrix $\mathbf{Q}(\lambda)$ if the eigenvalues of these matrices coincide. Thanks to this process, we can solve the QEP with the GEP. One possibility is to take

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{A}_0 \\ \mathbf{A}_0 & \mathbf{A}_1 \end{bmatrix}, \quad (28)$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}_2 \end{bmatrix}. \quad (29)$$

It can be verified that if $\mathbf{t}$ is an eigenvector of $\mathbf{Q}(\lambda)$ then

$$\mathbf{u} = \begin{bmatrix} \mathbf{t} \\ \lambda\mathbf{t} \end{bmatrix} \quad (30)$$

is an eigenvector of $\mathbf{A} - \lambda\mathbf{B}$. This process corresponds to the symmetric linearization since both $\mathbf{A}$ and $\mathbf{B}$ are Hermitian. The accuracy of this approach will depend on the conditioning of $\mathbf{A}_0$ and $\mathbf{A}_2$. Another possibility is to take

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & -\mathbf{I}_M \\ \mathbf{A}_0 & \mathbf{A}_1 \end{bmatrix}, \quad (31)$$

$$\mathbf{B} = \begin{bmatrix} -\mathbf{I}_M & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}_2 \end{bmatrix}. \quad (32)$$

This linearization is not symmetric. Different linearizations may give different results, so it is important to choose the right one.

Let us order the eigenvalues of $\mathbf{Q}(\lambda)$:

$$\Re(\lambda_1) \leq \Re(\lambda_2) \leq \cdots \leq \Re(\lambda_{2M}), \quad (33)$$

where $\Re(\cdot)$ is the real part of a complex number. Then, the controlled multichannel Wiener filter that will give exactly a noise reduction factor equal to ξ_0 is

$$\mathbf{h}_{\mathrm{CW}} = \left[\mathbf{\Phi_y} - \Re(\lambda_1)\mathbf{\Phi_v}\right]^{-1}\boldsymbol{\phi}_\mathbf{x}. \quad (34)$$

Now, if we consider the eigenvalues with the P smallest real parts, we can get another version of the controlled Wiener filter:

$$\mathbf{h}_{\mathrm{CW},P} = \left(\mathbf{\Phi_y} - \overline{\lambda}_P\mathbf{\Phi_v}\right)^{-1}\boldsymbol{\phi}_\mathbf{x}, \quad (35)$$

where

$$\overline{\lambda}_P = \frac{1}{P}\sum_{p=1}^{P}\Re(\lambda_p). \quad (36)$$

In general, $\xi_{\mathrm{n}}\left(\mathbf{h}_{\mathrm{CW},P}\right) \neq \xi_0$, but $\mathbf{h}_{\mathrm{CW},P}$ may lead to less speech distortion than $\mathbf{h}_{\mathrm{CW}}$ in practice.

An alternative approach is to use the distortion-based MSE criterion instead, which is given by

$$J_{\mathrm{d}}\left(\mathbf{h}\right) = E\left(\left|\mathbf{h}^H\mathbf{x} - X_1\right|^2\right). \quad (37)$$

The controlled Wiener filter is derived from the criterion:

$$\min_{\mathbf{h}} J_{\mathrm{d}}\left(\mathbf{h}\right) \quad \text{subject to} \quad \mathbf{h}^H\mathbf{\Phi_v}\mathbf{h} = \xi_0^{-1}\phi_{V_1}. \quad (38)$$

Following the same steps as above, the obtained QEP is

$$\left[\lambda^2\mathbf{\Phi_v} - 2\lambda\mathbf{\Phi_x} + \left(\mathbf{\Phi_x}\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x} - \frac{\boldsymbol{\phi}_\mathbf{x}\boldsymbol{\phi}_\mathbf{x}^H}{\xi_0^{-1}\phi_{V_1}}\right)\right]\overline{\mathbf{h}} = \mathbf{0}. \quad (39)$$

Now, the solution of (38) is $\mathbf{h} = \left(\mathbf{\Phi_x} - \lambda\mathbf{\Phi_v}\right)^{-1}\boldsymbol{\phi}_\mathbf{x}$, where λ is the smallest eigenvalue of (39).

V. Simulations

In this section, we study the performance of the developed controlled multichannel Wiener filter through simulations. We consider a reverberant enclosure of size 3 m × 3 m × 4 m [the coordinates are denoted as (x, y, z)], where a uniform linear array with 8 microphones is located along the x axis with an interelement spacing of 0.02 m, i.e., the position of sensors are (x, 2.5 m, 1.0 m) with $x = 1.64 : -0.02 : 1.50$. A source (the target) is located at (2.5 m, 1.0 m, 2.5 m), and an interference source is located at (1.0 m, 2.5 m, 0.5 m). The impulse responses from the source signal to the microphones (as well as that from the interference to the microphones) are generated with the image model method [6] with all the wall reflection coefficients being 0.6 (the corresponding reverberation time

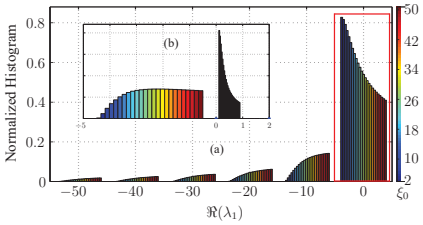


Fig. 1. The normalized histogram of $\Re(\lambda_1)$ as a function of ξ_0, where $\xi_0 = 2:2:50$: (a) the rough one and (b) details of the red box.

T_{60} is 140 ms approximately). The target source signal consists of 20 sentences extracted from the TIMIT database [7], with 10 from a male speaker and the other 10 from a female speaker. The interference is a babble noise taken from the NoiseX92 database [8]. Furthermore, a computer generated multichannel Gaussian noise is added with a specified interference-to-noise ratio (INR). The overall noise at the mth microphone is in the form of $V_m(f) = I_m(f) + \beta W_m(f)$, where $I_m(f)$ is the interference, $W_m(f)$ is the Gaussian white noise and β is a constant controlling INR. In this work, the value of β is adjusted so that the INR is 5 dB. Thus, the noise at different sensors are partially coherent. Then, $V_m(f)$ is properly scaled and added to the clean speech signal at an iSNR of 10 dB. To implement the algorithm, we choose a frame size of 32 ms with 75% overlap and a Kaiser window is used for data framing. The correlation matrix of $\mathbf{y}(f)$ and $\mathbf{v}(f)$ are computed recursively, i.e.,

$$\mathbf{\Phi}_{\mathbf{y}}(f,k) = \alpha_{\mathbf{y}}\mathbf{\Phi}_{\mathbf{y}}(f,k-1) + (1-\alpha_{\mathbf{y}})\mathbf{y}(f,k)\mathbf{y}^H(f,k), \tag{40}$$

$$\mathbf{\Phi}_{\mathbf{v}}(f,k) = \alpha_{\mathbf{v}}\mathbf{\Phi}_{\mathbf{v}}(f,k-1) + (1-\alpha_{\mathbf{v}})\mathbf{v}(f,k)\mathbf{v}^H(f,k), \tag{41}$$

where k is the frame index, and $\alpha_{\mathbf{y}}$ and $\alpha_{\mathbf{v}}$ are the forgetting factors, which are both set to 0.95 in this work.

As seen in (34) and (35), the values of $\Re(\lambda_1)$ and $\overline{\lambda}_P$ play an important role on the performances of the filters $\mathbf{h}_{\mathrm{CW}}$ and $\mathbf{h}_{\mathrm{CW},P}$. So, let us first examine the values of $\Re(\lambda_1)$ as a function of ξ_0. We set $\xi_0 = 2:2:50$. Figure 1 plots the histogram of $\Re(\lambda_1)$ and Fig. 1(b) is a zoomed part with $\Re(\lambda_1) \in [-5, 2]$. As discussed in Section IV, the controlled multichannel Wiener filter, $\mathbf{h}_{\mathrm{CW}}$, degenerates to the classical Wiener filter if $\Re(\lambda_1) = 0$ while it degenerates to an all-pass filter if $\Re(\lambda_1) = 0$. One can see from Fig. 1 that more than 80% values of $\Re(\lambda_1)$ lie in the range of $[0, 1]$ when $\xi_0 = 2$. However, as the value of ξ_0 increases, $\Re(\lambda_1)$ decreases towards negative values. This indicates that with ξ_0 increasing, $\mathbf{h}_{\mathrm{CW}}$ could achieve more noise reduction; but of course at the expense of more speech distortion as compared to the classical Wiener filter. Moreover, one can infer that as the value of P increases, the value of $\overline{\lambda}_P$ increases, in which case $\mathbf{h}_{\mathrm{CW},P}$ would achieve less noise reduction with less speech distortion as compared to $\mathbf{h}_{\mathrm{CW}}$.

Now, in the second simulation, we evaluate the performance of different noise reduction filters as a function of ξ_0. Again, we set $\xi_0 = 2:2:50$. Three performance measures are used: the output SNR, the speech distortion index, and the perceptual evaluation of speech quality (PESQ) score, which are defined, respectively in (9), (10), and [9]. Figure 2 plots the results

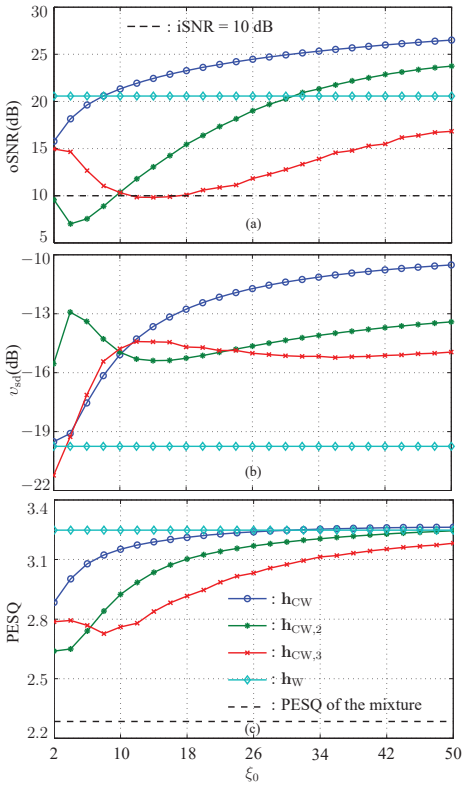


Fig. 2. Performance of $\mathbf{h}_{\mathrm{CW}}$, $\mathbf{h}_{\mathrm{CW},2}$, $\mathbf{h}_{\mathrm{CW},3}$, and the classical Wiener filter, $\mathbf{h}_{\mathrm{W}}$, as a function of the noise reduction factor, ξ_0. The input SNR is 10 dB and the PESQ score of the noisy signal is 2.28.

of the controlled Wiener filter, i.e., $\mathbf{h}_{\mathrm{CW}}$ in (34), $\mathbf{h}_{\mathrm{CW},P}$ in (35) with $P = 2$ and $P = 3$, as well as the classical Wiener filter in (14). The input SNR is 10 dB and the PESQ score of the noisy signal is 2.28. One can see from Fig. 2 that the oSNR, v_{sd}, and PESQ of the controlled Wiener filter, $\mathbf{h}_{\mathrm{CW}}$, $\mathbf{h}_{\mathrm{CW},2}$, and $\mathbf{h}_{\mathrm{CW},3}$, all increase with the increasing of ξ_0. The reason for this is that the value of $\overline{\lambda}_P$ decreases as the value of ξ_0 increases (as seen in Fig. 1), which leads to more noise reduction and more speech distortion. So, unlike the classical Wiener filter, the controlled noise reduction Wiener filter can make a compromise between the amount of noise reduction and the level of speech distortion by adjusting the value of ξ_0. Generally, as the value of P increases, the amount of speech distortion with $\mathbf{h}_{\mathrm{CW},P}$ decreases, but at the cost of less noise reduction as compared to the classical Wiener filter.

VI. Conclusions

In this paper, we proposed a controlled Wiener filter for multichannel noise reduction in the frequency domain. The filter is derived by minimizing the MSE criterion with a constraint on the amount of noise reduction. Using the Lagrange multiplier, we then transformed the constrained minimization problem into one of quadratic eigenvalue problem. The optimal and sub-optimal controlled Wiener filters were subsequently derived. Simulation results show that the proposed filters are able to control the amount of noise reduction as well as the tradeoff between noise reduction and speech distortion by adjusting a parameter that is equal to the average of the eigenvalues with smallest real parts.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang *Microphone Array Signal Processing.* Berlin, Germany: Springer-Verlag, 2008.

[2] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing.* Berlin, Germany: Springer-Verlag, 2009.

[3] W. Gander, "Least squares with a quadratic constraint," *Numer. Math*, vol. 36, pp. 291-307, 1981

[4] W. Gander, G. H. Golub, and U. von Matt, "A constrained eigenvalue problem," *Linear Algebra Appl.*, vol. 114/115, pp. 815-839, 1989

[5] F. Tisseur, and K. Meerbergen, "The quadratic eigenvalue problem," *SIAM Review*, vol. 43, no.2, pp. 235-286, 2001

[6] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993 [Online]. Available at: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[8] A. Varga and H. J. M. Steeneken, "Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–253, July 1993.

[9] *Mapping function for transforming raw result scores to MOS-LQO*, ITU-T Rec. P.862.1, 2003.

[10] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton Florida: CRC Press, 2007.

[11] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook on Speech Processing and Speech Communication*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., pp. 843–871. Berlin, Germany: Springer-Verlag, 2008.

[12] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech Enhancement: A Signal Subspace Perspective*, Academic Press, 2014.

[13] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement and noise statistics estimation for non-stationary voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 645–657, Apr. 2016.

[14] M. K. Becker and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 12, pp. 2251–2262, Dec. 2016.

[15] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, " Subjective intelligibility of deep neural network-based speech enhancement," in *Proc. Interspeech 2017*, Stockholm, Sweden, Aug. 2017, pp.2949–2955.

[16] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, pp. 1218–1234, July 2006.

[17] J. Benesty, J. Chen, and E. Habets, *Speech enhancement in the STFT domain.* Berlin, Germany: Springer-Verlag, 2011.

[18] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 16, pp. 957–968, July 2008.

[19] S. Doclo , A. Spriet , J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, pp. 636–656, July 2007.

[20] R. Serizel , M. Moonen , B. V. Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, pp. 785–798, Apr. 2014.

[21] R. Tong, G. Bao, and Z. Ye, "A higher order subspace algorithm for multichannel speech enhancement," *IEEE Signal Process. Lett.*, vol. 22, pp. 2004–2008, Nov. 2015.

[22] L. P. Yesenia, E. Habets, and J. Chen, "Multichannel noise reduction in the Karhunen-Loeve expansion domain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, pp. 923–936, May 2014.

[23] J. Tu and Y. Xia, "Fast distributed multichannel speech enhancement using novel frequency domain estimators of magnitude-squared spectrum," *Speech Commun.*, vol. 72, pp. 96–108, Sept. 2015.

[24] S. Kim and H. Kim , "Hybrid probabilistic adaptation mode controller for generalized sidelobe cancellers applied to multi-microphone speech enhancement," *Digit. Signal Process.*, vol. 25, pp. 123–133, Feb. 2014.

[25] Z. Wang and D. Wang, " Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 457–468, Feb. 2019.

[26] Z. Zhuo, T. Yoshioka, X. Xiao, L. Li, M. Seltzer, and Y. Gong, " Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,*, pp. 5384–5388, Apr. 2018.

[27] J. Rosca, R. Balan, and C. Beaugeant, " Multi-channel psycho acoustically motivated speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,*, vol. 3, pp. 217–220, May. 2003.

[28] J. Tu, Y. Xia, and S. Zhang, "A complex-valued multichannel speech enhancement learning algorithm for optimal tradeoff between noise reduction and speech distortion," *Neurocomputing*, vol. 267, pp. 333–343, June 2017.

Target Speaker Extraction with Attention Enhancement and Gated Fusion Mechanism

Wang Sijie*† , Askar Hamdulla*† , Mijit Ablimit*†
* School of Information Science and Engineering, Xinjiang University, Urumqi, China
† Key Laboratory of Signal Detection and Processing, Xinjiang, Urumqi, China
E-mail: himawari@stu.xju.edu.cn, askar@xju.edu.cn

***Abstract*—The objective of a target speaker extraction system is to extract the speech of the target speaker from a mixture of multiple speakers and noises using a certain amount of additional information of the target speaker. In this paper, we investigate the improvements of the baseline system by incorporating the light-weight CBAM module in the target extractor, and the gated fusion module (GFM) in the fusion layer. The CBAM introduces attention enhancement to baseline model with no significant increase in the number of parameters and complexity, and the previous concatenation-based fusion method used for speaker embedding and input mixture (or intermediate output) is replaced by GFM, enabling the model to better leverage the supplementary information provided by speaker embedding. Experimental results on datasets built from WSJ0-2mix and WHAM! demonstrate that both the CBAM module and the light-weight GFM module individually improve the model performance, and the GFM module shows better improvement on WHAM!. However, the combination of these two modules only exhibits mutually beneficial effects on the clean dataset WSJ0-2mix, while the performance of the combined module on the noisy dataset WHAM! is inferior to that of using the GFM module alone.**

***Index Terms*—target speaker extraction, attention, gated fusion, multi-task learning.**

I. Introduction

Target speaker extraction is a necessary and vital front-end task that has significant practical implications for voice related tasks, such as speech understanding and speaker verification [1]. It addresses challenging scenarios, including the cocktail party problem [2] and typical indoor environments with multiple speakers and non-speech noise interferences, such as electrical appliances and object collisions. Humans (animals alike) possess certain attentive biological mechanism which can employ selective attention to focus on specific audio sources. This filtering ability can be simulated with source separation techniques. However, the blind source separation (BSS) task and its specific applications like target speaker extraction (TSE) are difficult challenge for machines.

There are two approaches to addressing the target speaker extraction problem. The first approach involves utilizing speech separation (SS) techniques to separate the speech signals of all speakers and subsequently identifying the speech corresponding to the target speaker [3], [4]. The second approach, proposed by subsequent researchers, is the target speaker extraction (TSE) method which is developed based on speech separation. The research on target speaker extraction can be categorized into two stages. The first stage involves traditional signal processing methods, primarily employing microphone arrays, sound source localization, and other techniques to achieve the objective of target speaker extraction, these methods often require multiple microphones or prior knowledge of the source positions, and their separation performance may not be satisfactory when dealing with complex mixed speech signals. In the second stage, deep learning-based methods come into play, these methods primarily employ neural networks, speaker embedding features, attention mechanisms, and other techniques. Unlike the previous stage, these methods typically operate with a single channel and do not rely on prior knowledge of the source positions, achieving significant improvements in separation performance, especially when confronted with complex mixed speech signals.

Earlier methods for speaker extraction mainly focused on frequency domain, examples include SpeakerBeam [5]–[7] and Voice-Filter [8], etc. Similar to speech separation techniques, these frequency domain approaches suffer from the issue of inaccurate phase estimation during time domain signal reconstruction. To overcome this limitation, Luo et al. introduced Conv-TasNet [9], a method for speaker separation in the time domain, and some researchers use this architecture to serve for back-end tasks [3]. Building upon this idea, Spex [10] applied Conv-TasNet to speaker extraction, leading to significant improvements compared to the optimal baseline. Additionally, Spex+ [11] introduced twin speech encoders, where the same encoder is used to simultaneously encode the mixture and reference audio, this innovative approach further enhances the performance of speaker extraction.

In recent studies focusing on target speaker extraction, the typical module consists of two main components: embedding extractor and speech extractor. The embedding extractor is responsible for deriving an embedding that captures the characteristics of the target speaker from a pre-registered reference audio. This module can be trained independently or jointly with the speech extraction module, ensuring that resulting embedding is optimized for the extraction task. The speech extractor module comprises four key components: speech encoder, fusion layer, target extractor, and speech decoder. Speech encoder transforms the audio signal from the time domain into a latent space representation, fusion layer aims to integrate the speaker embedding with the mixture representation, target extractor serves as a masking network to obtain

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the latent space representation of the target speech, lastly, speech decoder reconstructs the representation back into the time domain.

Regarding the key module of the target extractor mentioned above, its core components have been evolving, primarily benefiting from continuous improvements and refinements in models related to the blind source separation (BSS) problem. The Spex series models utilize the TCN block from Conv-TasNet [10], [11] and propose the use of DPRNN as the core of the masking net, which has shown promising results. Additionally, transformer and conformer models have been applied to target speaker extraction tasks [12]. Attention mechanisms have been widely explored in related studies on target speaker extraction with two main aspects being investigated. The first aspect focuses on the use of attention mechanisms when combining embeddings with the mixture, for instance, there is a study introduces a dynamic attention mechanism that adjusts the embedding for each frame based on the acoustic similarity between the anchor speech and the mixture, leading to favorable results [13], [14]. Furthermore, multimodal research incorporates attention mechanisms to combine features from different modalities [15]. The second aspect concentrates on the use of attention mechanisms in the target extractor module. The ExARN model combines self-attention and recurrent neural networks (RNNs) to learn speaker embeddings [16], Av-Sepformer introduces Sepformer into the multimodal domain [17]. Exformer employs a separator network based on transformer encoder blocks, achieving superior speaker extraction performance compared to other time-domain models [18].

It is evident that relevant clues about the target speaker are crucial for selecting the target speech from a mixture. Commonly utilized clues include audio clues, visual clues, and spatial clues. In this paper, our focus is specifically on audio clues, and extensive research has been conducted on extracting target speakers based on audio clues. Typical audio clue embedding types include I-vectors [19], x-vectors [20], d-vectors [21] or even jointly-learned embedding [5], [10].

Various methods have been explored to fuse embeddings and speech encoder outputs. These methods include concatenation [22], [23], addition, multiplication [5], as well as more complex approaches such as combining addition and multiplication, and utilizing factorized layers which involve combining different mixture embedding weighted by a clue factor. Furthermore, some studies have investigated fusion styles based on gating mechanisms [24] or attention mechanisms. Another possibility is to use multiple clues and perform embedding fusion at different positions within the model.

This paper builds upon the widely recognized Spex+ model in the field of target speaker extraction. Specifically, we enhance the target extractor module by introducing a lightweight CBAM module after each TCN stack, thereby improving the representation of TCN output with no significant increase in the number of parameters and complexity. To fuse the embedding and encoded mixture (or intermediate result), we adopt a fusion mechanism based on gate units. The proposed improvement methods is evaluated on speaker extraction datasets constructed from WSJ0-2mix and WHAM!. Experimental results demonstrate a significant enhancement compared to the baseline TCN-based model.

II. Proposed Method

In this section, we introduce the overall framework of the target speaker extraction model adopted in this paper and propose improvement methods for the target extractor and embedding fusion layer. To address the issue that the target extractor relys solely on TCN blocks, we enhance the output of individual TCN stacks by incorporating the channel block attention mechanism (CBAM) with channel attention and spatial attention. For the fusion of speaker embedding and encoded mixture, we employ a gating mechanism, which allows the network to selectively utilize the auxiliary information in the speaker embedding. This enhances the focus on important information, improves the expressiveness and generalization capacity of the entire target speaker extraction model.

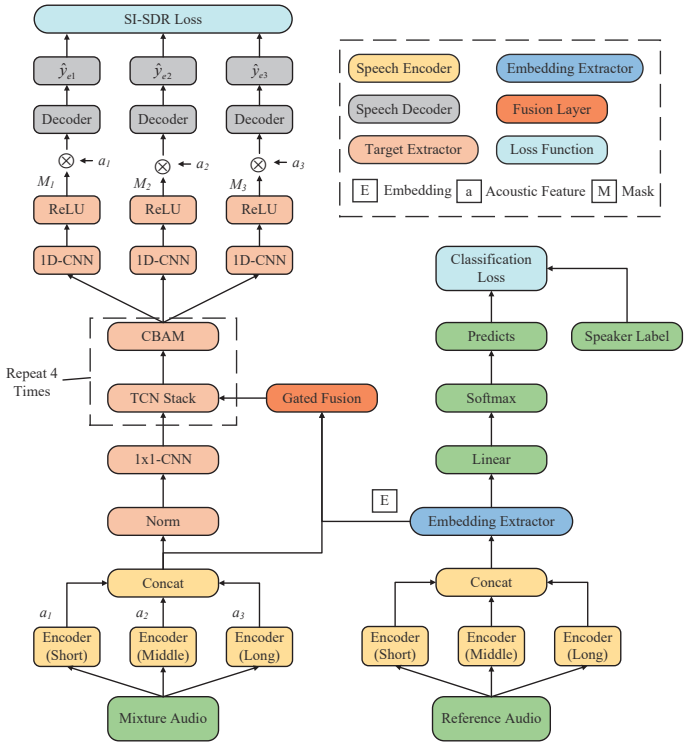


Fig. 1: The overall architecture of the proposed TSE system

A. The Overall Framework of the Proposed TSE Module

The overall framework used in this study, as shown in Fig. 1, is similar to the typical representative Spex+ model which consists of an embedding extractor and a speech extractor. After extracting the embedding, a speaker classification task is performed to enhance the discriminative features of the extracted embedding for different speakers. The speech extractor can be further divided into speech encoder, fusion layer, target

extractor, and speech decoder. In this study, we investigate the impact of attention and gated fusion mechanisms on model performance by incorporating CBAM and GFM in the target extractor and fusion layer, respectively.

First, we utilize the embedding extractor module to obtain an embedding vector $e \in \mathcal{R}^{N_e}$ that corresponds to the target speaker based on the reference audio $x_r \in \mathcal{R}^{N_r}$, this process can be represented by Equation (1):

$$e = f_{emb}\left(x_r; \theta_{emb}\right) \tag{1}$$

The speech encoder encodes the input mixture $x_m \in \mathcal{R}^{N_m}$ to obtain acoustic features a:

$$a = f_{enc}\left(x_m; \theta_{enc}\right) \tag{2}$$

The latent representation of the estimated target sound $\hat{y}_e$ can be derived based on the obtained embedding e and acoustic features a mentioned above:

$$\hat{y}_e = f_{tar}\left(a, e; \theta_{tar}\right) \tag{3}$$

The obtained results from the previous step can be passed through the speech decoder to obtain the final multi-scale time-domain representation of the target speech $\hat{y}_{el}$:

$$\hat{y}_{el} = f_{dec}\left(\hat{y}_e; \theta_{dec}\right)\left(l = 1, 2, 3\right) \tag{4}$$

where N_r, N_m and N_e are the dimensions of the reference audio, the mixture audio, and the embedding vector respectively. The f and θ refer to corresponding components and parameters.

B. Target Extractor with CBAM Attention Enhancement

Fig. 1 also illustrates the architecture of the target extractor incorporating the CBAM module for target speaker extraction. The fundamental concept behind this module is to employ the CBAM module to process the intermediate output of the TCN block. The Convolutional Block Attention Module (CBAM) is an light-weight attention mechanism that combines spatial and channel attention. As depicted in Fig. 2, the CBAM module sequentially generates attention feature maps in both channel and spatial dimensions. These two types of feature maps are then multiplied with the original input feature map for adaptive feature refinement, resulting in the final feature map.

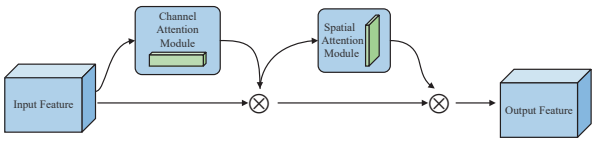


Fig. 2: Typical architecture of the CBAM module

In this study, we propose integrating the CBAM module after each TCN stack. This decision is motivated by the fact that the TCN module itself utilizes causal convolutions, dilated convolutions, and residual connections to extract features at different levels and expand the receptive field with attention mechanism absent. To address this limitation, we leverage the CBAM module's channel attention block to process the intermediate feature maps generated by the TCN stack. By calculating attention weights for each channel in a sequential manner, we can adaptively refine the feature representation by multiplying the weights with the input feature map. Furthermore, the CBAM module incorporates a spatial attention mechanism that operates on the intermediate feature maps. This mechanism computes attention weights for each spatial position, effectively emphasizing the importance of different positions and enhancing the expressiveness and discriminability of the features. It is important to note that in the context of the TCN stack output, the spatial attention mechanism is specifically applied to temporal chunks. This consideration is necessary due to the arbitrary length of time series compared to two-dimensional images. Therefore, our CBAM module operates solely along the chunk size dimension. In summary, the channel attention mechanism of CBAM allows for adjusting the importance of individual channels, while the spatial attention mechanism adjusts the importance of spatial positions. This combined attention mechanism facilitates adaptive feature optimization and improves the overall performance of the model with no significant increase in the number of parameters and complexity.

C. Gated Fusion Layer

The gated mechanism is a widely used technique in deep learning that selectively transmits or blocks information based on the relationship between input and output. It draws inspiration from the gate control theory of pain, which suggests that non-nociceptive inputs can close the "gate" for nociceptive inputs, thereby preventing pain signals from reaching the central nervous system. This technique has been extensively applied in recurrent neural networks such as LSTM and GRU. In addition, gated fusion layers are employed to selectively merge features from different modalities or sources by assigning appropriate weights to different features. This enables the transmission of more important features to deeper layers of the network.

As mentioned earlier, the gated unit was initially employed in network architectures such as LSTM and GRU [25], [26]. In 2017, Dauphin et al. introduced the Gated Linear Unit (GLU), a feedforward neural network that calculates the element-wise product of two linear projections, with one of the projections being processed by a sigmoid function. GLU has been widely adopted in natural language processing architectures, including the gated CNN [27]. The specific formula for GLU can be described as follows:

$$\text{SiGLU}\left(x, W, V, b, c\right) = \text{Sigmoid}\left(xW + b\right) \otimes \left(xV + c\right) \tag{5}$$

Later some researchers introduced several variations of GLU, including ReGLU, SwiGLU, and GEGLU, which employ different nonlinear (or even linear) functions instead of the sigmoid function. Through their experiments, they demonstrated that these GLU variants can significantly enhance

the performance of transformer models across various natural language understanding tasks [28]. Here are a few examples of the expressions utilized:

$$\begin{aligned} \text{ReGLU} &= \max(0, xW + b) \otimes (xV + c) \\ \text{GEGLU} &= \text{GELU}(xW + b) \otimes (xV + c) \\ \text{SwiGLU} &= \text{Swish}(xW + b) \otimes (xV + c) \end{aligned} \tag{6}$$

In this study, we utilize the idea of GEGLU for merging speaker embedding with the input mixture (or intermediate output). Our goal is to extract the target speech from the input mixture by considering it as the primary source while leveraging the auxiliary information provided by the target speaker embedding. To achieve this, we incorporate the input mixture (or intermediate output) as the "x" term and include a linear transformation of the speaker embedding as part of the bias term. The speaker embedding term remains fixed and non-trainable, to enhance the adaptability of the bias term to different target speakers, we choose to preserve the bias term of the convolutional operation itself. Additionally, we introduce a global layer normalization (GLN) module before the GELU activation function, this step aims to improve the data distribution, enhance the network's representation and learning capabilities, so as to enhance the overall generalization capacity and noise robustness of the model. The inclusion of the GLN layer also ensures consistency between the model's training and testing phases. Fig. 3 depicts a specific diagram illustrating these concepts, and the corresponding mathematical expressions are provided below:

$$\begin{aligned} \hat{a} = \widetilde{a} &+ (\widetilde{a} * W + b_1) \\ &\otimes \text{GELU}(\text{GLN}(\widetilde{a} * V + (b_2 + Ge))) \end{aligned} \tag{7}$$

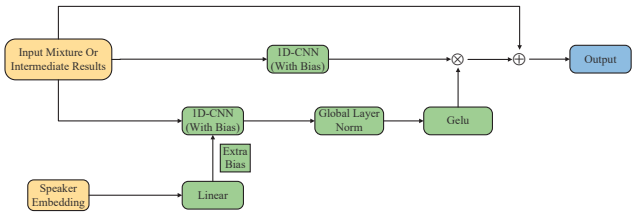


Fig. 3: Proposed gated fusion method

where $\widetilde{a}$ represents the encoded mixture (or intermediate output) of the model, W and V correspond to convolutional operations, the bias terms (b_1 and b_2) of the two convolutional operations are both retained, e represents the embedding vector extracted from the reference audio by the embedding extractor, G denotes the linear transformation applied to the embedding, GLN represents global layer normalization, GELU represents the non-linear activation function, the symbol $*$ represents the convolutional operation, and the symbol $\otimes$ denotes the Hadamard product.

The specific process of their combination is as follows: firstly, the encoded mixture $\widetilde{a}$ undergoes the first convolutional operation with a bias term of b_1, and the resulting output is denoted as c_1. Subsequently, $\widetilde{a}$ is processed by the second convolutional operation with a bias term of $b_2 + Ge$, the output of the second convolutional operation is sequentially passed through the global layer normalization and GELU layers, resulting in c_2. The Hadamard product is then computed between c_1 and c_2, and the resulting output is added to $\widetilde{a}$ using residual addition, this combined result is then passed as input to the subsequent layer of the network.

III. Experiments and Results

In this section, we focus on the experimental details, including the datasets used for training, validation, and testing. We also discuss the relevant hyperparameters of the model and explain the training process of the proposed target speech extraction system.

A. Dataset

We obtain our datasets from the widely used WSJ0-2mix dataset designed for speech separation task, as well as its noise-corrupted counterpart, the WHAM! dataset. To adapt these datasets for target speaker extraction, we perform additional preprocessing. Specifically, we assign the first speaker as the target speaker and the second speaker as the interference speaker. Moreover, we randomly select a distinct speech segment, which is different from the corresponding component in the mixture, as the reference audio for the target speaker. The training set consists of 20,000 samples, while the validation and testing sets comprise 5,000 and 3,000 samples, respectively.

B. Experimental setup

For all models, the training is conducted on 2-second segments for a total of 100 epochs. The initial learning rate is set to 1e-3, if the validation performance don't improve for two consecutive epochs, the learning rate is reduced by a factor of 0.5, in case of no improvement in the validation set for six consecutive epochs, early stopping is employed, and the Adam optimizer is utilized for optimization. The sampling rate for both mixtures and reference audio is set to 8kHz. The convolutional layers in the multi-scale speech encoder and decoder have filter lengths of L1=2.5ms, L2=10ms, and L3=20ms. The target extractor consists of 4 TCN stacks, in the benchmark model Spex+, each TCN stack has 8 TCN layers, for models that use lightweight attention module, each TCN stack is transformed into 7 TCN layers + 1 light-weight attention module. A gated fusion layer is introduced in the first layer of each TCN stack to integrate the speaker embedding with the input mixture (or intermediate output). The auxiliary embedding extractor network has the same structure as the Spex+ model, featuring three ResNet blocks and a speaker embedding dimension of 256. For multi-task training, the parameters of the encoders for reference audio and mixture are shared, with the scaling factors of SI-SDR loss and cross-entropy loss ratio set to 1, 10, respectively, the actual loss function is defined as follows:

$$L_{\text{total}} = L_{\text{SI - SDR}} + 10L_{\text{CE}} \tag{8}$$

where L_{total} represents the overall loss, $L_{\text{SI-SDR}}$ denotes the reconstruction loss for the target speaker's speech, and L_{CE} represents the speaker classification loss. The remaining hyperparameters involved in the loss function align with those of the Spex+ model.

C. Evaluation and results

TABLE I: Comparison of the number of parameters for each model

Model No.	Number of Parameters(Million)
1	11.41
2	10.1
3	12.45
4	11.42

TABLE II: Experimental results of various models on WSJ0-2mix dataset

Model No.	Methods	SI-SDR(dB)	PESQ
1	(Spex+)(Baseline)	15.39	3.16
2	(Spex+) + CBAM	15.77	3.23
3	(Spex+) + Gated Fusion	16.43	3.30
4	(Spex+) + CBAM + Gated Fusion	16.95	3.42

TABLE III: Experimental results of various models on WHAM! dataset

Model No.	Methods	SI-SDR(dB)	PESQ
1	(Spex+)(Baseline)	8.34	2.07
2	(Spex+) + CBAM	8.74	2.14
3	(Spex+) + Gated Fusion	9.75	2.26
4	(Spex+) + CBAM + Gated Fusion	9.36	2.20

We compared the baseline model with improved models using two evaluation metrics, SI-SDR and PESQ, Table I compares the number of parameters for each model, note that in the benchmark model Spex+, each TCN stack has 8 TCN layers, for models that use lightweight attention module CBAM, each TCN stack is transformed into 7 TCN layers + 1 light-weight attention module.

Table II presents the results of each model on the WSJ0-2mix test set, the histogram version is on the left side of Fig. 4 and Fig. 5. Model 1, Model 2, Model 3, and Model 4 represent Spex+ (baseline), the improved model incorporating the CBAM module, the improved model incorporating the GFM, and the improved model combining both modules, respectively. It can be observed that the number of parameters and complexity of the improved models don't significantly increase, and in

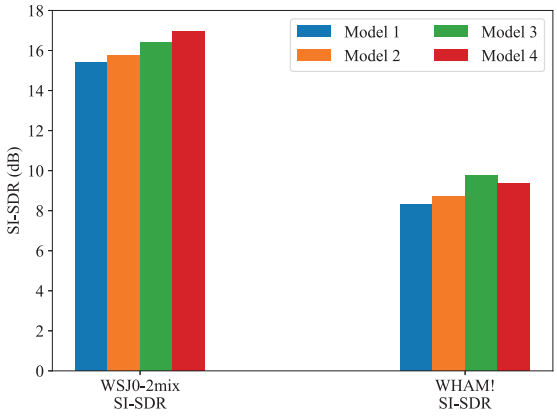


Fig. 4: Comparison of experiment results for SI-SDR

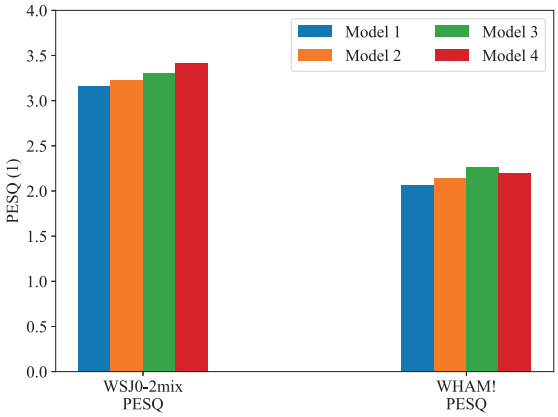


Fig. 5: Comparison of experiment results for PESQ

terms of SI-SDR, the addition of the CBAM module in the target extractor provides a gain of 0.38 dB compared to the baseline Spex+, this improvement is mainly attributed to the channel attention and spatial attention mechanisms introduced by the CBAM module, which are not present in the original model. Regarding the fusion of speaker embedding and input mixture (or intermediate result), the proposed GEGLU-based method brings a significant gain of 1.04 dB. During the training process, the performance gap between the validation and test sets of Model 3 is notably reduced compared to Model 1 and Model 2. This indicates that the gated fusion method used for speaker embedding and input mixture (or intermediate result) effectively selects more generalizable and important features from the embedding for fusion, resulting in improved generalization performance of the TSE model. Furthermore, Model 3 relative to Model 1 and Model 4 relative to Model 2 consistently demonstrate gains of 1.04 dB and 1.18 dB, respectively. This suggests the potential of applying the gated fusion method to different TSE models. Additionally, the combination of CBAM and GFM exhibits a synergistic effect, surpassing the individual contributions of each module, this indicates that advanced attention mechanisms can better utilize the more generalizable intermediate features obtained from

gated fusion, highlighting the amplifying effect of attention mechanisms in the TSE process. Moreover, the evaluation metric PESQ shows a similar trend to SI-SDR, specifically, the CBAM module or GFM module both lead to noticeable improvements in the quality of the extracted speech, and the combination of the two modules further enhances PESQ, exceeding the sum of the individual modules, which suggests that the CBAM and GFM mutually reinforce each other and jointly improve PESQ on the WSJ0-2mix dataset.

Table III presents the test results of various models on the WHAM! dataset, the histogram version is on the right side of Fig. 4 and Fig.5, and it is evident that the performance of all models in extracting target speech from noisy mixtures is significantly reduced. The improvement of Model 2 over Model 1 is 0.40 dB, while the improvement of Model 3 over Model 1 is 1.41 dB, which is higher than the improvement caused by GFM on the clean dataset WSJ0-2mix (1.04dB), this may be attributed to the following reasons: firstly, gated fusion can provide a richer representation of speech features by fusing outputs from different modules, in clean datasets, the target speech is usually clearer and easier to extract, and the model may already capture the features of the target speech effectively, however, in noisy datasets, the interference of noise reduces the effectiveness of the model in extracting the target speech, therefore the enhancement effect of gated fusion may be more pronounced on noisy datasets. Secondly, gated fusion introduces more parameters and complexity, which can provide more flexibility and expressive power to adapt to different speech scenarios and noise environments, this increased complexity may play a larger role on noisy datasets, helping the model better handle the competition between noise and target speech. In addition, the improvement generated by GFM (Model 3) is superior to the improvement generated by the combined module (Model 4) (1.41 dB vs. 1.02 dB, respectively). The possible reason for this is that noisy datasets typically contain various complex noise types and intensities, which can interfere with the extraction of the target speech, and a single module may have better performance in noise suppression, while the combination of CBAM and GFM may not effectively handle the complexity and interference of noise, thus affecting performance on noisy datasets. Also, combining CBAM and GFM increases the complexity and number of parameters in the model, while this can provide stronger feature representation and fusion capability, it may also lead to a decrease in the model's generalization ability on noisy datasets, the combined model may be more prone to fitting clean speech data and may not effectively adapt to the characteristics of noisy datasets. While PESQ exhibits a similar trend to SI-SDR, with improvements observed in Model 2 and Model 3 relative to Model 1, but the improvement of the combined module (Model 4) is inferior to that of the individual GFM module (Model 3).

IV. Conclusion

In this study, we focus on enhancing the key components of TSE model, particularly the target extractor and fusion layer. Specifically, we investigate the impact of integrating CBAM and GFM into the baseline model. CBAM adds a lightweight attention block after each TCN stack, introducing channel and spatial attention. The fusion layer replaces conventional concatenation with GFM, which better leverages embeddings, conveying crucial data to deeper layers. Experiments on WSJ0-2mix dataset exhibit noteworthy improvements in extraction performance, measured by SI-SDR and PESQ, without significant increase in the number of parameters and model complexity. GFM consistently enhances performance in various models, with combined attention and gated fusion yielding synergistic effects. Experiments on the WHAM! dataset show that the improved models incorporating CBAM module or GFM module achieve improvements in terms of both SI-SDR and PESQ on the test set. However, the combined module exhibit inferior performance than that of individual GFM, and the GFM module demonstrates better improvement on noisy dataset than that of clean dataset WSJ0-2mix. Future work entails refining models for superior performance on noisy datasets.

Acknowledgment

This work is supported by the Strengthening Plan of National Defense Science and Technology Foundation of China (2021-JCJQ-JJ-0059) and Natural Science Foundation of China (U2003207).

References

[1] R. Jin, M. Ablimit, and A. Hamdulla, "Speaker verification based on single channel speech separation," *IEEE Access*, pp. 1–1, 2023.

[2] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, Jul. 2015.

[3] R. Jin, M. Ablimit, and A. Hamdulla, "Speech separation and emotion recognition for multi-speaker scenarios," in *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, Chengdu, China: IEEE, Jul. 22, 2022, pp. 280–284.

[4] Z. Aysa, M. Ablimit, and A. Hamdulla, "Language identification in overlapped multi-lingual speeches," in *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, Chengdu, China: IEEE, Jul. 22, 2022, pp. 266–272.

[5] K. Zmolikova, M. Delcroix, K. Kinoshita, *et al.*, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, vol. 13, no. 4, 2019.

[6] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, May 2019, pp. 6965–6969.

[7] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 691–695.

[8] Q. Wang, H. Muckenhirn, K. Wilson, *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Interspeech 2019*, ISCA, Sep. 15, 2019, pp. 2728–2732.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[10] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[11] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Interspeech 2020*, ISCA, Oct. 25, 2020, pp. 1406–1410.

[12] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany: IEEE, Sep. 5, 2022, pp. 1–5.

[13] H. Li, X. Zhang, and G. Gao, "Dynamic-attention based encoder-decoder model for speaker extraction with anchor speech," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China: IEEE, Nov. 2019, pp. 297–301.

[14] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-SpEx: Localized target speaker extraction," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore: IEEE, May 23, 2022, pp. 7287–7291.

[15] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal attention fusion for target speaker extraction," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China: IEEE, Jan. 19, 2021, pp. 778–784.

[16] P. Shen, S. He, and X. Zhang, *ExARN: Self-attending RNN for target speaker extraction*, Dec. 2, 2022. arXiv: 2212.01106[eess].

[17] J. Lin, X. Cai, H. Dinkel, *et al.*, "Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 4, 2023, pp. 1–5.

[18] Z. Wang, R. Giri, S. Venkataramani, *et al.*, *Semi-supervised time domain target speaker extraction with attention*, Jun. 17, 2022. arXiv: 2206.09072[cs,eess].

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 5329–5333.

[21] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy: IEEE, May 2014, pp. 4052–4056.

[22] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-net: Target speaker separation via attention-based neural network," in *Interspeech 2020*, ISCA, Oct. 25, 2020, pp. 1411–1415.

[23] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, May 2019, pp. 6990–6994.

[24] Y. Du, Y. Liu, Z. Peng, and X. Jin, "Gated attention fusion network for multimodal sentiment classification," *Knowledge-Based Systems*, vol. 240, p. 108 107, Mar. 2022.

[25] K. Cho, B. Van Merrienboer, C. Gulcehre, *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734.

[26] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, vol. 2, 1999, 850–855 vol.2.

[27] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Place: Sydney, NSW, Australia, JMLR.org, 2017, pp. 933–941.

[28] N. Shazeer, *GLU variants improve transformer*, Feb. 12, 2020. arXiv: 2002.05202[cs,stat].

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Analysis of Speech Separation Performance Degradation on Emotional Speech Mixtures

Jia Qi Yip*†, Dianwen Ng*†, Bin Ma*, Chng Eng Siong†
* Speech Lab of DAMO Academy, Alibaba Group
† Nanyang Technological University, Singapore
E-mail: jiaqi006@e.ntu.edu.sg

***Abstract*—Despite recent strides made in Speech Separation, most models are trained on datasets with neutral emotions. Emotional speech has been known to degrade performance of models in a variety of speech tasks, which reduces the effectiveness of these models when deployed in real-world scenarios. In this paper we perform analysis to differentiate the performance degradation arising from the emotions in speech from the impact of out-of-domain inference. This is measured using a carefully designed test dataset, Emo2Mix, consisting of balanced data across all emotional combinations. We show that even models with strong out-of-domain performance such as Sepformer can still suffer significant degradation of up to 5.1 dB SI-SDRi on mixtures with strong emotions. This demonstrates the importance of accounting for emotions in real-world speech separation applications.**

***Index Terms*—speech separation, transformer, deep learning, emotional speech, emotion classification**

I. Introduction

Speech Separation is the task of obtaining single speaker speech from a mixture of speakers, also known as the cocktail party problem [1]. It has been the focus of much recent research, as many downstream speech models for tasks such as Automatic Speech Recognition (ASR) [2] are trained on single talker speech. For deployment, mixed speech received from the wild should ideally be separated before performing ASR.

The effect of a speaker's emotions on their speech [3] and speaker emotion recognition [4] are well studied research areas. The emotions of a speaker or the intention of a speaker to convey emotion can result in differences in speech. While some of the differences are semantic [5], there are clear differences also manifest in the prosody [6] and articulation [7] of words. Thus, emotional differences in speech can show up at the frame level, where speech separation models operate.

The difference in sound between neutral or read speech and emotional speech has been shown to lead to performance degradation on a number of speech tasks such as speaker verification [8] and ASR [9]. Conversely, a mixture of speakers with different emotions also results in a degradation of emotion recognition performance [10].

The separation of emotional speech is important because overlapping speech often occurs when speakers are at heightened emotional states, such as in a heated argument or excited conversation, where the typical decorum of turn taking is breached[11]. Speech Separation models are often trained on datasets built from neutral speech. The most common datasets used for speech separation are LibriMix[12] created from the LibriSpeech[13] corpus and WSJ0Mix[14] created from the WSJ0[15] corpus. Both datasets consist of read speech recorded under controlled conditions and neither are explicitly emotional datasets. This results in a potential mismatch between speech separation models trained primarily on neutral speech and the real-world emotional speech mixtures.

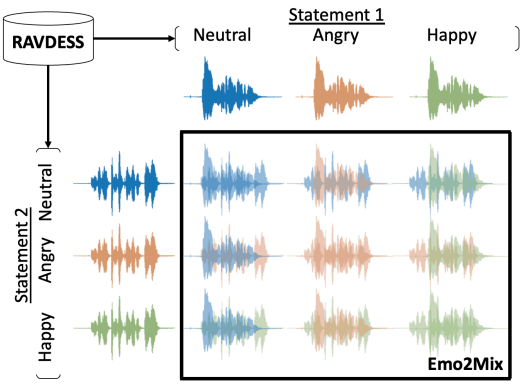


Fig. 1. Illustration of the mixing strategy. Only three emotions are shown here for brevity but the full set of available emotions are used in the dataset (8 at normal intensity and 7 at strong intensity). Each mixture consist of different speakers each vocalizing different statements. The same two statements are shared across emotions and speakers to control for semantic differences.

A. Our Approach

In this paper analyses the performance degradation of speech separation models using a custom speech separation test dataset, Emo2Mix, built based utterances from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [16]. To the authors' knowledge this is the first study to properly investigate the impact of emotions on speech separation with a test dataset that is balanced across multiple emotions at two different emotional intensities.

Using this dataset, we conduct a detailed analysis of the impact of emotional speech on state-of-the-art speech separation models that were trained on neutral speech. To control for the out-of-domain issue due to the unseen speakers in the Emo2Mix test dataset, we make use of the neural emotion utterances present within the RAVDESS dataset as a baseline for an out-of-domain speaker with an "in-domain" emotion.

We show that Emo2Mix presents a significant challenge to speech separation models trained on neutral speech with a

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

recent state-of-the-art Sepformer model [17][18] experiencing a performance degradation of 7.0dB compared to its Libri2Mix baseline. Of this, a performance difference of 1.9dB can be attributed to emotions alone, when comparing against the in-domain neutral emotion baseline.

B. Related Work

Recently, [19] also proposed an emotional mixture test dataset based on the RAVDESS [16] dataset, RAVDESS2Mix. The dataset consists of mixtures from the RAVDESS dataset and LibriSpeech [13] dataset, as well as enrollment speech from the RAVDESS dataset for the target speaker extraction task.

Compared to RAVDESS2Mix [19], our proposed Emo2Mix dataset does not consist of any utterances from the LibriSpeech corpus. This is for two reasons. Firstly, the LibriSpeech corpus consists exclusively of neutral speech, which limits the number of emotion combinations possible for the dataset. Secondly, many speech separation models are trained on the Libri2Mix dataset, which uses utterances from the LibriSpeech corpus. This makes the RAVDESS2Mix test dataset biased, as models pretrained on Libri2Mix will have and advantage over models pretrained on other datasets such as the commonly used WSJ02Mix dataset, which creates utterances from the WSJ0 corpus. Based on the RAVDESS2Mix [19] proposed that target speaker extraction is negatively impacted by emotional speech while blind source separation was not as significantly impacted. However, using Emo2Mix, we can show that the same performance degradation is observed once the in-domain advantage of Librispeech models are removed.

II. Methodology

A. Evaluation

The performance of the models are measured using Scale-invariant signal-to-distortion ratio improvement (SI-SDRi) which is measured by comparing the output waveforms of the models with the ground-truth waveforms. Since there are multiple outputs, the final performance is calculated using the permutation invariant approach.

B. The RAVDESS dataset

The RAVDESS dataset[16] is an emotional dataset consisting of 7356 recordings by 24 actors, gender balanced with 12 male and 12 female. The actors vocalize two lexically-matched statements in a neutral North American accent at two emotional intensities, normal and strong. At the normal emotional intensity, 8 emotions are expressed (neutral, calm, happy, sad, angry, fearful, disgust, surprised) while the strong intensity share the same emotions except for neutral.

RAVDESS has a number of advantages over other emotional speech databases because of its inclusion of two neutral emotions, "neutral" and "calm" and two emotional intensities "normal" and "strong". These features are useful for comparison with existing corpora like LibriMix and WSJ0Mix which also call in this emotional category. These baselines can help differentiate between the performance impact of out-of-domain inference from the unseen speakers and the actual performance impact of the emotional speech.

C. Emo2Mix

The Emo2Mix dataset draws from the RAVDESS dataset to produce two-speaker mixtures of emotional speech. The subset of RAVDESS selected for Emo2Mix excludes singing speech. It also excludes the audiovisual data. A breakdown of the features in the chosen subset of the RAVDESS dataset is shown in Table I.

TABLE I: Breakdown of the features of the selected subset of the RAVDESS dataset used to create Emo2Mix

Feature	**Normal**	**Strong**
Number of Speakers	24	24
Number of Emotions	8	7
Number of Statements	2	2
Number of Repetitions	2	2
Total Recordings	768	672

The mixtures are created with an even distribution across 64 combinations of 8 emotions at a normal emotional intensity and 49 combinations of 7 emotions at a strong emotion intensity. Each statement in the mixture is different, so the mixtures are produced as permutations and not combinations of emotions. While each actor will vocalize the same statement at the same emotion twice, to reduce superfluous mixtures we simply randomly select one of the two repetitions.

For each of the 113 emotion-intensity combinations, each actor vocalizes 2 different statements, from which we can create $N \times (N-1)$ combinations can be obtained, since we cannot mix a actor with themselves. For the full dataset, $N = 24$, which results in 552 mixtures per emotion-intensity permutation which would result in $552 \times 113 = 62,376$ mixtures overall. To reduce the number of testing mixtures, we reduce the number of speakers in the test set by setting $N = 8$, which results in 56 mixtures per emotion-intensity permutation. This gives $56 \times 64 = 3,584$ mixtures for the normal intensity test set, and $56 \times 49 = 2,744$ mixtures for the strong intensity test set. This is done by selecting only every third speaker (1,4,7,10,13,17,21,24) from the dataset. This approach, as opposed to randomly selecting 8 out of 24 of the speakers, is chosen so as to enable the possibility of future work on fine-tuning-models on the other 16 held-out speakers.

The utterances selected from the RAVDESS dataset are dynamically mixed [20] so that the mixture test set does not have to be separately stored. During mixing, a number of pre-processing steps are performed in line with the parameters used for dynamic mixing [20] implemented on the Speechbrain [21] framework. Two audio segments are first cropped to the length of the minimum of the two, then downsampled to 8kHz and normalized. These processed segments are then combined at equal weights if no clipping is observed, or reweighted if

TABLE II: Overall results across different datasets, with aggregated results reported for the emotional datasets. The performance degradation due to only emotional mixtures is calculated by (4)-(6) which is the difference in performance between the Strong emotion and the Neutral emotion baseline. Out-of-domain performance degradation, independent of emotions can be measured by the difference between the original baseline and the Emo2Mix (Neutral) mixtures, calculated by (1)-(4).

	Model:	Sepformer				ConvTasNet			
	Training Dataset:	Libri2Mix		WSJ0-2Mix		Libri2Mix		WSJ0-2Mix	
ID	**Testing Dataset**	**SI-SDRi**	**SDRi**	**SI-SDRi**	**SDRi**	**SI-SDRi**	**SDRi**	**SI-SDRi**	**SDRi**
1	Libri2Mix-Test	20.6	20.9	17.0	17.5	15.1	15.5	10.0	10.6
2	RAVDESS2Mix-Normal	20.2	20.7	15.1	16.0	15.0	15.5	8.6	9.4
3	RAVDESS2Mix-Strong	20.2	20.7	14.4	15.3	14.9	15.5	8.2	9.1
4	Emo2Mix (Neutral)	15.5	16.3	12.0	12.9	10.5	11.4	4.9	5.8
5	Emo2Mix-Normal	15.4	16.2	11.4	12.2	10.7	11.4	3.3	4.3
6	Emo2Mix-Strong	13.6	14.7	9.4	10.3	8.4	9.2	0.9	2.7
7	(1) - (4)	**5.1**	**4.6**	**5.0**	**4.6**	4.6	4.1	5.1	4.8
8	(4) - (6)	**1.9**	**1.6**	**2.6**	**2.6**	2.1	2.2	4.0	3.1

clipping is present. To ensure the reputability of the results we used fixed random seeds for all dataset generation and provide a standard script for generation which we release on GitHub[1].

D. Training Datasets

In this work we focus on models that have been trained on the Libri2Mix [12] and WSJ0-2Mix [14] datasets. The Libri2Mix training set consists of 50,800 mixtures while WSJ0-2Mix training set consists of 20,000 mixtures. Both have test sets of 3,000 mixtures. We note that 3,000 mixtures are comparable to the 3,584 and 2,744 mixtures in the Emo2Mix test set.

E. Baseline Models

We make use to two baseline models, Sepformer which adopts a dual-path speech separation architecture with transformer blocks, and ConvTasNet [22] which utilizes temporal convolution blocks. Both models are seminal in the field of speech separation, with Sepformer [17] [18] being the more recent state-of-the art model. We implement the two models on the Speechbrain framework as per their respective papers. For Sepformer we make use of pre-trained checkpoints available from the speechbrain huggingface[2]. Two versions of the model are available, one trained on Libri2Mix (Sepfromer-L2M) and one trained on WSJ0-2Mix (Sepformer-WJ2). For ConvTasNet we train the model from scratch and validate the model replicating the Libri2Mix-Test baseline. To align with the models available on hugging face, we also train the model on both the Libri2Mix (ConvTasNet-L2M) and WSJ0-2Mix (ConvTasNet-WJ2) datasets. Although newer models surpassing Sepformer have recently been released [23][24], Sepformer and ConvTasNet are currently still more widely used and replicated and thus serve as better baselines.

[1]https://github.com/Yip-Jia-Qi/EmoMix

[2]https://huggingface.co/speechbrain

III. Results

In Table II we compare the overall separation performance of the baseline models against three different datasets, the original Libri2Mix Test dataset, RAVDESS2Mix and Emo2Mix. For each of the baseline models, we report the results based on pre-training on two different sets of training data, Libri2Mix and WSJ02Mix. The results on the RAVDESS2Mix dataset are provided as a point of comparison between related emotional speech separation test sets.

A. Performance degradation due to emotional speech

The results presented in Table II for Emo2Mix-Normal and Emo2Mix-Strong are aggregated across all emotional combinations through simple averaging, since all emotional combinations are equally weighted. The Emo2Mix (Neutral) test is a subset of the Emo2Mix-Normal test data and consists of 56 mixtures of different speakers where both speakers are portraying neutral emotions. The purpose of Emo2Mix (Neutral) is to serve as a baseline where the speaker is out-of-domain, but has a neutral emotion matching the emotions in the training data. This allows us to measure the out-of-domain performance degradation, as per (7) in Table II, independent of the impact of emotions.

Any further performance degradation beyond the Emo2Mix (Neutral) baseline, i.e. (8) in Table II, is unlikely to be due to the out-of-domain inference. If emotions do not impact speech separation performance, we should see the values in row (8) approach zero. However, the results of our experiments in Table II show that this is not the case, as the values for (8) range between 1.6dB and 4.0dB. Thus, based on the comparisons of the out-of-domain and emotional degredation in (7) and (8) of Table II, we can conclude that emotions do result in performance degradation in speech separation models. However, the performance impact of performing inference on an unseen speaker is greater than the performance impact from the within-speaker variance resulting from emotions.

TABLE III: Performance of Sepformer-L2M on the Emo2Mix-Normal dataset broken down by emotion combination reported using SI-SDRi.

	RD2M	calm	happy	sad	angry	fearful	disgust	surprised	neutral
calm	**20.6**	16.0	15.6	16.0	**17.1**	15.0	14.9	14.6	15.9
happy	20.3	15.7	14.9	16.3	16.3	15.9	16.0	13.5	**16.4**
sad	20.4	15.6	15.8	16.1	**17.1**	16.1	15.7	15.7	16.3
angry	20.3	16.5	14.9	**17.2**	15.0	16.0	15.6	14.4	16.9
fearful	20.2	15.4	14.3	**16.9**	15.1	15.9	15.0	13.6	16.4
disgust	20.0	15.4	15.0	**16.3**	15.3	15.5	15.1	13.9	14.7
surprised	19.7	14.4	12.7	**15.3**	15.2	14.5	13.8	12.5	14.5
neutral	19.9	15.8	15.8	16.3	16.2	**16.8**	15.9	14.8	15.5

TABLE IV: Performance of Sepformer-L2M on the Emo2Mix-Strong dataset broken down by emotion combination reported using SI-SDRi.

	RD2M	calm	happy	sad	angry	fearful	disgust	surprised
calm	20.4	15.9	17.1	15.8	**18.6**	17.7	16.9	15.9
happy	**20.6**	**17.2**	12.4	13.8	10.9	11.2	14.8	12.2
sad	20.1	15.7	13.6	14.7	**15.8**	12.5	13.1	13.0
angry	20.4	**17.6**	11.6	15.9	9.3	10.6	15.7	13.6
fearful	20.3	**18.4**	11.9	14.1	11.6	9.7	14.9	11.7
disgust	19.8	**17.4**	11.7	14.4	13.9	11.8	13.4	10.7
surprised	19.7	**16.7**	12.4	14.3	13.0	11.9	12.4	10.4

Taking (8) divided by (7), we see across multiple models that strong emotions can cause an additional 35-55% performance drop on top of the out-of-domain performance drop (7) when compared to the in-domain baseline (1). This shows that although the performance degradation due to emotions (8) is less than that from out-of-domain inference (7), the impact is still significant.

We can further support the above conclusion using at the internal consistency in the results of the Emo2Mix test datasets at different emotional intensities. From the results of (4), (5) and (6) in Table II we see a significant performance gap between Emo2Mix-Normal and Emo2Mix-Strong, while the gap between Emo2Mix (Neutral) and Emo2Mix-Normal is much smaller. This suggests that emotional speech causes a deterioration in speech separation performance. If emotions do not impact speech separation performance, we should see that Emo2Mix-Normal and Emo2Mix-Strong have the similar performance. Since this is also a within-dataset comparison, the performance difference is not due to differences in speakers, since the speakers in both Emo2Mix-Normal and Emo2Mix-Strong are identical.

B. Comparison with RAVDESS2Mix

The RAVDESS2Mix dataset consists of mixtures of 2 speakers where one speaker is drawn from the RAVDESS dataset while the other speaker is drawn from the Libri2Mix-Test dataset. This means that for a model trained on the Libri2Mix dataset, one of the speakers would be in-domain while the other is out-of-domain. In a 2-speaker separation problem, knowing the mask for one speaker could allow a model to infer the mask for the other speaker. This gives a model trained on the Libri2Mix dataset a big advantage. This is observed across all models, where we see little to no performance degradation for Sepformer-L2M and ConvTasNet-L2M but a significant performance drop for Sepformer-WJ2 and ConvTasNet-WJ2.

Furthermore, when comparing the performance of RAVDESS2Mix-Normal and RAVDESS2Mix-Strong, the latter test set results in a lower performance for models trained on the WSJ0-2Mix dataset but this is not observed for the models trained on the Libri2Mix dataset. This lack of internal consistency across models trained on different training datasets, points to the fact that the use of the Libri2Mix-Test set in the RAVDESS2Mix mixtures allow for the model to bypass having to recognise emotional speech by inferring from the neutral speech from the in-domain Libri2Mix-Test utterance.

In comparison to RAVDESS2Mix, Emo2Mix does not suffer from such an issue and is able to accurately benchmark the impact to performance caused by emotional speech. Across all the models tested, for both Libri2Mix and WSJ0-2Mix trained versions, there is a significant performance difference. This shows that the test dataset we have developed for our analysis, Emo2Mix, is a better benchmark for emotional speech separation than the previously reported RAVDESS2Mix.

C. The benefits of a larger training dataset

It is common to find in various speech domains such as ASR [2] and Keyword spotting [25] [26], that a larger training dataset results in better performance. Since the Libri2Mix training dataset is significantly larger than WSJ0-2Mix we would expect to see a similar trend. Furthermore, Libri2Mix consists of LibriSpeech utterances which come from audio-

books which may be closer in domain to the RAVDESS datset. That said, the larger size of the Libri2Mix dataset likely improves the performance of the model on Emo2Mix. Looking only at the Sepformer experiments, we see that on Emo2Mix-Neutral the Sepformer-WJ2 model has a lower SDRi of 12.0 dB compared to 15.5 dB for Sepformer-L2M. The difference in SDRi performance for Emo2Mix-Strong is 1.9 dB on Sepformer-L2M compared to 2.6 dB for Sepformer-WJ2. A similar trend can be found for the ConvTasNet model, with the SDRi dropping to barely above noise at 0.9 dB on Emo2Mix strong. This mirrors the finding of [27] where emotional recognition performance benefited from non-emotional large scale pretraining.

D. Performance Comparisons by emotions

To streamline the performance comparison, we only report here the SI-SDRi metric of the best performing model, Sepformer-L2M. Table III shows the results of the model tested on the Emo2Mix-Normal dataset, while Table IV shows the results of the model tested on the Emo2Mix-Strong dataset. In each of the tables, the highest value in each row is listed in bold and while lowest value is underlined.

In the first column of each table, we also report the comparison with the RAVDESS2Mix dataset (RD2M) to which consists only of a combination of 1 emotional utterance from the RAVDESS dataset and 1 neutral utterance from the Libri2Mix Test dataset. In this comparison we also see that there is a minimal variation in the RD2M dataset across the various emotions of 0.9dB for both the Strong and Normal emotional intensities. Meanwhile the average difference between the min and max emotions across combinations is 2.5dB for the normal intensity and 6.1dB for the strong intensity, with the lowest being 1.5dB for the "sad" emotion row in the normal dataset. Across all combinations in the entire dataset, the difference between the min and max for Emo2Mix-normal is 4.7dB and for Emo2Mix-strong is 9.3dB.

Considering the results on the Emo2Mix-Normal dataset in Table III we see a trend that having an utterance with the surprise emotion consistently results in the worst performance compared to the other emotions, with the worst results arising from a combination of two surprise utterances. One possible explanation is that the actors in the dataset simply found it difficult to express surprise at a low emotional intensity and exhibited the most intense emotions in their surprised utterances compared to the other emotions expressed at normal intensity. Meanwhile sad and angry are emotions expressed at normal intensity could simply be close to neutral, resulting in better SI-SDRi results compared to the other emotions.

Finally, considering the results on the Emo2Mix-Strong dataset in Table IV we see a slightly different dynamic. The Calm utterances consistently have the best results in the Strong dataset compared to the Normal dataset. This could be that Calm utterances are closer to neutral at Strong emotional intensity and thus their contrast with the other strong emotions results makes it easier for the model to differentiate the speakers.

IV. Conclusion

In this work we have shown that, contrary to previous work [19], emotions play an important role in the speech separation performance. Our results show that models trained on neutral speech will suffer a performance degradation when the mixture contains strong emotional expressions at inference time. Thus, this work makes the case for the need of including emotional speech into speech separation training datasets.

V. Acknowledgements

This research is supported by ST Engineering Mission Software & Services Pte. Ltd under a collaboration programme (Research Collaboration No: REQ0149132). The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] D. Ng, R. Zhang, J. Q. Yip, Z. Yang, J. Ni, C. Zhang, Y. Ma, C. Ni, E. S. Chng, and B. Ma, "dehubert: Disentangling noise in a self-supervised model for robust speech recognition," 2023.

[3] R. Chakraborty, M. Pandharipande, and S. K. Kopparapu, "Analyzing emotion in spontaneous speech," in *Springer Singapore*, 2018.

[4] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639319302262

[5] T. Johnstone, "The effect of emotion on voice production and speech acoustics," Aug 2017. [Online]. Available: thesiscommons.org/qd6hz

[6] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1, pp. 227–256, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639302000845

[7] S. Lee, S. Yıldırım, E. A. Kazemzadeh, and S. S. Narayanan, "An articulatory study of emotional speech production," in *Interspeech*, 2005.

[8] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," 2020.

[9] R. Munot and A. Nenkova, "Emotion impacts speech recognition performance," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 16–21. [Online]. Available: https://aclanthology.org/N19-3003

[10] R. Jin, M. Ablimit, and A. Hamdulla, "Speech separation and emotion recognition for multi-speaker scenarios," in *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, 2022, pp. 280–284.

[11] F. Gervits and M. Scheutz, "Towards a conversation-analytic taxonomy of speech overlap," in *International Conference on Language Resources and Evaluation*, 2018.

[12] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[14] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[15] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a." 1993.

[16] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," Apr. 2018, Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. [Online]. Available: https://doi.org/10.5281/zenodo.1188976

[17] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, "On using transformers for speech-separation," *arXiv preprint arXiv:2202.02884*, 2022.

[18] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[19] J. Švec, K. Žmolíková, M. Kocour, M. Delcroix, T. Ochiai, L. Mošner, and J. Černocký, "Analysis of impact of emotions on target speech extraction and speech separation," 2022.

[20] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2020.

[21] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021. [Online]. Available: https://arxiv.org/abs/2106.04624

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, aug 2019. [Online]. Available: https://doi.org/10.1109%2Ftaslp.2019.2915167

[23] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," 2023.

[24] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *arXiv preprint arXiv:2211.12433*, 2022.

[25] D. Ng, R. Zhang, J. Q. Yip, C. Zhang, Y. Ma, T. H. Nguyen, C. Ni, E. S. Chng, and B. Ma, "Contrastive speech mixup for low-resource keyword spotting," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[26] D. Ng, J. Q. Yip, T. Surana, Z. Yang, C. Zhang, Y. Ma, C. Ni, E. S. Chng, and B. Ma, "I2cr: Improving noise robustness on keyword spotting using inter-intra contrastive regularization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 605–611.

[27] M. M. Gauy and M. Finger, "Pretrained audio neural networks for speech emotion recognition in portuguese," *arXiv preprint arXiv:2210.14716*, 2022.

Geometrically Constrained Blind Moving Source Extraction based on Constant Separation Vector and Auxiliary Function Technique

Ruifeng Zhang, Tetsuya Ueda and Shoji Makino
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan

***Abstract*—This paper focuses on the permutation problem in Blind Source Extraction (BSE) by employing beamforming-based Geometrical Constraints (GC). Specifically, we focus on the recently proposed auxiliary function-based Independent Vector Extraction (Aux-IVE) with Constant Separation Vector (CSV) mixing model. Building upon this mixing model, we provide the source signals' spatial information, utilizing GC based on beamforming. We facilitate the extraction of moving target source and improve the extraction performance. Furthermore, we discuss the effect of the number of constraints on the extraction performance. Experimental results confirmed that our proposed method, GC-CSV-Aux-IVE, exhibits superior extraction performance and achieves nearly 100% accuracy in extracting the moving target source.**

I. Introduction

Recently, the applications of Blind Source Separation (BSS) [1]–[3] span across a wide range of fields, contributing to advancements in speech recognition, audio processing, and speech enhancement. BSS algorithms can separate individual speech signals in scenarios where multiple speeches overlap, improving speech recognition accuracy and estimating the speaker's location.

In the field of BSS, Independent Component Analysis (ICA) [4], [5] and Independent Vector Analysis (IVA) [6], [7] have emerged as prominent techniques for separating mixed signals into their constituent sources. Auxiliary function-based Independent Vector Analysis (Aux-IVA) [8], [9] is an extension of IVA that incorporates auxiliary functions to improve the separation performance.

Moreover, in environments with noisy or degraded speech, Blind Source Extraction (BSE) methods [10]–[12] offer the capability to enhance the target speech by extracting it from the mixture, leading to improved speech intelligibility and quality. BSE techniques have emerged as effective methods for extracting specific source signals from mixtures. By utilizing various algorithms and signal processing methods [13], BSE focuses on extracting the Source of Interest (SOI) while suppressing or removing interference from other sources in the mixture.

In recent years, the Constant Separating Vector (CSV) model [14], [15] proposed in the BSE domain has demonstrated significant advancements in the moving source extraction. Based on the CSV model, [16] proposed CSV-Aux-IVE, which estimates a time-varying mixing vector and a time-invariant filter and utilizes an auxiliary function technique to enhance the extraction performance and the convergence speed of Independent Vector Extraction (IVE).

However, since the permutation problem exists, it is often difficult to determine which source will be extracted in the field of BSE. An important consideration is ensuring that the extracted signal corresponds to SOI. In [14], they addressed this problem by employing the pilot signal which is a signal that is mutually dependent with the corresponding source signal. This method uses prior information about the sound source to resolve the permutation ambiguity.

On the other hand, we propose to solve this ambiguity using Geometric Constraints (GC) [17]. Moreover, we aim to give several constraints for the moving speaker. By employing GC method [18]–[21], we can enhance the accuracy of signal extraction by only requiring approximate location information of the target signal and sensors positions. Consequently, it can be widely applied in scenarios where the signal's moving range is known, such as conferences and speech monitoring. GC method is based on beamforming, which utilizes spatial information and sensor positions to optimize BSS problems. It exploits spatial information to guide the separation matrix to obtain the signal in the desired direction. This method optimizes the constraint weights to ensure that extracted source signal keeps unit energy. Therefore, it is our objective to address the permutation problem and improve the performance of moving target source extraction by applying the GC method.

II. Problem Formulation

A. Blind Source Extraction

Considering the situation in BSE, we assume there are M microphones in the room. After the Short Time Fourier Transform (STFT), we obtain microphone signals $\boldsymbol{x}_{f,l}$ and source signals $\boldsymbol{s}_{f,l}$, where f and l represent the frequency bin and time frame index. Our main task is to extract SOI $s_{f,l}$ from M microphones, and $\boldsymbol{z}_{f,l}$ represent the remaining background signals. Generally, we can assume that $s_{f,l}$ corresponds to the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

first source of $\boldsymbol{s}_{f,l}$ without loss of generality. So the relationship of microphone signals $\boldsymbol{x}_{f,l} = [x_{f,l,1}, \ldots, x_{f,l,M}]^{\mathsf{T}} \in \mathbb{C}^M$ and source signals $\boldsymbol{s}_{f,l} = [s_{f,l}, \boldsymbol{z}_{f,l}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^N$ where $\boldsymbol{z}_{f,l} = [z_{f,l,1}, \ldots, z_{f,l,N-1}]$ can be written as

$$\boldsymbol{x}_{f,l} = \boldsymbol{A}_{f,l}\boldsymbol{s}_{f,l} = \boldsymbol{A}_{f,l}\begin{bmatrix} s_{f,l} \\ \boldsymbol{z}_{f,l} \end{bmatrix}, \tag{1}$$

where $\boldsymbol{A}_{f,l}$ stands for $M \times N$ mixing matrix and $(\cdot)^{\mathsf{T}}$ denotes the transpose. If M is equal to N, the inverse matrix of $\boldsymbol{A}_{f,l}$ exists and is called separation matrix $\boldsymbol{W}_{f,l} \in \mathbb{C}^{M\times M}$, so (1) can be written as:

$$\boldsymbol{s}_{f,l} = \boldsymbol{W}_{f,l}\boldsymbol{x}_{f,l} = \begin{bmatrix} \boldsymbol{w}_f^{\mathsf{H}} \\ \boldsymbol{B}_{f,l} \end{bmatrix}\boldsymbol{x}_{f,l}. \tag{2}$$

Here $(\cdot)^{\mathsf{H}}$ denotes Hermit transpose. Then we divide mixing matrix as $\boldsymbol{A}_{f,l} = [\boldsymbol{a}_{f,l}\ \boldsymbol{D}_{f,l}]$ which implies that $\boldsymbol{a}_{f,l}$ is the first column of $\boldsymbol{A}_{f,l}$ while $\boldsymbol{D}_{f,l}$ is the remaining part. Similarly, $\boldsymbol{w}_f^{\mathsf{H}}$ refers to the first row of the matrix, and $\boldsymbol{B}_{f,l}$ refers to the rest. Then the mixing and separation model (1) and (2) can be written as:

$$\boldsymbol{x}_{f,l} = \boldsymbol{a}_{f,l}s_{f,l} + \boldsymbol{D}_{f,l}\boldsymbol{z}_{f,l}, \tag{3}$$

and

$$\begin{bmatrix} s_{f,l} \\ \boldsymbol{z}_{f,l} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{x}_{f,l} \\ \boldsymbol{B}_{f,l}\boldsymbol{x}_{f,l} \end{bmatrix}. \tag{4}$$

B. CSV Mixing Model

In the CSV model, the signal after STFT is divided into T blocks, and the length of each block is defined as L_b. At this time, the index of each block is $t \in \{1, \ldots, T\}$. Thus, the microphone signal is divided into several blocks. We consider the signal in each block approximately stationary. Accordingly, the expressions of the corresponding mixing matrix $\boldsymbol{A}_{f,t}$ in (1) and separation matrix $\boldsymbol{W}_{f,t}$ in (2) will be rewrite as [12]

$$\boldsymbol{A}_{f,t} = [\boldsymbol{a}_{f,t}, \boldsymbol{D}_{f,t}] = \begin{bmatrix} \gamma_{f,t} & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & \frac{1}{\gamma_{f,t}}(\boldsymbol{g}_{f,t}\boldsymbol{h}_f^{\mathsf{H}} - \boldsymbol{I}_{M-1}) \end{bmatrix}, \tag{5}$$

$$\boldsymbol{W}_{f,t} = \begin{bmatrix} \boldsymbol{w}_f^{\mathsf{H}} \\ \boldsymbol{B}_{f,t} \end{bmatrix} = \begin{bmatrix} \beta_f^* & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & -\gamma_{f,t}\boldsymbol{I}_{M-1} \end{bmatrix}. \tag{6}$$

Here $(\cdot)^*$ denotes the complex conjugate, and $\boldsymbol{I}_M$ denotes to $M \times M$ identity matrix. In the remaining sections of this paper, we omit the frame index $l = \{1, \ldots, L\}$. Given that the mixing matrix and separation matrix are inverse to each other, we can obtain $\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{a}_{f,t} = 1$, which is called distortionless constraint. Similarly, we can get $\boldsymbol{B}_{f,t}\boldsymbol{a}_{f,t} = 0$, where $\boldsymbol{B}_{f,t}$ represents the block matrix. Using the assumption above, we can rewrite (1) and (2) as follows:

$$\boldsymbol{x}_{f,t} = \begin{bmatrix} \gamma_{f,t} & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & \frac{1}{\gamma_{f,t}}(\boldsymbol{g}_{f,t}\boldsymbol{h}_f^{\mathsf{H}} - \boldsymbol{I}_{M-1}) \end{bmatrix}\begin{bmatrix} s_{f,t} \\ \boldsymbol{z}_{f,t} \end{bmatrix}, \tag{7}$$

and

$$\begin{bmatrix} s_{f,t} \\ \boldsymbol{z}_{f,t} \end{bmatrix} = \begin{bmatrix} \beta_f^* & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & -\gamma_{f,t}\boldsymbol{I}_{M-1} \end{bmatrix}\boldsymbol{x}_{f,t}. \tag{8}$$

We assume that $\boldsymbol{s}_{f,t}$ and $\boldsymbol{z}_{f,t}$ are mutually independent across all time frequency bins. Based on this assumption, we can express the joint pdf of them as follows:

$$p(\{\boldsymbol{s}_t, \boldsymbol{z}_{f,t}\}_{f,t}) = \prod_t p(\boldsymbol{s}_t)\prod_{f,t} p(\boldsymbol{z}_{f,t}), \tag{9}$$

here $p(\boldsymbol{s}_t)$ denote the joint pdf of the SOI vector component $\boldsymbol{s}_t = [s_{1,t}, \ldots, s_{F,t}] \in \mathbb{C}^F$ and $p(\boldsymbol{z}_{f,t})$ denote the pdf of $\boldsymbol{z}_{f,t}$, respectively.

We consider the block-dependent variance of SOI, which can vary from block to block. We introduce pdf, denoted as $p(\boldsymbol{s}_t)$, that captures this block-dependent variance and accurately represents the characteristics of the source signal. Specifically, the pdf $p(\boldsymbol{s}_t)$ is defined as follows:

$$p(\boldsymbol{s}_t) = h\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right)\left(\prod_{f=1}^{F}\hat{\sigma}_{f,t}\right)^{-2}, \tag{10}$$

where $\hat{\sigma}_{f,t} = \sqrt{\boldsymbol{w}_f^{\mathsf{H}}\hat{\boldsymbol{C}}_{f,t}\boldsymbol{w}_f}$ and $\hat{\boldsymbol{C}}_{f,t} = \mathbb{E}\left[\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]$. $h(\cdot)$ is a pdf corresponding to a normalized non-Gaussian random variable:

$$h\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right) = C\exp(-G_R(r_t)), \tag{11}$$

where C is a coefficient and G_R is a continuous and differentiable function of a real variable r satisfying that $\psi(r) = \frac{G_R'(r)}{r}$ is continuous and monotonically decreasing in $r \geq 0$. The background noise $\boldsymbol{z}_{f,t}$ is assumed to follow a circular Gaussian distribution with zero mean and a variance matrix $\hat{\boldsymbol{C}}_{\boldsymbol{z}_{f,t}} = \mathbb{E}\left[\boldsymbol{z}_{f,t}\boldsymbol{z}_{f,t}^{\mathsf{H}}\right]$.

After applying the auxiliary function method described in [8], we can obtain

$$\hat{\mathbb{E}}\left[-\log h\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right)\right] \leq \frac{1}{2}\frac{1}{\hat{\sigma}_{f,t}^2}\boldsymbol{w}_f^{\mathsf{H}}\mathbb{E}\left[\psi(r_t)\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]\boldsymbol{w}_f + R_t, \tag{12}$$

where r_t is an auxiliary variable and R_t depends purely on r_t. Based on the above assumptions, we proceed to apply the auxiliary function method for updating the objective function. Consequently, we obtain the auxiliary function for CSV as follows:

$$\mathcal{L}_{\text{Aux}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{F}\left\{\frac{1}{2}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2} + \log\hat{\sigma}_{f,t}^2 + \mathbb{E}\left[\boldsymbol{z}_{f,t}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{z}_{f,t}}^{-1}\boldsymbol{z}_{f,t}\right] - (M-2)\log|\gamma_{f,t}|^2\right\} + R_t, \tag{13}$$

where

$$\boldsymbol{V}_{f,t} = \mathbb{E}\left[\psi(r_t)\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]. \tag{14}$$

To minimize the auxiliary function with respect to the normal variables, we calculate the partial derivative with respect

to $\boldsymbol{w}_f$. In order to minimize the auxiliary function (13), we derivate it and get

$$\frac{\partial \mathcal{L}_{\text{Aux}}}{\partial \boldsymbol{w}_f^*} = \frac{1}{2T}\sum_{t=1}^{T}\left\{\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\boldsymbol{w}_f - \frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right\}, \quad (15)$$

where we used the same technique as the previous researches [12], [16] to replace the derivative of the third and fourth term in (13) as $\sum_{f=1}^{F}\boldsymbol{a}_{f,t}$. Using conventional technique [16], we get an update rule for the separation filter $\boldsymbol{w}_f$ by taking linearized solution by fixing $\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}$ and $\hat{\sigma}_{f,t}^2$ as constant terms. Then the update rule of $\boldsymbol{w}_f$ can be written as

$$\boldsymbol{w}_f = \left(\sum_{t=1}^{T}\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1}\sum_{t=1}^{T}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}. \quad (16)$$

III. Proposed Method

So far, the CSV-Aux-IVE in Section II-B still encounters the permutation problem. In other words, without prior information, it is not determined whether the separate output corresponds to the SOI or Interference (IR). Therefore, we introduce GC [17] to ensure the extraction of the SOI. Then, let us consider GC that restrict the far-field response of the seperation filter$\boldsymbol{w}_f$ at the direction θ, which is described as

$$\mathcal{L}_{\text{GC}} = \lambda_{\text{GC}}\sum_{f=1}^{F}\sum_{\theta\in\Theta}|\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{d}_{f\theta} - 1|^2, \quad (17)$$

where $\boldsymbol{d}_{f\theta}$ is the steering vector pointing to the direction θ, λ_{GC} is a parameter weighting the importance of GC in the objective function, and we can preserve the SOI by designing this constraint when θ is the DOA (Direction of Arrival) of the target, which we call Unit Response (UR) constraints. The set Θ represents the collection of angles that will be taken into account.

The target source we aim to extract is moving, while the filter $\boldsymbol{w}_f$ remains time-invariant. Thus, a challenge lies in how to provide the filter with information regarding the movement of SOI. In this study, we make the assumption that the range of movement for the signals is known. We divide this range into p equally spaced angles and subsequently apply UR constraints to the signals at each angle.

We combine two loss functions (13) and (17) to apply UR constraints to SOI by summing them together to create a composite objective function, denoted as

$$\mathcal{Q}_{\text{CSV+GC}} = \mathcal{L}_{\text{Aux}} + \mathcal{L}_{\text{GC}}. \quad (18)$$

Then we calculate the partial derivative of the objective function. Thus the resulting derivative is as follows:

$$\frac{\partial \mathcal{Q}_{\text{CSV+GC}}}{\partial \boldsymbol{w}_f^*} = \frac{1}{2T}\sum_{t=1}^{T}\left\{\frac{\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2} - \frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right\} + \lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta}\boldsymbol{d}_{f\theta}^{\mathsf{H}}\boldsymbol{w}_f - \lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta}. \quad (19)$$

Subsequently, we take the linearized solution of $\boldsymbol{w}_f$:

$$\boldsymbol{w}_f = \left(\lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta}\boldsymbol{d}_{f\theta}^{\mathsf{H}} + \frac{1}{2T}\sum_{t=1}^{T}\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1}\left(\lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta} + \frac{1}{2T}\sum_{t=1}^{T}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right). \quad (20)$$

IV. Experiment

To evaluate the effectiveness of GC-CSV-Aux-IVE, we conducted a speech separation experiments. The evaluation criteria included extraction performance and order of the output signal, which we calculate its average as its accuracy. We compared our proposed method, GC-CSV-Aux-IVE, with the existing CSV-Aux-IVE method without being supervised.

A. Setup

In this experiment, we utilized a subset of the ATR Japanese Speech Database [22], consisting of recordings from three male speakers and three female speakers. We randomly selected two distinct speeches from this dataset to create mixtures. The indoor impulse response is generated through image method [23], with the observation signal generated using signal generator[1]. The layout of the simulated room is illustrated in Fig. 1. In total, we generated 100 pairs of original signals and mixed signals. The length of each signal is about 20 seconds. Among these, the first 50 signals had the SOI moving from 90° to 150°, while the remaining 50 signals had the SOI moving from 150° to 90°. Here SOI was set to move at a speed of 3° per second, while the IR is fixed at 45°. We have 6 microphones, 1 moving target source, and 1 fixed interference source.

To make this experiment more convincing, we added a white noise signal to the mixing signal. We adjusted the SNR to input SNR $= 10\log\frac{\lambda_1}{\lambda_2}$ as specified value, where λ_1 corresponds to the sample variance of mixing signal and λ_2 corresponds to that of noises. The other experiment conditions are shown in Table I.

In this experiment, we assumed that the movement range of SOI is known. We divided this movement range into p equal parts to set the collection of angles Θ. Then we applied UR constraints to SOI at each angle. Here is an example of how the Θ can be set:

- when p = 4, Θ = {90°, 110°, 130°, 150°},
- when p = 6, Θ = {90°, 102°, 114°, 126°, 138°, 150°},
- when p = 16, Θ = {90°, 94°, 98°, ..., 142° , 146°, 150°}.

For GC-CSV-Aux-IVE, we updated $\boldsymbol{w}_f$ 50 times while for CSV-Aux-IVE, we updated $\boldsymbol{w}_f$ 100 times. After applying GC, the convergence speed of the iterations is greatly improved, which results in the reduction of the number of convergence iterations required for convergence, compared to CSV-Aux-IVE. As a result, the convergence can be achieved in significantly fewer iterations.

[1]https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator

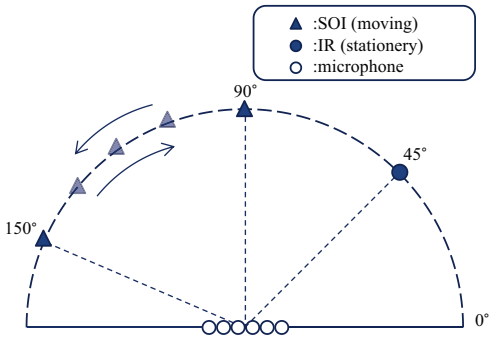


Fig. 1. Configurations of sources and microphones

TABLE I
EXPERIMENTAL CONDITIONS

Sampling rate	16 kHz
STFT window function	Hann
STFT window length	512 samples
STFT shift size	128 samples
Reverberation time	100 ms
Microphone spacing	2 cm
Input SNR	15 dB
L_b	100
λ_{GC}	0.2

TABLE II
AVERAGE SDR, SIR, AND ACCURACY OF EXTRACTING SOI (M=6)

method	SDR [dB]	SIR [dB]	Accuracy.
CSV-Aux-IVE	7.89	10.53	40.4%
GC-CSV-Aux-IVE (p=4)	11.42	15.73	97.4%
GC-CSV-Aux-IVE (p=6)	11.79	16.50	98.2%
GC-CSV-Aux-IVE (p=16)	**12.36**	**17.87**	**99.6%**

We exhibited the source extraction performance using the average source-to-distortion ratio (SDR), the source-to-interference ratio (SIR), and the accuracy of target source extraction. These evaluations were calculated by using BSS-eval [24] and setting the dry source as a reference. In order to obtain a more accurate extraction accuracy, we divided the signals into 20 blocks for comparison purposes. Hence we applied the BSS-eval to each block individually and calculated the average of the results from the 20 blocks. And we evaluated the accuracy of the extraction by assessing the output signal order.

B. Result

Table II shows the average SDR, SIR, and accuracy of the output signal. After applying GC, the accuracy of extracting SOI significantly improved from the initial 40% to over 98%. Additionally, the SDR experienced an improvement of approximately 4 dB, while the SIR showed an improvement of around 6 dB. Furthermore, we examined the performance of the extraction method in the different cases of p: $p > M$, $p = M$, and $p < M$ in Table II. Then we evaluated the extraction results under these three scenarios. Interestingly,

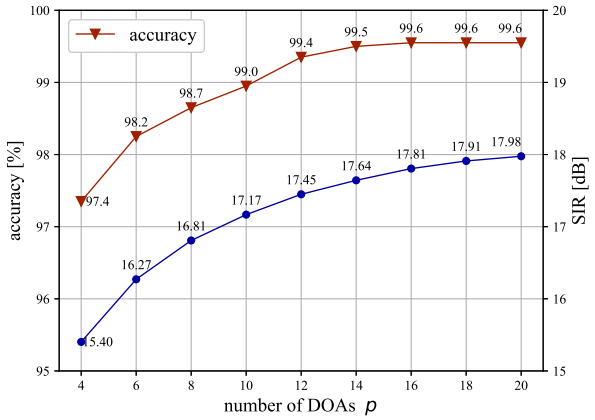


Fig. 2. Effect of increasing p on extraction performance

as the number of UR constraints p was increased, a notable improvement in the extraction performance was observed. The trend of this improvement is depicted in Fig. 2. The increasing trend of both the accuracy and SIR were observed with the increment of p.

V. CONCLUSION

This paper discovered a novel and effective BSE method for moving source. We apply a GC-based approach to address the permutation problem inherent in BSE. Our approach utilizes prior spatial information to aid in the filter updates. We extract the moving SOI by constraining multiple angles within the known range. Then we explore the effect of the changing number of constraint angles on the results. When the number of DOAs p is increased, the extraction effect is also improved. The results also demonstrated that, under similar conditions to the moving target source, GC-CSV-Aux-IVE algorithm achieved an average improvement of approximately 7dB in terms of SIR compared to the original algorithm and achieved an accuracy rate of nearly 100%.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] S. Makino, *Audio source separation*. Springer, 2018.
[2] S. Y. Low, R. Togneri, and S. Nordholm, "Spatio-temporal processing for distant speech recognition," in *Proc. ICASSP*, vol. 1, 2004, pp. I–1001.
[3] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.
[4] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[5] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. ISSPA*, vol. 2, 2003, pp. 411–414.
[6] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.
[7] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IWAENC*, 2012, pp. 1–4.

[8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.
[9] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, 2014, pp. 107–111.
[10] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. WASPAA*, 2019, pp. 185–189.
[11] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. SP*, vol. 69, pp. 3252–3267, 2021.
[12] Z. Koldovskỳ and P. Tichavskỳ, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. SP*, vol. 67, no. 4, pp. 1050–1064, 2018.
[13] J. Malek, J. Cmejla, and Z. Koldovsky, "Blind extraction of target speech source: Three ways of guidance exploiting supervised speaker embeddings," in *Proc. IWAENC*, 2022, pp. 1–5.
[14] Z. Koldovskỳ, J. Málek, and J. Janskỳ, "Extraction of independent vector component from underdetermined mixtures through block-wise determined modeling," in *Proc. ICASSP*, 2019, pp. 7903–7907.
[15] V. Kautský, Z. Koldovský, P. Tichavský, and V. Zarzoso, "Cramér–rao bounds for complex-valued independent component extraction: Determined and piecewise determined mixing models," *IEEE Trans. SP*, vol. 68, pp. 5230–5243, 2020.
[16] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP JASMP*, vol. 2022, no. 1, pp. 1–16, 2022.
[17] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.
[18] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.
[19] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. ICASSP*, 2020, pp. 846–850.
[20] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering," in *Proc. EUSIPCO*, 2022, pp. 757–761.
[21] A. H. Khan, M. Taseska, and E. A. P. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.
[22] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.
[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Universal Sound Separation Using Replay-based Data Sampling in Incremental Learning

Kanta Shimonishi*, Takahiro Fukumori* and Yoichi Yamashita*
* Ritsumeikan University, Japan

***Abstract*—In this paper, we investigate an incremental learning method in universal sound separation (USS). USS uses deep learning methods, which require learning large numbers of data. In most real-world situations, the number of available data increases, and it is necessary to learn additional data incrementally. One of the challenges of incremental learning is mitigating catastrophic forgetting, which is a performance degradation on past data. We used the replay-based method to reduce catastrophic forgetting in USS. We proposed three sampling methods for the replay-based method and compared their effectiveness for different properties of the reused data. We also compared the three proposed sampling methods with regularization-based and fine-tuning-based methods to investigate the effects of catastrophic forgetting. Experimental results show that many proposed methods reduce the performance degradation and improve the average separation performance by up to 0.36 dB compared with the fine-tuning-based method. The replay-based method is also found to be more effective than regularization-based methods.**

I. Introduction

In the society in which we live, we hear a mixture of various types of sound. Humans can selectively hear multiple sounds simultaneously occurring. This phenomenon is called the cocktail party effect. Sound separation (SS) is a technique that allows a machine to perform the cocktail party effect. SS makes it possible to separate each sound source from the mixture and listen to them separately, making it easier to recognize the sounds heard compared with the mixture. This technique is applied to noise-robust speech recognition systems and intelligible voice communication. Various methods have been used for sound source separation, such as independent component analysis (ICA) [1] and non-negative matrix factorization (NMF) [2]. In recent years, deep learning methods have been widely used to achieve high accuracy. Moreover, there has been a growing interest in SS for various sounds, not limited to speech [3], [4] and music [5], [6]. This universal sound separation (USS) [7], [8], [9] technique can separate environmental sounds around us. In this paper, USS means sound source separation for environmental sounds, and SS means sound source separation technology in general, including voice separation and music separation in addition to USS. USS is expected to enable listening to universal sounds individually, improve the accuracy of environmental sound recognition [10] and anomalous sound detection [11], [12], and assist people who are hard of hearing with difficulty distinguishing sounds.

The widespread use of smartphones and smartwatches has made data acquisition from various locations possible, and the number of available sound data may gradually increase. When training additional data, building a general-purpose model without performance degradation is possible if the previously trained data and the new data are trained together. Still, increased training costs associated with increased data number become a problem. On the other hand, if only new data are used for training, the training cost can be reduced, but overfitting to new data may cause performance degradation on past data, called catastrophic forgetting. Therefore, a low-cost training method and minimal performance degradation on past training data are needed.

In this paper, we propose three data sampling methods in the replay-based method that reuses parts of previously training data to reduce the degradation of separation performance due to catastrophic forgetting. We also compare our methods with a regularization-based method, which is widely used in incremental learning, and a fine-tuning-based method, in which only new data are trained sequentially. Our work will contribute to the realization of USS systems that take into account domain shifts caused by changes in the location of sound recordings (acoustic scenes). For the replay-based method, we compare three methods: random sampling, difficulty-based sampling, and clustering-based sampling. By comparing the three methods, we will verify what type of data can be reused to enable incremental learning with minimal performance degradation.

II. Related Works

A. USS

Most of the SS research that has been conducted so far has focused on speech and music. USS [7] is a sound separation technique for various sounds, not limited to speech and music. Previous reports on USS have proposed methods for conditioning acoustic event classification results [8] and pre-training feature extractors for each acoustic event class [9]. Defining the characteristics of sound is difficult for USS because there are many different types of sound compared with speech and music. The aforementioned previous reports have proposed methods to solve these problems.

USS mainly uses mask-based separation models based on deep learning [7], [13], [14], [15]. In mask-based SS, a mixed sound $\boldsymbol{x} \in \mathbb{R}^T$ of signal length T is first converted by an encoder into a feature representation $\boldsymbol{x'} \in \mathbb{R}^{D \times L}$, where the feature dimension is D and the time frame length is L. Using the mask $\mathcal{M}_n \in \mathbb{R}^{D \times L}$ predicted by the mask prediction network, we perform mask processing on $\boldsymbol{x'}$ to obtain a

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

feature representation of the separated sound $\boldsymbol{y}'_n \in \mathbb{R}^{D\times L}$. $n(=1,2,\cdots,N)$ is the index of the sound source, and N is the number of sound sources. The masking process is performed by solving equation (1), where $\odot$ is the element-wise multiplication.

$$\boldsymbol{y}'_n = \boldsymbol{x}' \odot \mathcal{M}_n. \tag{1}$$

$\boldsymbol{y}'_n$ is resynthesized to $\boldsymbol{y}_n \in \mathbb{R}^T$ with the decoder.

Research has also been conducted to separate particular sounds from a mixture of sounds. By embedding information about the sound to be separated into the SS model, we can isolate only particular sounds. Expressly, the model is conditioned on acoustic event labels [16], onomatopoeia [17], and descriptions of the sounds (captions) [18]. Note that the tasks in this research differ from these tasks because all sounds in the sound mixture are separated.

B. Incremental learning

Incremental learning is a framework in which a deep neural network (DNN) learns sequentially from multiple different data, also known as continuous learning or lifelong learning. Incremental learning aims to build a model that can handle newly added data while maintaining performance on previously learned data when a single DNN learns multiple data. When a DNN learns new data, it overfits the given data, which can degrade performance on previously known data. Performance degradation on past data is called catastrophic forgetting. Reusing and training all previously trained data minimize catastrophic forgetting but increases the cost and requires more time for training. In actual operation, the time required for re-training is a serious problem. Therefore, a training method using new data without reusing all previous data is necessary.

We focused on the replay-based method [19] because of the simplicity and the need to vary the conditions depending on the task. A previous study [20] has shown that the replay-based method produce different results depending on the reused data. The replay-based method reuses part of the previous training data and train the previous training data together with the new data. The regularization-based method called elastic weight consolidation (EWC) [21] adds information on previously trained model parameters to the loss function as a regularization term. Therefore, it is possible to learn without using previously trained data. The loss function $\mathcal{L}_{EWC}$ used in EWC is obtained as

$$\mathcal{L}_{EWC} = \mathcal{L}_{new} + \frac{\lambda}{2}F(\theta_{new} - \theta_{past})^2, \tag{2}$$

where $\mathcal{L}_{new}$ is the loss associated with the new task, F is the fisher information matrix of the previously learned task, θ_{new} is the parameter related to the new task, and θ_{past} is the parameter associated with the past task. Moreover, λ is the importance of the new task relative to the past tasks. The second term of equation (2) restricts the training on new data so as not to depart from the parameters related to the past tasks, thereby suppressing overfitting to new tasks.

In sound and speech, incremental learning methods have been proposed to suppress catastrophic forgetting [20], [22], [23]. In [20] and [22], a replay-based method has been proposed for multilingual speech synthesis and automatic speech recognition (ASR). In particular, in [20], multiple sampling methods are proposed on the basis of the appearance of each language. In [23], a method is proposed to reduce catastrophic forgetting in ASR by updating only a part of the ASR model.

III. Proposed data sampling methods in the replay-based method

In this section, we propose and describe in detail three methods of sampling reused data in the replay-based method.

A. Random sampling

The first proposed sampling method is random sampling. It randomly selects data to be reused from past training data. The data selection is based on uniformly distributed random numbers within the range of the numbers of training data. As shown in Fig. 1(a), the data are randomly selected from past training data, and it is difficult to determine the characteristics of the reused data compared with the other two methods. However, it is the simplest of the three proposed methods. In this paper, we define random sampling as **Method a**.

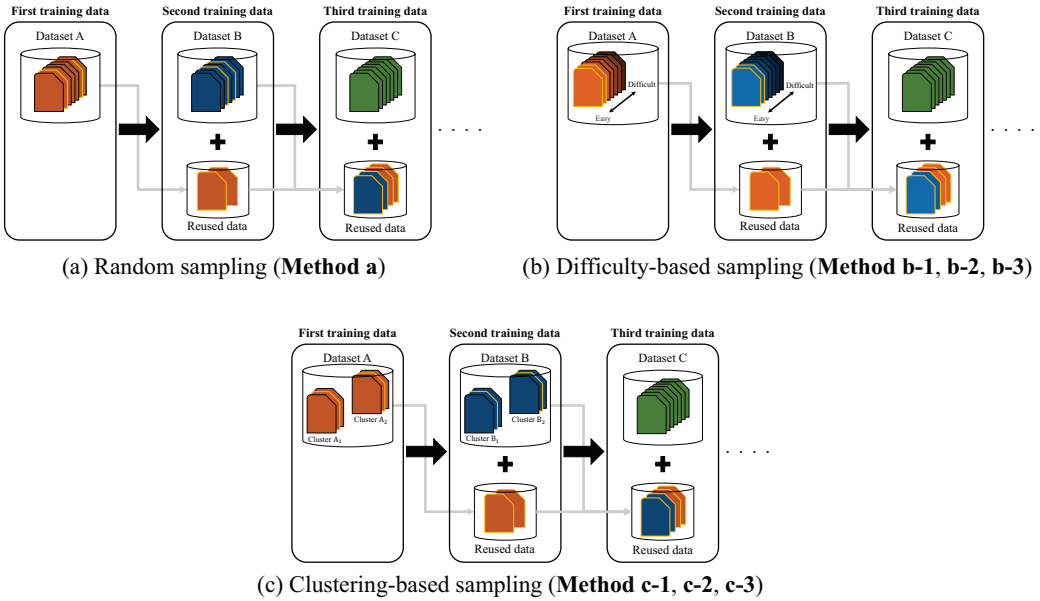
B. Difficulty-based sampling

The second method is difficulty-based sampling, which selects data according to the training difficulty of the training data. As shown in Fig. 1(b), the data are arranged in order of decreasing or increasing difficulty score and selected according to the degree of difficulty. A detailed description of difficulty-based sampling is given on the basis of Fig. 1(b). When training Dataset B newly, we prepare a difficulty-determining SS (DDSS) model trained only on Dataset A. The DDSS model is used to calculate the difficulty scores for the training data in Dataset A. In this work, difficulty scores are calculated by scale-invariant signal-to-distortion ratio improvement (SI-SDRi), which is a measure of separation performance. The details of SI-SDRi are described in Section IV-A3. The data with low separation performance are considered hard to learn. Conversely, the data with high separation performance are considered easy to learn, and the effectiveness of reused data selection is evaluated on the basis of the difference in learning difficulty. In difficulty-based sampling, the following three selection methods are compared in terms of the difference in separation performance due to the difference in learning difficulty.

Method b-1
: Reuse data with a low learning difficulty

Method b-2
: Reuse data with a moderate learning difficulty

Method b-3
: Reuse data with a high learning difficulty



(a) Random sampling (**Method a**)

(b) Difficulty-based sampling (**Method b-1**, **b-2**, **b-3**)

(c) Clustering-based sampling (**Method c-1**, **c-2**, **c-3**)

Fig. 1. Overview of reused data selection in the replay-based method

C. Clustering-based sampling

The third method is clustering-based sampling, which uses clustering to select data. The clustering is performed using the k-means method to obtain cluster centers. Using the obtained cluster centers, we select M data from the previous dataset. At this time, the reused data are selected in order of nearest to the cluster center of each class so that the total number of data is M. An overview of clustering-based sampling is shown in Fig. 1(c). The effectiveness of clustering-based sampling is verified on the basis of the degree of scatter of the reused data from the past training data, depending on the number of classes in the clustering. In this paper, we compare the following three patterns of separation performance for different numbers of classes.

Method c-1
: Clustering the training data into 1 class

Method c-2
: Clustering the training data into 4 classes

Method c-3
: Clustering the training data into M classes (the same number as that of reused data)

IV. Experiments

In this experiment, we investigate the effect of three proposed sampling methods in the replay-based method and compare them with the fine-tuning-based method and EWC.

A. Experimental setup

1) Datasets: USS with incremental learning requires multiple datasets with different characteristics. However, currently available corpora do not allow us to experiment with the problem set up in this study. Therefore, we generate multiple datasets by dividing the currently available corpora and train them sequentially. We use data from FUSS [24], a corpus for USS. FUSS contains a mixture of sounds and corresponding ground truths, each containing one background sound and 0 to 3 foreground sounds. Since the FUSS is not assigned any acoustic event labels indicating the type of sound, the data were segmented as follows.

1) Mel-frequency cepstrum coefficients (MFCC) of 20 dimensions are extracted from the background sound waveforms and averaged in the time direction. The waveforms are normalized to have a maximum amplitude of 1 and a minimum amplitude of -1.
2) The k-means clustering clusters the extracted MFCC to obtain three clusters. The number of clusters is determined by the elbow method [25].
3) Generate three datasets from the clustering results: Datasets A, B, and C.

The scatter plots of the first and second principal components extracted by principal component analysis (PCA) for the created datasets are shown in Fig. 2, and the number of training, validation, and evaluation data for the created data set is shown in Table I.

2) Separation model: We use Conv-TasNet [13] as our separation model. Conv-TasNet has high performance in the field of speech separation. Convolutional layers are used in the encoder, decoder, and mask prediction networks. The mask prediction network comprises convolutional blocks serially connected to multiple convolutional and linear layers. The outputs of the last convolutional block and the skip-connection of each convolutional block are summed, and the mask is predicted after passing through the activation function. The structure of Conv-TasNet is shown in Table II.

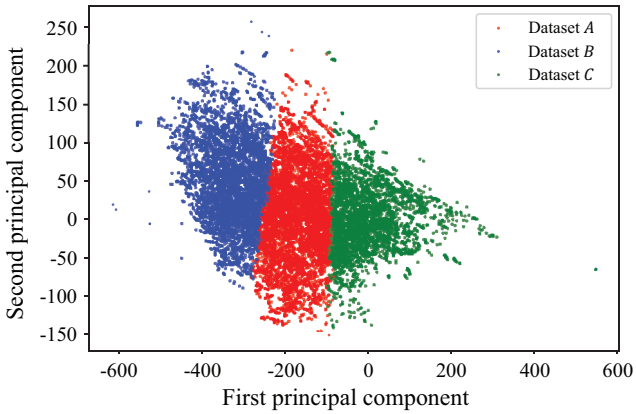


Fig. 2. PCA scatter plots of the created datasets

TABLE I
CONFIGURE OF THE CREATED DATASETS

Dataset	Train	Valid	Eval	Sum
A	7,759	435	391	8,585
B	7,683	384	306	8,373
C	4,558	181	303	5,042

TABLE II
DETAILS OF EACH PARAMETER OF CONV-TASNET

# filters in autoencoder	512
Length of filters in samples	16
# channels in bottleneck	128
# channels in skip-connection	128
# channels in convolutional blocks	512
Kernel size in convolutional blocks	3
# convolutional blocks in each repeat	8
# repeats	3
# output sources	4

3) Training and evaluation details: The training order of the datasets in incremental learning is $A \rightarrow B \rightarrow C$, and 80 epochs are trained in each training phase. The batch size is 4, Adam [26] is used as the optimizer, and the learning rate is 0.0001. SI-SDR loss [27] is used as the loss function. The SI-SDR is the distortion ratio of the separated sound to the ground truth and SI-SDR loss ($\mathcal{L}_{\text{SI-SDR}}$) is calculated as follows:

$$\text{SI-SDR}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = 10 \log_{10} \frac{\|\alpha \boldsymbol{s}\|^2}{\|\alpha \boldsymbol{s} - \hat{\boldsymbol{s}}\|^2} \quad (3)$$

$$\mathcal{L}_{\text{SI-SDR}} = -\text{SI-SDR}(\boldsymbol{s}, \hat{\boldsymbol{s}}), \quad (4)$$

where $\boldsymbol{s}$ is the ground truth and $\hat{\boldsymbol{s}}$ is the separated sound. Moreover, $\alpha = \langle \boldsymbol{s}, \hat{\boldsymbol{s}} \rangle / \|\boldsymbol{s}\|^2$ and $\langle \boldsymbol{s}, \hat{\boldsymbol{s}} \rangle$ are the inner products of $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$, respectively. Let M be 2% of the number of data in each dataset, and we compare the three sampling methods proposed in Section III. The M values for Datasets A and B are 155 and 154, respectively. In difficulty-based sampling, DDSS models of Datasets A and B have also trained 80 epochs, respectively. λ in EWC was set to 1000 by preliminary experiments.

Separation performance is evaluated on the basis of SI-SDRi. SI-SDRi is the difference between SI-SDR for separated sounds and ground truth and that for mixed sounds and ground truth. If the SI-SDRi is high, the signal can be separated with less distortion than the mixture, indicating high separation performance. In this experiment, we constructed five models on the basis of **Method a**. The average of SI-SDRi values for the five models was then employed as the result of Method a. This is because **Method a** cannot uniquely determine the sampling data compared with the other two sampling methods, and bias may occur. Therefore, to reduce the bias, we use the average of SI-SDRi values for the five models.

B. Results

The experimental results are shown in Tables III to VII. Table III shows the results of incremental learning without replay/regularization-based methods, A_L, B_L, and C_L are the training data, and A_E, B_E, and C_E are the evaluation data in Datasets A, B, and C, respectively.

Table III shows that the separation performance for A_E decreased from 6.64 dB to 6.37 dB without the replay-based and regularization-based methods, and the same is true for B_E. This indicates that catastrophic forgetting reduces the separation performance for past data sets. Tables IV to VI are the results of data selection by the proposed random sampling, difficulty-based sampling, and clustering-based sampling methods in the replay-based method, respectively. Looking at each proposed method, the separation performance for A_E and B_E is degraded, indicating that catastrophic forgetting has occurred. However, when compared with Table III, it can be seen that the separation performance for each evaluated data is reduced for most of the methods. Table VII also shows that even when EWC is used, a slight reduction in catastrophic forgetting is observed for each evaluation data. Still, the overall performance is lower than that of the replay-based method. This indicates that the replay and regularization-based methods are effective in USS with incremental learning.

Next, we analyze the three proposed data sampling methods in the replay-based method. In difficulty-based sampling, the best performance is achieved by **Method b-3** after $A_L \rightarrow B_L$ training, the best performance is achieved by **Method b-1** after $A_L \rightarrow B_L \rightarrow C_L$ training, and **Method b-2** has the lowest performance in all training phases. From the SI-SDRi histograms of the previous training data in Fig. 3, the SI-SDRi of the data selected by **Method b-2** is locally concentrated, and there is a slight variation in the SI-SDRi, whereas **Methods b-1** and **b-3** show the lowest separation performance. **Methods b-1** and **b-3** show some variation. Therefore, the variation of SI-SDRi of the reused data in difficulty-based sampling may affect the separation performance during continuous learning.

Clustering-based sampling shows that the highest separation performance is achieved using **Method c-1** at all training

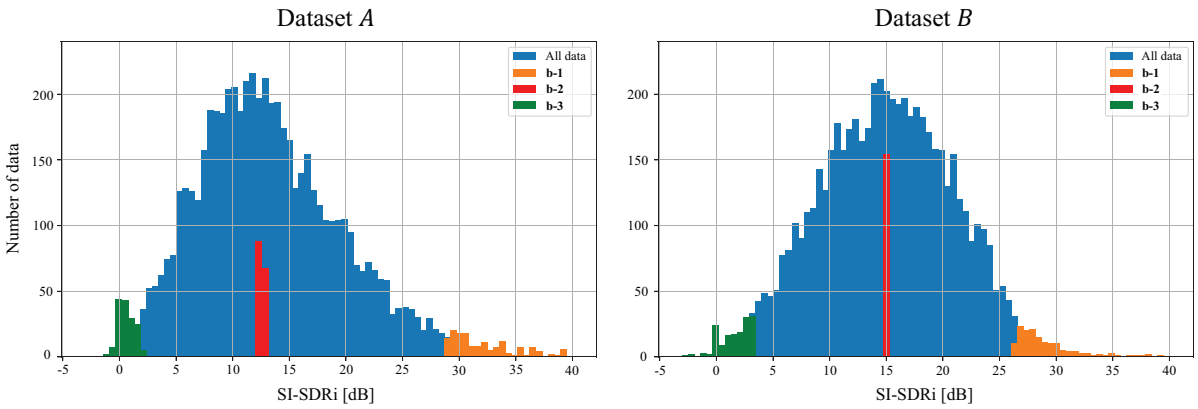


Fig. 3. Histograms of SI-SDRi for past training data for difficulty-based sampling

TABLE III
SI-SDRI IN FINE TUNING METHOD (BASELINE) [DB]

Training data	Evaluation data			
	A_E	B_E	C_E	Avg
(1) A_L	6.64	-	-	-
(2) $A_L \rightarrow B_L$	6.21	7.33	-	6.77
(3) $A_L \rightarrow B_L \rightarrow C_L$	6.37	6.16	8.17	6.90

TABLE IV
SI-SDRI IN RANDOM SAMPLING (**METHOD A**) [DB]

Training data	Evaluation data			
	A_E	B_E	C_E	Avg
(1) A_L	6.64	-	-	-
(2) $A_L \rightarrow B_L$	6.47	7.38	-	6.93
(3) $A_L \rightarrow B_L \rightarrow C_L$	6.44	6.54	8.23	7.07

TABLE V
SI-SDRI IN DIFFICULTY-BASED SAMPLING [DB]

Training data	Evaluation data											
	Method b-1				**Method b-2**				**Method b-3**			
	A_E	B_E	C_E	Avg	A_E	B_E	C_E	Avg	A_E	B_E	C_E	Avg
(1) A_L	6.64	-	-	-	6.64	-	-	-	6.64	-	-	-
(2) $A_L \rightarrow B_L$	6.57	7.43	-	7.00	6.37	7.38	-	6.88	6.74	7.52	-	**7.13**
(3) $A_L \rightarrow B_L \rightarrow C_L$	6.37	6.62	8.28	**7.09**	6.09	6.53	7.80	6.81	6.37	6.45	8.25	7.02

TABLE VI
SI-SDRI IN CLUSTERING-BASED SAMPLING [DB]

Training data	Evaluation data											
	Method c-1				**Method c-2**				**Method c-3**			
	A_E	B_E	C_E	Avg	A_E	B_E	C_E	Avg	A_E	B_E	C_E	Avg
(1) A_L	6.64	-	-	-	6.64	-	-	-	6.64	-	-	-
(2) $A_L \rightarrow B_L$	6.41	7.45	-	6.93	5.97	7.41	-	6.69	6.31	7.37	-	6.84
(3) $A_L \rightarrow B_L \rightarrow C_L$	6.19	6.52	8.44	7.05	6.26	6.52	8.15	6.98	6.37	6.59	7.95	6.97

TABLE VII
SI-SDRI IN EWC [DB]

Training data	Evaluation data			
	A_E	B_E	C_E	Avg
(1) A_L	6.64	-	-	-
(2) $A_L \rightarrow B_L$	6.23	7.29	-	6.74
(3) $A_L \rightarrow B_L \rightarrow C_L$	6.43	6.29	8.24	6.99

phases. By changing the number of clusters in clustering, we tested the effectiveness of selecting groups evenly or biasedly from the past training data. Table VI shows that selecting data biased toward a single location is practical. This may be because clustering was carried out using k-means when the datasets were generated. The cluster centers obtained by **Method c-1** are the average features of each dataset. These cluster centers are the same as those obtained when FUSS was clustered at the time of datasets generation. This is likely because data with average characteristics of each dataset were selected.

The separation performance of the seven data selection methods, **Method a**, **Methods b-1** to **b-3**, and **Methods c-1** to **c-3**, differed, and it was also found that the separation performance after incremental learning varied depending on the reused data. The separation performance after incremental learning also varied depending on the reused data. Among the seven methods, the difficulty-based sampling method

(**Methods b-1** and **b-3**) performed the best in each training phase. From the discussion of difficulty-based sampling mentioned above, it is clear that the method of selecting data with a variation in the SI-SDRi of the past training data is the best at each training phase. It is necessary to consider a method to select data with variations in the SI-SDRi of the past training data.

V. Conclusions

In this paper, we investigated USS methods using incremental learning and proposed sampling methods for reusing data in the replay-based method. We also compared our proposed methods with the fine-tuning-based method and regularization-based method. For the replay-based method, we proposed three reused data selection methods: random sampling, difficulty-based sampling, and clustering-based sampling, and verified what type of data is adequate for relearning. The experimental results showed that most of the methods could reduce catastrophic forgetting. Difficulty-based sampling, which was the most effective method, was even more effective when the data were sampled evenly without any bias in learning difficulty. In this study, we divided publicly available data on the basis of acoustic features to generate multiple datasets. The same experiments should be conducted with semantic data such as acoustic event and scene labels.

References

[1] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Two-stage blind source separation based on ica and binary masking for real-time robot audition system," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 2303–2308.

[2] A. Cichocki, R. Zdunek, and S.-i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, 2006, pp. V–V.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "Teacher-Student MixIT for Unsupervised and Semi-Supervised Speech Separation," in *Proc. INTERSPEECH*, 2021, pp. 3495–3499.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[6] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[7] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2019, pp. 175–179.

[8] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. W. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 96–100.

[9] H. Munakata, R. Takeda, and K. Komatani, "Multiple-embedding separation networks: Sound class-specific feature extraction for universal sound separation," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), 2021, pp. 961–967.

[10] N. Turpault, S. Wisdom, H. Erdogan, J. R. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving Sound Event Detection in Domestic Environments using Sound Separation," in *Proc. Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2020, pp. 205–209.

[11] A. Aiba, M. Yoshida, D. Kitamura, S. Takamichi, and H. Saruwatari, "Noise robust acoustic anomaly detection system with nonnegative matrix factorization based on generalized Gaussian distribution," *IEICE TRANSACTIONS on Information and Systems*, vol. 104, no. 3, pp. 441–449, 2021.

[12] K. Shimonishi, K. Dohi, and Y. Kawaguchi, "Anomalous Sound Detection Based on Sound Separation," in *Proc. INTERSPEECH*, 2023, pp. 2733–2737.

[13] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "DF-Conformer: Integrated architecture of Conv-TasNet and Conformer using linear complexity self-attention for speech enhancement," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2021, pp. 161–165.

[15] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (*MLSP*), 2020, pp. 1–6.

[16] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to What You Want: Neural Network-Based Universal Sound Selector," in *Proc. INTERSPEECH*, 2020, pp. 1441–1445.

[17] Y. Okamoto, S. Horiguchi, M. Yamamoto, K. Imoto, and Y. Kawaguchi, "Environmental sound extraction using onomatopoeic words," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 221–225.

[18] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate What You Describe: Language-Queried Audio Source Separation," in *Proc. INTERSPEECH*, 2022, pp. 1801–1805.

[19] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[20] M. Yang, S. Ding, T. Chen, T. Wang, and Z. Wang, "Towards lifelong learning of multilingual text-to-speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 8022–8026.

[21] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[22] B. Li, R. Pang, Y. Zhang, T. N. Sainath, T. Strohman, P. Haghani, Y. Zhu, B. Farris, N. Gaur, and M. Prasad, "Massively multilingual asr: A lifelong learning solution," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 6397–6401.

[23] Y. Takashima, S. Horiguchi, S. Watanabe, L. P. G. Perera, and Y. Kawaguchi, "Updating Only Encoders Prevents Catastrophic Forgetting of End-to-End ASR Models," in *Proc. INTERSPEECH*, 2022, pp. 2218–2222.

[24] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2021, pp. 186–190.

[25] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method," in *Journal of Physics: Conference Series*, vol. 1361, no. 1. IOP Publishing, 2019, p. 012015.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] S. Li, H. Liu, Y. Zhou, and Z. Luo, "A SI-SDR Loss Function based Monaural Source Separation," in *Proc. International Conference on Signal Processing* (*ICSP*), vol. 1, 2020, pp. 356–360.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multiple Sound Source Tracking Based on Generative Modeling and Recursive Bayesian Filtering of Spatial Gradient Spectra

Keisuke Takazawa*, Hirokazu Kameoka†, and Masahiro Yukawa*
* Keio University, Japan
E-mail: yukawa@elec.keio.ac.jp Tel: +81-45-566-1521
† NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan
E-mail: hirokazu.kameoka@ntt.com Tel/Fax: +81-46-240-3645

***Abstract*—This paper addresses the problem of tracking the positions of multiple sound sources using a compact-sized microphone array. Based on a generative model of the spatial gradient spectra of acoustic point sources, we propose incorporating the temporal continuity of source positions into the sequential posterior inference under a recursive Bayesian filter formulation. Simulation experiments with data generated by the image source model confirmed that the proposed idea had a certain effect on improving the accuracy of tracking moving sources.**

I. Introduction

This paper addresses the problem of tracking the positions of multiple "moving" sound sources using a compact array of microphones locally placed. Although the source is assumed to be moving in the present study, let us start by reviewing the sound source localization (SSL) methods which rely on the assumption that the source positions are "fixed". Indeed, many powerful SSL methods have already been proposed, representing the well-known multiple signal classification (MUSIC) algorithm [1] and generalized cross-correlation methods with phase transform (GCC-PHAT) [2] and their variants. Other methods have also been proposed, such as probabilistic generative model-based [3]–[7], deep neural network (DNN)-based [8], [9], and partial differential equation (PDE)-based [7], [10] methods.

Since most SSL methods enable SSL by means of information on the time-difference-of-arrival of each sound source measured by multiple microphones, it is generally preferable to use a large array with many microphones arranged in a spatially wide area. In contrast, the PDE-based methods, which utilize the PDE governing the propagation of the acoustic waves emitted by each sound source, are attractive because they can achieve SSL through a short-term observation of the sound pressure and its spatial gradient at a certain location, which can measured by a relatively compact array of locally placed microphones. This is an important feature, especially when it comes to realizing an SSL functionality on mobile objects such as robots and drones.

Although the DNN-based methods have proven to be very accurate under certain conditions, a potential drawback is the unpredictability of model behavior when used in acoustic environments that do not match the training conditions. In contrast, generative model-based methods are appealing in that they can be designed to flexibly adapt their models to various acoustic environments through model adaptation.

Focusing on the advantages of the PDE-based and generative model-based methods, an attempt to combine these ideas has already been made [7]. The idea is based on the assumption that only one source is dominant in each time-frequency (TF) bin of the observed signal and that its component is generated according to the PDE corresponding to the dominant source. Through simulation experiments, this method has been shown to work well with a compact microphone array even in reverberant environments. However, this method was designed to operate independently for each short-term frame, and the continuity of the location of each source was not considered in the inference. In order to enhance the accuracy and robustness of this method, this paper proposes an improvement that takes into account the continuity of the location of each sound source under the recursive Bayesian filter framework [11].[1]

II. Method

A. Sound Source Constraint PDE [10]

First, we briefly review the theory of the generative modeling of the spatial gradient spectra of a single acoustic point source. Let $\mathbf{r}_0 \in \mathbb{R}^3$ be the position of the point source of interest. By assuming the free-field (spherical-wave) propagation from that source, the sound pressure $f(\mathbf{r}, t)$ at an observation position $\mathbf{r} \in \mathbb{R}^3$ and time $t \in \mathbb{R}$ is given by

$$f(\mathbf{r}, t) = \frac{1}{\|\mathbf{r} - \mathbf{r}_0\|} g\left(t - \frac{\|\mathbf{r} - \mathbf{r}_0\|}{c}\right), \tag{1}$$

where g is the source signal, c is the speed of sound, and $\|\cdot\|$ denotes the Euclidean norm. From Eq. (1), the spatial gradient ∇f is given as

$$\nabla f(\mathbf{r}, t) = \rho^2 g\left(t - \frac{1}{\rho c}\right)\mathbf{n} + \frac{\rho}{c}\dot{g}\left(t - \frac{1}{\rho c}\right)\mathbf{n}, \tag{2}$$

[1]Temporal continuity of moving sources has been exploited in some other areas such as robotics [12].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

where $\rho := 1/\|\mathbf{r}-\mathbf{r}_0\|$, $\mathbf{n} := (\mathbf{r}-\mathbf{r}_0)/\|\mathbf{r}-\mathbf{r}_0\|$, and $\dot{g}$ denotes the time derivative of g. Namely, $\mathbf{n}$ and ρ correspond to the direction and reciprocal distance of the sound source from point $\mathbf{r}$, respectively. The time derivative $\dot{f}$ is given as

$$\dot{f}(\mathbf{r},t) = \rho \dot{g}\left(t - \frac{1}{\rho c}\right). \tag{3}$$

Substituting Eqs. (1) and (3) into Eq. (2), we obtain an equation governing the relationship between the sound pressure of the source signal observed at $(\mathbf{r},t)$, its spatial gradient, and its time-derivative

$$\nabla f(\mathbf{r},t) = \left(\rho f(\mathbf{r},t) + \frac{1}{c}\dot{f}(\mathbf{r},t)\right)\mathbf{n}. \tag{4}$$

Thus, if the sound pressure of the source signal, its spatial gradient, and its time-derivative can be measured at a certain position, the location of the sound source can be estimated by finding $\mathbf{n}$ and ρ that are as consistent as possible with Eq. (4). This equation is called the sound source constraint PDE (SSC-PDE) [10].

B. SSC-PDE-based Generaive Modeling [7]

Eq. (4) can be written in the frequency domain as

$$-\nabla F(\mathbf{r},\omega) + \left(\rho + \mathrm{j}\frac{\omega}{c}\right)F(\mathbf{r},\omega)\mathbf{n} = \mathbf{0}, \tag{5}$$

where $\mathbf{0}$ is a three-dimensional zero vector, F is the temporal Fourier transform of f, ω is the angular frequency, and j is the imaginary unit. ∇F is the spatial gradient spectrum of f. Let us use $\{\omega_m\}_m$ to denote the set of discrete frequencies, $F_{0,m}$ to denote $F(\mathbf{r},\omega_m)$, and $F_m^{(x)}$, $F_m^{(y)}$, and $F_m^{(z)}$ to denote the (x,y,z)-elements of $\nabla F(\mathbf{r},\omega_m)$. When $F_m^{(x)}$, $F_m^{(y)}$, and $F_m^{(z)}$ are approximated by the spatial differences measured using, for example, a microphone array shown in Fig. 1, Eq. (5) holds only approximately and the right-hand side of Eq. (5) can slightly deviate from $\mathbf{0}$. Here, we assume this deviation to be complex Gaussian noise. Namely, we replace the right-hand side of Eq. (5) with a random vector $\boldsymbol{\epsilon}_m = [\epsilon_m^{(x)}, \epsilon_m^{(y)}, \epsilon_m^{(z)}]^\mathsf{T}$ that follows a zero-mean complex normal distribution

$$\boldsymbol{\epsilon}_m \sim \mathcal{N}_\mathbb{C}(\boldsymbol{\epsilon}_m; \mathbf{0}, \mathbf{I}/\lambda_m), \tag{6}$$

where $\mathbf{I}$ denotes a 3×3 identity matrix and λ_m denotes the precision of each element of $\boldsymbol{\epsilon}_m$. For simplicity, we will henceforth assume that λ_m is an experimentally fixed constant. We further assume $F_{0,m}$ to follow a zero-mean complex normal distribution with source-dependent precision ζ_m:

$$F_{0,m} \sim \mathcal{N}_\mathbb{C}(F_{0,m}; 0, 1/\zeta_m). \tag{7}$$

Namely, $1/\zeta_m$ corresponds to the power spectral density of the source signal observed at point $\mathbf{r}$. Now, if we use $\mathbf{y}_m$ and $\mathbf{C}_m(\theta)$ to denote

$$\mathbf{y}_m = \left[F_m^{(x)}, F_m^{(y)}, F_m^{(z)}, F_{0,m}\right]^\mathsf{T}, \tag{8}$$

$$\mathbf{C}_m(\theta) = \begin{bmatrix} -1 & & & \left(\rho + \mathrm{j}\frac{\omega_m}{c}\right)n_x \\ & -1 & & \left(\rho + \mathrm{j}\frac{\omega_m}{c}\right)n_y \\ & & -1 & \left(\rho + \mathrm{j}\frac{\omega_m}{c}\right)n_z \\ & & & 1 \end{bmatrix}, \tag{9}$$

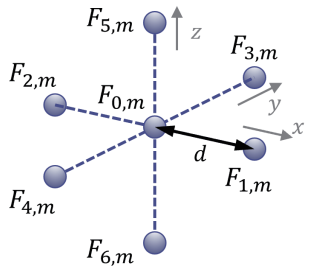


Fig. 1. Example of microphone placement. $F_m^{(x)}$, $F_m^{(y)}$, $F_m^{(z)}$ can be approximated by $(F_{1,m}-F_{2,m})/2d$, $(F_{3,m}-F_{4,m})/2d$, and $(F_{5,m}-F_{6,m})/2d$, respectively.

respectively, where $\theta = \{\mathbf{n}, \rho\}$, Eq. (5) can be rewritten as

$$\mathbf{C}_m(\theta)\mathbf{y}_m = [\boldsymbol{\epsilon}_m^\mathsf{T}, F_{0,m}]^\mathsf{T}. \tag{10}$$

Since $\mathbf{C}_m(\theta)$ is invertible and $\det \mathbf{C}_m(\theta) = -1$, from Eqs. (6), (7), and (10), the probability density function (PDF) of $\mathbf{y}_m$ given θ is obtained as

$$p(\mathbf{y}_m|\theta) = \mathcal{N}_\mathbb{C}(\mathbf{y}_m; \mathbf{0}, \mathbf{C}_m(\theta)^{-1}\boldsymbol{\Sigma}_m\mathbf{C}_m(\theta)^{-\mathsf{H}}), \tag{11}$$

where $\boldsymbol{\Sigma}_m = \mathrm{diag}(1/\lambda_m, 1/\lambda_m, 1/\lambda_m, 1/\zeta_m)$. Thus, given the observation $Y = \{\mathbf{y}_m\}_m$, $p(Y|\theta) = \prod_m p(\mathbf{y}_m|\theta)$ implies the likelihood of the source position θ [7].

C. Formulation of Multi-Source Case [7]

In the following, we will extend the above model to the case where there are multiple sound sources. Let $k \in \mathbb{N}$ denote the sound source index, where $k \neq 0$ corresponds to one of the point sources present and $k = 0$ corresponds to a mixture of all noise sources, including the reverberant components of the point sources.

Many real-world sound sources, such as speech and instrumental sounds, tend to have sparse TF representations. Hence, in many cases, each TF element of a mixture of multiple sources can be assumed to be dominated by at most one source. Recall that Eq. (11) describes a generative model that is justified in the frequency domain. We use it here as the generative model for each short-time frame signal, with frame index l. Assuming that the $z_{m,l}$-th source is active at the (m,l)-th TF bin, the PDF of the observed signal is given as

$$p(\mathbf{y}_{m,l}|z_{m,l}=k, \vartheta_l) = \mathcal{N}_\mathbb{C}(\mathbf{y}_{m,l}; \mathbf{0}, \boldsymbol{\Lambda}_{m,l}^{(k)}), \tag{12}$$

where

$$\boldsymbol{\Lambda}_{m,l}^{(k)} = \begin{cases} \mathbf{C}_{m,l}(\theta_l^{(k)})^{-1}\boldsymbol{\Sigma}_{m,l}^{(k)}\mathbf{C}_{m,l}(\theta_l^{(k)})^{-\mathsf{H}}, & (k \neq 0), \\ \mathbf{W}_m/\gamma_{m,l}, & (k = 0), \end{cases} \tag{13}$$

$$\boldsymbol{\Sigma}_{m,l}^{(k)} = \mathrm{diag}(1/\lambda_m, 1/\lambda_m, 1/\lambda_m, 1/\zeta_{m,l}^{(k)}). \tag{14}$$

Here, ϑ_l is used to denote the set of all the unknown parameters, $\{\theta_l^{(k)}\}_k$, $\{\zeta_{m,l}^{(k)}\}_{k,m}$, and $\{\gamma_{m,l}\}_m$, where $\theta_l^{(k)} = \{\mathbf{n}_l^{(k)}, \rho_l^{(k)}\}$. It should be noted here that the source position parameter $\theta_l^{(k)}$ is indexed also by time l, assuming that each source can move from moment to moment. As described above, the covariance matrix for the noise mixture is factorized

into channel-invariant and time-invariant factors, $1/\gamma_{m,l}$ and $\mathbf{W}_m$. The former and latter correspond to the power spectral density and normalized spatial covariance of the noise mixture, respectively. Note that if we assume a diffuse noise field, $\mathbf{W}_m$ can be set fixed using the diffuse model introduced in [13].

Modeling the prior distribution $p(z_{m,l})$ using a Dirichlet process allows us to handle an unknown number of sources [14]. The point is to first assume the existence of an infinite number of sound sources, one of which is drawn from a categorical distribution with infinite categories

$$p(z_{m,l} = k) = \pi_k \quad (k = 0, \ldots, \infty) \tag{15}$$

for each TF bin. The probability π_0, corresponding to the noise mixture, is assumed to be drawn according to a Beta distribution with hyperparameters α_{n} and β_{n}. The remaining probabilities $\pi_1, \ldots, \pi_\infty$, corresponding to the point sources, are assumed to be drawn according to a stick-breaking process with hyperparameter β_{s}:

$$\pi_k = v_k \prod_{j=0}^{k-1} (1 - v_j) \quad (k \geq 1), \tag{16}$$

$$v_k \sim \mathrm{Beta}(1, \beta_{\mathrm{s}}). \tag{17}$$

D. Posterior Inference

Let Θ_l denote the set consisting of $N_l = \{\mathbf{n}_l^{(k)}\}_k$, $\rho_l = \{\rho_l^{(k)}\}_k$, $\zeta_l = \{\zeta_{m,l}^{(k)}\}_{k,m}$, $\gamma_l = \{\gamma_{m,l}\}_m$, $V = \{v_k\}_k$, and $Z_l = \{z_{m,l}\}_m$. Given an observed signal $Y_l = \{\mathbf{y}_{m,l}\}_m$, the best approximating distribution $q(\Theta_l)$ that minimizes the Kullback-Leibler (KL) divergence between $q(\Theta_l)$ and the true posterior distribution $p(\Theta_l|Y_l)$ can be searched for using the variational Bayes inference (VB) algorithm [15]. By using an approximating distribution $q(\Theta_l)$ that factors as

$$q(\Theta_l) = q(N_l)q(\rho_l)q(\zeta_l)q(\gamma_l)q(V)q(Z_l), \tag{18}$$

each distribution becomes a parametric distribution of a particular form. For example, $q(N_l)$ is given in the form [7]

$$q(N_l) = \prod_k \mathrm{VMF}(\mathbf{n}_l^{(k)}; \bar{\mathbf{n}}_l^{(k)}, \kappa_l^{(k)}), \tag{19}$$

where $\mathrm{VMF}(\mathbf{x}; \mathbf{a}, b)$ denotes a von Mises-Fisher distribution with the mean direction $\mathbf{a}$ and the concentration parameter b. If we assume $p(N_l) = \prod_k \mathrm{VMF}(\mathbf{n}_l^{(k)}; \boldsymbol{\xi}_l^{(k)}, \tau_l^{(k)})$ to be the prior distribution of N_l (which becomes a uniform distribution when $\tau_l^{(k)} = 0$), the update equations of the VB algorithm for $\bar{\mathbf{n}}_l^{(k)}$ and $\kappa_l^{(k)}$ will be given as

$$\bar{\mathbf{n}}_l^{(k)} \leftarrow (\boldsymbol{\phi}_l^{(k)} + \tau_l^{(k)}\boldsymbol{\xi}_l^{(k)})/\|\boldsymbol{\phi}_l^{(k)} + \tau_l^{(k)}\boldsymbol{\xi}_l^{(k)}\|, \tag{20}$$

$$\kappa_l^{(k)} \leftarrow \|\boldsymbol{\phi}_l^{(k)} + \tau_l^{(k)}\boldsymbol{\xi}_l^{(k)}\|, \tag{21}$$

where

$$\boldsymbol{\phi}_l^{(k)} = \sum_m \bar{\varphi}_{m,l}^{(k)} \lambda_m \begin{bmatrix} \mathrm{Re}[(\bar{\rho}_l^{(k)} + \mathrm{j}\frac{\omega_m}{c}) F_{0,m,l} F_{m,l}^{(x)*}] \\ \mathrm{Re}[(\bar{\rho}_l^{(k)} + \mathrm{j}\frac{\omega_m}{c}) F_{0,m,l} F_{m,l}^{(y)*}] \\ \mathrm{Re}[(\bar{\rho}_l^{(k)} + \mathrm{j}\frac{\omega_m}{c}) F_{0,m,l} F_{m,l}^{(z)*}] \end{bmatrix}. \tag{22}$$

Here, $\bar{\varphi}_{m,l}^{(k)}$ and $\bar{\rho}_l^{(k)}$ represent $q(z_{m,l} = k)$ and the mean of $q(\rho_l{}^{(k)})$ obtained at the previous iteration, respectively. x^* denotes the complex conjugate of x. See [7] for the details of the derivation and the distribution form and update formula for each of the remaining parameters.

E. Recursive Bayesian Filter Formulation

The formulation up to this point did not consider the temporal continuity of the location of each sound source. In the following, we formulate the generative model of a mixture of moving sound sources under the recursive Bayesian filter framework [11], known as a generalization of the Kalman filter.

Let $x_{l_0:l_1}$ denote the abbreviation of the set $\{x_{l_0}, \ldots, x_{l_1}\}$. The main focus here is to infer the posterior $p(\Theta_l|Y_{1:l})$ under the assumption that $p(\Theta_{l-1}|Y_{1:l-1})$ is already given. First, we assume that $p(\Theta_{1:l})$ follows a Markov process such that

$$p(\Theta_l|\Theta_{1:l-1}) = p(\Theta_l|\Theta_{l-1}). \tag{23}$$

The assumption of the temporal continuity of the source positions can be imposed through the design of the transition probability $p(\Theta_l|\Theta_{l-1})$. By using $p(\Theta_{l-1}|Y_{1:l-1})$ and $p(\Theta_l|\Theta_{l-1})$, $p(\Theta_l|Y_{1:l-1})$ can be computed via

$$p(\Theta_l|Y_{1:l-1}) = \int p(\Theta_l|\Theta_{l-1}) p(\Theta_{l-1}|Y_{1:l-1}) \mathrm{d}\Theta_{l-1}. \tag{24}$$

This process corresponds to the prediction step in the Kalman filter. If we assume that Θ_l varies gradually over time, then it is reasonable to assume that $p(\Theta_l|\Theta_{l-1})$ is a unimodal distribution with a peak at Θ_{l-1}. In this case, Eq. (24) can be viewed as a convolution of $p(\Theta_l|\Theta_{l-1})$ and $p(\Theta_{l-1}|Y_{1:l-1})$, with $p(\Theta_l|\Theta_{l-1})$ serving to broaden $p(\Theta_{l-1}|Y_{1:l-1})$ around the peak.

Once $p(\Theta_l|Y_{1:l-1})$ is obtained, $p(\Theta_l|Y_{1:l})$ is finally obtained using the Bayes rule as

$$p(\Theta_l|Y_{1:l}) = \frac{p(Y_l|\Theta_l) p(\Theta_l|Y_{1:l-1})}{\int p(Y_l|\Theta_l) p(\Theta_l|Y_{1:l-1}) \mathrm{d}\Theta_l}, \tag{25}$$

with $p(\Theta_l|Y_{1:l-1})$ serving as the prior distribution. This process corresponds to the update step in the Kalman filter. Eqs. (24) and (25) imply that performing these two steps recursively allows us to infer the posterior distributions from $p(\Theta_1|Y_1)$ to $p(\Theta_l|Y_{1:l})$ sequentially. Note that the computation of Eq. (25) is analytically intractable, but as we saw in the previous section, a distribution that approximates it can be obtained with the VB algorithm.

In this study, we assume that only N_l follows a Markov process, and the remaining parameters are uniformly distributed and independent for each frame. Hence, $p(\Theta_l|\Theta_{l-1}) \propto \prod_k p(\mathbf{n}_l^{(k)}|\mathbf{n}_{l-1}^{(k)})$. Assuming that an approximation $q(\Theta_{l-1})$ of the posterior $p(\Theta_{l-1}|Y_{1:l-1})$ is already obtained using the VB algorithm, we define $p(\mathbf{n}_l^{(k)}|\mathbf{n}_{l-1}^{(k)})$ so that Eq. (24) can be written as

$$p(\Theta_l|Y_{1:l-1}) \propto \prod_k \mathrm{VMF}\left(\mathbf{n}_l^{(k)}; \bar{\mathbf{n}}_{l-1}^{(k)}, \delta\kappa_{l-1}^{(k)}\right), \tag{26}$$

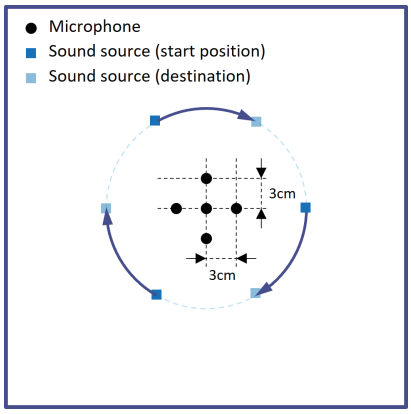


Fig. 2. Microphone and sound source arrangement

where $\bar{\mathbf{n}}_{l-1}^{(k)}$ and $\kappa_{l-1}^{(k)}$ are the mean direction and concentration parameter of $q(\mathbf{n}_{l-1}^{(k)})$, and δ is a non-negative hyperparameter that is responsible for sharpening or widening $q(\mathbf{n}_{l-1}^{(k)})$ around the peak. When $\delta = 0$, $p(\mathbf{n}_l^{(k)}|Y_{1:l-1})$ always becomes a uniform distribution, independent of the shape of $q(\mathbf{n}_{l-1}^{(k)})$.

As described in Subsection II-D, as long as the prior distribution of $\mathbf{n}_l^{(k)}$ is given in the form of a von Mises-Fisher distribution, the approximating distribution $q(\mathbf{n}_l^{(k)})$ in VB will be given in the same form. From Eq. (26), the update equations for $\bar{\mathbf{n}}_l^{(k)}$ and $\kappa_l^{(k)}$ in the VB algorithm for approximating Eq. (25) are given as

$$\bar{\mathbf{n}}_l^{(k)} \leftarrow (\boldsymbol{\phi}_l^{(k)} + \delta\hat{\kappa}_{l-1}^{(k)}\hat{\mathbf{n}}_{l-1}^{(k)})/\|\boldsymbol{\phi}_l^{(k)} + \delta\hat{\kappa}_{l-1}^{(k)}\hat{\mathbf{n}}_{l-1}^{(k)}\|, \quad (27)$$

$$\kappa_l^{(k)} \leftarrow \|\boldsymbol{\phi}_l^{(k)} + \delta\hat{\kappa}_{l-1}^{(k)}\hat{\mathbf{n}}_{l-1}^{(k)}\|, \quad (28)$$

where $\hat{\mathbf{n}}_{l-1}^{(k)}$ and $\hat{\kappa}_{l-1}^{(k)}$ represent the convergent values of $\bar{\mathbf{n}}_{l-1}^{(k)}$ and $\kappa_{l-1}^{(k)}$ in the VB algorithm performed at frame $l-1$, respectively. The update equations for the remaining parameters are the same as in the previous section. Note that when $\delta = 0$, the prior distribution will have no effect on inference at the current frame, and thus the above algorithm reduces to the frame-independent version described in Subsection II-D.

III. Simulation Experiments

A. Experimental Conditions

Simulation experiments were conducted to verify the effectiveness of the proposed method. Note here that the existing SSL methods cannot be applied to the present case because the source is unfixed and is moving over time. Experimental data were generated using the image source model under the microphone and sound source arrangement shown in Fig. 2. Each sound source was moved clockwise around the reference point of the microphone array at a speed of 6 degrees per second for 10s, with Fig. 2 as the initial and final placement. Each source was initially placed at 120 degree intervals. All the signals were sampled at 32kHz. See [7] for other experimental conditions.

In the experiment, several values of δ were tested to verify the effect of incorporating the temporal continuity of the source location into the inference.

B. Results

Figure 3 shows the estimated and true source positions for (a) $\delta = 0$, (b) $\delta = 1.0$, (c) $\delta = 6.4$, and (d) $\delta = 7.0$, respectively. As can be seen from Fig. 3(a), when the baseline method (frame-by-frame independent inference) was applied, the estimated source positions were temporally discontinuous and tended to deviate from the true positions. In contrast, Fig. 3(c) shows that the accuracy of inference was improved by incorporating the temporal continuity of the sound source location into the inference using the recursive Bayesian filtering approach. However, we also found, as shown in Fig. 3(d), that inference could instead deteriorate if δ is set too large. Therefore, finding an appropriate value for δ is an issue for future study.

IV. Conclusions

This paper addressed the problem of tracking the positions of multiple sound sources using a compact-sized microphone array. As a solution, we focused on the generative modeling formulation of the SSC-PDE-based SSL approach and proposed incorporating the temporal continuity of source positions into the inference based on a recursive Bayesian filter formulation. Simulation experiments with data generated by the image source model confirmed that the proposed idea had a certain effect on improving the accuracy of tracking moving sources.

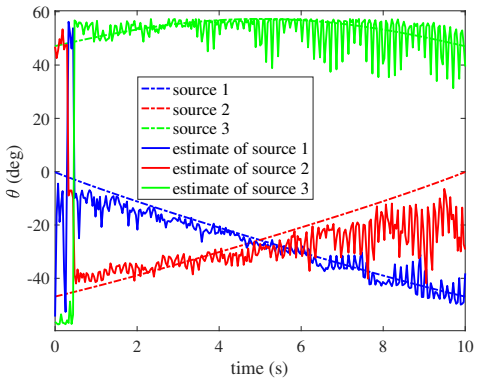
Acknowledgment

This work was partially supported by JSPS under Grants-in-Aid 22H01492.

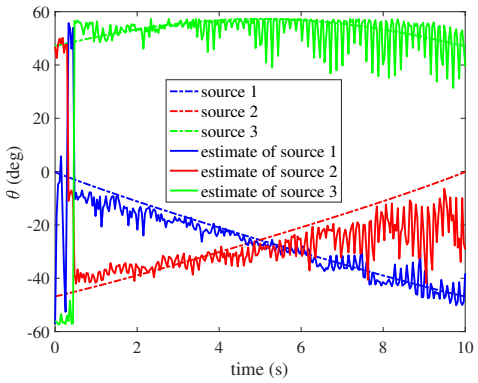
References

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.
[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.
[3] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.
[4] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 122, no. 2, pp. 392–402, 2014.
[5] Z. Zohny, S. Naqvi, and J. Chambers, "Variational EM for clustering interaural phase cues in MESSL for blind source separation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, p. 3966–3970.
[6] Y. Soussana and S. Gannot, "Variational inference for DOA estimation in reverberant conditions," in *Proc. European Signal Process. Conf.*, 2019, pp. 1–5.
[7] N. Ueno and H. Kameoka, "Multiple sound source localization based on stochastic modeling of spatial gradient spectra," in *Proc. European Signal Process. Conf.*, 2022, pp. 31–35.
[8] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

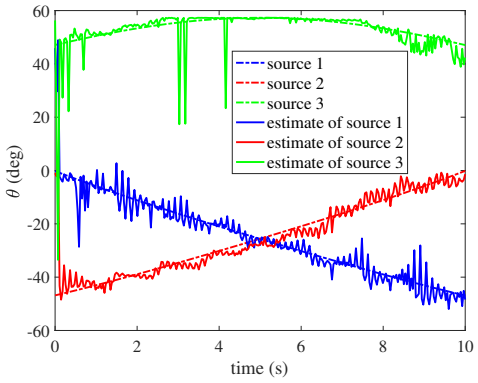
[9] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, 2019.
[10] S. Ando, N. Ono, and T. Nara, "Direct algebraic method for sound source localization with finest resolution both in time and frequency," in *Proc. Int. Congress Sound Vibration*, 2007, pp. 3273–3280.
[11] Z. Chen, "Bayesian filtering: From Kalman filters to particle filters, and beyond," *Statistics: A Journal of Theoretical and Applied Statistics*, vol. 182, no. 1, pp. 1–69, 2003.
[12] François Grondin and François Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robotics and Autonomous Systems*, vol. 113, pp. 63–80, 2017.
[13] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. T. Thompson Jr., "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.
[14] H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, "Bayesian nonparametric approach to blind separation of infinitely many sparse sources," *IEICE Trans. Fundamentals*, vol. E96-A, no. 10, pp. 1928–1937, 2013.
[15] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. Conf. Uncertainty Artif. Intell.*, 1999, pp. 21–30.



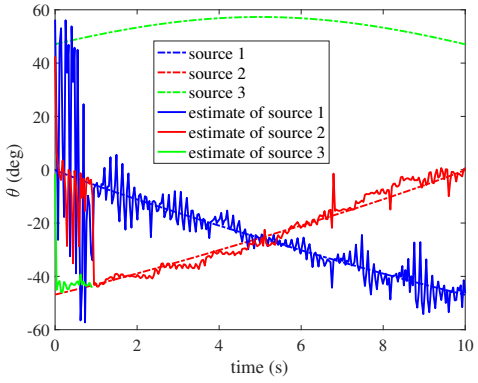
(a) $\delta = 0$



(b) $\delta = 1.0$



(c) $\delta = 6.4$



(d) $\delta = 7.0$

Fig. 3. Tracking behaviours.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Spatially-Regularized Switching Independent Vector Analysis

Tetsuya Ueda*, Tomohiro Nakatani†, Rintaro Ikeshita†, Shoko Araki†, and Shoji Makino*
* Waseda University, Japan and † NTT Coporation, Japan

***Abstract*—This paper proposes a novel algorithm that uses spatial regularization and a switching filter to enhance the separation accuracy of Independent Vector Analysis (IVA) with specified source permutation based on prior knowledge of sources' Directions-of-Arrival (DOAs). We call this algorithm Spatially-Regularized Switching Independent Vector Analysis (SR-SwIVA). Switching IVA (SwIVA) is a blind source separation (BSS) algorithm that builds on IVA and improves its separation accuracy with a switching filter. However, SwIVA requires relatively reliable estimates of acoustic transfer functions (ATFs) from sources to microphones for its initialization, which are difficult to obtain from DOAs in real reverberant environments. To overcome this limitation, we introduce spatial regularization to SwIVA, which has been shown to be effective for conventional BSS techniques to align permutation (or order) of separated sources with given DOAs. We conduct simulation experiments to show that our proposed SR-SwIVA can significantly improve the source separation accuracy using ATFs estimated from DOAs while completely eliminating the source permutation alignment errors. The achieved source separation accuracy is comparable to that of SwIVA with oracle ATFs.**

I. Introduction

Multichannel audio source separation has been actively studied as preprocessing of various speech applications to improve their performance. To develop effective source separation techniques, researchers often assume that speech signals' Directions of Arrival (DOAs) are available or can be estimated, e.g., from human look directions or using a camera. A promising source separation method for such a situation is Spatially-Regularized Source Separation (SRSS) that utilizes Blind Source Separation (BSS) techniques under regularization by DOA information [1], [2], [3], [4], [5], [6].

Independent Component Analysis (ICA) [7] is a BSS algorithm that estimates separation filters as ones that maximize the independence between the separated signals. To apply ICA to time-frequency domain audio signals, Independent Vector Analysis (IVA) [8], [9], [10], [11] has been proposed; it can group sources separated in different frequencies into individual sources using frequency-independent probabilistic source models. Several algorithms based on IVA have also been developed, including Independent Vector Extraction (IVE) [12], [13] and Convolutional beamformer extension of IVA (CIVA) [14], [15]. IVE can extract signals under the situations where microphones outnumber sources of interest in a computationally efficient way.

While BSS does not aim to determine the sources' order (or permutation) in the separated outputs, SRSS uses DOAs of sources as auxiliary information to align the separated source signals [1], [2], [3], [4], [5], [6]. We call this process source permutation alignment. In concrete, SRSS introduces to the BSS cost function a spatial regularization (also known as a geometric constraint [1]) using Acoustic Transfer Functions (ATFs) from sources to microphones estimated using DOAs (hereafter denoted as DOA-based ATFs). Note that DOA-based ATFs contain substantial estimation errors, e.g., due to reverberation in a real environment. Thus it is hard for classical beamformers like a Minimum Power Distortionless Response (MPDR) beamformer to extract speech signals accurately. In contrast, SRSS (e.g., Spatially-Regularized IVA (SR-IVA) [3], [4], [5], [6]) can estimate filters much more accurately by combining the BSS cost function with spatial regularization.

Although the above-mentioned multichannel audio source separation can separate signals effectively, the separation accuracy is limited, especially when the number of available microphones is small. A switching filter has recently been proposed to overcome this limitation [16], [17], [18], [19], [20]. With a switching filter, we cluster time frames of the microphone signals at each frequency so that each cluster contains relatively a small number of sources. Then, we estimate and apply different separation filters to respective clusters separately. It can achieve more accurate separation in noisy environments than simply applying a single separation filter to a whole captured signal because each cluster contains a smaller number of sources than a whole signal. The switching filter was initially introduced into a beamformer [16], [17] to enable it to handle underdetermined situations. It was then successfully introduced to IVA/CIVA [19] and IVE/CIVE [20] to improve their source separation accuracy. We denote IVA with switching filters as SwIVA, IVE with switching filters as SwIVE, and CIVA with switching filters as SwCIVA hereafter.

Despite the accurate source separation using a switching filter, SwIVA and algorithms based on SwIVA have a problem in that they are often trapped by a bad local optimum unless we carefully initialize their parameters at appropriate values. A reason for the problem to occur is that the permutation of separated sources can differ over different clusters. This problem is called inter-state permutation. To avoid the problem, previous research has proposed Blind Single-State initialization (*BSS init*) and SPatially-Guided initialization (*SPG init*) techniques [19]. Especially, *SPG init* avoids the inter-state permutation problem and aligns the source permutation when using relatively reliable ATFs estimated by a neural network. However, our preliminary experiments showed that using DOA-based ATFs was insufficient for *SPG init* in achieving

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

that.

This paper newly introduces spatial regularization to SwIVA to solve the above-mentioned inter-state permutation problem as well as to align the source permutation only using DOA-based ATFs. We call the algorithm Spatially-Regularized SwIVA (SR-SwIVA)[1]. In this algorithm, we use spatial regularization to align the source permutation in each cluster, thus the algorithm avoids the inter-state permutation problem during its filter optimization. Furthermore, we propose an initialization technique using SR-IVA with no switching mechanism. Because SR-IVA is free from the inter-state permutation problem, we can effectively avoid the problem at the filter initialization of SR-SwIVA. Our experiment shows that when using DOA-based ATFs, SR-SwIVA can perform as accurate source permutation alignment as SR-IVA and as much separation accuracy as SwIVA. In addition, the separation accuracy obtained by SR-SwIVA based on DOA-based ATFs is comparable to that obtained by SwIVA using accurately estimated ATFs.

In the remainder of this paper, we describe the problem formulation in Section II and the baseline algorithms in Section III. Our proposed algorithm is presented in Section IV. Experiment and conclusion are given in Sections V and VI.

II. Problem formulation

Suppose that N source signals are mixed and captured by M ($\geq N$) microphones. We represent observed signals $\boldsymbol{x}(f,t)$ and source signals $\boldsymbol{s}(f,t)$ at each time $t = 1,\ldots,T$ and frequency $f = 1,\ldots,F$ in the Short-Time Fourier Transform (STFT) domain as

$$\boldsymbol{x}(f,t) = [x_1(f,t),\ldots,x_M(f,t)]^{\mathsf{T}} \in \mathbb{C}^M, \tag{1}$$

$$\boldsymbol{s}(f,t) = [s_1(f,t),\ldots,s_N(f,t)]^{\mathsf{T}} \in \mathbb{C}^N, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes the transpose. We model the observed signals $\boldsymbol{x}(f,t)$ as a mixture of signals defined by

$$\boldsymbol{x}(f,t) = \sum_{n=1}^{N} \boldsymbol{d}_n(f,t) + \boldsymbol{r}(f,t), \tag{3}$$

$$\boldsymbol{d}_n(f,t) = \boldsymbol{h}_n(f)s_n(f,t), \tag{4}$$

where $\boldsymbol{d}_n(f,t) = [d_{n,1}(f,t),\ldots,d_{n,M}(f,t)] \in \mathbb{C}^M$ is the microphone image of the nth source and $\boldsymbol{h}_n(f)$ is the transfer function of the nth source. $\boldsymbol{r}(f,t)$ is the diffuse noise composed of the background noise.

We also suppose that we have an estimate of the transfer functions (hereafter called a steering vector). Assuming that the DOAs of the sources and the locations of microphones are given or estimated, we obtain a steering vector $\boldsymbol{a}_n(f)$ based on the plane wave assumption and the relative time-delay-of-arrival (TDOA) $\boldsymbol{\tau}_n \in \mathbb{R}^M$ from the nth source to M microphones:

$$\begin{aligned}\boldsymbol{a}_n(f) &= [a_{n,1}(f),\ldots,a_{n,M}(f)]^{\mathsf{T}} \in \mathbb{C}^M \\ &= \frac{1}{\sqrt{M}} \exp\left(-2\pi \frac{f}{N_{\mathrm{F}}} \boldsymbol{\tau}_n \sqrt{-1}\right) \text{ for } 1 \leq n \leq N. \end{aligned} \tag{5}$$

[1]Because it is straightforward to incorporate spatial regularization into SwIVE and SwCIVA, this paper omits this explanation for conciseness.

where each element $a_{n,m}(f)$ is an estimate of a transfer function from the nth source to the mth microphone and N_{F} is the length of Fourier transform. Note that $\boldsymbol{a}_n(f)$ contains substantial errors from the true transfer function $\boldsymbol{h}_n(f)$ in (4) because $\boldsymbol{a}_n(f)$ only contains a direct path component but does not include reflection paths.

The aim of this paper is to develop an effective algorithm to estimate a microphone image of each source at the first microphone as $[\hat{s}_1(f,t),\ldots,\hat{s}_N(f,t)]^{\mathsf{T}}$ given $\boldsymbol{x}(f,t)$ and $\{\boldsymbol{a}_n(f)\}_{n=1}^N$. It is important to note that we need to align the estimated sources with the same permutation as the sources in (2) during the estimation. We call this process source permutation alignment.

III. Baseline algorithms

This section briefly describes two of our baseline algorithms, SR-IVA and SwIVA.

A. Spatially-Regularized IVA (SR-IVA) [3], [4], [5], [6]

SR-IVA is an algorithm to estimate signals $[\hat{s}_1(f,t),\ldots,\hat{s}_N(f,t)]^{\mathsf{T}}$ with the same source permutation as in (2) by applying separation matrices $\{\boldsymbol{W}(f)\}_f$ to the observed signal:

$$\hat{\boldsymbol{s}}(f,t) = \boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,t), \tag{6}$$

where

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f),\ldots,\boldsymbol{w}_M(f)] \in \mathbb{C}^{M\times M}, \tag{7}$$

$$\hat{\boldsymbol{s}}(f,t) = [\hat{s}_1(f,t),\ldots,\hat{s}_M(f,t)]^{\mathsf{T}} \in \mathbb{C}^M. \tag{8}$$

Note that $\hat{s}_{N+1}(f,t),\ldots,\hat{s}_M(f,t)$ correspond to noise components. With BSS techniques such as IVA, we can obtain $\boldsymbol{W}(f)$ based on the maximum likelihood estimation assuming the statistical independence between each source $\hat{s}_n(f,t)$. However, we cannot determine the source permutation uniquely based only on the independence between sources. Thus, SR-IVA introduces a spatial regularization term to force each separation filter $\boldsymbol{w}_n(f)$ to extract a signal corresponding to $\boldsymbol{a}_n(f)$. The following is an example of a regularization term [5], which we use throughout this paper.

$$\mathcal{L}_{\text{regl}}(\{\boldsymbol{W}(f)\}_f) = \sum_{f=1}^{F}\sum_{n=1}^{N} \|\boldsymbol{w}_n(f) - \boldsymbol{a}_n(f)\|_2^2, \tag{9}$$

where $\|\cdot\|_2$ stands for l-2 norm. This regularization term enforces the nth separation filter $\boldsymbol{w}_n(f)$ to have a value close to the nth steering vector $\boldsymbol{a}_n(f)$. SR-IVA estimates signals $\hat{s}_n(f,t)$ by combining the likelihood function of BSS with the spatial regularization, thus simultaneously realizing source separation and source permutation alignment.

B. Switching IVA (SwIVA) [19]

Switching filter has recently been proposed to improve source separation accuracy. This section gives an overview of switching IVA (SwIVA). Fig. 1 shows the processing flow of SwIVA.

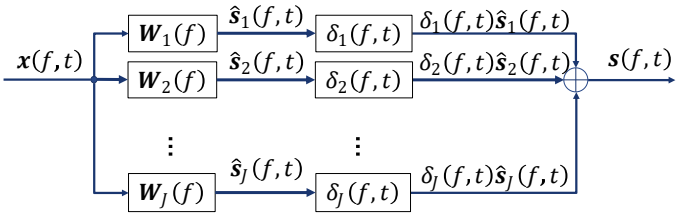


Fig. 1: Processing flow of switching IVA

SwIVA uses a multi-state separation model. It applies $J \geq 2$ separation matrices, $\boldsymbol{W}_j(f)$ for $1 \leq j \leq J$, to the observed signal to yield J different sets of separated sources $\hat{\boldsymbol{s}}_j(f,t) \in \mathbb{C}^M$:

$$\hat{\boldsymbol{s}}_j(f,t) = \boldsymbol{W}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,t) \text{ for } 1 \leq j \leq J, \tag{10}$$

$$\boldsymbol{W}_j(f) = [\boldsymbol{w}_{j,1}(f), \ldots, \boldsymbol{w}_{j,M}(f)] \in \mathbb{C}^{M \times M}. \tag{11}$$

Here j is an index of a switching state and J is the number of the switching states (or the number of separation matrices). Then, it uses a switch to select one of J separated sources at each time frequency and yields its final output:

$$\hat{\boldsymbol{s}}(f,t) = \sum_{j=1}^{J} \delta_j(f,t)\hat{\boldsymbol{s}}_j(f,t), \tag{12}$$

$$\sum_{j=1}^{J} \delta_j(f,t) = 1 \quad \text{and} \quad \delta_j(f,t) \in \{0,1\}. \tag{13}$$

The switch mechanism is implemented using a time-frequency dependent switching weight $\{\delta_j(f,t)\}_{j,f,t}$ in (12). In this paper, we only consider hard switches and allow $\delta_j(f,t)$ to take only binary values, 0 or 1. It has been reported that SwIVA achieved more accurate source separation than IVA that applies a single set of separation matrix in (6).

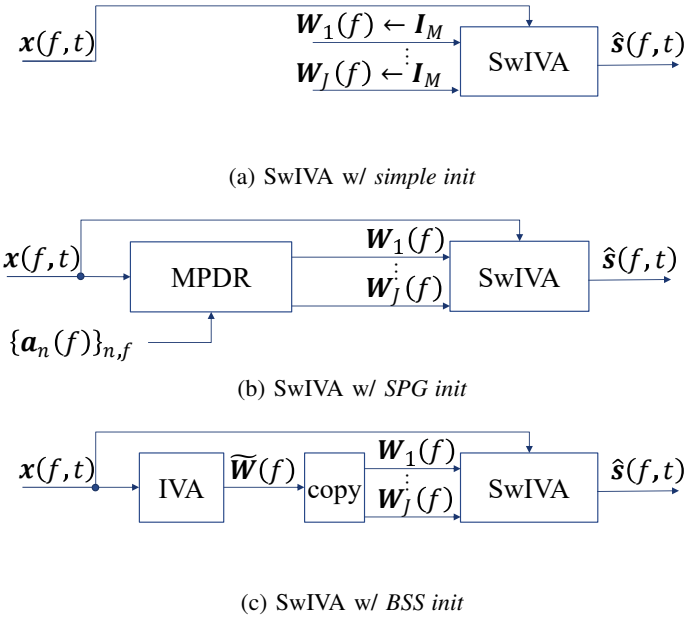
One important issue with SwIVA is that the separation matrices $\{\boldsymbol{W}_j(f)\}_j$ are often trapped by a bad local optimum that yields relatively less accurate source separation. This problem occurs because the permutations of separated sources can differ over different states. Thus, the order of sources changes when the switching state changes, resulting in large estimation errors. We call this problem inter-state permutation problem.

Previous research has proposed several initialization techniques for SwIVA to avoid the inter-state permutation problem [19]. We explain three of them in the following using Fig. 2.

1) Simple initialization (simple init): In this initialization, we set each separation matrix $\boldsymbol{W}_j(f)$ as an $M \times M$ identity matrix $\boldsymbol{I}_M$ as shown in Fig. 2a:

$$\boldsymbol{W}_j(f) \leftarrow \boldsymbol{I}_M \text{ for } 1 \leq j \leq J \text{ and } 1 \leq f \leq F. \tag{14}$$

It is reported that *simple init* does not avoid the inter-state permutation problem well for SwIVA.



(a) SwIVA w/ *simple init*

(b) SwIVA w/ *SPG init*

(c) SwIVA w/ *BSS init*

Fig. 2: Flow of SwIVA with three different initialization techniques

2) Spatially-guided initialization (SPG init): In this initialization, we employ the conventional MPDR beamformer to initialize the separation matrices $\boldsymbol{W}_j(f)$ as shown in Fig. 2b. In concrete, after randomly initializing $\delta_j(f,t)$, each filter $\boldsymbol{w}_{j,n}(f)$ is initialized:

$$\boldsymbol{w}_{j,n}(f) = \begin{cases} \frac{a^*_{n,1}(f)\boldsymbol{\Xi}_j^{-1}(f)\boldsymbol{a}_n(f)}{\boldsymbol{a}_n^{\mathsf{H}}(f)\boldsymbol{\Xi}_j^{-1}(f)\boldsymbol{a}_n(f)} & \text{for } 1 \leq n \leq N, \\ \boldsymbol{e}_n, & \text{for } N+1 \leq n \leq M, \end{cases} \tag{15}$$

where

$$\boldsymbol{\Xi}_j(f) = \frac{\sum_t \delta_j(f,t)\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)}{\sum_t \delta_j(f,t)}, \tag{16}$$

is a sample covariance matrix using time-frequency bins specified by the jth switch $\delta_j(f,t)$. $(\cdot)^*$ denotes complex conjugate, and $\boldsymbol{e}_n$ denotes the nth column of $\boldsymbol{I}_M$. The effectiveness of *SPG init* has been shown when we can obtain relatively accurate ATFs using a neural network.

3) Blind single-state initialization (BSS init): In this initialization, we use IVA to initialize the separation matrices $\boldsymbol{W}_j(f)$ as shown in Fig. 2c. Let $\tilde{\boldsymbol{W}}(f) \in \mathbb{C}^{M \times M}$ be a separation matrix estimated by IVA (or SwIVA with $J = 1$), then we initialize $\boldsymbol{W}_j(f)$ by $\tilde{\boldsymbol{W}}(f)$:

$$\boldsymbol{W}_j(f) \leftarrow \tilde{\boldsymbol{W}}(f) \text{ for } 1 \leq j \leq J \text{ and } 1 \leq f \leq F, \tag{17}$$

and randomly initialize $\delta_j(f,t)$. *BSS init* can avoid the inter-state permutation but is not intended to align the source permutation.

C. Drawbacks of SwIVA

Although using *SPG init* for SwIVA has been shown effective in avoiding the inter-state permutation problem when accurately estimated ATFs are available, it is not certain whether we can correctly align the source permutation based on only the initialization. In addition, our preliminary experiments showed that *SPG init* was insufficient when ATFs are estimated based on DOAs. This is probably because DOA-based ATFs contain substantial errors when estimated in real reverberant environments.

IV. Proposed algorithm: Spatially-Regularized SwIVA (SR-SwIVA)

This section proposes Spatially-Regularized SwIVA (SR-SwIVA) that controls SwIVA using DOA-based ATFs. To solve the problem discussed in Section III-C, we utilize DOA-based spatial regularization, which was shown effective for SR-IVA to align the source permutation. We first explain the estimation model and parameter optimization of SR-SwIVA in Section IV-A, and then present a new initialization technique for SR-SwIVA in Section IV-B. We expect that the algorithm effectively avoids the inter-state permutation problem and aligns the source permutation at the same time.

A. Estimation model and parameter optimization

First, we explain the estimation model and develop the parameter optimization of SR-SwIVA. SR-SwIVA uses the same separation model as SwIVA in (10). Then, similar to SR-IVA, we define the cost function as the sum of the negative log likelihood function $\mathcal{L}_{\text{NL}}(\mathcal{X};\Theta)$ used for SwIVA [19] and a spatial regularization term $\mathcal{L}_{\text{regl}}(\mathcal{W})$ that we design for SR-SwIVA:

$$\mathcal{L}(\Theta) = \mathcal{L}_{\text{NL}}(\mathcal{X};\Theta) + \lambda_{\text{regl}}\mathcal{L}_{\text{regl}}(\mathcal{W}). \tag{18}$$

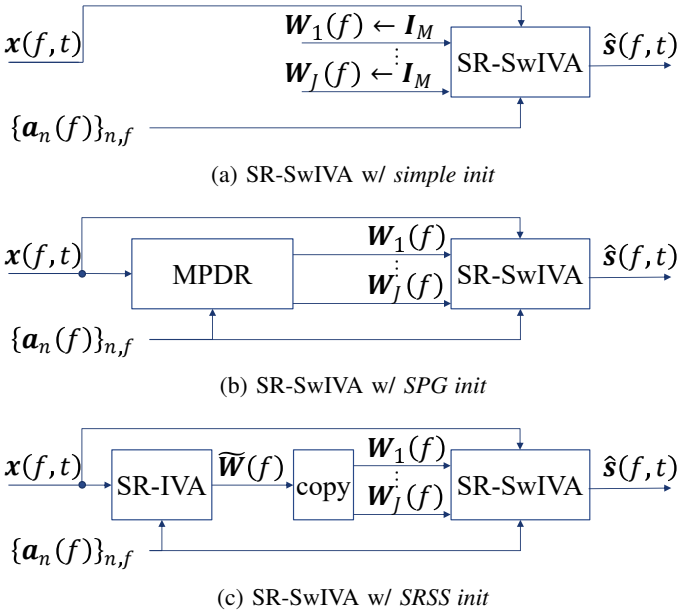
$$\mathcal{L}_{\text{NL}}(\mathcal{X};\Theta) = \sum_{j,f,t}\delta_j(f,t)\left\{\sum_{n=1}^{M}\left(\log v_n(f,t) + \frac{|\hat{s}_{j,n}(f,t)|^2}{v_n(f,t)}\right) - 2\log\left|\det \boldsymbol{W}_j(f)\right|\right\} \tag{19}$$

where $\hat{s}_{j,n}(f,t)$ is the nth element of $\hat{\boldsymbol{s}}_j(f,t)$ in (10), $\mathcal{X} = \{\boldsymbol{x}(f,t)\}_{f,t}$ is the set of observed signals, and $\Theta = \{\mathcal{D},\mathcal{W},\mathcal{V}\}$ is a set of model parameters composed of switching weights $\mathcal{D} = \{\delta_j(f,t)\}_{j,f,t}$, separation matrices $\mathcal{W} = \{\boldsymbol{W}_j(f)\}_{j,f}$, and the time-frequency dependent variance of the sources[2] $\mathcal{V} = \{v_n(f,t)\}_{n,f,t}$.

As for the spatial regularization term, $\mathcal{L}_{\text{regl}}(\mathcal{W})$, this paper adopts almost the same definition as that introduced for SR-IVA in (9). A major difference from SR-IVA is that we duplicate the same term to regularize the separation filters in all states. In concrete, the regularization term is formulated:

$$\mathcal{L}_{\text{regl}}(\mathcal{W}) = \sum_{f=1}^{F}\sum_{j=1}^{J}\sum_{n=1}^{N}\|\boldsymbol{w}_{j,n}(f) - \boldsymbol{a}_n(f)\|_2^2. \tag{20}$$

[2]Following the previous work for SwIVA [19], we use the frequency-independent source model only for updating the separation matrices as a practical technique to group the separated sources over different frequencies.



(a) SR-SwIVA w/ *simple init*

(b) SR-SwIVA w/ *SPG init*

(c) SR-SwIVA w/ *SRSS init*

Fig. 3: Flow of SR-SwIVA with three different initialization techniques

This term forces separation filters $\boldsymbol{w}_{j,n}(f)$ that share the same source index n at different states $1 \le j \le J$ to be close to the same steering vector of the nth source $\boldsymbol{a}_n(f)$. Therefore, as long as the spatial regularization successfully aligns the source permutation in the same way as SR-IVA, all the filters $\boldsymbol{w}_{j,n}(f)$ are expected to extract the same nth source for $1 \le j \le J$. As a consequence, we can also solve the inter-state permutation problem and align the source permutation.

The parameter optimization of SR-SwIVA can be straightforwardly derived by combining the existing algorithms (i.e., SR-IVA [5] and SwIVA [19]). We describe the derivation in Appendix A.

B. Spatially-Regularized single-state initialization (SRSS init)

Next, we introduce the DOA-based spatial regularization to the initialization for SR-SwIVA, referred to as Spatially-Regularized Single-State initialization (*SRSS init*). The aim of using *SRSS init* is to improve the estimation accuracy of SR-SwIVA better than *SPG init* that was used for SwIVA.

Fig. 3 shows flows of SR-SwIVA with three different initialization techniques that we compare in our experiments. In Figs. 3a and 3b, we use the same initialization techniques used in Figs. 2a and 2b. Fig. 3c shows our proposed *SRSS init*, which is a modification of *BSS init* in Fig. 2c for SwIVA. In *SRSS init*, we use SR-IVA (or SR-SwIVA with $J = 1$) instead of IVA to initialize all the parameters of SR-SwIVA. Because SR-IVA aligns the source permutation effectively, it can be a desirable initialization for SR-SwIVA using DOA-based ATFs.

V. Experiment

In this section, we compare the source separation and alignment accuracy of IVA, SwIVA, and SR-SwIVA combined with various initialization techniques, including *SPG init* and *SRSS init*. Our experiment shows the effectiveness of SR-SwIVA when using DOA-based ATFs.

A. Conditions of experiment

We conducted an experiment using TIMIT-ConvMix, which is composed of simulated noisy reverberant mixtures. To generate the mixtures, we first concatenated utterances extracted from the TIMIT corpus [21] to obtain a set of single-speaker clean utterance sequences, each of which is 30 s long. Then we mixed three utterance sequences and five different additive noise signals extracted from the CHiME-3 dataset [22] after individually reverberating them. We randomly selected 8 Room Impulse Responses (RIRs) out of 9 RIRs from the RWCP E2A dataset [23], corresponding to the azimuths of $10^\circ, 30^\circ, 50^\circ, \ldots, 170^\circ$. We then performed the convolution by assigning the RIRs to the utterances and noises. Here, we set the minimum angle difference between each utterance at 40°. RT60 of the RIRs was 0.3 s. We used four microphones labeled in the dataset: the 21st, 22nd, 23rd, and 24th. All four types of noise signals were used, BUS, STR, PED, and CAF, although each mixture contained only a single type. We set the power ratio of each reverberant utterance sequence to the sum of the additive noise signals to 10 dB.

We used a Hann window for a short-time analysis, where frame length N_{F} and shift size were set to 64 ms and 32 ms, respectively. The sampling frequency f_{s} was set to 16 kHz. We initialized switching weights $\delta_j(f,t)$ at a random value in a range of 1 ± 10^{-3}, and normalized it to satisfy $\sum_{j=1}^{J} \delta_j(f,t) = 1$. Note that after initialization, the weights $\delta_j(f,t)$ are updated to binary values during the parameter optimization. For all algorithms, including IVA, SR-IVA, SwIVA, and SR-SwIVA, we updated the separation filters 50 times and applied projection back [24] post-processing to solve the scale ambiguity.

Because we used linear arrays, we set relative TDOA $\boldsymbol{\tau}_n = [\tau_{n1}, \ldots, \tau_{nM}]$ in (5) for the DOA-based steering vectors:

$$\tau_{nm} = f_{\mathrm{s}} \frac{d(m-1)}{c} \cos\left(2\pi \frac{\theta_n}{360^\circ}\right), \tag{21}$$

where $c = 343$ m/s is the speed of sound and $d = 0.0281$ meter is the distance between adjacent microphones. We used the angle labels for RIRs in the RWCP dataset as the speakers' directions θ_n. Note that even using the angle labeled at the dataset, the steering vector in (5) should contain substantial errors because it does not include the effects of reflection paths. To confirm the performance using oracle information, we also evaluated the source separation accuracy of SRSS algorithms when replacing DOA-based ATFs $\boldsymbol{a}_n(f)$ with oracle steering vectors $\boldsymbol{h}_n(f)$ in (4). We created $\boldsymbol{h}_n(f)$ as the primary eigenvector of the spatial covariance matrix of the noiseless reverberant source image of $s_n(f,t)$.

In this evaluation, we adopted signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) as the metrics of source separation accuracy [25], which is widely used in source separation research. We used the MUSEVAL V4 toolkit [26] with its bss_eval_images configuration. As reference signals, we used clean utterance sequences that were convolved with the RIRs for generating the corresponding mixtures. We evaluated SDRs and SIRs using a correct permutation of estimated sources, not depending on the actual permutation estimated by each separation algorithm. We determined the correct permutation as the order of estimated sources that gave the best SIR score with the reference signals in their original order. In addition to SDR and SIR, we evaluated the correctness of the estimated source permutation alignment using the permutation error (permE) defined.

$$\text{permE} = \frac{\text{\# of mixtures separated with incorrect permutation}}{\text{Total \# of mixtures (=100)}}. \tag{22}$$

The estimated source permutation alignment was determined to be incorrect when it was not identical to the correct source permutation.

B. Result

We compared the SDR improvement, SIR improvement, and permE for each source separation algorithm in Table I. We conducted two experiments, one with three microphones and the other with four microphones, and show the results in Tables Ia and Ib. For each experiment, we set the regularization weight λ_{regl} as a value, with which SR-SwIVA achieved the highest SDRs while reducing permE to zero.

We first show the result of the baseline algorithms, SR-IVA and SwIVA. In both Tables Ia and Ib, although SR-IVA ($J = 1$) reduced permE to zero and achieved comparable SDRs and SIRs as IVA, SwIVA with *SPG init* showed 5 % permE when $M = 3$. Moreover, SwIVA with *SPG init* showed 100 % permE when $M = 4$. This is because DOA-based ATFs contained a lot of errors, as shown by the low source separation accuracy (3.83 dB SIR improvement) by MPDR beamformer. The above results show that when using DOA-based ATFs, SwIVA with *SPG init* cannot correctly align separated sources.

Next, we show the result of SR-SwIVA with several initialization techniques presented in Fig. 3 using the same tables. From the tables, SR-SwIVA could not achieve 0 % permE when using *simple init* or *SPG init* except for using *simple init* with $M = 4$. Also, their SDRs are comparable to or even worse than the baseline algorithms. On the other hand, SR-SwIVA with *SRSS init* achieved the highest SDR and 0 % permE when using DOA-based ATFs. Thus, we conclude that SR-SwIVA with *SRSS init* can separate signals and align the source permutation more effectively than the baseline algorithms.

Finally, we show the results using oracle ATFs. In both tables, the MPDR beamformer worked much better using oracle ATFs than using DOA-based ATFs, and SwIVA with *SPG init* attained high SDR and SIR and reduced permE to 0 %. It means that *SPG init* is effective for SwIVA only when

TABLE I: The improvements of SDR (SDRi), SIR (SIRi) [dB], and permutation error (permE) [%]. SDRs and SIRs were calculated using a correct source permutation that achieved the highest SIRs among all possible source permutations.

(a) $N = 3, M = 3, \lambda_{\text{regl}} = 0.4$

	initialization	J	SDRi	SIRi	permE
Blind Source Separation					
IVA [10]	-	1	7.73	19.97	87.5
SwIVA [19]	*BSS init*	2	8.55	22.43	87.5
DOA-based ATFs informed Source Separation					
MPDR	-	-	3.21	12.26	0.0
SR-IVA [3, .etc]	-	1	7.68	19.77	0.0
SwIVA [19]	*SPG init*	2	8.28	21.78	5.0
SR-SwIVA	*simple init*	2	7.99	20.64	7.5
	SPG init	2	7.34	19.44	6.0
	SRSS init	2	8.52	22.38	0.0
Oracle ATFs informed Source Separation					
MPDR	-	-	6.55	15.01	0.0
SwIVA [19]	*SPG init*	2	8.67	22.73	0.0
SR-SwIVA	*SRSS init*	2	8.50	22.32	5.0

(b) $N = 3, M = 4, \lambda_{\text{regl}} = 1.0$

	initialization	J	SDRi	SIRi	permE
Blind Source Separation					
IVA [10]	-	1	8.67	23.65	87.5
SwIVA [19]	*BSS init*	2	9.27	25.33	100.0
DOA-based ATFs informed Source Separation					
MPDR	-	-	2.31	3.83	80.0
SR-IVA [3, .etc]	-	1	8.70	23.63	0.0
SwIVA [19]	*SPG init*	2	8.75	23.52	100.0
SR-SwIVA	*simple init*	2	8.81	23.47	0.0
	SPG init	2	8.93	24.28	37.5
	SRSS init	2	9.30	25.31	0.0
Oracle ATFs informed Source Separation					
MPDR	-	-	7.63	18.22	0.0
SwIVA [19]	*SPG init*	2	9.05	24.47	0.0
SR-SwIVA	*SRSS init*	2	9.27	25.22	0.0

we have relatively accurate ATFs. In contrast, SR-SwIVA with *SRSS init* can achieve the same high scores regardless of using oracle ATFs or DOA-based ATFs. Thus we showed that SR-SwIVA could achieve as much accurate source separation and source permutation alignment as SwIVA based on relatively accurate ATFs.

VI. Conclusions

In this paper, we have proposed a novel algorithm called SR-SwIVA that uses spatial regularization to enhance SwIVA for accurate source separation and permutation alignment based on DOA-based ATFs. We have shown that DOA-based ATFs have significant estimation errors in real reverberant environments, which prevent conventional SwIVA from achieving high separation accuracy and correct permutation alignment. To address this challenge, we have developed SR-SwIVA by incorporating spatial regularization into SwIVA and presented a new effective initialization technique called *SRSS init* that uses DOA-based ATFs. We have conducted an experiment using noisy reverberant sound mixtures and demonstrated that our proposed SR-SwIVA can significantly improve the SDRs and SIRs compared to our baseline algorithms, SR-IVA and SwIVA, with DOA-based ATFs, and eliminate the source permutation alignment errors completely. Moreover, we have shown that the source separation accuracy achieved by SR-SwIVA is similar to that of SwIVA with oracle ATFs.

Acknowledgment

This work was supported by JSPS KAKENHI 23H03423 and JST SPRING JPMJSP2128.

Appendix A. parameter optimization of SR-SwIVA

We optimize the parameters Θ for SR-swIVA by minimizing the cost function in (18). Because no closed form solution has been obtained, we use iterative estimation based on a coordinate descent method. It alternately updates one of $\mathcal{W}$, $\mathcal{D}$, and $\mathcal{V}$ by fixing the other parameters, and iterates the update until convergence is obtained.

Because the update of $\mathcal{D}$ and $\mathcal{V}$ is not related with the regularization term, we can update $\mathcal{D}$ and $\mathcal{V}$ using the same update rules written in [19]. Hereafter we only explain the update of $\mathcal{W}$. For the update of the separation matrices $\mathcal{W}$, we adopt a frequency-independent source model following a practical technique proposed in [19] to group separated sources over different frequencies. In concrete, instead of using the frequency-dependent variance $v_n(f,t)$, we use a frequency-independent variance $v_n(t)$ for the update, which is obtained by $v_n(t) \leftarrow \frac{1}{F}\sum_{f=1}^{F} v_n(f,t)$. Then, fixing $\delta_j(f,t)$ and $v_n(t)$, we can rewrite the function $\mathcal{L}(\Theta)$ in (18) as

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_j(f)) \overset{c}{=} &-2\log\left|\det \boldsymbol{W}_j(f)\right| \\ &+ \sum_{n=1}^{M}\big(\boldsymbol{w}_{j,n}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f)\boldsymbol{w}_{j,n}(f) \\ &- \boldsymbol{w}_{j,n}^{\mathsf{H}}(f)\boldsymbol{p}_n(f) - \boldsymbol{p}_n^{\mathsf{H}}(f)\boldsymbol{w}_{j,n}(f)\big), \end{aligned} \tag{23}$$

where

$$\boldsymbol{\Pi}_{j,n}(f) = \begin{cases} \boldsymbol{\Sigma}_{j,n}(f) + \lambda_{\text{regl}}\boldsymbol{I}_M & \text{for } 1 \le n \le N, \\ \boldsymbol{\Sigma}_{j,n}(f) & \text{for } N+1 \le n \le M, \end{cases} \tag{24}$$

$$\boldsymbol{p}_n(f) = \begin{cases} \lambda_{\text{regl}}\boldsymbol{a}_n(f) & \text{for } 1 \le n \le N, \\ \boldsymbol{0}_M & \text{for } N+1 \le n \le M, \end{cases} \tag{25}$$

$$\boldsymbol{\Sigma}_{j,n}(f) = \frac{1}{T_j(f)}\sum_{t=1}^{T}\frac{\delta_j(f,t)}{v_n(t)}\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t), \tag{26}$$

$$T_j(f) = \sum_{t=1}^{T}\delta_j(f,t), \tag{27}$$

and $\boldsymbol{0}_M \in \mathbb{C}^M$ is a zero vector. Because the above objective has the same form as SR-IVA [4], [5], we can apply iterative optimization techniques proposed for it, such as Vectorwise Coordinate Decent (VCD) [2], and Iterative Source Steering [11], [6]. This paper employs VCD that updates $\boldsymbol{W}_j(f)$ with a sequence of $\boldsymbol{w}_{j,1}(f) \rightarrow \boldsymbol{w}_{j,2}(f), \ldots, \rightarrow \boldsymbol{w}_{j,M}(f)$ for $j =$

$1, \ldots, J$:

$$\boldsymbol{u}_{j,n}(f) = (\boldsymbol{W}_j^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f))^{-1}\boldsymbol{e}_n, \tag{28}$$

$$\tilde{\boldsymbol{u}}_{j,n}(f) = \boldsymbol{\Pi}_{j,n}^{-1}(f)\boldsymbol{p}_n(f), \tag{29}$$

$$h_{j,n}(f) = \boldsymbol{u}_{j,n}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f)\boldsymbol{u}_{j,n}(f), \tag{30}$$

$$\hat{h}_{j,n}(f) = \boldsymbol{u}_{j,n}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f)\tilde{\boldsymbol{u}}_{j,n}(f), \tag{31}$$

$$\tilde{h}_{j,n}(f) = \frac{\hat{h}_{j,n}(f)}{2h_{j,n}(f)}\left[-1 + \sqrt{1 + \frac{4h_{j,n}(f)}{|\hat{h}_{j,n}(f)|^2}}\right], \tag{32}$$

$$\boldsymbol{w}_{j,n}(f) = \begin{cases} \frac{1}{\sqrt{h_{j,n}(f)}}\boldsymbol{u}_{j,n}(f) + \tilde{\boldsymbol{u}}_{j,n}(f) & \text{if } \hat{h}_{j,n}(f) = 0, \\ h_{j,n}(f)\boldsymbol{u}_{j,n}(f) + \tilde{\boldsymbol{u}}_{j,n}(f) & \text{otherwise.} \end{cases} \tag{33}$$

REFERENCES

[1] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[2] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. ICASSP*, 2018, pp. 746–750.

[3] A. H. Khan, M. Taseska, and E. A. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.

[4] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. ICASSP*, 2020, pp. 846–850.

[5] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Processing*, vol. 68, pp. 3545–3558, 2020.

[6] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering," in *Proc. EUSIPCO*, 2022, pp. 757–761.

[7] A.Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2006.

[9] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[10] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *Proc. LVA/ICA*, 2010, pp. 165–172.

[11] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. ICASSP*, 2020, pp. 236–240.

[12] Z. Koldovsky and P. Tichavsky, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Processing*, vol. 67, no. 4, pp. 1050–1064, 2018.

[13] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. WASPAA*, 2019, pp. 185–189.

[14] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. ICASSP*, 2021, pp. 6129–6133.

[15] R. Ikeshita and T. Nakatani, "Independent vector extraction for fast joint blind source separation and dereverberation," *IEEE Signal Processing Letters*, vol. 28, pp. 972–976, 2021.

[16] K. Yamaoka, A. Brendel, N. Ono, S. Makino, M. Buerger, T. Yamada, and W. Kellermann, "Time-frequency-bin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios," in *Proc. EUSIPCO*, 2018, pp. 1582–1586.

[17] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 3461–3475, 2021.

[18] R. Ikeshita, N. Kamo, and T. Nakatani, "Blind signal dereverberation based on mixture of weighted prediction error models," *IEEE Signal Processing Letters*, vol. 28, pp. 399–403, 2021.

[19] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, "Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms," *IEEE/ACM Trans. ASLP*, vol. 30, pp. 1032–1047, 2022.

[20] ——, "Switching independent vector extraction and its joint optimization with weighted prediction error dereverberation," in *Proc. ICA*, 2022.

[21] L. Consortium *et al.*, "Timit acoustic-phonetic continuous speech corpus," 1993.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[23] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," 2000.

[24] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, 2001.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] "Museval," *https://github.com/sigsep/sigsep-mus-eval*.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

ASF-LLRDA: Locality-regularized Linear Regression Discriminant Analysis with Approximately Symmetrical Face Preprocessing for Face Recognition

Arya Widyadhana*, Shintami Chusnul Hidayati*, Dini Adni Navastara*, and Yeni Anistyasari†
* Institut Teknologi Sepuluh Nopember, Indonesia
E-mail: arya.206025@mhs.its.ac.id, shintami@its.ac.id, dini.navastara@its.ac.id
† Universitas Negeri Surabaya, Indonesia
E-mail: yenian@unesa.ac.id

***Abstract*—Face recognition is a crucial task in numerous applications, but it has difficulties because of the high-dimensional nature of facial photos, limited sample sizes, variations in illumination, and facial expressions. This paper presents a novel face recognition approach to overcome these challenges by combining the advantages of the Approximately Symmetrical Face (ASF) preprocessing strategy and the Locality-regulated Linear Regression Discriminant Analysis (LLRDA) method. The proposed method, called ASF-LLRDA, makes use of ASF to generate axis-symmetric face images for reducing the impact of illumination and variations in facial expressions, followed by LLRDA to extract discriminative features and project the data into a lower dimensional space. Locality-regulated Linear Regression (LLRC) is further utilized as the classifier. Experimental results on the Yale-B face dataset demonstrated the superiority of the proposed method compared to the baselines.**

I. Introduction

Face recognition has many applications in a variety of fields, such as biometric identification [1], action detection [2], marketing [3], and personalization [4], [5]. In real-world scenarios, face recognition systems often encounter challenges related to limited storage space and time constraints for capturing training samples [6], [7]. This is primarily due to the high-dimensional nature of facial images, which necessitates a larger storage capacity [8]. Additionally, when the available sample size is restricted, variations in illumination and facial expressions can lead to greater intra-subject variability compared to inter-subject variability [9]. Consequently, different faces may exhibit similar patterns, resulting in the incorrect recognition of individuals as the same person.

In order to deal with high-dimensional data effectively, dimension reduction has become a vital preprocessing step. The primary objective of dimensionality reduction is to concurrently solve two major problems: reducing intra-subject variability and enhancing inter-subject variability. This is achieved by eliminating irrelevant, redundant, and noisy data samples in extracting the underlying features inherent in high-dimensional data. As a result, dimensionality reduction has also been extensively explored to overcome the limitations encountered in real-world facial recognition applications, as previously discussed [10].

Among the various dimensionality reduction techniques utilized in face recognition algorithms, Principal Component Analysis (PCA) [11] and Latent Dirichlet Allocation (LDA) [12] are widely employed due to their effectiveness in extracting low-dimensional and discriminative features [13]. PCA is an unsupervised approach that aims to maximize the variance of the data by projecting it onto a new subspace. However, it does not incorporate any prior class information, thus ignoring potentially useful discriminative information [14]. Meanwhile, LDA is a supervised method that takes into account the prior class information. By considering class information, LDA can potentially enhance the separability of the data [14]. Nevertheless, both PCA and LDA primarily capture the global Euclidean structure of the data, disregarding the underlying manifold structure that may exist in the data [15]. Moreover, they do not explicitly consider the decision rule of the classifier, which can lead to subspaces with distinct characteristics that are not fully exploited by the classifiers, resulting in suboptimal classification performance.

To retain the local manifold structures of the data, several nonlinear-based techniques have been developed, including isometric mapping [16], [17], locally-linear embedding [18], [19], and multi-dimensional scaling [20]. However, these methods often exhibit poor performance when applied to new and unseen data [21]. Some existing approaches leverage representation-based classifiers to improve pattern recognition within dimensionality reduction. Examples include the sparse representation-based classifier (SRC) [22], collaborative representation-based classification (CRC) [23], and linear regression classification (LRC) [24]. Among these methods, LRC offers reasonable accuracy while maintaining lower complexity.

This work was supported in part by Institut Teknologi Sepuluh Nopember via 1148/PKS/ITS/2021. *(Corresponding Author: Shintami Chusnul Hidayati)*

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

In [14], Huang *et al.* have proposed the locality-regularized linear regression discriminant analysis (LLRDA) designed according to the decision rule of the expanded LRC using a locality-regularized procedure. LLRDA exhibits promising potential in capturing the most discriminant information from the data by leveraging the characteristics of the learned subspace. However, it should be noted that the original feature space may not guarantee high classification accuracy, particularly when faced with variations in illumination and limited training samples, which are common challenges in face recognition tasks.

Various preprocessing methods have been proposed to mitigate the effects of illumination on face images. Lambertian reflection-based methods have gained significant attention lately due to their effectiveness in handling unbalanced illumination conditions. Developing a robust model that can handle general real-world scenes, however, remains a challenge. One of the main difficulties lies in reliably estimating the surface orientations and light directions, which are crucial for accurate illumination compensation. The complexity of real scenes, with diverse lighting conditions and surface properties, adds to the complexity of this task.

A study proposed by Xu *et al.* [25] has introduced a non-Lambertian reflectance, called Axis-Symmetrical Face (ASF), to address the challenges of varying face appearances and their negative effects on face recognition. ASF leverages the symmetrical geometric structure of the face to generate an axis-symmetrical virtual face image by equalizing the illumination between the right and left sides of the face. In addition to its impact on face recognition accuracy, ASF also has the potential to produce more visually appealing face images. It is worth noting that ASF is a face image preprocessing method. Further research might open the possibility of using ASF to optimize the performance and computational efficiency of face recognition tasks.

Motivated by the promising properties of ASF and LLRDA, we propose a novel approach for dimensionality reduction designed to improve face recognition performance. The proposed method, called ASF-LLRDA, combines the advantages of the ASF method in handling varying illuminations and facial expressions of face images with the abilities of LLRDA to capture dominant trends in the data and project them to a lower-dimensional space. We further employ the locality-regularized linear regression classification (LLRC) classifier, known for its promising performance in face recognition, to enhance classification accuracy. As we demonstrate in the following discussions, the proposed method significantly enhances face recognition performance in small sample-size data with face images of varying illumination. We also conduct comprehensive experiments to show the outstanding performance of the method in recognizing face images.

The rest of this paper is organized as follows. Section II describes the details of our face recognition framework. The performance of the algorithm is thoroughly investigated in Section III. Finally, Section IV concludes our work with potential directions for future work.

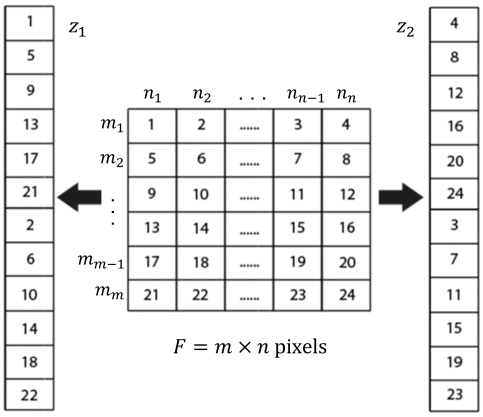


Fig. 1: An illustration of how to obtain the column vectors of z_1 and z_2.

II. Methodology

Our approach comprises three primary steps, namely: (1) facial image synthesis, (2) subspace construction, and (3) face recognition, as described below.

A. Facial Image Synthesis

To generate an approximate axis-symmetric model of a grayscale face image F with the size of $m \times n$ pixels, we first create the left column vector (z_1) and right column vector (z_2) of F. As illustrated in Figure 1, z_1 is formed by concatenating the first column of n, denoted by n_1, to the $n_{\frac{n}{2}}$-th column of F one by one. Similarly, z_2 is obtained by concatenating the last column of n, denoted by n_n, to the $n_{\frac{n}{2}+1}$-th column of F one by one.

In the following step, an axis-symmetric face image is generated to reduce the effect of varying illumination and facial expression. This involves iteratively updating each pixel value in the z_1 and z_2 using gradient descent:

$$x^{t+1} = x^t - \eta \triangle f(x^t), \tag{1}$$

where η represents the learning rate and x^t represents the pixel value at iteration t. Two types of f functions are used to minimize the difference in value between z_1 and z_2, which are defined as:

$$f_{z_1} = \|z_1 - z_2\|^2, \tag{2}$$

$$f_{z_2} = \|z_2 - z_1\|^2. \tag{3}$$

Then, the iterative update for the values of both z_1 and z_2 at each iteration t are as follows:

$$z_1^{t+1} = z_1^t - \eta(z_1^t - z_2^t), \text{and} \tag{4}$$

$$z_2^{t+1} = z_2^t - \eta(z_2^t - z_1^t). \tag{5}$$

As t increases, the learning rate η decreases. Thus, to facilitate easier convergence of the solution, the learning rate is adjusted using the following formula:

$$\eta = \alpha / t, \tag{6}$$

where α is a positive constant. The iteration continues until the following conditions are satisfied:

$$\|z_1^{t+1} - z_1^t\| < \epsilon(z_1^t), \text{and} \tag{7}$$

$$\|z_2^{t+1} - z_2^t\| < \epsilon(z_2^t). \tag{8}$$

Here, ϵ is a small positive constant.

B. Subspace Construction

This step is designed to highlight the distinctiveness of samples from different classes. To achieve this, the data is projected onto a subspace that minimizes the intra-class local reconstruction scatter, indicating the proximity of samples within the same class, while maximizing the inter-class local reconstruction scatter. The representation of the data in the new subspace is defined by:

$$y = \mathbf{A}^T F', \tag{9}$$

where F' represents the preprocessed F using the strategy described in Section II-A, and $\mathbf{A}$ denotes a matrix projection.

Specifically, the matrix projection $\mathbf{A}$ that describes how to perform the mapping from original space to lower dimensional space is formed by:

$$\mathbf{A} = \mathbf{A}_P \mathbf{A}_L. \tag{10}$$

Here, $\mathbf{A}_P$ is the result of PCA projection, and $\mathbf{A}_L$ is constructed by utilizing the intra-class local reconstruction scatter matrix S_w and the inter-class local reconstruction scatter matrix S_b. It is important to note that the dimension of the subspace for $\mathbf{A}_P$ does not exceed the number of samples N.

To obtain S_w and S_b, we first determine the nearest neighbors of F' within each class. The k-nearest neighbors of F belonging to class i can be expressed as

$$\tilde{F}_i = [\ F_i^1, F_i^2, ..., F_i^k]\ \in \mathbb{R}^{N \times k}, \tag{11}$$

where F_i^j represent the j-th sample of class i and F_l^j represent the j-th sample of class l, with l being a class different from i. Subsequently, S_w and S_b are calculated using

$$S_w = \sum_{i,j} (\ F_i^j - \tilde{F}_i^j \beta_i^j)\ (\ F_i^j - \tilde{F}_i^j \beta_i^j)\ ^T, \tag{12}$$

$$S_b = \sum_{i,j,l} (\ F_i^j - \tilde{F}_l^j \beta_l^j)\ (\ F_i^j - \tilde{F}_l^j \beta_l^j)\ ^T. \tag{13}$$

In these equations, β_i^j and β_l^j represent the reconstruction coefficient vectors for class i and class l, respectively. The linear reconstruction of y for class i is defined as

$$\tilde{\beta}_i = (\ \tilde{F}_i^T \tilde{F})\ ^{-1} \tilde{F}_i^T y. \tag{14}$$

By solving the following eigenvalue problem,

$$S_w^{-1} S_b^a = \lambda a, \tag{15}$$

we obtain $\mathbf{A}_L \in \mathbb{R}^{N \times d} (d < N)$, which is constructed from the eigenvectors of $S_w^{-1} S_b^a$ associated with the largest eigenvalues of $S_w^{-1} S_b^a$.



Fig. 2: Examples of images from a single subject within the Yale B dataset.

C. Face Recognition

In face recognition, it is commonly observed that samples within a class demonstrate local linearity rather than global linearity. This implies that neighboring samples within a class, rather than all samples, tend to lie in a linear subspace [14]. Consequently, in LLRC, only the k-nearest neighbors of the testing sample from each class are considered when performing the linear reconstruction. To generate a discriminant subspace for LLRC, the decision rule is formulated as:

$$class(y) = \min_i \|y - \tilde{X}_i (\ \tilde{X}_i^T \tilde{X}_i)\ ^{-1} \tilde{X}_i^T y\|^2. \tag{16}$$

III. Experiments

In this section, we provide a detailed description of the dataset utilized in the experiments, outline the parameters employed, baseline method for comparison, and subsequently present a comprehensive analysis of the experimental results.

A. Dataset

The primary objective of this study is to evaluate the performance of the proposed method in handling small sample-size datasets with variations in illumination and facial expressions. To achieve this, we utilize the Yale B dataset [26] as the dataset for our experiments. The Yale B dataset comprises a total of 16,128 grayscale face images captured from 28 distinct subjects. Each subject exhibits 9 different facial expressions, while the images are captured under 64 distinct illumination conditions. Figure 2 presents a visual representation of sample images from a single subject, illustrating the variations in facial expression and illumination within the dataset.

B. Experimental Setup

Three parameters were employed in the experiment. The first parameter, denoted by n_{tr}, represents the number of training samples utilized for each class or subject. Because our focus is on evaluating the performance of methods with small-sample sizes, a limited number of samples ranging from 3 to 6 were employed as training data. The second parameter relates to the number of nearest neighbors for each sample, denoted by *k*. To ensure a balanced consideration of neighboring samples, *k* was set to $\frac{n_{tr}}{2}$ in the case of an even n_{tr} and *k* was set to $\lfloor \frac{n_{tr}}{2} \rfloor + 1$ for an odd n_{tr}. Lastly, the third parameter corresponds to the number of features extracted from the data, denoted by d. For this study, d was consistently set to 20, allowing for a concise representation of the data while retaining relevant discriminatory information.

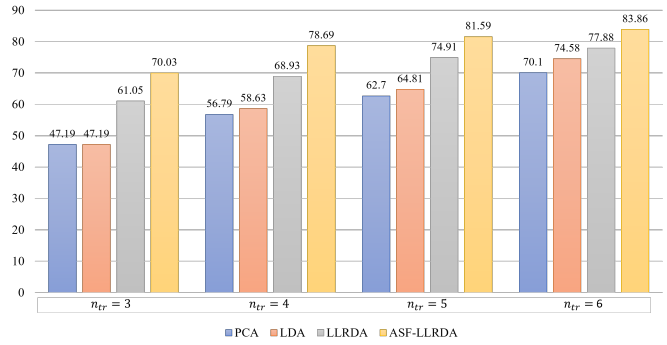


Fig. 3: The average accuracy of each method with different number of training samples n_{tr}.

C. Baselines

To assess the classification performance of the proposed method, we conducted a comparative analysis by evaluating its average classification accuracy against dimensionality reduction-based methods, namely PCA [11], LDA [12], and LLRDA [14]. The LLRC classifier was employed for each method to classify the testing samples. To ensure reliability, we performed five experiments for each n_{tr} and calculated the average classification accuracy. In each experiment, the training samples were randomly selected to assess the robustness of each method against variations in illumination and facial expressions depicted in the face images.

D. Experimental Results and Discussion

Table I summarizes the classification performance of baseline methods and the proposed ASF-LLRDA method. The average computational time for PCA, LDA, LLRDA, and ASF-LLRDA is 4.59, 4.42, 4.69, and 4.78 seconds, respectively. Note that, for a fair comparison across methods, we used the same data partition.

As shown in Table I and the bar chart in Figure 3, for $n_{tr} = 3$, the average classification accuracy of PCA and LDA was very low (less than 50%). However, as the number of training samples increased, the classification accuracy of both PCA and LDA markedly improved. For example, for $n_{tr} = 6$, the average classification accuracy exceeded 70%. We suggest this lack of accuracy is related to the extreme variations in illumination and facial expressions present in the images. Particularly, in PCA, this limitation arises because the focus is only on maximizing the variance of the data, without considering the previous class information and the geometric structure of the data. Therefore, in the generated new subspace, the samples from different classes tend to overlap, posing challenges in distinguishing between samples of different classes. On the other hand, LDA takes into account the prior class information, enabling it to capture the discriminant structure of the data and achieve a slightly improved average classification accuracy. However, even with this consideration, the variations in illumination and facial expression prove substantial, resulting in insufficient separability of samples from different classes within the new subspace generated by LDA. Similar to PCA, LDA disregards the geometric structure of the data, further contributing to lower classification accuracy compared to other methods.

In contrast to PCA and LDA, LLRDA demonstrated the ability to reveal the discriminant and local geometric structure of the data. Consequently, the generated new subspaces possess enhanced separability among samples from different classes, surpassing the subspaces produced by PCA and LDA. For $n_{tr} = 3, 4$, and 5, the average classification accuracy achieved by LLRDA was 61.05%, 68.93%, and 74.91% respectively, surpassing the average classification accuracy of PCA and LDA by a notable margin of 10%. For $n_{tr} = 6$, the difference in average classification accuracy between LLRDA and LDA decreases significantly to slightly over 3%. These results indicate that LLRDA performance is still affected by extreme variations in illumination and facial expressions present in images, although to a lesser degree compared to PCA and LDA.

Evidently, the experimental results demonstrate the superiority of ASF-LLRDA, which consistently outperformed PCA, LDA, and LLRDA by substantial margins with an average computation time that does not exceed 9%. Remarkably, even with a limited number of training samples, the average classification accuracy of ASF-LLRDA remained consistently above 70%. These results validate ASF-LLRDA as an effective face recognition technique when combined with the LLRC classifier, especially in scenarios with small sample sizes and face images characterized by extreme variations in illumination and facial expression. Nevertheless, it is worth noting that ASF-LLRDA still exhibits limitations in extracting meaningful information to comprehensively elucidate data characteristics. Consequently, promising directions for future research may entail exploring the application of Autoencoders to capture intricate data relationships, as demonstrated in [27]

IV. Conclusion

This paper has presented a novel method called ASF-LLRDA to address the challenges encountered in face recognition applications. By integrating the Approximately Symmetrical Faces (ASF) preprocessing technique with the Locality-regularized Linear Regression Discriminant Analysis (LLRDA) approach, ASF-LLRDA offers a robust solution for handling high-dimensional face images, limited sample sizes, and variations in illumination and facial expressions. The generation of axis-symmetric face images through ASF alleviated the impact of illumination and facial expression variations, while LLRDA extracted discriminative features and projected the data into a lower-dimensional space. Leveraging the locality-regularized linear regression classification (LLRC) classifier further enhanced the classification accuracy. Experimental evaluations conducted on the Yale-B face dataset confirmed the superiority of ASF-LLRDA over baselines, as it achieved significantly higher average classification accuracy.

TABLE I: Comparison of classification accuracy (%).

n_{tr}	Experiment No.	Method			
		PCA	LDA	LLRDA	ASF-LLRDA
3	1	47.54	47.60	59.34	69.34
	2	47.86	47.86	66.72	74.75
	3	48.00	48.16	57.38	72.79
	4	44.90	45.00	63.77	67.54
	5	47.37	47.37	58.03	65.74
	Average	47.19	47.19	61.05	**70.03**
4	1	55.33	55.50	63.50	72.66
	2	55.66	55.33	73.00	79.50
	3	60.00	63.00	68.66	80.16
	4	62.16	66.16	75.83	86.66
	5	50.83	53.16	63.66	74.50
	Average	56.79	58.63	68.93	**78.69**
5	1	61.52	59.66	70.33	80.16
	2	61.01	67.11	73.05	81.00
	3	60.68	62.88	76.44	83.22
	4	65.25	66.10	80.17	81.69
	5	65.08	68.30	74.57	81.86
	Average	62.70	64.81	74.91	**81.59**
6	1	74.48	73.79	78.96	83.10
	2	66.89	71.03	73.28	77.58
	3	72.93	78.79	81.03	82.58
	4	68.79	73.45	80.34	89.48
	5	71.90	75.86	75.82	86.55
	Average	70.10	74.58	77.88	**83.86**

REFERENCES

[1] Y.-C. Chen, S. C. Hidayati, W.-H. Cheng, M.-C. Hu, and K.-L. Hua, "Locality constrained sparse representation for cat recognition," in *Proc. International Conference on Multimedia Modeling*, 2016, pp. 140–151.

[2] Y.-H. Hsieh, S. C. Hidayati, W.-H. Cheng, M.-C. Hu, and K.-L. Hua, "Who's the best charades player? mining iconic movement of semantic concepts," in *Proc. International Conference on Multimedia Modeling*, 2014, pp. 231–241.

[3] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, "How artificial intelligence will change the future of marketing," *Journal of the Academy of Marketing Science*, vol. 48, pp. 24–42, 2020.

[4] S. C. Hidayati, K.-L. Hua, Y. Tsao, H.-H. Shuai, J. Liu, and W.-H. Cheng, "Garment detectives: Discovering clothes and its genre in consumer photos," in *Proc. IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 471–474.

[5] S. C. Hidayati, T. W. Goh, J.-S. G. Chan, *et al.*, "Dress with style: Learning style from joint deep embedding of clothing styles and body shapes," *IEEE Transactions on Multimedia*, vol. 23, pp. 365–377, 2021.

[6] L. Li, X. Mu, S. Li, and H. Peng, "A review of face recognition technology," *IEEE Access*, vol. 8, pp. 139 110–139 120, 2020.

[7] M. K. Rusia and D. K. Singh, "A comprehensive survey on techniques to handle face identity threats: Challenges and opportunities," *Multimedia Tools and Applications*, vol. 82, pp. 1669–1748, 2023.

[8] M. F. Aza, N. Suciati, and S. C. Hidayati, "Performance study of facial expression recognition using convolutional neural network," in *Proc. International Conference on Science in Information Technology*, 2020, pp. 121–126.

[9] I. Krisnahati, N. Suciati, and S. Chusnul Hidayati, "Face expression recognition with local ternary pattern images using convolutional neural network and extreme learning machine," in *Proc. International Conference on Information Technology, Information Systems and Electrical Engineering*, 2022, pp. 46–50.

[10] T. Thomas, A. P. Vijayaraghavan, and S. Emmanuel, *Dimensionality reduction and face recognition* (Machine Learning Approaches in Cyber Security Analytics). Springer Singapore, 2020.

[11] I. T. Jolliffe, *Principal component analysis*. Springer Series in Statistics, 2002.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[13] N. Zhang and Q. Wu, "Online learning for supervised dimension reduction," *Mathematical Foundations of Computing*, vol. 2, no. 2, pp. 95–106, 2019.

[14] H. Pu, L. Tao, S. Zhenqiu, G. Guangwei, Y. Geng, and Q. Chengshan, "Locality-regularized linear regression discriminant analysis for feature extraction," *Information Sciences*, vol. 429, pp. 164–176, 2018.

[15] W. Wei, H. Dai, and W. Liang, "Regularized least squares locality preserving projections with applications to image recognition," *Neural Networks*, vol. 128, 2020.

[16] G. Wen, L. Jiang, and N. R. Shadbolt, "Using graph algebra to optimize neighborhood for isometric mapping," in *Proc. International Joint Conference on Artificial Intelligence*, 2007.

[17] G. Pai, A. Bronstein, and R. K. R. Talmon, "Deep isometric maps," *Image and Vision Computing*, vol. 123, 2022.

[18] B. Ghojogh, F. Karray, and M. Crowley, "Locally linear image structural embedding for image structure manifold learning," in *Proc. International Conference on Image Analysis and Recognition*, 2019, pp. 126–138.

[19] P. He, T. Jing, X. Xu, L. Zhang, Z. Liao, and B. Fan, "Nonlinear manifold classification based on LLE," in *Proc. Advances in Computer Communication and Computational Sciences*, pp. 227–234.

[20] E. Demaine, A. Hesterberg, F. Koehler, J. Lynch, and J. Urschel, "Multidimensional scaling: Approximation and complexity," in *Proc. International Conference on Machine Learning*, 2021.

[21] S. Su, G. Zhu, and Y. Zhu, "An orthogonal locality and globality dimensionality reduction method based on twin eigen decomposition," *IEEE Access*, vol. 9, pp. 55 714–55 725, 2021.

[22] D. Zhang, J. He, and Y. Zhao, "Kernel sparse representation based dimensionality reduction with applications to image classification," in *Proc. International Conference on Intelligent Information Processing*, 2018, pp. 95–100.

[23] J. Gou, B. Hou, Y. Yuan, W. Ou, and S. Zeng, "A new discriminative collaborative representation-based classification method via l2 regularizations," *Neural Computing and Applications*, vol. 32, pp. 9479–9493, 2020.

[24] Z. Liu, G. Liu, L. Zhang, and J. Pu, "Linear regression classification steered discriminative projection for dimension reduction," *Multimedia Tools and Applications*, vol. 79, pp. 11 993–12 005, 2020.

[25] Y. Xu, Z. Zhang, G. Lu, and J. Yang, "Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification," *Pattern Recognition*, vol. 54, 2016.

[26] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[27] Z.-M. Liu, Y.-Y. Chen, S. Hidayati, S.-C. Chien, F.-C. Chang, and K.-L. Hua, "3d model retrieval based on deep autoencoder neural networks," in *Proc. International Conference on Signals and Systems*, 2017, pp. 290–296.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Joint Optimization Algorithm for Adaptive Bit Allocation Based on Temporal-Spatial Information

Shaokang Wang*†‡ and Songlin Sun*†‡
* School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
† Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, China
‡ Engineering Research Center of Blockchain and Network Convergence Technology, Ministry of Education, China
E-mail: {sk_wang, slsun}@bupt.edu.cn

***Abstract*—Versatile Video Coding (VVC) standard further improves the video compression ratio. Despite inter prediction and intra prediction effectively eliminating temporal and spatial residual information, there is still room for further improvement in rate control (RC) accuracy and rate distortion (R-D) performance. In view of the above issues, the joint optimization RC algorithm for adaptive bit allocation for VVC is proposed. In this paper, the proposed algorithm utilizes a specific pre-analysis mechanism so that the Temporal-Spatial Information (TSI) is obtained. And then the TSI is integrated into the RC model so as to optimize adaptive bit allocations for frame-level and coding tree unit level (CTU-level). Experimental results have shown that compared to the default RC under common test conditions (CTC), the proposed RC algorithm achieves obvious 2.26% Bjøntegaard Delta rate (BD-rate) saving, and the RC accuracy can be improved in the Random Access (RA) configuration.**

I. Introduction

Versatile Video Coding (VVC) [1] [2] is the latest internationally developed video coding standard to meet the growing demands of video services, improve video compression quality, and support today's diverse media content and emerging applications. Rate control (RC) [3] is widely used in the coding and transmission of video and is a significant module for controlling the bits output by encoder. It plays a role in allocating bitrate and adjusting the source output bit stream. In actual video coding, the generated bitrate also fluctuates due to differences in frame type, scene complexity, and coding configurations. The actual video transmission channel needs to introduce RC to ensure smooth transmission of the bit stream and full utilization of channel bandwidth. How to transmit various high-quality videos in different network environments and with limited bandwidth resources has put forward higher requirements for video coding technology. Therefore, the RC technology of general video coding has always been a problem worth studying.

B. Li et al. [4] [5] first proposed R-λ Model in High Efficiency Video Coding (HEVC) [6]. The Lagrange multiplier is directly related to the actual distortion and bitrate of the coding unit (CU). The R-λ model has also been further utilized and integrated into the VVC Test Model (VTM) [7]. The VTM encoder implements a single-pass RC algorithm that provides a constant bitrate (CBR) mode. Many scholars have optimized the RC model. Y. Chen et al. [8] jointly considered the content complexity and temporal dependence based on the CU Tree pre-analysis outputs and performed effective bit allocation. M. Zhou et al. [9] designed a decision tree based on a multiclass classification scheme by fully considering visual features for Ultra High Definition (UHD) videos. C. R. Helmrich et al. [10] proposed a scene change and noise-aware RC method for Versatile Video Encoder (VVenC). H. Liu et al. [11] developed two rate destruction solutions by determining the bit allocation factors for video frames based on their dependence. L. He et al. [12] proposed an adaptive parameter updating and bit allocation algorithm to control the bitrate of frames based on the Quasi-Newton algorithm. Besides the traditional rate distortion (R-D) models, this paper also summarizes schemes based on deep learning. Y. Li et al. [13] trained convolutional neural networks to estimate λ domain model parameters. Y. Mao et al. [14] adopted a pre-analysis framework to obtain the prediction residuals and derived the R-Q and D-Q model parameters by inferring from the prediction residuals with deep neural networks.

However, the RC algorithm in the VVC is not accurate enough at the frame-level and coding tree unit level (CTU-level) bit allocation for videos with some complex local textures, significant edge contours, and motion relationships, which can lead to a loss in coding quality. In the hierarchical coding of VVC, the frame-level bit allocation algorithm doesn't consider the texture complexity difference between frames at the same temporal level. In fact, different frame quality dependencies at the same temporal level also vary, and the consumed bits also vary. In the default CTU-level bit allocation algorithm of VVC, in order to reduce the computational complexity, each bit allocation weight directly follows each current frame-level Lagrange multiplier calculation without considering the difference in temporal importance. Due to the fact that bit allocation using the same Lagrange multiplier is not accurate enough, there is still room for further improvement in the RC accuracy and R-D performance.

In this paper, to address the above issues, a two-stage adaptive bit allocation joint optimization algorithm is proposed based on Temporal-Spatial Information (TSI) through pre-analysis. The contributions of this paper can be summarized as follows: We propose a frame-level bit allocation algorithm based on spatial information at the same temporal level. By

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

considering the impact of intra-frame texture complexity on bit allocation, the bit allocation weights between different frames at the same temporal level are adjusted. And we propose a CTU-level bit allocation algorithm based on temporal and spatial information. By considering the impact of motion and texture information of the CTUs on bit allocation, the CTU-level bit allocation weights are adjusted.

The rest of the paper is organized as follows: Section II briefly explains the workflow of the current integrated RC of VVC. Section III proposes the joint optimization algorithm for adaptive bit allocation based on TSI. Section IV provides the results and analysis of the experiments to verify the effectiveness of the proposed RC. Finally, Section V summarizes the conclusion.

II. Related Work

In this section, the workflow of the current integrated RC model of VVC is summarized [15].

A. Bit Allocation

Bit allocation is sequentially performed at the GOP-level, frame-level, and CTU-level. As the level deepens, the accuracy of RC is also improving. Initially, the encoder calculates the average bits per frame as follows,

$$R_{PicAvg} = \frac{R_{tar}}{f} \tag{1}$$

where R_{tar} is the target bitrate and f is the frame rate.

1) GOP-level: The value of target bits allocated for the current GOP is given as:

$$T_{GOP} = \frac{R_{PicAvg} \cdot (N_{coded} + SW) - R_{coded}}{SW} \cdot N_{GOP} \tag{2}$$

$$SW = \alpha * N_{GOP} / \max(32, IP) + \beta \tag{3}$$

where N_{coded} is the sum of frames already coded and R_{coded} is the bits consumed on these frames. *SW* and *IP* represent an adaptive smoothing window and $IntraPeriod$, respectively. N_{GOP} represents the current GOP size. α and β are set as 20 and 60, respectively.

2) Frame-level: The value of target bits allocated for the current frame is given as:

$$T_{CurrPic} = \frac{T_{GOP} - CODED_{GOP}}{\sum_{i \in \{non-coded\ frames\}} \omega_i} \cdot \omega_{CurrPic} \tag{4}$$

where $CODED_{GOP}$ represents the bits already coded inside the current GOP. $\omega_{CurrPic}$ and ω_i represent the weight of the current frame and the non-coded frames, respectively.

Frames at different temporal levels have quality dependent characteristics and risk of distortion propagation during the coding. In the Random Access (RA) configuration, those lower-level frames will consume more bits because they will be used as references for subsequent higher-level frames of the hierarchical GOP structure. On the contrary, those higher-level frames will consume fewer bits.

3) CTU-level: The value of target bits allocated for the current CTU is given as:

$$T_{CurrCTU} = \frac{\omega_{CurrCTU} - \left(\sum_{i \in \{non\text{-}coded\ CTUs\}} \omega_i - CODED_{CTU}\right)}{SW_{CTU}} \tag{5}$$

where $CODED_{CTU}$ represents the bits consumed while encoding previous CTUs. $\omega_{CurrCTU}$ and ω_i represent the weight of the current CTU and the non-coded CTUs, respectively. $SW_{CTU} = min\,(4, non\ coded\ CTUs)$ and represents a smoothing window used to spread the allocated bits over the next SW_{CTU} CTUs.

B. QP Derivation Based on the R-λ Model

After the available bits are allocated for the current CTU, it is possible to derive the Lagrange multiplier based on the *R-λ* model as follows,

$$\lambda = \alpha \cdot (bpp)^{\beta} \tag{6}$$

where *bpp* represents the bits per pixel in the current CTU. λ represents the slope of the line tangential to the R-D curve as follows,

$$D = CR^{-K} \tag{7}$$

$$\lambda = -\frac{\partial D}{\partial R} = CK \cdot R^{-K-1} \tag{8}$$

where λ, R, D are derived after encoding a frame and the value of C, K are estimated with (7) and (8). And then, the updated C and K can be utilized as the new R-D parameters.

Finally, QP can be derived as:

$$QP = \lfloor 4.2005 \cdot \ln(\lambda) + 13.7122 \rfloor \tag{9}$$

III. Proposed Algorithm

Under the constraints of GOP-level bit allocation and from the perspective of global rate distortion optimization (RDO), this paper improves the R-D performance by pre-analyzing TSI, and utilizing motion and texture based on TSI to jointly optimize the frame-level and CTU-level bit allocation for the reconstructed video more satisfies attention mechanism of human visual system (HVS).

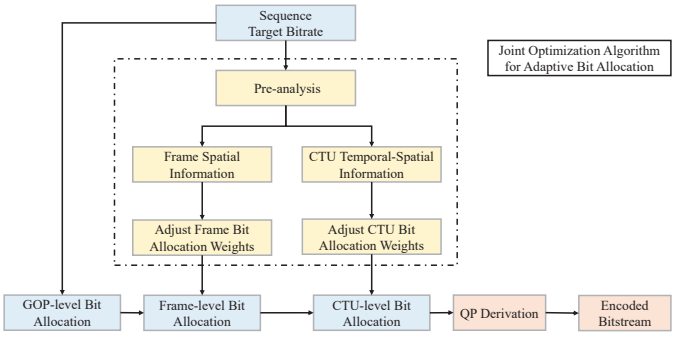


Fig. 1. The workflow of the proposed RC algorithm based on TSI.

Fig. 1 shows the workflow of the proposed joint frame-level and CTU-level bit allocation optimization algorithm. Add a pre-analysis mechanism before encoding to obtain the content information of the frame to be encoded. Specifically, the frames are first stored in the pre-analysis queue to calculate the TSI of the frames and CTUs. After the pre-analysis, the frame is sent to the encoder to complete the bit allocation and QP Derivation.

A. Frame-level Bit Allocation Algorithm

Before encoding, the pre-analysis mechanism for texture complexity is added to the temporal hierarchical structure. By considering the impact of intra-frame texture complexity on bit allocation, the bit allocation weights between different frames at the same temporal levels are adjusted and become more reasonable. Typically, compared with the chroma components (U and V), the luma component (Y) contains more texture features. As a result, the texture features of Y component are pre-analyzed for representing frame's spatial information.

Gradient is used to reflect the fluctuation range of intra-frame pixels, so it can be calculated to characterize the texture complexity of each frame. Take the calculation of Sobel operator with a 3×3 convolutional kernel as an example, convoluting the luma component of a frame with G_x and G_y, as shown in Fig. 2. And then calculate the Gradient of frame (G_{Pic}).

-1	0	1
-2	0	2
-1	0	1

G_x

1	2	1
0	0	0
-1	-2	-1

G_y

Fig. 2. The illustration of Sobel operator.

$$G_{Pic} = \sum_{I \in Pic} |I \odot G_x + I \odot G_y| \tag{10}$$

The bit allocation weight of the current frame is proportionally to each frame's Gradient as follows,

$$\omega_{CurrPic} = \frac{1}{\sum_{i=1}^{S_{Temp}} G_{Pic}\ (i)} G_{CurrPic} \tag{11}$$

$$T_{CurrPic} = \omega_{CurrPic} \cdot \left(\sum_{i=1}^{S_{Temp}} \omega(i) \cdot T_{GOP} - CODED_{Temp} \right) \tag{12}$$

where $\omega(i)$ and $\omega_{CurrPic}$ represent the default bit allocation weight of frames and the adjusted weight of the current frame, respectively. S_{Temp} and $CODED_{Temp}$ represent the sum of frames and the number of bits consumed at the same temporal level, respectively. $T_{currPic}$ represents the target bits of the current frame.

The bit allocation relationship between different temporal layers is still used according to the default algorithm.

B. CTU-level Bit Allocation Algorithm

The motion and texture information of each CTU are pre-analyzed in the current frame as its TSI. If the current frame is non-I frame, temporal information threshold is calculated and all the CTUs of the current frame are divided into important temporal CTUs (CTU_{imp}) and non-important temporal CTUs ($\text{CTU}_{non-imp}$). And then, according to these CTUs' respectively spatial information, the CTU-level bit allocation is performed. On the contrary, if the current frame is I frame, only the spatial information need to be pre-analyzed for CTU-level bit allocation.

1) Temporal-Spatial Information: In this paper, for optimizing CTU-level bit allocation of non-I frame, motion compensated temporary pre filtering (MCTF) is enabled.

Before temporal filtering, motion search will be conducted in multiple frames before and after to obtain the optimal matching block and the motion compensated frame. In the final motion search, each 8×8 block corresponding motion compensation (MC) distortions will be obtained. Based on MC distortion, the similarity between the current frame and the subsequent frames can be estimated, so the motion information in MCTF can be used to estimate the importance of CTUs within the current frame.

After performing MC on the 8×8 blocks, the last step of MCTF is to perform. The sum of the square differences (SSD) between the original block within the current frame and the best MC block within the search frame will be calculated. Based on above, the error (E) will be further calculated as follows,

$$E = 20 \cdot \frac{SSD + 5}{V + 5} + \frac{SSD}{3200} \tag{13}$$

where V is the original Variance within 8×8 blocks.

The E reflects the similarity between the original block within the current frame and the best MC block within the search frame. The more similar it is, the more likely the current block is to be referenced in subsequent encoding. Therefore, the E can be used to estimate the importance of the block, and a lower E value indicates a higher importance. $E1$ can be calculated for each CTU from frames with a distance of 1, and $E2$ can be calculated from frames with a distance of 2. Combining $E1$ and $E2$ can obtain the importance $\hat{E}$. As a result, the importance is pre-analyzed for representing CTU's temporal information as follows:

$$\hat{E} = 1/(\max(E1, E2) + abs(E2 - E1) * 3) \tag{14}$$

If both $E1$ and $E2$ values are small, it indicates that the importance duration is longer, and the current block is important for subsequent references, requiring allocating more bits to encode with the same coding quality or distortion. Due to the fact that the process of pre-analyzing the temporal information of CTU relies on motion information in MCTF,

this algorithm of non-I frame only adjusts the weight of CTU-level bit allocation for the filtered frames in temporal filtering, that is, the POC can only be divided by 8.

The average of temporal information is calculated to represent the temporal threshold (ξ) of this frame as follows,

$$\xi = \frac{1}{S_{CTU}} \sum_{i=1}^{S_{CTU}} \hat{E}(i) \tag{15}$$

Furtherly, according to the temporal threshold, these CTUs of the current frame are divided into two types such as CTU_{imp} and $\text{CTU}_{non-imp}$.

$$\text{CTU}(i) = \begin{cases} \text{CTU}_{imp}, & \hat{E}(i) > \xi \\ \text{CTU}_{non-imp}, & \hat{E}(i) \leq \xi \end{cases} \tag{16}$$

Specifically, if the current CTU's temporal information is less than ξ, it is classified as CTU_{imp}. Otherwise, the CTU is classified as $\text{CTU}_{non-imp}$. In the subsequent bit allocation, CTU_{imp} and $\text{CTU}_{non-imp}$ will be allocated different bits based on TSI.

For the non-I frames, due to the inter relationship, temporal information has a greater impact than spatial information. From this, the CTU-level bit allocation can be divided into two regions: T_{imp} and $T_{non-imp}$ represent important temporal and non-important temporal bits, respectively, as follows,

$$T_{imp} = T_{currPic} \cdot \frac{\sum_{j=x+1}^{S_{imp}} \hat{E}_{impCTU}(j)}{\sum_{i=coded+1}^{S_{CTU}} \hat{E}_{CTU}(i)} \tag{17}$$

$$T_{non-imp} = T_{currPic} \cdot \frac{\sum_{j=y+1}^{S_{non-imp}} \hat{E}_{non-impCTU}(j)}{\sum_{i=coded+1}^{S_{CTU}} \hat{E}_{CTU}(i)} \tag{18}$$

where S_{imp} and $S_{non-imp}$ are the sum of CTU_{imp} and $\text{CTU}_{non-imp}$ within the current frame, respectively. $\hat{E}_{impCTU}$ and $\hat{E}_{non-impCTU}$ are the importance of CTU_{imp} and $\text{CTU}_{non-imp}$, respectively. $coded$ represents the sum of all the CTUs that have been coded. x and y are the sum of coded CTU_{imp} and $\text{CTU}_{non-imp}$, respectively.

$T_{impCTU}(i)$ and $T_{non-impCTU}(i)$ represent the bits allocated to the CTU_{imp} and $\text{CTU}_{non-imp}$, respectively. As a result, furtherly, $T_{impCTU}(i)$ and $T_{non-impCTU}(i)$ can be calculated based on the proportion of the current CTU's Gradient in the sum of Gradient of remaining CTUs.

$$T_{impCTU}(i) = T_{imp} \cdot \frac{G_{impCTU}(i)}{\sum_{n=i}^{S_{imp}} G_{impCTU}(n)} \tag{19}$$

$$T_{non-impCTU}(i) = T_{non-imp} \cdot \frac{G_{non-impCTU}(i)}{\sum_{n=i}^{S_{non-imp}} G_{non-impCTU}(n)} \tag{20}$$

Algorithm 1 The Proposed RC Based on TSI.

Input:
Original video sequence and Target bitrate.
Output:
The TSI of frames and CTUs.
QPs of frames and CTUs.

```
Perform GOP-level model parameter initialization, using (1) to calculate R_PicAvg for each frame;
for i=1 to S_Pic do
  if the current frame is non-I frame and POC%8=0 then
    Allocate frame-level bits with (12);
    Calculate the each CTU's temporal information and the temporal threshold (ξ);
    if Ê(i) > ξ then
      The current CTU is classified as CTU_imp;
      Allocate CTU-level bits with (17) and (19);
    else
      Allocate CTU-level bits with (18) and (20);
    end if
  else
    Allocate frame-level bits with (12);
    for i=1 to S_CTU do
      Calculate G_CTU and allocate CTU-level bits with (21) and (5);
    end for
  end if
  Perform CTU-level and frame-level default QP Derivation and model parameter updating, respectively;
end for
```

where $G_{impCTU}(i)$ and $G_{non-impCTU}(i)$ represent the spatial information of CTU_{imp} and $\text{CTU}_{non-imp}$, respectively.

2) Spatial Information: For I frame, it doesn't have reference frame, so optimizing CTU-level bit allocation based on spatial information by calculating the Gradient of the CTU (G_{CTU}) according to the frame-level mentioned above (10) and (11).

$$\omega_{CurrCTU} = \frac{1}{\sum_{i=1}^{S_{CTU}} G_{CTU}(i)} G_{CurrCTU} \tag{21}$$

where S_{CTU} represents the sum of CTUs. $\omega_{CurrCTU}$ represents the adjusted bit allocation weight of the current CTU.

The target bits of the current CTU can be calculated with (5).

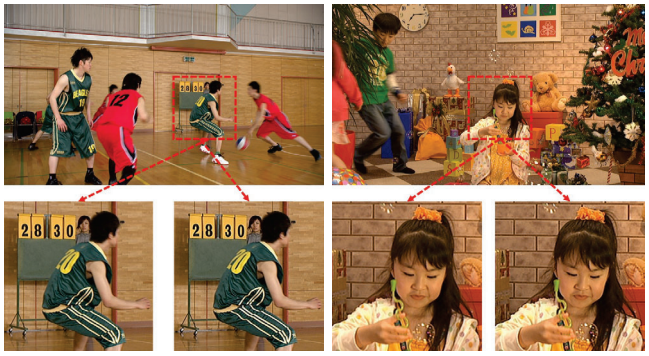
The whole proposed joint optimization algorithm for adaptive bit allocation based on TSI is summarized in Algorithm 1.

IV. Experimental Results and Analysis

Experiments are conducted on the VTM 18.0 under common test conditions (CTC) [16] for verifying the performance of the proposed RC algorithm based on TSI. The target bitrate of each test sequence can be obtained through encoding it in the RA configuration where QPs are set as $\{22,\ 27,\ 32,\ 37\}$.

TABLE I
PERFORMANCE COMPARISON BETWEEN THE PROPOSED RC ALGORITHM AND THE VTM 18.0 DEFAULT RC ALGORITHM

Classes	Sequences	Bitrate Error		BD-rate	$\Delta EncT$
		VTM RC	*Proposed RC*		
A1	Tango2	0.18%	0.10%	-3.15%	101.3%
	FoodMarket4	0.45%	0.49%	-1.71%	101.7%
	Campfire	0.27%	0.37%	-3.81%	103.8%
A2	CatRobot	0.23%	0.11%	-2.77%	101.2%
	DaylightRoad2	1.31%	1.22%	-4.14%	99.5%
	ParkRunning3	0.85%	0.87%	-1.80%	102.4%
B	MarketPlace	0.76%	0.85%	-2.87%	98.9%
	RitualDance	0.35%	0.35%	-1.54%	101.8%
	Cactus	0.37%	0.30%	-2.89%	104.9%
	BasketballDrive	0.11%	0.10%	-3.42%	107.1%
	BQTerrace	0.97%	0.94%	-2.49%	106.5%
C	BasketballDrill	0.60%	0.41%	-1.52%	107.8%
	BQMall	0.94%	0.63%	-1.96%	104.3%
	PartyScene	0.25%	0.23%	-2.27%	106.8%
	RaceHorses	0.28%	0.23%	-2.13%	104.7%
D	BasketballPass	0.56%	0.76%	-1.18%	108.2%
	BQSquare	1.08%	0.80%	-1.28%	111.5%
	BlowingBubbles	1.89%	0.96%	-0.83%	109.6%
	RaceHorses	0.19%	0.52%	-1.19%	107.7%
AVERAGE		0.61%	0.54%	-2.26%	104.7%



(a) (b)

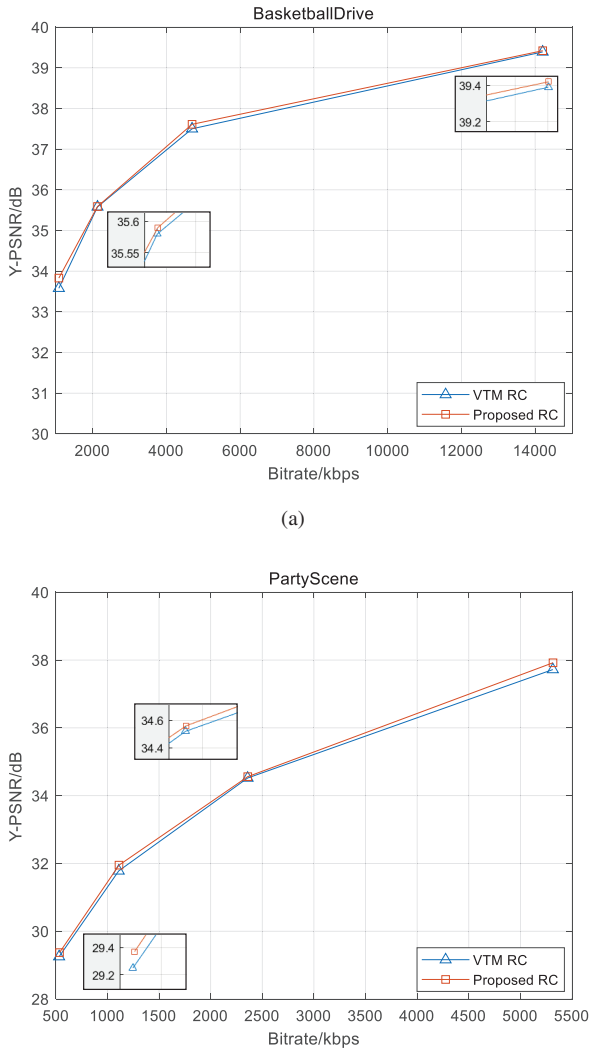
Fig. 3. The subjective quality comparison with the VTM 18.0 default RC and proposed RC of "BasketballDrive" for POC 201 (a) and of "PartyScene" for POC 51 (b). In (a) and (b), the left and right images are reconstructed with the VTM 18.0 default RC and Proposed RC, respectively.

Bitrate error is measured by $Error$. If the $Error$ is smaller, it indicates higher RC accuracy. Coding time complexity is measured by $\Delta EncT$.

$$Error = \frac{|R_{tar} - R_{act}|}{R_{tar}} \times 100\% \tag{22}$$

$$\Delta EncT = \frac{EncT_{proposedRC}}{EncT_{VTMRC}} \times 100\% \tag{23}$$

According to the experimental results shown in TABLE I, the average Bjøntegaard Delta rate (BD-rate) and the average bitrate error of classes A1, A2, B, C, D are reduced by 2.26% and 0.07%, respectively, while only increasing coding time complexity by 4.7% with the proposed RC based on STI in the RA configuration. The majority of the sequences have gained better R-D performance and higher RC accuracy. From the Fig. 3, it can be seen that the subjective quality using the proposed RC is better. As shown in Fig. 4, the "BasketballDrive" and "PartyScene" have obtained 3.42% and 2.27% R-D performance improvement, respectively.



(a)

(b)

Fig. 4. The R-D curves of sequence "BasketballDrive" (a) and "PartyScene" (b).

From above, this proposed algorithm can be verified and analyzed that it is effective in VVC and achieves a trade-off between bitrate and distortion. In addition, the increasing of coding time complexity is moderate, indicating its high practicality, and this algorithm can also be applied to other video coding standards.

V. Conclusions

In this paper, we proposed a frame-level and CTU-level adaptive bit allocation joint optimization algorithm based on TSI which is different from previous studies on the temporal-spatial domain. Our proposed algorithm is performed based on the pre-analysis information and more satisfies attention mechanism of HVS while improving the objective and subjective video coding quality. Specifically, texture complexity and temporal dependence are comprehensively considered to improve the compression performance of video coding in the RA configuration which can dynamically control and adjust local and global bit allocation.

References

[1] B. Bross, Y.-K. Wang, Y. Ye, *et al.*, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021. DOI: 10.1109/TCSVT.2021.3101953.

[2] B. Benjamin, C. Jianle, S. Liu, and W. Ye-Kui, "Versatile video coding (draft 10)," *document JVET-S2001, 19th JVET meeting by teleconference*, 22 June–1 July 2020.

[3] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for h.264/avc video coding and its application to rate control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1533–1544, 2005. DOI: 10.1109/TCSVT.2005.857300.

[4] B. Li, H. Li, L. Li, and J. Zhang, "λ Domain rate control algorithm for high efficiency video coding," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3841–3854, 2014. DOI: 10.1109/TIP.2014.2336550.

[5] L. Li, B. Li, H. Li, and C. W. Chen, "λ -domain optimal bit allocation algorithm for high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 130–142, 2018. DOI: 10.1109/TCSVT.2016.2598672.

[6] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012. DOI: 10.1109/TCSVT.2012.2221191.

[7] Y. Li, Z. Liu, Z. Chen, and S. Liu, "Rate control for versatile video coding," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1176–1180. DOI: 10.1109/ICIP40778.2020.9191125.

[8] Y. Chen, M. Wang, S. Wang, Z. Ni, and S. Kwong, "A ctu-level screen content rate control for low-delay versatile video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023. DOI: 10.1109/TCSVT.2023.3243225.

[9] M. Zhou, X. Wei, W. Jia, and S. Kwong, "Joint decision tree and visual feature rate control optimization for vvc uhd coding," *IEEE Transactions on Image Processing*, vol. 32, pp. 219–234, 2023. DOI: 10.1109/TIP.2022.3224876.

[10] C. R. Helmrich, C. Bartnik, J. Brandenburg, *et al.*, "A scene change and noise aware rate control method for vvenc, an open vvc encoder implementation," in *2022 Picture Coding Symposium (PCS)*, 2022, pp. 241–245. DOI: 10.1109/PCS56426.2022.10018041.

[11] H. Liu, S. Zhu, and B. Zeng, "Inter-frame dependency-based rate control for vvc low-delay coding," *IEEE Signal Processing Letters*, vol. 29, pp. 2727–2731, 2022. DOI: 10.1109/LSP.2023.3234774.

[12] L. He, X. He, S. Xiong, Z. Zhao, H. Xiao, and H. Chen, "Efficient rate control in versatile video coding with adaptive spatial-temporal bit allocation and parameter updating," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022. DOI: 10.1109/TCSVT.2022.3224723.

[13] Y. Li, B. Li, D. Liu, and Z. Chen, "A convolutional neural network-based approach to rate control in hevc intra coding," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4. DOI: 10.1109/VCIP.2017.8305050.

[14] Y. Mao, M. Wang, Z. Ni, S. Wang, and S. Kwong, "Neural network based rate control for versatile video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023. DOI: 10.1109/TCSVT.2023.3262303.

[15] B. Adrian, Y. Yan, and K. Seung Hwan, "Algorithm description for versatile video coding and test model 18 (vtm 18)," *document JVET-AB2002, 28th JVET meeting in Mainz*, 20–27 October 2022.

[16] B. Frank, L. Xiang, S. Karl, S. Vadim, and S. Karsten, "Jvet common test conditions and software reference configurations for sdr video," *document JVET-Y2010, 25th JVET meeting by teleconference*, 12–21 January 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Maximization of 2D Cross-Correlation Based on Auxiliary Function Method for Image Alignment

Yuma Kinoshita*, Kouei Yamaoka†, and Hitoshi Kiya†
* Tokai University, Kanagawa, Japan
† Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—In this paper, we propose an algorithm to maximize the generalized 2D cross-correlation function for image alignment, a cornerstone task in image signal processing and computer vision. As a fast image alignment method, 2D cross-correlation-based methods are often used in real-time applications. These approaches aim to optimize the 2D cross-correlation function between two images, inferring the displacement between them. Nevertheless, image misalignment commonly necessitates a maximization of the continuous cross-correlation function, and existing iterative and approximation-based approaches exhibit limited computational speed or accuracy. In response to this challenge, this paper proposes a novel algorithm to maximize the generalized 2D cross-correlation function. Inspired by Yamaoka et al.'s work for the 1D case, our approach maximizes a quadratic auxiliary function that bounds the 2D cross-correlation from below, instead of directly maximizing the cross-correlation. The accuracy and computational speed of our method, in comparison to three existing techniques, were assessed via a simulation experiment. Our findings confirm the superiority of the proposed method over a quadratic curve fitting approach in terms of image shift estimation accuracy. Remarkably, our proposed method proved approximately 6.8 and 4 times faster than gradient ascent and golden section search per iteration, respectively, competing with the speed of the non-iterative curve fitting.**

I. Introduction

Image alignment or matching is a fundamental problem in image signal processing and computer vision. Pre-alignment of images are required in many tasks: creating panoramas [1], medical imaging [2], image-based 3D modeling [3], high dynamic range imaging [4], [5], and more. Owing to the necessity, various image alignment methods have been proposed.

There are two main approaches to image alignment: homography-based methods and intensity-based methods. The former method endeavors to estimate a homography matrix between two images utilizing features such as SIFT [6], SURF [7], and A-KAZE [8]. RANSAC [9] is frequently employed to estimate the homography matrix. Contemporary research has explored the use of deep learning for estimating the homography matrix [10], [11]. The latter method typically focuses on estimating optical flow [12]–[15]. Nonetheless, these approaches require a significant amount of time for the estimation of either a homography matrix or optical flow.

Given these limitations, two-dimensional (2D) cross-correlation-based methods [16], [17] are often used for image alignment in real-time applications [5]. These strategies seek to maximize the 2D cross-correlation function between two images to deduce the displacement amongst them. The cross-correlation function for discrete pixel shifts can be computed expeditiously using the fast Fourier transform (FFT), and its maximum is easily found.

However, image misalignment usually manifests sub-pixel-wise necessitating maximization of the continuous cross-correlation function. Regrettably, a closed-form solution to this problem remains undiscovered. Although iterative algorithms and approximation-based solutions have been designed to tackle this challenge, these suffer from limited computational speed or accuracy, respectively. For example, the burst photography showcased in [5] adopts an alignment approach that maximizes the 2D cross-correlation function with subpixel accuracy by using a quadratic curve fitting method. It ensures rapid calculation on smartphones, but the limited accuracy of its estimations necessitates more resolution levels used for alignment in a progressive, coarse-to-fine strategy.

Because of such a situation, we aim to develop a fast and accurate maximization algorithm for the 2D cross-correlation function. For maximizing the generalized 1D cross-correlation function considering not only the standard correlation but also the phase-only correlation [18], Yamaoka et al. have proposed a fast and accurate algorithm based on an auxiliary function method which is also known as the majorization-minimization (MM) algorithm [19]. Following the approach, we derive an algorithm to maximize the generalized 2D cross-correlation function. The proposed method can find a (local) maximizer of the objective function by iteratively maximizing a quadratic auxiliary function that bounds the objective function from below. Since the maximizer of quadratic function can be obtained analytically, the proposed method can efficiently maximize the 2D cross-correlation.

We evaluate the accuracy and computational speed of the proposed method and other three methods, by a simulation experiment. Experimental results demonstrate the superior estimation accuracy of the gradient ascent method (GA) with backtracking, the golden section search (GSS), and the proposed method over a quadratic curve fitting approach for image shift estimation, corroborating the effectiveness of iterative methods including our one. Notably, the proposed method is approximately 6.8 and 4 times faster than GA and GSS per iteration, respectively, and could compete with the non-iterative curve fitting.

II. Problem Statement

In this paper, the focus is on the issue of aligning 2D discrete signals, essentially images, denoted as x and y. Their pixel

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

values at a position $\boldsymbol{p} = (p_1, p_2)^\top$, are represented as $x[\boldsymbol{p}]$ and $y[\boldsymbol{p}]$, respectively, where the superscript $\top$ indicates the transpose of a vector or a matrix. We denote the 2D discrete Fourier transforms of x and y, which are $N \times M$ in size, as $\hat{x}$ and $\hat{y}$, respectively. We assume that $\hat{x}$ and $\hat{y}$ are strictly band-limited, i.e., $\hat{x}(\boldsymbol{\omega}) = \hat{y}(\boldsymbol{\omega}) = 0$ for angular frequency $\boldsymbol{\omega}$ such that at least one element is equal to π. By using the 2D cross-spectrum $\hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl}) = \hat{x}^*(\boldsymbol{\omega}_{kl})\hat{y}(\boldsymbol{\omega}_{kl})$, the 2D cross-correlation function of x and y can be written as

$$\check{\Phi}_2^{(xy)}[\boldsymbol{p}] = \frac{1}{NM} \sum_{k \in K} \sum_{l \in L} \hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl}) \exp\left(j \boldsymbol{\omega}_{kl}^\top \boldsymbol{p}\right), \quad (1)$$

where j represents the imaginary unit and $\boldsymbol{\omega}_{kl} = (\omega_k, \omega_l)^\top = (\frac{2\pi k}{N}, \frac{2\pi l}{M})^\top$. The summation range for K and L is given as $K = \{-N/2+1, -N/2+2, \cdots N/2\}$ and $L = \{-M/2+1, -M/2+2, \cdots M/2\}$, respectively.

In applications like image matching, the peak of the cross-correlation function may be sharpened by applying weights to each term in (1). For this reason, we consider the 2D generalized cross-correlation function given as

$$\check{\Phi}_2^{(xy)}[\boldsymbol{p}] = \frac{1}{NM} \sum_{k \in K} \sum_{l \in L} w_{kl} \hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl}) \exp\left(j \boldsymbol{\omega}_{kl}^\top \boldsymbol{p}\right), \quad (2)$$

where $w_{kl} \in \mathbb{R}^+$ is an arbitrary weight. If the weight w_{kl} is set to 1, it is the same as the standard cross-correlation in (1). If $w_{kl} = |\hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl})|^{-1}$, it coincides with the phase-only correlation [18], also known as GCC-PHAT [16] in acoustic signal processing.

Usually, (2) is a discrete function calculated with each element of $\boldsymbol{p}$ being an integer. On the other hand, if $\boldsymbol{p}$ is a real vector, $\check{\Phi}_2^{(xy)}$ can be considered as a continuous function. The purpose of this paper is to estimate a real vector $\boldsymbol{p} \in \mathbb{R}^2$ that maximizes the continuous 2D generalized cross-correlation function, that is, to solve the following optimization problem:

$$\tilde{\boldsymbol{p}} = \underset{\boldsymbol{p} \in \mathbb{R}^2}{\arg\max}\, \check{\Phi}_2^{(xy)}(\boldsymbol{p}). \quad (3)$$

III. Maximizing 2D Cross-Corrlation based on Auxiliary Function Method

Yamaoka et al. have already proposed an algorithm to maximize the 1D generalized cross-correlation using an auxiliary function method [19]. In this paper, following Yamaoka's work, we derive an algorithm that maximizes the 2D generalized cross-correlation.

A. Auxiliary Function Method

The auxiliary function method is an iterative algorithm designed for maximizing an objective function [20]–[22]. It maximizes an auxiliary function corresponding to a lower bound of the objective function, rather than directly maximizing the objective function itself.

To employ the auxiliary function method for solving the optimization problem shown in (3), it is necessary to design an auxiliary function $Q(\boldsymbol{p}, \boldsymbol{\theta})$ that satisfies the following conditions:

- For any given vectors $\boldsymbol{p}$ and $\boldsymbol{\theta}$, the inequality $\check{\Phi}_2^{(xy)}(\boldsymbol{p}) \geq Q(\boldsymbol{p}, \boldsymbol{\theta})$ holds true.
- There exist a $\boldsymbol{\theta} = f(\boldsymbol{p})$ satisfying $\check{\Phi}_2^{(xy)}(\boldsymbol{p}) = Q(\boldsymbol{p}, \boldsymbol{\theta})$ for any $\boldsymbol{p}$.

In this context, $\boldsymbol{\theta} \in \mathbb{R}^2$ is referred to as an auxiliary variable. Given the presence of such a $Q(\boldsymbol{p}, \boldsymbol{\theta})$ and an initial estimate $\tilde{\boldsymbol{p}}^{(0)}$, we can assure the subsequent sequence of updates to converge towards a local maximum:

$$\boldsymbol{\theta}^{(i)} = f(\boldsymbol{p}^{(i)}),\ \boldsymbol{p}^{(i+1)} = \underset{\boldsymbol{p} \in \mathbb{R}}{\arg\max}\, Q(\boldsymbol{p}, \boldsymbol{\theta}^{(i)}). \quad (4)$$

Here, i stands for the iteration index.

B. Deriving Auxiliary Function for 2D Correlation

From the conjugate symmetry of $\hat{\Phi}_2^{(xy)}$, we can rewrite $\check{\Phi}_2^{(xy)}$ as a sum of cosines as

$$\check{\Phi}_2^{(xy)}[\boldsymbol{p}] = \frac{1}{NM} \sum_{k \in K^+} \sum_{l \in L} \alpha_{kl} \cos\left(\boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl}\right), \quad (5)$$

where $K^+ = \{0, 1, \cdots, N/2\}$,

$$\alpha_{kl} = w_{kl} \beta_{kl} \left|\hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl})\right|, \quad (6)$$

$$\beta_{kl} = \begin{cases} 0 & \text{if } k = 0 \text{ and } l < 0 \\ 1 & \text{if } k \in \{0, N/2\} \text{ and } l \in \{0, M/2\} \\ 2 & \text{otherwise} \end{cases}, \quad (7)$$

$$\varphi_{kl} = \angle \hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl}). \quad (8)$$

As shown in Proposition 1 in [19], for any real number θ, the following inequality holds

$$\cos\theta \geq -\frac{1}{2}\frac{\sin\theta_0}{\theta_0}\theta^2 + \left(\cos\theta_0 + \frac{1}{2}\theta_0 \sin\theta_0\right), \quad (9)$$

where $\|\theta_0\| \leq \pi$. Equality holds if and only if $|\theta| = |\theta_0|$ for $\theta_0 < \pi$ and $\theta = (2n+1)\pi, n \in \mathbb{Z}$ for $\theta_0 = \pi$.

Using (9) and the periodicity of $\cos\theta$, we obtain

$$\begin{aligned} &\alpha_{kl} \cos(\boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl}) \\ &\geq -\frac{\alpha_{kl}}{2}\frac{\sin\theta_{kl}}{\theta_{kl}} \left(\boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl} + 2n_{kl}\pi\right)^2 + C, \end{aligned} \quad (10)$$

where $n_{kl} \in \mathbb{Z}$ is such that $|\boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl} + 2n_{kl}\pi| \leq \pi$, and C is a constant term that does not include $\boldsymbol{p}$. Equality holds when

$$\theta_{kl} = \boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl} + 2n_{kl}\pi. \quad (11)$$

By applying (10) to each term of (5), an auxiliary function $Q(\boldsymbol{p}, \boldsymbol{\theta})$ of the objective function (5) is derived as follows:

$$\begin{aligned} &\check{\Phi}_2^{(xy)}[\boldsymbol{p}] \geq Q(\boldsymbol{p}, \boldsymbol{\theta}) \\ &= \sum_{k \in K^+} \sum_{l \in L} -\frac{\alpha_{kl}}{2NM}\frac{\sin\theta_{kl}}{\theta_{kl}} \left(\boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl} + 2n_{kl}\pi\right)^2 \\ &+ C. \end{aligned} \quad (12)$$

C. Update Rules

Since $Q(\boldsymbol{p}, \boldsymbol{\theta})$ is a quadratic function in terms of $\boldsymbol{p}$, the maximizer of Q is easily obtained by solving an equation

$$\begin{aligned}&\nabla Q(\boldsymbol{p}, \boldsymbol{\theta})\\ &= \sum_{k\in K^+}\sum_{l\in L} -s_{kl}\left(\boldsymbol{\omega}_{kl}^\top \boldsymbol{p} + \varphi_{kl} + 2n_{kl}\pi\right)\boldsymbol{\omega}_{kl}\\ &= \mathbf{0}. \end{aligned} \tag{13}$$

In the above, ∇ indicates the gradient for $\boldsymbol{p}$ and $s_{kl} = \alpha_{kl}\frac{\sin\theta_{kl}}{\theta_{kl}}$. This equation yields, $\mathbf{A}\boldsymbol{p} = \boldsymbol{b}$, defined as:

$$\mathbf{A} = \sum_{k\in K^+}\sum_{l\in L} s_{kl}\boldsymbol{\omega}_{kl}\boldsymbol{\omega}_{kl}^\top, \tag{14}$$

$$\boldsymbol{b} = \sum_{k\in K^+}\sum_{l\in L} s_{kl}\left(-\varphi_{kl} - 2n_{kl}\pi\right)\boldsymbol{\omega}_{kl}. \tag{15}$$

Considering the condition for equality (11), we can substitute $-\varphi_{kl} - 2n_{kl}\pi = \boldsymbol{\omega}_{kl}^\top \boldsymbol{p} - \theta_{kl}$, and obtain the following update rules:

$$n_{kl}^{(i)} \leftarrow \arg\min_{n\in\mathbb{Z}} \left|\boldsymbol{\omega}_{kl}^\top \boldsymbol{p}^{(i)} + \varphi_{kl} + 2n\pi\right|, \tag{16}$$

$$\theta_{kl}^{(i)} \leftarrow \boldsymbol{\omega}_{kl}^\top \boldsymbol{p}^{(i)} + \varphi_{kl} + 2n_{kl}^{(i)}\pi, \tag{17}$$

$$\boldsymbol{p}^{(i+1)} \leftarrow \left(\mathbf{A}^{(i)}\right)^{-1}\boldsymbol{b}^{(i)} = \boldsymbol{p}^{(i)} - \left(\mathbf{A}^{(i)}\right)^{-1}\boldsymbol{v}^{(i)}, \tag{18}$$

where

$$s_{kl}^{(i)} = \alpha_{kl}\frac{\sin\theta_{kl}^{(i)}}{\theta_{kl}^{(i)}}, \tag{19}$$

$$\mathbf{A}^{(i)} = \sum_{k\in K^+}\sum_{l\in L} s_{kl}^{(i)}\boldsymbol{\omega}_{kl}\boldsymbol{\omega}_{kl}^\top, \tag{20}$$

$$\boldsymbol{v}^{(i)} = \sum_{k\in K^+}\sum_{l\in L} s_{kl}^{(i)}\theta_{kl}^{(i)}\boldsymbol{\omega}_{kl}. \tag{21}$$

IV. Simulation

To evaluate the performance of the proposed method in terms of the estimation accuracy and computational cost, we conducted a simulation.

A. Experimental Conditions

In this simulation, we numerically prepared misaligned image pairs from 20 well-known standard test images such as *airplane, barbara, cameraman, mandrill*, and so on[1]. An example of the misaligned image pair is shown in Fig. 1. The size of the original images was 256×256 pixels. To generate shifted images, the original images were shifted both vertically and horizontally, with the cubic interpolation. Each shift amount $\boldsymbol{p}^*$ was randomly selected from the uniform distribution with a range of $[-10, 10]$ pixels. Areas that the image shift causes to fall outside the original image boundaries were cropped. Finally, we added Gaussian noise with a standard deviation of 0.1 to both original and shifted images, independently.

[1]They can be downloaded from https://sipi.usc.edu/database/ and http://www.ess.ic.kanagawa-it.ac.jp/app_images_j.html



(a) Original (b) Shifted

Fig. 1. Example of shifted image (*Pepper*)

TABLE I
Machine spec used for evaluating executing time

Processor	AMD Ryzen 7 5800X (3.8 GHz, 8 cores)
Memory	64 GB
OS	Ubuntu 20.04 LTS
Language	Python 3.9.11
Library	PyTorch 2.0.1

The proposed auxiliary function-based method (Aux.) was compared with three methods: the gradient ascent (GA), the 2D golden section search (GSS) in [23], and a quadratic curve fitting in [5]. GA iteratively updates the maximizer $\tilde{\boldsymbol{p}}$ as $\tilde{\boldsymbol{p}}^{(i+1)} = \tilde{\boldsymbol{p}}^{(i)} + \eta\nabla\check{\Phi}_2^{(xy)}(\tilde{\boldsymbol{p}}^{(i)})$, where the step size η is determined by the backtracking line search following the Armijo rule [24]. GSS is also an iterative algorithm, but it does not require the gradient of the objective function. For GA, GSS, and the proposed method, we set the number of iterations to 50 and the initial estimation $\boldsymbol{p}^{(0)}$ was set to the maximizer of the discrete cross-correlation $\check{\Phi}_2^{(xy)}[\boldsymbol{p}]$ with $w_{kl} = 1$.

The curve fitting can be found an approximate solution by fitting a quadratic curve to the discrete function and finding the maximum of the fitted curve. Although a quadratic curve is fit to the $L2$ distance in [5], we used $\check{\Phi}_2^{(xy)}[\boldsymbol{p}]$ with $w_{kl} = 1$ for the curve fitting similarly to the other methods.

We implemented all methods in Python with PyTorch. The estimation error between $\tilde{\boldsymbol{p}}$ estimated by methods above mentioned and true shift $\boldsymbol{p}^*$ was measured by the root mean square error (RMSE).

The simulation was run on a PC, with a 3.8 GHz processor and a main memory of 64 Gbytes (see Table I). For measuring the executing time, *time.perf_counter()* function in the Python standard library was used, and GPUs were not used in the simulation

B. Results

Table II shows the RMSE between the estimated and true shifts for five images in the test set and the average and standard deviation of the RMSE for all 20 images. From the result, we can see that GA, GSS, and the proposed method achieved the higher estimation accuracy than the curve fitting. For this reason, it is confirmed that the iterative methods are more effective to obtain the accurate shift estimation than the curve fitting.

TABLE II
RMSE BETWEEN ESTIMATED AND TRUE SHIFTS

Image	GA	GSS [23]	Curve fitting [5]	Aux. (ours)
Airplane	0.0158	0.0158	0.0489	0.0158
Barbara	0.0453	0.0453	0.1900	0.0453
Cameraman	0.0139	0.0140	0.0157	0.0140
Pepper	0.1489	0.1489	0.2241	0.1489
Woman	0.0448	0.0454	0.0820	0.0454
Average (20 images)	0.0679	0.0679	0.1202	0.0680
Std. (20 images)	0.0391	0.0391	0.0612	0.0391

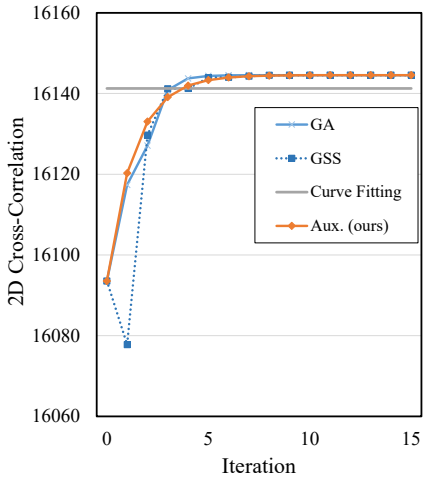


Fig. 2. Values of objective function during iterations for *Pepper*. Since curve fitting is not an iterative method, objective function value is drawn as line

TABLE III
ELAPSED TIME (MS) PER ONE ITERATION FOR ITERATIVE METHODS

Statistic	GA	GSS [23]	Curve fitting [5]	Aux. (ours)
Average	4.793	2.801	4.428	0.705
Std.	1.103	0.097	0.144	0.017

The values of the objective function during iteration are shown in Fig. 2. Here, the input image pair was *Pepper* shown in Fig. 1. From the figure, the objective values for GA and GSS converged mostly to the local maximum after six iterations and eight iterations for the proposed method. The trajectory of the estimated shift $\tilde{\boldsymbol{p}}^{(i)}$ for each method is shown in Figs. 3 to 5. Note that the maximizer of the 2D cross-correlation function $\check{\Phi}_2^{(xy)}(\boldsymbol{p})$ was slightly different from the true shift $\boldsymbol{p}^*$. A reason for this is that the input images were cropped after shifting.

Average elapsed times of the four methods for the 20 images are shown in Table III, where, for the iterative methods, the elapsed time for 50 iterations was measured and the elapsed time per iteration was calculated. As shown in the table, the proposed method is about 6.8 times faster than GA and about 4 times faster than GSS, and if the proposed method converges in about 8 iterations, as shown in Fig. 2, it is competitive with the non-iterative curve fitting.

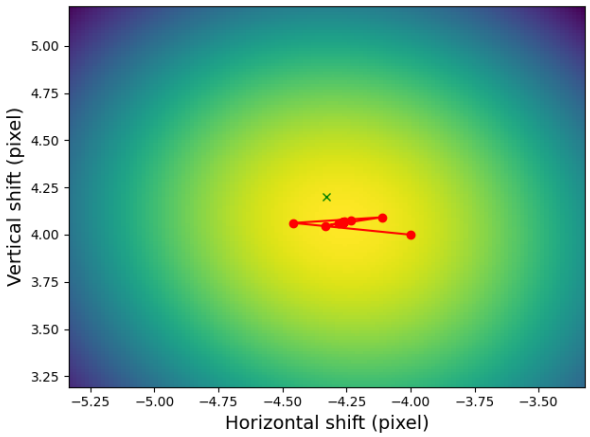


Fig. 3. Trajectories of estimated shifts by GA for image *Pepper* , where red points indicate estimations $\boldsymbol{p}^{(i)}$ and cross indicates true shift $\boldsymbol{p}^*$.

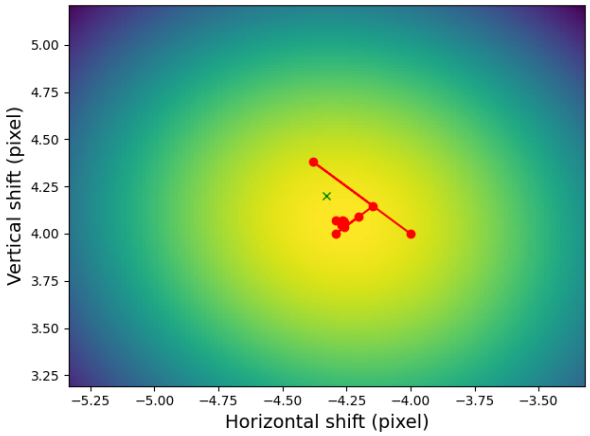


Fig. 4. Trajectories of estimated shifts by GSS [23] for image *Pepper* , where red points indicate estimations $\boldsymbol{p}^{(i)}$ and cross indicates true shift $\boldsymbol{p}^*$.

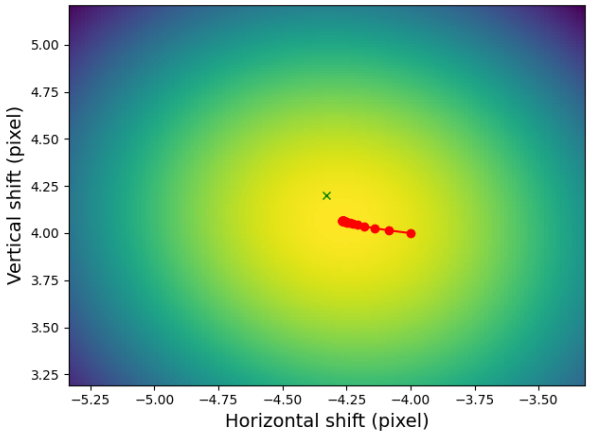


Fig. 5. Trajectories of estimated shifts by Aux. (ours) for image *Pepper* , where red points indicate estimations $\boldsymbol{p}^{(i)}$ and cross indicates true shift $\boldsymbol{p}^*$.

V. CONCLUSIONS

In this paper, we derived a maximization algorithm for the 2D generalized cross-correlation function, on the basis of the

auxiliary function based 1D cross-correlation maximization. The proposed method find the maximizer of the 2D cross-correlation function by iteratively maximizing a quadratic auxiliary function that bounds the 2D cross-correlation function from below. Experimental results showed that the proposed method can estimate the displacement between two images more accurately than a quadratic curve fitting method. In addition, the proposed method is faster than the gradient ascent method the golden section search, and the executing time is competitive with that of the curve fitting method.

REFERENCES

[1] M. Brown and D. G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features," *Int. J. Comput. Vision.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.

[2] G. Haskins, U. Kruger, and P. Yan, "Deep Learning in Medical Image Registration: A Survey," *Machine Vision and Applications*, vol. 31, no. 1-2, p. 8, Feb. 2020.

[3] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the World from Internet Photo Collections," *Int J Comput Vis*, vol. 80, no. 2, pp. 189–210, Nov. 2008.

[4] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust Patch-Based HDR Reconstruction of Dynamic Scenes," *ACM Trans. Graph.*, vol. 31, no. 6, 203:1–203:11, Nov. 2012.

[5] S. W. Hasinoff, D. Sharlet, R. Geiss, *et al.*, "Burst Photography for High Dynamic Range and Low-Light Imaging on Mobile Cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[6] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proc. ECCV*, vol. 3951, May 2006, pp. 404–417.

[8] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE Features," in *Proc. ECCV*, vol. 7577, Oct. 2012, pp. 214–227.

[9] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[10] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative Deep Homography Estimation," in *Proc. IEEE/CVF CVPR*, Jun. 2022, pp. 1869–1878.

[11] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao, and S. Liu, "Unsupervised Homography Estimation with Coplanarity-Aware GAN," in *Proc. IEEE/CVF CVPR*, Jun. 2022, pp. 17 642–17 651.

[12] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proc. DARPA Image Underst. Workshop*, Apr. 1981, pp. 121–130.

[13] A. Dosovitskiy, P. Fischer, E. Ilg, *et al.*, "FlowNet: Learning Optical Flow with Convolutional Networks," in *Proc. IEEE ICCV*, Dec. 2015, pp. 2758–2766.

[14] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks," in *Proc. IEEE/CVF CVPR*, Jul. 2017, pp. 2462–2470.

[15] H. Jung, Z. Hui, L. Luo, *et al.* "Anyflow: Arbitrary Scale Optical Flow with Implicit Neural Representation." arXiv: 2303 . 16493. (Mar. 29, 2023), [Online]. Available: http://arxiv.org/abs/2303.16493.

[16] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[17] K. Takita, T. Higuchi, and K. Kobayashi, "High-Accuracy Subpixel Image Registration Based on Phase-Only Correlation," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. 86, no. 8, pp. 1925–1934, Aug. 1, 2003.

[18] C. D. Kuglin, "The Phase Correlation Image Alignment Method," in *Proc. Int. Conf. Cybern. Soc.*, Sep. 1975, pp. 163–165.

[19] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-Sample Time Delay Estimation Via Auxiliary-Function-Based Iterative Updates," in *Proc. IEEE WASPAA*, Oct. 2019, pp. 130–134.

[20] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Trans. on Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[21] K. Lange, *MM Optimization Algorithms*. Philadelphia: SIAM, Jul. 2016.

[22] D. R. Hunter and K. Lange, "A Tutorial on MM Algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[23] C. He, Y. Zheng, and S. Ahalt, "Object Tracking Using the Gabor Wavelet Transform and the Golden Section Algorithm," *IEEE Trans. Multimed.*, vol. 4, no. 4, pp. 528–538, Feb. 2002, ISSN: 1941-0077.

[24] L. Armijo, "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, Jan. 1, 1966.

Multitask Record for Badminton Match

Jing-Ming Guo, Yu-Shun Huang, Ting-Yu Chang, Tai-Cyuan Ciou, Yun-Ching Yeh, and Jeffrey Chen
National Taiwan University of Science and Technology, Taiwan
E-mail: jmguo@mail.ntust.edu.tw Tel: +886- 227303241
E-mail: tw.yshuang@gmail.com Tel: +886-932832179
E-mail: tw.tychang@gmail.com Tel: +886-954079118
E-mail: tw.tcciou@gmail.com Tel: +886-908982776
E-mail: tw.ycyeh@gmail.com Tel: +886-989829968
E-mail: jeffrey97059@gmail.com Tel: +886-966313988

Abstract— A framework simplifies a novel multi-task for badminton, automating detection in tournament videos and game information prediction. The model combines EfficientNetv2 with distinct module networks for different tasks. This approach achieves a balance between efficiency and precision via a streamlined model architecture.

I. Introduction

Sports event analysis poses challenges in image analysis and machine learning. This task involves a sequential multi-task problem. While Transformer[1] and Multi-Level MTL[2] models are common choices, we've abstained due to suboptimal computational efficiency, which doesn't align with our speed requirements. Moreover, MTL encounters challenges with tasks of low interdependence, such as estimating ball coordinates and shot postures, leading to interference during training and performance decline. Given these constraints and task demands, we've adopted alternative approaches for an effective and efficient solution.

II. Related Work

EfficientNetV2[3] is an efficient neural network design aimed at enhancing model performance while maintaining efficiency. It offers variants of varying sizes to accommodate diverse applications. It employs the MBConv module, which decomposes convolutional operations into depth-wise and point-wise convolutions, reducing computation and parameters. To enhance the efficiency of representation learning, the authors introduce the Fused-MBConv module in shallower layers, combining depth-wise separable convolution and a linear bottleneck mechanism. EfficientNetV2 is acclaimed for its efficiency and robust performance, achieved through depth-wise separable convolution, linear bottleneck, and feature pyramid networks

III. Datasets

The dataset comprises competition videos and associated analysis data, denoted as L. These contest videos are sourced from prominent international badminton matches endorsed by the Badminton World Federation (BWF). The analysis data L encompasses various categories, as detailed in Table 1. The training dataset encompasses around 5,000 shots, while the testing dataset comprises approximately 1,000 shots.

Table 1. categories of $\mathcal{L}$

Categories	Meaning
HitFrame	The frame number when the *Hit* occurs in the video.
Hitter	Which player hit the current shot.
RoundHead	Whether the hitting player executed a round-the-head stroke.
Backhand	Whether the hitting player executed a backhand stroke
BallHeight	Whether the hitting point is above the net.
LandingX *LandingY*	The coordinates of the trajectory endpoint, projected onto the court plane.
HitterLocationX *HitterLocationY*	The coordinates of the hitter's dominant foot's tiptoe in the frame.
DefenderLocationX *DefenderLocationY*	The coordinates of the defender's dominant foot's tiptoe in the frame.
BallType	The type of shot played in each Hit. The ball types are sorted as follows: Shot serve, long serve, Clear, Drive, Smash, Net shot, Slice, Drop shot, Push or block shot.

The training dataset consists of approximately 5,000 shots, while the testing dataset consists of approximately 1,000 shots.

IV. Method

A. Processing

This section clips videos into *Tubes* and packages with the corresponding *Label*. *Tubes* are synthesized by a range of frames. based on the *Side*, which determines the number of frames extracted around the *Hit* frames. The options for *Side* are {1, 2, 3}. During the Training stage, the *HitFrame* is obtained from the $\mathcal{L}$ dataset, and 6 frames around the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

HitFrame are preserved. The required frames, $F_{require}$, are randomly selected while ensuring include the *Hit* frame using (1), shown in Fig. 1.

$$\mathcal{F}_{require} = random.choise(range(6 - Side, 6))) \quad (1)$$

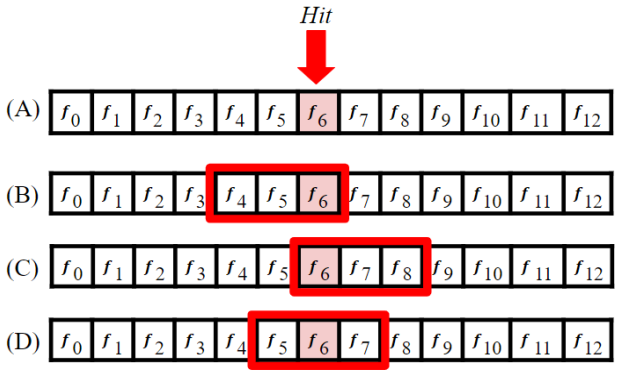


Fig. 1. (A) f_6 containing the detected Hit. (B), (C), and (D) depict the randomly selected 3 frames (*Side* = 1).

The *Labels* are categorized into two types: Regression and Classification [4]. For Regression, only consider six *X* and *Y* coordinates and normalize them into 0 to 1. Others belong to Classification and are converted into corresponding 0 and 1. Also, choose Classification for *HitFrame* because the given value represents which frame the *Hit* occurs, but the model only receives a few frames, lacking the information to predict. Therefore, the model is provided with several options and determines the probability P_t of each frame to be the *HitFrame*.

The index with the highest probability is selected, denoted as $P_{HitIndex}$, as in (2). By adding the starting frame number t, the *HitFrame* probability $P_{HitFrame}$ can be reconstructed using (3).

$$P_{HitIndex} = argmax(P_t, P_{t+1}, P_{t+2}) \quad (2)$$

$$P_{HitFrame} = t + P_{HitIndex} \quad (3)$$

During the Testing stage, a sliding window technique is employed to sequentially select frames, create *Tubes*, and predict, until the entire video is predicted.

B. Data Augmentation

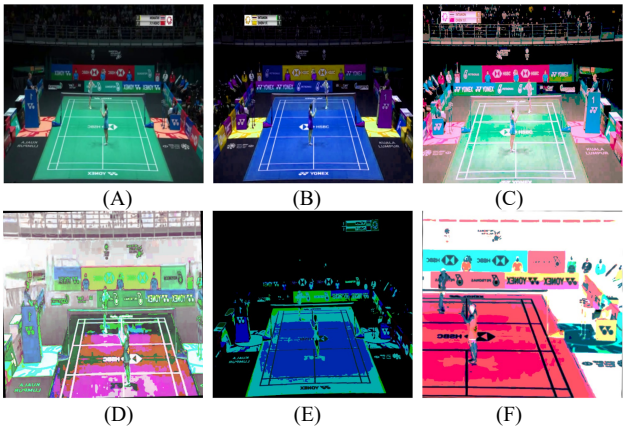


Fig. 1. (A) Original frame. (B) and (C) chosen augmentation parameter result. (D), (E) and (F) other parameter result.

Several algorithms including *Gaussian Blur*[5], *Color Jitter*[6], *Random Equalize*[7], *Random Horizontal Flip*, *Random Rotation*, and *Random Resized Crop* have been employed to enhance the dataset. Also, *Elastic Transform* has been applied to imitate the shacking of the net or the camera.

The augmentation parameters are specially designed so that the transformed image doesn't lose its feature, shown in Fig. 2.

C. Model

1. Architecture

As shown in the Fig. 2, the structure of the model is as follows:

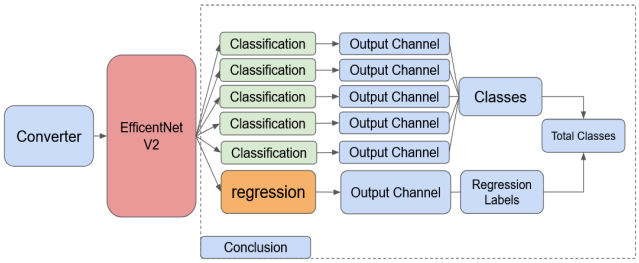


Fig. 2. The structure of the model established by Converter, EfficientNetV2 and Linear module.

1) Converter

The *Converter* comprises two components, each featuring a Conv3D[8] layer with ReLU activation. The distinction arises from an additional BN[9] layer in the first component, absent in the second. The primary function of the first component is to reduce the depth of input data to 3, followed by a further reduction to 1 in the second component. These convolutional operations perform dimensionality reduction and non-linear transformations on the input data, thereby endowing the model with the capability to comprehend sequential concepts while simultaneously achieving dimensionality reduction.

2) EfficientNetV2

EfficientNetV2 is chosen because of its ability to deliver better performance with fewer parameters. It adopts a non-uniform expansion strategy, gradually adding more layers in the later stages. The Fuse-MB Conv layer solves the problem of lightweight models not utilizing the GPU/CPU fully. The authors use NAS to search for the optimal combination of fused and regular MB Conv layers. Smaller models achieve better performance when some early MB Conv layers are replaced with fused layers. EfficientNetV2 provides a flexible network architecture that balances training speed and accuracy.

3) Linear module

The extracted outputs are propagated through a linear neural network module comprising three linear layers and two ReLU[10] activation functions. Within this module, Softmax

activation is employed for the classification task, while it is not utilized for regression task.

2. Backpropagation

According to the concept of backpropagation[11], the gradient vectors computed by each Cost Function are independent. When a network architecture is shared by different Cost Functions, the weights and biases within the network are independently calculated by each Cost Function, resulting in a weighted gradient.

3. Loss

Cross-Entropy is utilized as the loss function for the classification module, while MSE serves as the loss function for the regression module. As the labels in the regression module are normalized within the range of 0 to 1, the MSE-calculated gradient does not overpower the overall loss. This ensures a balanced weighting and similar significance between MSE and Cross-Entropy.

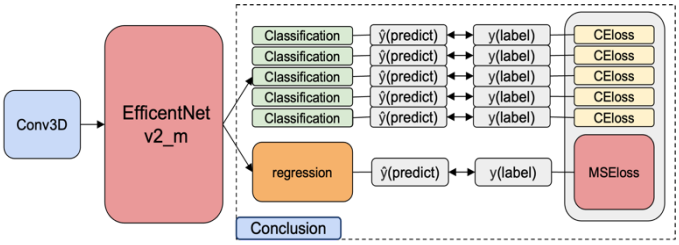


Fig. 4. The loss function design structure.

4. Optimizer

Various optimizers were employed, and a comparison of the differences in loss was conducted.

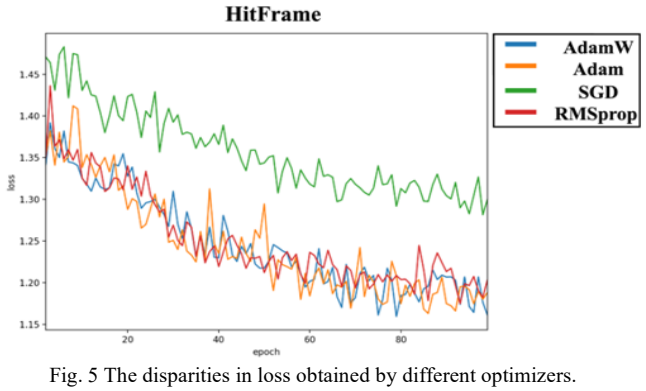


Fig. 5 The disparities in loss obtained by different optimizers.

In Figure 5, it is evident that AdamW[12] exhibited the best performance.AdamW[12] is employed as the optimizer, which receives gradients computed by all the loss functions. As each Linear module possesses its own specific loss function, the optimizer updates the parameters based on the corresponding loss function of each Linear module. However, in EfficientNetV2, the parameters are updated by aggregating a weighted combination of gradients all the loss functions.

V. Experiments

A. Category

Training: Our models are trained in PyTorch using the standard AdamW optimizer. We set a Side value of 1 per epoch and maintain a batch size of 15. We employ EfficientNetV2-M with either Fused-MBConv or MBConv at each stage of our architecture. To expedite training, we converted the data to pickle files.

Testing: Data is resized to 512x512 and normalized for testing.

Table 2. Accuracy of the original model, *Side*=1 and AdamW

Type	Accuracy	Type	Accuracy
Hitframe	56.6%	**LandingX**	89.1%
Hitter	99.9%	**LandingY**	88.1%
RoundHead	96.9%	**HitterLocationX**	93.0%
Backhand	97.2%	**HitterLocationY**	88.5%
BallHeight	96.1%	**DefenderLocationX**	95.0%
BallType	51.1%	**DefenderLocationY**	87.5%

Results in Table 2 presents data for nine categories, including *HitFrame* and *Hitter*. Among these, *Landing*, *DefenderLocation*, and *HitterLocation* predict coordinates individually. *HitFrame* and *BallType* exhibit relatively lower accuracy due to limited data for successful hits and the complexity of nine ball types. Increasing model depth might improve *BallType* classification. However, other categories demonstrate effective classification, affirming our use of distinct loss functions and optimizers. Accuracy and loss curves are in Figs. 6 and 7.

B. Ablation Studies

1. Side

In the Method section, we comprehensively described the Tube generation process using Side=1 as an example. We also examined Tubes for *Side*=2 and *Side*=3. We selected different frame configurations because they provide varying levels of information. However, we considered the model's overall learning and convergence capabilities. Increasing information doesn't guarantee improved performance. Our primary focus in the ablation experiments was to find the optimal data mixing ratio for each *Side*, aiming for the lowest loss and highest accuracy. Evaluation results in Fig. 8 and 9 showed that *Side*=1 had the highest recognition ability among all *HitFrames*.

VI. Conclusion

This article presents a study on badminton recognition using EfficientNetV2 as the training model. The focus is on efficiently and effectively categorizing each hit in a video. The model performs multi-task detection, encompassing regression

and classification of each Hit. We found that the label unbalanced caused the model unstable, it will be one of the main focuses in the future. This article is the first publication based on the dataset, contributing to the dissemination of the data and laying the foundation for the future development of a comprehensive badminton recognition system.

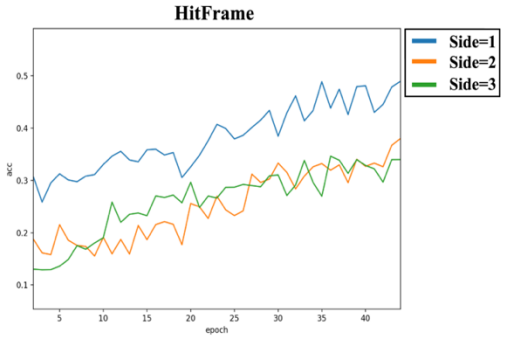


Fig. 6 Accuracy of different Side

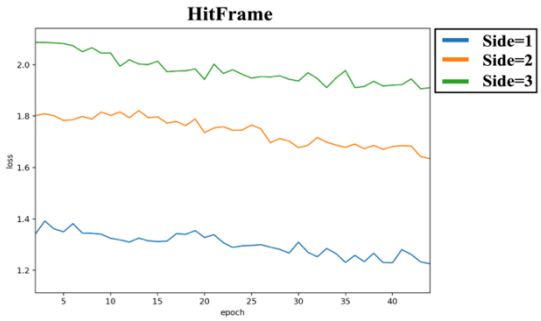


Fig. 7 Loss of different Side

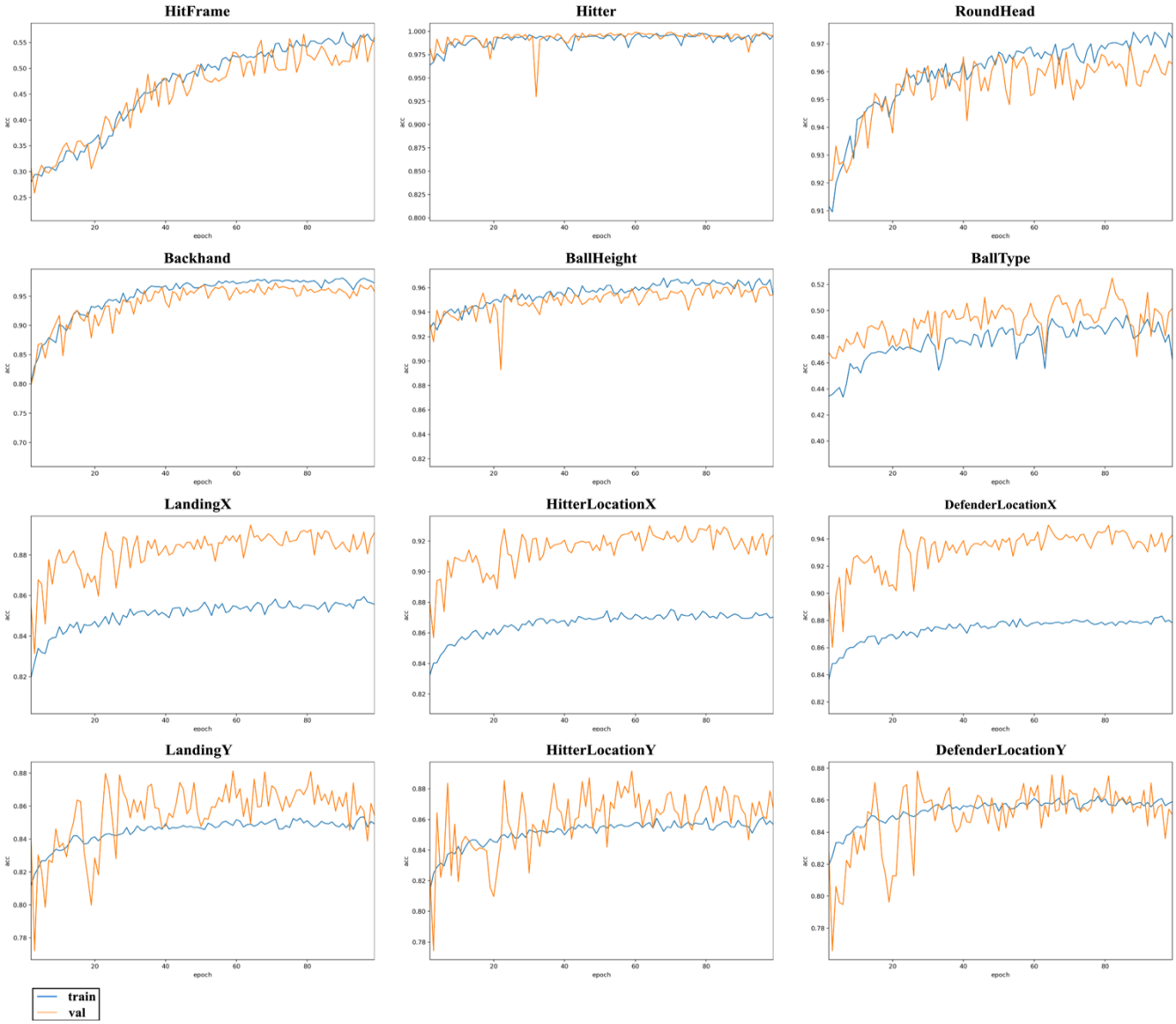


Fig. 8 Accuracy of each category, and Side=1 and AdamW

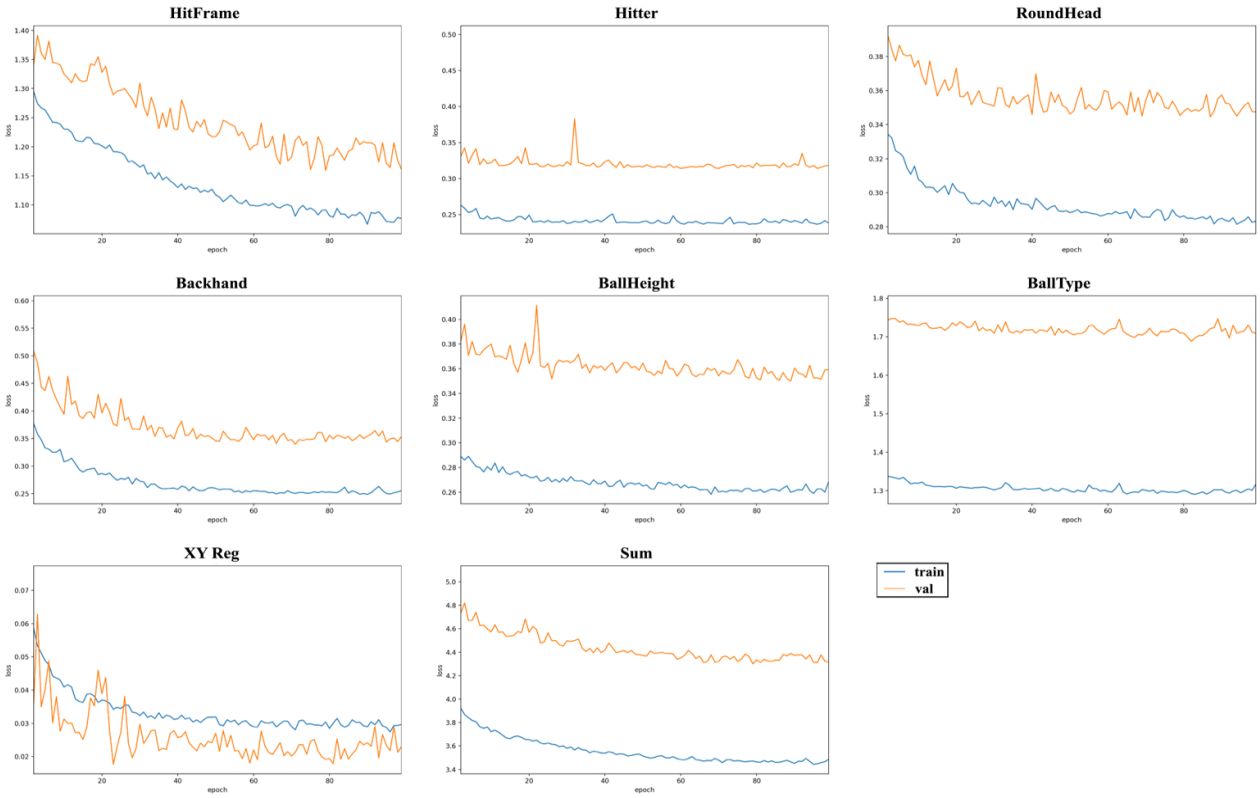


Fig. 9. Loss of each category and Side=1 and AdamW

REFERENCES

[1] F. Giuliari, I. Hasan, M. Cristani and F. Galasso, "Transformer Networks for Trajectory Forecasting," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 10335-10342, doi: 10.1109/ICPR48806.2021.9412190.

[2] O. A. Wahab, A. Mourad, H. Otrok and T. Taleb, "Federated Machine Learning: Survey, Multi-Level Classification, Desirable Criteria and Future Directions in Communication and Networking Systems," in IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 1342-1397, Secondquarter 2021, doi: 10.1109/COMST.2021.3058573.

[3] TAN, Mingxing; LE, Quoc. Efficientnetv2: Smaller models and faster training. In: International conference on machine learning. PMLR, 2021. p. 10096-10106.

[4] A. Landi, P. Piaggi, M. Laurino and D. Menicucci, "Artificial Neural Networks for nonlinear regression and classification," 2010 10th International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, 2010, pp. 115-120, doi: 10.1109/ISDA.2010.5687280.

[5] E. S. Gedraite and M. Hadad, "Investigation on the effect of a Gaussian Blur in image filtering and segmentation," Proceedings ELMAR-2011, Zadar, Croatia, 2011, pp. 393-396.

[6] Han, Ligong, et al. "Constructive Assimilation: Boosting Contrastive Learning Performance through View Generation Strategies." arXiv preprint arXiv:2304.00601 (2023).

[7] Somavarapu, N., Ma, C. Y., & Kira, Z. (2020). Frustratingly simple domain generalization via image stylization. arXiv preprint arXiv:2006.11207.

[8] TRAN, Du, et al. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 4489-4497.

[9] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. pmlr, 2015.

[10] AGARAP, Abien Fred. Deep learning using rectified linearunits (relu). arXiv preprint arXiv:1803.08375, 2018.

[11] RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by backpropagating errors. nature, 1986, 323.6088: 533-536.

[12] Ruder, Sebastian. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747, 2016.

Deep Residual and Classified Neural Networks for Inverse Halftoning

Jing-Ming Guo*, Sankarasrinivasan Seshathiri† and L. V. Hung#
*†#National Taiwan University of Science and Technology
* E-mail: jmguo@seed.net.tw Tel: +886-27303241
†#E-mail: D10507805@mail.ntust.edu.tw, m10607812@mail.ntust.edu.tw Tel: +886-978374981

Abstract—An effective approach to solve the ill-posed problem of Inverse Halftoning is presented in this study using a combination of deep residual neural network and variance classification. The proposed method involves a two-stage process that utilizes the concept of progressive learning. In the first stage, the proposed deep neural network extracts intrinsic features from the image and eliminates halftone patterns. In the second stage, the network incorporates multiple deep residual blocks to restore fine details with high accuracy. The model employs various neural nets trained over different statistical ranges, taking into account the statistics of halftone patches. The proposed approach outperforms other methods and achieves superior reconstruction quality.

Keywords: Inverse halftoning, progressive learning, residual neural networks, variance classification, convolutional neural network

I. INTRODUCTION

Digital halftoning [1] is used to obtain the two-tone or binary equivalent of the continuous tone image. The technique is predominantly used in the printing industry specifically to obtain black and white printouts. When the printed image is viewed from a certain distance, it is perceived as a continuous tone image due to the low pass nature of the human eyes. This conversion helps us to create the illusion of the continuous tone image (various gray shades) using a single black cartridge ink. Briefly, the objective of halftoning is to convert a high bit-depth image (0-255) into a low bit-depth image (0-1). Many image rendering schemes have been developed in the past, likely ordered dithering (OD) [2] [3], error diffusion (ED) [4] [5], dot diffusion (DD) [6] [7], and direct binary search (DBS) [8] [9].

Inverse halftoning is a reverse process of halftoning that reconstructs a continuous-tone image of grayscale image (0-255) from its halftone version (0-1). Besides image restoration, inverse halftoning is also useful for printed image processing, image compression, and image enhancement.

For example, printed photos from old books or magazines can be scanned and recovered into the original grayscale image, which is extremely important for historical photos restoration.

Moreover, continuous-tone images are required to reconstruct in many applications, such as image sharpening, size adjustment, image rotation, or even color correction to o favorable results. Nevertheless, inverse halftoning is an ill-posed problem since there would be numerous feasible solutions.

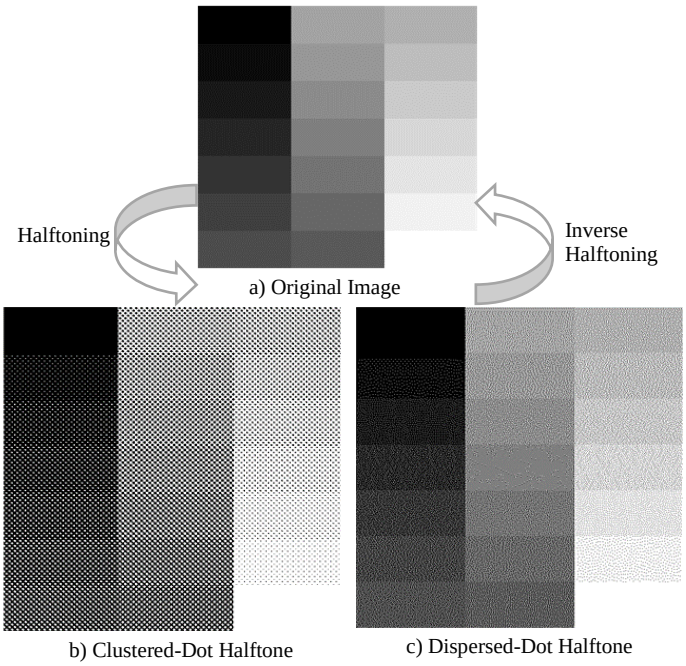


Fig. 1. Types of halftones

In this paper, we take the advantages of deep Convolutional Neural Network (CNN), which has been widely used for many computer vision problem [10] [11]. The training based on CNN can learn corresponding pairs of continuous-tone and halftone images efficiently. In general, the proposed scheme includes two steps: first removing the halftone effects to obtain the raw approximation of the image and then enhance the fine details using progressive learning. This is an end-to-end mapping solution to recover not only the gray-scale version but also the color halftone image. The experimental results show that the proposed deep learning approach is superior to the inverse halftoning benchmarks and achieves a new state-of-the-art performance.

We also show that statistical analysis on the dataset is an important factor to optimize the deep learning task. The halftone patterns are classified according to their variance values and then is used to train the model sets. We consider variance as a major factor because it has a close relationship with edge reconstruction.

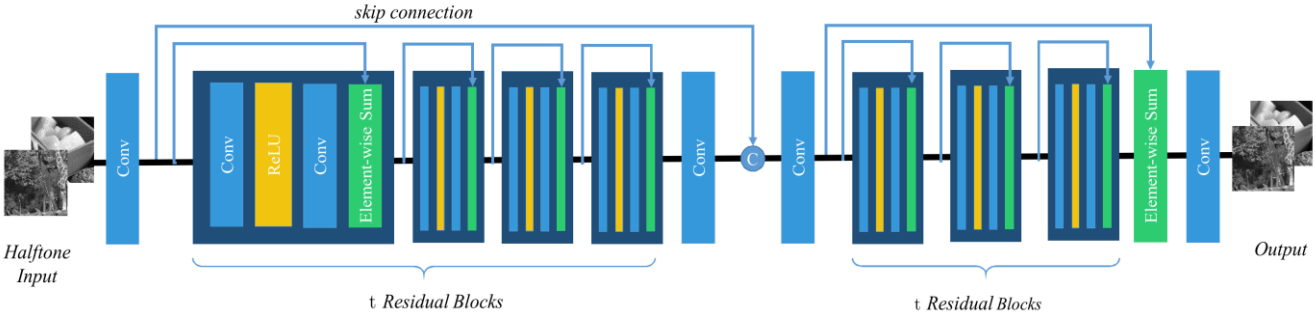


Fig. 2. Proposed deep neural network architechture

Section II briefly introduces different inverse halftoning approaches including conventional methods and new deep learning techniques. Section III explains the proposed method and the highlights of our proposed architecture. Section IV focuses on the experimental results and comparison to conventional techniques and deep learning-based methods as well.

II. PROPOSED WORK

A. *Conventional Inverse Halftoning*

The simplest approach to perform inverse halftoning is to use a simple gaussian low-pass kernel that can remove the halftone patterns. The approximation results in elimination of high frequency information and textures, and hence the resultant image is often blurry and shows poor reconstruction result. Chang et. al. [12] proposed a method based on the least-mean-square adaptive filtering algorithm which attempts to estimate the optimal mask shapes and filter weights. The method achieves better image quality with reduced image artifacts. Further, Mese et. al. [13] utilize a look up table (LUT) strategy to which approximates the halftone image to its closest gray scale equivalent image. The reconstruction performance completely relies on LUT construction and has severe performance limitations. Chung-Wu et. al [14] proposed an improved version of the Mese [13] method incorporating the classified edge features to achieve a sharper edge which slightly improve the visual quality. The main problem associated with the LUT methods is it memory requirement and time consumption to search and obtain the optimal gray-scale image. Guo. et. al [15] solved the memory issue using the LUT based on the multi-classifier optimized inverse halftone approach. The approach also uses the edge-based LUT which results in better reconstruction, but it is trained on a set of 30 image which resulted in limited performance.

B. *Deep Learning-Based Methods*

Deep convolutional neural networks have become the state-of-the-art method for numerous computer visions tasks including image segmentation [16], image classification [11], image denoising [17], super resolution [18], image restoration [19], and image-to-image transformation [20]. Especially, the inverse halftoning task is almost identical to the image-to-image transformation issue that transforms a halftone into the desired continuous-tone image. Qiu et. al [21] successfully deployed the deep learning-based solution to inverse halftoning. The method trained a deep convolutional neural network inspired by U-Net architecture [22] and leveraged another pre-trained deep CNN as a feature extractor. The U-Net network is similar to encoder-decoder that primarily used in biomedical semantic segmentation. In latest, Xia et. al [27] also proposed an encoder-decoder approach similar to U-Net architechture containing up- and down-sampling. Along with this progressive learning approach is also adopted for inverse halftoning. However, a lot of important features are lost due to the spatial dimension reduction of feature maps to obtain the latent representation. Therefore, edges and textures information cannot be recovered very accurately.

In contrast to this, the proposed model is straightforward method that doesn't affect the spatial dimension of inputs. Furthermore, our network is deeper, we immensely investigate the efficient residual blocks in the network design. To summarize, a novel residual neural network inspired by the progressive learning concept to enhance the subtle information such as textures, edges, and shape of objects is proposed. Our deep CNN is considered as a nonlinear mapping function that approximate the relationship between halftone images and their continuous-tone version. The model architecture and loss function are presented in section A and B respectively.

C. *Model Architectures*

At present, the existing CNN architectures are found to be very effective in solving many image processing tasks. Among different CNN models, the proposed model is constructed based on the combination of VGG16 [10] and SR-ResNet [18] model. As the primary VGG model is proposed for the image classification problem, it has been modified optimally to perform for the inverse halftoning problem. As shown in Fig. 2, the proposed network comprises of two stages: the first part of DRCNN starts with a normal convolutional layer and followed by t residual blocks. From our experiment results, it has been verified that at $t = 18$, our model has best performance. Besides, the concatenation is also provided to extract the contextual semantic information to the network. The first stage extracts a variety of features and simply remove the halftone patterns.

After the first step, essential information such as shapes, colors, tones, and contextual information are extracted. In the next stage, consecutive *t* residual blocks are integrated network to extract the subtle information by learning a residual image in order to restore the fine details similar to the concept of SR-ResNet [18]. The addition of *t* residual blocks into the first part has two benefits. Firstly, increasing the number of residual blocks can boost the performance of the model significantly as reported in [23]. Secondly, the concept of progressive learning is applied to synthesize subtle details by the leveraging prior knowledge from the first stage.

The proposed progressive framework gradually exploits the intrinsic features from the first part of the network and those are then fed into the next part to synthesize subtle structures incrementally. This process is progressive so that that numerous essential information extracted from the first part network can have an impact on later enhancement process. In such a progressive manner, our approach utilizes the features learned in the first progressive step to benefit the detail enhancement stage.

D. Loss Functions

Our goal is to reconstruct the continuous-tone images I_r from a given halftone images I_h that looks like the ground truth image I_{gt} with good perceptual quality. The halftone image of size $W \times H \times C$ can be transformed by the residual neural network into the continuous-tone image of size $W \times H \times C$ where W, H, and C are the width, height, and the number of channels of the image respectively. The proposed network can learn directly features from halftone without downsizing spatial dimension in U-Net architecture.

To recover I_{gt} from I_h , the network firstly trained to reduce the pixel-wise difference between I_r and I_{gt}. Pixel-wise loss reproduces I_{gt} in terms of PSNR. The $\mathcal{L}_1$ loss is a pixel-based reconstruction loss function which is widely used for many optimization problems. Optimizing $\mathcal{L}_1$ loss is the cheapest way to offer high PSNR performance. However, in the absence of high frequency information, images are usually blurry and noisy. We employ the $\mathcal{L}_1$ by computing the pixel-wise value between the ground-truth image I^{gt} and reconstructed image I^r.

$$\mathcal{L}_1 = \frac{1}{W \times H \times C} \sum_{i}^{W} \sum_{j}^{H} \sum_{k}^{C} \left| I^r_{i,j,k} - I^{gt}_{i,j,k} \right| \qquad (1)$$

Furthermore, we adopt the ideas of Gats et al. [24], Bruna et al. [25] and Johnson et al. [26] to integrate the VGG based perceptual loss $\mathcal{L}_{VGG}$, that is defined as

$$\mathcal{L}_{VGG/i,j} = \frac{1}{W_{i,j} \times H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{gt})_{x,y} - \phi_{i,j}(I^r)_{x,y})^2 \qquad (2)$$

This loss encourages the network to generate a realistic result and enhance visual quality by optimizing the difference between high-level features. In Eq. 2, the $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the corresponding feature maps. $\mathcal{L}_{VGG}$ loss measures the Euclidean distance between the feature representations of a ground truth image I_{gt} and a recovered image I_r. Moreover, $\phi_{i,j}$ indicates the feature map obtained by the j^{th} convolution (after activation) before the i^{th} max-pooling layer. Similar to [27], we choose $\phi_{i,j} = \phi_{4,4}$ that means the feature map obtained by the fourth convolution before the fourth max-pooling layer within the VGG19 network.

Structural similarity index (SSIM) is one of the popular full reference quality indexes which evaluates image quality based on the structural similarity. It can be used as an assessment of the perceptual similarity between two images with respect to a human viewer. We adopt SSIM to be a loss function that helps network learn the local structures in the image. μ and σ are means and standard deviations that are computed with a Gaussian filter kernel size of 7x7. According to [28], the SSIM loss function can be written by:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \qquad (3)$$

To sum up, the final loss of our network is the weighted sum of pixel-wise $\mathcal{L}_1$, perceptual loss, $\mathcal{L}_{VGG}$ and $\mathcal{L}_{SSIM}$:

$$\mathcal{L}_{total} = \alpha_1 \times \mathcal{L}_1 + \alpha_2 \times \mathcal{L}_{VGG} + \alpha_3 \times \mathcal{L}_{SSIM} \qquad (4)$$

Here $\alpha_1 = 1, \alpha_2 = 4 \times 10^{-4}$ and $\alpha_3 = 1.5$

The α_2 is set in reference with the paper [27]. The $\mathcal{L}_{VGG/i,j}$ is calculated based on the difference between two feature maps of I^{gt} and I^r from pre-trained VGG19 network on another dataset. Therefore, the $\mathcal{L}_{VGG}$ from our experiments are extremely large. The weight α_2 will scale down the impact of $\mathcal{L}_{VGG}$ since $\mathcal{L}_1$ and $\mathcal{L}_{SSIM}$ are our objective for optimization. The final value α_2 is 4×10^{-4}. $\mathcal{L}_1$ and $\mathcal{L}_{SSIM}$ have correlated weights and are chosen based on practical experiments.

In this study, we select three popular datasets including MSCOCO [29], VOC2012 [30] and Place365 [31]. Due to a large number of images in each dataset, we select the small portion of images to do the variance analysis. MSCOCO [29] subset contains 20,000 image while VOC2012 [30] and Place365 [31] subset consist of 17,000 and 36,500 images respectively. For each subset, we randomly select 5 million patches with size 33x33 and attempt to sketch the histogram of variances for all subsets as in Fig. 3.

III. RESULTS AND DISCUSSION

In this part, the proposed DRCNN is applied for quantitative and qualitative performance evaluation. Firstly, we compare our proposed DRCNN with many state of the art methods and the results are provided in Fig. 4 and Table I. From the reconstruction quality assessment using SSIM and PSNR it can been seen that the proposed work has achieved a better results in terms of better edge reconstruction and achieved better perceptual quality than the existing works.



(a) Low Variance 0 – $2x10^3$ (b) Medium Variance $2x10^3$ – $4x10^3$ (c) High Variance Above $4x10^3$

Fig. 3. Statistical Analysis of halftone patches

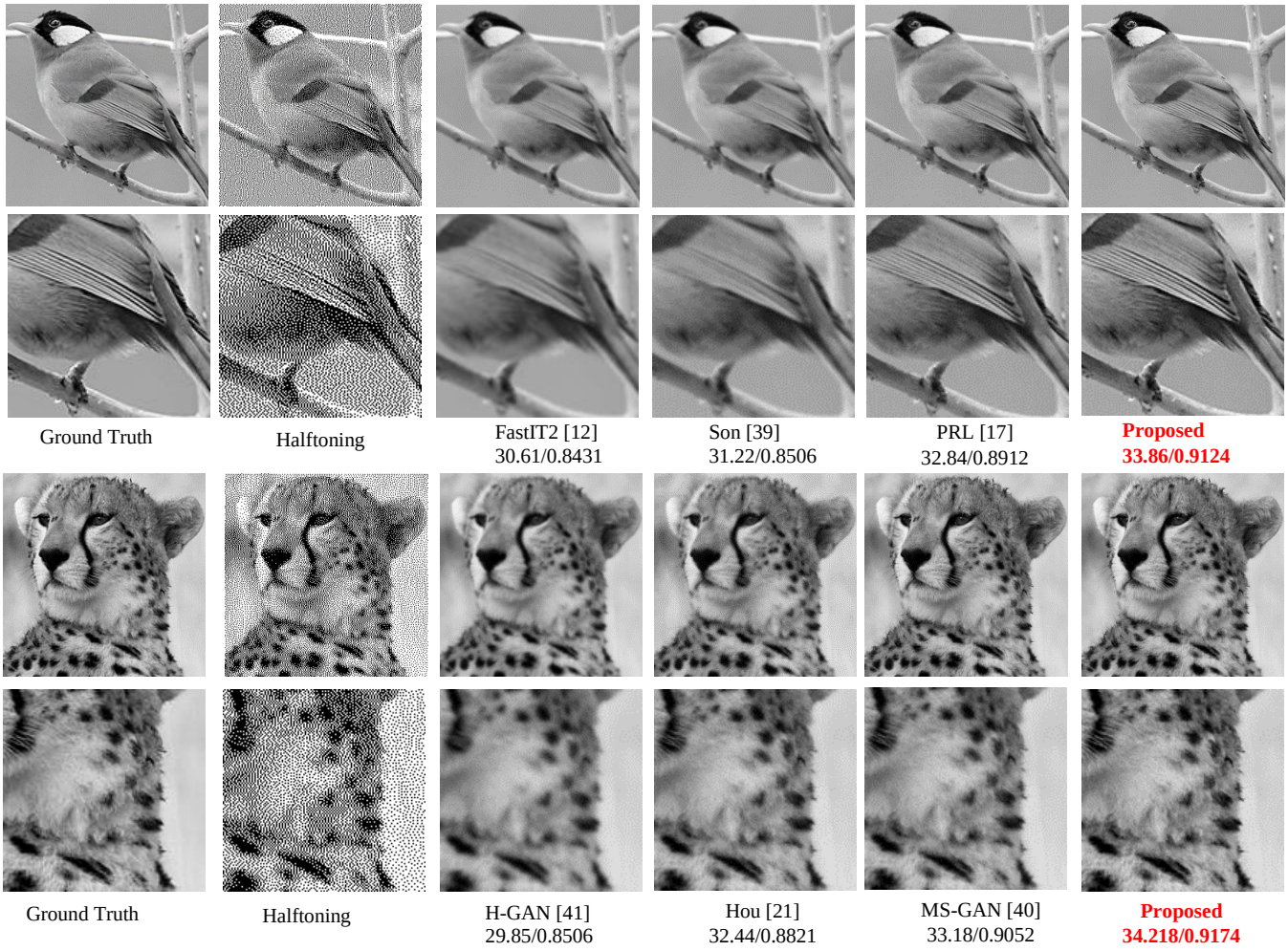


Fig. 4. Comparison results of the various halftoning techniques.

TABLE I. PERFORMANCE OF DIFFERENT LOSS FUNCTIONS ON RANDOM 100 IMAGES OF VOC2012 DATASET

VOC100	$\mathcal{L}_1$	$\mathcal{L}_{SSIM}$	$\mathcal{L}_{VGG}$	$\mathcal{L}_1 + \mathcal{L}_{SSIM}$	$\mathcal{L}_{SSIM} + \mathcal{L}_{VGG}$	$\mathcal{L}_1 + \mathcal{L}_{VGG}$	$\mathcal{L}_1 + \mathcal{L}_{SSIM} + \mathcal{L}_{VGG}$
PSNR	30.33	30.24	29.87	30.37	30.09	30.44	**30.47**
SSIM	0.8820	0.8835	0.8689	0.8826	0.8816	0.8767	**0.8787**

IV. CONCLUSION

In this paper, a deep residual and classified neural network (DRCNN) is proposed to perform inverse halftoning.

The proposed architecture is comprising of the CNN network integrated with residual learning blocks to perform the optimal reconstruction. For training, various databases such as VOC2012, Place365 and DHD are used, and this helps to

improve the reliability and robustness of the proposed scheme. From initial experiments, it has been inferred that the statistical properties of the halftones have a substantial influence over the reconstruction accuracy. Consequently, the proposed method is an integration of many sub-models each of which performs for the halftone pattern with specific variance. In addition, various loss function such as PSNR, SSIM, and VGG loss are integrated to accurately reconstruct the average and structural information effectively. From our comprehensive results, it has been concluded that the proposed architecture performs superior over the existing models in terms of image quality and achieves good generalization to images of various types.

REFERENCES

[1] R. Ulichney, "Digital halftoning and the physical reconstruction function," 2005.

[2] B. E. Bayer, "An optimum method for two-level rendition of continuous tone pictures," in *IEEE International Conference on Communications*, 1973, vol. 26.

[3] R. Ulichney, *Digital halftoning*. MIT press, 1987.

[4] R. W. Floyd and L. S. Steinberg, "An adaptive algorithm for spatial gray scale," 1975.

[5] J. Jarvis and C. Roberts, "A new technique for displaying continuous tone images on a bilevel display," *IEEE Transactions on Communications*, vol. 24, no. 8, pp. 891-898, 1976.

[6] R. A. Ulichney, "Dithering with blue noise," in *Proceedings of the IEEE*, 1988, vol. 76, no. 1, pp. 56-79.

[7] D. E. J. A. T. o. G. Knuth, "Digital halftones by dot diffusion," vol. 6, no. 4, pp. 245-273, 1987.

[8] D. J. Lieberman and J. P. Allebach, "Efficient model based halftoning using direct binary search," in *Proceedings of International Conference on Image Processing*, 1997, vol. 1, pp. 775-778: IEEE.

[9] Q. Lin and J. P. Allebach, "Color FM screen design using DBS algorithm," in *Color Imaging: Device-Independent Color, Color Hardcopy, and Graphic Arts III*, 1998, vol. 3300, pp. 353-361: International Society for Optics and Photonics.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," presented at the ICLR, 2015.

[11] I. Sutskever, G. E. Hinton, and A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," presented at the Advances in neural information processing systems, 2012.

[12] P.-C. Chang and C.-S. Yu, "Neural net classification and LMS reconstruction to halftone images," in *Visual Communications and Image Processing'98*, 1998, vol. 3309, pp. 592-602: International Society for Optics and Photonics.

[13] M. Mese and P. P. Vaidyanathan, "Look-up table (LUT) method for inverse halftoning," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1566-1578, 2001.

[14] K.-L. Chung and S.-T. Wu, "Inverse halftoning algorithm using edge-based lookup table approach," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1583-1589, 2005.

[15] J.-M. Guo, Y.-F. Liu, J.-H. Chen, and J.-D. Lee, "Inverse Halftoning With Context Driven Prediction," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1923-1924, 2014.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[17] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736-3745, 2006.

[18] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.

[19] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2774-2781.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125-1134.

[21] X. Hou and G. Qiu, "Image companding and inverse halftoning using deep convolutional neural networks," 2017.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241: Springer.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[24] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in neural information processing systems*, 2015, pp. 262-270.

[25] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," 2015.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016, pp. 694-711: Springer.

[27] M. Xia and T.-T. Wong, "Deep Inverse Halftoning via Progressively Residual Learning," in *Asian Conference on Computer Vision*, 2018, pp. 523-539: Springer.

[28] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47-57, 2016.

[29] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740-755: Springer.

[30] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," presented at the International journal of computer vision, 2012.

[31] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. J. I. t. o. p. a. Torralba, and m. intelligence, "Places: A 10 million image database for scene recognition," vol. 40, no. 6, pp. 1452-1464, 2017.

[32] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136-144.

[33] J. M. Guo and S. Sankarasrinivasan, "Digital Halftone Database (DHD): A Comprehensive Analysis on Halftone Types," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1091-1099: IEEE.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[35] T. D. Kite, N. Damera-Venkata, B. L. Evans, and A. C. Bovik, "A fast, high-quality inverse halftoning algorithm for error diffused halftones," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1583-1592, 2000.

[36] R. Neelamani, R. D. Nowak, and R. G. J. I. T. o. I. P. Baraniuk, "WInHD: Wavelet-based inverse halftoning via deconvolution," 2002.

[37] P.-C. Chang, C.-S. Yu, and T.-H. J. I. T. o. I. P. Lee, "Hybrid LMS-MMSE inverse halftoning technique," vol. 10, no. 1, pp. 95-103, 2001.

[38] Y.-F. Liu, J.-M. Guo, and J.-D. J. I. T. o. I. P. Lee, "Inverse halftoning based on the Bayesian theorem," vol. 20, no. 4, pp. 1077-1084, 2010.

[39] C.-H. Son and H. Choo, "Local learned dictionaries optimized to edge orientation for inverse halftoning," IEEE Transactions on Image Processing, vol. 23, no. 6, pp. 2542-2556, 2014.

[40] E. Zhang, et. al., "An inverse halftoning method for various types of halftone images based on multi-scale generative adversarial network," *Engineering Applications of Artificial Intelligence*, 2023.

[41] J. M. Guo and S. Sankarasrinivasan, "H-GAN: Deep learning model for halftoning and its reconstruction," *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1-2. IEEE, 2020.

DOA-Aware Audio-Visual Self-Supervised Learning for Sound Event Localization and Detection

Yoto Fujita*†, Yoshiaki Bando†, Keisuke Imoto‡†, Masaki Onishi†, and Kazuyoshi Yoshii*
*Graduate School of Informatics, Kyoto University, Japan, {fujita.yoto.26m@st.kyoto-u.ac.jp, yoshii@i.kyoto-u.ac.jp}
†National Institute of Advanced Industrial Science and Technology, Japan, {y.bando,onishi-masaki}@aist.go.jp
‡Faculty of Science and Engineering, Doshisha University, Japan, keisuke.imoto@ieee.org

***Abstract*—This paper describes sound event localization and detection (SELD) for spatial audio recordings captured by first-order ambisonics (FOA) microphones. In this task, one may train a deep neural network (DNN) using FOA data annotated with the classes and directions of arrival (DOAs) of sound events. However, the performance of this approach is severely bounded by the amount of annotated data. To overcome this limitation, we propose a novel method of pretraining the feature extraction part of the DNN in a self-supervised manner. We use spatial audio-visual recordings abundantly available as virtual reality contents. Assuming that sound objects are concurrently observed by the FOA microphones and the omni-directional camera, we jointly train audio and visual encoders with contrastive learning such that the audio and visual embeddings of the same recording and DOA are made close. A key feature of our method is that the DOA-wise audio embeddings are *jointly* extracted from the raw audio data, while the DOA-wise visual embeddings are *separately* extracted from the local visual crops centered on the corresponding DOA. This encourages the latent features of the audio encoder to represent both the classes and DOAs of sound events. The experiment using the DCASE2022 Task 3 dataset of 20 hours shows non-annotated audio-visual recordings of 100 hours reduced the error score of SELD from 36.4 pts to 34.9 pts.**

***Index Terms*—Sound event localization and detection, audio-visual contrastive learning, self-supervised learning**

I. Introduction

Sound event localization and detection (SELD) is a task that aims to estimate the activations, classes, and directions of arrival (DOAs) of sound events from multichannel audio recordings [1]. It is one of the foundations of computational intelligence for understanding acoustic environments. The current standard approach to this task is to train a deep neural network (DNN) in a supervised manner using pairs of audio recordings with ground-truth annotations [2]–[4]. In general, it is difficult to collect a sufficient amount of annotated audio data covering a wide range of acoustic environments.

A popular solution to this problem is data augmentation [5], [6]. Specifically, one can synthesize multichannel audio data by convolving source signals of various classes with room impulse responses (RIRs) that simulate various acoustic environments and DOA conditions. This approach is known to effectively improve the performance of SELD for many common sound events (e.g., music and speech). However, SELD for some complex sound events (e.g., wildlife sound) remains an open problem because it is difficult to isolate and capture their individual sound source signals.

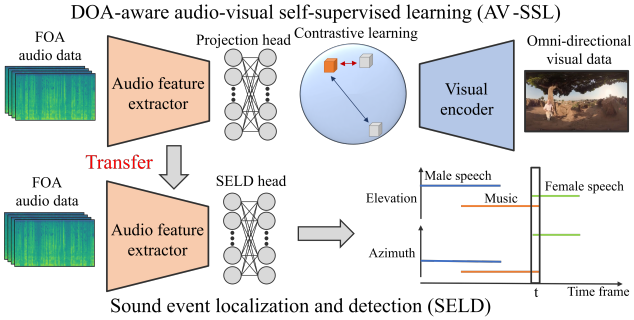


Fig. 1. AV-SSL of an audio feature extractor for SELD.

Another solution is to pretrain an audio feature extractor that constitutes the front end of a DNN used for SELD with audio-visual self-supervised learning (AV-SSL) (Fig. 1) [7]. This approach can make effective use of abundant virtual reality (VR) contents, each of which consists of multichannel audio data recorded by first-order ambisonics (FOA) microphones and 360° equirectangular visual data recorded by an omni-directional camera. Considering the cross-modal co-occurrence between the sounds and appearances of the same objects, one can jointly train audio and visual encoders in a contrastive fashion such that the audio and visual embeddings are made close to each other if they correspond to the same DOA and recording (positive sample) and far apart otherwise (negative sample). The front end of the trained audio encoder is then used for initializing the audio feature extractor for SELD.

Audio-visual spatial alignment (AVSA) [7] is one of the latest AV-SSL methods. It takes advantage of the FOA format, in which the single-channel audio signal with an arbitrary DOA can be computed from the observed FOA data. The audio embedding extracted from this enhanced signal are made close to the visual embedding extracted from the visual crop centered on the same DOA in the equirectangular visual data. However, the audio feature extractor trained in this way is insufficient for SELD because the DOA features of sound events cannot be extracted from the enhanced signal. On the other hand, the features useful for DOA estimation of sound events are not learnable with such non-DOA-aware contrastive learning equivalent in principle to AVC [8].

In this paper, we propose a DOA-aware extension of AVSA. Our method differs from the conventional AVSA [7] in that it jointly extracts audio embeddings over a DOA grid from raw

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

FOA audio data without DOA-wise signal enhancement. This encourages the latent features of the audio encoder to represent both the classes and DOAs of sound events. In addition, this paper also tackles one of the remaining challenges in AVSA that the cross-modal co-occurrence between sound and appearance does not hold for visible silent objects and occluded sound objects. In general, meaningful sound events exist only in a small number of DOAs, which limits this *local* contrastive learning based on the DOA-wise similarity between audio and visual embeddings.

To mitigate this problem, we also propose *global* contrastive learning based on recording-wise audio-visual similarity obtained by averaging the DOA-wise similarities over all the DOAs. Our goal is to maximize the similarity when the audio and visual embeddings correspond to the same recording (positive) and minimize it otherwise (negative). To encourage the audio encoder to extract DOA information as the latent features, we introduce a data augmentation technique that randomly spatially rotates only the equirectangular visual data to generate negative samples from the same recording.

II. Related Work

This section reviews existing SELD and AV-SSL methods.

A. Sound event localization and detection (SELD)

The modern approach to SELD is to train a DNN in a supervised manner. For instance, the activations of sound events are estimated with a DNN, while the DOAs are estimated geometrically [2]. Both the activations and DOAs can be estimated using a DNN [3]. Recently, an end-to-end approach to SELD has been proposed for directly estimating a DOA vector whose length corresponds to the duration of sound events in the Cartesian coordinate system [4].

To improve the robustness of SELD with a limited amount of annotated audio data, data augmentation tequniques such as SpecAugment [9] and Mixup [10] can be employed. Particularly for audio data provided in the FOA format, rotation-based augmentation is known to be effective [11]. Multichannel audio data with ground-truth annotations can be synthesized by convolving audio samples of various classes with arbitrary transfer functions [5], [12]. Alternatively, a general-purpose audio model [13] pretrained with a large dataset of audio signals called AudioSet [14] in an unsupervised manner can be used for sound event detection [15].

B. Audio-visual self-supervised learning (AV-SSL)

At the heart of AV-SSL is contrastive learning, which aims to train audio and visual encoders based on the cross-modal co-occurrence between the sounds and appearances of objects. The information encoded in the audio and visual embeddings may vary depending on the design of positive and negative samples. AVC [8], for example, uses standard video recordings and considers a pair of audio and visual data from the same recording as a positive sample and a pair of those from different recordings as a negative pair. The audio and visual embeddings are thus encouraged to represent the features of sound event classes. AVSA [7] uses spatial video recordings, which are originally made for VR application that requires DOA-based audio-visual rendering. AVSA, however, is essentially identical to AVC because it can be regarded as AVC for *non-spatial* video standard recordings obtained by decomposing *spatial* video recordings in a DOA-wise manner.

Audio-visual temporal synchronization (AVTS) [16] is another AVC-like self-supervised method. A key feature of AVTS is that it incorporates audio and visual data in the same recording, but at different moments in time, to negative pairs. The embeddings are thus encouraged to represent both the classes and activations of sound events.

III. Proposed Method

This section describes two variants of DOA-aware AV-SSL to improve SELD. The first variant employs DOA-wise contrastive learning, and the other employs recording-wise contrastive learning (Fig. 2). In both variants, an audio feature extractor $\mathcal{A}$ is used to transform raw FOA data into the latent audio features that represent sound event classes and DOAs. In this study, K discrete points on the Fibonacci lattice [17] are considered as potential DOAs. Each DOA $k \in [1, K]$ is defined by an azimuth angle θ_k and an elevation angle ϕ_k. A projection head $\mathcal{H}$ is then used to *jointly* convert the latent features to the audio embeddings over the DOA grid. A visual encoder $\mathcal{V}$, in contrast, is used to *separately* convert visual crops of a 360° equirectangular visual data over the DOA grid to the visual embeddings. Let I be the total number of recordings.

In DOA-wise contrastive learning, we maximize the local similarity between DOA-wise audio and visual embeddings when they correspond to the same DOA (positive sample), and minimize it otherwise (negative sample). In recording-wise contrastive learning, in contrast, we maximize global similarity obtained by averaging the local similarities over the DOA grid when the audio and visual embeddings correspond to the same recording (positive), and minimize it otherwise (negative).

Once $\mathcal{A}$, $\mathcal{H}$, and $\mathcal{V}$ are trained jointly in a self-supervised manner, $\mathcal{A}$ is connected to another head $\mathcal{H}'$ for SELD and the entire network is fine-tuned in a supervised manner.

A. Audio encoding

Let $\mathbf{X}_i^{(a)} \in \mathbb{R}^{(4+3)\times F\times T_a}$ be the multichannel audio spectrogram of spatial recording $i \in [1, I]$ obtained by stacking the mel spectrograms of the four-channel audio signals of the FOA format and the three intensity spectrograms on the orthogonal axes, where F is the number of mel frequency bins, and T_a is the number of frames. A series of frame-wise audio latent features denoted by $\mathbf{Y}_i^{(a)} \triangleq \{\mathbf{y}_{it}^{(a)}\}_{t=1}^{T_a}$ is obtained with the audio feature extractor $\mathcal{A}$ as follows:

$$\mathbf{Y}_i^{(a)} \leftarrow \mathcal{A}\left(\mathbf{X}_i^{(a)}\right). \tag{1}$$

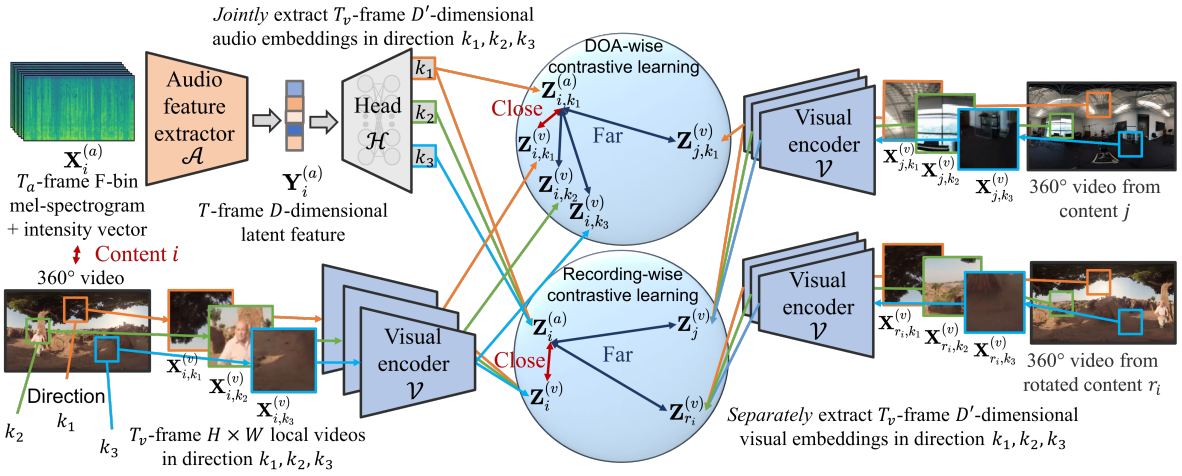


Fig. 2. The proposed DOA-wise and recording-wise DOA-aware contrastive learnings for pretraining the audio feature extractor.

A series of DOA- and frame-wise audio embeddings $\mathbf{Z}_i^{(a)} \triangleq \{\mathbf{z}_{ikt}^{(a)}\}_{k=1,t=1}^{K,T_v}$ is then obtained with the projection head $\mathcal{H}$:

$$\{\mathbf{z}_{ikt}^{(a)}\}_{k=1}^{K} \leftarrow \mathcal{H}\left(\mathbf{y}_{it}^{(a)}\right), \tag{2}$$

where the features $\mathbf{y}_{it}^{(a)}$ at each frame t are independently transformed for temporally localizing information, followed by an adaptive average pooling to match the visual data of T_v frames.

B. Visual encoding

Let $\mathbf{X}_i^{(v)} \triangleq \{\mathbf{x}_{ikt}^{(v)}\}_{k=1,t=1}^{K,T_v}$ be the series of DOA-wise local images cropped from the 360° equirectangular visual data of recording $i \in [1, I]$, where $\mathbf{x}_{ikt}^{(v)} \in \mathbb{R}^{H \times W}$ is a local image that corresponds to DOA $k \in [1, K]$ at time $t \in [1, T]$, H and W are the height and width of each image (cropping size), and T_v is the number of frames. Note that $\mathbf{x}_{ikt}^{(v)}$ is centered on DOA k on the Gnomonic projection [18] of the original spatial visual data. A series of DOA- and frame-wise visual embeddings $\mathbf{Z}_i^{(v)} \triangleq \{\mathbf{z}_{ikt}^{(v)}\}_{k=1,t=1}^{K,T_v}$ is obtained with the visual encoder $\mathcal{V}$ as follows:

$$\{\mathbf{z}_{ikt}^{(v)}\}_{t=1}^{T_v} \leftarrow \mathcal{V}\left(\{\mathbf{x}_{ikt}^{(v)}\}_{t=1}^{T_v}\right), \tag{3}$$

where the same visual encoding is independently applied to each DOA k to obtain the embeddings that represent the classes of visible sound objects.

C. Self-supervised learning (pretraining)

We describe the two variants of contrastive learning.

1) Similarity measures: Since the audio and visual data of the same spatial recording usually have the DOA-wise correspondence, we define the similarity between recordings i and j for each DOA k as the cosine similarity as follows:

$$\mathcal{S}_{\mathrm{DOA}}(\mathbf{Z}_{ik}^{(a)}, \mathbf{Z}_{jk}^{(v)}) = \frac{1}{T_v} \sum_{t=1}^{T_v} \frac{\mathbf{z}_{ikt}^{(a)\mathsf{T}} \mathbf{z}_{jkt}^{(v)}}{\|\mathbf{z}_{ikt}^{(a)}\| \|\mathbf{z}_{jkt}^{(v)}\|}, \tag{4}$$

where $\mathbf{Z}_{ik}^{(a)} \triangleq \{\mathbf{z}_{ikt}^{(a)}\}_{t=1}^{T_v}$ and $\mathbf{Z}_{ik}^{(v)} \triangleq \{\mathbf{z}_{ikt}^{(v)}\}_{t=1}^{T_v}$.

We here focus on the InfoNCE loss [19] defined as follows:

$$\mathcal{I}(\mathbf{Z}, \mathbf{Z}_p, \mathbf{U}_n) = -\frac{\exp(\mathcal{S}(\mathbf{Z}, \mathbf{Z}_p)/\tau)}{\sum_{\mathbf{Z}_n \in \mathbf{U}_n} \exp(\mathcal{S}(\mathbf{Z}, \mathbf{Z}_n)/\tau)}, \tag{5}$$

where $\mathbf{Z}$ is an anchor, $\mathbf{Z}_p$ is a positive sample, $\mathbf{U}_n$ is a set of negative samples, and τ is a temperature hyperparameter. Minimizing $\mathcal{I}$ encourages the similarity of $\mathbf{Z}$ to $\mathbf{Z}_p$ to be larger than to $\mathbf{U}_n$ in a contrastive manner.

2) DOA-wise contrastive learning: The loss $\mathcal{L}_{\mathrm{DOA}}$ used in this variant is calculated as follows:

$$\mathcal{L}_{\mathrm{DOA}} = \sum_{i,k=1}^{I,K} \mathcal{I}\left(\mathbf{Z}_{ik}^{(a)}, \mathbf{Z}_{ik}^{(v)}, \{\mathbf{Z}_{jk}^{(v)}\}_{j=1}^{I} \cup \{\mathbf{Z}_{ik'}^{(v)}\}_{k'=1}^{K}\right), \tag{6}$$

where the similarity between the audio and visual embeddings is maximized when they correspond to the same DOA.

3) Recording-wise contrastive learning: The loss $\mathcal{L}_{\mathrm{REC}}$ used in this variant is calculated as follows:

$$\mathcal{L}_{\mathrm{REC}} = \sum_{i=1}^{I} \mathcal{I}\left(\mathbf{Z}_i^{(a)}, \mathbf{Z}_i^{(v)}, \{\mathbf{Z}_j^{(v)}\}_{j=1}^{I} \cup \{\mathbf{Z}_{r_i}^{(v)}\}\right), \tag{7}$$

where $\mathbf{Z}_{r_i}^{(v)}$ is a series of visual embeddings that corresponds to the visual data from the rotated recording r_i, which is obtained by rotating the spatial information of the recording i. The recording-wise similarity used in $\mathcal{L}_{\mathrm{REC}}$ is obtained by averaging the DOA-wise similarities over all the directions:

$$\mathcal{S}_{\mathrm{REC}}(\mathbf{Z}_i^{(a)}, \mathbf{Z}_j^{(v)}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{S}_{\mathrm{DOA}}(\mathbf{Z}_{ik}^{(a)}, \mathbf{Z}_{jk}^{(v)}). \tag{8}$$

D. Supervised learning (fine-tuning)

Using annotated data, the audio feature extractor $\mathcal{A}$ is fine-tuned for SELD based on activity-coupled Cartesian DOA representation (ACCDOA) [20]. Specifically, multiple sound

events of the same class can be separately assigned to different tracks. Let $\hat{\mathbf{P}} \in \mathbb{R}^{T' \times C \times N \times 3}$ be a multi-ACCDOA vector of T' frames, C classes, N tracks given by

$$\hat{\mathbf{P}} \leftarrow \mathcal{H}'\left(\mathcal{A}\left(\mathbf{X}\right)\right), \tag{9}$$

where $\mathbf{X} \in \mathbb{R}^{(4+3)\times F \times T'_a}$ is the multichannel spectrogram computed in the same way as $\mathbf{X}_i^{(a)}$, where T_a is the number of frames. The projection head $\mathcal{H}'$ consists of several fully-connected layers, followed by an adaptive average pooling to suit the target time resolution T'. Since the SELD label for frame t, class c, and track n is given as a pair of the activity $a^*_{tcn} \in \{0, 1\}$ and the Cartesian DOA vector $\mathbf{R}_{tcn}$, the target ACCDOA vector is given by

$$\mathbf{P}^*_{tcn} = a^*_{tcn}\mathbf{R}_{tcn}. \tag{10}$$

The overall network is trained to minimize:

$$\mathcal{L}^{\mathrm{PIT}} = \frac{1}{TC}\sum_{t,c=1}^{T',C} \min_{\alpha \in \mathrm{Perm}(ct)} \mathcal{L}^{\mathrm{ACCDOA}}_{\alpha,tc}, \tag{11}$$

$$\mathcal{L}^{\mathrm{ACCDOA}}_{\alpha,tc} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{MSE}\left(\mathbf{P}^*_{\alpha,tcn}, \hat{\mathbf{P}}_{tcn}\right), \tag{12}$$

where $\alpha \in \mathrm{Perm}(t)$ is a possible frame-level permutation of M tracks at the frame t and $\mathrm{Perm}(t)$ is a set of all possible permutations. $\mathrm{MSE}(\cdot, \cdot)$ is the mean square error function.

IV. Evaluation

This section reports a comparative experiment conducted for evaluating the proposed AV-SSL methods with the two variants of contrastive learning.

A. Network configuration

The STFT spectrograms with a shifting interval of 480 samples and a window size of 1024 were converted to the mel spectrograms with $F = 64$ mel bins. The number of DOAs K was 220.

We used a ResNet-Conformer [12] for the audio feature extractor $\mathcal{A}$ as shown in Figs. 3 (a)–(c). The architecture of the projection head $\mathcal{H}$ consisted of two linear+SiLU layers followed a linear layer which output dimension was 220×128 as shown in Figs. 3 (d). The FOA audio data were resampled at a sampling rate of 24 kHz and split into 2-second clips in the training.

The visual encoder $\mathcal{V}$ consisted of 9-layer $R(2+1)D$ convolution layers [21] as shown in Fig. 4. The input equirectangular visual data were extracted at the time resolution T_v of 16 (8 Hz). The field of view ψ of visual crops for each direction was $40°$, and the resolution $H \times W$ was 16×16. The horizontal and vertical flipping was applied to the visual crops as data augmentation.

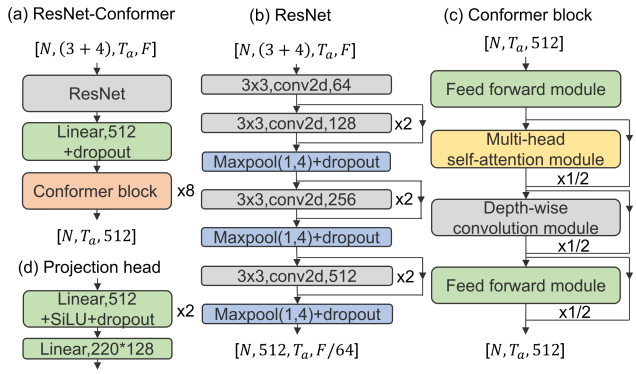


Fig. 3. The architecture of the audio encoder.

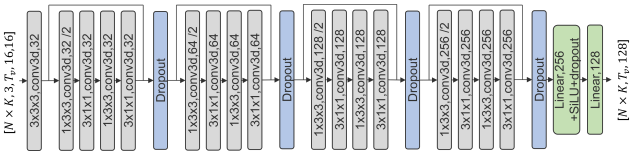


Fig. 4. The architecture of the visual encoder.

B. Pretraining

We used the YT-360 dataset [7] for pretraining. It contains VR contents collected from YouTube, each of which consists of a synchronized pair of FOA audio data and equirectangular visual data, including 246 hours of 10-second-long recordings on diverse genres including music and sports. After removing recordings with some missing channels, 104 hours of the training data and 20 hours of the validation data were used.

By the proposed AV-SSL methods, the ResNet-Conformer was trained for 100 epochs using AdamW optimizer [22] with a batch size of 4, a learning rate of 10^{-4}, a weight decay of 10^{-5}. Rotated recordings for the negative sample were obtained by randomly rotating the equirectangular visual data around the z-axis. The temperature hyper-parameter τ was 0.1. The SpecAugment [9] and the dropout with a rate of 0.1 were used. The model having the lowest validation loss was used for the downstream SELD task.

The conventional AV-SSL method called AVC [8] was also evaluated as a baseline for comparison. Specifically, the loss $\mathcal{L}_{\mathrm{AVC}}$ was calculated with the recording-wise similarity $\mathcal{H}(\{\mathbf{z}^{(a)}_{it}\}_{t=1}^{T_v}, \{\mathbf{z}^{(v)}_{it}\}_{t=1}^{T_v})$, where $\mathbf{z}^{(a)}_{it} \in \mathbb{R}^{128}$ was obtained by max-pooling the DOA-wise visual embeddings $\{\mathbf{z}^{(v)}_{ikt}\}_{k=1}^{K}$ and $\mathbf{z}^{(a)}_{it} \in \mathbb{R}^{128}$ was directly obtained from the output layer $\mathcal{H}$ with output dimension 128.

As in [7], curriculum learning was introduced for the proposed methods, where two audio feature extractors $\mathcal{A}$ were first trained with AVC for 50 epochs, then trained with the two different proposed AV-SSLs for 50 epochs.

C. Fine-tuning

The SELD performance was evaluated on the STARSS22 and Synth1 datasets [5]. The STARSS22 is a dataset of actual recorded FOA data annotated with the activations, classes, and

TABLE I
CLASS-WISE ACTIVITIY IN THE STARSS22 DATASET [5].

	Fem. speech	Male speech	Clap	Phone	Laugh	Dom. sounds	Footsteps	Door	Music	Music. instr.	Faucet	Bell	Knock
Frame coverage (% total frames)	20.4	37.6	0.7	1.4	2.7	17.9	1.3	0.6	29.4	4.0	1.7	1.5	0.1

TABLE II
EVALUATION RESULT FOR THE TWO DATASETS.

Fine-tuning dataset	Pretraining method	$ER_{\leq 20^\circ} \downarrow$	$F_{\leq 20^\circ} \uparrow$	$LE \downarrow$	$LR \uparrow$	$SELD \downarrow$
STARSS22+Synth	None	0.53	48.9 %	18.2°	68.7 %	0.364
	AVC[8]	0.52	49.7 %	17.9°	69.0 %	0.359
	AV-SSL with DOA-wise contrastive learning	**0.51**	50.5 %	17.9°	**70.1 %**	0.351
	AV-SSL with recording-wise contrastive learning	**0.51**	**51.6 %**	**17.1**°	69.0 %	**0.349**
STARSS22	None	0.65	**37.9 %**	**22.1**°	**58.4 %**	**0.452**
	AVC[8]	**0.63**	37.7 %	22.4°	55.2 %	0.458
	AV-SSL with DOA-wise contrastive learning	0.68	35.8 %	22.9°	53.7 %	0.478
	AV-SSL with recording-wise contrastive learning	0.67	36.3 %	23.2°	56.3 %	0.467

TABLE III
CLASS-WISE INCREASE OR DECREASE OF THE THREE SELD METRICS BY EACH PRETRAINING.

Fine-tuning dataset	Pretraining method	Δ Metrics	Dom. sounds	Door	Music	Music. instr.	Bell	Knock
STARSS22+Synth	AVC [8]	$\Delta F_{\leq 20^\circ}$	+0.03	−0.02	−0.02	+0.02	+0.04	+0.01
		$\Delta(1-\bar{LE}/180)$	+0.01	+0.01	−0.01	0.0	+0.00	−0.0
		ΔLR	+0.02	−0.01	+0.05	+0.01	+0.08	−0.0
	DOA-wise	$\Delta F_{\leq 20^\circ}$	+0.05	+0.08	+0.02	−0.12	+0.11	−0.1
		$\Delta(1-\bar{LE}/180)$	+0.0	+0.0	−0.01	−0.03	+0.02	−0.0
		ΔLR	+0.07	+0.07	+0.09	−0.02	+0.01	−0.05
	recording-wise	$\Delta F_{\leq 20^\circ}$	+0.1	+0.06	−0.04	−0.01	+0.12	−0.03
		$\Delta(1-\bar{LE}/180)$	+0.01	+0.01	−0.02	−0.0	+0.02	−0.0
		ΔLR	+0.04	−0.01	+0.02	−0.0	+0.06	−0.0
STARSS22	AVC [8]	$\Delta F_{\leq 20^\circ}$	−0.02	+0.09	+0.0	−0.05	−0.26	−0.18
		$\Delta(1-\bar{LE}/180)$	−0.01	+0.01	+0.0	−0.01	−0.09	+0.01
		ΔLR	+0.07	+0.07	−0.05	+0.06	−0.17	−0.38
	DOA-wise	$\Delta F_{\leq 20^\circ}$	−0.0	+0.04	+0.02	−0.14	−0.2	−0.07
		$\Delta(1-\bar{LE}/180)$	−0.0	+0.0	−0.01	−0.05	−0.06	+0.02
		ΔLR	+0.06	+0.07	−0.03	+0.06	−0.1	−0.31
	recording-wise	$\Delta F_{\leq 20^\circ}$	+0.03	+0.0	−0.0	−0.1	−0.16	−0.2
		$\Delta(1-\bar{LE}/180)$	−0.0	+0.02	+0.02	−0.02	−0.05	−0.0
		ΔLR	+0.05	+0.14	−0.14	−0.06	−0.09	−0.46

DOAs of 13 sound events every 0.1 seconds, where the class labels follow the Audioset ontology [14]. This dataset consists of 2.9 hours of training data and 2 hours of validation data. Note that some classes have very low frame coverages in STARSS22 as shown in Table I. The Synth1 is a dataset of synthesized FOAs annotated in the same way as STARSS22, obtained by remixing source signals from FSD50K [23] with room RIRs from TAU-SRIR database [24]. Two datasets of different sizes constructed from these datasets were used for fine-tuning. The first one was a dataset made up of all training data from STARSS22 and Synth1 (STARSS22+Synth1). The second one consisted only of all training data from STARSS22 (STARSS22). A dataset that consists of all validation data from STARSS22 and Synth1 was used as the validation dataset in both conditions. The original FOA data were split every 5 seconds, and they were sampled uniformly over all 13 classes to deal with the class imbalance in STARSS22.

Each pretrained model was fine-tuned with a head $\mathcal{H}'$ to output 3-track multi-ACCDOAs [20]. The architecture of $\mathcal{H}'$ was the same as the head $\mathcal{H}$ shown in the Sec. IV-A except that the output dimension was changed to $13 \times 3 \times 3$. AdamW with a learning rate of 10^{-4} and a weight decay of 10^{-5} was used for training. The models were trained for 1000 epochs. For data augmentation, the input FOA was randomly rotated. The dropout rate was 0.1 for ResNet-Conformer and 0.05 for $\mathcal{H}'$.

Models pretrained with our methods were compared with a non-pretrained model and a model pretrained with the conventional AVC method based on SELD score [1] on the vali-

dation dataset. The SELD score is obtained by averaging over four metrics as follows:

$$SELD = \frac{1}{4}\left(ER_{\leq 20^\circ} + (1 - F_{\leq 20^\circ}) + \frac{LE}{180} + (1 - LR)\right),$$

where the smaller score indicates better performance. $ER_{\leq 20^\circ}$ is the location-dependent error rate, where the prediction is counted as a true positive only if the estimated class activity is correct and the distance between the estimated and reference DOAs is smaller than 20°. $F_{\leq 20^\circ}$, LE and LR are the averages of class-wise metrics $F_{c,\leq 20^\circ}$, LE_c and LR_c, respectively. $F_{c,\leq 20^\circ}$ is the location-dependent F1-score for class c. LE_c is the average DOA estimation error over only the correctly detected events of class c. LR_c is the location-independent recall for class c. LE_c and LR_c can be considered to be the performance of detection and localization, respectively. The performance of each model was measured by averaging the metrics over the epochs with the first to tenth highest SELD scores.

D. Experimental results

Table II shows the overall SELD performances in terms of the five metrics, i.e., $ER_{\leq 20^\circ}$, $F_{\leq 20^\circ}$, LE, LR, and SELD score for each dataset and each pretraining method. In the STARSS22+Synth1 dataset, both the AV-SSL with DOA-wise contrastive learning and the AV-SSL with recording-wise contrastive learning improved the SELD scores by about 1.5 pts, while the conventional AVC improved only about 0.5 pts. This result indicates that our proposed approach is more suitable for the pretraining of SELD than AVC.

Table III shows the performance gaps in the class-wise metrics, $F_{c,\leq 20^\circ}$, LE_c and LR_c, between the non-pretrained model and each pretrained model, where all the metrics were transformed to the range of $[0, 1]$ for readability, and only some of the classes required for the following discussion were picked up due to the page limitation. The proposed AV-SSLs degraded the SELD scores more than AVC when only a few hours of imbalanced data were used for training (STARSS22). This would be partly due to the domain mismatch between YT-360 and STARSS22 and the class imbalance of STARSS22. As for the knock class, which was not included in the pretraining dataset (YT-360), for example, the detection performance LR_c got particularly worse. While in the classes related to home sounds (e.g., domestic sounds and door), which frequently appear in YT-360, the detection performance was improved.

Another reason for the degradation would be that a large amount of background music data independent from the visual data are included in YT-360. Background music was considered to have a negative impact on the localization of sound events related to music because it did not spatially correspond to the paired visual data. In fact, the localization performance LE_c was significantly degraded by the pretraining for the musical instruments and bell classes.

V. Conclusion

We proposed two variants of pretraining an audio feature extractor useful for SELD using spatial audio-visual recordings. To obtain the latent audio features representing not only the classes but also the DOAs of sound events, the audio encoder takes the FOA data as input, and outputs the audio embeddings over the DOA grid. The transfer learning with a sufficient amount of data showed the effectiveness of the proposed AV-SSLs as pretraining for SELD. For future work, one should deal with the deterioration in the SELD performance when sufficient labeled data is unavailable. One of the promising approaches would be to prepare the dataset of spatial audio-visual recordings covering various domains with good correspondence between audio and visual data.

.

Acknowledgment

This work was supported in part by NEDO, JSPS KAKENHI Nos. 20H00602, 19H04137, and ISHIZUE 2023 of Kyoto University.

References

[1] A. Politis *et al.*, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM TASLP*, vol. 29, pp. 684–698, 2020.

[2] T. N. T. Nguyen *et al.*, "A sequence matching network for polyphonic sound event localization and detection," in *IEEE ICASSP*, 2020, pp. 71–75.

[3] S. Adavanne *et al.*, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE JSTSP*, vol. 13, no. 1, pp. 34–48, 2018.

[4] K. Shimada *et al.*, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE ICASSP*, 2021, pp. 915–919.

[5] A. Politis *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, pp. 1–6, 2022.

[6] Q. Wang, L. Chai, H. Wu, *et al.*, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," DCASE2022 Challenge, Tech. Rep., Jun. 2022.

[7] P. Morgado *et al.*, "Learning representations from audio-visual spatial alignment," *NeurIPS*, vol. 33, pp. 4733–4744, 2020.

[8] R. Arandjelovic *et al.*, "Look, listen and learn," in *IEEE ICCV*, 2017, pp. 609–617.

[9] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, pp. 1–6, 2019.

[10] N. Takahashi *et al.*, "Deep convolutional neural networks and data augmentation for acoustic event detection," in *INTERSPEECH*, 2016, pp. 2982–2986.

[11] L. Mazzon *et al.*, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," in *DCASE Workshop*, 2019, p. 154.

[12] Q. Wang, J. Du, H. Wu, J. Pan, F. Ma, and C. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *CoRR*, vol. abs/2101.02919, 2021. arXiv: 2101.02919. [Online]. Available: https://arxiv.org/abs/2101.02919.

[13] Y. Gong *et al.*, "SSAST: Self-supervised audio spectrogram transformer," in *AAAI*, vol. 36, 2022, pp. 10 699–10 709.

[14] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017, pp. 776–780.

[15] R. Scheibler *et al.*, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel separation network," in *DCASE Workshop*, 2022, pp. 1–5.

[16] B. Korbar *et al.*, "Cooperative learning of audio and video models from self-supervised synchronization," *NeurIPS*, vol. 31, pp. 7774–7785, 2018.

[17] Á. González, "Measurement of areas on a sphere using fibonacci and latitude–longitude lattices," *Mathematical Geosciences*, vol. 42, pp. 49–64, 2010.

[18] E. W. Weisstein, "Gnomonic projection," *https://mathworld. wolfram. com/*, 2001.

[19] A. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, pp. 1–13, 2018.

[20] K. Shimada *et al.*, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE ICASSP*, 2022, pp. 316–320.

[21] L. Perotin *et al.*, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *IEEE IWAENC*, 2018, pp. 241–245.

[22] I. Loshchilov *et al.*, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[23] E. Fonseca *et al.*, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM TASLP*, vol. 30, pp. 829–852, 2021.

[24] A. Politis *et al.*, *TAU spatial room impulse response database (TAU-SRIR DB)*, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios

Shi Cheng†, Jun Du†, Qing Wang† *, Ya Jiang†,
Zhaoxu Nian†, Shutong Niu†, Chin-Hui Lee‡, Yu Gao§ and Wenbin Zhang§
† University of Science and Technology of China, Hefei, Anhui, PR China
‡ Georgia Institute of Technology, Atlanta, GA, USA
§ AI Innovation Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China

***Abstract*—In this study, we propose a novel approach to sound event localization and detection (SELD) by using sound separation (SS) models to tackle key challenges of a high percentage of overlapped segments between sound events and imbalanced distributions of sound event classes in real-world scenarios. Specifically, we introduce class-dependent SS models to deal with overlapping mixtures and extract features from the SS model as prompts for SELD of a specific event class. The proposed SS-SELD method enhances the overall performance of the SELD system, resulting in improved accuracy and robustness in real-world scenarios. In contrast to many other classification methods that can be affected by the interference events, the proposed class-dependent SS framework enhances the overall performance of the SELD system, resulting in improved accuracies and robustness in real-world scenarios. When evaluated on the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset, we demonstrate significant improvements in both sound event detection (SED) and direction-of-arrival (DOA) estimation. Our findings suggest that sound separation is a promising strategy to enhance the performance of SELD systems, particularly in scenarios with high overlaps between sound events and imbalanced distributions of event classes. In addition, our proposed framework had contributed building to our champion systems submitted to the Challenge of DCASE 2023 Task 3.**

I. Introduction

Sound event localization and detection (SELD) is an essential task in various audio processing applications, including surveillance systems [1]–[3], environmental monitoring [4], [5], and augmented reality [6], [7]. It was initially introduced as Task 3 in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge [8], [9]. The goal of SELD is to accurately estimate the temporal onsets and offsets, spatial locations, as well as categories of sound events simultaneously within an audio recording. Over the past few years, significant advancements have been made in SELD techniques, driven by the increasing demand for robust and efficient sound analysis algorithms.

SELDnet was proposed to map the input to the sound event detection (SED) and the 3-D Cartesian coordinates of direction-of-arrival (DOA) with two separated branches [10], [11]. In this paper, we use SEDDOA to replace the two-branch network. Shimada et al. introduced an output format called Activity-coupled Cartesian DOA (ACCDOA) for SELD [12]. ACCDOA associates event activity with the length of the Cartesian DOA vector, enabling the solution of SELD task without the requirement of separate branches [13], [14]. However, ACCDOA assumes that each event occurs only once within a frame. When multiple sources of the same event class exist in different directions at the same time, ACCDOA becomes unreliable, leading to angular errors. To address this issue, multi-ACCDOA was proposed [15], which adds a track for concurrence of sources in the prediction results, thereby yielding more accurate localization results. Wang et al. employed data augmentation techniques and performed model fusion with the previous SELD methods, achieving the state-of-the-art performance and ranking the first place in the DCASE 2022 Task 3 Challenge [16]. While SELD models have achieved impressive performance in real-world scenarios, obtaining excellent results in the DCASE 2019-2022 Task 3 Challenge series [8], [9], [13], [17], they face challenges due to the unbalanced distribution of sound events and severe overlapping in real-world scenes. As a result, pure SELD classification systems struggle to effectively model all sound event classes. Therefore, we aim to explore a frontend approach to enable the SELD system to better model each individual class. It has been proven that the separation method is effective in weakly labeled sound event detection (SED) tasks [18]–[20]. However, the effectiveness of sound separation in SELD tasks has been less extensively studied. In DCASE 2022, Jin-Young and his team attempted to use separation as a branch to handle SED information, ultimately merging it with DOA information for further processing [21]. However, such an approach led to the waste of spatial information contained in the 4-channel recordings within the SED branch and resulted in SED information relying entirely on the performance of separation, which could easily lead to unstable results.

In this paper, we introduce a new framework that employs sound separation (SS) as the front-end of SELD model, named SS-SELD. By combining the separated sound with the original mixture, better results can be achieved compared to the traditional SELD methods which only use the mixture. We evaluate our approach on the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [22] and demonstrate significant

* Corresponding author: qingwang2@ustc.edu.cn

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

improvements in both SED and DOA estimation. Our findings suggest that sound separation is a promising strategy to enhance the performance of SELD systems. Besides, the newly proposed methods made a great contribution to our system ranking the first place in the DCASE 2023 Task3 Challenge.

II. Proposed Methods

In real-world scenarios, sound events often exhibit a high degree of overlap, as depicted in Table I. It can be observed that all sound events exhibit an overlap rate of over 50%, with certain classes such as clapping and laughter having overlap rates exceeding 80%. The Bell class, in particular, has a 100% overlap rate, indicating that all instances of this class overlap with other sound events in the dataset. Traditional SELD models focus on classification-based approaches, facing difficulties in adequately modeling these overlapping events. Building upon this foundation, we propose a novel SS-SELD method. Fig.1 illustrates this new framework.

A. Sound Separation

The newly proposed method consists of two main components. The first introduced component of SS-SELD is the SS method as depicted by the blue dashed box in Fig.1. SS illustrates the inference process of the separation model. As mentioned earlier, due to the uneven distribution of sound events, some events have short durations. Therefore, we sought to enhance the richness of the data by sourcing the required sound events from open-source datasets. Given a mixed audio signal $x(t)$, where t is the sample index in time domain. $x(t)$ contains target sound event $s(t)$ and multiple Interfering sound classes $x_1(t), \cdots, x_m(t)$:

$$x(t) = s(t) + \cdots + x_m(t), \quad m = 1, \ldots, n \tag{1}$$

where n represents the number of events in the mixed audio. The separation process is performed as follows:

$$f_\theta(x(t)) = \hat{s}(t) \tag{2}$$

Here, $f_\theta(\cdot)$ denotes the separation model, θ represents the model parameter set. $\hat{s}(t)$ is the extracted sound. In this process, we train a widely used time-domain model architecture, namely, Conv-TasNet [23] on the dataset simulated from both STARSS23 and Audioset. We first segment data into single-event segments based their labels, which are then added in the time domain to generate mix audios. Simulating training data for the sound separation model also considers the proportion of overlap in natural conditions. The separation model is capable of filtering out the mixed audios and providing separated event distributions, which can serve as guidance for further analysis and modeling. It is worth mentioning that due to significant differences between different event classes, we constructed a dedicated training set for each class with that class as the target. This allows us to achieve class-dependent sound separation. This approach effectively mitigates detection errors of low-intensity classes with significant overlap.

During the fusion process of SS-SELD, the mixture $x(t)$ will be first passed through Audio Channel Swapping (ACS) to obtain augmented signal $x'(t)$. Then for each class, $x'(t)$ will be sent into the separation model to obtain separated sound $\hat{s}(t)$ that only contains the specific class of interest. Afterward, audio signal $\hat{s}(t)$ and $x'(t)$ undergo feature extraction modules separately. $S_{\text{lm}}(k,l)$ denotes log Mel-spectrogram extracted from $\hat{s}(t)$, $X_{\text{lm}}(k,l)$ and $X_{\text{iv}}(k,l)$ denote log Mel-spectrogram and intensity vectors extracted from $x'(t)$, respectively. These features are concatenated to train the SS-SELD model, resulting in output O for detection and localization. We can represent the entire process using the following equations, corresponding representations in Fig.1:

$$x'(t) = \text{ACS}(x(t)) \tag{3}$$

$$\hat{s}(t) = \text{SS}(x'(t)) \tag{4}$$

$$S_{\text{lm}}(k,l) = \text{FE1}(\hat{s}(t)) \tag{5}$$

$$\{X_{\text{lm}}(k,l), X_{\text{iv}}(k,l)\} = \text{FE2}(x'(t)) \tag{6}$$

$$O = \text{SELD}(S_{\text{lm}}(k,l), X_{\text{lm}}(k,l), X_{\text{iv}}(k,l)) \tag{7}$$

where ACS$(\cdot)$, SS$(\cdot)$, FE1$(\cdot)$, FE2$(\cdot)$ and SELD$(\cdot)$ denote the audio channel swapping, sound separation, separated sound feature extraction, mixed sound features extraction and SELD predicting operations, respectively. The final output O contains both frame-level sound event labels and the angle information.

B. Sound Event Localization and Detection

Fig.1 (b) illustrates the training process of the traditional SELD model. Mixture from STARSS23 undergo the data augmentation technique ACS [24] to increase the data size by approximately 8 times. The augmented mixture are then fed into the feature extraction module, which extracts log Mel-spectrogram containing event label information and intensity vector features containing spatial location information. These features are used to train the SELD which follows a similar principle to the traditional SELD approach, consistent with the SELD methodology employed in our winning solution for DCASE 2022 TASK 3. For more detailed information, please refer to reference [16].

III. Experiment Setup

We evaluate SELD on the official development set of the DCASE 2023 Task 3, which is collected in realistic spatial soundscapes [22]. The set totals 168 recording clips (about 9 hours), which can be split into a training part (dev-train, 90 clips) and a testing part (dev-test, 78 clips) according to the official set. We have utilized labeled segments of individual sound events from Audioset [25], in addition to the DCASE 2023 Task 3 dev-train data, to construct training data for sound separation. The total duration of single-event segments amounts to approximately 39.42 hours. Out of this duration, 6.21 hours are obtained from the STARSS23 dev-train dataset, while the remaining 33.21 hours are sourced from the Audioset dataset. For each sound event class, we utilize its corresponding single-event segments as targets and treat the remaining sound events as noise. During the data simulation

TABLE I
OVERLAPPING PERCENTAGE (%).

	Woman speaking	Man speaking	Clap	Telephone	Laughter	Domestic sounds	Footsteps	Door	Music	Musical instrument	Water tap	Bell	Knock
Overlap	59.98	54.25	83.19	51.89	84.79	69.56	54.79	66.77	67.20	58.62	79.79	100.00	82.22

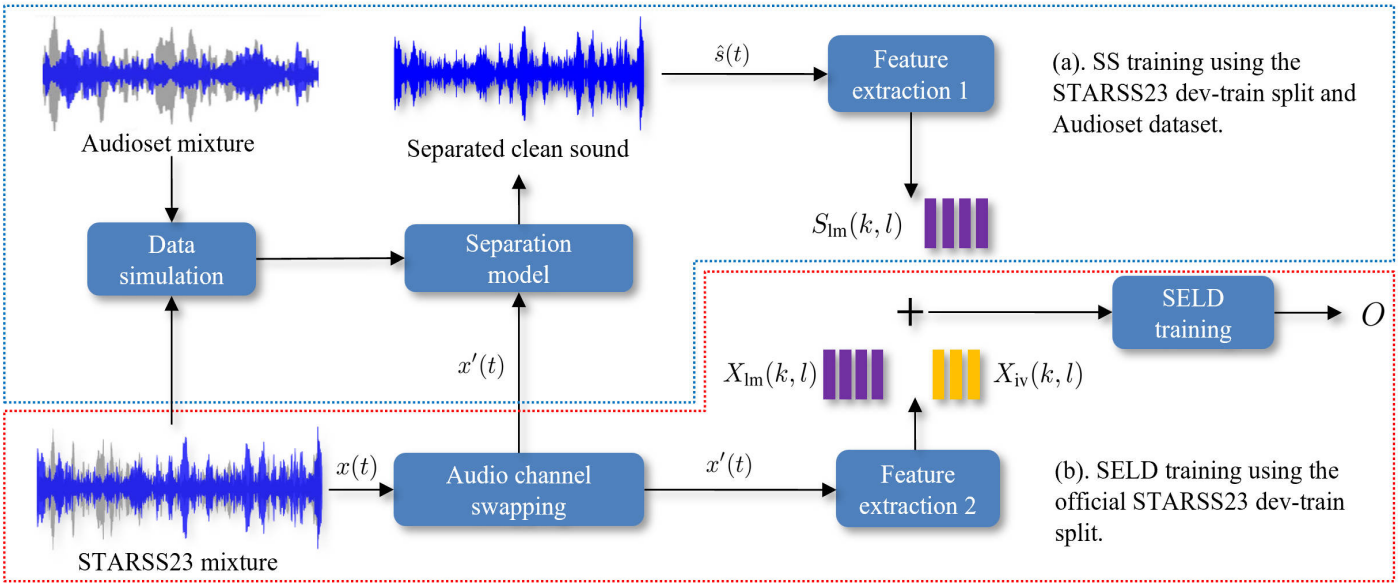


Fig. 1. Framework of the proposed SS-SELD system. Note that (a) in the blue dashed box represents the inference process of SS. while (b) in the red dashed box represents the inference process of the SELD model. In the data simulation part, the data from the dev-train set and open-source dataset are split into single-event segments based on their labels. Once the target class is determined, other classes are considered as interference, and simulation is carried out accordingly.

phase, we overlay the target sound event segments with other sound event segments in the time domain to create mixed audio. This allows us to simulate approximately 40 hours of data, which is used for training the corresponding separation model. All recordings from STARSS23 are 4-channel with a 24 kHz sampling rate. Similarly, single-channel speech from Audioset are also downsampled to 24 kHz. There are 13 sound classes in total. We apply the short-term Fourier transform (STFT) with 40 ms frame length and 20 ms frame hop on 4-channel first-order Ambisonics (FOA) [26] audios to extract log Mel-spectrogram. We concatenate them to get the 11-channel feature at each frame. Considering its outstanding performance in real-world scenarios in recent years, the separation models is trained on ConvTasNet architecture and the model parameter settings can be referenced in [23].

For the SS-SELD training, the data size can be augmented to 8 times by applying the ACS strategy, which results in about 192 hours of data. We use the ResNet-Conformer as our main model architecture [27]. The model employs 8 attention heads, with input, key, and value vectors having dimensions of 256, 32, and 32, respectively. It consists of 8 Conformer layers. The optimizer used is Adam, and a warm-up learning schedule is employed with a maximum learning rate of 0.001. Batch-size is set to 32. The training process is conducted for a maximum of 180,000 steps (which is around). All experiments are conducted on a NVIDIA Tesla V100 Graphic Card-32GB GPU.

All methods are evaluated using SELD_{score} [28], which is calculated as follows:

$$\text{SELD}_{score} = \frac{\text{ER}_{20^\circ} + (1 - \text{F}_{20^\circ}) + \text{LE}'_{\text{CD}} + (1 - \text{LR}_{\text{CD}})}{4} \tag{8}$$

$$\text{LE}'_{\text{CD}} = \frac{\text{LE}_{\text{CD}}}{\pi} \tag{9}$$

where ER_{20° and F_{20° are location-dependent error rate and F-score when the spatial error is within 20°. Note that the F_{20°, LE_{CD} and LR_{CD} are calculated through macro-averaging follow the official definition.

IV. RESULTS AND ANALYSIS

The dev-train split of the official development set of the DCASE 2023 Task 3 is used for training the SELD model as

TABLE II
EXPERIMENTAL RESULTS FOR DEVELOPMENT SET OF DCASE 2023 TASK 3 BASELINE AND SOME SELD METHODS, AS WELL AS THE PROPOSED SS-SELD METHOD. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$SELD_{score} \downarrow$
Baseline-FOA [22]	0.57	0.30	21.60	0.48	0.478
SEDDOA	0.41	0.59	14.05	0.70	0.304
Multi-ACCDOA	0.44	0.58	13.75	0.74	0.303
ACCDOA	0.42	0.59	13.72	0.72	0.300
SS-SELD	**0.40**	**0.64**	**13.40**	**0.74**	**0.279**

well as the newly proposed SS-SELD framework. The dev-test split is used to validate the performance of the models.

Table II shows the overall experimental results of the proposed method for development dataset. "ACCDOA", "SEDDOA" represent the ACCDOA- and SEDDOA-based modeling method, where SEDDOA adopts a two-branch (SED and DOA) approach for SELD modeling. "Multi-ACCDOA" represents the multi-ACCDOA-based method. SS-SELD denotes the proposed SELD model with SS front-end, with the same output format as SEDDOA. As shown in the table, each proposed single model outperforms the two baseline systems by a large margin. It can also be observed that compared to the baseline, traditional SELD models (SEDDOA, ACCDOA, Multi-ACCDOA) exhibit a decrease of approximately 0.18 in $SELD_{score}$, dropping from 0.478 to around 0.300. Despite attempting to use more data and iterative optimization, there was no significant improvement in the performance of the SELD models. However, with the assistance of the separation method, our SS-SELD system further reduces $SELD_{score}$ to 0.279, breaking through the bottleneck around 0.30 for SELD performance. Additionally, the results of the ER_{20° and F_{20° metrics demonstrate that the inclusion of the separation method significantly improves the model's performance in the detection aspect. As the localization results rely on the accuracy of the detection, the introduction of separation indirectly enhances the localization performance, leading to improvements in LE_{CD} and LR_{CD} as well.

Furthermore, we present the results for each individual class as shown in Table III. It can be observed that the separation method achieves significant improvements in the majority of classes, especially for classes like 'class11' (i.e., *Bell*). These classes have short durations within the entire test set but poss distinct features. The separation method effectively distinguishes them from other classes. However, for the class Water tap, the separation method is ineffective. Through specific analysis, we discovered that Water tap sounds are often low in volume, distant from the microphones (the official dataset provides auxiliary video information, but it is not allowed to be used during the testing process in the Audio-only track. Therefore, we only utilized the video information for analysis purposes.) More importantly, there exists interference noise in the test set that closely resembles the sound of Water tap. This makes it challenging for the separation model to accurately isolate the target Faucet sounds, thereby misleading

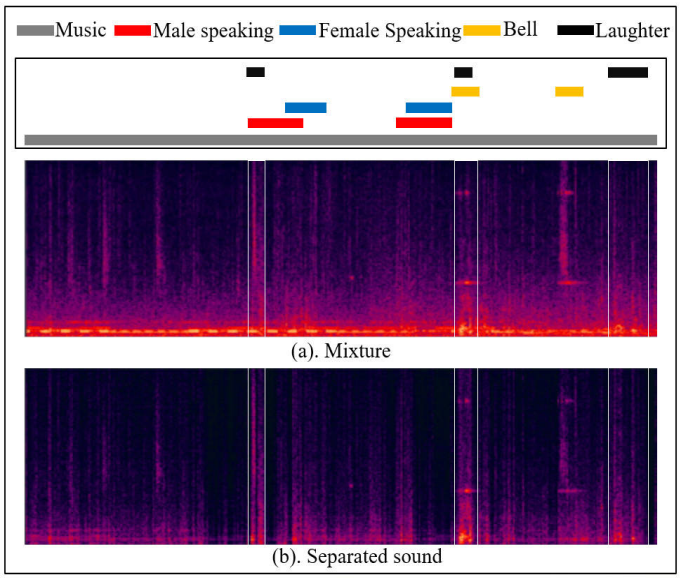


Fig. 2. Comparison of spectrograms before and after sound separation, where the colored rectangular bars at the top represent temporal event labels. The target class in this case is laughter. The model is able to preserve the laughter while effectively removing other sounds such as music.

the training of the subsequent SS-SELD model.

Fig.2 illustrates the spectrogram comparison of the input mixture before and after separation. The colored rectangular tags at the top represent different sound events, with laughter being the target class in this example. The white regions represent the target sound event, which is *Laughter*. It can be observed that the input mixture contains background 'class 8' (*Music*) with high energy in the low-frequency range throughout the entire audio segment, as well as overlapping conversations between different speakers ('class 0' and 'class 1', i.e., *Female speaking* and *Male speaking*), and occasional instances of bell sounds. After passing through the separation model, the target class, laughter, is preserved completely while effectively removing the background music and other interfering sound events. Additionally, in the first white rectangular region, 'class 4' (*Laughter*) in the original mixed audio is almost completely masked by 'class 8' (*Music*). This would cause significant interference for the laughter class in the SELD methods. However, after applying the separation process, it can be observed that the music sound is almost completely removed, allowing the laughter to be clearly revealed. This ensures 'class 4' can be detected more accurately. All of the above indicate that the application of the separation model can eliminate non-target classes with high energy, allowing the extraction of target classes that may have been masked by other sounds with lower energy. This provides a cleaner reference for the SELD model.

In order to analyze the performance of the model more comprehensively, we visualized the SED and DOA estimation, presented in Fig.3. The horizontal axis in all subplots

TABLE III
SELD$_{score}$ COMPARISON BETWEEN SELD SYSTEM AND THE PROPOSED SS-SELD METHOD OF ALL 13 CLASSES FOR DEVELOPMENT DATASET. THE BETTER RESULTS ARE HIGHLIGHTED IN BOLD.

	Woman speaking	Man speaking	Clap	Telephone	Laughter	Domestic sounds	Footsteps	Door	Music	Musical instrument	Water tap	Bell	Knock
SELD	0.243	0.231	0.254	0.345	0.377	0.341	0.458	0.318	0.241	0.257	**0.276**	0.338	0.278
SS-SELD	**0.227**	**0.217**	**0.231**	**0.313**	**0.359**	**0.303**	**0.421**	**0.282**	**0.236**	**0.254**	0.314	**0.248**	**0.244**

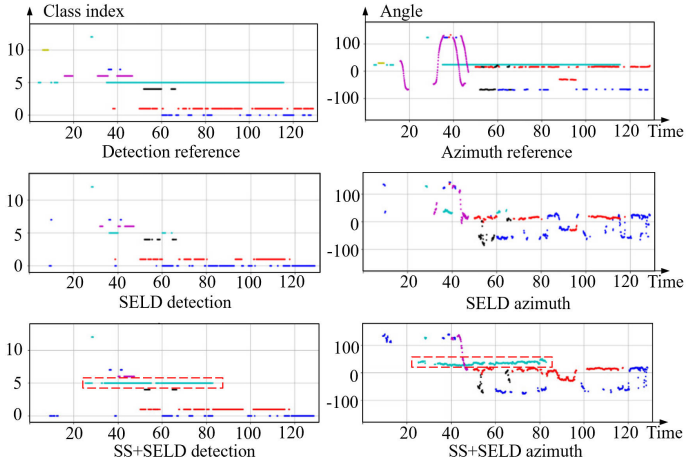


Fig. 3. The visualized results comparing SELD and SS-SELD, with the two columns representing the detection results and azimuth angle results, respectively. Different colors represent different events. The horizontal axis represents time (s) for all subplots. The vertical axis in the left column represents event class index, while the vertical axis in the right column represents angles in degrees (°).

represents time in seconds, while different colors represent different sound events. The first column represents the SED results, where the vertical axis represents 13 sound event classes. The occurrence of a color indicates the presence of the corresponding sound event at that time. The second column represents the DOA results. It is worth mentioning that the presented information pertains to the azimuth angle. The trend in the elevation angle is similar to that of the azimuth angle. To avoid clutter and congestion in the figure, the elevation angle information is not shown. The azimuth angle ranges from -180° to 180°, illustrating the relative angles of sound events with respect to the origin of the microphone array. The first row represents the ground truth results, while the second and third rows represent the results of SELD and SS-SELD, respectively. From the figure, it can be observed that SS-SELD exhibits a significant improvement over SELD in the 'class 5' (light blue line, *Domestic sounds*), as shown in the dashed boxes. SELD almost entirely missed the home-related sounds, thus failing to provide angle information for this class. In contrast, SS-SELD can detect the majority of these sounds and accurately provide angle information. This further demonstrates that our SS-SELD not only enhances the model's detection capability but also brings improvements in localization performance.

V. CONCLUSIONS

In this study, we explored the effectiveness of the sound separation method in sound event localization and detection task. By using a separation model as a front-end to obtain separated audio for each event class, we were able to extract the features of both the sound event mixtures and the separated sound. The combined SS-SELD model trained on these features effectively improved the performance of single SELD model. In future work, we will continue to explore more effective and universally applicable separation strategies to address the existing issues and enhance the performance of the SELD model.

VI. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427.

REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[2] A. Greco, N. Petkov, A. Saggese, and M. Vento, "Aren: A deep learning approach for sound event recognition using a brain inspired representation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3610–3624, 2020. DOI: 10.1109/TIFS.2020.2994740.

[3] V. Suruthhi, V. Smita, R. Gini J, and K. Ramachandran, "Detection and localization of audio event for home surveillance using crnn," *International Journal of Electronics and Telecommunications*, vol. 67, no. 4, 2021.

[4] F. Angulo, S. Essid, G. Peeters, and C. Mietlicki, "Cosmopolite sound monitoring (cosmo): A study of urban sound event detection systems generalizing to multiple cities," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095833.

[5] W. Xiong, X. Xu, L. Chen, and J. Yang, "Sound-based construction activity monitoring with deep learning," *Buildings*, vol. 12, no. 11, p. 1947, 2022.

[6] A. Härmä, J. Jakka, M. Tikander, *et al.*, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[7] B. S. Liang, A. S. Liang, I. Roman, *et al.*, "Reconstructing room scales with a single sound for augmented reality displays," *Journal of Information Display*, vol. 24, no. 1, pp. 1–12, 2023.

[8] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019.

[9] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[10] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019. DOI: 10.1109/JSTSP.2018.2885636.

[11] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," *arXiv preprint arXiv:1908.00766*, 2019.

[12] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 915–919. DOI: 10.1109/ICASSP39728.2021.9413609.

[13] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[14] O. Slizovskaia, G. Wichern, Z.-Q. Wang, and J. Le Roux, "Locate this, not that: Class-conditioned sound event doa estimation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 711–715. DOI: 10.1109/ICASSP43922.2022.9747604.

[15] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 316–320. DOI: 10.1109/ICASSP43922.2022.9746384.

[16] Q. Wang, L. Chai, H. Wu, *et al.*, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," *DCASE2022 Challenge, Tech. Rep.*, 2022.

[17] A. Politis, K. Shimada, P. Sudarsanam, *et al.*, "Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, 2022.

[18] N. Turpault, S. Wisdom, H. Erdogan, *et al.*, *Improving sound event detection in domestic environments using sound separation*, 2020. arXiv: 2007.03932 [cs.SD].

[19] Q. Zhou, Z. Feng, and E. Benetos, "Adaptive noise reduction for sound event detection using subband-weighted nmf," *Sensors*, vol. 19, no. 14, p. 3206, 2019.

[20] S. Liu, F. Yang, F. Kang, and J. Yang, "A multi-task learning method for weakly supervised sound event detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8802–8806. DOI: 10.1109/ICASSP43922.2022.9746947.

[21] J.-Y. Park, D.-H. Kim, B. H. Ku, *et al.*, "Sound event localization and detection based on cross-modal attention and source separation,"

[22] A. Politis, K. Shimada, P. Sudarsanam, *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, Nov. 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019. DOI: 10.1109/TASLP.2019.2915167.

[24] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023. DOI: 10.1109/TASLP.2023.3256088.

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 776–780.

[26] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[27] S. Niu, J. Du, Q. Wang, *et al.*, "An experimental study on sound event localization and detection under realistic testing conditions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[28] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337. DOI: 10.1109/WASPAA.2019.8937220.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Joint Analysis of Acoustic Scenes and Sound Events Based on Semi-Supervised Approach

Ami Igarashi*, Shunsuke Tsubaki*, Daisuke Niizumi†, Daiki Takeuchi†, Yasunori Ohishi†,
Noboru Harada†, and Keisuke Imoto*
*Doshisha University, Japan
E-mail: keisuke.imoto@ieee.org
†NTT Corporation, Japan

***Abstract*—Considering that acoustic scenes and sound events are closely related to each other, their joint analysis in some previous studies has been explored utilizing multitask learning (MTL)-based neural networks. In conventional approaches, either strongly or weakly supervised schemes are applied to sound event detection in MTL models. However, annotating strong event labels is quite time-consuming or using only weak labels achieves low performance in sound event detection. In this paper, we thus propose a method for the joint analysis of acoustic scenes and sound events based on the MTL framework using both strong and weak labels of sound events. Our experimental results, which were obtained using subsets of the TUT Acoustic Scenes 2016/2017 and TUT Sound Events 2016/2017 datasets, show that the proposed method using both strong and weak labels achieves reasonable performance even with a small number of strongly labeled data and a large number of weakly labeled data.**

I. Introduction

In recent years, there has been great attention to research and development in the computational analysis of environmental sounds. The computational analysis of environmental sounds is applicable to a wide variety of applications including media retrieval, hearing assistance, machine condition monitoring, and biomonitoring systems [1], [2], [3].

There are several significant tasks in the computational analysis of environmental sounds, such as acoustic scene classification (ASC) and sound event detection (SED). Among these tasks, ASC, which predicts a predefined acoustic scene label from an input audio signal, is one of the most major tasks. Moreover, sound event detection (SED), in which sound event labels and their corresponding time stamps within an audio signal are detected, is also a popular task in this area. Many neural-network-based methods have been recently employed in ASC and SED systems, leading to marked advancements in performance. For example, Valenti et al. [4] and Raveh and Amar [5] have proposed ASC methods based on the convolutional neural network (CNN) and ResNet, respectively. Çakır et al. have proposed a SED method using a convolutional recurrent neural network (CRNN) [6]. Kong et al. [7] and Miyazaki et al. [8] have recently proposed a SED method using the Transformer and Conformer encoders, respectively.

A majority of conventional methods analyze acoustic scenes and sound events separately. However, acoustic scenes and sound events are closely related, and it would be inefficient to analyze them separately. For example, in the acoustic scene "park," the sound events "bird singing" and "wind blowing" are likely to co-occur, whereas the sound events "cutlery" and "keyboard typing" rarely co-occur. Considering this relationship, Mesaros et al. [9] and Heittola et al. [10] proposed methods utilizing information on acoustic scenes for SED. Imoto and co-workers proposed ASC methods considering the relationship between sound events and acoustic scenes using Bayesian generative models [11], [12]. Moreover, Mesaros et al. [9], Heittola et al. [10], and Imoto et al. [11], [12] have proposed ASC or SED methods considering the relationship between sound events and acoustic scenes. More recently, Bear et al. [13], Tonami et al. [14], and Jung et al. [15] have proposed the joint analysis of acoustic scenes and sound events based on multitask learning-based neural networks of ASC and SED.

The neural network-based methods for the computational analysis of environmental sounds typically rely on models trained using large numbers of strongly annotated data, which has information of sound types and their time boundaries. However, especially in SED, the process of annotating time boundaries is very time-consuming. One way to alleviate this difficulty is to apply weakly supervised SED, which only utilizes weakly labeled data to train the SED model [16], [17], [18], where weak labels in SED only have information on whether sound events are active or inactive per sound clip. By applying a weakly supervised approach to the MTL framework, Tsubaki et al. [19] have proposed the joint analysis of acoustic scenes and sound events only with weak labels of sound events. This approach achieves a higher performance in ASC than a single-task ASC method, but it results in performance degradation in SED compared with the MTL method using strong labels of sound events.

To address the limitations of the conventional MTL framework using either strong or weak labels of sound events, in this paper, we propose a method of joint analysis of acoustic scenes and sound events in a situation where the strong labels of sound events are partially available, that is, in a semi-supervised scenario [20]. We then investigate ASC and SED experimental results in detail, including performance changes depending on the ratios of strong and weak labels of sound events.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

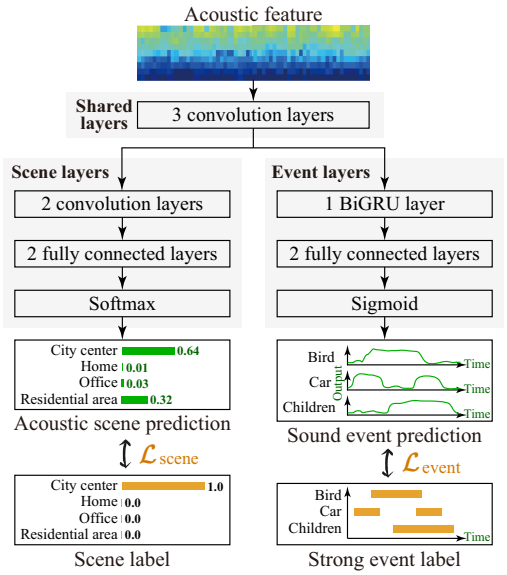


Fig. 1. Network structure of conventional MTL-based method [14]

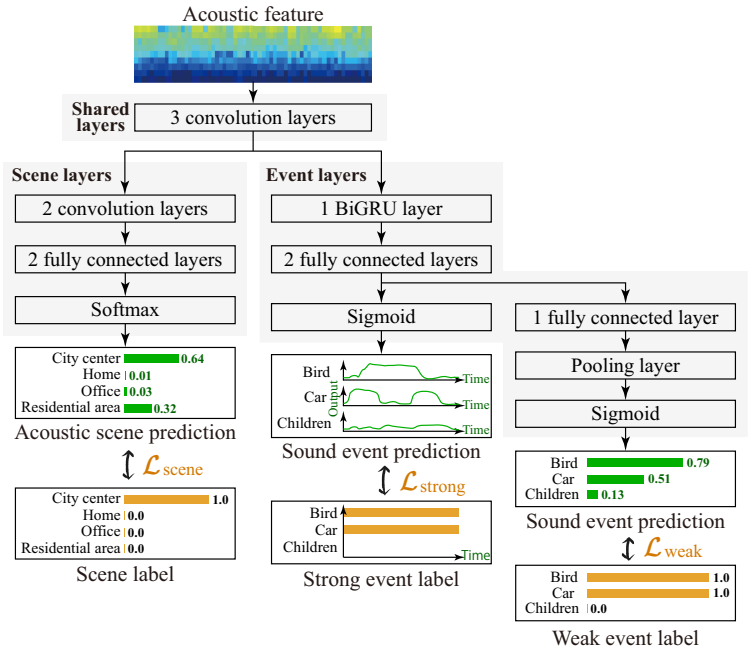


Fig. 2. Network structure of MTL-based method using weak labels of sound events

The remainder of this paper is organized as follows. In section 2, we introduce the conventional methods of ASC, SED, and the joint analysis of scenes and events based on multitask learning. In section 3, we propose a method for the joint analysis of acoustic scenes and sound events using both strong and weak labels of sound events. In section 4, we discuss the experiments conducted to evaluate the ASC and SED performance characteristics. Finally, we conclude our work in section 5.

II. Conventional Methods

A. Acoustic Scene Classification and Event Detection

In this section, we review conventional systems for ASC and SED based on neural networks, such as CNN [4], [21], CRNN [6], and the Transformer encoder [7]. Many ASC and SED systems calculate a time–frequency feature of the acoustic signal $\mathbf{X} \in \mathcal{R}^{D \times T}$, such as the log mel-band spectrogram or a time series of mel frequency cepstrum coefficients (MFCCs), as the acoustic feature. We here denote D and T as the numbers of frequency bins and time frames, respectively. The acoustic feature is fed to the ASC and SED networks, which calculate logits $\mathbf{y}$ for scene classification and event detection, respectively. The model parameters of the ASC network are trained using the logits and the cross-entropy (CE) loss function $\mathcal{L}_{\text{scene}}$ as

$$\mathcal{L}_{\text{scene}} = -\sum_{n=1}^{N} \Big\{ z_n \log\big(\sigma\left(y_n\right)\big) \Big\}, \tag{1}$$

where N, $z_n \in \{0, 1\}$, and σ are the number of acoustic scene classes, the target one-hot scene label, and softmax function, respectively.

The model parameters of the SED network are trained using the logits and the binary cross-entropy (BCE) loss function $\mathcal{L}_{\text{event}}$ as

$$\begin{aligned}\mathcal{L}_{\text{event}} &= -\sum_{t=1}^{T} \Big\{ \mathbf{z}_t \log(\mathbf{y}_t) + (1-\mathbf{z}_t) \log\big(1-s(\mathbf{y}_t)\big) \Big\} \\ &= -\sum_{t,m=1}^{T,M} \Big\{ z_{t,m} \log s(y_{t,m}) + (1-z_{t,m}) \log\big(1-s(y_{t,m})\big) \Big\},\end{aligned} \tag{2}$$

where T, M, $z_{t,m}$, and s are the number of time frames, the number of sound event classes, the target event label for sound event m in time frame t, and sigmoid function, respectively.

B. Joint Analysis of Acoustic Scenes and Sound Events Based on Multitask Learning

Many methods for environmental sound analysis handle scene classification and event detection as separate tasks. Moreover, some recent works employ methods that jointly analyze acoustic scenes and sound events, which are based on the idea that information on acoustic scenes and sound events is mutually beneficial for ASC and SED. In particular, several works apply a multitask (MTL)-based neural network for the joint analysis of ASC and SED [13], [14], [15], [22].

Figure 1 shows a typical implementation of the MTL-based method, which shares part of the network and information on acoustic scenes and sound events. In the shared layers, acoustic features are embedded in the feature space common to acoustic scenes and sound events. This acoustic embedding is then fed to the scene and event-specific layers that perform ASC and SED, respectively. As the scene and event-specific

layers, CNN, the recurrent neural network (RNN), and the Transformer encoder are often applied.

To train the model parameters, the following linear combination of Eqs. (1) and (2) is adopted as the loss function to train the MTL model:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{scene}} + \beta\mathcal{L}_{\text{event}}, \tag{3}$$

where α and β are the constant weights of scene classification and event detection losses, respectively. In this paper, $\beta = 1.0$ without loss of generality.

C. Joint Analysis of Scenes and Events Based on Multitask Learning With Weak Labels of Sound Events

Annotating strong labels for sound events is very time-consuming. To alleviate this problem of annotation, many SED methods utilize a weakly supervised approach with weak labels of sound events.

In particular, Tsubaki *et al.* have proposed a method for the joint analysis of acoustic scenes and events [19] using weak labels of sound events. This method applies the multiple-instance learning (MIL) framework [23] to the weakly supervised training of SED. As shown in Fig. 2, the network of this method has two branches in the event layers. One event branch has the MIL-based pooling layer immediately before the sigmoid function, which enables the weakly supervised training of a SED model. The other branch holds temporal information to predict sound events and their timestamps as in conventional methods.

The model parameters of the weakly supervised method are trained using Eq. (3), whereas the loss function $\mathcal{L}_{\text{event}}$ is expressed as

$$\begin{aligned}\mathcal{L}_{\text{event}} &= \gamma\mathcal{L}_{\text{strong}} + \zeta\mathcal{L}_{\text{weak}} \\ &= -\gamma \sum_{m,t=1}^{M,T} \Big\{ z_{m,t} \log s(y_{m,t}) + (1 - z_{m,t}) \log\big(1 - s(y_{m,t})\big) \Big\} \\ &\quad - \zeta \sum_{m=1}^{M} \Big\{ z_m \log s(y_m) + (1 - z_m) \log\big(1 - s(y_m)\big) \Big\},\end{aligned} \tag{4}$$

where y_m and z_m are the bag (clip) level prediction of the sound event m and the weak label of the target event, respectively. γ and ζ are the constant weights of event detection losses. Here, the strong event label $z_{m,t}$ is generated from the weak label z_m as a pseudo-label as follows:

$$\mathbf{Z}_{\mathsf{pseudo_strong}} = \begin{pmatrix} \overbrace{\begin{matrix} z_1 & \cdots & z_1 & \cdots & z_1 \end{matrix}}^{T} \\ \begin{matrix} \vdots & \ddots & \vdots & & \vdots \end{matrix} \\ \begin{matrix} z_m & \cdots & z_m & \cdots & z_m \end{matrix} \\ \begin{matrix} \vdots & & \vdots & \ddots & \vdots \end{matrix} \\ \begin{matrix} z_M & \cdots & z_M & \cdots & z_M \end{matrix} \end{pmatrix} \tag{5}$$

TABLE I
DETAILED STRUCTURE OF MTL NETWORK OF ASC AND SED

Shared layers		
Log-mel energy (500 frames × 64 mel bin)		
3×3 kernel size/128 ch. Batch norm., Leaky ReLU 1×8 Max pooling		
(3×3 kernel size/128 ch. Batch norm., Leaky ReLU 1×2 Max pooling) × 2		
Scene layers	**Event layers**	
3×3 kernel size/256 ch. Batch norm., Leaky ReLU 25×1 Max pooling	BiGRU w/ 64 units	
3×3 kernel size/256 ch. Batch norm., Leaky ReLU Global max pooling	FC w/ 48 units, Leaky ReLU	
FC w/ 32 units, Leaky ReLU	FC w/ 25 units	
FC w/ 4 units, Softmax	Sigmoid	FC w/ 16 units
		Leaky ReLU
		Global max pooling
		Sigmoid

TABLE II
EXPERIMENTAL CONDITIONS

Acoustic feature	Log-mel energy (64 dim.)
Frame length/shift	40 ms/20 ms
Length of sound clip	10 s
Optimizer	RAdam [24]
α, β	0.001, 1.0
γ, ζ in first strategy	1.0, 0.01
ρ_{GTC}, ρ_{DTC}	0.1, 0.1

III. PROPOSED METHOD

The conventional MTL-based method using weak labels of sound events [19] alleviates the problem of collecting strong labels of sound events. However, the joint analysis of ASC and SED using weak labels results in a lower SED performance than that using strongly labeled data [14]. In the proposed method, we then apply a semi-supervised approach to the joint analysis of ASC and SED. In general, a method that combines supervised and unsupervised training schemes is called semi-supervised learning. In this paper, however, we refer to the method that uses both strong and weak labels for the model training of SED as semi-supervised learning, following [20].

For the semi-supervised learning approach, we construct the following acoustic feature set $\mathcal{X}_{\mathsf{semi}}$ and label set $\mathcal{Z}_{\mathsf{semi}}$ from $\mathcal{X}_{\mathsf{strong}}$, $\mathcal{X}_{\mathsf{weak}}$, $\mathcal{Z}_{\mathsf{strong}}$, and $\mathcal{Z}_{\mathsf{weak}}$.

$$\mathcal{X}_{\mathsf{semi}} = \{\mathcal{X}_{\mathsf{strong}}, \mathcal{X}_{\mathsf{weak}}\}, \tag{6}$$

$$\mathcal{Z}_{\mathsf{semi}} = \{\mathcal{Z}_{\mathsf{strong}}, \mathcal{Z}_{\mathsf{weak}}\}, \tag{7}$$

where $\mathcal{X}_{\mathsf{strong}}$, $\mathcal{X}_{\mathsf{weak}}$, $\mathcal{Z}_{\mathsf{strong}}$, and $\mathcal{Z}_{\mathsf{weak}}$ indicates the acoustic feature set with strong labels, the acoustic feature set with weak labels, the string label set, and weak label set, respectively.

TABLE III
OVERALL PERFORMANCE CHARACTERISTICS OF ASC AND SED

Method	Scene		Event (Frame-based)		Event (IS-based)	
	Micro-Fscore	Macro-Fscore	Micro-Fscore	Macro-Fscore	Micro-Fscore	Macro-Fscore
Strong MTL	90.82% ± 3.38	91.13% ± 3.47	**52.94% ± 2.62**	**23.91% ± 1.67**	**25.89% ± 1.31**	**16.24% ± 1.11**
Weak MTL	90.85% ± 2.11	91.26% ± 2.10	36.07% ± 0.85	16.57% ± 0.53	25.51% ± 1.65	14.89% ± 0.87
Semi MTL w/ const. weights	**92.62% ± 2.50**	**92.88% ± 2.98**	44.44% ± 1.73	18.75% ± 0.77	19.29% ± 1.21	11.51% ± 0.96
Semi MTL w/ delta func.	90.53% ± 3.45	90.67% ± 3.91	52.04% ± 2.47	22.38% ± 1.53	23.94% ± 2.57	15.28% ± 1.76

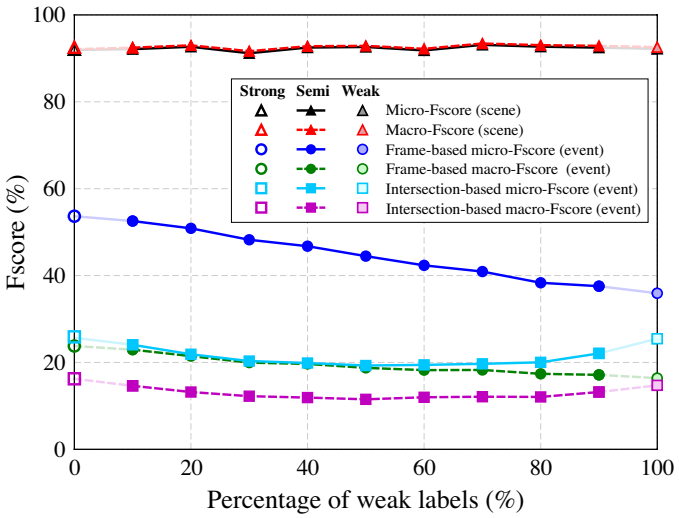


Fig. 3. ASC and SED performance characteristics for various ratios of weakly labeled data of sound events with constant weights γ and ζ

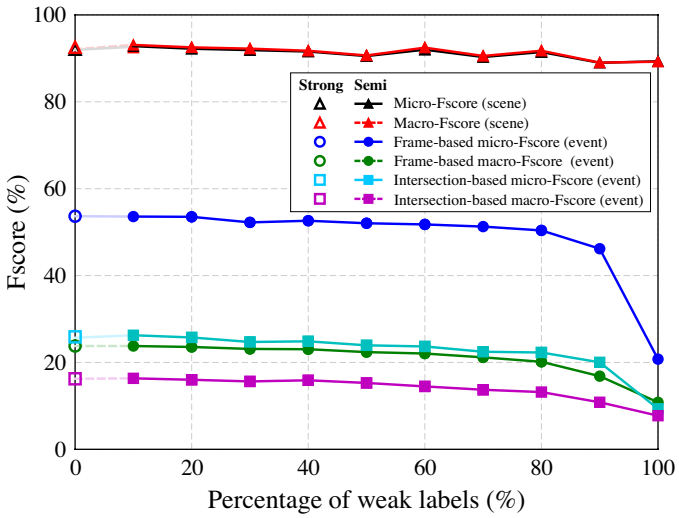


Fig. 4. ASC and SED performance characteristics for various ratios of weakly labeled data of sound events with Kronecker delta function δ_γ and δ_ζ

In the proposed method, we apply the same network architecture as that for the conventional weakly supervised MTL approach, as shown in Fig. 2. We then apply two strategies to train the MTL model. In the first one, we prepare both strong and weak labels for each training data using Eq. (5) if the acoustic feature has only weak labels. The model parameters are then trained using the same Eqs. (3) and (4) as in the conventional MTL-based method using weakly labeled data. In the second strategy, we respectively replace γ and ζ with the following Kronecker delta functions:

$$\delta_\gamma = \begin{cases} 1 & \text{if } \mathbf{z} \text{ is strong label} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$\delta_\zeta = \begin{cases} 1 & \text{if } \mathbf{z} \text{ is weak label} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

IV. EVALUATION EXPERIMENTS

A. Experimental Conditions

We conducted evaluation experiments to determinate how effective the proposed method can analyze environmental sounds using a limited number of strong labels. In our experiments, we used datasets composed of the TUT Acoustic Scenes 2016/2017 and TUT Sound Events 2016/2017 [25], [26]. Among these datasets, we selected four acoustic scenes, namely *city center*, *home*, *office*, and *residential area*, and 25 sound events, such as *bird singing*, *car*, and *keyboard typing*. These sounds comprised a total of 192 min of sounds for the training set and 74 min for the evaluation set. The weak labels for sound events were created using data from a part of the training set with strong labels, which were selected randomly. The details of the dataset used for the evaluation experiments are in [27]. We also evaluated the MTL-based method using only strong label data [14] and that using only weak label data [19] as comparison methods.

As an acoustic feature, we applied the 64-dimensional log mel-band spectrogram, which was calculated with a frame length of 40 ms and a hop size of 20 ms. The model architecture for this experiment is shown in Table I, which is originally proposed in [14]. We performed the evaluation experiments 10 times under each experimental condition with random initial values of model parameters. The detailed experimental conditions are listed in Table II, which were determined by referring to [19].

TABLE IV
NUMBERS OF TN, FP, AND FN FOR FRAME-BASED AND INTERSECTION-BASED METRICS (SEMI MTL WITH CONSTANT WEIGHTS γ AND ζ)

Percentage of weak label	Frame-based			Intersection-based		
	# TP	# FP	# FN	# TP	# FP	# FN
10%	90,216	86,220	78,388	578	2,753	900
30%	95,260	133,124	71,717	628	4,091	851
50%	97,662	175,527	69,315	622	4,338	872
70%	101,961	229,783	65,017	620	4,184	872
90%	105,319	289,086	61,658	576	3,138	924

B. Experimental Results

1) Overall Performance characteristics of ASC and SED: The average performance characteristics of ASC and SED in terms of Fscore are shown in Table III. Here, we refer to the MTL-based methods of joint analysis of acoustic scenes and sound events using only strong, only weak, and both strong and weak labels as strong MTL, weak MTL, and semi MTL, respectively. As the evaluation metrics for SED, we applied the frame-based and intersection-based metrics [28]. For the experimental results shown in this paper, except for Figs. 3 and 4, we conducted the experiments with half of the strongly labeled data and half of the weakly labeled data under the semi MTL condition. The results show that the proposed method based on semi MTL achieves micro- and macro-Fscores for ASC that are similar to those of the conventional strong and weak MTL methods. The performance of the proposed method in the frame-based SED metric is between those of the strong and weak MTL methods, whereas that using the constant weights in SED losses is lower than those of the strong and weak MTL method in the intersection-based SED metric.

To investigate the detailed behavior of the strong, weak, and semi MTL methods, we show the ASC and SED performance characteristics with various ratios of weakly labeled data of sound events in Figs. 3 and 4. The results show that the ASC performance does not degrade when we use more weak labels of sound events. This result indicates that ASC does not require temporal information on sound events, but requires only information on sound event tags in acoustic signals. In the proposed method with the constant weights γ and ζ, the frame-based Fscore decreases as the percentage of weak labels is increased. On the other hand, in the proposed method with the Kronecker delta functions δ_γ and δ_ζ, the Fscore does not decrease markedly until the percentage of weak labels reaches around 80%. This means that the semi MTL approach with the delta functions achieves reasonable performance even with a small number of strongly labeled data and a large number of weakly labeled data. Thus, the proposed method mitigates the problem of annotating sound event labels. Figure 3 also shows that the intersection-based Fscores of the proposed method with the constant weights decrease as the percentage of weak labels is increased, but gradually increase as the percentage of weak labels exceeds 50%.

We then investigate the numbers of total true positives (TPs), total false positives (FPs), and total false negatives (FNs) to explore why the intersection-based Fscores of semi MTL with the constant weights gradually increase as the percentage of weakly labels exceeds 50%, as shown in Table IV. The table shows that in the frame-based metric, the numbers of TPs and FPs increase with the percentage of weak labels. The large increase in the number of FPs compared with the number of TPs indicates that the model is trained to detect more false frames as the percentage of weak labels increases. On the other hand, in the intersection-based metric, the number of FPs increases with the percentage of weak labels; however, it decreases when the percentage of weak labels exceeds 50%. As shown in Fig. 5, this is because the number of FPs increases as the percentage of weak labels increases, and then it decreases as multiple FP instances are integrated. That is, the intersection-based Fscore tends to improve with increasing percentage of weak labels; however, a large number of FP frames may appear. Thus, the experimental results corollarily revealed that the intersection-based Fscore and polyphonic sound detection score (PSDS) [28] may not appropriatly evaluate the SED performance with many FPs.

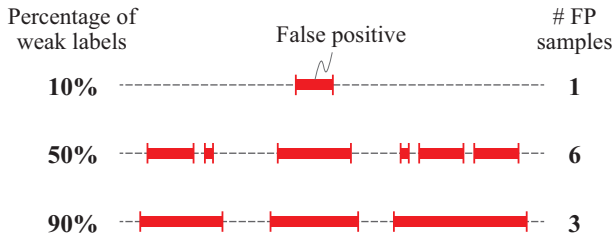


Fig. 5. Number of FPs for ratio of weakly labeled data of sound events

TABLE V
ASC PERFORMANCE FOR EACH SCENE IN TERMS OF FSCORE

Method	city center	home	office	residential area
Strong MTL	89.94% ±4.28	94.28% ±4.61	95.08% ±4.22	85.20% ±6.08
Weak MTL	89.60% ±3.50	95.41% ±1.45	95.50% ±3.46	84.53% ±3.03
Semi MTL w/ const. weights	**92.10%** **±1.44**	**96.66%** **±1.37**	96.67% ±1.43	**88.76%** **±2.47**
Semi MTL w/ delta func.	90.48% ±2.57	93.90% ±3.70	**96.69%** **±1.10**	81.59% ±10.68

2) Detailed performance evaluation for each acoustic scene and event: The ASC performance for each acoustic scene is shown in Table V. The results show that there is no significant difference in ASC performance among the strong, weak, and semi MTL methods. These results also indicate that the temporal information on sound events is not important for each scene classification and that information on sound event tags should be sufficient.

The SED performance for each sound event is shown in Table VI. The results show that the conventional and proposed methods exhibit similar trends. In particular, these methods

TABLE VI
SED PERFORMANCE FOR EACH SOUND EVENT IN TERMS OF FRAME-BASED FSCORE

Method	bird singing	brakes squeaking	car	cutlery	dishes	drawer	fan	keyboard typing	large vehicle	mouse clicking	people walking	washing dishes
Strong MTL	43.35% ±6.13	51.30% ±4.82	**52.03% ±1.62**	9.13% ±6.46	10.65% ±4.88	0.19% ±0.52	**96.82% ±1.08**	**57.61% ±3.02**	16.87% ±1.80	**70.78% ±1.83**	19.01% ±4.23	3.56% ±3.27
Weak MTL	43.13% ±5.93	**51.75% ±4.30**	50.86% ±3.56	**11.76% ±9.63**	12.56% ±6.64	0.07% ±0.20	96.53% ±1.70	56.26% ±4.78	18.04% ±2.41	70.32% ±3.46	20.19% ±4.13	6.07% ±4.78
Semi MTL w/ const. weights	**44.00% ±4.25**	38.14% ±6.23	50.30% ±1.25	4.46% ±5.37	**14.32% ±5.00**	**1.34% ±2.26**	95.44% ±2.01	25.83% ±3.51	16.34% ±2.21	13.41% ±3.96	15.68% ±4.42	**14.68% ±9.32**
Semi MTL w/ delta func.	39.95% ±8.08	40.75% ±12.45	50.22% ±2.92	8.70% ±7.13	12.75% ±4.31	0.06% ±0.21	96.46% ±1.71	54.01% ±4.46	**18.37% ±3.31**	69.81% ±3.21	**20.23% ±5.02**	7.93% ±5.81

show high SED performance for sound events with long durations, such as *bird singing* and *fan*, whereas they show low performance for instantaneous sounds, such as *cutlery* and *drawers*. For a detailed discussion on the reason for these results, refer to [29]. More investigation on SED performance needs to be conducted in a future work.

V. CONCLUSIONS

We proposed a method for the joint analysis of acoustic scenes and sound events using the semi-supervised approach for SED. In the proposed method, we applied the MTL-based framework including the shared layers with acoustic scenes and sound events, ASC layers, and strong and weak SED layers. The experiments using the TUT Acoustic Scenes 2016/2017 and TUT Sound Events 2016/2017 datasets show that the semi MTL approach with the Kronecker delta functions δ_γ and δ_ζ achieves reasonable performance even with a small number of strongly labeled data and a large number of weakly labeled data. The experimental results corollarily revealed that the intersection-based Fscore and the PSDS metric may not approximately evaluate the SED performance with many FPs.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP20H00613, 22H03639, 23K16908, and ROIS-DS-JOINT (006RP2023) to K. Imoto.

REFERENCES

[1] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Jordi, and X. Serra, "General-purpose tagging of freesound audio with AudioSet labels: Task description, dataset, and baseline," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 69–73, 2018.

[2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *Proc. Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 81–85, 2020.

[3] V. Morfi, R. F. Lachlan, and D. Stowell, "Deep perceptual embeddings for unlabelled animal sound," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 150, no. 1, pp. 2–11, 2020.

[4] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," *Proc. International Joint Conference on Neural Networks* (*IJCNN*), pp. 1547–1554, 2017.

[5] A. Raveh and A. Amar, "Multi-channel audio classification with neural network using scattering transform," *Tech. Rep. DCASE Challenge 2018 Task5*, pp. 1–4, 2018.

[6] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2450–2460, 2020.

[8] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," *Tech. Rep. DCASE Challenge 2020 Task4*, pp. 1–4, 2020.

[9] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 1307–1311, 2011.

[10] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–13, 2013.

[11] K. Imoto and S. Shimauchi, "Acoustic scene analysis based on hierarchical generative model of acoustic event sequence," *IEICE Transactions on Information and Systems*, vol. E99-D, no. 10, pp. 2539–2549, 2016.

[12] K. Imoto and N. Ono, "Acoustic topic model for scene analysis with intermittently missing observations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 2, pp. 367–382, 2019.

[13] H. L. Bear, I. Nolasco, and E. Benetos, "Towards joint sound scene and polyphonic sound event recognition," *INTERSPEECH*, pp. 4594–4598, 2019.

[14] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, "Joint analysis of acoustic events and scenes based on multitask learning," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 333–337, 2019.

[15] J. Jung, H. J. Shim, J. H. Kim, and H. J. Yu, "DCASENET: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 621–625, 2021.

[16] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," *Proc. ACM International Conference on Multimedia* (*ACMMM*), pp. 1038–1047, 2016.

[17] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 253–257, 2019.

[18] X. Zheng, H. Chen, and Y. Song, "Zheng USTC team's submission for DCASE2021 task4 - semi-supervised sound event detection," *Tech. Rep. DCASE Challenge 2021 Task4*, pp. 1–3, 2021.

[19] S. Tsubaki, K. Imoto, and N. Ono, "Joint analysis of acoustic scenes and sound events with weakly labeled data," *Proc. International Workshop on Acoustic Signal Enhancement* (*IWAENC*), pp. 1–5, 2022.

[20] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 19–23, 2018.

[21] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C.

Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 131–135, 2017.

[22] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 621–625, 2020.

[23] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[24] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *Proc. International Conference on Learning Representations* (*ICLR*), pp. 1–13, 2020.

[25] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 1128–1132, 2016.

[26] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 85–92, 2017.

[27] https://www.ksuke.net/dataset.

[28] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 61–65, 2020.

[29] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance," *Applied Acoustics*, vol. 196, no. 108882, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Cross-domain Sound Recognition for Efficient Underwater Data Analysis

Jeongsoo Park*, Dong-Gyun Han†‡§, Hyoung Sul La‡, Sangmin Lee*, Yoonchang Han*, and Eun-Jin Yang‡
* Cochl Inc, USA
E-mail: {jspark, ychan, smlee}@cochl.ai
† Division of Ocean Sciences, Korea Polar Research Institute, Incheon, Republic of Korea
E-mail: dghandg@gmail.com, {hsla, ejyang}@kopri.re.kr
‡ Research Center for Ocean Security Engineering and Technology, Hanyang University ERICA, Republic of Korea
§ Oceansounds Inc, Republic of Korea

***Abstract*—This paper presents a novel deep learning approach for analyzing massive underwater acoustic data by leveraging a model trained on a broad spectrum of non-underwater (aerial) sounds. Recognizing the challenge in labeling vast amounts of underwater data, we propose a two-fold methodology to accelerate this labor-intensive procedure.**

The first part of our approach involves PCA and UMAP visualization of the underwater data using the feature vectors of an aerial sound recognition model. This enables us to cluster the data in a two dimensional space and listen to points within these clusters to understand their defining characteristics. This innovative method simplifies the process of selecting candidate labels for further training.

In the second part, we train a neural network model using both the selected underwater data and the non-underwater dataset. We conducted a quantitative analysis to measure the precision, recall, and F1 score of our model for recognizing airgun sounds, a common type of underwater sound. The F1 score achieved by our model exceeded 84.3%, demonstrating the effectiveness of our approach in analyzing underwater acoustic data.

The methodology presented in this paper holds significant potential to reduce the amount of labor required in underwater data analysis and opens up new possibilities for further research in the field of cross-domain data analysis.

I. Introduction

Underwater acoustics is a complex and vast field with a wide range of applications, such as ocean survey, marine life monitoring, and submarine detection. In general, the underwater sound data is collected via Passive Acoustic Monitoring (PAM), which is a non-invasive technique aiming to record cetacean sounds. However, the analysis of underwater acoustic data is challenging due to the unique characteristics of underwater sound. The underwater sound propagating through the acoustic waveguide interacts with the boundaries (i.e., reflection and scattering) while also undergoing refraction due to variations in the sound speed in the water medium. Additionally, the kinds of animals that reside underwater and the kinds of artificial noise that humans make are different from the aerial noise. As a result, even if we have a massive sound dataset in advance of the underwater sound analysis, it can be difficult to reuse it directly to analyze the underwater data.

Hence, the studies concerning underwater acoustic data have been focusing on making a sound recognizer from the scratch. Many research efforts used conventional rule-based methods such as Erbe and King's method using information theory [1] and Baumgartner *et al.*'s spectrogram filtering-based method [2]. In Yack *et al.*'s study [3], six detection algorithms were tested and compared. In other studies including Halkias *et al.*'s [4] and Thomas *et al.*'s work [5], deep neural network architecture was used. Usman *et al.* summarized the recent research trends and their methods [6].

To train a neural network model that is capable of recognizing underwater sounds, we need an enormous amount of labeled data. However, labeling underwater acoustic data is a labor-intensive task due to the limited available labeled data and the complexity of the underwater environment. Additionally, the people who can distinguish the underwater sound events are very rare because they need to have expertise in the underwater environments and acoustics.

To overcome this problem, several studies tried to reduce the amount of data required to train a model. Both Shiu *et al.* [7] and Webber *et al.* [8] generalized a model trained on a small underwater data to analyze massive database consists of several years of recording. Thomas *et al.* utilized a transfer learning approach to generalize a neural network model to detect other sources [9]. However, none of the previous studies tried to utilize aerial sounds to analyze the underwater data even though they are easy to collect and guarantee diversity of the data.

In this paper, we propose a novel approach to address the challenges of underwater acoustic data analysis. Our methodology is a cross-domain approach, which is about applying state-of-the-art aerial sound recognition AI to underwater data to accelerate the entire analysis process. Our methodology is two-fold, involving the visualization of underwater data using the feature vectors of a sound recognition model trained on non-underwater sounds and quantitative analysis using this same model.

By using a deep learning model trained on a broad array of aerial sounds, we can navigate around the limited availability

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

of labeled underwater data. Furthermore, visualizing and listening to the clustered data allows for efficient labeling and selection of candidate labels for additional training.

Following this approach, we conducted an in-depth analysis of our model's performance with underwater acoustic data, specifically focusing on the recognition of airgun sounds. The results highlight the model's high precision, recall, and F1 score, demonstrating the effectiveness and robustness of our approach. Through this research, we aim to pave the way for more efficient underwater acoustic data analysis, reducing labor-intensive processes, and fostering further innovations in the field.

The rest of this paper is organized as follows. In Section II, we explain our underwater data and how we collected it. Section III covers our method for the visualization of the underwater data. Section IV provides the empirical measures that confirm the success of our model in detecting underwater sounds. In Section V, we conclude the paper with our final thoughts.

II. Underwater Dataset and Collection Method

A. Underwater sound data

The underwater dataset used in this study was collected in the East Siberian Sea, Arctic Ocean, using a hydrophone over one year [10]. The data collection period spanned from August 2017 to August 2018.

The hydrophone was programmed to record 600 seconds of sound every hour, with a sample rate of 32,768Hz, a common rate for high-quality underwater audio capture. Over the one-year period, this method provided a comprehensive audio snapshot of the underwater environment, capturing a broad range of underwater sound events.

B. Airgun labels

Among the various sound events in the dataset, specific labels for airgun sounds were included. Airguns, widely used in marine seismic surveys, produce unique acoustic signatures and are among the more recognizable sound events in underwater acoustic data. Unlike other sounds, such as mammal sounds, airgun sounds are steady in terms of the sound pattern. Hence, collecting labels for this sound is relatively straightforward with the help of automated algorithms.

In particular, we focused on the September 2017 data, since most of the airgun sounds occur in the summer season [11]. For the initial identification of airgun sounds in our dataset, we employed a matched filter algorithm as described in [10] and [12]. Following this semi-automated identification process, an expert in underwater acoustics reviewed the potential airgun events. This step ensured the accuracy of the labels, verifying that the sounds identified by the algorithm were indeed airgun sounds. Additionally, the expert examined if there were any missing airgun sounds.

As a result, the dataset we obtained covered a wide range of underwater sounds and also contained a verified set of labeled airgun sounds, serving as a valuable resource for our deep learning model training and evaluation.

III. Underwater Sound Visualization for Initial Identification

The manual labor involved in the categorization and identification of large, unlabeled underwater acoustic datasets can be overwhelming. This challenge calls for a method that can aid in the initial identification and selection of underwater sound characteristics without relying on extensive manual labeling. This section introduces our innovative approach to visualizing underwater sounds, which streamlines this process and provides a launching point for further analysis.

Our methodology relies on a deep learning model trained on a broad array of aerial sounds. The basis for this approach is the understanding that a variety of aerial sounds can capture a wide range of acoustic features, which can be extended to the underwater environment.

A. Principal component analysis

Our process begins by feeding the underwater acoustic data through our trained sound recognition model. The audio data is segmented into one-second snippets, with no overlap between adjacent snippets, to match our model's input requirement of a 22,050Hz audio snippet lasting one second. These segmented chunks are then fed into the model. For this experiment, we removed the final layer, which functions as the classifier, to obtain the feature vectors of 1,280 dimensions. These contain intricate information about the acoustic properties of the underwater sounds. These feature vectors serve as the coordinates in a high-dimensional space where similar sounds cluster together.

However, visualizing and interpreting high-dimensional data is neither intuitive nor straightforward. Therefore, we employ Principal Component Analysis (PCA), which is a dimensionality reduction technique, to project the high-dimensional feature vectors into a 2D space. In the visualized space, clusters of similar sounds emerge based on their acoustic features, making them discernible to the human eye and ear.

This visualization allows researchers to isolate and listen to individual points or clusters within the projected data. By doing so, we can qualitatively assess the sound characteristics of these clusters and choose suitable candidate labels for the sounds within them.

Fig. 1 shows various visualizations of the underwater sounds. Each subfigure represents the data of different dates and times. It can be seen that the data points spread in different locations on the plane. For example, the data points of **August 21, 2017, 22:00:00** are condensed around (0, 0), as in Fig. 1a. Meanwhile, those of **December 04, 2017, 09:00:00** in Fig. 1c have a higher $PrincipalComponent2$. In the **June 05, 2018, 05:00:00** data, we can observe that the data points are spread across various $PrincipalComponent1$ as in Fig. 1f.

This difference in distribution may be caused by the sound characteristics, which in turn are influenced by the environmental changes due to the changing seasons. Deeper investigation by an expert revealed that $PrincipalComponent1$ is related to the presence of animal sounds such as those of

TABLE I: Sound Types and Characteristics of the data with $PrincipalComponent1 > 40$.

Sound Type	Number of Data	Occurrence Date
Bearded Seal	1033	Mar. or May 2018
Walrus	9	Sept. or Oct. 2017
Airgun	275	Sept. 2017
Sea ice	1	Aug. 25 2017 00:00:00
Whales	12	Sept. 11 or 19 2017
Mammal	7	Sept. 23 2017 17:00:00

bearded seals. Besides, $PrincipalComponent2$ is associated with sea ice concentration and fluid velocity.

For further analysis, we extracted 84,499 one-second long sound data that had a $PrincipalComponent1$ larger than 40. Then, the underwater data expert screened out the ones that contain sound events from those which do not. As a result, we obtained the labeled dataset shown in Table I. It is interesting to observe that most of the sound events are either bearded seals or airguns. Also, each sound is concentrated in a certain season. For example, all the airgun sounds were detected in September 2017, and bearded seal sounds were detected in either March or May 2018.

B. Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection, often referred to as UMAP, is a technique to reduce data dimension [13]. It is often used for high-dimensional data visualization. Even though PCA is a simple approach to visualize data, we can further improve the visualization by applying UMAP and obtain further insights.

Fig. 2a and Fig. 2b show the visualizations using PCA and UMAP, respectively. To reduce computational complexity and enhance graph visibility, we used 25,650 randomly chosen data points, which constitute 0.5% of the entire data in quantity. It is interesting to see that the data points of a season are gathered together in both plots. In the PCA visualization, the sound cluster boundaries are relatively unclear and the data points are scattered. Based on the clear clusters of data points that represent specific types of sounds, we can argue that the UMAP algorithm better visualizes our data.

The sound visualization method drastically simplifies the process of labeling, offering an intuitive, visual, and auditory understanding of the characteristics and distribution of the underwater acoustic data. In the subsequent section, we will outline the process of training our new deep learning model with selected candidate labels and the aerial dataset, followed by a rigorous quantitative analysis of the model's performance.

IV. Quantitative Analysis of Underwater Sound Recognition

A. Model training

This section applies quantitative analysis to underwater data, focusing specifically on the recognition of airgun sounds. In the previous chapter, candidate labels for the underwater sounds were selected based on the results of our visualization and on listening to the clustered data. Bearded seal and airgun labels served as starting points for a more comprehensive training process. They were combined with the aerial sound dataset with the aim of creating a robust model capable of recognizing a variety of underwater sounds. Other underwater sound event labels, such as whales, were discarded due to the small quantity of data available for training.

The trained model employs a Convolutional Neural Network (CNN)-based architecture with an argmax classifier. An audio snippet with a sample rate of 22,050Hz and 22,050 samples in length is fed into the model as input and transformed into a melspectrogram. Because the original data has a sample rate of 32,768Hz, downsampling to 22,050Hz is necessary prior to training.

A significant advantage of our approach is its ability to work with relatively small amounts of labeled underwater data due to the use of the aerial sound dataset. This substantially reduces the burden of manual labeling and allows us to quickly move into the training phase. Cochl's aerial sound dataset, which consists of 586,343 labels across 242 classes, helps the new model learn the feature representation of the sounds. Both aerial and underwater sound data are split into train, validation, and test sets, with the validation and test sets each containing 10% of the data.

B. Internal evaluation

After training, a comprehensive quantitative analysis of the model's performance is conducted, focusing specifically on the recognition of airgun sounds. Precision, recall, and the F1 score are used as the primary metrics for evaluating the performance of our model.

Precision measures the model's ability to correctly identify positive instances out of all instances it classified as positive. Recall, on the other hand, determines the model's ability to identify all positive instances from the total actual positives. The F1 score harmonizes these two metrics, providing an overall assessment of the model's accuracy in recognizing airgun sounds.

In the first part of the evaluation, an internal evaluation is performed using the test split of the dataset. It consists of 28 labels of airgun, 103 labels of bearded seal, and 173,844 labels of the rest of the aerial sound data. Notably, the underwater ambient noise data is not included in the test set.

Interestingly, both the recall and precision of airgun and bearded seal sounds are estimated as 1.0, which in turn causes the F1 scores also to be 1.0. This high performance can be attributed to the fact that the training and test data share the same characteristics because they are randomly split. Also, the exclusion of ambient underwater sound from the dataset may have made the evaluation easier because it is closer to the target sounds than the aerial sounds.

C. External evaluation

The second part of the evaluation uses the airgun labels mentioned in Section II-B. Along with the labeled airgun sounds, the remaining underwater sounds are also used to check for false positives. Preliminarily, labels overlapping with

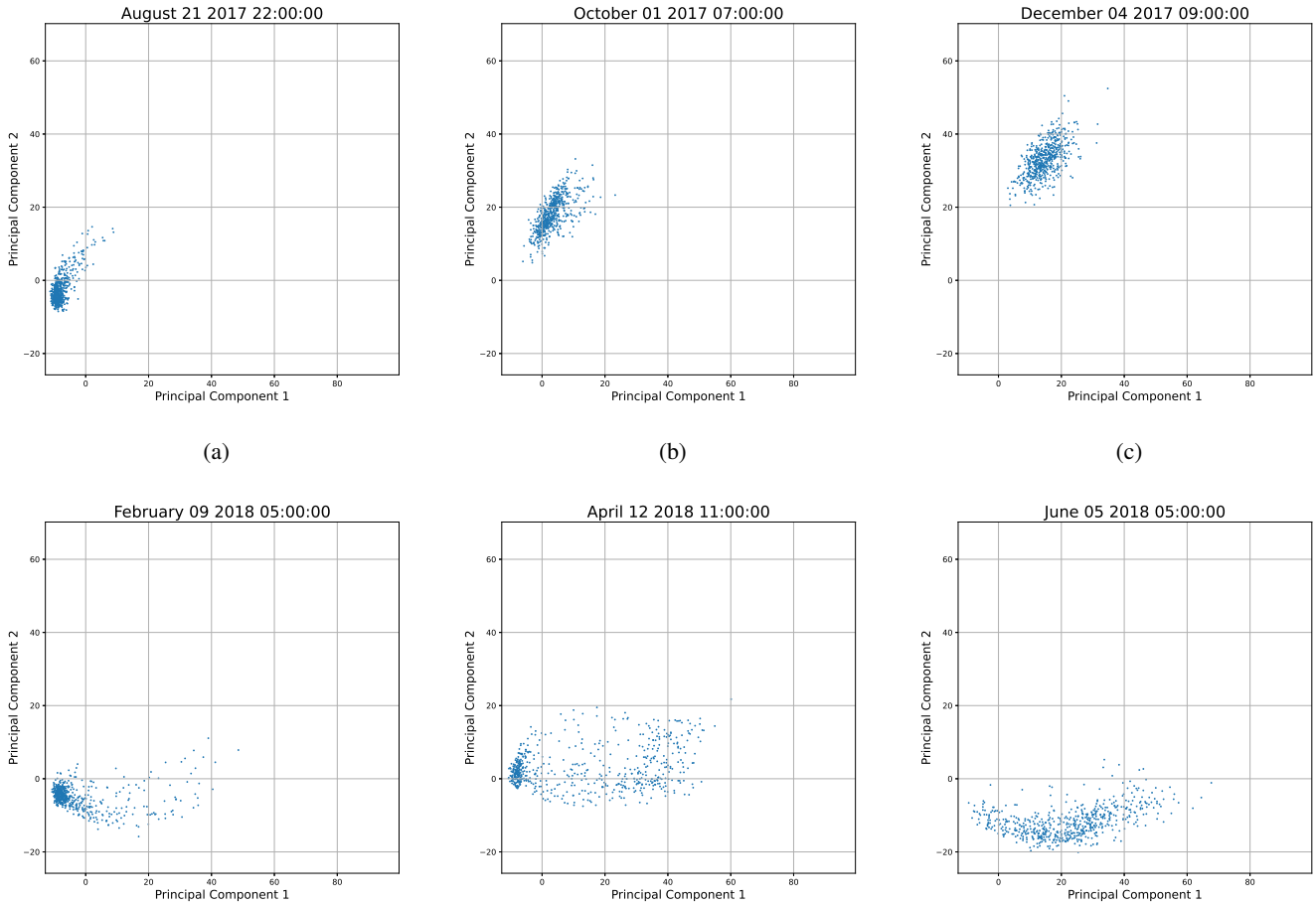


Fig. 1: Visualizing and analyzing the distributions of audio data based on seasonal sound characteristic variations.

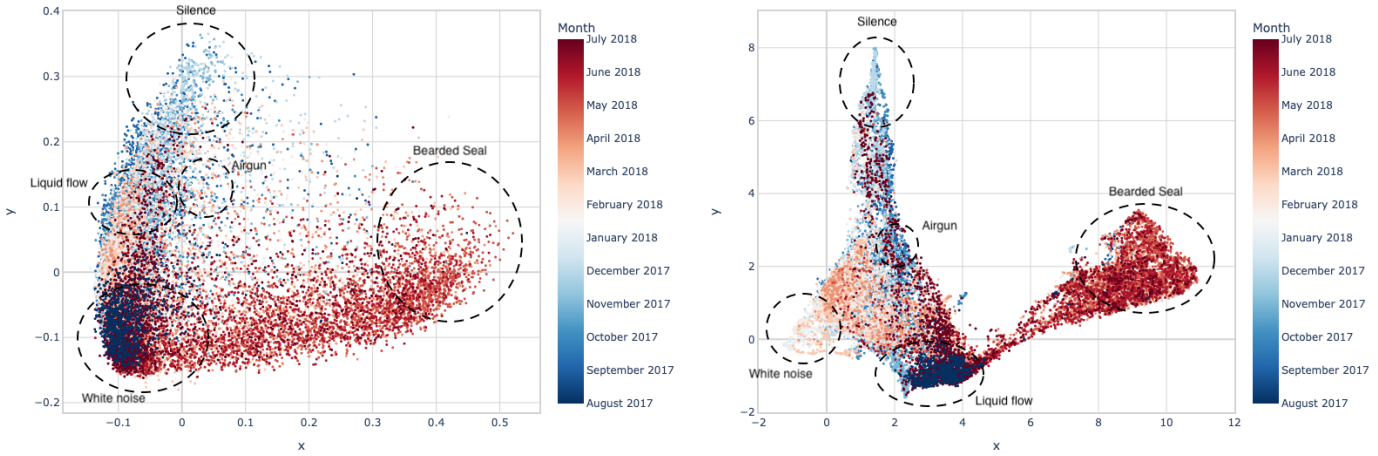
the training data are eliminated to avoid the influence of the training and internal test data on the performance. As a result, the total count of the airgun data is 20,397, whereas that of the ambient underwater data is 259,238. The sound snippets are cut into 1-second chunks without overlap with the neighboring chunk and fed into the model.

We initially used the argmax criteria, as in the training phase. However, we also observed changes in performance according to the threshold level. As the model returns the probability for three classes, the probability of the detected class has to exceed 0.3333. Hence, lowering the threshold with the threshold-based detection criteria increases the recall.

Fig. 3 shows the evaluation results. The precision under the argmax criteria is noticeably high, with the model only detecting 4 false positives out of a total of 259,238 snippets. These misidentified chunks were walrus moan sounds, which may be hard to distinguish from airgun sounds for non-experts. Nevertheless, the model struggled to achieve a high recall rate, potentially attributable to the constrained volume and diversity of the data. The lack of underwater ambient sound data in the training set might have also contributed to this issue. Moreover, the model's sensitivity to the start time of input audio chunks could have complicated results, as airgun sounds are short and can be challenging to identify if split between two consecutive chunks. Using the threshold-based detection criterion, the recall rate escalates as the threshold descends, thereby accommodating more false positives. The optimal F1-score can be achieved at a threshold value of 0.05.

Upon observing the threshold and probability values, we can infer that the large amount of aerial data might have influenced the probability of the airgun sounds. This is particularly relevant considering the argmax layer's characteristic that normalizes the sum of probabilities to 1.0, which could have magnified the effect.

In conclusion, our deep learning model showcases an effective and efficient method for recognizing underwater sounds. This quantitative analysis underscores the value of our approach in streamlining labor-intensive processes in underwater acoustic data analysis, and it emphasizes its potential for future improvements. However, some types of post-processing, such as threshold optimization, may be necessary to maintain a high recall rate.



(a) PCA

(b) UMAP (n_neighbors=10)

Fig. 2: Comparison of PCA and UMAP visualization.

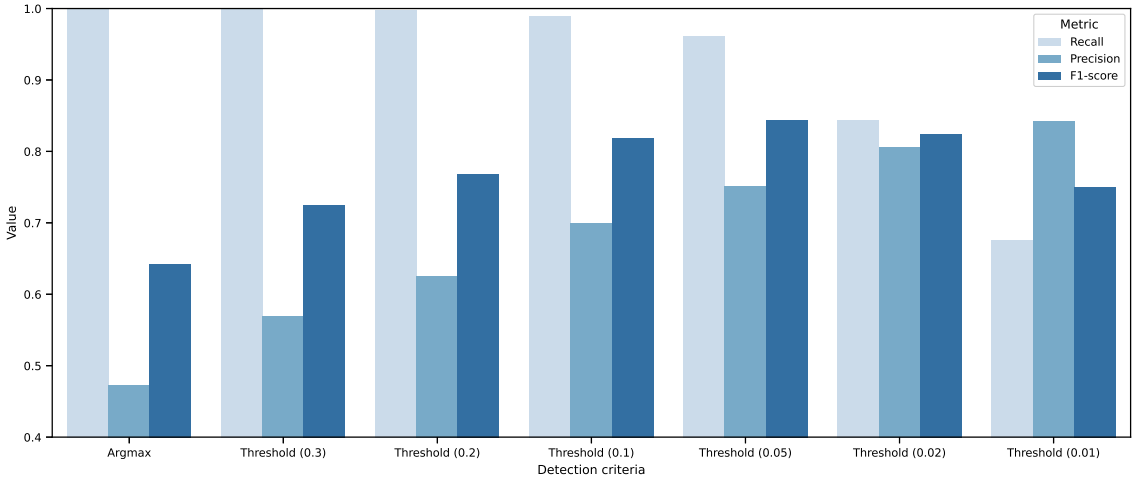


Fig. 3: Performance metrics for airgun detection: Precision, Recall, and F1-score. Optimal F1-score is achieved when the detection threshold is established at 0.05.

V. Conclusion

In this paper, we have presented a novel approach to accelerate underwater sound analysis by leveraging an aerial sound recognition model. Successful visualization and labeling of underwater sound data were achieved, and a sound recognition model that achieves F1 scores of 1.0 in internal evaluation was trained. In external evaluation, although the overall probability of airgun detection was low, we managed to obtain an F1 score of 84.3% when we set the threshold to a lower value. This study is the first to utilize an aerial sound recognition model to accelerate the process of underwater sound analysis. Our approach holds significant potential to improve the efficiency of underwater sound analysis substantially. This efficiency could be utilized to a variety of applications, such as monitoring marine life and detecting underwater threats.

References

[1] C. Erbe, and A. R. King, "Automatic detection of marine mammals using information entropy," The Journal of the Acoustical Society of America, Vol. 124, Issue 5, pp. 2833-2840, 2008.

[2] M. F. Baumgartner, and S. E. Mussoline, "A generalized baleen whale call detection and classification system," The Journal of the Acoustical Society of America, Vol. 129, Issue 5, pp. 2889-2902, 2011.

[3] T. M. Yack, J. Barlow, M. A. Roch, H. Klinck, S. Martin, D. K. Mellinger, and D. Gillespie, "Comparison of beaked whale detection algorithms," Applied Acoustics, Vol. 71, Issue 11, pp. 1043-1049, 2010.

[4] X. C. Halkias, S. Paris, and H. Glotin, "Classification of mysticete sounds using machine learning techniques," The Journal of the Acoustical Society of America, Vol. 134, Issue 5, pp. 3496-3505, 2013.

[5] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, “Marine mammal species classification using convolutional neural networks and a novel acoustic representation,” In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III, pp. 290-305, Springer International Publishing, 2020.
[6] A. M. Usman, O. O. Ogundile, and D. J. Versfeld, “Review of automatic detection and classification techniques for cetacean vocalization,” IEEE Access, Vol. 8, pp. 105181-105206, 2020.
[7] Y. Shiu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, E. M. Nosal, and H. Klinck, “Deep neural networks for automated detection of marine mammal species,” Scientific Reports, Vol. 10, Issue 1, Article 607, 2020.
[8] T. Webber, D. Gillespie, T. Lewis, J. Gordon, T. Ruchirabha, and K. F. Thompson, “Streamlining analysis methods for large acoustic surveys using automatic detectors with operator validation,” Methods in Ecology and Evolution, Vol. 13, Issue 8, pp. 1765-1777, 2022.
[9] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, “Marine mammal species classification using convolutional neural networks and a novel acoustic representation,” In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III, pp. 290-305, Springer International Publishing, 2020.
[10] D. G. HAN, et al. “Effects of geophony and anthrophony on the underwater acoustic environment in the East Siberian Sea, Arctic Ocean,” Geophysical Research Letters, 2021, 48.12: e2021GL093097.
[11] S. B. Blackwell , C. S. Nations, T. L. McDonald, A. M. Thode, D. Mathias, K. H. Kim, C. R. Greene Jr, A. M. Macrander, “Effects of airgun sounds on bowhead whale calling rates: Evidence for two behavioral thresholds,” *PloS one,* vol. 10, no. 6, 2015.
[12] C. W. Clark, P. Marler, and K. Beeman, “Quantitative analysis of animal vocal phonology: an application to swamp sparrow song,” Ethology, Vol. 76, Issue 2, pp. 101-115, 1987.
[13] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” arXiv preprint arXiv:1802.03426 (2018).
[14] C. W. Clark, P. Marler, and K. Beeman, “Quantitative analysis of animal vocal phonology: an application to swamp sparrow song,” Ethology, Vol. 76, Issue 2, pp. 101-115, 1987.
[15] P. J. Dugan, A. N. Rice, I. R. Urazghildiiev, and C. W. Clark, “North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms,” In 2010 IEEE Long Island Systems, Applications and Technology Conference, pp. 1-6, IEEE, 2010.
[16] D. Gillespie, M. Caillat, J. Gordon, and P. White, “Automatic detection and classification of odontocete whistles,” The Journal of the Acoustical Society of America, Vol. 134, Issue 3, pp. 2427-2437, 2013.
[17] D. K. Mellinger, S. W. Martin, R. P. Morrissey, L. Thomas, and J. J. Yosco, “A method for detecting whistles, moans, and other frequency contour sounds,” The Journal of the Acoustical Society of America, Vol. 129, Issue 6, pp. 4055–4061, 2011.
[18] X. Lurton. “An introduction to underwater acoustics: principles and applications,” London: springer, 2002.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Augmentation of Various Speed Data by Controlling Frame Overlap for Acoustic Traffic Monitoring

Tomohiro Takahashi*, Yuma Kinoshita*, Natsuki Ueno*, Yukoh Wakabayashi*, Nobutaka Ono*,
Jun Honda†, Seishi Fukuma†, Aoi Kitamori† and Hiroshi Nakagawa†
*Tokyo Metropolitan University, Tokyo, Japan
†NEXCO-EAST ENGINEERING, Tokyo, Japan

***Abstract*—In this study, we present a data augmentation method for machine-learning-based acoustic traffic monitoring that estimates traffic speed from the running-vehicle sound observed by microphones. Acoustic traffic monitoring is superior to existing traffic detectors using loop coils in terms of installation and maintenance costs, and machine-learning-based acoustic traffic monitoring has demonstrated high performance with sufficient training data. However, it is often difficult or burdensome to collect a large amount of running-vehicle sound with the corresponding labels of traffic speed over many days as training data. In addition, the tendency of the traffic speed changes significantly from day to day, depending highly on the traffic situation, such as congested or free flowing. This may result in biased traffic speed in the training data, particularly the absence of low-speed data. To overcome this problem, we propose a data augmentation method for artificially generating training data with a sufficient range of traffic speed from data with a limited range of traffic speed by controlling frame overlap in the time-frequency domain. In the experimental evaluation, a deep neural network was trained using only high-speed real data with and without the proposed data augmentation method, and the estimation performance characteristics for high-speed and low-speed test data were compared. The results showed that the estimation accuracy for low-speed data was greatly improved with the proposed data augmentation method, while that for high-speed data was kept almost the same, compared with the condition without data augmentation.**

I. Introduction

The measurement of road traffic conditions, such as traffic volume, speed, and time occupancy, is essential to understand traffic situations and provide them to road traffic control systems and users. For this reason, traffic detectors are used for the real-time, automatic, and constant monitoring of traffic conditions [1], [2].

Loop-coil-based detectors are the most commonly used traffic detectors that detect the passage of vehicles electromagnetically. In spite of their high observation accuracy, they require high installation and maintenance costs because they use loop coils embedded in the roadway surface [3]. As a low-cost alternative to loop-coiled-based detectors, acoustic traffic monitoring, which uses the acoustic information obtained by microphones installed at the roadside, has been proposed and investigated in the literature. Unlike vision-based traffic monitoring, it can also cope with dark areas and overlapping vehicles. Approaches to acoustic traffic monitoring are classified roughly into two types: rule-based detection [4]–[8] and machine-learning-based detection [9]–[13]. In our previous study [13], it was demonstrated that our proposed machine-learning-based method using a deep neural network (DNN) with sufficient training data achieved a higher estimation accuracy than the rule-based method [8].

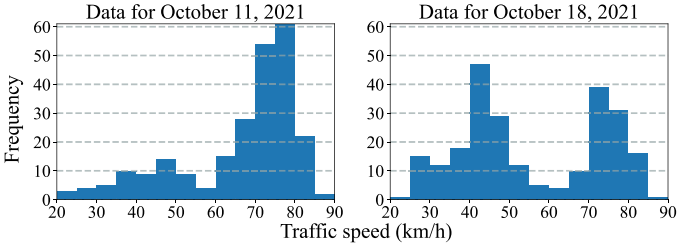


Fig. 1. Histogram of traffic speed for two different days

As in other general tasks in machine learning, a large amount of vehicle-running sound and corresponding labels of traffic conditions must be collected as training data in machine-learning-based acoustic traffic monitoring. In addition, the training data are desired to cover a sufficient variety of traffic situations, e.g., congested and free flowing conditions. However, the tendency of the traffic conditions varies significantly from day to day depending on the road and weather conditions. For example, Fig. 1 shows a histogram of the traffic speed for every minute during the same four hours but on two different days along a certain road. One can see that low-speed data below $60\,\mathrm{km/h}$, corresponding to traffic congestion, are rarely included in the left figure. As shown in this example, the absence of low-speed data is one of the typical difficulties that can occur in collecting training data, because they tend to appear only under somewhat rare weather or road conditions, such as heavy rain, traffic congestion, or regulated traffic. Such biases in the training data may degrade the estimation accuracy, which is indeed confirmed by our experiment described in Section IV.

In this study, we propose a data augmentation method for machine-learning-based acoustic monitoring to improve the estimation accuracy even when a sufficient variety of data are unavailable. To focus on the problem of the absence of low-speed data mentioned above, we here deal with speed estimation when only high-speed data are available as training data, leaving the esimation of other traffic conditions, such as traffic volume and time occupancy, for future work. Data augmentation is one typical strategy to widen the variety of training data, and various data augmentation methods have

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

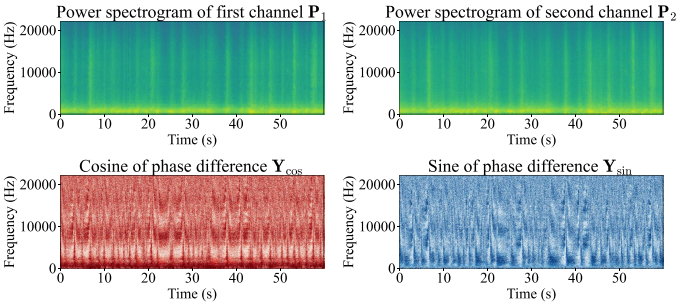


Fig. 2. 2-channel power spectrograms and cosine and sine of phase difference

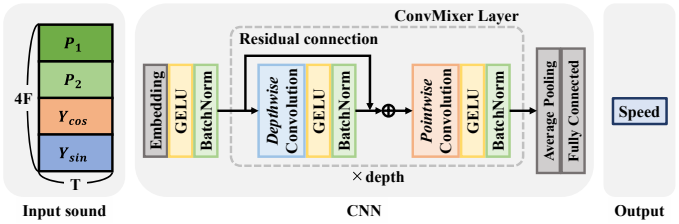


Fig. 3. Network architecture used in acoustic traffic monitoring

been proposed for acoustic data [14]–[22]. Note that these methods are designed mainly to prevent overfitting and are not applicable in the specific problem of our interest, i.e., the augmentation of various speed data in acoustic traffic monitoring. Instead, in the proposed method, low-speed data are generated artificially from high-speed data by controlling frame overlap in the time–frequency domain. The proposed method is evaluated by experiment, where a DNN was trained only on the basis of high-speed data under two conditions: with and without data augmentation. Then, the estimation performance characteristics for low-speed and high-speed data were compared. The experimental results indicate that the proposed data augmentation method improved the estimation accuracy significantly for low-speed test data that were not included in the training data, while maintaining good estimation accuracy for high-speed data.

II. Previous Work on Our Approach

In this section, we introduce our previous work on acoustic traffic monitoring by machine learning. Important points regarding input and output features and network architecture are summarized briefly below.

A. Input and Output Features

The objective of acoustic traffic monitoring is to estimate traffic conditions from acoustic signals. Our acoustic traffic monitoring method uses the stereo running-vehicle sound observed by two microphones aligned parallel to the direction of the road as the input acoustic signal because spatial acoustic information is expected to be useful in the detection of the passage of vehicles. Although four variables, i.e., the total number of vehicles passing, the number of large vehicles passing, traffic speed, and time occupancy, were defined as output features in our previous work [13], we use only traffic speed as the single output feature in this study.

The strict definitions of the input and output features are described below. First, the input acoustic signal is divided into one-minute segments in accordance with the specification of the loop-coil-based traffic detector used as a teacher. Then, the 2-channel power spectrograms $\mathbf{P}_1$ and $\mathbf{P}_2$ and the cosine and sine of the phase differences $\mathbf{Y}_{\cos}$ and $\mathbf{Y}_{\sin}$ (see Fig. 2) are used as input features, which are calculated using the following equations:

$$\mathbf{X}_i = \mathrm{STFT}(\mathbf{x}_i), \tag{1}$$

$$P_i(k,l) = |X_i(k,l)|^2, \tag{2}$$

$$\Delta\phi(k,l) = \arg(X_1(k,l)/X_2(k,l)), \tag{3}$$

$$Y_{\cos}(k,l) = \cos(\Delta\phi(k,l)), \tag{4}$$

$$Y_{\sin}(k,l) = \sin(\Delta\phi(k,l)). \tag{5}$$

Here, $\mathbf{x}_i$ is the ith-channel acoustic signal ($i = 1, 2$), and $\mathrm{STFT} : \mathbb{C}^N \to \mathbb{C}^{F\times T}$ in (1) represents the short-time Fourier transform that transforms an N-sample time signal into a complex-valued spectrogram with F frequency bins and T time frames, where redundant components for negative frequencies are omitted. k and l are indices in the frequency and time directions, respectively, and $X_i(k,l)$ denotes the (k,l) element of $\mathbf{X}_i$ (the same notation applies to other matrices). Finally, the arrays $(\mathbf{P}_1^\top, \mathbf{P}_2^\top, \mathbf{Y}_{\cos}^\top, \mathbf{Y}_{\sin}^\top)^\top$, which are concatenated along the frequency index direction, are used as an input feature. The output feature, i.e., traffic speed, is defined as the space-mean speed for every minute, which corresponds to the harmonic mean of the speed of individual vehicles.

B. Network Architecture

We use ConvMixer [23], [24] as a network architecture of the estimator in our approach. ConvMixer has a repeating network structure consisting of a pointwise convolution, which mixes the features for each time frame using a fully connected layer, and a depthwise convolution, which mixes the features by convolving in the direction of the time frame (see Fig. 3). Patch embedding is applied to input features by handling short time frames in the input as patches. The output of the ConvMixer layer is finally aggregated in the time-frame direction by the average pooling layer and then transformed into a one-dimensional vector representing the traffic speed label by the fully connected layer.

III. Proposed Data Augmentation Method

As mentioned in Section I, it is often difficult to collect training data with a sufficient variety of traffic speed, especially low-speed data, within a limited amount of time. To compensate the lack of data variety, we propose a data augmentation method based on the control of the frame overlap in the time–frequency domain. As previously mentioned, the inputs to our DNN are the power spectrum and phase difference of the stereo-recorded signal. Our idea is to virtually generate their

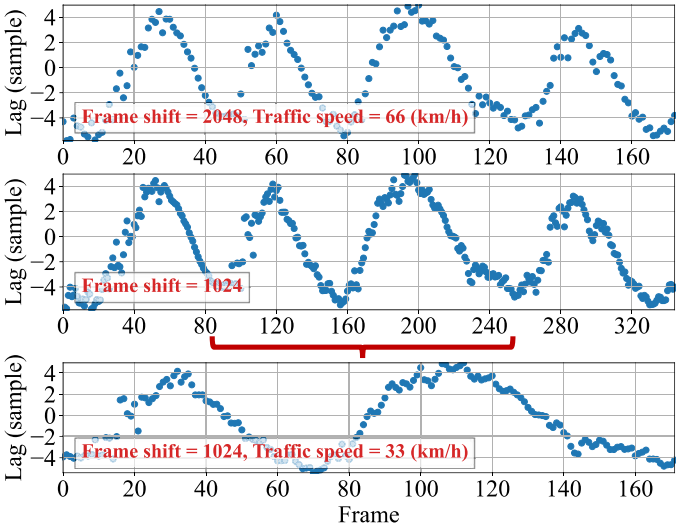


Fig. 4. Conceptual diagram of the proposed data augmentation method

features of slowly-moving vehicles from the original high-speed data by changing the frame shift of STFT.

Let W and M respectively denote the frame length and frameshift of the short-time Fourier transform in (1) for unaugmented data, e.g., $M = W/2$ in the case of half-overlap. Then, the proposed data augmentation algorithm is described as follows. First, a positive integer $M' \leq M$ is determined randomly. Next, a complex-valued spectrogram is obtained from the original acoustic signals $\mathbf{x}_1$ and $\mathbf{x}_2$ via the short-time Fourier transform with M'-sample frameshift instead of M-sample frameshift, and the central T frames, denoted by $\mathbf{X}'_1$ and $\mathbf{X}'_2$, are extracted so that the size of augmented data becomes equal to that of unaugmented data. Finally, input features, i.e., the 2-channel power spectrograms and the cosine and sine of the phase differences, $\mathbf{P}'_1, \mathbf{P}'_2, \mathbf{Y}'_{\cos}, \mathbf{Y}'_{\sin}$ are obtained from $\mathbf{X}'_1$ and $\mathbf{X}'_2$ by (2), (3), (4), and (5). For these input features, the label of the traffic speed is obtained by multiplying M'/M to the original label.

Figure 4 shows the concept of the proposed data augmentation, where the time difference of arrival (TDOA) is plotted for the original data (top), augmented data before T-frame extraction (center), and augmented data after T-frame extraction (bottom). In this example, W, M, and M' are 4096, 2048, and 1024, respectively, and the original label of the traffic speed is $66\,\mathrm{km/h}$; therefore, the artificial label for the augmented data is $33\,\mathrm{km/h}$. One can see that the change in TDOA for the augmented data is stretched twice in the horizontal direction from the original data, as would be expected to occur also for actual vehicles traveling at $33\,\mathrm{km/h}$.

IV. Experiments

To evaluate the proposed method, we investigated the traffic speed estimation performance using a DNN with and without the proposed data augmentation method.

A. Experimental Conditions

The dataset used in this experiment was composed of $48\,\mathrm{h}$ of vehicle-running sound recorded by an IC recorder and



Fig. 5. Recording scenery. A traffic detector is embedded in road pavement in the direction of travel and is not shown in the picture.

corresponding traffic speed labels collected by a loop-coil-based traffic detector near the $6.7\,\mathrm{km}$ post on the Yokohama Shindo in Yokohama City, Kanagawa Prefecture, Japan, which was also used in our previous work [13]. The acoustic signals were recorded with a Panasonic RR-XS470 IC recorder located near the traffic detector (see Fig. 5). The sampling frequency of the IC recorder was $44.1\,\mathrm{kHz}$, and the high-pass filter equipped in the IC recorder was activated during the recording to reduce the effect of wind noise. Four-hour data from 6:00 am to 10:00 am per day were collected over eleven weekdays. The data for the eleven days were first split into seven days for training, two days for validation, and two days for test. Furthermore, the data were split into high-speed data above $60\,\mathrm{km/h}$ and low-speed data below the low-speed data were omitted from the training and validation data to simulate an extreme case in which any low-speed data were not available in training.

We compared the estimation performance characteristics under the following two conditions: (1) without and (2) with proposed data augmentation method. In the absence of data augmenation, only high-speed data were used in the training and validation. For the proposed augmentation method, high-speed data were used with data augmentation in the training and validation, where the frameshift M' was determined independently of each one-minute data and each epoch in the optimization as follows. First, data augmentation was performed with a probability of $1/2$ so that the original data were kept in the training data while the total number of iterations was equal to that in the case without data augmentation to ensure a fair comparison. In the case of performing data augmentation, a random integer p between 10 and 100 was generated from the uniform distribution. Then, the frameshift M' was defined as $M' = \mathrm{round}(pM/100)$, where round denotes rounding to the nearest integer.

Parameters for the ConvMixer and its training were determined as follows. The number of embeddings was set to $T = 1029$. The dimension of feature vectors in the ConvMixer layer was 529, and the kernel size of the depthwise convolution was 5. The number of iterations of the ConvMixer layer was set to 5. The mean absolute error was used for the loss function, and Adam [25] was used for the optimizer. Training was performed for 100 epochs with an initial learning rate of

TABLE I
AVERAGE RMSE ± STD. (km/h) FOR 5 TRIALS

	High speed (Above 60 km/h)	Low speed (Below 60 km/h)
Without data augmentation	3.13±0.13	26±3.85
Proposed	4.12±0.2	**14.7**±1.39
All	3.42±0.11	7.42±0.62

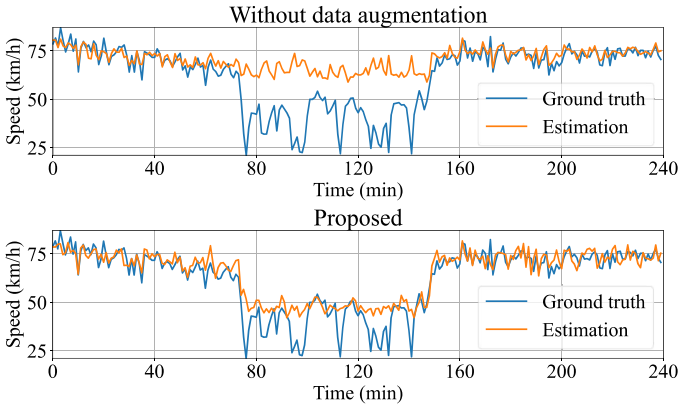


Fig. 6. Temporal changes in true and estimated traffic speeds with and without data augmentation

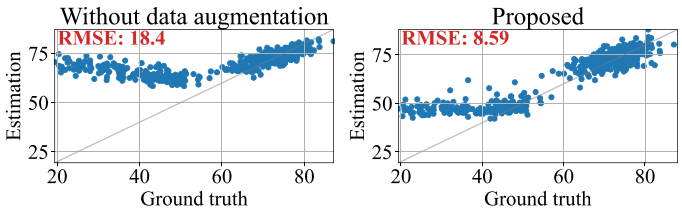


Fig. 7. Scatter plots of true and estimated traffic speeds with and without data augmentation

0.0005.

The estimation performance was evaluated for the test data including both high-speed and low-speed data. The evaluation criterion was the root mean square error (RMSE) between the estimated and true traffic speeds. We additionally evaluated binary classification of low- and high-speed because the characteristics of traffic monitoring do not require perfect estimation for, e.g., detecting congestion.

B. Experimental Results

Table I shows the estimation accuracies obtained with and without the proposed data augmentation method. We also evaluated the estimation accuracy in the case where all data, including both high-speed and low-speed data, were used in the training and validation (without data augmentation) as a reference, denoted by All. The obtained results indicate that the proposed data augmentation method significantly improves the estimation performance for low-speed data compared with the case where no data augmentation is applied. Although the estimation accuracy became slightly degraded for high-speed data, the amount of degradation was much smaller than that of improvement for low-speed data. We also provide the temporal changes in the estimated and true traffic speeds within the designated four hours in Fig. 6 and the scatter plot for all test

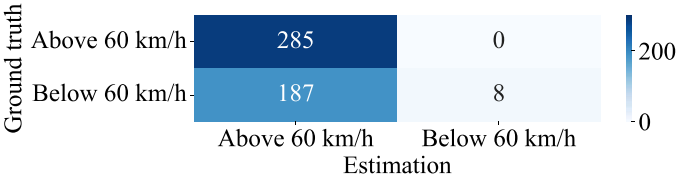


Fig. 8. Confusion matrix in discrimination of high-speed and low-speed data without data augmentation

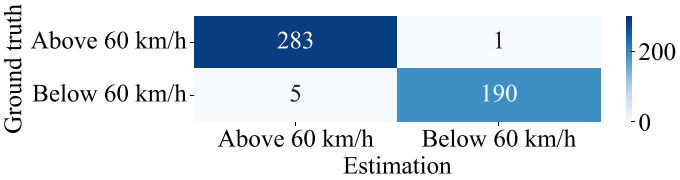


Fig. 9. Confusion matrix in discrimination of high-speed and low-speed data with data augmentation

TABLE II
DISCRIMINATION ACCURACIES FOR HIGH-SPEED AND LOW-SPEED DATA

	Accuracy	**Precision**	**Recall**	**F1-score**
Without data augmentation	61.0 %	100 %	4.1 %	7.9 %
Proposed	98.7 %	99.5 %	97.4 %	98.4 %

data in Fig. 7. From the results in these figures, it is confirmed that the proposed data augmentation method did not simply induce bias on all data toward low speed but indeed improved the estimation accuracy to some extent for the low-speed range while maintaining good accuracy for the high-speed range.

Figures 8 and 9 show the confusion matrices for high-speed (above 60 km/h) and low-speed (below 60 km/h) test data, which represent the total number of data whose true/estimated traffic speed was above/below 60 km/h. In addition, Table II shows the accuracy, precision, recall, and F1-score values calculated from the confusion matrices in Figs. 8 and 9. As shown in these results, the DNN was seriously affected by the bias of traffic speed in the training data without data augmentation; the estimated traffic speed was almost always above 60 km/h for the test data. On the other hand, the proposed data augmentation method enabled the DNN to discriminate high-speed and low-speed data.

V. CONCLUSIONS

We proposed a data augmentation method for machine-learning-based acoustic traffic monitoring, which widens the variety of traffic speed in the training data by controlling frame overlap in the time-frequency domain. The experimental evaluation revealed that the proposed data augmentation method improved the accuracy of traffic speed estimation in the low-speed range, which was not included in the original training data, compared with the case without data augmentation.

Future work includes data augmentation for other features of the traffic condition, such as the number of vehicles.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP20H00613.

References

[1] M. Won, "Intelligent traffic monitoring systems for vehicle classification: A survey," *IEEE Access*, vol. 8, pp. 73 340–73 358, 2020.

[2] P. T. Martin, Y. Feng, and X. Wang, "Detector technology evaluation," Mountain-Plains Consortium, MPC Report 03-154, 2003.

[3] B. Coifman and S. Neelisetty, "Improved speed estimation from single-loop detectors with high truck flow," *Journal of Intelligent Transportation Systems*, vol. 18, no. 2, pp. 138–148, 2014.

[4] M. A. Sobreira-Seoane, A. Rodríguez Molares, and J. L. Alba Castro, "Automatic classification of traffic noise," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3823–3823, 2008.

[5] G. Szwoch and J. Kotus, "Acoustic detector of road vehicles based on sound intensity," *Sensors*, vol. 21, no. 23, 7781 (18 pages), 2021.

[6] J. Kotus and G. Szwoch, "Estimation of average speed of road vehicles by sound intensity analysis," *Sensors*, vol. 21, no. 16, 5337 (18 pages), 2021.

[7] T. Toyoda, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Traffic monitoring with ad-hoc microphone array," in *Proc. 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 318–322.

[8] S. Ishida, K. Mimura, S. Liu, S. Tagashira, and A. Fukuda, "Design of simple vehicle counter using sidewalk microphones," in *Proc. ITS European Congress*, 10 pages, 2016.

[9] A. Y. Nooralahiyan, M. Dougherty, D. McKeown, and H. R. Kirby, "A field trial of acoustic signature analysis for vehicle classification," *Transportation Research Part C: Emerging Technologies*, vol. 5, no. 3–4, pp. 165–177, 1997.

[10] J. George, L. Mary, and K. S. Riyas, "Vehicle detection and classification from acoustic signal using ANN and KNN," in *Proc. International Conference on Control Communication and Computing (ICCC)*, 2013, pp. 436–439.

[11] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram, "Vehicular traffic density state estimation based on cumulative road acoustics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1156–1166, 2012.

[12] T. Shinohara, Y. Wakabayashi, R. Scheibler, N. Ono, N. Aizawa, and H. Nakagawa, "Sound-based speed estimation using a neural network," in *Proc. Spring meeting of the Acoustical Society of Japan*, 2020, pp. 385–386, (in Japanese).

[13] T. Takahashi, Y. Kinoshita, Y. Wakabayashi, *et al.*, "Acoustic traffic monitoring based on deep neural network trained by stereo-recorded sound and sensor data," in *Proc. 31st European Signal Processing Conference (EUSIPCO)*, 2023, (to appear).

[14] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[15] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conference on Artificial Intelligence*, 2020, pp. 13 001–13 008.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. International Conference on Learning Representations*, 2018, 13 pages.

[17] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6022–6031.

[18] G. Kim, D. K. Han, and H. Ko, "SpecMix: A mixed sample data augmentation method for training with time-frequency domain features," in *Proc. Interspeech*, 2021, pp. 546–550.

[19] H. Wang, Y. Zou, and W. Wang, "SpecAugment++: A hidden space data augmentation method for acoustic scene classification," in *Proc. Interspeech*, 2021, pp. 551–555.

[20] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4308–4312.

[21] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3008–3017.

[22] J. Han, M. Matuszewski, O. Sikorski, H. Sung, and H. Cho, "Randmasking augment: A simple and randomized data augmentation for acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 5 pages.

[23] A. Trockman and J. Z. Kolter, "Patches are all you need?" *arXiv preprint arXiv:2201.09792*, 2022.

[24] R. Baidya and H. Jeong, "YOLOv5 with ConvMixer prediction heads for precise object detection in drone imagery," *Sensors*, vol. 22, no. 21, 8424 (17 pages), 2022.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations*, 2015, 15 pages.

Distributed Computation of Heat Kernel Smoothing Using Series Expansion Method

Chien-Cheng Tseng* and Su-Ling Lee
* National Kaohsiung University of Science and Technology, Taiwan
E-mail: tcc@nkust.edu.tw Tel: +886-7-6011000 EXT. 32029
Chang Jung Christian University, Taiwan
E-mail: lilee@mail.cjcu.edu.tw Tel: +886-6-2785123 EXT. 6155

Abstract— **Heat kernel smoothing (HKS) is one of important methods to reduce noise corrupted in the irregular data collected from various complex network. In this paper, the distributed computation of HKS method using series expansion is studied. First, the relation between HKS computation and matrix exponential is presented and the centralized computation of HKS method is derived. Then, two series expansion methods are used to implement the HKS operator in a distributed manner. They include Taylor series expansion (TSE) and Laguerre series expansion (LSE). The TSE method corresponds to a maximally-flat graph filter method, and LSE method corresponds to the weighted-least-squares graph filter method. Next, the realization structures of two series expansion methods are described by choosing the Laplacian matrix as graph shift operator. Finally, the denoising experiments of data collected from sensor network and social network are used to demonstrate the effectiveness of HKS methods in terms of the improvement of signal to noise ratio.**

I. INTRODUCTION

Heat kernel smoothing (HKS) is a mathematical technique that is commonly used in data analysis and signal processing to reduce the noise in a dataset. The technique is based on the idea of heat diffusion, where heat is allowed to flow from hotter areas to cooler areas. HKS method has applications in various fields such as image processing, computer vision, signal processing and brain network etc. In [1], the graph spectral image smoothing using heat kernel is presented. In [2], heat kernels of generalized Laplacians and its application to color image smoothing are studied. In [3], discrete heat kernel smoothing in irregular image domains is investigated and used to filter out noisy data in the lung blood vessel trees obtained from computed tomography. In [4], the investigating heritability across resting state brain networks via heat kernel smoothing on persistence diagrams is provided. In [5], the denoising method of temperature data collected from sensor network is derived by using directed Laplacian matrix and heat kernel smoothing. In [6], the distributed implementation of heat kernel smoothing for graph signal denoising is described. Owing to the success of HKS method in the denoising of irregular dataset, this paper aims to study the efficient distributed computation method of HKS.

On the other hand, series expansion is a method used in mathematics and physics to represent a function as a sum of terms in a series. This is often done to simplify the function or to approximate it with a simpler function. Moreover, series expansion and filters are often used together in signal processing and control systems. Series expansions can be used to approximate the transfer function of a system, which describes the relationship between the input and output signals. In [7], evaluation and design of filters using a Taylor series expansion is presented. In [8], the closed-form design of maximally flat FIR Hilbert transformer, differentiators, and fractional delays by power series expansions is studied. In [9], the series expansion design of variable fractional order integrator and differentiator using logarithm is investigated. In [10], the series expansions in Hermite polynomials and Laguerre polynomials have been studied in science and engineering. In [11], a simple design of fractional delay FIR filter based on Binomial series expansion theory is described. In [12], the Fourier series expansion based on filter parametrization for equivariant convolutions is investigated. Due to the success of series expansion methods in the signal processing area, the purpose of this paper is to use series expansion to develop the distributed computation methods of heat kernel smoothing.

This paper is organized as follows. In section II, the relation between HKS computation and matrix exponential is presented and the centralized computation of HKS method is derived. In section III and IV, two series expansion methods are used to implement the HKS operator including Taylor series expansion

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

and Laguerre series expansion. The realization structures of two series expansion methods are described by choosing the Laplacian matrix as graph shift operator. In section V, the denoising experiments of irregular data collected from sensor network and social network are used to show the effectiveness of HKS methods. Finally, a conclusion is made.

II. HEAT KERNEL SMOOTHING

In this section, the basics of heat kernel smoothing (HKS) which is based on graph signal processing (GSP) theory in [13]-[15] are described. Let $G=(V,E,\boldsymbol{A})$ be an undirected weighted graph, where $V=\{v_1,v_2,\cdots,v_N\}$ is a set of vertices, $E\subseteq V\times V$ is a set of edge and matrix $\boldsymbol{A}$ is the weighted adjacency matrix. If we denote $\overline{v_nv_m}$ as the edge between vertices v_n and v_m, then elements of matrix $\boldsymbol{A}$ is defined by

$$A(n,m)=\begin{cases}a_{n,m} & \overline{v_nv_m}\in E\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

where $a_{n,m}>0$, i.e., vertices are connected with positive edge weights. For an undirected graph, the $\boldsymbol{A}$ is a symmetric matrix. Moreover, let the degree of vertex v_n is defined as

$$d_n=\sum_{m=1}^{N}a_{n,m} \tag{2}$$

then degree matrix $\boldsymbol{D}$ of graph is given by the diagonal matrix $\boldsymbol{D}=diag[d_1,d_2,\cdots,d_N]$. Based on the above definitions, the Laplacian matrix of graph G is defined as $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$ which is a positive semi-definite operator. Let the eigen-decomposition of $\boldsymbol{L}$ be given by $\boldsymbol{L}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda}$ is an diagonal matrix containing the eigenvalues $\lambda_1\le\lambda_2\le\cdots\le\lambda_N$ and orthonomal matrix $\boldsymbol{U}$ is composed of the eigenvectors $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_N$. If $z(n)$ is the signal value associated with the vertex v_n, the graph signal is the vector $\boldsymbol{z}=[z(1),z(2),\cdots,z(N)]^T$ defined on the vertices V. The graph Fourier transform (GFT) of signal $\boldsymbol{z}$ is defined as $\hat{\boldsymbol{z}}=\boldsymbol{U}^T\boldsymbol{z}$ and its corresponding inverse graph Fourier transform (IGFT) is given by $\boldsymbol{z}=\boldsymbol{U}\hat{\boldsymbol{z}}$, where spectrum $\hat{\boldsymbol{z}}=[\hat{z}(\lambda_1),\hat{z}(\lambda_2),\cdots,\hat{z}(\lambda_N)]^T$. After reviewing the GSP basics, the heat kernel smoothing (HKS) method is described below. The well-known heat equation is given by

$$\frac{\partial\phi}{\partial t}=\nabla^2\phi \tag{3}$$

where ϕ is the temperature function. Replacing operator ∇^2 by $-\boldsymbol{L}$ and function ϕ by graph signal $\boldsymbol{z}$, the discrete heat equation over a graph is then given by

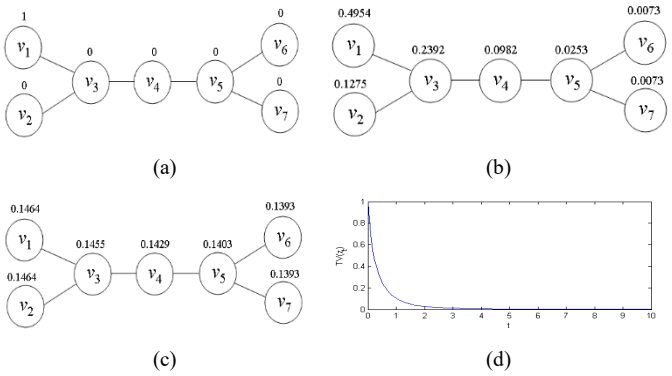


Fig.1 An example of heat kernel smoothing or diffusion. (a) The topology of a simple graph with 7 nodes. The initial signal is $\boldsymbol{z}_0=[1,0,0,0,0,0,0]^T$. (b) The graph signal $\boldsymbol{z}_t$ at $t=1$. (c) The graph signal $\boldsymbol{z}_t$ at $t=15$. (d) The relation between total variation $TV(\boldsymbol{z}_t)$ and t.

$$\frac{\partial\boldsymbol{z}}{\partial t}=-\boldsymbol{L}\boldsymbol{z}\quad\text{with initial}\quad\boldsymbol{z}_0 \tag{4}$$

The solution of this differential equation is given by

$$\boldsymbol{z}_t=e^{-t\boldsymbol{L}}\boldsymbol{z}_0=\boldsymbol{H}_d\boldsymbol{z}_0 \tag{5}$$

where $\boldsymbol{H}_d=e^{-t\boldsymbol{L}}$ is matrix exponential. Clearly, the diffusion process of heat energy is determined completely by the initial value $\boldsymbol{z}_0$. The matrix exponential $\boldsymbol{H}_d=e^{-t\boldsymbol{L}}$ can be computed by using the eigen-decomposition of $\boldsymbol{L}$, that is, substituting $\boldsymbol{L}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ into $\boldsymbol{H}_d$, we have

$$\boldsymbol{H}_d=\boldsymbol{U}\boldsymbol{\Lambda}_d\boldsymbol{U}^T \tag{6}$$

where $\boldsymbol{\Lambda}_d=diag[h_d(\lambda_1),h_d(\lambda_2),\cdots,h_d(\lambda_N)]$ with ideal spectral response $h_d(\lambda)=e^{-t\lambda}$. Now, one example is illustrated. Fig.1(a) shows the graph topology with $N=7$ vertices. The edge weights are all equal to unity. The initial graph signal is chosen as $\boldsymbol{z}_0=[1,0,0,0,0,0,0]^T$. Fig.1(b) and (c) depicts the graph signals $\boldsymbol{z}_t$ for $t=1$ and $t=15$. Clearly, the heat energy flows from vertex v_1 to all other vertices $v_2,v_3,\cdots,v_7$ gradually until signal values at all vertices are the same. Moreover, the total variation (TV) that measures the smoothness of signal $\boldsymbol{z}_t$ is defined by

$$TV(\boldsymbol{z}_t)=\boldsymbol{z}_t^T\boldsymbol{L}\boldsymbol{z}_t \tag{7}$$

Fig.1(d) shows the relation between $TV(\boldsymbol{z}_t)$ and t. It can be seen that the larger t is, the smaller TV signal has. That is, the larger t is, the smoother signal $\boldsymbol{z}_t$ is.

Finally, two remarks are made. First, if $\boldsymbol{z}_t=\boldsymbol{H}_d\boldsymbol{z}_0$ in (5) is used to compute smoothed signal $\boldsymbol{z}_t$, the matrix-vector multiplication is needed. Thus, this method is only suitable for

centralized computation and has high communication cost in the computation process. Second, the eigen-decomposition of $\boldsymbol{L}$ is needed to compute the matrix exponential $\boldsymbol{H}_d = e^{-t\boldsymbol{L}}$. Because the computation load of eigen-decomposition of a matrix is high, the eigen-decomposition free method is wanted. In next sections, the series expansion methods are presented to compute $\boldsymbol{z}_t$ from the initial $\boldsymbol{z}_0$ such that the distributed computation can be achieved and the computation of eigen-decomposition of matrix $\boldsymbol{L}$ can be also avoided.

III. TAYLOR SERIES EXPANSION METHOD

The Taylor series expansion (TSE) of exponential function is given by

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{8}$$

Replacing x with $-t\boldsymbol{L}$, the matrix exponential $\boldsymbol{H}_d = e^{-t\boldsymbol{L}}$ can be expressed as

$$\boldsymbol{H}_d = e^{-t\boldsymbol{L}} = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} \boldsymbol{L}^k \tag{9}$$

Truncating this infinite series to a finite series with truncation order M, it yields the result

$$\boldsymbol{H}_d \approx \sum_{k=0}^{M} \frac{(-t)^k}{k!} \boldsymbol{L}^k \tag{10}$$

Let the coefficients $b_k = \frac{(-t)^k}{k!}$ be chosen, then the following polynomial graph filter $\boldsymbol{H}_1$ will approximate $\boldsymbol{H}_d$ well:

$$\boldsymbol{H}_1 = \sum_{k=0}^{M} b_k \boldsymbol{L}^k \tag{11}$$

Substituting the eigen-decomposition $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ into filter $\boldsymbol{H}_1$ in (11), we have

$$\boldsymbol{H}_1 = \sum_{k=0}^{M} b_k (\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T)^k = \boldsymbol{U}\boldsymbol{\Lambda}_1\boldsymbol{U}^T \tag{12}$$

where diagonal matrix $\boldsymbol{\Lambda}_1 = diag[h_1(\lambda_1), h_1(\lambda_2), \cdots, h_1(\lambda_N)]$ with filter spectral response

$$h_1(\lambda) = \sum_{k=0}^{M} b_k \lambda^k \tag{13}$$

Now, let us show that the filter spectral response $h_1(\lambda)$ satisfies the following property:

Fact 1: The spectral responses $h_1(\lambda)$ and $h_d(\lambda) = e^{-t\lambda}$ have the same m-th order derivatives at $\lambda = 0$ for $m = 0,1,\cdots,M$, that is, the following equalities hold:

$$\left.\frac{d^m h_1(\lambda)}{d\lambda^m}\right|_{\lambda=0} = \left.\frac{d^m h_d(\lambda)}{d\lambda^m}\right|_{\lambda=0} \qquad m = 0,1,\cdots,M \tag{14}$$

Proof: Because $h_d(\lambda) = e^{-t\lambda}$, the derivatives of $h_d(\lambda)$ at $\lambda = 0$ can be computed as

$$\left.\frac{d^m h_d(\lambda)}{d\lambda^m}\right|_{\lambda=0} = (-t)^m e^{-t\lambda}\Big|_{\lambda=0} = (-t)^m \tag{15}$$

Moreover, the derivatives of $h_1(\lambda)$ can be calculated as

$$\left.\frac{d^m h_1(\lambda)}{d\lambda^m}\right|_{\lambda=0} = m!b_m = m!\frac{(-t)^m}{m!} = (-t)^m \tag{16}$$

From the above results in (15) and (16), it can be seen that equation (14) is valid. The proof is completed.

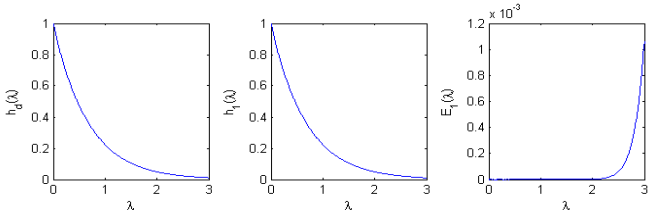
Based on the fact 1, it is clear that the larger order M is, the responses $h_1(\lambda)$ and $h_d(\lambda)$ have more same derivatives at frequency $\lambda = 0$. Therefore, the approximation error $E_1(\lambda) = |h_d(\lambda) - h_1(\lambda)|$ decreases when order M increases. Moreover, fact 1 tells us that polynomial graph filter $\boldsymbol{H}_1$ is a maximally flat approximation of HKS filter $\boldsymbol{H}_d = e^{-t\boldsymbol{L}}$ at $\lambda = 0$. When $t = 1.5$ and $M = 15$ are chosen, Fig.2(a)-(c) show the spectral responses $h_d(\lambda)$, $h_1(\lambda)$ and error $E_1(\lambda)$. Clearly, the errors around $\lambda = 0$ are very small, so the Taylor series expansion method corresponds to a maximally flat graph filter design.

Next, the distributed computation of HKS using TSE is investigated. Because $\boldsymbol{H}_d \approx \boldsymbol{H}_1$ holds, the HKS equation in (5) can be rewritten as

$$\boldsymbol{z}_t \approx \boldsymbol{H}_1\boldsymbol{z}_0 = \left(\sum_{k=0}^{M} b_k \boldsymbol{L}^k\right)\boldsymbol{z}_0 \tag{17}$$

Let us define $\boldsymbol{z}_k = \boldsymbol{L}^k\boldsymbol{z}_0$, then $\boldsymbol{z}_k$ satisfies the recursive equation $\boldsymbol{z}_{k+1} = \boldsymbol{L}\boldsymbol{z}_k$ and (17) reduces to

$$\boldsymbol{z}_t = \sum_{k=0}^{M} b_k \boldsymbol{z}_k \tag{18}$$



(a) (b) (c)

Fig.2 The spectral responses and error of TSE method. (a) Ideal response $h_d(\lambda)$. (b) The actual response $h_1(\lambda)$. (c) The error $E_1(\lambda)$.

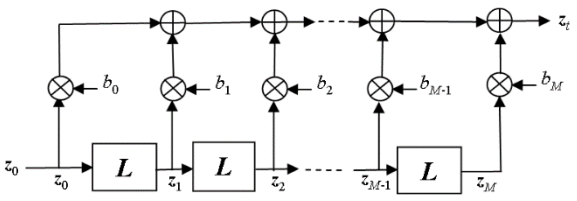


Fig.3 An implementation structure of Taylor series expansion (TSE) method.

Based on the above results, an implementation structure of TSE method is shown in Fig.3. Using $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ and $\boldsymbol{z}_{k+1} = \boldsymbol{L}\boldsymbol{z}_k$, the signal $\boldsymbol{z}_{k+1}$ can be calculated from $\boldsymbol{z}_k$ by

$$z_{k+1}(n) = d_n z_k(n) - \sum_{m \in \mathrm{N}_n} a_{n,m} z_k(m) \tag{19}$$

where notation N_n of neighbor denotes the set of vertices connected to the vertex v_n. When the signal z_1, z_2, ..., z_M are computed in a distributed manner by (19), the HKS signal z_t is easily obtained by the weighted sum in (18). Thus, the distributed computation of $\boldsymbol{z}_t = \boldsymbol{H}_d \boldsymbol{z}_0$ can be accomplished by the TSE structure in Fig.3.

IV. LAGUERRE SERIES EXPANSION METHOD

In this section, the Laguerre series expansion (LSE) method is studied. The k-th order Laguerre polynomials $p_k(x)$ are the solutions of the following differential equation:

$$x\frac{d^2 p_k(x)}{dx^2} + (1-x)\frac{dp_k(x)}{dx} + kp_k(x) = 0 \tag{20}$$

The closed-form solution of $p_k(x)$ is given by

$$p_k(x) = \sum_{r=0}^{k} (-1)^r \frac{k!}{(r!)^2 (k-r)!} x^r \tag{21}$$

Some typical solutions of $p_k(x)$ are given by $p_0(x) = 1$, $p_1(x) = 1 - x$, and $p_2(x) = \frac{1}{2}(2 - 4x + x^2)$. In the literature, it has been shown that $p_k(x)$ satisfies the following properties:
(i) The $p_k(x)$ can be obtained from the Rodriguez formula:

$$p_k(x) = \frac{e^x}{k!}\frac{d^k(x^k e^{-x})}{dx^k} \tag{22}$$

(ii) The $p_k(x)$ satisfies the following recursion relation:

$$p_{k+1}(x) = \frac{(2k+1-x)}{k+1} p_k(x) - \frac{k}{k+1} p_{k-1}(x) \tag{23}$$

with initial $p_0(x) = 1$ and $p_1(x) = 1 - x$.
(iii) The $p_k(x)$ satisfies the orthogonal relation below:

$$\int_0^\infty e^{-x} p_k(x) p_m(x) dx = \begin{cases} 1 & k = m \\ 0 & k \neq m \end{cases} \tag{24}$$

Based on the above orthogonal property, we have the fact:
Fact 2: Let $f(x)$ be a sufficiently smooth real function in the interval $[0,\infty)$, then it can be expanded in terms of the Laguerre polynomials $p_k(x)$ as

$$f(x) = \sum_{k=0}^{\infty} \alpha_k p_k(x) \tag{25}$$

If the α_k are sought to minimize the following error

$$J_1 = \int_0^\infty e^{-x} \left[f(x) - \sum_{k=0}^{\infty} \alpha_k p_k(x) \right]^2 dx \tag{26}$$

then the coefficients α_k are given by

$$\alpha_k = \int_0^\infty e^{-x} f(x) p_k(x) dx \tag{27}$$

The proof of this fact can be found in [10]. Now, let us choose the function $f(x)$ as the exponential function below:

$$f(x) = e^{-\beta x} \qquad \beta > 0 \tag{28}$$

then it can be shown that α_k are given by

$$\alpha_k = \frac{\beta^k}{(\beta+1)^{k+1}} \tag{29}$$

Thus, we have the following Laguerre series expansion:

$$f(x) = e^{-\beta x} = \sum_{k=0}^{\infty} \frac{\beta^k}{(\beta+1)^{k+1}} p_k(x) \tag{30}$$

After changing the variables $\beta \to t$ and $x \to \lambda$, the HKS spectral response $h_d(\lambda) = e^{-t\lambda}$ can be written as

$$h_d(\lambda) = e^{-t\lambda} = \sum_{k=0}^{\infty} \frac{t^k}{(t+1)^{k+1}} p_k(\lambda) \tag{31}$$

Based on this LSE result, the matrix exponential $\boldsymbol{H}_d = e^{-t\boldsymbol{L}}$ can be approximated by the polynomial graph filter:

$$\boldsymbol{H}_2 = \sum_{k=0}^{M} \frac{t^k}{(t+1)^{k+1}} p_k(\boldsymbol{L}) = \sum_{k=0}^{M} c_k p_k(\boldsymbol{L}) \tag{32}$$

The above coefficients c_k are given by

$$c_k = \frac{t^k}{(t+1)^{k+1}} \tag{33}$$

Substituting the eigen-decomposition $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ into graph filter $\boldsymbol{H}_2$ in (32), it yields the result

$$\boldsymbol{H}_2 = \sum_{k=0}^{M} c_k p_k(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T) = \boldsymbol{U}\boldsymbol{\Lambda}_2\boldsymbol{U}^T \tag{34}$$

where diagonal matrix $\boldsymbol{\Lambda}_2 = diag[h_2(\lambda_1), h_2(\lambda_2), \cdots, h_2(\lambda_N)]$ with filter spectral response

$$h_2(\lambda)=\sum_{k=0}^{M}c_k p_k(\lambda) \tag{35}$$

Based on error function in (26), it means that the c_k are obtained by minimizing the following spectral response error

$$\begin{aligned} J_2 &= \int_0^{\infty} e^{-\lambda}\left[h_d(\lambda)-\sum_{k=0}^{M}c_k p_k(\lambda)\right]^2 d\lambda \\ &= \int_0^{\infty} e^{-\lambda}\left[h_d(\lambda)-h_2(\lambda)\right]^2 d\lambda \end{aligned} \tag{36}$$

Therefore, the Laguerre series expansion method corresponds to a weighted-least-squares (WLS) graph filter design method. When $t=1.5$ and $M=15$ are chosen, Fig.4(a)-(c) depict the spectral responses $h_d(\lambda)$, $h_2(\lambda)$ and response error $E_2(\lambda)=|h_d(\lambda)-h_2(\lambda)|$. Clearly, the errors are small in the frequency band $\lambda\in[0,3]$. Next, the distributed computation of HKS using LSE is investigated. Let input signal $\boldsymbol{z}_0$ pass through the graph filter $\boldsymbol{H}_2$ in (32), its output signal $\boldsymbol{z}_t$ is given by

$$\boldsymbol{z}_t \approx \boldsymbol{H}_2\boldsymbol{z}_0=\left(\sum_{k=0}^{M}c_k p_k(\boldsymbol{L})\right)\boldsymbol{z}_0 \tag{37}$$

Using (23), the following recurrence relation can be used to get an efficient implementation structure:

$$p_{k+1}(\boldsymbol{L})=\frac{(2k+1)\boldsymbol{I}-\boldsymbol{L}}{k+1}p_k(\boldsymbol{L})-\frac{k}{k+1}p_{k-1}(\boldsymbol{L}) \tag{38}$$

where $\boldsymbol{I}$ is the identity matrix, initial value $p_0(\boldsymbol{L})=\boldsymbol{I}$ and $p_1(\boldsymbol{L})=\boldsymbol{I}-\boldsymbol{L}$. Multiplying both sides of (38) by the $\boldsymbol{z}_0$ and defining $\boldsymbol{y}_k=p_k(\boldsymbol{L})\boldsymbol{z}_0$, we have

$$\boldsymbol{y}_{k+1}=\frac{2k+1}{k+1}\boldsymbol{y}_k-\frac{1}{k+1}\boldsymbol{L}\boldsymbol{y}_k-\frac{k}{k+1}\boldsymbol{y}_{k-1} \tag{39}$$

with $\boldsymbol{y}_0=\boldsymbol{z}_0$ and $\boldsymbol{y}_1=\boldsymbol{z}_0-\boldsymbol{L}\boldsymbol{z}_0$. Based on equation (39), an implementation structure is shown in Fig.5. Clearly, the key step in this structure is the computation of the signal $\frac{1}{k+1}\boldsymbol{L}\boldsymbol{y}_k$. Using $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$, signal $\frac{1}{k+1}\boldsymbol{L}\boldsymbol{y}_k$ can be computed by

$$\left(\tfrac{1}{k+1}\boldsymbol{L}\boldsymbol{y}_k\right)(n)=\frac{1}{k+1}\left(d_n\boldsymbol{y}_k(n)-\sum_{m\in \mathrm{N}_n}a_{n,m}\boldsymbol{y}_k(m)\right) \tag{40}$$

So, the signals $\boldsymbol{y}_2$, $\boldsymbol{y}_3$, ..., $\boldsymbol{y}_M$ are easily computed in a distributed manner by (39) and (40). The above notation N_n of neighbor denotes the set of vertices connected to the vertex v_n. Using (37) and $\boldsymbol{y}_k=p_k(\boldsymbol{L})\boldsymbol{z}_0$, the HKS signal $\boldsymbol{z}_t$ can be obtained by the following weighted sum

$$\boldsymbol{z}_t=\sum_{k=0}^{M}c_k\boldsymbol{y}_k \tag{41}$$

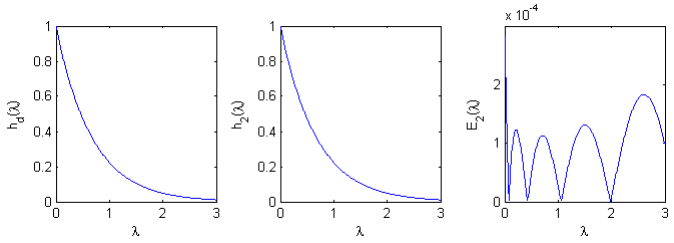


Fig.4 The spectral responses and error of LSE method. (a) Ideal response $h_d(\lambda)$. (b) The actual response $h_2(\lambda)$. (c) The error $E_2(\lambda)$.

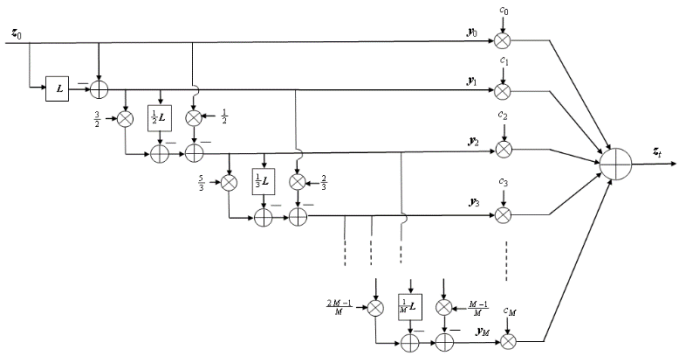


Fig.5 A distributed implementation structure of Laguerre series expansion (LSE) method.

Therefore, the distributed computation of $\boldsymbol{z}_t=\boldsymbol{H}_d\boldsymbol{z}_0$ can be accomplished by the LSE structure in Fig.5.

V. APPLICATION EXAMPLES

In this section, the denoising experiments of data collected from sensor network and social network are used to show the effectiveness of heat kernel smoothing method. To achieve this purpose, the initial signal $\boldsymbol{z}_0$ is chosen as the given noisy signal below:

$$\boldsymbol{z}_0=\boldsymbol{s}+\boldsymbol{\varepsilon} \tag{42}$$

where $\boldsymbol{s}=[s(1),s(2),\cdots,s(N)]^T$ is the clean data and $\boldsymbol{\varepsilon}=[\varepsilon(1),\varepsilon(2),\cdots,\varepsilon(N)]^T$ is the un-correlated white noise. Because the clean signal $\boldsymbol{s}$ is usually a low-frequency component and noise $\boldsymbol{\varepsilon}$ is a high-frequency component, the low-pass heat kernel smoothing signal $\boldsymbol{z}_t=e^{-t\boldsymbol{L}}\boldsymbol{z}_0=\boldsymbol{H}_d\boldsymbol{z}_0$ is the desirable denoised signal. The reason is explained below: From the spectral responses $h_d(\lambda)$ in Fig.2(a) and Fig.4(a), it can be seen that the smoothing filter $\boldsymbol{H}_d$ is a low-pass filter. The response $h_d(\lambda)$ at $\lambda=0$ is exactly equal to one and $h_d(\lambda)$ is a monotonically decreasing function with respect to

the frequency variable λ. The 3-dB cutoff frequency λ_c of $h_d(\lambda)$ needs to satisfy the following condition:

$$h_d(\lambda_c) = e^{-\lambda_c t} = \frac{1}{\sqrt{2}} \tag{43}$$

Solving the above equation, the value of λ_c is given by

$$\lambda_c = \frac{\ln\sqrt{2}}{t} \tag{44}$$

Thus, the variable t can be used to control the 3-dB bandwidth of low-pass smoothing filter $\boldsymbol{H}_d$. The larger t is, the smaller passband width filter has. Moreover, to make comparison, the denoised signal $\boldsymbol{s}_r$ of conventional matrix inversion method is given by

$$\boldsymbol{s}_r = (\boldsymbol{I} + \rho\boldsymbol{L})^{-1}\boldsymbol{z}_0 = \boldsymbol{H}_c\boldsymbol{z}_0 \tag{45}$$

where $\boldsymbol{H}_c = (\boldsymbol{I} + \rho\boldsymbol{L})^{-1}$, $\boldsymbol{I}$ is the identity matrix, and ρ is a prescribed number. The spectral response of the filter $\boldsymbol{H}_c$ is given by

$$h_c(\lambda) = \frac{1}{1+\rho\lambda} \tag{46}$$

Obviously, $h_c(\lambda)$ is also a low-pass filter. To evaluate the performance of the denoised methods, the signal to noise ratio (SNR) of noisy and denoised signals are defined by

$$SNR_i = 10\log_{10}\left(\frac{\|\boldsymbol{s}\|_2^2}{\|\boldsymbol{z}_0 - \boldsymbol{s}\|_2^2}\right) \quad SNR_o = 10\log_{10}\left(\frac{\|\boldsymbol{s}\|_2^2}{\|\boldsymbol{s}_r - \boldsymbol{s}\|_2^2}\right) \tag{47}$$

where $\boldsymbol{s}_r$ is the denoised signal in the following four cases:

(a) In case 1, $\boldsymbol{s}_r$ is the denoised signal of the conventional matrix inversion (MI) method, i.e., $\boldsymbol{s}_r = \boldsymbol{H}_c\boldsymbol{z}_0$. Because matrix inverse needs to be computed, this method is only suitable for centralized computation.

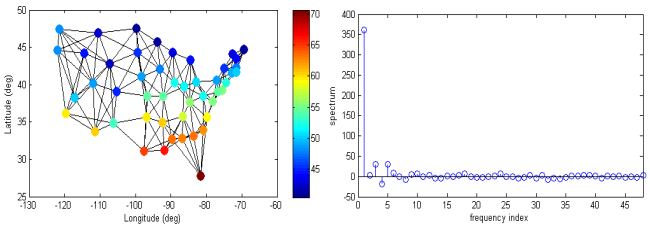
(b) In case 2, $\boldsymbol{s}_r$ is the denoised signal of the original heat kernel smoothing (HKS) method, i.e., $\boldsymbol{s}_r = \boldsymbol{H}_d\boldsymbol{z}_0$. Because eigen-decomposition of matrix needs to be computed, this method is only suitable for centralized computation.

(c) In case 3, $\boldsymbol{s}_r$ is the denoised signal of the Taylor series expansion (TSE) method, i.e., $\boldsymbol{s}_r = \boldsymbol{H}_1\boldsymbol{z}_0$. Because this method can be implemented by the structure in Fig.3, it is suitable for distributed computation.

(d) In case 4, $\boldsymbol{s}_r$ is the denoised signal of the Laguerre series expansion (LSE) method, i.e., $\boldsymbol{s}_r = \boldsymbol{H}_2\boldsymbol{z}_0$. Because this method can be implemented by the structure in Fig.5, it is suitable for distributed computation.

For these four cases, the larger SNR_o is, the better performance the denoising method has. In the following, the numerical examples of sensor network and social network are presented to study the performance.

Example 1: In this example, the temperature data of sensor network at USA is used to make experiment [16]. Fig.6(a) depicts the graphical representation of the sensor network. The number of vertices is $N = 48$. And, the edge is constructed by the nearest-neighbor rule. The unit of temperature data is given by Fahrenheit degree. Fig.6(b) shows the GFT spectrum of this temperature data. Clearly, it is a low-frequency signal. The noise $\boldsymbol{\varepsilon}$ is zero-mean white Gaussian noise. The SNR_i of noisy signal $\boldsymbol{z}_0$ is chosen as 17.29 dB. When the parameters are selected as $\rho = 0.7$, $M = 15$ and $t = 0.4$, Table I lists the SNR_o of the denoised signals $\boldsymbol{s}_r$ in four cases. It can be seen that the heat-kernel-smoothing related methods in cases 2, 3, 4 have better performance than conventional matrix inversion method in case 1. Moreover, the TSE and LSE methods in cases 3 and 4 almost provide the same SNR_o as the original HKS method in case 2.

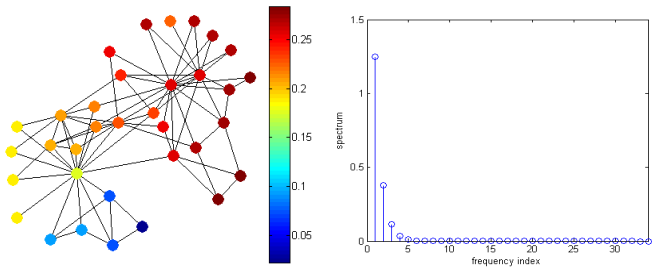


(a) (b)

Fig.6 The denoising example of sensor network at USA. (a) Irregular temperature data. (b) The GFT spectrum of temperature data.

TABLE I: THE SNR_o OF THE DENOISED SIGNAL IN SENSOR NETWORK.

	case 1	case 2	case 3	case 4
SNR_o	22.88 dB	23.16 dB	23.15 dB	23.16dB



(a) (b)

Fig.7 The denoising example of the karate-club social network. (a) Irregular graph data. (b) The GFT spectrum of irregular data.

TABLE II: THE SNR_o OF THE DENOISED SIGNAL IN SOCIAL NETWORK.

	case 1	case 2	case 3	case 4
SNR_o	17.49 dB	18.05 dB	18.04 dB	18.05dB

Example 2: In this example, the data of karate-club social network is used to make experiment [17]. Fig.7(a) depicts the graphical representation of the karate-club social network. The number of vertices is $N = 34$. And, the edge is constructed by using the communication relation of individuals in the club. Fig.7(b) shows the GFT spectrum of this irregular data. Clearly, it is a low-frequency signal. The noise $\boldsymbol{\varepsilon}$ is zero-mean white Gaussian noise. The SNR_i of noisy signal $\boldsymbol{z}_0$ is chosen as 14.49 dB. When the parameters are selected as $\rho = 0.1$, $M = 15$ and $t = 0.1$, Table II lists the SNR_o of the denoised signals $\boldsymbol{s}_r$ in four cases. It is clear that the HKS related methods in cases 2, 3, 4 have better performance than the MI method in case 1. Moreover, the TSE and LSE methods in cases 3 and 4 almost yields the same SNR_o as the original HKS method in case 2.

VI. CONCLUSIONS

In this paper, the distributed computations of heat kernel smoothing using Taylor and Laguerre series expansion methods have been presented. The TSE method corresponds to a maximally-flat graph filter method, and LSE method corresponds to the weighted-least-squares graph filter method. The denoising experiments of data collected from sensor network and social network are also used to demonstrate the effectiveness of HKS methods in terms of the improvement of SNR. However, only Taylor and Laguerre series expansions are studied here. Thus, it is interesting to use other series expansions to implement the heat kernel smoothing method in a distributed manner in the future.

REFERENCES

[1] F. Zhang, and E. R. Hancock, "Graph spectral image smoothing using heat kernel," *Pattern Recognition*, vol. 41, pp. 3328-3342, Nov. 2008.

[2] T. Batard, and M. Bertheir, "Heat kernels of generalized Laplacians – application to color image smoothing," *in Proc. of IEEE ICIP*, pp.461-464, Nov. 2009.

[3] M. K. Chung, Y. Wang, and G. Wu, "Discrete heat kernels smoothing in irregular image domains," *in Proc. of IEEE EMBC*, pp.5101-5104, July 2018.

[4] A. P. Kulkarni, M. K. Chung, B. B. Bendlin, and V. Prabhakaran, "Investigating heritability across resting state brain networks via heat kernel smoothing on persistence diagrams," *in Proc. of IEEE ISBI Workshops*, pp.1-4, Apr. 2020.

[5] C. C. Tseng, and S. L. Lee, "Temperature data denoising based on directed Laplacian matrix and heat kernel smoothing," *in Proc. of IEEE ICCE-TW*, pp.223-224, July 2022.

[6] C. C. Tseng, and S. L. Lee, "Distributed implementation of heat kernel smoothing for graph signal denoising," *in Proc. of IEEE MWSCAS*, pp.1-4, Aug. 2022.

[7] T. Moller, R. Machiraju, K. Mueller, and R. Yagel, "Evaluation and design of filters using a Taylor series expansion," *IEEE Trans. on Visualization and Computer Graphics*, vol. 3, pp. 184-199, Apr. 1997.

[8] S. C. Pei, and P.H. Wang, "Closed-form design of maximally flat FIR Hilbert transformer, differentiators, and fractional delays by power series expansion," *IEEE Trans. on Circuits and Systems-I*, vol. 48, pp. 389-398, Apr. 2001.

[9] C. C. Tseng, "Series expansion design of variable fractional order integrator and differentiator using logarithm," *Signal Processing*, Vol.88, pp.2278-2292, Sep. 2008.

[10] S. S. Bayin, *Essentials of Mathematical Methods in Science and Engineering*, John Wiley & Sons, 2008.

[11] T. Bensouici, A. Charef, and A. Imen, "A simple design of fractional delay FIR filter based on binomial series expansion theory," *Circuits, Systems, and Signal Processing*, Vol.38, pp.3356-3369, 2019.

[12] Q. Xie, Q. Zhao, Z. Xu, and D. Meng, "Fourier series expansion based filter parametrization for equivariant convolutions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.45, pp.4537-3551, Apr. 2023.

[13] P.M. Djuric and C. Richard, *Coorperative and Graph Signal Processing: Principle and Applications*, Academic Press, 2018.

[14] A. Ortega, *Introduction to Graph Signal Processing,* Cambridge University Press, 2022.

[15] C.C. Tseng and S.L. Lee, "Minimax design of graph filter using Chebyshev polynomial approximation," *IEEE Trans. on Circuits and Systems-II*, vol.68, iss.5, pp.1630-1634, May 2021.

[16] S. Furutani, T. Shibahara, M. Akiyama, K. Hato and M. Aida, "Graph signal processing for directed graphs based on the Hermitian Laplacian," in *Machine Learning and Knowledge discovery in Databases*, pp.447-463, 2020.

[17] W.W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol.33, no.4, pp.452-473, 1977.

In-Air Handwriting for Chinese Character Recognition from Monocular Camera: A Deep Learning based Approach with Fingertip Detection and Virtual Strokes Elimination

Chih-Chang Yu*, Zi-Hang Huang and Hsu-Yung Cheng
*Chung Yuan Christian University, Taiwan
E-mail: ccyu@cycu.edu.tw Tel: +886-3-2659999
National Central University, Taiwan
E-mail: chengsy@ncu.edu.tw Tel: +886-3-4227151

***Abstract* — This study presents a deep learning-based approach for recognizing Chinese characters written in the air. Chinese characters, which are composed of numerous strokes in a square shape, pose a challenge for recognizing real and virtual strokes when writing in the air. Existing Optical Character Recognition (OCR) models often result in inaccurate recognition due to this difficulty. Depth cameras can improve recognition accuracy, but they are more expensive than traditional optical cameras.**

To address these challenges, this study employs a deep learning model to track fingertips from a single optical camera. The approach incorporates an intuitive interface that distinguishes between real and virtual strokes while writing. Gesture recognition serves as commands to enable users to start and stop writing and to make corrections during input. This approach makes in-air handwriting behavior similar to writing on paper. To evaluate the effectiveness of the proposed approach, a comparative experiment was conducted to evaluate the model's performance for recognizing characters with and without virtual strokes. Eliminating virtual strokes enabled the OCR model to achieve over 90% recognition accuracy, which was 37% higher than the accuracy with virtual strokes. Moreover, the recognition accuracy of characters with virtual strokes fluctuated significantly in a stability test conducted one week later, while the input method without virtual strokes remained stable at over 90% recognition accuracy. These experimental results indicate that the proposed method can not only effectively remove virtual strokes in air writing but also make in-air handwriting easier for people.

I. INTRODUCTION

The development of contactless handwriting recognition input methods has gained significance in the post-COVID-19 era. Unlike traditional input methods, voice recognition and air handwriting can be input naturally and intuitively without physical contact with devices. In certain circumstances, air writing may be more favorable than voice recognition. However, recognizing air handwriting for Chinese characters arise a challenge due to the lack of clear differentiation between writing down (real strokes) and lifting up (virtual strokes) actions, resulting in the written characters have both types of strokes. In addition, In contrast to English, there are 4,808 frequently used traditional Chinese characters, making the problem be more complicated. The square shape of Chinese characters could make strokes become indistinguishable when there are many real strokes with virtual strokes, resulting in a significant difference in recognition success rates between air handwriting and writing on paper. Therefore, this study aims to alleviate the problem of connected strokes when air writing, allowing the results to more closely resemble those of regular writing results and reducing the difficulty and scale of data collection for air handwriting recognition models of traditional Chinese, which is a significant challenge in this field.

There are various approaches for in-air handwriting recognition. Chen et al. [1] proposed a fingertip detection method based on smart glasses that returned the coordinates of the detected fingertips as input. They utilized Google Input API's Chinese handwriting recognition model for recognition, achieving over 80% accuracy. Hayakawa et al. [2] proposed a method that employs the PS Move motion controller device for air handwriting. This method added a handheld device for precise detection of the writing position and lifting and laying down of the pen. It effectively eliminated virtual strokes during

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

the input stage and used the Japanese recognition model for recognition. However, its use of the PS Move motion controller device made it less convenient and portable as compared to other methods. Additionally, it could not circumvent the disadvantage of needing an additional handheld device in public places.

Another method to eliminate virtual strokes in air handwriting is to perform post-processing after writing. Wu [3] proposed an effective method for eliminating virtual strokes in English air handwriting. This method compared the input letters with all 26 English letters and converts connected letters into individual letters. By doing so, virtual strokes can be eliminated, and results similar to regular writing can be obtained. This method reduced the model's complexity and can use general handwriting data to build the model. However, users need to pre-enter the letters, and the processing time will increase proportionally as the word length increases, affecting the efficiency of each writing procedure. Additionally, this method cannot eliminate virtual strokes within a letter. It was mentioned in the literature that distinguishing between "i" and "l" can be challenging. Traditional Chinese characters usually have virtual strokes within them, making this method unsuitable for traditional Chinese character recognition.

Zou [4] proposed a method that utilized a Kinet and PrimeSense skeleton tracking kit to record the trajectories of writing movements. This approach required users to execute writing movements where the wrist was positioned higher than the shoulder, and the input data did not eliminate virtual strokes. Trajectory normalization, stroke segmentation, trajectory direction, and shape context were employed for recognition. However, this method necessitated pre-entering the writing trajectories of each character and was susceptible to stroke order, which could impact recognition accuracy. The character recognition model collected 300 handwriting characters from four testers to establish a recognition database of 2,400 trajectory data. 150 trajectory data (out of a total of 600) were collected for testing and compared with the data in the database. Although this method satisfied the requirements for contactless devices and did not necessitate additional equipment, the accuracy of the model was influenced by each user's stroke or writing habits. Furthermore, this method's recognition model could not use general writing characters for training. Each character's handwriting data had to be re-collected, increasing the difficulty of data collection for building the model. Chen et al. [5] proposed a method that used dual cameras. The recognition accuracy was impressive but the hardware requirement is relatively high.

After reviewing the above studies, this study presents three contributions:

1. An intuitive interface is designed that effectively eliminates virtual strokes in Chinese in-air handwriting characters, resulting in high recognition accuracy.
2. The proposed method only relies on an optical camera to reduce hardware requirements.
3. The recognition model can be built using regular handwritten data without the need to record writing order, thereby reducing modeling complexity.

II. Proposed Method

This study employs a Convolutional Neural Network (CNN)-based hand pose detection model proposed by Google [6], in conjunction with an optical camera, to detect hand poses and track fingertips frame by frame. In order to distinguish between real and virtual strokes, an intuitive method of marking a region of interest (ROI) on the image is proposed. The ROI moves with the movement of the hands, requiring the user to maintain a writing behavior in order to keep the fingertip inside the ROI. This approach enables handwriting results similar to writing on a plane. Finally, the fingertip trajectories inside the ROI are recorded to form the written character. The system architecture is illustrated in Figure 1.

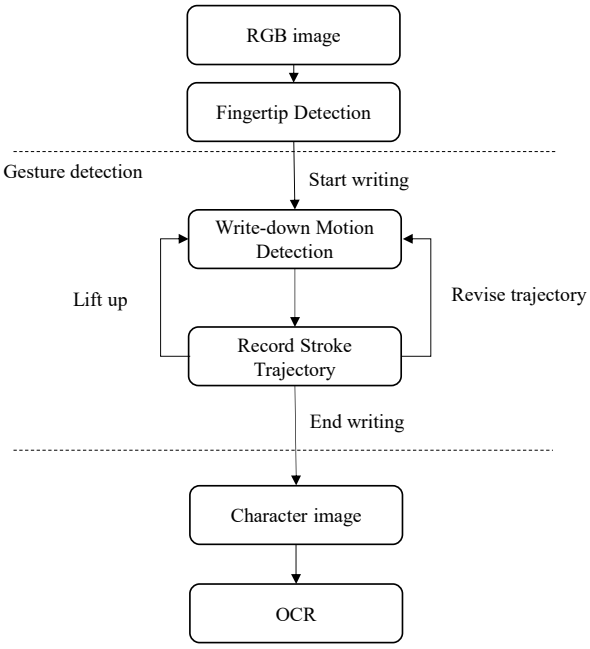


Fig. 1 System architecture

A. *Distinguishing real and virtual strokes*

To ensure that the results of aerial handwriting are similar to those of physical handwriting on a plane, an analysis of the actual way of writing with fingers was conducted, resulting in the following conclusions: When writing on a physical plane, touching the plane with the fingertips is considered to be the real pen, and lifting the fingertips off the plane is considered not writing (Fig. 2). Based on this concept, when writing in the air, an virtual plane is assumed. When the index finger touches the virtual plane, it is considered a real stroke, and when it leaves, it is considered a virtual stroke (Fig. 3). This allows for the differentiation between real and virtual strokes. When observed from the perspective of a camera, the distance between the plane coordinates of the fingertips and the plane coordinates of the proximal interchangeable joint of the index finger is shorter when touching the imaginary plane (real stroke) than when not touching it (virtual stroke). This rule is used as the core, and the ROI range is calculated for determining real and virtual strokes.

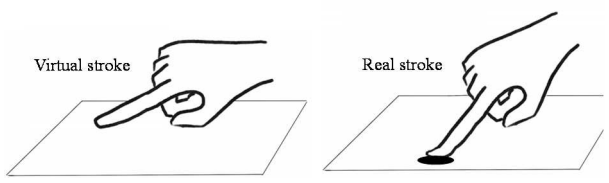


Fig. 2 Scenario of writing on a physical plane using fingertip.

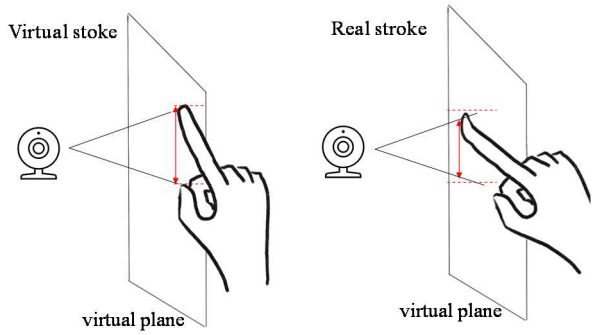


Fig. 3 Detecting write-down behavior when using a single non-depth camera

B. *Gesture recognition and OCR*

To accurately determine the timing of users' writing, a multi-layer perceptron [7-9] was designed as a model for gesture recognition. Three gestures were designed, representing the start of writing, the end of writing, and the correction function. This design allows users to communicate directly with the device without touching any other equipment, controlling the state of air handwriting in real-time, and achieving the desired input results.

After distinguishing between real and virtual strokes, the obtained trajectories are similar to flat writing. Then the trajectory image is converted into binary format and passed to any handwriting OCR model to obtain recognition results. Due to hardware and Mediapipe-hands calculation speed limitations, when writing speed is fast, recorded strokes appear as dotted intervals, which can affect the writing quality. Therefore, after each stroke is completed (lifting action), each point coordinate is calculated, and gaps are filled if they are too large.

III. Experimental Results

In this study, a comparative experimental analysis was conducted to evaluate the effectiveness of different methods for recognizing Chinese characters written in the air. The experiment utilized different handwriting input trajectories, namely, those with and without virtual strokes, and tested them with different handwriting recognition models, including Google Input API and Google Cloud Vision, for cross-comparison. Ten traditional Chinese words were designed with varying levels of difficulty to compare the recognition accuracy and input speed of each method. Since the deep learning models used in this study were all pre-trained models, we don't have to prepare additional dataset for training.

A. *Comparative results*

This study focuses on enhancing input methods by enabling users to avoid generating virtual strokes before recognition. The recognition results are compared using a series of comparative tests that involve different methods, as shown in Table 1. A fingertip detection method proposed by Chen et al. [1] was re-implemented in this study for comparison purposes. This method was chosen because it is highly similar to the approach used in this study, with both involving in-air fingertip tracking and subsequent processing by handwriting recognition models. The fundamental difference between the two methods is the presence or absence of virtual strokes, which allows us to effectively determine the impact of virtual strokes on the recognition model.

Three participants with varying levels of understanding of the presented system were arranged to conduct the test. The first participant was proficient in all four methods used in the experiment, having practiced specifically for many days. The second participant was informed of the four methods in

advance and had one day of practice. The third participant was tested directly after being explained the four methods during the initial experiment. During the test, all three participants used the four methods to input the samples designed in Table 2. In addition to the accuracy of the model, the stroke input time was calculated by dividing the input time of each question by the stroke count of that question. This was used as the evaluation indicator for the accuracy and input efficiency of this cross-comparison test. Results are shown in Table 3 to Table 5.

Table 1. Combinations of stroke type and OCR model.

Methods	Combinations
Method 1	Real stroke + virtual stroke + Google Input API
Method 2	Real stroke + virtual stroke + Google Cloud Vision
Method 3	Real stroke + Google Input API
Method 4	Real stroke + Google Cloud Vision

Table 2. Samples used in the experiment. Note that the input samples are Chinese characters in practice.

Input samples	Characteristics	# of strokes
我	Single character	7
學生	Multiple characters	21
夢想	Multiple characters	27
繁星	Dense strokes	26
水平	Few strokes	9
自信心	Consecutative three characters	19
中文字	Consecutative three characters	14
人工智慧	Consecutative four characters	32
空中手寫	Consecutative four characters	30

Table 3. Recognition results of the first subject.

	Method 1	Method 2	Method 3	Method 4
Accuracy	98%	72.6%	100%	90%
time (s)	2.74	1.53	2.67	2.22
	With virtual stroke		Without virtual stroke	
Accuracy	85%		95%	
time(s)	2.13		4.89	

Table 4. Recognition results of the second subject.

	Method 1	*Method 2*	*Method 3*	*Method 4*
Accuracy	81%	20%	100%	89%
time (s)	4.31	2.92	4.80	6.92
	With virtual stroke		Without virtual stroke	
Accuracy	51%		95%	
time(s)	3.61		5.86	

Table 5. Recognition results of the third subject.

	Method 1	Method 2	Method 3	Method 4
Accuracy	59%	20%	95%	97%
time (s)	4.46	2.08	5.47	6.64
	With virtual stroke		Without virtual stroke	
Accuracy	40%		96%	
time(s)	3.27		6.05	

B. Discussions

The results of this study demonstrated that Method 1 outperformed Method 4 with an average increase of 12.7% in accuracy across three participants. Furthermore, the Google Input API exhibited better recognition results than the Google Cloud Vision, with a minimum recognition accuracy of 59% when virtual strokes were included. Contrasively, eliminating virtual strokes significantly improved accuracy compared to including them, with an increase of 37%. This was particularly true for inexperienced subjects, who experienced an increase of 56% in accuracy. However, writing without virtual strokes was significantly slower than writing with virtual strokes, with an increase of approximately 2.6 seconds in input time per stroke. This was due to the subjects needing to lift their index fingers to avoid generating virtual strokes.

A follow-up experiment was conducted with the same subjects one week later. In the second experiment, accuracy fluctuated greatly when virtual strokes were included across the three participants, with Personnel 1 and Personnel 3 experiencing fluctuations of over 20% and high uncertainty in recognition. However, when eliminating virtual strokes, all participants achieved an accuracy rate of 90% or higher in the experiments, which was more stable. After training, Personnel

2 and Personnel 3 significantly reduced their input time, with an average decrease of 1 second and an average stroke time of 5.87 seconds, while maintaining an accuracy rate of 90% or higher. These results demonstrated that the proposed method can maintain recognition accuracy for those who are unfamiliar with in-air handwriting.

IV. CONCLUSIONS

This study presents an approach to resolve the challenge of recognizing Chinese characters written in the air. The proposed method utilizes an intuitive interface that effectively eliminates virtual strokes, allowing users to clearly distinguish between real and virtual strokes while writing. Handwriting trajectories almost exclusively include real strokes, resulting in high recognition accuracy even for inexperienced users of air handwriting input methods. The proposed method also allows users to correct stroke errors, reducing user stress and providing a more natural and intuitive input method. The virtual stroke-free approach is close to flat writing and is not affected by stroke order, reducing error rates and lowering the data collection threshold for building a recognition model for traditional Chinese air handwriting. Furthermore, the proposed method utilizes RGB images, reducing the hardware threshold for reproducing this technology. Overall, this study presents a significant contribution to the field of contactless handwriting recognition input methods, particularly in the post-COVID-19 era, where such methods have gained importance due to their natural and intuitive input capabilities without physical contact with devices.

ACKNOWLEDGEMENT

This study is partially supported by National Science and Technology Council under grant no. 112-2221-E-033 -041-.

REFERENCES

[1] Y. H. Chen, P. C. Su, F. T. Chien, "Air-Writing for Smart Glasses by Effective Fingertip Detection, " IEEE 8th Global Conference on Consumer Electronics (GCCE), 2019

[2] S. Hayakawa, I. Goncharenko,Y. Gu, "Air Writing in Japanese:A CNN-based character recognition system using hand tracking, " IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), 2022.

[3] C. H. Wu, Air Writing Recognition. Master thesis of Institute of Multimedia Engineering, National Chiao-Tung University, 2014.

[4] P. S. Tsou, In-air Handwriting Chinese Character Recognition. Master thesis of Department of Computer Science and Information Engineering, National Central University, 2018.

[5] Y. N. Chen, E. T. Yang, K. C. Fan, "Air-Handwriting Chinese Character Recognition Based on Dual Camera," *Journal of Advanced Technology and Management*, vol. 8, no. 2, pp. 27-57, 2018.

[6] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking, " *arXiv: 2006.10214*, 2020

[7] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation".

[8] D. E. Rumelhart, J. L. McClelland, *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1: Foundations. MIT Press, 1986.

[9] R.Y.M. Li, B. Tang, K.W. Chau, "Sustainable Construction Safety Knowledge Sharing: A Partial Least Square-Structural Equation Modeling and A Feedforward Neural Network Approach," *Sustainability*, 11(20), 5831, 2019

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

EffSegmentNet: Efficient Design for Real-time Semantic Segmentation

Cyun-Bo Wang and * Jian-Jiun Ding
National Taiwan University, Taiwan
E-mail: r10942041@ntu.edu.tw Tel: +886-956178869
* National Taiwan University, Taiwan
E-mail: jjding@ntu.edu.tw Tel: +886-233669652

***Abstract*— This paper represents the EffSegmentNet, which is a powerful real-time semantic segmentation model. It consists of two segments: (1) A novel MetaFormer-based encoder, termed the EffVisionFormer, is introduced. It captures multiscale image features efficiently. (2) A lightweight decoder which utilizes multiscale image features from the encoder is applied to conduct rapid yet accurate segmented result. The proposed EffSegmentNet achieves remarkable performance which takes the inference speed, accuracy, and model parameters into account. On the Cityscapes test set, we attain 71.9% mIoU with 195.1 frames per second (FPS) on a NVIDIA RTX 2080Ti card. Furthermore, the proposed EffSegmentNet utilizes only 4.4 million parameters, which demonstrates its advantage on real-time segmentation.**

I. INTRODUCTION

Semantic segmentation, a fundamental task in the field of computer vision, involves assigning pixel-level labels to images. Its applications span video surveillance, autonomous driving, medical image processing, etc., which demand the models to inference rapidly and accurately.

To meet these demands, researchers have proposed various real-time models based on Fully Convolutional Networks (FCNs), such as DFANet [4], BiSeNetV1 [5], and BiSeNetV2 [6]. While these methods satisfy the requirements for fast inference speed and up to par segmentation accuracy, they struggle to capture long-range image information due to their limited receptive fields.

Recently, transformer-based methods have demonstrated great prospects and have outperformed FCN-based methods in some cases. The encoders of transformer-based methods offer larger receptive fields, enabling the extraction of global contextual information. Due to their powerful ability to capture image information, their number of parameters have the potential to be compressed while fulfilling the demands for low-latency and satisfactory segmentation accuracy.

However, the Multi Head Self Attention (MHSA) mechanism at the core of transformers is computationally expensive, which is quadratic to the number of tokens. Models like Swin Transformer [11] and Twins [12] have addressed this challenge by introducing window-based self-attention, reducing computational complexity from quadratic to linear while maintaining performance. Although such designs are much more efficient theoretically, they are difficult to be deployed with TensorRT [1] or mobile devices.

In 2021, the introduction of spatial MLPs [13] as token mixers emerged as an alternative to attention mechanisms, achieving competitive results on image classification benchmarks. This development raised questions about the necessity of attention-based token mixers in image understanding. In 2022, a paper introduced the concept of the MetaFormer [14], a general structure derived from the Transformer, omitting the specification of the token mixer component but still achieving competitive performance.

This paper introduces EffSegmentNet, a powerful semantic segmentation model which considers FPS, number of parameters and accuracy. Our main contributions can be concluded as follows:

- EffVisionFormer, an innovative MetaFormer-based encoder, is introduced to capture robust multiscale information of images. It features a unique token mixer design that integrates local spatial MLP and global efficient attention mechanism, enabling the efficient extraction of rich contextual information.
- A thoughtfully-crafted lightweight decoder which aggregates hierarchical image information from the encoder to conduct

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

precise segmented prediction rapidly.

- Our method demonstrates its remarkable performance on the number of parameters, FPS, and segmentation accuracy. EffsegmentNet attains 71.9% mIoU on the Cityscapes test set, while maintaining a prompt inference speed of 195.1FPS on one NVIDIA GTX 2080Ti card. Moreover, the EffsegmentNet leverages only 4.4 million parameters, showcasing its superiority over other real-time alternatives.

II. Related Work

Semantic segmentation aims to categorize each pixel in an image into its corresponding class. The pioneering work of deep learning methods for semantic segmentation is FCN [2], which is a fully convolutional network that enables end-to-end pixel-to-pixel classification. Since then, extensive research [3, 4, 5, 6, 7, 16, 17] has been dedicated to improve FCN [2], which includes non-real-time and real-time methods. Real-time methods like ICNet [3], DFANet [4], BiSeNet [5], and BiseNetV2 [6] consider both the inference speed and the segmentation accuracy, making them suitable for applications requiring fast inference.

In 2020, the transformer architecture was introduced to vision tasks by Dosovitskiy et al [8]. Subsequently, numerous studies have emerged aiming to improve the transformer architecture specifically for vision tasks. DeiT [9] enhanced the training pipeline by distillation, thereby eliminating the requirement for extensive pretraining. PVT [10] introduced the pyramid structure into the transformer, creating the opportunity to leverage the Transformer model instead of the CNN model for dense vision prediction tasks like semantic segmentation. Building upon PVT's foundation [10], Swin Transformer [11] and Twins [12] adopted the pyramid structure and alleviated the computational complexity of the attention mechanism. ResMLP [13] challenged the necessity of the attention-based token mixer and introduced an alternative known as the spatial MLP token mixer. Drawing inspiration from the spatial MLP token mixer and other research works, the MetaFormer [14] concept emerged, asserting that token mixer design can be versatile and not limited to attention-based token mixers. In line with this concept, PoolFormer [14] was introduced, utilizing pooling as a simple token mixer and achieving competitive performance compared to other approaches that employ more intricate attention-based token mixers. Building upon the MetaFormer [14] concept and existing research, following works such as EfficientFormer [28] and EfficientFormerV2 [29] were proposed. These approaches leverage the pyramid structure and employ distinct token mixers at different stages of the network pipeline. Early stages utilize lightweight token mixers, such as pooling or depthwise convolution, for extracting low-level features (local information). This sophisticated design enables these models to excel in image classification tasks, boasting low latency and high parameter efficiency. These transformer-derived methods showcase their remarkable ability to extract image features, making them highly proficient in image classification tasks and well-suited as encoders for semantic segmentation.

III. Proposed Method

In this section, we introduce the EffSegmentNet, a powerful, and efficient real-time framework for semantic segmentation. Our encoder, EffVisionFormer, employs the pyramid structure and utilizes various token mixers at different stages of the network pipeline, aligning with the concepts of previous works. The core aspect of the encoder architecture lies in the meticulous design of token mixers in the late stages of the network pipeline, enabling efficient capture of global information. Our compact decoder architecture fuses spatial feature maps and contextual feature maps effectively to generate the semantic segmentation result. The overall architecture design is illustrated in Fig. 1.

A. Encoder

As shown in Fig. 1(a), our sophisticated designed encoder, EffVisionFormer, includes convolutional stems, embedding layers, and two distinct block types: local blocks and global blocks. The convolutional stem efficiently downsamples the input using a 3×3 convolution with stride 2, followed by batch normalization (BN) and ReLU. After two convolutional stems, the resolution is reduced to 1/4 of the input resolution. The embedding layer is located between two adjacent stages. It reduces the resolution by half through a 3×3 convolution with a stride of 2 and BN.

The local block, as introduced in [29], is a variant of the MetaFormer block [14]. This variant [29] employs the depthwise convolution as a token mixer and positions it between two 1×1 convolutional layers to further enhance the ability of extracting local information. This configuration can be seen as integrating the local token mixer between two feedforward layers of the channel MLP in the MetaFormer block [14]. The essence of our design lies in the global block, which is placed in the last two stages of the encoder. The global block follows a similar architecture to the MetaFormer block [14], comprising two stages: the first stage features a meticulously-designed global token mixer, and the second stage employs the local block. In the global block, we employ

the local block as the improved version of the channel MLP, allowing the global block to extract a greater amount of information.

The token mixer of the global block contains two components. The first component is Local Spatial MLP. The original spatial MLP introduced in [13] utilizes a fully connected linear layer to exchange information across patches (tokens). Given a feature map F^{3D} of dimension $h \times w \times C$, where h represents height, w denotes the width, and C indicates the number of channels, the feature map F^{3D} needs to be flattened into F^{2D} with dimension $N \times C$ in order to feed it to the original spatial token mixer [13]. Here, N is defined as the number of input tokens (nodes) and is calculated as $N = h \times w$. The original spatial token mixer linearly projects N nodes to N nodes, resulting in the computational complexity of $O(N^2)$, which is computationally expensive. To solve this issue, our Local Spatial MLP design partitions the input tokens into R groups, with each group containing N/R input tokens. We assume that the total number of input tokens, N, are divisible by R. Within each group, a local fully-connected linear layer, which projects N/R nodes to N/R nodes, is applied to enable information exchange among N/R patches. The computational complexity of a local fully-connected linear layer for each group is $O(N^2/R^2)$, resulting in a total computational complexity of $O(N^2/R^2 \times R) = O(N^2/R)$ for R groups. This represents a reduction in computational complexity compared to the original spatial token mixer [13].

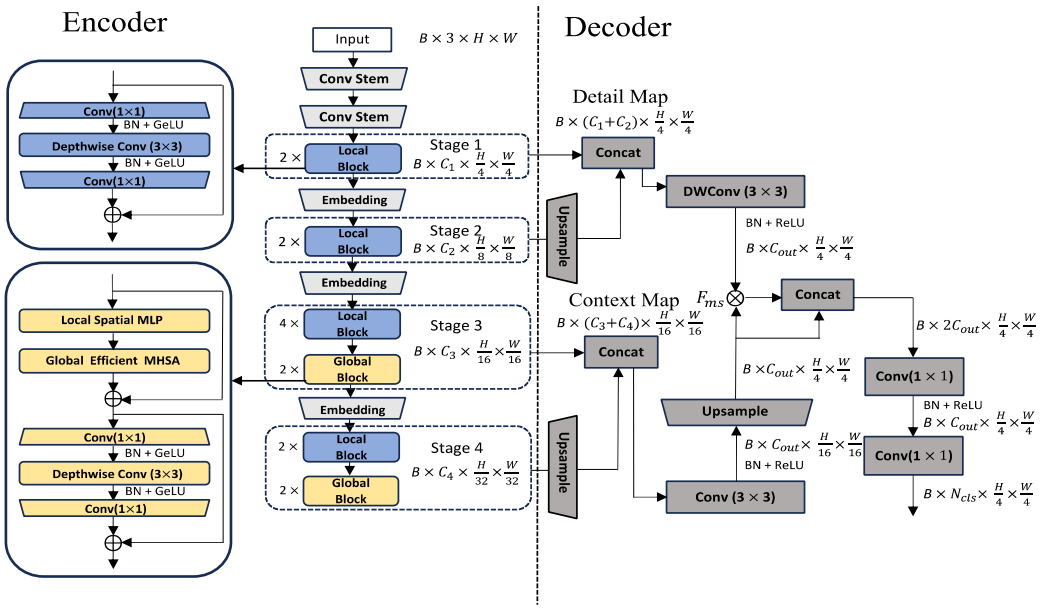
Moreover, when processing high-resolution inputs with the Local Spatial MLP design, we employ three local fully-connected linear layers, denoted as $f_{1,i}$, $f_{2,i}$, and $f_{3,i}$ for each group to further decrease the total computational complexity, where i denotes the i^{th} group. The first layer, $f_{1,i}$, linearly projects N/R nodes to N_c/R nodes, where $N_c = \alpha \times N$ with $0 < \alpha < \sqrt{2} - 1$ and $N_c \in \mathbb{N}$. The computational complexity of this layer is $O(NN_c/R^2)$. The second layer, $f_{2,i}$, linearly projects N_c/R nodes to N_c/R nodes. The computational complexity of this layer is $O(N_cN_c/R^2)$. The third layer, $f_{3,i}$, linearly projects N_c/R nodes to N/R nodes with the computational complexity of $O(NN_c/R^2)$. In a generic scenario, we assume that $N\%R = 0$ and $N_c\%R = 0$. The combined complexity of the three fully connected layers for each group is $O(N_c \times (N_c + 2N)/R^2)$, which leads to the total computational complexity of $O(N_c \times (N_c + 2N)/R) = O(\alpha N \times (\alpha N + 2N)/R) = O((\alpha^2 + 2\alpha)N^2/R)$. Based on the mathematical deduction, it becomes evident that the computational complexity can be further diminished when the condition $0 < \alpha^2 + 2\alpha < 1$ is satisfied. Solving the equation, we find that the requirement holds true when $0 < \alpha < \sqrt{2} - 1$.

Despite the computational efficiency provided by the Local Spatial MLP design, its information exchange is limited to tokens within the same groups. This limitation results in a smaller receptive field, affecting overall performance. To overcome this challenge, we introduce the Global Efficient MHSA as the second component of the token mixer of the global block. The Global Efficient MHSA enables the exchange of information between different groups, effectively resolving the issue. The attention process of the multi-head self-attention is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V, \quad (1)$$

where Q, K, V have dimensions of $n_{head} \times N \times d_{head}$. Here, n_{head} denotes the number of heads, N represents the number of tokens, and d_{head} is defined as the number of channels per head. The product of n_{head} and d_{head} is equal to C, which represents the total number of channels. The computational complexity of this process is $O(N^2)$, which indicates that it has a quadratic relationship with the number of tokens N. This quadratic complexity can be computationally expensive, especially for large values of N. In the Local Spatial MLP component, the N input tokens are divided into R groups, and the relationships among input tokens within each group are computed, generating N output tokens. These N output tokens are then utilized as Q in the Global Efficient MHSA module, with the dimensions of $n_{head} \times N \times d_{head}$. Additionally, the N/R output tokens from each group are averaged to obtain a group representative for each group. With the presence of R groups, R group representatives are produced, which fullfill the roles of K and V. As a result, the dimensions of K and V are reduced from $n_{head} \times N \times d_{head}$ to $n_{head} \times R \times d_{head}$. By employing this approach, the computational complexity of Global Efficient MHSA is $O(NR)$, which exhibits significant reduction in computational cost when $R \ll N$. In summary, the total computational complexity of the Local Spatial MLP and the Global Efficient MHSA is $O((\alpha^2+2\alpha)N^2/R + NR)$. To reduce the computational complexity, we choose $0 < \alpha < \sqrt{2} - 1$ and ensure that R is less than N. Selection of α and R directly influences the extent of computational complexity reduction.

Table 1 illustrates the architectural setting of the encoder designed for ImageNet-1K [31], which is a prominent image classification dataset. The input resolution of this structure is



(a) The architecture of the encoder (b) The architecture of the decoder

Fig. 1 The overall architecture of the EffSegmentNet. (a) The hierarchical encoder of the EffSegmentNet that extracts low-level and high-level image information. (b) The lightweight decoder of the EffSegmentNet that fuses multi-level features to conduct the semantic segmentation prediction.

224 × 224, and due to the relatively smaller image resolution, we employ a single fully connected linear layer for each group in the Local Spatial MLP Module.

On the other hand, Table 2 showcases the architectural configuration of the encoder tailored for Cityscapes [32], a prevalent dataset used for semantic segmentation tasks. The input resolution of this architecture is 512 × 1024, and owing to the higher image resolution, we utilize three fully connected linear layers for each group in the Local Spatial MLP module to further reduce the computational complexity. Although the two architectures share a similar design, the notable difference lies in the number of fully connected linear layers employed per group in the Local Spatial MLP Module. The hyper-parameters of our encoder are listed as follows:

- K: the kernel size utilized for Patch Embed;
- S: the stride value utilized for Patch Embed;
- C: the number of output channels at each stage;
- Q: the quantity of each type listed in Table;
- E: the expansion ratio of the local block;
- R: the number of groups in the Local Spatial MLP Module;
- F: the number of fully connected linear layers for each group in the Local Spatial MLP Module;
- α: the reduction ratio as illustrated above, its value falls within the open interval $(0, \sqrt{2}-1)$;
- N_c: $\alpha \times N$, where N is the number of tokens inputted into the Local Spatial MLP Module and $N_c \in \mathbb{N}$.

B. *Decoder*

The decoder plays a crucial role in aggregating the multiscale features generated by the encoder. Inspired by the Bilateral Guided Aggregation Layer in [6], we have developed a more streamlined decoder architecture. In order to capture detailed spatial information, we employ a combination of the output from stage 1 and stage 2 of the encoder. While the output from stage 1 remains unaltered, the output from stage 2 is subjected to a *2x* upsampling. These two outputs are then concatenated together to form the detail map. To further enhance the detail information, we apply depthwise convolution, batch normalization (BN) and ReLU operations to the detail map, allowing us to capture plentiful spatial details.

To capture abundant context information, our approach involves utilizing both the output from stage 3 and stage 4 of the encoder. While the output from stage 3 remains unchanged, the output from stage 4 undergoes a *2x* upsampling. These two outputs are then concatenated together to form the context map. To further strengthen the contextual information, we apply convolution, batch normalization (BN) and ReLU operations to the context map, enabling us to capture a rich array of semantic information. Next, we perform element-wise multiplication between the detail information and the contextual information to obtain the feature map $F_{multiscale}$ (F_{ms}), which contains a wealth of multiscale information. To further enrich the feature map F_{ms}, we concatenate it with the contextual information

Table 1. The configuration of the encoder designed for ImageNet-1K.

Stage	Output size	Type	Parameters		
Image	224 × 224				
Stem 1	112 × 112	Patch Embed	K = 3, S = 2, Q = 1, C = 16		
Stem 2	56 × 56	Patch Embed	K = 3, S = 2, Q = 1, C = 32		
1	56 × 56	Local Block	Q = 2, C = 32, E = [4,4]		
2	28 × 28	Patch Embed	K = 3, S = 2, Q = 1, C = 64		
		Local Block	Q = 2, C = 64, E = [4,4]		
3	14 × 14	Patch Embed	K = 3, S = 2, Q = 1, C = 128		
		Local Block	Q = 4, C = 128, E = [4,3,3,3]		
		Global Block	Q = 2, C = 128	Token Mixer	R = 4, F = 1
				Local Block	E = [4,4]
4	7 × 7	Patch Embed	K = 3, S = 2, Q = 1, C = 224		
		Local Block	Q = 2, C = 224, E = [4,3]		
		Global Block	Q = 2, C = 224	Token Mixer	R = 1, F = 1
				Local Block	E = [3,4]

and apply the 1 × 1 convolution, BN and ReLU operations.

Finally, we utilize a 1 × 1 convolutional layer to reduce the number of channels to match the number of classes N_{cls} in the dataset. In the case of the Cityscapes dataset, N_{cls} is set as 19. The overall architecture design of the decoder is illustrated in Fig. 1(b). Regarding the hyper-parameter configuration of the decoder, we set C_{out} to 320, and remarkably, the decoder consists of only 1.2 million parameters.

IV. Experimental Results

A. Datasets

We employ two datasets for evaluation: ImageNet-1K [31] and Cityscapes [32]. The ImageNet-1K [31] dataset is used to evaluate the effectiveness of our proposed encoder, EffVisionFormer. On the other hand, the Cityscapes [32] is utilized to assess the performance of our semantic segmentation framework, EffSegmentNet.

The ImageNet-1K [31] dataset is widely regarded as the most popular image classification dataset, which contains 1.28 million training images and 50000 validation images distributed across 10000 different classes.

Cityscapes [32] is an urban segmentation dataset specifically designed for driving scenarios. It contains a total of 5000 images with fine annotations, out of which 2975 images are used for training, 500 images for validation and 1525 images for testing.

Table 2. The configuration of the encoder designed for Cityscapes.

Stage	Output size	Type	Parameters		
Image	512 × 1024				
Stem 1	256 × 512	Patch Embed	K = 3, S = 2, Q = 1, C = 16		
Stem 2	128 × 256	Patch Embed	K = 3, S = 2, Q = 1, C = 32		
1	128 × 256	Local Block	Q = 2, C = 32, E = [4,4]		
2	64 × 128	Patch Embed	K = 3, S = 2, Q = 1, C = 64		
		Local Block	Q = 2, C = 64, E = [4,4]		
3	32 × 64	Patch Embed	K = 3, S = 2, Q = 1, C = 128		
		Local Block	Q = 4, C = 128, E = [4,3,3,3]		
		Global Block	Q = 2, C = 128	Token Mixer	R = 4, F = 3, $\boldsymbol{\alpha}$ = 0.096, $\boldsymbol{N_c}$ = 196
				Local Block	E = [4,4]
4	16 × 32	Patch Embed	K = 3, S = 2, Q = 1, C = 224		
		Local Block	Q = 2, C = 224, E = [4,3]		
		Global Block	Q = 2, C = 224	Token Mixer	R = 1, F = 3, $\boldsymbol{\alpha}$ = 0.096, $\boldsymbol{N_c}$ = 49
				Local Block	E = [3,4]

B. Implementation Details

We train the model on a NVIDIA RTX 2080Ti card and a NVIDIA GTX 1080Ti card using CUDA 11.8, CUDNN 8.6.0. During inference, the model is run on a NVIDIA RTX 2080Ti card, utilizing CUDA 11.8, CUDNN 8.6.0, and TensorRT 8.5.3.1. A batch size of 1 is used to evaluate the computational capacity of our method.

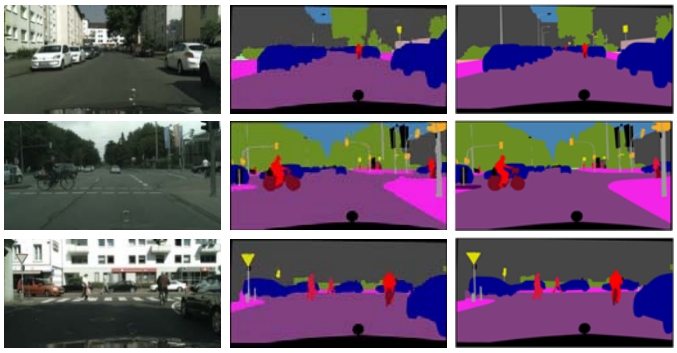
Initially, we pretrain the encoder on the ImageNet-1K dataset with an input image resolution of 224 × 224. The pretraining process involves training the encoder for 300 epochs using an AdamW optimizer, a cosine annealing rate scheduler, and a linear warmup period of 5 epochs. A batch size of 1024, an initial learning rate of 0.001, and a weight decay of 0.025 are employed. Due to limited computational resources, the batch size on each card is set to 128. To simulate a batch size of 1024, we utilize gradient accumulation with accumulation steps of 4. During training, we incorporate various data augmentation and regularization strategies from [9], including Rand Augment, random erasing, Mixup, Cutmix, and label smoothing. However, due to limited computational resources, we do not employ the distillation approach of [9] to further enhance the accuracy of the model.

Table 3. Comparisons with other state-of-the-art real time semantic segmentation frameworks on Cityscapes. The table provides information on the input resolution and the GPU utilized for inference of each method.

Model	Encoder	Resolution	mIoU(%)		FPS	Params	GPU
			val	test			
ENet [19]	-	512 × 1024	-	58.3	76.9	-	Titan X
ICNet [3]	PSPNet50	1024 × 2048	-	69.5	30.3	26.5M	Titan X
ESPNet [30]	ESPNet	512 × 1024	-	60.3	112.9	0.4M	Titan X
DFANet A [4]	Xception A	1024 × 1024	-	71.3	100	7.8M	Titan X
DFANet B [4]	Xception B	1024 × 1024	-	67.1	120	4.8M	Titan X
FastSCNN [20]	-	1024 × 2048	68.6	68.0	123.5	1.1M	Titan Xp
BiSeNetV1[5]	Xception39	768 × 1536	69.0	68.4	105.8	5.8M	GTX 1080Ti
BiSeNetV1-L [5]	ResNet18	768 × 1536	74.8	74.7	65.5	49M	GTX 1080Ti
BiSeNetV2 [6]	-	512 × 1024	73.4	72.6	156	-	GTX 1080Ti
BiSeNetV2-L [6]	-	512 × 1024	75.8	75.3	47.3	-	GTX 1080Ti
SwiftNet [21]	ResNet18	1024 × 2048	75.4	75.5	45.4	11.8M	RTX 2080Ti
GAS [22]	-	769 × 1537	-	71.8	108.4	-	Titan Xp
MSFNet [34]	-	512 × 1024	-	71.3	117	-	RTX 2080Ti
FasterSeg [23]	-	1024 × 2048	73.1	71.5	163.9	4.4M	GTX 1080Ti
EffSegmentNet	EffVisionFormer	512 × 1024	73.1	71.9	**195.1**	4.4M	RTX 2080Ti

Table 4. Comparisons with other compact networks on ImageNet-1K.

Model	Top1 Acc	FLOPs	Params
ResNet-18 [24]	69.0%	1.8G	11.2M
DF1 [25]	69.8%	0.746G	8.0M
DF2 [25]	73.9%	1.77G	17.5M
MobileNetV2 [26]	72.0%	0.3G	3.4M
GhostNet(x1.0) [27]	73.9%	0.141G	5.2M
STDC1 [7]	73.9%	0.813G	8.4M
EfficientFormerV2-S0 [29]	75.7%	0.4G	3.5M
MiT-B0 [18]	70.5%	0.6G	3.7M
EffVisionFormer	71.3%	0.4G	**2.9M**



(a) Input (b) EffSegmentNet (c) Groundtruth

Fig. 2 Visualizations of multiple semantic segmentation results generated by the EffSegmentNet on the Cityscapes validation set. The figure consists of three columns: (a) shows the input images, (b) displays the predicted results obtained by the EffSegmentNet, and (c) represents the corresponding groundtruths of the input images.

Following the pretraining, we proceed to train the pretrained encoder in combination with the decoder on Cityscapes [32], using an input resolution of 512 × 1024. The EffSegmentNet is trained for 160000 iterations, employing an AdamW optimizer, a poly linear rate scheduler, and a linear warmup period of 1500 iterations. A batch size of 8, an initial learning rate of 0.00006, and a weight decay of 0.01 is used. During the training phase, we utilize multiple data augmentation methods to enhance the model's performance. In addition, we also adopt OHEM (Online Hard Example Mining) [33] and auxiliary losses to further boost the model's performance. It is notable that the model is trained on the training set and validation set

when evaluating on the test set.

C. Evaluation Metrics

To assess the performance of the backbone on the ImageNet-1K dataset [31], we employ top-1 accuracy, parameters and floating point operations (FLOPs) as the evaluation metrics. For evaluating the performance of the segmentation framework, we utilize mean intersection over union (mIoU), parameters, and frames per second (FPS).

D. Comparisons to State-of-the-Arts

In this section, we evaluate our backbone, EffVisionFormer, against existing methods on ImageNet-1K [31], and we compare the EffSegmentNet with other methods on Cityscapes [32].

In Table 3, we provide comparisons between the EffSegmentNet and other state-of-the-art real-time semantic segmentation frameworks on Cityscapes [32]. The images in the Cityscapes dataset [32] have a resolution of 1024 × 2048. During the inference phase, we first resize the image resolution to 512 × 1024 and input it into EffSEgmentNet. Then we upsample the prediction of EffSEgmentNet to restore it to the original resolution of 1024 × 2048. This entire process is accounted for in the measurement of inference time.

Upon analyzing the data provided in Table 3, it becomes apparent that our method strikes a commendable equilibrium between multiple evaluation metrics, including mIoU, FPS, and parameters. The EffSegmentNet attains a decent mIoU score of 71.9% on the Cityscapes test set, accompanied by a high FPS of 195.1, indicating its efficiency in processing images in real-time scenarios. Notably, these impressive results are achieved with a relatively small parameter count of 4.4 million, showcasing the effectiveness of our approach in efficiently utilizing computational resources. When compared to other methods with similar number of parameters, our method outperforms in terms of mIoU and FPS.

Examining Table 4 reveals that while our backbone does not attain the highest accuracy when compared to other lightweight backbones, it still delivers competitive performance with its relatively small parameters size.

To provide additional insights, Fig. 2 visualizes multiple semantic segmentation results, which show the model's ability in performing semantic segmentation tasks.

V. Conclusions

This paper introduces EffSegmentNet, a lightweight and powerful semantic segmentation model which comprises a hierarchical encoder, EffVisionFormer, and a compact decoder. The innovative design of the token mixer, combining Local Spatial MLP and Global Efficient MHSA, enables the encoder to capture global information efficiently. Moreover, the decoder is meticulously designed to effectively fuse the multiscale information obtained from the encoder. As a result, EffSegmentNet achieves an excellent equilibrium between mIoU, FPS, and parameters when compared to other state-of-the-art real time methods, making it a promising solution for various real-world semantic segmentation applications.

References

[1] TensorRT. https://developer.nvidia.com/tensorrt

[2] J. Long, E. Shelhamer, and T..Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015.

[3] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *European Conf. Computer Vision,* pp. 405-420, 2018.

[4] H. Li, P. Xiong, H. Fan, and J. Sun. "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 9522-9531, 2019.

[5] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *European Conf. Computer Vision*, pp. 325-341, 2018.

[6] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang. "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Computer Vision*, vol. 129, pp. 3051-3068, 2021.

[7] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. "Rethinking bisenet for real-time semantic segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 9716-9725, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learning Representations*, 2020.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. "Training data-efficient image transformers & distillation through attention," in *Int. Conf. Machine Learning*, PMLR, vol. 139, pp. 10347-10357, 2021.

[10] W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE Int. Conf. Computer Vision*, pp. 568-578, 2021.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE Int. Conf. Computer Vision*, pp. 10012-10022, 2021.

[12] X. Chu, Zhi Tian, Y. Wang, Bo Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355-9366, 2021.

[13] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou. "ResMLP: Feedforward networks for image classification with data-efficient training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, issue 4, pp. 5314-5321, 2023.

[14] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan. "Metaformer is actually what you need for vision," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 10819-10829, 2022.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is all you need," *Advances in Neural Information Processing Systems*, article 30, pp. 1-11, 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation," in *Int Conf. Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, pp. 234-241, 2015.

[17] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, issue 4, pp. 834-848, 2017.

[18] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077-12090, 2021.

[19] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. "Enet: Deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv*:1606.02147, 2016.

[20] R. P. K. Poudel, S. Liwicki, and R. Cipolla. "Fast-SCNN: Fast semantic segmentation network," *arXiv preprint arXiv*: 1902.04502, 2019.

[21] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic. "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 12607-12616, 2019.

[22] P. Lin, P. Sun, G. Cheng, S. Xie, X. Li, and J. Shi, "Graph-guided architecture search for real-time semantic segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4203-4212, 2020.

[23] W. Chen, X. Gong, X. Liu, Q. Zhang, Y. Li, and Z. Wang, "Fasterseg: Searching for faster real-time semantic segmentation," in *Int. Conf. Learning Representations,* pp. 1-14, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[25] X. Li, Y. Zhou, Z. Pan, and J. Feng. "Partial order pruning: for best speed/accuracy trade-off in neural architecture search," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 9145-9153, 2019.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.

[27] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu. "Ghostnet: More features from cheap operations," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1580-1589, 2020.

[28] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12934-12949m 2022.

[29] Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov, and J. Ren, "Rethinking vision transformers for mobilenet size and speed," *arXiv preprint arXiv:2212.08059*, 2022.

[30] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi. "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *European Conf. Computer Vision*, pp. 552-568, 2018.

[31] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. p. 248-255.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3213-3223, 2016.

[33] A. Shrivastava, A. Gupta, and R. Girshick. "Training region-based object detectors with online hard example mining," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 761-769, 2016.

[34] H. Si, Z. Zhang, F. Lv, G. Yu, and F. Lu, "Real-time semantic segmentation via multiply spatial fusion network," *arXiv preprint arXiv*:1911.07217, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Universal Optimal Parameters of the Closed-Form Linear Canonical Wigner Distribution

Zhichao Zhang*
* School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
E-mail: zzc910731@163.com Tel: +86-13376073017

***Abstract*—The closed-form linear canonical Wigner distribution (CFLCWD), equipped with nine linear canonical transform (LCT) free parameters, is an effective time-frequency method suitable for extracting the principal features of linear frequency-modulated signals from an extreme strong stationary Gaussian noise background. However, the appropriate parameters for non-stationary non-Gaussian noise jamming nonlinear frequency-modulated signal processing are still unknown. To address this issue, we reveal the equivalence of the uncertainty product in CFLCWD domains and those in LCT domains, based on which we deduce the CFLCWD's uncertainty inequality, whose equality is achieved by a Gaussian signal. We further derive the universal optimal LCT free parameters with which the CFLCWD reaches the highest-resolution in time-frequency plane. In particular, the strong uncertainty inequality of the classical Wigner distribution is also revisited.**

I. Introduction

Wigner distribution (WD) is a mother distribution of the celebrated Cohen's class distribution [1], [2]. It acts a fundamental role in non-stationary signal time-frequency analysis [3], [4]. Many fundamental results of the WD are presently established, for example uncertainty principles [5]–[12] which are of importance in time-frequency analysis. Uncertainty inequalities of the WD given by [5], [11] are essentially the classical Heisenberg's results [2], [13]–[16]. The attainable lower bound derived is none other than Planck constant which imposes limits on the WD's resolution in time-frequency plane.

With the techniques of parametric time-frequency analysis, many attempts [17]–[28] have recently been made to introduce free parameters of the linear canonical transform (LCT) into the ordinary WD to break the traditional time-frequency resolution limit. The closed-form linear canonical Wigner distribution (CFLCWD) is a representative variation among them [29]. It equips three parameter matrices, i.e., nine LCT free parameters. Due to more degrees of freedom, the CFLCWD exhibits better detection performance on linear frequency-modulated signals corrupted with an extreme strong stationary Gaussian noise [30]–[33]. Namely, it is able to achieve higher resolution for this kind of noisy signals in time-frequency plane.

Despite the CFLCWD's advantage in time-frequency analysis, there remains a serious challenge in the use of it to non-stationary non-Gaussian noises jamming nonlinear frequency-modulated signals processing. Uncertainty principle of a time-frequency distribution yields a lower bound that depends only on the distribution itself [34]. For the CFLCWD, the corresponding lower bound relies merely on the embedded LCT free parameters. Generally speaking, a smaller lower bound yields a higher time-frequency resolution [28]. Uncertainty principle of the CFLCWD thus meets the intrinsic requirement of non-stationary non-Gaussian noise and nonlinear frequency-modulated signal time-frequency superresolution analysis.

In brief, to tackle a challenge in non-stationary non-Gaussian noise jamming nonlinear frequency-modulated signal processing, uncertainty principle of the CFLCWD is a critical foundational technology. The main purpose of the paper is therefore to formulate a universal strategy suitable for general noisy signals through the research of the strong uncertainty inequality of the CFLCWD. The academic contributions of the paper are concluded as follows:

- The paper establishes uncertainty principle of the CFLCWD and obtains an attainable lower bound that relies only on the LCT free parameters embedded in the CFLCWD.
- The paper proposes a universal optimal LCT free parameters selection strategy on non-stationary non-Gaussian noise jamming nonlinear frequency-modulated signal processing.

The rest of the paper is structured as follows. In Section II we collect many primary conceptions and notations on the CFLCWD. In Section III we define the uncertainty product in CFLCWD domains and expound the research question, method, purpose and significance. In Section IV we explore the CFLCWD's uncertainty principle. In Section V we discuss the optimal LCT free parameters of the CFLCWD. In Section VI we draw a conclusion.

II. Closed-form linear canonical Wigner distribution (CFLCWD)

This section mainly reviews the CFLCWD's definition.

Definition 1: The LCT of a signal $f(t) \in L^2(\mathbb{R})$ with the parameter matrix $\mathbf{A} = (a, b; c, d)$ is given by [35]–[39]

$$\widehat{f}_{\mathbf{A}}(u) = \begin{cases} \int_{\mathbb{R}} f(t)\mathcal{K}_{\mathbf{A}}(u,t)\mathrm{d}t, & b \neq 0 \\ \sqrt{d}\mathrm{e}^{\mathrm{j}\frac{cd}{2}u^2} f(du), & b = 0 \end{cases},$$

where

$$\mathcal{K}_{\mathbf{A}}(u,t) = \frac{1}{\sqrt{\mathrm{j}2\pi b}} \mathrm{e}^{\mathrm{j}\left(\frac{d}{2b}u^2 - \frac{1}{b}ut + \frac{a}{2b}t^2\right)}$$

This work was supported by the National Natural Science Foundation of China under Grant No 61901223 and the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant No 2021K205B.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

denotes the kernel, and where a, b, c, d are real parameters satisfying $ad - bc = 1$.

The LCT is just a combination of scaling operations and square frequency modulation operations for $b = 0$. This specific case is of no interest to us. The paper thus focuses merely on the LCT with $b \neq 0$. Because of $c = \frac{ad-1}{b}$, three free parameters of the LCT are a, b, d.

Let $\widehat{f}_{\mathbf{A}}(v)$ be the LCT of $f(t) \in L^2(\mathbb{R})$ with $\mathbf{A}$, and $\mathcal{K}_{\mathbf{A}_1}(u, v)$ and $\mathcal{K}_{\mathbf{A}^{-1}}(t, v)$ be the kernels with $\mathbf{A}_1$ and $\mathbf{A}^{-1} = (d, -b; -c, a)$, respectively. The LCT is additive:

$$\widehat{f}_{\mathbf{A}_1\mathbf{A}}(u) = \int_{\mathbb{R}} \widehat{f}_{\mathbf{A}}(v)\mathcal{K}_{\mathbf{A}_1}(u, v)\mathrm{d}v$$

and is also invertible:

$$f(t) = \int_{\mathbb{R}} \widehat{f}_{\mathbf{A}}(v)\mathcal{K}_{\mathbf{A}^{-1}}(t, v)\mathrm{d}v.$$

The LCT with $\mathbf{A} = (0, 1; -1, 0)$ reduces to the ordinary Fourier transform (FT) $\widehat{f}(\omega)$.

Proposition 1: Let $\widehat{f}(\omega)$ be the FT of $f(t)$, and $\widehat{f}_{\mathbf{A}}(u)$ be the LCT of $f(t)$ with $\mathbf{A}$. Assuming $f(t) \in L^2(\mathbb{R})$. Then it has $\|f\|_2 = \left\|\widehat{f}\right\|_2 = \left\|\widehat{f}_{\mathbf{A}}\right\|_2$.

Definition 2: Let $\widehat{f}_{\mathbf{A}_k}(u)$ be the LCT of a signal $f(t) \in L^2(\mathbb{R})$ with $\mathbf{A}_k = (a_k, b_k; c_k, d_k)$, $k = 1, 2$, and $\mathcal{K}_{\mathbf{A}}(u, \tau)$ be the kernel with $\mathbf{A} = (a, b; c, d)$. The CFLCWD of the signal $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$ is given by [24], [29]–[32]

$$\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u) = \int_{\mathbb{R}} \widehat{f}_{\mathbf{A}_1}\left(t + \frac{\tau}{2}\right)\overline{\widehat{f}_{\mathbf{A}_2}\left(t - \frac{\tau}{2}\right)}\mathcal{K}_{\mathbf{A}}(u, \tau)\mathrm{d}\tau.$$

Remark 1: The CFLCWD with $\mathbf{A}_1 = \mathbf{A}_2 = (1, 0; 0, 1)$ and $\mathbf{A} = (0, 1; -1, 0)$ reduces to the ordinary WD $\mathrm{W}_f(t, \omega)$.

III. Problem formulation

This section introduces some preparatory conceptions on the uncertainty product in CFLCWD domains. It also clarifies the motive and the meaning of the topic research.

A. Definitions of CFLCWD domains spreads and moments

Definition 3: Let $\widehat{f}_{\mathbf{A}}(u)$ be the LCT of $f(t)$ with $\mathbf{A}$. Assuming $f(t), u\widehat{f}_{\mathbf{A}}(u) \in L^2(\mathbb{R})$. Then the LCT domain spread and moment are defined as [40], [41]

$$\triangle\, u_{\mathbf{A}}^2 = \frac{\int_{\mathbb{R}} \left(u - u_{\mathbf{A}}^0\right)^2 \left|\widehat{f}_{\mathbf{A}}(u)\right|^2 \mathrm{d}u}{\|f\|_2^2}, u_{\mathbf{A}}^0 = \frac{\int_{\mathbb{R}} u \left|\widehat{f}_{\mathbf{A}}(u)\right|^2 \mathrm{d}u}{\|f\|_2^2}.$$

Remark 2: The LCT domain spread $\triangle\, u_{\mathbf{A}}^2$ and moment $u_{\mathbf{A}}^0$ reduce respectively to the time domain spread $\triangle\, t^2$ and moment t^0 for $\mathbf{A} = (1, 0; 0, 1)$, and to the FT domain spread $\triangle\, \omega^2$ and moment ω^0 for $\mathbf{A} = (0, 1; -1, 0)$.

Definition 4: Let $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u)$ be the CFLCWD of $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$. Assuming $f(t), t\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u), u\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u) \in L^2(\mathbb{R})$. Then the time-CFLCWD domain spread and moment are defined as

$$\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} = \frac{\iint_{\mathbb{R}^2} \left(t - t^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\right)^2 \left|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u)\right|^2 \mathrm{d}t\mathrm{d}u}{\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2},$$

$$t^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} = \frac{\iint_{\mathbb{R}^2} t \left|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u)\right|^2 \mathrm{d}t\mathrm{d}u}{\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2}.$$

The LCT-CFLCWD domain spread and moment are defined as

$$\triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} = \frac{\iint_{\mathbb{R}^2} \left(u - u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\right)^2 \left|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u)\right|^2 \mathrm{d}t\mathrm{d}u}{\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2},$$

$$u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} = \frac{\iint_{\mathbb{R}^2} u \left|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t, u)\right|^2 \mathrm{d}t\mathrm{d}u}{\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2}.$$

Remark 3: The time(LCT)-CFLCWD domain spread $\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}(\triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}})$ and moment $t^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}(u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}})$ reduce respectively to the time(FT)-WD domain spread $\triangle\, t^2_{\mathrm{W}}(\triangle\, \omega^2_{\mathrm{W}})$ and moment $t^0_{\mathrm{W}}(\omega^0_{\mathrm{W}})$ for $\mathbf{A}_1 = \mathbf{A}_2 = (1, 0; 0, 1)$ and $\mathbf{A} = (0, 1; -1, 0)$.

B. The highest time-frequency resolution of the CFLCWD

Uncertainty principle of the CFLCWD is essentially an inequality that provides an attainable lower bound for the uncertainty product $\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} \triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}$ in CFLCWD domains. This lower bound is usually applied to characterize the CFLCWD's time-frequency resolution limit which is independent of both signal types and noisy types. Uncertainty principle of the CFLCWD is therefore the basic mathematical principle which can reveal the physical mechanism behind the highest-resolution in time-frequency plane triggered by the LCT free parameters.

The paper aims to first formulate an attainable lower bound for $\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} \triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}$, and then investigates the universal optimal parameters with which the CFLCWD reaches the highest-resolution in time-frequency plane by solving the lower bound minimization problem.

IV. Uncertainty principle of the CFLCWD

This section first explores Parseval equality in the CFLCWD domain and the reexpression of the CFLCWD in LCT domains. Then it translates the time(LCT)-CFLCWD domain spread into two LCT domain spreads, based on which it translates the uncertainty product in CFLCWD domains into four uncertainty products in LCT domains. Finally it uses the uncertainty inequality in two LCT domains to deduce the uncertainty inequality in CFLCWD domains.

A. Parseval's relation of the CFLCWD

Lemma 1: Let $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)$ be the CFLCWD of $f(t)\in L^2(\mathbb{R})$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$. Then it has $\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2 = \|f\|_2^2$.

Proof: From Definition 2, there is

$$\begin{aligned}\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2 =& \iiint_{\mathbb{R}^3} \widehat{f}_{\mathbf{A}_1}\left(t+\frac{\tau}{2}\right)\overline{\widehat{f}_{\mathbf{A}_2}\left(t-\frac{\tau}{2}\right)}\\ &\times\overline{\widehat{f}_{\mathbf{A}_1}\left(t+\frac{\theta}{2}\right)}\widehat{f}_{\mathbf{A}_2}\left(t-\frac{\theta}{2}\right)\\ &\times\left(\int_{\mathbb{R}}\mathcal{K}_{\mathbf{A}}(u,\tau)\overline{\mathcal{K}_{\mathbf{A}}(u,\theta)}\mathrm{d}u\right)\mathrm{d}t\mathrm{d}\tau\mathrm{d}\theta.\end{aligned} \tag{1}$$

Thanks to

$$\int_{\mathbb{R}}\mathcal{K}_{\mathbf{A}}(u,\tau)\overline{\mathcal{K}_{\mathbf{A}}(u,\theta)}\mathrm{d}u = \mathrm{e}^{\mathrm{j}\frac{a}{2b}(\tau^2-\theta^2)}\delta(\tau-\theta)$$

and the sifting of Dirac delta operator δ, Eq. (1) becomes

$$\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2 = \iint_{\mathbb{R}^2}\left|\widehat{f}_{\mathbf{A}_1}\left(t+\frac{\tau}{2}\right)\right|^2\left|\widehat{f}_{\mathbf{A}_2}\left(t-\frac{\tau}{2}\right)\right|^2\mathrm{d}t\mathrm{d}\tau. \tag{2}$$

By taking the change of variables $t=\frac{u+v}{2}$ and $\tau=u-v$, there is

$$\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2 = \left\|\widehat{f}_{\mathbf{A}_1}\right\|_2^2\left\|\widehat{f}_{\mathbf{A}_2}\right\|_2^2. \tag{3}$$

Using Proposition 1 yields $\left\|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}\right\|_2^2 = \|f\|_2^4$. ■

Remark 4: From Remark 1, Lemma 1 reduces to Parseval's relation of the WD: $\|\mathrm{W}_f\|_2 = \|f\|_2^2$.

B. Reexpression of CFLCWD in LCT domains

Lemma 2: Let $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)$ be the CFLCWD of $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$, and $\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_1}(u)$ and $\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_2}(u)$ be the LCTs of $f(t)$ with $\breve{\mathbf{A}}\mathbf{A}_1$ and $\breve{\mathbf{A}}\mathbf{A}_2$, respectively, where $\breve{\mathbf{A}}=(a,b/2;2c,d)$ and $\breve{\mathbf{A}}=(a,-b/2;-2c,d)$. Assuming $f(t)\in L^2(\mathbb{R})$. Then, the CFLCWD can be reexpressed in LCT domains as

$$\begin{aligned}&\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\\ =&2\int_{\mathbb{R}}\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_1}(v)\overline{\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_2}(v-u)}\overline{\mathcal{K}_{\mathbf{A}}(2v-u,2t)}\mathrm{d}v.\end{aligned}$$

Proof: From Definition 2, taking the variable substitution $\tau=2(t-\theta)$ gives

$$\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u) = 2\int_{\mathbb{R}}\widehat{f}_{\mathbf{A}_1}(2t-\theta)\overline{\widehat{f}_{\mathbf{A}_2}(\theta)}\mathcal{K}_{\mathbf{A}}(u,2(t-\theta))\mathrm{d}\theta. \tag{4}$$

Thanks to the additivity and invertibility of the LCT, there is

$$\widehat{f}_{\mathbf{A}_1}(2t-\theta) = \int_{\mathbb{R}}\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_1}(v)\mathcal{K}_{\breve{\mathbf{A}}^{-1}}(2t-\theta,v)\mathrm{d}v,$$

where the parameter matrix $\breve{\mathbf{A}}^{-1}=(d,-b/2;-2c,a)$. By substituting into (4), and because of

$$\mathcal{K}_{\mathbf{A}}(u,2(t-\theta)) = \frac{1}{\sqrt{2}}\mathrm{e}^{-\mathrm{j}\left(\frac{d}{2b}u^2+\frac{2a}{b}t^2-\frac{2}{b}ut\right)}\mathrm{e}^{\mathrm{j}\frac{a}{b}\theta^2}\mathcal{K}_{\breve{\mathbf{A}}}(u,2t-\theta),$$

it follows that

$$\begin{aligned}&\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\\ =&\sqrt{2}\mathrm{e}^{-\mathrm{j}\left(\frac{d}{2b}u^2+\frac{2a}{b}t^2-\frac{2}{b}ut\right)}\int_{\mathbb{R}}\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_1}(v)\\ &\times\left(\int_{\mathbb{R}}\overline{\widehat{f}_{\mathbf{A}_2}(\theta)}\mathrm{e}^{\mathrm{j}\frac{a}{b}\theta^2}\mathcal{K}_{\breve{\mathbf{A}}}(u,2t-\theta)\mathcal{K}_{\breve{\mathbf{A}}^{-1}}(2t-\theta,v)\mathrm{d}\theta\right)\mathrm{d}v.\end{aligned} \tag{5}$$

Due to

$$\begin{aligned}&\mathrm{e}^{\mathrm{j}\frac{a}{b}\theta^2}\mathcal{K}_{\breve{\mathbf{A}}}(u,2t-\theta)\mathcal{K}_{\breve{\mathbf{A}}^{-1}}(2t-\theta,v)\\ =&\frac{1}{\sqrt{\mathrm{j}\pi b}}\frac{\overline{\sqrt{-\mathrm{j}\pi b}}}{\sqrt{-\mathrm{j}\pi b}}\mathrm{e}^{\mathrm{j}\frac{d}{b}u^2}\mathrm{e}^{-\mathrm{j}\frac{4}{b}ut}\\ &\times\mathrm{e}^{-\mathrm{j}\frac{d}{b}v^2}\mathrm{e}^{\mathrm{j}\frac{4}{b}vt}\mathrm{e}^{-\mathrm{j}\frac{d}{b}(v-u)^2}\overline{\mathcal{K}_{\breve{\mathbf{A}}}(v-u,\theta)}\end{aligned}$$

and the additivity of the LCT, Eq. (5) turns into

$$\begin{aligned}&\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\\ =&\frac{2}{\sqrt{\mathrm{j}2\pi b}}\frac{\overline{\sqrt{-\mathrm{j}\pi b}}}{\sqrt{-\mathrm{j}\pi b}}\mathrm{e}^{-\mathrm{j}\left(-\frac{d}{2b}u^2+\frac{2a}{b}t^2+\frac{2}{b}ut\right)}\\ &\times\int_{\mathbb{R}}\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_1}(v)\overline{\widehat{f}_{\breve{\mathbf{A}}\mathbf{A}_2}(v-u)}\mathrm{e}^{-\mathrm{j}\frac{d}{b}v^2}\mathrm{e}^{\mathrm{j}\frac{4}{b}vt}\mathrm{e}^{-\mathrm{j}\frac{d}{b}(v-u)^2}\mathrm{d}v.\end{aligned}$$

With a relation

$$\begin{aligned}&\mathrm{e}^{-\mathrm{j}\left(-\frac{d}{2b}u^2+\frac{2a}{b}t^2+\frac{2}{b}ut\right)}\mathrm{e}^{-\mathrm{j}\frac{d}{b}v^2}\mathrm{e}^{\mathrm{j}\frac{4}{b}vt}\mathrm{e}^{-\mathrm{j}\frac{d}{b}(v-u)^2}\\ =&\overline{\sqrt{\mathrm{j}2\pi b}}\,\overline{\mathcal{K}_{\mathbf{A}}(2v-u,2t)},\end{aligned}$$

we arrive the conclusion. ■

Remark 5: From Remark 1 and Lemma 2, the WD can be reexpressed in the FT domain as $\mathrm{W}_f(t,\omega) = \sqrt{\frac{2}{\pi}}\int_{\mathbb{R}}\widehat{f}(v)\overline{\widehat{f}(2\omega-v)}\mathrm{e}^{2\mathrm{j}(v-\omega)t}\mathrm{d}v$.

C. Time-CFLCWD domain spread

Lemma 3: Let $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)$ be the CFLCWD of $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$. Assuming $f(t), u\widehat{f}_{\mathbf{A}_k}(u), t\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\in L^2(\mathbb{R})$, $k=1,2$. Then the time-CFLCWD domain spread can be expressed by the LCT domain spreads:

$$\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} = \frac{\triangle\, u^2_{\mathbf{A}_1}+\triangle\, u^2_{\mathbf{A}_2}}{4}. \tag{6}$$

Proof: Similar to (1)–(3), there is

$$\begin{aligned}&\iint_{\mathbb{R}^2} t\left|\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\right|^2\mathrm{d}t\mathrm{d}u\\ =&\iint_{\mathbb{R}^2} t\left|\widehat{f}_{\mathbf{A}_1}\left(t+\frac{\tau}{2}\right)\right|^2\left|\widehat{f}_{\mathbf{A}_2}\left(t-\frac{\tau}{2}\right)\right|^2\mathrm{d}t\mathrm{d}\tau\\ =&\frac{\iint_{\mathbb{R}^2}(u+v)\left|\widehat{f}_{\mathbf{A}_1}(u)\right|^2\left|\widehat{f}_{\mathbf{A}_2}(v)\right|^2\mathrm{d}u\mathrm{d}v}{2}.\end{aligned}$$

Thanks to Lemma 1, it follows that

$$\begin{aligned}t^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} =&\frac{\iint_{\mathbb{R}^2}(u+v)\left|\widehat{f}_{\mathbf{A}_1}(u)\right|^2\left|\widehat{f}_{\mathbf{A}_2}(v)\right|^2\mathrm{d}u\mathrm{d}v}{2\|f\|_2^4}\\ =&\frac{u^0_{\mathbf{A}_1}+u^0_{\mathbf{A}_2}}{2}.\end{aligned} \tag{7}$$

Similarly, we have

$$\begin{aligned}&\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\\ =&\frac{\iint_{\mathbb{R}^2}\left(\frac{u+v}{2}-t^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\right)^2\left|\widehat{f}_{\mathbf{A}_1}(u)\right|^2\left|\widehat{f}_{\mathbf{A}_2}(v)\right|^2\mathrm{d}u\mathrm{d}v}{\|f\|_2^4},\end{aligned}$$

and then it turns into the conclusion (6) because of (7). ■

Remark 6: From Remarks 2 and 3, Eq. (6) reduces to $\triangle\, t^2_{\mathrm{W}}=\frac{\triangle t^2}{2}$.

D. LCT-CFLCWD domain spread

Lemma 4: Let $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)$ be the CFLCWD of $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$, and $\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(u)$ and $\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(u)$ be the LCTs of $f(t)$ with $\check{\mathbf{A}}\mathbf{A}_1$ and $\check{\mathbf{A}}\mathbf{A}_2$, respectively. Assuming $f(t),u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(u),u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(u),u\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\in L^2(\mathbb{R})$. Then the LCT-CFLCWD domain spread can be expressed by the LCT domain spreads:

$$\triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}=\triangle\, u^2_{\check{\mathbf{A}}\mathbf{A}_1}+\triangle\, u^2_{\check{\mathbf{A}}\mathbf{A}_2}. \tag{8}$$

Proof: From Lemmas 1 and 2, there is

$$\begin{aligned}&u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\\ =&2\iiint_{\mathbb{R}^3}u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(v)\overline{\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(v-u)\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(w)}\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(w-u)\\ &\times\left(\int_{\mathbb{R}}\mathcal{K}_{\mathbf{A}}(2w-u,t)\overline{\mathcal{K}_{\mathbf{A}}(2v-u,t)}\mathrm{d}t\right)\mathrm{d}u\mathrm{d}v\mathrm{d}w\Big/\|f\|_2^4.\end{aligned} \tag{9}$$

By using

$$\begin{aligned}&\int_{\mathbb{R}}\mathcal{K}_{\mathbf{A}}(2w-u,t)\overline{\mathcal{K}_{\mathbf{A}}(2v-u,t)}\mathrm{d}t\\ =&\mathrm{e}^{\mathrm{j}\frac{d}{2b}\left[(2w-u)^2-(2v-u)^2\right]}\delta(2w-2v)\end{aligned}$$

and the sifting of Dirac delta operator δ, Eq. (9) turns into

$$u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}=\frac{\iint_{\mathbb{R}^2}u\left|\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(w)\right|^2\left|\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(w-u)\right|^2\mathrm{d}u\mathrm{d}w}{\|f\|_2^4}.$$

Taking the change of variable $u=-\mu+w$ yields

$$\begin{aligned}u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}=&\frac{\iint_{\mathbb{R}^2}(w-\mu)\left|\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(w)\right|^2\left|\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(\mu)\right|^2\mathrm{d}w\mathrm{d}\mu}{\|f\|_2^4}\\ =&u^0_{\check{\mathbf{A}}\mathbf{A}_1}-u^0_{\check{\mathbf{A}}\mathbf{A}_2}.\end{aligned} \tag{10}$$

Similarly, we have

$$\begin{aligned}\triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}=&\iint_{\mathbb{R}^2}\left(w-\mu-u^0_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\right)^2\\ &\times\left|\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(w)\right|^2\left|\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(\mu)\right|^2\mathrm{d}w\mathrm{d}\mu\Big/\|f\|_2^4,\end{aligned}$$

and then it becomes the conclusion (8) because of (10). ■

Remark 7: From Remarks 2 and 3, Eq. (8) reduces to $\triangle\, \omega^2_{\mathrm{W}}=\frac{\triangle\omega^2}{2}$.

E. Uncertainty product of the time-CFLCWD domain and the LCT-CFLCWD domain spreads

Lemma 5: Let $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)$ be the CFLCWD of $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$, and $\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(u)$ and $\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(u)$ be the LCTs of $f(t)$ with $\check{\mathbf{A}}\mathbf{A}_1$ and $\check{\mathbf{A}}\mathbf{A}_2$, respectively. Assuming $f(t),u\widehat{f}_{\mathbf{A}_1}(u),u\widehat{f}_{\mathbf{A}_2}(u),u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(u),u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(u),t\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u),u\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\in L^2(\mathbb{R})$. Then the uncertainty product in CFLCWD domains can be expressed by uncertainty products in LCT domains:

$$\begin{aligned}&\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\\ =&\Big(\triangle\, u^2_{\mathbf{A}_1}\triangle\, u^2_{\check{\mathbf{A}}\mathbf{A}_1}+\triangle\, u^2_{\mathbf{A}_1}\triangle\, u^2_{\check{\mathbf{A}}\mathbf{A}_2}\\ &+\triangle\, u^2_{\mathbf{A}_2}\triangle\, u^2_{\check{\mathbf{A}}\mathbf{A}_1}+\triangle\, u^2_{\mathbf{A}_2}\triangle\, u^2_{\check{\mathbf{A}}\mathbf{A}_2}\Big)\Big/4.\end{aligned} \tag{11}$$

Proof: From Lemmas 3 and 4, multiplying (6) and (8) together yields the conclusion (11). ■

Remark 8: As stated in Remarks 6 and 7, Eq. (11) can reduce to $\triangle\, t^2_{\mathrm{W}}\triangle\, \omega^2_{\mathrm{W}}=\frac{\triangle t^2\triangle\omega^2}{4}$.

F. Uncertainty inequality of the CFLCWD

Theorem 1: Let $\widehat{f}(\omega)$ be the FT of a signal $f(t)$, $\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)$ be the CFLCWD of $f(t)$ with $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}$, and $\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(u)$ and $\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(u)$ be the LCTs of $f(t)$ with $\check{\mathbf{A}}\mathbf{A}_1$ and $\check{\mathbf{A}}\mathbf{A}_2$, respectively. Assuming the conventional derivative $\frac{\mathrm{d}f}{\mathrm{d}t}$ exists at any point $t\in\mathbb{R}$, and $f(t),tf(t),\omega\widehat{f}(\omega),u\widehat{f}_{\mathbf{A}_1}(u),u\widehat{f}_{\mathbf{A}_2}(u),u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_1}(u),u\widehat{f}_{\check{\mathbf{A}}\mathbf{A}_2}(u),t\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u),u\mathrm{W}_f^{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}(t,u)\in L^2(\mathbb{R})$. Then it has

$$\triangle\, t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\triangle\, u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}\geq\mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}, \tag{12}$$

where

$$\begin{aligned}\mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}=&\frac{b^2}{32}+\frac{1}{16}\left[a\left(a_1b_2-a_2b_1\right)-\frac{b}{2}\left(a_1d_2-b_1c_2\right)\right]^2\\ &+\frac{1}{16}\left[a\left(a_2b_1-a_1b_2\right)+\frac{b}{2}\left(a_2d_1-b_2c_1\right)\right]^2.\end{aligned}$$

Assuming $f(t)\neq 0$ almost everywhere, the equality holds iff $f(t)$ is a Gaussian signal $\mathrm{e}^{-\frac{1}{2\zeta}\left(t-t^0\right)^2+\sigma}\mathrm{e}^{\mathrm{j}\left(\omega^0t+\kappa\right)}$, where $\zeta>0$ and $\sigma,\kappa\in\mathbb{R}$.

Proof: $\check{\mathbf{A}}\mathbf{A}_1=\begin{pmatrix}aa_1+\frac{bc_1}{2} & ab_1+\frac{bd_1}{2}\\ 2ca_1+dc_1 & 2cb_1+dd_1\end{pmatrix}$ and $\check{\mathbf{A}}\mathbf{A}_2=\begin{pmatrix}aa_2-\frac{bc_2}{2} & ab_2-\frac{bd_2}{2}\\ -2ca_2+dc_2 & -2cb_2+dd_2\end{pmatrix}$. Thanks to Lemma 5 and the uncertainty inequality $\triangle\, u^2_{\mathbf{A}_1}\triangle\, u^2_{\mathbf{A}_2}\geq\frac{\left(a_1b_2-a_2b_1\right)^2}{4}$ in two LCT domains introduced by J. Zhao, R. Tao and their collaborators [40], [41], we arrive the conclusion (12). The equality of (12) is achieved iff the equality $\triangle\, u^2_{\mathbf{A}_1}\triangle\, u^2_{\mathbf{A}_2}=\frac{\left(a_1b_2-a_2b_1\right)^2}{4}$ is achieved, or equivalently, the equality of the classical Heisenberg's uncertainty inequality $\triangle\, t^2\triangle\, \omega^2\geq\frac{1}{4}$ [13], [14] is achieved. The equality $\triangle\, t^2\triangle\, \omega^2=\frac{1}{4}$ holds iff $f(t)$ is a Gaussian signal $\mathrm{e}^{-\frac{1}{2\zeta}\left(t-t^0\right)^2+\sigma}\mathrm{e}^{\mathrm{j}\left(\omega^0t+\kappa\right)}$. ■

Remark 9: As stated in Remark 8, the inequality (12) can reduce to the uncertainty inequality of the WD: $\triangle\, t^2_{\mathrm{W}}\triangle\, \omega^2_{\mathrm{W}}\geq\frac{1}{16}$.

V. TIME-FREQUENCY ANALYSIS OF THE CFLCWD

This section treats the lower bound minimization problem $\min_{a_kd_k-b_kc_k=1,k=1,2} \mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}$ as a conditional extremum problem and solves it to generate the universal optimal LCT free parameters with which the CFLCWD reaches the highest-resolution in time-frequency plane.

Theorem 2: The lower bound $\mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}$ is a function of a_k, b_k, c_k, d_k, $k = 1, 2$ and a, b. Its minimum is unattainable.

Proof: By using Lagrange multipliers method to solve the conditional extremum problem: Minimize $\mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}$; Subject to $a_kd_k - b_kc_k = 1$, $k = 1, 2$, the auxiliary function follows $\mathcal{L} = \mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}} + \sum_{k=1}^{2} \rho_k(a_kd_k - b_kc_k - 1)$.

For the case of $a_1b_2 - a_2b_1 = 0$, there are $a_1d_2 - b_1c_2 = \frac{b_1}{b_2}$ and $a_2d_1 - b_2c_1 = \frac{b_2}{b_1}$, based on which the equation $\frac{\partial\mathcal{L}}{\partial b} = 0$ becomes a contradictory equation $2 + \frac{b_1^2}{b_2^2} + \frac{b_2^2}{b_1^2} = 0$. As for the case of $a_1b_2 - a_2b_1 \neq 0$, combining $\frac{\partial\mathcal{L}}{\partial c_1} = 0$ and $\frac{\partial\mathcal{L}}{\partial d_1} = 0$ gives $a(a_2b_1 - a_1b_2) + \frac{b}{2}(a_2d_1 - b_2c_1) = 0$; combining $\frac{\partial\mathcal{L}}{\partial c_2} = 0$ and $\frac{\partial\mathcal{L}}{\partial d_2} = 0$ gives $a(a_1b_2 - a_2b_1) - \frac{b}{2}(a_1d_2 - b_1c_2) = 0$; then the equation $\frac{\partial\mathcal{L}}{\partial b} = 0$ becomes a contradictory equation $b = 0$. In a word, $\nabla\mathcal{L} = \overrightarrow{0}$ has no solution, where ∇ denotes gradient operator. ■

Remark 10: In general, there are no universal optimal LCT parameters with which the CFLCWD reaches the highest-resolution in time-frequency plane.

Theorem 3: Assuming b is a nonzero constant. Then the CFLCWD reaches the highest-resolution in time-frequency plane iff $\frac{a_1d_2-b_1c_2}{a_1b_2-a_2b_1} = \frac{a_2d_1-b_2c_1}{a_1b_2-a_2b_1} = \frac{2a}{b}$.

Proof: Since b is a nonzero constant, the lower bound $\mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}$ reaches its minimum $\frac{b^2}{32}$ iff $a(a_1b_2 - a_2b_1) - \frac{b}{2}(a_1d_2 - b_1c_2) = a(a_2b_1 - a_1b_2) + \frac{b}{2}(a_2d_1 - b_2c_1) = 0$. Note that $a_1b_2 - a_2b_1 \neq 0$, otherwise, there is a contradictory continued equality $\frac{b_1}{b_2} = \frac{b_2}{b_1} = 0$. ■

Remark 11: When b is a nonzero constant, there is a universal optimal LCT free parameters set

$$\left\{a_k, b_k, d_k, a, b, d, k = 1, 2 \middle| \frac{d_2}{b_2} + \frac{b_1}{b_2(a_1b_2 - a_2b_1)} = -\frac{d_1}{b_1} + \frac{b_2}{b_1(a_1b_2 - a_2b_1)} = \frac{2a}{b}, b \text{ is constant}, d \in \mathbb{R}\right\}$$

with which the CFLCWD reaches the highest-resolution in time-frequency plane.

VI. CONCLUSIONS

Uncertainty principle of the CFLCWD was formulated, based on which its time-frequency analysis theory was established. It turns out that the uncertainty product $\triangle\ t^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}} \triangle\ u^2_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A},\mathrm{W}}$ in CFLCWD domains equals to the sum of four uncertainty products $\triangle\ u^2_{\mathbf{A}_1} \triangle\ u^2_{\breve{\mathbf{A}}\mathbf{A}_1}, \triangle\ u^2_{\mathbf{A}_1} \triangle\ u^2_{\breve{\mathbf{A}}\mathbf{A}_2}, \triangle\ u^2_{\mathbf{A}_2} \triangle\ u^2_{\breve{\mathbf{A}}\mathbf{A}_1}, \triangle\ u^2_{\mathbf{A}_2} \triangle\ u^2_{\breve{\mathbf{A}}\mathbf{A}_2}$ in LCT domains. The lower bound on the uncertainty product in CFLCWD domains is shown as $\mathrm{B}_{\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}}$, which is attainable by a Gaussian signal and depends only on the LCT free parameters $a_k, b_k, d_k, a, b, k = 1, 2$. The universal optimal LCT free parameters set of the CFLCWD is given by $\left\{a_k, b_k, d_k, a, b, d, k = 1, 2 \middle| \frac{d_2}{b_2} + \frac{b_1}{b_2(a_1b_2-a_2b_1)} = -\frac{d_1}{b_1} + \frac{b_2}{b_1(a_1b_2-a_2b_1)} = \frac{2a}{b}, b \text{ is constant}, d \in \mathbb{R}\right\}$, in which the highest time-frequency resolution is achieved.

REFERENCES

[1] L. Cohen, "Time-frequency distributions-a review," *Proc. IEEE*, vol. 77, no. 7, pp. 941–981, Jul. 1989.

[2] L. Cohen, *Time-Frequency Analysis: Theory and Applications.* New Jersey, NJ, USA: Prentice Hall, 1995.

[3] B. Boashash, "Note on the use of the Wigner distribution for time-frequency signal analysis," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 9, pp. 1518–1521, Sep. 1988.

[4] L. Stanković, S. Stanković, and M. Daković, "From the STFT to the Wigner distribution [lecture notes]," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 163–174, May 2014.

[5] G. B. Folland and A. Sitaram, "The uncertainty principle: a mathematical survey," *J. Fourier Anal. Appl.*, vol. 3, no. 3, pp. 207–238, May 1997.

[6] K. Gröchenig, "Uncertainty principles for time-frequency representations," in: *Advances in Gabor Analysis.* H. G. Feichtinger and T. Strohmer, Eds., New York, NY, USA: Springer Science+Business Media, pp. 11–30, 2003.

[7] P. Korn, "Some uncertainty principles for time-frequency transforms of the Cohen class," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 523–527, Feb. 2005.

[8] P. Boggiatto, G. De Donno, and A. Oliaro, "Time-frequency representations of Wigner type and pseudo-differential operators," *Trans. Am. Math. Soc.*, vol. 362, no. 9, pp. 4955–4981, Sep. 2010.

[9] P. Boggiatto, E. Carypis, and A. Oliaro, "Windowed-Wigner representations in the Cohen class and uncertainty principles," *J. Geom. Anal.*, vol. 23, no. 4, pp. 1753–1779, Oct. 2013.

[10] P. Boggiatto, E. Carypis, and A. Oliaro, "Local uncertainty principles for the Cohen class," *J. Math. Anal. Appl.*, vol. 419, no. 2, pp. 1004–1022, Nov. 2014.

[11] Y. G. Li, B. Z. Li, and H. F. Sun, "Uncertainty principles for Wigner-Ville distribution associated with the linear canonical transforms," *Abstr. Appl. Anal.*, vol. 2014, Art. No. 470459, May 2014.

[12] N. C. Dias, M. A. de Gosson, and J. N. Prata, "A refinement of the Robertson-Schrödinger uncertainty principle and a Hirschman-Shannon inequality for Wigner distributions," *J. Fourier Anal. Appl.*, vol. 25, no. 1, pp. 210–241, Feb. 2019.

[13] W. Heisenberg, "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik," *Z. Phys.*, vol. 43, no. 3, pp. 172–198, Mar. 1927.

[14] D. Gabor, "Theory of communication," *J. Inst. Electr. Eng.*, vol. 93, no. 26, pp. 429–457, Nov. 1946.

[15] P. Dang, G. T. Deng, and T. Qian, "A sharper uncertainty principle," *J. Funct. Anal.*, vol. 265, no. 10, pp. 2239–2266, Nov. 2013.

[16] Z. C. Zhang, X. Y. Shi, A. Y. Wu, and D. Li, "Sharper N-d Heisenberg's uncertainty principle," *IEEE Signal Process. Lett.*, vol. 28, pp. 1665–1669, Jul. 2021.

[17] S. C. Pei and J. J. Ding, "Relations between fractional operations and time-frequency distributions and their applications," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1638–1655, Aug. 2001.

[18] D. Y. Wei and Y. M. Li, "Linear canonical Wigner distribution and its application," *Optik*, vol. 125, no. 1, pp. 89–92, Jan. 2014.

[19] R. F. Bai, B. Z. Li, and Q. Y. Cheng, "Wigner-Ville distribution associated with the linear canonical transform," *J. Appl. Math.*, vol. 2012, Art. No. 740161, Jul. 2012.

[20] Y. E. Song, X. Y. Zhang, C. H. Shang, H. X. Bu, and X. Y. Wang, "The Wigner-Ville distribution based on the linear canonical transform and its applications for QFM signal parameters estimation," *J. Appl. Math.*, vol. 2014, Art. No. 516457, May 2014.

[21] Z. C. Zhang, "New Wigner distribution and ambiguity function based on the generalized translation in the linear canonical transform domain," *Signal Process.*, vol. 118, pp. 51–61, Jan. 2016.

[22] H. C. Xin and B. Z. Li, "On a new Wigner-Ville distribution associated with linear canonical transform," *EURASIP J. Adv. Signal Process.*, vol. 2021, Art. No. 56, Aug. 2021.

[23] Z. C. Zhang, "Unified Wigner-Ville distribution and ambiguity function in the linear canonical transform domain," *Signal Process.*, vol. 114, pp. 45–60, Sep. 2015.
[24] Z. C. Zhang and M. K. Luo, "New integral transforms for generalizing the Wigner distribution and ambiguity function," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 460–464, Apr. 2015.
[25] W. C. Zhu and Z. C. Zhang, "Instantaneous cross-correlation function-τ-Wigner distribution: theory and application," *Signal Process.*, vol. 208, Art. No. 108993, Jul. 2023.
[26] Z. C. Zhang and X. Y. Shi, "Kernel function-τ-Wigner distribution associated with the linear canonical transform," *IEEE Signal Process. Lett.*, vol. 29, pp. 1764–1768, Aug. 2022.
[27] Z. C. Zhang and Y. F. He, "Wigner distribution associated with the symplectic coordinates transformation," *Signal Process.*, vol. 204, Art. No. 108846, Mar. 2023.
[28] Z. C. Zhang, D. Li, Y. F. He, Y. J. Chen, J. W. Zhang, and C. X. Zhou, "K-Wigner distribution: definition, uncertainty principles and time-frequency analysis," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, pp. 2722–2736, Apr. 2023.
[29] I. Djurović and V. N. Ivanović, "Analysis of closed-form Wigner distribution in linear canonical domains for continuous-time noisy signals," in: *Proc. 9th Mediterranean Conf. Embedded Comput. (MECO)*, pp. 1–4, Jul. 2020.
[30] Z. C. Zhang, "Linear canonical Wigner distribution based noisy LFM signals detection through the output SNR improvement analysis," *IEEE Trans. Signal Process.*, vol. 67, no. 21, pp. 5527–5542, Nov. 2019.
[31] Z. C. Zhang, "The optimal linear canonical Wigner distribution of noisy linear frequency-modulated signals," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1127–1131, Aug. 2019.
[32] Z. C. Zhang, D. Li, Y. J. Chen, and J. W. Zhang, "Linear canonical Wigner distribution of noisy LFM signals via multiobjective optimization analysis involving variance-SNR," *IEEE Commun. Lett.*, vol. 25, no. 2, pp. 546–550, Feb. 2021.
[33] X. Y. Shi, A. Y. Wu, Y. Sun, S. Z. Qiang, X. Jiang, P. Y. Han, Y. J. Chen, and Z. C. Zhang, "Unique parameters selection strategy of linear canonical Wigner distribution via multiobjective optimization modeling," *Chinese J. Electron.*, vol. 32, no. 3, pp. 453–464, May 2023.
[34] Y. C. Eldar, "Uncertainty relations for shift-invariant analog signals," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5742–5757, Dec. 2009.
[35] R. Tao, J. M. Ma, B. Deng, and Y. Wang, *Fractional Fourier Transform and its Applications (2nd Ed.)*. Beijing, China: Tsinghua Univ. Press, 2022.
[36] B. Z. Li, Y. N. Sun, Y. N. Zhang, Y. Guo, and Q. Feng, *Mathematical Transformation Methods in Signal Processing*. Beijing, China: Beijing Inst. Tech. Press, 2020.
[37] J. Shi, X. P. Liu, F. G. Yan, and W. B. Song, "Error analysis of reconstruction from linear canonical transform based sampling," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1748–1760, Apr. 2018.
[38] J. Shi, X. P. Liu, Y. N. Zhao, S. Shi, X. J. Sha, and Q. Y. Zhang, "Filter design for constrained signal reconstruction in linear canonical transform domain," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6534–6548, Dec. 2018.
[39] D. Y. Wei, Y. J. Zhang, and Y. M. Li, "Linear canonical Stockwell transform: theory and applications," *IEEE Trans. Signal Process.*, vol. 70, pp. 1333–1347, Feb. 2022.
[40] J. Zhao, R. Tao, Y. L. Li, and Y. Wang, "Uncertainty principles for linear canonical transform," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2856–2858, Jul. 2009.
[41] R. Tao and J. Zhao, "Uncertainty principles and the linear canonical transform," in: *Linear Canonical Transforms: Theory and Applications*. J. J. Healy, M. A. Kutay, H. M. Ozaktas, and J. T. Sheridan, Eds., New York, NY, USA: Springer, pp. 97–111, 2016.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Autoencoder-Enhanced Federated Learning with Reduced Overhead and Lower Latency

Chi-Kai Hsieh*, Feng-Tsun Chien* and Min-Kuan Chang†
* National Yang Ming Chiao Tung University, Taiwan
E-mail: {mee29616243.ee08, ftchien}@nycu.edu.tw
† National Chung Hsing University, Taiwan
E-mail: minkuanc@dragon.nchu.edu.tw

***Abstract*—This paper investigates the application of autoencoder (AE) in supporting the training process of federated learning by reducing communication overhead and latency. We propose a scheduling algorithm to determine when and how to use autoencoder during training. Our simulation shows that federated learning with an autoencoder significantly reduces communication overhead without compromising testing accuracy. Moreover, the testing accuracy curve shows a more consistent increase over training rounds in federated learning with an autoencoder than in federated learning without an autoencoder. Additionally, the latency of federated learning with an autoencoder is lower than that of federated learning without an autoencoder.**

***Index Terms*—Federated learning, autoencoder, scheduling.**

I. Introduction

In recent years, federated learning has become a popular method of distributed learning. To protect the privacy of local devices and to reduce the communication resources, transmitting local model weights to a server is more feasible than transmitting data in a distributed learning setting [1]. Nonetheless, the complexity of transmitting model weights to the centralized server could still be very high, especially for applications demanding stringent performance measures (e.g. high accuracy or low mean-square error), consuming a significant amount of communication resources.

Several works have studied the problem of reducing required communication resources in wireless federated learning [1], [2], [3], [4]. In [1], the authors propose FedAvg, which leverages over-the-air computation to reduce the computation cost of aggregation. In [2], the authors optimize the number of iterations to reduce the computation consumption of both local devices and the server. Alternatively, compressing the required dimension of transmitting signals is a promising approach to reducing transmission overheads in wireless federated learning. While classical signal compression can directly address this issue, using lossless compression may not be the most effective way here since the updated information is only from the local model weights at each communication round and not all model weights are equally informative. In [3], [4], the authors replace the local model weights with the prompt to reduce the communication overhead between the local devices and the server. However, the server needs to have the pretrained global model before federated learning.

On the other hand, autoencoder (AE) offers a more efficient way of compression from the machine learning perspective, even though it is lossy, which features signal denoising capability. As an unsupervised learning model, autoencoder consists of an encoder and a decoder, where dimensionality reduction is realized at the encoder and the compressed data after the encoder is reconstructed at the decoder, e.g., [5], [6], [7] for image recovery applications, [8] for channel state information (CSI) recovery and [9] for beamforming vector design.

Despite the benefits of using an autoencoder for dimension reduction and denoising in various applications, little research effort is invested in applications for federated learning. Only one work can be found in the literature [10], which presents a novel approach for detecting and mitigating the impact of malicious model updates in federated learning using a conditional variational autoencoder. However, the paper does not discuss the specifics of when and how to train the autoencoder for this application. Thanks to the properties of dimensionality reduction and denoising, we are therefore motivated to employ AE in the wireless federated learning to reduce the transmission overheads and to alleviate the effect of heterogeneity in data and local devices by the denoising property. In particular, we develop protocols that allow the AE to update its encoder/decoder parameters during the federated learning training process and establish a local device scheduling algorithm to improve data reconstruction performance with low latency

The contributions of this research can be summarized as follows:

- To our best knowledge, this work presents the first study to use AE in federated learning for more efficient model parameters transmissions and for consistently improving learning performance by leveraging its denoising capability. Specifically, this lowers the number of model parameters needed for transmission compared to vanilla federated learning, thereby reducing communication overhead overall latency of the federated learning process.
- We introduce a scheduling for AE-enhanced FL (SAEL) algorithm to determine when to utilize the AE during the federated learning training process to validate the feasibility of this approach with reduced overall latency. This allows the central

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

server to make decisions on whether or not local devices should activate the autoencoder.
- We conduct extensive simulations that demonstrate the significant decrease in communication consumption with the use of an autoencoder, highlighting its potential for application in real-world federated learning scenarios.

The remainder of this paper is organized as follows. Section II introduces the structure of federated learning and autoencoder. Section III proposes the protocol and scheduling algorithm of federated learning with an autoencoder. Reduced communication overhead of the proposed protocol is analyzed in Section IV. In Section V, we present simulation results to verify the feasibility and effectiveness of our proposed method. Finally, Section VI provides concluding remarks and suggests future research directions.

II. System Model and Problem Formulation

Federated learning is a distributed learning architecture where a central server coordinates multiple local devices to train a shared machine learning model without sharing their raw data. Instead, the local devices train their own models using their own data and upload the trained model weights to the server. Then, the central server aggregates the model parameters from each local device to update the shared model [11]. More specifically, a *communication round* in a federated learning consists of three steps: global model broadcasting, local model updates, and global model aggregation. The communication round goes on until convergence.

In this paper, we combine federated learning with AE to improve the efficiency of the training process and the effectiveness of the training result. In particular, AE can reduce the dimensionality of model parameters being transmitted to the server in federated learning, which facilitates communication overhead reduction. Additionally, denoising capability in AE can be utilized to ensure consistency of model updates during federated learning training. By leveraging the strengths of both techniques, the resulting federated learning process can achieve better performance and be more applicable in real-world scenarios.

Autoencoder is a type of neural network that consists of an encoder and a decoder [12]. The encoder compresses the input data into a low-dimensional representation, while the decoder reconstructs the original input from the compressed representation.

A. Federated learning

The federated learning process starts by initializing the global model weights $\mathbf{w}$ on the server. In each communication round, a subset of local devices may be selected to participate in the training process. Each selected local device k updates its own model weights $\mathbf{w}_k$ by minimizing the local loss function $f_k\left(\mathbf{w}, \mathbf{x}_{k,m}, y_{k,m}\right)$ over its dataset D_k, where $\mathbf{x}_{k,m}$ is the mth data sample associated with label $y_{k,m}$. Then, the updated local model weights are transmitted to the server for aggregation, which computes the new global model weights as a weighted average of the received local model weights, where the weight for the kth edge device is the ratio of the number of data samples in its dataset to the total number of data samples from all devices. This process is repeated until convergence.

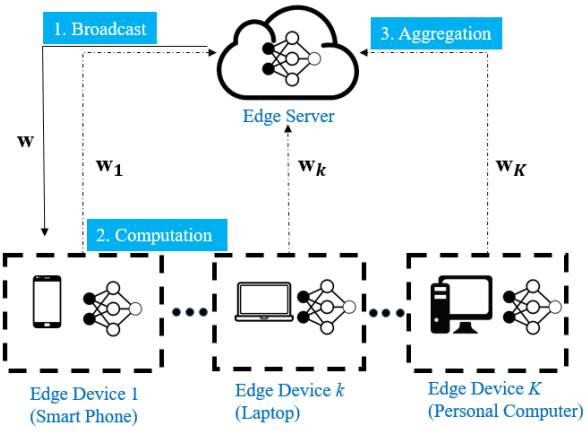


Fig. 1. An illustration for a federated learning network.

Examples of loss function

Model	Loss function
Linear regression	$\frac{1}{2}\left\|y_m - \mathbf{w}^T\mathbf{x}_m\right\|^2$
Cross-Entropy	$-\sum_{class} y_m \log(p(y_m \mid \mathbf{x}_m, \mathbf{w}))$

The global loss function in federated learning can be computed as a weighted sum of local loss functions from each edge device. Specifically, the global loss function can be expressed as

$$F(\mathbf{w}) = \frac{1}{|D|}\sum_{k=1}^{K}\sum_{m=1}^{|D_k|} f_k\left(\mathbf{w}, \mathbf{x}_{k,m}, y_{k,m}\right) \quad (1)$$

$$= \sum_{k=1}^{K}\frac{|D_k|}{\sum_{k'=1}^{K}|D_{k'}|}F_k(\mathbf{w}), \quad (2)$$

where $|D|$ is the total number of data samples in all local datasets and $F_k(\mathbf{w})$ is the overall local loss function of the kth edge node accounting for the entire local data.

The objective of federated learning is to minimize the global loss function by collaboratively training the model weights across all local devices. Mathematically, the learning problem can be expressed as

$$\min_{\{\mathbf{w}\}} \sum_{k=1}^{K}\frac{|D_k|}{\sum_{k'=1}^{K}|D_{k'}|}F_k(\mathbf{w}). \quad (3)$$

Figure 1 illustrates the process of training the global model weights by federated learning, which can be summarized as follows:

1) The central server broadcasts the current global model weights to all local devices.
2) Each local device computes its local model weights using its own local training dataset.
3) The local devices send their updated model weights to the central server, which aggregates the received model weights and produces an updated global model weights.

The above steps are repeated until convergence or that a predefined stopping criterion is met.

B. Autoencoder

An autoencoder is a type of neural network used for unsupervised learning, which is composed of two parts: an encoder and a decoder. The encoder takes the input data and maps it to a lower-dimensional representation, while the decoder takes the encoded representation and maps it back to the original input space. The objective of an autoencoder is to minimize the reconstruction error between the input and the output of the decoder. By doing so, an autoencoder can learn a compressed representation of the input data that captures its most important features.

III. Autoencoder-enhanced Federated Learning

In this section, we introduce the proposed AE-enhanced federated learning, in which an AE encoder is built in each local device to compress the parameters of the learning model before sending to the server. Particularly, instead of training the AE encoder on each local device, we relegate the training process to the centralized server. Therefore, a complete AE encoder and decoder have to be built in the server where the decoder is used to reconstruct the encoded model parameters sent from local devices and the AE encoder at the server is used for the purpose of training AE. The training process of AE at the server requires uncompressed model parameters sent from each device, and the trained AE model will be broadcast to all local devices from the server. In other words, an initial training process of the AE must take place before the AE on the local device can serve to compress the model parameters in federated learning.

The training process in the beginning for training AE and the global model parameters in federated learning is illustrated in Fig. 3. The server collects uncompressed model weights from local devices as in a typical federated learning process. The AE is not activated in this phase. Having trained with sufficient amount of data[1], the AE on the local device can be activated to compress the model parameters before sending to the server. Fig. 4 illustrates the training process of federated learning enhanced by an AE. Once each local device starts using the AE encoder during the training process of federated learning, the server can still continue to collect the model weights from local devices and employ these weights to update the AE.

A. Scheduling for AE-Enhanced FL (SAEF)

In the beginning phase, to update the autoencoder, original uncompressed local model weights are required to feed the AE at the server for training. To continuously improve the reconstruction performance of the AE, the server can utilized the reconstructed model parameters to futher update the AE, which however has limited

[1]This can be notified by the server when the loss value of the AE training starts to be lower than a predefined threshold

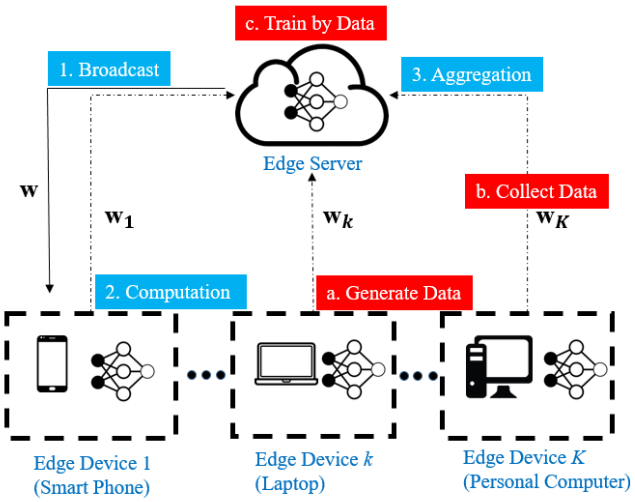


Fig. 2. An illustration of updating AE where the AE training takes place at the server. In AE-enhanced federated learning AE encoder compresses the local model parameters at local devices and AE decoder performs model reconstruction at the server.

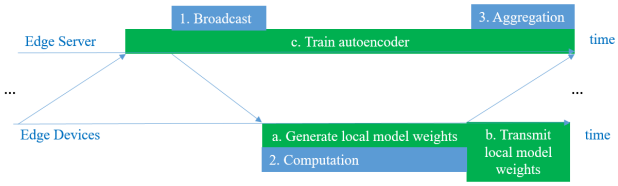


Fig. 3. Training process of the autoencoder. (Time slot.)

performance gain. Alternatively, a selected set of the local devices is scheduled to send uncompressed local model parameters to the server, while the remaining local devices are designated to transmit the compressed local model parameters using AE (after AE is activated) in federated learning.

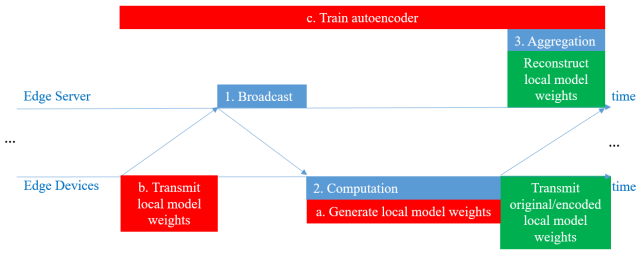


Fig. 4. Training process of federated learning.

The decision of whether a local device is selected to utilize AE for data compression in federated learning is made with consideration of latency in each communication round. Let the latency in the jth communication round be denoted by t^j_{round}. The scheduling mechanism of the AE-enhanced FL (SAEF) determines whether the kth local device should utilize the AE to compress the model parameters or send the uncompressed directly to the server such that t^j_{round} can be minimized. For the kth

local device, the latency can be evaluated by

$$t_{j,k}^{\text{cp}} + \frac{S_{j,k}}{r_{j,k}},$$

where, all in the jth communication round, $t_{j,k}^{\text{cp}}$ is the computation latency of the kth device, $S_{j,k}$ is the data size to be transmitted to server (either compressed or uncompressed) transmission data of the kth device, and $r_{j,k}$ is the transmission rate of the kth device

$$r_{j,k} = B \cdot \log_2 \left(1 + \frac{P_k |h_{j,k}|^2}{B N_0}\right), \tag{4}$$

with B the bandwidth, P_k the transmitted power, $h_{j,k}$ the corresponding channel between the kth local device and server, and N_0 is the noise power spectral density. The device scheduling result determines the data size $S_{j,k}$ and consequently has impact on the overall latency in each round.

B. SAEF Algorithm

In this paper, we propose the following SAEF algorithm to select local devices to transmit compressed local model parameters.

Algorithm 1 SAEF Algorithm

Input: $t_{j,k}^{\text{cp}}$, $r_{j,k}$, $S_{j,k}$,S^{AE} for $k = 1, 2, \ldots, K$
Output: $\hat{t}_{\text{round}}^{j}$, $S_{j,k}$
Find the maximum $t_{j,k'}^{\text{cp}}$ at k'-th local device
Let $t_{max} = t_{j,k'}^{\text{cp}}$
$\hat{t}_{\text{round}}^{j} = t_{max} + \frac{S^{AE}}{r_{j,k'}}$
parallel $k = 1, 2, \ldots, K$ **do**
    if $t_{j,k}^{\text{cp}} + \frac{S}{r_{j,k}} \geq \hat{t}_{\text{round}}^{j}$ **then**
        $S_{j,k} = S^{AE}$
    else
        $S_{j,k} = S_{j,k}$
end parallel

The SAEF algorithm is designed to address latency issues caused by stragglers, which are the primary contributors to overall latency. In particular, the local devices with the longer latency are (and should be) more likely to be selected to utilize AE encoder for data compression, where the data size of the encoded model parameters are denoted by S_{AE}. If the total time spend by a local device for data transmission is greater than $\hat{t}_{\text{round}}^{j}$, which is defined in the pseudo-code Algorithm 1 and calculated according to the longest computation latency among all local devices, the local device needs to utilize AE to reduce the communication latency. Specifically, we first identify the local device k' with the largest $t_{j,k'}^{\text{cp}}$ and compute $t_{max} \triangleq t_{j,k'}^{\text{cp}} + \frac{S^{AE}}{r_{j,k'}}$. We then can decide whether to apply autoencoder or not to other devices based on t_{max}.

IV. Performance Analysis

Because the autoencoder is trained by the server, the local devices only need to receive the autoencoder from the server. This does not raise any privacy concerns for the local devices. The autoencoder can help in reducing the uploaded communication cost. Therefore, we analyze the impact of using AE-enhanced federated learning on the communication overhead, specifically with respect to three AE-training strategies: 1) Separately training AE and global model (STAG) , 2) and 3) the SAEL scheduling. Let the number of hidden nodes in the AE be denoted by L, and the total number of local model weights is M.

A. Separately Training AE and Global Model (STAG)

A properly trained AE ensures that the reconstructed input of the autoencoder closely resembles the original input. One way of training the AE-enhanced FL is to train AE and global model separately (STAG), as illustrated in Fig. 5. The server collects the local model weights from federated learning and updates the model weights of the autoencoder until the AE is properly trained.

The STAG protocol can be summarized as follows:

1) The server broadcasts the global model parameters along with the AE encoder parameters to all local devices.
2) Each local device updates its local model weights using its local dataset.
3) The local devices use the encoder to compress their local model weights and transmit the encoded weights to the server.
4) The server receives the encoded local model weights from all devices.
5) After decoding the encoded local model weights, the server updates the global model by the reconstructed local model weights.

Steps 2 to 5 are repeated until all local datasets have been used in training. The training process of federated learning with the STAG protocol is illustrated in Fig. 6, where the AE parameters are not updated in this phase.

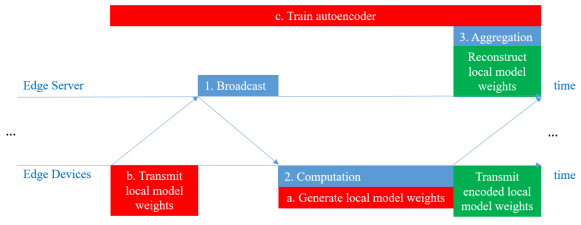


Fig. 5. Training process of the autoencoder. (Time slot of STAG.)

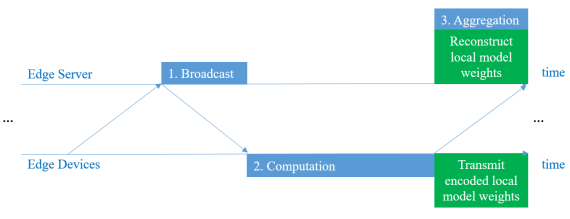


Fig. 6. Training process of federated learning with STAG autoencoder.

In federated learning without an autoencoder, the total communication overhead is $C_1 KM$, where C_1

is the number of communication rounds. On the other hand, using AE can reduce the upload communication overhead per round by a factor of $1 - \frac{L}{M}$, resulting in a total communication overhead of $C_1KL + C_2KM$, where C_2 is the number of communication rounds in federated learning with an AE. Therefore, the reduced total communication overhead can be calculated as $1 - \frac{(C_1KL+C_2KM)}{C_1KM}$.

B. Train AE and Global Model Simultaneously (TAGS)

The TAGS learning protocol is a special case of SAEF learning process. In TAGS all local devices activate AE encoders when AE is properly trained, whereas in SAEF only a subset of the local devices turn on the AE encoders. As such, the amount of information that needs to be uploaded in each communication round is reduced by a factor of $1 - \frac{L}{M}$. This reduction is due to the compression of local model parameters by the autoencoder before they are transmitted to the server.

C. SAEF Scheduling

In this case, the reduced upload communication overhead per round is $1 - \frac{LK_{on}}{MK}$, where K_{on} is lower than K. This means that the autoencoder not only compresses the local model weights but also reduces the number of non-zero elements, resulting in further reduction in the communication overhead.

V. Simulation Results

To evaluate the efficacy of our proposed method, we conduct simulations using deep neural network (DNN) on the MNIST and Fashion-MNIST dataset and evaluate testing accuracy. The simulation setup involves five local nodes in the network. For the dataset, each node collects a fraction of labeled training datasets $D_1, D_2, \ldots, D_5$, with $12,000$ training data at each device. We consider $t_{j,k}^{\text{cp}}$ to be uniformly distributed from 1s to 20 sec, while the available bandwidth B is 1 MHz. To simulate a realistic communication environment, we assume small scale Rayleigh fading $h_{j,k} \sim \mathcal{CN}(0,1)$ and set N_0 to be -107 dBm. The path-loss model we adopt is $PL[dB] = 128.1 + 37.6\log_{10}(d)$, where d in km is the edge-to-server distance. We assume a transmit power of 30 dBm for each local device.

The format of floating numbers in the storage can be either 32 or 64 bits. During transmission, we use 16 bits to transmit one element of model weights, so the size of transmission is the product of the number of weights and 16 bits.

The autoencoder model consists of a neural network with one hidden layer and 500 hidden nodes. The global model weights are represented by a neural network with one hidden layer and 100 hidden nodes, resulting in a total of $784\times100+1\times100+100\times10+1\times10 = 79,510$ elements in the global model weights. Two performance metrics are used to evaluate the proposed protocol: (1) separate training the autoencoder and the global model weights (STAG), and (2) train the autoencoder and the global model weights at the same time (TAGS).

In the simulation, we use $30,000$ data to construct $3,000$ model weights (training data) for the AE. Then, we use the remaining $30,000$ data to train the global model weights of STAG. To efficiently train the global model weights, training the AE and the global model weights at the same time offers a more practical option. However, the AE cannot be used in the beginning without initial training.

A. Performance Comparison

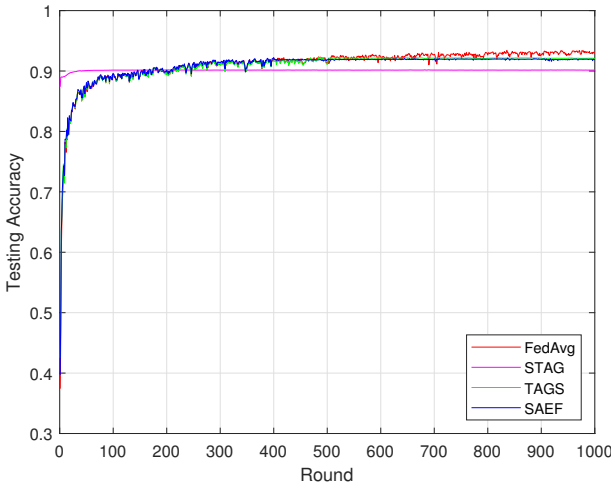


Fig. 7. Comparison of testing accuracy. (MNIST)

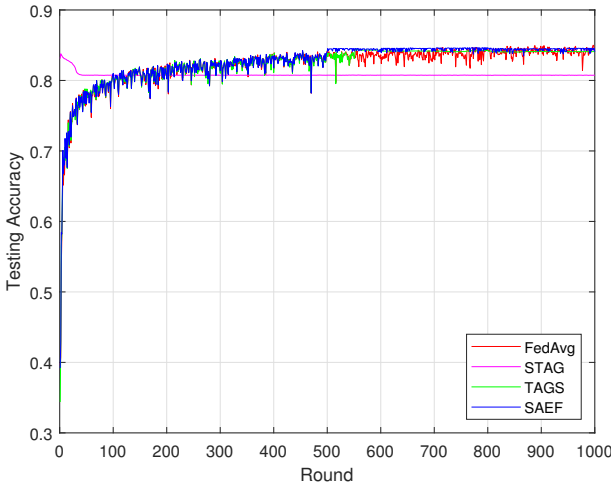


Fig. 8. Comparison of testing accuracy. (Fashion-MNIST)

Fig. 7 and Fig. 8 compare the proposed SAEL scheduling algorithm with FedAvg, STAG, and TAGS. The SAEL and TAGS update the AE every 50 communication rounds. It shows that STAG converges faster than the other methods because the AE in STAG has been trained with 3000 training data, allowing it to capture more information to update the global model weights. However, the AE in STAG is not updated during the training process of federated learning, resulting in worse testing accuracy when the model weights converge. In TAGS, local devices start to use the AE encoder in the 500th communication round. In SAEF, we use the testing accuracy to decide whether to use the AE or not, and we can observe that the server starts using the AE

TABLE I

LATENCY COMPARISON

$\hat{t}^j_{\text{round}}$	t^j_{round}	Latency without Autoencoder
73.09	2,440.67	386,912.57

at the 400th communication round. We can also observe from the figure that the AE-enhanced FL is more stable in the testing accuracy whereas the testing accuracy of FedAvg fluctuates more, which demonstrates the denoising property of the AE. Since all local devices use AE to transmit the local model weights, STAG has the best stability during the training process.

In TABLE I, we compare latency of different methods in one communication round. $\hat{t}^j_{\text{round}}$ represents the estimated time latency of the scheduling algorithm in the communication round, while t^j_{round} is the actual time latency computed by the server with the autoencoder. The latency without an autoencoder is the time latency of federated learning without an autoencoder. The unit of time in this table is seconds. It is evident that $\hat{t}^j_{\text{round}}$ and t^j_{round} are different, as the algorithm cannot predict the exact value of t^j_{round}. Our primary objective is to schedule the use of the autoencoder to reduce communication consumption, and the latency is a secondary target. The server only needs to wait for the last local device to finish transmission, and then it can get t^j_{round}. Overall, the proposed method achieves lower latency than other methods in one communication round.

In TABLE II, we present a comparison of the complexity between federated learning and federated learning with an autoencoder. The communication consumption column represents the number of elements transmitted between the server and local devices per communication round. In federated learning with an autoencoder, each local device needs to transmit the encoded local model weights, which is much smaller than the original local model weights. Thus, using an autoencoder reduces the communication consumption in federated learning. The reduced overhead column shows the difference in communication consumption between federated learning and federated learning with an autoencoder. The accuracy column displays the testing accuracy of federated learning and federated learning with an autoencoder after the training process. From the table, we can observe that federated learning with an autoencoder can significantly reduce the overhead and power consumption of local devices during the training process. Moreover, it achieves comparable testing accuracy to federated learning without an autoencoder.

TABLE II

COMPLEXITY COMPARISON (MNIST)

Method	Communication Consumption	Reduced Overhead	Accuracy
Federated Learning	79,510	0	92.97%
Federated Learning with Autoencoder	50	79,460(99.9%)	92.04%
	500	79,010(99.3%)	92.68%
	5,000	74,510(93.7%)	92.49%

VI. CONCLUSION

In conclusion, we proposed a novel approach for reducing communication consumption in federated learning using an autoencoder. We developed a scheduling algorithm to minimize the latency of the communication round. Our experimental results on the MNIST and Fashion-MNIST dataset demonstrate the effectiveness of our proposed method in terms of reducing communication consumption and improving the testing accuracy of the global model weights. In addition, we provided a complexity analysis of our approach compared to traditional federated learning. Our work contributes to the field of federated learning by providing a practical solution to reduce communication overhead and improve the efficiency of the training process. Future research can explore the use of autoencoders in other federated learning scenarios and investigate the potential advantages of our approach in real-world applications.

REFERENCES

[1] J. Konečný, H. B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," 2015.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[3] T. Guo, S. Guo, J. Wang, X. Tang, and W. Xu, "Promptfl: Let federated participants cooperatively learn prompts instead of models - federated learning in age of foundation model," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2023.

[4] H. Zhao, W. Du, F. Li, P. Li, and G. Liu, "Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1930–1943, 2013.

[6] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 119–130, 2016.

[7] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659–5670, 2015.

[8] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, 2018.

[9] J. Guo, C.-K. Wen, and S. Jin, "Deep learning-based csi feedback for beamforming in single- and multi-cell massive mimo systems," *IEEE Journal on Selected Areas in Communications*, pp. 1872–1884, 2021.

[10] Z. Gu and Y. Yang, "Detecting malicious model updates from federated learning on conditional variational autoencoder," in *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021, pp. 671–680.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[12] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Deep Unfolding-based Distributed MIMO Detection

Masaya Kumagai*, Ayano NAKAI-KASAI*, and Tadashi WADAYAMA*
* Nagoya Institute of Technology, Gokiso, Nagoya, Aichi 466-8555, Japan,
E-mail: clp14056@stn.nitech.ac.jp, {nakai.ayano,wadayama}@nitech.ac.jp

***Abstract*—This paper proposes a distributed multiple-input multiple-output signal detection algorithm based on DU (deep unfolding), which is known as the optimizing method of parameters involved in an iterative algorithm by using commonly used deep learning techniques such as backpropagation and stochastic gradient descent. In the distributed system, multiple distributed SPUs (sub processing units) having multiple antennas execute an iterative signal processing process to detect transmitted signals in cooperation with a CPU (central processing unit). In applications where SPUs are located distributedly, SPUs are placed under heterogeneous SNR (signal-to-noise ratio) environments depending on geographical location and distances from transmitters. We introduce weighting parameters corresponding to each SPU into the aggregation process of the proposed method to incorporate the heterogeneous SNR environments into the detection algorithm. In addtion, we apply DU to optimize the parameters and obtain improvement of the detection perfomance. Results of numerical experiments indicate that the proposed method with the parameters learned by DU improve detection performance by appropriately tuning the weighting parameters according to the heterogeneous SNR environments.**

I. Introduction

Towards the 6G (6th Generation mobile communication system) for the next-generation infrastructure for wireless communications, intensive research activities are currently underway [1]. In the 6G, it is expected to achieve faster and more reliable communication with lower latency, and to accommodate a tremendous number of mobile devices. Massive MIMO (Multiple-Input Multiple-Output) with hundreds of antennas is an essential system for dealing with increasing data traffic and achieving high spectral efficiency [2].

In massive MIMO systems, centralized signal detection methods such as ML (Maximum Likelihood), ZF (Zero-Forcing), and linear MMSE (Minimum Mean Squared Error) are commonly used [3]. However, these centralized signal processing methods require high data throughput and high computing power that highly likely exceeds existing hardware performance [4] in the 6G era due to the increased number of mobile devices. Specifically, the computational demands of a BS (base station) cannot be satisfied with the current BS architecture in the near future. The *load balancing* technologies to mitigate the concentration of computational load in a BS will be indispensable for achieving the goal of the 6G system.

Distributed MIMO signal detection methods have been proposed [2], [4]–[6], which can be regarded as a promising load balancing technique. The distributed methods cooperatively solve a signal detection problem with multiple computing units, consequently reducing computational loads in the BS. In [4]–[6], distributed MIMO detection algorithms composed of multiple SPUs (Sub Processing Units) and a single CPU (Centralized Processing Unit) were proposed. Zhao et al. [5] and Sanchez et al. [6] proposed a signal processing architecture where the SPUs sequentially execute processing tasks. Li et al. [4] proposed algorithms based on the CG (Conjugate Gradient) and ADMM (Alternating Direction Method of Multipliers) where SPUs executed in parallel under the orchestration of CPU. Although these methods [4]–[6] provide us useful insights for distributed MIMO methods, there are still several weak points. The distributed methods have a trade-off between the load balancing and performance in general. The performance of these methods can become close enough to that of the centralized method when the number of SPUs is small or under some system conditions. However, they cannot achieve exactly the same performance because each SPU can access only a limited channel state information due to the distributed nature.

Furthermore, in applications where SPUs can be regarded as being located distributedly [7]–[9], the SPUs with several receive antennas are placed under heterogeneous SNR (Signal to Noise Ratio) environments depending on the geographical location and the distances from the transmitters. A SPU located far from the transmitter receives relatively weak signals due to the path loss effect. Such a SPU might suffer from worse SNR and have possibly poor detection performance than other SPUs located closer to the transmitter. In other words, the heterogeneous SNR causes difference in detection performance of SPUs. The conventional methods [4]–[6] do not incorporate the heterogeneity into the detection algorithms. The appropriate exploitation of the difference in SNR among SPUs may yield an improvement of detection performance of the distributed MIMO detection methods.

In this paper, we propose a cooperative signal detection method for distributed MIMO systems, *LWSP (Learned Weighting Soft Projection)* method, which incorporates the difference in SNR among SPUs. The proposed LWSP method introduces trainable *weighting parameters* corresponding to each SPU, which are used in a message averaging process. These weighting parameters play a key role to *adjust* the message from SPUs under different SNR. The adjustment of the weighting parameters may lead to better detection performance and faster convergence of the detection processes that are the main feature of the proposed method. We apply DU (deep unfolding) technique to the proposed method to optimize the trainable weighting parameters. DU is an optimization technique for trainable parameters included in iterative algorithms [10], [11]. It has been widely used

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

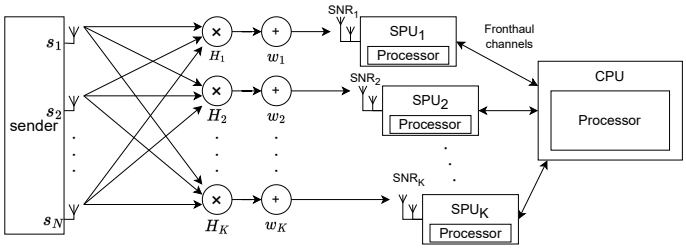


Fig. 1. Distributed MIMO system.

to improve signal detection performance in MIMO systems [12]–[15]. The proposed method is based on the gradient descent-based MIMO detection method with soft projection, which are reported to achieve high detection performance in a DU-based centralized detection method [16].

II. Preliminaries

A. Notation

We denote the zero and one vectors with boldface letters $\mathbf{0}$ and $\mathbf{1}$. The matrix $\boldsymbol{I}_n$ denotes the identity matrix of size $n \times n$. For a complex matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{A}^T$ denotes a transpose and $\boldsymbol{A}^H$ denotes the Hermitian transpose of $\boldsymbol{A}$, respectively. We define $\mathcal{CN}(\mathbf{0}, \sum)$ as the circularly symmetric complex Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\sum$. For vector $\boldsymbol{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$, $||\boldsymbol{w}||$ represents the ℓ_2 norm of $\boldsymbol{w}$.

B. Distributed MIMO System Model

In this paper, we consider a distributed MIMO system between a sender with N transmit antennas and a receiver composed of the CPU and the multiple SPUs [4]–[6], which is shown in Fig. 1. We assume that the CPU and the SPUs have computation capability required for the signal detection tasks. It is expected that an appropriate cooperating detection process of the CPU and the SPUs provides computational load balancing among them. A vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_N)^T \in \mathbb{C}^N$ represents transmit signals from N antennas of the sender. The transmit signal $\boldsymbol{s}$ is received by K SPUs, each of which has m receive antennas.

We assume that each SPU is placed at geographically different locations [7]–[9]. This means that the receive power for each SPU might be different because of the different path loss effect. It is natural to assume that each SPU has distinct SNR, i.e., SPUs are under a heterogeneous SNR environment. In the subsequent formulation, we will use the following *equivalent heterogeneous SNR model*, which includes the SPU-dependent noise variances. This is because it allows a simpler discussion.

Throughout the paper, the kth SPU represents the SPU$_k$($k = 1, 2, \ldots, K$). As a received vector, SPU$_k$ receives

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{s} + \boldsymbol{w}_k \in \mathbb{C}^m, \tag{1}$$

where $\boldsymbol{H}_k \in \mathbb{C}^{m\times N}$ is the channel matrix of SPU$_k$. A complex vector $\boldsymbol{w}_k \in \mathbb{C}^m$ is the noise vector that follows $\mathcal{CN}(\mathbf{0}, \sigma_k^2 \boldsymbol{I}_m)$.

Algorithm 1 CG-based Detection Method [4]

Input: $\boldsymbol{H}_k (k = 1, 2, \ldots, K), y_k, \rho$

$\boldsymbol{y}_k^{\mathrm{MRC}} = \boldsymbol{H}_k^H \boldsymbol{y}_k$
$\boldsymbol{y}^{\mathrm{MRC}} = \sum_{k=1}^{K} \boldsymbol{y}_k^{MRC}$
$\boldsymbol{r}^{(0)} = \boldsymbol{y}^{MRC}, \boldsymbol{p}^{(0)} = \boldsymbol{r}^{(0)}, \boldsymbol{x}^{(0)} = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  SPU$_k$($k = 1, 2, \ldots, K$):
  $\boldsymbol{w}_k^{(t)} = \boldsymbol{H}_k^H \boldsymbol{H}_k \boldsymbol{p}^{(t-1)}$
  CPU:
  $\boldsymbol{w}^{(t)} = \sum_{k=1}^{K} \boldsymbol{w}_k^{(t)}$
  SPU$_k$($k = 1, 2, \ldots, K$):
  $\boldsymbol{e}^{(t)} = \rho \boldsymbol{p}^{(t-1)} + \boldsymbol{w}^{(t)}$
  $\alpha = ||\boldsymbol{r}^{(t-1)}||^2 / ((\boldsymbol{p}^{(t-1)})^H \boldsymbol{e}^{(t)})$
  $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha \boldsymbol{p}^{(t-1)}$
  $\boldsymbol{r}^{(t)} = \boldsymbol{r}^{(t-1)} - \alpha \boldsymbol{e}^{(t-1)}$
  $\beta = ||\boldsymbol{r}^{(t)}||^2 / ||\boldsymbol{r}^{(t-1)}||^2$
  $\boldsymbol{p}^{(t)} = \boldsymbol{r}^{(t)} + \beta \boldsymbol{p}^{(t-1)}$
end for

Output: $\hat{\boldsymbol{x}} = \boldsymbol{x}^{(T)}$

SNR of SPU$_k$ is defined as $\mathrm{SNR}_k = 10 \log_{10}(N/\sigma_k^2)$. In addition, $\boldsymbol{w}_i$, $\boldsymbol{w}_j (i \neq j)$ are independent.

Since the SPUs cooperatively execute iterative processes to detect the transmitted signals, the total number of receiving antennas can be virtually considered as $M = mK$. It is assumed that the SPU$_k$ has only own received signal $\boldsymbol{y}_k$ and own channel state information $\boldsymbol{H}_k$. To detect the transmitted signal at each SPU, SPU needs to communicate with CPU, which can receive and aggregate the messages sent from all the SPUs. We assume that all the SPUs are connected to CPU with a fast fronthaul channel and that no transmission delays and error-free links between each SPU and the CPU.

C. CG-based Detection Method

We here briefly review the conventional CG-based MIMO detection method. The CG method is a numerical algorithm for solving linear equations [17]. Li et al. proposed the CG-based signal detection method for the distributed MIMO systems [4]. The CG-based detection method can achieve performance close to that of the centralized MMSE method by combining with error correcting codes.

The CG-based detection method is summarized in Algorithm 1. This algorithm includes a hyper parameter ρ to be determined. The method has two processes to calculate the estimated signal $\hat{\boldsymbol{x}}$. The first process is the preprocessing phase to calculate the temporary vector $\boldsymbol{y}^{\mathrm{MRC}}$. The second process is the iteration phase to calculate the estimated vector $\hat{\boldsymbol{x}}$. In the preprocessing phase, the method generates the temporary vectors $\boldsymbol{y}_k^{\mathrm{MRC}}$ and $\boldsymbol{y}^{\mathrm{MRC}}$ used for the iteration phase. Each SPU calculates the temporary vector $\boldsymbol{y}_k^{\mathrm{MRC}} = \boldsymbol{H}_k^H \boldsymbol{y}_k$ and CPU calculates $\boldsymbol{y}^{\mathrm{MRC}}$, where $\boldsymbol{y}^{\mathrm{MRC}}$ is the sum of the temporary vectors $\boldsymbol{y}_1^{\mathrm{MRC}}, \ldots, \boldsymbol{y}_K^{\mathrm{MRC}}$. In the iteration phase, the temporary vectors $\boldsymbol{e}^{(t)}, \boldsymbol{x}^{(t)}, \boldsymbol{r}^{(t)}, \boldsymbol{p}^{(t)}$ and state vector $\boldsymbol{x}^{(t)}$ are updated on the basis of the CG method at SPUs. The resulting estimated

signal $\hat{\boldsymbol{x}}$ can be obtained after T iterations.

In this paper, we will use the CG-based detection method for benchmark purpose.

D. Deep Unfolding

1) Overview of DU processes: DU is a method to improve performance of iterative algorithms by using deep learning techniques such as stochastic gradient descent (SGD) and back propagation. One can expand iterative processes of an algorithm to an unfolded feed-forward network similar to feed-forward neural network, where each layer of the network consists of the process of the algorithm. Next, some trainable parameters can be embedded in the unfolded network and the embedded parameters are optimized by using deep learning techniques.

2) Update of Trainable Parameters: In an iterative algorithm, several parameters included in the algorithm have a critical effect on performance of the algorithm. Let a set of parameters in an algorithm be the trainable parameters $\boldsymbol{\Theta}^{(t)}$, which depend on the time index $t = 1, \ldots, T$ corresponding to the iterative steps of the algorithm. The trainable parameters can be optimized by using deep learning techniques. The training data used for the training processes generally consist of mini-batches of input-output pairs and supervised learning is performed by using the training data. We denote a set of all trainable parameters as $\tilde{\boldsymbol{\Theta}} = \{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(T)}\}$. A loss function $L(\tilde{\boldsymbol{\Theta}})$ and its gradient $\nabla L(\tilde{\boldsymbol{\Theta}})$ are required to be set to reduce error between outputs of the algorithm and the supervised signals. The cth update $(c = 1, \ldots, C)$ of the trainable parameters in the case of SGD is given by

$$\tilde{\boldsymbol{\Theta}}_c = \tilde{\boldsymbol{\Theta}}_{c-1} - \epsilon \nabla L(\tilde{\boldsymbol{\Theta}}_{c-1}), \tag{2}$$

where ϵ (> 0) is a learning rate.

3) Incremental Training: In DU, a training method, called incremental training, is known to provide excellent training results in many cases [18]. Generally, in deep learning, all trainable parameters are updated simultaneously. On the other hand, the incremental training trains parameters while increasing the number of layers one by one instead of the simultaneous update.

The detail of the incremental training processes is as follows. First, in the first layer, we optimize the trainable parameters $\boldsymbol{\Theta}^{(1)}$ by DU processes. Second, we add the second layer and optimize the trainable parameters $\boldsymbol{\Theta}^{(1)}$ and $\boldsymbol{\Theta}^{(2)}$. In this way, we optimize the trainable parameters up to Tth iteration while adding layers per the number of iterations and then obtain the trained parameters $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(T)}$. The method allows us to concentrate on optimizing the trainable parameters of the added layers on each process. For this reason, we can mitigate the vanishing gradient problem [10].

III. Learned Weighting Soft Projection Method

In this section, we introduce the proposed detection method named *LWSP method*. Section III-A describes iterative processes of the proposed algorithm and Sect. III-B summarizes training procedure of the parameters included in the algorithm.

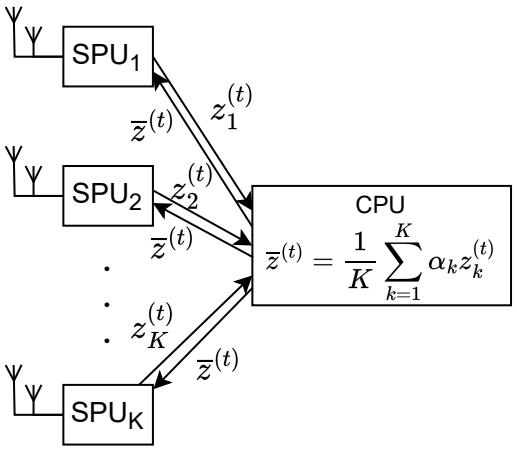


Fig. 2. Message exchange between CPU and SPUs

In addition, we summarize computational complexity of the proposed and conventional detection methods in Sect. III-C.

A. Learned Weighting Soft Projection (LWSP) Method

LWSP method is based on the gradient descent method with soft projection, which can be seen as a variant of *projected gradient descent*. The proposed procedure consists of parallel processing at SPUs and *weighted message aggregation* with the CPU. The proposed algorithm iterates the following three steps:

1) Gradient descent step in parallel at all SPUs,
2) Weighted message aggregation at CPU,
3) Soft projection step in parallel at all SPUs.

Figure 2 illustrates the message exchange between the SPUs and CPU of the proposed method. Each SPU$_k$ executes a gradient descent step and calculates a temporary estimated signal $\boldsymbol{z}_k^{(t)}$, which is sent to the CPU. The CPU then receives $\boldsymbol{z}_1^{(t)}, \ldots, \boldsymbol{z}_K^{(t)}$ from all the SPUs, and executes a weighted message aggregation, which is a process to evaluate the weighted average of the incoming messages. As a result, the CPU obtains the weighted average $\bar{\boldsymbol{z}}^{(t)}$ and send it back to all the SPUs. Each SPU$_k$ receives the weighted average $\bar{\boldsymbol{z}}^{(t)}$ and then executes a soft projection step. The above three steps are iterated T times, and the state vectors of SPUs at the end of the Tth iteration are the resulting estimated signal. A detailed description of each step is shown below.

Each SPU$_k$ has a state vector $\boldsymbol{x}_k^{(t)}$ at tth iteration. In the gradient descent step, each SPU executes a gradient descent step aiming at minimizing the global quadratic objective function:

$$L(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} ||\boldsymbol{y}_k - \boldsymbol{H}_k \boldsymbol{x}||^2. \tag{3}$$

The negative gradient of the objective function (3) is given by

$$-\nabla L(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{H}_k^H (\boldsymbol{y}_k - \boldsymbol{H}_k \boldsymbol{x}). \tag{4}$$

The global gradient, however, cannot directly be used because each SPU can access only its own received signal and channel matrix. Therefore, each SPU$_k$ obtains a temporary estimated signal $\boldsymbol{z}_k^{(t)}$ by using the vector $\boldsymbol{H}_k^H(\boldsymbol{y}_k - \boldsymbol{H}_k\boldsymbol{x}^{(t)})$ as a local gradient instead of the global gradient. The gradient descent step for SPU$_k$ becomes

$$\boldsymbol{z}_k^{(t)} = \boldsymbol{x}^{(t)} + \mu^{(t)}\boldsymbol{H}_k^H(\boldsymbol{y}_k - \boldsymbol{H}_k\boldsymbol{x}^{(t)}). \tag{5}$$

The parameter $\mu^{(t)}(>0)$ is a step-size parameter that affects convergence performance.

The gradient of (3) should be obtained from the sum of the $\{\boldsymbol{z}_k^{(t)}\}$. In the weighted aggregation step, the CPU receives the temporary estimated signals $\boldsymbol{z}_1^{(t)}, \ldots, \boldsymbol{z}_K^{(t)}$ from all the SPUs. The full gradient of $L(\boldsymbol{x})$, $\boldsymbol{z}^{(t)}$, is thus given by

$$\boldsymbol{z}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{z}_k^{(t)}. \tag{6}$$

In the proposed algorithm, we will use the the following weighted average $\bar{\boldsymbol{z}}^{(t)}$ instead of $\boldsymbol{z}^{(t)}$:

$$\bar{\boldsymbol{z}}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\alpha_k\boldsymbol{z}_k^{(t)}. \tag{7}$$

The parameter α_k represents the weight assigned for SPU$_k$. The evaluation process of $\bar{\boldsymbol{z}}^{(t)}$ is called weighted aggregation.

The average $\bar{\boldsymbol{z}}^{(t)}$ can be seen as a *modified gradient*. The reason why we introduced the weighted average is the following. The weighting parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^T \in \mathbb{R}_+^K$ allow us to weigh partial gradients according to its reliability. We may be able to increase the weight of SPU$_k$ if the SNR is enough high. It is expected that such an assignment leads to a better convergence behavior. These weighting parameters should be determined appropriately by incorporating the SNR differences among SPUs. The adjustment of the weighting parameters is of critical importance to achieve appropriate detection performance. The optimization of the weighting parameters is to be discussed later.

In the soft projection step, the state vector $\boldsymbol{x}_k^{(t)}$ is updated by the following soft projection step:

$$\boldsymbol{x}_k^{(t+1)} = \boldsymbol{x}_k^{(t)} - \eta(\boldsymbol{x}_k^{(t)} - \bar{\boldsymbol{z}}^{(t)}). \tag{8}$$

The parameter $\eta(>0)$ controls the strength of the soft projection. This step implies a smooth projection of $\boldsymbol{x}_k^{(t)}$ onto the weighted average $\bar{\boldsymbol{z}}^{(t)}$ instead of a hard projection $\boldsymbol{x}_k^{(t)} = \bar{\boldsymbol{z}}^{(t)}$. The soft projection was reported to yield performance improvement when combining DU [19].

The pseudo code of the proposed LWSP method is described in Algorithm 2.

The proposed algorithm includes free parameters $\mu^{(t)}, \boldsymbol{\alpha}, \eta$. The setting of these parameters causes significant influences to the detection performance of the proposed method. Finding an appropriate set of suitable parameters is not a straightfoward problem. We here employ DU to optimize these parameters. Namely, the values of these parameters are to be optimized by DU processes described in the next subsection.

Algorithm 2 LWSP method

Input: $\boldsymbol{H}_k(k=1,2,\ldots,K), \alpha_k, \boldsymbol{y}_k, \mu^{(t)}(t=1,2,\ldots,T)$, η, $\boldsymbol{x}_k^{(1)} = \boldsymbol{0}$ $(k=1,2,\ldots,K)$

1: **for** $t = 1, \ldots, T$ **do**
2: SPU$_k(k=1,2,\ldots,K)$:
$\boldsymbol{z}_k^{(t)} = \boldsymbol{x}_k^{(t)} + \mu^{(t)}\boldsymbol{H}_k^H(\boldsymbol{y}_k - \boldsymbol{H}_k\boldsymbol{x}_k^{(t)})$
3: CPU:
$\bar{\boldsymbol{z}}^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\alpha_k\boldsymbol{z}_k^{(t)}$
4: SPU$_k(k=1,2,\ldots,K)$:
$\boldsymbol{x}_k^{(t+1)} = \boldsymbol{x}_k^{(t)} - \eta(\boldsymbol{x}_k^{(t)} - \bar{\boldsymbol{z}}^{(t)})$
5: **end for**

Output: $\hat{\boldsymbol{x}} = \boldsymbol{x}_k^{(T+1)}$

TABLE I
SUMMARY OF COMPUTATIONAL COMPLEXITY

Detection method	CPU	SPU
LWSP (proposed)	$O(NT)$	$O(mNT)$
CG-based	$O(NT)$	$O(mNT)$
Centralized MMSE	$O(N^3)$	-

B. Parameter Learning by DU

The parameters $\mu^{(t)}$, η and $\boldsymbol{\alpha}$ affect the convergence performance of the proposed method. We thus need to determine these parameters properly. In order to optimize the parameters, we consider these parameters as trainable parameters $\tilde{\boldsymbol{\Theta}}$ and apply DU.

We here explain the procedure to optimize these parameters. In this study, we use training data generated from the probabilistic model of the distributed MIMO system defined in subsection II-B. A transmit signal $\boldsymbol{s}$ and noise vector $\boldsymbol{w}_k$ are randomly generated, and the received signal $\boldsymbol{y}_k$ is obtained by (1). The estimated signal $\hat{\boldsymbol{x}}_k^{(T+1)}$ is obtained by Algorithm 2. The transmit signal $\boldsymbol{s}$ is used as the supervised signal. The trainable parameters are updated to decrease the squared error loss function between the supervised signals $\boldsymbol{s}$ and estimated signals $\hat{\boldsymbol{x}}$. In addtion, we exploit incremental training described in subsection II-D3.

C. Computational Complexity

The required computational complexity for the proposed LWSP method, conventional CG-based detection method, and centralized (typical) MMSE detection method [3] is summarized in Table I. The LWSP and CG-based detection methods are distributed algorithms that consist of processing on multiple SPUs and CPU. The main computational factor for each SPU is the matrix-vector multiplication included in each iteration. CPU only calculates the sum so that it requires $O(N)$ in each iteration. On the other hand, the centralized MMSE method includes matrix inverse calculation that requires $O(N^3)$.

Table I reveals that the proposed LWSP method achieves load balancing into SPUs and CPU. The LWSP method is more computationally efficient than the centralized MMSE detection method.

TABLE II
PARAMETER SETTINGS OF EXPERIMENTS

Experiment	N	m	K	SNR_k[dB]
1	16	20	2	0.0, variable
2	16	8	5	0.0, 5.0, 10.0, 15.0, 20.0
3	16	8	5	0.0, 0.0, 0.0, 20.0, 20.0
4	16	4	10	0.0, 2.22, 4.44, 6.66, 8.88, 11.11, 13.33, 15.55, 17.77, 20.0

IV. NUMERICAL EXPERIMENTS

In this section, we show results of numerical experiments to evaluate the detection performance of the proposed method by comparing with that of the conventional CG-based detection method [4] as the benchmark.

A. Experimental Settings

For the numerical experiments, the LWSP method was implemented using PyTorch 1.12 [20]. We assumed QPSK transmit signal points:

$$\left\{\frac{1+i}{\sqrt{2}}, \frac{-1+i}{\sqrt{2}}, \frac{1-i}{\sqrt{2}}, \frac{-1-i}{\sqrt{2}}\right\} \tag{9}$$

and the transmit symbol s_i was a randomly generated according to the probability:

$$\begin{aligned} P\left(s_i = \frac{1+i}{\sqrt{2}}\right) &= P\left(s_i = \frac{-1+i}{\sqrt{2}}\right) \\ &= P\left(s_i = \frac{1-i}{\sqrt{2}}\right) = P\left(s_i = \frac{-1-i}{\sqrt{2}}\right) = \frac{1}{4}. \end{aligned} \tag{10}$$

Each element of the channel matrix $\boldsymbol{H}_k$ was randomly generated according to $\mathcal{CN}(0,1)$. A channel matrix was generated before the training process and the same matrix was used during the numerical simulation. The trainable parameters were updated by Adam optimizer using mini-batches whose size was 1000 and a training process is executed with 10000 rounds. The initial values of $\mu^{(1)}, \ldots, \mu^{(T)}$ were set to 0.01 and that of η was 1.0, respectively. The initial vector of weighting parameters $\boldsymbol{\alpha}$ was set to $\mathbf{1}$. The parameter ρ of the CG-based detection method [4] was set to $\rho = 0$. The learning rate of Adam optimizer for $\mu^{(t)}$, η and α_k were 10^{-3}, 10^{-5} and 10^{-5}, respectively. The number of iterations of the LWSP method and conventional method was set to $T = 10$.

We prepared four experimental conditions presented in Table II. The experiment 1 is the case where the number of the SPUs is $K = 2$. We present the detection performance when SNR of SPU_2, SNR_2, varies relative to SNR_1 which is fixed to $\mathrm{SNR}_1 = 0$dB. In the experiments 2–4, we evaluate detection performances under various environments when the number of the SPUs is $K = 5$ or 10. The system parameters and SNR settings of each SPU are also summarized in Table II.

To confirm the effectiveness of index-dependent step-size parameters $\mu^{(t)}$, we also evaluated detection performance of the LWSP method that uses fixed, i.e., index-independent, step-size parameter μ.

B. Performance Evaluation

The results of the experiment 1 are shown in Figs. 3, 7, 11. Figure 3 presents SER (Symbol Error Rate) of the detection methods as functions of SNR_2. The LWSP method with the index-dependent step-size parameters $\mu^{(t)}$ yields the smallest SER when SNR_2 is larger than 10dB. This result implies that the LWSP method with index-dependent step-size performs better than the one with fixed step-size. The larger degree of freedom on the step-size parameters appears beneficial to improve the convergence speed of the LWSP method. The difference in SER between the LWSP method with index-dependent step-size parameters and the CG-based method is small when the SNR_2 is in the range from 0dB to 5dB. However, in the regime where SNR_2 is greater than 5dB, the LWSP method achieves significantly smaller SER than that of the conventional CG-based method. An appropriate setting for the weighting parameters in the LWSP method result in this huge difference.

Figure 7 shows the values of the trained index-dependent step-size parameters for $\mathrm{SNR}_2 = 20$dB. We can see that the value varies over a wide range depending on the iteration index. Figure 11 displays the trained parameter values of $\boldsymbol{\alpha}$ when SNR_2 is 20dB. The value of α_2 is much larger than that of α_1. It indicates that the message from SPU_2 with higher SNR is regarded as more important than that from SPU_1.

Figures 4–6, 8–10, and 12–14 show SER performance versus the number of iterations, the values of the trained index-dependent step-size parameters, and the values of the trained parameter $\boldsymbol{\alpha}$ in the experiments 2–4, respectively. From Figs. 4–6, the proposed LWSP method with index-dependent step-size parameters achieves the lowest SER in each iteration. We can find that the proposed method can provides superior detection performance to the conventional method in a variety of SNR setting. Figures 8–10 show the same tendency as the result in the experiment 1. These index-dependent step-size parameters yield the improvement in detection performance of the proposed method.

The trained weighting parameters $\boldsymbol{\alpha}$ shown in Figs. 12–14 can be interpreted as a reflection of the SNR setting. From these figures, we can see the correspondence between the value of trained weighting parameter α_k and SNR_k, i.e., a larger SNR_k leads to a larger α_k. This means that the weighting parameters in the proposed method can be appropriately learned by using DU according to the value of SNR_k.

In summary, the proposed LWSP method can achieve excellent signal detection performance under various heterogeneous SNR environments by introducing the weighting parameters

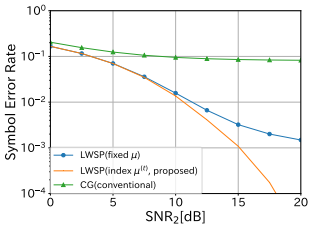


Fig. 3. SER, experiment 1

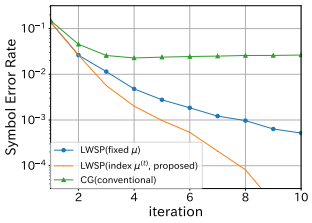


Fig. 4. SER, experiment 2

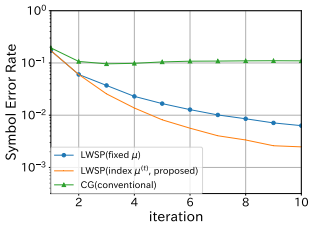


Fig. 5. SER, experiment 3

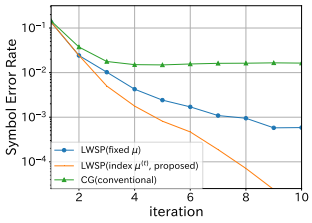


Fig. 6. SER, experiment 4

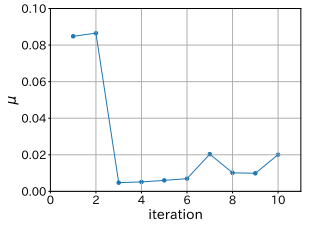


Fig. 7. Step-size $\mu^{(t)}$, experiment 1, $\mathrm{SNR}_2 = 20\mathrm{dB}$

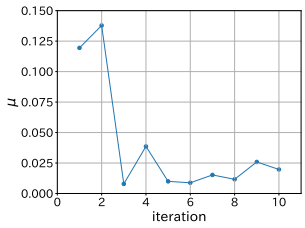


Fig. 8. Step-size $\mu^{(t)}$, experiment 2

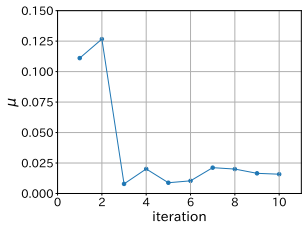


Fig. 9. Step-size $\mu^{(t)}$, experiment 3

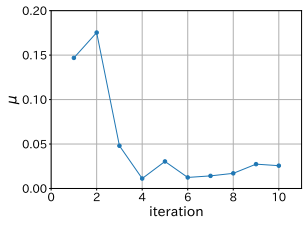


Fig. 10. Step-size $\mu^{(t)}$, experiment 4

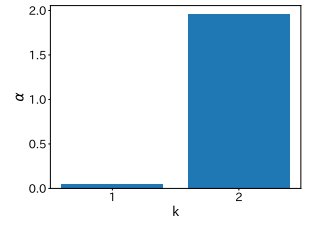


Fig. 11. Parameter $\boldsymbol{\alpha}$, experiment 1, $\mathrm{SNR}_2 = 20\mathrm{dB}$

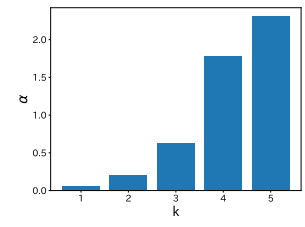


Fig. 12. Parameter $\boldsymbol{\alpha}$, experiment 2

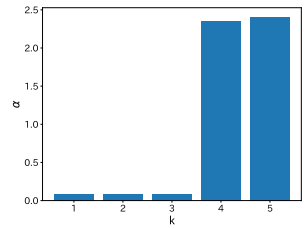


Fig. 13. Parameter $\boldsymbol{\alpha}$, experiment 3

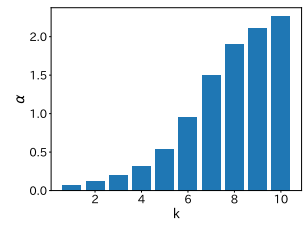


Fig. 14. Parameter $\boldsymbol{\alpha}$, experiment 4

optimized by DU.

V. Concludeing Remarks

In this paper, we proposed the LWSP method for distributed MIMO signal detection. The proposed method is based on the gradient descent method with soft projection. The trainable weighting parameters are introduced to adapt heterogeneous SNR environments of SPUs. These weighting parameters are optimized by using deep unfolding and used for the weighting aggregation process at CPU. The numerical results show that the proposed method outperforms the conventional method if SPUs are in heterogeneous SNR environments. In addition, further performance improvement can be achieved by using index-dependent step-size parameter, which improves the convergence behavior of the proposed method.

The proposed method assumed the existance of CPU for weighting aggregation. A decentralized variant without CPU may be able to be developed based on the idea of the trainable weighting parameters. Such a direction remains as an interesting open problem.

Acknowledgment

This work was supported by KAKENHI Grant Number JP22H00514.

References

[1] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6G Wireless Communication Systems: Applications, Requirements, Technologies, Challenges, and Research Directions," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 957–975, 2020. DOI: 10.1109/OJCOMS.2020.3010270.

[2] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation Propagation Detector for Extra-Large Scale Massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2036–2051, 2020. DOI: 10.1109/TWC.2019.2961892.

[3] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017, ISSN: 1932-8346. DOI: 10.1561/2000000093. [Online]. Available: http://dx.doi.org/10.1561/2000000093.

[4] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized Baseband Processing for Massive MU-MIMO Systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 491–507, 2017. DOI: 10.1109/JETCAS.2017.2775151.

[5] X. Zhao, X. Guan, M. Li, and Q. Shi, "Decentralized Linear MMSE Equalizer Under Colored Noise for Massive MIMO Systems," in *Proc. 2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 01–06. DOI: 10.1109/GLOBECOM46510.2021.9685271.

[6] J. R. Sanchez, F. Rusek, M. Sarajlic, O. Edfors, and L. Liu, "Fully Decentralized Massive MIMO Detection Based on Recursive Methods," in *Proc. 2018 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2018, pp. 53–58. DOI: 10.1109/SiPS.2018.8598321.

[7] J. Liu, N. B. Shroff, and H. D. Sherali, "Optimal Power Allocation in Multi-Relay MIMO Cooperative Networks: Theory and Algorithms," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 331–340, 2012. DOI: 10.1109/JSAC.2012.120212.

[8] Y. Lv, Z. He, and Y. Rong, "Two-Way AF MIMO Multi-Relay System Design Using MMSE-DFE Techniques," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 389–405, 2021. DOI: 10.1109/TWC.2020.3024898.

[9] T. Okuyama, S. Suyama, N. Nonaka, and T. Asai, "Millimeter-Wave Base Station Cooperation Technologies for High-Mobility Environments," in *Proc. 2021 IEEE VTS 17th Asia Pacific Wireless Communications Symposium (APWCS)*, 2021, pp. 1–5. DOI: 10.1109/APWCS50173.2021.9548717.

[10] A. Balatsoukas-Stimming and C. Studer, "Deep Unfolding for Communications Systems: A Survey and Some New Directions," in *Proc. 2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2019, pp. 266–271. DOI: 10.1109/SiPS47522.2019.9020494.

[11] Q. Wang, H. Hai, K. Peng, B. Xu, and X.-Q. Jiang, "A Learnable Gauss-Seidel Detector for MIMO Detection," in *Proc. 2020 IEEE/CIC International Conference on Communications in China (ICCC)*, 2020, pp. 107–111. DOI: 10.1109/ICCC49849.2020.9238938.

[12] M.-W. Un, M. Shao, W.-K. Ma, and P. C. Ching, "Deep Mimo Detection Using ADMM Unfolding," in *Proc. 2019 IEEE Data Science Workshop (DSW)*, 2019, pp. 333–337. DOI: 10.1109/DSW.2019.8755566.

[13] L. Pellaco, M. Bengtsson, and J. Jaldén, "Matrix-Inverse-Free Deep Unfolding of the Weighted MMSE Beamforming Algorithm," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 65–81, 2022. DOI: 10.1109/OJCOMS.2021.3139858.

[14] M. Miyoshi, W. Tsujimoto, and T. Nishimura *et al*, "Parameter-Learned AMP for MIMO Signal Detection," in *Proc. 2022 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)*, 2022, pp. 99–103. DOI: 10.1109/APWCS55727.2022.9906492.

[15] S. Takabe, M. Imanishi, T. Wadayama, R. Hayakawa, and K. Hayashi, "Trainable Projected Gradient Detector for Massive Overloaded MIMO Channels: Data-Driven Tuning Approach," *IEEE Access*, vol. 7, pp. 93 326–93 338, 2019. DOI: 10.1109/ACCESS.2019.2927997.

[16] S. Takabe, M. Imanishi, T. Wadayama, and K. Hayashi, "Deep Learning-Aided Projected Gradient Detector for Massive Overloaded MIMO Channels," in *Proc. ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6. DOI: 10.1109/ICC.2019.8761049.

[17] M. R. Hestenes and E. Stiefel *et al*, "Methods of conjugate gradients for solving linear systems," *Journal of research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.

[18] D. Ito, S. Takabe, and T. Wadayama, "Trainable ISTA for sparse signal recovery," *IEEE Transactions on Signal Processing*, vol. 67, no. 12, pp. 3113–3125, 2019. DOI: 10.1109/TSP.2019.2912879.

[19] S. Takabe and T. Wadayama, "Deep Unfolded Multicast Beamforming," in *Proc. GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6. DOI: 10.1109/GLOBECOM42002.2020.9322114.

[20] *PyTorch*, https://pytorch.org/.

A Comparative Analysis of the Yolo Models for Intelligent Lobster Surveillance Camera

Fityanul Akhyar*, Ledya Novamizanti, Koredianto Usman, Ghanes Mahesa Aditya, Farhan Nur Hakim, Mukhamad Zidni Ilman, Ferdi Ramdhon and Chih-Yang Lin
*Universiti Tunku Abdul Rahman, Malaysia
E-mail: fityanul@gmail.com Tel: +60-1137625514
Institut Teknologi Bandung, Telkom University, Indonesia
E-mail: ledyaldn@telkomuniversity.ac.id Tel: +62-81363090822
Telkom University, Indonesia
E-mail: korediantousman@telkomuniversity.ac.id Tel: +62-81294191520
Telkom University, Indonesia
E-mail: ganeshmahesa@student.telkomuniversity.ac.id Tel: +62-81386919942
Telkom University, Indonesia
E-mail: farhannurhakim@student.telkomuniversity.ac.id Tel: +62-82119151861
Telkom University, Indonesia
E-mail: zidnibinwakhidin@student.telkomuniversity.ac.id Tel: +62- 82265511122
PT. Aruna Jaya Nusantara, Indonesia
E-mail: ferdi.ramdhon@aruna.id Tel: +62- 87885288777
National Central University, Taiwan
E-mail: andrewlin@ncu.edu.tw Tel: +886- 935685662

Abstract — **In lobster farming, monitoring of lobsters is essential and must be performed. Aruna, one of the lobster trading startups in Indonesia, is still using manual monitoring by sending divers underwater to observe the progress of the lobsters. This way is considered less effective because it is done intermittently, only at certain times and in good weather conditions. In this study, we propose to investigate object detection for underwater surveillance cameras as a means of real-time monitoring and counting of lobsters. Underwater video data is fed into the proposed system based on a deep learning-based model. This system consists of three processes: image enhancement with WaterNet as noise reduction, object counting with the YOLOv6/7/8 model, and object tracking using StrongSORT. For details, we annotated the lobster data by head only and full body. Our test results show that YOLOv7, with head annotation, has stable mean Average Precision (mAP), making it suitable for the proposed system. Then, we implemented a combination of WaterNet, YOLOv7 and StrongSORT to the proposed system architecture for the Intelligent surveillance camera.**

I. INTRODUCTION

Monitoring livestock is one of the duties that must be performed daily in order to maintain livestock quality. Humans require nutritious sustenance to maintain their nutritional status. Consequently, monitoring of food constituents such as plants and animals is required. Occasionally, monitoring is problematic due to great distances, extreme weather, and the diverse nature of fresh/seawater organisms, such as lobsters. This can result in the loss of animal's cause preyed upon by other animals or environmental degradation.

A camera is required to monitor livestock. Drone Sheep Livestock [1] is one of the many tools developed to monitor livestock using a camera, and it is used to count the quantity of sheep. What if these animals are aquatic? This is one of the greatest challenges for humans, given that water has pressure, turbidity, etc. The objects are affected by the reflection of light near the surface of the water, and when they are in deep water, it is difficult to detect their margins or boundaries due to their dull colour and poor lighting [5]. Object detection requires clear images to capture the object distinctly and feed it into deep learning-based object detection systems such as YOLOv6

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

[7], YOLOv7 [13], and YOLOv8 [10]. Therefore, image enhancement is necessary in this circumstance.

Object detection one of many essential computer vision tasks that involves detecting instances of visual objects belonging to specific classes in digital images [15]. You Only Look Once, or YOLO, is a well-known and extensively utilized algorithm. YOLO is synonymous with its object detection function's characteristics [11]. In recent years, academics have published YOLOV6, YOLOV7, and YOLOV8 as successors to YOLO. The discussion in this research paper concentrates on combining Image Enhancement with these three YOLO variants. In the final phase, it will be necessary to demonstrate and justify the selection of appropriate unifying Image Enhancement or WaterNet [6], YOLOv7 [13] and StrongSORT [3] for smart surveillance system cameras on lobster detection, counting and tracking.

II. Related works

There have been numerous YOLO variations over the past few years that and be the state-of-the-art object detection model. In 2022, YOLOv6 was released [7]. This version is one of the still-popular YOLO versions in the field of object detection. YOLOv6 was designed as a rapid and precise single-stage system for object detection. It utilizes a singular convolution neural network to predict the type and position of image objects. The new architecture and optimization techniques employed render YOLOv6 more effective and accurate than its predecessor. Additionally, previous research compares the outcomes of YOLOv6 tests on prominent datasets with those of other object detection systems. The results indicate that YOLOv6 has excellent efficacy and industry application potential. Furthermore, YOLOv7 [13] is a variant of the YOLO model that has been enhanced. YOLOv7 employs the new architecture of real-time object detector, which is extended efficient layer aggregation networks (E-ELAN) based on ELAN and the corresponding model scaling technique and proposes the trainable bag-of-freebies method to improve object detection accuracy. YOLOv8 [10] is one of the most recent real-time object detection families in the YOLO family. The company Ultralytics has just published this version at the beginning of 2023. YOLOv8 has several enhancements and optimizations over the previous version, including an improved architecture, more efficient optimization techniques, and the use of a larger training data set. The efficacy of YOLOv8 in terms of speed and accuracy is dependent on the hardware specifications and image size employed during the detection process. However, YOLOv8 has excellent precision and relatively high speed, allowing it to be used in a variety of real-time object detection applications. Furthermore, kind of YOLO models have been utilized in the underwater applications task. For instance, the work in used YOLOv3 as the baseline for an underwater robot or remotely operated underwater vehicle (ROV) that proposed to monitor the kind of fish. Similarly, the work in [8] utilized the YOLOv3 baseline combined with data augmentation for underwater object detection. This research proves that the improved version of YOLO achieves good performance under various water conditions. On the other hand, the work in [1] employs the one version before the work in [8, 12] or YOLOv2 with the modified MobileNetV2 as the backbone and gets very high accuracy in kind of fish detection.

Marine engineering and aquatic technologies have paid significant attention to the enhancement of underwater images. In recent years, numerous algorithms for enhancing underwater images have been proposed [2]. However, these algorithms have been evaluated primarily using computer-generated datasets or a small number of actual images. As a result, it remains uncertain how well these algorithms would perform on images captured in natural underwater environments, and it is challenging to evaluate the progress in this field. To address this deficiency, the work in [6] present the first exhaustive perceptual investigation and analysis of underwater image enhancement using a large collection of actual images. In their study, they establish an Underwater Image Enhancement Benchmark (UIEB) consisting of 950 real-world underwater images and 890 reference images. The remaining sixty underwater images for which adequate reference images cannot be obtained are classified as problematic data. They conduct a comprehensive qualitative and quantitative analysis of the state-of-the-art algorithms for underwater image enhancement using this dataset. In addition, they propose the WaterNet network, which is trained on this benchmark and serves as an indication of the generalization capabilities of the proposed UIEB for training CNNs. The benchmark evaluations and WaterNet model results demonstrate the strengths and limitations of the current state-of-the-art algorithms, thereby providing valuable insights for future research in the field of underwater image enhancement.

In this work, we proposed to combine the current version of YOLO or YOLOv6,7,8, image enhancement and Simple Online and Realtime Tracking (SORT) to perform the application for lobster detection and counting in under the sea cage. Moreover, SORT is a pragmatic approach to tracking multiple objects with an emphasis on basic and efficient algorithms [14]. In addition, SORT methods include three

steps: detection with the Faster Region CNN detection framework [9], estimation model with the Kalman Filter [4], and classification with the Hungarian Algorithm [5]. DeepSort is a method for object tracking that integrates object detection and probabilistic data association (PDA)-based tracking techniques [14]. DeepSort utilizes object detection information to approximate the object's location in subsequent frames and updates this estimate whenever an object is detected. StrongSORT [3] is an algorithm for object tracking that is an enhancement over DeepSORT, or you could say that it is a Stronger DeepSORT. StrongSORT has three advantages over DeepSORT: IDF1, HOTA, and MOTA. StrongSORT has IDFI 82.3, MOTA 77.1, and HOTA 69.6, whereas DeepSORT has IDFI 77.3, MOTA 76.7, and HOTA 69.6. Performance-wise, StrongSORT is superior to DeepSORT.

III. PROPOSED SYSTEM

The method utilized in this investigation for lobster detection is the YOLOv6,7,8 [7, 10, 13] model. In this experiment, lobster images were trained with and without enhanced images using these detection models, and then the mAP of each was compared. The training results were then applied to the detection and counting of lobsters from the video input, which was then transmitted to the tracking algorithm, namely StrongSORT [14]. There are three stages in the monitoring procedure flow for lobsters, as illustrated in Fig 1.

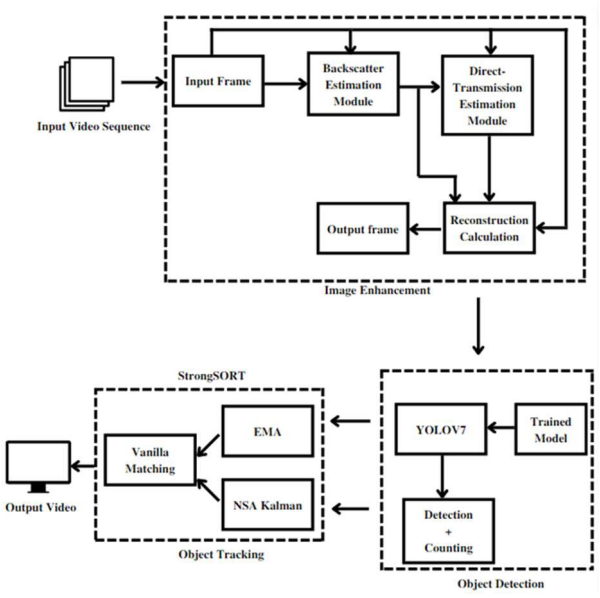


Fig. 1 The proposed intelligent camera for lobster monitoring system.

The first stage involves acquiring the frame from the underwater camera and involves enhancing the frame by using the WaterNet as the model for image enhancement. Next, the processing stage is the frame for the detection of lobster objects using YOLO, which has been trained, and the final stage is counting the number of lobsters detected on the frame. In addition, the third stage involves object tracking on detected objects using the StrongSORT algorithm and assigning an ID to each lobster object detected in the frame. The results of the final stage display the results as a video output for user monitoring.

IV. RESULTS AND DISCUSSIONS

Fig 2 shows the general architecture of the proposed monitoring system in the real environment of the cultivation site. The dataset used to create this lobster detector is a captured of a video in Aruna[1] lobster farm where the cage is about 15 meters under the sea. In total, there are 22 images of lobster captured, which are then processed and augmented to create 45 images, of which 40 are used for training, 2 for validation, and 3 for testing. We are separating the annotation into two scenarios. In the first scheme, we only annotate the main showing part of or the head of the lobsters, whereas, in the second scheme, we annotate the entire body. Additionally, to solve the limitation of the dataset, we conducted the experiment by fine-tuning the detection models by using a pre-trained weight that has been trained on another lobster dataset captured from public videos of a lobster farm[2]. This investigation focuses on comparing the mAP of each image enhancement. The device is divided into training and testing procedures, yet the calculation of the accuracy focuses on the testing part.

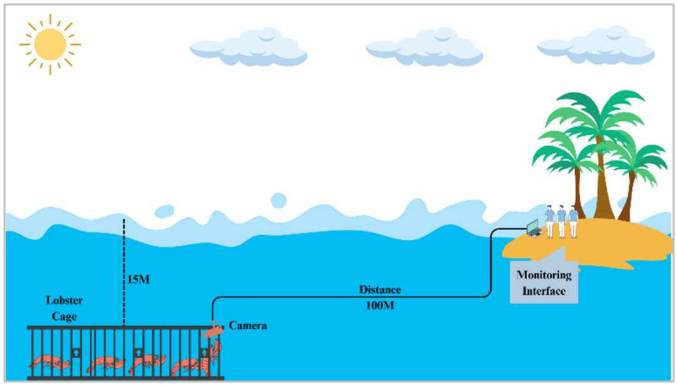


Fig. 2 The proposed of the architecture for lobster monitoring system.

[1] https://aruna.id/

[2] https://youtu.be/a5HpRgDwEgg

We utilized the Tesla T4 graphics processing unit (GPU) with 16GB of GPU memory and 12GB of RAM for the training procedure. The testing procedure for this endeavour employs the Windows 11 operating system and GPU NVIDIA GTX 1650 TI 4GB VRAM as well as the interface for using monitoring where the connection between the camera and this PC is about 100 Meters, as shown in Fig 2.

Table 1. The testing accuracy with head annotation of YOLOv6

Testing Scenario	mAP 0.5	mAP 0.5-0.95
YOLOv6	0.771	0.33
YOLOv6+Water Net	0.771	0.33
YOLOv6+Fine Tuning	0.836	0.405
YOLOv6+WaterNet+Fine Tuning	**0.836**	**0.405**

Table 2. The testing accuracy with head annotation of YOLOv7

Testing Scenario	mAP 0.5	mAP 0.5-0.95
YOLOv7	0.995	0.603
YOLOv7+Water Net	0.995	0.548
YOLOv7+Fine Tuning	0.996	0.51
YOLOv7+WaterNet+Fine Tuning	**0.996**	**0.515**

Table 3. The testing accuracy with head annotation of YOLOv8

Testing Scenario	mAP 0.5	mAP 0.5-0.95
YOLOv8	0.963	0.477
YOLOv8+Water Net	0.957	0.475
YOLOv8+Fine Tuning	0.935	0.499
YOLOv8+WaterNet+Fine Tuning	**0.995**	**0.5**

Table 4. The testing accuracy with full body annotation of YOLOv6

Testing Scenario	mAP 0.5	mAP 0.5-0.95
YOLOv6	0.618	0.21
YOLOv6+Water Net	0.618	0.21
YOLOv6+Fine Tuning	**0.845**	**0.358**
YOLOv6+WaterNet+Fine Tuning	**0.845**	**0.358**

Table 5. The testing accuracy with full body annotation of YOLOv7

Testing Scenario	mAP 0.5	mAP 0.5-0.95
YOLOv7	0.982	0.388
YOLOv7+Water Net	0.926	0.437
YOLOv7+Fine Tuning	0.996	0.483
YOLOv7+WaterNet+Fine Tuning	**0.996**	**0.514**

Table 6. The testing accuracy with full body annotation of YOLOv8

Testing Scenario	mAP 0.5	mAP 0.5-0.95
YOLOv8	0.995	0.614
YOLOv8+Water Net	**0.995**	**0.618**
YOLOv8+Fine Tuning	0.995	0.557
YOLOv8+WaterNet+Fine Tuning	0.995	0.551

Based on the experimental results in Table 1 to Table 6, several scenarios are: original YOLO models, added image enhancement or WaterNet, and added finetuning. It is evident that the combination of underwater image enhancement or WaterNet and YOLOv7 is useful for achieving the mean Average Precision or mAP 0.5 and also 0.5-0.95 with 99.6% and 51.5 %, respectively. It is only displayed for YOLOv7 head annotation, and mAP increases when underwater image enhancement is added. It is also particularly useful for YOLOv6 and YOLOv8. Moreover, Fig 3 shows the detected lobster from the surveillance system with the counted number.

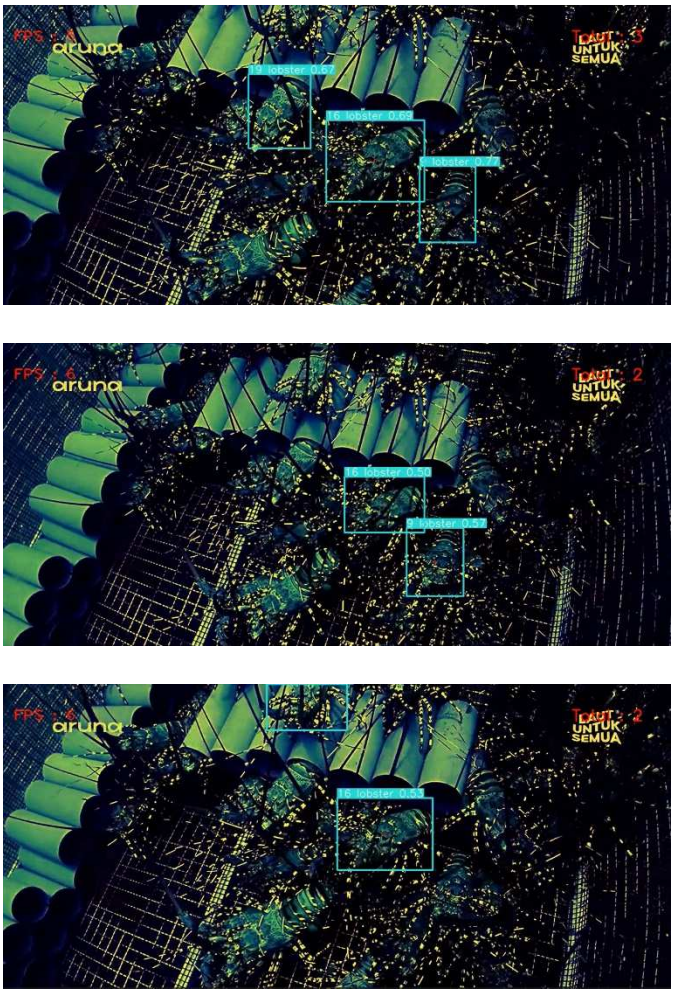


Fig. 3 The captured lobster from the monitoring system.

V. CONCLUSIONS

Our experiments have demonstrated that underwater image enhancement increases the mAP in some scenarios on YOLOv6, YOLOv7, and YOLOv8, with the most significant benefit in mAP 0.5-0.95 with YOLOv7 when the finetuning scenario is added. This result shows that the proposed work is suitable for the intelligent system of lobster surveillance camera.

ACKNOWLEDGMENT

This work was partially supported by Indonesian Endowment Fund for Education (LPDP) Scholarships and financially supported by the Doctoral Dissertation Research Grant from the Ministry of Research and Technology of the Republic of Indonesia, the Directorate of Research and

Community Service at Telkom University, and Institut Teknologi Bandung.

REFERENCES

[1] A. Ayob, K. Khairuddin, Y. Mustafah, A. Salisa, and K. Kadir, "Analysis of pruned neural networks (MobileNetV2-YOLO v2) for underwater object detection," *Proceedings of the 11th National Technical Seminar on Unmanned System Technology 2019: NUSYS'19*, pp. 87-98, 2021.

[2] X. Chen, P. Zhang, L. Quan, C. Yi, and C. Lu, "Underwater image enhancement based on deep learning and image formation model," *arXiv preprint arXiv:2101.00991*, no. 1, 2021.

[3] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "Strongsort: Make deepsort great again," *IEEE Transactions on Multimedia*, no. 1, 2023.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," no., 1960.

[5] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83-97, 1955.

[6] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 4376-4389, 2019.

[7] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, and W. Nie, "YOLOv6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, no. 1, 2022.

[8] H. Liu, P. Song, and R. Ding, "WQT and DG-YOLO: Towards domain generalization in underwater object detection," *arXiv preprint arXiv:2004.06333*, no. 1, 2020.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, no. 1, 2015.

[10] J. Solawetz, "What is YOLOv8? The Ultimate Guide," Jan, 2023.

[11] F. Sultana, A. Sufian, and P. Dutta, "A review of object detection models based on convolutional neural network," *Intelligent computing: image processing based applications*, no. 1, pp. 1-16, 2020.

[12] C.-C. Wang and H. Samani, "Object detection using transfer learning for underwater robot," *2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, pp. 1-4, 2020.

[13] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475, 2023.

[14] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *2017 IEEE international conference on image processing (ICIP)*, pp. 3645-3649, 2017.

[15] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, no. 1, 2023.

A UAV Indoor Obstacle Avoidance System Based on Deep Reinforcement Learning

Chun-Huang Lo and Chung-Nan Lee

Department of Computer Science and Engineering

National Sun Yat-Sen University, Kaohsiung, Taiwan

E-mail: cnlee@mail.cse.nsysu.edu.tw Tel: +886-7-5252000 ext. 4335

***Abstract*—This paper presents a vision-based unmanned aerial vehicle (UAV) indoor obstacle avoidance using a deep reinforcement learning (DRL). The system consists of two parts a depth map compression and a UAV control. For the depth map compression part, the pre-trained Variational AutoEncoder model is used to improve obstacle avoidance and reduce the training time. The states include the UAV information, gray images and compressed features. A dueling double deep recurrent Q network model is used to control the UAV. The network is trained in the AirSim simulation. To validate the performance of the proposed algorithm some simulations are conducted. The results show that the proposed algorithm can avoid obstacles at a fast speed in a narrow space, and fly through a difficult L-shaped corner in an indoor simulation.**

Keywords—Unmanned Aerial Vehicle, Obstacle Avoidance, Deep Reinforcement Learning, AirSim, Variational AutoEncoder

I. INTRODUCTION

In recent years, the development of unmanned aerial vehicles (UAVs) has been advancing rapidly, and they are increasingly being used in various fields to complete tasks. For example, UAVs are used for a large-scale crop spraying, soil and crop analysis, or search and rescue mission, product delivery. In addition to the common outdoor applications, UAVs are also used indoors, such as indoor cleaning and disinfection, as well as patrolling factories, campuses, offices and managing warehouses.

Regardless of indoor or outdoor scenarios, during the execution of tasks, UAVs will face the situation that needs to avoid obstacles in front of them. The basis for a UAV to perform obstacle avoidance is to be able to know the distance between itself and the obstacle. Currently, there are various technologies that can be used to measure distance, including ultrasonic, infrared, laser and visual, each has its advantages and disadvantages. However, if we consider costs, power consumption, and weight, using only visual sensors for obstacle avoidance could be a better choice on different UAVs. The difference of indoor environment and outdoor environment is that the indoor environment is mostly an enclosed space that makes the control of UAVs more stringent. Also the indoor environment has more types of textures and colors than the outdoor environment, and there are also some objects without textures, such as white walls or glass that increases the difficulty of obstacle avoidance indoors. However, indoor flying is not affected by weather changes and has less interference than outdoor scenes.

Traditional obstacle avoidance methods, such as simultaneous localization and mapping that requires high computing power and accuracy of device, and it cannot detect the walls or glasses without texture. With the flourishing development of deep learning, more and more researches have applied it to the obstacle avoidance, path planning, safe landing, height control for UAV. This paper presents a UAV obstacle avoidance system based on deep reinforcement learning that can avoid the obstacle rapidly in both outdoor and indoor environments.

The contributions of this paper are mainly as follows:

- To add a VAE model to the DRL model to improve the training speed, collision rate, and UAV movement speed.
- To propose an algorithm to adjust the number of training times for different difficulty levels of initial points in DRL that can improve training effectiveness.
- To propose a quick and safe indoor DRL model and a reward algorithm to allow to pass through narrow indoor spaces or right-angled bend successfully.

The organization of this paper is as follows. Section 2 reviews the current obstacle avoidance using deep learning and the architecture of DRL. Section 3 presents the proposed method. In Section 4 the performance evaluation is conducted in the simulator. Conclusions are given in the last section.

II. RELATED WORK

A. Reinforcement Learning

Reinforcement learning is a part of machine learning, but it is different from supervised learning and unsupervised learning. The goal of reinforcement learning is to maximum the future reward. The element of reinforcement learning are shown in Figure 1. The agent selects the best action (A_t) based on the state (S_t) provided by the environment at each time step. The environment then provides a reward (R_{t+1}) and the next state (S_{t+1}) based on the action (A_t) selected by the agent. The design of the reward algorithm is important in reinforcement learning because it directly affects the agent's learning based on what reward should be given in different states.

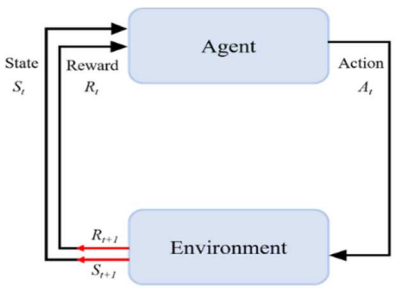


Figure 1. Reinforcement learning overview

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Markov decision processes (MDPs) are the idealized mathematical form for reinforcement learning problems. All processes in MDPs are stochastic, and the next state (S_{t+1}) depends only on the current state (S_t). Reinforcement learning can be transformed into the format of MDPs (S,A,P,R, γ), where S is the set of state space, A is the set of actions, $P = P(s'|s,a), for\ all\ s \in S, a \in A$, indicates the probability of transitioning from s to s' after taking action a; $R = R(s,a,s')$ represents the reward obtained when transitioning from state s to state s' after performing action a; $\gamma \in [0,1]$ is the discount factor. The goal of reinforcement learning is to maximize future rewards, also known as the expected return. Therefore, a policy (π) is used to determine what action to take in a given state. As shown in equation (1), the probability of taking action a in state s is determined by the policy. This formula defines the behavior of a policy.

$$\pi(a|s) = \mathrm{P}(A = a|S = s) \tag{1}$$

An agent can have many policies, and to find the best one requires evaluating different policies using a value function. Equation (2) shows the expected return in the future G_t, where γ is a discount factor representing the degree of emphasis on the future rewards, with $0 \leq \gamma \leq 1$. To find the best policy, the state-value function and the action-value function are used for evaluation. Equation (3) represents the state-value function, which calculates the expected return from state s based on policy π. Equation (4) represents the action-value function, which calculates the expected return from taking action a in state s based on policy π.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2}$$

$$V_\pi(s) = E[G_t|S_t = s] \tag{3}$$

$$Q_\pi(s,a) = E[G_t|S_t = s, A_t = a] \tag{4}$$

The Bellman equation is one of the key of many reinforcement learning algorithms and is used to simplify reinforcement learning or MDP problems. The Bellman equation divides the value function into two parts, immediate reward and future reward weighted by γ. Equations (5) and (6) represent the state-value function and action-value function using the Bellman equation. By using these two value functions, one can find the optimal policy that maximizes the expected return. The optimal policy is denoted as π^*, and the two optimal value functions are represented in equations (7) and (8). With these two optimal value functions, one can find the best policy.

$$V_\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s'} p(s', r|s, a)[r(s, a, s') + \gamma V_\pi(s')] \tag{5}$$

$$Q_\pi(s,a) = \sum_{s'} p(s', r|s, a)[r(s, a, s') + \gamma \sum_{a'} \pi(a'|s') \cdot Q_\pi(s', a')] \tag{6}$$

$$V_{\pi^*}(s) = max_{a \in A} \sum p(s', r|s, a)[r + \gamma V_{\pi^*}(s')] \tag{7}$$

$$Q_{\pi^*}(s,a) = \sum_{s'} p(s', r|s, a)[r + \gamma max_{a'} Q_{\pi^*}(s', a')] \tag{8}$$

But in reality, the probability of state transitions is unknown, and the environment is not fully understood. Therefore, it is necessary to estimate the value function through experience by Monte Carlo (MC) and Temporal Difference (TD) that are called model-free methods, as they do not model the environment, but instead estimate V(s) by sampling enough rounds. With the MC method, updates are made only after a complete episode is run, while TD updates can be made after each state is completed through the concept of bootstrapping.

B. DQN

Mnih et al. [1] first proposed using Deep Q-Network (DQN) to play Atari games and beat human. By combining deep learning with reinforcement learning, the Q-table used in Q-learning, which is limited in recording states and actions, is replaced with a Q-network deep learning model to approximate the optimal policy, thus solve the limited capacity problem of the Q-table. Hasselt et al. [2] proposed Double DQN to reduce the problem of overestimation of Q values in DQN. Wang et al. [3] proposed Dueling DQN, which modified the architecture of DQN to improve the training performance and reduce the overestimation of Q values. Hausknecht et al. [4] replaced the fully connected layer in DQN with a Long Short-Term Memory (LSTM) model, called DRQN, which performed better than DQN in Atari games with partially observable Markov decision processes. Schaul et al. [5] proposed Prioritized Experience Replay by modifying Experience Replay, By using a prioritized sampling structure, the probability of important experiences being sampled is increased, making the model learning more efficient.

C. Actor-Critic

Actor-Critic is a model-free, off-policy RL algorithm, where the actor is based on policy gradient and the critic is based on Q-learning. The actor chooses an action and the critic judges the goodness of the action. Commonly used actor-critic based algorithms in UAV obstacle avoidance include Deep Deterministic Policy Gradient (DDPG), Twin Delayed DDPG (TD3) and Soft Actor-Critic (SAC).

Lillicrap et al. [6] combined DPG and DQN to propose DDPG, which addresses the inability of DQN to deal with continuous control problems by directly outputting deterministic actions. Fujimoto et al. [7] optimized DDPG and proposed TD3, which includes two additional models compared to DDPG. It uses two critic networks and selects the smaller one as the target Q-value for updating. Additionally, the critic network is updated more frequently than the actor network to ensure the stability of the critic network before training the actor network, reducing the potential for incorrect updates. Furthermore, TD3 adds random noise to the action to improve the stability of the model.

D. UAV obstacle avoidance using DRL

The literature related to deep reinforcement learning are mainly divided into indoor and outdoor scenes. In terms of model architecture, they can be divided into models based on Deep Q Network (DQN) and Actor-Critic. During training, models mostly use two types of data for training: depth maps and color images.

Singla et al. [8] proposed an indoor obstacle avoidance using cGAN+DRQN to train the cGAN model to convert color images to depth maps, then use real photos to do the fine-tuned. This model performs better than DQN and D3QN in obstacle avoidance and power consumption. Doshi et al. [9] trained a Double DQN model using different urban environments, including areas with stationary obstacles and areas with moving obstacles. The UAV is set to fly at a constant speed, and the model outputs five different yaw rates.

Roghair et al. [10] proposed improvements to exploration algorithms using D3RQN, where object detection was performed using RGB images to enhance obstacle range in depth maps. They proposed two exploration algorithms: Convergence Exploration and Guidance Exploration. Ou et al. [11] proposed DispNet to combine with D3RQN. DispNet is used to convert RGB images into depth maps that are then used as input for the D3RQN model. The effectiveness of DDQN, D3QN, DDRQN, and D3RQN were compared, and all models were tested in various scenarios. It achieves a success rate of over 90%. Thomas et al. [12] used the D3QN-SAM model, which showed good obstacle avoidance performance in different environments. The model takes a sequence of 5 consecutive depth maps as input. Kim et al. [13] used the FCRN model to convert color images into depth maps, which were then used as inputs for the D3QN model to perform indoor navigation.

III. PROPOSED METHOD

There are many research about obstacle avoidance based on DRL, but most of them only test in the simulate environment. To test the effectiveness of DRL in real-world scenarios, this system was tested on a 450-axis quadrotor UAV through actual experiments.

We extend the framework based on the previous research [14] and make some improvements. The system architecture is shown in Figure 2. The system uses Nvidia Jetson Nano as an onboard edge device instead of Raspberry Pi and transmits images and control commands through 4G/5G. Image information is obtained through the Intel RealSense D415 depth camera. The control package is ROS.

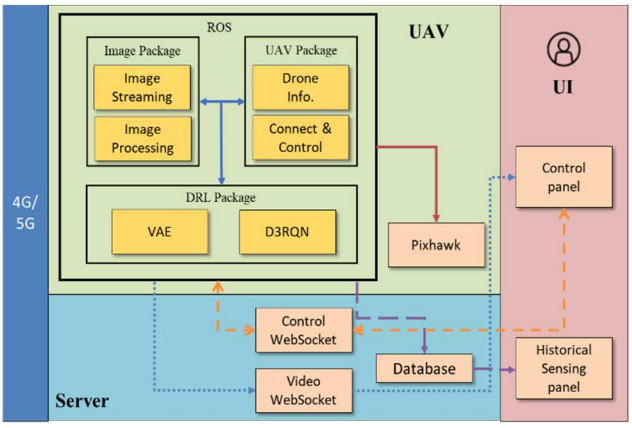


Figure 2. System architecture

Figure 3 shows the system model. The inputs include depth maps, RGB images, and UAV information. The image part is obtained through a depth camera, while the UAV information includes speed, yaw angular velocity, and so on. The RGB images are converted to grayscale, and both images are cropped to remove the regions without depth information. VAE is used to reduce the dimensionality of the depth map, and then the grayscale image, the reduced depth data, and UAV information, along with the past 7 data, are combined as the input to the deep reinforcement learning model. Finally, the output UAV action is obtained.

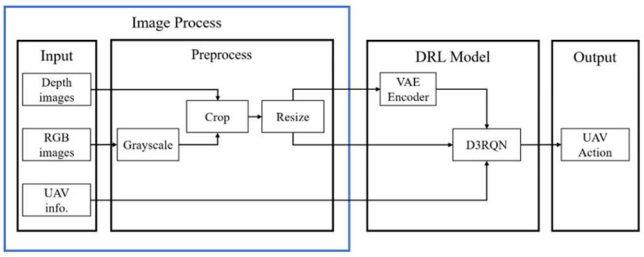


Figure 3. The proposed model

A. Image Processing

The acquired depth maps from the depth camera has some area with holes. In order to reduce the impact of holes on the actual flight test, it is necessary to preprocess the images to remove the holes before they can be used as inputs for the DRL model. Initially, a depth map and a color image with a size of 640x480 are obtained through the depth camera. The color image is aligned with the depth map during acquisition, because the camera positions are different, there are invalid depth areas around the depth map, as shown in the left image of Figure 4. To speed up the obstacle avoidance algorithm, the hole-filling algorithm is not used to repair the depth map, and only the areas around the image that have no depth information are removed.

To ensure the safety of personnel and equipment, a safety threshold is set at the end of the image preprocessing process to prevent collisions of the UAV. Before taking action based on the input image, we first determine whether the area of invalid depth pixels in the image exceeds 50% of the entire depth map. If it does, we judge that the UAV is too close to an obstacle in front, and command it to hover in place. In the deep depth map in the lower left of Figure 4, after cropping the surrounding area, the area of invalid depth pixels is 58%, indicating that the UAV is too close to the pillar on the right, and immediate hovering is necessary to avoid collision.

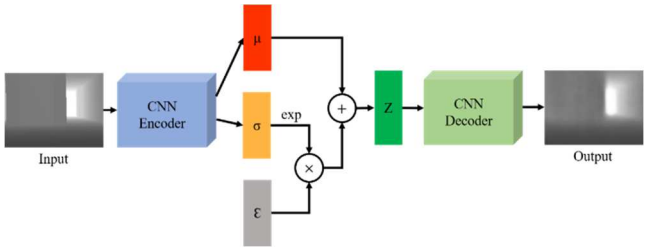


Figure 4. Depth images and Gray images

B. Variational AutoEncoder

VAE is a kind of Autoencoder and its architecture is shown in Figure 5. After passing through the encoder, the original encoded μ and the σ controlling the interference of noise are output. σ is exponentiated to ensure its value is positive, and multiplied by a normal distribution noise ε. The final encoded result is obtained by adding the original encoded result with the noise. Table 1 shows the parameters of each layer in the encoder and decoder, using Leaky ReLU as the activation function, stride is 2, and padding is 1. It was shown that adding VAE to DRL training can speed up the

training process [15]. Finally, the dimensionality-reduced



data Z is 64-dimensional, which reduces the data amount by 1200 times compared to the original 320x240 depth map. The decoder layer is not required in the subsequent D3RQN training and actual use.

Figure 5. The VAE architecture

Table 1 Parameters of each layer in the Encoder and Decoder

Layer	Encoder		Decoder	
	Channels	Kernel size	Channels	Kernel size
Input	1		512	
#1	16	[4, 4]	256	[5, 4]
#2	32	[4, 4]	128	[4, 2]
#3	64	[4, 4]	64	[4, 5]
#4	128	[4, 4]	32	[6, 8]
#5	256	[4, 4]	16	[8, 7]
#6	512	[4, 4]	1	[4, 6]

C. *Dueling Double Deep Recurrent Q Network*

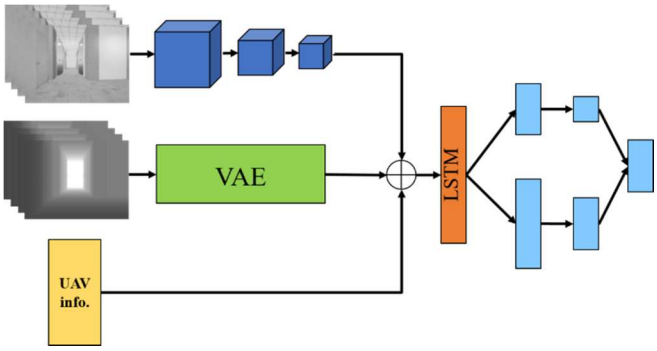
The D3RQN model shown in Figure 6 combining Multi-step learning, Double Q-learning, Dueling network and RNN is used. Multi-step learning can improve the model's effectiveness and train better in tasks with the sparse reward. Double Q-learning and Dueling Network can effectively avoid overestimation of the Q value and accelerate the training speed and performance of the model. Adding RNN to assist in memory can achieve better results in partially-observable markov decision process.

The input of the model consists uav information, the gray images which is genrated by convolution network and depth maps from the VAE model. The data consists eight continous time steps that include the current time and the previous seven time steps. The model will output seven actions that are shown in Table 2. Note that V is velocity(m/s) and ω is rotation angle.

Table 2. UAV action

Choice	$[V, \omega]$
Action 1	[0.2, 0]
Action 2	[0.5, 0]
Action 3	[1.0, 0]
Action 4	$[V, 15^{\circ}]$
Action 5	$[V, -15^{\circ}]$
Action 6	$[V, 30^{\circ}]$
Action 7	$[V, -30^{\circ}]$

.



D. *Training Environment*

AirSim simulator is used as the training environment to provide highly realistic virtual scenes. Through the API of AirSim, UAVs can be controlled to perform different actions, and also can detect the collision of UAVs. The training environment is shown in Figure 7.



Figure 7. Two simulation environments

In the training scenario, there are 20 different starting points with varying difficulty levels. For example, when the starting point is one of the four corners of map, the UAV only needs to move forward without getting too close to the side walls throughout the entire episode. There are no obstacles and no need to turn, making it easier and faster to train. However, for other starting points, after moving forward for a distance, the UAV needs to dodge obstacles or turn at a right angle, resulting in a higher probability of collision and requiring more training and exploration. Therefore, during training, different training frequencies should be given according to the training situation of each starting point, while at least one training frequency is allocated for simpler starting points.

E. *Reward Function*

The reward function is divided into two parts, Reward_speed and Reward_depth, both of them are in the range of [0, 0.5]. During the training phase, three distance sensors located in the middle and on the left and right sides are added to detect obstacles in front. The distances obtained by the distance sensor are defined as Dis_{mid}, Dis_{left}, and Dis_{right} .The termination conditions for an episode are either exceeding max steps or collision of the UAV. The collision of the UAV is determined through the AirSim API or one of the distance sensors is less than 0.5 meter and set the reward to -2 when collision occurs.

The Reward_speed depends on the obstacles in front of UAV, if there is no obstacle in front of UAV, the UAV should go straight with higher speed, also the UAV should use

smaller displacement to avoid obstacles. The UAV's speed and yaw rate are defined as V and ω, respectively, and a threshold value δ is set as the UAV's safety distance. If all three distance sensors are greater than the threshold value δ, the UAV should use a faster speed and smaller angular velocity to avoid obstacles. Otherwise, a larger angular velocity should be used. The calculation of speed reward in the former case is given in equation (9) and the calculation of speed reward in the latter case is given in equation (10).

$$Reward_speed = V \times cos(\omega \times 2)/2 \quad (9)$$

$$Reward_speed = |\omega| \quad (10)$$

The reward_depth is calculated with the depth maps, the depth maps is divided to three parts :left, right and middle as shown in Figure 8. If the action of the UAV is to turn right, the depth map on the right side is used, the same goes for turning left and go straight. If the model focuses on the area that has an obstacle, the reward calculated from that area will be smaller, while a reward from an open area will be larger. This way, the model can learn from the depth map to know which direction it should face to obtain a higher reward. To calculate the reward, the depth map in each of the three parts is first normalized, averaged, and defined as D_{mid}, D_{left}, and D_{right}. Each of these is then divided by the average depth D_mean of the entire depth map and multiplied by 70 to obtain their respective rewards, defined as R_{mid}, R_{left} and R_{right} The calculation formulas are as follows:

$$D_{mid} = \text{mean}(D[90{:}150, 120{:}200]/255) \quad (11)$$

$$R_{mid} = (D_{mid}/D_{mean}) * 70 \quad (12)$$

IV. Experimental Result

This section describes the architecture of the simulation environment, as well as the design of the training and experimentation. The effectiveness of the depth reinforcement learning model in obstacle avoidance is verified in various simulation environments and compared with works in the existing papers. Finally the real test on 450 mm Quadcopter with Jetson Nano is conducted.

A. Variational AutoEncoder model

The hyperparameters of VAE model are shown in Table 3. The depth maps is acquired by random position and angle in the simulate environment . The input and output of VAE model is shown in Figure 9. The encoder layer can extract important features from depth map and the image can be restored through the decoder layer. One can see that the depth map can be roughly restored after going through the VAE, which proves that the VAE can effectively extract important features from the depth map as input for the D3RQN model.

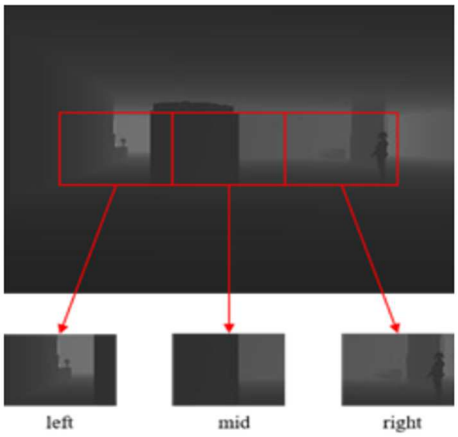


Figure 8. Divided depth image

Table 3. VAE hyperparameter

Hyperparameter	Value
No of images	20,000
Optimizer	Adam
Learning rate	0.001
Epochs	50
Batch size	64

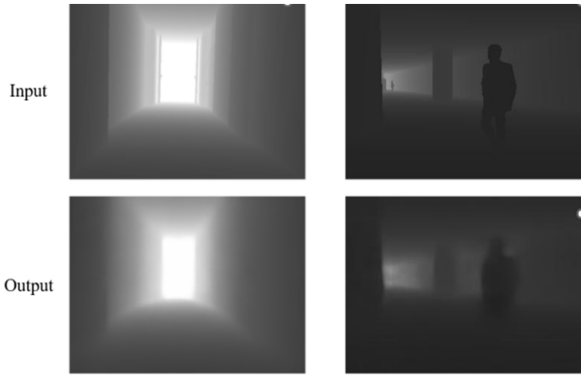


Figure 9. The input and output of VAE

B. Dueling Double Deep Recurrent Q Network

The pre-trained VAE model and the D3RQN model described in Section 3 are used to train in two simulation environments. The initial speed of UAV is set to 0.2m/s, and the height is set to 1 meter above the ground. The termination for each episode is when the UAV collides with an obstacle or the UAV exceeded 150 steps. The hyperparameters of D3RQN are shown in Table 4.

Table 4. D3RQN hyperparameter

Hyperparameter	Value
Max step	150
N-step	4
Optimizer	Adam
Learning rate	0.0001
Episode	4,000
Batch size	16
Target network update	5000
Initial epsilon	0.999
Final epsilon	0.001
Epsilon decaying	10,000
Replay memory size	4,000

C. *Training result*

First, we compared the method of adjusting the number of training times for each starting point and the method of fixing the training for 5 times for each starting point. The result is shown in Figure 10. We can see that the fixed number of training don`t show the significant improvement after training for 2,000 episodes. Giving different training times for different starting points can significantly improve the training effect of the model, allow the model to have more training and exploration opportunities for starting points with poorer performance.

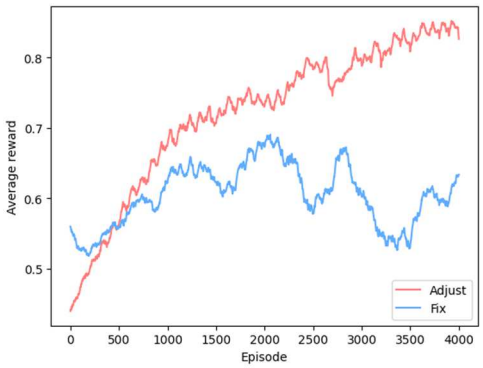


Figure 10. The result of the fixed and the adjust the number of training

Next, we tested whether using VAE can improve the training performance of the model. Using the same model architecture, we modify the VAE model to have the same convolutional layers as the grayscale image. Table 5 shows the results of testing 100 times. Using the VAE model can reduce the collision rate and make the model more likely to move at a faster speed.

Table 5. Comparison with and without the VAE models

model	Collision rate	Avg total step	Minimum step	Avg velocity
w/ VAE	8%	145.62	43	0.97
w/o VAE	13%	143.61	77	0.43

Finally, the different combinations of images as input were compared : using only the depth map, the depth map with color image, and the depth map with grayscale image. Comparison results are given in Figure 11 and Table 6. Using only depth map and using depth map with grayscale image can improve the performance before 1000 episodes, but the grayscale provided additional information for better training of the model. Using the depth map with color image does not improve the performance of model. Table 6 shows that using the depth map with grayscale image as input had better results in terms of collision rate, average total steps, average velocity than the other two combinations.

Table 6. Comparison with different input

Input	Collision rate	Avg total steps	Minimum steps	Avg velocity
Grayscale + Depth	8%	145.62	43	0.97
Depth	11%	143.56	41	0.49
Depth + Color	23%	140.51	59	0.55

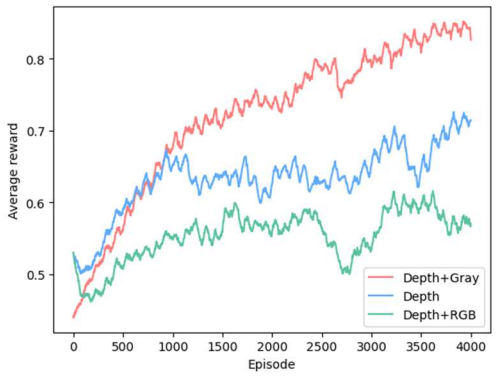


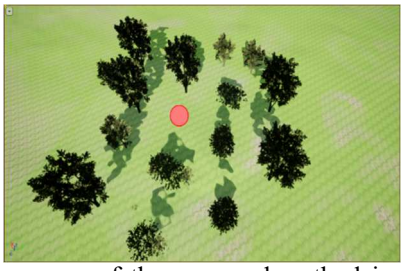
Figure 11. The average reward with different input

D. *Test the model and comparisons*

In order to verify that the model can successfully avoid obstacles in unseen environments without any prior training, three completely different test environments were built in the AirSim simulator. Each test map had 10 different starting points, and 10 episodes were tested at each starting point before moving on to the next one, for a total of 100 tests. The maps are shown in Figure 12.



Figure 12. Test environments



Performance of the proposed method is compared with some existing works. The training scene is similar to the one in [13], which is an indoor scene with different obstacles and turns. Therefore, the reward algorithm and hyperparameters in that paper were used, with only the action method modified to match the experiment here. A total of 10,000 episodes were trained. The training scene in [11] is an outdoor scene with only straight paths and no turns, and the reward algorithm used

in that paper may not be suitable for the scene here. Therefore, both the reward algorithm and the action method were modified to match our experiment. And a total of 4,000 episodes were trained. After training in their respective scenes, both methods were tested 100 times in both training and testing scenes, and compared in terms of collision rate, average total steps, minimum steps, and average velocity, as shown in Tables 7 and 8. The proposed method performs better than other two papers in terms of collision rate and velocity.

Table 7. Comparisons with different inputs on training

	Training environment			
Agent	Collision rate	Avg total steps	Minimum steps	Avg velocity
Random	92%	80.44	23	0.54
D3RQN [17]	26%	138.82	52	0.56
D3QN [21]	11%	142.24	56	0.67
The proposed algorithm	8%	145.62	43	0.97

Table 8. Comparisons with different inputs on test

	Test environment			
Agent	Collision rate	Avg total steps	Minimum steps	Avg velocity
Random	85%	86.57	25	0.54
D3RQN [17]	20%	134.77	33	0.51
D3QN [21]	17%	144.15	86	0.49
The proposed algorithm	14%	141.35	45	0.92

V. CONCULUSION

This paper implements an indoor obstacle avoidance system for UAVs based on deep reinforcement learning. A depth camera installed in front of the UAV provides grayscale images and depth maps, both of them are used as inputs for a deep reinforcement learning model that combines a variational autoencoder with a D3RQN algorithm. The proposed method was tested in simulated environments for training and testing, and compared to some existing works. The results show that the proposed algorithm can outperform Q3DN about 17% and 87% for collision rate and velocity.

Figure 13. Outdoor simulation environment

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement

[2] H. v. Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-learning," *Association for the Advancement of Artificial Intelligence*, pp. 2094-2100, 2016.

[3] Z. Wang, T. Schaul, M. Hessel, H. v. Hasselt, M. Lanctot, and N. d. Freitas, "Dueling network architectures for deep reinforcement learning," *Proc. Int. Conf. Learning Representations*, 2016.

[4] M. Hausknecht and P. Stone, "Deep Recurrent Q-Learning for Partially Observable MDPs," *2015 AAAI fall symposium series*, 2015.

[5] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized Experience Replay," *International Conference on Learning Representations*, 2016.

[6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous Control with Deep Reinforcement Learning.," *arXiv preprint arXiv:1509.02971*, 2015.

[7] S. Fujimoto, H. v. Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-critic Methods," *International Conference on Machine Learning*, pp. 1587-1596, 2018.

[8] A. Singla, S. Padakandla, and S. Bhatnagar, "Memory-based Deep Reinforcement Learning for Obstacle Avoidance in UAV with Limited Environment Knowledge," *arXiv:1811.03307*, 2018.

[9] N. Doshi, S. Gujar, A. Sutvani, and C. Morato, "Double DQN based Autonomous Obstacle Avoidance for Quadcopter Navigation," 2019.

[10] J. Roghair, K. Ko, A. E. N. Asli, and A. Jannesari, "A Vision Based Deep Reinforcement Learning Algorithm for UAV Obstacle Avoidance," *Proceedings of SAI Intelligent Systems Conference*, pp. 115-128, 2021.

[11] J. Ou, X. Guo, M. Zhu, and W. Lou, "Autonomous Quadrotor Obstacle Avoidance Based on Dueling Double Deep Recurrent Q-learning with Monocular Vision," *Neurocomputing, 441*, pp. 300-310, 2021.

[12] D.-G. Thomas, D. Olshanskyi, K. Krueger, T. Wongpiromsarn, and A. Jannesari, "Interpretable UAV Collision Avoidance using Deep Reinforcement Learning," *arXiv preprint arXiv:2105.12254,* 2021.

[13] M. Kim, J. Kim, M. Jung, and H. Oh, "Towards Monocular Vision-based Autonomous Flight through Deep Reinforcement Learning," *Expert Systems with Applications, 198, 116742*, 2022.

[14] Y. C. Lin, A UAV Precision Landing and Obstacle Avoidance System Based on Point Cloud and Artificial Potential Field, *Master Thesis, National Sun Yat-sen University*, 2021.

[15] Z. Xue and T. Gonsalves, "Vision Based UAV Obstacle Avoidance by Deep Reinforcement Learning," *AI 2021, 2(3), 366-380,* 2021.

Approximate modeling of malware diffusion on wireless mobile devices

Hideyoshi Miura*, Shoya Abukawa*, Tomotaka Kimura†, and Kouji Hirata‡
* Graduate School of Science and Engineering, Kansai University, Osaka, Japan
E-mail: {k846996, k507593}@kansai-u.ac.jp
† Faculty of Science and Engineering, Doshisha University, Kyoto, Japan
E-mail: tomkimur@mail.doshisha.ac.jp
‡ Faculty of Engineering Science, Kansai University, Osaka, Japan
E-mail: hirata@kansai-u.ac.jp

***Abstract*—In recent years, malware infecting mobile devices has become serious issues. In this paper, we propose an approximate model representing the behavior of malware diffusion on wireless mobile devices with ordinary differential equations. The proposed model divides a target area into some fixed-size compartments. Then, for each compartment, it provides ordinary differential equations considering the movement of devices between the compartments. By doing so, we can reduce the number of ordinary differential equations to represent the malware diffusion on mobile devices, compared with conventional epidemic models. As a result, the computation time to obtain solutions showing infection dynamics can be reduced. Through numerical calculations, we show that our proposed epidemic model well approximates results obtained from a conventional epidemic model with low computation time.**

I. Introduction

In recent years, cyber attacks using malware has become serious issues [4]. Furthermore, malware infecting mobile devices has appeared [2]. For example, BlueBorne [5] exploits vulnerabilities of Bluethooth connections to attack mobile devices. In response to these issues, some epidemic model considering malware diffusion on mobile devices have been proposed.

In [3], the authors have introduced a stochastic epidemic model of malware on mobile devices using Bluetooth connections. They have discussed the behavior of malware spreading on mobile devices with the use of results obtained from simulation experiments. The authors in [6] have proposed a deterministic epidemic model that indicates the infection dynamics of malware spreading on mobile devices, with the use of ordinary differential equations. They have shown that the results of the numerical calculation of the epidemic model are very close to those obtained from simulation experiments. Although this model can obtain results in short time compared with simulation experiments, the computation time is still large. This is because this epidemic model requires a lot of ordinary differential equations to represent infection dynamics according to the scale of assumed environments.

To resolve this problem, in this paper, we propose an approximate model representing infection dynamics on mobile devices based on the deterministic epidemic model. The proposed model divides a target area into some fixed-size compartments. It then provides ordinary differential equations for each compartment, considering the movement of devices between the compartments. By doing so, the proposed model aims to reduce the number of ordinary differential equations to represent the malware diffusion on mobile devices. As a result, the computation time to obtain solutions showing infection dynamics can be reduced. We show that the results of our proposed epidemic model properly approximate those of a conventional epidemic model with low computation time through numerical calculations.

II. System model

In [6], the authors have proposed the deterministic epidemic model of malware diffusion on mobile devices, assuming the following system model. Note that our approximate model proposed in this paper also assumes the same system model.

Fig. 1 illustrates the system model assumed in this paper. In this model, there is a two-dimensional space represented by grid $\mathcal{G}$ of $M \times N$ cells as a target area where there exist a set $\mathcal{H}$ of mobile hosts. Each mobile host is located on one of the cells and can move to one of the adjacent cells. Let $\mathcal{C}_{(i,j)}$ denote the neighborhood cells of cell $(i,j) \in \mathcal{G}$ ($1 \leq i \leq M$, $1 \leq j \leq N$, $i, j \in \mathbb{N}$), which is given by

$$\mathcal{C}_{(i,j)} = \{(m,n) \mid |m-i| \leq 1, |n-j| \leq 1\}.$$

Furthermore, we denote the adjacent cells of cell $(i,j) \in \mathcal{G}$ as $\mathcal{A}_{(i,j)} = \mathcal{C}_{(i,j)} \setminus \{(i,j)\}$.

As shown in Fig. 1, the communication range of each host corresponds to the neighborhood cells $\mathcal{C}_{(i,j)}$. Specifically, each mobile host on cell (i,j) can communicate with other mobile hosts on the neighborhood cells $\mathcal{C}_{(i,j)}$. On the other hand, the adjacent cells $\mathcal{A}_{(i,j)}$ indicates the moving range of hosts at cell (i,j). In this model, we consider discrete timesteps t_d ($t_d = 0, 1, 2, \dots$). We denote the location (i.e., cell) of host $h \in \mathcal{H}$ at time t_d as $c_h(t_d) = (i,j)$. Each mobile host on each cell $(i,j) \in \mathcal{G}$ moves to one of adjacent cells $\mathcal{A}_{(i,j)}$ randomly selected at each time step. For instance, $c_h(t_d+1)$ is equal to $(i, j+1)$ if host h on cell (i,j) moves to the right cell at time t_d.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

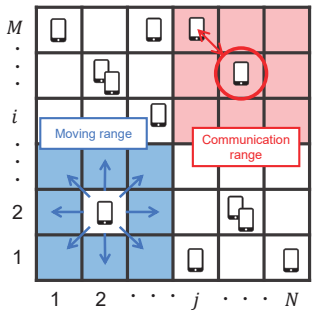


Fig. 1: System model.

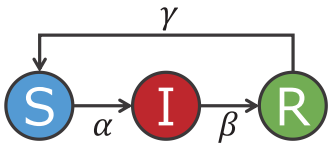


Fig. 2: SIRS model.

As a basic epidemic model that represents the state of each host, we assume an SIRS model shown in Fig. 2. Each host on the grid $\mathcal{G}$ has one of three states, i.e., Susceptible (S), Infected (I), and Recovered (R) states. The state S means that the host is not infected with malware but it has vulnerabilities that can get infected with malware. The state I means that the host is infected with malware. The state R means that the host is protected against malware. At each time step, the state transition of each host occurs based on the following events:

1) With the probability α_d, an infected host on each cell $(i,j) \in \mathcal{G}$ infects a susceptible host located on the neighborhood cells $\mathcal{C}_{(i,j)}$. In the next time step, the susceptible host transitions to an infected host.
2) With the probability β_d, an infected host removes the malware from itself. In the next time step, the infected host transitions to a recovered host.
3) With the probability γ_d, a recovered host is discovered a new vulnerability. In the next time step, the recovered host transitions to a susceptible host.

III. Epidemic Model of Malware Diffusion on Mobile Devices [6]

We here explain the epidemic model introduced in [6]. The epidemic model considers the continuous time environment instead of discrete timesteps discussed in the system model. In this case, discrete timestep t_d corresponds to continuous time t (≥ 0). Accordingly, the probabilities α_d, β_d, and γ_d correspond α, β, and γ, respectively, where α, β, and γ are parameters indicating the infection rate, the elimination rate, and the vulnerability discovery rate, respectively.

Let $S_{(i,j)}(t)$, $I_{(i,j)}(t)$, and $R_{(i,j)}(t)$ denote the number of susceptible hosts, infected hosts, and recovered hosts, respectively, on cell (i,j) at time t. The epidemic model based on the SIRS model represents the infection dynamics with the following differential equations, where $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$:

$$\begin{aligned}\frac{d}{dt}S_{(i,j)}(t) =& -\alpha S_{(i,j)}(t)\left\{\sum_{(m,n)\in\mathcal{C}_{(i,j)}} I_{(m,n)}(t)\right\} \\ &+ \gamma R_{(i,j)}(t) - \mu S_{(i,j)}(t) \\ &+ \mu \sum_{(m,n)\in\mathcal{A}_{(i,j)}} \frac{1}{|\mathcal{A}_{(n,m)}|} S_{(m,n)}(t),\end{aligned} \tag{1}$$

$$\begin{aligned}\frac{d}{dt}I_{(i,j)}(t) =& \alpha S_{(i,j)}(t)\left\{\sum_{(m,n)\in\mathcal{C}_{(i,j)}} I_{(m,n)}(t)\right\} \\ &- \beta I_{(i,j)}(t) - \mu I_{(i,j)}(t) \\ &+ \mu \sum_{(m,n)\in\mathcal{A}_{(i,j)}} \frac{1}{|\mathcal{A}_{(m,n)}|} I_{(m,n)}(t),\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{d}{dt}R_{(i,j)}(t) =& \beta I_{(i,j)}(t) - \gamma R_{(i,j)}(t) - \mu R_{(i,j)}(t) \\ &+ \mu \sum_{(m,n)\in\mathcal{A}_{(i,j)}} \frac{1}{|\mathcal{A}_{(m,n)}|} R_{(m,n)}(t),\end{aligned} \tag{3}$$

where μ denotes the moving rate of mobile hosts, which implies the migration speed of hosts to adjacent cells. The detailed discussion about the infection dynamics has been provided in [6].

This epidemic model formulates ordinary differential equations for each cell $(i,j) \in \mathcal{G}$ to represent infection dynamics on mobile devices. As a result, it requires $3 \times M \times N$ ordinary differential equations, and thus it spends a lot of time to calculate solution. Our approximate model proposed in this paper aims to reduce the computation time by decreasing the number of ordinary differential equations.

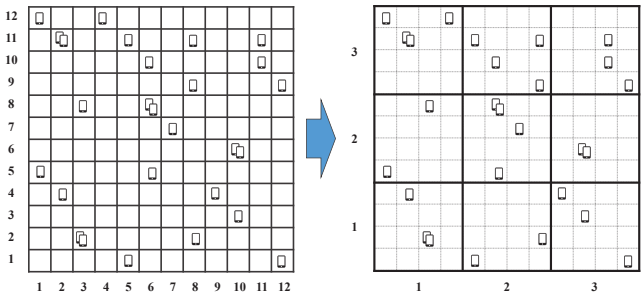


Fig. 3: Example of the proposed approximate model ($G = 12, d = 3$).

IV. Proposed Approximate Model

A. Assumption

In what follows, for simplicity, we consider scenarios where a target area consists of $G \times G$ cells, i.e., $M = N = G$. The proposed approximate model divides the target area into a set $\mathcal{D}$ of $d \times d$ compartments as shown in Fig. 3. In this case, each compartment consists of g^2 cells, where $g = \frac{G}{d}$. The proposed approximate model formulates ordinary differential

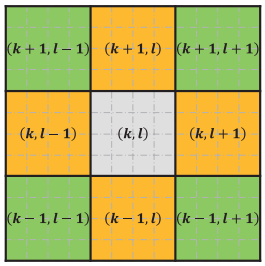


Fig. 4: $\mathcal{A}^{\text{side}}_{(k,l)}$ and $\mathcal{A}^{\text{diag}}_{(k,l)}$ of compartment (k,l).

equations for each compartment instead of each cell to reduce the number of equations.

Let $\mathcal{A}^{\text{side}}_{(k,l)}$ and $\mathcal{A}^{\text{diag}}_{(k,l)}$ denote sets of adjacent compartments of compartment $(k,l) \in \mathcal{D}$ $(1 \le k \le d, 1 \le l \le d, k, l \in \mathbb{N})$, which is defined as

$$\mathcal{A}^{\text{side}}_{(k,l)} = \{(k-1,l),(k,l-1),(k,l+1),(k+1,l)\},$$
$$\mathcal{A}^{\text{diag}}_{(k,l)} = \{(k-1,l-1),(k-1,l+1),(k+1,l-1),(k+1,l+1)\}.$$

Fig. 4 shows an example of $\mathcal{A}^{\text{side}}_{(k,l)}$ and $\mathcal{A}^{\text{diag}}_{(k,l)}$. In this figure, $\mathcal{A}^{\text{side}}_{(k,l)}$ is yellow-colored compartments and $\mathcal{A}^{\text{diag}}_{(k,l)}$ is green-colored compartments. Let K denote the number of mobile hosts in the target area. Let $S_{(k,l)}(t)$, $I_{(k,l)}(t)$, and $R_{(k,l)}(t)$ denote the number of susceptible hosts, infected hosts, and recovered hosts, respectively, on compartment (k,l) at time t. Note that $S(t) + I(t) + R(t) = K$ holds where $S(t) = \sum_{(k,l)\in\mathcal{D}} S_{(k,l)}(t)$, $I(t) = \sum_{(k,l)\in\mathcal{D}} I_{(k,l)}(t)$, and $R(t) = \sum_{(k,l)\in\mathcal{D}} R_{(k,l)}(t)$.

B. Formulation

Our proposed approximate model represents the infection dynamics for each compartment $(k,l) \in \mathcal{D}$ with the following differential equations, where $k, l = 1, 2, \ldots, d$:

$$\begin{aligned}\frac{d}{dt}S_{(k,l)}(t) =& -\frac{\alpha(3g-2)^2}{g^4}S_{(k,l)}(t)I_{(k,l)}(t)\\ &-\frac{\alpha}{g^4}S_{(k,l)}(t)\left\{\sum_{(m,n)\in\mathcal{A}^{\text{diag}}_{(k,l)}} I_{(m,n)}(t)\right.\\ &\left.+\sum_{(m,n)\in\mathcal{A}^{\text{side}}_{(k,l)}}(3g-2)I_{(m,n)}(t)\right\}\\ &+\gamma R_{(k,l)}(t)+\mu\left\{S^{\text{in}}_{(k,l)}(t)-S^{\text{out}}_{(k,l)}(t)\right\},\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{d}{dt}I_{(k,l)}(t) =& \frac{\alpha(3g-2)^2}{g^4}S_{(k,l)}(t)I_{(k,l)}(t)\\ &+\frac{\alpha}{g^4}S_{(k,l)}(t)\left\{\sum_{(m,n)\in\mathcal{A}^{\text{diag}}_{(k,l)}} I_{(m,n)}(t)\right.\\ &\left.+\sum_{(m,n)\in\mathcal{A}^{\text{side}}_{(k,l)}}(3g-2)I_{(m,n)}(t)\right\}\\ &-\beta I_{(k,l)}(t)+\mu\left\{I^{\text{in}}_{(k,l)}(t)-I^{\text{out}}_{(k,l)}(t)\right\},\end{aligned} \tag{5}$$

$$\begin{aligned}\frac{d}{dt}R_{(k,l)}(t) =& \beta I_{(k,l)}(t)-\gamma R_{(k,l)}(t)\\ &+\mu\left\{R^{\text{in}}_{(k,l)}(t)-R^{\text{out}}_{(k,l)}(t)\right\}.\end{aligned} \tag{6}$$

$X^{\text{in}}_{(k,l)}(t)$ $(X \in \{S, I, R\})$ denotes the average number of hosts in state X that come into the compartment (k,l) from the adjacent compartments. On the other hand, $X^{\text{out}}_{(k,l)}(t)$ denotes the average number of hosts in state X that move from the compartment (k,l) to the adjacent compartments, which will be defined later.

In (4), the first term indicates the average number of susceptible hosts that get infected with the malware in compartment (k,l). Susceptible hosts on a cell can get infected by contact with infected hosts in the neighborhood cells as discussed in the system model. We thus consider three cases of the contact shown in Fig. 5, namely, corner, edge, and internal cells. For susceptible hosts at a corner cell of the compartment, the number of neighborhood cells is 4. Thus, the average number of contacts between the susceptible hosts and infected hosts in the neighborhood cells is given by $\frac{S_{(k,l)}(t)}{g^2}\frac{4I_{(k,l)}(t)}{g^2}$, where $\frac{S_{(k,l)}(t)}{g^2}$ and $\frac{I_{(k,l)}(t)}{g^2}$ indicates the number of susceptible hosts and infected hosts, respectively, per cell on compartment (k,l). Also, for susceptible hosts at an edge cell and an internal cell, the average number of contacts between the susceptible hosts and infected hosts in the neighborhood cells is given by $\frac{S_{(k,l)}(t)}{g^2}\frac{6I_{(k,l)}(t)}{g^2}$ and $\frac{S_{(k,l)}(t)}{g^2}\frac{9I_{(k,l)}(t)}{g^2}$. Because the numbers of corner, edge, and internal cells are 4, $4(g-2)$, and $(g-2)^2$, respectively, the total number of contacts between the susceptible hosts and infected hosts in the neighborhood cells can be calculated as follows:

$$\begin{aligned}&\left\{\frac{S_{(k,l)}(t)}{g^2}\frac{4I_{(k,l)}(t)}{g^2}\times 4\right\}\\ &+\left\{\frac{S_{(k,l)}(t)}{g^2}\frac{6I_{(k,l)}(t)}{g^2}\times 4(g-2)\right\}\\ &+\left\{\frac{S_{(k,l)}(t)}{g^2}\frac{9I_{(k,l)}(t)}{g^2}\times (g-2)^2\right\}\\ &=\left\{\frac{9g^2-12g+4}{g^4}\right\}S_{(k,l)}(t)I_{(k,l)}(t)\\ &=\frac{(3g-2)^2}{g^4}S_{(k,l)}(t)I_{(k,l)}(t).\end{aligned} \tag{7}$$

The first term in (4) is given by the product of the infection rate α and (7).

The second term in (4) indicates the average number of susceptible hosts getting infected by contact with infected hosts located in adjacent compartments $\mathcal{A}^{\text{diag}}_{(k,l)} \cup \mathcal{A}^{\text{side}}_{(k,l)}$. This can be obtained from similar discussion to the first term. The third term means the average number of recovered hosts that transition to the susceptible state due to the discovery of a new vulnerability. The fourth term represents the average number of susceptible hosts that come into the compartment (k,l) or move to adjacent compartments.

Similarly, in (5), the first term and second term indicate the average number of susceptible hosts getting infected by contact with infected hosts, i.e., the average increase in infected hosts.

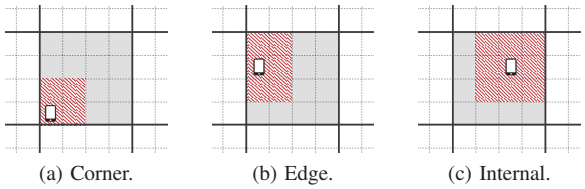
The third term means the average number of infected hosts eliminating the malware and moves to the recovered state. The fourth term represents the average number of infected hosts that come into the compartment (k,l) or move to adjacent compartments. In (6), the first term and second term correspond to the third term of (5) and the third term of (4), respectively. The third term indicates the average number of recovered hosts that come into the compartment (k,l) or move to adjacent compartments.

We here discuss $X^{\mathrm{in}}_{(k,l)}(t)$ and $X^{\mathrm{out}}_{(k,l)}(t)$ $(X \in \{S,I,R\})$ in (4)-(6). We suppose that hosts randomly move to adjacent cells. Therefore, we consider the cases of migration between compartments shown in Fig. 6. We first consider the case where hosts in adjacent compartments come into compartment (k,l), which is shown in Fig. 6(a). For each compartment $(m,n) \in \mathcal{A}^{\mathrm{diag}}_{(k,l)}$, only hosts at a corner cell can come into compartment (k,l). Furthermore, the probability that hosts at the corner cell of compartment (m,n) moves to the adjacent cell of compartment (k,l) is $1/8$ because the hosts randomly select a cell to move from among 8 adjacent cells. Therefore, the average number of such hosts is $\frac{1}{8}\frac{X_{(m,n)}(t)}{g^2}$. On the other hand, for each compartment $(m,n) \in \mathcal{A}^{\mathrm{side}}_{(i,j)}$, hosts at corner or edge cells can come into compartment (k,l). The numbers of those cells are 2 and $g-2$, respectively. The probabilities are $2/8$ and $3/8$, respectively. Thus, the average number of hosts that come from the compartment is given by $\frac{2}{8}\frac{X_{(m,n)}(t)}{g^2} \times 2 + \frac{3}{8}\frac{X_{(m,n)}(t)}{8g^2} \times (g-2) = \frac{(3g-2)X_{(m,n)}(t)}{8g^2}$. Therefore, $X^{\mathrm{in}}_{(k,l)}(t)$ is given by

$$X^{\mathrm{in}}_{(k,l)}(t) = \frac{1}{8g^2}\left\{ \sum_{(m,n)\in\mathcal{A}^{\mathrm{diag}}_{(k,l)}} X_{(m,n)}(t) + \sum_{(m,n)\in\mathcal{A}^{\mathrm{side}}_{(k,l)}} (3g-2)X_{(m,n)}(t) \right\}. \tag{8}$$

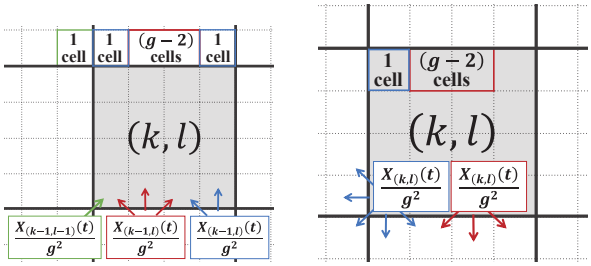
We then consider the case where hosts in compartment (k,l) move to adjacent compartments. In this case, hosts at corner or edge cells can move to adjacent compartments as shown in Fig. 6(b). The numbers of those cells in the compartment is 4 and $4(g-2)$, respectively. The probabilities are $5/8$ and $3/8$, respectively. Thus, $X^{\mathrm{out}}_{(k,l)}(t)$ is given by

$$\begin{aligned} X^{\mathrm{out}}_{(k,l)}(t) &= \left\{\frac{5}{8}\frac{X_{(k,l)}(t)}{g^2} \times 4\right\} + \left\{\frac{3}{8}\frac{X_{(k,l)}(t)}{g^2} \times 4(g-2)\right\} \\ &= \left\{\frac{12g-4}{8g^2}\right\} X_{(k,l)}(t) \\ &= \frac{3g-1}{2g^2} X_{(k,l)}(t). \end{aligned} \tag{9}$$



(a) Corner. (b) Edge. (c) Internal.

Fig. 5: Contact area.



(a) Move-in. (b) Move-out.

Fig. 6: Migration between compartments.

V. Numerical Calculations

A. Model

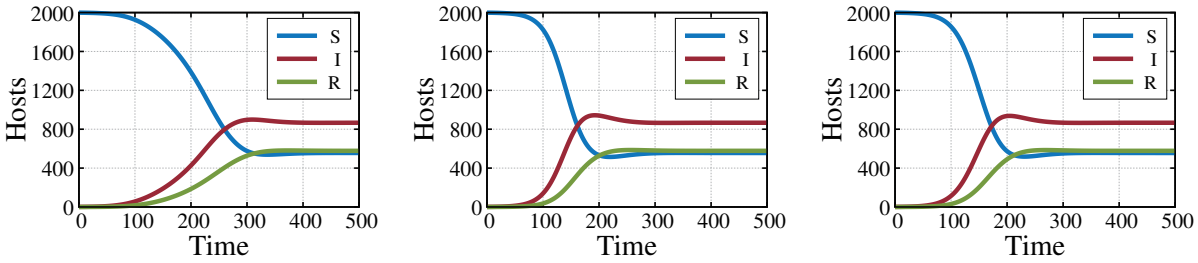
In this paper, we conduct numerical calculations to discuss the effectiveness of our proposed approximate model. In the numerical calculations, we use DELL Precision 5820 with an Intel Xeon Sliver 5110 2.10 GHz CPU and 16 GB memory, running Ubuntu 20.04. In order to obtain solutions of our proposed approximate model, we calculate the ordinary differential equations (4)-(6) with the use of MATLAB [1]. We also calculate the ordinary differential equations (1)-(3) based on the epidemic model proposed in [6] for the sake of comparison (named ODE-all in the result figures).

We consider scenarios where a target area consists of 100×100 cells, i.e., $G = 100$, and the target area is divided into 10×10 and 20×20 compartments, i.e., $d = 10$ and 20. The number K of mobile hosts in the target area is set to $2,000$ and $3,000$. We assume that there exist one infected host and $K-1$ susceptible hosts at the initial state. Specifically, we set the initial value as $S(0) = K-1$, $I(0) = 1$, and $R(0) = 0$. We place each susceptible host on randomly selected cells in the area. We set the value of the system parameters as follows: $\alpha = 0.04$, $\beta = 0.02$, $\gamma = 0.03$, and $\mu = 1$.

B. Results

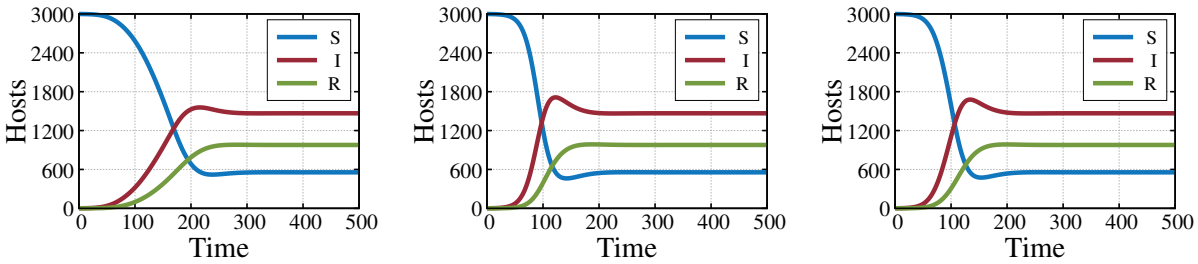
In Fig. 7, we show the change in the number of hosts in each state (i.e., $S(t)$, $I(t)$, and $R(t)$) against elapsed time t in the case of $K = 2,000$. Fig. 7(a) is the result of (1)-(3) based on the epidemic model proposed in [6]. Figs. 7(b) and (c) are the results obtained by calculating (4)-(6) in our proposed epidemic model, where $d = 10$ and 20, respectively.

From these figures, we observe that the number $S(t)$ of susceptible hosts decreases with time elapsed. Accordingly, the number $I(t)$ of infected hosts increases with time elapsed. The number $R(t)$ of recovered hosts also increases with time



(a) ODE-all. (b) $d = 10$. (c) $d = 20$.

Fig. 7: Number of hosts against time elapsed in the case of $K = 2,000$.



(a) ODE-all. (b) $d = 10$. (c) $d = 20$.

Fig. 8: Number of hosts against time elapsed in the case of $K = 3,000$.

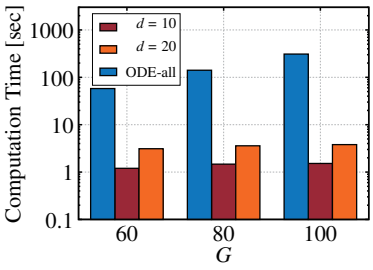


Fig. 9: Computation time.

elapsed. Finally, those values become stable at a certain level. We also observe that although the the proposed approximate model is more sensitive to the time elapsed, the proposed approximate model has similar characteristic to the ODE-all, regardless of the value of d. Specifically, they are stable at almost the same level.

In Fig. 8, we show the change in the number of hosts in each state against elapsed time t in the case of $K = 3,000$. As we can see from this figure, the result of our proposed model properly approximates that of ODE-all, similar to the result in the case of $K = 2,000$.

Finally, we measure the computation time of ODE-all and our proposed approximate model. Fig. 9 shows the computation time against the value of G, i.e., the size of the target area. As we can see from this figure, the computation time of our proposed approximate model is much smaller than that of ODE-all, regardless of the values of d and G. This is because the proposed approximate model reduces the number of ordinary differential equations to represent infection dynamics.

VI. Conclusion

In this paper, we proposed an approximate model representing the behavior of malware diffusion on wireless mobile devices, using ordinary differential equations. We showed that the results of our proposed epidemic model properly approximate those obtained from a conventional epidemic model with low computation time through numerical calculations. As future work, we will consider heterogeneous situations where mobile devices have different communication ranges and mobility characteristics (e.g., moving speed and patterns).

Acknowledgement This research was partially supported by JSPS KAKENHI Grant No. 21K11857 and 23K11077.

References

[1] MATLAB, https://jp.mathworks.com/products/matlab.html, accessed Jun. 15, 2021.

[2] N. Atanassov and M. M. Chowdhury, "Mobile Device Threat: Malware," *2021 IEEE International Conference on Electro Information Technology (EIT)*, USA, pp. 007-013, 2021.

[3] G. Gonzalez, M. E. Larraga, L. Alvarez-Icaza, and J. Gomez, "Bluetooth worm propagation in smartphones: modeling and analyzing spatio-temporal dynamics," *IEEE Access*, vol. 9, pp. 75265–75282, 2021.

[4] J. Jang-Jaccard, S. Nepal, "A survey of emerging threats in cyber security," *J. Comput. Syst. Sci.*, vol. 80, no. 5, pp. 973–993, 2014.

[5] A. M. Lonzetta, P. Cope, J. Campbell, B. J. Mohd, and T. Hayajneh, "Security vulnerabilities in Bluetooth technology as used in IoT," *J. Sens. Actuator Netw.*, vol. 7, no. 3, 2018.

[6] H. Miura, S. Abukawa, T. Kimura, and K. Hirata, "Modeling of malware diffusion with mobile devices in intermittently connected networks," *in Proc. APSIPA ASC 2022*, Chang Mai, Thailand, Nov. 2022.

Impacts of 5G-TDD Time Slot Configurations on the Downlink and Uplink Data Rates

Wen-Ping Lai*, Wen-Ru Chen, Hong-Lun Lai and Hong-Yi Li
Department of Electrical Engineering, Yuan Ze University, Taoyuan City, Taiwan
*Email: wpl@saturn.yzu.edu.tw

***Abstract*— Diverse and differential service types of 5G trigger the needs of novel technologies in radio resource allocation and management, and it is a challenging problem to enhance flexibility of radio resource allocation while improving radio spectrum efficiency. To tackle this problem, *numerology* and *time division duplex* (TDD) have been introduced accordingly. This paper presents analyses of asymmetric performance impacts of flexible 5G-TDD time slot configurations on the *downlink* (DL) and *uplink* (UL) data rates, using our lab system design and realization based on the hardware and software reference designs of the *OpenAirInterface* (OAI) 5G *stand-alone* (SA) system. For 5G *new radio* (5G-NR) at *frequency range 1* (FR1), aka sub-6 GHz, the adopted numerology scheme is $\mu = 1$, ensuring an enough *sub-carrier spacing* (SCS) of *30 kHz* for supporting higher mobility than 4G, in particular suitable for those bands above *3 GHz* in FR1. For this study, the adopted band is *n78* (3.3 - 3.8 GHz) due to its dominance among the 5G operators. In addition, six scenarios of TDD time slot configurations are studied for their impacts on the DL and UL data rates, in terms of TCP goodputs. Our measurement results show two effects: (1) asymmetric trend slopes of DL and UL data rates, and (2) symmetric performance degradation at the DL/UL extremes. Analyses of these two effects are proposed with respect to DL/UL traffic orientations, including DL-Heavy, UL-Heavy and Balanced applications, such as *internet of humans* (IoH), *internet of things* (IoT) and video conferencing respectively.**

Keywords— 5G-SA, 5G-NR, Numerologies, TDD Slot Configurations, Downlink and Uplink Performance Asymmetry

I. INTRODUCTION

The *fifth generation* (5G) communication technology [1, 2] has revolutionized the telecommunication industry, providing faster data rates, lower latency, and enhanced connectivity. Such diverse and differential service types trigger the needs of novel technologies in radio resource allocation and management [3]. Considering the sky-high cost of radio spectrum licensing, how to improve radio spectrum efficiency while enhancing flexibility of radio resource allocation is a challenging problem. To tackle this problem, both *numerology* [4, 5] and *time division duplex* (TDD) [6] have been introduced. These two technologies work together to optimize radio resource utilization and meet the diverse requirements of differential services.

The 5G numerology defines the fundamental parameter sets which characterize the time-frequency structure of the 5G radio resources. It includes flexible parameter setting such as *sub carrier spacing* (SCS) in the frequency domain, and *symbol duration*, *slot duration* and *frame structure* in the time domain. Such flexible parameter setting in different numerology schemes allow for *adaptive granularity* and *low-latency timing* of data communications in the 5G system.

On the other hand, the 5G TDD allows for asymmetric and flexible *slot-level* or even *symbol-level* configurations of radio resource allocation between the *downlink* (DL) and *uplink* (UL), where more radio resource should be allocated if the traffic is DL-heavy, and vice versa if UL-heavy. Not only *flexible adjustment*, but also *real-time adaptation* to dynamic DL/UL traffic demands can be expected to be realized in the 5G system. Certainly, the realization is challenging since it involves a good understanding of the flexible and yet complicated TDD configurations, including both the DL/UL *transmission periodicity* and *slot configuration mapping patterns*. Moreover, *synchronization* between the *5G base station* (gNB) and *5G new-radio user equipment* (nrUE) is another technical challenge [7-9]. In addition, an advanced TDD concept called *sub-band full-duplex* (SBFD) is being studied by 3GPP R.18, aka 5G-Advanced [10-13].

In fact, the 5G numerology and the 5G TDD configurations are closely intertwined, working together to enable efficient resource allocation, meet the diverse traffic demands, and optimize the overall performance of the 5G system. The flexibility provided by the flexible TDD configurations within the framework of the various 5G numerology schemes allows for adaptive resource management and customization to meet differential and dynamic service requirements.

From the aspect of performance impact, this paper presents a better reasoning on the flexible and yet complex TDD configurations for the major 5G-numerology scheme ($\mu = 1$), particularly designed for the *5G new radio* (5G-NR) [14] around *frequency range 1* (FR1), aka sub-6 GHz. This paper also presents analyses of performance impacts of TDD time slot configurations (in terms of various DL/UL ratio scenarios) on the DL and UL data rates, from DL-heavy scenarios to UL-heavy ones. Analyses of DL and UL data rates and their sums can thus potentially lead to proposing a promising user-demand-oriented strategy between *internet of humans* (IoH) and *internet of things* (IoT) [15] traffic demands, where the IoH is more characteristic of DL-heavy traffic while the IoT is more featured by UL-heavy traffic.

The remainder of this paper is organized as follows. Section II presents our lab system design implemented with the OAI-5G reference designs in both the software and

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

hardware, and explains the working mechanisms of the adopted 5G TDD time slot configurations, together with the major numerology scheme of FR1 as aforementioned, aiming at carrying out the performance impacts of flexible TDD slot configurations on the DL and UL data rates. Section III presents the experimental results and the analyses of the aforementioned performance impacts. Section IV concludes and outlooks this paper.

II. System Design and Mechanisms

A. 5G-NR Based Lab System Design and Realization

OpenAirInterface (OAI) [16] provides the first open-source reference designs of 4G/LTE and 5G-NR, including both the UE end and the network end. The UE end could be implemented with a generic PC equipped with a *software defined radio* (SDR) unit or a commercial cellular phone. The network end consists of the base station (eNB for 4G or gNB for 5G) and the core network (EPC for 4G or 5GC for 5G). For phase I of 5G, the gNB can co-exist with the 4G network end, namely the eNB and the EPC, to form the *non-stand-alone* (NSA) scheme, aka EN-DC. As for phase II of 5G, the gNB can directly work with the 5GC in the *standalone* (SA) scheme, realized with the novel design concepts such as *software defined networking* (SDN) [17, 18] and *network functions virtualization* (NFV) [19, 20], where SDN improves the network management by separating its control- and user-planes in terms of south-bound interface whereas NFV automates and enriches service deployments of 5GC by decoupling software and hardware via the virtualization technology. The management plane of the 5GC is compliant to the 3GPP *service based architecture* (SBA) [21, 22] so as to be capable of deploying rich *micro services* in terms of *virtualized* or *containerized network functions*, aka VNFs or CNFs, which can interact with each other via a single common message bus or interface to greatly simplify the design of communication protocols among multiple component functions or novel future unknown add-on services, aka *service based interface* (SBI) [23].

Our 5G-NR lab design and realization follows the OAI 5G-NR reference design for the 5G SA scheme. Fig. 1 demonstrates the system architecture of our 5G-NR lab system, which consists of one OAI nrUE, one OAI gNB and one OAI 5GC. Each of these three entities is realized and hosted on an independent generic x86 *personal computer* (PC), and the details of hardware and operating system are summarized in Table 1. The adoption of i7-based CPUs follows the hardware recommendation of OAI. Considering the need of enough computing power for real-time signal processing between the nrUE and the gNB, the PCs with a CPU of 8 physical cores are provisioned, running with a real-time Linux kernel. On the other hand, a PC with a CPU of 4 physical cores, running with a generic non-real-time Linux kernel, can turn into 8 logic cores with the help of the Intel *virtualization technology for execution* (VTx), aka *hyper-threading* of CPU (i.e., each *physical* core can be hyper-threaded into 2 *logic* cores, with the benefit of improved computing power typically by around 140% increase). In other words, such a hardware virtualization technology can thus enable software virtualization for the micro-services based management-plane design (NFV-based) on top of the software-defined separation of *control* and *user* planes (SDN-based).

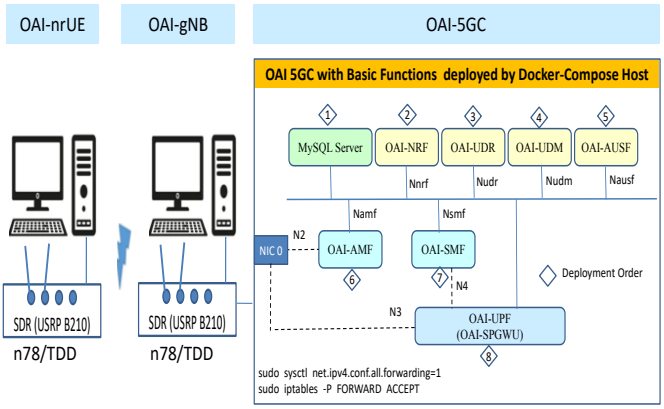


Fig. 1 System architecture of our 5G-NR lab design.

Table I: Configuration of Hardware and Software

Hardware & Software	urUE	gNB	5GC
CPU Type	i7-10700	i7-10700	i7-6700K
CPU Cores	8 (physical)	8 (physical)	8 (logic)
Memory Size (GB)	16	16	32
RF Unit	USRP B210	USRP B210	
Max. Frequency Bandwidth	40 MHz	40 MHz	
Tx/Rx switching for TDD	supported	supported	
Antenna Type for 5G FR1	*n78*	*n78*	
Realtime of Linux Kernel	low latency	low latency	generic
Version of Linux Kernel	5.15.0-67	5.15.0-67	5.15.0-71
Version of Ubuntu Linux	20.04.6	20.04.6	20.04.5

B. 5G-NR Channel Bandwidth and Resource Blocks

5G-NR is expected to support up to 100 MHz channel bandwidth for lower bands less than 6 GHz, aka sub-6 GHz or FR1 as aforementioned, and 400 MHz channel bandwidth for higher bands in the mm-Wave ranges. This paper focuses on how to calculate the number of *resource blocks* (RBs) in FR1. As given in Table II, 3GPP TS 38.101-1 [24] specifies the *maximum number of RBs* (N_{RBs}^{max}), available in different numerology configurations, such as $\mu = 0, 1, 2$, where the corresponding *sub carrier spacing* (SCS), *supported channel bandwidth* ($\Delta f_{s.ch}$) with a guard band (Δf_{gb}) on each side are also specified so as to deliver the corresponding *max number of RBs* (N_{RBs}^{max}). Note that the SCS is a function of μ, denoted as $SCS(\mu)$ and defined by Eq. (1), and the corresponding N_{RBs}^{max} can thus be carried out based on Eq. (2).

Table II: N_{RBs}^{max} **as a function of the numerology scheme** μ.

μ	$SCS(\mu)$	$\Delta f_{s.ch}$	Δf_{gb}	N_{RBs}^{max}
0	15 KHz	50 MHz	692.5 kHz	270
1	30 KHz	100 MHz	845.0 kHz	273
2	60 KHz	100 MHz	1370.0 kHz	135

$$SCS(\mu) = 2^{\mu} \cdot 15 \text{ kHz} \quad (1)$$

$$N_{RBs}^{max} = (\Delta f_{s.ch} - 2\Delta f_{gb}) / 12 \cdot SCS(\mu) \quad (2)$$

The main focus of this paper is to study the effect of TDD on the DL/UL throughput performances. In particular, only the major numerology scheme $\mu = 1$ for above 3 GHz of FR1 is considered, particularly designed for higher mobility than 4G-LTE. Technically speaking, however, one still has to take into consideration the factor of $\Delta f_{s.ch}$. In other words, $\Delta f_{s.ch}$ could potentially vary in reality in two cases: (1) the supported channel bandwidth of the actually licensed band owned by an operator could be less and different depending on the price paid, and (2) the supported channel bandwidth of the adopted RF unit. For this lab-based study, case (2) is the main concern since the adopted RF unit is USRP B210 and its $\Delta f_{s.ch}$ is limited to 40 MHz. Considering the $\mu = 1$ numerology scheme, Table II together with Eqs. (1) and (2) can thus work together to deliver the value of N_{RBs}^{max} to be 106.

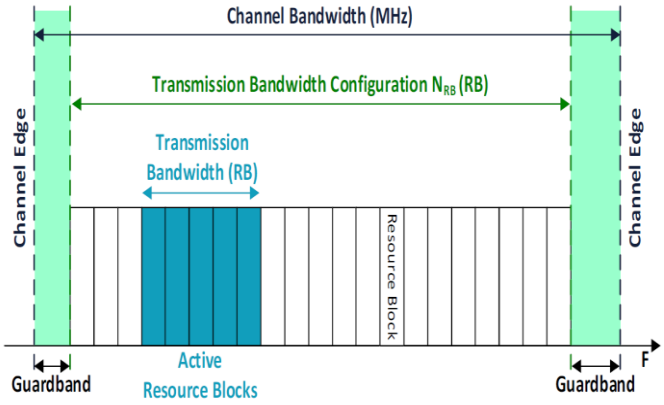


Fig. 2 5G-NR frame structure and time slot durations as a function of the numerology scheme (μ = 0, 1, 2, 3, 4) [2]

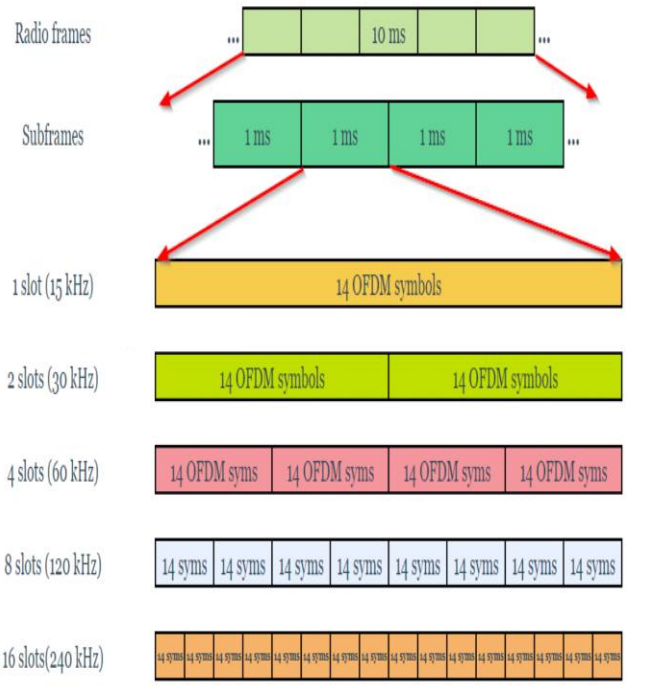


Fig. 3 5G-NR frame structure and time slot durations as a function of the numerology scheme (μ = 0, 1, 2, 3, 4) [5]

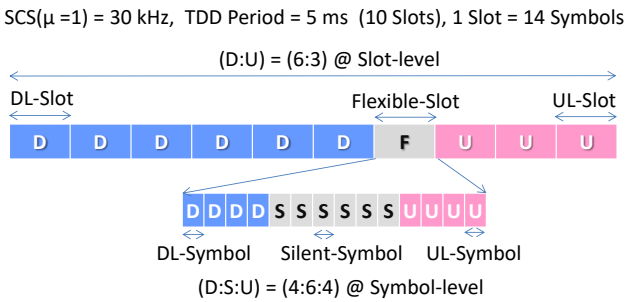


Fig. 4 A typical DL/UL time slot configuration of 5G-NR TDD for DL-Heavy applications, in the major numerology scheme (μ = 1) for RF1.

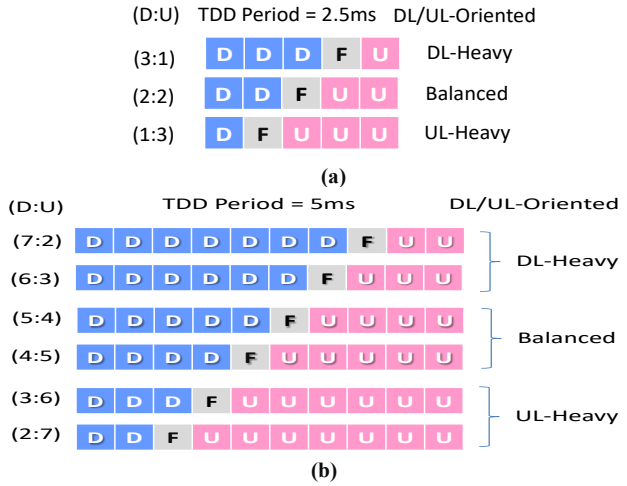


Fig. 5 5G-NR TDD slot configurations for different TDD periods and DL/UL orientations: (a) period = 2.5 ms, and (b) period = 5 ms.

C. Flexible Numerology Schemes for 5G-NR

As aforementioned in Section I, the 5G numerology defines the fundamental parameter sets which characterize the time-frequency structure of the 5G radio resources. Fig. 3 depicts the 5G-NR frame structure and the variation of time slot duration with the numerology scheme (μ = 0, 1, 2, 3, 4) [5], where it is clearly seen that the time slot duration decreases linearly as the numerology scheme index μ increases, meaning that the SCS(μ) also increases linearly, based on Eq. (1).

Compared to the concept of *fixed* SCS (15 kHz) for 4G-LTE, the concept of *flexible* SCS is new to 5G-NR. The most fundamental motivation behind this new concept is to support *higher mobility* for 5G-NR, where the Doppler frequency drift becomes larger with the higher speed demand of the UE, and also linearly with the higher frequency band adopted. To compensate the Doppler Effect, a wider SCS is thus needed. For frequency bands in between 3GHz and 6GHz, the μ = 1 scheme adopts a SCS of 30 kHz, namely 2 times the fixed SCS (i.e., 15 kHz) designed for those bands below 3 GHz, is already enough and thus becomes the major scheme for 5G FR1. As for mm-Waves, aka 5G FR2, those schemes with $\mu \geq 2$ will serve for the needs. Another benefit coming with a wider SCS is a shorter time duration for both time slots and symbols, which can thus generate a lower latency for time-

sensitive applications. However, the low-latency issues are out the scope of this paper. In this paper, the main focus is on the performance impact of TDD onto DL/UL data rates for the *n78* band (3.3 – 3.8 GHz), and the adopted numerology scheme is thus μ = 1, which will be the default scheme throughout the rest of this paper.

D. Flexible TDD DL/UL Slot Configurations for 5G-NR

Both 5G FDD and TDD have their advantages and are suitable for different scenarios. FDD provides good coverage, better signal propagation, and is suitable for a wide range of environments, including rural areas. On the other hand, TDD offers flexibility, efficient spectrum usage, and lower latency, making it ideal for dense urban environments with asymmetric traffic demands. Recall that, from Section 1, the 5G TDD allows for an asymmetric configuration of radio resource allocation between the DL and UL, where more radio resource should be allocated in the DL if the traffic is DL-heavy, and vice versa if UL-heavy.

Fig. 4 illustrates a typical DL/UL time slot configuration of 5G-NR TDD for DL-Heavy applications, in the major numerology scheme (μ = 1) for the *n78* band of FR1, where the ratio of DL-Slots to UL-Slots is denoted as (D:U), is (6:3). Note that one *flexible slot* (F-Slot) is needed for a special purpose: it contains 4 DL-Symbols and 4 UL-Symbols, separated by 6 Silent-Symbols as the *guard period* when switching from the DL to the UL. The physics for this guard period has been well explained in [25].

Fig. 5 demonstrates more 5G-NR TDD slot configuration scenarios for different TDD periods and DL/UL orientations, where (a) is for period = 2.5 ms, and (b) for period = 5 ms. In the scenario of (a), three (D:U) slot configurations are depicted and classified into three different DL/UL-orientations, namely (3:1) for DL-Heavy, (2:2) for Balanced, and (1:3) for UL-Heavy. The scenario of (b) presents more configurations, i.e., six (D:U) slot configurations classified into three different DL/UL-orientations: (7:2) and (6:3) for DL-heavy, (5:4) and (4:5) for Balanced, and eventually (3:6) and (2:7) for UL-Heavy.

In this study, periodicity (b) is adopted for its rich configurations so as to carry out performance impacts of 5G TDD slot configuration scenarios on the DL and UL data rates for different traffic-demand orientations.

III. Results and Analyses

A. Measurement Setup for TDD DL/UL Data Rates

Recall that Fig. 1 describes the system architecture of our 5G-NR lab design, and Table 1 summarizes the configuration of the associated hardware and software. In addition, Section 2.1 also details of such a lab design and its realization.

For this study, it is interesting to understand the aforementioned performance impacts on the achievable DL and UL data rates, together with their summed rates, due to the flexible time slot configurations of 5G TDD. As aforementioned and demonstrated by Fig. 5(b), the six slot configurations, classified into three DL/UL user-demanded traffic orientations, were adopted in the study.

In order to measure the achievable DL and UL data rates, the *iperf* toolset was utilized. The single software of *iperf* can play two roles, namely the traffic generator (client) or traffic receiver (server) when respectively followed with a command-line option of -c or -s. Both TCP and UDP traffic types are supported in iperf.

In this study, due to TCP's flow control, the TCP traffic type is more suitable and faster to explore the achievable DL and UL data rates than UDP. For the DL, the *iperf client* should be setup on the 5G network side, and the *iperf server* on the 5G UE side; and vice versa for the UL. In the telecom network, the achievable DL and UL data rates are the concepts based on *overlay networking* (a logical network over a physical network), realized by the GTP Tunnel. Hence, the 5G network side is referred to as the 5G-CN. Considering the fact that our 5G-CN is fully virtualized in terms of its component functions, it takes some tricks to locate and enter the container-based OAI-UPF (i.e., OAI-SPGWU), and eventually identify the GTP Tunnel End as the virtual network interface that either the *iperf server* or *client* has to bind. In addition, note that the measured data rates should be recognized as the TCP Goodputs, instead of the MAC Throughputs. In other words, the TCP Goodputs are the achievable maximum data rates at the layer of TCP, somewhat smaller than the MAC Throughputs.

B. Performance Impacts of TDD Slot Configurations

Fig. 6 summarizes the performance impacts of 5G TDD slot configuration scenarios on the DL and UL data rates (in terms of TCP Goodputs) for different traffic-demand orientations, as depicted in Fig. 5(b) and described at the end of the previous section. Fig. 6 delivers two distinguished effects as follows.

- Asymmetric trend slopes of DL and UL data rates

Generally speaking, one should expect that the DL data rate increases with the number of allocated DL-slots, and vice versa for the UL. In fact, the measured results show that such an expectation is somewhat naive: the expected increasing trends do exist, but their trend slopes are quite different. In other words, the much larger DL-trend slope supports that more DL-slots do help increasing the DL data rate substantially. However, such a substantial help is not seen in the UL due to the much smaller UL-trend slope. However, this asymmetry is not totally surprising and can be well explained by the huge difference of transmission power capabilities between the gNB (along the DL) and the UE (along the UL), even given the same channel condition. Obviously, the DL is much stronger, and thus much more sensitive and effective in its data rate enhancement given more DL-slots. On the other hand, although one more UL-slot does help the UL data rate effectively from D7U2 to D6U3, but not so much given any more. Eventually, this asymmetry effect also leads to asymmetric total data rates, i.e., the summed rates of both the links. Again, it seems that the small enhancement of the UL data rate is at the cost of the large degradation of the DL data rate. However, whether or not such a cost is worthwhile really depends on application scenarios.

- Symmetric performance degradation at DL/UL extremes

Another visible and interesting phenomenon happens, more or less symmetrically, at both the DL and UL extremes, namely D7U2 and D2U7. In other words, symmetric performance degradation is clearly seen at both the two extremes. However, it violates our basic intuitions. The question is why the performance on either link degrades when given more TDD slots. Presumably, it may come from TCP itself. If true, the most promising guess is due to the required ACK packets in the returned direction, for acknowledgements of the sent packets. In other words, if some ACK packets are late enough or simply lost, they will generate resent packets, and equivalently degrade the achievable data rate. Actually, it is quite possible for such an effect to occur at both the extremes, where the link direction of sent packets are allocated with too many slots so that too few slots are left for the ACKs, which will thus face much longer latencies or even congestions. Hence, for both the DL-Heavy and UL-Heavy applications over TCP, such a performance degradation effect should be skipped by avoiding going into both the extremes. However, it should not be the case for UDP where no ACK packets are required. For an application which is interaction based (such as video conferencing), Balanced configurations seem to be a good strategy no matter it is over TCP or UDP.

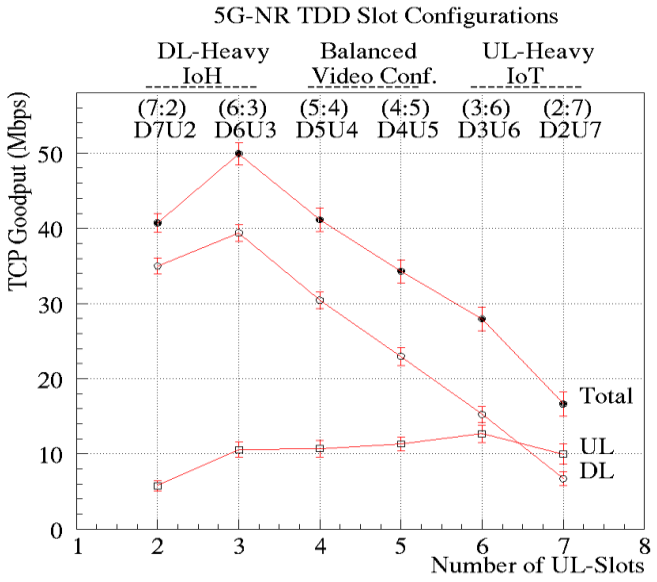


Fig. 6 Asymmetry effect of TDD DL/UL data rates under different 5G-NR TDD slot configurations (period = 5 ms).

IV. Conclusion and Outlook

In this paper, a 5G-NR based lab system design together with its realization has been demonstrated. The main focus has been placed on the performance impacts of 5G-TDD time slot configurations on the DL and UL data rates for the major band of 5G-NR, namely *n78* (3.3 – 3.8 GHz), suitable for the adoption of the $\mu = 1$ numerology (with a SCS of 30 kHz, 2 times larger than 4G-LTE) for higher mobility. The performance measurement results show two major effects: (1) asymmetric trend slopes of DL and UL data rates, and (2) symmetric performance degradation at the DL/UL extremes. Analyses of these two effects have also been proposed with respect to the DL/UL traffic orientations, including DL-Heavy, UL-Heavy and Balanced applications, such as *internet of humans* (IoH), *internet of things* (IoT) and video conferencing respectively.

As an outlook into the future, there are some works worth doing. First of all, the performance effects and their analyses carried out in this study are mostly based on the view point of TCP-Goodput based data rates. It is certainly interesting to extend the work to cover the view point of UDP, where one might still expect to see the asymmetry phenomenon similar to Effect (1) given the fact that the transmission power of the gNB is always stronger than that of the UE; but one might not see Effect (2) any longer, due to no returned ACK packets in the opposite link of sent packets.

Another useful view point can also be provided by a study on packet-based *latency* and *jitter* spectra, respectively important to *interactive* and *one-way* applications. However, the *ping*-based packet's *round-trip time* (RTT) information only provides two-way knowledge. Hence, a special time synchronization technique between the CN and the UE with a time resolution much more precise than the generic NTP protocol (only precise to the level of *ms*) is certainly wanted for resolving one-way knowledge down to the level of sub-μs, either along the DL or UL.

V. Acknowledgement

This work was partly supported by Taiwan's National Science and Technology Council under grant 112-2221-E-155-010 and Taiwan's Ministry of Education under grant PEE1110344.

References

[1] 3GPP TS 23.501 v15.00, System architecture for the 5G system, Stage 2, Dec. 2017.
[2] *View on 5G Architecture*, 5GPPP White Paper v2.0, Dec. 2017.
[3] 3GPP TS 38.133 v16.5.0, Requirements for support of radio resource management, Dec. 2020.
[4] ShareTechnote, 5G/NR – Numerology / SCS (Sub Carrier Spacing), [Online]. Available: https://www.sharetechnote.com/html/5G/5G_Phy_Numerology.html (Accessed: Apr. 1, 2023)
[5] GaussianWaves, 5G NR Resource Block [Online]. Available: https://www.gaussianwaves.com/2022/02/5g-nr-resource-block (Accessed: Apr. 2, 2023)
[6] *5G TDD Uplink*, NGMN White Paper v1.0, Jan. 2022.
[7] 5G TDD Synchronisation Guidelines and Recommendations for the Coexistence of TDD Networks in the 3.5 GHz Range, GSMA White Paper, Apr. 2020.
[8] ShareTechnote, 5G/NR – SS Block, [Online]. Available: https://www.sharetechnote.com/html/5G/5G_SS_Block.html (Accessed: Jun. 1, 2023)
[9] Matlab, 5G Technology Explained: Synchronization Signal Blocks (SSB) in the 5G New Radio (NR), Jan. 2020, [Online]. Available: https://www.youtube.com/watch?v=9Y8ajltjiII (Accessed: May 30, 2023)
[10] 3GPP R.18. [Online]. Available: https://www.3gpp.org/specifications-technologies/releases/release-18 (Accessed: Apr. 5, 2023)

[11] Becoming 5G-Advanced: the 3GPP 2025 Roadmap, 5G America White Paper, Dec. 2022.
[12] 5G-Advanced explained, Nokia. [Online]. Available: https://www.nokia.com/about-us/newsroom/articles/5g-advanced-explained/ (Accessed: Apr. 7, 2023)
[13] 5G Advanced: Evolution towards 6G, Ericsson. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/5g-advanced-evolution-towards-6g (Accessed: Apr. 8, 2023)
[14] 3GPP TS 36.300 v17.4.0, Overall description, Stage 2 (R.17), Mar. 2023.
[15] 3GPP TS 22.368 v17.0.0, Service requirements for Machine-Type Communications (MTC), Stage 1 (R.17), Dec. 2022.
[16] N. Nikaein, M. K. Marina, S. Manickam et al., "OpenAirInterface: a flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33-38, 2014.
[17] X. Foukas, N. Nikaein, M. M. Kassem et al., "FlexRAN: A flexible and programmable platform for software-defined radio access networks," *Proc. 12th International Conference on Emerging Networking Experiments and Technologies* (CoNEXT '16), Irvine, CA, USA, Dec. 2016, pp. 427–441.
[18] R Schmidt, C.Y. Chang and N. Nikaein, "FlexVRAN: A flexible controller for virtualized RAN Over Heterogeneous Deployments," *Proc. 2019 IEEE International Conference on Communications* (ICC'19)*,* Shanghai, China, 2019.
[19] *Network Functions Virtualization* (*NFV*) *Release 2 on Testing;* , *Guidelines on Interoperability Testing for MANO,* ETSI GR NFV-TST 007 v2.6.1, Jan, 2020.
[20] ETSI-NFV, Network Functions Virtualization. [Online]. Available: https://www.etsi.org/technologies/nfv (Accessed: July 14, 2021)
[21] 3GPP, 5G System Overview [Online]. Available: https://www.3gpp.org/technologies/5g-system-overview (Accessed: May 3, 2023)
[22] 3GPP TS 32.240 v17.4.0, Overall description, Stage 2 (R.17), Mar. 2023.
[23] 3GPP TS 23.501 v18.1.0, System architecture for the 5G System (5GS), Stage 2 (R.18), Mar. 2023.
[24] 3GPP TS 38.101-1 v17.7.0, User Equipment (UE) radio transmission and reception, Part 1: Range 1 Standalone (R.17) Sep. 2022.
[25] Techtrained, TDD: Why a Guard Period only in DL-to-UL Switching? [Online]. Available: https://www.. gaussianwaves.com/2022/02/5g-nr-resource-block (Accessed: Jun. 21, 2023)

Bearing Fault Diagnosis and Interpretation Based on 2D Images and Convolutional Neural Network

Zhenzhen Tian[1,†], Xinyu Zhang[1,‡], Wei Yan[1,§] and Jihua Wang[1,*]
[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250399, China
E-mail: †17860548251@163.com, ‡xinyuzhang20210228@163.com
Corresponding author E-mail: §wyaninsa@gmail.com, *jihuaaw@126.com

***Abstract* — In intelligent manufacturing, fault diagnosis of rotating machinery is indeed a critical and complex task. Most of the current research only classifies faults without considering the interpretability of classification. To fill this gap, inspired by the interpretable technology of computer vision, we propose 2DImages-CNN, a method combining 2D images and a convolutional neural network (CNN) for mechanical fault diagnosis and interpretation. First, we transform time series signals into distinct 2D images through two techniques: the Gramian angular field (GAF) and the Wavelet transform (WT). Then, we input 2D images into the ResNet34 model for fault recognition. We conducted experiments using transfer learning on the Case Western Reserve University (CWRU) bearing dataset. The results provide compelling evidence that, when compared to renowned methods, the 2DImages-CNN method achieves superior fault recognition accuracy. Finally, we use gradient-weighted class activation mapping (Grad-CAM) to interpret deep fault features. The results show that, for GAF images, the network shows a strong degree of activation in the area with large signal fluctuations. For WT images, the network shows a strong degree of activation in the area with large amplitude.**

***Keywords* —Convolutional Neural Network (CNN), Fault Diagnosis, Gramian Angular Field (GAF), Wavelet Transform (WT), Interpretation.**

I. INTRODUCTION

Bearings hold an indispensable role within the field of rotating machinery, serving as a critical component. They account for a substantial proportion of failures, estimated at 45% to 55%. Therefore, accurate and timely diagnosis of bearing issues is significant in improving the reliability of this machinery [1]-[2].

As intelligent technology continues to advance, various data-driven techniques, particularly those centered on deep learning, have gained extensive traction. This popularity is attributed to their robust data processing capabilities and remarkable learning prowess. CNNs automatically extract relevant features according to learning objectives and do not rely on expert experience. Therefore, CNNs have found broad utility in fault diagnosis. Zhang et al. [3] introduced a CNN model that utilizes time-frequency images generated through short-time Fourier transform (STFT) as input for fault diagnosis. In reference [4], the authors used the WPT method to generate spectral images of vibration signals and then used CNN for fault identification. Meanwhile, this paper employs simulated degradation algorithms to determine the optimal parameters. In reference [5], the authors developed a CNN-based model for bearing fault diagnosis that incorporates time-frequency image analysis. They investigated the model's performance with respect to three distinct time-frequency analysis techniques. Liang et al. [6] introduced the WT-MLCNN method. Initially, they applied the WT to transform vibration signals into 2D time-frequency images. Following this transformation, the subsequent step involved feeding these images into the MLCNN model, thereby enabling comprehensive fault diagnosis for gearboxes. Liang et al. [7] additionally presented the WT-GAN-CNN approach, where they applied the WT to convert the signals into time-frequency images. Following this, they employed GAN to generate additional training image samples, which were then used in conjunction with a CNN model for fault detection. The above researchers have combined 2D time-frequency images with CNNs for fault diagnosis. Time-frequency images describe the overall characteristics of signals. In reference [8], the authors used the GAF and MTF models to encode time series as images, thereby proving that the CNN could obtain excellent performance. Han et al. [9] introduced a fault diagnosis approach that incorporates the GAF, MTF, and capsule network. GAF and MTF images were input into the capsule network as two-channel images so that the network could obtain more complete fault characteristics. GAF images

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

describe the instantaneous characteristics of signals.

The above research indicates that inspired by computer vision, fault diagnosis methods using a combination of 2D images and CNNs are widely used. However, the above research only classifies faults without considering the interpretability of neural networks. In recent times, there has been a significant surge in interest surrounding the interpretability of deep learning methods. Researchers and practitioners alike have devoted substantial attention to unraveling the inner workings of deep learning models. The interpretable methods of neural networks include CAM [10], Grad-CAM [11], and Grad-CAM++ [12]. However, the disadvantage of CAM is that the original model structure needs to be modified and retrained. The Grad-CAM++ model is not clean enough visually, and a lot of background information is marked. Therefore, this paper uses the Grad-CAM method for visualization, which can establish the relationship between the key activation areas of the neural network and target categories, and solve the problem of black-box operation of the neural network.

This paper's contributions include:

(1) First, the GAF and WT methods encode time series signals into two types of 2D images. GAF and WT images describe the instantaneous and overall features of signals, respectively. This paper analyzes signals from different perspectives.

(2) Second, GAF and WT images are used as inputs to the Resnet34 model for fault detection. The experimental results demonstrate that the 2DImages-CNN method outperforms many existing methods.

(3) Finally, the black-box problem of fault diagnosis is explained through Grad-CAM visualization. The experimental results show that for GAF images, the network makes classification decisions in areas with large fluctuations in signal values; For WT images, the network makes classification decisions in areas with large amplitudes.

Section II of this paper presents the fundamental theories behind the GAF, WT, and 2DImages-CNN methods. Section III validates the effectiveness of 2DImages-CNN for fault diagnosis through experiments. Finally, Section IV offers the conclusion.

II. METHODS

A. Gramian Angular Field

The GAF [13] method can convert temporal signals into images with rich features, which contain the temporal correlation between any two data points. GAF coding steps are as follows:

(1) First, the time series $X = \{x_1, x_2, \ldots, x_n\}$ in Cartesian product coordinates are scaled to [0,1] by Equation (1).

$$\bar{x}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_i is the ith value in the original time series; $\bar{x}_i$ is the ith value in the reconstructed time series; and x_{min}, x_{max} are the minimum and maximum values of X, respectively.

(2) Then, the scaled value is encoded as the angular cosine $\emptyset$, and the time stamp is encoded as the radius r. We can represent the rescaled time series in the polar coordinate system. Equation (2) is as follows:

$$\begin{cases} \emptyset = \arccos(\bar{x}_i);\ 0 \le \bar{x}_i \le 1, \bar{x}_i \in \widetilde{X} \\ r = \frac{t_i}{N}; t_i \in N \end{cases} \tag{2}$$

where t_i is the time stamp, and N is the constant factor.

(3) Finally, we can leverage the angular perspective by assessing the trigonometric distinctions between each data point to discern time correlations within various time intervals. The GAF coding method can encode time series into feature images, which is defined as Equation (3).

$$G_{GAF} = \begin{bmatrix} \sin(\emptyset_1 - \emptyset_1) & \cdots & \sin(\emptyset_1 - \emptyset_n) \\ \sin(\emptyset_2 - \emptyset_1) & \cdots & \sin(\emptyset_2 - \emptyset_n) \\ \vdots & \ddots & \vdots \\ \sin(\emptyset_n - \emptyset_1) & \cdots & \sin(\emptyset_n - \emptyset_n) \end{bmatrix}$$

$$= \sqrt{I - \left(\widetilde{X}^2\right)^T} \cdot \widetilde{X} - \widetilde{X}^T \cdot \sqrt{I - \widetilde{X}^2} \tag{3}$$

where I is the unit row vector; $\tilde{X}^T$ is the transpose of X. GAF images of 10 different fault types of the CWRU dataset are shown in Fig. 1.

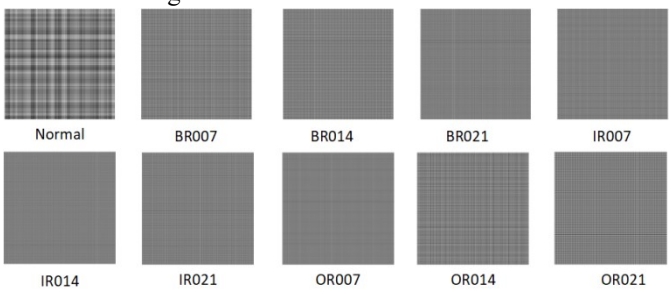


Fig. 1: GAF images of different fault types of the CWRU dataset.

B. Wavelet transform

WT is a modern approach to transform analysis. It builds upon the concept of STFT localization, addressing the limitation of fixed window size with respect to frequency. WT offers a "time-frequency" window that adapts to different frequencies and is defined by Equations (4) and (5).

$$W(a,\tau) = [x(t), \psi_{a,\tau}(t)] = \frac{1}{\sqrt{a}} \int x(t) \psi\left(\frac{t-\tau}{a}\right) dt \tag{4}$$

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right) \qquad a, \tau \in R, a > 0 \tag{5}$$

where, a is the scale factor; τ is the translation factor; $\psi_{a,\tau}(t)$ is the wavelet basis function. WT images of 10 different fault types of the CWRU dataset are shown in Fig. 2.

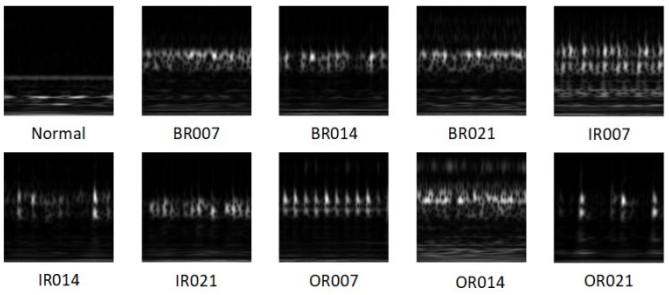


Fig. 2: WT images of different fault types of the CWRU dataset.

C. *Convolutional Neural Network*

This section provides a detailed introduction to the CNN framework used for fault diagnosis [14]. ResNet34 was chosen for the fault classification task in this paper for several reasons. First, ResNet34 contains residual structure [10], which can simplify network training and reduce network training parameters. Second, ResNet34 uses the batch normalization layer, which can avoid gradient disappearance and explosion. Finally, the CNNs are capable of extracting translation-invariant features, thereby improving robustness and mitigating the influence of noise. The specific algorithm for the CNN-based method is as follows:

Algorithm CNN-Based Fault Diagnosis Method

Input: GAF or WT images.

Step 1: Feature extraction.

The feature extractor can extract features from the original data and connect them to form a global representation.

Convolutional layer: $C^q = \sum_{p=1}^{P} W^{p,q} * X^p + b^q$

Batch normalization layer: $\widehat{c_i} = \frac{c_i - u}{\sqrt{\sigma^2 + \varepsilon}}$, $h_i = \gamma \widehat{c_i} + \beta$

ReLU activation layer: $R^r = ReLU(H^r) = max(H^r, 0)$

Max pooling layer: $M^r = \max_{T^r \in S} R^r$

Step 2: Fault classification.

Fault classifier identifies fault type according to the global representation.

Average pooling layer: $Z^r = \operatorname{avg}_{M^r \in S} M^r$

Full connected layer: $Z = [Z^1, Z^2, \ldots, Z^r, \ldots, Z^R]$, $G = \mathrm{ReLU}(W_g Z + b_g)$

Softmax function: $O_j = P(y = j|G;\theta) = \frac{exp(\theta^{(j)} G)}{\sum_{k=1}^{K} exp(\theta^{(k)} G)}$, $j = 1,2,\ldots,K$

Output: the target categories(Normal,BR007,.....,OR021.).

In the above algorithm, $q = 1,2,3,\ldots,Q$, Q represents the predefined number of filters; $p = 1,2,3,\ldots,P$, P is the number of convolution channels; $W^{p,q}$ and b^q represent the weight and bias of the filter, respectively; c is a minibatch of feature C, $c = c_1, c_2, c_3, \ldots, c_t$, σ^2 and u are the variance and mean of the minibatch, respectively; ε is the minimum value close to 0; H^r is result of the BN layer of channel r; S represents the range of the pooling function; W_g and b_g represent the weight and bias, respectively. O_j represents the estimated probability of category j; $\theta^{(j)}$ is the parameter of the softmax function; and K is the number of fault categories..

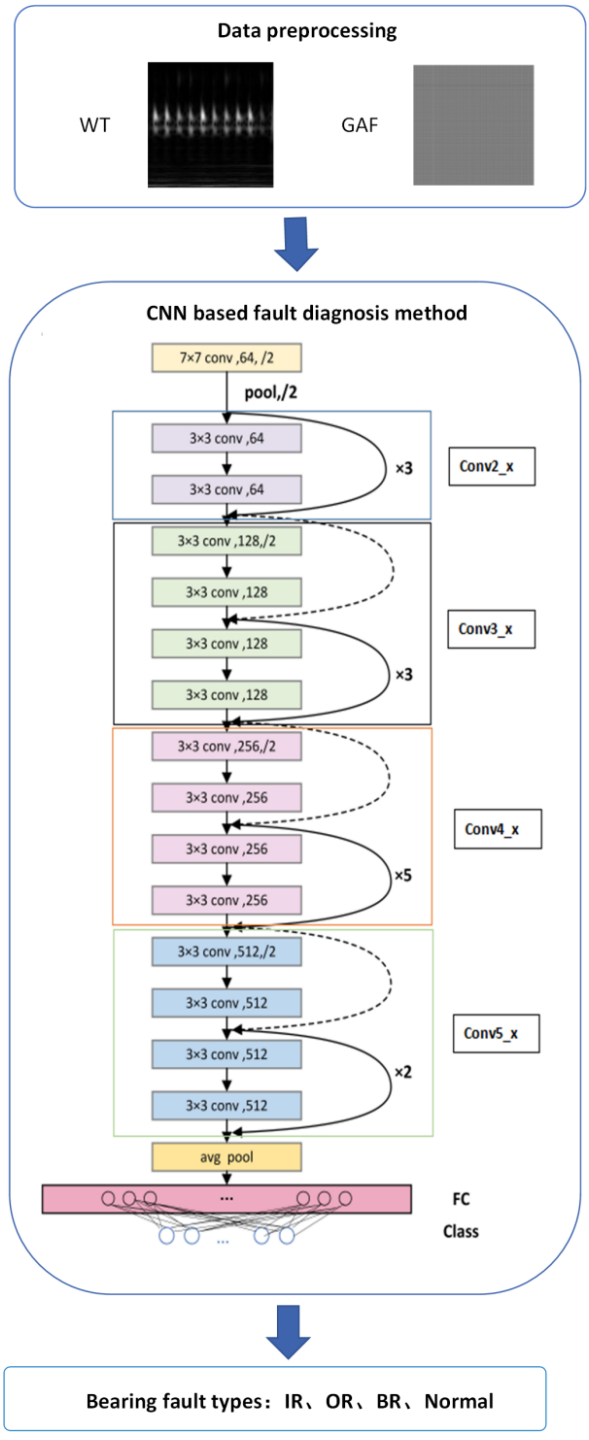


Fig. 3: 2DImages-CNN fault diagnosis method.

The 2DImages-CNN method is shown in Fig. 3. Firstly, the GAF and WT methods encode time series signals into two types of 2D images. Next, we input the 2D images into the ResNet34 model to perform feature extraction. The first layer convolution operation of the network uses 64 convolution kernels with a size of (7,7) and a moving step size of 2. After each convolution operation, the output sequentially undergoes batch normalization and activation layers. The remaining convolution layer operations are similar to those of the first layer. Each curved arrow represents a residual structure, and the number next to it represents the number of residuals stacked. We input the size of images as (224,224,1), after the operation of the first convolution layer the output feature map size becomes (112,112,64), and then through the Conv2_x, Conv3_x, Conv4_x, and Conv5_x layers, output feature map size changes in turn for (56,56,64) → (28,28,128) → (14,14,256) → (7,7,512). After a series of convolution operations, fault types are determined using successive layers including average pooling, fully connected layers, and the softmax function. Bearing failures can be categorized into four main types: inner race fault (IR), outer race fault (OR), bearing roller fault (BR), and normal (Normal).

D. Evaluation index of fault diagnosis

We further evaluated the classification results using evaluation indicators commonly used in machine learning. The commonly used evaluation indexes for multi-classification problems are Macro-precision (macro-P), Macro-recall (macro-R), and Accuracy. Evaluation indicators are defined as follows:

$$\mathrm{macro}-\mathrm{P}=\frac{1}{\mathrm{n}}\sum_{1}^{\mathrm{n}}\mathrm{P}_i\,,\ \text{where}\ \mathrm{P}_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i}\times 100\% \quad (6)$$

$$\mathrm{macro}-\mathrm{R}=\frac{1}{\mathrm{n}}\sum_{1}^{\mathrm{n}}\mathrm{R}_i,\ \text{where}\ \mathrm{R}_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i}\times 100\% \quad (7)$$

$$\mathrm{Accuracy}=\frac{\mathrm{Correct}}{\mathrm{Total}}\times 100\% \quad (8)$$

where the abbreviations TP_i, FP_i, TN_i, and FN_i represent the number of True Positives, False Positives, True Negatives, and False Negatives in the i-th category, respectively [15]. *Correct* denotes the count of samples correctly classified, while *Total* represents the overall number of samples.

E. The interpretability of fault diagnosis

For the black-box operation of the neural network, Grad-CAM is used to interpret the fault diagnosis results. Grad-CAM is a class-discriminative localization technique that assigns a score to each class based on gradient information from backpropagation and convolution activation values [11]. Grad-CAM provides a way to determine which specific parts of the image influence model decisions. Mathematically, the Grad-CAM calculation of the class discriminant positioning map is as follows:

$$\mathrm{L}^{\mathrm{c}}=\mathrm{f}\left(\sum_{\mathrm{k}}\mathrm{a}_{\mathrm{k}}^{\mathrm{c}}\mathrm{A}^{\mathrm{k}}\right) \quad (9)$$

where A^k is the k-th feature map of a CNN layer; a_k^c is the weight of A^k; and $f(\cdot)$ is a ReLU activation function.

III. Experimental Study

To evaluate the effectiveness of the 2DImages-CNN method we have proposed, we conducted a series of experiments utilizing the CWRU dataset. This dataset was chosen due to its well-established reputation for providing a robust and comprehensive basis for our evaluation process.

A. Experimental equipment and data description

Bearing failure stands out as a primary reason behind failures in rotating machinery. Therefore, the detection of bearing faults holds immense significance in preventing failures in rotating machinery. Precise bearing fault diagnosis contributes to enhancing equipment reliability and availability.

The CWRU dataset [16] includes the following:

(1) The fault points on the bearings are manually processed, and vibration acceleration signals from the faulty bearings are collected using an acceleration sensor.

(2) The vibration signals of the IR, OR, and BR with different fault diameters under different loads are collected by an accelerometer. The diameters are 0.007, 0.014, and 0.021 inches respectively. The loads are 0, 1, 2, and 3 HP respectively.

(3) We selected the bearing fault data of the drive end with a sampling frequency of 12 kHz and the normal data and used the DE-driving end acceleration data to experiment.

(4) The original dataset comprises samples with varying numbers of time course measurements. To augment the dataset and facilitate the training of a more precise model, we standardize the samples to ensure each one contains a consistent set of 1200 time course measurements.

The details of the CWRU dataset can be found in Table 1. In Table 1, Normal, IR, OR, and BR indicate the different states of the bearing. BR007 indicates a failure of a bearing roller with a diameter of 0.007 inches, IR007 indicates a failure of a bearing inner race with a diameter of 0.007 inches, while other categories follow a similar pattern.

Table 1: The details of the CWRU dataset.

Health states descriptions					Dataset
Types/Labels			Diameter (inch)	Load (HP)	Train/Test
Normal	Normal	Normal	-	0-3	1120/280
Failure	BR	BR007	0.007	0-3	320/80
		BR014	0.014	0-3	320/80
		BR021	0.021	0-3	320/80
	IR	IR007	0.007	0-3	320/80
		IR014	0.014	0-3	320/80
		IR021	0.021	0-3	320/80
	OR	OR007	0.007	0-3	960/240
		OR014	0.014	0-3	320/80
		OR021	0.021	0-3	960/240

B. Diagnosis results and comparative analysis

The 2D images generated by both GAF and WT techniques are randomly split into training and test sets, with an 8:2 ratio. The training set is used exclusively to train the ResNet34 model, while the test set is reserved for evaluating the model's performance.

For bearing fault diagnosis, three models are trained: the two-classifications model (Normal and Failure), the four-classifications model (Normal, IR, OR, BR), and the ten-classifications model (Normal, IR007, IR014, IR021, OR007, OR014, OR021, BR007, BR014, and BR021). According to different classification and prediction targets, the parameters are adjusted on the basis of the pretraining ResNet34 model. During training, we froze the parameters of Conv2_x, Conv3_x, and Conv4_x layers. The learning rate of the Adam optimization algorithm is set to 0.0001, the number of training iterations is set to 50, and the number of categories is set to 2/4/10. The experimental results of different classification models of bearing faults are shown in Table 2.

The findings presented in Table 2 demonstrate that employing both GAF and WT images in conjunction with the ResNet34 network model yields a high fault recognition rate. It's noteworthy that while the fault diagnosis rate of GAF images is slightly lower compared to WT images, this can be attributed to the more complex nature of fault information contained in GAF images. In contrast, WT images offer more intuitive fault information, resulting in superior network performance when classifying WT images. Our approach doesn't depend on prior knowledge; the measured data and their corresponding labels alone suffice for precise fault type classification.

Table 2: Experimental results of different classification models of bearing faults.

Types	Classification	macro-P	macro-R	Accuracy
GAF	Two-classifications	100%	100%	**100%**
	Four-classifications	98.78%	98.63%	**98.79%**
	Ten-classifications	94.13%	94.22%	**96.29%**
WT	Two-classifications	100%	100%	**100%**
	Four-classifications	100%	100%	**100%**
	Ten-classifications	100%	100%	**100%**

To more intuitively display the results and perform misclassification analysis, Fig. 4 shows the confusion matrix of GAF images in three models [17]. The vertical axis corresponds to the predicted label of the samples, while the horizontal axis represents the true label of the samples. As depicted in Fig. 4, in the case of GAF images and the bearing two-classifications fault diagnosis task, all samples are accurately categorized into their respective classes. In the four-classifications, a few samples of BR and OR categories were misclassified into adjacent categories. In the ten-classifications, most of the misclassification samples of the BR007 category are classified into the BR021 category. Because both BR007 and BR021 belong to the BR category, their fault characteristics have certain similarities, so there are certain challenges in network classification.

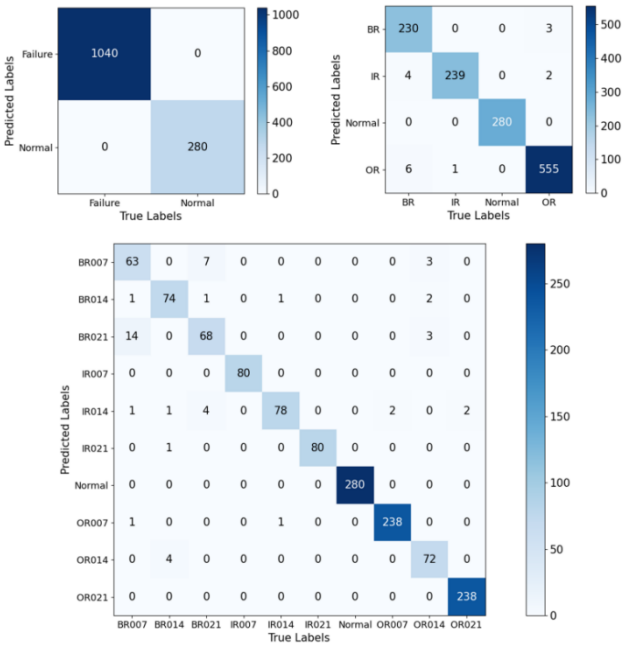


Fig. 4: Confusion matrix of GAF images in different classification tasks.

Fig. 5 shows the confusion matrix of WT images in three models. As can be seen from Fig. 5, WT images are correctly classified into corresponding categories in the tasks of bearing two-classifications, four-classifications and ten-classifications. Because the signal information expressed by WT image is more intuitive and simple. The experimental outcomes affirm the applicability of our approach within the domain of fault diagnosis.

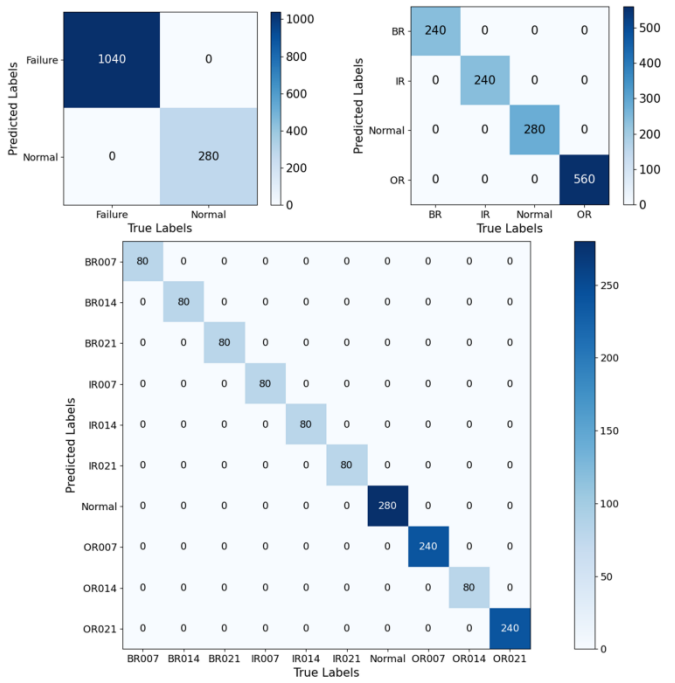


Fig. 5: Confusion matrix of WT images in different classification tasks.

To illustrate the performance of our proposed 2DImages-CNN framework, we compare it with other well-known approaches using the same dataset. The comparison results are shown in Table 3. The accuracy of the 2DImages-CNN model proposed by us is 100%, 98.79%, and 96.29% in two-classifications, four-classifications, and ten-classifications of GAF images, respectively. Our method achieves 100% accuracy on WT images, which is superior to many traditional methods. Additionally, compared with ANN, DBN, or SVM methods, which are applicable to only one or two kinds of classification tasks, the 2DImages-CNN method proposed in this paper can handle three kinds of diagnosis tasks in a unified framework and thus has wide applicability.

C. Grad-CAM explains deep fault features

To interpret the fault classification results, we use Grad-CAM to locate and highlight the key active areas at the final convolution layer of the Resnet34 model.

Select one sample from each of the four types of bearing failure GAF images for visualization and the visualization results are shown in Fig. 6. The first column of Fig. 6. shows the bearing original vibration signals, and it can be observed that the fluctuation of the normal signal is small, while the fluctuation of fault signals is significant. Therefore, bearing faults can be classified based on signal fluctuations. The red part of the image in the third column in Fig. 6 is the main part that affects the bearing fault classification. The experimental results show that the vibration signal of normal bearing is weak, and there is no obvious impact peak. The focus area of the network model is scattered on most segments of the input signal, which indicates that most signal sequences have basically the same contribution to the output; For bearings with outer race, inner race, and roller faults, the part with greater activation degree of the network is basically concentrated in the part with stronger vibration signal fluctuations, indicating that this part contains more fault feature information. The visualization results can prove that the area of concern of our proposed method in fault classification is correct, and solve the problem of black-box operation of neural network.

Table 3: Comparison of 2DImages-CNN method with other methods in bearing fault tasks.

Methods	Categories	Accuracy
2DImages-CNN(GAF)	**2**	**100%**
2DImages-CNN(WT)	**2**	**100%**
ANN [18]	2	92%
2DImages-CNN(GAF)	**4**	**98.79%**
2DImages-CNN(WT)	**4**	**100%**
LS-SVM [19]	4	89.50%
WT-CNN [6]	4	95.09%
RNN [20]	4	95.60%
2DImages-CNN(GAF)	**10**	**96.29%**
2DImages-CNN(WT)	**10**	**100%**
ConvRNN[21]	10	94.74%
DBN[22]	10	90%
WDCNN [23]	10	95.9%
CapsNet [9]	10	92.39%

Select one sample from each of the four types of bearing failure WT images for visualization and the visualization results are shown in Fig. 7. The second column of Fig. 7 shows that large amplitudes for different fault types are distributed over different frequency ranges. Therefore, networks can be classified based on areas with large

amplitudes. The third column in Fig. 7 shows that the experimental results show that our network pays attention to the right key area in fault classification, which solves the problem of black-box operation in fault diagnosis.

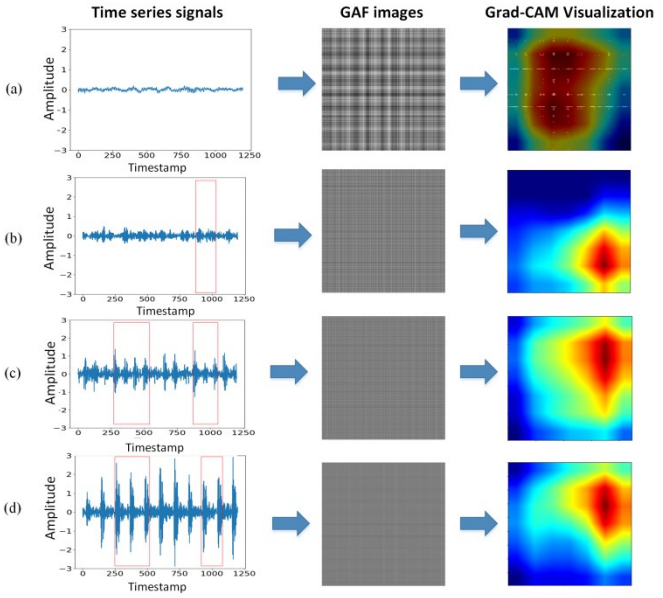


Fig. 6. Visualization results of GAF images: (a) Normal, (b) BR007, (c) IR007, (d) OR007.

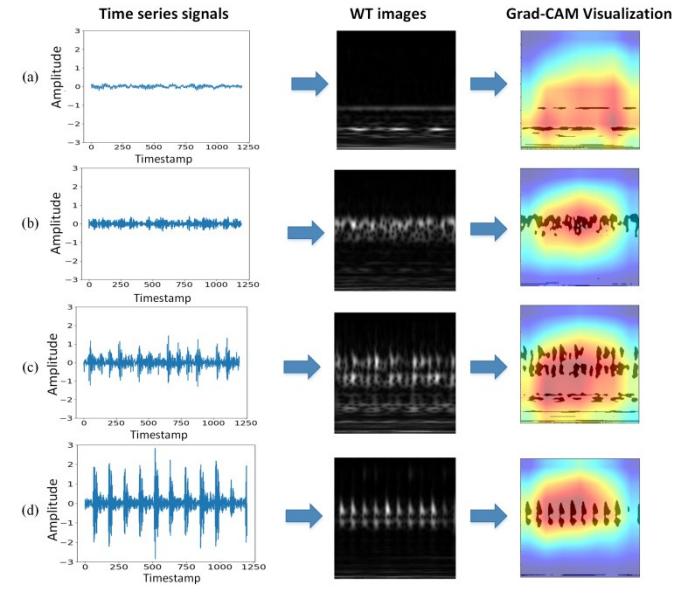


Fig. 7. Visualization results of WT images: (a) Normal, (b) BR007, (c) IR007, (d) OR007.

For commonly used deep learning networks, it is generally believed that they are a black-box and have weak interpretability. In response to this issue, many researchers are trying to explore its internal connections. However, the traditional method only classifies different fault types without considering the interpretability of fault classification. To fill this gap, we use Grad-CAM visualization. The experimental results show that for GAF images, the network makes classification decisions in areas with large fluctuations in signal values; For WT images, the network makes classification decisions in areas with large amplitudes.

IV. Conclusions

This paper introduced a bearing fault diagnosis approach utilizing 2DImages-CNN. Firstly, vibration signals are transformed into 2D images using the GAF and WT techniques. Then, the 2D images are input into the ResNet34 model for fault identification. The experimental results of the CWRU bearing dataset shows that the accuracy of our method is 100%, 98.79%, and 96.29% for two-classifications, four-classifications, and ten-classifications of GAF images, respectively. Our method achieves 100% accuracy in two-classifications, four-classifications, and ten-classifications of WT images. Our method demonstrates superior feature extraction capabilities and higher accuracy when compared to conventional approaches. This performance enhancement underscores the practical utility of our approach in fault diagnosis applications. Finally, Grad-CAM is used for visualization, and the visualization results show that when GAF images are visualized, the network classifies the bearing faults according to the correlation between areas with large signal fluctuations. When visualizing WT images, the network classifies bearing faults according to the areas with large amplitude.

In the future, we will attempt to apply the 2DImages-CNN model to the field of fault prediction. Secondly, we will conduct bearing failure experiments using our own dataset.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant No. 61472233) and the Natural Science Foundation of Shandong Province of China (Grant No. ZR2021MF048) and the Natural Science Foundation of Shandong Province of China (Grant No. ZR2022LZH004) and the Natural Science Foundation of China (No.62002207).

References

[1] X. Chen, B. Zhang, and D. Gao, “Bearing fault diagnosis base on multi-scale CNN and LSTM model,” Journal of Intelligent Manufacturing, vol. 32, no. 4, pp. 971-987, 2021.

[2] J. A. Lucena-Junior et al., “Chaos theory using density of maxima applied to the diagnosis of three-phase induction motor

bearings failure by sound analysis," Computers in Industry, vol. 123, p. 103304, 2020.

[3] Y. Zhang et al., "An enhanced convolutional neural network for bearing fault diagnosis based on time-frequency image," Measurement, vol. 157, p. 107667, 2020.

[4] F. He, and Q. Ye, "A bearing fault diagnosis method based on wavelet packet transform and convolutional neural network optimized by simulated annealing algorithm," Sensors, vol. 22, no. 4, p.1410, 2022.

[5] D. Verstraete et al., "Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings," Shock and Vibration, vol. 2017, 2017.

[6] P. Liang et al., "Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform," Computers in Industry, vol. 113, p. 103132, 2019.

[7] P. Liang, C. Deng, J. Wu, and Z. Yang, "Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network," Measurement, vol. 159, p. 107768, 2020.

[8] C.-L Yang, Z.-X Chen, and C.-Y Yang, "Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images," Sensors, vol. 20, no. 1, p. 168, 2019.

[9] B. Han, H. Zhang, M. Sun, and F. Wu, "A new bearing fault diagnosis method based on capsule network and Markov transition field/Gramian angular field," Sensors, vol. 21, no. 22, p. 7762, 2021.

[10] B. Zhou et al., "Learning deep features for discriminative localization," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921-2929, 2016.

[11] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.

[12] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," 2018 IEEE winter conference on applications of computer vision (WACV), pp. 839-847, 2018.

[13] Z. Wang, and T. Oates, "Imaging time-series to improve classification and imputation," arXiv preprint arXiv1506.00327, 2015.

[14] K. Huang et al., "Fault diagnosis of hydraulic systems based on deep learning model with multirate data samples," IEEE Transactions on neural networks and learning systems, vol. 33, no. 11, pp. 6789-6801, 2021.

[15] G. Tsoumakas, and I. Katakis, "Multi-label classification: An overview," International Journal of Data Warehousing and Mining (IJDWM), vol. 3, no. 3, pp. 1-13, 2007.

[16] "Bearing Data Center — Case Western Reserve University." https://engineering.case.edu/bearingdatacenter.

[17] Y. Yuan et al., "A general end-to-end diagnosis framework for manufacturing systems," National Science Review, vol. 7, no. 2, pp. 418-429, 2020.

[18] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection," Engineering applications of artificial intelligence, vol. 16, no. 7-8, pp. 657-665, 2003.

[19] W. Deng et al., "A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm," Soft Computing, vol. 23, no. 7, pp. 2445-2462, 2019.

[20] R. Zhao et al., "Machine health monitoring using local feature-based gated recurrent unit networks," IEEE Transactions on Industrial Electronics, vol. 65, no. 2, pp. 1539-1548, 2017.

[21] Y. Jin, L. Hou, and Y. Chen, "A time series transformer based method for the rotating machinery fault diagnosis," Neurocomputing, vol. 494, pp. 379-395, 2022.

[22] H Shao et al., "Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing," Mechanical Systems and Signal Processing, vol. 100, pp. 743-765, 2018.

[23] W. Zhang et al., "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," Sensors, vol. 17, no. 2, p. 425, 2017.

Study on Reduction of Background Fringes for Defect Detection of Specular Surface

An-Chi Wei*, Yi-Cheng Chang, and Jyh-Rou Sze
* National Central University, Taiwan, R.O.C.
E-mail: acwei@ncu.edu.tw Tel: +886-4227151#34378
National Central University, Taiwan, R.O.C.
E-mail: ilu0711@yahoo.com.tw Tel: +886-4227151#34389
Taiwan Instrument Research Institute, Taiwan, R.O.C.
E-mail: sze@narlabs.org.tw Tel: +886-5779911#618

***Abstract* — For the defect detection of specular surfaces, the image of the light source and other images of surrounding objects tend to be recorded, increasing the complexity of recognition. The method of combined dark-and-bright field illumination has been shown to improve the image quality of the defects on specular surfaces. However, there is still an issue of background fringes. In this study, we design a compensating pattern for diminishing the background fringes. The experimental results show that the proposed method can effectively dilute these fringes and increase the detection efficiency by fewer capture images.** *

I. INTRODUCTION

Detecting defects on specular surfaces, such as metal, glass, or glossy surfaces, tends to shoot the image of the light source and other images of surrounding objects. As a result, more procedures of subsequent computer processing are required, and the complexity of defect recognition increases. Additionally, some defects are only apparent at certain angles of incidence. When they are under test, the adjustable range of the instruments is limited. To improve the defect detection for the samples with specular surfaces, Forte et al. reported a system of combined dark and bright field illumination. The illumination in that system is provided by a monitor, displaying dark and bright lines to perform the dark-and-bright-field illumination [1-2]. After each image of the sample is captured, the illumination pattern changes. Although that method successfully improves the image quality, the background fringe is still an issue. To further improve the image quality of defect detection for specular surfaces, we propose a formula based on Forte's system for designing a compensating pattern. With the proposed compensating pattern, the detected image of defects is optimized by reducing the background fringes, leading to a higher defect-detection efficiency.

II. THEORY

According to Forte's approach, the illumination is a sequent of patterns with shifting dark and bright lines, as shown in Fig. 1. By their algorism, the gray value at (i, j) of the output image is expressed as [1]:

$$I_{i,j}^{out} = \sum_{k=1}^{N} \left| I_{i,j}^{k} - \frac{1}{N} \sum_{l=1}^{N} I_{i,j}^{l} \right|, \quad (1)$$

where N is the number of images and is the ratio between the pixels of one period, T, to the number of shifted pixels per pattern change, P:

$$N = \mathrm{T/P}. \quad (2)$$

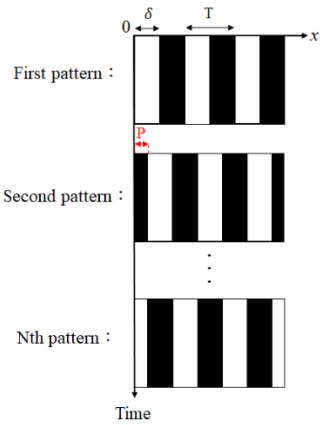


Fig. 1 Sequence of image patterns. δ, T, and P are the width of a bright line, the width of one period, and the number of shifted pixels between the adjacent patterns, respectively.

* The authors appreciate the financial support from the National Science and Technology Council, Taiwan, R.O.C (Grant Nos. 112-2221-E-008-053 and 112-2221-E-492-012).

The second term on the right side of Eq. (1) denotes the average gray values of these *N* images at *(i, j)*. The final output image I_{Final} is normalized by the formula:

$$I_{i,j}^{final} = I_{i,j}^{out} \times \frac{255}{I_{max}^{out}}. \quad (3)$$

The compensating image is for reducing the background dark fringes. These fringes are derived from the interface of each pattern shifting. Assume that in a row, the number of bright pixels per period, equivalent to the width of the bright line, is δ. Take a case with T = 3δ, P = 0.5δ as an example. The first background dark fringe occurs at the *x* coordinate of 0.5δ and 0.5δ+1, while the second one occurs at δ and δ+1. Thus, the compensating pattern is a set of bright lines in these regions. The modified algorism is

$$I_{i,j}^{out2} = \sum_{k=1}^{N} \left| I_{i,j}^{k} - \frac{1}{N} \sum_{l=1}^{N} I_{i,j}^{l} \right| + c \times I_{i,j}^{com} \ , \quad (4)$$

where $I_{i,j}^{com}$ is the gray value of the compensating pattern at *(i, j)*. *c* is a ratio factor. The final image is also normalized based on the following equation:

$$I_{i,j}^{final2} = I_{i,j}^{out2} \times \frac{255}{I_{max}^{out}}. \quad (5)$$

III. Results and Discussion

The experimental setup consists of an illumination source, a camera, and a sample, as shown in Fig. 2. The illumination source is an LCD monitor (ASUS) with 840 × 1050 active pixels. The camera is a CMOS camera (acA5472-17, Basler) with a lens (TUS-1220, TUSS). The illumination source and the camera are connected to a computer with codes for automatedly changing illumination patterns, triggering the camera, receiving the captured images, and processing images. The image processing codes include the algorithms above. The sample is a plate made of stainless steel, and the size of the specular surface is 12 cm × 11 cm. There are 14 hand-made defects, as shown in Fig. 3.



(a) (b)

Fig. 2 System setup (a) illumination pattern (b) sample, camera, and illumination source

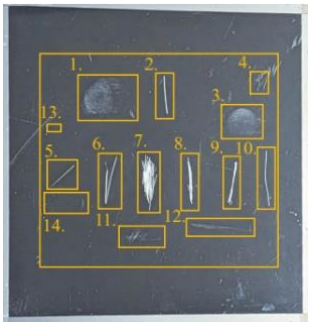


Fig. 3 Sample made of stainless steel with 14 hand-made defects

The compensating pattern can be predesigned by calculating the interface regions during pattern shifting and setting these regions with high gray values. The gray values of other regions are assigned as zero. However, the exact gray values for the interface regions are difficult to define by calculation. Thus, a pretest of the no-defect sample is done to examine the background fringes' gray values and to assign the proper gray values for the compensating pattern. An example of the compensating pattern is shown in Fig. 4.

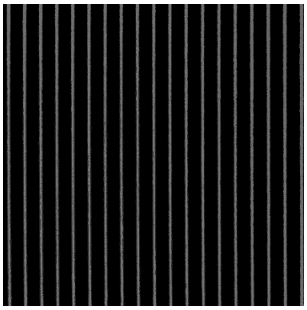


Fig. 4 Example of the compensating pattern

Firstly, the system is operated with three configurations whose numbers of shifted pixels P are δ, $\delta/2$, and $\delta/3$, respectively, to test the algorithm without compensating pattern. The width of the bright line of each configuration is kept the same: δ = 60 pixels. Then, the background fringes in the captured images are observable, as illustrated in Fig. 5. The results exhibit that with the decreasing number of shifted pixels, P, the background fringes become less apparent. However, according to Eq. (2), the required number of images, N, is increased when the period, T, is fixed.

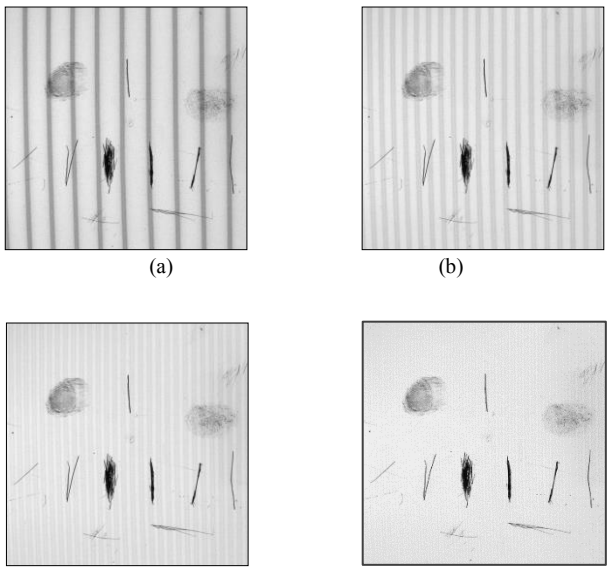


Fig. 5 Captured images without the compensating patterns in the configurations of P = (a) δ, (b) $\delta/2$, (c) $\delta/3$ and (d) $\delta/4$.

If the compensating patterns are applied, the background fringes of these three configurations become dilute, as shown in Fig. 6. Thus, the proposed method has effectively improved the image quality. The parameters of these configurations are listed in Table I. Because of the appended compensating pattern, the required number of images is N+1.

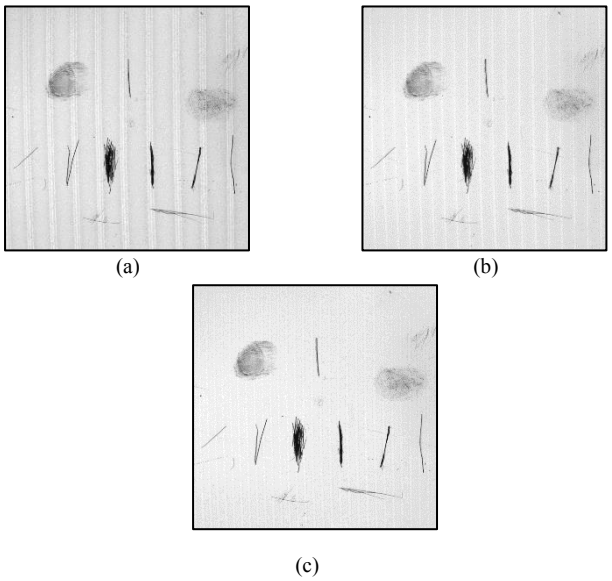


Fig. 6 Captured images with the compensating patterns in the configurations of P = (a) δ, (b) $\delta/2$, and (c) $\delta/3$

Table I Parameters of configurations

δ = 60 pixels				
T = 3δ				
Trial		1	2	3
P		δ	$\delta/2$	$\delta/3$
Required number of images	Without compensation	3	6	9
	With compensation	4	7	10
c (ratio factor)	With compensation	1.6	1	0.8

To enhance the contrast between the defects and the background, we processed the captured images by a modified binary technique [1,3]. Unlike the conventional Niblack's binarization using standard deviation and local mean to count the threshold [4-6], the modified binary technique consists of another ratio factor for adjusting the strength of the standard deviation. The processed image B_{Final} satisfies the the following equation [1]:

$$B_{Final,i,j} < A_{Temp,i,j} - Y \times S_{Temp,i,j}, \quad (6)$$

where Y is a ratio factor. $A_{Temp,\,i,j}$ and $S_{Temp,\,i,\,j}$ are the average gray value and the standard deviation of a region with the center coordinate of (i, j), respectively. The size of each region is (2M + 1) × (2N + 1). Meanwhile, $A_{Temp,\,i,j}$ and $S_{Temp,\,i,\,j}$ can be expressed as:

$$A_{i,j}^{Template} = \frac{1}{(2M+1)(2N+1)} \sum_{k=-N}^{N} \sum_{l=-M}^{M} \left[B_{i-k,j-l}\right] \quad (7)$$

and

$$S_{i,j}^{Template} = \left\{\sum_{k=-N}^{N} \sum_{l=-M}^{M} \left[(I_{i-k,j-l} - A_{i,j}^{Template})^2\right]\right\}^{1/2} \quad (8)$$

The defect images after binarization are illustrated in Fig. 7. These configurations are also all under δ = 60 pixels and T= 3δ. Figures 7(a) and 7(b) are based on the images without compensation, while Fig. 7(c) and 7(d) are from the compensated ones. The Y factor for these four configurations are 0.82, 0.6, 0.55, and 0.45, respectively. In addition to the configurations with P = δ, $\delta/2$, and $\delta/3$, as shown in Fig. 6, the configuration of P = $\delta/4$ without compensation was included in Fig. 7 for comparison. As shown in Fig. 5(d), the background black fringes of this configuration are unapparent. However, because the number of shifted pixels P is reduced to δ /4, the required number of images increases to N = 12, which adds to the processing time. The analyzed results show that the ratio of the dark pixels to the total pixels of the compensated output is

slightly higher than that without compensation, as shown in Fig. 8.

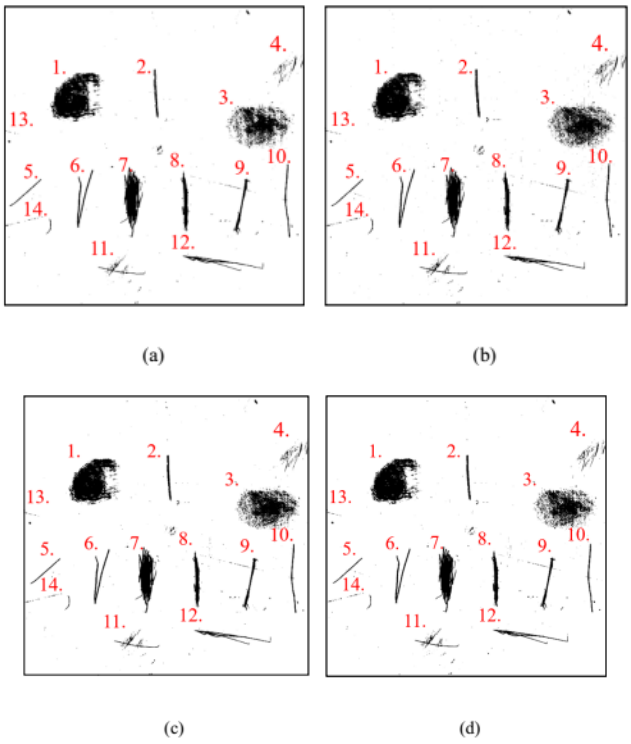


Fig. 7 Defect images after binarization in the configurations with δ = 60 pixels, T = 3δ, and P = (a) $\delta/3$ without compensation, (b) $\delta/4$ without compensation, (c) $\delta/2$ with compensation, and d) $\delta/3$ with compensation.

In other words, the defects in the compensated images are more obvious than those in the uncompensated images. It proves that the image quality of the compensated result is acceptable. Moreover, since the detection speed is one of the critical parameters of automatic optical inspection (AOI) techniques [7-8], fewer captured images can bring about faster detection. For the demonstrated configurations, the required number of images for the configurations in Fig. 7(a) to 7(d) are 9, 12, 7, and 10, respectively. Thus, the compensation configuration shown in Fig. 7(c) requires fewer capture images, implying the increment of the detection efficiency.

IV. Conclusions

This study is based on an image-processing method proposed by Forte et al. and aims to solve the background fringes of that image-processing method. According to the analyses of the captured background dark fringes, it is concluded that the junction of the bright and dark areas of the captured image induced these background dark fringes. Therefore, this study proposes a scheme that compensates for the illumination. From the experimental results, this scheme can effectively compensate for the dark fringes. The defects on the image after compensation do not have obvious fading or cracking. Finally, this paper uses the dark ratio as an indicator to compare the binarized images of several configurations. The final result shows that the configuration of P = $\delta/2$ with compensation has a better dark ratio, and its required number of images is fewer than other illustrated configurations, equivalent to the increment of the detection efficiency.

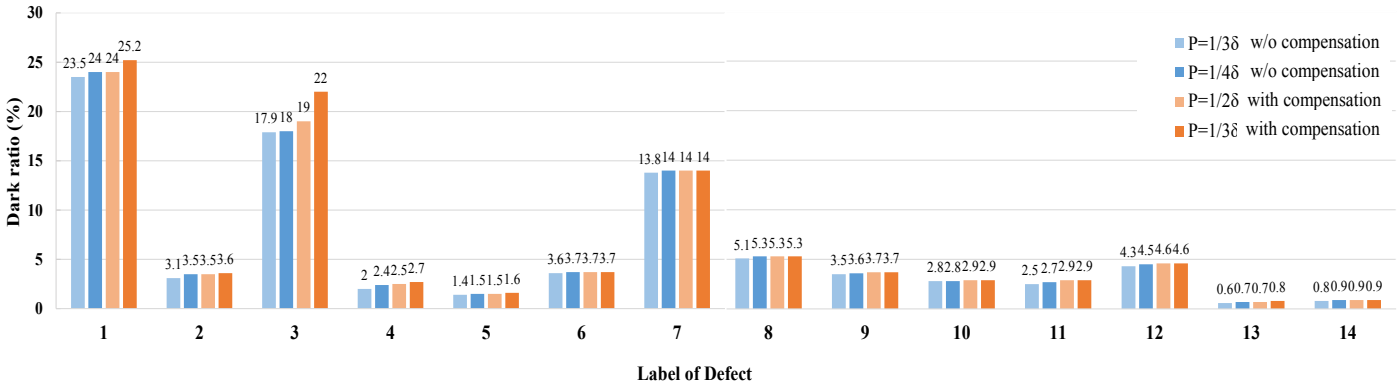


Fig. 8 Comparison of dark ratios of the configurations illustrated in Fig. 7.

REFERENCES

[1] P. M. F.Forte, P. E. R. Felgueiras, F. P. Ferreira, M. A. Sousa, E. J. Nunes-Pereira, B. P. J. Bret, M. S. Belsley, "Exploring combined dark and bright field illumination to improve the detection of defects on specular surfaces." Optics and Lasers in Engineering, 88, 120–128, **2017**.

[2] Q. Sun, K. Xu, H. Liu, and J. Wang, "Unsupervised surface defect detection of aluminum sheets with combined bright-field and dark-field illumination." Optics and Lasers in Engineering, 168, 107674, **2023**.

[3] P. L. Saxena, "Niblack's binarization method and its modifications to realtime applications: a review," Artificial Intelligence Review, 51, 673-705, **2019**.

[4] W. Niblack, "An introduction to digital image processing," Prentice-Hall International, **1986**.

[5] T. Kalaiselvi, P. Nagaraja, and V. Indhu, "A comparative study on thresholding techniques for gray image binarization," International Journal of Advanced Research in Computer Science, 8, 1168-1172, **2017**.

[6] L. P. Saxena, "Niblack's binarization method and its modifications to real-time applications: a review," Artificial Intelligence Review, 51, 673-705, **2019**.

[7] J. Karangwa, L. Kong, D. Yi, and J. Zheng, "Automatic optical inspection platform for real-time surface defects detection on plane optical components based on semantic segmentation." Applied Optics, 60, 5496-5506, **2021**.

[8] Y. Du, J. Chen, H. Zhou, X. Yang, Z. Wang, J. Zhang,et al., "An automated optical inspection (AOI) platform for three-dimensional (3D) defects detection on glass micro-optical components (GMOC)," Optics Communications, 545, 129736 **2023**.

On the Optimal Self-Supervised Multi-Fault Detector for Temperature Sensor Data

Latifa Nabila Harfiya[1], Yan-Cheng Hsu[2], Yung-Hui Li[3], and Jia-Ching Wang[1]
[1] Department of Computer Science and Information Engineering, National Central University, Taiwan
[2] Department of Computer Science and Engineering, University of California San Diego, USA
[3] AI Research Center, Hon Hai Research Institute, Taiwan
E-mail: latifanaharfiya@g.ncu.edu.tw, bill.yancheng.hsu@gmail.com, yunghui@foxconn.com, jcw@csie.ncu.edu.tw

***Abstract* — Accurate detection of faults in sensor data is essential for monitoring and controlling industrial processes, environmental conditions, and infrastructure to ensure reliability and enable informed decision-making. Resolving these faults ensures measurement quality and unlocks process optimization opportunities, resulting in improved performance, energy efficiency, and cost savings. We present a transformer-based fault detection model which adopts the anomaly-attention mechanism. Experiments have been performed on the benchmark faults injected Intel temperature sensor datasets using precision, recall, and F1-score metrics. The result outperforms the other classical and complex algorithms, proving our method's effectiveness.**

I. INTRODUCTION

Currently, there is a significant shift happening in the manufacturing industry. Integrating big data and machine learning, especially deep learning, is revolutionizing traditional manufacturing practices, ushering in the era of smart manufacturing known as Industry 4.0 (I4.0). This shift presents new opportunities for manufacturers. Within this context, the Industrial Internet of Things (IIoT) emerges and connects machines through various sensors, RFID tags, software, and electronics, allowing real-time data collection. The proliferation of intelligent sensors and the IoT serves as crucial facilitators for collecting and storing vast amounts of data in industrial settings [1, 2]. Sensor data is often used to detect abnormal conditions or events requiring immediate attention or action. It is crucial for accurately monitoring and controlling industrial processes, environmental conditions, and infrastructure. Detecting faults in sensor data ensures the reliability and quality of the measurements, which is vital for making informed decisions and maintaining process integrity. By identifying and resolving these faults, process optimization opportunities can be realized, leading to improved performance, reduced energy consumption, and cost savings [3, 4].

Sensor data can encompass various types of information, including numerical readings, categorical data, or even event-based data. Time series data refers to data collected and recorded over a sequence of time intervals or timestamps, where the order of the observations is essential. This is commonly seen in sensor data that captures measurements at regular intervals, such as temperature readings, humidity levels, or vibration intensity recorded at specific time points. Many works address the temporal variation of time series data, which is more meaningful and can reveal intrinsic properties of time series, such as continuity, regularity or trends, because one single point in time typically cannot give enough information for analysis [5].

A fault detection framework has been presented in [3] using a Support Vector Machine (SVM), which is optimized with a biologically-inspired technique, Grey Wolf Optimization (GWO). The results suggest that when the failure rate stays at a low level, an accurate diagnosis of sensor fault can be made. Recurrent Neural Network (RNN) is widely considered a powerful tool for analyzing and modeling time series data due to its ability to capture sequential patterns and dependencies. In [3] and [6], Long Short-Term Memory (LSTM), a variant of RNN, is utilized to make an end-to-end prediction fault diagnosis to prognosis. On the other hand, a stacked LSTM has fully exploited the long-term dependencies information and improved the fault diagnosis performance in [7]. Although using LSTM might be sufficient, architectures like Transformers, originally introduced for natural language processing tasks, have also gained attention in the time series domain, especially for tasks involving long sequences or capturing global dependencies. Several research studies [8-10] and practical applications have demonstrated the effectiveness of transformers in time series anomaly detection. Using transformers, one can capture long-range dependencies, exploit global context, and leverage self-attention mechanisms to

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

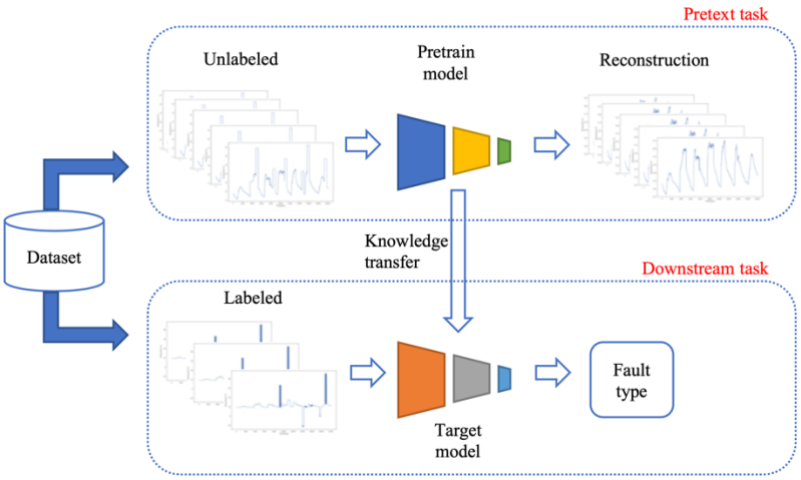


Fig 1. Proposed framework

detects anomalies in complex and high-dimensional time series data.

In order to obtain a progressive detection method to help the model distinguish sensor conditions, we adopt a Transformer-based network architecture from [11]. The literature utilizes the network to solve an unsupervised anomaly detection task. Inspired by a Self-Supervised Learning (SSL) approach, we first train the network model as a denoising network that attempts to remove the faults, which we consider as noise, from the one-dimensional input time series signal and reconstruct the output like the original clean input. In SSL, this first step is known as the pretext task. Next, given less quantity of labeled data, we finetuned the pretrained network model to detect faults based on their class. This step is known as the downstream task or target task in SSL. Furthermore, the experimental results are examined in a comparative manner to confirm the effectiveness of the method presented in this research paper for detecting temperature sensor faults.

II. METHOD

The proposed framework, shown in Fig. 1, is divided into two phases: the pretext task to direct the model toward acquiring intermediary representations of data and the downstream task to transfer the knowledge from the finetuned pretext model. We split our dataset for the pretext task and the downstream task. We use the interpolated time-series signal for each task to help the model observe the learnable pattern. We use the clean signal as the ground truth for the first task and the class label as the ground truth for the latter. Then, we apply the preprocessing techniques outlined in [11], including dividing the dataset into consecutive non-overlapping segments using a sliding window approach. The adopted network architecture introduces association discrepancy and incorporates an Anomaly-Attention mechanism that models the prior-association and series-association to the vanilla Transformer as the core element.

A. Anomaly-Attention Block

Instead of using point-wise representation, the original self-attention mechanism properties, using prior-association, which benefiting the unimodal property of Gaussian kernel, and Series-association, which is used to learn the associations from raw series, are found to be more informative. The adjacent-concentration prior and the learned real associations are also reflected in this approach and their differences are meant to be distinguishable between normal and abnormal cases [11]. Specifically, this association-based criterion attention block has demonstrated the ability to consistently yield smaller values for the normal component, which starkly contrasts the point-contextual and pattern-seasonal scenarios.

B. Focal Loss

Focal Loss [12] was proposed to address the problem of imbalanced data by introducing a new loss control function assigning higher weights to challenged or incorrectly classified cases. To reduce the loss contribution of highly classified examples and focus on more difficult examples is an important idea behind this loss function, shown in Eq. (1)

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log p_t \qquad (1)$$

where $FL(p_t)$ represents the focal loss for a specific example with predicted probability p_t. α_t is a balancing factor to address class imbalance. It is often defined as the inverse of the class frequency or adjusted based on the dataset characteristics. γ is the focusing parameter that controls the rate of downweighing for well-classified examples and pt represents the predicted probability for the ground truth class. By incorporating Focal Loss into the training process, the model can focus more on the challenging examples, improving its ability to handle class imbalance and boost performance, particularly in scenarios where the number of abnormal examples is significantly lower than the number of normal examples. Since faults in sensor data are expected to happen infrequently, thus, we utilize focal loss for our target model's loss function.

III. Experiment and Analysis

A. Dataset and Evaluation Metrics

Data-driven methods are highly prevalent and successful because they are based on data extracted from historical processes and large quantities of such data generated by modern production. However, for security reasons, this information is generally kept private. Thus, we conduct experiments using the public simulation data published by Intel Labs with injected faults from [13]. The injected fault types include bias, drift, malfunction, random, polynomial drift, and mixed, as shown in Fig. 2, along with the normal and interpolated original data. The original data is obtained from 10 sensors and a total data sample of 3,980,192. This fault detection task is evaluated using precision, recall, and the F1-score.

B. Implementation Details

The temperature sensor data obtained are univariate, indicating that each time series example consists of only one channel. Consequently, we convert the time series into a multivariate format with a single channel by employing a straightforward reshaping technique using the Numpy library. The model has 3 layers and is trained for 10 epochs for each phase with batch size 64 and mean squared error (MSE) as the loss function for the pretext task. This approach will enable the data to fit the model with multivariate time series input. The experiments are implemented in PyTorch with a single NVIDIA GeForce GTX 1080Ti 11GB GPU.

C. Performance Comparison

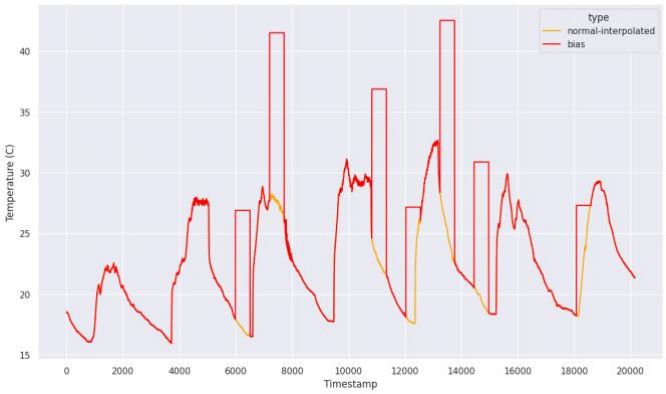
The fault detection task's performance evaluation relies on a critical metric, i.e., the probability of occurrence for the positive class containing faults. We require the commonly used performance metrics for the experiment: precision, recall, and F1-Score. Table I shows the performance comparison of our proposed method with other competing methods. FOCUSeR is based on fog computing that uses online learning and constant updating over time to avoid faults. The second method is a sensor fault diagnosis method based on α Grey Wolf Optimization Support Vector Machine (α-GWO-SVM). We compare our results with the reported results in the literature for the first two comparisons due to the code's unavailability. Lastly, the original Anomaly-Transformer is also tested on the dataset. In this experiment, we use the binary class (normal and fault) of the signal as the ground truth to the original model instead of the multi-class label. It can be seen that our method outperforms other methods on the benchmark dataset. Moreover, compared with the best performance of other methods, our method has a significant improvement on the F1-Score by 1.69% over the best results of other methods, which clearly shows that our method has good performance for the task.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT METHODS

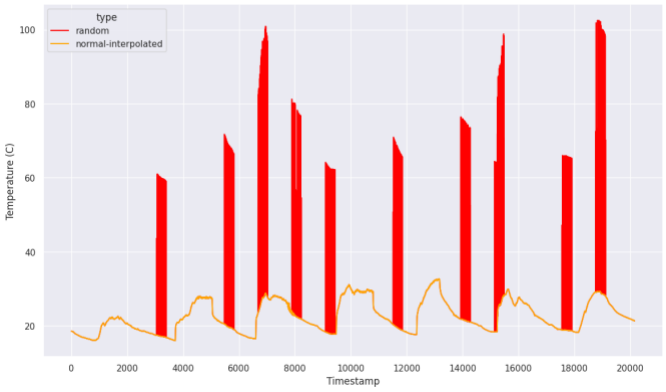
Method	Precision	Recall	F1-Score
FOCUSeR [14]	96.1%	97.7%	96.9%
α-GWO-SVM [3]	97.29%	-	-
Anomaly-Transformer [11]	90.43%	92.18%	91.29%
Ours	98.07%	99.11%	98.59%

(-), not known, the authors did not evaluate the results of their methods using certain metrics nor make their codes available.

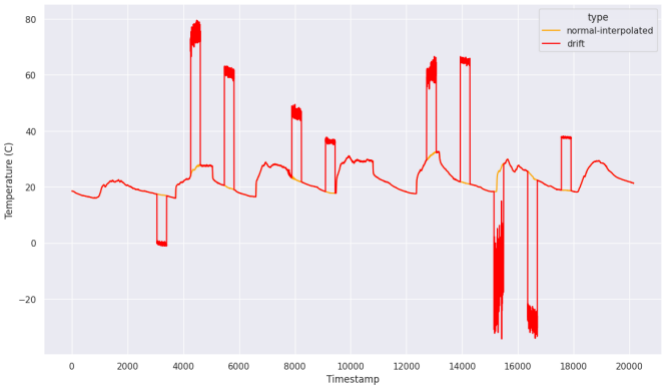
Our proposed method learns the pattern of the signals by reconstructing the clean signal from the fault-injected signal. This imposes the model to identify anomalous patterns that have never appeared in the clean signal. Moreover, the performance of the model indicates that the difference between normal and abnormal time points is effectively intensified by adapting the anomaly-attention block. Thus, our proposed method achieves the state-of-the-art result on an exhaustive set of empirical studies.



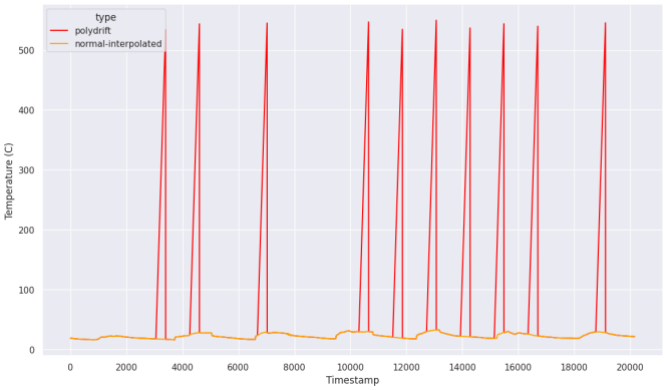
(a) Example of bias fault



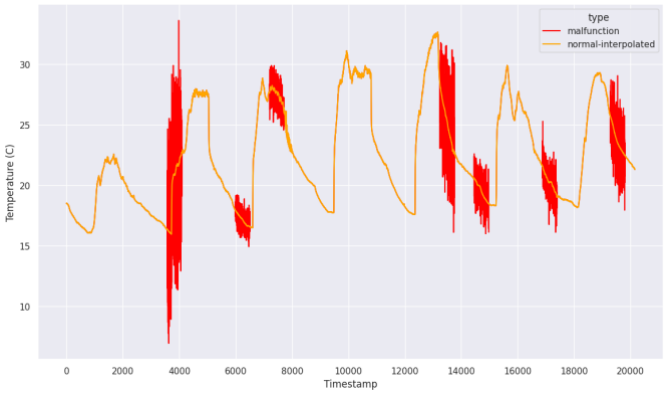
(b) Example of random fault



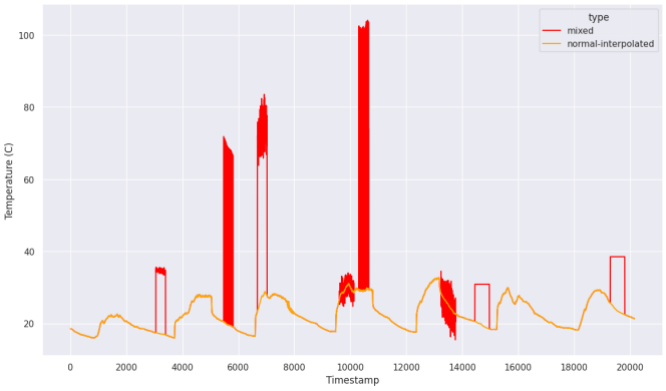
(c) Example of drift fault



(d) Example of polynomial drift fault



(e) Example of malfunction fault



(f) Example of mixed fault

Fig. 2 Examples of injected fault

From a statistical standpoint, faults are infrequent occurrences, that happen less often than the normal class. The primary way to describe this behavior is by considering the anticipated occurrence rate of anomalies. Thus, the specific advantage of applying Focal loss for the experiment is observed. Table II demonstrates the effectiveness of using Focal loss as the target model's loss function. Focal loss can differentiate various signal conditions, enabling the model to adapt more effectively. This adaptation facilitates improved fault detection accuracy by enhancing the model's ability to accurately identify faults in diverse scenarios.

TABLE II. PERFORMANCE OF PROPOSED MODEL UNDER DIFFERENT LOSS FUNCTION

Method	Loss Function	F1-Score
Pretrain	MSE	96.58%
Target	MSE	
Pretrain	FL	94.87%
Target	FL	
Pretrain	MSE	98.59%
Target	FL	

IV. CONCLUSIONS

In this research paper, we propose an effective self-supervised approach for detecting multi-fault in temperature sensor data. Firstly, we enhance the discriminative power of the anomaly attention block. Next, we incorporate attribute information of sensor data into the model training through a customized loss, enabling the detector to differentiate better and adapt to diverse sensor conditions, resulting in improved accuracy in detecting faults. Lastly, experimental results validate the effectiveness of our proposed method, and our future work involves developing a more efficient model structure for temperature sensor data multi-fault detection.

REFERENCES

[1] R. Rai, M. K. Tiwari, D. Ivanov, and A. Dolgui, "Machine learning in manufacturing and industry 4.0 applications," *International Journal of Production Research,* vol. 59, no. 16, pp. 4773-4778, 2021/08/18 2021, doi: 10.1080/00207543.2021.1956675.

[2] M. Sujatha *et al.*, "IoT and Machine Learning-Based Smart Automation System for Industry 4.0 Using Robotics and Sensors," *Journal of Nanomaterials,* vol. 2022, p. 6807585, 2022/09/15 2022, doi: 10.1155/2022/6807585.

[3] X. Cheng, D. Wang, C. Xu, and J. Li, "Sensor Fault Diagnosis Method Based on -Grey Wolf Optimization-Support Vector Machine," *Computational Intelligence and Neuroscience,* vol. 2021, pp. 1-15, 09/10 2021, doi: 10.1155/2021/1956394.

[4] S. U. Jan, Y. D. Lee, and I. S. Koo, "A distributed sensor-fault detection and diagnosis framework using machine learning," *Information Sciences,* vol. 547, pp. 777-796, 2021/02/08/ 2021, doi: https://doi.org/10.1016/j.ins.2020.08.068.

[5] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis," *arXiv preprint arXiv:2210.02186,* 2022.

[6] Q. Wu, K. Ding, and B. Huang, "Approach for fault prognosis using recurrent neural network," *Journal of Intelligent Manufacturing,* vol. 31, pp. 1621-1633, 2020.

[7] Q. Zhang, J. Zhang, J. Zou, and S. Fan, "A Novel Fault Diagnosis Method based on Stacked LSTM," *IFAC-PapersOnLine,* vol. 53, no. 2, pp. 790-795, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.ifacol.2020.12.832.

[8] Y. Liu, H. Wu, J. Wang, and M. Long, "Non-stationary transformers: Rethinking the stationarity in time series forecasting," *arXiv preprint arXiv:2205.14415,* 2022.

[9] H. Zhou *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, no. 12, pp. 11106-11115.

[10] S. Tuli, G. Casale, and N. Jennings, "TranAD: deep transformer networks for anomaly detection in multivariate time series data," *Proceedings of the VLDB Endowment,* vol. 15, pp. 1201-1214, 02/01 2022, doi: 10.14778/3514061.3514067.

[11] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv preprint arXiv:2110.02642,* 2021.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

[13] B. Bruijn, T. A. Nguyen, D. Bucur, and K. Tei, *Benchmark Datasets for Fault Detection and Classification in Sensor Data*. 2016, pp. 185-195.

[14] F. Costa, S. Nassar, and M. Dantas, "FOCUSeR: A Fog Online Context-Aware Up-to-Date Sensor Ranking Method," *Journal of Sensor and Actuator Networks,* vol. 11, p. 25, 05/17 2022, doi: 10.3390/jsan11020025.

Application of Wafer Defect Pattern Classification Model in the Semiconductor Industry

Chin-Wei Lee[1], Daniel Hládek[2], Matúš Pleva[3], Yuan-Fu Liao[4] and Ming-Hsiang Su[5]
[1] Soochow University, Taiwan
E-mail: vivibank888@gmail.com
[2] Technical University in Košice, Slovakia
E-mail: daniel.hladek@tuke.sk
[3] Technical University in Košice, Slovakia
E-mail: matus.pleva@tuke.sk
[4] National Yang Ming Chiao Tung University, HsinChu , Taiwan
E-mail: yfliao@nycu.edu.tw
[5] Soochow University, Taiwan
E-mail: huntfox.su@gmail.com

***Abstract* — Deep learning (DL) methods are widely employed in the semiconductor manufacturing process to enhance pattern recognition and classification accuracy, specifically for addressing defect patterns. However, the classification performance of the current models is hindered by the imbalanced distribution of defect data within the test dataset. To tackle this issue, this study presents a feature extraction approach utilizing data transformation, and ensemble learning techniques aiming to enhance the model's classification performance. The primary objective of this study is to mitigate selection and imbalance problems in the dataset through random sampling and assigning distinct weights to individual classifiers. The results demonstrate that the proposed method achieves an impressive accuracy rate of 95.09%, thus substantiating its efficacy in improving the robustness of both the classification model and wafer classification.**

I. INTRODUCTION

Semiconductors, also known as integrated circuits (ICs), are essential to the economy and often display the crucial components of electronics. During these few years, the chip requirement of electronic devices become smaller, faster, and cheaper, which increases the degree of complexity [1]. To organize the manufacturing activities, the main production processes in the supply chain are ICs design in upstream, ICs manufacturing in mid-stream, and outsourced assembly and testing (OSAT) in downstream [2]. In wafer processing, wafers are composed of silicon extracted from sand. Then, it becomes the silicon rod called an ingot, which needs to be heated to get high-purity silicon [3]. An ingot gets cut into sliced wafers called a die.

After passing the Chip Probing (CP) test on manufactured wafers, engineers obtain some degree of Wafer Bin Map (WBM) on defective or non-defective chips and Systematic Defect or Random Defect [4]. While analyzing the result of pattern recognition, some particular figures could be the cause of machine abnormality. At the same time, Wafer Bin Map patterns' diversity and complexity are hard to recognize depending on domain knowledge from engineers, which wastes time and lacks systematic approaches. Most semiconductor companies are required to shorten cycle time and maximize the benefits. Therefore, how to use deep learning (DL) and data analysis methods to increase the yield rate, optimize the parameters adjustment, and respond to the defective pattern immediately and accurately is the core problem for the semiconductor industry.

DL is more mathematically complex in algorithms, which can be regarded as a subset of ML based on Artificial Neural Networks (ANNs) [5]. It is one of the approaches to learning good functions and is used in various NNs to solve complex problems. The main point of the DL process is that the feature extractions in the model are all self-learning by computer. The three most used neural networks in DL are Feedforward Neural Networks (FNNs) [6], Convolution Neural Networks (CNNs) [7], and Recurrent Neural Networks (RNNs) [8].

In the semiconductor industry, defects are a critical issue during the manufacturing process. Any defect can significantly impact the quality of the wafer, and it is essential to identify the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

subtle differences among them. Traditional methods often lack random sampling, which fails to ensure equal opportunities for every minority class in imbalanced problems and may lead to biased inferences. To address these challenges, we focus on ensemble learning methods that prevent training on noisy data and reduce model instability. Moreover, we propose a novel approach that combines the soft voting method with assigning different weights to individual classifiers, thereby increasing the likelihood of selecting minority classes to solve the imbalance problem. In this study, we incorporate the best-performing individuals into a weighted soft voting scheme, calculated using an optimization function. The primary objectives of this research are as follows: (1) employing random sampling in classification problems to minimize selection bias and sampling errors, and (2) combining the classification results of individual classifiers with different weights to enhance overall classification performance.

II. Related Work

A. K-nearest neighbors (KNN)

KNN uses the k-nearest neighbors to determine which group the new data point belongs to [9]. The algorithm adopts a majority vote and adds the data point into the specific group by calculating the distance between new data and the adjacent data. The benefits of KNN are its high accuracy and suitability for multiclass classification problems. The limitation of KNN was due to its requirement of huge processing data, which would weaken when met with small datasets.

In [10], a brief review of the KNN method and its modified versions, such as weighted adjusted KNN and SVM-KNN were described, and those variants removed the weakness of KNN. In [11], a new method called MSML-KNN by adding Matching Coordinate Rotation (MCR) and matching scratches based on KNN was proposed. It improved the KNN algorithms and provided valuable information about the scratches for semiconductor manufacturers. The research indicated that some KNN methods needed high computational costs, and noise might cause confusion.

B. Decision Trees Classification (DTs)

DTs classify the data using the greedy rule and generate the prediction result according to different conditional judgments of each layer [12]. The strengths of DTs are their robustness to noise and are good to handle large-scale data. In [13], a new DT-based wafer map failure pattern method with a radon transform feature was proposed, and [14] generated an Alternative Decision Tree (ADTree) classifier combined with cluster recognition. The combination of ADTree and cluster overcame the low signal-to-noise problem. In [15], the comparison between KNN, Naïve Bayes, and DTs classification techniques. The comparative parameter includes speed, accuracy, effectiveness, etc. The result indicated that the DTs have the highest accuracy and less error rate than KNN and Naïve Bayes.

C. Support Vector Machine (SVM)

SVM is used to find a decision boundary to maximize the margins between the two classes, and the support vector classification (SVC) is used for the classification problem [16]. The advantages of SVM are its high dimensionality and its suitability for small samples. In [17], it provided a comprehensive introduction to SVM and has proven that it was one of the best-supervised learning methods in the application. The weakness of SVM was selecting the parameters and the performance in unbalanced data sets. In [18], the SVM with multiple selected features combined with a CNN model gets a better result than only one model, which pointed out CNN model with SVM could yield improved consequences.

III. Proposed Method

A. Feature Extraction Method

Four feature extraction methods used in this research: density-based features, radon-based features, geometry-based features, and convolutional-based models. For example, we have selected labels 36, 255, 11, 151, 94, 39, 24, and 65, corresponding to eight wafer defect pattern types (Center, Donut, Edge-Local, Edge-Ring, Local, Random, Scratch, and Near-full) as shown in Fig. 1.

We evenly divided the wafer map into 20 parts and re-divided 13 regions using a density-based feature extraction technique [19]. It contains the top, right, left, bottom, and the same nine inner regions. Then, calculated the failure density of each region is. The radon-based features are related to the Radon transform, which can compute tomography for the medical field and gain a tomographic image from projections

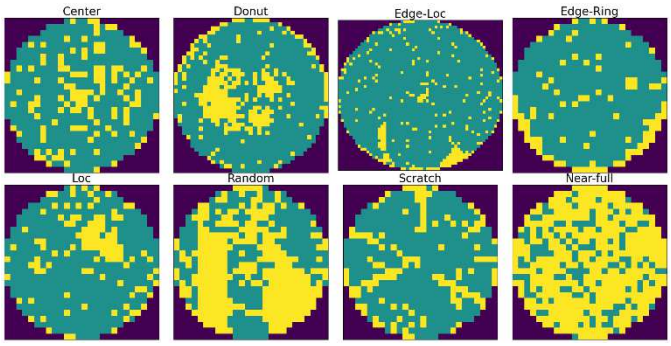


Fig. 1 An Example of Eight Wafer Defect Pattern Types.

[20]. Mathematically defined, the radon transform converts a two-dimensional planar function into a linear function in two-dimensional space. The concept of projection in the Radon transform transfers the 2D grayscale wafer map image at the specified angle by theta, and the Radon transform signal after projection is called a sinogram. Due to the multiple sizes of the wafer, we implement cubic interpolate to calculate the row mean and row standard deviation and each of them extracts 20 dimensions, which means the total is 40 dimensions.

The technique of geometry-based features is to evaluate the geometric characteristics of each wafer map [21]. Based on the salient region and the variety of wafer maps, the geometry-based features used to be obtained by computing local, statistical, and linear characteristics. The scale and rotation of these characteristics are invariant, and the region-labeling algorithm chooses the most salient region, which is the max region area of the wafer defect pattern. The six characteristics of geometry features choose by calculating area, perimeter, length of major axes, length of minor axes, solidity, and eccentricity. Solidity represents a measure of the concentration of an object, and it is the proportion of pixels in the object to its convex hull of defective areas. Eccentricity represents a measure of the elongation or stretching of an object and sets 0 as a round object and positive otherwise.

B. Classification of the Data

Fig. 2 shows the architecture of the proposed model framework. The individual classifier includes ML classifiers: KNN, SVM, CategoricalNB [22], DTs, and Multilayer Perceptron (MLP) in DNN. The ensemble classifier includes soft voting, hard voting, stacking, bagging, and AdaBoost.

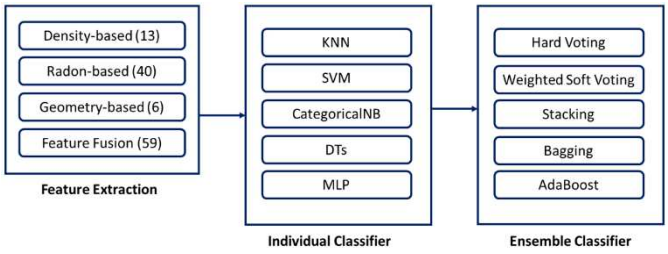


Fig. 2 The architecture of the proposed model framework.

KNN is an algorithm that determines which group the new data belongs to in the classification problem. Three steps of the algorithm process include deciding the value of *k*, finding the distance between each neighbor and itself, and joining the group which has the most neighbor. The algorithm compares the similarity between existing and new labels, then classifies them into the corresponding group. The classification accuracy of the KNN algorithm is high due to its robustness to noisy training data, and huge datasets, and is suitable for the multi-class problem.

MLP uses the concept of perceptron, a learning mechanism that imitates the neural network model in the human cerebral cortex, and the neurons transfer signals by layers. The structure with an input layer, one or multiple layers, and an output layer and uses back-propagation to learn from the previous errors. The hidden layer sizes are defined as the number of hidden neurons in the hidden layer. Less hidden layer sizes lead to underfitting and high bias, and more hidden layer sizes lead to overfitting and high variance. Three steps of the algorithm process, include creating the input layer and the hidden layer, recalculating after passing hidden layers, and creating the output layer.

C. Ensemble Classifiers

Hard voting is to use different models and aggregates the prediction results of these models for the major vote. The method needs to ensure that each model is independent and that the errors are irrelevant to achieve the best performance. Soft voting calculates the average probability of each model and ensures that all classifiers can estimate the probability of the category. Weighted soft voting is an approach to assign different weights on averaging, which depends on the importance of a model to the final prediction. Stacking shorts for stacked generalization, also called meta ensembling. Compared to bagging and boosting, stacking usually combines heterogeneous weak learners, which are different learning algorithms. The stacking method is a layered model integration framework containing a few layers, in which the meta-learner in level 1 learns the output features of the base learner in level 0 and improves the prediction effect. The meta-learner provides an interpretation of the prediction for the base learners. For regression tasks that predict a numeric value using linear regression as the meta-learner. Classification tasks that predict a class label using logistic regression as the meta-learner.

Bagging refers to Bootstrap Aggregating. The method is to use sampling with replacement and build the models for the subsets. Although the bias of each sub-model is low, the individual variance is high and prone to overfit. The bagging approach which is executed in parallel uses Bootstrap to produce training samples and feed them to a weak learner. The final classification result is produced by a majority vote and aggregates the weak learner output that can reduce the variance. Due to the similarity of the subset and using the same model, the based learners are independent and identically distributed. AdaBoost is aimed to strengthen the samples misclassified by the previous basic classifier and trains the next basic classifier

after the weighted sample. A weak classifier will add in each time, and end up with the best pair with a small error rate or the maximum number of sequential iteration. The repeat iterations reduce the bias of the model.

IV. The Experiment Results and Discussion

A. WM-811K Dataset

The public semiconductor dataset, WM-811K (LSWMD) from Kaggle [23], contains huge real-world wafer maps. With 811,457 wafer maps that provide information about the wafer map, die size, lot name, training or testing label, and failure pattern type. The dataset is divided into training and testing sets, and the defect types have been labeled by domain experts [24]. The dataset consists of 172,950 labeled wafers and 638,507 unlabeled wafers. There are only 25,519 wafer maps with failure patterns and 147,431 wafer maps with no failure pattern among the label wafers.

B. Classification with Individual Classifiers

Individual classifiers include KNN, SVC, CategoricalNB, DTs, and MLP. The parameter that chooses to fine-tune is the most influential one in the classifier, and the result will be different depending on the parameter. Feature fusion combined density-based feature, radon-based feature (mean), radon-based feature (std), and geom-based feature. In Table 1, the accuracy of the KNN model is 83%, where the parameter setting is based on the radon-based feature (mean) and the number of neighbors is 6. The accuracy of the SVC model is 82%, where the parameter settings are fusion features and gamma is 0.001. The accuracy of the CategoricalNB model is 62%, where the parameter setting is based on density features. The accuracy of the DTs model is 85%, where the parameter settings are fusion features and the max depth is 14. The accuracy of the MLP model is 86%, where the parameter settings are fusion features and the hidden layer size is 150.

Table 1 The Experiment Results with Individual Classifiers

Model	Feature	parameter	Accuracy	F1-Score
KNN	radon	n=6	0.83	0.82
SVC	fusion	gamma=0.001	0.82	0.80
CategoricalNB	density	default	0.62	0.62
DTs	fusion	max depth=14	0.85	0.85
MLP	**fusion**	**hidden layer=150**	**0.86**	**0.86**

C. Classification with Emsemble Learning

Ensemble classifiers include hard voting, soft voting, stacking, bagging, and AdaBoost. In Table 2, the accuracy of Hard Voting is 0.50%, where the parameter setting is based on feature fusion and SVC+DT+MLP model. The accuracy of Weighted Soft Voting is 0.95%, where the parameter setting is based on feature fusion and KNN+DT+MLP model. The accuracy of Stacking is 0.67%, where the parameter setting is based on the radon-based feature (mean) and KNN+DT+MLP model. The accuracy of Bagging is 0.83%, where the parameter setting is based on feature fusion and KNN model. The accuracy of AdaBoost is 0.62%, where the parameter setting is based on feature fusion and CategoricalNB model.

Table 2 The Experiment Results with Emsemble Learning

Mothod	Feature	Classifier	Accuracy	F1-Score
Hard Voting	fusion	SVC+DT+MLP	0.50	0.47
Weighted Soft Voting	**fusion**	**KNN+DT+MLP**	**0.95**	**0.95**
Stacking	radon	KNN+DT+MLP	0.67	0.62
Bagging	fusion	KNN	0.83	0.82
AdaBoost	fusion	CategoricalNB	0.62	0.60

V. Conclusions

As the advanced technique of finding the wafer map defect, it is crucial to discover how to use the DL method to improve the accuracy and increase the yield rate. The result may lead to bias due to the unequal selection and imbalance problem. We present the comparison of multiple classification models on wafer defect pattern classification. The research focuses on using voting, bagging, and Adaboost to solve the problem. Among them, the ensemble method using Weighted Soft Voting achieves the best performance, with an accuracy of 95%.

Acknowledgment

This work was supported by the Slovak Research and Development Agency (APVV) under contract no. SK-TW-21-0002 (MOST-SRDA contract no. 111-2927-I-031-501), by Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under the research projects VEGA 1/0753/20 & VEGA 2/0165/21 and by National Science and Technology Council of Taiwan under the research project: Technologies to Support Response Generation for Multilingual Intelligent Agent.

References

[1] H. P. Rapp, J. Möbert, S. Schneider, & D. B. AG, Extraordinary semiconductor cycle triggered by one-time events, cyclical and geopolitical effects. Germany Monitor. Deutsche Bank Research, 2022.

[2] F. OLIVIERI, A secondary sources analysis of the business models and supply chain strategies of semiconductor manufacturers. Master of Science thesis, School of Industrial and Information Engineering. 2021.

[3] I. L. Y. B. TARMIZI, Design and Fabrication of Semi-Auto Wafer Scriber for 4-in and 8-in Wafers. Doctoral dissertation, Universiti Sains Malaysia. 2018.

[4] S. Manivannan, "An ensemble-based deep semi-supervised learning for the classification of Wafer Bin Maps defect patterns," *Computers & Industrial Engineering*, vol. 172, pp. 108614, October 2022.

[5] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, & F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11-26, April 2017.

[6] D. Svozil, V. Kvasnicka, & J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and intelligent laboratory systems*, vol. 39, no. 1, pp. 43-62, November 1997.

[7] S. Albawi, T. A. Mohammed, & S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, pp. 1-6, August 2017.

[8] M. Kaur, & A. Mohta, "A review of deep learning with recurrent neural network," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 460-465, November 2019.

[9] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, pp. 1883, 2009.

[10] K. Taunk, S. De, S. Verma, & A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1255-1260, May 2019.

[11] T. Pan, J. Yang, W. Wu, Z. Dong, & B. O. Ayinde, "Artificial intelligent matching for scratches of semiconductor wafers based on a K-NN algorithm," *Surface Topography: Metrology and Properties*, vol. 7, no. 2, pp. 025012, May 2019.

[12] B. Charbuty, & A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20-28, March 2021.

[13] M. Piao, C. H. Jin, J. Y. Lee, & J. Y. Byun, "Decision tree ensemble-based wafer map failure pattern recognition based on radon transform-based features," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 2, pp. 250-257, May 2018.

[14] M. P. L. Ooi, H. K. Sok, Y. C. Kuang, S. Demidenko, & C. Chan, "Defect cluster recognition system for fabricated semiconductor wafers," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 3, pp. 1029-1043, March 2013.

[15] S. D. Jadhav, & H. P. Channe, "Comparative study of K-NN, naive Bayes and decision tree classification techniques," *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842-1845, January 2016.

[16] S. Suthaharan, & S. Suthaharan. Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, pp. 207-235, 2016.

[17] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, & A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189-215, September 2020.

[18] J. L. Gómez-Sirvent, F. L. de la Rosa, R. Sánchez-Reolid, A. Fernández-Caballero, & R. Morales, "Optimal feature selection for defect classification in semiconductor wafers," *IEEE Transactions on Semiconductor Manufacturing*, vol. 35, no. 2, pp. 324-331, January 2022.

[19] E. E. Schadt, C. Li, B. Ellis and W. H. Wong, "Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data," *Journal of Cellular Biochemistry*, vol. 84, no. S37, pp. 120–125, 2001.

[20] R.C. Gonzalez & R. E. Woods, *Digital Image Processing*, 3rd ed. Harlow, U.K.:Prentice-Hall, 2008.

[21] R. M. Haralock & L. G. Shapiro, "Computer vision: Overview," in *Computer and Robot Vision,* California, USA: Addison-Wesley Longman Publishing Co., Inc., 1992.

[22] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).

[23] X. Han, X. Jia, D. Y. Ji, & J. Lee, "Designing Robust Topological Features for Wafer Map Pattern Classification," in *the 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pp. 1-6, May 2023.

[24] Wu, M. J., Jang, J. S. R., & Chen, J. L. (2014). Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing, 28*(1), 1-12.

QUESTION ANSWERING SYSTEM BASED ON PRE-TRAINING MODEL AND RETRIEVAL RERANKING FOR INDUSTRY 4.0

Ta-Fu Chen[1], Yi-Xing Lin[1], Ming-Hsiang Su[2], Po-Kai Chen[1], Tzu-Chiang Tai[3], Jia-Ching Wang[1]

[1]Department of CSIE, National Central University, Taoyuan City, Taiwan
[2]Department of Data Science, Soochow University, Taipei City, Taiwan
[3]Department of CSIE, Providence University, Taichung, Taiwan

ABSTRACT

By providing enhanced knowledge retrieval capabilities, real-time decision support, and efficient information exchange, question answering (QA) systems play a crucial role in driving productivity, efficiency, and innovation in Industry 4.0. Today's most reliable knowledge-based QA systems require a large knowledge base, which tends to consume more reasoning time. In order to improve the inference speed and response accuracy of the system, this paper adds a Reranker between the Retriever and Reader of the traditional two-stage mechanism. This study uses pretrained Roberta to perform system retrieval and improve data processing and training methods. Experiments on Chinese Wikipedia show that the proposed system significantly reduces the system response time and improves the accuracy and scope of the response.

Index Terms— QA system, Pre-training model, Reranking

1. INTRODUCTION

Question answering (QA) systems fulfill the demands in Industry 4.0 with the following features: (1) Enabling rapid and accurate retrieval of relevant information from vast amounts of structured and unstructured data sources. (2) Providing real-time access to up-to-date information, enabling decision-makers to make timely and data-driven decisions. (3) Enabling effective communication and collaboration by providing a common platform for sharing knowledge, best practices, and troubleshooting information. Thus, QA systems is important to Industry 4.0.

QA system has been one of the most popular fields of natural language processing recently. Although the large-scale pre-training language model brings strong comprehension and answer extraction capabilities to the question answering system, its huge number of parameters result in long inference time. Inspired by RE^{3}QA [1], and IRRR [2], the proposed QA system has three stage architecture: Retriever, Reader, and Reranking. Different from previous works, this paper locates Reranking mechanism between Retriever and Reader. This approach would greatly reduce the number of paragraphs which need to be input into the QA model which greatly improves the response speed. It also reduces the possibility of the QA model being misled due to the large text content searched. Therefore, the proposed system can achieve better performance in experiments.

In addition, in practical applications, the input in various types such as spoken input cannot be included in the training data. It is easy to cause problems in the inference process and training gap. In order to provide model with better performance which can deal with diverse input, this paper introduces WebQA [3] dataset for training. The proposed system also adjusts and combines existing datasets in pre-processing, and adjusts the sample length during training, so as to further improve the adaptability of the system to different types of questions.

The contributions of this paper are as follows: First, the proposed system searches for answers more precisely and makes the answers more accurate. Second, the maximum system response time is significantly reduced, 30% less than the system using only TF-IDF without Reranker mechanism. Third, the proposed system response exact match (EM) score improves by 3~4% compared to traditional methods. Finally, the generalization ability of the model is improved.

2. PROPOSED METHOD

This paper proposes a Chinese question answering system with high accuracy and fast response. The architecture is shown in Fig. 1. The overall architecture can be divided into three parts: Retriever, Reranker and Reader.

2.1. Retriever

The goal of Retriever is to quickly find documents that are highly relevant to the question in a huge document database, so this study uses the TF-IDF algorithm to achieve this goal. Even though the Retriever using the TF-IDF algorithm has some shortcomings, it is still used in the Retriever tasks. The reason is that it is fast and the computational costs is not high after the statistic table of TF-IDF is constructed. The TF-IDF algorithm can easily find the document with the special keywords in the absence of Chinese corpus. To implement TF-IDF, a TF-IDF matrix must be created for all documents in the database, and a TF-IDF vector must be created for questions. This study first segmented all sentences, and then established bigrams that outperformed unary in experiments. Then this study uses MurmurHash3 to map these bigrams to

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

2^{24} hash fields, and finally this study uses TF-IDF for calculation. The reason why hash is used is that it is of great help in accelerating acceleration and saving memory.

2.2. Reader

After dividing the paragraphs, the reader needs to understand each paragraph and find the answer based on the question. This study uses the reading comprehension model of extractive question answering for Reader. The reason is that the document content in the database is reliable, and the model burden is smaller and the speed is fast enough even if it is slower than traditional model such as RNN. Under the same model specification, the answer can be given more quickly than the non-extractive question answering. In view of the powerful capabilities of large-scale pre-training language models, this paper adopts the pretrained Roberta [4] as the main model for Reader and the pre-training parameters is from Cui et al. [5]. The system architecture is shown in Fig. 1. A fully connected layer is added to the output of the encoder, and outputs the probability of the start index and the end index of the answer fragment. Each word in the input sequence will output these two probabilities. This study then uses (1) to multiply the output probability of the starting position and the ending position for all words, and take the index pair corresponding to the maximum value of the product as the start index and the end index.

$$P_{start}(i) * P_{end}(j), \text{ where } 0 \leq i, j \leq W \qquad (1)$$

where i, j are token indexes in the input sequence, $P_{start}(.)$ and $P_{end}(.)$ denote probability of the start position and the end position, respectively, and W is number of words in the input sequence. Since the proposed goal is to establish a Chinese question answering system, English vocabulary and letters or symbols of other languages in the Chinese training corpus are regarded as specific keywords. For such keywords, this study disassembles each non-Chinese word individually into the smallest unit, so as to force the model to understand that this type of text will be considered as keywords. The same way is used for Arabic numerals.

2.3. Reranker

In this paper, Retriever using the TF-IDF algorithm is very fast and has a certain degree of accuracy, but it lacks semantic information and cannot distinguish the meaning of the same word in different sentences. Therefore, we propose to add Reranker for re-scoring mechanism between Retriever and Reader in the system. This study can filter out the paragraphs where the answer is more likely to be located, according to the semantic meaning of the documents output by Retriever. This approach also greatly reduces the problem of slow reading caused by Reader using a large language model as the main body. Reranker in this paper use retrieval model with dual-encoder architecture. Both encoders use the pretrained Roberta model, which saves a lot of training time. Similar to the training methods of DPR and Siamese neural network, the independent encoders is to filter out paragraphs that are high relevant to the question. This approach achieves the transition from document-level filtering to paragraph-level filtering which greatly reduce the time of reading a large number of unrelated documents, and is also reduce the probability of Reader being misled. At the same time, the processing time for each question is controlled within a certain range. Reducing the number of paragraphs solves the problem of long and unfixed reading comprehension time due to large language models.

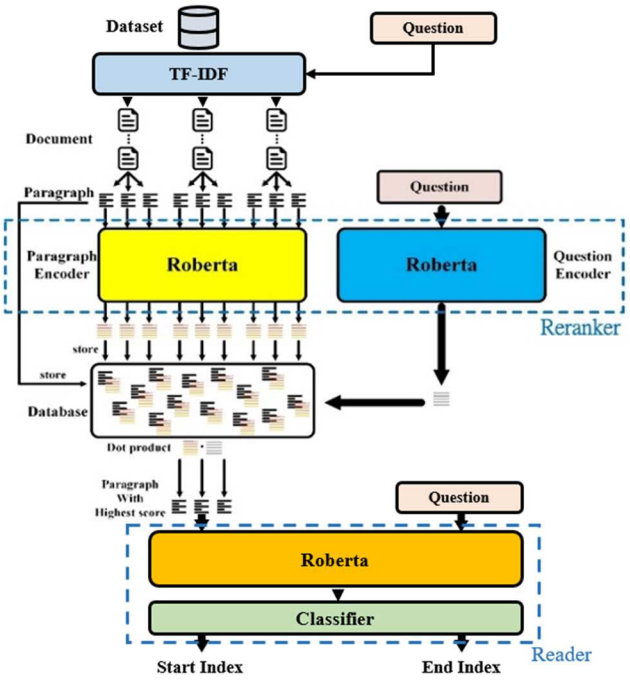


Fig. 1. Architecture of the proposed system.

2.4. Inverse cloze task training

Due to the dataset for training Reranker is the same as the dataset for training Reader, it was found that the data distribution learned by the two models is similar in the experiment. Although the system responds much faster after adding Reranker, the accuracy of answer is not improved. In order to further enhance the system capabilities, we need to expand the data distribution which is used to train Reranker and Reader, so as to help Reader more effectively. Wikipedia is used in the proposed system as an unlabeled document. Before training with the reading comprehension dataset, Reranker is pre-trained on the Inverse Cloze Task [6]. The Inverse Cloze Task is originally used to train Retriever and can increase the model's sensitivity and distinguishing ability on the correlation between sentences and paragraphs.

This study divides each document into paragraphs, and randomly selects sentences with 5 or more words from the paragraphs as the excavated sentence. Then, the excavated sentence and the corresponding excavated paragraph are formed as a pair. The question encoder and paragraph encoder of the dual encoders in Reranker is used to encode the excavated sentence and the excavated paragraph, respectively. During training, an excavated sentence will be trained together with its corresponding excavated paragraph and excavated paragraphs of other sentences. The model is

trained with Cross Entropy as the loss function to select the corresponding excavated paragraphs.

3. EXPERIMENT

This study uses AdamW with 2e-5 learning rate and 1e-4 weight decay. During training, the batch size is 4. This study trains for 500k steps, adopts a linear scheduler as the warm-up scheduler, and sets the number of warm-up steps to 25k.

3.1. Data

The document database used in the experiment is the Chinese Wikipedia database released in 2021. This study uses WikiExtractor [7] and extracts the title and content from it. Since the Chinese Wikipedia database is a mixture of traditional and simplified Chinese, this study uses the Opencc suite [8] to integrate all simplified and traditional Chinese characters, and then converts all punctuation marks to full-type fonts. Finally, we use the sqlite3 suite [9] to create the database and use the jieba suite [10] for tokenization. Regarding to the training dataset, in addition to using the DRCD dataset [11] and the CMRC dataset [12], this study has added WebQA [3] as the extended data. The reason is that all questions in WebQA are from real questions raised by real users. Thus, the questions are more consistent with the actual application situation. The paragraphs of the documents that reply to the question can be regarded as from the knowledge database on the Internet. That is, the answer only appears after the question. This feature is exactly the opposite of the two original datasets. In WebQA dataset, this study removes a small number of fill-in-the-blank questions and the tautology in the question and answer sentence is also removed, such as "question:", "answer:". In order to improve the generalization ability of the model, this study adopts the strategy of preserving punctuation marks and discarding samples without question marks when processing training data. After removing duplicate title paragraphs, the average paragraph length on Chinese Wikipedia is about 95 words. Therefore, the training strategy used in this study is to use question sentences and paragraphs with a total length of less than 256 words for 80% of the training steps. For the remaining 20% of training steps, train with question sentences and paragraphs with a total length of less than 512 words. Such an approach can improve the answering ability for documents of average length and consider the answering ability for long documents.

3.2. System accuracy improvement

This study first evaluated the improvement in system accuracy after adding Reranker. Exact match (EM) score was used as evaluation criteria. The results are shown in Table 1. After adding Reranker, the accuracy of the system increased by 3~4% on average. In addition to the capabilities of Reranker and Reader, the performance of question answering systems is also affected by Retriever tokenization. The score can be improved if the accuracy of the word segmentation and the defined stop words are more refined.

3.3. Effect of ICT training on Reranker

Then, this study validated the effect of the reverse cloze task (ICT) training. After Retriever selects the top five documents, this study uses Reranker for paragraph-level filtering and Top-K as the evaluation criterion. The results are shown in Table 2. This study verified that similarity filtering between question and answer passages is improved by Reranker after expanding learning data distribution for Reranker.

Table 1. EM-score results for Reranker evaluation.

	DRCD	CMRC
Retriever + Reader	33.2	26.7
Retriever + Reranker + Reader	**37.5**	**30.0**

Table 2. Top-K results for ICT training evaluation.

	Top-5	Top-20	Top-100
Reranker	40.6	75.2	86.0
Reranker+ICT	**43.4**	**77.9**	**87.8**

3.4. System response time improvement

In this experiment, this study evaluated the improvement of system response time after adding Reranker. When the system is implemented without Reranker, the paragraphs entering the Reader are all paragraphs in the top five documents found by Retriever. So, the number of paragraphs entering the Reader is uncontrollable. Sometimes there are thousands of paragraphs entering the Reader in the experiments, and the system takes more than 30 seconds to response. The advantage of adding Reranker is that the number of paragraphs received by Reader is controllable. This study seted different restrictions on the number of paragraphs for Reranker, and measure the system response time. The results are shown in Table 3. Note that in Table 3, 512 is the number of paragraphs received by Reader for most systems without Reranker. Subsection 3.3 shows that Reranker can output 87% of the paragraphs containing the correct answer under the conditions of Top-100. If the output of Reranker is limited to 128 paragraphs, we can ensure that the paragraphs received by the Reader have a high probability of containing the correct answer. Besides, this can also save more than 3 times the system response time. The paragraph restriction in Table 3 is the maximum number of paragraphs. If the number of paragraphs for the Reranker is less than the maximum number, the required system response time will be less than the value in Table 3.

Table 3. System response time for Reranker evaluation.

Reranker	**w/o**	**w/**	**w/**	**w/**	**w/**	**w/**
Paragraph restriction	512*	128	64	32	16	8
Response time (seconds)	15.4	4.60	3.02	2.26	2.01	1.83

3.5 Evaluation of dataset transfer performance

In order to test the performance of the system in different datasets and test the adaptability of the training results of the

Table 4. Results of datasets-transfer-analysis.

	DRCD			CMRC			WebQA				
	train	**dev**	**test**	**train**	**dev**	**test**	**train**	**dev A**	**test A**	**dev B**	**test B**
DRCD	97.0	78.6	**76.0**	34.7	54.0	20.4	61.6	69.8	70.0	68.7	70.3
CMRC	35.8	37.6	37.5	80.1	62.3	**51.2**	33.6	32.7	35.4	34.9	34.7
WebQA	37.6	38.1	36.3	15.5	27.8	8.1	96.6	78.2	**79.9**	78.8	**78.6**

system under different datasets, this study trained and tested the proposed system in DRCD, CMRC and WebQA datasets. The experiments used EM score as the evaluation criteria. Table 4 has shown the experimental results. the testing results on CMRC were generally unsatisfactory. We speculated there are mainly two reasons. First, CMRC dataset is originally constructed on simplified Chinese, but the proposed system was based on traditional Chinese, so we need to convert the simplified Chinese to traditional Chinese first, and then transformed into dictionary code, which increase the appearance possibility of the token [UNK]. Therefore, in order to prevent it from having too much influence on the performance, this study removed the paragraphs and questions with more than two [UNK]. The second reason is that the paragraphs of CMRC were usually longer than other datasets, after removing the paragraphs with more than 512 length, the size of the dataset was greatly decreased, which maked the training difficult. The performance on DRCD dataset was normal, but it was worth noting that when using DRCD dataset as the training set, the model performed well in WebQA dataset while it did not perform well in DRCD dataset when WebQA dataset was used as the training set. This study speculated that since the text lengths in WebQA datasets were usually short, it was difficult to learn the features of long text samples.

4. CONCLUSION

Question answering systems play a vital role in Industry 4.0. This study proposes a three-stage Chinese extractive question answering system. Reranking mechanism is added between Retriever and Reader. In addition to providing information judgment at the semantic level, the Reranking mechanism can effectively improve the system speed due to the use of less hardware resources. Besides, the proposed system also controls the processing time of each question to a fixed time. This study attempts to add datasets and process training data to improve the model's ability to generalize to spoken habits. Although its performance has improved in the experiment, there is no significant difference in numerical value. This research will continue with further research on the practical application scenario experience of open-domain question answering systems, including generalization, performance, and anthropomorphic responses.

REFERENCES

[1] M. Hu, Y. Peng, Z. Huang, and D. Li, "Retrieve, Read, Rerank: Towards End-to-End Multi-Document Reading Comprehension," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 2285-2295, 2019.

[2] P. Qi, H. Lee, O. Sido, C. D. Manning, "Answering Open-Domain Questions of Varying Reasoning Steps from Text," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 3599-3614, 2021.

[3] P. Li, W. Li, Z. He, X. Wang, Y. Cao, J. Zhou, and W. Xu, "Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering," *arXiv preprint* arXiv:1607.06275, 2016, doi: https://doi.org/10.48550/arXiv.1607.06275.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *Proceedings of International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia, 2020.

[5] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-Training with Whole Word Masking for Chinese BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 3504-3514, November 2021.

[6] K. Lee, M. W. Chang, and K. Toutanova, "Latent Retrieval for Weakly Supervised Open Domain Question Answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 6086-6096, 2019.

[7] G. Attardi, WikiExtractor, 2015, [online] Available: https://github.com/attardi/wikiextractor.

[8] X. Yang, X. Zhao, G. Tjio, C. Chen, L. Wang, B. Wen, and Y. Su, "Opencc–an open Benchmark data set for Corpus Callosum Segmentation and Evaluation," in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, Abu Dhabi, pp. 3020-3024, 2020.

[9] X. Liu, X. Fu, and G. Sun, "Recovery of deleted record for SQLite3 database," in *Proceedings of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, IEEE, Hangzhou, China, pp. 183-187, 2016.

[10] Y. Ding, F. Teng, P. Zhang, X. Huo, Q. Sun, and Y. Qi, "Research on Text Information Mining Technology of Substation Inspection Based on Improved Jieba," in *Proceedings of 2021 International Conference on Wireless Communications and Smart Grid (ICWCSG)*, IEEE, Hangzhou, China, pp. 561-564, 2021.

[11] C. C. Shao, T. Liu, Y. Lai, Y. Tseng, and S. Tsai, "Drcd: a chinese machine reading comprehension dataset," *arXiv preprint* arXiv:1806.00920, 2018, doi: https://doi.org/10.48550/arXiv.1806.00920.

[12] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang, and G. Hu, "A span-extraction dataset for Chinese machine reading comprehension," *arXiv preprint* arXiv:1810.07366, 2018, doi: https://doi.org/10.48550/arXiv.1810.07366.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Deepfake-speech Detection with Pathological Features and Multilayer Perceptron Neural Network

Anuwat Chaiwongyen*†, Suradej Duangpummet‡,
Jessada Karnjana‡, Waree Kongprawechnon†, and Masashi Unoki*
* Japan Advanced Institute of Science and Technology, Ishikawa, Japan
† Sirindhorn International Institute of Technology, Thammasat University, Pathumthani, Thailand
‡ NECTEC, National Science and Technology Development Agency, Pathumthani, Thailand
E-mail: {anuwat, unoki}@jaist.ac.jp, {suradej.dua, jessada.kar}@nectec.or.th, waree@siit.tu.ac.th

***Abstract*—Deepfake speech, a misuse of speech technology, is of great concern since it seems natural and is difficult to detect. Although many methods using various speech features have been proposed, deepfake-speech detection accuracy must be improved, especially in real-world scenarios. Therefore, this paper presents a method for detecting deepfake speech on the basis of pathological features used by pathologists for assessing voice quality. The six-pathological features, including jitter, shimmer, harmonics-to-noise ratio, cepstral-harmonics-to-noise ratio, normalized noise energy, and glottal-to-noise excitation ratio, are fed to a multilayer perceptron neural network. We evaluated the proposed method using the Audio Deep Synthesis Detection Challenge dataset. The results indicate that the proposed model can be used for detecting deepfake speech. The proposed method's accuracy, precision, recall, and F1-score were over 98% on the development set, and it outperformed the baseline method on the adaptation set.**

I. Introduction

Deepfake speech is a misuse of speech technologies, such as voice conversion or text-to-speech techniques [1], [2], to synthesize fake speech. Since deepfake speech seems natural and is challenging to detect, it poses a significant threat to economies and societies. For example, criminals used deepfake speech to impersonate a CEO's voice and successfully swindled over USD $243,000$ [3]. Automatic speaker verification to authenticate personal voice is also vulnerable to such attack [4].

Several methods have been proposed to detect deepfake speech [5]–[9]. Many classifiers have been utilized, such as the Gaussian mixture model (GMM), deep neural networks (DNNs) [10], recurrent neural networks (RNNs) [11], convolution neural networks (CNN) [12], and residual neural network (ResNet) [13]. Also, many speech features, including spectrograms, linear-frequency cepstral coefficients (LFCCs) [14], mel-frequency cepstral coefficients [15], constant-Q transform [16], and constant-Q cepstral coefficients [17], have been used. For example, Yi [18] and Wang [19] independently proposed a method based on GMM that takes LFCCs as the input feature [20].

These features are represented in phase, power spectrum, and cepstral coefficients. However, to the best of our knowledge, there is a lack of attention on pathological features in deepfake-speech detection. Since pathological features are widely used to analyze voice disorders, we hypothesize that the voice quality of deepfake speech can be considered as a disordered voice. Hence, pathological features might be clues for deepfake-speech detection.

We, therefore, investigated six pathological features, including jitter, shimmer, harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio (CHNR), normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE). These features are fed to a classifier for detecting deepfake speech.

The rest of this paper is organized as follows. Section 2 briefly describes the pathological features mentioned above. Section 3 presents the proposed method. Sections 4 and 5 present the evaluation of the proposed method, results, and discussion. Finally, Section 6 summarizes this work.

II. Pathological Features

Pathological features can be used to distinguish between normal and pathological voices [21] and diagnose diseases such as Parkinson's disease [22] neck and head cancers [23]. This study investigated whether the following pathological features can be used to recognize the deepfake-speech signal.

A. Jitter features

Jitter is the measure of the cycle-to-cycle variations of the fundamental frequency [24], [25]. As the characteristics of jitter can be identified by several methods, this work focuses on four types as follows.

1) Jitter (local): Jitter ($local$) is the average absolute difference between consecutive periods divided by the average period, that is:

$$\text{Jitter } (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100, \qquad (1)$$

where T_i represents the extracted f_0 period lengths, and N is the number of extracted f_0 periods [25].

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

2) Jitter (rap): Jitter (rap) is the average absolute difference between a period of its average and its two neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (rap) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - (\frac{1}{3}\sum_{i=i-1}^{i+1} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \quad (2)$$

3) Jitter ($ppq5$): Jitter ($ppq5$) is the average absolute difference between a period and its average and its five closest neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (ppq5) = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|T_i - (\frac{1}{5}\sum_{i=i-2}^{i+2} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \quad (3)$$

4) Jitter ($ppq55$): Jitter ($ppq55$) is the average absolute difference between a period and its average and its 55 closest neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (ppq55) = \frac{\frac{1}{N-1}\sum_{i=27}^{N-27}|T_i - (\frac{1}{55}\sum_{i=i-27}^{i+27} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \quad (4)$$

B. Shimmer features

Shimmer is a variation of a signal in amplitude that results from irregular vocal fold vibrations. There are various ways to identify shimmer characteristics. We focused on two types of shimmer features.

1) Shim ($local$): Shim ($local$) refers to the average of absolute differences between the source-signal-amplitude-related in each index (A_i) and its next neighbor (A_{i+1}), divided by the average of the signal amplitudes. It is defined as:

$$\text{Shim } (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \quad (5)$$

where N is the number of fundamental frequency periods, and A_i denotes the signal amplitude at index i.

2) Shim (x-point amplitude perturbation quotients): Shim x-point amplitude perturbation quotients (APQx) is also defined similarly as Shim ($local$). However, it considers the absolute difference between the amplitude of each index and its $x-1$ closest neighbors. It is defined as:

$$\text{Shim } (\text{APQ}x) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - (\frac{1}{x}\sum_{n=i-m}^{i+m} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \quad (6)$$

where $m = \frac{x-1}{2}$.

C. Harmonics-to-noise ratio (HNR)

The HNR is a measure of the proportion of the harmonic and noise components of speech. The noise (ι_{En}) is computed as the energy of the residual produced after subtracting the average waveform from each individual cycle. The harmonic energy (γ_{En}) is determined as the energy of an average waveform of a frame pitch built synchronously around ten consecutive glottal cycles. Hence, this feature requires an earlier f_0 estimation [26]. The HNR is defined as:

$$\text{HNR} = 20\log\frac{\gamma_{En}}{\iota_{En}}. \quad (7)$$

The HNR is calculated for each frame of analysis. The output HNR is the average of each frame.

D. Cepstral-harmonics-to-noise ratio (CHNR)

The CHNR is used to calculate HNR as the level difference between the total energy of the spectrum and noise energy, with the noise component being considered as the energy that cannot be related to the spectrum of the original signal [26]. The CHNR-calculation process is illustrated in Fig. 1.

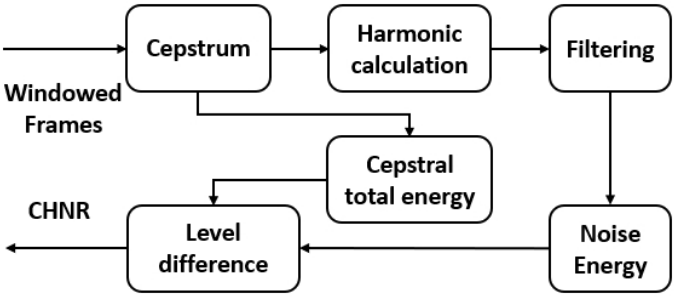


Fig. 1. CHNR calculation process [26].

E. Normalized noise energy (NNE)

NNE is another feature used to quantify the amount of additive noise, and it is defined as the ratio of the energy of the noise to the total energy of the signal for each frame of analysis [26]. The NNE-calculation process is illustrated in Fig. 2.

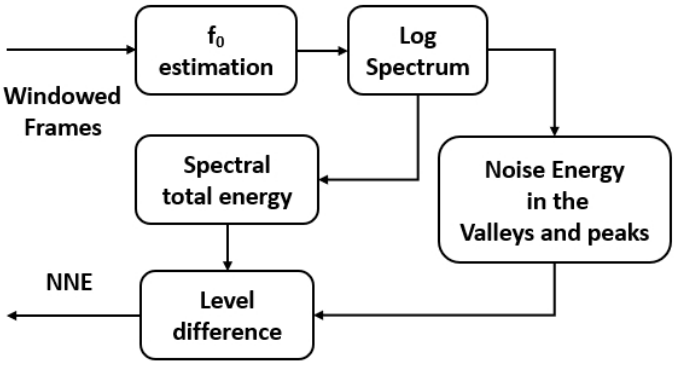


Fig. 2. NNE calculation process [26].

F. Glottal-to-noise excitation ratio (GNE)

The GNE is used to describe turbulent noise while disregarding modulation effects [27]. It is assumed that glottal pulses produce a simultaneous and synchronous excitation of multiple frequency channels, which is indicated by the correlation between Hilbert envelopes of multiple frequency bands [26]. The GNE-calculation process is illustrated in Fig. 3.

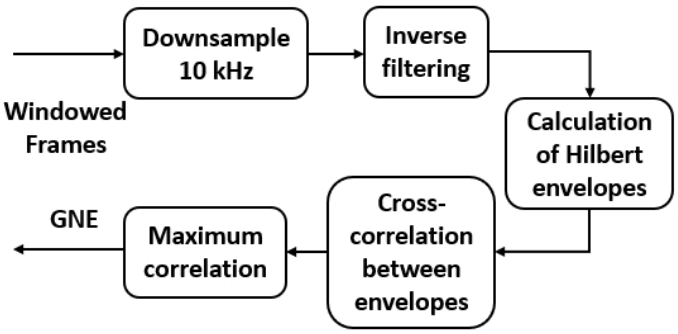


Fig. 3. GNE calculation process [26].

III. Proposed Method

The proposed method detects deepfake speech using a multilayer perceptron (MLP) neural network that takes a combination of the above pathological features as its input, as shown in Fig. 5.

A. Feature analyses

We first investigated those features that can be used to distinguish between genuine and fake speech signals. We conducted an experiment by using the dataset from the Audio Deep Synthesis Detection (ADD) 2022 Challenge [18].

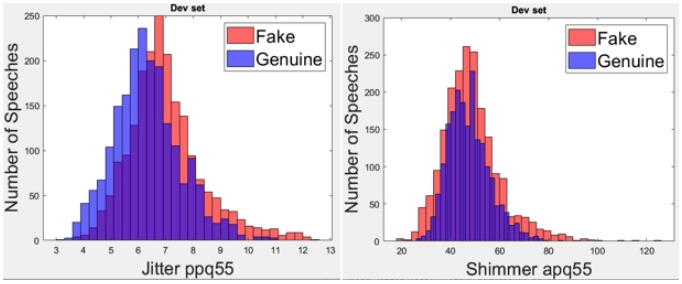


Fig. 4. Histograms of Jitter ($apq55$) and Shim (apq55) in the development set.

Figure 4 shows the histograms of the difference between genuine and fake speech signals for each pathological feature that is not useful in detecting fake speech signals. We randomly selected $2,000$ samples of both genuine and fake speech signals.

Figure 6 shows histograms of the difference between genuine and fake speech signals. for each pathological feature used for detecting fake speech in this paper.

For distinguishing genuine and fake speech, we found that Jitter ($local$), Jitter (rap), Jitter ($ppq5$), CHNR, NNE, and GNE help discriminate between genuine and fake speech signals.

Shim ($local$), Shim (APQ3), Shim (APQ5), Shim (APQ11), and HNR are less effective in discriminating between genuine and fake speech signals, as shown in Fig. 6.

On the other hand, it was found that Jitter ($ppq55$) and Shim (APQ55) are unsuitable for the discrimination because their genuine and fake speech histograms overlap, as shown in Fig. 4.

TABLE I
Number of utterances in ADD 2022 dataset [18]

Dataset	Number of utterances		
	Fake	Genuine	Total
training set	24,072	3,012	27,084
dev. set	26,017	2,307	28,324
adp. set	700	300	1,000

Therefore, we used Jitter ($local$), Jitter (rap), Jitter ($ppq5$), Shim ($local$), Shim (APQ3), Shim (APQ5), Shim (APQ11), HNR, CHNE, GNE, and NNE with an MLP neural network for detecting deepfake speech.

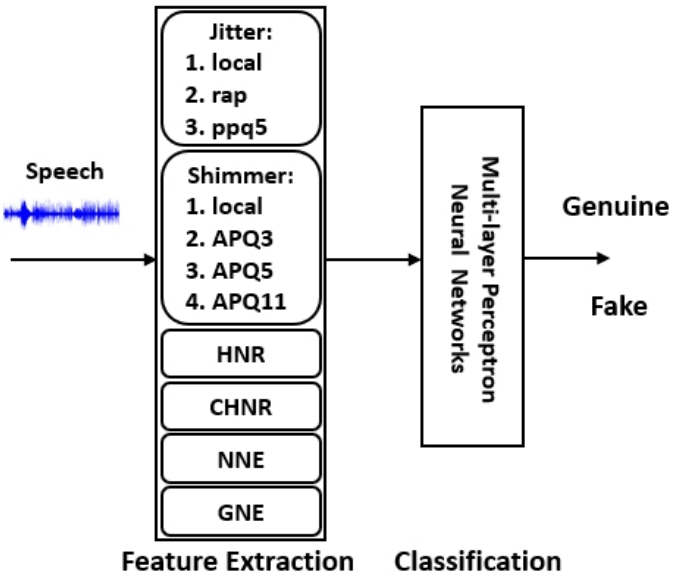


Fig. 5. Proposed method.

B. Experimental setup

The ADD 2022 dataset was used to evaluate the performance of the proposed method. The dataset is divided into three sets: training, development, and adaptation, as shown in Table I. From our observation, both training and development sets have a high signal-to-noise ratio (SNR), whereas the adaptation set has a low SNR with real-world background noise, such as background music and people chatting. In this research, our method was trained from the training set. We then evaluated the proposed method from the development and adaptation sets. For jitter and shimmer extraction, we used the perturbation analysis code implemented in MATLAB [28]. For HNR, CHNR, NNE, and GNE extraction, we used the AVCA-ByO MATLAB toolbox [26].

For classification, we applied the MLP neural network, 3 layers, 11 input features, 11 hidden-layer nodes, and one output layer. The number of training epochs was set to $1,000$. The model used an Adam optimizer, and a learning rate was set to 0.0001. The total number of parameters of our model were only 288.

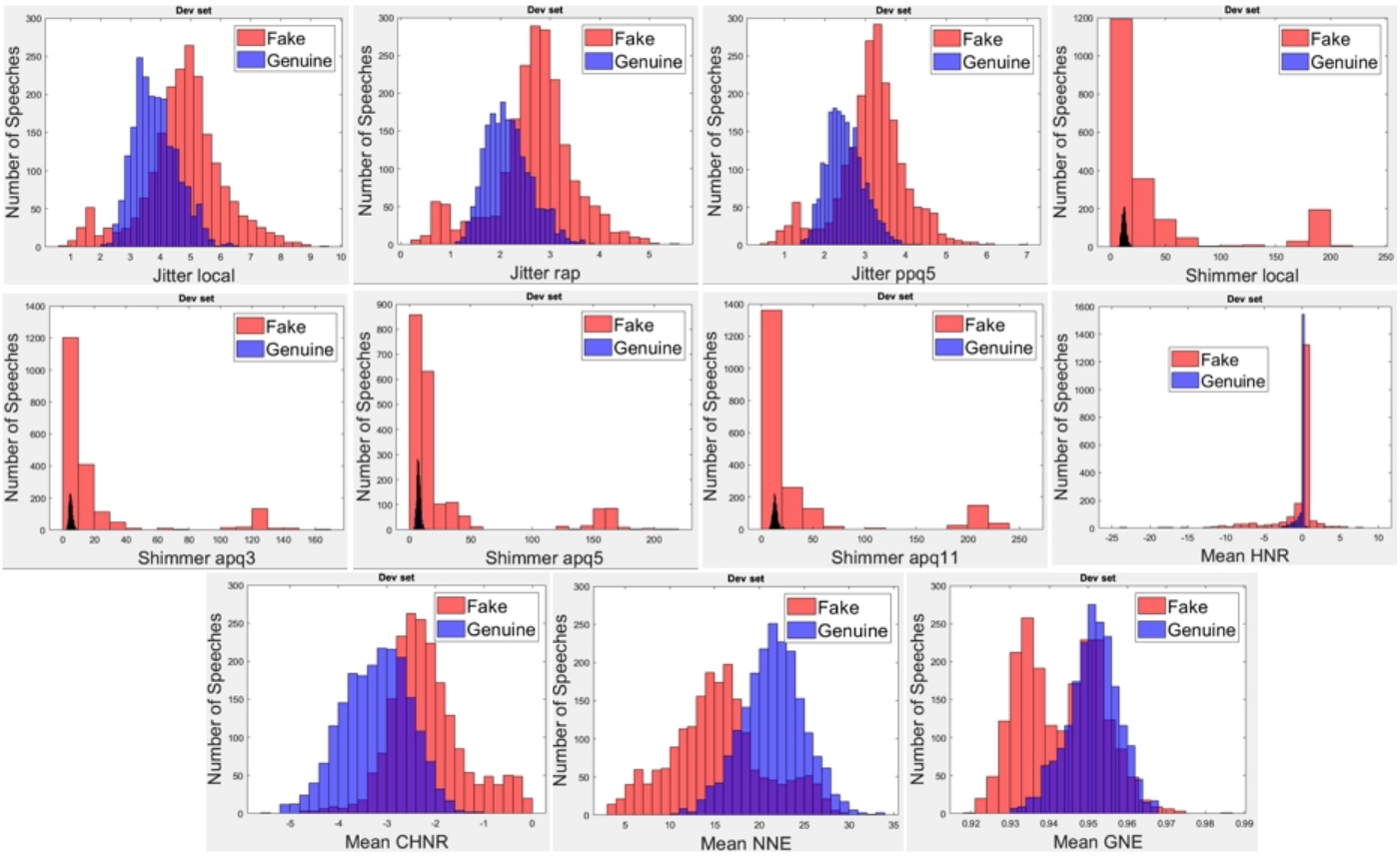


Fig. 6. Histograms of pathological features used in the development set.

IV. Results

We compared our method with two methods using the LFCC feature with two classifiers: GMM and CNN. The method using gammatone cepstral coefficients (GTCCs) with Resnet34 was also compared [29]. The results are shown in Table II. They indicate that even though the accuracy, precision, recall, and F1-score of the proposed method were slightly lower than those of LFCC with GMM, LFCC with CNN, and GTCC with ResNet34 in the development set, the accuracy, recall, and F1-score of our method were better than those of the three other methods in the adaptation sets. It means that the proposed method may be robust against speech signals with background noise, even though the proposed method has never been trained with speech signals with background noise before. The reason for this might be that the proposed method incorporates features such as GNE, HNR, and CHNR, which are related to the ratio between speech and noise, resulting in a particularly high recall rate of 97.14% and an F1-score of 85.16%

It can be seen from the table that using the pathological features alone is not good in terms of balanced accuracy. However, the proposed method can be improved further by incorporating these pathological features with others, e.g., features obtained from a CNN that takes LFCCs as its input, as shown in Fig. 7. The 11 pathological features and the flattened feature obtained from the CNN are combined and fed to an MLP neural network. The results indicate that the efficiency improved in both the development set and adaptation set, as shown by the fifth method of Table II. It can be concluded that the flattened feature from the CNN with LFCCs may marginally enhance the performance of the proposed method for the clean speech signal but considerably improve the balanced accuracy and precision for the noisy signal. However, such improvement costs a poorer recall.

V. Discussion

Although our method could successfully detect deepfake speech in the development set, the following problems and limitations should be discussed. First, the experimental results indicate that the proposed method could equally efficiently detect deepfake speech in the development set without background noise compared to the other methods regarding the accuracy, precision, recall, and F1 score. However, its balanced accuracy is poorer to some extent. The reason might be that the unbalanced dataset was used to train the proposed model. The investigation of the proposed method to be trained with a balanced dataset will be explored further. Second, the feature size of our proposed method is relatively small in comparison with the other methods. For example, for the same utterance, the size of the pathological features was only 11, whereas these

TABLE II
RESULTS OF PROPOSED METHODS COMPARED WITH EXISTING METHODS

Method	Dev. set evaluation (%)					Adp. set evaluation(%)				
	Accuracy	Precision	Recall	F1-score	Balanced accuracy	Accuracy	Precision	Recall	F1-score	Balanced accuracy
1. LFCC with GMM[18]	**99.99**	**99.99**	**100**	**99.99**	**99.97**	68.80	57.73	96.63	72.04	76.38
2. LFCC with CNN	99.92	99.98	99.94	99.96	99.84	73.00	95.74	64.29	78.81	78.81
3. GTCC with ResNet34	99.96	99.98	99.98	99.98	99.86	65.10	59.71	86.19	70.55	68.69
4. Pathological features with MLP (Proposed method)	98.09	98.47	99.47	98.97	91.02	76.30	75.81	**97.14**	**85.16**	62.40
5. Pathological features, LFCC and CNN features with MLP (Proposed method)	99.76	99.92	99.82	99.87	99.45	**79.90**	**96.46**	74.00	83.75	**83.83**

sizes for GTCC with ResNet34, LFCC with CNN, and LFCC with GMM were larger, which are 60×128, 60×128, and 57×100, respectively. Thus, the pathological features have discrimination potential for recognizing the deepfake speech. Suppose they are used well and smartly with other conventional features. In that case, they might reduce misclassification in decision-making in some cases with higher uncertainty. Third, the performances of our proposed method, LFCC with GMM, and GTCC with ResNet34, gradually degraded in noisy environments because all methods, trained from only the training set, were fragile under noisy conditions in the adaptation set. One possible way that we can try to solve this problem is to apply augmentation techniques so that the model can be trained with a noisy version of clean speech. Finally, in this study, we experimentally explore the usefulness of the pathological features by combining them with some conventional features, as shown in the fifth row of Table II. The overall performance is remarkably improved, especially in the balanced accuracy and precision. However, there is a trade-off between those improved aspects and recall. The reason for the recall decrease will be studied further. Also, the features used to combine with the pathological features in this study were chosen to prove the concept. They are not crafted in such a way that they go hand in hand with the pathological features. Thus, feature selection is left from this study and will be done in the future.

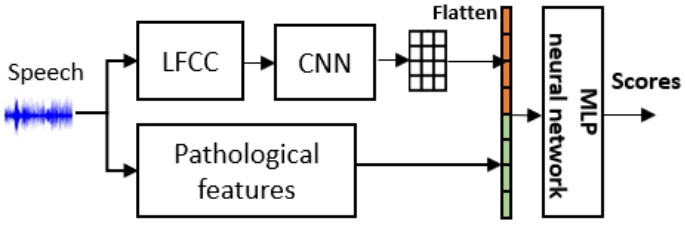


Fig. 7. Structure of proposed method using pathological features with LFCC and CNN feature, and MLP neural networks.

VI. CONCLUSION

This paper highlighted the study of pathological features to detect deepfake speech. We conducted experiments to investigate pathological features of genuine and fake speech. We then proposed a method that uses three jitter features, four shimmer features, HNR, CHNR, NNE, and GNE, for deepfake-speech detection. These pathological features, consisting of only 11 numbers, were classified using an MLP neural network to determine whether the speech was genuine or fake. The proposed method was evaluated on the basis of the ADD 2022 Challenge. The results indicate that the proposed method could effectively detect deepfake speech. Although the efficiency of the proposed method was still lower than the baseline in the development set, it was better than the baseline in the adaptation set. Thus, these pathological features can consistently account for synthetic voices and be used for deepfake-speech detection. In addition, we can improve the performance of the proposed method by combining these pathological features with some conventional ones (e.g., flattened features obtained from a CNN that takes LFCCs as its input) and then classifying the integrated features with an MLP neural network or other classifiers.

ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), I-O DATA foundation, and KDDI foundation (Research Grant Program). This work was supported by a grant from the SIIT-JAIST-NSTDA Dual Doctoral Degree Program. Also, this work is the output of the ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) project, titled Spoof Detection for Automatic Speaker Verification, and financially supported by NICT (http://www.nict.go.jp/en/index.html).

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *2019 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[3] K. Hartmann and K. Giles, "The next generation of cyber-enabled information warfare," in *2020 12th International Conference on Cyber Conflict (CyCon)*, IEEE, vol. 1300, 2020, pp. 233–250.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[5] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[6] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," *arXiv preprint arXiv:1906.09890*, 2019.

[7] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for add 2022," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 9226–9230.

[8] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9231–9235.

[9] J. M. Martıén-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9241–9245.

[10] S. Duraibi, W. Alhamdani, and F. T. Sheldon, "Replay spoof attack detection using deep neural networks for classification," in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2020, pp. 170–174.

[11] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[12] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495, 2017.

[13] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[14] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, 2015, pp. 2087–2091.

[15] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, IEEE, 2019, pp. 1–3.

[16] J. Yang and R. K. Das, "Low frequency frame-wise normalization over constant-Q transform for playback speech detection," *Digital Signal Processing*, vol. 89, pp. 30–39, 2019.

[17] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients.," in *Odyssey*, vol. 2016, 2016, pp. 283–290.

[18] J. Yi, R. Fu, J. Tao, *et al.*, "Add 2022: The first audio deep synthesis detection challenge," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022.

[19] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.

[20] S. S. Nidhyananthan and R. S. S. Kumari, "Language and text-independent speaker identification system using GMM," *WSEAS Transactions on Signal processing*, vol. 9, no. 4, pp. 185–194, 2013.

[21] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2017, pp. 1243–1247.

[22] D. Meghraoui, B. Boudraa, T. Merazi-Meksen, and P. G. Vilda, "A novel pre-processing technique in pathologic voice detection: Application to parkinson's disease phonation," *Biomedical Signal Processing and Control*, vol. 68, p. 102 604, 2021.

[23] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66 749–66 776, 2020.

[24] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," International Speech Communication Association (ISCA), 2007, pp. 778–781.

[25] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013, ISSN: 2212-0173.

[26] J. Gómez-Garcıéa, L. Moro-Velázquez, J. D. Arias-Londoño, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part iii: Review of acoustic modelling strategies," *Biomedical Signal Processing and Control*, vol. 66, p. 102 049, 2021.

[27] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[28] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar als detection based on analysis of voice perturbation and vibrato," in *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2019, pp. 267–272.

[29] H. Choudhary, D. Sadhya, and V. Patel, "Automatic speaker verification using gammatone frequency cepstral coefficients," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, 2021, pp. 424–428.

Temporal and Type Correlation in Digital Phenotyping for Bipolar Disorder State Prediction Using Multitask Self-Supervised Learning

Jia-Hao Hsu[1], Hua-Wei Tseng[1], Chung-Hsien Wu[1], Esther Ching-Lan Lin[2], and Po See Chen[3]
[1]Department of Computer Science and Information Engineering; [2]Department of Nursing; [3]Department of Psychiatry, National Cheng Kung University, Taiwan
E-mail: jiahaoxuu@gmail.com.tw, tmty210857@gmail.com, chunghsienwu@gmail.com, chinglan@ncku.edu.tw, chenps@mail.ncku.edu.tw Tel: +886-2757575-62530-2801

***Abstract* — Bipolar disorder is a prevalent mental illness characterized by a high relapse rate. In this study, we propose an early warning system that utilizes digital phenotyping to collect various data points from bipolar patients, including location information, self-assessment scales, daily mood reports, sleep patterns, and multimedia records, through a mobile application. These collected data are utilized to develop a predictive model for assessing the risk state of bipolar disorder. Compared to traditional recurrence prediction methods, this study incorporates medical records, medication data, and emergency records, as suggested by medical professionals, to define the five states of bipolar disorder, leading to enhanced accuracy. To account for data type correlation and temporal correlation, we employ a multitask self-supervised learning mechanism. The proposed method is trained on a Gated recurrent unit and demonstrates an improved prediction accuracy of 88.2% on the collected test data, as compared to the baseline accuracy of 85.1%. These findings highlight the significant importance of considering data type and temporal correlations in digital phenotyping for predicting the state of bipolar disorder.**

I. INTRODUCTION

Bipolar disorder (BD) is a chronic psychiatric condition characterized by episodes of depression, mania, or mixed states. It is estimated that BD affects approximately 435-544 million individuals worldwide, with a prevalence rate of 0.7% (0.6%-0.8%) [1]. Alarmingly, 25% - 50% of BD patients have attempted suicide, leading to significant social and economic consequences [2]. The World Health Organization recommends a psychiatrist-patient ratio of 1 doctor per 10,000 patients [3]. However, many regions struggle to meet this staffing requirement. While long-term monitoring can aid in managing BD [4], it also consumes substantial medical resources and personnel. Therefore, the objective of this study is to develop an early warning system to assist healthcare professionals in predicting the state of BD.

Numerous studies have explored early warning systems for bipolar disorder, with some focusing on monitoring patients' emotional or activity status [5, 6]. Digital phenotyping, involving the collection and analysis of various patient data, has become a common practice in this field [5, 7, 8]. Rating scales have been widely used for evaluating and diagnosing patients [9, 10]. Moreover, some studies have incorporated multimedia data, such as images and sounds, to detect potential BD symptoms [11, 12]. Building upon these methods, this study utilizes smartphones to collect digital phenotyping data from patients, including sleep patterns, Global Positioning System (GPS) information, emotional states, and weekly self-assessments, aiming for comprehensive monitoring.

However, most existing studies have not adequately assessed the temporal and data type correlations within long-term digital phenotyping data. To address this limitation, this study proposes the incorporation of multitask self-supervised learning [13], which allows the detection model to gain a more comprehensive understanding of the data by considering both temporal and data type correlations. Additionally, while many studies have focused on analyzing BD-related scale scores or determining whether a user has been diagnosed with BD [6, 14], this study takes a step further by observing patients' long-term medication and medical information to predict their future risk state. By integrating this information, the proposed model aims to provide doctors with a more intuitive basis [15].

This study makes two significant contributions. Firstly, it introduces a multitask self-supervised learning approach to enhance the prediction model's ability to capture the correlation between different data types and temporal sequences. The

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

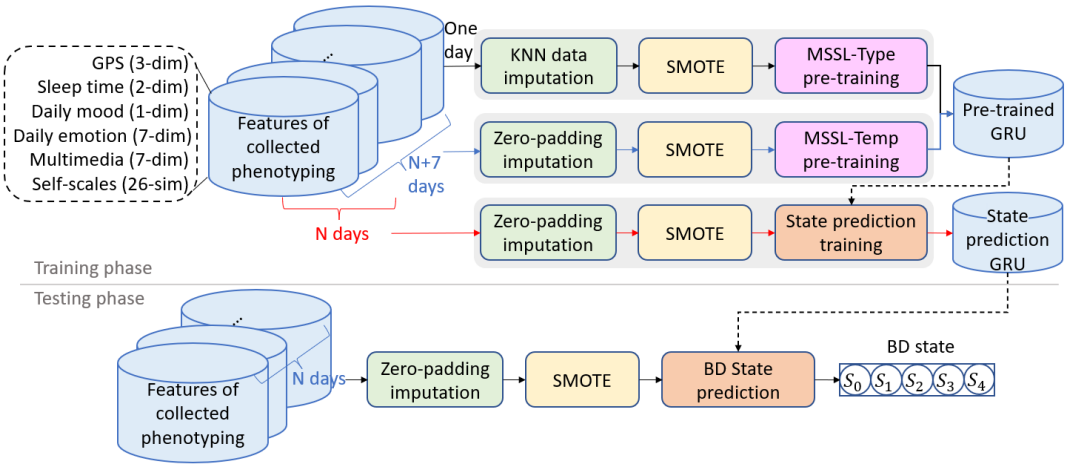


Fig.1 The system architecture proposed in this study.

model is first pre-trained using a pretext task and subsequently fine-tuned for bipolar disorder state prediction through downstream training. This method effectively leverages the benefits of self-supervised learning to improve the model's understanding of complex data patterns. Secondly, this study proposes the inclusion of medication and medical information of BD patients to differentiate their risk states. By considering these additional factors, the predictive model gains valuable insights that can aid in distinguishing different risk levels and informing clinical decisions.

II. Data Collection and Data Features

This study collaborated with the National Cheng Kung University Hospital to collect digital phenotyping data from patients diagnosed with bipolar disorder (BD). The project obtained approval from the Human Research Ethics Review Committee of the hospital (meeting number: A-ER-106-229). To be included in the study, patients had to meet the diagnostic criteria outlined in the DSM-5 Manual of Mental Disorders. The study involved a total of 166 participants, including 75 men and 91 women, with an average age of 34.3 years.

The data collection method employed in this study follows the approach used in a previous study [7][16]. A mobile application was used to gather data directly from the patients. The collected data can be categorized into two types: daily records and weekly records. The daily records encompassed information such as the patient's daily mood, emotions, location (GPS), multimedia records, and sleep duration. The weekly records included the Depression Anxiety Stress Scales (DASS-21) [17] and the Altman Self-Rating Mania Scale (ASRM) [18]. These records provided valuable insights into the patients' mental state and served as crucial data for analysis and prediction in this study.

In this study, 3-dimensional features were extracted from the latitude and longitude of the patients' GPS. The location features are the Local Variance, the Transition Time, and the Total Distance [19]. We calculate the variability of longitude and latitude separately, add them and take the logarithm to get the Local Variance. The Transition Time is the proportion of points with speed in all recorded GPS points in one day. The Total Distance is the sum of the distances between all recorded GPS points in one day. The self-scales are DASS-21 and ASRM. There are 21 questions in DASS-21, and each question is 0-3 points for depression, anxiety, and stress-related questions. The ASRM has 5 questions, and it is a mania scale of 0-4 points per question. By taking out the scores of each question of the two scales, the 26-dimensional self-scale features can be obtained. The 1-dimensional feature of patients' daily mood is the scores of emotional valences (-3 to 3) recorded by patients every day. The 7 daily recorded emotions are anger, boredom, disgust, nervousness, happiness, sadness, and surprise. And the corresponding scores are (0-3) for each emotion, which we use as the 7-dimensional emotional feature. The patient's sleep time is the wake-up and bedtime, and we convert it into numbers in minutes as a 2-dimensional sleep time feature. In the multimedia records, we extract the emotion profile of the utterance as features. Users can choose to record text, speech, or audio-visual records every day. We use emotion recognition models to classify multimedia data into 7 emotion classes, including anger, sadness, neutral, happiness, fear, surprise, and disgust. The feature of the audio-visual data is the averaged emotional profile of the obtained speech emotion and facial expression. The speech emotion recognition

model is wav2vec-2.0, the text emotion recognition model is BERT, and the facial expression recognition model is ResNet18.

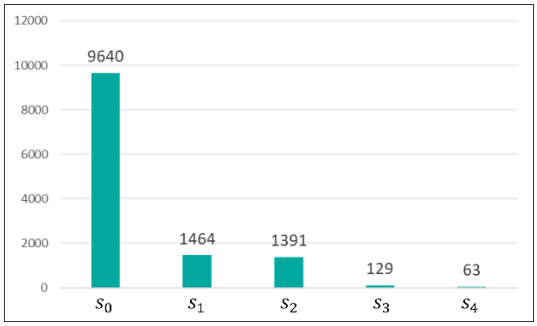


Fig. 2 The amount of data in each state

The goal of this study was to use data from the past N days to predict the BD states of patients in the coming week. The state is defined by extracting the inpatient emergency records and hospitalized records from the medical records. It can be divided into five states. s_0 is in a stable state, in which the patient's consultation interval and medication dose in the next week will be predicted as unchanged. s_1 is a state of relief, which is defined as a predicted increase in the interval between visits in the next week or a decrease in the dosage of medication, indicating that the condition of patients is relieved. s_2 is an aggravated state, and the patient's visit interval for the next week decreases or the dosage of medication increases, indicating that the patient may have a worsening condition. s_3 is in a serious state, and the patient has emergency records in the next week. This state means that the patient needs to be notified to return to the clinic immediately, and there may be a risk of recurrence. s_4 is the most serious state. The patient has a hospitalization record in the next week, and the patient should be returned to the clinic immediately. If multiple states occur at the same time, the data will be marked with the most severe state. The final data distribution is shown in Fig. 2. In the data statistics, as the patients must meet the aforementioned DSM-5 criteria to participate in data collection, and their status is relatively stable, the number of s_4 is relatively low.

III. Proposed Methods

The proposed system is divided into two main phases, as depicted in Fig. 1. The input to the system consists of the extracted data features, as described in Section 2 of the paper.

In the training phase, several steps are performed. Firstly, data imputation and data balancing are conducted for the consecutive data features spanning 1 day, N+7 days, and N days. This is followed by multitask self-supervised learning based on type correlation, multitask self-supervised learning based on temporal correlation, and BD state prediction training, respectively. Here, N represents the value of 21 in this study.

In the testing phase, the data features for the consecutive N days are utilized for state prediction after undergoing data imputation. Given the limited amount of data and the necessity to capture temporal correlation, the Gated recurrent unit (GRU) model [20] is employed as the prediction model. This model is chosen due to its ability to comprehend temporal dependencies effectively. Overall, this system framework combines data imputation, multitask self-supervised learning, and state prediction using a GRU model.

A. Multitask Self-Supervised Learning based on Type Correlation (MSSL-Type)

Missing data issues in the collected dataset are addressed using K-nearest neighbors (KNN) imputation. This method fills in missing values by estimating them based on nearby data points, approximating the true raw features. To handle class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) [21] is employed. SMOTE generates synthetic data points for the minority class, effectively balancing the dataset and preventing overfitting. The multitask self-supervised learning approach (MSSL-Type) enables the model to understand the relationships between different data types. By training the model with this method, it can infer the target data type using information from other types of data. Fig. 3 illustrates how the trained model leverages different data types to improve overall understanding of the dataset.

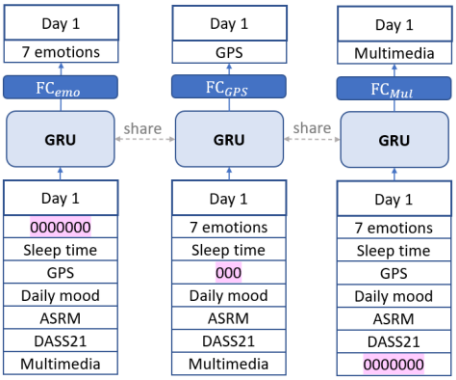


Fig. 3 MSSL-Type training process.

For the training process, the data features of a single day are utilized. Each type of feature is masked (replaced with zero) individually, and the remaining unmasked data is used to predict the masked data. This is done for various tasks, as outlined in Tab. 1. These tasks involve classification, using the cross-entropy loss function, and are trained on GRU and fully connected layers (FC). The tasks include predicting 7 daily

emotions, the midpoint of sleep time, and clustering the 3-dimensional GPS features into K groups (K=3 in this study). The classes for these tasks are defined based on the distribution of data in each class. Since the training data consists of only one-day data, the model is trained for a single time step, and all training shares the same GRU model parameters. In summary, the training process involves masking and predicting various data features for classification tasks, including daily emotions, sleep time midpoint, and GPS clustering, utilizing a single time step with shared GRU model parameters.

Tab. 1: The self-supervised training tasks.

Type	Classification Task
Daily 7 emotions	7 classes (emotions)
Daily mood	7 classes (-3~+3)
Sleep time duration	3 classes (≤ 320mins ≤ 450 mins <)
Sleep time median	3 classes (≤ 2 a.m. ≤ 4 a.m. <)
DASS-21	5 classes (≤ 9 ≤ 13 ≤ 20 ≤ 27 <)
ASRM	2 classes (≤ 5 <)
Multimedia emotion	7 classes (emotions)
GPS	3 classes (K-means clusters)

B. *Multitask Self-Supervised Learning with Temporal Correlation (MSSL-Temp)*

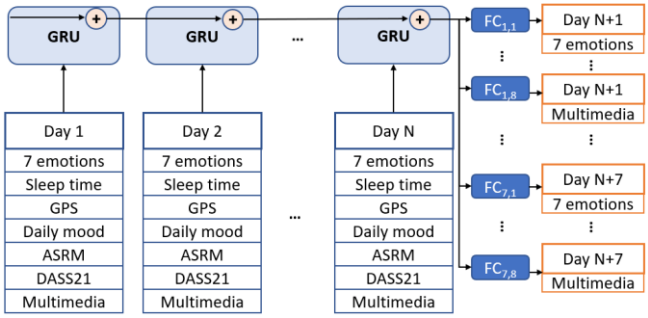


Fig. 4 MSSL-Temp training process.

This study employs multitask self-supervised learning with temporal correlation (MSSL-Temp) to predict future data based on past data, as depicted in Fig. 4. The aim is to predict the types of data for each of the next seven days using data from the previous N days. Zero-padding is utilized as an imputation method to simulate missing data, capturing the influence of missing data as a feature that affects the results. For instance, patients often neglect to record data during unstable periods, indicating potential fluctuations in their BD state. To address class imbalance, SMOTE is employed for data balancing. The training data consists of digital phenotyping data for N days encompassing seven types of features, with the target being the eight classification results mentioned in Tab. 1. The N days of data features are arranged sequentially based on the days and recursively input into the GRU model. The GRU model performs eight classification tasks for each day from the N+1st day to the N+7th day. Each training iteration involves 56 tasks (8 tasks * 7 days) and shares the same model parameters. In summary, the study uses MSSL-Temp to predict future data based on past data, considering missing data through zero-padding. SMOTE is applied for data balancing, and the GRU model performs eight classification tasks for each day from N+1 to N+7, with shared model parameters across the tasks.

C. *Bipolar state prediction training*

In this supervised training, digital phenotyping from the past N days is used to predict the next weekly BD state. This training uses zero-padding as the data imputation method to simulate the real situation. The SMOTE method is used to balance the data. The GRU pre-trained by MSSL-Type and MSSL-Temp is used as the initial model, and it is trained by the state prediction task together with a fully connected layer. This training is the final downstream task, and we feed the data recursively into the GRU model.

IV. Experimental Results

A. *Emotion Recognition of Multimedia Records*

The multimedia emotion recognition models are the emotion recognition models of text, speech, and image, respectively. The recognition results of each modality are shown in Tab. 2. We used the pre-trained Chinese BERT-base model as the text emotion recognition model. The Ren-CECps corpus [22] was used for fine-tuning, and the accuracy on the Ren-CECps testing set was 59.2%. We fine-tuned the pre-trained wav2vec-2.0 using the RAVDESS corpus [23]. After fine-tuning, the model had a recognition accuracy of 93.8% on the RAVDESS testing set. Resnet18 was used as the facial expression recognition model, and we used the RAF-DB database [24] for training. The recognition accuracy of the model on the RAF-DB testing set achieved 89.0%. It can be seen that these models have acceptable emotion recognition ability.

Tab. 2 Performance of emotion recognition models.

Modalities	Models	Datasets	Accuracy
Text	BERT	Ren-CECps	59.2%
Speech	Wav2vec-2.0	RAVDESS	93.8%
Facial	ResNet18	RAF-DB	89.0%

B. *Performance on the data used for evaluation*

The data collected in this study are long-term data, and the duration in days of the data collection period would affect the model performance. According to the recommendations of the cooperating doctors, the duration of the transition of BD patients in principle is no more than one month, and the short one is one week [25]. The data collected in this study consists of weekly self-scales, so the data unit is weekly data. Different numbers of weeks were used as the input lengths for this experiment, including 7 days, 14 days, 21 days, and 28 days. We trained the data on a GRU, and the 5-fold cross-validation results are shown in Fig. 5. It showed that too short an input length may contain insufficient information, and too long input may contain too much noise. Subsequent experiments will be performed with 21 days as the input length.

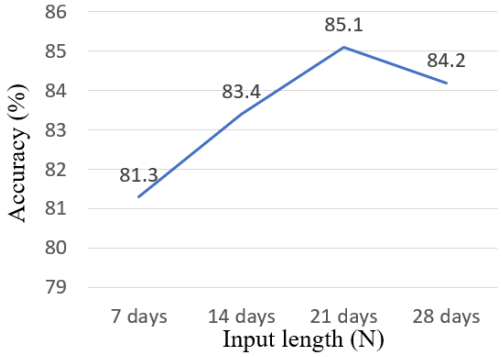


Fig. 5 Performance of different input lengths on GRU.

We also compared the effect of the amount of data balanced by SMOTE on GRU model performance. In Fig. 6, the optimal amount of data is 3500. That is, balancing each state of the training set to 3500 data can get the best performance on the testing set. We used this SMOTE setting for subsequent experiments.

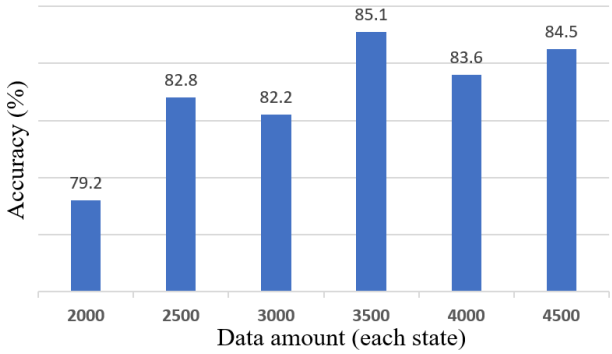


Fig. 6 Performance of different amount of balanced data.

C. *Performance of Different Detection Models*

This experiment compared other commonly used models in the research of BD, including logistic regression, random forest, SVM, and GRU. Tab. 3 shows the 5-fold cross-validation results of each method in the five BD state prediction tasks. Evaluation metrics included unweighted accuracy, macro-precision, macro-recall, and macro-F1 score. GRU had the best performance among various evaluation metrics, with an accuracy of 85.1%. GRU can better predict the long-term tracking of BD data by considering the time series relationship. Subsequent experiments would be based on GRU.

Tab. 3 Performance of different detection models.

	Recall	**Precision**	**F1**	**Accuracy**
Logistic	27.3	22.6	21.0	39.9%
Random Forest	52.3	50.5	50.8	77.9%
SVM	35.7	34.4	34.6	66.6%
GRU	**86.8**	**65.5**	**73.5**	**85.1%**

Tab. 4 Ablation studies.

	Recall					**Precision**					**F1 score**					**Acc.**
State / **Methods**	s_0	s_1	s_2	s_3	s_4	s_0	s_1	s_2	s_3	s_4	s_0	s_1	s_2	s_3	s_4	
GRU	85.0	86.4	84.0	86.8	92.0	96.7	65.5	60.1	48.9	56.3	90.5	74.5	70.0	62.5	69.8	85.1
GRU with autoregressive	86.0	85.7	84.2	84.4	92.0	96.6	66.8	61.0	58.2	52.7	91.0	75.1	70.7	68.9	67.0	85.9
GRU with MSSL-Type	87.0	83.8	83.5	86.8	96.8	96.4	66.0	63.9	56.2	53.5	91.5	73.8	72.4	68.2	68.9	86.3
GRU with MSSL-Temp	87.3	**87.3**	85.9	**89.1**	95.2	96.8	68.4	64.3	**69.6**	59.4	91.8	76.7	73.5	**78.2**	73.1	87.2
GRU with both MSSL	**88.4**	86.4	**87.8**	88.3	**96.8**	**97.0**	**69.1**	**67.5**	67.4	**67.7**	**92.5**	**76.8**	**76.3**	76.5	**79.7**	**88.2**

D. Ablation studies

This study compares the effect of adding different self-supervised learning methods on the model effect, including the original GRU, autoregressive self-supervised learning, our proposed MSSL-Type, and MSSL-Temp. Among them, the training objective of autoregressive self-supervised learning is to restore data values and train the model with MSE loss. This study is a prediction task, so the autoregressive objective is set to predict the data value of the next 7 days. The 5-fold cross-validation results are shown in Tab. 4. The methods proposed in this study all outperformed the traditional autoregressive self-supervised learning methods. MSSL-Temp may be more similar to downstream tasks, and the prediction task is also more detailed, which can bring the most improvement. Finally, the two self-supervised learning methods proposed in this study achieved the best accuracy.

V. Conclusions

his study makes two main contributions: predicting the BD risk state and considering the impact of correlations between temporal and data type in digital phenotyping. By using self-supervised learning for pre-training, the model's understanding of the data is enhanced. Temporal correlations are also crucial due to the long-term nature of digital phenotyping. The combination of both self-supervised learning methods improves BD state prediction accuracy from 85.1% to 88.2%. Specifically, MSSL-Type improves accuracy by 1.2%, and MSSL-Temp improves accuracy by 2.1%.

The digital phenotyping data used in this study was collected directly by the BD patients themselves, which introduces subjectivity to the information. However, considering that people with BD interact with others in their daily lives, the connections between individuals may also impact their condition. In future research, collecting relatively objective data from the observations and interactions of the patient's family and friends will be explored to enhance the prediction of BD states and improve overall prediction performance.

Acknowledgment

This work was supported in part by the National Science and Technology Council of Taiwan, under Contract No. 111-2221-E-006-150-MY3.

References

[1] A.J. Ferrari, E. Stockings, J.P. Khoo, H.E. Erskine, L. Degenhardt, T. Vos, and H.A. Whiteford, "The prevalence and burden of bipolar disorder: findings from the Global Burden of Disease Study 2013," Bipolar disorders, vol. 18, no. 5, pp. 440-450, 2016.

[2] L. Sher, "Manic-depressive illness: bipolar disorders and recurrent depression," American Journal of Psychiatry, vol. 165, no. 4, pp. 541-542, 2008.

[3] P. Burvill, "Looking beyond the 1: 10,000 ratio of psychiatrists to population," Australian & New Zealand Journal of Psychiatry, vol. 26, no. 2, pp. 265-269, 1992.

[4] I. Grande, M. Berk, B. Birmaher, and E. Vieta, "Bipolar disorder," The Lancet, vol. 387, no. 10027, pp. 1561-1572, 2016.

[5] Y.-C. Tseng, E.C.-l. Lin, C.H. Wu, H.-L. Huang, and P.S. Chen, "Associations among smartphone app-based measurements of mood, sleep and activity in bipolar disorder," Psychiatry Research, vol. 310, p. 114425, 2022.

[6] I. Perez Arribas, G.M. Goodwin, J.R. Geddes, T. Lyons, and K.E. Saunders, "A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder," Translational psychiatry, vol. 8, no. 1, pp. 1-7, 2018.

[7] H.-Y. Su, C.-H. Wu, C.-R. Liou, E.C.-L. Lin, and P.S. Chen, "Assessment of Bipolar Disorder Using Heterogeneous Data of Smartphone-Based Digital Phenotyping," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: IEEE, pp. 4260-4264.

[8] L. Orsolini, M. Fiorani, and U. Volpe, "Digital phenotyping in bipolar disorder: Which integration with clinical endophenotypes and biomarkers?," International Journal of Molecular Sciences, vol. 21, no. 20, p. 7684, 2020.

[9] R. Jadhav, V. Chellwani, S. Deshmukh, and H. Sachdev, "Mental disorder detection: Bipolar disorder scrutinization using machine learning," in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019: IEEE, pp. 304-308.

[10] Y.S. Liu, S. Chokka, B. Cao, and P.R. Chokka, "Screening for bipolar disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study," Journal of Affective Disorders Reports, vol. 6, p. 100215, 2021.

[11] Q.-B. Hong, C.-H. Wu, M.-H. Su, and C.-C. Chang, "Exploring Macroscopic and Microscopic Fluctuations of Elicited Facial Expressions for Mood Disorder Classification," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 989-1001, 2019.

[12] K.-Y. Huang, C.-H. Wu, M.-H. Su, and Y.-T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 393-404, 2018.

[13] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 2, pp. 1-30, 2019.

[14] J. Tomasik, S.Y.S. Han, G. Barton-Owen, D.-M. Mirea, N.A. Martin-Key, N. Rustogi, S.G. Lago, T. Olmert, J.D. Cooper, and S. Ozcan, "A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental

health questionnaire and blood biomarker data," Translational psychiatry, vol. 11, no. 1, pp. 1-12, 2021.

[15] R. Borges-Júnior, R. Salvini, A.A. Nierenberg, G.S. Sachs, B. Lafer, and R.S. Dias, "Forecasting depressive relapse in Bipolar Disorder from clinical data," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018: IEEE, pp. 613-616.

[16] J.-H. Hsu, C.-H. Wu, W.-K. Wang, H.-Y. Su, E. C.-L. Lin, and P. S. Chen, "Digital Phenotyping-Based Bipolar Disorder Assessment Using Multiple Correlation Data Imputation and Lasso-MLP," IEEE Trans. Affective Computing, DOI: 10.1109/TAFFC.2023.3299607, 2023.

[17] P.J. Norton, "Depression Anxiety and Stress Scales (DASS-21): Psychometric analysis across four racial groups," Anxiety, stress, and coping, vol. 20, no. 3, pp. 253-265, 2007.

[18] E.G. Altman, D. Hedeker, J.L. Peterson, and J.M. Davis, "The Altman self-rating mania scale," Biological psychiatry, vol. 42, no. 10, pp. 948-955, 1997.

[19] N. Palmius, A. Tsanas, K.E. Saunders, A.C. Bilderbeck, J.R. Geddes, G.M. Goodwin, and M. De Vos, "Detecting bipolar depression from geographic location data," IEEE Transactions on Biomedical Engineering, vol. 64, no. 8, pp. 1761-1771, 2016.

[20] R. Dey and F.M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), 2017: IEEE, pp. 1597-1600.

[21] A. Fernández, S. Garcia, F. Herrera, and N.V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal of artificial intelligence research, vol. 61, pp. 863-905, 2018.

[22] C. Quan and F. Ren, "A blog emotion corpus for emotional expression analysis in Chinese," Computer Speech & Language, vol. 24, no. 4, pp. 726-749, 2010.

[23] S.R. Livingstone and F.A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PloS one, vol. 13, no. 5, p. e0196391, 2018.

[24] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2852-2861.

[25] C.D. Schneck, D.J. Miklowitz, S. Miyahara, M. Araga, S. Wisniewski, L. Gyulai, M.H. Allen, M.E. Thase, and G.S. Sachs, "The prospective course of rapid-cycling bipolar disorder: findings from the STEP-BD," American Journal of Psychiatry, vol. 165, no. 3, pp. 370-377, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Data Selection Based on Phoneme Affinity Matrix for Electrolarynx Speech Recognition

I-Ting Hsieh*, Chung-Hsien Wu* and Shu-Wei Tsai†
* Graduate Program of Multimedia Systems and Intelligent Computing,
National Cheng Kung University, Tainan, Taiwan
E-mail: nd8081026@gs.ncku.edu.tw, chunghsienwu@gmail.com
†Dept. of Otolaryngology, National Cheng Kung University Hospital, Taiwan
E-mail: tsaisuwei@gmail.com

Abstract— **Electrolarynx (EL) is a communicative aid for the patient after laryngectomy to generate communicable speech. Since EL speech exhibits low speech intelligibility and produces loud noise, understanding the content of the speech remains challenging for listeners, even if the patient is proficient in using the EL device. Accordingly, it is important to develop the tools that offer additional communication methods. Automatic speech recognition (ASR) of EL speech emerges as a method worth considering in this regard. However, the problem of under-resourced data dramatically degrades the recognition performance of EL speech. Data augmentation is one of the viable solutions for addressing the issue of under-resourced speech data. However, even with an increased health training corpus, the improvement in EL speech recognition may not be satisfactory. Because the characteristics of the EL speech still differ significantly from those of health speech. This paper proposes a data selection method using the phoneme affinity matrix to prioritize the selection of health speech that closely resembles EL speech for data augmentation. The affinity between two phonemes is defined as the similarity of the Phone Posteriorgrams(PPGs) of the two phonemes, considering the phoneme models. The experimental results demonstrate that the approach utilizing data selection based on the phoneme affinity matrix yields superior results compared to both the baseline and the method employing random sampling to select the augmented health speech corpus.**

INTRODUCTION

For patients undergoing laryngectomy, some may only be able to produce disordered speech due to surgery or lesions on their vocal organs. This can result in communication barriers between the patient and others [1, 2]. Since laryngectomy involves the removal of the vocal cords, patients typically rely on the use of the Electrolarynx (EL) [3], a communication device that serves as a substitute for the vocal cords. However, the EL speech has low speech intelligibility [4], even the listener is familiar with the EL device. The listener typically struggles to fully comprehend the content of EL speech. There are two primary reasons for the poor quality of EL speech: (1) The EL produces a mechanical tone that sounds unnatural and lacks variation in pitch. (2) The proficiency of the patient in using the EL device. Furthermore, the proper use of the EL device requires full hand control from the patient. Although EL devices with manual tone control [5] have been introduced, they are still challenging to popularize.

In previous research, the majority of speech processing applications related to EL focused on voice conversion (VC). VC aims to transform EL speech into normal speech in order to enhance the communication abilities of patients [6-11]. Although VC has approached maturity in converting normal speech, learning the mapping relationship between EL speech and normal speech remains challenging due to the variability of EL speech. On the other hand, automatic speech recognition (ASR) is a choice worth investigating.

Despite the fact that EL-ASR can produce reasonable transcriptions after post-processing, it still faces the challenge of under-resource. Since collecting the EL speech corpus is more challenging compared to gathering the health speech, addressing the problem of under-resource primarily relies on novel training methods or data augmentation techniques [12-17]. In the case of ASR applied to health speech, novel training methods such as meta-learning [18-21] aim to construct the target ASR task by training on related tasks. For data augmentation, classical methods include perturbation of speed and volume [22] or SpecAugment [23]. To address the EL-ASR challenge, a related method [24] involves prolonging voiceless phonemes that exhibit high discrimination in EL speech.

Even if we augment the amount of EL speech data by incorporating more health speech, the augmented speech corpus remains inadequate for training the EL-ASR model. This is because the characteristics of EL speech differ not only in terms of volume or rhythm, but more importantly, in the acoustic characteristics specific to each patient based on their proficiency in using the EL device. It is crucial to propose a data selection method that can effectively select speech data beneficial for EL-ASR. Additionally, due to the scarcity of previous research on EL, it is challenging to determine the characteristics of EL speech for the purpose of data selection.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Given the limited prior knowledge available, utilizing the existing ASR model for data selection is a viable approach. In this paper, we propose a data selection method that utilizes the phoneme affinity matrix to identify speakers whose speech data closely resemble EL speech. In a similar method based on dysarthria, another speech disorder, researchers have used an ASR model trained on dysarthric speech to select similar speakers based on speech tempo [25] or convert the phoneme distribution into a Gaussian distribution to select similar utterances [26]. Due to the variable severity of dysarthria, the pronunciation of each patient may differ. Therefore, in [26], the entire phoneme distribution was used for data selection. Although EL speech may have lower intelligibility compared to dysarthric speech, the symptoms are roughly similar for each patient (i.e., complete removal of vocal cords). Selecting data based on individual phonemes is a valuable approach. This paper proposes a data selection method that utilizes the phoneme affinity matrix to prioritize the selection of health speech that is similar to EL speech for data augmentation.

PROPOSED METHOD

A. Data description

The EL speech is primarily generated by patients with speech disorders using the EL device, which produces excitation signals to replace the vibrations of the human vocal folds. The spectrograms of the EL speech are shown in Fig. 1. In Fig. 1, there are three speech samples recorded: one from a healthy individual and two from EL patients. From the three spectrograms, it is evident that the components with frequencies above 2500Hz are noise, resulting from the mechanical sound radiation noise produced by the EL device. For the healthy speaker using an EL device, there is a distinct speech pattern present below 2500Hz, since the healthy speaker possesses intact vocal organs and is proficient in operating the EL device. Patient speaker 1, who is acquainted with the EL device, produces a less clear speech pattern compared to the one produced by the healthy speaker. Patient speaker 2, who is unfamiliar with the EL device, produces a more muffled speech pattern with significant noise in the lower frequency. This can be attributed to their lack of familiarity with the EL device and their current condition shortly after surgery.

The EL speech corpus is collected by the Department of Otolaryngology at National Cheng Kung University Hospital. The speakers consist of 6 EL patients and 4 healthy speakers. Each patient has undergone laryngectomy surgery. The language of the transcription of the EL speech corpus is Mandarin. The EL speech corpus consists of 4280 utterances with a total duration of approximately 4 hours. Each utterance in the EL speech corpus consists of 10 Mandarin characters. The details of the EL speech corpus are shown in Table 1.

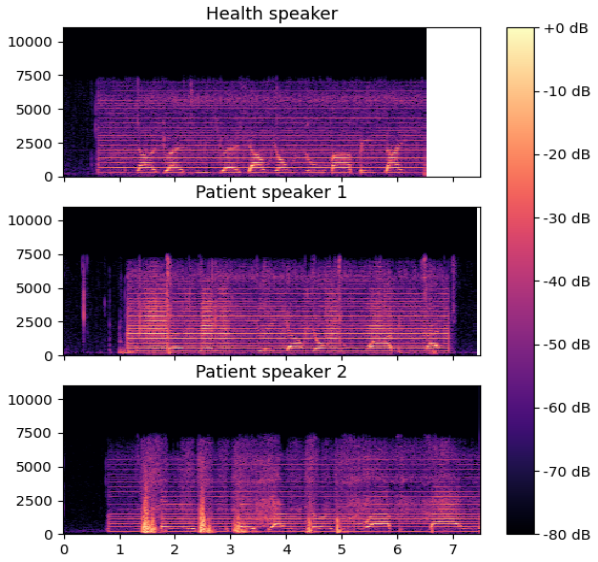


Fig. 1 The spectrograms of the EL speech (For details about the speech samples, please visit https://nd8081026.github.io/).

Table 2 The specification of the EL speech corpus.

	No. of utt.	No. of spk.	Total dur.	Dur. per utt.
Patient	1070	6	2.5 hrs.	2.89 sec.
Healthy	3210	4	1.5 hrs.	7.42 sec.
Total	4280	10	4 hrs.	

B. Data selection based on phoneme affinity matrix

Since the data augmented by a random selection mechanism for training the EL-ASR without knowing the characteristics of the EL cannot achieve the expected performance, a feasible data selection method is expected to assist in selecting useful data through the existing EL-ASR. The proposed method is shown in Fig. 2. The idea is to select speakers with similar articulation by utilizing Phonetic PosteriorGrams (PPGs) derived from a speaker-independent automatic speech recognition system. It is assumed that these PPGs can represent the articulation of speech sounds in a speaker-normalized space and correspond to spoken content independently. Firstly, the trained EL-ASR is applied to generate a phoneme affinity matrix for each EL or healthy speaker. If the phoneme affinity matrix of a specific healthy speaker's speech is similar to that of the EL speaker, the speech data of the healthy speaker is preferentially selected and added to the training corpus. The large health speech corpus is decoded by the trained EL-ASR model to generate the phoneme affinity matrix for each healthy speaker. The entire EL corpus is divided into two parts. The first part is used to train the EL-ASR model depicted in Fig 2. The second part is used for generating the phoneme affinity matrix. However, because the EL corpus is extremely scarce to accurately build a phoneme affinity matrix based on each

speaker, the phoneme affinity matrix of the EL speech is constructed for all EL speakers.

Before obtaining the phoneme affinity matrix, the PPGs and alignment information of the speech need to be obtained. The PPGs are extracted by the acoustic model of the trained EL-ASR. They represent the probability of phoneme frames in the speech. The alignment information is obtained from the Gaussian mixture models-hidden Markov models (GMM-HMM) of the trained EL-ASR model, which establishes the correspondence between the GMM and the frames within each phoneme.

The phoneme affinity matrix represents the reference phoneme as rows and the predicted phoneme as columns. Given a set of N phonemes $\{p_1, p_2, \ldots, p_N\}$, it can construct a similarity graph by computing similarities of each phoneme p_i with respect to all phoneme models, M_j $(1 \leqq j \leqq N)$. Let $M \in R^{N \times N}$ be the affinity matrix of the graph. The affinity of two phonemes p_i and p_j is defined as the PPGs of the phoneme p_i aligned to phoneme models, M_j. The speech of each healthy speaker will be traversed frame by frame to generate the affinity matrix for each speaker. Next, these phoneme affinity matrices need to be averaged by the total frame of each reference phoneme of the alignment information due to the imbalanced phoneme duration of each phoneme.

After averaging all the phoneme affinity matrices, the differences between the phoneme affinity matrix of each healthy speaker and that of all the EL speakers are estimated. The speech of the top-K healthy speakers with the smallest matrix differences is selected as the supplementary data to be added to the training database for data augmentation. The estimation of the difference matrices between each phoneme affinity matrix of the healthy (normal) speakers and global phoneme affinity matrix of the EL speakers is shown in Fig 3.

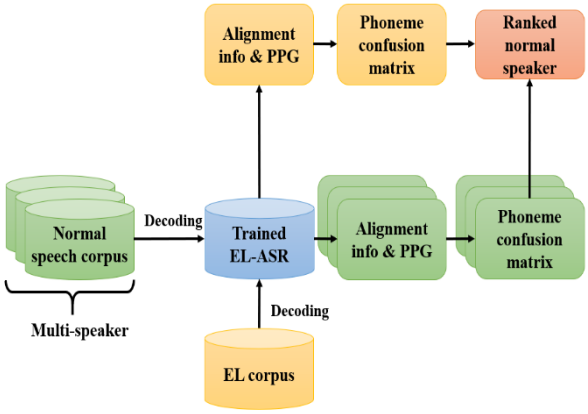


Fig. 2 The data selection mechanism based on phoneme affinity matrix.

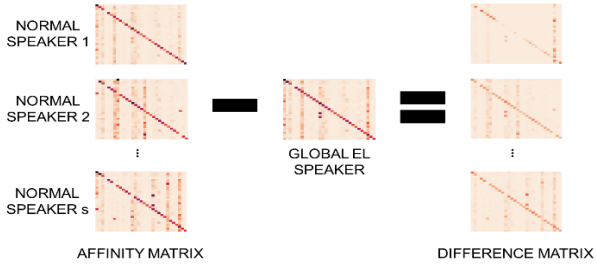


Fig. 3 Estimation of the difference matrices between each phoneme affinity matrices of the normal (healthy) speaker and the global phoneme affinity matrix of the EL speakers.

Experimental Results

A. Experimental Settings

This paper utilized the Kaldi speech recognition toolkit [27] as the foundational framework, which encompasses feature extraction, utterance alignment using the GMM-HMM acoustic model (AM), factorized time delay neural network (TDNN-F) AM [28], and a linguistic decoder. Since the EL speech corpus is very scarce, in order to find the best ASR parameter setting, this paper divided the EL speech corpus into a training set, a validation set and a test set. The validation set and test set each consisted of 100 randomly selected utterances, while the remaining utterances in the EL corpus were used for training. The AM utilized in this study consisted of 8 layers of TDNN-F, and a 5-gram language model [29] trained on a dataset of 1,643,267 utterances. The details of the parameters of the AM of the EL-ASR are listed in Table 2. For the subsequent experiments, to mitigate the issue of speaker dependency, this paper recalibrated the distribution of the EL corpus. For fairness, the leave-one-speaker-out verification method was employed to conduct the subsequent experiments. In each round of experiments, the corpus of one speaker was used as the test set, while the remaining speakers were used as the training set. After traversing all speakers, the final test result can be obtained by averaging all the traversed results.

Table 2 Specification of the AM of the EL-ASR.

Category	Hyperparameters	Values
Feature extraction	Dimension of MFCC	40
	Dimension of i-vector	100
GMM-HMM	Gaussian mixture components	1000
	Number of training iterations	40
TDNN-HMM	Dimension of all hidden layer	512
	Number of layers	8
	Number of epochs	4
	Learning Rate	10^{-3} to 10^{-4}

Table 3 Specification of the augmented speech corpus.

Corpus name	KING-ASR-044
Language	Mandarin
Number of the utterance	1,641,659
Duration (hours)	2,181.70
Number of the speaker	2,352 Male and 2,874 Female
Sampling rate	16000Hz

Next, for data selection, the data pool used for selection is the KING-ASR-044 [30]. The content of the augmented speech corpus is shown in Table 3. Due to the abundant data and speakers in KING-ASR-044, it is expected that our proposed data selection method can be used to select highly similar healthy speakers.

B. Evaluation of the EL-ASR

The first experiment investigated whether the EL-ASR relies on tone for phonemes. Due to the removal of the larynx after laryngectomy surgery, the patients are unable to adjust the pitch themselves, resulting in a very flat quality of the generated speech. There are five tones in Mandarin, and the Mandarin phoneme set comprises a total of 31 phonemes. If a phoneme contains a tone, the tonal phoneme set consists of 155 phonemes. For this issue, this paper applied the parameter settings obtained from Table 2 and conducted evaluations using only the EL speech corpus. The results are shown in Table 4.

Table 4 Comparison of the EL-ASR with/without tone.

Phoneme with/without tone	Test corpus	CER (%)
Phoneme with tone	Healthy	44.35
	Patient	84.28
Phoneme without tone	Healthy	41.88
	Patient	77.56

From the results shown in Table 4, it can be observed that both healthy and patient speakers achieved a lower character error rate (CER) for phonemes without tone. For the healthy speakers, the phonemes without tone yielded a CER improvement of approximately 2.47% compared to the phonemes with tone. For the patient speakers, the phonemes without tone showed an improvement of approximately 6.72% in CER compared to the phonemes with tone. From the analysis of the experimental results, the patients exhibited the most significant improvement in their results. It is speculated that the speech characteristics of the patients exhibit a wide range of diversity. Some patients have low speech intelligibility, some patients struggle to pronounce specific phonemes smoothly, and some patients are not familiar with the use of the EL device. Due to these diverse reasons, the patients have extremely low recognition ability, and incorporating tones into the phoneme set makes it even more challenging for the EL-ASR to generalize the features of each phoneme. The experimental results of healthy speakers using the EL device showed a slight improvement in the recognition ability of the EL-ASR, whether tones were included or not in the phoneme set. Due to the involuntary adjustment of speech tone or the absence of vocal cords, the EL speech, even when imitating the patient's usage, exhibits slight differences. Therefore, subsequent experiments were conducted using phoneme sets without tones.

For the evaluation, considering that the original EL corpus is very limited, if a large number of selected corpus are added, it may offset the recognition objects of the EL-ASR. Accordingly, only a small amount of data from selected speakers were added to the training set. In the experiment, data from the top 50, top 100, top 200, and top 400 healthy speakers were selected and added to the training set. Additionally, to demonstrate the effectiveness of the phoneme affinity matrix in selecting similar speakers, randomly selected samples from the healthy speakers were added to the experiment for comparison. The experimental results are shown in Table 5.

Table 5 The CER (%) of the EL-ASR with data augmentation.

Train corpus	Healthy speaker	Patient speaker
baseline	41.88	77.56
+top 50 speakers	36.72	69.45
+random 50 speakers	37.43	73.35
+top 100 speakers	33.56	65.30
+random 100 speakers	34.89	68.26
+top 200 speakers	**31.49**	**62.93**
+random 200 speakers	32.04	67.36
+top 400 speakers	31.58	63.44
+random 400 speakers	32.08	67.42

The experimental results demonstrated the effectiveness of the data selection method using the phoneme affinity matrix. The recognition results of the data selection method using the phoneme affinity matrix outperformed those of the random selection method.

Regarding the recognition results of the patient speakers, compared with the baseline results, the methods using the data

from the top-50, top-100, top-200, and top-400 speakers showed improvements of 8.11%, 12.26%, 14.63%, and 14.12% respectively. When compared with the previous set, the recognition results increased by 8.11%, 4.15%, 2.37%, and -0.51% respectively. For the recognition results of the healthy speakers, the corresponding improvements compared to the baseline were 5.16%, 8.32%, 10.39% and 10.3%, respectively. Compared to the previous set of quantifications were 5.16%, 3.16%, 2.07% and -0.09%, respectively. The improvement in the experimental results of the patient speakers was significantly greater than that of the healthy speakers. And the largest improvement was achieved when the top-50 speakers were added, as the phoneme characteristics of the top 50 healthy speakers closely resembled those of the EL speakers. When the top-200 speakers were added, both healthy and EL speakers reached their peak CER. However, after adding the top-400 speakers, the CER of both healthy and EL speakers decreased. Therefore, it can be concluded that the phoneme characteristics of speech pronounced by the top-200 speakers were similar to those of the EL speakers.

For the experimental results of healthy speakers, the difference between the results obtained from the two selection mechanisms was not greater than that observed for the EL patients. Because the EL speech of healthy speakers is closer to health speech in terms of speech intelligibility, tonal characteristics, etc., compared to the EL speech of the patients.

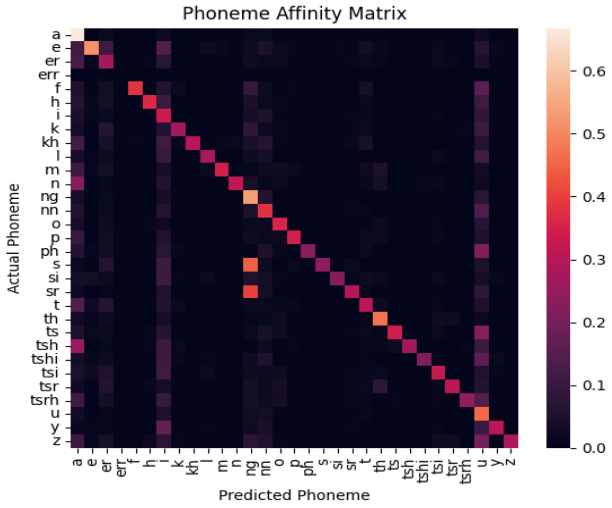


Fig 4. The phoneme affinity matrix of the EL patient.

Fig. 4 shows an example of the patient's phoneme affinity matrix. It can be observed that the diagonal PPGs of the phoneme affinity matrix exhibit high similarity. Apart from that, the PPGs of vowels such as /a/, /i/, and /u/, which are influenced by the vocal cords, are relatively more scattered compared to other phonemes.

Based on previous experiments, it is known that EL speech poses significant challenges in the ASR task. However, our proposed data selection method has shown considerable improvement in EL speech recognition performance. To demonstrate the data selection method proposed in this paper can select similar acoustic characteristics, we further utilize a VC model to transform healthy speech into EL speech. The converted speech is generated using STAR-GAN [31], an unsupervised many-to-many voice conversion model. By reducing the difference between healthy speech and EL speech, the data selection method aims to obtain EL converted speech, thus enhancing the ASR's recognition ability. The corresponding results are presented in Table 6.

Table 6 The CER (%) of the EL-ASR with data selection by the health speech and the converted speech.

Train corpus	Health speaker	Patient speaker
Baseline	41.88	77.56
+top 200 speakers	31.49	62.93
+STAR-GAN (top 200 speakers)	**30.19**	**62.46**

Based on the previous best experimental settings, the top-200 condition is used for data selection. From Table 6, it can be observed that selecting the converted speech yields the best CER for both healthy and patient speakers. And it proves that the data selection method proposed in this paper can indeed select the speech similar to the target speech.

Conclusion

The goal of this paper is to enhance the accuracy and stability of EL-ASR. In this paper, considering that the EL corpus is a low-resourced corpus, the method of data augmentation is used to improve the recognition ability of EL-ASR. Furthermore, due to the scarcity of resources similar to EL speech, randomly selecting data may not contribute to improving the recognition ability of EL-ASR. Therefore, this paper proposed a data selection method based on the phoneme affinity matrix. From a large Mandarin corpus, speakers with a highly similar phoneme recognition distribution to EL speech are added to the EL training set to augment the available data. The experimental results demonstrated that when incorporating the speech of the top-200 most similar healthy speakers, both patients and healthy speakers exhibited a significant improvement in recognition ability compared to the baseline. After, in order to reduce the difference between the characteristic of the health and patient, VC model was applied to generate the converted speech, and selected from the data selection method, more robust recognition ability of the ASR was obtained.

However, the expansion of the corpus through this data selection method will eventually reach its limit. The research on improving EL-ASR encompasses not only data augmentation but also exploring methods to extract speech

features that are beneficial to EL-ASR, which is another direction worthy of investigation.

ACKNOWLEDGMENT

This work was funded in part by Qualcomm through a Taiwan University Research Collaboration Project.

REFERENCES

[1] R.B. Bell, P. Andersen, R. Fernandes, "Oral, Head and Neck Oncology and Reconstructive Surgery-E-Book," Elsevier Health Sciences, 2017.

[2] P.M. Stell, "The first laryngectomy," The Journal of Laryngology & Otology 89.4 (1975): 353-358.

[3] S.R. Cox, "Review of the electrolarynx: the past and present." Perspectives of the ASHA Special Interest Groups 4.1 (2019): 118-129.

[4] G.S. Meltzner, and R.E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech." (2005).

[5] B. Shakya, V. Bharam and A. Merchen, "Development of an electrolarynx capable of supporting tonal distinctions in mandarin." (2014).

[6] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 2115-2119). IEEE.

[7] K. Tanaka, T. Toda, G. Neubig, S. Sakti and S. Nakamura, (2013, September). "A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion," In INTERSPEECH (pp. 3067-3071).

[8] H. Doi, K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5136-5139). IEEE.

[9] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, (2009). "Electrolaryngeal speech enhancement based on statistical voice conversion".

[10] M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. R. Jang and H.-M. Wang, "Mandarin Electrolaryngeal Speech Voice Conversion with Sequence-to-Sequence Modeling," In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 650-657). IEEE.

[11] E. Urabe, R. Hirakawa, H. Kawano, K. Nakashi and Y. Nakatoh, "Electrolarynx System Using Voice Conversion Based on WaveRNN," In 2020 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1-2). IEEE.

[12] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech communication, 56, 85-100, (2014).

[13] H.B. Sailor, A.T. Patil and H.A. Patil, "Advances in Low Resource ASR: A Deep Learning Perspective." SLTU. 2018.

[14] B. Thai, R. Jimerson, D. Arcoraci, E. Prud'hommeaux and R. Ptucha, "Synthetic data augmentation for improving low-resource asr," In 2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW) (pp. 1-9). IEEE.

[15] Y. Sharma, B. Abraham, K. Taneja and P. Jyothi, "Improving low resource code-switched ASR using augmented code-switched TTS," arXiv preprint arXiv:2010.05549, (2020).

[16] R. Gokay and H. Yalcin, "Improving low resource Turkish speech recognition with data augmentation and TTS," In 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 357-360). IEEE.

[17] I. -T. Hsieh, C. -H. Wu and C. -H. Wang, "Acoustic and Textual Data Augmentation for Code-Switching Speech Recognition in Under-Resourced Language," 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 302-307.

[18] C. Finn, P. Abbeel, and S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” 2017, p. 10.

[19] G.I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu and P. Fung, "Meta-Transfer Learning for Code-Switched Speech Recognition." In ACL 2020.

[20] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu and P. Fung, "Learning Fast Adaptation on Cross-Accented Speech Recognition." In Interspeech 2020.

[21] J.-Y. Hsu, Y.-J. Chen and H.-Y. Lee, "Meta learning for end-to-end low-resource speech recognition," In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7844-7848). IEEE.

[22] T. Ko, V. Peddinti, D. Povey and S. Khudanpur, (2015), "Audio augmentation for speech recognition". In Sixteenth annual conference of the international speech communication association.

[23] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk and Q.V. Le, (2019). "Specaugment: A simple data augmentation method for automatic speech recognition". arXiv preprint arXiv:1904.08779.

[24] P. Stanislav, J.V. Psutka and J. Psutka, (2020, October), "Increasing the Accuracy of the ASR System by Prolonging Voiceless Phonemes in the Speech of Patients Using the Electrolarynx". In International Conference on Speech and Computer (pp. 562-571). Springer, Cham.

[25] F. Xiong, J. Barker and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalized dysarthric speech recognition," In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5836-5840). IEEE.

[26] F. Xiong, J. Barker, Z. Yue and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7424-7428). IEEE.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcˇek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. vesely, “The Kaldi Speech Recognition Toolkit,” p. 4, 2011.

[28] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, S. Khudanpur (2018, September). "Semi-orthogonal low-rank matrix factorization for deep neural networks." In Interspeech (pp. 3743-3747).

[29] A. Stolcke, “SRILM – an Extensible Language Modeling Toolkit,” Denver, Colorado, USA, 2002, p. 4.

[30] “King-ASR-044, speechocean,” Tech. Rep., Aug. 2014. [Online]. Available: https://en.speechocean.com/datacenter/details/1409.html

[31] Kaneko, T., Kameoka, H., Tanaka, K., & Hojo, N. (2019). "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion". Proc. Interspeech 2019, 679-683.

Reduction of Annotation Effort in Medical Image Analysis Based on Self-supervised Learning

Kai-Hsuan Chan[1] and Yi-Chong Zeng[2]
Software Technology Institute, Institute for Information Industry, Taipei, Taiwan, R.O.C.
[1]E-mail: kaihsuanchan@iii.org.tw Tel: +886-2-66072930
[2]E-mail: yichongzeng@iii.org.tw Tel: +886-2-66072958

***Abstract* — Self-supervised learning is increasingly valued in medical imaging, with the hope of solving the problem of scarce high-quality labeled medical image datasets. This study proposes a self-supervised learning training framework for single-view and multi-view images, which combines multi-task and semi-supervised loss applied to medical image classification and detection problems. It can save more than 50% of labeled training data and significantly reduce the demand for medical images and labeled data. Meanwhile, the proposed method achieves high performance in the classifications.**

I. INTRODUCTION

In recent years, the field of artificial intelligence and deep learning has flourished. In addition to natural image processing, the analysis of medical images has also attracted widespread research interest. Based on the significant achievements of deep learning models in many medical imaging tasks, including, X-ray images, computational topology (CT), magnetic resonance imaging (MRI), ultrasound, pathological sections, and various medical data [1], the technology assisted doctors in disease detection and diagnosis [2], and had been gradually deployed in various countries to obtain medical device licenses and commercialization. However, in the development process of artificial intelligence models for medical imaging, it often requires a considerable cost to collect images and annotate lesions or diseases. Unlike general natural images, which can be annotated through relatively simple observation, medical images must be annotated by experienced radiologists or doctors in related fields. What is more difficult is that these professional doctors are already very busy with their tasks in hospitals and clinics. It is not easy to obtain a large amount of medical image annotation data. How to reduce the impact of insufficient medical image annotation data is a very important issue. Self-supervised learning has received considerable attention in academia in recent years. It can be analogous to the effectiveness of training supervised learning with a large number of images, such as the well-known benchmark ImageNet in the field of natural images. The pre-training model obtained by self-supervised learning can be used to improve the effectiveness of supervised learning with relatively few images using transfer learning technology [2][3].

Compared to natural images, there has been relatively less research on the application of self-supervised learning in medical imaging in recent years. However, in order to solve the problem of scarce high-quality labeled medical image datasets, more and more research is using self-supervised learning technology to seek effective solutions [4]. For example, Ghesu et al. proposed using more than 100 million medical images of various modalities [3], including X-ray images, CT, MRI images, and ultrasound images, to train pre-training models using self-supervised learning and further transfer learning to the required application images to improve overall model performance. In [4], Shurrab and Duwairi mentioned that the use of pre-training models trained on natural image ImageNet in many past medical image studies has always been a controversial issue. Therefore, 40 recent research papers on self-supervised learning in the field of medical image analysis are organized to reveal recent innovations in this field and serve as possible reference directions for future research. In many cases, medical images collected usually contain more unlabeled images than labeled images. Therefore, how to use unlabeled data more effectively to improve the training efficiency of labeled images is very important. In [5] and [6], the authors proposed more effective self-supervised learning strategies applied to 2D ultrasound images, CT images, MR images, skin lesion images, and chest X-ray images. In [7], Taleb et al. presented a method specifically for self-supervised learning of 3D medical image processing. Furthermore, Chen et al. used self-supervised learning techniques to deal with the problem of difficulty in obtaining medical image annotation data [8].

In addition to common 2D single images (such as X-rays,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

pathological sections, and ultrasound), 3D images (such as CT, and MRI), multi-view image processing (such as mammography, and multi-view skeletal images) is also a common research topic in medical imaging. How to use deep learning to process multi-view images is based on the concept of taking different perspectives with a camera in natural images to deal with the problems caused by dead angle occlusion in individual perspectives. For example, in [9] and [10], the researchers utilized multi-view convolutional neural network (CNN) architecture to deal with surveillance detection or object recognition problems. More detailed introductions to the principles of related multi-view techniques can be obtained from [11]. In [12], Yan et al. summarized past research on multi-view image processing and summarizes the CNN architecture into two mechanisms: one-view-one-net and multi-view-one-net for analysis. Through these two mechanisms, some research has also been used to process multi-view medical images. In [13] and [14], the authors extracted different perspective cross-sectional images of computer tomography and input them into a CNN model for analysis using the concept of multi-view to perform preliminary screening for COVID-19 pneumonia and identification of benign and malignant lung nodules. Even a single medical image can be pre-processed in multiple ways and analyzed using a multi-view CNN architecture [15]. It can even be combined with a multi-task CNN architecture to process multi-view mammography images and predict radiological evaluations related to cancer [16].

In this study, we propose a method that combines multi-task learning, semi-supervised loss, and self-supervised learning to deal with the problem of difficulty in obtaining medical image annotation data. The main contributions are: (1) analyzing how much self-supervised learning can reduce the amount of medical image annotation data, and (2) evaluating the impact of reducing the insufficient amount of annotation data for single images and multi-view images respectively.

The structure of this paper is as follows: Section II briefly introduces self-supervised learning techniques and the concept of Spatial Transformer Networks used in multi-view image analysis; Section III describes our proposed self-supervised learning methods for single images and multi-view images respectively; Section IV presents the results of reducing the impact of insufficient data using our proposed self-supervised learning methods on single image and multi-view image datasets respectively. Finally, Section V summarizes this study.

II. Related Works

A. *Self-supervised Learning*

Self-supervised learning is a robust training method in the field of deep learning in recent years. It can train models without manual annotation to solve the problem of difficulty in obtaining high-quality labeled medical image datasets. It can usually be classified into three approaches: predictive, generative, and contrastive. The main concept is to first use a large number of unlabeled image datasets to generate pseudo labels for training and produce a pre-training model. Then use a small amount of labeled data that has been obtained to perform transfer learning with this pre-training model and fine-tune the entire CNN model. This allows a small amount of labeled data to achieve robust performance in model training [4] [17], as shown in Fig. 1.

This study mainly adopts the generative approach, such as the Models Genesis method [18]. The main concept is to randomly pre-process the input image through transformations such as non-linear, local-shuffling, out-painting, and in-painting. After feature extraction by the CNN encoder, it is connected to the decoder to generate the final image, so that the generated image is restored to the original image. Therefore, a large number of unlabeled datasets can be used to learn the CNN model and produce a pre-training model. This pre-training model can further provide a small amount of labeled data for transfer learning and fine-tune the final model.

B. *Spatial Transformer Networks*

Spatial Transformer Networks (STN) is a learnable module that can be connected to CNN. It is used to solve the problem that CNN cannot handle spatial invariance, including translation, scale, rotation and more generic warping [19]. This method is mainly divided into two parts: Localisation Network and Sampling grid generator, as shown in Fig. 2. The Localisation Network can generate the angle θ required to estimate the parameters of the transformation $\mathcal{T}_\theta$ from the input feature map, where $\mathcal{T}_\theta$ can be a 2D affine transformation A_θ.

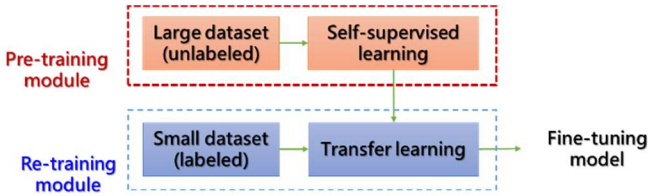


Fig. 1. Self-supervised learning main workflow

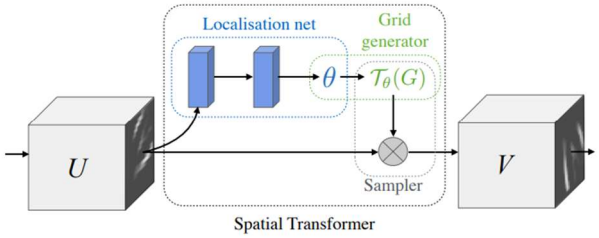


Fig. 2. Architecture of a spatial transformer module [18]

The sampling grid generator uses this estimated 2D affine transformation A_θ to resample the input feature map into the result after affine transformation, allowing the input image to be automatically corrected and making the CNN model recognition more accurate [19].

III. The Proposed Method

This study analyzes self-supervised learning and transfer learning for single images and multi-view images in medical imaging in order to reduce the amount of annotation data required for medical images. The following introduces the self-supervised learning methods for single images and multi-view images respectively.

A. Self-Supervised Learning in Single-view Image

The most commonly taken X-ray and ultrasound images in medical imaging are usually used to check the health of various parts of the body, such as bones, lungs, and abdomen. They are often used as preliminary examination tools and a large part of them are single image shooting and processing. If these single images are used for the purpose of auxiliary screening or diagnosis, and AI model development and training are carried out, in the case of insufficient image or annotation data, we propose a method that combines self-supervised learning and a multi-task architecture that can simultaneously perform image classification and detection.

As shown in Fig. 3, a classifier and a decoder are connected to a backbone architecture at the same time. The classifier performs image classification, while the decoder is connected to the backbone in a U-Net-like architecture and is trained to generate suspicious lesion locations. Since this is an encoder-decoder architecture, the generative approach in Self-Supervised Learning (SSL) can be used at the same time, such as the Models Genesis method proposed in [18], which uses a large number of unlabeled images to train and generate pre-trained models. The input images are subjected to various random preprocessing transformations, and the results generated by the decoder are expected to be restored to the original image.

Artificial labeling usually includes identification labeling of image anomalies, labeling of lesions or organ areas, and not necessarily all kinds of labeling data exist in every case. Therefore, in the design of the loss function, a semi-supervised loss must be used. When each type of labeling data is used as ground truth, individual loss L_i is multiplied by a weight w_i so that when the label exists, w_i is 1, otherwise w_i is 0 when it does not exist. We can use a large number of unlabeled images to train a pre-trained model with SSL, and then use this with artificial labeling data to perform transfer learning and fine-tuning to produce the final model with the goal of improving the overall performance of the model.

B. Self-supervised Learning in Multi-view Images

In addition to taking a single image, many diseases often require examination through images from different perspectives to reduce the problem of missed interpretation due to partial obstruction by body tissues or bones. Multiple perspective images taken include chest X-rays, limb skeletal lesions, mammography, etc. Therefore, if image analysis is performed, the CNN model must be able to input multiple perspective images simultaneously and cannot directly use the architecture shown in Fig. 3 for analysis.

Therefore, we adjusted the CNN architecture for multi-view image analysis, as shown in Fig. 4. The CNN architecture can input multiple view images at the same time. Each image is input into the same backbone structure and then connected to a classifier and a decoder at the same time as Fig. 3. The images from the second view and above are connected to the Spatial Transformer Network (STN) after the decoder and provide more classification information to the classifier.

During SSL training, the approach is similar to that of single-view image, using a generative approach. Images from each perspective undergo various random preprocessing transformations. The results generated by the decoder are expected to be restored to the original image or restored to the first perspective image after STN processing to generate a pre-trained model. Then, transfer learning is performed with manually labeled data, and like single-view image, training is done using a semi-supervised loss. The decoder outputs the lesion area for each perspective separately, allowing the model to simultaneously produce case classification results and suspicious lesion locations.

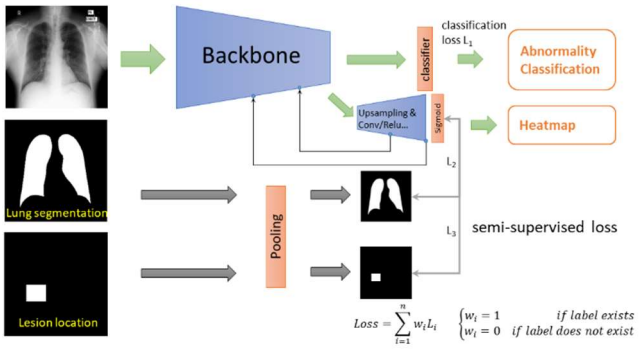


Fig. 3. CNN architecture with self-supervised learning in single-view image

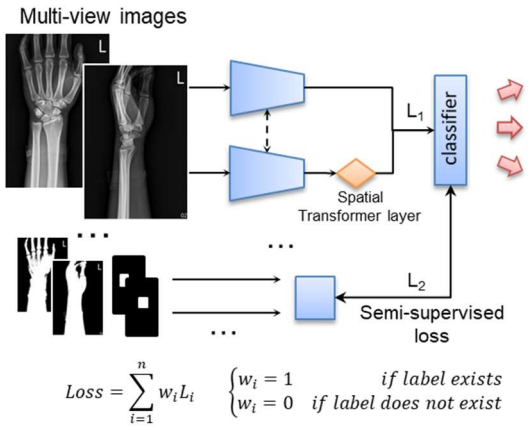


Fig. 4. CNN architecture with self-supervised learning in multi-view images

C. Datasets

In this study, the single-view image experiment mainly performs tuberculosis classification. The datasets used for tuberculosis are taken from Montgomery [21], Shenzhen [21], Kaggle [22], TB Portal [23], TBX11K [24], and TBXpredict [25]. After cleaning, it contains 4,988 tuberculosis positive images and 11,633 tuberculosis negative images (including non-tuberculosis abnormal and normal). This dataset is divided into a test dataset, which includes 530 tuberculosis positive images and 1,152 tuberculosis negative images. The chest X-ray datasets used for self-supervised learning include NIH Chest X-ray [26] with 112,120 images and CheXpert Dataset with 224,316 images [27].

The images used in the multi-view images experiment include 158,626 cases of mammography images and 12,071 cases of two-view wrist fracture datasets obtained through the hospital Institutional Review Board (IRB). In the experiment, the wrist fracture datasets were divided into a test dataset, which includes 347 wrist fracture positive images (radius fracture) and 451 negative images.

IV. The Experiment Results

This experiment analyzes single-view images for tuberculosis classification in chest X-rays. For multi-view images, it uses wrist fracture classification in two-view X-ray images for experimental analysis.

A. Tuberculosis Classification

As listed in Table I, the baseline, named Model A, uses ImageNet as a pre-trained model to train a tuberculosis classification model with the tuberculosis classification dataset we collected and organized. Model B uses two large chest X-ray datasets, NIH Chest X-ray and CheXpert Dataset, with over 330,000 images to train a pre-trained model for single-view images using the SSL method introduced in Section III. Transfer learning is performed with different proportions of training data and compared with the performance of Model A, including evaluating the Area under the Curve of ROC (AUC), sensitivity (SEN), and specificity (SPE) for tuberculosis classification. The suspicious tuberculosis heatmap images are shown in Fig.5. From the experimental results, it can be found that when we use only 50% of the training data for transfer learning on the SSL pre-trained model, its performance is still better than Model A. The training data can be saved by more than 50%, greatly reducing the demand for medical images and annotation data.

B. Wrist Fracture Classification

As listed in Table II, Model A still uses ImageNet as a pre-trained model and trains on a single-view image architecture. The performance evaluation is based on the ensemble results of individual performances from multiple perspectives. Model B uses mammography images and wrist fracture images obtained through the hospital IRB, with two views as one case, to train a pre-trained model for multi-view images using the SSL method introduced in Section III. Transfer learning is performed with different proportions of training data and compared with the performance of Model A. The suspicious wrist fracture heatmap images are shown in Fig.6. From the experimental results, it can be found that when we use only 40-50% of the training data for transfer learning on the SSL pre-trained model, its performance is still better than Model A. Similar to the results of single-view images experiments, the training data can be saved by more than 50%, greatly reducing the demand for medical images and annotation data.

V. Conclusions

In this study, we propose a CNN architecture that combines SSL with semi-supervised loss and multi-task in medical

Table I. Performance comparison of tuberculosis classification data with different proportions

Model	Proportion	Data	AUC	SEN	SPE
Model A (baseline)	100%	13,316	0.945	75.8%	94.7%
Model B (SSL)	100%	13,316	0.968	84.9%	97.3%
	80%	10,825	0.957	83.3%	93.3%
	50%	**6,737**	**0.949**	**80.3%**	**95.5%**
	30%	4,042	0.907	74.2%	91.5%
	20%	2,692	0.803	65.2%	79.3%
	10%	1,344	0.823	69.7%	79.8%

Table II. Performance comparison of wrist fracture classification data with different proportions

Model	Proportion	Data	AUC	SEN	SPE
Model A (baseline)	100%	11,226	0.925	82.1%	92.9%
Model B (SSL)	100%	11,226	0.943	87.6%	92.6%
	80%	8,988	0.940	84.6%	93.1%
	50%	5,619	0.934	82.5%	92.6%
	40%	**4,494**	**0.934**	**82.4%**	**92.7%**
	30%	3,370	0.927	79.0%	92.7%
	20%	2,248	0.922	79.6%	92.6%
	10%	1,124	0.906	69.6%	92.6%

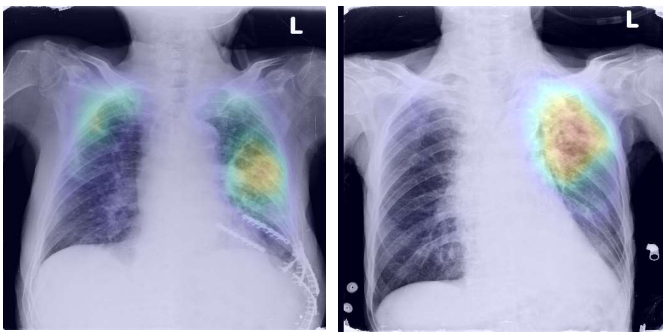


Fig.5. The heatmaps show that some suspected tuberculosis lesions are present on the left lung in both chest X-ray images. The higher the heat in the heatmaps, the higher the level of suspicion.

imaging and is applied to single-view images and multi-view images. A large number of unlabeled medical images are used to train the SSL pre-training model, and the final model is fine-tuned using transfer learning. It is evaluated on single-view images for tuberculosis classification and multi-view images for wrist fracture classification. Both can save more than 50% of labeled training data and can be used to solve the problem of scarce high-quality labeled medical image datasets. Furthermore, the sensitivities were, respectively, 82.4% and 80.3% for wrist fracture classification and tuberculosis classification.



Fig. 6. The heatmap shows wrist fracture X-ray images. The left image shows the radius fracture in a frontal view X-ray image, while the right image shows the radius fracture in a lateral view X-ray image.

ACKNOWLEDGMENT

This study is conducted under the "FAST AI Technologies and System (2/4)" of the Institute for Information Industry which is subsidized by the Administration for Digital Industries of moda of the Republic of China.

REFERENCES

[1] X. Zeng, S. Lin, and C. Liu, "Multi-view deep learning framework for predicting patient expenditure in healthcare," IEEE Open J. Comput. Soc., vol. 2, pp. 62–71, 2021.

[2] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," Med. Image Anal., vol. 79, Jul. 2022.

[3] F. C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, D. Neumann, P. Patel, R. S. Vishwanath, J. M. Balter, Y. Cao, S. Grbic, and D. Comaniciu, "Self-supervised learning from 100 million medical images," arXiv:2201.01283, 2022.

[4] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: a survey," PeerJ Computer Science, vol. 8, no. 19, Jul. 2022.

[5] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," Med Image Anal, vol. 58, pp. 101539, Dec, 2019.

[6] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, Deaton J., ... and M. Norouzi, “Big self-supervised models advance medical image classification,” Proc. of the IEEE/CVF Int. Conf. on Computer Vision, pp. 3478-3488, 2021.

[7] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, “3D self-supervised methods for medical imaging,” Proc. Adv. Neural Inf. Process. Syst., pp. 18158–18172, 2020.

[8] S. Chen, G. Bortsova, A. G.-U. Juárez, G. van Tulder, and Y. M. de Bruijne, “Multi-task attention-based semi-supervised learning for medical image segmentation,” Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., pp. 457–465, 2019.

[9] Y. Hou, L. Zheng, and S. Gould, “Multiview Detection with Feature Perspective Transformation,” In European Conf. on Computer Vision, pp. 1–18, 2020.

[10] M. Seel, P. Mäder, “Multi-view classification with convolutional neural networks,” PLoS ONE, vol. 16, no. 1, pp. 1-17, 2021.

[11] V. O. B. S. P. Dip, “Multi-view Camera synthesis using Convolutional Neural Network,” A Dissertation Presented to the University of Dublin, Trinity College in partial fulfillment of the requirements for the degree of Master of Science in Computer Science, 2019.

[12] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, “Deep multi-view learning methods: A review,” Neurocomputing, vol. 448, pp. 106–129, Aug. 2021.

[13] X. Wu, H. Hui, M. Niu, L. Li, L. Wang, B. He, X. Yang, L. Li, H. Li, J. Tian, and Y. Zha, “Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study,” Eur. J. Radiol., vol. 128, Jul. 2020.

[14] S. Hussein, R. Gillies, K. Cao, Q. Song, U. Bagci, “TumorNet: lung nodule characterization using multi-view convolutional neural network with Gaussian process,” Proc. of the IEEE Int. Symposium on Biomedical Imaging, pp. 1007–1010, 2017.

[15] J.R. Ferreira, D. Armando Cardona Cardenas, R.A. Moreno, M. de Fatima de Sa Rebelo, J.E. Krieger, M. Antonio Gutierrez, “Multi-view ensemble convolutional neural network to improve classification of pneumonia in low contrast chest X-ray images,” Annu. Inte. Conf. of the IEEE Engin. in Medic. & Bio. Soc., pp. 1238-1241, 2020.

[16] T. Kyono, F. J. Gilbert, and M. Schaar, “Multi-view multi-task learning for improving autonomous mammogram diagnosis,” Proc. of the Machine Learning for Healthcare Conf., vol. 106, pp. 571–591, August. 2019.

[17] G. Huang, I. Laradji, D. Vazquez, S. Lacoste-Julien, and P. Rodriguez,“A survey of self-supervised and few-shot object detection,”arXiv:2110.14711, 2021.

[18] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, “Models genesis: Generic autodidactic models for 3D medical image analysis,” Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, pp. 384–393, 2019.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” Proc. NIPS, pp. 2017–2025, 2015.

[20] A. M. Raisuddin, E. Vaattovaara, M. Nevalainen, M. Nikki, E. J¨arvenp¨a¨a, K. Makkonen, P. Pinola, T. Palsio, A. Niemensivu, O. Tervonen, et al., “Critical evaluation of deep neural networks for wrist fracture detection,” Scientific reports, vol. 11, no. 1, pp. 1–11, 2021.

[21] S. Jaeger, S. Candemir, S. Antani, Y.-X. Wang, P.-X. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” Quant. Imag. Med. Surg., vol. 4, no. 6, p. 475, 2014.

[22] Chest Xray Masks and Labels, Jun. 2020, [online] Available: https://www.kaggle.com/nikhilpandey360/chest-xray-masks-and-labels.

[23] A. Rosenthal, A. Gabrielian et al., “The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis,” Journal of Clinical Microbiology, vol. 55, no. 11, pp.3267–3282, 2017.

[24] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, “Rethinking computer-aided tuberculosis diagnosis,” IEEE CVPR, pp. 2646–2655, Jun. 2020.

[25] A. Chauhan, TBXpredict dataset, 2014. [online] Available: https://sourceforge.net/projects/tbxpredict/.

[26] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, “ChestXRay8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” IEEE CVPR,. pp. 3462–3471, Jul. 2017.

[27] J. Irvin et al., “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” CoRR, vol. abs/1901.07031, pp. 4–7, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1901.07031

STUA-Net: A Fingerprint Reconstruction with Swin Transformer and Soft Collective Attention

Farchan Raswa Hakim[1] Prabowo Yoga Wicaksana[1,2] Wenny Ramadha Putri[1] Agus Harjoko[2] Jia-Ching Wang[1]
[1]*Department of Computer Science and Information Engineering, National Central University, Taiwan*
[2]*Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia*
email: farchan.hakim.r@g.ncu.edu.tw

***Abstract*—Fingerprints play a vital role in person authentication and verification. To achieve accurate recognition, fingerprint images should contain 25 to 80 minutiae points, which define the unique characteristics of a fingerprint. However, due to various factors such as changes in the environment, the fingerprint structure can become corrupted, resulting in low-quality fingerprints. This corruption leads to a limited number of extractable minutiae points, making it challenging to establish the uniqueness of an individual. In this paper, we propose STUA-Net, a novel approach that incorporates Swin Transformer into the encoding and decoding layers to effectively map corrupted regions. Additionally, we introduce Soft Collective Attention to suppress the activation of relevant features. Our proposed method serves as a foundation for future research to improve recognition accuracy, particularly in scenarios involving low-quality fingerprints. It addresses an important problem in the field and contributes to the advancement of fingerprint recognition technology**

***Index Terms*—Fingerprint Reconstruction, Corrupted Region, Swin Transformer, Soft Collective Attention**

I. Introduction

Automated Fingerprint Identification Systems (AFIS) have been widely used in embedded and mobile applications [1], [2]. Due to changes in both inner and outer environments, the fingerprint structure is often corrupted, resulting in low-quality fingerprints. Wetness, oiliness, and dirtiness affect the internal environment, thereby destroying the fingerprint structure. When a person has oily fingers, the contrast between ridges and valleys diminishes, making the pattern less visible [3], [4]. Consequently, we cannot find enough minutiae points that can be used to uniquely identify the person. The occurrence of corruption in the fingerprint structure is also highly dependent on an individual's occupation. People who are required to frequently disinfect their hands at work, such as doctors, chemical workers, and lab workers, experience corruption and damage to their fingerprints. Therefore, the reconstruction or recovery of low-quality fingerprint images is a significant issue that needs to be addressed before recognition and matching can be performed [5].

Recently, learning-based methods have emerged as an alternative solution and have been successfully deployed in fingerprint reconstruction [6], [7], [8]. Machine and deep learning approaches, such as convolutional networks, generative adversarial networks, and autoencoder networks, have been employed to learn the probability structure of the damaged areas in fingerprints. These approaches can efficiently and automatically estimate numerous parameters and attributes.

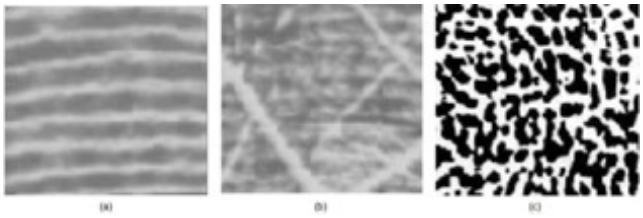


Fig. 1. Fingerprint areas of (a) Well-defined area, (b) Re- coverable corrupted region, and (c) Unrecoverable corrupted region.

Unlike face, iris, and voice data, fingerprint structures are highly detailed and susceptible to damage, requiring high-quality data collection. Consequently, directly applying ordinary deep learning models to reconstruct fingerprints with strict constraints is challenging.

The Transformer architecture, known for its attention mechanism, has gained popularity in language processing. However, adapting the Transformer to vision tasks faces challenges due to the differences between the two domains. These differences include significant variations in the size of visual entities and the higher pixel resolution in images compared to words in text. To address these challenges, the hierarchical Swin (Shifted Window) Transformer has been proposed. This approach limits self-attention computations to non-overlapping local windows while allowing cross-window connections, resulting in improved efficiency [9], [10].

They presented a novel approach based on the Swin Transformer, which showcased superior performance compared to CNN-based methods, leading to state-of-the-art results. Inspired by their work, we integrate the Swin Transformer [9] and Soft Collective Attention within the U-Net [11] Attention framework, called STUA-Net. The contributions of our work can be summarized as follows:

1) Due to the highly detailed and susceptible nature of fingerprint images, integrating the strengths of the Swin Transformer into the Encoding and Decoding layers allows for effective learning of long-range dependencies and contextual information related to rigid fingerprint features.
2) We introduce Soft Collective Attention (SCA) based on soft-attention, which actively suppresses activation at both coarse and fine levels when connecting the encoder and decoder stages.
3) We combine per-pixel loss and perceptual measurement

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

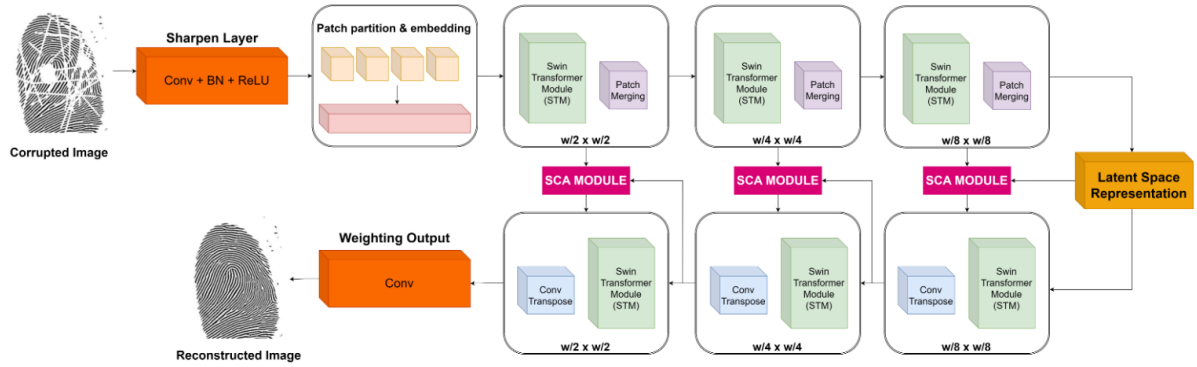


Fig. 2. The network architecture of our proposed method.

as a loss function to select a perceptually correlated metric.

However, our proposed method has limitations for selecting which corrupted image can be reconstructed, as seen in Fig 1.

II. Proposed Method

A. Datasets

Our study was conducted by generating a dataset using the Synthetic Fingerprint Generator (SFinGe) software, followed by preprocessing stages. Subsequently, the training data was prepared for training, validation, and testing purposes. Initially, the training dataset was separated from the testing dataset. Image quality assessment metrics were then calculated, using a split of 90% for training, 10% for validation, and 81 images from the FVC2004 dataset [12] for testing.

Generation of Corrupted Region aims to simulate the noise and corruption that occur in real fingerprints. This is achieved by applying a random configuration that includes line scratches and ellipses. Specifically, for the ellipses representing wrinkly scars, a set of random parameters, denoted as $\{x_0, y_0, E_w, E_l\}$, is generated. Here, (x_0, y_0) represents the center of the ellipse, which can be any point within the fingerprint area, while E_w and E_l correspond to the width and length of the ellipse, respectively. Using these parameters, a closed elliptical mask is created as follows:

$$E(x,y) = \begin{cases} x = x_0 + \frac{E_w}{2}\cos\theta \\ y = y_0 + \frac{E_l}{2}\cos\theta \end{cases} \quad (1)$$

B. Proposed Architecture

Convolutional Autoencoder. Motivated by the architecture employed in previous works [6], [7], [8], we employ a convolutional autoencoder to incorporate a corrupted region. The proposed architecture follows a deep network structure with four routes and includes encoder and decoder stages. The STUA-Net model (see in 2 is composed of three main modules: 1) Sharpen Layer, 2) Swin Transformer Layer, and 3) Soft Collective Attention (SCA). This architecture has been designed to address the challenges of reconstructing corrupted fingerprint images, offering a novel approach for improved performance.

Sharpen Layer. Given the corrupted image $I_c \in \mathbb{R}^{H\times W\times 3}$, where $H \times W$ represents the image resolution, we apply a harmonic extraction [Conv+BN+ReLU] to obtain sharp information. The convolutional filter size is set to $[3\times3]$, which determines the size of the connecting regions. Batch Normalization [13] and ReLU [14] activation are employed to enhance gradient performance and reduce the dependency of gradients on size. The output, denoted as $L_{sharpen} \in \mathbf{R}^{H\times W\times 3}$, is defined as follows:

$$L_{sharpen} = ReLU(BN(Conv(I_c))) \quad (2)$$

Swin Transformer Layer. The typical U-Net system learned spatial information on both the encoder and decoder stages using a convolutional (Conv) layer. In this work, we upgraded each Conv layer with a Swin Transformer Module (STM) [9]. Swin Transformer is constructed by replacing the multi-head self-attention (MSA) module in a Transformer block by a Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA). The consecutive equation 3 and 4 are used for W-MSA and SW-MSA, respectively. The $\hat{z}^l$ and z^l denote the output features of the SW-MSA module and the MLP module in l^th.

$$\hat{z}^l = W - \text{MSA}(\text{LN}(\hat{z}^{l-1})) + z^{l-1} \quad (3)$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \quad (4)$$

$$\hat{z}^{l+1} = SW - \text{MSA}(\text{LN}(\hat{z}^{l-1})) + \hat{z}^l \quad (5)$$

$$z^{l+1} = MLP(LN(\hat{z}^l)) + \hat{z}^l + 1 \quad (6)$$

Encoder Stage. The output from the sharpen layer is initially divided into non-overlapping patches of size $W/s\times W/s$, where $s(=2)$ represents the patch size. Each patch is treated as a "token" and passed through a linear embedding layer, projecting it to a dimension of C. These patch tokens are then fed into the four-stage Swin Transformer, which consists of a series of W-MSA and SW-MSA blocks at different levels. As the network's depth increases, the number of tokens decreases to provide a hierarchical representation. Thus, the output resolutions are $W/s\times W/s$, $W/2s\times W/2s$, and $W/4s\times W/4s$.

TABLE I
EXPERIMENTS WITH DIFFERENT CONFIGURATIONS

Configuration	SCA Enable	Up Sampling	PSNR	SSIM	MSE
Config-1	No	UpSampling2D	16.655	0.916	0.0221
Config-2	No	Conv2DTranspose	17.076	0.932	0.0202
Config-3	Yes	UpSampling3D	18.261	0.936	0.0214
Config-4	Yes	Conv2DTranspose	18.419	0.914	0.0190

TABLE II
COMPARISONS WITH STATE-OF-THE-ART

Methods	PSNR	SSIM	MSE
Corrupted Image	11.561	0.535	0.0373
U-Net [11]	15.871	0.772	0.0202
U-Finger [7]	16.205	0.802	0.0235
FDP-M-Net [8]	16.205	0.863	0.0264
FDP-M-Net [8] w/SCA	16.655	0.916	0.0256
STUA-Net w/o SCA	**17.075**	**0.923**	**0.0202**
STUA-Net w SCA	**18.419**	**0.914**	**0.0190**

Decoder Stage. In each stage of the decoder, the input features are upsampled by 2 and then concatenated with the corresponding skip connection feature maps from the encoder in the same stage. The output is fed into Swin Transformer block. We modified the skip connection by adding Soft Collective Attention (SCA). It can build the dense interac- tion between encoder and decoder to achieve decoding performance. The attention will highlight the important feature and ignore the irrelevant feature. After this stage, it will produce the reconstructed image $I_r \in \mathbb{R}$

Soft Collective Attention: In the decoder stage, when performing upsampling, the recreated spatial information may lack precision. To address this issue, we incorporate skip connections that capture spatial context from both the down and up sampling paths. To further enhance this process, we introduce Soft Collective Attention (SCA) to selectively activate relevant regions and reduce the transmission of redundant features. SCA effectively suppresses noise at both coarse and fine levels in the connections between the encoder and decoder stages. SCA takes two input vectors: $x \in \mathbb{R}^C$, which is copied from the down sampling path, and $g \in \mathbb{R}^C$, obtained from the up sampling path, where C represents the dimension of the feature maps.

On both vectors, the Conv + Dropout $[1 \times 1]$ layers, denoted as ${W_T}^x$ and ${W_T}^g$, are utilized to weight the information in their respective feature maps. The feature maps are cross-weighted to produce feature dependencies q', and a Leaky ReLU activation function $(\sigma(x) = \max(0.1x, x))$ is applied to prevent computational growth. Finally, the output is passed through Conv + Dropout $[1 \times 1]$ $({W_n}^T)$ and Sigmoid activation function $(\sigma_2(x) = \frac{1}{1+e^{-x}})$ to obtain the final output, expressed as:

$$q' = \sigma({W_T}^x * g + {W_T}^g * x) \tag{7}$$

$$q = \sigma_2({W_n}^T * q') \tag{8}$$

Loss Function ($\mathcal{L}$). Per-pixel loss functions like MSE and MAE have become a major concern because these functions only analyze loss per pixel [6]. To address this problem, we combine per-pixel loss with Structural Similarity Index Measurement (SSIM). Then, we determine the optimal combination of loss functions for our experiment setting. The equation of the loss function with weight δ is shown as follows:

$$\mathcal{L} = \delta * \mathcal{L}_{SSIM}(\delta) + (1 - \delta) * \mathcal{L}_{MSE}(\delta) \tag{9}$$

III. EXPERIMENTS AND DISCUSSION

A. Ablation Study

We conducted experiments to evaluate the impact of different components in our proposed method. Specifically, we focused on the SCA attention module and different up-sampling techniques to determine the most suitable architecture for our problem. The experimental scenarios are summarized in Table I.In Config-3, we placed emphasis on the skip connection with SCA as the recommended method. The inclusion of SCA resulted in significant improvements, with an increase in PSNR by 18.261, SSIM by 0.936, and a reduction in MSE by 0.0214. The SCA module effectively highlighted relevant features during training, allowing the model to focus more on key regions, leading to improved performance.Furthermore, in Config-4, we updated the up-sampling technique to Conv2DTranspose, which incorporates a kernel for optimizing the up-sampling process during training. This modification yielded the best evaluation metrics among the different configurations, with a PSNR of 18.419, SSIM of 0.941, and MSE of 0.0190.

These findings demonstrate the importance of the SCA attention module and the choice of up-sampling technique in achieving superior performance for our fingerprint reconstruction problem.

B. Comparisons with State-of-the-art

We conducted a comprehensive comparison of our proposed technique with state-of-the-art methods [7], [8]. In order to ensure a fair evaluation, we adopted their published methods and applied them to our specific scenario and dataset. The results of the comparisons are presented in Table II.

Compared to the previous methods, our STUA-Net demonstrates proficiency in understanding long-range relationships and capturing the contextual information of rigid fingerprint features. It effectively addresses the coarse- and fine-level connections between the encoder and decoder stages. The quantitative experimental results clearly indicate the superiority of our STUA-Net, as we achieved a significant improvement

in the overall evaluation metrics compared to the corrupted image. Specifically, we observed a 37.2% increase in PSNR, a 43.2% increase in SSIM, and a 3.7% decrease in MSE.

To further evaluate the subjective impression of our reconstruction, we provide visualizations of each arrangement in Fig. 3. To facilitate a more precise comparison, we magnified the region within the red box in Fig. 4 by a factor of four for better visualization. Additionally, we trimmed the size of the image to reveal the correct pixel values. Our approach effectively reconstructs fingerprint patterns while eliminating irrelevant patterns, resulting in improved image quality.

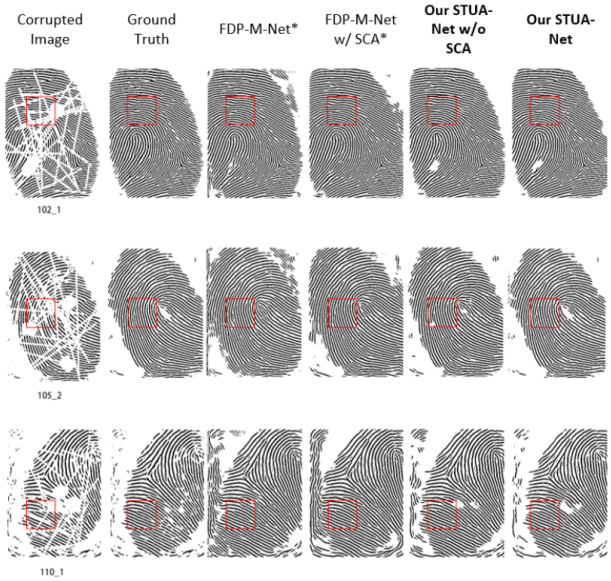


Fig. 3. The reconstruction result of corrupted image..

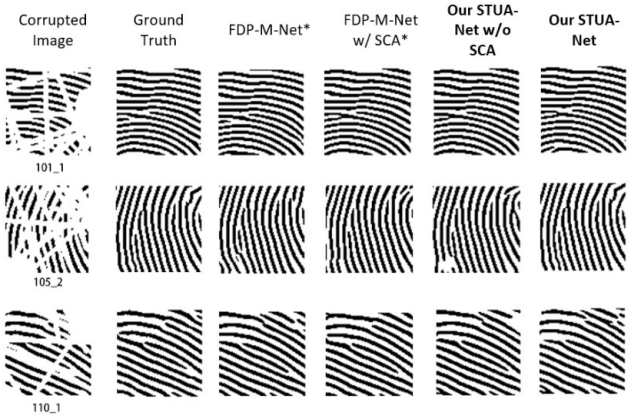


Fig. 4. The reconstruction result of corrupted image at 4 scale.

These comparisons with state-of-the-art methods, along with the subjective visualizations, collectively demonstrate the superiority and effectiveness of our STUA-Net in fingerprint image reconstruction.

IV. Conclusion and Future Work

Fingerprint reconstruction aims to restore the missing regions in fingerprint images. In our approach, we leverage the combination of Swin Transformer and Soft Collective Attention within the U-Net architecture, which we refer to as STUA-Net. Our proposed method yields satisfactory results in both objective and subjective evaluations. The introduced techniques significantly enhance the assessment metrics when comparing the reconstructed images to the corrupted ones.

When it comes to deep learning, depth plays a crucial role. However, deeper networks often come with increased sequential competition and higher latency. To address this issue, we consider the application of a low-cost and efficient fingerprint system. In this regard, A. Goyal et al. [15] proposed the PareNet block as a potential solution. In our future work, we intend to explore and integrate the PareNet block into our fingerprint reconstruction model to further enhance its performance.

References

[1] D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar *et al.*, *Handbook of fingerprint recognition*. Springer, 2009, vol. 2.

[2] F. H. Raswa, F. Halberd, A. Harjoko, C.-T. Lee, Y.-H. Li, P.-C. Chang, and J.-C. Wang, "Leverage limited features of partial fingerprint recognition using improved siamese network with self-spatial attention."

[3] S. H. Hur, H. Kim, Y.-K. Kim, J. M. An, J. H. Lee, S. Kim, Y.-H. Choi, and H. J. Kim, "Discrimination of the geographical origin of dry red pepper using inorganic elements: A multielement fingerprinting analysis," *Journal of Food Composition and Analysis*, vol. 116, p. 105076, 2023.

[4] M. A. Olsen, M. Dusio, and C. Busch, "Fingerprint skin moisture impact on biometric performance," in *3rd International Workshop on Biometrics and Forensics (IWBF 2015)*. IEEE, 2015, pp. 1–6.

[5] F. H. Raswa, F. Halberd, A. Harjoko, C.-T. Lee, Y.-H. Li, J. C. Wang *et al.*, "Multi-loss function in robust convolutional autoencoder for reconstruction low-quality fingerprint image," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 428–431.

[6] J. Bae, H.-S. Choi, S. Kim, and M. Kang, "Fingerprint image denoising and inpainting using convolutional neural network," *Journal of the Korean Society for Industrial and Applied Mathematics*, vol. 24, no. 4, pp. 363–374, 2020.

[7] R. Prabhu, X. Yu, Z. Wang, D. Liu, and A. Jiang, "U-finger: Multi-scale dilated convolutional network for fingerprint image denoising and inpainting," in *Inpainting and Denoising Challenges*. Springer, 2019, pp. 45–50.

[8] S. Adiga V and J. Sivaswamy, "Fpd-m-net: Fingerprint image denoising and inpainting using m-net based convolutional neural networks," in *Inpainting and Denoising Challenges*. Springer, 2019, pp. 51–61.

[9] Z. Liu, Y. Lin, Y. Cao, H. Han, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Hierarchical vision transformer using shifted windows," 2021.

[10] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[12] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, "Fvc2004: Third fingerprint verification competition," in *Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004. Proceedings*. Springer, 2004, pp. 1–7.

[13] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[14] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[15] A. Goyal, A. Bochkovskiy, J. Deng, and V. Koltun, "Non-deep networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6789–6801, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Coarse-Age Loss: A New Training Method Using Coarse-Age Labeled Data for Speaker Age Estimation

Yuki Kitagishi*, Hosana Kamiyama*, Naohiro Tawara*,
Atsunori Ogawa*, Noboru Miyazaki* and Taichi Asami*
* NTT Corporation

***Abstract*—We propose *Coarse-age loss* for a new training method that uses coarse-age labeled speech samples for numerical speaker age estimation. The estimation performance of existing approaches can be degraded by a mismatch in the acoustic characteristics between the training and test phases. The age estimation model must be fine-tuned to suit the test condition. It is easier to collect coarse-age labeled samples, such as ten-year step size, than exact-age labeled speech samples due to privacy concerns. However, special care is needed in using coarse-age labeled samples because they cannot be used as is for training numerical age estimation models. A conventional method converts the coarse-age labels into their median value as tentatively-correct labels (e.g., twenties is converted to 24.5 years old). However, errors between the correct and the tentatively-correct labels negatively influence the model training phase. A key idea of our method proposed here is that the coarse-age labels indicate continuous correct- and incorrect-age ranges. Coarse-age loss trains the model so that it does not misestimate the incorrect range by minimizing a sum of posteriors corresponding to the incorrect range without using exact-age labels. In experiments, the proposed method reduces the mean absolute errors by up to 30.1 % and 10.8 % compared to the baseline and the conventional method, respectively. Also, the results indicated that somewhat broad coarse-age labels, such as the ten-year step, are suitable for the proposed method.**

***Index Terms*—Speaker age estimation, coarse-age label, exact-age label, x-vectors**

I. Introduction

The performance of speaker age estimation is increasing by use of deep neural networks (DNNs) [1]–[11], and application is expected in various service fields[12]–[14]. Speake age estimation can be categorized into two types; categorical and numerical estimations. The categorical estimation uses a classifier that estimates some age groups, such as Child, Young, Adult, and Senior [12]. On the other hand, the numerical estimation uses a regressor or classifier that estimates age values. Modern DNN-based speaker age estimation techniques tackled numerical age estimation [1]–[11].

Most DNNs-based models can achieve impressive high performance, however, the performance can seriously degrade due to the differences in acoustic characteristics between training and test phases. For example, some conventional studies reported that the mismatches in speaking styles and noise conditions caused a serious performance degradation [8], [9], [15], [16]. In spite of this matter, most current studies focus on constructing a general age-estimation model using publicity available speech corpora annotated with exact-age labels, such as TIMIT [17], Fisher English [18], NIST Speaker Recognition Evaluation (SRE) 08/10 [19], [20], and AgeVoxCeleb [7], [21]. Therefore, it is essential to fine-tune or adapt the general model to the target domain using the data collected in the target condition for practical usage.

In actual use, there exists a possibility wherein only speech samples labeled with coarse-age categories, such as five-year step [22], ten-year step size [23], [24], or categorical age labels [12], are available for the model training due to privacy concerns. However, we cannot directly utilize these coarse-age labeled samples as training data in a numerical estimation model. Therefore, careful considerations are required if we intend to utilize the coarse-age labeled samples for training the numerical age-estimation model.

A simple solution is to convert the coarse-age labels into tentatively-correct age labels. For example, Kwasny et al. use median values of ten-year step labels as tentatively-correct age labels (e.g., twenties is converted into 24.5 years old) [3], [25] [1]. However, the error between the correct and the tentatively-correct age labels can degrade performance. Another solution is to make the model train the order relation of the coarse-age labels by using the ranking learning method [26]. However, this approach also handles the samples with different exact-age labels as the same age if they have the same coarse-age label. Therefore, a new training method is necessary that can well handle the difference between the tentatively- and correct age label.

We propose a new loss function, *Coarse-age loss*, to train a DNN-based speaker age-estimation model using coarse-age labeled speech samples without setting tentatively exact-age labels. In the model training phase, Coarse-age loss minimizes the sum of posterior probabilities corresponding to the incorrect-age range obtained by coarse-age labels. This method cannot train the model to estimate the exact-age label. However, using the tentatively exact-age label for the model training phase is unnecessary. Therefore, we expect that our proposed loss will reduce the degradation of the estimation

[1] We confirmed with the authors of [3] that they use the median value as the tentatively-correct age even though it is not specified in their papers

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

performance.

II. Related works

A. Related works about weakly-supervised learning

Our Coarse-age loss is related to a complementary label used in weakly-supervised learning.

In weakly-supervised learning, complementary labels that indicate incorrect classes have been explored [27]–[29]. In typical supervised learning, a DNN classifier is trained to estimate the correct class. Conversely, in weakly-supervised learning with complementary labels, the objective is to train the classifier to avoid misestimating the incorrect classes.

Motivated by this concept of weakly-supervised learning, we propose the Coarse-age loss in our approach. The coarse-age labels serve to indicate the correct-age range as well as the incorrect-age range. Since human age spans a finite range and age values are continuous, the coarse-age labels can effectively represent all the incorrect-age ranges. For instance, a coarse-age label indicating the twenties age range signifies that the correct range is from 20 to 29 years old, while any other age below 19 or above 30 is considered an incorrect range. Thus, the coarse-age label can be viewed as a form of a complementary label.

However, it is important to note that while general classification tasks that use complementary labels often involve discrete relationships among classes, age labels exhibit continuous relationships among themselves. Therefore, we design the Coarse-age loss by specifically considering this characteristic of age labels.

B. Related works about EM-algorithm based method

Kashima et al. proposed an expectation-maximization (EM) algorithm based method using coarsely labeled samples [2] for general regression tasks [30]. In their method, a regressor is optimized so that the predicted output falls within the range of the coarse label based on the EM algorithm. This method is similar to our approach; both their and our approaches aim to have the estimated value/probabilities are fallen within the range of the coase label.

Their experimental results showed the tendency that the estimation performance of their conventional method, which converts the coarse label to the predicted value, and their proposed method were dependent on the datasets. Also, they focused on only adding the coarsely labeled samples to the exact labeled samples.

In this paper, we apply such method to spekaer age estimation for the first time and propose a new loss function for DNN-based models. Alos, we conduct speaker age experiments using both exact- and coarse-age labeled dataset for cross-corpus domain adaptation, in addition to the data addition experiment. Furthermore, we reveal optimal combination of various kinds of ranges of the coarse-age label and our conventional/proposed loss functions.

[2]Kashima et al. called interval target values

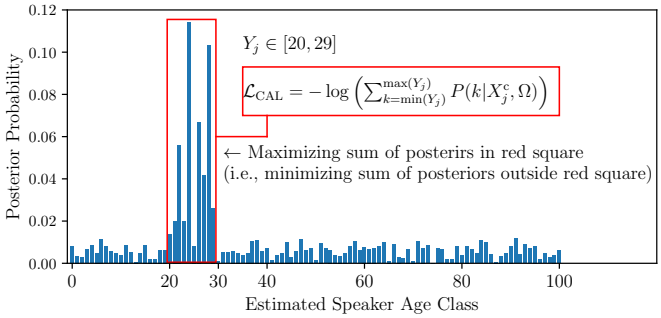


Fig. 1. Proposed Coarse-age loss $\mathcal{L}_{\mathrm{CAL}}$ minimizes sum of posterior probabilities outside red square

III. Speaker Age Estimation

A. Model training using exact-age labeled speech samples

We define a DNN-based speaker age estimation model as a classification model that estimates posterior probabilities for each age value of 0 to 100 years old. DNN-based speaker age estimation models can be categorized into two types; regression [1]–[4] and classification [5]–[11] We use the classification approach because classification models tend to yield better estimation performance than regression models [5].

The classification model is formulated as the estimation of the speaker's actual age y_i from input an acoustic feature sequence X_i. It is defined as,

$$\hat{y}_i = \sum_{k=0}^{100} P(k|X_i, \Omega)k, \tag{1}$$

where $\hat{y}_i$ is the estimated speaker age that is computed as the expected value of speaker age using the posterior probabilities [6], [10], [16], [31], [32], $P(k|X_i, \Omega)$ is the posterior probabilities for each age value $k \in [0, 100]$, and Ω is a set of model parameters. The sum of posterior probabilities is normalized to 1.0 using a softmax function. The set of model parameters Ω is optimized by a stochastic gradient descent algorithm. Recently, some works have trained the model considering the label distribution. For example, they define the actual target as the normal distribution $\mathcal{N}(y_i, \sigma^2)$, where σ is the hyper-parameter [6], [10], [16]. In this paper, we use soft-target cross entropy (SCE) loss [6], defined as,

$$\mathcal{L}_{\mathrm{SCE}} = \frac{1}{I}\sum_{i=1}^{I}\left(-\sum_{k=0}^{100}\mathcal{N}(k|y_i, \sigma^2)\log P(k|X_i, \Omega)\right), \tag{2}$$

where I is the number of exact-age labeled samples in a mini-batch.

B. Model training using coarse-age labeled speech samples

We define a new loss function, *Coarse-age loss* $\mathcal{L}_{\mathrm{CAL}}$, that trains the model as preventing estimation for the incorrect-age classes. In the training phase, the posterior probabilities $P(k|X_j^{\mathrm{c}}, \Omega)$ are estimated from the coarse-age labeled training sample X_j^{c}. Since its exact-age label y_j is unknown, the loss function for the exact-age labeled samples, such as the

SCE loss, cannot be used. Our Coarse-age loss minimizes the posterior probabilities corresponding to the incorrect-age range.

Referring to binary cross-entropy loss, Coarse-age loss is defined as,

$$\mathcal{L}_{\mathrm{CAL}} = \frac{1}{J}\sum_{j=1}^{J} -\log\left(\sum_{k=\min(Y_j)}^{\max(Y_j)} P(k|X_j^{\mathrm{c}},\Omega)\right), \quad (3)$$

where J is the number of coarse-age labeled samples in the mini-batch, X_j^{c} is coarse-age labeled acoustic feature sequence, and Y_j is the set of age values corresponding to the correct-age range indicated by the coarse-age label of X_j^{c}. Coarse-age loss maximizes the sum of posterior probabilities corresponding to the correct-age range. It means minimizing the sum of posterior probabilities corresponding to the incorrect-age ranges since the sum of the posteriors of all age values is become 1.0 by the softmax function. Coarse-age loss realizes the model training without defining tentatively-exact age labels.

Figure 1 shows an example of a case where the coarse age label is defined as twenties (i.e., $Y_j \in [20, 29]$). The posterior probabilities outside a red square correspond to the incorrect-age range. Coarse-age loss is calculated from the posterior probabilities in the red square and minimizes the sum of the posterior probabilities outside the red square.

There are two types of training methods; one is *single-step* model training that uses mixed exact- and coarse-age labeled training samples at the same time. The SCE loss is computed from the exact-age labeled samples, and Coarse-age loss is computed from the coarse-age labeled samples. We use the weighted sum of the SCE and Coarse-age loss; it is defined as,

$$\lambda_{\mathrm{CAL}}\mathcal{L}_{\mathrm{SCE}} + (1-\lambda_{\mathrm{CAL}})\mathcal{L}_{\mathrm{CAL}}, \quad (4)$$

where λ_{CAL} is the weighted value.

The other type is *double-step* model training; first, a model is pre-trained using the exact-age labeled samples, then the model is fine-tuned using the coarse-age labeled samples. In practical use, only the coarse-age labeled speech samples can be collected reliably due to privacy concerns. Training the model by mixing exact- and coarse-age labeled speech samples for each application environment is too expensive. In terms of the cost, it is reasonable to fine-tune the pre-trained model to each application environment using only the coarse-age labeled samples. Then, the SCE and Coarse-age loss are used separately.

IV. Speaker Age Estimation Experiment

A. Experimental setup

1) Dataset: We used SRE08/10 [19], [20], Fisher [18] and AgeVoxCeleb (AVC, [7] as adjusted using [21]). SRE08/10 and Fisher contain channel-separated telephone conversation speech samples generated by two speakers quantized at 16 bits with sampling frequency of 8 kHz. AVC contains human speech extracted from interview videos and was converted into single-channel, 16 bit stream with 16 kHz sampling rate and downsampled to 8 kHz. Table I shows details of each dataset, and Figures 2, 3, and 4 show histograms of the number of speakers per age in each dataset. We divided SRE08 and AVC's training set into training and validation sets; SRE08 was separated using the same method as [5], and AVC were divided randomly at a ratio of 8:2 to avoid duplication of the speakers. The training datasets were augmented by adding noise [33] and convolving room reverberations [34] following the technique of [5].

TABLE I
Details of each dataset

Corpus	Speaking styles	Data subset: train	validation	test
Fisher	telephone conversation	√		
SRE08	generated by	√	√	
SRE10	two speakers			√
AVC (train)	human speech	√	√	
AVC (test)	from interview videos			√

To evaluate the proposed method, we quantized the exact-age labels of each dataset to yield three kinds of coarse-age labels: five-year steps, ten-year steps, and categorical consisting of *Young* (12–24 years old), *Adult* (25–54 years old) and *Senior* (55–90 years old) following [12].

2) System configurations: We employed a time delay neural networks (TDNN)-based model (a.k.a. x-vector architecture [35]); details are shown in Table II. Referring to [5], 23-dimensional MFCCs were extracted with energy-based voice activity detection and normalized by cepstral mean normalization with three-second windows. We used the MFCCs clipped into the chunk size of 3000 frames (30 sec.) randomly in each training phase and full chunk size in each test phase. For model training, we used RAdam optimizer [36] by setting its initial learning rate, β_1, and β_2 to 1e$-$4, 0.9 and 0.999, respectively. The mini-batch size was 32, and the minimum learning rate was 1e$-$6. The learning rate was fell, and training was deemed completed following the method in [5]. The weighted value λ_{CAL} in Equation (4) was set to 0.6.

We employed two kinds of loss functions for the exact-age labeled samples; one is the SCE loss defined in Equation (2). The other is a label distribution learning (LDL) loss consisting of the soft-target, L1, and variance loss functions [10], because [10] indicated that the LDL loss can yield better performance than the soft-target loss. Both the SCE and the LDL loss used the soft target $\mathcal{N}(k|y_i,\sigma^2)$ as the actual target, and their σ values were set to 3.6 and 1.5, respectively. The weight values of LDL loss λ_1, λ_2, and λ_3 in Equation (6) of [10] were set to 1.0, 1.0, and 0.01, respectively.

As the conventional methods, we used the median value of the coarse-age label as the tentatively-correct age [3], and the ranking training method [26]. As the hyper-parameters of the ranking method, λ in Equation (8) of [26] was set to 0.01 with α in Equation (3) of [26] was 2.5, 5.0, and 10.0 for the coarse-age label size of five-years, ten-years and categorical, respectively. All hyper-parameters were decided based on the performance of the validation set.

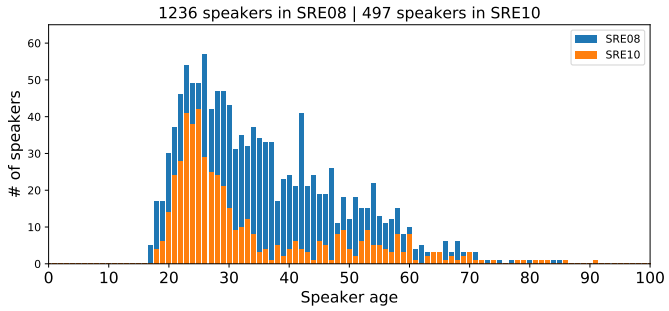


Fig. 2. # of speakers of SRE08/10

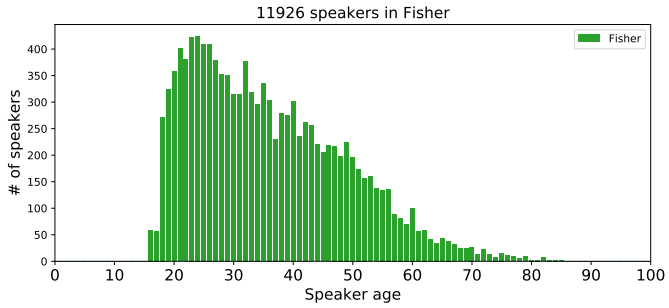


Fig. 3. # of speakers of Fisher

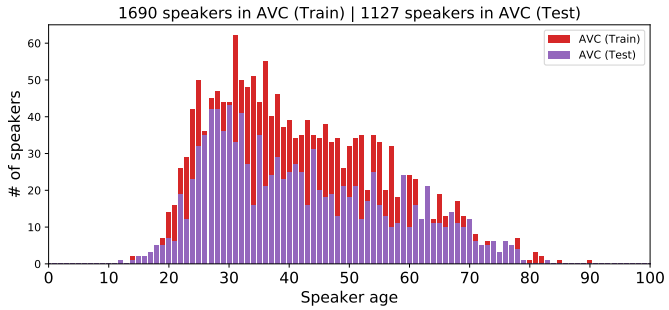


Fig. 4. # of speakers of AgeVoxCeleb

TABLE II
MODEL ARCHITECTURE

	context	output dim.	activation
TDNN 1	$[t-2, t+2]$	512	ReLU + BN
TDNN 2	$\{t-2, t, t+2\}$	512	ReLU + BN
TDNN 3	$\{t-3, t, t+3\}$	512	ReLU + BN
TDNN 4	$\{t\}$	512	ReLU + BN
TDNN 5	$\{t\}$	1500	ReLU + BN
stats. pool	$[0, T)$	1500+1500	—
Dense 1	$\{0\}$	512	ReLU + BN
Dense 2	$\{0\}$	512	ReLU + BN
Output	$\{0\}$	101	Softmax

"BN" denotes Batch Normalization [37]

We used mean absolute error (MAE) and Pearson's correlation coefficient ρ to evaluate estimation performance.

B. In-domain data-addition experiments

We assumed that the exact- and coarse-age labeled datasets were collected in the same domain and made a simulation using exact-age labeled SRE08/10 and coarse-age labeled Fisher. SRE08/10 and Fisher are same-domain datasets since both datasets consist of the telephone conversation (to see Table I). Speech samples of both datasets are used as training samples at the same time. The loss was defined as the weighted sum of the SEC and Coarse-age loss, as defined in Equation (4).

Table III shows the lower bounds of MAE and the upper bounds of ρ values for SRE10 and AVC obtained by the model that was trained using exact-age labeled SRE08 and Fisher. Table IV shows the estimation performance for SRE10 obtained by the model that was trained using exact-age labeled SRE08 and coarse-age labeled Fisher.

1) Results and discussions: The proposed method achieved high performance by using the SCE loss for the exact-age labeled samples, while the coarse-age labels of ten-year and categorical is used. When using the coarse-age labels of ten-year step, the MAE yielded by the proposed method was significantly reduced, by 19.3 % and 3.8 % (6.53/5.48→5.27), relative to the results of upper-bound of the baseline (Table III) and the conventional method using median loss (Table IV), respectively. However, the conventional median loss yielded higher performance than the proposed method when using the coarse-age labels of five-year step size. There is a possibility of a small negative effect on model training because the error between correct and tentatively-correct age was small due to the narrow coarse-age label of five-year step. Our proposed method is effective for coarse-age labels with wider coarse-age windows, such as the ten-year step size or more.

The MAEs were not reduced by the conventional ranking method. However, ρ values were higher than the other two methods; for example, in Table III, ρ was 0.73 when the MAE was 7.96. On the other hand, ρ values were 0.82 and 0.81 for MAEs of 7.12 and 7.54, respectively, see Table IV. It is considered that this method yielded model training that utilized the order relationship shown by the coarse-age labels. Hereafter, we use only median loss [3] as the conventional method in assessing the proposed method.

The proposed method yielded worse results by using the LDL loss for exact-age labeled samples. The LDL loss trains the model by minimizing the variance of the posterior distribution (i.e., rigorous model training). However, our Coarse-age loss trains the model to avoid misestimating the speaker age using coarse-age labeled speech samples (i.e., ambiguous model training). Therefore, our proposed method is incompatible with rigorous training methods, such as LDL loss.

2) Summary of Experiments in Section IV-B: In the in-domain additional-corpus condition, our proposed method for coarse-age labeled samples yielded better estimation performance. There is a tendency that suitable windows of the coarse-age labels are different in the proposed method and the conventional method that uses the median value of the coarse-age labels. The narrow coarse-age label, such as five-year, is suitable for the conventional method. On the other hand, the wider coarse-age labels, such as ten-year or more, are suitable

TABLE III
LOWER BOUND OF MAES AND UPPER BOUND OF ρ VALUES; ONLY EXACT-AGE LABELED SAMPLES WERE USED AS TRAINING SET

Loss for exact-age	Training Dataset SRE08	Fisher	AVC	Test (MAE/ρ) SRE10	AVC
SCE	√			6.53/.80	11.29/.39
	√	√		**5.22/.91**	9.98/.54
			√	7.96/.73	**8.81**/.64
	√		√	6.56/.81	8.83/**.67**
LDL	√			6.59/.80	—
	√	√		6.10/.88	—

TABLE IV
EXPERIMENTAL RESULTS FOR SRE10; EXACT-AGE LABELED SRE08 AND COARSE-AGE LABELED FISHER WERE USED AS TRAINING SET

Loss for exact-age	Loss for coarse-age	Window of coarse-age label 5-year step	10-year step	Categorical
SCE	median [3]	**5.49/.89**	5.48/**.89**	6.40/.87
	ranking [26]	7.12/.82	6.47/.85	7.54/.81
	ours	6.02/.86	**5.27/.89**	**5.71/.88**
LDL	median [3]	**5.47/.90**	**5.84/.88**	8.85/.69
	ranking [26]	6.59/.82	6.58/.82	**6.79**/.80
	ours	6.69/.81	6.66/.80	**6.79/.81**

(MAE/ρ)

for the proposed method. In particular, the proposed method yielded comparable performance to the lower bound (5.27 vs. 5.22 in MAE) when using the coarse-age labels of ten-year step. However, rigorous training methods such as LDL loss, and the ranking training method [26] are not suitable for the proposed method.

C. Experiments on Coarse-age loss and distribution of actual target

We verify the relationship between the label distribution of the actual target and Coarse-age loss, to be clear why the performance was degraded when the LDL and Coarse-age loss were combined. In the previous section, using the LDL loss degraded the estimation performance compared to using the SCE loss. We used exact-age labeled SRE08, and exact- and coarse-age labeled Fisher as training samples. The window of the coarse-age label was set to ten-years. We used only the SCE loss for the exact-age labeled samples. Table V shows the MAEs for SRE10 with σ value of SCE loss ranging from 0.0 to 5.0.

The MAEs obtained by the model that trained using exact-age labeled SRE08 and coarse-age labeled Fisher were degraded compared to the baseline (2nd row of Table V) as the σ value decreased. The LDL loss optimizes the model by minimizing the variance of the posterior distribution using a small σ value and the variance loss to yield rigorous speaker age estimations. On the other hand, Coarse-age loss optimizes the model so as not to misestimate speaker age, that is, it is a form of ambiguity model training. The performance was degraded because rigorous training methods, such as SCE loss with small σ values, and LDL loss and ambiguous training methods, such as Coarse-age loss, are incompatible.

Hereafter, we use SCE loss to train the model with exact-age labeled speech samples.

D. Domain adaptation learning experiments

1) Experimental configurations: We assume a practical situation where only a coarse-age labeled speech dataset is available for the training phase and verify the effectiveness of Coarse-age loss. We use SRE08/10 and AVC as mutually out-of-domain datasets since these datasets consist of different speaking styles. Experiments have shown that the estimation performance falls in the cross-corpus condition using SRE08/10 and AVC [9], [15], [16]. We evaluate the proposed method using two training methods; the single- and the double-step model training, as described in Section III-B.

Experimental results are shown in Table VI and VII. The baseline performance is the upper bound obtained the model that was trained using only exact-age labeled samples, shown in Table III. Table VI shows the performance for AVC obtained by the model that was trained using exact-age labeled SRE08 and coarse-age labeled AVC. Table VII shows the performance for SRE10 obtained by the model that was trained using exact-age labeled AVC and coarse-age labeled SRE08.

2) Results and Discussions of single-step model training: The MAEs for the out-of-domain dataset of the exact-age labeled samples were reduced by using our proposed method. Comparing the upper bound, the proposed method reduced MAEs by up to 31.1 % in AVC when the model was trained using exact-age labeled SRE08 and coarse-age labeled AVC (11.29→8.77/7.78/9.51, to see Table III and 2nd row of Table VI), and by up to 21.23 % in SRE10 when the model was trained using exact-age labeled AVC and coarse-age labeled SRE08 (7.96→6.98/6.27/7.74, to see Table III and 2nd row of Table VII). Also, the conventional method reduced the MAEs by up to 13.1 % in AVC (11.29→8.59/8.76/9.81, to see Table III and 1st row of Table VI), and by up to 21.2 % in SRE10 (7.96→6.28/6.40/7.89, to see Table III and 1st row of Table VII), in same training condition above, respectively.

Similar to the experiment in Section IV-B and Table IV, there is the tendency that the conventional method outperformed the proposed method when using the coarse-age labels with five-year step size. The MAEs for AVC obtained by the conventional and proposed method using exact-age labeled SRE08 and coarse-age labeled AVC were comparable (8.59 vs. 8.77, to see 1st and 2nd rows of Table VI) The proposed method could not reduce the MAE for SRE10 relative to the conventional method using exact-age labeled AVC and coarse-age labeled SRE08 (6.28→6.98, to see 1st and 2nd rows of Table VII).

Similarly, there is the tendency that the proposed method outperformed the conventional method when using coarse-age labels of ten-year step and categorical. The conventional and proposed methods achieved comparable MAEs for AVC using exact-age labeled SRE08 and coarse-age labeled AVC of the ten-year step (8.76 vs. 8.78, to see Table VI). The proposed method yielded reduced MAE for AVC compared to the conventional method using exact-age labeled SRE08 and coarse-age labeled AVC of categorical (9.81→9.51, reduced by 3.1 %). The proposed method reduced the MAEs for SRE10

TABLE V
MAEs FOR SRE10 WITH DIFFERENT σ VALUES OF SOFT-TARGET CROSS-ENTROPY LOSS

Training datasets			σ value of SCE loss										
Method	SRE08	Fisher	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.6	4.0	4.5	5.0
baseline	exact-age label	—	9.65	9.24	8.71	8.68	6.64	6.54	7.22	**6.53**	7.25	6.56	7.07
baseline	exact-age label	exact-age label	5.57	5.43	**5.18**	5.24	5.23	5.30	5.32	5.22	5.29	5.28	5.29
ours	exact-age label	coarse-age label	7.43	7.23	6.99	5.53	5.43	5.46	5.38	**5.27**	5.42	5.35	5.37

TABLE VI
MAEs AND ρ VALUES FOR AVC; EXACT-AGE LABELED SRE08 AND COARSE-AGE LABELED AVC WERE USED AS TRAINING SET

Training step	Loss for coarse-age label	Window of coarse-age label		
		5-year step	10-year step	categorical
single	median [3]	**8.59/.65**	**8.76/.69**	9.81/.57
	ours	8.77/**.65**	8.78/.64	**9.51/.59**
double	median [3]	**8.77/.66**	9.36/**.64**	10.00/.52
	ours	8.97/.63	**8.94/.64**	**9.47/.58**

TABLE VII
MAEs AND ρ VALUES FOR SRE10; EXACT-AGE LABELED AVC AND COARSE-AGE LABELED SRE08 WERE USED AS TRAINING SET

Training step	Loss for coarse-age label	Window of coare-age label		
		5-year step	10-year step	categorical
single	median [3]	**6.28/.83**	6.40/.81	7.89/.76
	ours	6.98/.76	**6.27/.83**	**7.74/.78**
double	median [3]	**7.50**/.72	8.16/.72	13.36/.58
	ours	8.07/**.75**	**7.34/.76**	**11.64/.59**

compared to the conventional method using exact-age labeled AVC and coarse-age labeled SRE08 with the ten-year step size (6.40→6.27, reduced by 2.0 %) and categorical (7.89→7.74, reduced by 1.9 %, to see Table VII).

3) Results and Discussions of the double-step model training: This experiment yielded the same trend as single-step model training. In particular, the proposed method attained the greatest reduction in MAE when using the coarse-age labels of the ten-year step size. The MAEs were significantly better, 20.8 % in AVC and 7.8 % in SRE10 than the results of baseline (11.29/7.96→8.94/7.34, to see Table III and 4th row of Tables VI and VII), and 4.5 % in AVC and 10.0 % in SRE10 than the results of the conventional method in the target domain (9.36/8.16→8.94/7.34, to see 3rd and 4th row of Tables VI and VII).

However, the performance obtained by the proposed method with double-step training was degraded compared to single-step training. This runs counter to the expectation that double-step model training should be superior as it used only the target-domain dataset in fine-tuning the model. This is explained by the catastrophic forgetting caused by using only coarse-age labeled speech samples with Coarse-age loss. The model became unable to estimate the age rigorously.

4) Summary of Experiments in Section IV-D: In the out-of-domain cross-corpus condition, both the proposed and conventional methods indicated that they could adapt the model to the condition of coarse-age labeled dataset. Similar to the in-domain additional-corpus experiment in Section IV-B, suitable windows of the coarse-age labels for the conventional and proposed method are different; the narrow window is suitable for the conventional method and the wider windows are suitable for the proposed method.

We conducted two training methods considering model training cost; one was the single-step that uses both exact- and coarse-age labeled datasets at the same time. The other was the double-step that fine-tuned the model using only the coarse-age labeled dataset. The model has been pre-trained using only the exact-age labeled samples. Both training methods indicated reducing the MAEs for the domain of the coarse-age labeled dataset. However, the estimation performance obtained by the single-step were better than the double-step because of catastrophic forgetting.

V. CONCLUSIONS AND FUTURE WORKS

We proposed a new model training method with *Coarse-age loss* for numerical speaker age estimation using coarse-age labeled speech samples. Most current works using deep neural networks train a general age-estimation model using publicly available exact-age-labeled speech samples. In practical environments, collecting coarse-age labeled samples is easier than the exact-age labeled speech samples due to speakers' privacy concerns. However, using coarse-age labeled samples makes special care in training the model necessary since existing numerical speaker age estimation techniques require exact-age labels.

We focused on coarse-age labels as they indicate the correct- and incorrect-age ranges. Coarse-age loss minimizes the sum of posterior probabilities corresponding to the incorrect-age ranges. Therefore, unlike conventional methods, Coarse-age loss does not require exact-age labels. We mainly conducted two experiments; one used both exact- and coarse-age labeled speech samples in the same speaking style. Our proposed method reduced the mean absolute errors (MAEs) by 19.3 % and 3.8 % compared to the baseline and the conventional method, respectively. The other used the exact- and coarse-age labeled speech samples in different speaking styles. This simulated the situation in which acoustic characteristics are an out-od-domain condition between the training and actual use environments. In this experiment, the proposed method reduced the MAEs by up to 30.1 % and 10.8 % compared to the baseline and the conventional method, respectively. Also, the results indicated that the proposed method is more effective for somewhat broad age windows such as ten-year, or more step size. In future work, we verify by this mixing various types of coarse-age labeled samples.

ACKNOWLEDGMENT

We thank Mr. Fuga Nakamura who has completed Graduate School of Nagaoka University of Technology in 2023. He had been an intern at NTT and conducted the experiments.

REFERENCES

[1] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, pp. 22 524–22 530, 2018.

[2] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "A deep neural network based end to end model for joint height and age estimation from short duration speech," in *ICASSP. IEEE*, 2019, pp. 6580–6584.

[3] D. Kwasny and D. Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors*, vol. 21, no. 14, 2021.

[4] T. Gupta, D.-T. Truong, T. T. Anh, and C. E. Siong, "Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model," in *INTERSPEECH*, 2022, pp. 1978–1982.

[5] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, "End-to-end deep neural network age estimation," in *INTERSPEECH*, 2018, pp. 277–281.

[6] Y. Kitagishi, H. Kamiyama, A. Ando, N. Tawara, T. Mori, and S. Kobashikawa, "Speaker age estimation using age-dependent insensitive loss," in *APSIPA. IEEE*, 2020, pp. 319–324.

[7] N. Tawara, A. Ogawa, Y. Kitagishi, and H. Kamiyama, "Age-vox-celeb: Multi-modal corpus for facial and speech estimation," in *ICASSP. IEEE*, 2021, pp. 6963–6967.

[8] R. Takeda and K. Komatani, "Age estimation with speech-age model for heterogeneous speech datasets," in *INTERSPEECH*, 2021, pp. 4164–4167.

[9] A. Saraf, G. Sivaraman, and E. Khoury, "Confidence measure for automtic age estimation from speech," in *INTERSPEECH*, 2022, pp. 2033–2037.

[10] S. Si, J. Wang, J. Peng, and J. Xiao, "Towards speaker age estimation with label distribution learning," in *ICASSP. IEEE*, 2022, pp. 4618–4622.

[11] Z. Kang, J. Wang, J. Peng, and J. Xiao, "SVLDL: Improved speaker age estimation using selective variance label distribution learning," in *SLT. IEEE*, 2022, pp. 1037–1044.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardr, L. Devillers, C. Muller, and S. Narayanan, "The Interspeech 2010 paralinguistic challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[13] D. C. Tanner and M. E. Tanner, *Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection*. Lawers & Judges Publishing, 2004.

[14] T. Pellegrini, V. Hedayati, I. Trancoso, A. Hämäläinen, and M. S. Dias, "Speaker age estimation for eldery speech recognition in european portuguese," in *INTERSPEECH*, 2014, pp. 2962–2966.

[15] N. Tawara, A. Ogawa, Y. Kitagishi, H. Kamiyama, and Y. Ijima, "Robust speech-age estimation using local maximum mean discrepancy under mismatched recording conditions," in *ASRU. IEEE*, 2021.

[16] A. Saraf and E. Khoury, "Distribution learning for age estimation from speech," in *ICASSP. IEEE*, 2022, pp. 8552–8556.

[17] J. S. Garofolo, L. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[18] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *LREC*, 2004.

[19] S. S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," in *ICASSP. IEEE*, 2009, pp. 4205–4208.

[20] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *INTERSPEECH*, 2010, pp. 2726–2729.

[21] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "Voxceleb enrichment for age and gender recognition," in *ASRU. IEEE*, 2021, pp. 687–693.

[22] K. Maekawa, H. Kikuchi, and W. Tsukahara, "Corpus of spontaneous Japanese: Design, annotation, and xml representation," in *LKR*, 2004.

[23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, *Common voice: A massively-multilingual speech corpus*, arXiv:1912.06670, 2019.

[24] NII-SRC, *ASJ Japanese newspaper article sentences read speech corpus (JNAS)*, http://research.nii.ac.jp/src/en/JNAS.html.

[25] D. Kwasny and D. Hemmerling, *Joint gender and age estimation based on speech signals using x-vectors and transfer learning*, arXiv:2012.01551, version 1, 2020.

[26] H. Pan, H. Han, S. Shan, and X. Chen, "Revised contrastive loss for robust age estimation from face," in *ICPR*, 2018, pp. 3586–3591.

[27] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," in *NIPS*, 2017.

[28] L. Feng, T. Kaneko, B. Han, G. Niu, and M. Sugiyama, "Learning with multiple complementary labels," in *ICML*, 2020.

[29] D. Liu, J. Ning, J. Wu, and G. Yang, "Extending ordinary-label learning losses to complementary-label learning," *IEEE Signal Process Lett.*, vol. 28, pp. 852–856, 2021.

[30] H. Kashima, K. Yamasaki, A. Inokuchi, and H. Saigo, "Regression with interval output values," in *ICPR*, 2008, 4 pages. DOI: 10.1109/ICPR.2008.4761270.

[31] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep expectation of apparent age from a single image," in *ICCV. IEEE*, 2015, pp. 10–15.

[32] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognit.*, vol. 72, pp. 15–26, 2017.

[33] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[34] *Simulated room impulse response database*, https://www.openslr.org/26/.

[35] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP. IEEE*, 2018, pp. 5329–5333.

[36] L. Liu, H. Jiang, P. He, C. Weizhu, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *ICLR*, 2020.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of International Conference on International Conference on Machine Learning*, 2015, pp. 448–456.

Contribution of modulation spectral features for cross-lingual speech emotion recognition under noisy reverberant conditions

Taiyang Guo, Sixia Li, Shunsuke Kidani, Shogo Okada and Masashi Unoki
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {guotaiyang, lisixia, kidani, okada-s, unoki}@jaist.ac.jp

***Abstract*—Handling multiple languages under noisy reverberant conditions has become increasingly important for speech emotion recognition (SER). Previous studies found that modulation spectral features (MSFs) are robust to noisy reverberant conditions for SER. However, they mainly focused on specific languages; the universality of MSFs among languages is still unclear. To address this issue, we compared MSFs, hand-crafted features, Wav2Vec2.0-based features, MSFs+hand-crafted features for SER on four languages under 12 noisy reverberant conditions. Intra-lingual results showed that MSFs+hand-crafted features performed best on most conditions of all languages. Inter-lingual results showed that MSFs performed best on most conditions of test languages except training on a tonal language and testing on others. The results demonstrate that MSFs are robust to multilingual SER under noisy reverberant conditions and suggest that MSFs are potentially language-independent features for nontonal languages.**

I. Introduction

Speech-emotion recognition (SER) has become an increasingly important technique in a wide range of fields, including psychology for detecting emotional disorders. In real-world applications, noise and reverberation can have a significant impact on the quality and clarity of speech signals [1]–[5].

To address the challenge of the lack of robustness against noise and reverberation in SER, researchers have proposed methods based on speech signal processing to improve SER. Modern physiological [6] and psychological [7] models suggest that the human auditory system contains an auditory filterbank that decomposes speech signals into channel signals, such as the temporal amplitude envelope (TAE) and temporal fine structure. Several studies have proved that the TAE and its modulation cues play an important role in speech recognition [8]–[10]. Based on these findings, a previous study proposed modulation spectral features (MSFs) that were extracted based on the modulation analysis processing of the TAE and showed that MSFs are important cues for human emotion recognition [11]. Furthermore, MSFs were shown to be more robust on SER under noisy and reverberant conditions compared to widely used feature sets IS09 and eGeMAPS [12], [13].

So far, studies on MSFs were mainly focused on specific languages, such as Japanese [12], [13] and German [14]. However, handling multiple languages is also important for SER [15], [16]. Whether or not MSFs retain their robustness in multilingual SER and remain unaffected by noise and reverberation is still unclear. On the other hand, inspired by human perception, researchers have recently proposed to use spectral features with the three-layer model for multilingual emotion recognition tasks [17]–[19]. This suggests that spectral features inspired by human perception, such as MSFs, may reflect the essential characteristic of emotions under noisy reverberant conditions among languages.

Accordingly, this study investigated the potential universality of MSFs in intra-lingual and inter-lingual SER under noisy and reverberant conditions. To do so, we used support vector machine (SVM) to compare the SER performance of MSFs, hand-crafted feature sets, MSFs + hand-crafted feature sets, and Wav2Vec2.0-based features on different languages under noisy reverberant conditions. Specially, we investigated intra-lingual and inter-lingual SER performances on recognizing four common emotions (happy, neutral, sad, and angry) in four languages (Japanese, German, English and Chinese) under 12 noisy reverberant conditions. To the best of our knowledge, we are the first to investigate the robustness of MSFs in multilingual and cross-lingual SER under noisy reverberant conditions. The findings of this study will not only advance the understanding of the importance of MSFs in the speech perception mechanism, but also contribute to the development of robust SER systems, such as speech-based human-machine interaction under noisy reverberant conditions.

II. Modulation spectral features

We utilized the same MSFs that were shown to be robust to noise and reverberation in previous studies [11]–[13]. To extract modulation spectral cues such as MSFs of modulation spectrogram, it is necessary to calculate the modulation spectrogram using the modulation filterbank first. Figure 1 shows the auditory-based process used in this study to calculate the MSFs of modulation spectrograms.

Emotional speech signals s were divided into several frequency bands by using an auditory-based band-pass filterbank,

$$s_k(n) = s(n) * h_k(n), \tag{1}$$

where * denotes the convolution operator, $h_k(n)$ is the impulse response of the k^{th} channel, and n is the sample number in

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

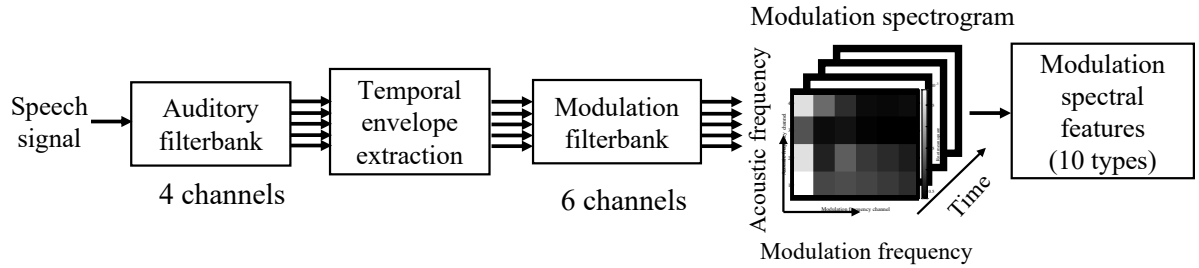


Fig. 1. Process of extracting MSFs [13].

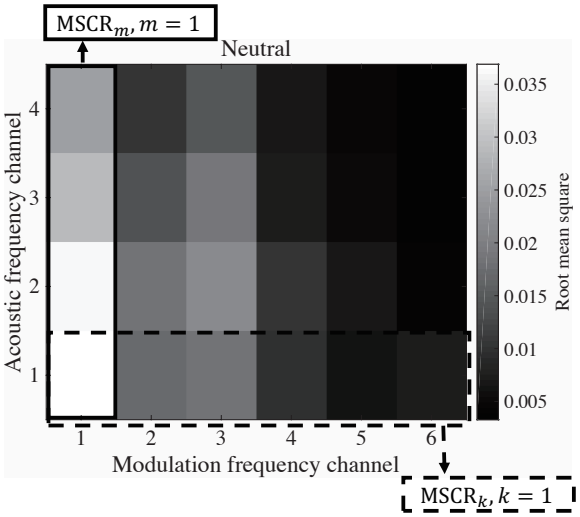


Fig. 2. Calculation example of MSCR$_m$ and MSCR$_k$ [11].

the time domain. The 6^{th}-order Butterworth infinite impulse response (IIR) band-pass filterbank was used as the auditory filterbank. The bandwidth of each filter was the bandwidth of the human auditory filter, and the order of the filters was determined in accordance with the equivalent rectangular bandwidth (ERB_N) and ERB_N-number scale, where the unit of ERB_N-number is Cam. The boundary frequencies of the band-pass filters (BPFs) were defined as ERB_N-number from 3 to 35 Cam with an 8 ERB_N bandwidth, and the number of channels was 4.

The temporal envelope of the output signal from each BPF $s_k(n)$ was extracted using the Hilbert transformation, and a 2^{nd}-order Butterworth IIR low-pass filter (LPF) (cut-off frequency is 64 Hz) as follows,

$$e_k(n) = \text{LPF}\left[|s_k(n) + j\mathcal{H}[s_k(n)]|\right], \tag{2}$$

where $\mathcal{H}$ denotes the Hilbert transform.

The next step involved decomposing the temporal envelope into several modulation-frequency bands by using a modulation filterbank,

$$E_{k,m}(n) = g_m(n) * (e_k(n) - \overline{e_k(n)}), \tag{3}$$

where m is the channel number of the modulation filter, $g_m(n)$ is the impulse response of the modulation filterbank, and $\overline{e_k(n)}$ is the time-averaged amplitude of $e_k(n)$. The modulation filterbank consisted of six filters (one LPF and five BPFs). The boundary frequencies of the filters were spaced on an octave frequency band from 2 to 64 Hz.

The root-mean-square of $E_{k,m}(n)$ is calculated as the modulation spectrogram,

$$\bar{E}_{k,m}(n) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} E_{k,m}^2(n)}, \tag{4}$$

where the N is the length of the speech signal $s(n)$.

Then, we calculated the high-order statistics of the three-dimensional modulation spectrograms (time, acoustic frequency, modulation frequency) in the acoustic frequency direction and modulation frequency direction as MSFs. The ten types of MSFs are the modulation spectral features in the acoustic frequency domain (the subscript is m) and in the modulation frequency domain(the subscript is k): the modulation spectral centroid (MSCR$_{m/k}$), modulation spectral spread (MSSP$_{m/k}$), modulation spectral skewness (MSSK$_{m/k}$) and modulation spectral kurtosis (MSKT$_{m/k}$), which are defined as follows.

$$\text{MSCR}_\text{m} = \frac{\sum_{k=1}^{K} k\bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{5}$$

$$\text{MSSP}_\text{m} = \frac{\sum_{k=1}^{K} [k - MSCR_m]^2 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{6}$$

$$\text{MSSK}_\text{m} = \frac{\sum_{k=1}^{K} [k - MSCR_m]^3 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{7}$$

$$\text{MSKT}_\text{m} = \frac{\sum_{k=1}^{K} [k - MSCR_m]^4 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{8}$$

$$\text{MSCR}_\text{k} = \frac{\sum_{m=1}^{M} m\bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \tag{9}$$

$$\text{MSSP}_\text{k} = \frac{\sum_{m=1}^{M} [m - MSCR_k]^2 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \tag{10}$$

$$\mathrm{MSSK_k} = \frac{\sum_{m=1}^{M}[m - MSCR_k]^3 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \quad (11)$$

$$\mathrm{MSKT_k} = \frac{\sum_{m=1}^{M}[m - MSCR_k]^4 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \quad (12)$$

The final two MSFs, the modulation spectral tilts ($\mathrm{MSTL}_{m/k}$) are the linear regression coefficients obtained by fitting the first-degree polynomial to the modulation spectrograms. Figure 2 shows an example of calculating MSCR_m and MSCR_k.

III. Noisy reverberant conditions

In this study, we used the same method as in previous studies [11]–[13] to add noise and reverberation to speech to simulate noisy, reverberant, and noisy reverberant conditions.

For noisy conditions, we added white Gaussian noise and the adjusted noise to the speech at SNR levels of 20, 10, and -5 dB. The noise conditions were three in total. For reverberant conditions, we used the convolution method and a statistical room impulse response (Schroeder model). We convoluted the original speech with two types of room-impulse responses, with reverberation times (T_{R}) of 1.0 and 2.0 s, respectively. The reverberation conditions were two in total. For noisy reverberant conditions, we used the combination of the above three noise conditions and two reverberation conditions. The noisy reverberant conditions were six in total. With the clean condition (original speech), we conducted experiments under 12 conditions in total.

IV. Experiments

A. Datasets

We used datasets of Japanese, German, Chinese, and English in this study. We chose happy, neutral, sad, and angry emotions in these datasets, because these emotions are common emotion categories that guarantee the emotion categories in each dataset are the same for experiments. Table 1 shows the statistics of utterances of emotions in each dataset.

Fujitsu dataset: The Fujitsu Japanese Emotional Speech Database is a Japanese dataset that was used for emotion recognition [11]–[13]. **Berlin EmoDB dataset**: This is a German dataset, which has also been widely used in previous research [18]–[20]. For convenience, we will refer to Berlin EmoDB as Berlin for short. **IEMOCAP dataset**: IEMOCAP is an interactive emotional dyadic motion capture database in English [21], [22]. For emotion categories, we combine the emotions excited and happy together as the emotion happy, as was done in previous studies [23], [24]. **CASIA dataset**: The CASIA emotional speech dataset [18], [19], [22] is a Mandarin emotional speech dataset. This dataset does not contain the emotion of happy, but it has joy, a closely related emotion. Accordingly, we use joy as the emotion happy for this dataset.

TABLE I
Statistics of utterances of emotions in each dataset

Dataset	Happy	Neutral	Sad	Angry	Total
Fujitsu	20	20	20	20	80
Berlin	71	79	62	127	339
IEMOCAP	1636	1708	1084	1130	5531
CASIA	1600	400	1599	1597	5196

B. Baseline feature sets

Hand-crafted feature sets: InterSpeech2009 (IS09) [25] and eGeMAPS [26] are two widely used hand-crafted feature sets in the emotion recognition area. We used two feature sets separately in our experiments.

Wav2Vec2.0-based feature: Wav2Vec2.0 [27] is a self-supervised learning pre-trained model for representing speech. We used Wav2Vec2.0-large-xlsr models that were fine-tuned on each language to make fair comparisons. These models were obtained from jonatasgrosman/Wav2Vec2.0-large-xlsr-53-[Japanese, German, English, Chinese] from HuggingFace. To make a fair comparison with MSFs, we did not fine-tune Wav2Vec2.0 models on SER tasks but froze those models to use them as feature extractors only. For feature extractions, as Wav2Vec2.0's output is a sequence of vectors corresponding to frames sequence, we refered to previous studies [28], [29] to obtain fixed-length features for each speech. Specifically, we computed the mean of the last layer, the sum of the last layer, the mean of all layers, and the sum of all layers as four different features. Each feature is a 1024-dimension vector.

C. MSFs

In this study, we investigated the performance of each of the ten types of MSFs and the four combinations of MSFs. Combination features include the combination of all types of MSFs, the combination of acoustic-frequency domain MSFs, the combination of modulation-frequency domain MSFs, and the modulation spectral tilts (MSKT_k and MSTL_k). The last two features have been demonstrated to be robust to all daily noise and reverberation conditions in a previous study [13]. We concatenated corresponding MSFs together to obtain the combination of features.

Moreover, since MSFs are hand-crafted features, we also used MSFs + hand-crafted feature sets to investigate whether or not MSFs are complementary to conventional feature sets. Specifically, we concatenated all types of MSFs with IS09 and all types of MSFs with eGeMAPS as two kinds of MSFs + hand-crafted features.

D. Intra-lingual and inter-lingual experiment setting

We used SVM as a classifier to conduct experiment recognition experiments, which has also been widely used in many studies related to SER (as examples, see [30] and [31]). The input was the feature vector of MSFs, MSFs + hand crafted, hand crafted, and Wav2Vec2.0-based features. The output was the emotion from the four emotion classes.

For the intra-lingual experiment, we performed five-fold experiments to comprehensively evaluate the performance. Specifically, we split the data of each emotion into a five-fold average for each dataset. Then, we performed five experiments by using four of the folds as the training set and the remaining

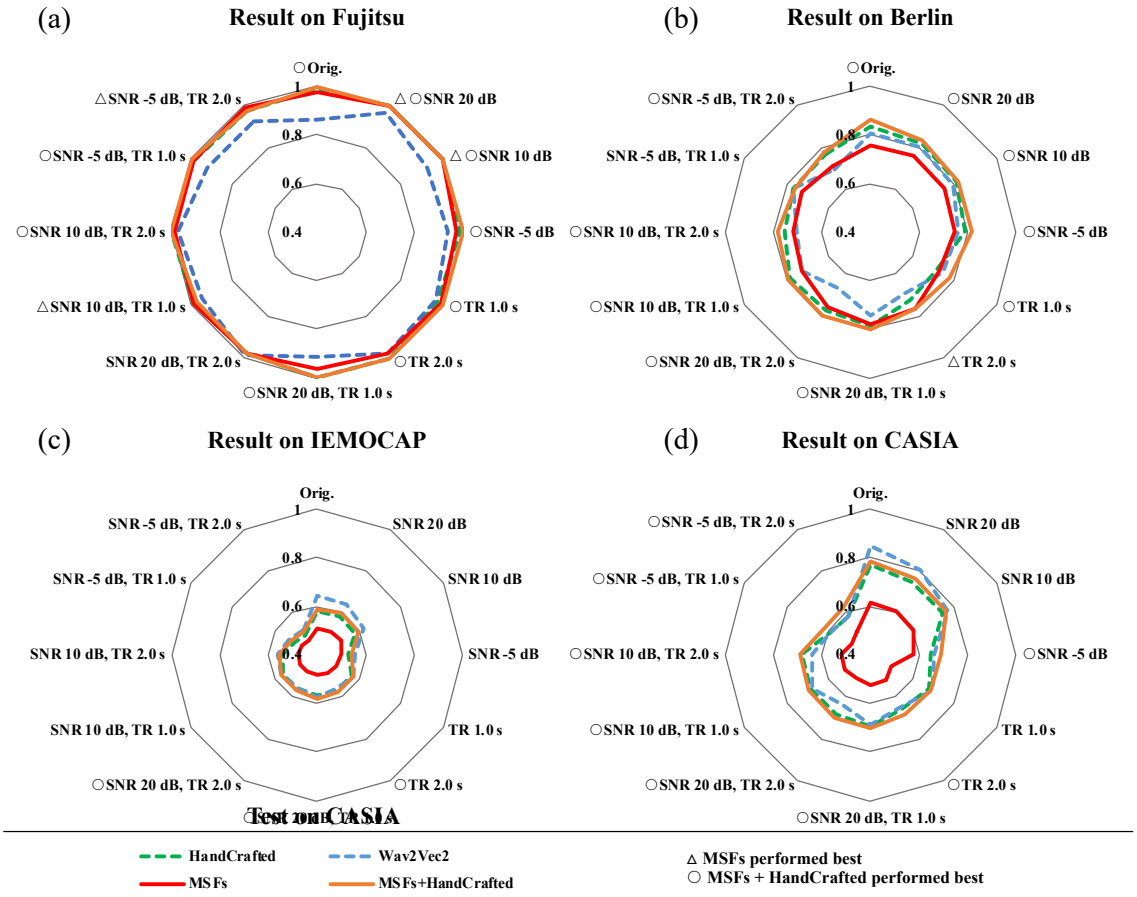


Fig. 3. Result of intra-lingual SER on four datasets: (a) result of train on Fujitsu and test on Fujitsu, (b) result of train on Berlin and test on Berlin, (c) result of train on IEMOCAP and test on IEMOCAP, and (d) result of train on CASIA and test on CASIA.

one as the test set. Moreover, we splited 80% of the training set for training and the remaining 20% for validation. After using each fold as the test set, we used the average performance of the five experiments for evaluation.

For the inter-lingual experiment, we trained the model on one dataset and tested it on the other three datasets separately. Specifically, we used 80% of each emotion's data from the training dataset as the training set and the remaining 20% of each emotion's data for validation. All data from test datasets were used for the test. We conducted experiments by training the model on each dataset and testing it on other datasets to evaluate the performance.

We used weighted accuracy that considers the imbalance of each emotion as the evaluation metric.

V. Results and Discussion

Figure 3 and 4 show radar charts of intra-lingual and inter-lingual results. Different axes represent the original speech and 12 noise reverberant conditions; values on axes represent the weighted accuracy. To show comparisons between four feature sets clearly, and due to space limitations, we only show the best result in categories of hand-crafted feature sets, Wav2Vec2.0-based features, MSFs, and MSFs + hand-crafted feature sets. For convenience, we hereafter refer to these categories as hand-crafted, Wav2Vec2.0-based, MSFs, and MSFs + hand-crafted. In Figure 3 and 4, different colored lines represent different feature categories: the green dashed line, blue dashed line, red line, and orange line indicate the highest weighted accuracy from hand-crafted, Wav2ec2-based, MSFs, and MSFs + hand-crafted, respectively. Conditions with △ indicate conditions where MSFs performed the best. Conditions with ◯ indicate conditions where MSFs + hand-crafted performed the best.

As seen in Fig. 3, MSFs + hand-crafted performed best on 9 of 12 conditions on Fujitsu, 10 of 12 conditions on Berlin, and 8 of 12 conditions on CASIA, with average improvements from the best baseline of 0.69%, 2.03%, and 1.85% on Fujitsu, Berlin, and CASIA, respectively. The reason MSFs + hand-crafted improved relatively less is because the best baselines and MSFs + hand-crafted have 100% weighted accuracy on many conditions in the Fujitsu dataset. On the other hand, MSFs + hand-crafted performed best on 3 of 12 conditions on IEMOCAP, while Wav2Vec2.0-based performed best on more conditions in IEMOCAP. The reason could be that Wav2Vec2.0 was originally pre-trained in English, and it captures characteristics of emotions in English well. However,



Fig. 4. Result of inter-lingual SER that train the model on a single language and test on other languages: (a) result of train on Fujitsu and test on other datasets, (b) result of train on BERLIN and test on other datasets, (c) result of train on IEMOCAP and test on other datasets, and (d) result of train on CASIA and test on other datasets.

although Wav2Vec2.0 models were fine-tuned on languages other than English, MSFs + hand-crafted performed better than Wav2Vec2.0-based on most conditions of other languages. Meanwhile, Wav2Vec2.0 is a large model that needs more computational resources, while MSFs + hand-crafted can be effectively obtained. These results demonstrated that MSFs are complementary with conventional hand-crafted features and are effective features that are robust in noisy reverberant conditions in different languages.

As seen in Fig. 4, MSFs performed best on most of the conditions when training on Fujitsu and testing on other datasets, training on Berlin and testing on Fujitsu and CASIA, and training on IEMOCAP and testing on other datasets, training on CASIA and testing on Fujitsu. MSFs + hand-crafted performed best on most conditions when training on Berlin and testing on IEMOCAP. These results demonstrated that MSFs are robust in emotion recognition in noisy reverberant conditions across languages. On the other hand, MSFs or MSFs

+ hand-crafted performed best on relatively fewer conditions when training on CASIA and testing on Berlin and IEMOCAP. Since Chinese was the only tonal language in our experiments, the emotion perception in Chinese may be influenced by tone and other reasons. These results suggested that MSFs potentially describe common characteristics of emotions under noisy reverberant conditions among non-tonal languages.

For the MSFs combination that performed best in the results, among all 48 conditions (12 conditions $\times$ 4 datasets) in multilingual SER, features containing MSFs performed best on 33 conditions, and among them, 30 combinations are MSFs + hand-crafted features. On the other hand, among all 144 conditions (12 conditions $\times$ 4 datasets $\times$ 3 cross-lingual) in cross-lingual SER, features containing MSFs performed best on 120 conditions, and among them, 29 of the best performing features are $\mathrm{MSSK_m}$. Therefore, $\mathrm{MSSK_m}$ is considered the universal feature in cross-lingual SER.

VI. Conclusion

This study investigated the robustness of MSFs in multilingual and cross-lingual SER under noisy reverberant conditions. We utilized SVM to compare the SER performance for recognizing four common emotions. These comparisons were conducted on intra-lingual and inter-lingual SER in four languages under 12 noisy reverberant conditions. The results demonstrate that MSFs are complementary with conventional hand-crafted features and they are robust under noisy reverberant conditions in different languages. The results also suggest that MSFs potentially describe common characteristics of emotions in noisy reverberant conditions among non-tonal languages.

For future work, it would be interesting to investigate the applicability of MSFs in recognizing emotions in other tonal languages to further examine the language-dependency of these features. Additionally, it would be worthwhile to explore the use of deep learning techniques to further enhance the performance of SER systems utilizing MSFs.

Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research (B) (21H03463), a Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 22K21304 and 22H00536), JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237).

References

[1] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 177–188, 2018.

[2] S. R. Kshirsagar and T. H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1892–1905, 2022.

[3] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 5795–5799.

[4] P. Heracleous, K. Yasuda, F. Sugaya, A. Yoneyama, and M. Hashimoto, "Speech emotion recognition in noisy and reverberant environments," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 262–266.

[5] A. R. Avila, J. Monteiro, D. O'Shaughneussy, and T. H. Falk, "Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks," in *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, 2017, pp. 360–365.

[6] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: System design, integration, and evaluation," *IEEE reviews in biomedical engineering*, vol. 1, pp. 115–142, 2008.

[7] J. Xiang, D. Poeppel, and J. Z. Simon, "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, EL7–EL12, 2013.

[8] R. O. Tachibana, Y. Sasaki, and H. Riquimaroux, "Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech," *Acoustical Science and Technology*, vol. 34, no. 4, pp. 263–270, 2013.

[9] L. Xu and B. E. Pfingst, "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hearing research*, vol. 242, no. 1-2, pp. 132–140, 2008.

[10] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.

[11] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 6, pp. 379–386, 2018.

[12] T. Guo, S. Li, M. Unoki, and S. Okada, "Investigation of noise-reverberation-robustness of modulation spectral features for speech-emotion recognition," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 39–46.

[13] T. Guo, Z. Zhu, S. Kidani, and M. Unoki, "Contribution of common modulation spectral features to vocal-emotion recognition of noise-vocoded speech in noisy reverberant environments," *Applied Sciences*, vol. 12, no. 19, p. 9979, 2022.

[14] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[15] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, pp. 1–10, 2021.

[16] M. Neumann *et al.*, "Cross-lingual and multilingual speech emotion recognition on english and french," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5769–5773.

[17] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[18] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Communication*, vol. 110, pp. 1–12, 2019.

[19] X. Li and M. Akagi, "A three-layer emotion perception model for valence and arousal-based detection from multilingual speech," 2018.

[20] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on hmm and ann," in *2009 WRI World congress on computer science and information engineering*, IEEE, vol. 7, 2009, pp. 225–229.

[21] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[22] S. Mao, D. Tao, G. Zhang, P. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6715–6719.

[23] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[24] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning.," in *Interspeech*, vol. 2021, 2021, pp. 4508–4512.

[25] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," 2009.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[28] O. Mohamed and S. A. Aly, "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset," *arXiv preprint arXiv:2110.04425*, 2021.

[29] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[30] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proceedings of 2011 international conference on electronic & mechanical engineering and information technology*, IEEE, vol. 2, 2011, pp. 621–625.

[31] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Vocal Tract Length Perturbation-based Pseudo-Speaker Augmentation for Speaker Embedding Learning

Tomoka Wakamatsu* Sayaka Shiota* and Hitoshi Kiya*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: wakamatsu-tomoka@ed.tmu.ac.jp

***Abstract*—Data augmentation is essential for constructing reliable automatic speaker verification (ASV) systems. It is well known that data augmentation increases the number of utterances by adding noise and is effective in most methods. However, the number of speakers in the training data also plays an important role in enhancing ASV system performance. The robustness of speaker embedding networks, which are used in ASV systems, relies on the number of speakers present in the training data. To address this, we propose a method called pseudo-speaker augmentation, which utilizes a technique called vocal tract length (VTL) warping. By changing a parameter, the VTL warping technique alters speaker characteristics, allowing us to easily increase the number of speakers. Since the speaker embedding network aims to classify speakers, having a larger number of speakers enhances its robustness. In our experiments, the pseudo-speaker augmentation method improved the performance of the speaker embedding-based ASV system, achieving an equal error rate of 4.058% on the JTubeSpeech database.**

I. Introduction

With the spread of spoken dialogue systems, automatic speaker verification (ASV), a biometric authentication technology using voice, has been actively researched in recent years. As many deep learning-based approaches advance, various deep neural network (DNN)-based ASV methods, such as x-vector [1] and ECAPA-TDNN [2], have been proposed and achieved remarkable performances. In these methods, a fixed-length feature vector called speaker embedding, extracted from the embedding layer of the DNN, is used to represent the personal characteristics of a speaker. In order to construct a reliable model in such a deep learning-based ASV system, a large amount of annotated data is required for model training. The most famous solution for insufficient annotated data is data augmentation by adding noise. Additionally, self-supervised learning (SSL) is also paid attention for ASV systems [3].

Data augmentation is an essential technique for constructing reliable speaker embedding networks. It is well known that data augmentation increases the number of utterances by adding noise and is effective in most methods [4], [5]. This technique also improves the generalization ability and robustness of the model by augmenting not only the amount of data but also the diversity of the data. In ASV tasks, data augmentation techniques to increase the number of utterances are generally used, such as the addition of noise or echoes [1], specaugment [5]. On the other hand, in a speaker embedding-based ASV system, it is known that the classification performance of the speaker embedding networks depends on the number of speakers in training data. Therefore, in constructing an ASV system, not only the number of utterances in the training data but also the number of speakers is considered an essential factor for improving the performance [6]. However, it is hard to collect large amounts of annotated data, including speaker labels. In [6], an augmentation method using speed perturbation [7] is proposed to increase the number of speakers at a small cost. It is reported that generating new speakers increases the range of speaker representation in the embedding space and improves the performance of ASV systems.

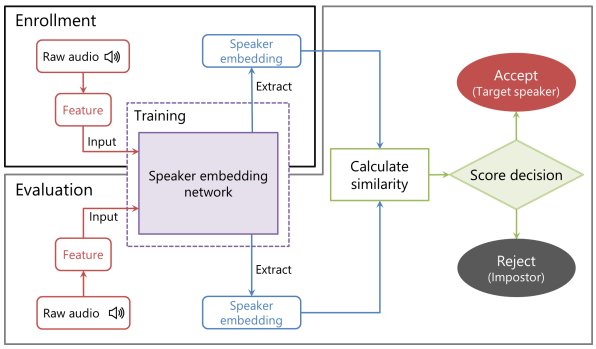


Fig. 1. Framework of speaker embedding-based ASV

Vocal tract length normalization (VTLN) [8] is used to counteract the effects of differences in vocal tract length. This technique can be reversely adapted to add variation to speech data. By changing a parameter, the vocal tract length warping technique alters speaker characteristics, allowing us to increase the number of speakers easily. Since the speaker embedding network aims to classify speakers, having a larger number of speakers enhances its robustness. Therefore, in this paper, we propose pseudo-speaker augmentation for speaker-embedding-based ASV by using the vocal tract length warping technique. In the experiments, we evaluated the effectiveness of the proposed pseudo-speaker augmentation with JTubeSpeech database [9]. The experimental results show that the best performance was obtained in the system with data augmentation for both the number of utterances and the number of speakers, and the effectiveness of speaker augmentation was confirmed.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

The rest of the paper is organized as follows. Section 2 describes a general speaker embedding-based ASV system. Section 3 will show the conventional utterance number augmentation by adding noise and the proposed pseudo-speaker augmentation with VTLP. Section 4 gives the detailed experimental setup and results, and the paper is concluded in Section 5.

II. Speaker Embedding-based Automatic Speaker Verification

ASV is a binary classification task that determines whether the given voice is a registered speaker or not. Fig. 1 shows the framework of a speaker embedding-based ASV system, which has become mainstream in recent years. The system is constructed in the following flow:

First, a speaker embedding network is trained using a large amount of labeled data to extract features known as speaker embeddings, which represent the speaker's characteristics. The speaker embedding network is a neural network model trained for speaker classification and used as a feature extractor. The speaker embeddings are extracted from its embedding layer. Next, in the enrollment part, enrolled utterances are converted into acoustic feature vectors such as MFCC. Then, the features are fed into the trained speaker embedding network, and speaker embeddings are extracted. In the evaluation part, the speaker embeddings of test utterances are also extracted in the same manner. Subsequently, the similarity scores between the extracted speaker embedding vectors are computed, and the scores are used for determining the decision whether to accept or reject.

III. Pseudo-speaker Augmentation

A. Data augmentation using simulated noises

In most ASV systems in real environments, test utterances contain background noise. To construct a robust ASV system, it is necessary to include training data similar to the test environment. Also, a variety of training data is essential to build a more robust system that can adapt to different or unknown environments. However, it is very costly that collect large amounts of speech data recorded in many kinds of different environments. Data augmentation techniques are widely used in various tasks to increase simulated variation in training data and to improve the robustness of the trained model. Adding noise or background sounds to original data is a common data augmentation method Incorporating different types and levels of noise into the training data allows for better performance in real-world scenarios with varying acoustic conditions.

B. Pseudo-speaker augmentation using Vocal Tract Length Warping

There are various factors that cause different speakers' voices to have different acoustic features, one of which is called vocal tract length (VTL). If the vocal tract is modeled as a uniform tube, it is possible to scale the frequency axis of the spectra of two speakers with different vocal tract lengths using a warping factor. It is known that such differences

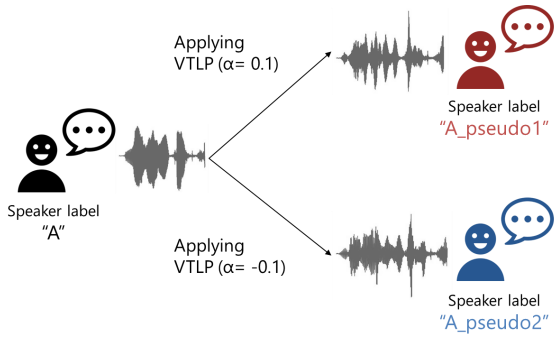


Fig. 2. Flow of making pseudo-speaker by VTLP

in vocal tract length affect recognition accuracy in speech recognition, and vocal tract length normalization (VTLN) [8] is used as a method to counteract this effect. VTLN removes inter-speaker variation by linearly warping the frequency axis of each speaker's spectrogram using the optimal warp factor for each speaker. A reverse adaptation of the VTLN method is the vocal tract length perturbation (VTLP) [10] approach, which adds variation to the input data by adding variation. In this paper, these frequency warping techniques are used as a data augmentation method to increase the number of speakers by attaching a different speaker ID to the processed speech than the original voice, like Fig. 2. If ω is the normalized frequency of the original voice and ω' is the frequency after augmentation and contraction, it can be expressed as:

$$\omega' = \omega + 2\arctan\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)}, \tag{1}$$

where α is the VTL warping factor.

IV. Experiment

We conducted ASV experiments to show the effectiveness of the proposed pseudo-speaker augmentation using VTLP.

A. Database

JTubeSpeech [9] was used to train and evaluate the ASV systems. This is a database of mainly spoken Japanese speakers extracted from videos uploaded to YouTube. It is noted that since the videos are automatically collected, JTubeSpeech also includes other languages, such as English and Chinese. A dataset of JTubeSpeech is designed for ASV with a large number of speakers. In the experiments, we used the single-speaker dataset of the JTubeSpeech subset as the training dataset. The training dataset contains 502 hours of utterances from 1,795 speakers, and the development dataset contains 138 hours of utterances from 1,795 speakers. A total of 640 hours of utterances were used as the training dataset. The test dataset contained 20,976 trials by 92 speakers. The evaluation trials consist of 228 pairs with target speakers and 20,748 pairs with impostors. The speakers in the test set are all manually annotated as Japanese speakers.

MUSAN [11] was used for utterance augmentation. It contains 42 hours of music of various genres, 60 hours of

TABLE I
ARCHITECTURE OF CNN USED AS SPEAKER EMBEDDING EXTRACTOR

	Layer	Kernel size	Input x Output
1	Conv1d	5	40 x 128
2	Conv1d	3	128 x 128
3	Conv1d	3	128 x 128
4	Conv1d	3	128 x 64
5	Fc	-	(64x3) x 512
6	Fc	-	512 x (*n_classes*)

conversations in 12 languages, and more than 900 noise types. Noise type which contains approximately 6 hours of various types of noise, including mechanical, non-mechanical, and environmental sounds.

B. Experimental setup

The architecture of the convolutional neural network (CNN) used as the speaker embedding extractor is shown in Table I. The output vectors of the first fully connected layer in Table I were regarded as the speaker embedding vectors. The input features of CNN were a 40-dimensional MFCC with a frame length of 25 ms. The network was trained for 100 epochs with the Adagrad optimizer [12] and Cross-Entropy Loss.

For the JTubeSpeech training dataset, we applied both data augmentation increasing the number of utterances by adding noise and increasing the number of speakers by VTLP described in section 3. In this paper, we refer the noise-adding data augmentation as utterance augmentation. The comparison methods were shown as follows:

(A). None
No data augmentation was applied to the training data. This was referenced as the baseline system.

(B). Noise
For utterance augmentation, noise data was randomly selected from the MUSAN noise data set, and the signal-to-noise ratio (SNR) was randomly set from [-5, 0, 5, 10, 15]. We applied utterance augmentation by using different noises and SNRs for each utterance. In this condition, the number of utterances was augmented by a factor of three by adding two types of noise data.

(C). Noise (large)
The utterance augmentation was applied in the same manner as condition (B). To compare the performance of a larger amount of data expansion only by adding noise, the SNRs were set to all of [-5, 0, 5, 10, 15], and the number of utterances was augmented by a factor of six.

(D). VTLP (all)
The pseudo-speaker augmentation was applied to all speakers in the training data, and the number of speakers was augmented by a factor of three. The warping factor α for VTLP was set to -0.1 or +0.1 for all utterances.

(E). VTLP (select)
Pseudo-speaker augmentation was performed in the same manner as condition (D). Figure 3 shows the distribution of the degree of changing speaker characteristics before and after adapting VTLP. Therefore, we decided to select the pseudo-speakers whom can use as different speakers from the original one adequately. Finally, the number of speakers is about twice as large, which is a smaller number than in condition (D).

(F). VTLP (select) + Noise
To investigate the effectiveness of combining both utterance number augmentation and speaker number augmentation, almost the same number of utterance augmentation was applied to condition (E) as in condition (B).

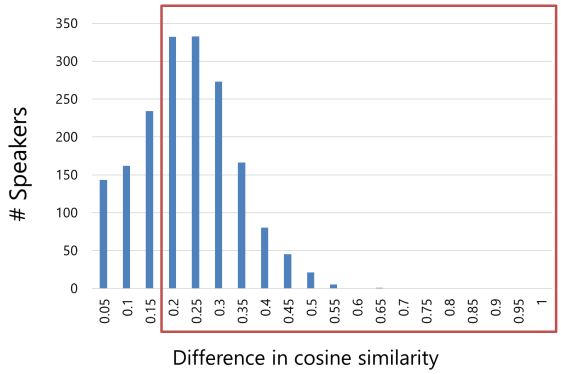


Fig. 3. Distribution of variation in speakability by VTLP

In creating pseudo-speakers, it is important that the individual characteristics after VTLP differ significantly from the original speaker. Therefore, we prepared condition (D) to test whether the degree of change in personal characteristics due to VTLP affects the accuracy of the ASV system. Specifically, we extracted the speaker embeddings of both the pseudo-speaker and the original speaker using one of our trained speaker classification models, and computed the average of the difference in the cosine similarity between them for each speaker. The average of the differences in similarity per speaker for a single warping factor is shown in Fig. 3. This confirms that the change in personal characteristics at a given value of the warping factor varies from speaker to speaker. For some speakers, however, the warping coefficients set in this study were not sufficient to treat the pseudo-speaker as a different speaker. Therefore, we used as VTLP (select) only pseudo-speakers with a difference in cosine similarity greater than 0.2. By implementing these conditions, we examined the usefulness of each data augmentation method in terms of the number of speakers and utterances. The amount of data in each condition is shown in the table II.

C. Evaluation metrics

To evaluate the comparison methods, we used two kinds of test trials provided in the JTubeSpeech dataset. One trial is defined as a core task, whose segments are set to under

TABLE II
DATA AMOUNT OF THE TRAINING DATASET IN EACH CONDITION

	Augmentation	# Speakers	Total hours
(A)	None	1,795	502.49
(B)	Noise	1,795	1,510.17
(C)	Noise (large)	1,795	3,014.90
(D)	VTLP (all)	5,385	1,506.32
(E)	VTLP (select)	3,681	957.59
(F)	VTLP (select) + Noise	3,681	2,875.49

TABLE III
EERS (%) FOR EACH AUGMENTATION METHOD.

	Augmentation	EER (%)	
		Core	Short
(A)	None	6.140	7.456
(B)	Noise	5.104	8.333
(C)	Noise (large)	5.104	8.333
(D)	VTLP (all)	7.018	9.211
(E)	VTLP (select)	6.579	7.822
(F)	VTLP (select) + Noise	**4.058**	**6.777**

five seconds, and another one is referenced as a short task, whose segments are set to under two seconds. These trials were released as the evaluation tasks of JTubeSpeech-ASV[1]. The evaluation metric was the equal error rate (EER) based on a comparison of cosine similarity between the speaker embeddings.

D. Experimental results

Table III shows the EERs under the core and short trials for each comparison condition. First, we discuss the results under the core trials. Comparing the results of (A) with (B) and (C) indicates that (B) and (C) improved performance compared to the baseline, attributed to the inclusion of utterance augmentation. On the other hand, when comparing (B) with (C), since only utterance augmentation was performed, the EERs were the same. Since both augmented data were generated by adding noise, the data diversity was limited, and it seems to lead to the saturated result. Comparing (D) with (E), both using the pseudo-speaker augmentation, the EER of (E) is 0.439 points better than that of (D). In both (D) and (E), pseudo-speaker augmentation was applied, and in (E), speakers were selected for which the change in their characteristics due to VTLP was significant. This indicates that when the speakers were chosen to apply VTLP, as in (E), VTLP can properly function as a data augmentation method. However, both (D) and (E) perform worse than the baseline. One reason for this result is that whereas the diversity of the speakers was increased, the number of utterance data for each speaker was not enough. (F) achieved the highest performance among all conditions. Especially comparing (F) with (C), while the EER of (C) seems saturated, the EER of (F) was 1.046 points better than that of (C) with almost the same amount of data. It denotes that the combination of utterance augmentation and pseudo-speaker augmentation contributes to expanding the data diversity, and it helps to improve performance. The short trials are challenging tasks, as stated in the Short-duration Speaker Verification Challenge(SdSVC) 2021[13]. In the results under these short trials, the performances were getting lower than the performances of the core trial. Comparing the results of (A) with those from (B) to (E), it appears that data augmentation is not effective in improving performance. However, comparing (E) with (B) and (C), the performance with the pseudo-speaker augmentation by VTLP(select) is better than the utterance augmentation. This trend is in contrast to the core trials. For short utterance trials, the speaker embedding quality is more important to represent speaker features accurately. The results suggest that the augmentation of the number of speakers may be effective in spreading the representation space of the speaker embedding. Consequently, the proposed pseudo-speaker augmentation was found to be effective when combined with the utterance augmentation in both the core and short trials.

[1]https://github.com/sarulab-speech/jtubespeech

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed pseudo-speaker augmentation by using VTLP for robust ASV systems. Data augmentation by adding noise is basically used for the state-of-the-art speaker embedding models. On the other hand, the larger the number of enrollment speakers for the speaker embedding models, the more robust the performances. Therefore, the pseudo-speaker augmentation was performed by using VTLP. Additionally, we described that the pseudo-speaker selection was effective for the ASV systems. The experimental results showed that the best performance was obtained when both the pseudo-speaker augmentation and the utterance augmentation were applied simultaneously. In the future, we will investigate the effectiveness of the proposed method on state-of-the-art systems such as x-vector and ECAPA-TDNN.

ACKNOWLEDGMENT

This work was supported in part by SECOM Science and Technology Foundation.

REFERENCES

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Proc. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, 2018.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech 2020*, pp. 3830–3834, 2020.

[3] Z. Chen, S. Chen, Y. Wu, *et al.*, "Large-scale self-supervised speech representation learning for automatic speaker verification," *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, 2022.

[4] Y. Yang, S. Wang, M. Sun, Y. Qian, and K. Yu, "Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification," *Proc. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 205–209, 2018.

[5] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[6] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," *Proc. Interspeech 2019*, pp. 406–410, 2019.

[7] T.Ko, V.Peddinti, D.Povey, and S.Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[8] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on speech and audio processing*, vol. 6, no. 1, pp. 49–60, 1998.

[9] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "Jtubespeech: Corpus of japanese speech collected from youtube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[10] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, p. 21, 2013.

[11] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[12] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization.," *Journal of machine learning research*, vol. 12, no. 7, 2011.

[13] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (sdsv) challenge 2021: The challenge evaluation plan," *arXiv preprint arXiv:1912.06311*, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Automatic Call Classification of Autism Model Marmosets by Deep Learning and Analysis of Their Vocal Development

Minato Uesaka*, Hideto Kawauchi*, Kouei Yamaoka*, Yukoh Wakabayashi*,
Yuma Kinoshita*, Nobutaka Ono*, Jun Noguchi‡, Satoshi Watanabe‡,
Noritaka Ichinohe‡, Seico Benner§¶, and Hidenori Yamasue§
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: uesaka-minato@ed.tmu.ac.jp
† Tokai University, Kanagawa, Japan
‡ National Center of Neurology and Psychiatry, Tokyo, Japan
§ Hamamatsu University School of Medicine, Shizuoka, Japan
¶ National Institute for Environmental Studies, Ibaraki, Japan

***Abstract*—In this study, we aim to automatically classify calls of autism model marmosets to analyze their vocal development. As a nonhuman primate model of autism spectrum disorder (ASD), valproic-acid-exposed (VPA) marmosets have recently been developed. This marmoset model has been shown to have abnormal vocal development similar to ASD symptoms compared with VPA-unexposed (UE) marmosets. Automatic classification of vocalizations and analysis of calls can be used to study the VPA marmosets. In the past, there have been no studies on the automatic classification of VPA marmoset calls. Therefore, we focus on classifying calls of both UE and VPA marmosets automatically and applying the automatic call classification to the analysis of their vocal development. We construct a call classifier based on deep learning, which distinguishes four types of marmoset call from audio signals. By experiment, we confirmed that the classifier trained on calls of UE marmosets could accurately classify two main classes of calls: 'phee' and 'twitter'. The classifier also worked well for VPA marmosets. Furthermore, we found that the classifier was also helpful for analyzing the vocal development of UE and VPA marmosets.**

I. Introduction

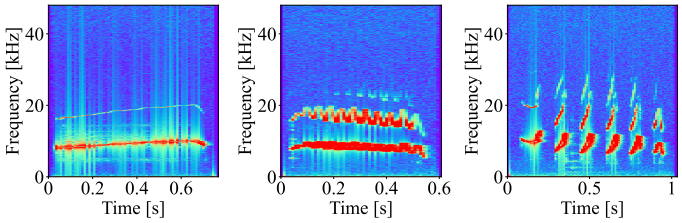
Autism spectrum disorder (ASD) is a developmental disorder characterized by symptoms such as social interaction and communication deficits. One notable symptom involves vocalization, which can exhibit specific features such as flat intonation and peculiar speech timing [1], [2]. To understand the pathogenic mechanisms of ASD and develop potential treatments, research has been conducted using animal models that replicate ASD-like pathologies and symptoms [3], as well as actual ASD diagnostic data [4]. Specifically, nonhuman primate models of ASD have attracted attention because they are more closely related to humans than rodent models [5].

As one of nonhuman primate model, an ASD marmoset model has been developed by administering valproic acid (VPA), which increases the risk of ASD in offspring, to the mothers of common marmosets (i.e., *Callithrix jacchus*) [6]. Common marmosets (hereafter simply referred to as marmosets) are small primates known for their ease of handling, small size, and high reproductive capacity, which provide several advantages for research purposes [7], [8]. Because of their high sociability and diverse calls used for communication, they have been extensively utilized in the field of psychiatric and neurological disorder research [9]–[11]. Consequently, marmosets are expected to be suitable for studying communication impairments in ASD. Indeed, VPA marmosets, which are marmosets exposed to VPA, exhibit anomalies in their calls [6], and they are expected to show abnormalities in their vocal development. Developing a system that efficiently and accurately captures call characteristics of the VPA and normal unexposed (UE) marmosets would aid in elucidating the mechanisms underlying communication impairments in ASD and would contribute to the development of treatments for ASD.

However, the classification of marmoset calls imposes a significant workload because it is performed manually by researchers. For this reason, we aim to classify marmoset calls automatically. Various methods have been proposed for this automatic classification, including signal-processing-based approaches, such as the threshold processing of the amplitude of acoustic signals [12], and deep-learning-based approaches [13]–[15] that demonstrate high efficacy for classification by considering relationships among short-time frames of acoustic signals. However, existing methods have primarily targeted UE marmosets and have not addressed the call classification of VPA marmosets.

Because of this situation, we construct a deep-learning-based call classifier for not only UE marmosets but also VPA marmosets. For training of the classifier, calls of VPA and UE marmosets are recorded, and the recorded signals are annotated by 12 different call types and two special classes, *unknown* and *no call*. Using the collected dataset, our classifier is trained to classify acoustic signals of marmoset calls into three main classes (*phee*, *trill*, and *twitter*) and two other classes (*other calls* and *no call*). For the architecture of our classifier, we

979-8-3503-0067-3/23/$31.00 ©2023 IEEE



(a) phee (b) trill (c) twitter

Fig. 1. Example of call spectrograms

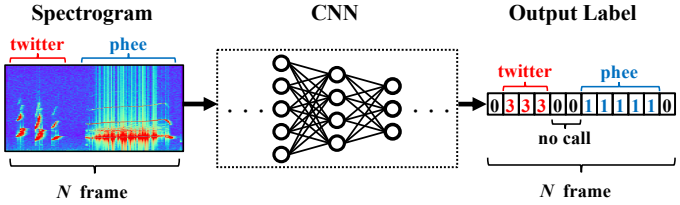


Fig. 2. Flow of call classification

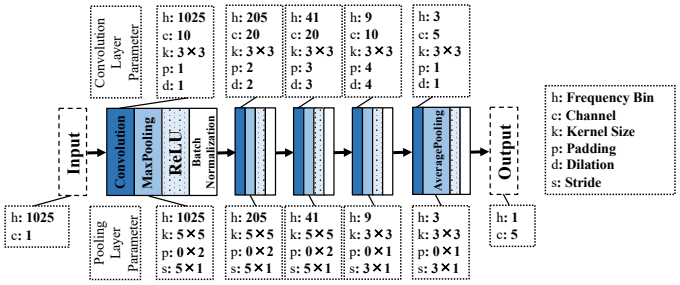


Fig. 3. Network architecture and hyperparameters of our classifier for marmoset call classification

use a convolutional neural network (CNN) because it will capture spectrotemporal characteristics of marmoset call signals in spectrograms obtained by short-time Fourier transform (STFT).

To objectively evaluate the classification accuracy of the proposed classifier, we conducted experiments using the data obtained from both UE and VPA marmosets. Experimental results show that the classifier can classify calls of VPA marmosets into the five classes with a high accuracy of 88.0 % in terms of the macro F1-score, even when trained only on calls of UE marmosets. Furthermore, we examined whether the trained classifier helps analyze the call changes observed during marmoset development.

II. Deep-learning-based Classification of Marmoset Calls

Aiming to analyze the vocal development of UE and VPA marmosets, we train a marmoset call classifier to distinguish call types of not only UE marmosets but also VPA ones. Here, we briefly summarize how to collect training data for the classifier. After that, we describe the problem statement for the marmoset call classification and the analysis method including the network architecture of our classifier.

A. Data Collection

To train the marmoset call classifier, we use a dataset of marmoset calls recorded and annotated by the Ichinohe group at the National Center of Neurology and Psychiatry. The following is the recording procedure. We collected calls of 37 marmosets including 24 UE, 11 VPA, and two other types of marmosets. The recordings were conducted for each marmoset from the first week to the 20th week after birth. During each recording, a marmoset was individually housed in a birdcage in a soundproof room, and its calls were monaurally recorded for approximately 5 min using a recording device (Olympus LS-100) placed in front of the birdcage. Marmoset calls include frequency components higher than human audible ones. Therefore, we set the sampling frequency to 96 kHz with a bit depth of 24 bits. The recorded audio signals include not only marmoset calls but also incidental noises, such as the sounds produced when the marmosets step on the pet sheet on the birdcage floor or grasp the wire mesh. For this reason, we classified the recorded marmoset calls into 12 different call classes (*phee, trill, twitter, tsik, ek, ek-tsik, cough, cry, chatter, breath, phee-trill,* and *trill-phee*) and two additional classes (*unknown* and *no call*). If calls were unidentifiable by annotators, they were labeled as *unknown*. If no marmoset call was included in a time segment, it was labeled as *no call*. The duration of each call was also annotated manually.

B. Problem Statement

The problem that we tackle in this study is detecting time segments including marmoset calls from audio signals in our dataset and classifying them into a number of classes.

Although 14 different labels are assigned to the signals in our dataset, as an initial study of VPA marmoset call classification, we focus on three main classes among them, *phee*, *trill*, and *twitter*, which are often analyzed in related studies. Figure 1 shows an example of spectrograms for calls of the three classes. Audio signals labeled as *phee-trill* or *trill-phee* are ignored because they are difficult to distinguish from *phee* and *trill*, and we also omit utterances labeled as unknown. We treat the other eight call classes as *other calls*. For signals that do not include any calls, we classify them as a *no call*. As a result, we consider a five-class classification problem in which audio signals are classified into one of the above classes.

C. Analysis Methods

The marmoset audio signals are fed into a deep neural network (DNN) to estimate the class for each short time frame of approximately 10 ms. The classification process is illustrated in Fig. 2.

We utilize the spectrogram obtained by applying the STFT to an input acoustic signal as an input feature. For the STFT, we use a frame length of 2048 samples, a frame shift of 1024 samples, and the Hamming window as the window function. As a result, the size of input feature matrices is the number of frequency bins (1025) $\times$ the number of time frames (N).

TABLE I
PERFORMANCE EVALUATION OF CALL CLASSIFICATION ON THE UE SET

	recall	precision	F1-score
no call	98.76 %	98.49 %	98.63 %
phee	99.06 %	96.93 %	97.99 %
trill	76.57 %	95.19 %	84.87 %
twitter	95.61 %	97.34 %	96.47 %
other calls	72.69 %	84.94 %	78.34 %

TABLE II
PERFORMANCE EVALUATION OF CALL CLASSIFICATION ON THE VPA SET

	recall	precision	F1-score
no call	99.07 %	98.45 %	98.76 %
phee	97.59 %	97.28 %	97.44 %
trill	87.58 %	85.65 %	86.60 %
twitter	96.49 %	97.37 %	96.93 %
other calls	44.78 %	92.63 %	60.38 %

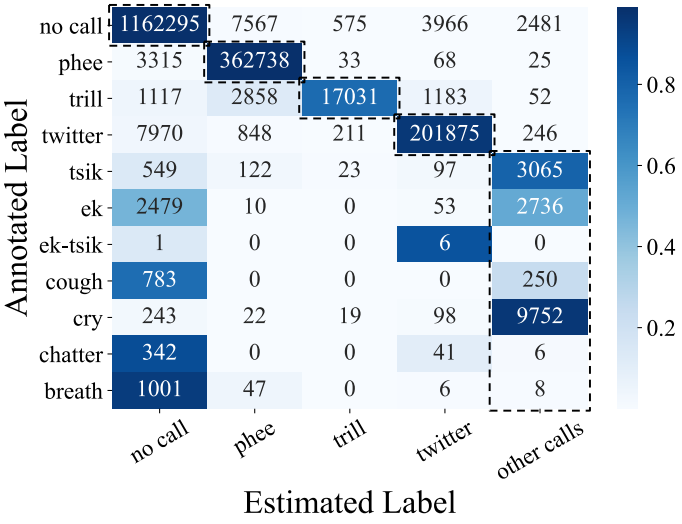


Fig. 4. Confusion matrix of classification results for the UE set (the estimated label of the *other call* class displayed the call types, *tsik, ek, ek-tsik, cough, cry*, and *chatter*, which belong to the annotated label)

The output of the classifier is a $5 \times N$ matrix, where the nth column in the matrix represents a call class as a five-dimensional one-hot vector for the nth short-time frame.

Since the marmoset call is generally longer than the length of the time frame, we employed a CNN. This network architecture can consider relationships among near-time frames. The CNN architecture and its hyperparameters are presented in Fig. 3.

A kernel size of 3×3 is utilized in the convolution layers, and three additional dilated convolution layers are employed to capture a wide range of temporal and spectral features. Furthermore, to enable framewise call classification, the network shown in Fig. 3 is designed to preserve the temporal size N of the feature maps. Specifically, the strides of the pooling layers are set to 1, and the padding widths of the convolution layers are determined on the basis of the kernel size and dilation rate, in order to maintain the output size.

III. EXPERIMENTAL EVALUATION OF CALL CLASSIFICATION

We conducted a numerical experiment to evaluate the performance of marmoset call classifiers that we trained. Here, we summarize the experimental conditions used and the results obtained.

A. Experimental Conditions

For the experiments, the dataset explained in Section II-A, consisting of UE marmosets recorded from the 3rd week to the 14th week of postnatal life, was divided into three subsets: training, validation, and test sets. The total signal lengths of the training, validation, and test subsets are 14 h 58 min 52 s, 3 h 20 min 2 s, and 5 h 24 min 27 s, respectively. Recordings and labels with deficiencies were excluded. The subsets consist of signals of 15, four, and five individuals, respectively.

The CNN model was trained using the training and validation subsets for learning analysis. To train the proposed CNN model, stochastic gradient descent was utilized as the optimization method, with an initial learning rate of 0.01 and a batch size of two. The model was trained for 100 epochs. Learning rate scheduling was also applied, which reduces the learning rate by a factor of 0.1 at the 50th and 75th epochs. The cross-entropy loss function was employed as the loss function.

To verify that the CNN trained only on the UE marmoset data can classify calls of VPA marmosets, the data from VPA marmosets (11 individuals, 11 h 21 min 13 s) were used solely for evaluation purposes.

B. Experimental Results

The performance of the proposed method on the UE set was evaluated using recall, precision, and F1-score, and the results are presented in Table I. Table I shows that the F1-scores for the *phee*, *trill*, and *twitter* classes are higher than that for the *other calls* class, which is 78.34 %.

Figure 4 displays the confusion matrix of the classification results on the UE set. The vertical axis represents the manually annotated labels, while the horizontal axis represents the labels estimated by the proposed method. Each element in the matrix represents the number of frames in which the label estimated by the proposed method corresponds to the annotated label. The background color of each element varies depending on the proportion it represents in each row. To examine the misclassification patterns in detail, the annotated labels are presented individually, without grouping *tsik, ek, ek-tsik, cough, cry, chatter*, or *breath* labels as *other calls*. The results show many correctly estimated frames along the diagonal, indicating accurate estimations for most frames.

Figures 5 and 6 show the spectrograms of a specific individual from the UE and VPA sets, along with the annotated and estimated labels for a 20 s segment. Although there are

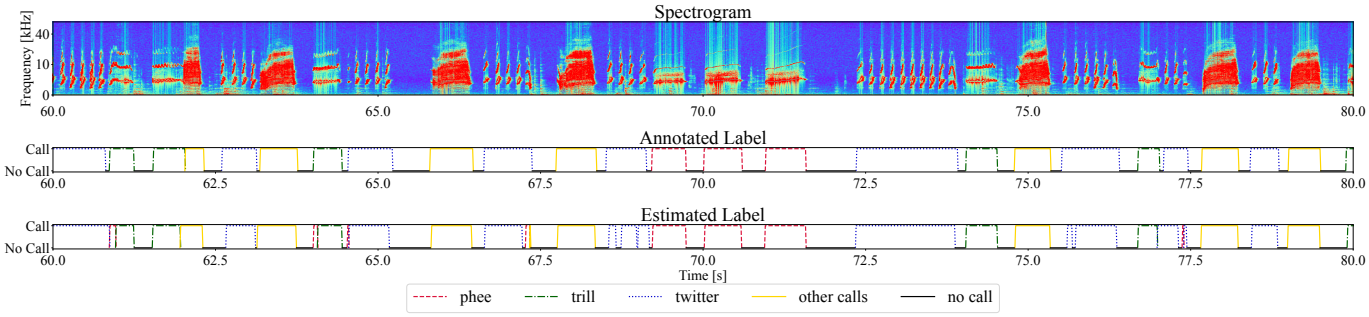


Fig. 5. Example of call classification results for the UE set (spectrograms of audio signals, annotated labels per time frame, and estimated labels)

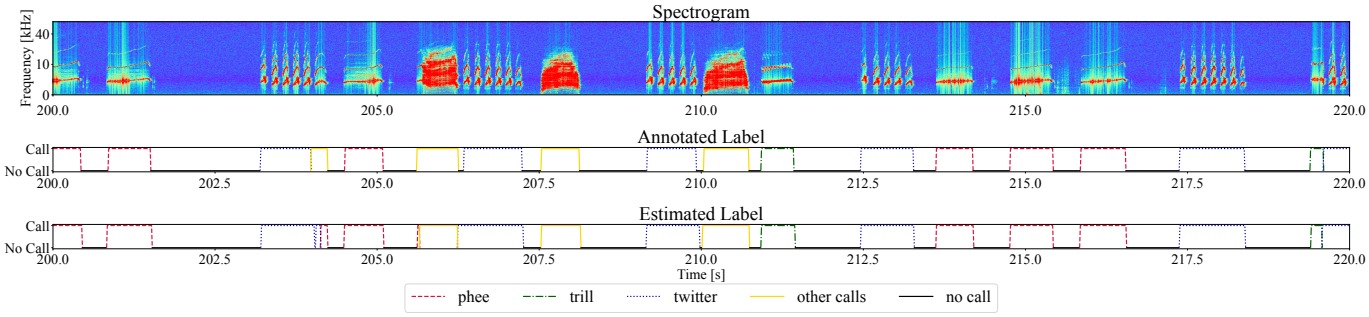


Fig. 6. Example of call classification results for the VPA set (spectrograms of audio signals, annotated labels per time frame, and estimated labels)

some misclassifications, it is evident that the classifications are generally accurate at the frame level.

The classification performance on the VPA set is presented in Table II. The results for the *phee*, *trill*, and *twitter* classes indicate that similar scores were achieved, as shown in Table I. This suggests that the proposed method can classify calls of VPA marmosets even when trained solely on the data of UE marmosets.

IV. Analysis of Vocal Changes Associated with Development

We examine whether our marmoset call classifier is helpful for analyzing changes in call proportions during marmoset development and for identifying differences between the calls of VPA and UE marmosets.

First, we analyze the changes in call patterns based on the annotated labels in the sets of UE and VPA individuals. The call labels considered here are the same as those described in Section III. Figure 7(a) represents the stacked bar graph of the proportion of each annotated label per frame over weeks for the UE Test individuals. Similarly, Fig. 7(b) shows the same graph for the VPA individuals. From Fig. 7(a) and Fig. 7(b), it can be observed that in both UE and VPA individuals, the proportion of *phee* calls increases with development. In contrast, *trill* and *twitter* calls tend to become less frequent. A comparison of UE and VPA data shows that VPA individuals have a higher proportion of *phee* calls than UE individuals. Furthermore, VPA individuals tend almost to cease *trill* calls with development.

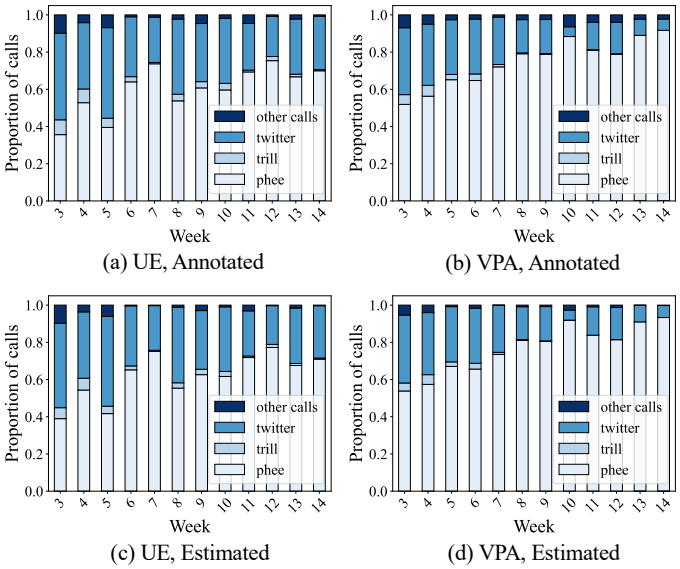


Fig. 7. Changes in proportion of calls for annotated and estimated labels by week (the UE set, the VPA set)

Figure 7(c) represents the stacked bar graph of the proportion of each estimated label per week for the UE Test individuals, obtained by applying the proposed method for call classification. Similarly, Fig. 7(d) shows the same graph for the VPA individuals. Comparing Figs. 7(c) and 7(d) with Figs. 7(a) and 7(b), we can see that even when using the proposed method to estimate call labels, the trends in call changes associated with development are still evident for both UE and

VPA individuals. This suggests that the proposed method will also be helpful for analyzing vocal development.

V. Conclusion

This study was aimed at developing a foundation for analyzing the developmental process of VPA marmosets by focusing on the automatic classification of marmoset calls. For marmoset call classification, marmoset audio signals were transformed into spectrograms by the STFT, and these spectrograms were used as acoustic features input to the CNN. The proposed method could classify four call types (*phee, trill, twitter*, and *other calls*) for each time frame by training the CNN on audio data from multiple individuals. Through objective evaluation experiments, high classification accuracy was confirmed.

Furthermore, it was found that even when training the deep learning model using only audio data from UE marmosets, high classification accuracy could still be achieved for the calls of VPA marmosets. Additionally, by comparing the annotated labels with the estimated labels for each developmental stage, it was determined that the proposed method will also be helpful for analyzing the vocal development of UE and VPA marmosets.

Future works include conducting individual-specific analyses of marmosets and exploring additional call labels not included in the current classification scheme, such as the *ek* and the *trill-phee* calls.

Acknowledgment

This work was supported by JSPS KAKENHI Grants (JP20H00613 to N.O. and JP22K07363 to J.N.), an Intramural Research Grant (Nos. 23-7, 26-9, and 29-6 to N.I.) for Neurological and Psychiatric Disorders from the NCNP, and AMED (JP23dm0207066 to N.I.).

References

[1] J. McCann and S. Peppe, "Prosody in autism spectrum disorders: A critical review," *International Journal of Language and Communication Disorders*, vol. 38, no. 4, pp. 325–350, 2003.

[2] A. Kondo and T. Deguchi, "Prosodic features in the speech of children with autism spectrum disorders," in *Proc. The Japanese Society of Hearing and Language Disorders*, vol. 42, 2013, pp. 23–30.

[3] S. S. Moy, J. J. Nadler, R. T. Magnuson, and N. J. Crawley, "Mouse models of autism spectrum disorders: The challenge for behavioral genetics," *American Journal of Medical Genetics*, vol. 142C, no. 1, pp. 40–51, 2006.

[4] K. Ochi, N. Ono, K. Owada, *et al.*, "Discrimination of people with autism spectrum disorder and with typical development using speech feature," in *Proc. Fall Meeting of Acoustical Society of Japan*, in Japanese, Sep. 2016, pp. 631–632.

[5] M. D. Bauman and C. M. Schumann, "Advances in nonhuman primate models of autism: Integrating neuroscience and behavior," *Exp. Neurol.*, vol. 299, no. Pt A, pp. 252–265, 2018.

[6] S. Watanabe, T. Kurotani, T. Oga, *et al.*, "Functional and molecular characterization of a non-human primate model of autism spectrum disorder shows similarity with the human disease," *Nature Communications*, vol. 12, no. 1, p. 5388, 2021.

[7] Y. Kobayashi, Y. Okada, G. Itakura, *et al.*, "Pre-evaluated safe human iPSC-derived neural stem cells promote functional recovery after spinal cord injury in common marmoset without tumorigenicity," *PloS One*, vol. 7, no. 12, e52787, 2012.

[8] H. Okano, K. Hikishima, A. Iriki, and E. Sasaki, "The common marmoset as a novel animal model system for biomedical and neuroscience research applications," *Seminars in Fetal and Neonatal Medicine*, vol. 17, no. 6, pp. 336–340, 2012.

[9] Y. Yamazaki and S. Watanabe, "Marmosets as a next-generation model of comparative cognition," *Japanese Psychological Research*, vol. 51, no. 3, pp. 182–196, 2009.

[10] D. Cyranoski, "Marmosets are stars of Japan's ambitious brain project," *Nature*, vol. 514, no. 7521, p. 151, 2014.

[11] J. M. Burkart and C. Finkenwirth, "Marmosets as model species in neuroscience and evolutionary anthropology," *Neuroscience Research*, vol. 93, pp. 8–19, 2015.

[12] J. A. Agamaite, C. J. Chang, M. S. Osmanski, and X. Wang, "A quantitative acoustic analysis of the vocal repertoire of the common marmoset (*Callithrix jacchus*)," *Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2906–2928, 2015.

[13] A. Wisler, L. J. Brattain, R. Landman, and T. F. Quatieri, "A framework for automated marmoset vocalization detection and classification," in *Proc. Interspeech*, 2016, pp. 2592–2596.

[14] Y. J. Zhang, J. F. Huang, N. Gong, Z. H. Ling, and Y. Hu, "Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks," *Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 478–487, 2018.

[15] T. Oikarinen, K. Srinivasan, O. Meisner, *et al.*, "Deep convolutional network for animal sound classification and source attribution using dual audio recordings," *Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 654–662, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Cross-Domain adaptation in Distance Space for Speaker Verification

Lu YI and Man Wai MAK
The Hong Kong Polytechnic University, Hong Kong SAR, China
E-mail: lu-louisa.yi@connect.polyu.hk, enmwmak@polyu.edu.hk

***Abstract*—Significant performance degradation often occurs when a well-trained speaker verification system is applied to an unseen domain. Data augmentation and domain adaptation are two common approaches to tackling this problem. However, data augmentation would not be helpful when language mismatch rather than environment noise causes the domain shift. Domain adaptation also suffers from a label mismatch problem, making feature distribution alignment unreliable. We propose incorporating a distance metric space into model adaptation to address these issues. The idea is to align not only the embeddings across domains but also the distributions of their pairwise distances, resulting in embeddings tolerant to domain shift. To validate the idea, we used the non-Chinese utterances in VoxCeleb2 and the Chinese utterances in CN-Celeb2 as the source and target domain training data, respectively. Results show that the alignments reduce the EER on the CN-Celeb1 test set by 15.2%.**

Index Terms: Domain shift, speaker embedding, distance metric space, distance alignment

I. Introduction

Speaker verification (SV) has become a prevalent authentication method in remote services and financial transactions primarily because of its convenience and acceptable accuracy. However, the technology still has limitations under some practical scenarios. For example, an English system could fail when deployed to a Chinese environment, primarily due to the domain shift. Because collecting sufficient labeled data for all possible environments is time-consuming and labor-intensive, exploring more efficient and practical approaches to addressing domain shift is essential.

Domain mismatch and domain shift occur when speech data are collected from different acoustic environments or the speakers have various demography and speak different languages. Data augmentation can partly alleviate these problems. In particular, a wide variety of acoustic environments can be emulated by adding noise and reverberation to clean speech. However, if languages cause the domain shift, data augmentation would not be helpful. In this case, domain adaptation would be a better solution.

Most traditional approaches to domain adaptation adapt the PLDA models to fit the target data [1], [2]. However, the recent trend in speaker embedding is to train a speaker embedding network (the front end) and replace PLDA scoring with cosine distance scoring. Because the latter does not have any parameters for adaptation, all adaptations must be performed on the front-end network. Alternatively, we may adopt a training strategy that enables the network to find an embedding space that works well across different domains. Previous attempts in this line of thought align the embedding vectors across multiple domains by minimizing the inter-domain cosine distances or maximum mean discrepancy (MMD) [3], [4]. More recent research applied adversarial learning to extract domain-invariant features [5]–[7]. However, aligning the distributions of embedding vectors across domains without losing speaker discrimination requires the domains to have a common set of speakers (or at least many speakers have speech in all domains). Unfortunately, this requirement is not achievable in practice. To tackle this problem, Mekhazni et al. [8] proposed aligning the pairwise dissimilarity between domains for person re-identification. Inspired by this idea, Han and Song [9] found that within-speaker and between-speaker can be considered classes, eliminating the need for the source and target domains to share the same set of speakers. Then, the distributions of the within-speaker distances from different domains can be aligned, and similarly for the distributions of the between-speaker distances.

Motivated by [8] and [9], we propose incorporating the idea of aligning the within- and between-speaker distance distributions into the domain adversarial neural network (DANN) [10] and the domain separation and adaptation network (DSAN) [11]. The DSAN can disentangle the domain-specific and domain-invariant speaker features, ensuring that the domain-invariant features contain fewer redundant features. However, the original DANN and DSAN only focus on feature alignments, which may align speaker distributions instead of domain distributions, causing even higher speaker confusion. Different from [9], the extended DANN and DSAN in this paper align the feature distributions and the within- and between-class distance distributions, resulting in more robust and discriminative speaker embeddings. Our experimental results on the CN-Celeb dataset demonstrate the effectiveness of combining feature and pair-wise distance distribution alignments for domain adaptation.

II. Methodology

In speaker verification, aligning the feature distributions between multiple domains cannot produce domain-invariant speaker discriminative features because different domains have different speaker sets. The feature alignment may just be speaker alignment rather than domain alignment. To tackle this problem, we propose introducing pairwise-distance distribution

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

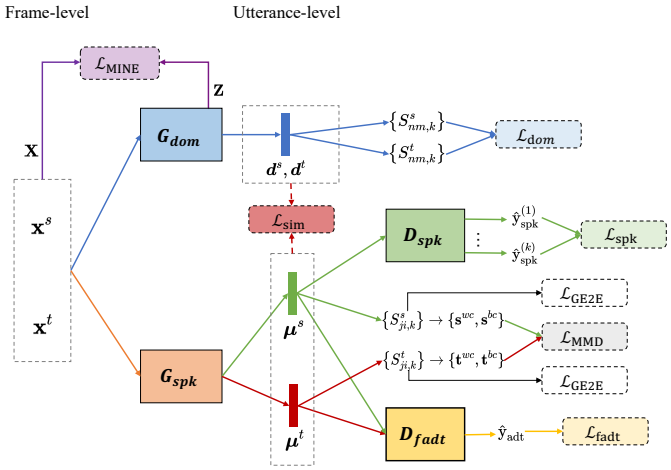


Fig. 1: Overview of the proposed domain separation and adaptation network (DSAN) with both feature distribution alignment and pairwise-distance distribution alignment.

alignment to the domain separation and adaptation network (DSAN) [11], as Fig. 1 shows.

Denote a labelled dataset from the source domain $\mathcal{D}^s$ as $\mathbf{X}^s = \{(\mathbf{x}_i^s, y_i)\}_{i=1}^{N_s}$, and an unlabelled dataset from the target domain $\mathcal{D}^t$ as $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$. The domain-specific features ($\boldsymbol{d}^s$ and $\boldsymbol{d}^t$) and domain-invariant features ($\boldsymbol{\mu}^s$ and $\boldsymbol{\mu}^t$) are extracted from the domain feature extractor G_{dom} and G_{spk}, respectively.

A. Domain-Invariant Speaker Features

To make the domain-invariant features speaker discriminative, AM-softmax loss [12] and generalized end-to-end (GE2E) loss [13] are adopted:

$$\mathcal{L}_{spk} = -\frac{1}{T}\sum_{i=1}^{T}\log\frac{e^{s\cdot\left(\cos\theta_{y_i}-m\right)}}{e^{s\cdot\left(\cos\theta_{y_i}-m\right)}+\sum_{c=1,c\neq y_i}^{C}e^{s\cdot\cos\theta_c}}, \tag{1}$$

where T denotes the batch size, s is the scaling factor, m is the margin, y_i denotes the speaker label of the i-th segment, and C is the total number of speakers from the source domain.

For the GE2E loss, we assume that each batch has $N^u \times M^u$ utterances, where N^u is the number of speakers and M^u is the number of utterances per speaker. Denote $\boldsymbol{\mu}_{ji}$ (which can be $\boldsymbol{\mu}^s$ or $\boldsymbol{\mu}^t$ in Fig. 1) as the domain-invariant features extracted from the i-th utterance of the j-th speaker. The centroids of the embedding vectors for the j-th speaker can be obtained by [13]:

$$\mathbf{c}_j = \frac{1}{M^u}\sum_{m=1}^{M^u}\boldsymbol{\mu}_{jm} \tag{2}$$

or

$$\mathbf{c}_j^{(-i)} = \frac{1}{M^u-1}\sum_{\substack{m=1\\m\neq i}}^{M^u}\boldsymbol{\mu}_{jm}. \tag{3}$$

Then, the similarity matrix between the embedding vectors $\boldsymbol{\mu}_{ji}$'s to all centroids is

$$S_{ji,k} = \begin{cases} w\cdot\cos\left(\boldsymbol{\mu}_{ji}, \mathbf{c}_j^{(-i)}\right)+b & \text{if } k=j \\ w\cdot\cos\left(\boldsymbol{\mu}_{ji}, \mathbf{c}_k\right)+b & \text{otherwise} \end{cases}, \tag{4}$$

where j and k index speakers and w and b are learnable parameters. Therefore, the GE2E loss is defined as:

$$L_{\text{GE2E}} = \frac{1}{N^u M^u}\sum_{j,i}\left(-S_{ji,j}+\log\sum_{k=1}^{N^u}\exp\left(S_{ji,k}\right)\right). \tag{5}$$

1) Feature Distribution Alignment: To explore the common features across multiple domains, we train G_{spk} adversarially to align the feature distributions. Specifically, the feature adaptation network D_{fadt} is trained to distinguish whether the speaker features come from the source domain or the target domain, while the speaker feature extractor G_{spk} is trained to produce domain-invariant features to confuse the D_{fadt}. Therefore, G_{spk} and D_{fadt} are trained with the following minimax loss:

$$\begin{aligned}\min_{G_{\text{spk}}}\max_{D_{\text{fadt}}}\mathcal{L}_{\text{fadt}} &= \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[\log D_{\text{fadt}}\left(G_{\text{spk}}\left(\mathbf{x}^s\right)\right)\right] \\ &+ \mathbb{E}_{\mathbf{x}^t\sim\mathcal{D}^t}\left[\log\left[1-D_{\text{fadt}}\left(G_{\text{spk}}\left(\mathbf{x}^t\right)\right)\right]\right].\end{aligned} \tag{6}$$

2) Pairwise Distance Distribution Alignment: As discussed in Section 1, only aligning the feature distributions may not truly learn domain-invariant features because of the label mismatch between the source and target domains. To address this problem, we also train G_{spk} to align the similarity matrices (Eq. 4) obtained from the source and target domains. Denote $\mathbf{s}^{wc}$ and $\mathbf{s}^{bc}$ as the source domain's within-class and between-class distance vectors, respectively. The elements of $\mathbf{s}^{wc}$ and $\mathbf{s}^{bc}$ are obtained from the first and second lines of Eq. 4, respectively, as Fig. 2 shows. Similarly, we denote $\mathbf{t}^{wc}$ and $\mathbf{t}^{bc}$ as the within-class and between-class distance vectors from the *target* domain. One advantage of using the similarity matrix (Eq. 4) is that it can save computation.

Given many similarity matrices (one for each utterance), we may consider the distance vectors as random vectors generated from some distributions. We use maximum mean discrepancy (MMD) [14] to measure the distance between the distributions from the source and target domains:

$$\begin{aligned}\text{MMD}(p(\mathbf{s}), p(\mathbf{t})) &= \frac{1}{{T_s}^2}\sum_{n=1}^{T_s}\sum_{m=1}^{T_s}k(\mathbf{s}_n, \mathbf{s}_m) \\ &+ \frac{1}{{T_t}^2}\sum_{n=1}^{T_t}\sum_{m=1}^{T_t}k(\mathbf{t}_n, \mathbf{t}_m) \\ &- \frac{2}{T_s T_t}\sum_{n=1}^{T_s}\sum_{m=1}^{T_t}k(\mathbf{s}_n, \mathbf{t}_m),\end{aligned} \tag{7}$$

where T_s and T_t are the numbers of samples from the source and target domains in a mini-batch, respectively, and $k(.,.)$ is a kernel. A mixture of radial basis function (RBF) kernels was

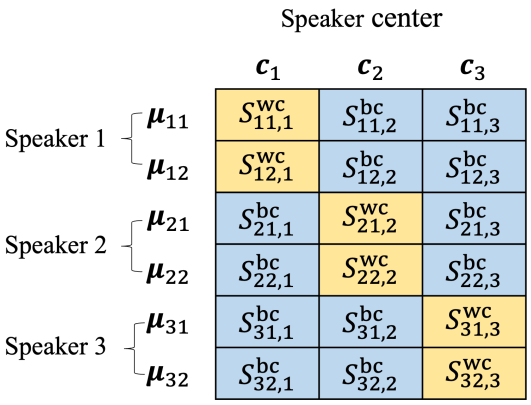


Fig. 2: Extracting $\mathbf{s}^{wc}$ and $\mathbf{s}^{bc}$ from the similarity matrix $\{S_{ji,k}\}$, where $j,k = 1,\ldots,3$ and $i = 1,2$, by stacking the gold and blue elements, respectively.

applied, i.e., $k(\mathbf{x},\mathbf{y}) = \sum_{q=1}^{K} \exp\left(-\frac{1}{2\sigma_q^2}\|\mathbf{x}-\mathbf{y}\|^2\right)$. Note that vectors $\mathbf{s}$ and $\mathbf{t}$ can be either within-class (wc) or between-class (bc) distance vectors. The loss function for distance distribution alignment is

$$\begin{aligned}\mathcal{L}_{\text{MMD}} = \text{MMD}(p(\mathbf{s}^{wc}), p(\mathbf{t}^{wc})) \\ + \text{MMD}(p(\mathbf{s}^{bc}), p(\mathbf{t}^{bc})).\end{aligned} \tag{8}$$

B. Domain-Specific Features

Instead of using a discriminator to distinguish the domain-specific features ($\mathbf{d}^s$ and $\mathbf{d}^t$ in Fig. 1), the proposed network leverages the similarity matrices in Eq. 4 to make these features distinguishable between domains. The loss has the same form as Eq. 5, except that G_{dom} is trained to produce domain-specific features close to the centroid of their respective domain and far away from the centroids of other domains.

Assume that there are $N^d \times M^d$ utterances in a batch, where the utterances are from N^d domains and each domain has M^d utterances. The feature vector $\mathbf{d}_{nm} (1 \le n \le N^d \text{ and } 1 \le m \le M^d)$ represents the domain-specific features extracted from the m-th utterance of the n-th domain. To extract domain-specific features, we train G_{dom} to minimize the loss:

$$\mathcal{L}_{\text{dom}} = \frac{1}{N^d M^d} \sum_{n,m} \left(-S_{nm,n} + \log \sum_{k=1}^{N^d} \exp\left(S_{nm,k}\right) \right), \tag{9}$$

where $\{S_{nm,k}\}$ are obtained by replacing $\boldsymbol{\mu}$ by $\mathbf{d}$ in Eq. 4 and replacing $\mathbf{c}$ and $\mathbf{c}^{(-1)}$ by the centroids of $\mathbf{d}$'s.

In addition, a mutual information neural estimator (MINE) is introduced, as proposed in [11], to maximize the mutual information between the frame-level domain features and the input acoustic features. The objective of adding MINE is to ensure that the domain-specific features $\mathbf{d}^s$ and $\mathbf{d}^t$ are meaningful, i.e., they learn useful domain information from speech. The MINE loss is defined as

$$\mathcal{L}_{\text{MINE}} = -\text{MINE}(\mathbf{X}, \mathbf{Z}), \tag{10}$$

where $\mathbf{X}$ and $\mathbf{Z}$ contain acoustic vectors and frame-level domain representation (before statistics pooling) of an utterance or a speech segment. Note that $\mathbf{X}$ and $\mathbf{Z}$ can be from the source or target domain.

MINE was initially proposed by Belghazi *et al.* [15] to approximate the information shared between random variables. However, we used a MINE based on the Jensen-Shannon divergence [16] instead of the KL divergence.

Apart from finding the domain-specific and domain-invariant features, we need to disentangle the domain-specific features from the speaker features. To achieve this, we introduce a constraint to make the speaker embeddings ($\boldsymbol{\mu}$) and the domain embeddings ($\mathbf{d}$) different, i.e., contain less mutual information. This goal can be achieved by minimizing a similarity loss:

$$\mathcal{L}_{\text{sim}} = \text{MINE}(\boldsymbol{\mu}^s, \mathbf{d}^s) + \text{MINE}(\boldsymbol{\mu}^t, \mathbf{d}^t), \tag{11}$$

where $\text{MINE}(a,b)$ is the MINE between the distributions of a and b [16]. G_{dom} and G_{spk} are trained to minimize $\mathcal{L}_{\text{sim}}$.

C. Overall Optimization

The overall objective of the proposed model is

$$\begin{aligned}\min_{G_{\text{dom}},G_{\text{spk}},D_{\text{spk}},T_\theta} \max_{D_{\text{fadt}}} \mathcal{L}_{\text{spk}} + \mathcal{L}_{\text{GE2E}} + \mathcal{L}_{\text{fadt}} + \lambda\mathcal{L}_{\text{MMD}} \\ + \mathcal{L}_{\text{dom}} + \mathcal{L}_{\text{MINE}} + \mathcal{L}_{\text{sim}},\end{aligned} \tag{12}$$

where T_θ is a MINE neural network parameterized by θ. λ is a hyper-parameter controlling the contribution of a sub-loss to the total loss, and we set it to 1 in the experiments. Noted that $\mathcal{L}_{\text{spk}}$ uses supervised learning, so only source domain data $\mathbf{x}_s$ is contributed to this loss.

III. Experimental Setting

A. Datasets

We used two datasets from different languages to perform domain adaptation. The source domain data comprises utterances from the development set of VoxCeleb2 [17], and the target domain data contains utterances from CN-Celeb2 [18]. There are 5950 speakers from the source domain, where most of the utterances are in English, and the Chinese utterances were removed. For the target domain data, around 1996 Mandarin speakers were used for training the domain adaptation models without using their speaker labels. Since CN-Celeb covers 11 genres, such as singing, speech, interview, movie, and so on, we considered them 11 domains when training the DSAN. In other words, we further divided the target domain data into 11 sub-domains according to the speech genres, i.e., $N^d = 11$ in Eq. 9. This division of domains is helpful in domain-specific feature extraction. No data augmentation was applied when training the domain adaptation models.

We evaluate the performance on the Voxceleb1 evaluation set [19] and the CN-Celeb1 evaluation set [20].

B. Data Preprocessing

For all audio files, 40-dimensional filter-bank features were extracted using Kaldi, with a 25ms window and 10ms frame shift, and the silence frames were removed using Kaldi's energy-based voice activity detector (VAD).

C. Implementation Details

Due to the complexity of the DSAN, we used a lightweight TDNN [21] instead of the bulky ECAPA-TDNN [22] as the backbone to reduce computation cost. G_{spk} has one fully-connected (FC) layer with 256 nodes after the statistics pooling layer, while D_{spk} comprises one FC layer with 64 nodes and a softmax output layer. The domain feature extractor G_{dom} shares the same layer structure and initial weights with G_{spk}. The adaptation network D_{fadt} has two FC layers with 256 nodes in each layer. The mutual information neural estimator (MINE) is a three-layer FC network with 512 nodes in each layer. The dimension of domain-specific embeddings $\mathbf{d}$'s is 64, while that of the speaker embeddings $\boldsymbol{\mu}$'s is 256.

The batch size is 1024, with 512 from the source domain and 512 from the target domain. Each batch has 128 speakers with four utterances per speaker from the source domain data. For the target domain data, each batch contains eight different genres with 16 utterances per genre and four chunks per utterance. Since we cannot access the speaker labels in the target domain, we considered segments sampled from the same utterance to be from the same speaker, and segments from different utterances to be from different speakers. As a result, the speaker classification in Eq. 5 was minimized in a self-supervised manner. For source domain data, we used Eq. 1 to compute the speaker classification loss.

A lightweight TDNN, which can be considered as the combination of G_{spk} and D_{spk}, was pre-trained using the source domain data (VoxCeleb2) with the additive margin softmax (AM-softmax) loss. The scaling factor s and the margin m in Eq. 1 are 30 and 0.2, respectively. The pre-trained model was then used to initialize G_{dom}, G_{spk}, and D_{spk}. An SGD optimizer with an initial learning rate of 0.01 was used to train the pre-trained and other models. For adversarial training (Eq. 6), we followed the training strategy in [10] by increasing the hyperparameter of the gradient reversal layer from 0 to 1 during training.

Probabilistic linear discriminant analysis (PLDA) [23] was used as the backend for scoring. Before PLDA training, LDA was applied to project the speaker embeddings onto a 150-dimensional space. We used the development set of VoxCeleb2 and the training set of CN-Celeb2 to train the PLDA backend, respectively. The former setting meets the condition of unsupervised domain adaptation (UDA), i.e., no access to the labels of the target domain. Although the latter setting slightly violates the UDA condition, it allows us to compare with the results of [24] and [25].

IV. Results and Discussions

We first compare the results of different models on the CN-Celeb1 evaluation set in the unsupervised setting, as Table I shows. The first four rows show the performance of the baseline models trained with different training sets. The front-end models trained with VoxCeleb2 perform better than the ones trained with CN-Celeb2, and data augmentation can help improve the performance slightly. The performance improved when training the front-end models with VoxCeleb2 and CN-Celeb2 (Rows 4 and 5). Row 5 shows the results obtained from a TDNN-based model trained with self-supervised fine-tuning using VoxCeleb2 and CN-Celeb2 as the training set. These promising results suggest that combining two different datasets extends the diversity of speech samples, encouraging the TDNN model to find more robust speaker embeddings.

The domain adaptation models, DANN and DSAN, were initialized using the parameters of the TDNN model from Row 1. Each DANN model comprises G_{spk}, D_{spk} and/or D_{fadt} in Fig. 1. If focusing only on the feature distribution alignment (Row 6), the performance is even worse than the baseline. This may be due to the negative transfer, meaning that the network cannot leverage the knowledge from the source domain to explore discriminative speaker features in the target domain. This result is consistent with our previous statement that only aligning the feature distributions would not be helpful in open-set domain adaptation. Row 7 shows the results from a DANN system that only considers the pairwise-distance distribution alignment. The results demonstrate the effectiveness of distance metric adaptation. However, the performance can be further improved when training the DANN with both kinds of alignments (Row 8). This suggests that with the constraint of distance metric alignment, the feature distribution alignment can bring a positive transfer from the source domain to the target domain.

Most domain adaptation techniques only focus on finding the common features while ignoring the domain-specific features. However, our results in Rows 9-11 demonstrate that finding the domain-specific features (Section II-B) is essential. Though the label sets of the source and target domains are different, robust speaker embeddings can be obtained by exploring the domain-specific features (the speech genres in our case). The performance is even better if we incorporate the pairwise distance distribution alignment into the DSAN models, as Row 11 shows. These positive results show that aligning the distance metrics across multiple target domains and considering the domain-specific features can help find more invariant and discriminative speaker features.

Although no domain adaptation is needed when the evaluation data are from the source domain, except for Row 8, the performance is even better after domain adaptation, demonstrating the robustness of the proposed DA method.

Table II demonstrates the performance of the CN-Celeb1 evaluation set in the supervised DA setting, i.e., using the target domain data (CN-Celeb2) to train the backend models. The DSAN with PDA can outperform the model with entirely supervised learning (Row 4). This further suggests that learning domain invariant speaker embeddings can enhance speaker verification. We also compare our results with other well-

TABLE I: Performance on the VoxCeleb1 evaluation set and CN-Celeb1 evaluation set in the unsupervised domain adaptation setting. For comparison, we also trained a DANN [10], which can be considered as the lower part of Fig. 1. The domain adaptation networks were trained to conduct feature alignment (FA) or/and pairwise distance distribution alignment (PDA).

Row	**System**	**Training data**	**Vox1 eval.**		**CN1 eval.**	
			EER(%)	minDCF	EER(%)	minDCF
1	TDNN	VoxCeleb2	2.047	0.2042	14.35	0.8843
2		VoxCeleb2 (+augmentation)	1.910	0.2201	14.27	0.7978
3		CN-Celeb2	7.492	0.6510	15.56	0.7731
4		VoxCeleb2 + CN-Celeb2 (with label)	2.227	0.2416	13.81	0.7289
5		VoxCeleb2 + CN-Celeb2 (without label)	2.863	0.3066	13.92	0.7513
6	DANN (FA) [10]	VoxCeleb2 + CN-Celeb2 (without label)	2.752	0.2927	15.42	0.7888
7	DANN (PDA) [9]		1.994	0.2136	13.81	0.7862
8	DANN (FA + PDA)		1.978	0.2107	13.49	0.7982
9	DSAN (FA) [11]	VoxCeleb2 + CN-Celeb2 (without label)	1.898	0.1996	13.69	0.7738
10	DSAN (PDA)		1.914	**0.1806**	13.33	0.7746
11	DSAN (FA + PDA)		**1.893**	0.1984	**13.04**	**0.6673**

TABLE II: Performance on the CN-Celeb1 evaluation set in the supervised domain adaptation setting.

Row	**System**	**EER (%)**	**minDCF**
1	TDNN ($\mathcal{D}^s$)	12.63	0.5649
2	TDNN ($\mathcal{D}^s$ + aug.)	12.24	0.5724
3	TDNN ($\mathcal{D}^t$)	14.23	0.6678
4	TDNN ($\mathcal{D}^s, \mathcal{D}^t$)	11.26	0.5659
5	TDNN ($\mathcal{D}^s, \widetilde{\mathcal{D}}^t$)	12.02	0.5729
6	TDNN + CORAL+ [24]	11.92	0.5654
7	TDNN + AS-Norm [25]	12.25	0.5812
8	DANN (FA) [10]	13.59	0.6204
9	DANN (PDA) [9]	11.57	**0.5242**
10	DANN (FA + PDA)	11.37	0.5244
11	DSAN (FA) [11]	11.46	0.5281
12	DSAN (PDA)	10.76	0.5465
13	DSAN (FA + PDA)	**10.71**	0.5364

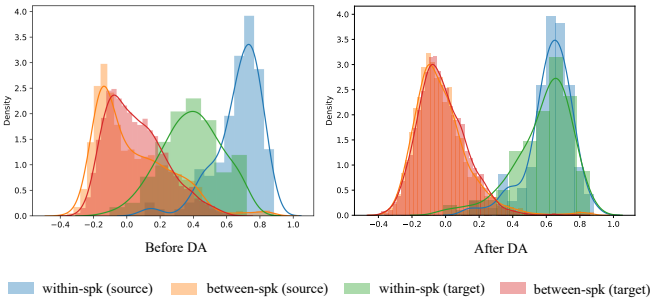


Fig. 3: Visualization of pairwise distance distributions.

established unsupervised domain adaptation baselines (Rows 6 and 7), which further demonstrate the effectiveness of our proposed DSAN model.

The basic assumption of applying the distance metric alignment is that the pairwise distance distributions will differ when employing a well-trained system on an unseen domain. We plot the distance distributions derived from a well-trained TDNN model (Row 1) to verify this assumption, as the left figure in Fig. 3 shows. It can be observed that the within-class distance distribution of the target domain is different from that of the source domain. There is a significant overlap between the within- and between-speaker distance distributions, indicating that the speaker vectors from different speakers are intertwined. After domain adaptation, the distributions of the target domain become similar to those of the source domain, as the right figure in Fig. 3 shows. The less overlapping between the within- and between-speaker distributions suggests that more discriminative speaker embeddings are produced.

V. Conclusions

In this paper, we propose aligning both feature and pairwise-distance distributions to produce domain-invariant speaker embeddings. We applied this idea to a domain adversarial neural network and our proposed domain adaptation and separation network. Our experiments demonstrate that the adaptation in a distance metric space helps find domain-invariant features for open-set domain adaptation tasks. The combination of feature distribution alignment and pairwise-distance distribution alignment can help find more speaker discriminative and domain invariant features.

Acknowledgment

This work was in part supported by the National Natural Science Foundation of China under Grant 61971371.

References

[1] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4047–4051.

[2] M. W. Mak and J. T. Chien, *Machine Learning for Speaker Recognition*. Cambridge University Press, 2020.

[3] J. Huang and T. Bocklet, "Intel far-field speaker recognition system for VOiCES challenge 2019," in *Proc. Interspeech*, 2019, pp. 2473–2477.

[4] W. W. Lin, M. W. Mak, and J. T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

[5] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4889–4893.

[6] Y. Tu, M. W. Mak, and J. T. Chien, "Variational domain adversarial learning for speaker verification," in *Proc. Interspeech*, 2019, pp. 4315–4319.

[7] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6006–6010.

[8] D. Mekhazni, A. Bhuiyan, G. Ekladious, and E. Granger, "Unsupervised domain adaptation in the dissimilarity space for person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, Springer, 2020, pp. 159–174.

[9] J. Li, J. Han, and H. Song, "CDMA: Cross-domain distance metric adaptation for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7197–7201.

[10] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] L. Yi and M. W. Mak, "Disentangled speaker embedding for robust speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7662–7666.

[12] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[13] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4879–4883.

[14] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[15] M. I. Belghazi, A. Baratin, S. Rajeshwar, *et al.*, "Mutual information neural estimation," in *Proc. International Conference on Machine Learning*, 2018, pp. 530–539.

[16] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, *et al.*, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[18] L. Li, R. Liu, J. Kang, *et al.*, "CN-Celeb: Multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[20] Y. Fan, J. Kang, L. Li, *et al.*, "CN-Celeb: A challenging chinese speaker recognition dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7604–7608.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[23] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. European Conference on Computer Vision*, 2006, pp. 531–542.

[24] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5821–5825.

[25] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, "Analysis of Score Normalization in Multilingual Speaker Recognition," in *Proc. Interspeech*, 2017, pp. 1567–1571.

Urban Noise Monitoring using Edge Computing with CNN-LSTM on Jetson Nano

Bo Peng*, Waleed H. Abdulla†, and Kevin I-Kai Wang‡
*†‡The University of Auckland, Auckland, New Zealand
E-mail: *bpen264@aucklanduni.ac.nz, †w.abdulla@auckland.ac.nz, ‡kevin.wang@auckland.ac.nz

***Abstract*— With the negative impact of noise pollution on human health and well-being, distributed wireless sensor network with edge devices like Jetson Nano as the centre of the network for noise data classification is an effective approach to urban noise monitoring. However, most classification systems were not developed for edge computing or did not include timing performance as a metric during development. The weaknesses can limit edge computing devices' real-time ability, stability, etc. This study addresses the weaknesses by proposing a classification system using MFCC and CNN-LSTM for Jetson Nano's low-latency computation. We consider both accuracy and timing performance as critical metrics in this study. The proposed approach is based on MFCC with a window length of 0.128 seconds and overlap of 50%, classified by a CNN-LSTM trained with 30% dropout probabilities, which has obtained an accuracy of 94.94% and an execution time of 35 ms to classify 6.144 seconds audio at an audio segment length of 1.024 seconds on a Jetson Nano. The experiment demonstrates that the proposed system has outstanding accuracy and timing performance compared to the existing urban sound classification systems.**

Keywords—CNN, Jetson Nano, LSTM, Sound Classification, Urbansound8k

I. INTRODUCTION

Urban noise monitoring, including noise classification, has become increasingly important due to the negative impact of noise pollution on human health and well-being. Excessive noise levels in urban areas can lead to hearing impairment, sleep disturbance, cardiovascular diseases, and other health problems [1]–[4]. Thus, monitoring and controlling noise levels in urban areas have become essential for maintaining a healthy living environment.

A distributed wireless acoustic sensor network (WASN) approach can be used in noise monitoring systems, where edge devices like the Jetson Nano can be deployed as the centre of the sensor network for classifying noise data from sensors. Jetson Nano is compact, low-power, and capable of processing data at low latency, making them ideal for monitoring noise levels in urban areas. Using machine learning algorithms and GPU acceleration, Jetson Nano can categorise noise at high speed.

Therefore, we propose a classification system based on the Mel-scale frequency cepstral coefficients (MFCC) feature and convolutional neural networks-long short term memory (CNN-LSTM) model for urban noise classification on Jetson Nano. The system is designed to provide low-latency noise classification, enabling quick responses to noise-related issues.

In this study, our main contributions include:

- A classification system based on MFCC and CNN-LSTM is proposed to give Jetson Nano a fast and highly accurate noise classification capability.
- Experiments and comparisons are conducted to evaluate the accuracy and timing performance of different feature extraction methods, classifiers, and parameters.

The work is organised as follows. Section II briefly introduces work related to our study. A detailed description of the dataset preparation, the proposed classifier, and the experiment settings are given in Section III. In Section IV, the experimental results are given. Finally, the paper is concluded in Section V.

II. RELATED WORK

Urban noise classification is a well-researched field, with researchers investigating various classification systems using different features and classifiers. Dong et al. achieved a 94.97% accuracy using log-Mel spectrogram and raw audio classified by 2D and 1D CNN, respectively, followed by network fusion [5]. Su et al. proposed a fused two-stream CNN network, called TSCNN-DS, to classify fused-feature input, which included MFCC, log-Mel spectrogram, chroma, spectral contrast, and Tonnetz. This system achieved 97.2% accuracy on the UrbanSound8k dataset [6]. Das et al. improved the accuracy to 98.81% using MFCC and chroma short-time Fourier transform (STFT) features classified by LSTM with data augmentation, and the result was 93.76% without augmentation [7].

However, the systems mentioned above with ultra-high accuracy were not explicitly designed for edge computing devices. Besides, some of them used fused features and/or models to achieve better accuracy, which may impose a computational burden on edge devices. Vidaña-Vila et al. used Raspberry Pi to classify the spectrogram by SqueezeNet18 and obtained 80.1% accuracy [8]. Shah et al. got 88% accuracy in classifying spectrogram from European Broadcast Union normalised audio with CNN on Raspberry Pi [9]. Furthermore, Maijala et al. classified MFCC by Artificial Neural Network using Raspberry Pi and obtained 93.8% accuracy [10]. These classification systems are more lightweight than the previously introduced systems for running on an edge device, with lower cost and classification accuracy.

Although these studies worked on running classification models with edge devices, they still considered accuracy the most critical metric but ignored the execution efficiency during

This work was supported in part by the China Scholarship Council under Grant 202208250025.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

development. This renders their study insufficiently convincing, given that classification systems with higher accuracy rates are already available. In addition, few studies explored the effects of the different number of feature vectors included in each feature segment as classifier inputs on accuracy and classification timing performance. Therefore, it is necessary to develop a classification system for edge devices that considers the accuracy and timing performance as evaluation metrics and addresses the weaknesses mentioned above.

III. METHODOLOGY

A. Dataset

The most well-known datasets for noise classification include UrbanSound8K, ESC-50, and NoiseX92. UrbanSound8K contains 8732 labelled slices of audio in ten classes commonly found in urban areas; ESC-50 has 2000 slices of audio in 50 classes; NoiseX92 has fifteen audios, and each audio records a type of sound. Many studies, such as [7], have indicated that a more extensive dataset leads to a better result in deep learning models, so UrbanSound8K is the best of the three datasets totalling 8 hours and 45 minutes audio length, compared to ESC-50 with 2 hours 46 minutes and NoiseX92 with 57 minutes. Thus, we use UrbanSound8K for training and testing classification models. The ten classes in this dataset are air conditioners, car horns, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, police sirens, and street music.

Audio slices in UrbanSound8K have a sampling rate of 44.1 kHz. Since the classification system running as a part of WASN samples sounds at 8 kHz, the dataset needs to be downsampled to match the system's real-world conditions. We concatenate the resampled UrbanSound8k audios according to classes in random order into ten audios, followed by manually seeking and removing the silence and wrong labelled segments. The corrected UrbanSound8K dataset has an audio length of 7 hours and 12 minutes. Then, the dataset is scaled using the following equation

$$x' = \frac{x}{\max(\mathrm{abs}(x))}, \tag{1}$$

where x is the raw audio waveform and x' is the new audio waveform. Hence, the maximum absolute value in this dataset is 1, which is consistent with the real scenarios. Then, each audio is segmented into the training, validation, and testing datasets in a ratio of 8:1:1. We normalise the training dataset using Z-score normalisation, as in

$$Z_{train} = \frac{x'_{train} - \mu_{train}}{\sigma_{train}}, \tag{2}$$

where x'_{train}, μ_{train}, and σ_{train} represent the training dataset, the training dataset's mean value, and the standard deviation, respectively. Finally, we obtain normalised validation and testing datasets using the following equation

$$Z_{validation/test} = \frac{x'_{validation/test} - \mu_{train}}{\sigma_{train}}. \tag{3}$$

B. Features Extraction

Considering the classification system is expected to run on a Jetson Nano which has a limited computation power, the feature extraction process should be less time-consuming, and the extracted features should be as abstract as possible. The more abstract feature can express sound information with fewer coefficients to reduce the burden on the classifiers. According to this guideline, spectrograms, including Mel and log-Mel spectrograms, are not considered as they usually have more than 128 coefficients in each feature vector. We select four static features that conducted good classification accuracy in previous studies, which are MFCC, perceptual linear prediction cepstral coefficients (PLPCC), Gamma-PLP cepstral coefficients (Gamma-PLPCC), and Gamma-cepst coefficients (Gamma-CC). Each window is represented by a 16-coefficient static feature vector.

Studies such as [11]–[13] have proven that using static features alone is not robust enough, especially with low-SNR sounds, compared to the combination of static features and dynamic features (first and second derivatives of the static features). The first derivative of a static feature vector is typically defined as

$$d_t = \frac{\sum_{k=1}^{K} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2}, \tag{4}$$

where d_t represents the t-th coefficient, c_t represents the t-th static coefficient, and K represents the maximum interval to consider, which is 2 in this paper. The second-derivative vector can be obtained by substituting the first-derivative feature vector back to (4). Thus, each window can be represented with a 48-

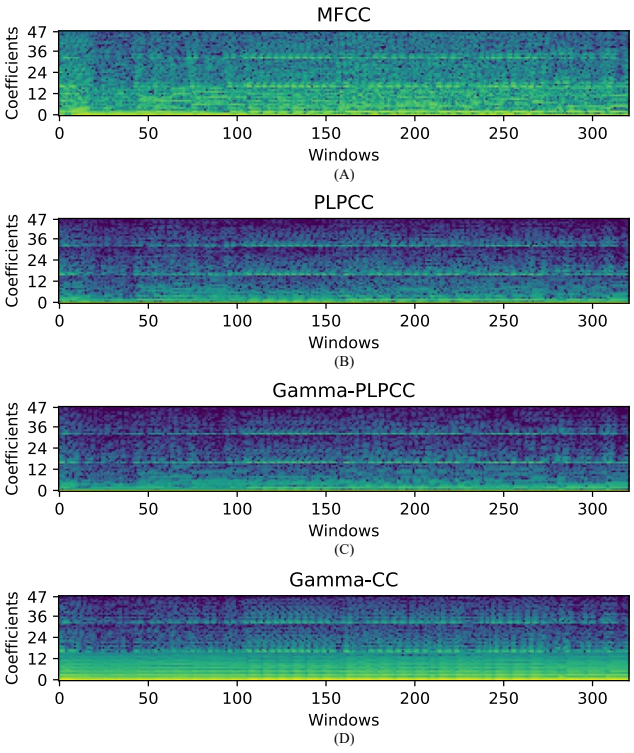


Fig. 1. Features extracted from dog-bark audio.

coefficient feature vector when the dynamic feature vector (first and second derivatives) is attached to the static feature vector.

1) MFCC (Fig. 1 (A)): MFCC is a popular and easy-to-calculate feature with good performance often used in speech recognition [14] by simulating the behaviour of human perception. Human perception of frequency is not linear and is more sensitive to low-frequency than high-frequency signals. Hence the Mel scale, a non-linear transformation of frequency, allows people to perceive signals with the same frequency difference almost equally well.

2) PLPCC (Fig. 1 (B)): Hermansky proposed PLP in 1990 based on LP because LP has the main flaw beyond 800 Hz where the spectral resolution disagrees with human hearing [15]. Unlike MFCC, which uses Mel scale filter banks, PLP uses Bark scale filter banks to simulate human hearing behaviours. Methods for computing PLPCC have been proposed in [16]–[18], which are by cepstral recursion based on the PLP results.

3) Gamma-PLPCC (Fig. 1 (C)) and Gamma-CC (Fig. 1 (D)): Gamma-PLP and Gamma-CC are proposed by Abdulla [19], which are based on Gammatone filter banks. When an external speech signal enters the cochlea's basilar membrane, it is decomposed according to frequency and produces a travelling wave vibration that stimulates the auditory receptor cells. Gammatone filters are a set of filter models that simulate the cochlea's frequency decomposition characteristics and simulate the acoustic performance of the human ear by modelling the cochlea and auditory stimuli [20]. Gamma-PLPCC is obtained from Gamma-PLP with cepstral recursion.

C. Classifiers

1) GMM: Gaussian mixture model (GMM) can be considered as a model consisting of a certain number of single Gaussian models, which are the mixture model's hidden variables. A mixture model can generally use any probability distribution, but the Gaussian distribution is chosen because it has good mathematical properties and computational performance. GMM can be described as a clustering algorithm where each cluster is modelled according to different Gaussian distributions, and any distribution with the corresponding probability may generate each data point.

2) ResNet18: We made some changes to ResNet18 to make it fit the features. Since each window is represented by a feature vector containing 48 coefficients that are highly abstract, we removed the max pool layer at the beginning of ResNet18 to ensure no coefficient is abandoned. Additionally, the convolutional layers are modified to work similarly to 1D layers along each feature vector (e.g., kernel size of (3, 1)) to process an input vector by vector. The average pool before the fully connected layer is kept for converting the data shape to 512*1*1.

3) ShuffleNetV2 1x: Similar to the modifications made on ResNet18, we removed the max pool layer and modified convolutional layers in ShuffleNet V2 for noise classification.

4) CNN-LSTM (proposed): Different to GMM and pure CNN models, CNN-LSTM considers the temporal relationship between feature vectors and has been proven by many studies to provide excellent classification accuracy. Since we use highly abstract features that barely include temporal information within each feature vector, the temporal relationship between vectors

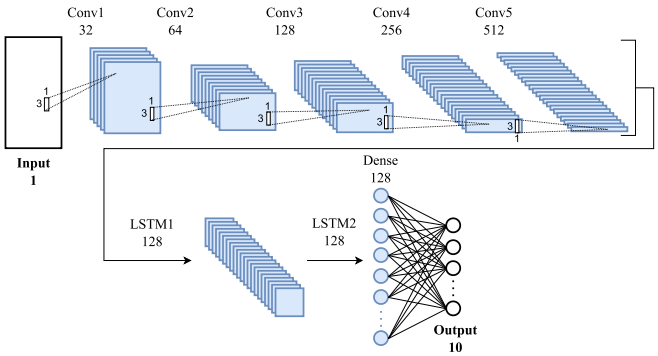


Fig. 2. Proposed CNN-LSTM structure.

TABLE I. PROPOSED CNN-LSTM STRUCTURE IN DETAIL

Layers	Hyperparameters	Output shape
Input (Feature segment)	-	(1, 48, n)
Conv-1	Output channel = 32; Kernel size = (3, 1); Stride = (2, 1); Padding = (1, 0)	(32, 24, n)
BN2D-1	Channel number = 32	(32, 24, n)
Act-1	Leaky ReLU	(32, 24, n)
Conv-2	Output channel = 64; Kernel size = (3, 1); Stride = (2, 1); Padding = (1, 0)	(64, 12, n)
BN2D-2	Channel number = 64	(64, 12, n)
Act-2	Leaky ReLU	(64, 12, n)
Conv-3	Output channel = 128; Kernel size = (3, 1); Stride = (2, 1); Padding = (1, 0)	(128, 6, n)
BN2D-3	Channel number = 128	(128, 6, n)
Act-3	Leaky ReLU	(128, 6, n)
Conv-4	Output channel = 256; Kernel size = (3, 1); Stride = (2, 1); Padding = (1, 0)	(256, 3, n)
BN2D-4	Channel number = 256	(256, 3, n)
Act-4	Leaky ReLU	(256, 3, n)
Conv-5	Output channel = 512; Kernel size = (3, 1); Stride = 1; Padding = 0	(512, 1, n)
BN2D-5	Channel number = 512	(512, 1, n)
Act-5	Leaky ReLU	(512, 1, n)
Reshape, Swap axes	-	(n, 512)
LSTM-1	Hidden size = 128	(n, 128)
Dropout	Dropout probability = 30%	(n, 128)
LSTM-2	Hidden size = 128	(n, 128)
Data select	-	(128)
BN1D	Feature number = 128	(128)
FC	Output feature = 10; Activation = Softmax	(10)

Conv = Convolutional Layer
BN = Batch Normalisation Layer
Act = Activation Layer
FC = Fully Connected Layer

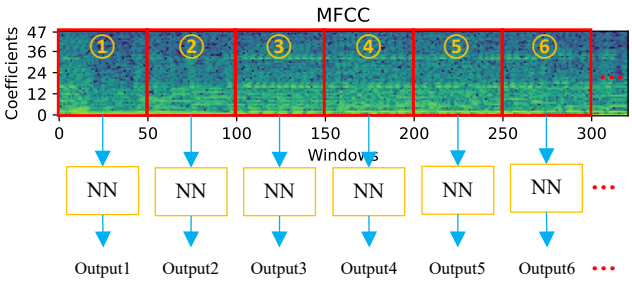


Fig. 3. Example of NNs inputs and outputs

can be critical for classification. Therefore, we propose a CNN-LSTM-based classifier (structure shown in Fig. 2) for classifying 48 coefficient-per-vector feature segments. The classifier contains 5 convolutional blocks, 2 LSTM layers, and a fully connected layer. TABLE I shows a detailed description of the classifier where "n" is the number of feature vectors in each input (feature segment) which should be greater than 1. The dropout probability at the dropout layer is set to 30% as the default unless otherwise stated.

D. Classifier Compilation in Training

1) Programming language and libraries: In its years of development, Python has provided convenient programming capabilities for numerous fields, including machine learning. Hence, we use Scikit-learn and PyTorch in Python to construct GMM and neural networks (NNs), respectively. We utilise the PyTorch with the GPU-supported version to speed up the NN computations to improve timing performance.

2) The Classifier inputs: The inputs for the classifiers, namely NNs like ResNet, ShuffleNet, and CNN-LSTM, are shaped as (1, 48, n), where "48" stands for the number of coefficients in each feature vector, and "n" represents the number of vectors in each feature segment. Fig. 3 illustrates an example of using MFCC, where each feature segment, shown by a red box, makes up 50 feature vectors representing 50 windows. Therefore, the inputs for NNs have a shape of (1, 48, 50).

The feature is calculated with a window length of 0.256 seconds and a hop length of 0.128 seconds. As a result, each feature segment is an audio segment of approximately 6.4 seconds, calculated as 0.128 seconds multiplied by 50. To represent audio segments of different lengths, we can calculate the number of vectors in each feature segment, denoted by "n", using the equation

$$n = \text{ceil}\left(\frac{L_s}{L_h}\right), \tag{5}$$

where L_s, L_h, and n represent the audio segment length, hop length, and number of vectors in each feature segment, respectively.

Unlike NNs, the GMM inputs are a sequence of independent feature vectors having a minimal number of coefficients to avoid problems such as the curse of dimensionality, computational complexity, and data sparsity. Therefore, each input in the GMM modelling is introduced by one feature vector, i.e., 48 coefficients.

3) Settings in NN training:

- Loss function: Cross entropy is used as the loss function, considering it can penalise a model more heavily for making larger errors in predicting the wrong class probabilities.
- Optimiser: We use Adam with 1e-4 weight decay as the optimiser. Adam can adjust learning rates adaptively to make changes on each parameter. We will use 1e-4 weight decay to balance the overfitting and accuracy of the models.
- Learning rate and decay: 1e-2 is set as an initial learning rate of Adam. Even though Adam can adjust the learning rate, we halve the learning rate manually when the validation accuracy is not increased for two epochs.
- Metric: We use accuracy as the evaluation metric during training. The following equation calculates accuracy:

$$\text{Accuracy} = \frac{N_{correct}}{N_{total}} \times 100\%. \tag{6}$$

- Training process: We train models using the training dataset and evaluate the models with the evaluation dataset in each epoch. The training process will terminate if the evaluation accuracy is not increasing for ten epochs. Then, the training and evaluation datasets are concatenated to train the model for one more epoch to get the final models for testing.

E. Experiment Environment and Methods

The advantage of choosing Jetson Nano over other edge devices in this study is that it supports GPU acceleration. To maximise the use of GPU acceleration, as much of the work as possible is scheduled to run using the GPU.

The classification process at edge devices should take as few computation resources as possible to avoid multiple issues, such as overheating, running out of memory, and system crashes. Hence, unlike the training process where accuracy is the only evaluation metric, we consider the execution time for classifying 6.144-second audio on a Jetson Nano as a critical metric in addition to accuracy. Detailed steps are as follows:

- Sampling a 6.144-second raw audio at 8 kHz.
- Pre-process the audio following the similar normalisation technique introduced in Section 3.1:

$$Z = \frac{x - \mu_{train}}{\sigma_{train}}, \tag{7}$$

  where x is the sampled audio, μ_{train} is the training dataset's mean value, σ_{train} is the training dataset's standard deviation, and Z is the normalised audio.
- Extract feature from the normalised audio.
- Segment the feature into feature segments.
- Classify the feature segments as input to the classifiers.

IV. Experiment Results

In this section, testing datasets are used to evaluate the classifiers. We start the experiment with an audio segment length of 4096 samples (0.512 seconds), a window length of 8192

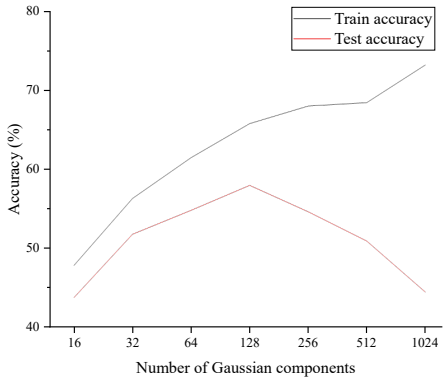


Fig. 4. Accuracy vs GMM component number

samples (1.024 seconds), and an overlap of 50%, so each input feature segment contains only one feature vector according to (5).

MFCC is used for exploring how many GMM components can provide the highest classification accuracy. Fig. 4 shows the result where 128-component GMM can provide the highest classification accuracy at 57.96%.

TABLE II. COMPARISON USING SINGLE-VECTOR INPUT

Classifiers		ResNet	ShuffleNet	GMM
MFCC	Accuracy (%)	88.47	89.78	57.96
	Time taken (ms)	104.05	88.57	-
PLPCC	Accuracy (%)	85.19	86.27	58.65
	Time taken (ms)	379.87	356.86	-
Gamma-PLPCC	Accuracy (%)	83.62	81.76	56.34
	Time taken (ms)	334.81	310.46	-
Gamma-CC	Accuracy (%)	81.09	84.49	61.42
	Time taken (ms)	109.09	88.4	-
Total NN parameters		3849034	1263422	-

Different combinations of features (MFCC, PLPCC, Gamma-PLPCC, and Gamma-CC) and classifiers (ResNet18, ShuffleNetV2 1x, and GMM) are tested, and the results are shown in TABLE II. The timing performance tests were not conducted with those combinations using GMM since GMM has significantly lower classification accuracy than ResNet18 and ShuffleNetV2. Results show that MFCC provides the highest accuracy for both NNs at around 89%. Considering the Jetson Nano's execution time and the numbers of NNs' parameters, the combination of ShuffleNetV2 1x and MFCC (highlighted) has more advantages than others.

Then, we keep the audio segment length at approximately 4096 samples and use MFCC to conduct more experiments. Multiple window lengths and overlaps are tested to determine which combination can produce better accuracy and timing performance.

TABLE III. Comparison using multiple-vector input

Window length (samples)	Overlap (%)	CNN-LSTM		ShuffleNet	
		Accuracy (%)	Time taken (ms)	Accuracy (%)	Time taken (ms)
1024	25	93.26	33.15	92.76	81.85
	50	94.21	36.91	92.55	83.29
	75	92.77	36.85	92.89	80.81
2048	25	91.96	41.82	92.08	81.9
	50	92.74	35.21	92.52	81.43
	75	93.39	32.99	91.83	83.61
4096	50	91.65	38.08	89.25	82.65
	75	92.64	43.94	92.47	89.98

CNN-LSTM can be tested since there are multiple vectors in each feature segment now. According to TABLE III, MFCC, with a window length of 1024 samples and an overlap of 50%, classified by CNN-LSTM, has the highest accuracy at 94.21%. At the same time, this combination has a short execution time, which is lower than half of those using ShuffleNetV2 1x, as shown in TABLE III. One of the reasons why CNN-LSTM has a much shorter computation time than ShuffleNet is that CNN-LSTM has 0.99M trainable parameters compared to 1.26M parameters in ShuffleNet.

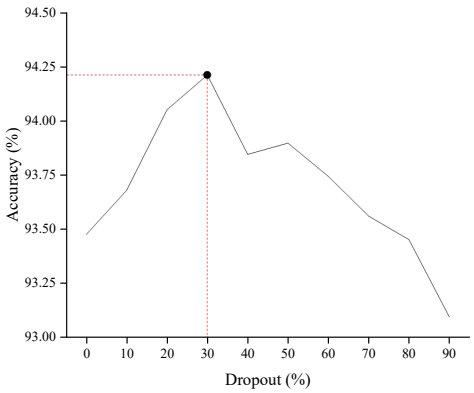


Fig. 5. Accuracy vs CNN-LSTM dropout probability

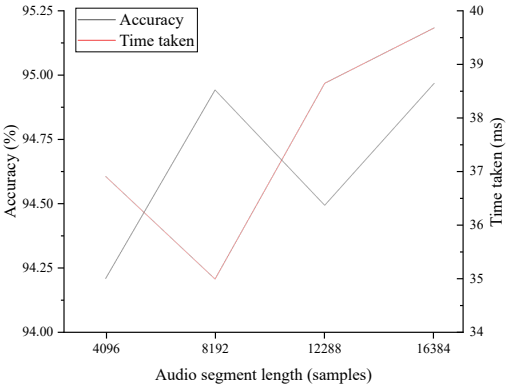


Fig. 6. Classification performance vs audio segment length

In addition, the proposed classifier, CNN-LSTM, with a dropout probability from 0% to 90%, is also tested using MFCC with a window length of 1024 samples, an overlap of 50%, and an audio segment length of 4096 samples. Results shown in Fig. 5 indicate that the classifier trained with a 30% dropout probability can obtain the highest classification accuracy at 94.21%.

Furthermore, audio segment lengths of 4096 samples (0.512 seconds), 8192 samples (1.024 seconds), 12288 samples (1.536 seconds), and 16384 samples (2.048 seconds) are tested using MFCC with a window length of 1024 samples and overlap of 50%, classified by CNN-LSTM. According to (5), each feature segment contains 8, 16, 24, and 32 feature vectors, corresponding

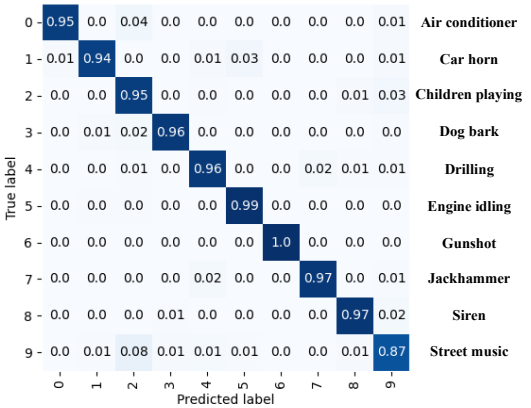


Fig. 7. Confusion matrix of the CNN-LSTM with 94.94% accuracy

to the audio segment lengths, respectively. Fig. 6 shows the results, where the execution time reaches the bottom at the audio segment length of 8192 samples. At the same time, the accuracy reaches a peak of 94.94%, and the highest accuracy of 94.966% is obtained at 16384 samples. Still, considering that the 0.026% improvement comes at the cost of a reduction in temporal resolution (e.g., to classify 49152-sample audio, 3 results are given with an audio segment length of 16384 samples versus 6 results with 8192 samples) and an increment in the execution time, we prefer the audio segment length of 8192 samples (1.024 seconds).

Fig. 7 is a normalised confusion matrix of CNN-LSTM trained with a 30% dropout probability, classifying MFCC with a window length of 1024 samples, an overlap of 50%, and an audio segment length of 8192 samples. 0 to 9 in the confusion matrix represent classes of air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music, respectively. Street music has the lowest accuracy at 87%, and it has an 8% chance of being classified into the class of children playing, which is the highest classification error rate.

TABLE IV. Proposed system vs previous systems

Network	Feature	Number of classes	Accuracy (%)	Used on edge devices
SqueezeNet 18 [8]	Spectrogram	10	80.1	Yes
ANN [10]	MFCC	10	81	Yes
D-2-DensNet [20]	MFCC + GFCC	10	84.83	No
ResNet-50 [21]	Log-Mel spectrogram	10	86.49	No
CNN [9]	Spectrogram	10	88	Yes
ShuffleNet [22]	Log-Mel spectrogram	2	93.1	No
LSTM [7]	MFCC + Chroma STFT	10	93.76	No
CNN [5]	log-Mel spectrogram + Raw audio	10	94.97	No
TSCNN-DS [6]	Chroma + Spectral Contrast + Tonnetz + Log-Mel spectrogram + MFCC	10	97.2	No
CNN-LSTM (ours)	**MFCC**	**10**	**94.94**	**Yes**

TABLE IV compares the proposed system and existing urban noise classification systems without data augmentation. Since some systems are not explicitly designed for edge devices, they may be computationally costly and low in timing performance. The proposed system has the second highest accuracy, but the highest one uses seven features and two stream CNN, which is inefficient to run on edge devices. Therefore, the proposed classification system has competitive accuracy with better timing performance.

V. CONCLUSION

This paper proposed a novel environmental noise classification system that targets running on Jetson Nano. We used a 7-hour-and-12-minute manually checked UrbanSound8K dataset to test different combinations of features, classifiers, window lengths, overlap, audio segment lengths, etc., using accuracy and timming performance as evaluation metrics, enabling the system to be accurate and efficient. The experiment results confirmed that the proposed system using MFCC with a window length of 1024 samples (0.128 seconds), overlap of 50%, and audio segment length of 8192 samples (1.024 seconds), classified by the CNN-LSTM that is trained with 30% dropout probabilities, can achieve an accuracy of 94.94% and an execution time of 35 ms for classifying a 6.144-second raw audio on a Jetson Nano. The proposed classification system shows outstanding accuracy in existing urban noise classification systems that do not use data augmentation.

REFERENCES

[1] S. A. Stansfeld and M. P. Matheson, 'Noise pollution: non-auditory effects on health', Br. Med. Bull., vol. 68, no. 1, pp. 243–257, Dec. 2003, doi: 10.1093/bmb/ldg033.

[2] M. Basner et al., 'Auditory and non-auditory effects of noise on health', The Lancet, vol. 383, no. 9925, pp. 1325–1332, Apr. 2014, doi: 10.1016/S0140-6736(13)61613-X.

[3] J. Gubbi, S. Marusic, A. S. Rao, Y. W. Law, and M. Palaniswami, 'A pilot study of urban noise monitoring architecture using wireless sensor networks', in International Conference on Advances in Computing, 2013. doi: 10.1109/ICACCI.2013.6637321.

[4] Committee on Environmental Health, 'Noise: A Hazard for the Fetus and Newborn', PEDIATRICS, vol. 100, no. 4, pp. 724–727, Oct. 1997, doi: 10.1542/peds.100.4.724.

[5] X. Dong, B. Yin, Y. Cong, Z. Du, and X. Huang, 'Environment Sound Event Classification With a Two-Stream Convolutional Neural Network', IEEE Access, vol. 8, pp. 125714–125721, 2020, doi: 10.1109/ACCESS.2020.3007906.

[6] Y. Su, K. Zhang, J. Wang, and K. Madani, 'Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion', Sensors, vol. 19, no. 7, Jan. 2019, doi: http://dx.doi.org/10.3390/s19071733.

[7] J. K. Das, A. Ghosh, A. K. Pal, S. Dutta, and A. Chakrabarty, 'Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features', in 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), Oct. 2020, pp. 1–9. doi: 10.1109/ICDS50568.2020.9268723.

[8] E. Vidaña-Vila, J. Navarro, C. Borda-Fortuny, D. Stowell, and R. M. Alsina-Pagès, 'Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring', Electronics, vol. 9, no. 12, 2020, doi: http://dx.doi.org/10.3390/electronics9122119.

[9] S. K. Shah, Z. Tariq, and Y. Lee, 'IoT based Urban Noise Monitoring in Deep Learning using Historical Reports', in 2019 IEEE International Conference on Big Data (Big Data), Dec. 2019, pp. 4179–4184. doi: 10.1109/BigData47090.2019.9006176.

[10] P. Maijala, Z. Shuyang, T. Heittola, and T. Virtanen, 'Environmental noise monitoring using source classification in sensors', Appl. Acoust., vol. 129, pp. 258–267, Jan. 2018, doi: 10.1016/j.apacoust.2017.08.006.

[11] N. Hammami, M. Bedda, and F. Nadir, 'The second-order derivatives of MFCC for improving spoken Arabic digits recognition using Tree distributions approximation model and

HMMs', in 2012 International Conference on Communications and Information Technology (ICCIT), Jun. 2012, pp. 1–5. doi: 10.1109/ICCITechnol.2012.6285769.

[12] K. Kumar, C. Kim, and R. M. Stern, 'Delta-spectral cepstral coefficients for robust speech recognition', in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4784–4787. doi: 10.1109/ICASSP.2011.5947425.

[13] V. Z. Këpuska and H. A. Elharati, 'Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions', J. Comput. Commun., vol. 3, no. 6, Art. no. 6, May 2015, doi: 10.4236/jcc.2015.36001.

[14] S. Jin, X. Wang, L. Du, and D. He, 'Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN', Appl. Acoust., vol. 172, p. 107562, Jan. 2021, doi: 10.1016/j.apacoust.2020.107562.

[15] H. Hermansky, 'Perceptual linear predictive (PLP) analysis of speech', J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752, Apr. 1990, doi: 10.1121/1.399423.

[16] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, 'Effectiveness of PLP-based phonetic segmentation for speech synthesis', in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 270–274. doi: 10.1109/ICASSP.2014.6853600.

[17] B. B. Vachhani and H. A. Patil, 'Use of PLP Cepstral Features for Phonetic Segmentation', in 2013 International Conference on Asian Language Processing, 2013, pp. 143–146. doi: 10.1109/IALP.2013.47.

[18] M. Cao, J. Wang, J. Cao, and H. Zeng, 'Acoustics recognition of excavation equipment based on MF-PLPCC features and RELM', in 2017 36th Chinese Control Conference (CCC), Jul. 2017, pp. 5400–5404. doi: 10.23919/ChiCC.2017.8028211.

[19] W. Abdulla, 'Auditory based feature vectors for speech recognition systems', Adv. Commun. Softw. Technol., pp. 231–236, Jan. 2002.

[20] Z. Huang, C. Liu, H. Fei, W. Li, J. Yu, and Y. Cao, 'Urban sound classification based on 2-order dense convolutional network using dual features', Appl. Acoust., vol. 164, p. 107243, Jul. 2020, doi: 10.1016/j.apacoust.2020.107243.

[21] A. Nasiri and J. Hu, 'SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification'. arXiv, Mar. 02, 2021. doi: 10.48550/arXiv.2103.01929.

[22] I. J. Ananya, S. Suad, S. H. Choudhury, and M. A. Khan, 'A Comparative Study on Approaches to Acoustic Scene Classification Using CNNs', in Advances in Computational Intelligence, I. Batyrshin, A. Gelbukh, and G. Sidorov, Eds., in Lecture Notes in Computer Science, vol. 13067. Cham: Springer International Publishing, 2021, pp. 81–91. doi: 10.1007/978-3-030-89817-5_6.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Random forest of Classification and Regression Tree (CART) in the estimation of SWC based on meteorological inputs and hydrodynamics behind

Tsung-Hsi Wu*, Pei-Yuan Chen†, Chien-Chih Chen*‡, Meng-Ju Chung†, Zheng-Kai Ye and Ming-Hsu Li†
* Department of Earth Sciences, National Central University, Taiwan
† Graduate Institute of Hydrological & Oceanic Sciences, National Central University, Taiwan
‡ Earthquake-Disaster & Risk Evaluation and Management Center, National Central University, Taiwan
Tsung-Hsi Wu: tsung.hsi@g.ncu.edu.tw; Pei-Yuan Chen: pychen@ncu.edu.tw
Chien-Chih Chen: chienchih.chen@g.ncu.edu.tw; Meng-Ju Chung: cc105626007@g.ncu.edu.tw
Zheng-Kai Ye: supon3060@gmail.com; Ming-Hsu Li: mli@ncu.edu.tw

***Abstract*—This study employs the random forest algorithm of Classification and Regression Trees (CART) to estimate soil water content (SWC) at shallow depths in a grassland terrain site. Leveraging meteorological parameters, the random forest model demonstrates efficient and effective SWC estimation in a 12-folds time series cross-validation. The results reveal distinct strategies employed by the CART model for different SWC depths, addressing seasonal variations and SWC sensitivity to precipitation. Additionally, the study highlights limitations of CART in extrapolating beyond training data, leading to misfit in certain scenarios. These findings contribute valuable insights for improved SWC estimation and agricultural practices.**

I. Introduction

Understanding and estimating soil water content (SWC) is crucial for various applications such as agriculture and green space management [1], [2]. Shallow Machine learning (ML) methods, such as decision trees and random forests, have been used for drought monitoring, irrigation management, and satellite-based SWC map enhancement [3]–[5]. Although deep learning (DL) methods have shown promise in numerical investigations and SWC modeling, tree-based shallow ML methods are preferred in scenarios where interpretability is important and when dealing with small datasets [6], [7].

In our previous study [8], the feasibility of using Classification and Regression Tree (CART) for SWC estimation is investigated. CART's feature selection ability allows for the inclusion of various meteorological parameters. The study uses data from a grassland terrain in Taiwan and demonstrates the effectiveness of CART in estimating SWC for agricultural irrigation and flood risk assessment.

In this study, we extend our investigation by employing a random forest of CART instead of a single tree to estimate SWC using the same dataset and feature sets. The emphasis is placed on understanding the role of rainfall, evapotranspiration, and soil texture in ML-based SWC estimation.

II. Data and Method

The study area is located at the Atmospheric and Hydrological Observatory site [9], characterized by a grassland terrain with specific meteorological and SWC measurements [10]. Meteorological data, including precipitation, air temperature, relative humidity, wind speed, atmospheric pressure, and solar radiation, were collected from various instruments. SWC was measured at different depths using specific sensors. The data covered the entire year of 2018 and underwent down-sampling and interpolation to fill in missing values.

In this study, we utilize a random forest of CART models for ensemble prediction of SWC. Each forest consists of 100 CART models, with a bagging ratio of 0.8. CART, a decision tree-based method, is a supervised machine learning algorithm that performs regression tasks by splitting the data based on features and thresholds. This splitting process is performed recursively, resulting in the creation of nested binary branches with the objective of minimizing the sum of squared errors (SSE) for each subset of data on the branch for every split [11]:

$$\min_{j,s}[\min_{c_L} \sum_{x_i \in R_L(j,s)} (y_i - c_L)^2 + \min_{c_R} \sum_{x_i \in R_R(j,s)} (y_i - c_R)^2]. \quad (1)$$

In this equation, x_i represents the input variables and y_i represents the target variable for the i_{th} observation. c_R and c_L represent the average responses (in this case, SWC) over the right and left subsets of data (R_R and R_L) respectively, after a binary split at a node using feature j with threshold s.

Table I presents the list of input and target features used for SWC estimation using CART. To incorporate the information of time points from the past, each meteorological parameter is shifted by a specific number of hours (0, -12, -24, -36, and -48). Models were trained using these shifted variants to estimate SWC at 5 depths for one hour ahead, and tested with 12-folds time-series cross validation (TSCV) as displayed in Fig. 1. The input and target features were listed in Table II;

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I
BASIC DESCRIPTIONS FOR ALL FEATURES.

variable	mean	min	median	max	unit
hour	11.5	0	11.5	23	hour
windspeed_CWB	2.46664	0.0	2.3	9.4	m/s
solar_radiation	163.532	0.0	3.467	1259.0	W/m^2
humidity_CWB	77.8678	25.0	79.0	98.0946	% RH*
pressure_CWB	998.503	977.6	1000.43	1016.0	hPa
air_temperature	22.5717	5.9393	23.0683	35.648	$^\circ C$
precipitation	0.0330226	0.0	0.0	9.0	mm
precipitation_1hr	0.16955	0.0	0.0	25.0	mm
precipitation_12hr	1.9041	0.0	0.0	68.7	mm
precipitation_1d	3.78865	0.0	0.0	113.5	mm
precipitation_2d	7.54727	0.0	0.2	178.0	mm
precipitation_3d	11.3072	0.0	1.4	226.0	mm
SWC_10cm	15.0258	6.145	15.08	38.71	% V/V†
SWC_30cm	28.3645	21.5	28.57	44.0	% V/V
SWC_50cm	29.4681	22.78	29.47	46.35	% V/V
SWC_70cm	31.1118	23.4	30.57	49.21	% V/V
SWC_100cm	20.8358	14.35	20.02	41.79	% V/V

* Relative Humidity; † Volumetric percentage

TABLE II
INPUT/TARGET FEATURES AND THEIR TIME-SHIFTED VARIANTS.

	Group tag/Feature name		Time shift (lag) of input features
input features	NP	hour	0
		windspeed_CWB	0 ,-12, -24, -36, -48
		solar_radiation	0 ,-12, -24, -36, -48
		air_temperature	0 ,-12, -24, -36, -48
		pressure_CWB	0 ,-12, -24, -36, -48
		humidity_CWB	0 ,-12, -24, -36, -48
	P	precipitation	0 ,-12, -24, -36, -48
	AP	precipitation_1hr	0 ,-12, -24, -36, -48
		precipitation_12hr	0 ,-12, -24, -36, -48
		precipitation_1d	0 ,-12, -24, -36, -48
		precipitation_2d	0 ,-12, -24, -36, -48
		precipitation_3d	0 ,-12, -24, -36, -48
target feature	soil_water_content_10cm		1
	soil_water_content_30cm		1
	soil_water_content_50cm		1
	soil_water_content_70cm		1
	soil_water_content_100cm		1

it is important to note that only one target feature is used for training and validation at a time.

III. RESULTS AND DISCUSSION

In this study, we examine the impact of individual parameters on SWC estimation by utilizing ten parameter selection sets for the random forest model of CART. Fig. 4 presents the mean absolute error (MAE) loss of these sets in a 12-folds TSCV. The ten sets consist of variations that exclude a single parameter, exclude all AP parameters, use only AP and P parameters, and include all parameters, where AP denotes accumulated precipitation, P denotes precipitation and NP denotes non-precipitation parameters, as listed in Table II.

Results reveal that excluding a single NP or P parameter has minimal impact on the model's loss, suggesting the compensatory effect of other parameters. In contrast, removing all AP parameters significantly reduces the model's performance. Notably, the random forest model struggles to provide accurate SWC estimates without NP parameters, as demonstrated by the "only AP & P" set. Additionally, the findings indicate that AP parameters are crucial for estimating shallow SWC depths, while NP parameters play a vital role in estimating deeper SWC depths, as evident from the "w/o AP" and "only AP & P" results. In addition, the loss in MAE of the previous study [8] is demonstrated as the size of green hollow circles in the same plot for comparison. The results show a high consistency between the two studies, with ensemble prediction consistently bringing modest improvements for all parameter selection sets.

The fitting results of the 12-folds TSCV for different depths of SWC measurements are presented in Fig. 2. Overall, the trained models perform well, with MAE in general of less than 4% (v/v) except for the 3rd and 8th fold, as demonstrated in Fig. 3. In the case of the 3rd fold, the SWC of depths ≥ 50 cm reached historically low levels not present in the training data. Since the CART algorithm is intrinsically unable to give predicted values beyond the range of those in the training data, the misfit is high.

Hydrological factors also play a role. These factors are manifested in Fig. 2 from the 3rd to the 7th fold, where the responsiveness of SWC to non-intensive precipitation decreases at deeper depths. In contrast, SWC from the 8th to the 9th fold shows a clear response to precipitation at all depths. The different behavior in terms of the responsiveness of SWC to precipitation between deeper and shallow depths is a result of the evapotranspiration (ET) mechanism at this site. ET often reaches only to depth ≤ 50 cm, and only in relatively large or consecutive rainfall events can the precipitation reach a greater depth. Accordingly, during the 8th fold period, intense rainfall events result in a sudden change in the responsiveness of SWC to precipitation at deeper depths. As a consequence, the model trained with smaller rainfall amounts and more frequent drying periods gave erroneous predictions of SWC, leading to a distinctively larger misfit in this fold.

IV. CONCLUSION

In conclusion, our study demonstrates the effectiveness of random forest models based on the CART algorithm for estimating SWC at shallow depths in a grassland terrain in northern Taiwan. The random forest approach proves to be efficient in both training and prediction, yielding acceptable errors for agricultural purposes. The 12-fold TSCV analysis conducted throughout the entire year of 2018 reveals that the variation in misfit across TSCV can be attributed to the learning capabilities of the CART model in different layers and the characteristics of rainfall and evapotranspiration during different seasons. Notably, the CART model developed distinct strategies for estimating SWC at the shallowest depth (10 cm) and the deepest depth (100 cm). At 10 cm, the model relies more on AP parameters, while at 100 cm, it relies more on NP parameters. This accounts for the inconsistent sensitivity of SWC to precipitation between dry and wet seasons. These findings not only provide insights into the near-surface soil structure of the study area but also have

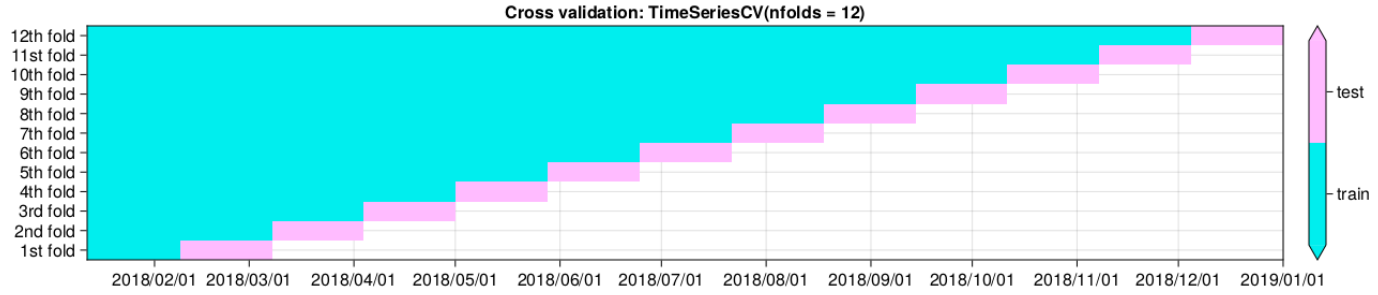


Fig. 1. Subsets of training-testing data for cross validation in this study.

implications for improving SWC estimation using other techniques beyond CART. Furthermore, our results exhibit a high level of consistency with the previous study, with ensemble predictions consistently showing modest improvements across various parameter selection sets.

Our work has resulted in the development of TWAISWCF.jl (Fig. 5 and Ref. [12]), a Julia package in early stage specifically designed for estimation of SWC in agricultural applications. In terms of the machine learning workflow, TWAISWCF.jl currently covers major stages such as data preprocessing, data splitting, and model training. It incorporates DecisionTree.jl [13] in the framework of MLJ [14] utilizing the CART algorithm based random forest strategy as the underlying foundation, and leverages the results of this study in processing the meteorological data. In terms of design pattern, TWAISWCF.jl stands out for its utilization of polymorphism through Julia's multiple dispatch mechanism, enabling a highly customizable and extensible API for integration with other collaborative applications. Furthermore, TWAISWCF.jl has been made freely available to the public, allowing researchers and practitioners to explore and experiment with the package for SWC estimation. We envision that its utilization will contribute to advancements in agricultural irrigation practices and enhanced flood risk assessment.

ACKNOWLEDGEMENT

The authors of the manuscript gratefully acknowledge the laboratories providing valuable data, namely the Surface Hydrology Lab., Aerosol and Cloud Lab., and Planetary Boundary Layer and Air Pollution Lab. of National Central University. The authors also would like to express their gratitude to the National Science and Technology Council (R.O.C) for providing funding for this research through grant number MOST 110-2634-F-008-008.

REFERENCES

[1] A. Garg, V. K. Gadi, Y.-C. Feng, *et al.*, "Dynamics of soil water content using field monitoring and AI: A case study of a vegetated soil in an urban environment in China," *Sustainable Computing: Informatics and Systems*, vol. 28, p. 100 301, Dec. 2020, ISSN: 2210-5379. DOI: 10.1016/j.suscom.2019.01.003.

[2] R. Sobayo, H.-H. Wu, R. Ray, and L. Qian, "Integration of convolutional neural network and thermal images into soil moisture estimation," in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, IEEE, 2018, pp. 207–210, ISBN: 1-5386-5762-7.

[3] Y. Tramblay and P. Quintana Seguí, "Estimating soil moisture conditions for drought monitoring with random forests and a simple soil moisture accounting scheme," *Natural Hazards and Earth System Sciences*, vol. 22, no. 4, pp. 1325–1334, Apr. 2022, ISSN: 1561-8633. DOI: 10.5194/nhess-22-1325-2022.

[4] Y. Huang, H. Jiang, W.-f. Wang, W. Wang, and D. Sun, "Soil moisture content prediction model for tea plantations based on SVM optimised by the bald eagle search algorithm," *Cognitive Computation and Systems*, vol. 3, no. 4, pp. 351–360, 2021, ISSN: 2517-7567. DOI: 10.1049/ccs2.12034.

[5] Y. Liu, Y. Yang, W. Jing, and X. Yue, "Comparison of Different Machine Learning Approaches for Monthly Satellite-Based Soil Moisture Downscaling over Northeast China," *Remote Sensing*, vol. 10, no. 1, p. 31, Jan. 2018, ISSN: 2072-4292. DOI: 10.3390/rs10010031.

[6] S. Chauhan, L. Vig, M. De Filippo De Grazia, M. Corbetta, S. Ahmad, and M. Zorzi, "A Comparison of Shallow and Deep Learning Methods for Predicting Cognitive Performance of Stroke Patients From MRI Lesion Images," *Frontiers in Neuroinformatics*, vol. 13, 2019, ISSN: 1662-5196.

[7] J. Padarian, B. Minasny, and A. B. McBratney, "Using deep learning to predict soil properties from regional spectral data," *Geoderma Regional*, vol. 16, e00198, Mar. 2019, ISSN: 2352-0094. DOI: 10.1016/j.geodrs.2018.e00198.

[8] T.-H. W. et al., "Classification and regression tree (cart)-based estimation of swc based on meteorological inputs and hydrodynamics behind," Manuscript submitted to SUSCOM for publication, Jan. 2023.

[9] National Central University, NCU, *Atmospheric and Hydrological Observatory*, http://naho.atm.ncu.edu.tw/Home.html.

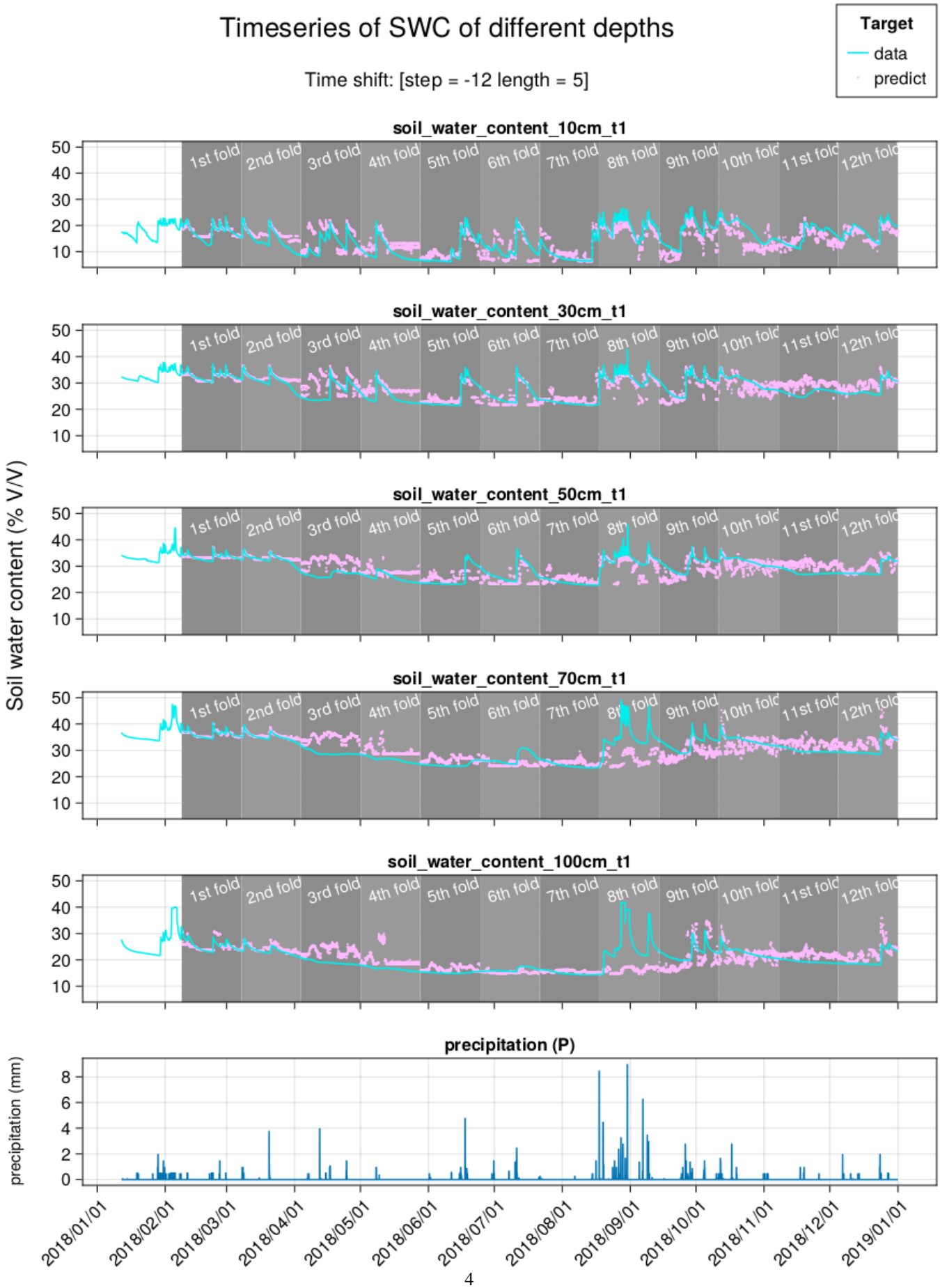


Fig. 2. Time-series of soil water content predictions in the testing phases of the 12 folds (scattered points), referencing real observations (soild line); with results of soil water content of different depths separately demonstrated in each subplot.

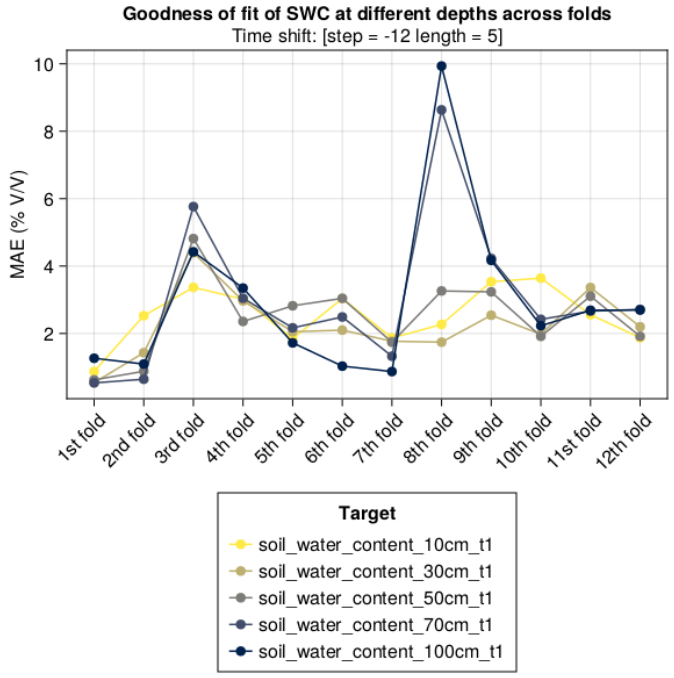


Fig. 3. Loss in MAE across 12 folds for all SWC depths, referring the fitting goodness of the results demonstrated in Fig. 2.

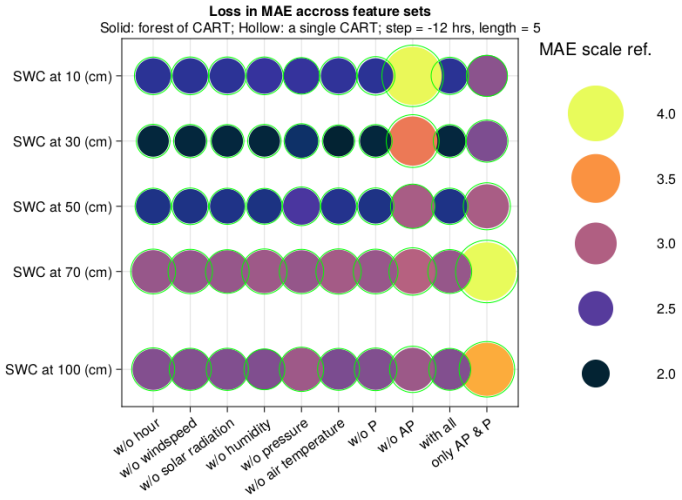


Fig. 4. MAE in the 12-fold TSCV of ten sets of parameter selection for estimating SWC at five depths. The magnitude of MAE is represented by the colors or sizes of circles, with solid circles representing results of random forest, and hollow circles on top of the solid circles representing results of the previous study [8] which uses only a single CART as the model.



Fig. 5. TWAISWCF.jl is an open-source Julia package for the research of SWC estimation in agriculture that utilizes the CART algorithm-based random forest strategy and common meteorological parameters. Following the instructions on the website https://okatsn.github.io/TWAISWCF/dev/ to train and test the tree-based model for SWC estimation with your own data or the example dataset.

[10] Surface Hydrology Lab. of NCU, *Aerial view of NCU Atmospheric and Hydrological Observatory*, https://hydro.ihs.ncu.edu.tw/.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY: Springer, 2009, ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7.

[12] W. Tsung-Hsi, *TWAISWCF.jl, a julia package for soil water content estimation*, https://github.com/okatsn/TWAISWCF. [Online]. Available: https://okatsn.github.io/TWAISWCF.

[13] B. Sadeghi, P. Chiarawongse, K. Squire, *et al.*, *DecisionTree.jl - A Julia implementation of the CART Decision Tree and Random Forest algorithms*, version 0.11.3, Nov. 2022. DOI: 10.5281/zenodo.7359268. [Online]. Available: https://doi.org/10.5281/zenodo.7359268.

[14] A. D. Blaom, F. Kiraly, T. Lienart, Y. Simillides, D. Arenas, and S. J. Vollmer, "MLJ: A julia package for composable machine learning," *Journal of Open Source Software*, vol. 5, no. 55, p. 2704, 2020. DOI: 10.21105/joss.02704. [Online]. Available: https://doi.org/10.21105/joss.02704.

A Framework for Reusing Earth Science Data on Data and Model Marketplaces

Chung-I Huang, Jih-Sheng Chang, Chen-Kai Sun, Taichi Wang, Wei-Yu Chen, Hui Hung Yu, Wen-Yi Chang, Fang-Pang Lin
National Center for High-performance Computing (NCHC) ,Taiwan

Abstract: The AI-based Management Strategy for Earth and Environmental Disaster project proposes a framework for integrating AI technology into Earth environmental disaster prediction models. The project emphasizes the collaboration of multidisciplinary experts, such as information engineering and network engineering, to explore the effectiveness of AI technology in enhancing Earth environmental disaster prediction models. The project particularly highlights the concepts of data reuse and model reuse, combining the data marketplace and model marketplace platforms of the NCHC to establish a national-level data service platform. Through the data marketplace platform, data reuse is achieved, encouraging cross-disciplinary collaboration and fostering innovative applications and decision-making recommendations. Simultaneously, the model marketplace platform allows AI models from Earth science sub-projects to be launched, realizing model reuse and providing a strong reference for Taiwan's disaster response decision-making system, contributing to sustainable development.

Introduction: Natural disasters, such as earthquakes, typhoons, and floods, can have a significant impact on people's lives and the environment. The prediction and management of such disasters have become increasingly critical. The use of AI technology in disaster prediction models has shown great potential in improving accuracy and efficiency. This paper presents the AI-based Management Strategy for Earth and Environmental Disaster project, which aims to integrate AI technology into Earth environmental disaster prediction models and establish a framework for data and model reuse (1-4).

Several studies have shown the potential of AI technology in improving the accuracy and efficiency of disaster prediction models (5-7). These studies also suggest that collaboration among experts from different fields is crucial for enhancing the effectiveness of such models. The concept of data and model reuse has also been discussed in the literature, with several studies suggesting that it can improve resource utilization efficiency and promote the development of innovative applications (8-12).

Methods:

The research methodology for this project includes a literature review, case studies, and interviews with experts. The case studies focus on AI technology's application in Earth environmental disaster prediction models, while the interviews with experts explore their perspectives on the effectiveness of AI technology in enhancing such models and the importance of data and model reuse.

The concept of data and model reuse is a key aspect of this project. Data and model reuse helps avoid duplicative efforts in creating similar databases or models, saving time and economic costs while improving resource utilization efficiency. Sharing data and models enables experts from different fields to easily access the necessary resources, thus strengthening cross-disciplinary communication and collaboration and

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

promoting the development of innovative applications. By utilizing this framework, secondary innovations can be made based on existing foundations, leading to the development of new application services and decision-making recommendations, and promoting the advancement of relevant fields.

The data and model marketplace platforms of the NCHC will be crucial for achieving data and model reuse in this project. The data marketplace platform enables the sharing of diverse types of data, including meteorological data, geological data, and demographic data, among others. Through this platform, cross-disciplinary collaboration can be encouraged, and innovative applications and decision-making recommendations can be developed. The model marketplace platform, on the other hand, enables AI models from Earth science sub-projects to be launched, providing a strong reference for Taiwan's disaster response decision-making system. This model reuse can lead to significant improvements in the accuracy and efficiency of disaster prediction models.

In addition to data and model reuse, interdisciplinary collaboration is also a key aspect of this project. Collaboration among experts in Earth environmental research, information engineering, and network engineering will be crucial for enhancing the effectiveness of Earth environmental disaster prediction models. Through cross-disciplinary collaboration, new perspectives and innovative solutions can be developed.

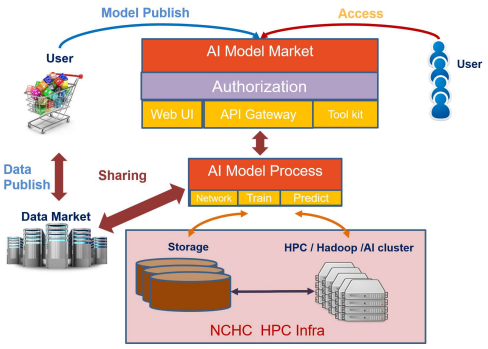


Fig.1 AI-based Management Strategy for Earth and Environmental Disaster Project Flowchart and Architecture

The flowchart and architecture of the AI-based Management Strategy for Earth and Environmental Disaster project highlights the importance of data and model reuse in disaster prediction models. Users can upload data to the data marketplace, and AI models can be launched on the model marketplace. The AI models can directly stream data from the data marketplace, and other authorized users can subscribe to the model marketplace to access the relevant data and models. This framework promotes cross-disciplinary collaboration and resource utilization efficiency, as users can easily access the necessary resources and develop innovative applications and decision-making recommendations. The data and model marketplace platforms of the NCHC enable data and model reuse and can significantly improve the accuracy and efficiency of disaster prediction models. By utilizing this framework, secondary innovations can be made based on existing foundations, leading to the development of new application services and decision-making recommendations, and promoting the advancement of relevant fields.

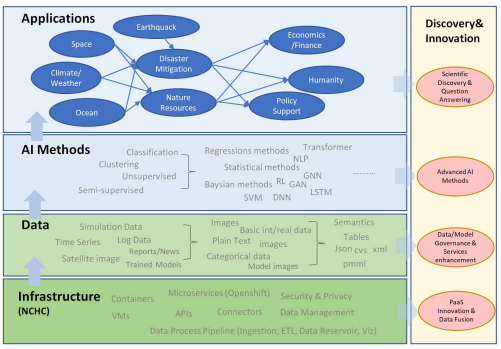


Fig.2 End-to-End AI-based Disaster Management System Infrastructure and Workflow

The "End-to-End AI-based Disaster Management System Infrastructure and Workflow" is a comprehensive framework for disaster prediction and management. The system includes multiple components, starting from the infrastructure, data, AI models, and application. The infrastructure component focuses on developing a robust and scalable system that can handle large amounts of data. The data component includes diverse types of data, including meteorological, geological, and demographic data, among others. The AI models component focuses on developing accurate and efficient models for disaster prediction and management, utilizing advanced machine learning algorithms. The application component includes a variety of innovative applications and decision-making recommendations, developed based on the outputs of the AI models. The end-to-end workflow of this system provides a comprehensive approach to disaster prediction and management, utilizing the latest technologies and promoting interdisciplinary collaboration. This framework has the potential to significantly improve the accuracy and efficiency of disaster prediction models, while promoting resource utilization efficiency and innovative applications.

Results: The results show that the integration of AI technology into Earth environmental disaster prediction models can significantly improve their accuracy and efficiency. The project's emphasis on data and model reuse is crucial for avoiding duplicative efforts in creating similar databases or models, saving time and economic costs while improving resource utilization efficiency. Sharing data and models enables experts from different fields to easily access the necessary resources, thus strengthening cross-disciplinary communication and collaboration and promoting the development of innovative applications. The AI-based Management Strategy for Earth and Environmental Disaster project's outcomes will comprehensively assess the impact on legislation, society, and humanities, ensuring that environmental protection and social equity and justice are considered while pursuing technological progress. The TWAI project will promote more breakthroughs in Earth environmental research and related fields in Taiwan, showcasing Taiwan's technical capabilities internationally. Through the implementation of this project, Taiwan will be able to establish a more robust disaster response decision-making system and enhance the nation's overall capacity in dealing with Earth environmental disasters. Additionally, the project's outcomes will benefit the research and development of related fields both domestically and internationally, making a significant contribution to Taiwan's sustainable development.

Conclusion: The AI-based Management Strategy for Earth and Environmental Disaster project proposes a framework for integrating AI technology into Earth environmental

disaster prediction models and establishing a framework for data and model reuse. The project emphasizes the collaboration of multidisciplinary experts to enhance the effectiveness of Earth environmental disaster prediction models. The results suggest that the integration of AI technology and the emphasis on data and model reuse can significantly improve the accuracy and efficiency of disaster prediction models, while promoting resource utilization efficiency and innovative applications. This project also underscores the importance of interdisciplinary collaboration, particularly among experts in Earth environmental research, information engineering, and network engineering. The project's outcomes will contribute to sustainable development and showcase Taiwan's technical capabilities internationally.

Acknowledgements:
We would like to express our sincere gratitude to the National Science Council of Taiwan (NSC) for their support and funding of this project. This research was supported by NSC grant number NSTC 111-2634-F-008-001. We would also like to thank all the experts who participated in the interviews and provided valuable insights on the effectiveness of AI technology in enhancing Earth environmental disaster prediction models. Additionally, we would like to express our appreciation to the National Center for High-performance Computing (NCHC) for providing the data and model marketplace platforms that were crucial for achieving data and model reuse in this project. Finally, we would like to thank all the members of our team for their hard work and dedication throughout this project.

References:

1. Chang, L. C., Chen, Y. H., & Lee, C. T. (2018). A study of typhoon prediction using artificial intelligence models. Applied Sciences, 8(7), 1145.
2. Duan, J., Song, Z., Wang, C., Liu, Y., & Yang, L. (2019). Multi-model ensemble prediction of flood inundation with artificial neural networks and support vector machines. Remote Sensing, 11(17), 1947.
3. Han, D., & Hwang, D. (2020). Integrating artificial intelligence and GIS for disaster management: A review. International Journal of Disaster Risk Reduction, 45, 101465.
4. Hossain, M. S., Sultana, S., & Purohit, N. (2019). Application of artificial intelligence in disaster management: A review. Sustainable Cities and Society, 48, 101518.
5. Alizadeh, A. H., & Ghaffarian, H. R. (2019). Prediction of water quality parameters using artificial neural network (ANN): A case study of the Karoon River, Iran. Environmental Earth Sciences, 78(18), 543.
6. Chen, K. H., Chiang, C. H., & Kao, S. H. (2019). Deep learning for typhoon precipitation prediction. Journal of Hydrology, 572, 153-162.
7. Chen, K. H., Chiang, C. H., & Kao, S. H. (2020). Deep learning for typhoon wind prediction. Natural Hazards, 104(2), 1491-1507.
8. Jang, J., & Lim, C. (2019). An application of data mining and machine learning techniques for water resources management: A review. Water, 11(3), 414.

9. Jia, P., Sun, P., Wang, P., & Hu, K. (2019). Flood inundation prediction using artificial intelligence models: A review. Water, 11(10), 2084.
10. Kim, T., & Lee, Y. (2020). A review of AI-based approaches for natural disaster management: Challenges and future directions. Sustainability, 12(11), 4518.
11. Li, Y., Qin, H., Liu, G., & Sun, X. (2019). Application of machine learning methods in flood disaster risk assessment: A review. International Journal of Disaster Risk Science, 10(4), 429-440.
12. Qu, J., & Li, X. (2020). Review of artificial intelligence for disaster management: Trends, challenges, and future directions. Journal of Cleaner Production, 263, 121551.

Impact of the weighted loss function on the innovative CMAQ-CNN $PM_{2.5}$ forecasting model

Yi-Ju Lee*, Fang-Yi Cheng, Chih-Yung Feng, and Zhih-Min Yang
* National Central University, Taiwan
E-mail: lili82628@gmail.com Tel: +886-3-4227151 #65556
E-mail: bonniecheng18@gmail.com Tel: +886-3-4227151 #65508
Manysplendid Infotech, Taipei, Taiwan
E-mail: cyfeng@manysplendid.com Tel: +886-2-29334260
E-mail: zhiminyang@manysplendid.com Tel: +886-2-29334260

Abstract — **Developing an accurate air quality forecasting model with a lead time longer than 72 hours is important for predicting hazardous pollution events. An air quality forecasting system (AQF) conducted with the WRF-CMAQ numerical model has been developed. This study used the predicted $PM_{2.5}$ from AQF, observed $PM_{2.5}$, longitude, latitude, the land use index of the corresponding grid, and indexes of synoptic weather patterns to develop a forecasting model through the CNN algorithm to predict 72 hours of $PM_{2.5}$ concentrations at 75 surface air quality monitoring sites. The training dataset was from October 2019 to September 2021, and the data from October 2021 to September 2022 was divided into validation and test datasets in order of date. Since the $PM_{2.5}$ varies in a wide range, the weighted loss function was applied to improve the model performance. In general, the CMAQ-CNN models developed by this study show better performance than the AQF, especially during high pollution events when the CMAQ-CNN model utilizes the weighted loss function.**

I. Introduction

Previous research has demonstrated the adverse impacts of fine particular matter ($PM_{2.5}$) on human health, visibility, and global climate change. Developing a reliable air quality forecasting model with a lead time length longer than 72 hours is crucial for environmental agencies to predict severe pollution events in advance, control the concentration of ambient pollutants, and reduce the population's exposure to health dangers.

In recent years, various $PM_{2.5}$ forecasting models based on algorithms linked to machine learning have been developed. The convolutional neural network (CNN) contains one or several convolutional layers that can extract local characteristics from the input parameters. The forecasting model based on CNN with a large number of characteristic parameters through 1-D convolution could tackle the time-series data and showed better performance than using neural networks and a numerical model [1,2].

Since the machine learning algorithms are a statistical model, the developed models would highly depend on the training dataset. In the one-year data of each monitoring station, the observed high-level $PM_{2.5}$ concentration only accounts for a minor portion, which leads to a poor prediction of high concentration events. The weighted loss functions were often used to construct classification models with imbalanced input data. Ref. [3] applied the weighted loss function to solve the imbalanced rainfall levels and improve the precipitation nowcasting.

Considering the practical operation, this study develops a CMAQ-CNN forecasting model with as few auxiliary features as possible to avoid increasing the time cost of data preprocessing. Moreover, the impact of the weighted loss function in the air quality forecasting model will be discussed.

II. Data and Data Preprocessing

A. Observation datasets

The hourly $PM_{2.5}$ concentration observed at 75 surface air quality monitoring stations (Fig. 1) operated by the Taiwan Environmental Protection Agency (EPA) from October 2019 to September 2022 was collected. According to the characteristics of the emission and meteorological conditions, the EPA divided Taiwan into seven air quality zones (AQZs), which are: northern Taiwan (NT), the Chu-Miao area (CM), central Taiwan (CT), the Yun-Chia-Nan area (YCN), the Kao-Ping area (KP), the Yilan (YI), and the Hua-Dong area (HD) (refer to Fig 1). In this study, the YI and HD areas are combined into the Yi-Hua-Dong area (YHD), where the air quality tends

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

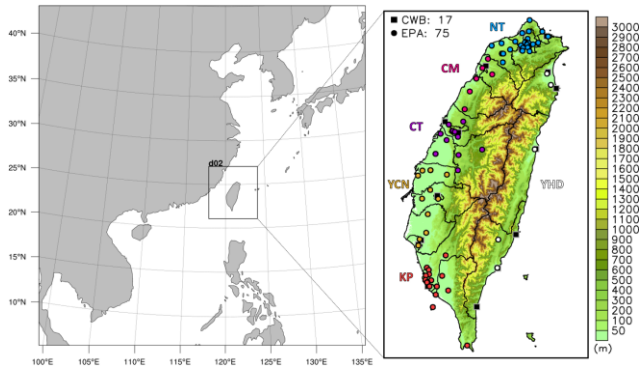


Fig. 1 The left panel displays the simulation domains. The locations of CWB and EPA surface monitoring stations are shown in the right panel. The shaded color is the elevation. The six AQZs (NT, CM, CT, YCN, KP, and YHD) are identified with different colors.

to be in good condition.

In addition, the meteorological parameters, including surface pressure, wind speed, and wind direction, were obtained from 17 surface weather stations (Fig. 1) developed by the Central Weather Bureau (CWB) in Taiwan. These daily data were used to identify the synoptic weather pattern of each day during the study period with the cluster analysis method [4].

This study objectively removed the abnormal values and outliers for each station to ensure the dataset's quality. The distribution of $PM_{2.5}$ concentration in each year would not be close to the normal distribution. After redistributing the data by taking the n-th root to make the absolute value of the skewness close to 0, we remove the data not in the interval of [$\overline{x}$-4*x_{std}, $\overline{x}$+4*x_{std}], where $\overline{x}$ and x_{std} are the average and standard deviation of one-year data, respectively. This quality assurance method was also applied to meteorological parameters.

B. Community Multiscale Air Quality forecasting system (AQF)

To forecast the concentration of ambient pollutants in Taiwan for a lead time of 96 hours, an air quality forecasting system (AQF) using Weather Research and Forecasting (WRF) version 3.8.1 [5] and Community Multiscale Air Quality (CMAQ) version 5.2 [6,7] was developed [8]. Two nested domains with resolutions of 15 km and 3 km are used to capture the synoptic weather pattern and background concentrations and long-range transported air pollutants in the coarse domain (d01) and local formation and transportation of air pollutants in the fine domain (d02) (Fig. 1). The biogenic and anthropogenic emissions are provided by the Model of Emissions of Gases and Aerosols from Nature (MEGAN) [9] in both domains and the Model Inter-Comparison Study for Asia (MICS-Asia) based on 2010 [10] in d01 and the Taiwan Emission Data System version 10 (TEDS10) based on 2016 [11] in d02, respectively. Besides, the CMAQ model executed with the Carbon Bond (CB05) [12] and aerosol module version six (AERO6) [13] chemical mechanisms.

The AQF is conducted with numerical models that are based on the atmospheric governing equations, and is able to predict the air pollutants concentrations. The prediction of 72-hr $PM_{2.5}$ concentration from d02 of the AQF during the study period were collected for developing the CMAQ-CNN forecasting model.

III. CMAQ-CNN Model

A. Input features

The input features can be split into time-dependent and time-independent components for each sample. The time-dependent elements include 24-hr observed $PM_{2.5}$ from the previous day and 72-hr predicted $PM_{2.5}$ from AQF.

For the time-independent parameters, the longitude and latitude of each air quality monitoring station are selected to depict the spatial variation of air pollutants. The land use (LU) index of the corresponding grid used in the AQF is also chosen to indirectly provide information on emission sources.

Ref. of [4] indicated that the synoptic weather pattern has a significant impact on air pollutant concentrations. The six synoptic weather patterns are identified with the cluster analysis method. This study replaced a substantial number of meteorological parameters with the synoptic weather patterns indexes. Four transformed indexes, including those for the previous day and the next three days (forecasting 1-24 hr: Day 1, 25-48 hr: Day 2, and 49-72 hr: Day 3), are included in the input sample to represent the changing pattern of the synoptic weather system.

The scales of input parameters vary in different ranges, which may lead to the misjudge of the importance of variables and also affect the efficiency of finding the best solution in the algorithms. Thus, the Min-Max normalization rescales each input feature to the range [0, 1].

B. Construction of model

Fig. 2 illustrates the structure of the model that was built to predict 72 hours of $PM_{2.5}$ concentrations at 75 surface air quality monitoring sites. Because the characteristics of concentration variations might differ in observation and simulation results, two datasets utilized various convolutional layers. The 1-D convolutional layers with a kernel size of 3 are used to address the time-varying properties of the air pollutants. Then, the convolutional layer outputs are combined with time-independent parameters. To capture the non-linear relationship

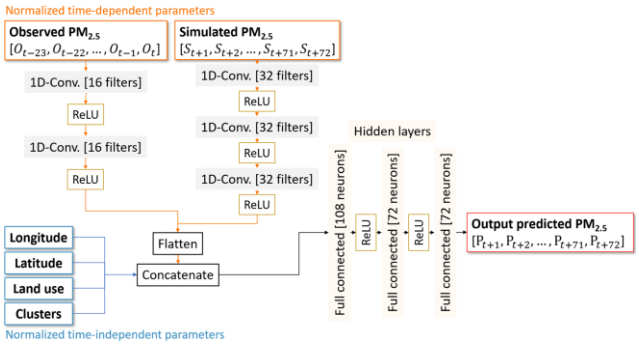


Fig. 2 The structure of the CMAQ-CNN forecasting model in this study.

between the input features and true $PM_{2.5}$ concentration, three full-connected layers, which had 108, 72, and 72 neurons with a Rectified Linear Unit (ReLU) activation function, were built.

The two-year data from October 2019 to September 2021 was used to train the CMAQ-CNN model, while the data from October 2021 to September 2022 was divided into validation and test datasets. The early-stopping was applied to avoid overfitting.

C. Experimental designs

This study constructed two CMAQ-CNN models: CNN_Cluster8 and CNN_Cluster8_WMSE. Table 1 shows the occurrence percentages of different $PM_{2.5}$ level in each AQZ. The occurrence frequency of high $PM_{2.5}$ concentrations is very low, which can degrade the performance of severe $PM_{2.5}$ events. As a result, the loss function of the Cluster8_WMSE experiment was applied with a weighted mean squared error as expressed by (1) and (2):

$$WMSE = \left(\sum_{i=1}^{72} W_i \times (P_i - O_i)^2\right)/72 \quad (1)$$

$$W_i(O_i) = \begin{cases} 1, & O_i < 20 \\ 1.5, & O_i \geq 20 \; and \; O_i < 30 \\ 2, & O_i \geq 30 \; and \; O_i < 40 \\ 3, & O_i \geq 40 \; and \; O_i < 50 \\ 4, & O_i \geq 50 \; and \; O_i < 60 \\ 5, & O_i \geq 60 \; and \; O_i < 70 \\ 6, & O_i \geq 70 \end{cases} \quad (2)$$

where P_i and O_i are the prediction and observation of i-th hour, respectively. The weight (W_i) of each forecasting hour is the function of observation value.

IV. Results and Discussions

varies from 7.63 to 14.36 μg/m^3, while the RMSE in eastern Taiwan is lower between 4.54 and 7.7 μg/m^3 due to the low air

The evaluation was conducted for the test datasets from two CMAQ-CNN models, and the AQF system. The RMSE of AQF predicted 72-hr $PM_{2.5}$ concentrations in western Taiwan

Table 1 The occurrence percentage (%) of hourly $PM_{2.5}$ concentration in each $PM_{2.5}$ levels for each AQZ from the training dataset.

Observed $PM_{2.5}$ (μg/m^3)	AQZ					
	NT	CM	CT	YCN	KP	YHD
<20	79.7	76.4	63.4	57.1	51.4	92.7
20–30	12.7	14.8	20.2	22.2	23.5	4.4
30–40	3.9	4.9	9.2	10.8	14.3	0.6
40–50	1.2	1.5	3.4	4.4	5.5	0.1
50–60	0.4	0.5	1.3	1.9	1.9	---
60–70	0.1	0.2	0.4	0.9	0.7	---
≥70	0.1	0.1	0.3	0.8	0.4	---

pollutants concentration (Fig. 3). The largest predicted bias exists in the KP AQZ and followed by the YCN AQZ. Both the CNN_Cluster8 and CNN_Cluster8_WMSE models had better performance than AQF, with the RMSE in western Taiwan ranging from 5.3 to 9.7 μg/m^3 and 5.88 to 12.07 μg/m^3, respectively. When the CNN model tends to focus on the high $PM_{2.5}$ level through the weighted loss function (Cluster8_WMSE experiment), it is prone to having larger biases. The spatial patterns of RMSE of the two CMAQ-CNN models are similar, which is higher in southern Taiwan than in northern Taiwan. The average RMSE from AQF is 10.48 μg/m^3, which can be reduced to 7.24 μg/m^3 by the CNN_Cluster8 and 8.3 μg/m^3 by the CNN_Cluster8_WMSE. Table 2 provides the performances of each model during different forecasting periods in each AQZ.

The AQF is based on the numerical model of atmospheric dynamic, physical, and chemical processes to describe the variations of air pollutant concentrations. Fig. 4 compares the mean RMSE and correlation coefficient (r) between AQF and two CMAQ-CNN models during 72 forecasting hours. The

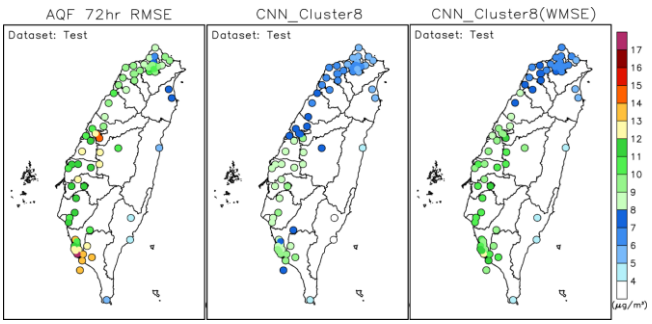


Fig. 3 The RMSE of 72-hr forecasted $PM_{2.5}$ concentration of the AQF (left), CNN_Cluster8 (center), and CNN_Cluster8_WMSE (right) models.

Table 2 The RMSE of 72-hr, Day1 (1-24 hr), Day 2 (25-48 hr), and Day 3 (49-72 hr) forecasted hourly $PM_{2.5}$ concentration of each model in each AQZ and also the average of all stations. The number in the brackets represent the amount of monitoring stations in that area. Unit: μg/m^3.

Model	Period	AQZ						ALL
		NT (26)	CM (6)	CT (11)	YCN (11)	KP (16)	YHD (5)	(75)
AQF	72-hr	9.33	9.26	11.86	11.35	12.66	5.91	10.48
	Day 1	9.37	9.27	12.18	10.66	12.04	5.79	10.29
	Day 2	9.46	9.41	11.79	11.86	13.16	5.82	10.70
	Day 3	9.17	9.10	11.60	11.54	12.80	6.10	10.44
CNN_Cluster8	72-hr	6.15	6.67	8.09	8.69	8.53	4.45	7.24
	Day 1	5.66	6.28	7.29	7.59	7.49	4.17	6.52
	Day 2	6.42	6.91	8.45	9.28	8.96	4.51	7.59
	Day 3	6.38	6.83	8.54	9.22	9.13	4.67	7.61
CNN_Cluster8_WMSE	72-hr	6.71	7.55	9.62	10.24	10.00	4.80	8.30
	Day 1	6.14	7.02	8.17	8.48	8.30	4.33	7.19
	Day 2	7.05	7.82	10.2	10.94	10.77	4.77	8.79
	Day 3	6.95	7.80	10.50	11.29	10.92	5.29	8.91

mean RMSE of AQF falls between 8.79 and 12.33 μg/m^3 and exhibits significant diurnal variation with higher value in the early morning and lower value in the afternoon for each forecasting day. This variation allowed the application of only one weight-sharing convolutional layer to cover the entire 72 hours rather than independently building the convolutional layers for each forecasting day. For the Cluster8 experiment, the mean RMSE gradually increases from 4.93 to 7.12 μg/m^3, during the first 12-hr forecasting time, then remains stable at around 7.5 μg/m^3 (varying from 6.77 to 8.33 μg/m^3) during the next 60-hr forecasting hours. The Cluster8_WMSE experiment has a higher RMSE than the Cluster8 experiment, with the mean RMSE increasing from 5.66 to 7.96 μg/m^3 during the first 12-hr prediction time and falling between 7.44 to 9.83 μg/m^3 during the next 60-hr prediction time.

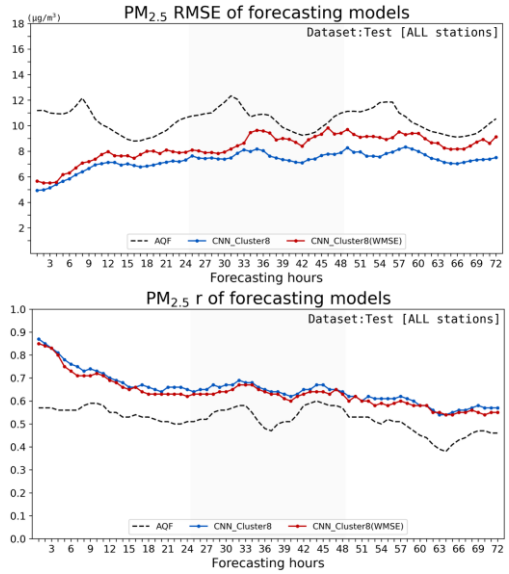


Fig. 4 The comparison of mean RMSE (top) and r (bottom) of all stations during 72-hr forecasting time between the AQF (black), CNN_Cluster8 (blue), and CNN_Cluster8_WMSE (red) models.

For the correlation coefficient (r) between observed and predicted concentration from AQF during 72 forecasting hours, it ranges from about 0.38 to 0.60 and has smaller r in Day 3 (49-72 hr). On the other hand, both CMAQ-CNN models show high correlation during the first 12-hr forecasting time ($r \geq 0.7$) and the average r of all sites exceed 0.54 during the 72-h forecasting time. These findings demonstrate that the CMAQ-CNN model has the ability to predict the variations of air pollutant concentrations up to 72 hrs.

To discuss the impact of the weighted loss function for high $PM_{2.5}$ event day (with daily mean $\geq$ 35 μg/m^3), this study developed two CMAQ-CNN models with different loss

functions. For each forecasting day, we collected the samples that meet the criteria from the test dataset and calculated the average of observed and predicted daily concentration and RMSE in each AQZ (Fig. 5). The YHD area does not have high $PM_{2.5}$ events during the study period. Instead, more than 57% of the high $PM_{2.5}$ events are found in the KP region, with 22% found in the YCN region. Under the influence of the northeasterly monsoonal flow, the YCN and KP AQZs are located in the leeside of the mountain, which exhibits stagnant and stable atmospheric conditions [8]. Moreover, the northerly wind might transport the air pollutants from the NT, CM, and CT areas toward the YCN and KP areas, deteriorating the air quality in southwestern Taiwan.

Overall, the AQF underestimates the $PM_{2.5}$ concentration in all AQZs for each forecasting day, especially in the YCN and KP areas. Both CMAQ-CNN models have worse performance than AQF in the NT area due to the relatively small sample sizes. The CNN_Cluster8 underestimates the

$PM_{2.5}$ concentrations during the high $PM_{2.5}$ events, but has significantly better performance than the AQF in southern Taiwan, with the average RMSE of Day 1 reducing from 13.59 to 8.33 μg/m^3. The overall performance of the high $PM_{2.5}$ events is enhanced with the use of the weighted loss function. The predicted $PM_{2.5}$ concentration from the CNN_Cluster8_WMSE is close to the observation for each of the three forecasting days.

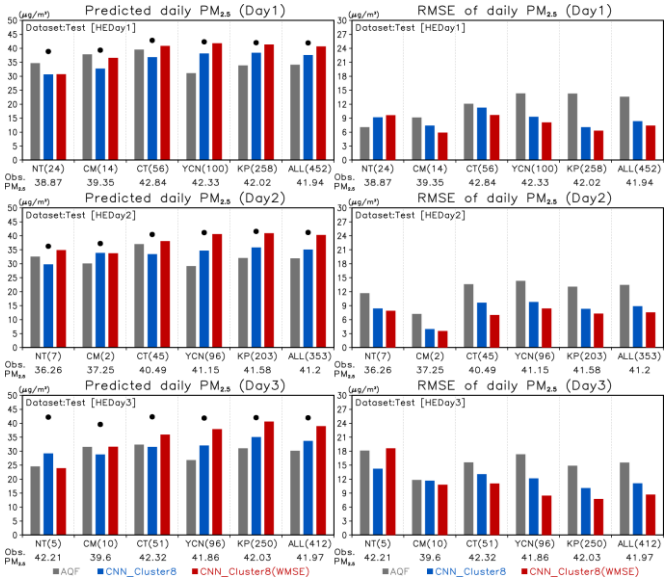


Fig. 5 The observed (black dots) and forecasted (bar) daily $PM_{2.5}$ concentration from the AQF (gray), CNN_Cluster8 (blue), and CNN_Cluster8_WMSE (red) models in each AQZ when Day 1, Day 2, and Day 3 are high pollution events (left panels, from top to bottom row). The right panels display the RMSE in each AQZ and also the average of all samples. The number in the brackets represent the sample size of high pollution events in that area.

V. Conclusions

This study utilized the predicted $PM_{2.5}$ from a numerical AQF system conducted with WRF-CMAQ and other auxiliary features to develop two forecasting models through CNN algorithm (CMAQ-CNN). The indexes of synoptic weather patterns were used in place of the meteorological parameters to decrease the number of input features and to increase the computing efficiency for air quality forecasting operations. To solve the problem of an imbalanced $PM_{2.5}$ concentration value, one of the CMAQ-CNN models (CNN_Cluster8_WMSE) utilized the weighted loss function to put more weight to high-level $PM_{2.5}$ conditions.

Overall, the CMAQ-CNN models perform better than the AQF system. The mean RMSE of all stations for the predicted 72-hr $PM_{2.5}$ from the AQF, Cluster8 experiment and Cluster8_WMSE experiment are 10.48, 7.24, and 8.3 μg/m^3, respectively. For each of the 72 forecasting hours, the RMSE of both CMAQ-CNN models slightly increases during the first 12-hr forecasting time, and the r is larger than 0.7.

According to the comparison of predictions for high pollution events (daily average of $PM_{2.5} \geq 35$ μg/m^3), the AQF underestimated the concentration, while the predicted $PM_{2.5}$ from the Cluster8_WMSE experiment was close to the observation for each of the three forecasting days. The mean RMSE of all sites from the AQF, Cluster8 experiment, and Cluster8_WMSE experiment are 13.59, 8.33, and 7.4 μg/m^3 in Day 1, 13.44, 8.89, and 7.56 μg/m^3 in Day 2, and 16.56, 11.09, and 8.68 μg/m^3 in Day 3, respectively. These results confirm that the weighted loss function has certain capabilities to solve the problem of imbalanced input features.

Acknowledgment

This work was supported by the National Science and Technology Council, Taiwan (contract no: MOST 111-2634-F-008-001-). We also thank the Taiwan EPA and CWB for providing the surface station datasets.

References

[1] Kow, P.-Y., Chang, L.-C., Lin, C.-Y., Chou, C. C.-K., & Chang, F.-J. (2022). Deep neural networks for spatiotemporal $PM_{2.5}$ forecasts based on atmospheric chemical transport model output and monitoring data. *Environmental Pollution*, **306**, 119348.

[2] Sayeed, A., Lops, Y., Choi, Y., Jung, J., & Salman, A. K. (2021). Bias correcting and extending the PM forecast by CMAQ up to 7

days using deep convolutional neural networks. *Atmospheric Environment*, **253**, 118376.

[3] Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, **30**.

[4] Hsu, C.-H., & Cheng, F.-Y. (2019). Synoptic weather patterns and associated air pollution in Taiwan. *Aerosol and Air Quality Research*, **19**(5), 1139-1151.

[5] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., ... & Powers, J. G. (2008). *A description of the advanced research WRF. National Center for Atmospheric Research, Boulder, CO), Version, 3*.

[6] Byun, D., & Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Applied mechanics reviews*, **59**(2), 51-77.

[7] Ching, J., & Byun, D. (1999). Introduction to the Models-3 framework and the Community Multiscale Air Quality model (CMAQ). *Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System*, 20460.

[8] Cheng, F.-Y., Feng, C.-Y., Yang, Z.-M., Hsu, C.-H., Chan, K.-W., Lee, C.-Y., & Chang, S.-C. (2021). Evaluation of real-time $PM_{2.5}$ forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted $PM_{2.5}$ forecasts in Taiwan. *Atmospheric Environment*, **244**, 117909.

[9] Guenther, A., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L., & Wang, X. (2012). The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geoscientific Model Development*, **5**(6), 1471-1492.

[10] Li, M., Zhang, Q., Kurokawa, J.-i., Woo, J.-H., He, K., Lu, Z., Ohara, T., Song, Y., Streets, D. G., & Carmichael, G. R. (2017). MIX: a mosaic Asian anthrouopogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP. *Atmospheric Chemistry and Physics*, **17**(2), 935-963.

[11] TEDS-10 (2019). *Taiwan Emission Data System Version 10*, Environmental Protection Administration, Taipei, Taiwan, Republic of China.

[12] Yarwood, G., Rao, S., Yocke, M., & Whitten, G. (2005). Updates to the carbon bond chemical mechanism: CB05. *Final report to the US EPA, RT-0400675*, **8**, 13.

[13] Appel, K. W., Napelenok, S. L., Foley, K. M., Pye, H. O., Hogrefe, C., Luecken, D. J., ... & Young, J. O. (2017). Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1. *Geoscientific model development*, **10**(4), 1703-1732.

Jointly Modelling Transcriptions and Phonemes with Optimal Features to Detect Dementia from Spontaneous Cantonese

Xiaoquan Ke*, Man-Wai Mak*, and Helen M. Meng†
* The Hong Kong Polytechnic University, Hong Kong SAR
† The Chinese University of Hong Kong, Hong Kong SAR
xiaoquan.ke@connect.polyu.hk, enmwmak@polyu.edu.hk, hmmeng@se.cuhk.edu.hk

***Abstract*—Dementia is a severe cognitive impairment that affects the health of older adults. This paper analyzes diverse speech-based features extracted from spoken languages and selects the most discriminative ones for dementia detection. We propose a two-step feature selection method to handle the circumstance where the feature dimension is far larger than the number of training samples. Recently, the performance of dementia detection has been significantly improved by utilizing Transformer-based models that automatically capture the linguistic properties of spoken languages. We combine features extracted from BERT with selected speech-based features to enhance dementia detection performance. We propose a novel strategy to model the transcriptions and their phonemes using BERT and phoneme-BERT. The proposed method is evaluated on a Cantonese dataset called CU-Marvel, which contains 185 healthy older adults, 98 older adults having minor neurocognitive disorders (minor NCD), and 26 older adults suffering from major NCD. Experimental results show that simultaneously fine-tuning the BERT and phoneme-BERT can leverage information from the recognized phonemes and make the detection performance robust to automatic speech recognition errors. Simultaneous fine-tuning of the BERT and phoneme-BERT models results in a 6% improvement in F_1 scores, compared to fine-tuning the BERT model alone.**

Index Terms: Dementia detection, feature selection, phoneme-BERT

I. Introduction

Dementia is a severe cognitive impairment that may seriously affect the health and daily lives of the afflicted individuals. The most common form of dementia is Alzheimer's Disease (AD). According to a report from the World Health Organization,[1] more than 55 million people live with dementia worldwide, and there are nearly 10 million new cases every year. The disease has a huge impact on the quality of life of not only the patients but also their families and caretakers. Fortunately, with effective detection of early dementia, disease-modifying medications and interventions are possible [1].

A. Related Works

Recently, automatic detection of dementia through speech and language analyses has gathered attention in the research community. Some studies investigated different types of speech-based features for dementia detection. For example, Haider *et al.* [2] compared different types of paralinguistic features – including the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [3], ComParE 2013 [4], Emobase [4], and MRCG [5] – for dementia detection. Because the paralinguistic features are high-dimensional, Pearson's Correlation (PeaCorr) tests were performed to reduce the feature dimensions.

In addition to speech-based features, transcription-based features have also been used for dementia detection [6–8]. For example, Qiao *et al.* [8] combined disfluency and linguistic complexity features with features extracted from Transformer-based models for AD detection. In [8], BERT [9] and ERNIE [10] models were fine-tuned to capture the language characteristics of the AD patients. The Transformer-based models were also extensively investigated by Syed *et al.* [6].

More recently, the fully-automatic assessment of dementia from spontaneous speech has gathered more attention as it does not require labor-intensive annotations or manual transcriptions. In such a case, an automatic speech recognition (ASR) system is utilized to transcribe the patients' speech. However, the erroneous transcriptions produced by the ASR system could lead to high detection errors. To mitigate the errors in ASR systems, the research community has developed three strategies:

- *Adapting ASR systems.* In the Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge [11], the conventional Kaldi-based ASR system was adapted using multiple datasets containing spontaneous speech [12]. Pappagari *et al.* [13] first used a pre-trained ASpIRE model from Kaldi to transcribe target domain data. To improve transcription quality, they interpolated the language model (LM) of ASpIRE with an LM trained on automatic transcriptions of the target domain.
- *Utilizing ASR lattices.* An ASR lattice can provide time alignments, recognized words, and confidence scores for different hypotheses. Usually, we only consider the best hypothesis that has the highest confidence score.

[1] https://www.who.int/news-room/fact-sheets/detail/dementia

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

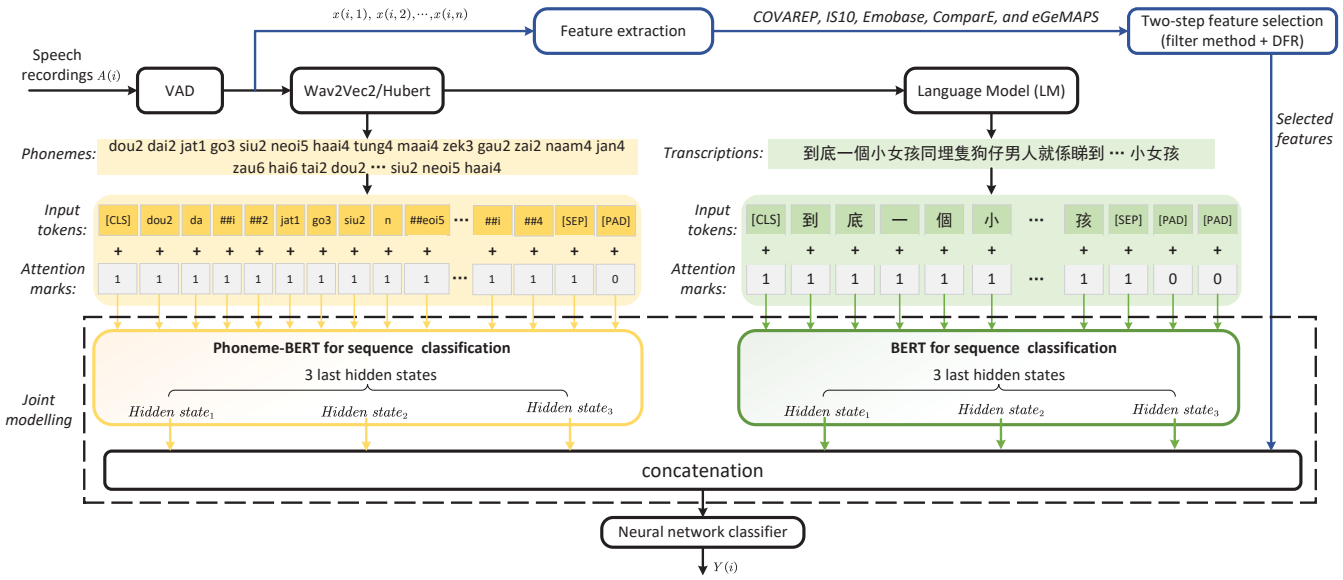


Fig. 1. The architecture of our system for detecting dementia of Cantonese speakers. The recording $A(i)$ is segmented using voice activity detection into n segments. A two-step FS method identifies the most discriminative speech-based features. BERT and phoneme-BERT are jointly fine-tuned on ASR transcriptions and their corresponding phonemes and are combined with the speech-based features to screen dementia patients. We concatenated the last three hidden states of the BERT model, the last three hidden states of the phoneme-BERT model, and the selected features. The neural network classifier was adopted for classification.

However, in addition to the best hypothesis, Pan *et al.* [12] used multiple hypotheses to augment the input of BERT to recognize AD patients. More importantly, they concatenated the confidence scores and hidden states of the BERT model, where the confidence scores were used as a proxy measure of accuracy to inform the classifier about the transcription quality.

- *ASR correction.* To extract more robust linguistic features to distinguish AD patients, the transcribed words with low confidence scores can be removed from the automatic transcriptions [14]. In [12], the automatically transcribed words with the confidence scores lower than 0.87 were replaced by the unknown token <unk>.

B. Modeling Approach

The modeling approach presented in this paper is based on the key insights from the above studies. Importantly, it combines the most discriminative speech-based features and the features extracted from Transformer-based models for fully-automatic dementia screening. To overcome the data sparsity problem, we propose a two-step feature selection (FS) method to identify the most discriminative speech-based features. Additionally, to mitigate the errors in ASR systems, we propose a new strategy to jointly model the ASR transcripts and their corresponding phonemes using the Transformer-based models. The whole architecture is shown in Fig. 1. The proposed system is evaluated on a Cantonese corpus called CU-Marvel. The main contributions of this work are summarized as follows:

1) We propose a two-step FS method to identify the most discriminative speech-based features to screen dementia patients.
2) Extensive experiments are conducted to explore the possibilities of modelling phonemes using phoneme-BERT to detect dementia. Results show that information underlying the ASR phonemes is also indicative of dementia.
3) We jointly model transcriptions and phonemes using BERT and phoneme-BERT. Experimental results show that simultaneously fine-tuning the BERT and phoneme-BERT models can leverage information from the recognized phonemes, making the dementia classifier more robust to ASR errors.

II. Methods

A. Selection of Speech-based Features

While various types of features have been used for dementia detection, it is still unclear which features or their combinations are more effective. We built on key insights from the previous studies and utilized FS methods to find the most effective speech-based features for dementia detection. The extracted speech-based features are as follows.

- *INTERSPEECH 2010 Paralinguistic Challenge Features (IS10).* IS10 [15] is a feature set for emotion recognition and bipolar disorder recognition. In addition to the 32 low-level descriptors (LLDs) in INTERSPEECH 2009 Emotion Challenge, 44 LLDs were added to IS10, including PCM loudness, eight log Mel-frequency bands, eight line-spectral frequency pairs, fundamental frequency (F0) envelope, voicing probability, jitter, and shimmer. Twelve statistics (minimum, maximum, mean, range, etc.) of the LLDs were computed, leading to a 1582-dimensional feature vector per recording. IS10 was adopted as the

baseline feature set for the AD2021 Alzheimer's Dementia Recognition Challenge.[2]

- *COVAREP.* COVAREP [16] provides comprehensive acoustic features, which include prosodic features (F0 and voicing), voice quality features, and spectral features. We extracted COVAREP features at 100Hz; for each recording, the mean, maximum, minimum, median, standard deviation, skew, and kurtosis of the features were computed, leading to a 518-dimensional feature vector per recording. Rohanian *et al.* [17] used the COVAREP features for cognitive impairment detection.
- *eGeMAPS.* The eGeMAPS [3] contains 88 features that are selected based on their potential for characterizing physiological changes in voice production.
- *ComParE 2013.* The ComParE 2013 [4] feature set was adopted as a baseline feature set for AD detection in the ADReSS [18] and ADReSSo [11] challenges, which contains energy, spectral, mel-frequency cepstral coefficients (MFCC), and voicing related LLDs.
- *Emobase.* The Emobase feature set [4] comprises MFCC, F0, F0 envelope, line spectral pairs, etc. Wang *et al.* [19] used the Emobase feature set in multi-modal attention network for AD detection.
- *MRCG.* The MRCG features are multi-resolution cochleagram features, which were used by Haider *et al.* [2] to identify AD patients.

We combined all the feature listed above and adopted a two-step FS approach to selecting dementia features. When the feature dimension is very high, filter methods are indispensable for obtaining a reduced feature set for the expensive FS methods. Therefore, in Step 1, filter methods are utilized to pre-screen the original features. Three filter methods were evaluated in the experiments: Fisher's discriminant ratio (FDR) [20], PeaCorr tests, and mutual information (MutInfo). In Step 2, a deep-learning-based feature ranking method called dual-net feature ranking (DFR) [21] is applied to rank the remaining features. Features with high feature importance were then selected. DFR utilizes a dual-net architecture, where two networks (called operator and selector) are trained to simultaneously perform FS and dementia detection. Specifically, the selector is trained to find multiple subsets of features with minimal cardinality to predict the operator's performance, and the operator uses these feature subsets to minimize classification errors. DFR uses all of the selector's parameters to determine the contribution of individual features to the selector's predictions. DFR shows good performance on a dementia-related Cantonese dataset called JCCOCC-MoCA [22].

B. Automatic Speech Recognition

In order to build a fully-automatic dementia screening system, ASR was used for transcribing speech. Wav2vec 2.0 [23] (denoted as Wav2vec2 from now on) and HuBERT [24] are self-supervised pre-trained models that can be utilized for end-to-end ASR. They can also learn powerful representations from a large amount of unlabeled speech data. By fine-tuning the models on a small amount of transcribed speech, they can achieve similar performance as the traditional fully-supervised ASR systems [23]. As there is no Cantonese pre-trained version of Wav2vec2 or HuBERT, we adopted multilingual and Chinese pre-trained versions of Wav2vec2 and HuBERT from the Transformer Python library, including *Wav2vec2-large-xlsr*,[3] *Wav2vec2-large-Chinese*,[4] and *Hubert-large-Chinese*.[5]

The Cantonese version of Common Voice Speech dataset [25] (common-voice-zh-HK) was used for fine-tuning. The PyCantonese library was utilized to convert the transcriptions to corresponding phonemes.[6] The acoustic models were end-to-end fine-tuned on phone-level using connectionist temporal classification loss. The fine-tuned acoustic models were tested on common-voice-zh-HK test data. The phone error rate for *Wav2vec2-large-xlsr*, *Wav2vec2-large-Chinese*, and *Hubert-large-Chinese* were 0.087, 0.145, and 0.148, respectively. Therefore, *Wav2vec2-large-xlsr* was selected for producing phonemes using the CU-Marvel corpus as inputs. To generate the ASR transcriptions, the phonemes were decoded using a beam search decoder with a 4-gram KenLM trained on common-voice-zh-HK.[7]

C. Modelling with Optimal Features

The BERT models were pre-trained on a large amount of transcriptions using masked language modeling (MLM) as the pre-training objective. As there is no pre-trained version of BERT on Cantonese phonemes, we pre-trained our phoneme-BERT using the common-voice-zh-HK Cantonese phonemes. To form a better representation of the phonemes, we first trained a Word Piece Encoder (WPE) with a vocabulary that consists of 600 sub-word units[8] using the common-voice-zh-HK phonemes. The WPE was shown to work well on the Cantonese phonemes. The trained WPE was then utilized to encode the phonemes into multiple WPE-tokens. In our pre-training settings, the token embeddings of WPE-tokens were randomly initialized and the default configuration of the original BERT was adopted. Following the pre-training regime of BERT, we randomly masked 20% of WPE-tokens using `[MASK]`. Finally, we pre-trained our phoneme-BERT using the masked WPE-tokens as inputs and MLM as the learning objective.

After pre-training our phoneme-BERT, we combined it with a well pre-trained BERT model[9] from the Transformer library. We followed the strategy in [12] and concatenated the last three hidden states of the BERT model with the last three hidden states of the phoneme-BERT model, see

[2]https://github.com/THUsatlab/AD2021

[3]https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[4]https://huggingface.co/TencentGameMate/chinese-wav2vec2-large

[5]https://huggingface.co/TencentGameMate/chinese-hubert-large

[6]https://pycantonese.org/

[7]Decoding parameters: language model's weight $= 3.0$, word score $= 0.0$, number of best hypotheses $= 1$, and beam size $= 500$.

[8]We set the size of the vocabulary to 600 according to training procedure of a byte pair encoder in [26].

[9]https://huggingface.co/bert-base-chinese

TABLE I
CHARACTERISTICS OF THE CU-MARVEL DATASET.

Variable	HCs ($n = 185$)	Dementia ($n = 124$)
No. of Female	122	70
No. of Male	63	54
Age (years)	70.0 (66.0, 75.0)	78.0 (68.8, 83.0)
Education (years)	7.0 (6.0, 11.0)	6.0 (2.0, 9.25)
MoCA scores	23.0 (21.0, 26.0)	19.0 (15.0, 21.0)
Language	Cantonese	
Task	Rabbit-story picture description	
Manual transcriptions	No	

HCs: healthy older adults; MoCA: Montreal Cognitive Assessment.
The values are presented as *median (interquartile range)*.

Fig. 1. To utilize information from the speech, we combined the concatenated hidden states with the selected speech-based features. Note that the selected speech-based features were normalized to make them compatible with the characteristics of the hidden states. A fully-connected neural network (FCNN) classifier was built on top of the concatenated hidden states and the selected features. Finally, the phoneme-BERT and BERT were end-to-end fine-tuned on the ASR transcriptions and their corresponding phonemes.

III. DATASETS AND EXPERIMENTAL SETTINGS

A. The CU-Marvel Cantonese Corpus

Cantonese is one of the major Chinese dialects that has over 80 million native speakers in Southern China. The CU-Marvel corpus was collected for the research on the screening and monitoring of neurocognitive disorders based on spoken language technologies. A series of cognitive tests, including Montreal Cognitive Assessment (MoCA) tests and picture description tests, were given to each participant for assessing the mild cognitive impairment and dementia in older adults. According to the assessment results, 309 participants were divided into three groups: 185 healthy older adults (HCs), 98 older adults having minor neurocognitive disorders (minor NCD), and 26 older adults suffering from major NCD. For detecting dementia, we combined minor NCD and major NCD into one category called "possible dementia". A rabbit story picture description task was selected for the experiments. Table I shows the corpus's characteristics.

B. Implementation Details

The phoneme-BERT was pre-trained on the common-voice-zh-HK phonemes with a batch size of 512 for 3,000 steps. We observed an increase in the validation loss when the number of steps exceeded 3,000. 10-fold cross-validation (CV) was applied to the CU-Marvel dataset to determine the performance of dementia detection. For each fold, we fine-tuned the models with a batch size of 4 for 10 epochs. The maximum length for word tokens and phoneme tokens was set to 512. This setting aims to cover as many words and phonemes as possible because most participants spoke a lot.

TABLE II
DETECTION PERFORMANCE (F_1 SCORES) ON THE CU-MARVEL. THE NUMBERS IN THE BRACKETS ARE FEATURE DIMENSIONS.

Feature set	SVM	DT	KNN	LDA	FCNN
IS10 (1582)	0.521	0.559	0.528	0.517	0.590
COVAREP (518)	0.567	0.552	0.564	0.562	0.574
eGeMAPS (88)	0.571	0.582	0.525	0.539	0.614
ComparE (6373)	0.556	0.526	0.555	0.569	0.603
Emobase (988)	0.497	0.540	0.521	0.503	0.630
MRCG (768)	0.529	0.533	0.543	0.495	0.605
Mean	0.540	0.549	0.539	0.531	***0.603***

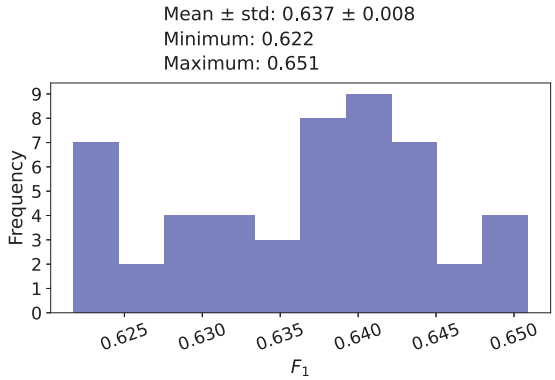


Fig. 2. Variation in detection performance across the repeated CVs.

IV. RESULTS AND DISCUSSIONS

A. Performance of Different Feature Types

We first evaluated the recognition performance of all the feature sets *before* FS. Five classifiers were adopted, including k-nearest neighbor (k-NN) classifier ($k = 1$), linear support vector machines (SVM, box constraint $= 0.1$), decision trees (DT, leaf size $= 20$), linear discriminant analysis (LDA), and FCNN classifier.[10] We ran 50 repetitions of the 10-fold CV based on different data splittings and averaged the performance (F_1 scores). The corresponding results are reported in Table II. Table II shows that the FCNN classifier achieves the best classification performance among all classifiers, that is, it achieves the highest averaged F_1 scores. Therefore, subsequent experiments adopted the FCNN classifier.

We then combined all of the features to form the *combined features*, which are 10317-dimensional vectors. When conducting 10-fold CV on the combined features, large variation in detection performance across the repeated CVs was observed, as illustrated in Fig. 2. The variation is caused by applying random splitting on a small dataset for each CV, inducing significantly different training partitions (TR) for each run. To reduce the variation, we propose using an ensemble procedure to stabilize the detection performance. Specifically, we ran 10 repetitions of CV based on different data splittings.

[10]Except for the FCNN classifier, all other classifiers were adopted from [2]. FCNN classifier settings: network architecture 'feature_dimension–128–32–2", batch size $= 4$, and epochs $= 10$.

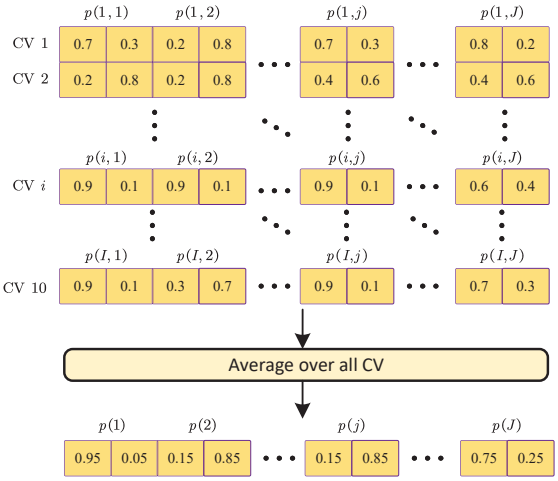


Fig. 3. The ensemble procedure to stabilize the classification performance estimated by CV. We ran I repetitions of CV based on different data splittings and averaged the predicted scores $p(i,j)$ over all the CV for each of the J subjects.

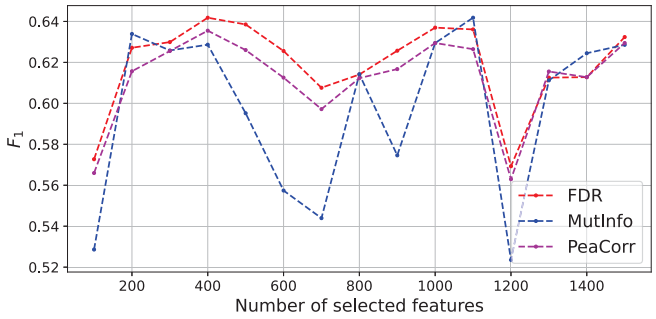


Fig. 4. Classification performance of the filter methods.

We then produced the predicted scores $p(i,j)$ for subject j in CV i, as shown in Fig. 3. Finally, we averaged the predicted scores $p(j) = (1/10)\sum_{i=1}^{10} p(i,j)$ over all the 10 CV for each of the J subjects. After utilizing the ensemble procedure, we obtained 0.670 accuracy (ACC) and 0.645 F_1 scores, which boosts the minimum F_1 scores from 0.622 (Fig. 2) to 0.645. Therefore, the subsequent experiments adopted the ensemble procedure to stabilize classification performance.

B. Performance of Two-step FS

We followed the procedure described in Section II-A to evaluate the performance of the filter methods (FDR, PeaCorr, and MutInfo) on the combined features (10317-dimensional vectors). On the TR of individual folds, we applied the filter methods to reduce the feature dimension to $n = \{100, 200, 300, 400, \ldots, 1500\}$, as shown in Fig. 4. It shows that FDR achieves the highest F_1 scores (0.645) when the feature dimension was reduced to 400. We then applied DFR on the remaining 400 features to further select more discriminative features, as shown in Table III. It shows that when we reduced the feature dimension from 400 to 300, we further improved the ACC. This two-step procedure

TABLE III
CLASSIFICATION PERFORMANCE OF TWO-STEP FS. *ACC*: ACCURACY; *PRE*: PRECISION; *REC*: RECALL.

	Dimension	10-fold CV ACC	PRE	REC	F_1
FDR + DFR (Step 1 + Step 2)	150	0.650	0.634	0.631	0.632
	200	0.663	0.648	0.626	0.627
	250	0.628	0.623	0.628	0.622
	300	***0.680***	***0.668***	***0.643***	***0.645***
	350	0.634	0.622	0.624	0.623
	400	0.663	0.648	***0.643***	***0.645***

significantly reduces the feature dimension and outperforms the combined features.

C. Performance of Fine-tuning

We evaluated the performance on fine-tuning BERT, phoneme-BERT, and the combination of BERT and phoneme-BEET, as shown in Table IV (Row 2 to Row 4). Row 3 shows that fine-tuning phoneme-BERT on ASR phonemes is beneficial to dementia detection. This indicates that some abnormal information (e.g., phoneme disorder, phoneme repetition, phoneme scarcity, and phoneme deficiency) in the patients' ASR phonemes is also indicative of dementia. Row 2 and Row 3 show that fine-tuning phoneme-BERT is better than fine-tuning BERT. This maybe due to the insufficient training of the LM. If we adapt the LM with an elderly corpus containing similar picture description tasks, we may reduce the ASR errors and improve detection performance.

Row 4 shows that it is better to fine-tune the phoneme-BERT and BERT models jointly instead of fine-tuning the BERT model only, indicating that the recognized phonemes contain useful information for dementia detection. Our ASR system produces phonemes, which are subsequently decoded into words using a 4-gram KenLM. In Cantonese, a single set of phonemes may correspond to multiple words, i.e., homophones. Despite the ASR system producing the correct phonemes, in some instances, these phonemes are misinterpreted, resulting in erroneous words. Though the ASR system decodes some erroneous words, simultaneous fine-tuning of BERT and phoneme-BERT enables the system to leverage the corresponding correct phonemes, enhancing its overall robustness to ASR errors.

D. Fine-tuning with Optimal Features

Finally, we fine-tuned BERT, phoneme-BERT and the combination of phoneme-BERT and BERT with the optimal features selected in Section IV-B. Table IV (Row 5 to Row 7) shows that fine-tuning with the optimal features can improve performance, no matter on which model. This strategy leverages the optimal features to enhance the inputs to the classifier. Row 7 shows that with the optimal feature set, simultaneously fine-tuning the phoneme-BERT and BERT models achieves the best detection performance.

TABLE IV
CLASSIFICATION PERFORMANCE ACHIEVED BY FINE-TUNING DIFFERENT MODELS. *ACC*: ACCURACY; *PRE*: PRECISION; *REC*: RECALL.

Row	FDR + DFR	Fine-tuning		10-fold CV			
		BERT	Phoneme-BERT	ACC	PRE	REC	F_1
1	✔	✘	✘	0.680	0.668	0.643	0.645
2	✘	✔	✘	0.608	0.591	0.591	0.591
3	✘	✘	✔	0.650	0.635	0.634	0.634
4	✘	✔	✔	0.667	0.654	0.655	0.654
5	✔	✔	✘	0.693	0.680	0.666	0.669
6	✔	✘	✔	0.683	0.670	0.671	0.671
7	✔	✔	✔	***0.696***	***0.685***	***0.687***	***0.686***

V. CONCLUSIONS

We proposed a two-step feature selection method to identify the most effective speech-based features to screen Cantonese-speaking dementia patients. Fine-tuning Transformer-based models with the selected speech-based features improved performance further. We also fine-tuned BERT and phoneme-BERT on transcriptions and their corresponding phonemes, which made the dementia classifier more robust to ASR errors. Fine-tuning both the BERT and phoneme-BERT models together resulted in a 6% improvement in F_1 scores, compared to fine-tuning the BERT model alone.

VI. ACKNOWLEDGMENTS

This work was in part supported by Research Grants Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N).

REFERENCES

[1] J. L. Cummings, R. Doody, and C. Clark, "Disease-modifying therapies for Alzheimer disease: Challenges to early intervention," *Neurology*, vol. 69, no. 16, pp. 1622–1634, Oct. 2007.

[2] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 272–281, Feb. 2020.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia Int. Conf.*, Oct. 2010, pp. 1459–1462.

[5] F. Haider and S. Luz, "Attitude recognition using multi-resolution Cochleagram features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2019, pp. 3737–3741.

[6] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Tackling the ADRESSO challenge 2021: The MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech*, Aug. 2021, pp. 3815–3819.

[7] J. Li, J. Yu, Z. Ye, S. Wong, M. W. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6423–6427.

[8] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, "Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models," in *Proc. Interspeech*, Aug. 2021, pp. 3805–3809.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805

[10] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 8968–8975.

[11] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," in *Proc. Interspeech*, Aug. 2021, pp. 4211–4215.

[12] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based Alzheimer's dementia detection through spontaneous speech." in *Proc. Interspeech*, Aug. 2021, pp. 3810–3814.

[13] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios," in *Proc. Interspeech*, Aug. 2021, pp. 3825–3829.

[14] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *Proc. Interspeech*, Oct. 2020, pp. 159–164.

[15] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, Sep. 2010, pp. 2794–2797.

[16] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.

[17] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," 2021, *arXiv:2106.09668*. [Online]. Available: https://doi.org/10.48550/arXiv.2106.09668

[18] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," 2020, *arXiv:2004.06833*. [Online]. Available: https://arxiv.org/abs/2004.06833

[19] N. Wang, Y. Cao, S. Hao, Z. Shao, and K. Subbalakshmi, "Modular multi-modal attention network for Alzheimer' disease detection using patient audio and language data," in *Proc. Interspeech*, Aug. 2021, pp. 3835–3839.

[20] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8696–8702, Jul. 2011.

[21] X. KE, M.-W. Mak, and H. M. Meng, "Automatic selection of discriminative features for dementia detection in Cantonese-speaking people," in *Interspeech 2022*, Sep. 2022, pp. 2153–2157.

[22] S. S. Xu, M. W. Mak, K. H. Wong, H. Meng, and C. Y. Kwok, "Speaker turn aware similarity scoring for diarization of speech-based cognitive assessments," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC*. IEEE, 2021, pp. 1299–1304.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. neural inf. proces. syst. (NIPS)*, Dec. 2020, pp. 12 449–12 460.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," 2021, *arXiv:2106.07447*. [Online]. Available: https://doi.org/10.48550/arXiv.2106.07447

[25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2020, *arXiv:1912.06670*. [Online]. Available: https://doi.org/10.48550/arXiv.1912.06670

[26] M. N. Sundararaman, A. Kumar, and J. Vepa, "PhonemeBERT: Joint language modelling of phoneme sequence and ASR transcript," in *Proc. Interspeech*, Aug. 2021, pp. 3236–3240.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Combining multiple end-to-end speech recognition models based on density ratio approach

Keigo Hojo*, Daiki Mori*, Yukoh Wakabayashi*, Kengo Ohta†, Atsunori Ogawa‡, and Norihide Kitaoka*
* Toyohashi University of Technology, Japan
† National Insitute of Technology, Anan College, Japan
‡ Nippon Telegraph and Telephone Corporation, Japan

***Abstract*—Automatic speech recognition (ASR) systems which involve the use of multiple ASR models, known as "system combination" approaches, have been confirmed by many studies to be effective for improving speech recognition performance. In this study, we propose a system combination method using multiple end-to-end (E2E) ASR models trained under different conditions (i.e., using data from different source domains), in order to perform ASR on an unknown target domain without the need for additional, expensive model training. In order to efficiently adapt existing ASR models to the target domain, we exploit a density ratio approach (DRA), and experimentally demonstrate that by combining ASRs trained with multiple types of acoustic information with an external language model for the target domain, i.e., the standard DRA, our method can more robustly perform ASR for an unknown target domain than a single ASR model trained under one domain condition.**

I. Introduction

Combining the outputs of multiple automatic speech recognition (ASR) systems has been shown to be an effective technique for improving ASR performance. For example, the NIST Recognizer Output Voting Error Reduction (ROVER) method integrates the output of multiple Large Vocabulary Continuous Speech Recognition (LVCSR) systems using a voting module [1], [2]. Other methods integrate the outputs of multiple LVCSR systems in a machine-learning manner, e.g., using a Support Vector Machine [3], [4]. These multiple model ASR systems have outperformed single model ASR approaches, demonstrating the effectiveness of this approach. Robust acoustic models have also been created by integrating large-scale acoustic models, acoustic models trained with speech recorded in noisy environments, and acoustic models trained with data containing different styles of speech, in a Gaussian mixture model (GMM) space [5]. Researchers have also proposed integrating a hybrid ASR model and an end-to-end (E2E) ASR model at the hypothesis level to improve recognition performance [6], while other investigators have integrated ASR models with completely different model architectures, such a hybrid ASR model and a GMM-HMM ASR model [7].

In contrast, only a few studies have explored combining multiple E2E ASR models [8]. This is because it is generally believed that the vast training parameters of a single E2E ASR model allow a sufficient degree of freedom to process speech data from a variety of domains, including speech data containing noise and multiple styles of speech, and thus there is no need to combine multiple ASR models. While the recognition performance of large-scale E2E ASR models has improved over the years, researchers have also explored the use of E2E ASR models in combination with external language models. Shallow Fusion (SF) is a method frequently used to integrate an ASR model and an external language model [9]–[12], allowing easy integration without the need for retraining, unlike other integration methods such as Deep Fusion [13], Cold Fusion [14], and Component Fusion [15]. The Density Ratio Approach (DRA) [16], [17], which is a stochastic improvement of SF, is a more recently proposed integration method which considers the "implicit language information" contained inside the trained ASR model. It has been reported that the language information contained in English [16], [17], Spanish, Italian [18], and Japanese [19] ASR models can be approximated by external language models using DRA, which separates the acoustic and language information acquired by the ASR model.

Therefore, in this study, we combine the acoustic information contained within multiple existing ASR models with language information from an external language model for the target domain using DRA, under the assumption that there may be situations in which it is not feasible to create a large E2E ASR model for a particular target domain. By combining acoustic information from multiple existing E2E ASR models with language information from the target domain language model, we thought it might be possible to create an ASR model that is acoustically robust for the target domain, improving speech recognition performance. Therefore, our proposed method extracts multi-domain acoustic information from multiple E2E ASR models and combines it with knowledge about the target domain obtained from the external language model.

Multiple E2E ASR and language models are combined also in [8]. The authors conduct their experiments in one domain scenario. In contrast, we conduct our experiments in a domain adaptation scenario.

II. Language model fusion methods

A. Shallow fusion

The standard method used to integrate information from an ASR model and an external language model is called shallow fusion (SF). The output probabilities of the ASR model

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

are added with those of the external language model in the logarithmic domain follows:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\{\log P(y|x) + \lambda \log Q(y)\}, \tag{1}$$

where $P(y|x)$ is the output probability of the ASR model, i.e., the probability of inferring symbol label y when acoustic feature x is given, and where $Q(y)$ is the prior output probability of the external language model. When using the SF method, the external language model is only used during inference. Note also that the language and ASR models are trained independently.

B. Density ratio approach (DRA)

More recently, research have attempted to estimate the "implicit language information" contained with ASR models to facilitate integration with external language models [20], [21]. The original DRA is presented below.

The density ratio approach (DRA) is one of the method s proposed. The original DRA method is similar to SF, but DRA also takes this implicit language information into account during inference. Conventional ASR models attempt to infer sequence $\hat{y}$ as follows:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\{\log P_{\mathrm{src}}(y|x)\}, \tag{2}$$

where $P_{\mathrm{src}}(y|x))$ is the probability of output sequence y obtained from the ASR model for the source domain when given input sequence x. Here, "source domain" refers to the task being performed when the speech data used for model training was recorded, e.g., dialog speech, presentation speech, etc. Output probability $P_{\mathrm{src}}(y|x)$ from the ASR model can be expanded using Bayes' rule as follows:

$$\log P_{\mathrm{src}}(y|x) = \log P_{\mathrm{src}}(x|y) + \log P_{\mathrm{src}}(y) - \log P_{\mathrm{src}}(x), \tag{3}$$

where $P_{\mathrm{src}}(x)$ is a term commonly calculated for all y and does not affect the relative relation of the probability. Therefore (3) can be expressed as follows:

$$\log P_{\mathrm{src}}(y|x) = \log P_{\mathrm{src}}(x|y) + \log P_{\mathrm{src}}(y) + \mathrm{Const}. \tag{4}$$

On the right side of (4), we can see that the ASR model includes the acoustic information term $P_{\mathrm{src}}(x|y)$ and the language information term $P_{\mathrm{src}}(y)$, the letter of which represents the "implicit language information" refereed to above, i.e., the statistics of language obtained from the speech data used for training the ASR model. This "implicit language information" is considered to be approximately the same as the language information that would be possessed by a language model trained in the same domain, i.e., $P_{\mathrm{src}}(y) \approx Q_{\mathrm{src}}(y)$.

ASR models take advantage of implicit language information to improve decoding when the test data domain is the same as the training domain. But if the test domain differs from the training domain, this implicit language information can degrade the ASR model's decoding performance. When using DRA, this implicit language information can be removed by subtracting the output probability of the language model from the output probability of the ASR model for the source domain, and then adding the output probability of the target domain language model as follows:

$$\begin{aligned}\log P_{\mathrm{src}}(y|x) &- \lambda_{\mathrm{sub}} \log Q_{\mathrm{src}}(y) + \lambda_{\mathrm{add}} \log Q_{\mathrm{tgt}}(y) \\ &\approx \log P_{\mathrm{src}}(x|y) + \lambda_{\mathrm{add}} \log Q_{\mathrm{tgt}}(y) + \mathrm{Const} \\ &\propto \log P_{\{\mathrm{src,tgt}\}}(y|x),\end{aligned} \tag{5}$$

where $Q_{\mathrm{tgt}}(y)$ is the probability of the external language model for the target domain, and λ_{add} is an addition weight. The DRA assumes that the source domain acoustic information $P_{\mathrm{src}}(x|y)$ is roughly equal to target domain acoustic information $P_{\mathrm{tgt}}(x|y)$. Therefore, $P_{\{\mathrm{src,tgt}\}}(y|x)$ in (5) indicates that the acoustic information is the source domain and the language information is the target domain. It can also be said that $P_{\{\mathrm{src,tgt}\}}(y|x)$ is the ASR model, where the acoustic information is the source domain and the language information is the target domain. By using (2) to (5), we can successfully replace only the ASR model's language information allowing the create of an ASR model for any target domain simply by preparing text data for that domain. Furthermore, DRA integrates the external language model and ASR model only at the time of inference, as in SF.

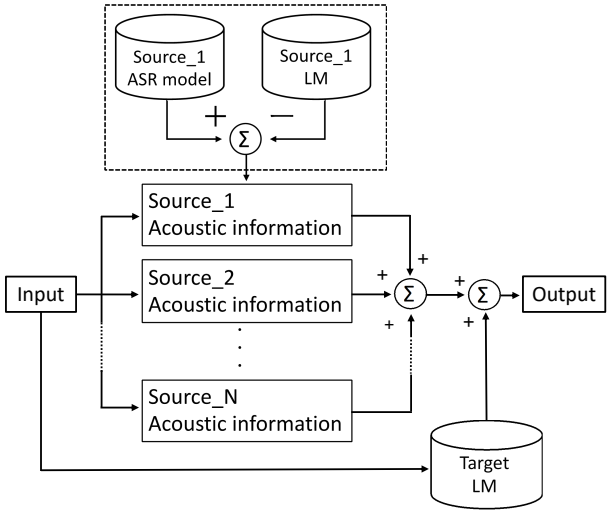


Fig. 1. Overview of the proposed method.

III. Proposed method

In this study, we assumed that only pure acoustic information remains in the E2E ASR model when the implicit language information is removed, and that it is possible to construct a ASR model for the target domain by combining acoustic information from multiple source domains with language information from the target domain. By integrating the acoustic information extracted from multiple ASR models, it should be possible to enhance the acoustic information even when the ASR models are trained with only a small amount of data.

We first extract acoustic information from multiple E2E ASR models at each decoding step, using the process shown

in (6):

$$\log \tilde{P}_{\mathrm{src}}(x|y) \approx \sum_{i=1}^{N} \{\log P_{\mathrm{src_}i}(y|x) - \lambda_{\mathrm{sub_}i} \log Q_{\mathrm{src_}i}(y)\}, \tag{6}$$

where N is the number of E2E ASR models used, $P_{\mathrm{src_}i}(y|x)$ is the i-th source domain ASR model, and $Q_{\mathrm{src_}i}(y)$ is the i-th source domain language model. The acoustic information for each domain is extracted using the external language model's language information for each domain, which is then subtracted from the E2E ASR models for those domains. We then add extracted acoustic information $\tilde{P}_{\mathrm{src}}(x|y)$ and external language model information for the target domain $Q_{\mathrm{tgt}}(y)$, as shown in (7) below, confirming in theory that a target domain ASR model can be created with acoustic information obtained from multiple E2E ASR models and language information from an external language model trained with a large amount of target domain text data.

Figure 1 provides an overview of our proposed method. We integrated the acoustic information from each E2E ASR model with the language information from the external language model during the beam search step as follows:

$$\hat{y} = \underset{y}{\mathrm{argmax}} \{\log \tilde{P}_{\mathrm{src}}(x|y) + \lambda_{\mathrm{add}} \log Q_{\mathrm{tgt}}(y)\}. \tag{7}$$

We expect that our proposed method will be more acoustically robust due to the use of acoustic information from multiple, existing E2E ASR models, potentially improving ASR performance while reducing development cost. Since our method does not require re-training, such as fine-tuning the existing ASR model, this type of ASR system can be easily developed without incurring the cost of dataset construction (e.g., pairing speech data with its transcription). We only need to prepare an external language model trained with text data for the target domain.

IV. Experiment

A. Setup

We evaluated the performance of the proposed method experimentally using an ASR task, and compared its recognition accuracy with results when using E2E ASR models with SF, DRA, and an without external language model. The speech recognition accuracy for each method was evaluated using character error rates (CER) for the test set.

We used data from the Corpus of Spontaneous Japanese (CSJ) dataset [22], the Mainichi Shimbun (MS) dataset, the Japanese Newspaper Article Sentences (JNAS) dataset [23], and Corpus of Everyday Japanese Conversation (CEJC) [24] dataset in our experiments. Table I shows the details of the datasets used in the experiment. Two domains were used from the CSJ dataset: Academic Presentation Speech (APS), which includes presentations on the subjects of engineering, humanities, and sociology, and Simulated Public Speech (SPS), which did not contain academic terminology. The APS and SPS domain consist of 214 and 257 hours of speech data, respectively. These speech data was randomly divided into a training set (Train-set) and a development set (Dev-set), representing the training and development data used when training the E2E ASR models. The Dev-set was used to monitor learning, and early stopping was applied to prevent overfitting with the patience value set to 3. We expanded the training datasets using speaking speed perturbation, adding voice data converted to 0.9 and 1.1 times the speed of the original voice data. The MS dataset is a text dataset consisting of 58 million characters from 11 years of Japanese newspaper articles (1991 to 2002), and was only used for training the external language models. The JNAS dataset was created by having speakers read excerpts from daily newspaper articles aloud, thus the JNAS speech data is from the same domain as the MS text data. The JNAS corpus was used for the Dev-set and Test-set. The Dev-set was used when selecting hyperparameters for the model integration methods, while the Test-set was used to evaluate the proposed method and the other integration methods. The CEJC dataset contains natural daily conversations in various situations by speakers of a wide range of ages. This data was divided into a Train-set, Dev-set, and Test-set. Train-set was used for training the external language model, and Dev-set and Test-set were used for the same purposes as JNAS.

TABLE I
Details of the data sets used in the experiment.

Domain	Split	Speakers	Utterances	Characters	Duration
APS (CSJ)	Train	889	139,396	5,041,328	206h
	Dev	49	5,272	195,416	8h
SPS (CSJ)	Train	1,555	221,994	5,503,402	245h
	Dev	80	8,612	258,651	12h
MS	Train	—	—	64,864,516	—
JNAS	Dev	23	500	230,821	0.9h
	Test	23	500	238,717	0.9h
CEJC	Train	—	—	2,792,530	—
	Dev	279	1,161	30,567	1h
	Test	279	1,161	30,396	1h

As our E2E ASR models, we used both the hybrid CTC/attention architecture [25] and the joint CTC/Attention Transformer [26] provided by ESPnet (version: v.202205) [27], [28]. We prepared three versions of the E2E ASR models: E2E ASR models trained with the CSJ's APS domain, E2E ASR models trained with CSJ's SPS domain, and E2E ASR models trained with both the APS and SPS domain of the CSJ. We also prepared five external language models, which were trained with APS transcription data, SPS transcription data, both APS and SPS transcription data, MS text data, and CEJC transcription data, respectively. The vocabulary size of dictionary used for training the ASR and language models was set to 3,262, which is the total number of characters in the CSJ.

Based on the theoretical foundation established in (6) and (7), we first extracted the acoustic information for the APS and SPS speech data from the APS E2E ASR and SPS E2E ASR models, respectively, using the method shown in (8). The extracted acoustic information was then integrated with the MS

trained external language model, as shown in (9):

$$\log P_{\{\mathrm{APS,SPS}\}}(x|y) \approx \log P_{\mathrm{APS}}(y|x) - \lambda_{\mathrm{sub_1}} \log Q_{\mathrm{APS}}(y) + \log P_{\mathrm{SPS}}(y|x) - \lambda_{\mathrm{sub_2}} \log Q_{\mathrm{SPS}}(y), \quad (8)$$

$$\hat{y} = \underset{y}{\mathrm{argmax}} \{\log P_{\{\mathrm{APS,SPS}\}}(x|y) + \lambda_{\mathrm{add}} \log Q_{\mathrm{MS}}(y)\}, \quad (9)$$

where $\hat{y}$ represents the most probable character sequence obtained using a beam search with the beam size set to 20. $\lambda_{\mathrm{sub_1}}$ and $\lambda_{\mathrm{sub_2}}$ in (8) correspond to the subtraction weights of the APS and SPS language models, respectively. λ_{add} in (9) corresponds to the addition weight of the MS language model. These parameters were selected from grid search with values from 0.1 to 1.1 by 0.2 increments, and the most optimal parameters for the Dev-set were used. The weights of the language models used with the SF, DRA, and proposed methods were tuned by grid search. In addition, we conducted an experiment using CEJC as another target domain, and this is formulated as follows:

$$\hat{y} = \underset{y}{\mathrm{argmax}} \{\log P_{\{\mathrm{APS,SPS}\}}(x|y) + \lambda_{\mathrm{add}} \log Q_{\mathrm{CEJC}}(y)\}. \quad (10)$$

B. Training

Following ESPnet's CSJ recipe, the ASR and language models were trained under the following conditions.

1) CTC/Attention model based BLSTM: The hybrid CTC/Attention architecture is an RNN-based speech recognition model that utilizes the advantages of both Connectionist Temporal Classification (CTC) and an Attention mechanism during training and inference for E2E ASR. The acoustic features of the input speech for the ASR models were processed using Mean and Variance Normalization (MVN) with an 80 mel-scale filter bank. The ASR model consisted of an encoder with VGGnet and four Bidirectional Long Short Term Memory (BLSTM) layers, and an Attention-based decoder with one Long Short Term Memory (LSTM) and a linear layer. A CTC module-based decoder was also used as an auxiliary, in which case the weight of the CTC decoder was set to 0.3. We trained the entire ASR model end-to-end using the Ada-delta optimizer with a batch size of 24.

2) CTC/Attention model based Transformer: The joint CTC/Attention Transformer is a Transformer-based ASR model that combines CTC and Attention, as in the hybrid CTC/Attention architecture described above. The model first performs SpecAugment (time contraction, time masking, and frequency masking) [29] and MVN when acoustic features are input. The model consisted of 18 Transformer encoders, 6 Transformer decoders and a CTC module-based decoder with the weight set to 0.3. The number of Multi-Head Attention heads used by each encoder and decoder was set to 8. We then trained the entire model end-to-end using Adam.

TABLE II
PERPLEXITIES FOR EACH LANGUAGE MODEL AND TEST SET.

Language Model	Perplexity for Test Set				
	APS	SPS	APS+SPS	JNAS	CEJC
APS	17.33	33.37	24.23	147.66	43.07
SPS	34.79	17.38	24.39	100.64	32.07
APS+SPS	15.78	16.77	16.28	80.53	29.06
MS	95.82	56.58	73.18	33.43	50.78
CEJC	99.15	49.51	69.50	208.29	23.80

3) RNN language model: The character-based RNN language model consisted of an embedding layer, two LSTM layers, and a linear layer. The input to the embedding layer was a one-hot vector of characters. The batch size was 256 and stochastic gradient descent (SGD) was used as the optimization function.

C. Results

1) Perplexity: To compare the differences in the language information of each domain, we calculated the perplexities of the external language models for each domain, which are shown in Table II. The APS and SPS test sets consist of 117,252 and 121,465 characters, respectively, and the APS + SPS test set is simply a combination of these two sets of text data. The perplexities shown in Table II confirm that the language information of each domain is different. In particular, we can confirm that there is a large difference between the language information in the MS and CEJC domains and that in the APS and SPS domains.

2) CERs with CTC/Attention model based BLSTM: Using the ASR and language models obtained after training in each domain, we compared and evaluated the performance of the proposed method with that of the conventional integration methods.

First, the results of our comparative experiment when using MS as the target domain are shown in Table III. where, we used RNN-based models as the ASR model. These results reveal that the proposed method outperformed the single ASR model methods (ASR models without an external language model, ASR models using SF or DRA). Results when using the APS+SPS E2E ASR models, which are relatively large-scale E2E ASR models trained with both the APS and SPS datasets, were CERs of 18.9% without the external language model, 16.6% with SF, 15.9% with DRA. The CER when using the proposed method was 16.3%, which was slightly higher than the CER when using the DRA, however our proposed method achieved comparable performance without the need to train the ASR model with the APS and SPS speech data. The relative difference between the CER of the proposed model and the collective CER of the large-scale E2E ASR model was only 2.45%. Despite its higher CER, we believe the proposed method is very effective given the fact that it can integrate multiple existing E2E ASR models without re-training. On the other hand, when comparing the CERs of the proposed method with those of the integrated methods using two domains (APS and SPS) for acoustic information, the

TABLE III
CERs (%) FOR EACH METHOD WHEN USING RNN-BASED ASR MODELS WITH JNAS AS THE TARGET DOMAIN.

ASR Model	N	Integration method			
		w/o LM	SF (λ_{add})	DRA (λ_{sub} / λ_{add})	Proposed ($\lambda_{\mathrm{sub_1}}$, $\lambda_{\mathrm{sub_2}}$ / λ_{add})
APS	1	26.3	23.9 (0.5)	22.4 (0.3 / 0.3)	—
SPS	1	22.0	21.7 (0.5)	19.9 (0.3 / 0.3)	—
APS+SPS	1	18.9	16.6 (0.5)	**15.9** (0.3 / 0.5)	—
APS, SPS	2	19.4	17.9 (0.7)	—	**16.3** (0.5, 0.3 / 0.9)

TABLE IV
CERs (%) FOR EACH METHOD WHEN USING RNN-BASED ASR MODELS WITH CEJC AS THE TARGET DOMAIN.

ASR Model	N	Integration method			
		w/o LM	SF (λ_{add})	DRA (λ_{sub} / λ_{add})	Proposed ($\lambda_{\mathrm{sub_1}}$, $\lambda_{\mathrm{sub_2}}$ / λ_{add})
APS	1	42.7	42.1 (0.3)	41.1 (0.3 / 0.5)	—
SPS	1	38.9	38.4 (0.1)	38.3 (0.1 / 0.3)	—
APS+SPS	1	36.6	36.2 (0.3)	**36.2** (0.1 / 0.3)	—
APS, SPS	2	35.4	35.3 (0.1)	—	**35.1** (0.5, 0.1 / 0.7)

TABLE V
CERs (%) FOR EACH METHOD WHEN USING TRANSFORMER-BASED ASR MODELS WITH JNAS AS THE TARGET DOMAIN.

ASR Model	N	Integration method			
		w/o LM	SF (λ_{add})	DRA (λ_{sub} / λ_{add})	Proposed ($\lambda_{\mathrm{sub_1}}$, $\lambda_{\mathrm{sub_2}}$ / λ_{add})
APS	1	21.5	17.5 (0.5)	15.7 (0.3 / 0.7)	—
SPS	1	14.7	12.9 (0.3)	11.6 (0.5 / 0.7)	—
APS+SPS	1	12.8	11.0 (0.5)	**10.5** (0.3 / 0.5)	—
APS, SPS	2	15.2	13.2 (0.7)	—	**11.6** (0.7, 0.3 / 1.1)

proposed method achieved higher ASR accuracy. Therefore, we can confirm that the DRA is effective even when multiple E2E ASR models are used together.

Next, the experimental results when the target domain was CEJC is shown in Table IV. We can see that the CER achieved by the proposed method is 35.1%, which outperformed the single ASR model methods. On the other hand, unlike when using MS as the target domain, the accuracy was higher than APS+SPS. We assumed that when a single model is trained, the probability distribution converges to fit to the source domains, but when multiple models are integrated, multiple distributions are combined to form an abstract representation. This distribution covers the feature space robustly and improves recognition performance in the domains different from the source domains. Therefore, it indicates that the proposed method is effective even when the speech styles of the source and target domains are different.

3) CERs with CTC/Attention model based Transformer: Experimental results when using Transformer-based ASR models are shown in Table V. We can see that the CER achieved by the proposed method is 11.6%, which is much better than that of the RNN-based version (16.3%). However, the relative relationship between the performance of the proposed and conventional methods remained almost the same. The accuracy of DRA methods with SPS and that of the proposed method were the same however, which is presumably because the performance of the Transformer model is already very high, and is not improved by the combination of multiple ASR models. But these results show that the proposed method achieved performance comparable to that of the best performing conventional model, regardless of which type of ASR model was used.

V. CONCLUSION

In this study we proposed a method of combining multiple E2E ASR models using a density ratio approach. We used subtraction through DRA to extract pure acoustic information from multiple existing E2E ASR models, and then integrated the extracted acoustic information with an external language model for the target domain to create an ASR model for the target domain. The proposed method achieved better performance than most of the conventional methods using a single E2E ASR model. By using two different, existing ASR models in our proposed method, we have successfully created an acoustically robust ASR model for the target domain without the need to retrain the ASR model.

In our comparative experiment, we only combined two E2E ASR models in our proposed method, but in future work we would like to evaluate the effectiveness of using three or more models for acoustic information. We would also like to compare our method's performance with ROVER, a conventional combination method. Furthermore, we plan to conduct experiments using ASR models trained with noisy data and with data containing multiple speech styles.

REFERENCES

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, 1997, pp. 347–354.

[2] H. Schwenk and J.-L. Gauvain, "Combining multiple speech recognizers using voting and language model information," in *Proc. ICSLP*, 2000, pp. 915–918.

[3] T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki, and S. Nakagawa, "An empirical study on multiple LVCSR model combination by machine learning," in *Proc. NAACL-HLT*, 2004, pp. 13–16.

[4] M. Matsushita, H. Nishizaki, T. Utsuro, Y. Kodama, and S. Nakagawa, "Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven web retrieval task," in *Proc. Eurospeech*, 2003, pp. 1205–1208.

[5] B. O. Kang and O.-W. Kwon, "Combining multiple acoustic models in GMM spaces for robust speech recognition," *IEICE Transactions on Information and Systems*, vol. 99, no. 3, pp. 724–730, 2016.

[6] J. Wong, Y. Gaur, R. Zhao, *et al.*, "Combination of end-to-end and hybrid models for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 1783–1787.

[7] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proc. ICASSP*, 2013, pp. 6744–6748.

[8] Z. Tuske, G. Saon, and B. Kingsbury, "On the limit of English conversational speech recognition," in *Proc. INTERSPEECH*, 2021, pp. 2062–2066.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, 2010, pp. 1045–1048.

[10] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," in *Proc. INTERSPEECH*, 2017, pp. 523–527.

[11] T. Hori, S. Watanabe, and J. R. Hershey, "Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition," in *Proc. ASRU*, 2017, pp. 287–293.

[12] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *Proc. ICASSP*, 2018, pp. 5824–5828.

[13] C. Gulcehre, O. Firat, K. Xu, *et al.*, "On using monolingual corpora in neural machine translation," in *Proc. WMT*, 2018, pp. 144–155.

[14] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold Fusion: Training Seq2Seq models together with language models," in *Proc. INTERSPEECH*, 2018, pp. 387–391.

[15] C. Shan, C. Weng, G. Wang, *et al.*, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *Proc. ICASSP*, 2019, pp. 5361–5635.

[16] E. McDermott, H. Sak, and E. Variani, "A density ratio approach to language model fusion in end-to-end automatic speech recognition," in *Proc. ASRU*, 2019, pp. 434–441.

[17] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid Autoregressive Transducer (HAT)," in *Proc. ICASSP*, 2020, pp. 6139–6143.

[18] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing RNN transducer technology for speech recognition," in *Proc. ICASSP*, 2021, pp. 5654–5658.

[19] D. Mori, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "Advanced language model fusion method for encoder-decoder model in Japanese speech recognition," in *Proc. APSIPA ASC*, 2021, pp. 503–510.

[20] Z. Meng, S. Parthasarathy, E. Sun, *et al.*, "Internal language model estimation for domain-adaptive end-to-end speech recognition," in *Proc. SLT*, 2021, pp. 243–250.

[21] Z. Meng, N. Kanda, Y. Gaur, *et al.*, "Internal language model training for domain-adaptive end-to-end speech recognition," in *Proc. ICASSP*, 2021, pp. 7338–7342.

[22] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & SSPR*, 2003, pp. 7–12.

[23] K. Itou, M. Yamamoto, K. Takeda, *et al.*, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[24] H. Koiso, H. Amatani, Y. Den, *et al.*, "Design and Evaluation of the Corpus of Everyday Japanese Conversation," in *Proc. LREC2022*, 2022, pp. 5587–5594.

[25] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[26] S. Karita, N. Chen, T. Hayashi, *et al.*, "A comparative study on transformer vs rnn in speech applications," in *Proc. ASRU*, 2019, pp. 449–456.

[27] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[28] S. Watanabe, F. Boyer, X. Chang, *et al.*, "The 2020 ESPnet update: New features, broadened applications, performance improvements, and future plans," in *Proc. DSLW*, 2021, pp. 1–6.

[29] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Speech-Emotion Control for Text-to-Speech in Spoken Dialogue Systems Using Voice Conversion and x-vector Embedding

Shunichi Kohara*†, Masanobu Abe* and Sunao Hara*
* Okayama University, Okayama, Japan
† E-mail: pdaepw2@s.okayama-u.ac.jp

***Abstract*—In this paper, we propose an algorithm to control both speaker individuality and emotional expressions in synthesized speech, where the most important feature is the controllability of intensity in emotional expressions. An aim of the proposed algorithm is to generate various responses including emotions in text-to-speech (TTS) for spoken dialogue systems (SDS), which results in making the system more human-like. An idea is to control emotion and its intensity in line with the user's utterances. For example, when a user happily talks to SDS, the agent of the SDS responses with happy voice. Generally, voice quality of a user and the agent are different. Therefore, the proposed algorithm consists of two steps: (1) voice conversion to change speaker individuality including emotional expressions and (2) TTS with x-vector acting as an embedding vector to mainly control speech quality related to the intensity of emotions. Evaluation experiments are carried out using a scenario of a spoken dialogue system, where a teacher system of TTS encourages or cheers up students according to students' utterances. The experiment results showed that TTS can successfully reproduce the emotion and its intensity that are extracted from students' utterances, while maintaining the teacher's speaker individuality.**

I. Introduction

Spoken dialogue systems (SDS) are systems that accomplish a range of tasks using speech, and consist of speech recognition, dialogue processing, and speech synthesis. Important features of SDS is the function to appropriately synthesize speech that conveys rich information such as focus and emotions, and generate various responses according to situations. Monotonous speech responses makes the systems boring, and repetitions of the same sentences are not interesting to use. To address these problems, several methods are proposed such as the usage of linguistic and/or prosodic contexts of dialogue history [1], [2], and the modeling of speaking styles [3], [4].

In this paper, in terms of expressing emotional information, we propose a different approach, i.e., to reproduce emotion and its intensity in synthesized speech to match the user's utterances. Figure 1 shows examples of dialogues. When a user uttered a sentence with sad, happy, and very happy emotions, the agent of SDS responds with the same emotion, sad, happy, and very happy, respectively.

For emotional speech synthesis, extensive studies have been carried out using deep neural networks(DNN) [5]–[7] which showed that emotional IDs work well to control emotional information in DNN. Since emotional IDs are categorical feature vectors, it is not always suitable to control the intensity of emotions. [8] reported that emotional information can be represented by utterance-level x-vectors [9] that were originally proposed to extract features of speakers. In addition, x-vectors can be applied to emotion recognition [10]. [11] proposed an embedding vector called global style tokens (GST) and showed that GST are related to emotional information. Although x-vectors or GST is suitable to control the level of intensity unlike emotional IDs, it is difficult to disentangle the feature representation from the emotion and speaker's individuality.

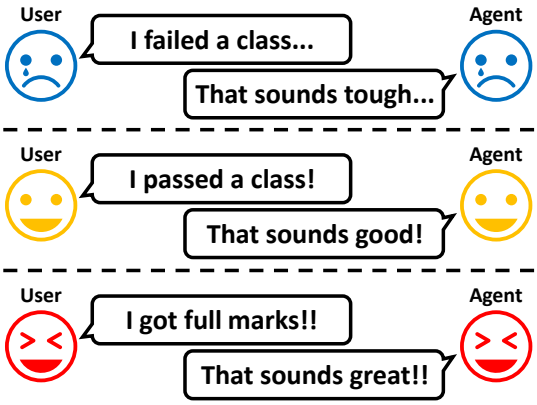


Fig. 1. Examples of dialogues. Our SDS aim to generate response speech in line with the user's speech taking into account nonverbal information, especially emotion, in the user's speech.

In this paper, we propose an algorithm to control both speaker individuality and emotional expressions in synthesized speech; the most important feature is the controllability of intensity in emotional expressions [12]. A promising usage of the proposed algorithm is to generate responses in line with the user's emotions in spoken dialogue systems which makes it possible to generate various responses with emotions and its their intensity making the system interesting or pleasant. This strategy might not always work in SDS systems, but occasionally be reasonable. The proposed algorithm consists of two steps: (1) voice conversion to change speaker individuality including emotional expressions and (2) TTS with x-vector extracted from the converted speech as the embedding vector

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

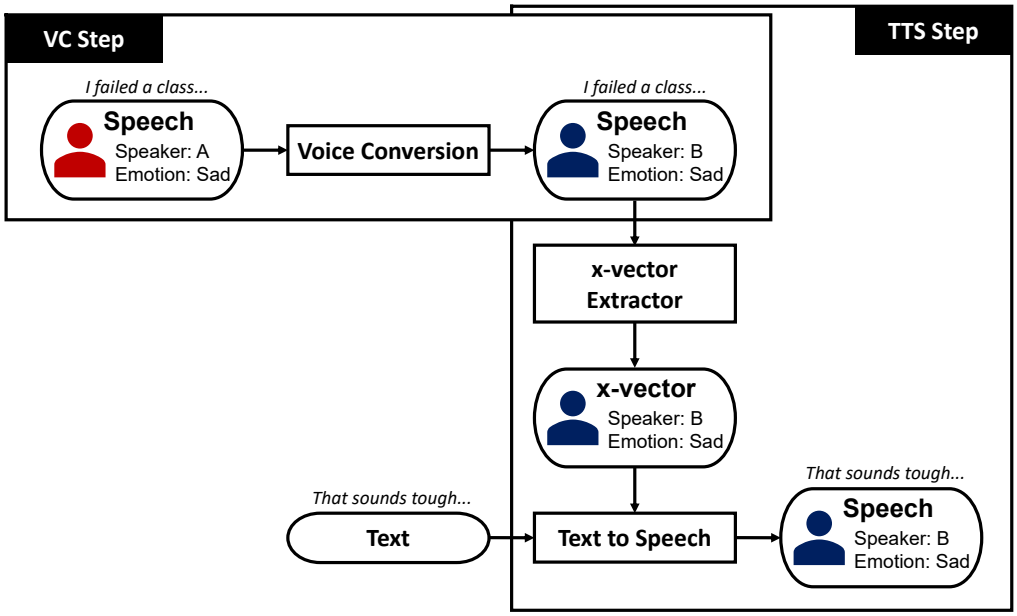


Fig. 2. Outline of the proposed algorithm. The proposed algorithm consists of a VC Step and a TTS Step.

to control the speech quality of the emotion and its intensity.

The rest of this paper is organized as follows: Section 2 describes the outline of the proposed algorithm, Section 3 explains the setup for carrying out evaluation experiments of the proposed algorithm, Section 4 presents the experiment methods and evaluation results, and Section 5 discusses our conclusions and suggests avenues for future work.

II. Proposed Algorithm

Figure 2 shows the outline of the proposed algorithm, which consists of 2 steps: 1. VC Step and 2. TTS Step.

A. Training Procedure

1) VC Step: In the VC Step, we train the multi-speaker setting VITS[13][1] as a voice conclusion model. VITS architecture can apply many-to-many voice conversions, so we use this architecture to convert the speaker's individuality. The speakers of multi-speaker VITS are controlled by speaker ID. At this time, we assume that voice conversion changes the speaker's individuality by inheriting emotional expressions.

2) TTS Step: In the TTS Step, we train a single-speaker (an agent of SDS) TTS model using VITS with x-vector (X-VITS) on ESPnet2-TTS [14][2]. The training data of the speaker contains various instances of emotional speech, and we expect it to express emotion and its intensity by conditioning x-vectors extracted from various instances of emotional speech. The x-vector extractor is trained by ECAPA-TDNN [15] architecture, using large-scale multi-speaker datasets VoxCeleb [16] by SpeechBrain[17].

[1]VITS: https://github.com/jaywalnut310/vits
[2]ESPnet: https://espnet.github.io/espnet/

B. Inference Procedure

1) VC Step: In the VC Step, we convert the speaker individuality of input speech to the TTS model (the agent of SDS). A process of voice conversion by VITS is formulated as follows;

$$z \sim \quad q_\phi(z \mid x_{lin}, s) \tag{1}$$

$$e = \quad f_\theta(z \mid s) \tag{2}$$

For a given speaker identity s and an utterance of the speaker, VITS needs a linear spectrogram x_{lin} from the corresponding utterance. VITS can transform x_{lin} into a speaker-independent representation e through the posterior encoder and the normalizing flow [18] in the prior encoder. Then, the converted speech $\hat{y}$ of a target speaker identity $\hat{s}$ from the representation e is synthesized through the inverse transformation of the normalizing flow f_θ^{-1} and decoder G:

$$\hat{y} = G(f_\theta^{-1}(e \mid \hat{s}) \mid \hat{s}). \tag{3}$$

The speech converted speaker individuality as inheriting emotional expressions.

2) TTS Step: In the TTS Step, we synthesize speech conditioned x-vectors extracted from the converted speech at the VC Step. x-vectors trained with large-scale multi-speaker datasets can encode information such as speaker individuality and emotion and its intensity. Therefore, the TTS conditioned with x-vectors can synthesize the agent of SDS speech including a variety of emotions.

III. Experiments Setup

A. Dataset

The dataset used in evaluation experiments of this paper is STUDIES [19]. STUDIES is the Japanese emotional speech

TABLE I
SEVERAL UTTERANCES BY EMOTION LABEL (WITHOUT PHONETICALLY-BALANCED SENTENCES SPEECH).

Speaker	Neutral	Happy	Sad
Teacher	1796	615	212
M-student	351	381	329
F-student	300	417	340

corpus and contains speech by three voice actors (one male and two female). This dataset consists of simulated dialogue speech between a female teacher who is an agent of SDS in this paper (hereinafter called the Teacher) and a male student (hereinafter called the M-student) and between the Teacher and a female student (hereinafter called the F-student). Also, STUDIES has speech of phonetically-balanced sentences of Japanese, ITA corpus[3], by the Teacher. All audio data are 48kHz/16bit wav files and consist of the attached text, monophone labels, and full-context labels with phoneme alignments.

Each utterance is annotated with emotion labels of four types: Neutral, Happy, Sad, and Angry. In evaluation experiments in this paper, we evaluate three emotions: Neutral, Happy, and Sad, because the Angry utterances are less necessary for our proposed dialogue strategy. Table I shows the number of utterances by emotion label. As a result, the total number of hours for each speaker is approximately 5.6 hours for the Teacher, 1.3 hours for the M-student, and 1.5 hours for the F-student. In this paper, we treat the Teacher of STUDIES as a agent of SDS and the M-student and the F-student as users of SDS.

B. Training Models

We trained each of the models for evaluation experiments, for which speech was down-sampled to 22.05 kHz in advance. Also, 80% of each speaker's utterances and each emotion of the dataset were allocated to the training data, and the remaining 20% to the validation and evaluation data.

In the VC Step, we trained a VITS models for the three-speakers (the Teacher, the M-student, and the F-student) and assigned speaker ID to the three speakers of STUDIES. As a side note, the speech of each speaker contains all emotions. For training multi-speaker VITS, phonemes are derived from full-context labels of the dataset, and we used the method of Kurihara et al. [20] to obtain prosody symbols. The model was trained for 20k epochs using the AdamW optimizer [21] with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay $\lambda = 0.01$. Settings of networks like hyperparameters are not changed from official implementation[4].

In the TTS Step, we trained a X-VITS of the Teacher, which can control emotions by x-vectors. The phoneme and prosody symbols are obtained in the same way as in the VC Step. This model was trained for 20k epochs using the AdamW optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay $\lambda = 0.01$. The settings of networks, like hyperparameters, conform to the VCTK dataset[5] Recipe of ESPnet. Also, to compare with the proposed algorithm, we trained a VITS of the Teacher using one-hot emotional ID as a multi-emotion model. All conditions without one-hot settings are the same as in the X-VITS model.

[3]ITA corpus: https://github.com/mmorise/ita-corpus
[4]Official implementation: https://github.com/jaywalnut310/vits

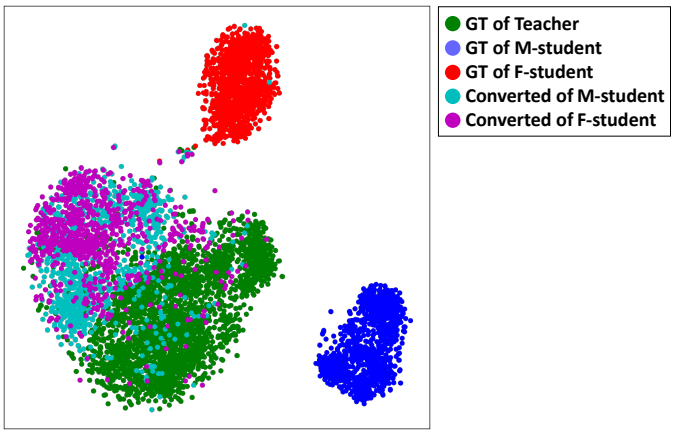


Fig. 3. Distribution of x-vectors. • is x-vector extracted from one utterance. "Converted M-student" and "Converted F-student" are converted speech from the Studnets to the Teacher.

IV. EVALUATION EXPERIMENTS

A. Analysis of VC Step

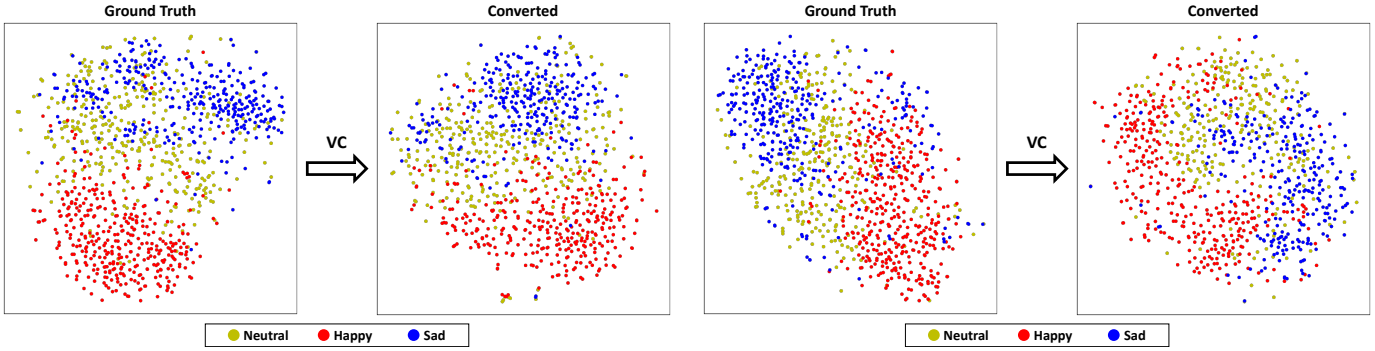
Figure 3 shows the distribution of x-vectors extracted from the ground truth of three speakers in STUDIES and the converted speech of the M-student and the F-student to the Teacher. In this figure, • means each x-vector extracted from one utterance. The dimensions of the x-vectors were reduced by t-SNE [22]. Although the converted speech of the M-student and the F-student are not exactly same as the Teacher's speech itself, this figure shows that the converted speech of the M-student and the F-student are closer to the characteristics of the Teacher than the ground truth of the M-student and the F-student.

To examine emotions in detail, Figure 4 shows the distributions of x-vectors for the M-student and the F-student before and after applying voice conversion. In the figure, • means the x-vector extracted from one utterance. About ground truth, this figure shows that the happy speech of both of the M-student and the F-student is far from Neutral and Sad, and Neutral and Sad are distributed nearby. These relationships are also generally inherited for the converted speech. To summarize voice conversion, the conversion to the speaker individuality while retaining emotional expressions, was adjudged successful.

B. Subjective Evaluation Experiments

As subjective evaluation experiments of the proposed algorithm, we carried out emotional identification tests and paired-comparison experiments using synthesized speech, and these experiments were conducted in a soundproofed room (background noise level: 25 dB) using Sony CD900ST headphones.

[5]VCTK dataset: https://datashare.ed.ac.uk/handle/10283/3443



(a) M-student (b) F-student

Fig. 4. Distributions of x-vectors of each Student before and after voice conversion. ● is the x-vector extracted from one utterance.

TABLE II
CONFUSION MATRICES OF EXPERIMENTAL RESULTS.

(a) OneHot-EmoID (accuracy: 0.903)

Correct emorion	Subject-perceived emotions Neutral	Happy	Sad
Neutral	**0.940**	0.000	0.060
Happy	0.120	**0.880**	0.000
Sad	0.110	0.000	**0.890**

(b) Teacher-GT (accuracy: 0.963)

Correct emorion	Subject-perceived emotions Neutral	Happy	Sad
Neutral	**0.970**	0.020	0.010
Happy	0.000	**1.000**	0.000
Sad	0.080	0.000	**0.920**

(c) M-student-Converted (accuracy: 0.830)

Correct emorion	Subject-perceived emotions Neutral	Happy	Sad
Neutral	**0.860**	0.030	0.110
Happy	0.000	**1.000**	0.000
Sad	0.370	0.000	**0.630**

(d) F-student-Converted (accuracy: 0.690)

Correct emorion	Subject-perceived emotions Neutral	Happy	Sad
Neutral	**0.870**	0.040	0.090
Happy	0.040	**0.960**	0.000
Sad	0.730	0.030	**0.240**

TABLE III
PRECISION AND RECALL OF EXPERIMENTAL RESULTS.

(a) OneHot-EmoID

	Neutral	Happy	Sad
precision	0.803	1.000	0.937
recall	0.940	0.880	0.890

(b) Teacher-GT

	Neutral	Happy	Sad
precision	0.924	0.980	0.989
recall	0.970	1.000	0.920

(c) M-student-Converted

	Neutral	Happy	Sad
precision	0.699	0.971	0.851
recall	0.860	1.000	0.630

(d) F-student-Converted

	Neutral	Happy	Sad
precision	0.530	0.932	0.727
recall	0.870	0.960	0.240

The participants are ten native Japanese speakers. Note that examples of generated speech can be found on our webpage[6].

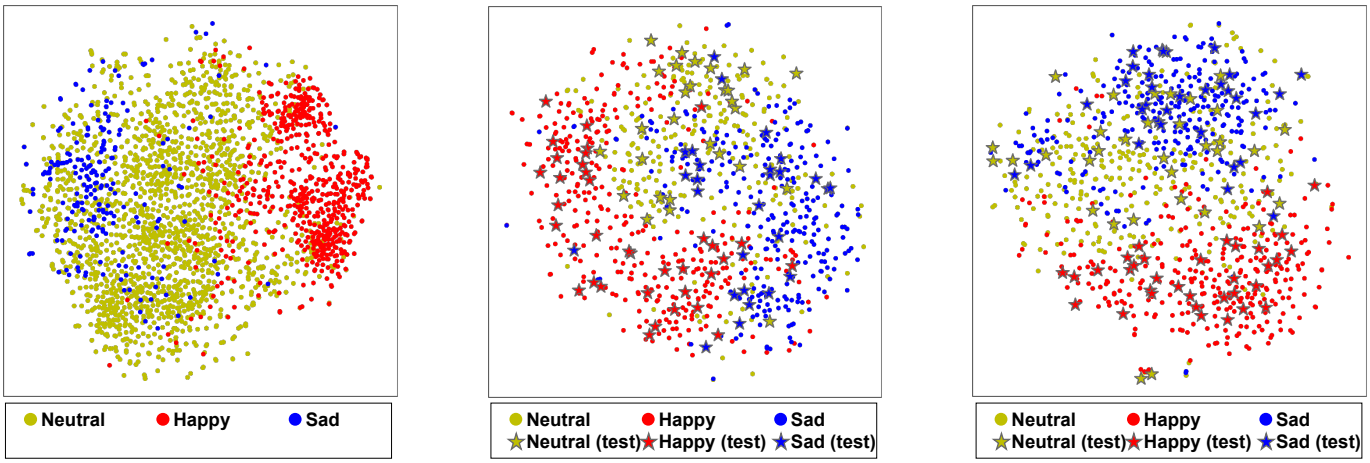
1) Emotion Identification Tests: We asked participants to listen to synthesized speech and select an emotion representing each stimulus from Neutral, Happy, and Sad, and the speech was synthesized under the following four conditions.

1) The Teacher model with one-hot emotional ID (OneHot-EmoID)
2) Averaged x-vectors extracted from the ground truth of the Teacher (Teacher-GT)
3) Averaged x-vectors extracted from the converted speech of the M-student (M-student-Converted)
4) Averaged x-vectors extracted from the converted speech of the F-student (F-student-Converted)

In these conditions, OneHot-EmoID means the very basic emotional speech control algorithm by one-hot ID, and Teacher-GT, such that synthesized speech with x-vectors extracted from the Teacher's ground truth, denotes the ideal conditions for using x-vectors. Compared with the ideal condition for the ground truth of the Teacher, we verify how much control is possible with x-vectors extracted from the converted speech. Incidentally, x-vectors of the Teacher-GT are the mean of "all" data of the Teacher and x-vectors of the M-student-Converted and the F-student-Converted are the mean of "test" data of each converted speech. For each condition, 15 stimuli (3 emotions $\times$ 5 texts) were generated. In terms of texts for TTS, sentences with neutral emotion were selected from the J-set of ATR Japanese speech database [23]. In total, 60 stimuli were presented to the participants in random order. Each stimulus was presented twice.

Table II shows the confusion matrices of the emotion recognition experimental result. About Table II, the rows show the correct emotions, while the columns show subject-perceived emotions.

[6]Demo page: https://oohara1202.github.io/apsipa2023-sample/



(a) Teacher-GT (b) M-student-Converted (c) F-student-Converted

Fig. 5. Distributions of x-vectors used to calculate averaged x-vectors. Note that ★ of (b) and (c) means x-vectors extracted from "test" data, are used to calculate averaged x-vectors.

First of all, compared with OneHot-EmoID and the Teacher-GT, the Teacher-GT is not less than OneHot-EmoID, so speech-emotion control by x-vectors seems valid. Second, compared with the Teacher-GT and the proposed algorithm, the M-student-Converted and the F-student-Converted, the identification rate of Neutral and Happy in the proposed algorithm is almost the same as the Teacher-GT, but that of Sad is inferior. Table III shows precision and recall, however, the precision of Sad at M-student-Converted and F-student-Converted is over 70%, which suggests that the proposed algorithm can express Sad itself.

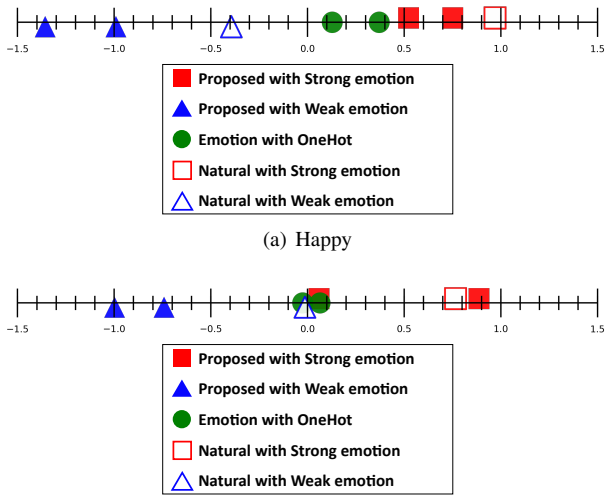
The identification rate of the F-student-Converted's Sad has only 24%, and this might come from the averaged x-vectors used in synthesis. So, we plot x-vectors used to calculate averaged x-vectors in Figure 5. Judging from this figure, the boundary between Neutral and Sad is blurred by the order of the Teacher-GT, MStudent-Converted, and FStudent-Converted, and the identification rate of Sad in each condition is the same as the order. Therefore, we can say that x-vectors caused poor judgments in Sad. At the same time, we have to pay attention the fact that there is little differences in grand truth utterances between Neutral and Sad as shown in Figure 4.

2) Paired-Comparison Experiments: Paired-comparison experiments are carried out to evaluate whether the proposed algorithm can express the emotion's intensity. First, we selected four utterances from the M-student utterances; strong-Happy, weak-Happy, strong-Sad and weak-Sad. Those are assumed to be input utterances to a SDS. Then, as response utterances of the SDS, the proposed algorithm synthesized following two types of speech.

1) Using x-vectors extracted from converted speech with strong emotion
2) Using x-vectors extracted from converted speech with weak emotion

As references, we also prepared following synthesized speech.

3) The synthesized speech generated by OneHot-EmoID



(a) Happy

(b) Sad

Fig. 6. Thurstone's Case V scale of experimental results. Outlined square and outlined triangle indicate natural speech uttered by the M-student with strong emotion and with weak emotion.

Moreover, we used following natural speech for comparisons.

4) Ground truth uttered by the M-student with strong emotion
5) Ground truth uttered by the M-student with weak emotion

In terms of emotions, we evaluated for Happy and Sad, and listening tests were carried out separately using Happy stimuli and Sad stimuli. As response texts of synthesized speech, for Happy and Sad, two sentences are used. As stimuli pairs, all combinations of the five types of speech are used. Participants were asked to select an utterance with stronger emotions from presented pair. The total number of the pairs was 28, and each pair was presented twice.

Figure 6 shows the values of Thurstone's Case V scale calculated from the results of paired-comparison, where larger

values indicate a stronger emotional expression. In Figure 6, outlined square and outlined triangle indicate natural speech uttered by the M-student with strong emotion and with weak emotion, respectively.

First of all, in Figure 6, OneHot-EmoID is always placed in the similar position, which indicates that participants consistently judged the emotional intensity of the OneHot-EmoID. This is quite reasonable. In the results for Happy, it is clearly observed that the proposed algorithm can control intensities of the emotion as good as the natural speech. In the results for Sad, although performance is inferior to the case of Happy, the same tendencies are observed. Judging from these results, we can say that the proposed algorithm can successfully control the intensities of the emotion.

V. Conclusions

In this paper, we proposed a algorithm to control speech-emotion including the intensity of emotions to fit the emotions of input speech in SDS using voice conversion and x-vector embedding. First, the proposed algorithm converts the speaker individuality of input speech to the TTS model speaker including emotional expressions, and second, a TTS model synthesizes speech with x-vectors extracted from the converted speech to add emotions and their intensities. In evaluation experiments, the results showed speech-emotions are controllable by the proposed algorithm and its intensity is expressed appropriately against input-speech.

In future, we have a plan to evaluate that the proposed algorithm makes a SDS system interesting or pleasant by generating various responses with emotions and its their intensity. We also need to investigate a voice conversion method that uses a small amount of the source speaker's data to enlarge the area of the proposed algorithm and to carry out a listening experiment about the speaker individuality of speech synthesized by the proposed algorithm. In addition, we would like to investigate to use a categorical classification of emotions for controlling the emotional intensity instead of VC steps and to compare the proposed algorithm with TTS that can be controlled by speaker and emotion IDs.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K11963.

References

[1] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end tts for voice agents," in *Proceedings of 2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 403–409. DOI: 10.1109/SLT48900.2021.9383460.

[2] Y. Nishimura, Y. Saito, S. Takamichi, K. Tachibana, and H. Saruwatari, "Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History," in *Proceedings of Interspeech 2022*, 2022, pp. 3373–3377. DOI: 10.21437/Interspeech.2022-403.

[3] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems," in *Proceedings of Interspeech 2021*, 2021, pp. 246–250. DOI: 10.21437/Interspeech.2021-381.

[4] K. Mitsui, T. Zhao, K. Sawada, Y. Hono, Y. Nankaku, and K. Tokuda, "End-to-End Text-to-Speech Based on Latent Representation of Speaking Styles Using Spontaneous Dialogue," in *Proceedings of Interspeech 2022*, 2022, pp. 2328–2332. DOI: 10.21437/Interspeech.2022-259.

[5] A. Triantafyllopoulos, B. W. Schuller, G. İymen, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, pp. 1–27, 2023. DOI: 10.1109/JPROC.2023.3250266.

[6] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model architectures to extrapolate emotional expressions in DNN-based text-to-speech," *Speech Communication*, vol. 126, pp. 35–43, 2021, ISSN: 0167-6393.

[7] K. Matsumoto, S. Hara, and M. Abe, "Speech-like emotional sound generation using wavenet," *IEICE Transactions on Information and Systems*, vol. E105.D, no. 9, pp. 1581–1589, 2022. DOI: 10.1587/transinf.2021EDP7236.

[8] J. Williams and S. King, "Disentangling Style Factors from Speaker Representations," in *Proceedings of Interspeech 2019*, 2019, pp. 3945–3949. DOI: 10.21437/Interspeech.2019-1769.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[10] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173. DOI: 10.1109/ICASSP40776.2020.9054317.

[11] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 5180–5189.

[12] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2023.

[13] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end

text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 5530–5540.

[14] T. Hayashi, R. Yamamoto, K. Inoue, *et al.*, "Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit," in *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7654–7658. DOI: 10.1109/ICASSP40776.2020.9053512.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proceedings of Interspeech 2020*, 2020, pp. 3830–3834. DOI: 10.21437/Interspeech.2020-2650.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech and Language*, vol. 60, p. 101 027, 2020, ISSN: 0885-2308.

[17] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, *Speechbrain: A general-purpose speech toolkit*, 2021. DOI: 10.48550/ARXIV.2106.04624. [Online]. Available: https://arxiv.org/abs/2106.04624.

[18] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 1530–1538.

[19] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent," in *Proceedings of Interspeech 2022*, 2022, pp. 5155–5159. DOI: 10.21437/Interspeech.2022-300.

[20] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural tts," *IEICE Transactions on Information and Systems*, vol. E104.D, no. 2, pp. 302–311, 2021. DOI: 10.1587/transinf.2020EDP7104.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of International Conference on Learning Representations*, 2019.

[22] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[23] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990, ISSN: 0167-6393.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Narrow-edged Acoustical Beamforming Utilizing Phase Inversion for Frequency Modulation-based Parametric Array Loudspeaker

Yuting Geng*, Masato Nakayama† and Takanobu Nishiura*
* Ritsumeikan University, Japan
E-mail: {geng@fc, nishiura@is}.ritsumei.ac.jp
† Osaka Sangyo University, Japan
E-mail: nakayama@ise.osaka-sandai.ac.jp

***Abstract*—Parametric array loudspeakers (PALs) can achieve narrow directivity with the straightness of ultrasonic waves. However, the acoustic beam formed by small-scale PALs has a tendency to spread easily. We have previously proposed a narrow-edged beamforming based on phase inversion for common amplitude modulation-based PALs; however, it decreased the sound pressure significantly when forming a narrow acoustic beam. In this study, we focus on frequency modulation-based PALs, in which a modulated wave has a constant envelope and can hold more energy in a limited amplitude. The phase inversion strategy is discussed to use phase cancellation in both ultrasound and audible sound for a narrow beam. The experimental results show that the proposed method can achieve narrow-edged acoustical beam while reducing the decrease in sound pressure.**

I. Introduction

Commonly, loudspeakers are used to reproduce acoustic signals. A typical electro-dynamic loudspeaker emits audible sound directly; therefore, the sound it reproduces has a wide directivity. In contrast, a parametric array loudspeaker(PAL) [1], [2], [3] emits ultrasounds and reproduces an audible sound indirectly; as a result, the reproduced audible sound inherits the straightness of the emitted ultrasounds and achieves a sharp directivity. One basic principle of the PAL is the nonlinearity of air, which causes the generation of new frequency components, including the sum frequency and difference frequency of emitted components [4], [5]. This phenomenon can be exploited to reproduce an audible component at frequency f_{d} by emitting a pair of ultrasonic components at frequencies f_{c} and $f_{\mathrm{c}} - f_{\mathrm{d}}$. The emission signal for the PAL can be generated by modulating an ultrasonic carrier wave at frequency f_{c} with target sound at frequency f_{d}. Here, the audible sound generated as the difference frequency of the emitted ultrasounds can be called demodulated sound [6].

A typical PAL, which consists of dozens of ultrasonic transducers, can be regarded as a piston source under the conditions that the number of transducers is sufficiently large and the interval is adequately small. Because the demodulated sound inherits the straightness of emitted ultrasounds, PALs have been used in many applications [7], [8], [9], [10]. In contrast, small-scale PALs tend to suffer the problem that the beam of reproduced audible sound spreads widely. Directivity control methods for PAL's reproduced sound have been proposed based on transducer configuration [11], [12], [13] and array signal processing [14], [15], [16]. Moreover, in the study of phase cancellation excitation [17], it is reported that phase cancellation in ultrasounds can widen the directivity of the demodulated sound of a PAL. In the study of length-limited PALs [18], [19], [20], it is reported that the phase cancellation of audible sound limits the acoustic beam in the length direction.

In our previous works [21], [22], we focused on controlling the acoustical beam in the width direction, to form a narrow-edged acoustical beam. Based on the phase cancellation in both ultrasounds and audible sound, we have proposed narrow-edged beamforming using individual phase inversion in amplitude-modulated waves for PALs [21], [22]. In this method, the transducers are divided into interior and exterior zones. Either the carrier or the sideband is phase inverted in the signal fed to the exterior transducers, while all components hold their original phase in the signal fed to the interior transducers. Both numerical and experimental results demonstrated that a narrower beam can be formed with this method [22]. However, this conventional method suffers from two main drawbacks: one is the trade-off between the directivity and the sound pressure, and the other is the occurrence of side lobes in directivity.

Our goal is to form a narrow-edged acoustical beam with less degradation of the sound pressure, which is the solution to the first drawback of the conventional method. The conventional method was proposed for the amplitude modulation (AM) methods for PAL, whose modulated wave has a simple spectral structure consisting of a carrier wave and only one or two sideband wave(s). In this paper, we focus on the frequency modulation (FM) method for PAL [23], which has a complicated spectral structure consisting of a carrier wave and multiple sideband waves. We propose a narrow-edged acoustical beamforming method using phase inversion for FM-based PAL. This proposed method shares part of the theoretical basis with the conventional method [22], and keeps the same division of transducers as in the conventional method. However, the phase inversion strategy is re-discussed

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

based on the characteristics of FM for PAL. Experimental evaluation was carried out to demonstrate the effectiveness of the proposed method.

II. Principle of frequency modulation-based parametric array loudspeaker

A. Basic Principles of Parametric Array Loudspeakers

The PAL is an application of a parametric array [24] for audible sound reproduction in the air. In the simplest case, the signal emitted from a PAL consists of two ultrasonic components and can be expressed as

$$p(t) = \sum_{n=1,2} P_n \cos\left(2\pi f_n t\right), \tag{1}$$

where t denotes time and P_n and f_n denote the initial pressure and frequency of the ultrasound, respectively. The emitted ultrasound observed at distance r on the propagation axis can be expressed as

$$p_r(t) = \sum_{n=1,2} P_n e^{-\alpha_n r} \cos\left(2\pi f_n t - k_n r\right), \tag{2}$$

where α_n and k_n denote the attenuation rate and wavenumber of the ultrasound, respectively.

Based on Westervelt's theory, virtual sources are generated as frequencies $2f_1$, $2f_2$, $f_1 + f_2$ and $|f_1 - f_2|$. Considering the demodulation process of the ultrasound emitted from PAL, we focus on the difference frequency $f_\mathrm{d} = |f_1 - f_2|$, whose intensity can be expressed as

$$q_\mathrm{d}(t) = \frac{2\pi\beta P_1 P_2 f_\mathrm{d}}{\rho_0^2 c_0^4 e^{(\alpha_1+\alpha_2)r}} \sin(2\pi f_\mathrm{d} t - k_\mathrm{d} r), \tag{3}$$

where β, ρ_0 and c_0 denote a nonlinear coefficient, the density and the speed of sound in the air, respectively, and k_d denotes the wavenumber of the difference frequency.

The demodulated sound observed in the far field can be expressed as a volume integral of virtual sources at difference frequency f_d, and can be further approximated into a line integral along the propagation axis as follows

$$\begin{aligned} p_\mathrm{d}(t) &= \frac{\rho_0}{4\pi} \int_V \frac{e^{-jk_\mathrm{d}R}}{R} \frac{\partial q_\mathrm{d}(t)}{\partial t} \mathrm{d}V \\ &\approx \frac{\rho_0 S_0}{4\pi} \int_0^{l_0} \frac{e^{-jk_\mathrm{d}R(r)}}{R(r)} \frac{\partial q_\mathrm{d}(t)}{\partial t} \mathrm{d}r, \end{aligned} \tag{4}$$

where $R(r)$ denotes the distance from the virtual source at distance r. Here, the integral interval is approximated as a cylinder with volume V, base area S_0 and height l_0.

The signal emitted from the PAL can be generated by modulating an ultrasonic carrier $c(t)$ with the target sound $s(t)$. In the case of AM methods, which are widely used for PAL, the amplitude of the carrier wave holds the information of the amplitude of the target sound. For example, the modulated wave of double sideband (DSB) modulation [1], [6], [25] can be given as

$$u_\mathrm{DSB}(t) = \left(1 + m_\mathrm{DSB} s(t)\right) c(t), \tag{5}$$

where $m_\mathrm{DSB} (0 < m_\mathrm{DSB} \leq 1)$ denotes the modulation factor. In general, the carrier wave $c(t)$ can be expressed as

$$c(t) = A_\mathrm{c} \cos\left(2\pi f_\mathrm{c} t\right), \tag{6}$$

where A_c and f_c denote the amplitude and frequency of the carrier wave, respectively. When the target sound $s(t)$ is a pure tone, it can be given as

$$s(t) = A_\mathrm{s} \cos\left(2\pi f_\mathrm{s} t\right), \tag{7}$$

where A_s and f_s denote the amplitude and frequency of the target sound, respectively.

Substituting (6) and (7) into (5), it can be derived that the DSB-modulated wave $u_\mathrm{DSB}(t)$ contains three frequency components: carrier wave f_c, upper sideband wave $f_\mathrm{c} + f_\mathrm{s}$, and lower sideband wave $f_\mathrm{c} - f_\mathrm{s}$. Because the difference frequency between the carrier wave and either sideband wave is the same as that of the target sound, the target sound at frequency f_s is expected to be generated with a self-demodulation process.

B. Frequency Modulation for Parametric Array Loudspeaker

In the case of FM, the instantaneous frequency of the carrier wave holds the amplitude information of the target sound. A frequency-modulated wave [23], [26], [27] can be given as

$$u_\mathrm{FM}(t) = A_\mathrm{c} \cos\left(2\pi f_\mathrm{c} t + 2\pi f_\mathrm{f} \int_0^t s(\tau)\mathrm{d}\tau\right), \tag{8}$$

where f_f denotes the frequency fluctuation. The instantaneous frequency $f_\mathrm{ins}(t)$ varies with time and can be derived as

$$f_\mathrm{ins}(t) = \frac{1}{2\pi}\frac{\mathrm{d}}{\mathrm{d}t}\left(2\pi f_\mathrm{c} t + 2\pi f_\mathrm{f} \int_0^t s(\tau)\mathrm{d}\tau\right) = f_\mathrm{c} + f_\mathrm{f} s(t). \tag{9}$$

Substituting (7) into (9), derives the instantaneous frequency $f_\mathrm{ins}(t)$ as sinusoidal with bias f_c and amplitude $A_\mathrm{s} f_\mathrm{f}$. In this case, the frequency-modulated wave $u_\mathrm{FM}(t)$ can be expanded with Bessel functions of the first kind, which can be expressed as

$$\begin{aligned} u_\mathrm{FM}(t) &= A_\mathrm{c} \cos\left(2\pi f_\mathrm{c} t + \frac{A_\mathrm{s} f_\mathrm{f}}{f_\mathrm{s}} \sin(2\pi f_\mathrm{s} t)\right) \\ &= A_\mathrm{c} \sum_{i=-\infty}^{\infty} J_i(m_\mathrm{FM}) \cos\left(2\pi\left(f_\mathrm{c} + i f_\mathrm{s}\right) t\right), \end{aligned} \tag{10}$$

where $J_i(\cdot)$ denotes the i-th order Bessel function of the first kind and m_FM denotes the modulation factor of FM which is defined as

$$m_\mathrm{FM} = \frac{A_\mathrm{s} f_\mathrm{f}}{f_\mathrm{s}}. \tag{11}$$

It is known that $J_{-i}(\cdot) = (-1)^i J_i(\cdot)$ holds for Bessel functions; therefore, (10) can be further expanded as

$$\begin{aligned} u_\mathrm{FM}(t) = {} & A_\mathrm{c} \sum_{i=1}^{\infty} J_i(m_\mathrm{FM}) \cos\left(2\pi\left(f_\mathrm{c} + i f_\mathrm{s}\right) t\right) \\ & + A_\mathrm{c} \sum_{i=1}^{\infty} (-1)^i J_i(m_\mathrm{FM}) \cos\left(2\pi\left(f_\mathrm{c} - i f_\mathrm{s}\right) t\right) \\ & + A_\mathrm{c} J_0(m_\mathrm{FM}) \cos\left(2\pi f_\mathrm{c} t\right). \end{aligned} \tag{12}$$

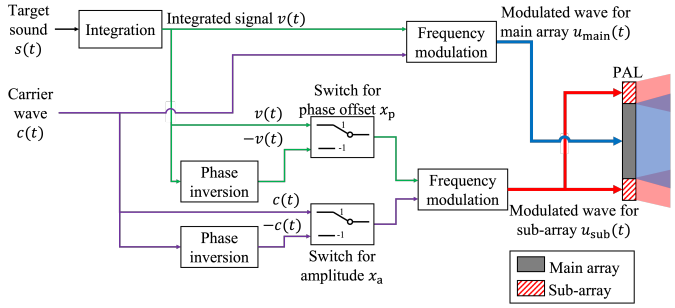


Fig. 1. Overview of the proposed method.

It is obvious from (12) that the first term indicates multiple upper sideband components, the second term indicates multiple lower sideband components, and the last term indicates a carrier component.

When the modulation factor m_{FM} is a very small value, for example, when $f_{\mathrm{f}} \ll f_{\mathrm{s}}$, the energy of high-order components ($|i| > 1$) can be approximated to zero. In this case, the frequency-modulated wave $u_{\mathrm{FM}}(t)$ has a spectral structure similar to that of the DSB-modulated wave $u_{\mathrm{DSB}}(t)$. In contrast, if the above condition does not hold, the high-order components cannot be ignored, and the frequency-modulated wave $u_{\mathrm{FM}}(t)$ will have a complicated structure. However, in both cases, the target sound at frequency f_{s} can be demodulated as the difference frequency between the carrier component and sideband component.

III. Proposed Narrow-edged Beamforming Using Phase Inversion for Frequency Modulation

A. Overview of the Proposed Method

Fig. 1 gives an overview of the proposed method. The proposed method uses phase cancellation in both ultrasounds and audible sounds by inverting the phase of components in the frequency-modulated wave. The PAL is divided into a main array in the interior and a sub-array in the exterior. The modulated wave for the main array is generated by common FM, as expressed in (8). First, the target sound $s(t)$ is integrated by the time index into signal $v(t)$ which can be given as

$$v(t) = \int_0^t s(\tau)\mathrm{d}\tau. \tag{13}$$

The integrated signal $v(t)$ is then used to determine the phase offset for a sinusoidal carrier wave at frequency f_{c}, and the output signal can be expressed as

$$u_{\mathrm{main}}(t) = A_{\mathrm{c}} \cos\left(2\pi f_{\mathrm{c}} t + 2\pi f_{\mathrm{f}} v(t)\right). \tag{14}$$

The modulated wave for the sub-array is generated by FM with phase inversion. The output signal can be expressed as

$$u_{\mathrm{sub}}(t) = x_{\mathrm{a}} A_{\mathrm{c}} \cos\left(2\pi f_{\mathrm{c}} t + 2\pi f_{\mathrm{f}} x_{\mathrm{p}} v(t)\right), \tag{15}$$

where we set two phase inversion switches x_{a} and x_{p} ($x_{\mathrm{a}} = \pm 1$, $x_{\mathrm{p}} = \pm 1$), which indicate whether the corresponding component is phase-inverted.

In the case of $x_{\mathrm{a}} = 1$ and $x_{\mathrm{p}} = 1$, (15) becomes equivalent to (14), which indicates that the same signal is fed to the main array and sub-array. In other cases, the signal for the sub-array differs from that for the main array, because the phase of some components is different in these two signals. To realize narrow-edged beamforming, the values of phase inversion switches should be determined properly. Detailed discussion on the determination of phase inversion switches is given in Section III-B.

B. Phase Inversion Strategy for Frequency Modulation

As shown by (15), the proposed method has two phase inversion switches: x_{a} controls the sign of the amplitude in signal $u_{\mathrm{sub}}(t)$, and x_{p} controls the sign of the phase offset in signal $u_{\mathrm{sub}}(t)$. The phase inversion patterns can be categorized into four cases: one case is regarded as a reference; one case can form a wider beam; and the other two cases can form a narrower beam, which correspond to our goal in this paper. A detailed description is as follows:

1) Reference – inverting none of the components ($x_{\mathrm{a}} = 1$, $x_{\mathrm{p}} = 1$): As described in Section III-A, the signal for sub-array $u_{\mathrm{sub}}(t)$ is equivalent to the signal for main array $u_{\mathrm{main}}(t)$ in this case, and can be expanded into the same expression with (10). The instantaneous frequency of signal $u_{\mathrm{sub}}(t)$ can be expressed with (9). This case can be regarded as a common FM-based PAL emitting the same frequency-modulated wave from all transducers. Therefore, this case is the reference case for comparison.

2) Inverting only the amplitude ($x_{\mathrm{a}} = -1$, $x_{\mathrm{p}} = 1$): In this case, the signal for the sub-array can be expanded as

$$\begin{aligned} u_{\mathrm{sub}}^{\mathrm{A}}(t) &= -A_{\mathrm{c}} \cos\left(2\pi f_{\mathrm{c}} t + 2\pi f_{\mathrm{f}} v(t)\right) \\ &= -A_{\mathrm{c}} \sum_{i=1}^{\infty} J_i(m_{\mathrm{FM}}) \cos\left(2\pi \left(f_{\mathrm{c}} + i f_{\mathrm{s}}\right) t\right) \\ &\quad - A_{\mathrm{c}} \sum_{i=1}^{\infty} (-1)^i J_i(m_{\mathrm{FM}}) \cos\left(2\pi \left(f_{\mathrm{c}} - i f_{\mathrm{s}}\right) t\right) \\ &\quad - A_{\mathrm{c}} J_0(m_{\mathrm{FM}}) \cos\left(2\pi f_{\mathrm{c}} t\right). \end{aligned} \tag{16}$$

Note that all the frequency components are phase-inverted compared with signal $u_{\mathrm{main}}(t)$, but the instantaneous frequency is the same as expressed (9). Based on the related studies [17], [22], we can predict that the beam will become wider than that in the reference case.

3) Inverting only the phase offset ($x_{\mathrm{a}} = 1$, $x_{\mathrm{p}} = -1$): In this case, the signal for the sub-array can be expanded as

$$\begin{aligned} u_{\mathrm{sub}}^{\mathrm{P}}(t) &= A_{\mathrm{c}} \cos\left(2\pi f_{\mathrm{c}} t - 2\pi f_{\mathrm{f}} v(t)\right) \\ &= A_{\mathrm{c}} \sum_{i=1}^{\infty} (-1)^i J_i(m_{\mathrm{FM}}) \cos\left(2\pi \left(f_{\mathrm{c}} + i f_{\mathrm{s}}\right) t\right) \\ &\quad + A_{\mathrm{c}} \sum_{i=1}^{\infty} J_i(m_{\mathrm{FM}}) \cos\left(2\pi \left(f_{\mathrm{c}} - i f_{\mathrm{s}}\right) t\right) \\ &\quad + A_{\mathrm{c}} J_0(m_{\mathrm{FM}}) \cos\left(2\pi f_{\mathrm{c}} t\right). \end{aligned} \tag{17}$$

Note that only the sideband components of the odd-number order are phase-inverted compared with signal $u_{\mathrm{main}}(t)$. More-

over, the instantaneous frequency becomes the following expression:

$$f_{\mathrm{ins}}^{\mathrm{P}}(t) = f_{\mathrm{c}} - f_{\mathrm{f}} s(t). \tag{18}$$

This case can be regarded as only inverting the sideband component in an amplitude-modulated wave when the high-order sideband components are ignored. Based on the conclusion in our previous work [22], this case is expected to form a narrower beam than that of the reference case.

4) Inverting both the amplitude and the phase offset ($x_{\mathrm{a}} = -1$, $x_{\mathrm{p}} = -1$): In this case, the signal for the sub-array can be expanded as

$$\begin{aligned} u_{\mathrm{sub}}^{\mathrm{AP}}(t) &= -A_{\mathrm{c}} \cos\left(2\pi f_{\mathrm{c}} t - 2\pi f_{\mathrm{f}} v(t)\right) \\ &= -A_{\mathrm{c}} \sum_{i=1}^{\infty} (-1)^i J_i(m_{\mathrm{FM}}) \cos\left(2\pi \left(f_{\mathrm{c}} + i f_{\mathrm{s}}\right) t\right) \\ &\quad - A_{\mathrm{c}} \sum_{i=1}^{\infty} J_i(m_{\mathrm{FM}}) \cos\left(2\pi \left(f_{\mathrm{c}} - i f_{\mathrm{s}}\right) t\right) \\ &\quad - A_{\mathrm{c}} J_0(m_{\mathrm{FM}}) \cos\left(2\pi f_{\mathrm{c}} t\right). \end{aligned} \tag{19}$$

Note that the sideband components of the even-number order and the carrier component are phase-inverted compared with signal $u_{\mathrm{main}}(t)$. Moreover, the instantaneous frequency becomes the following expression:

$$f_{\mathrm{ins}}^{\mathrm{AP}}(t) = f_{\mathrm{c}} - f_{\mathrm{f}} s(t). \tag{20}$$

This case can be regarded as only inverting the carrier component in an amplitude-modulated wave when the high-order sideband components are ignored. Based on the conclusion in our previous work [22], this case is expected to form a narrower beam than that of the reference case.

We here introduce a comparison of the time-domain waveforms in the above four cases in Fig. 2. The parameters are set as $A_{\mathrm{c}} = 1$, $A_{\mathrm{s}} = 1$, $f_{\mathrm{c}} = 40$ kHz, $f_{\mathrm{s}} = 4$ kHz, and $f_{\mathrm{f}} = 2$ kHz. The instantaneous frequency can be estimated using the Hilbert transform [28], and the analysis results are given in Fig. 3. The spectra were also estimated and the results are shown in Fig. 4. In Fig. 4, red dashed lines show the estimated amplitude spectra while blue solid lines show the estimated phase spectra. It can be noticed from the amplitude spectra that the frequency-modulated waves consist of carrier component and sideband components as expressed in (12), (16), (17) and (19). In the estimation of the phase spectra, circular-shift is carried out to prevent the original phase from being affected by the window function [29]. Although Fig. 3 shows several analysis errors, it can be seen that the analytical results on phases of components correspond to the theoretical expansion in (12), (16), (17) and (19).

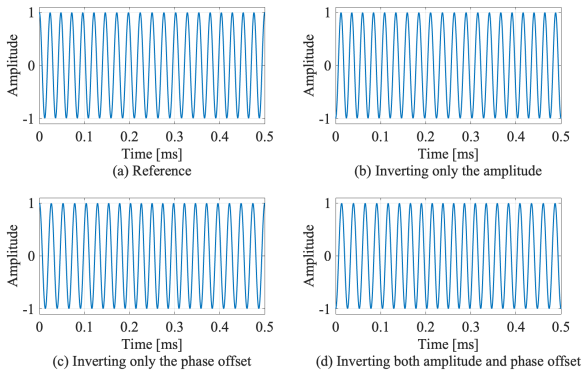


Fig. 2. Time-domain waveforms of the comparison cases.

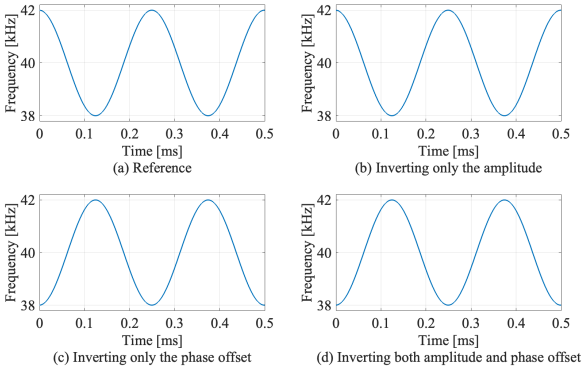


Fig. 3. Instantaneous frequencies of the comparison cases.

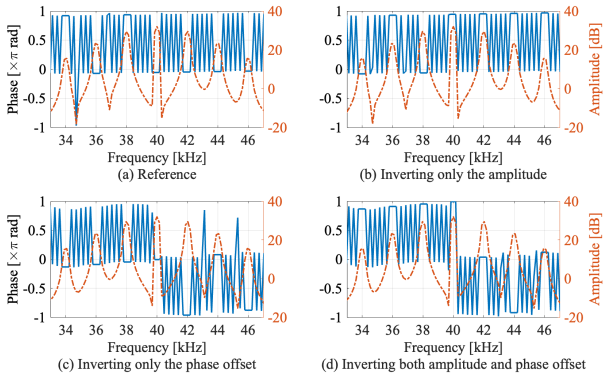


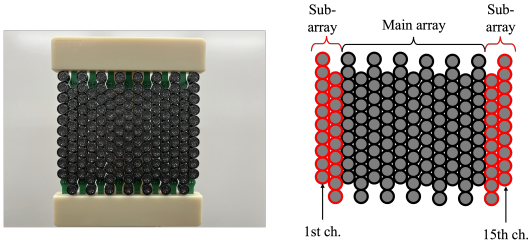
Fig. 4. Phase spectra and amplitude spectra of the comparison cases.

IV. Experiments and Results

A. Experimental Setup

We conducted experiments to demonstrate the effectiveness of the proposed method. We used a prototype PAL with a column-wise configuration of piezoelectric ultrasound transducers (MURATA, MA40S4S), which consists of 15 channels in total, as shown in Fig. 5. As in previous work [22], we assigned $M = 2$ or 3 channels on each side for the sub-array, and the remaining 11 or 9 channels on the interior for the main array. The PAL was set on a pan-rotation unit (Manfrotto, 300 N), by which the direction of the PAL could be rotated manually with an interval of 5 degs. to measure the directivity characteristics. Other conditions and equipment are shown in Table I and Table IV-A, respectively.

Time-stretched pulses (TSPs) [30] with a frequency range of



(a) Photograph of prototype PAL

(b) Transducer assignment (in case of $M = 2$)

Fig. 5. Column-wise prototype PAL used in the experiments.

TABLE I
EXPERIMENTAL CONDITIONS.

Temperature	20.2°C
Humidity	33.7%
Ambient noise level	L_{A} = 25.2 dB
Reverberation time	T_{60} = 650 ms
Sampling	96 kHz / 24 bits

TABLE II
EXPERIMENTAL EQUIPMENT.

D/A converter	RME, M-32 DA
Audio interface	RME, MADIface XT
A/D converter	RME, M-32 AD
Microphone amplifier	RME, OctaMic II
Microphone	SENNHEISER, MKH8020

1 to 8 kHz were used as target signal $s(t)$ in the experiments. The signal fed to the main array was generated based on (14), while the signal fed to the sub-array was generated based on (15). The parameters were set as $f_{\mathrm{c}} = 40$ kHz, and $f_{\mathrm{f}} = 2$ kHz. As discussed in Section III-B, the four cases of phase inversion for the signal fed to the sub-array, are case Ref., case A, case P and case A/P, respectively. Cases P and A/P were expected to form a narrower beam than that of case Ref.

The directivity of demodulated sound is known to depend on the frequency of the target signal [31], so we measured the frequency-dependent directivity in the experiments. The observed TSP signals were convolved with the inverse TSP signal to estimate the impulse response of demodulated sound in each direction. Analyzing with Fourier transform, we derived sound pressure $p_{\mathrm{idx}}(f_{\mathrm{s}}, \theta)$ in the θ direction when the target frequency was f_{s}, where idx = Ref., A, P, or A/P denotes the cases index. The on-axis ($\theta = 0$ deg.) sound pressures were expected to be different among the comparison cases [22]. Moreover, a PAL holds a frequency-dependent response of +12 dB/octave in theory [1]. To eliminate the influence of these factors, we calculated the frequency-dependent directivity by normalizing the sound pressure at each frequency, which can be given as:

$$d_{\mathrm{idx}}(f_{\mathrm{s}}, \theta) = p_{\mathrm{idx}}(f_{\mathrm{s}}, \theta) - p_{\mathrm{idx}}(f_{\mathrm{s}}, 0). \tag{21}$$

B. Experimental Results and Discussions

Figs. 6 and 7 show the measured frequency-dependent directivity characteristics when $M = 2$ and 3, respectively. It can be seen that cases Ref. and A achieved significantly

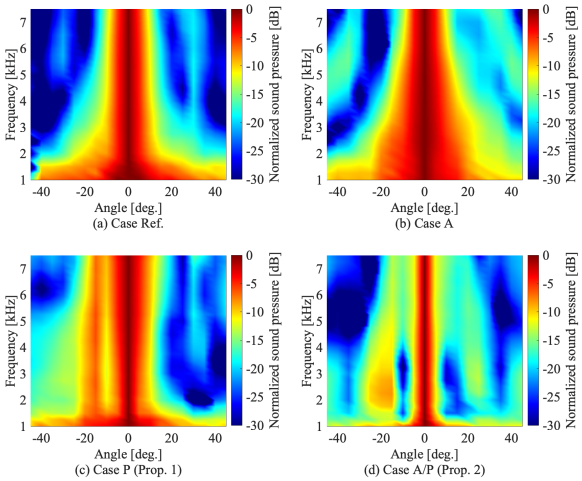


Fig. 6. Measured frequency-dependent directivity characteristics (M=2).

frequency-dependent directivity, the beam becoming narrower as frequency f_{s} increased. In contrast, cases P and A/P achieved an almost constant beamwidth in the main lobe, despite having larger side lobes. It can be derived that the proposed method (cases P and A/P) has an advantage over the reference (case Ref.) when target frequency f_{s} is lower, because the proposed method can form a much narrower beam than can the reference case. However, when target frequency f_{s} increased, the difference in beamwidth became smaller between the proposed method and the reference case. Comparing Fig. 6 with Fig. 7 shows that the beam of case A becomes wider as M increases. In the results of cases P and A/P, the main lobe remains the same or becomes narrower while the side lobes increase as M increases.

The results of the reference case correspond to the theoretical model of Westervelt's directivity [24], [32], [31], while the others can be explained by the phase cancellation in ultrasound and audible sound. Especially, phase cancellation as used in the proposed method contributes to a narrow beam at low-frequencies.

Fig. 8 shows the measured directivity characteristics when the target frequency f_{s} is 2 kHz. Table III shows the corresponding measured results at a beamwidth of -6 dB. It can be easily observed that cases P and A/P achieved a much narrower main lobe than did case Ref., while case A achieved a wider beam than did case Ref. However, cases P and A/P achieved more significant side lobes than did case Ref. Comparison of Fig. 8(a) with (b) shows that the directivity of the proposed method (cases P and A/P) becomes narrower while the sound pressure becomes lower as M increases.

The above conclusions generally correspond to those of phase inversion for AM-based PAL [22] and can be explained by the enhancement of phase cancellation because more channels are assigned to the sub-array.

The on-axis ($\theta = 0$ deg.) relative sound pressures of these results are listed in Table IV, where case Ref. was set as 0

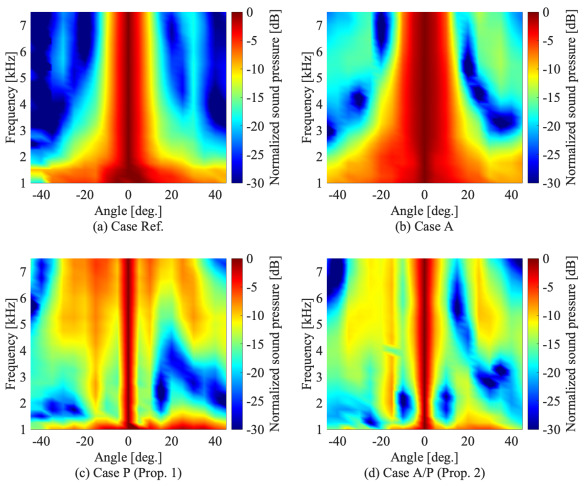


Fig. 7. Measured frequency-dependent directivity characteristics (M=3).

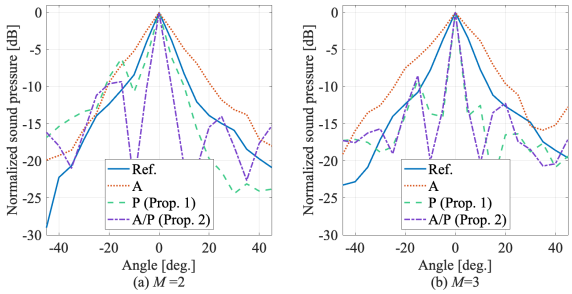


Fig. 8. Measured directivity characteristics at 2 kHz.

TABLE III
MEASURED HALF-POWER DIRECTIVITY ANGLE AT 2 KHZ.

Case	-6dB beamwidth [deg.]	
	$M = 2$ ch.	$M = 3$ ch.
Ref.	15.8	15.8
A	24.9	28.4
P (Prop. 1)	9.6	4.3
A/P (Prop. 2)	5.8	4.6

dB. It can be derived that the phase inversion for FM-based PAL decreases the sound pressure by a very small amount. The phase inversion for AM-based PAL causes a decrease in sound pressure of over 10 dB at 2 kHz [22], while the proposed phase inversion for FM-based PAL only causes a decrease of about 3 dB. Therefore, our results demonstrate that the proposed method can achieve an equivalent level of narrow directivity with a much lower reduction of sound pressure compared with that of the phase inversion for AM-based PAL.

V. CONCLUSIONS

In this paper, we propose a narrow-edged beamforming method based on phase inversion for FM-based PAL. The phase inversion strategy is developed to realize phase cancellation in both ultrasound and audible sound, aiming to form a narrow beam of demodulated sound. The experimental

TABLE IV
MEASURED RELATIVE VALUES OF THE ON-AXIS SOUND PRESSURE AT 2 KHZ.

Case	Relative sound pressure [dB]	
	$M = 2$ ch.	$M = 3$ ch.
Ref.	0	0
A	-0.4	-1.3
P (Prop. 1)	-0.3	-3.4
A/P (Prop. 2)	-1.6	-3.3

results demonstrated the effectiveness of the proposed method, which shows that the proposed method can achieve a level of narrow directivity equivalent to that of the phase inversion for AM-based PAL. Moreover, the decrease of sound pressure is mitigated significantly, which is one of the main drawbacks of the phase inversion for AM-based PAL. In the future, we will work on the elimination or the exploitation of the side lobes for the proposed method.

ACKNOWLEDGMENT

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP21H03488, JP21H04427, and JP23H03425.

REFERENCES

[1] W. S. Gan, J. Yang, and T. Kamakura, "A review of parametric acoustic array in air," *Applied Acoustics*, vol. 73, no. 12, pp. 1211–1219, 2012.
[2] M. Yoneyama, J. Fujimoto, Y. Kawamo, and S. Sasabe, "The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design," *The Journal of the Acoustical Society of America*, vol. 73, no. 5, pp. 1532–1536, 1983.
[3] T. Kamakura, M. Yoneyama, and K. Ikegaya, "Developments of parametric loudspeaker for practical use," in *Proc. 10th International Symposium on Non-linear Acoustics, 1984*, 1984, pp. 147–150.
[4] A. Thuras, R. Jenkins, and H. O'Neil, "Extraneous frequencies generated in air carrying intense sound waves," *The Journal of the Acoustical Society of America*, vol. 6, no. 3, pp. 173–180, 1935.
[5] F. J. Pompei, "Sound from ultrasound: The parametric array as an audible sound source," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.
[6] Y. Geng, M. Nakayama, and T. Nishiura, "Demodulated sound quality improvement for harmonic sounds in over-boosted parametric array loudspeaker," *Applied Acoustics*, vol. 186, p. 108460, 2022.
[7] S. Aoki, M. Toba, and N. Tsujita, "Sound localization of stereo reproduction with parametric loudspeakers," *Applied Acoustics*, vol. 73, no. 12, pp. 1289–1295, 2012.
[8] K. Tanaka, C. Shi, and Y. Kajikawa, "Binaural active noise control using parametric array loudspeakers," *Applied Acoustics*, vol. 116, pp. 170–176, 2017.
[9] Y. Ogami, M. Nakayama, and T. Nishiura, "Virtual sound source construction based on radiation direction control using multiple parametric array loudspeakers," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1314–1325, 2019.
[10] Y. Geng, S. Sayama, M. Nakayama, and T. Nishiura, "Movable virtual sound source construction based on wave field synthesis using a linear parametric loudspeaker array," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, p. e14, 2023.
[11] U. Sayin, P. Artís, and O. Guasch, "Realization of an omnidirectional source of sound using parametric loudspeakers," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 1899–1907, 2013.
[12] D. Ikefuji, M. Nakayama, T. Nishiura, and Y. Yamashita, "Robust sound image localization for moving listener with curved-type parametric loudspeaker," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC 2015)*. IEEE, 2015, pp. 1045–1049.

[13] M. Arnela, O. Guasch, P. Sánchez-Martín, J. Camps, R. M. Alsina-Pagès, and C. Martínez-Suquía, "Construction of an omnidirectional parametric loudspeaker consisting in a spherical distribution of ultrasound transducers," *sensors*, vol. 18, no. 12, p. 4317, 2018.

[14] W. S. Gan, J. Yang, K. S. Tan, and M. H. Er, "A digital beamsteerer for difference frequency in a parametric array," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 3, pp. 1018–1025, 2006.

[15] C. Shi, Y. Kajikawa, and W. S. Gan, "Generating dual beams from a single steerable parametric loudspeaker," *Applied Acoustics*, vol. 99, pp. 43–50, 2015.

[16] C. Shi, R. Bai, J. Gou, and J. Liang, "Multi-beam design method for a steerable parametric array loudspeaker," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC 2020)*. IEEE, 2020, pp. 416–420.

[17] T. Kamakura, S. Sakai, H. Nomura, and M. Akiyama, "Parametric audible sounds by phase-cancellation excitation of primary waves," *Acoustics' 08 in Paris*, pp. 4365–4370, 2008.

[18] C. Hedberg, K. Haller, and T. Kamakura, "A self-silenced sound beam," *Acoustical Physics*, vol. 56, no. 5, pp. 637–639, 2010.

[19] H. Nomura, C. M. Hedberg, and T. Kamakura, "Numerical simulation of parametric sound generation and its application to length-limited sound beam," *Applied Acoustics*, vol. 73, no. 12, pp. 1231–1238, 2012.

[20] E. Skinner, M. Groves, and M. K. Hinders, "Demonstration of a length limited parametric array," *Applied Acoustics*, vol. 148, pp. 423–433, 2019.

[21] Y. Geng, H. Wang, M. Nakayama, and T. Nishiura, "Narrow-edged beamforming using masked parametric array loudspeakers," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC 2021)*. IEEE, 2021, pp. 968–973.

[22] Y. Geng, M. Nakayama, and T. Nishiura, "Narrow-edged beamforming based on individual phase inversion in amplitude-modulated wave for parametric array loudspeaker," *Applied Acoustics*, vol. 200, p. 109060, 2022.

[23] Y. Hatano, C. Shi, and Y. Kajikawa, "Compensation for nonlinear distortion of the frequency modulation-based parametric array loudspeaker," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1709–1717, 2017.

[24] P. J. Westervelt, "Parametric acoustic array," *The Journal of the Acoustical Society of America*, vol. 35, no. 4, pp. 535–537, 1963.

[25] C. Shi and Y. Kajikawa, "A comparative study of preprocessing methods in the parametric loudspeaker," in *2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC 2014)*. IEEE, 2014, pp. 1–5.

[26] M. Nakayama and T. Nishiura, "Synchronized amplitude-and-frequency modulation for a parametric loudspeaker," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC 2017)*. IEEE, 2017, pp. 130–135.

[27] Y. Geng, K. Asahi, M. Nakayama, and T. Nishiura, "Demodulated sound pressure enhancement based on frequency-modulation utilizing square-wave carrier for small parametric array loudspeaker," in *Proceedings of the 24th International Congress on Acoustics (ICA 2022)*, 2022, p. 0210.

[28] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, 1992.

[29] B. T. Nguyen, Y. Wakabayashi, K. Iwai, and T. Nishiura, "Inter-frequency phase difference for phase reconstruction using deep neural networks and maximum likelihood," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1667–1680, 2023.

[30] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1484–1488, 1981.

[31] C. Shi and R. Bai, "Design of a constant beamwidth beamformer for the parametric array loudspeaker," in *Proceedings of the 23rd International Congress on Acoustics (ICA 2019)*, 2019, pp. 3077–3081.

[32] K. S. Tan, W. S. Gan, J. Yang, and H. E. Meng, "Constant beamwidth beamformer for difference frequency in parametric array," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5. IEEE, 2003, pp. V–361.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Corpus Construction for Deaf Speakers and Analysis by Automatic Speech Recognition

Akio Kobayashi* and Keiichi Yasu†
* Yamato University, Japan
E-mail: akiok@ieee.com
† Tsukuba University of Technology, Japan
E-mail: k-yasu@a.tsukuba-tech.ac.jp

***Abstract*—This study explores automatic speech recognition (ASR) for the deaf and hard-of-hearing. Despite the recent progress in ASR for dysarthric speakers, existing research primarily focuses on people with motor speech disorders. Thus, the effect of speech diversity on the performance of ASR is not considered for ambiguous deaf speech owing to a lack of auditory feedback. Therefore, we compiled a corpus of speech of many profoundly deaf speakers to compare the ASR performance with that of normal-hearing speakers. The performance analysis is reported through a set of phoneme recognition experiments. Furthermore, we show that additional phonological features that reflect deaf speakers' articulation can improve performance in phoneme recognition for deaf speech.**

I. Introduction

In communication between normal-hearing and deaf or hard-of-hearing people, accessibility supports, such as written conversation, sign language, and automatic speech recognition (ASR), are commonly used. Notably, it has been reported that only 25 % of deaf people in Japan use sign language daily, and more than half of them communicate using spoken language. Moreover, smartphones and tablets have recently diversified the applications of ASR, such as through interactive information retrieval and the control of electrical appliances through Internet of Things (IoT) devices. Hence, in addition to considering the disparity in ASR-based operations as an urgent social issue, considering accessibility and inclusion for deaf people is necessary.

In recent years, many ASR studies have been on atypical speakers, such as those with speech disorders [1]–[6]. Many such studies deal with motor speech disorders; as part of these studies, hearing impairment is often included as a subject. In [1], [2], the ASR performance was improved by multitask learning of voice conversion from disorders of a normal-hearing speaker and ASR of speech of people with disorders. In [3], [4], the authors employed articulatory motion data collected using electromagnetic articulography. Then, they combined acoustic and articulatory features to improve ASR performance. Because it is generally difficult to collect a large amount of speech data from people with speech disorders, data augmentation methods have been proposed to improve performance [5], [6].

Many studies mentioned above have focused on dysarthric people, i.e., people with motor speech disorders caused by amyotrophic lateral sclerosis (ALS) and other disorders. These studies focused on various speech disorders but have not intensively analyzed the speech of deaf people. It is challenging to focus on ASR for deaf speakers because of their small population. This means the details of the ASR performance of deaf speech remain unclear. This study first attempts to construct a corpus to examine the ASR performance for deaf speakers. The differences in performance caused by the diversity of their speech were clarified through a series of phoneme recognition experiments. From a phonological perspective, segmental units, such as phonemes, can be further classified based on features like articulation. For example, [7] showed that deaf speakers' articulation has different phonological characteristics than normal-hearing people. If neural networks can capture such characteristics of deaf speakers, ASR performance would be significantly improved.

II. Deaf Speech and Corpus

A. Deaf Speech

A review of deaf speech based on phonetics and phonology, particularly concerning articulation errors, can be found in [8]. [8] reported that substitutions, deletions, and insertions of phonemes are frequently observed in deaf speakers' articulation, especially as substitutions in place and manner of articulation for consonants. In a recent study [7], the speech of cochlear implant users was analyzed and compared with that of normal-hearing speakers for use in rehabilitation. Similar substitutions are observed in front- and close-vowels. The vowel space of the deaf speaker is generally narrower than that of normal-hearing speakers, and the formants (F_1, F_2) that characterize vowels tend to fluctuate significantly [9]. The cause of these errors is the difficulty in acquiring articulation owing to a lack of auditory feedback [10].

B. Deaf Speech Corpus

It is essential to construct a speech corpus for deaf speakers. A corpus targeting deaf speakers was described in [11], which gives a set of detailed metadata, including hearing abilities and communication methods in addition to speech. However, [11] is a corpus for analysis from a phonetic perspective and is not designed for spoken language processing.

In the field of spoken language processing, there are many speech corpora for individuals with normal-hearing. Even for the Japanese, there are many widely used corpora, such as the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

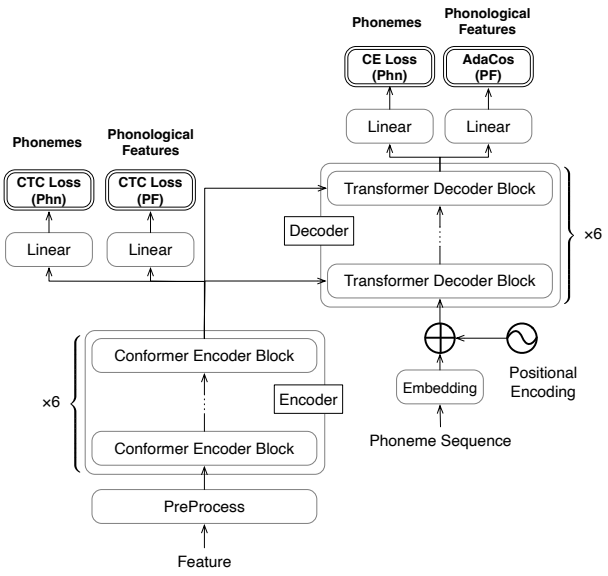


Fig. 1. Neural network and loss functions

Corpus of Spontaneous Japanese (CSJ) [12], ASJ Continuous Speech Corpus for Research [13], and Japanese Newspaper Article Reading Corpus (JNAS) [14]. Some of these corpora contain a set of phoneme-balanced sentences, which is the de facto standard for Japanese spoken language processing. Thus, we must develop a corpus that can compare ASR performance with normal-hearing while collecting standard sentences spoken by deaf speakers. Therefore, the authors are eager to collect a variety of utterances from deaf speakers [15]. The corpus includes phoneme-balanced sentences, utterances from JNAS newspaper articles, and spontaneous speech as conversations. During compilation, the corpus comprised the speeches of 46 deaf students.

III. Phonological Features

A. Phonemes and Articulation

A phoneme is the smallest unit used to distinguish a word from others; hence it is convenient for investigating the error tendency in deaf speech. Alternatively, each phoneme can be characterized by articulation in terms of phonology. Phonological features are smaller units than phonemes used to capture features spanning among the phonemes. For example, comparing the phonemes /s/ with /z/, both are alveolar fricatives but are in conflict for voicing. For example, phonological features have often been introduced when evaluating the pronunciation of L2 speakers [16], [17]. Therefore, using phonological features allows phonemes to be evaluated from other measures and, as a result, may provide robust discrimination of phonemes in ambiguous deaf speech.

B. Phonological Features

The phonological features in this study are defined as multiple classes that characterize phonemes. Therefore, phonemes and classes representing phonological features are output symbols in a neural network that performs phoneme recognition using articulation features. First, phonological features are divided into five groups; any feature in an opposing relationship within a group is output. However, if none of the features in the group are activated, the symbol none is output. The five groups of articulatory features are as follows:

i. Voice activity features
   This group includes three classes—vowels, consonants, and silence.

ii. Voicing features
   This group includes two categories—voiced and unvoiced.

iii. Vowel features
   This group includes the following three categories
   a. Rounding features
      This group represents the degree of lip rounding, i.e., rounded and unrounded.
   b. Height features
      This group represents the height of vowels or tongue. The group has four categories: close, close-mid, open-mid, and open. These are defined by the International Phonetic Association (IPA) [18].
   c. Backness features
      This group represents the position of the tongue. This group has three categories: front, central, and back tongue. The IPA also defines these categories.

iv. Consonant features
   The following two groups characterize consonants:
   a. Place-of-articulation features
      This group represents the 13 places of articulation of consonants, e.g., bilabial, dental, alveolar, velar, uvular, and alveolo-palatal.
   b. Manner-of-articulation features
      This group represents the nine manners of articulation for consonants, e.g., nasal, plosive, tap/flap, fricative, and affricate.

v. Other features
   These features include palatalization and devoicing.

Since the phoneme environment is affected by coarticulation, it is impossible to determine the articulatory features uniquely. However, in this study, one typical feature is used for simplicity.

C. Loss Functions

For phonological features defined as multiclass, the loss functions of the encoder/decoder of the neural network are determined. As described later, we employ a network structure based on the conformer–transformer model. We use the CTC loss functions [19] for the encoder outputs and AdaCos for the decoder outputs. AdaCos[20] is a function employed in deep metric learning. Therefore, it can effectively improve

performance while correcting the discriminative hyperplanes among phonemes with phonological features.

IV. Experiments

In the experiments, we used a relatively small-scale neural network with conventional structures to investigate the differences in ASR performance between normal hearing and deaf speakers in a preliminary study. The deaf speech was first evaluated using a neural network trained with only normal-hearing speech, referred to as the *NH network*. Then, the NH network parameters were fine-tuned using deaf speech to evaluate the speech, referred to as the *DF network*.

A. Training and Test Data

To evaluate deaf speech, we used 699 phonetically balanced sentences spoken by 14 students (nine males and five females) from the Tsukuba University of Technology. Specifically, the B set, composed of 50 sentences from ATR phonetically-balanced sentences, was used[21]. For comparison, 5,996 phonetically-balanced sentences from 120 normal-hearing speakers from JNAS were used. 77.9k (123.7 h) balanced sentences of normal-hearing from ASJ, JNAS, and ATR multispeaker speech databases were used to train the NH network. As the fine-tuned DF network, we used 3,525 balanced sentences (6.3 h) spoken by the abovementioned 14 students.

B. Neural Networks

We employed encoder-decoder networks with a Conformer [22] as the encoder and a Transformer [23] as the decoder (see Fig. 1). The encoder comprised six conformer blocks with 512 input dimensions, eight heads of multi-head attention for convolution modules, and 1024 dimensions for feed-forward modules. The kernel size of the depth-wise convolution block was set to 31. The decoder comprised six transformer blocks with similar configurations for input dimensions, number of heads, and dimensions for feed-forward modules. Convolution layers for preprocessing were prepended to the encoder, reducing the temporal length to 1/4, similar to [22]. We used four convolution layers, and their kernel sizes were set to 3 for 256 channels. The Adam optimizer was used to train the network, and 50 epochs were configured with a learning rate of 4.0e-5.

We conducted experiments with different configurations of the four loss functions listed in Table I for the ablation studies.

TABLE I
Network configurations

	Encoder Loss		Decoder Loss	
	CTC (Phn)	CTC (PF)	CE (Phn)	AdaCos (PF)
Baseline			✓	
CTC	✓		✓	
Proposed-A			✓	✓
Proposed-B		✓	✓	✓
Proposed-C	✓		✓	✓
Proposed-D	✓	✓	✓	✓

TABLE II
Overall results for normal-hearing on NH networks (%)

	PER	Sub.	Del.	Ins
Baseline	8.9	5.4	1.7	1.8
CTC	7.3	4.5	1.3	1.5
Proposed_A	7.5	4.6	1.5	1.4
Proposed_B	7.7	4.8	1.8	1.1
Proposed_C	7.1	4.6	1.4	1.1
Proposed_D	**6.0**	**4.0**	**1.2**	**0.9**

TABLE III
Overall results for deaf speech on NH networks (%)

	PER	Sub.	Del.	Ins.
Baseline	41.6	24.4	4.9	12.3
CTC	38.8	**23.2**	5.0	10.6
Proposed_A	37.9	23.6	4.9	9.5
Proposed_B	39.9	24.5	4.3	11.1
Proposed_C	37.5	23.5	4.8	9.2
Proposed_D	**36.9**	23.6	**4.2**	**9.0**

C. Evaluation on normal-hearing(NH) Network

First, we conducted experiments on NH networks trained with only normal-hearing data. Table II lists the phoneme error rates (PERs) for 120 normal-hearing subjects, and Table III lists the PERs for the 14 deaf subjects. The tables include the substitution, deletion, and insertion rates. As shown in Table II, for the normal-hearing, **Proposed_C** and **D** reduced the PERs against **CTC**, with relative reduction rates of 2.7 % and 17.8 %. According to the matched-pair test, these were statistically significant ($p < 0.05$). However, **Proposed_A** and **B** that were trained without CTC loss did not achieve **CTC** results. The results show that the CTC loss of phonemes for the encoder is effective for gaining performance and that the loss functions of phonological features play supplemental roles in network training.

Table 2 shows that the reduction in PER for deaf speakers is limited, while it is statistically significant at $p < 0.05$ for **Proposed_C** and **D**. It is because deaf speech differs significantly from normal-hearing speech. The detailed results show that insertion errors are reduced, whereas substitution errors are not reduced. We found fewer vowel insertions in the results. Introducing phonological features would lead to robustness in suppressing articulatory fluctuations of vowels.

D. Evaluation on deaf(DF) Network

Next, we conducted experiments employing DF networks, whose parameters were fine-tuned to those of NH networks with deaf speech. The DF network parameters were updated with the speech of 13 speakers so that the evaluated speakers were excluded, i.e., in a *leave-one-speaker-out* manner. Therefore, the PER of the evaluated speaker would be reduced if the neural network reflected the speech characteristics common to deaf speakers. Table IV lists the PERs. In the configuration of **Baseline**, only a small error reduction from its counterpart of the DF network was achieved. When adding CTC loss functions, PER was reduced from 38.8 % (NH) to 34.1 %, achieving a 12.3 % relative reduction. The additional CTC

TABLE IV
OVERALL RESULTS FOR DEAF ON DF NETWORK (%)

	PER	Sub.	Del.	Ins.
Baseline	41.0	23.9	9.8	7.3
CTC	34.1	21.5	7.0	5.6
Proposed_A	35.2	22.2	6.9	6.1
Proposed_B	35.6	22.6	6.7	6.3
Proposed_C	33.7	21.7	6.5	5.5
Proposed_D	**31.4**	**20.2**	**6.3**	**4.9**

TABLE V
VOWEL/CONSONANT ERROR RATES ON DF NETWORK (%)

	Vowels	Consonants
CTC(NH)	30.7	47.1
CTC(DF)	25.4	43.2
Proposed_A	26.6	44.1
Proposed_B	26.1	45.4
Proposed_C	24.6	43.0
Proposed_D	23.0	40.1

loss on the encoder effectively improved the performance for deaf speech and normal-hearing [24]. **Proposed_D**, which is a network using all loss functions, reduced the PER from 36.9 % to 31.4 %, achieving a 14.9 % relative reduction. The network reduced the substitution errors, which were difficult to achieve in its counterpart NH network.

The PERs for the vowels and consonants are listed in Table V. Compared with the **CTC(DF)** results, the error reduction rate for vowels in **Proposed_D** was 9.4 %, whereas that for consonants was 7.2 %. Generally, vowels are less complicated to articulate than consonants, even for deaf speakers. Thus, the phonological features compensated for the narrowness of their vowel space to reduce vowel errors.

Next, we investigated error reduction for consonants. Table VI lists the top five error-reduction rates for consonant errors, comparing **CTC(NH)** and **CTC(DF)**. The table shows that nasal /N/ achieved the highest reduction rate. Nasalization is one of the characteristics of deaf speech and commonly appears within it. /w/ and /y/ are approximants and close to vowels in articulation, so they probably benefit from error reduction, similar to vowels.

Comparing **CTC(DF)** and **Proposed_D**, the voiced plosive /b/achieved the highest error reduction (Table VII). This reduction is because the confusion with the nasal /m/ is resolved. The nasal /n/ and approximants /w/ and /y/ were also recovered from errors.

V. CONCLUSION

We described automatic speech recognition for deaf speakers. We showed that phonological features could capture characteristics common to deaf speakers and significantly reduce phoneme errors.

The corpus is still being developed due to the difficulty of collecting deaf speech, especially from communication with deaf people. In the future, we intend to assemble more speech recordings, including spoken dialogues from over 100 deaf speakers, and to make them widely available to the public.

TABLE VI
TOP-5 ERROR-REDUCED CONSONANTS: CTC(NH) TO CTC(DF)

	CTC(NH)	CTC(DF)	reduction
/N/	51.2	37.4	26.9
/w/	32.8	24.4	25.7
/y/	66.0	50.5	23.5
/s/	63.4	50.7	20.0
/ts/	86.5	70.4	18.7

TABLE VII
TOP-5 ERROR-REDUCED CONSONANTS: CTC(DF) TO PROPOSED_D(DF)

	CTC(DF)	Proposed_D(DF)	reduction
/b/	45.6	37.9	16.9
/y/	50.5	42.4	16.0
/n/	29.9	26.0	13.0
/w/	24.4	21.5	12.0
/t/	28.6	25.3	11.6

ACKNOWLEDGEMENT

This study was supported by JSPS KAKENHI (grant number JP23H00995, JP20H01716 and JP21H00901).

REFERENCES

[1] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech*, 2019, pp. 4115–4119.

[2] R. Doshi, Y. Chen, L. Jiang, *et al.*, "Extending Parrotron: An end-to-end, speech conversion and speech recognition model for atypical speech," in *Proc. ICASSP*, 2021, pp. 6988–6992.

[3] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition," in *Proc. ICASSP*, 2022, pp. 7372–7376.

[4] S. Hu, S. Liu, X. Xie, *et al.*, "Exploiting cross domain acoustic-to-articulatory inverted features for disordered speech recognition," in *Proc. ICASSP*, 2022, pp. 6747–6751.

[5] L. Wu, D. Zong, S. Sun, and J. Zhao, "A sequential contrastive learning framework for robust dysarthric speech recognition," in *Proc. ICASSP*, 2021, pp. 7303–7307.

[6] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *Proc. ICASSP*, 2021, pp. 6428–6432.

[7] T. Arias-Vergara, J. R. Orozco-Arroyave, M. Cernak, S. Gollwitzer, M. Schuster, and E. Nöth, "Phone-Attribute Posteriors to Evaluate the Speech of Cochlear Implant Users," in *Proc. Interspeech 2019*, 2019, pp. 3108–3112.

[8] M. J. Osberger and N. S. McGarr, "Speech production characteristics of the hearing impaired," in *Speech and Language*, vol. 8, Elsevier, 1982, pp. 221–283.

[9] R. B. Monsen, "Normal and reduced phonological space: The production of English vowels by deaf adolescents," *Journal of Phonetics*, vol. 4, no. 3, pp. 189–198, 1976.

[10] A. C. Coelho, D. M. Medved, and A. G. Brasolotto, "Hearing loss and the voice," *An update on hearing loss*, pp. 103–128, 2015.

[11] L. L. Mendel, S. Lee, M. Pousson, *et al.*, "Corpus of deaf speech for acoustic and speech production research," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, EL102–EL107, 2017.

[12] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, 2000.

[13] T. A. S. of Japan, *ASJ continuous speech corpus for research (ASJ-JIPDEC)*, https://doi.org/10.32130/src.ASJ-JIPDEC, 2007.

[14] T. A. S. of Japan, *ASJ Japanese newspaper article sentences read speech corpus (JNAS)*, https://doi.org/10.32130/src.JNAS.

[15] A. Kobayashi, K. Yasu, H. Nishizaki, and N. Kitaoka, "Corpus design and automatic speech recognition for deaf and hard-of-hearing people," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 17–18.

[16] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, "Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information," in *Proc. Interspeech*, 2016, pp. 2686–2690.

[17] S. Naijo, A. Ito, and T. Nose, "Improvement of automatic English pronunciation assessment with small number of utterances using sentence speakability," in *Proc. Interspeech*, 2021, pp. 4473–4477.

[18] I. P. Association, *The international phonetic alphabet*, https : / / www . internationalphoneticassociation . org / IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf, 2020.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[20] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "Adacos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 823–10 832.

[21] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[22] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[23] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

Ensemble of Transformer and Convolutional Recurrent Neural Network for Improving Discrimination Accuracy in Automatic Chord Recognition

Hikaru Yamaga*, Toshifumi Momma†, Kazunori Kojima*, Yoshiaki Itoh*
* Iwate Prefectural University, Japan, E-mail: y-itoh@iwate-pu.ac.jp
† Graduated from Iwate Prefectural University, Japan

***Abstract*—Automatic chord recognition is a task of recognizing and transcribing chords from music data such as popular music. Manual chord transcription requires highly technical knowledge and great effort. A chord is a typical musical feature. Realization of automatic chord recognition can enable their use for many purposes such as musical notation and structural analysis. For this reason, automatic chord recognition has become a major research task in the field of music information retrieval. In recent years, automatic chord recognition has widely used deep learning models. Convolutional Recurrent Neural Network (CRNN) and Transformer have achieved high accuracy. For this study, we focus on the differences in feature extraction approaches used by CRNN and Transformer, and propose an ensemble learning method using the two models. Additionally, we adopt an original overlap inference method to improve their accuracy by complementing the lack of temporal information. Results show that we achieved average accuracy of 78.92% under the traditional evaluation metrics, which are, respectively, 1.64% and 2.43% higher than those of CRNN and Transformer.**

I. INTRODUCTION

When playing a musical instrument, a chord is essential information: a symbol expressing the musical scale used in each section. In recent years, chord information for various musical pieces has become accessible on the internet, but in many cases, chord information for new music or independent music is not publicly available. Although transcribing chords from the music by oneself is necessary, this process requires highly technical knowledge and great effort. Therefore, the task of automatic chord recognition (ACR), which recognizes and transcribes a series of chords from music data, has been studied actively in the research field of music information processing. A chord is a typical musical feature. Chord information obtained by ACR is useful for many purposes such as music notation and music structure analysis.

Most traditional ACR systems consist of feature extraction and chord sequence decoding. Among input features, chromagram [3, 4, 5] is widely used: it reflects the 12 different pitch classes. Other attempts at handcrafted features have been made, such as chroma vectors and tonnetz [6]. The most common sequence decoding method is based on Hidden Markov Model (HMM) [7, 8]. Dynamic Bayesian Network (DBN) is also used for decoding [3, 5].

Recently, deep learning has been developed and used in various fields [9]. In ACR, Convolutional Neural Network (CNN) [10, 11] and Recurrent Neural Network (RNN) [12, 13] have been used particularly. Moreover, Vaswani et al. proposed Transformer, which is based solely on Attention mechanisms, dispensing entirely with recurrence and convolutions [14]. Transformer was first proposed for application to a machine translation task, but in recent years, it has been applied to various domains because of its high performance and the very low training costs. In the field of music, Huang et al. applied it to a music generation task and obtained state-of-the-art results [15].

In the dataset used for this study, Jiang et al. used the Convolutional Recurrent Neural Network (CRNN) [1] and recorded 67.09% in tetrads. Park et al. used Transformer [2] and recorded 66.00% in tetrads. A tetrad is an evaluation metric that targets all chords. However, it is not easy to infer chords that theoretically include more than 4,000 types. Therefore, some room remains for improving accuracy.

For this study, we focus on the differences of feature extraction approaches between CRNN with convolutional filter and Transformer with Self-Attention mechanism, and propose an ensemble learning method using these two models and our own neural network (NN) model. Additionally, we adopt an original overlap inference method to improve accuracy by complementing the lack of temporal information in inference.

The remainder of this paper is organized as follows: Section 2 explains chords in the field of music. Section 3 introduces two earlier studies. Section 4 presents a description of ensemble learning and overlap inference proposed herein. Section 5 presents results of experiments using the proposed method. Finally, the conclusion is given in Section 6.

II. CHORD

In the field of music, a chord is a sound that is formed by multiple and different overlapping notes. Generally, a chord comprises three or more notes. However, some notes might be omitted. Chords might consist of two or fewer notes.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

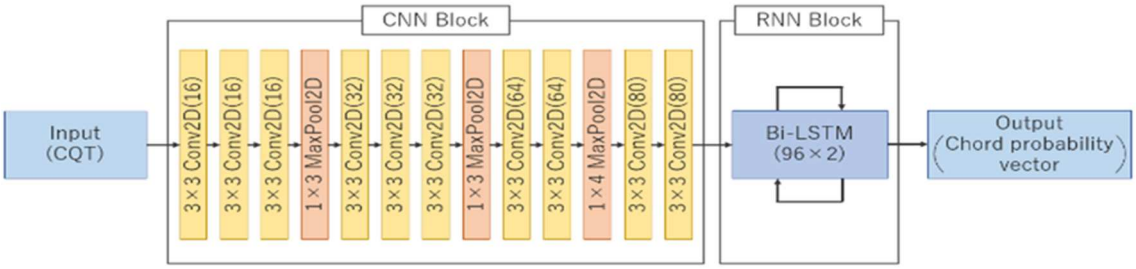


Fig. 1 Model architecture of the CRNN by Jiang et al. [1].

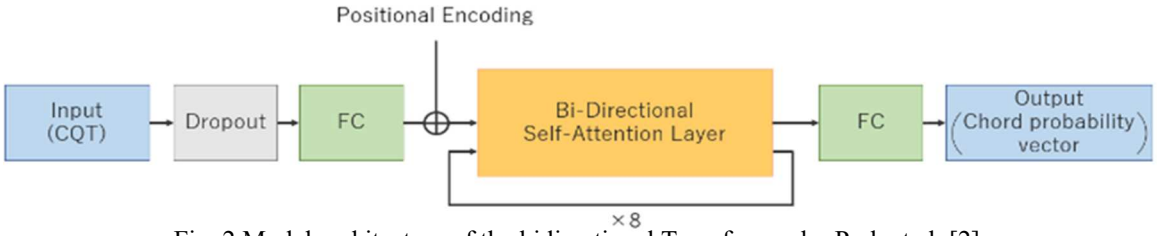


Fig. 2 Model architecture of the bidirectional Transformer by Park et al. [2].

Table 1 Pitch and the classification in the music theory

Semitones	Pitch	Classification
0	P1, 12	Root
+1	11, m2	-
+2	M2, 13	-
+3	12, m3	Third
+4	M3, 14	Third
+5	13, P4	-
+6	14, 15	Fifth
+7	P5, 16	Fifth
+8	15, m6	Fifth
+9	M6, 17	Sixth, Seventh
+10	16, m7	Seventh
+11	M7, 18	Seventh
+12	17, P8	(Root)

A chord consists of a single note (Root) over multiple notes (Quality). Various variations exist for chords, such as triads and tetrads. Each note in the Quality is interpreted in terms of its musical distance relative to the Root, based on the pitch relations shown in Table 1. The type of chord is determined by the pitch relations of all notes in the chord. In popular music, chords called triads or tetrad are mainly used. In this paper, ACR discriminates evaluation metrics of seven types.

A triad is a chord made up of one note from the Root, Third, and Fifth in the classification shown in Table 1, which is a predetermined combination. For example, the major chord consists of P1 (Root), M3 (Third), and P5 (Fifth). The minor chord consists of P1 (Root), m3 (Third), and P5 (Fifth).

A tetrad is a chord made by depositing a single note classified as a Sixth or Seventh on the Triad. For example, the M7 chord consists of P1 (Root), M3 (Third), P5 (Fifth), and M7 (Seventh). The m7 chord consists of P1 (Root), m3 (Third), P5 (Fifth) and m7 (Seventh).

III. RELATED WORK

Actually, ACR is a task that has been studied for more than two decades. Various methods have been proposed [16]. For most recent work, ACR systems based on deep learning models have been used. This section introduces two earlier studies using deep learning models [1, 2] because we use them in proposed ensemble learning.

A. ASR with CRNN

As a deep learning model for ACR, Jiang et al. used CRNN, a combination of CNN and RNN. It consists of the CNN part with convolutional and pooling layers, and the RNN part with Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU).

Fig. 1 portrays the model architecture of the CRNN as described by Jiang et al. The input is the fixed-length Constant-Q Transform (CQT) spectrogram extracted from a musical piece, which is a widely used two-dimensional feature in music analysis. The horizontal axis represents time. The vertical axis is the strength of each music scale, as calculated using CQT transformation at each time. The output is the probability vectors of each chord for the same frame as the input. Jiang et al. used 10 convolutional layers and three pooling layers with optimization in the CNN part. The RNN part uses Bi-directional LSTM (Bi-LSTM).

B. ASR with Transformer

Park et al. used bidirectional Transformer, a model based on the Attention mechanism, which is used widely in natural language processing fields. The Attention mechanism computes the output vector using three vectors obtained from the input: Query, Key and Value. In a sequence task, Query corresponds to a specific element of the target sequence, whereas Key and Value correspond to an element of the source sequence. The output vector is a weighted sum of Values. The weights are calculated using Query and Key. When Query, Key, and Value are obtained from the same input, it is called

the Self-Attention mechanism. Transformer is a model with the Self-Attention mechanism.

Fig. 2 shows the model architecture of the bidirectional Transformer for ACR by Park et al. Similarly to CRNN, the input is a fixed-length CQT spectrogram. The output is probability vectors of each chord for the same frame as the input. In the middle, the process called positional encoding is performed before the bidirectional Self-Attention layer to consider temporal information.

IV. PROPOSED METHOD

Because it has a convolutional filter, CRNN can capture local features, whereas the self-attention mechanism of Transformer enables it to capture global features. For this study, we aim to improve the ACR discrimination accuracy by performing ensemble learning using two models having different feature extraction approaches. We also propose a novel overlap inference method to resolve the lack of temporal musical information at the beginning and end of clips during inference.

Because the source chord of the CRNN model is not publicly available for one earlier study [1], we reproduced it based on descriptions in the report of that study. Considering the pre-experiments, we used CRNN without the normalization process. We used the Transformer model source chord published by its authors [2].

First, ensemble learning is explained. Then an overlap inference method is described.

A. Ensemble learning

Ensemble learning, a machine learning method, uses each output value of multiple models. For this study, we use the stacking method, which can deal with probability sequences. In stacking, the output values obtained in the training process of multiple models (lower models) are used to create datasets for training upper models. By constructing upper models that capture the tendency of each lower model, we can expect to improve the accuracy. For this study, we use the full-connected NN as upper models, for which the inputs and outputs are sequences. Fig. 3 shows a training flow for the lower and upper models. Regarding lower models, the input is a CQT spectrogram (F in the frequency direction × T frames in the time direction). The output is a sequence of chord probability vectors (chord classes C × T frames). N_Train is the number of spectrograms used for training.

The model is trained by application of k-Fold Cross-Validation, by which 1/k of the training data are used as validation data. The rest are used as training data. Training is performed k times. For this study, the number of Folds for the lower and upper models is set as five. K-Fold Cross-Validation is generally used to verify generalization performance accurately. However, it also creates datasets for training upper models of stacking. At the end of model training for each Fold of lower models, the output values for validation data are stored as input features for upper models. As the result of 5-Fold, five trained lower models and output values for the validation data are obtained for each fold.

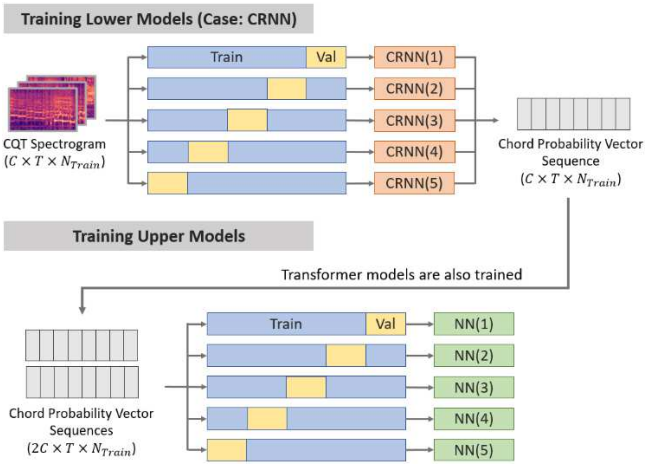


Fig. 3 Training flow for the lower and upper models.

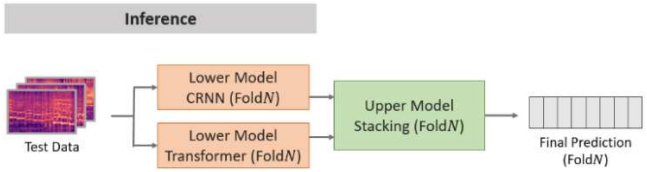


Fig. 4 Inference flow.

The sequence of chord probability vectors, which is the output value obtained in training lower models, CRNN and Transformer, is input to upper models as two-dimensional features (2 × C × T). The upper models are also trained with k-Fold Cross-Validation because of the inference described later. Therefore, five trained upper models are obtained.

Fig. 4 shows the inference flow. Along with the training, the CQT spectrogram is input to each lower model. The obtained two sequences of chord probability vectors are used to infer upper models. The lower and upper models used are of the same Fold.

B. Overlap inference

Both the earlier research models and the proposed models use the fixed-length spectrogram as input. For earlier studies [1, 2], the target music piece was divided into short audio data (clip) before input to the model. The predicted chord sequences obtained for each inference were arranged and treated as inference results for the entire piece. In this case, as presented in Fig. 5, the beginning and end of each clip often lacks the former and the latter part of temporal musical information, resulting in unstable inference. To resolve this difficulty, we adopt an original overlap inference method. An overview of overlap inference is presented in Fig. 6. The music is divided into clips with overlap. The middle parts of each inference result, shown as green in Fig. 6, are used as a result of the entire song. Presumably, this method generates stable inference results overall and engenders improved accuracy.

Our experiments compare two types of overlap inference method. Hereinafter, N_Test is the number of clips in T frames. In Pattern 1, the clips are overlapped by 1/2T and inference is

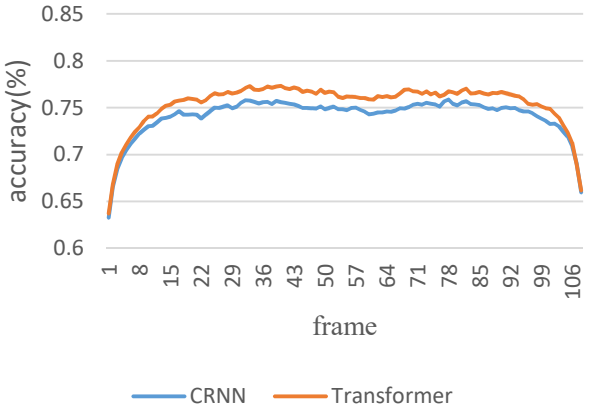


Fig. 5 Average accuracy per frame for validation data.

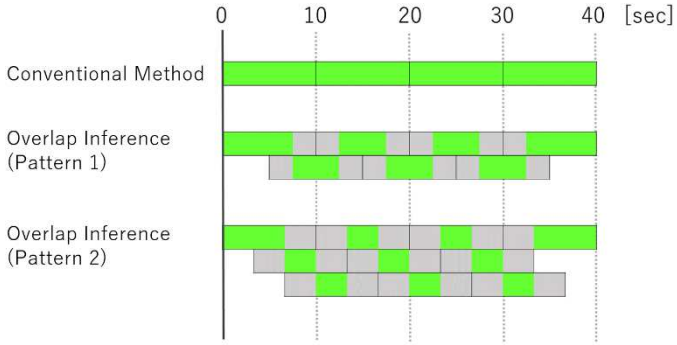


Fig. 6 Overview of overlap inference.

performed. Then the central 1/2 part of the inference result is extracted. Considering the non-overlapped part, the first clip is the 0 to 3/4T frame part of the inference result, each second to N_Test-1 clip is the 1/4T to 3/4T frame part of the inference result, and the N_Test clip is the 1/4T to T frame part of the inference result, are extracted and used as inference results of the entire song. In Pattern 2, the clips are overlapped by 2/3T; the central 1/3 part of the result is extracted. Considering the non-overlapped part, the first clip is the 0 to 2/3T frame part of the inference result, each second to N_Test-1 clip is the 1/3T to 2/3T frame part of the inference result, and the N_Test clip is the 1/3T to T frame part of the inference result, are extracted and used as inference results of the entire song.

V. EXPERIMENTS

This section presents comparison of the proposed model with CRNN [1] and Transformer [2], which are the earlier research models, with verification of the effectiveness of overlap inference. To make fair comparisons, experiments are conducted under the common conditions of input–output format and dataset used.

A. Dataset

For this study, we use 390 songs for model training and testing (80% for training and 20% for testing) as shown in

Table 2 Dataset used

Dataset	Annotated by	No. songs
Beatles	Isophonics [19]	180
Queen	Isophonics [19]	20
Zweieck	Isophonics [19]	18
Carole King	Isophonics [19]	7
Robbie Williams	Bruno et al. [20]	65
RWC POP [18]	MARL [21]	100

Table 2. Each annotation datum includes the type of chord that appears in the song, the start time, and the end time.

B. Pre-processing

We use the CQT spectrogram as an input feature for the lower models in ensemble learning. Because the input of the lower model must have fixed length, each piece of music is divided into clips of short duration. The number of seconds in each clip was set as 10 s (overlap 5 s), which preserves sufficient temporal information and which can be extracted efficiently from the music. We extract the CQT spectrogram from the audio by Librosa [17], with a sampling rate of 22,050 Hz, an analysis window length of 2,048 sampling points (approximately 0.093 s), and a scale of 8 octaves from C1 and 24 bins per octave. The analysis window is shifted without overlap so that each clip is a two-dimensional feature of 108 frames and 192 frequency bins.

We applied data augmentation using the 11-step pitch shift operation (from -5 semitones to -1 semitone and from +1 semitone to +6 semitones) on the training set to avoid overfitting. Annotation data corresponding to the music data generated by this extension process are also created and used.

Chords in the annotation data are based on 170 labels (Root{C, C#, D, D#, E, F, F#, G, G#, A, A#, B} × Quality{maj, min, dim, aug, sus2, sus4, maj6, min6, maj7, minmaj7, min7, hdim7, dim7} + No Chord + Unclassified) and are assigned to each frame of the feature to obtain the correct data.

C. Model training

We trained the CRNN, Transformer, and NN used for ensemble learning under the conditions shown in Table 3. If the validation loss did not improve by more than 0.001 in three epochs, then we decreased the learning rate by 80%.

Table 3 Training conditions

Model	CRNN	Transformer	NN
Programming language	Python 3.8.10		
Framework	Pytorch 1.10.0		
Input dim	108 × 192		108 × 340
Output dim	108 × 170		
Batch size	128		
Loss Function	Categorical CrossEntropy		
Optimizer	Adam [22]		
Initial learning rate	0.0001		

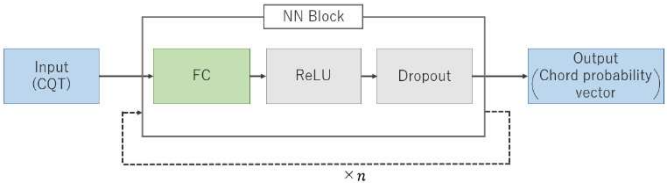


Fig. 7 NN model structure.

Table 4 NN model details

Model name	No. units	No. NN blocks	No. parameters
NN256_1	256	1	130,986
NN256_2	256	2	196,778
NN512_1	512	1	261,802
NN512_2	512	2	524,458
NN1024_1	1024	1	523,434
NN1024_2	1024	2	1,573,034

D. Upper model structure

The NN structure used as upper models is portrayed in Fig. 7. The chord probability vector output by the lower model (CRNN, Transformer) is used as input. The probability of each chord per frame is output after passage through the intermediate block consisting of the full-connected layer, ReLU function, and the dropout layer (rate 0.2). For this study, we trained several models with different numbers of units in full-connected layers and intermediate blocks for experiments. Table 4 presents the models used in experiments.

E. Comparative results

We evaluate the performance of an ensemble learning and an overlap inference. We use seven evaluation metrics: root only, root and thirds, major–minor chords, triads, sevenths, tetrads, and MIREX (more than three correct notes between reference and estimated chords). All evaluation metrics are calculated as the number of correct frames out of the total number of frames. They are calculated using the python library mir_eval [23].

First, we compare the results obtained for the six proposed models without overlap inference shown in Table 4 with CRNN and Transformer, the earlier research models. Table 5 shows inference results of each model. The proposed model (NN1024_1), compared to CRNN, improves the accuracy by 0.87% on average. Compared to Transformer, the proposed model improves accuracy by 1.66%. It confirms that an ensemble learning model in ACR is effective for improving the accuracy because the upper NN model was able to incorporate consideration of the tendency of the CRNN flexibly, which is strong for local features, and Transformer, which is strong for global features. In the case of model comparisons without overlap, neither increasing the number of full-connected layer units nor increasing the number of intermediate blocks improved accuracy in many cases. Most information had already been processed by the lower model. The upper model needs only to judge the accuracy of the information: it needs no large network structure.

Table 5 Results of the earlier studies' model and the six proposed NN models without overlap inference

	Root	**Thirds**	**Maj-min**	**Triads**	**Sevenths**	**Tetrads**	**MIREX**
CRNN	82.64	80.21	81.08	78.74	70.01	67.09	81.19
Trans-former	81.88	79.53	80.37	78.03	68.87	66.00	80.72
NN256_1	**83.34**	81.15	81.87	79.56	70.71	67.88	82.09
NN512_1	83.28	81.12	81.85	79.55	70.64	67.85	82.10
NN1024_1	**83.34**	**81.19**	**81.96**	**79.65**	70.77	**67.94**	**82.18**
NN256_2	83.27	81.11	81.87	79.56	70.62	67.78	82.10
NN512_2	83.25	81.06	81.86	79.52	**70.79**	67.93	82.01
NN1024_2	83.21	81.05	81.86	79.52	70.77	67.90	82.01

Table 6 Results of overlap inference (Pattern 1)

with OL1	**Root**	**Thirds**	**Maj-min**	**Triads**	**Sevenths**	**Tetrads**	**MIREX**
CRNN	83.10	80.73	81.60	79.28	70.56	67.63	81.72
Trans-former	82.46	80.25	81.02	78.73	69.44	66.54	81.46
NN256_1	83.85	81.81	82.51	80.23	71.37	68.55	82.80
NN512_1	83.89	81.84	82.56	80.27	71.52	68.69	82.83
NN1024_1	83.88	81.81	82.55	80.26	71.61	68.79	82.82
NN256_2	83.88	81.79	82.56	80.26	71.50	68.64	82.77
NN512_2	83.83	81.80	82.59	80.28	71.53	68.66	82.79
NN1024_2	**83.93**	**81.88**	**82.70**	**80.37**	**71.77**	**68.90**	**82.87**

Table 7 Results of overlap inference (Pattern 2)

with OL2	**Root**	**Thirds**	**Maj-min**	**Triads**	**Sevenths**	**Tetrads**	**MIREX**
CRNN	83.09	80.74	81.61	79.29	70.53	67.57	81.71
Trans-former	82.41	80.14	80.95	78.65	69.54	66.63	81.33
NN256_1	83.87	81.77	82.50	80.22	71.40	68.57	82.78
NN512_1	83.86	81.77	82.51	80.21	71.43	68.59	82.80
NN1024_1	**83.89**	**81.83**	82.57	80.27	71.58	68.74	**82.81**
NN256_2	83.84	81.78	82.54	80.24	71.55	68.68	82.75
NN512_2	83.84	81.75	82.52	80.20	71.60	68.73	82.70
NN1024_2	83.86	**81.83**	**82.61**	**80.30**	**71.72**	**68.84**	82.79

Next, we compare the results of overlap inference. The Pattern 1 and Pattern 2 results are presented, respectively, in Table 6 and Table 7. In Pattern 1, NN1024_2 had the highest average accuracy of 78.92% for all indicators. In Pattern 2, NN1024_2 had the highest average accuracy of 78.85%. Apparently, the inference performance in overlap inference depends on the number of parameters. Comparison of Pattern 1 and Pattern 2 shows that Pattern 1 had higher accuracy. Pattern 1 is a long interval extraction. Therefore, less discrepancy exists in the integration of each inference result.

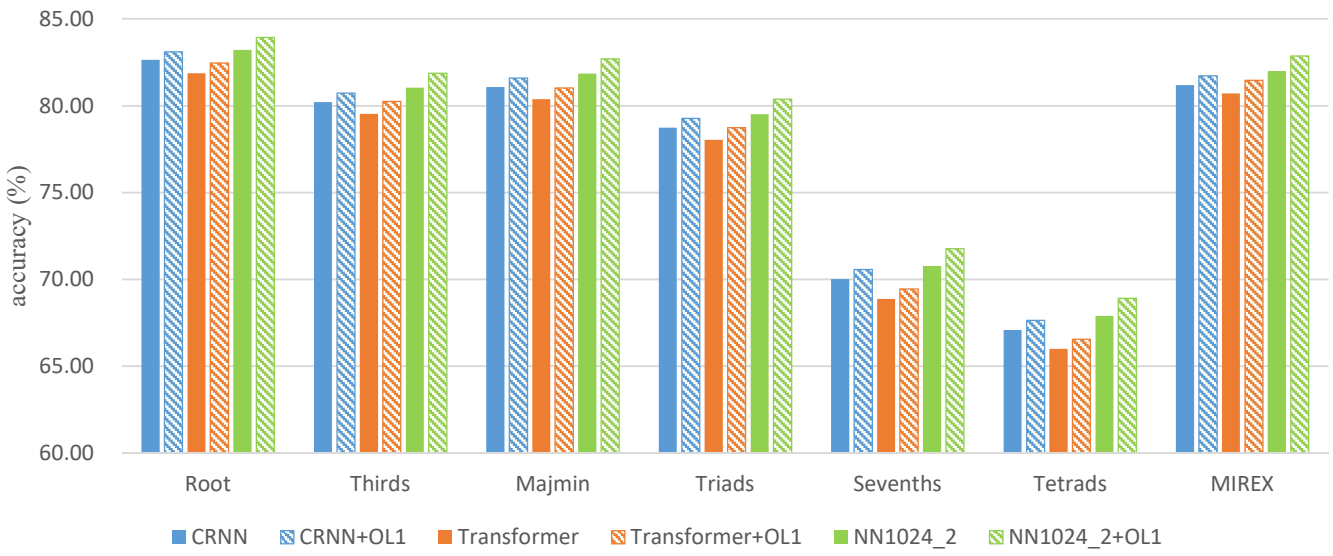


Fig. 8 Results of proposed NN model compared to CRNN and Transformer.

We confirmed the effectiveness of overlap inference (Pattern 1).

Finally, we compare the results obtained using the NN model (NN1024_2), which has the highest average accuracy with overlap inference with CRNN and Transformer. Fig. 8 presents the inference results obtained using each model. Pattern 1, which shows the greatest improvement in average accuracy, was used for overlap inference for all models. For inference without overlap, the NN model (green) improved the average accuracy by 0.77% and 1.56%, respectively, compared to the CRNN (blue) and Transformer (orange). Particularly Tetrads, the most difficult evaluation metric, showed the greatest improvement, respectively, of 0.81% and 1.90%. When we used overlap inference, CRNN (shaded blue) and Transformer (shaded orange) improved the average accuracy by 0.52% and 0.64%, respectively, compared to the same models without overlap inference. The NN model with overlap inference (shaded green) improved the average accuracy by 0.87% compared to the same model without overlap inference. It also improved the average accuracy by 1.64% and 2.43%, respectively, compared to the CRNN (blue) and Transformer (orange) without overlap inference and improved the average accuracy by 1.12% and 1.79%, respectively, compared to the CRNN (shaded blue) and Transformer (shaded orange) with overlap inference. We confirmed ensemble learning effectiveness and the overlap inference proposed in this study.

VI. Conclusions

As described in this paper, we propose ensemble learning models for automatic chord recognition, particularly addressing differences in feature extraction approaches: The CRNN has a convolutional filter, which can capture local features. By contrast, Transformer has a Self-Attention mechanism, which can capture global features. We also proposed original overlap inference methods to solve unstable inference because of the lack of temporal musical information in the inference of each clip.

In the inference without overlap, the NN model with 1024 full-connected layer units and one intermediate block had the highest average accuracy of 78.15%, which are 0.87% and 1.66% higher, respectively, than the CRNN and Transformer. In the overlap inference (Pattern 1), CRNN and Transformer improved the average accuracy by 0.52% and 0.64%, respectively, compared to inference without overlap. The NN model with 1024 full-connected layer units and two intermediate blocks in the overlapped inference caused the highest average accuracy of 78.92% among all models.

This study confirmed the effectiveness of ensemble learning and an overlap inference proposed in this study. In future studies we plan to improve the accuracy further by increasing the number of intermediate blocks and the number of full-connected layer units of the proposed NN model, improving the model structure, and adopting multi-task learning and Conditional Random Fields (CRF) used with other methods.

Acknowledgment

This work was supported by JSPS (C), KAKENHI Grant Number 21K12611.

References

[1] Junyan Jiang, Ke Chen, Wei Li and Gus Xia, “Large-vocabulary Chord Transcription via Chord Structure Decomposition,” 20th ISMIR Conference, pp. 644-651, 2019.

[2] Jonggwon Park, Kyoyun Choi, Sungwook Jeon, Dokyun Kim and Jonghun Park, “A Bi-Directional Transformer for Musical Chord Recognition,” 20th ISMIR Conference, pp. 620-627, 2019.

[3] Matthias Mauch and Simon Dixon, “Approximate note transcription for the improved identification of difficult chords,” 11th ISMIR Conference, pp. 135-140, 2010.

[4] Matt McVicar, Raúl Santos-Rodríguez, Yizhao Ni, and Tijl De Bie, “Automatic chord estimation from audio: A review of the state of the art,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(2):556-575, 2014.

[5] Yizhao Ni, Matt McVicar, Raul Santos-Rodriguez, and Tijl De Bie, “An end-to-end machine learning system for harmonic analysis of music,” IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1771-1783, 2012.

[6] Eric J. Humphrey, Taemin Cho, and Juan P. Bello, “Learning a robust tonnetz-space transform for automatic chord recognition” In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 453-456. IEEE, 2012.

[7] Hélene Papadopoulos and Geoffroy Peeters, “Large scale study of chord estimation algorithms based on chroma representation and HMM,” In 2007 International Workshop on Content-Based Multimedia Indexing, pp. 53-60. IEEE, 2007.

[8] Yushi Ueda, Yuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama, “HMM-based approach for automatic chord detection using refined acoustic features,” In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5518-5521. IEEE, 2010.

[9] Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” Nature, 521(7553):436-444, 2015.

[10] Eric J Humphrey and Juan P Bello, “Rethinking automatic chord recognition with convolutional neural networks,” In 2012 11th International Conference on Machine Learning and Applications, volume 2, pp. 357-362. IEEE, 2012.

[11] Filip Korzeniowski and Gerhard Widmer, “A fully convolutional deep auditory model for musical chord recognition,” In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6. IEEE, 2016.

[12] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, “Audio chord recognition with recurrent neural networks,” 14th ISMIR Conference, pp. 335-340, 2013.

[13] S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon, “Audio chord recognition with a hybrid recurrent neural network,” 16th ISMIR Conference, pp. 127-133, 2015.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 6000-6010, 2017.

[15] C.-Z. Anna Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, and D. Eck, ”Music transformer: Generating music with long-term structure,” arXiv preprint, arXiv:1809.04281, 2018

[16] Johan Pauwels, Ken O’Hanlon, Emilia Gómez and Mark B. Sandler, “20 Years of Automatic Chord Recognition from Audio,” 20th ISMIR Conference, pp. 54-63, 2019.

[17] librosa 0.8.1 documentation
https://librosa.org/doc/latest/index.html, 11 July 2022.

[18] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura and Ryuichi Oka, “RWC Music Database: Popular, Classical, and Jazz Music Databases,” ISMIR Conference, pp.287-288, 2002.

[19] Datasets | Isophonics
http://isophonics.net/datasets, 11 July 2022.

[20] Bruno Di Giorgi, Massimiliano Zanoni, Augusto Sarti and Stefano Tubaro, “Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony,” Eighth NDS Conference, 2013.

[21] GitHub - tmc323/Chord-Annotations
https://github.com/tmc323/Chord-Annotations, 11 July 2022.

[22] Diederik Kingma and Jimmy Ba. Adam, “A method for stochastic optimization,” In ICLR, 2015.

[23] mir_eval 0.6 documentation - Colin Raffel
https://craffel.github.io/mir_eval/, 11 July 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Construction of Automatic Speech Recognition Model that Recognizes Linguistic Information and Verbal/Non-verbal Phenomena

Nagito Shione, Yukoh Wakabayashi and Norihide Kitaoka
Toyohashi University of Technology, Japan

***Abstract*—This study addresses an automatic speech recognition (ASR) model that simultaneously recognizes linguistic information and various verbal/non-verbal (VNV) phenomena: fillers, laughter, interrogative ascending tones, end of speech, pronunciation errors, word fragments, etc. Although ASR technology has rapidly advanced in recent years, most ASR systems can only recognize linguistic information and can still not VNV phenomena. Recognition of VNV phenomena is important because speakers often convey their intentions during spontaneous speech using these types of utterances. We achieve VNV recognition by annotating VNV phenomena tags on learning transcribed text when the ASR model is trained. Our experimental results revealed that the accuracy of speech recognition is improved by also recognizing VNV phenomena. The ASR model in which the VNV phenomena tags appeared before the linguistic information in the transcribed text showed the best performance when using data from the Corpus of Everyday Japanese Conversation.**

I. Introduction

In recent years, conversational AI assistants like Apple's Siri and Google Assistant have been widely installed in smartphones and smart speakers, and have become increasingly popular for helping people in their daily lives. These conversational AI assistants have been realized through the use of Automatic Speech Recognition (ASR), which is used to interpret user commands. With the rise of End-to-End (E2E) models, ASR performance has surpassed that of the previously dominant Deep Neural Network-Hidden Markov Model (DNN–HMM) approaches in terms of recognition accuracy. However, conversational speech often includes fillers like "uh" and "um", missing words, repetitions, and other disfluencies, which degrade speech recognition [1] since ASR models are typically trained using fluent speech. Furthermore, conversational speech includes not only verbal phenomena, such as fillers, but also non-verbal phenomena, such as laughter, changes in intonation and variations in voice intensity, which can play an important role in communication. For example, fillers are used to give the speaker additional time when choosing words, or to organize one's thoughts [2]. Laughter is used to soften the meaning of the previous utterance or to express emotions such as happiness or discomfort [3]. Thus, speakers try to communicate their intentions to others by using not only verbal information but also verbal/non-verbal (VNV) phenomena, and therefore it is important to be able to recognize VNV phenomena in speech.

Several studies on recognizing VNV phenomena in speech have been previously conducted. Kaushik *et al.* built an ASR model that could detect fillers and laughter in speech by combining a DNN, a Convolutional Neural Network (CNN), and a Low-pass filter (LPF) [4]. Chatziagapi *et al.* built an ASR model to detect fillers in both speech and linguistic information by combining a CNN and a Recurrent Neural Network (RNN) [5]. However, these studies only focused on recognizing VNV phenomena in speech, and not on recognizing the linguistic information, which is our primary method of understanding conversation. Therefore, it is important to recognize both language information and VNV phenomena simultaneously in order to understand a speaker's intended message.

Some studies on simultaneously recognizing linguistic information and verbal phenomena have been conducted. Mendelev *et al.* built an ASR model to simultaneously recognize linguistic information and hesitations in speech using a RNN Transducer (RNN-T) [6]. Horii *et al.* proposed an ASR model capable of simultaneously recognizing linguistic information, fillers and hesitations in speech using a Joint/CTC Attention Transformer [7], and reported that the simultaneous recognition of linguistic information and along with fillers and hesitations improved the accuracy. These studies suggest that the simultaneous recognition of verbal phenomena and linguistic information improves the accuracy of speech recognition. However, these studies only made use of fillers and hesitations as verbal phenomena. In order to better understand natural human speech, it will be necessary to recognize a wider variety of VNV phenomena. Therefore, in this study we propose an ASR model that simultaneously recognizes linguistic information and multiple types of VNV phenomena. To achieve this, the proposed ASR model is trained using a conversational corpus containing spontaneous utterances with VNV phenomena. The proposed model can more accurately interpret user intentions by simultaneously recognizing linguistic information and a wide variety of VNV phenomena

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

TABLE I: VNV tags and transcription examples from CEJC. The nine in bold type are phenomena recognized in this study.

Tag	Description	Example of transcription
(W)	**pronunciation errors such as hesitation and incorrect pronunciation**	**(W コエ｜これ)，(W ギーツ｜技術)** **(W koe｜kore)，(W gītsu｜gijutsu)**
(D)	**word fragments**	**(D コ) 明日から，(D ＃)** **(D ko) starting tomorrow，(D ＃)**
?	**interrogative ascending tones other than emphatic ascending tones**	**行きます?，コップ?** **coming?，cup?**
(T)	**small voice**	**(T これじゃないのか)** **(T I thought this was it)**
(L)	**speech with laughter or laughter**	**(L なんですけど)，(L)** **(L but)，(L)**
<>	**speech related to the flow of conversation**	**<舌打ち>，<咳>，<口笛>** **<click one's tongue>，<cough>，<whistle>**
(F)	**filler**	**(F あの)，(F そーの)** **(F well)，(F let's see)**
。	**end of speech**	**食べます。，やったけど。，うん。** **I eat it.， I did it.， yes.**
(O)	**speech in dialects or foreign languages**	**(O ボッソワー)，(O ＃＃＃)** **(O bossowā)，(O ＃＃＃)**
:	prolongation	すご：い，けれども： ama:zing， but:
%	clogged sound	す%ごい，解%析 a%mazing，analy%sis
(C)	speech with crying or crying	(C なにが)，(C) (C what)，(C)
(S)	speech with singing or singing without lyrics	(S ヘイヘイホー)，(S) (S heiheihō)，(S)
(U)	speech with uncertain comprehension/word judgment	(U ジャック) に，(U 国産/特産) (U Jack) to，(U domestic production/special production)
(X)	speech in which words are not understood	(X フンジン) 中に，(X ＃＃＃) (X hunjin) in，(X ＃＃＃)
(K)	speech that cannot be written in Kanji for tags, resulting in reduced readability	(K シ: ツ｜質) 問，(K ナ%シ｜梨) (K qu:es｜ques)tion，(K pe%ar｜pear)
(M)	speech that is difficult to grasp when transcribed, such as (W).	すごいを (M すっごい) と発音する pronouncing with sugoi as (M suggoi)
(B)	baby babble	(B アー)，(B バ＃＃＃) (B ah)，(B ba ＃＃＃)
(Y)	speech that is pronounced differently from the common reading of the kanji	(Y ゼツ｜舌)，(Y センゲン｜浅間) (Y zetsu｜舌)，(Y sengen｜浅間)
(G)	colloquial expressions with low readability	(G 嫌｜や:)，(G もう｜も) (G iya｜ya:)，(G mou｜mo)
(I)	speech that is not divided into speech units due to interjection	最近 (I あっ) 付いてるみたいな。 it is (I oh) attached.
+	speech in which one short unit is divided by a perceptible pause	す+ごい，神+ 田川 ama+zing，kan+da river
(R)	speech with pseudonyms and foreshadowing related to personal information	(R 国語) 大学の (R 佐藤) さん Mr. (R Sato) from (R japanese)university
@	comments on the transcription	お願いしま：す。@店員への応答 plea:se. @reply to clerk

II. Corpus of Everyday Japanese Conversation

The Corpus of Everyday Japanese Conversation (CEJC) [8] is spontaneous speech corpus which contains 200 hours of audio, 577 conversations, and approximately 2.4 million words uttered by 1675 speakers. It includes everyday conversations occurring in a variety of situations, such as a conversation with a friend at a restaurant, and a meeting at work, etc. Unique characteristics of the CEJC are that it includes natural everyday conversations which occur in daily life, and that audio and video are both recorded. It also contains tagged transcriptions of the utterances, identifying VNV phenomena as well as other phenomena. There are 24 types of VNV phenomena which are tagged in the CEJC in total, a list of which is shown in Table I. As shown, the CEJC supplements linguistic information with tags for verbal phenomenon, such as fillers (F) and pronunciation errors (W), as well as tags for non-verbal phenomenon tags, such as laughter (L) and interrogative ascending tones (?), and other phenomenon tags such as uncertain hearing or word judgment (U) and pseudonyms and foreshadowing related to personal (R). In the CEJC, the VNV and other information phenomenon tags are generally added before the linguistic information, as a supplement to the linguistic information, but some VNV phenomenon tags and other phenomenon tags are handled differently. For example, (W), which indicates pronunciation errors, is transcribed with the linguistic information of the mispronunciation and the language information of the correction. The tag (L), which indicates laughter, is transcribed as (L) if there is only a laugh.

TABLE II: Positions of the VNV phenomenon tags

Positions of VNV tags	Example of transcription
Before	まず (D ソ) 何を (L 目的にするか)。
After	まず (ソ D) 何を (目的にするか L)。
Enclosing	まず (D) ソ (/D) 何を (L) 目的にするか (/L)。

III. Proposed method

In this study, we propose an ASR model that simultaneously recognizes linguistic information and 9 types of VNV phenomenon tags. In addition, we investigate the effect on speech recognition accuracy of changing the position of VNV phenomenon tags. The proposed method is described below.

A. Recognition of verbal and non-verbal phenomena

The VNV phenomena that we recognize in this study are fillers (F), laughter (L), interrogative ascending tones (?), end of speech (。), pronunciation errors (W), word fragments (D), small voice (T), speech related to the flow of conversation (<>) and speech in dialects or foreign languages (O). Of the twenty-four types of phenomena listed in Table I, the nine VNV phenomena tags with the largest number of occurrences were selected as recognition targets. By simultaneously recognizing not only linguistic information but these nine types of VNV phenomena, we can build an ASR model that is able to use a richer amount of information for recognition than conventional ASR models.

B. Changing the position of VNV phenomenon tags

In this study, the positions of the VNV phenomenon tags in the transcription were changed, and a different ASR model constructed for each variation. The three tag position protocols used for the VNV phenomena tags are shown in Table II. These are before, after, and enclosing the targeted linguistic information. Within the CEJC, the end of speech (。), interrogative ascending tones (?) and voice, speech related to the flow of conversation (<>) are not supplementary to the linguistic information, and are assigned independently. Therefore, the positions of these three types of VNV phenomenon tags were not changed under any of the three conditions. By building and evaluating ASR models under each of these three tag position conditions, we can clarify the effect of the position of the VNV phenomenon tags on speech recognition.

IV. Experimental settings

We used the CEJC as the dataset to implement the proposed ASR method. We excluded from the dataset utterances that contained unrecognized VNV phenomenon tags, and other phenomenon tags other than the nine we selected. Because the CEJC take place daily in a variety of settings, some of the speech is relatively clean (daily conversation in a staff room), while other conversations contain considerable noise (daily conversation in a restaurant). In fact, the CEJC contains more noisy than clean audio, therefore we performed speech enhancement on all of the speech samples using the Conv-TasNet model [9] which was pre-trained with CHiME [10] as provided by ESPnet [11]. This process denoised somewhat. The dataset was divided into three sets: a Train set for training data, a Dev set for development data and an Eval set for evaluation data. Details of the datasets used in this study are shown in Table III.

We used ESPnet to conduct the experiments in this study. ESPnet is an open source toolkit for speech processing which is specialized for E2E ASR models. An ASR model used in this study is the Joint CTC/Attention Transformer model provided by ESPnet, which is a Transformer Encoder–Decoder model that combines Attention and Connectionist Temporal Classification (CTC) [12]. Speed Perturbation (SP) [13] was applied to the speech in the training set to create additional speech data, converted at 0.9x and 1.1x the speed of the original speech data. The input acoustic features were log-mel spectrograms calculated using 80 mel-scale filter banks. The input acoustic features were further augmented with Spec Augment [14], and then Mean and Variance Normalization (MVN). Next, embedding and positional encoding were performed to convert the features into feature vectors with embedded positional information. This data is then input to the Transformer Encoder, which consisted of 18 layers, with each layer consisting of 2 sub-layers; a Multi-Head Attention layer and a Feed-Forward Network layer. Each sub-layer was followed by residual connection and normalization. The output of the last layer was input to the Transformer Decoder and the CTC Decoder, the latter of which has a fully-connected layer. The Transformer Decoder consisted of six layers, each of which consisted of 3 sub-layers; a Masked Multi-Head Attention layer, a Multi-Head Attention layer and a Feed-Forward Network layer. Each sub-layer was followed by residual connection and normalization, like the Encoder. For Multi-Head Attention, the number of heads in the Encoder and Decoder was 8. We trained the model using the Adam optimizer. For inference, the output of the Transformer Decoder and the output of the CTC Decoder were used to output the sequence of label $\hat{y}$, which was obtained as follows:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \{\lambda \log p_{\mathrm{ctc}}(y|x) + (1-\lambda) \log p_{\mathrm{tf}}(y|x)\}, \quad (1)$$

where $\log p_{\mathrm{ctc}}(y|x)$ and $\log p_{\mathrm{tf}}(y|x)$ are the output probabilities of the CTC Decoder and the Transformer Decoder, respectively, that output the sequence of label y when input acoustic features x are input, and λ is the weight of the CTC Decoder. In this study, λ was set to 0.3, which is the ESPnet default.

We used Character Error Rate (CER) and Slot Error Rate (SER) [15] to evaluate and compare the performance of the ASR models as measures. CER was used to evaluate speech recognition accuracy for linguistic information while SER was used to evaluate VNV phenomenon tag detection accuracy. SER is the ratio of the total number of deletion errors, insertion errors, and substitution errors divided by the total number of VNV phenomenon tags correctly detected, as shown in Eq. (2). SER is obtained as follows:

TABLE III: Details of the datasets used in this study

Dataset	Number of utterances	Duration	Number of VNV tags								
			(W)	(D)	?	(T)	(L)	<>	(F)	。	(O)
Train	476590	116:04:57	12743	25140	34654	8786	46917	169	13308	374066	16
Dev	59529	15:55:41	1855	3702	3466	1222	4737	54	1819	46450	15
Eval	60618	16:15:26	1775	4048	2809	1711	7289	27	2284	44937	209

$$\text{SER} = \frac{\text{\# of insertion errors + \# of deletion errors + \# of substitution errors}}{\text{\# of target VNV phenomenon tags}}. \quad (2)$$

SER is greater than 100% if the total number of targeted VNV phenomenon tags that were correctly detected is less than the number of insertion errors.

V. Experimental results

We built three different ASR models by changing the position of the VNV phenomenon tags in the training data, as described in Sec. Ⅲ-B, and obtained the CER of the linguistic information and SER for tag detection. The results are shown in Table IV. We can see that the ASR model for the "enclosing" VNV tag condition produced a higher CER than the other two ASR models. This is because the number of characters to be recognized was larger. For example, the "before" and "after" tag conditions require estimation of four characters because "(" and ")" is one character when the recognition target is (D ソ) or (ソ D). On the other hand, the "enclosing" tag condition requires estimation of eight characters when the recognition target is (D) ソ (/D). Therefore, there is a difference in the number of characters even if the recognition target is the same. As a result, the "enclosing" tag condition had a larger negative impact on the CER when mis-estimation occurred, compared to the other ASR models. Table IV also shows that the SERs for speech related to the flow of conversation (<>) and for the speech in dialects or foreign languages ((O)) were greater than 100% in all conditions. This was due to a lack of sufficient data for training. However, the SERs for interrogative ascending tones (?) and small voice (T) were also high, even though the number of utterances used for training. For (T), this was because the input acoustic feature is a non-verbal phenomenon (the loudness of a sound). Since we only used the log-mel spectrogram as the acoustic features input, we did not use a feature related to the loudness of the sound, so the SER of (T) was high. In future work, we will add features that are sensitive to changes in sound intensity, such as power features. The log-mel spectrogram used as input acoustic features contains information on pitch although the SER for (?) was also high. It is highly possible that the change in pitch due to (?) was not captured by the spectrogram. In future work, we will add features that are more sensitive to changes in pitch, such as F0 features. The SERs for pronunciation errors (W) and word fragments (D) were much higher than fillers (F) even though these three are the same verbal phenomena. This is because VNV phenomenon tags contain a variety of linguistic information; (F) contains a limited variety of linguistic information, such as (F あの) and (F その) while both (W) and (D) both include a huge variety of linguistic information unlike (F). As a result, recognition of the (W) and (D) tags was not successful.

Table IV shows that the SERs were different for each VNV tag position condition, so the position of the VNV phenomenon tags effected VNV phenomena detection in the CEJC data. The Transformer model uses self-attention mechanisms to calculate the relationships between tokens within an input sequence. Since the tag indicates positional information, so the model takes that positional information into account. For example, when tag is placed before tokens, the model is more likely to associate them with the preceding tokens. Therefore, the position of tags can be considered to alter the importance of positional information, potentially leading to differences in accuracy. The ASR model trained under the "before" tag condition achieved a lower CER and SERs than those trained under the other tag conditions. Therefore, the best ASR model for the CEJC data was the one that recognizes linguistic information by prefixing it with a VNV phenomenon tag.

We also evaluated ASR performance when only linguistic information was used for recognition, in order to compare the performance of our proposed method with that of a standard ASR model that does not use VNV tag information. To perform this comparison, we built an additional ASR model trained without VNV tags as a baseline model. The training conditions and architecture of the baseline model were the same as those of the other models, as described in Section Ⅳ, except no VNV tags were used. To compare performance of the proposed methods with the baseline model, we post-processed the recognition results of the proposed ASR models by removing the VNV phenomenon tags, and then obtained CERs when using only linguistic information.

The flow of VNV phenomenon tag removal is shown in Fig. 1. Performance results for each model are shown in Table V. We can see that the CERs achieved by the proposed ASR models were lower (i.e., better) compared to the baseline model. Therefore, the proposed ASR modeling method improves speech recognition accuracy by simultaneously recognizing not only linguistic information but also VNV phenomena tags in the CEJC.

VI. Conclusion

In this study, we proposed an ASR model that simultaneously recognizes linguistic information and a wide variety of VNV phenomena, in contrast to conventional ASR models that only recognize linguistic information. We also investigated the effect of changing the position of the VNV phenomenon tags on speech recognition accuracy, and on the accuracy of VNV phenomena detection, as well as the effect of simultaneously recognizing VNV phenomena and linguistic information on

TABLE IV: Character Error Rates (%) and Slot Error Rates (%) when VNV tags are located in various positions. CER was calculated by including the VNV phenomenon tags as characters to be recognized.

Position of VNV tags	CER (%)	SER (%)								
		(W)	(D)	?	(T)	(L)	<>	(F)	。	(O)
Before	25.3	98.9	77.0	73.1	93.8	27.9	100.0	24.6	16.7	100.0
After	25.4	99.3	77.6	73.4	94.5	28.5	103.7	25.9	17.1	100.0
Enclosing	30.4	100.9	81.6	74.1	95.9	27.8	107.4	28.1	16.9	100.0

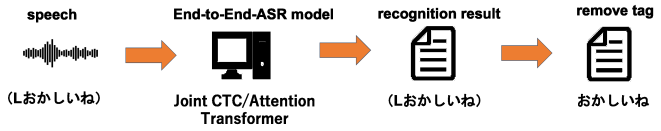


Fig. 1: Flow of VNV phenomenon tag removal

TABLE V: Character Error Rates (%) when VNV phenomenon tags are removed from recognition results. "Baseline" is an ASR model trained without VNV tags.

Position of VNV tags	CER (%)
Baseline (w/o tags)	26.3
Before	23.5
After	23.6
Enclosing	24.1

the accuracy of speech recognition. Our experimental results showed that the best performing ASR model for CEJC data was one that recognizes linguistic information by prefixing it with VNV phenomenon tags. We also demonstrated that simultaneous recognition of linguistic information and VNV phenomena improves the accuracy of speech recognition, in comparison to a conventional ASR method using only linguistic information. In the future, we would like to investigate methods of improving both the speech recognition accuracy and VNV phenomena detection accuracy of the proposed method by using it in combination with a language model, F0 features and power features.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22H04863.

References

[1] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates," in *Proceedings of ACL-08: HLT*, 2008, pp. 380–388.

[2] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[3] J. Vettin and D. Todt, "Laughter in conversation: Features of occurrence and acoustic structure," *Journal of Nonverbal Behavior*, vol. 28, pp. 93–115, 2004.

[4] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Laughter and filler detection in naturalistic audio," in *Interspeech*, 2015, pp. 2509–2513.

[5] A. Chatziagapi, D. Sgouropoulos, C. Karouzos, *et al.*, "Audio and ASR-based Filled Pause Detection," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2022, pp. 1–7.

[6] V. Mendelev, T. Raissi, G. Camporese, and M. Giollo, "Improved robustness to disfluencies in RNN-Transducer based speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6878–6882.

[7] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-End Spontaneous Speech Recognition Using Disfluency Labeling," in *Interspeech*, 2022, pp. 4108–4112.

[8] H. Koiso, H. Amatani, Y. Den, *et al.*, "Design and Evaluation of the Corpus of Everyday Japanese Conversation," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 5587–5594.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[11] S. Watanabe, F. Boyer, X. Chang, *et al.*, "The 2020 ESPnet Update: New Features, Broadened Applications, Performance Improvements, and Future Plans," in *2021 IEEE Data Science and Learning Workshop (DSLW)*, IEEE, 2021, pp. 1–6.

[12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, 2019, pp. 2613–2617.

[15] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance Measures for Information Extraction,"

in *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, 1999, pp. 249–252.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Exploring Isolated Musical Notes as Pre-training Data for Predominant Instrument Recognition in Polyphonic Music

Lifan Zhong*, Erica Cooper†, Junichi Yamagishi†, and Nobuaki Minematsu*
* Graduate School of Engineering, The University of Tokyo, Japan
E-mail: {zhong_lifan, mine}@gavo.t.u-tokyo.ac.jp Tel/Fax: +81-3-5841-6662
† National Institute of Informatics, Japan
E-mail: {ecooper, jyamagis}@nii.ac.jp Tel: +81-3-4212-2576

***Abstract*—With the growing amount of musical data available, automatic instrument recognition, one of the essential problems in Music Information Retrieval (MIR), is drawing more and more attention. While automatic recognition of single instruments has been well-studied, it remains challenging for polyphonic, multi-instrument musical recordings. This work presents our efforts toward building a robust end-to-end instrument recognition system for polyphonic multi-instrument music. We train our model using a pre-training and fine-tuning approach: we use a large amount of monophonic musical data for pre-training and subsequently fine-tune the model for the polyphonic ensemble. In pre-training, we apply data augmentation techniques to alleviate the domain gap between monophonic musical data and real-world music. We evaluate our method on the IRMAS testing data, a polyphonic musical dataset comprising professionally-produced commercial music recordings. Experimental results show that our best model achieves a micro F1-score of 0.674 and an LRAP of 0.814, meaning 10.9% and 8.9% relative improvement compared with the previous state-of-the-art end-to-end approach. Also, we are able to build a lightweight model, achieving competitive performance with only 519K trainable parameters.**

I. Introduction

Music is an essential part of our lives. Nowadays, online music streaming services such as Apple Music and Spotify enable us to access an abundance of music recordings much more easily than before. At the same time, the massive amount of data brings new challenges: how can we find the music we want efficiently? Music Information Retrieval (MIR) has emerged as a research field that focuses on extracting and inferring meaningful features from music, indexing music using these features, and developing various search and retrieval methods to make the world's vast store of music easily accessible by individuals [1].

As one of the important problems in MIR, instrument recognition shows great potential to contribute to real-world music applications, such as music recommendation and source separation [2]. Though the past 30 years have seen great progress in the automatic classification of monophonic phrases or isolated notes, the task of instrument classification for polyphonic multi-instrument music, where the sound mixture becomes more complex but similar to the real-world scenario, remains rather difficult. In this work, we tackle predominant instrument recognition, which focuses on identifying the main instruments in a given polyphonic music segment.[1]

Generally speaking, training a robust statistical model for classification requires a large amount of high-quality annotated data. However, in the field of predominant instrument recognition, a lack of such annotated polyphonic musical data has been a constraint for training a robust classifier. First, domain knowledge is necessary for annotation in polyphonic musical mixtures, especially for complex ones. Second, well-produced music recordings help generalization but have copyright issues. On the contrary, monophonic sounds and isolated notes require relatively less effort to collect and label. But in terms of timbral characteristics, these are far from the music we listen to daily, which is consecutive, melodic, polyphonic, and not as "clean" as the isolated samples. Therefore, it would be interesting to discuss how monophonic musical data can help instrument recognition in a polyphonic setting.

Some studies have touched on this. Bosch *et al.* [3] made an early attempt by using source separation as a pre-processing step to divide polyphonic music signals into stems, and classified each stem through a separate model. Kratimenos *et al.* [4] experimented with four random mixing augmentation methods using the IRMAS training set [3], which contains monophonic samples and polyphonic samples but all labeled with one predominant instrument, to improve the polyphonic instrument recognition system.

In this paper, we look into this topic from another perspective. We investigate the effectiveness of using isolated musical notes as pre-training data for improving the predominant instrument recognition system. We adapt the model proposed in [5] to classify instrument classes from raw musical waveforms and experiment with various data augmentation techniques to bridge the domain gap between isolated notes and polyphonic music. Furthermore, we investigate a weight-sharing method [6] for reducing the parameters of our instrument encoder.

The remainder of the paper is structured as follows: Section II gives a non-exhaustive review of the relevant research in the

[1] In this paper, "polyphonic" means both polyphonic and multi-timbral, following the literature in this field.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

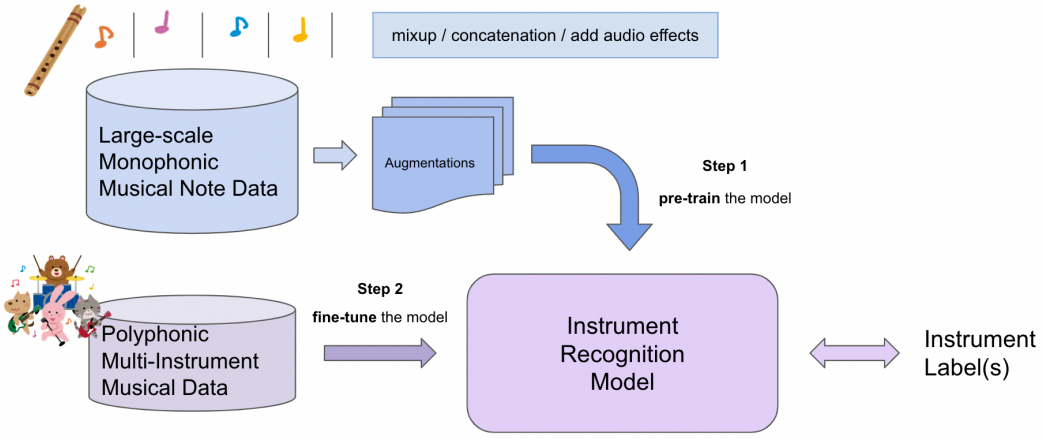


Fig. 1. Overview of the proposed method.

field of instrument recognition. Section III describes the model architecture and data augmentation pipeline. Section IV introduces the details of our experiment settings. Section V presents our experimental results. Finally, Section VI summarizes this work.

II. Related Work

Automatic classification of musical instruments has been studied since the mid-1990s. In 1995, Kaminsky & Materka [7] trained instrument classifiers for four instruments in isolated notes. Eggink & Brown [8][9] conducted two consecutive studies on instrument classification in polyphonic mixtures with models trained on monophonic audio. Livshin & Rodet [10] developed a real-time solo instrument classifier and also applied it directly to performance duets, demonstrating that the solo instrument classifier has the potential for multi-instrument classification.

As the automatic classification of isolated musical notes has been considered a solved problem [11], and monophonic music is not very applicable to real-world scenarios, more and more attention has shifted to instrument recognition in a polyphonic context. Fuhrmann *et al.* [12][13] took the first step towards the problem of predominant instrument recognition in polyphonic music by collecting a database and using hand-crafted features and support vector machines (SVM) to classify the music segments. Bosch *et al.* [3] extended this research by using source separation as a pre-processing step. The dataset used in their research, the IRMAS (Instrument Recognition in Musical Audio Signals) dataset, later became the benchmark for predominant instrument recognition.

Recently, deep learning techniques have fostered more and more advances in polyphonic instrument recognition. In the task of predominant instrument recognition, Han *et al.* [14] continued the research on the IRMAS dataset by using VGG-styled convolutional neural networks (CNNs), and their system outperformed the previous SVM methods, marking a pioneering work in deep models for predominant instrument recognition. Pons *et al.* [15] made an attempt to modify the neural network structures based on domain knowledge, achieving similar scores to previous work with much fewer model parameters. Yu *et al.* [16] extended the CNN-based predominant instrument recognizer by combining additional spectral features and a multi-task approach based on instrument groups. Reghunath and Rajan [17] applied Swin-Transformers to this task, and their input incorporates images of Mel-spectrograms as well as phase and tempo features. With an ensemble voting technique, a micro F1-score of 0.66 was achieved. Kratimenos *et al.* [4] proposed to randomly mix monophonic audio clips as augmented data samples to train neural networks for polyphonic instrument recognition. Avramidis *et al.* [18] compared different architectures in an end-to-end fashion, using raw waveforms as model input. Zhong *et al.* [19] investigated a cross-modal transfer learning approach for predominant instrument recognition using an ImageNet pre-trained vanilla ResNet-50 model.

III. Methodology

A high-level scheme of our proposed method is presented in Fig. 1. The training pipeline is as follows: First, we augment the monophonic musical note data by mixing, concatenating, and adding effects, to alleviate the domain gap between isolated notes and polyphonic recordings. The augmentations are done on-the-fly during the training process. Then we use the augmented data to pre-train the instrument recognition model. Finally, we fine-tune the pre-trained model for predominant instrument recognition using polyphonic, multi-instrument musical recordings. An overview of the instrument recognition model is shown in Fig. 2. We will give a detailed explanation of the model and the augmentation methods in the following section.

A. Model Architecture

We adapt the instrument encoder proposed by Shi *et al.* in [5] as the main backbone. It consists of a learnable front-end, a CNN feature extractor, and a learnable pooling layer. The components of the backbone are described next.

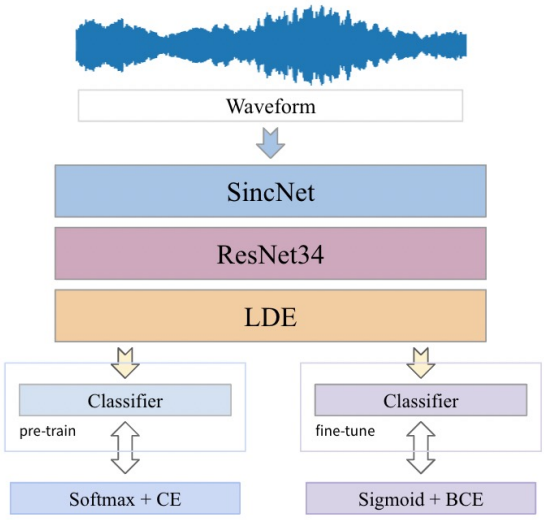


Fig. 2. Model Architecture.

A SincNet [20] layer is used as the trainable front-end to extract the low-level features from the raw waveform. SincNet is implemented as a 1-D CNN layer. Instead of learning all the elements of the kernels, SincNet constrains its kernels to be a set of sinc functions in the time domain. Suppose g is one SincNet kernel; then the operation is written as:

$$g[n, f_1, f_2] = 2f_2 \,\mathrm{sinc}(2\pi \mathrm{f}_2 \mathrm{n}) - 2\mathrm{f}_1 \,\mathrm{sinc}(2\pi \mathrm{f}_1 \mathrm{n}), \quad (1)$$

where $\mathrm{sinc}(\mathrm{x}) = \sin(\mathrm{x})/\mathrm{x}$. Then its Fourier transform is a rectangular bandpass filter in the frequency domain, which can be described as the difference between two low-pass filters:

$$G[f, f_1, f_2] = \mathrm{rect}(\frac{\mathrm{f}}{2\mathrm{f}_2}) - \mathrm{rect}(\frac{\mathrm{f}}{2\mathrm{f}_1}), \quad (2)$$

where $\mathrm{rect}(\cdot)$ is the rectangular function. For such a parameterized filter, the only parameters to learn are f_1 and f_2, which are the low and high cutoff frequencies, respectively. Although in [5], a constant Q transform (CQT)-initialized SincNet provided better results, we did not find results showing improved polyphonic instrument recognition performance. Thus, the SincNet layer is initialized using the Mel-scale cutoff frequencies in our experiments, following the original SincNet paper.

We then use 34-layer residual convolutional networks (ResNets) [21] to encode the low-level acoustic features to meta-level feature vectors. ResNets are CNNs which have identity mapping by shortcuts in every few stacked layers.

Finally, we use a learnable dictionary encoding (LDE) layer [22] to aggregate CNN-encoded features as the instrument embeddings. In a nutshell, LDE learns a dictionary as a set of clustering centroids based on the weighted distance between every frame-level feature and every centroid.

B. Data Augmentation

We use the following three data augmentation strategies at the pre-training stage:

1) Concatenation: Some instruments cannot produce sustained notes. Thus, following [5], we randomly concatenate trimmed samples of the same instruments. By doing so, we also introduce note changes within an instrument in the generated samples, which is common in real-world recordings.

2) Mixup: Since multiple instruments may play at the same time in multi-timbral polyphonic music, in the pre-training phase, we adopt a mixup style training process [23] [24], which trains the networks on convex combinations of pairs of examples and their labels. This process can be described as:

$$x_{mix} = \lambda x_i + (1 - \lambda) x_j, \quad (3)$$

$$y_{mix} = \lambda y_i + (1 - \lambda) y_j, \quad (4)$$

where x_i and x_j are two randomly selected audio waveforms of the same length, y_i and y_j are their labels respectively, $\lambda \in [0, 1] \sim \mathrm{Beta}(\alpha, \alpha)$ is the random mixing ratio, x_{mix} is the generated waveform, and y_{mix} is the corresponding mixed label. In our implementation, we randomly select the audio samples regardless of the instrument class for mixing. Before combing the samples, we normalized them to the same Loudness Units relative to Full Scale (LUFS) level [25][2]. LUFS is a weighted loudness measure mimicking the relative sensitivity of the human ear to different frequencies in terms of perceived loudness.

3) Audio effects: As in reality, professionally-produced music often involves different kinds of audio effects such as reverb, delay, etc., we apply audio effect augmentation, including adding noise, delay, reverb, gain, pitch shift, and high-low pass for pre-training[3]. According to [26], augmentation with audio effects commonly used in electronic music production improves the instrument recognition accuracy in both processed and unprocessed test sets.

C. Fine-tuning for Polyphonic Multi-instrument Recordings

Pre-training uses augmented monophonic note data, and its optimization criterion is to select only one most plausible instrument among multiple instruments. We therefore update the model with the softmax layer based on the cross-entropy loss with corresponding instrument labels.

The criterion of the final task, on the other hand, is to select one or more principal instruments in a segment of the input polyphonic multi-instrument music recording, which is different from the pre-training criterion. Hence, for fine-tuning, we replace the softmax layer with a Sigmoid layer where each dimension corresponds to a probability for each instrument class and compute an element-wise binary cross-entropy loss so that more than one class may be predicted by setting a threshold value for the predicted probability of each instrument class.

[2]https://github.com/csteinmetz1/pyloudnorm
[3]https://github.com/Spijkervet/torchaudio-augmentations

TABLE I
SUMMARY OF THE NSYNTH DATASET AND THE IRMAS DATASET

Dataset	NSynth	IRMAS - train	IRMAS - test
# Instruments	1,006	11	11
# Samples	305,979	6,705	2,874
Duration per sample	4 seconds	3 seconds	5 - 20 seconds
Total duration	340.0 hours	5.6 hours	13.5 hours

IV. EXPERIMENTS

A. Data Preparation

In our experiments, we pre-train our models on the NSynth dataset [27], fine-tune the models with the IRMAS training data [3], and perform the evaluation on the IRMAS testing data. Table I shows a summary of the two datasets.

1) NSynth: The NSynth dataset is a collection of 305,979 audio snippets in 16 kHz, each lasting 4 seconds, containing an isolated monophonic musical note. Each note was held for the first 3 seconds, and for the last second, it was allowed to decay. Every note is different in instrument, pitch, or velocity and generated using one of 1,006 instruments from a commercial sample library. These 1,006 instrument labels are fine-grained, and each belongs to one of 11 instrument families: bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, synth lead, and vocal. Since not all instruments play a full range of 88 pitches, NSynth contains an average of 65.4 pitches per instrument. However, in NSynth, some notes contain a pitch unusual for the particular instrument, thus producing nearly silent sounds that are hazardous to the classification. We therefore filter out these notes by applying an energy threshold before pre-training. NSynth was released with an official split of train, validation, and test sets, and the instruments in the train set do not overlap with the ones in the validation set or test set. In our work, we formulate the pre-training process as a multi-class classification task in which the labels of samples in the validation set should be seen in the training phase. So we combine the train and validation sets in the official splits, resulting in a new train set with all 1,006 instruments. Then, 4% of the new train set is separated as a new validation set.

2) IRMAS: We use the IRMAS dataset to fine-tune and evaluate our instrument recognition model. This dataset consists of stereo recordings of professionally produced music excerpts with a sampling rate of 44.1 kHz, featuring varying instrumentation, playing styles, genres, audio quality, and production styles. Every sample in the dataset contains the segment-wise label of eleven pitched instruments, specifically cello (cel), clarinet (cla), flute (flu), acoustic guitar (gac), electric guitar (gel), organ (org), piano (pia), saxophone (sax), trumpet (tru), violin (vio), and human singing voice (voi). Note that the human voice was labeled as one of the instruments, while bass, drums, and percussion instruments were not annotated. Synthesizers, now essential ingredients in popular music, are not included in the annotations, either. IRMAS provides an official split of the dataset into training and testing sets. The training set comprises 6,705 3-second audio files, some are monophonic and others are polyphonic. But they all have only one predominant instrument label. The testing set includes 2,874 audio files, each lasting from five to twenty seconds and labeled with one or more predominant instruments. All the samples in the IRMAS dataset are downsampled to 16 kHz and converted to mono by taking the mean of the two stereo channels. Then, they are normalized to -12 LUFS.

We use the raw waveform signals as the input to our model. In the pre-training phase, only the first second of each sample in the augmented NSynth dataset is used as the fixed-length input for training and validation. There are two main reasons for doing so: first, as found in previous research [14], a 1-second duration is sufficient for the CNNs to capture the features of the instruments involved, and it was also found to be the optimal setting; second, in our preliminary experiments, we found that longer duration produced worse results. In the fine-tuning phase, for the IRMAS dataset, every 3-second audio sample in the training set is divided into three 1-second clips, resulting in 20,115 training samples. 15% of the training samples are separated to form a validation set. However, the 3-second samples in the IRMAS training set are from 2,000 distinct songs. Obviously, within each song, the samples share similar timbral characteristics. So, in creating a validation set, we make sure that there are no overlaps with songs in the training split.

B. Training Configuration

In this study, all experiments are implemented with the PyTorch framework. The source code is available online[4].

1) Pre-training: The models are pre-trained using vanilla categorical cross-entropy loss on 2 GPUs with synchronized batch normalization [28]. We set the mini-batch size to 128 and trained the model for 30 epochs. We use Adam optimizer [30], and a weight decay of 0.0005. The maximum value of the learning rate is set to 0.001. In training, the learning rate increases linearly from zero to the maximum value for the first three epochs and decreases to zero following a cosine curve. We set label smoothing [29] to 0.05. Mixup is applied with a probability of 0.5 and $\alpha = 0.3$. Concatenation augmentation was applied with a probability of 0.5, and we sliced the first second of the generated samples as the input. The effect augmentation ratio was set to 0.3. We use the model weights of the last epoch as the starting point for fine-tuning.

2) Fine-tuning: We fine-tune the models with the IRMAS training data. We use binary cross entropy loss and Sigmoid as the final activation. This is because our testing samples have one or more than one instrument labels. We set the mini-batch size to 64 and trained the model for 40 epochs on a single GPU. We use Adam optimizer and a weight decay of 0.0005. The maximum value of the learning rate is set to 0.00025. To examine the effects of pre-training, we also train models from scratch for comparison. For the models trained from scratch, we set the learning rate to 0.0035, which is the optimal setting based on the results of the preliminary experiments. We

[4]https://github.com/nii-yamagishilab/predominant-instrument-recognition

believe this leads to a fairer comparison. Again, the "warm-up" training strategy is adopted: the learning rate increases linearly from zero for the first five epochs and decreases following a cosine curve. No data augmentations were implemented during fine-tuning since we didn't observe any improvements using augmentations. We use the model weights of the last epoch for evaluation.

C. Evaluation Methodology

We use the IRMAS testing set to evaluate our model, following the evaluation settings of previous works [14][4][18]. The testing samples in the IRMAS dataset are of varying lengths. To handle this, we divide each of them into 1-second fixed-length clips with 50% overlap and take the average of clip-wise logits as the sample-wise predictions. As previously mentioned, there can be one or more predominant instrument labels, so generally, predominant instrument recognition is formulated as a multi-class multi-label classification task. We use the F1-score and Label Ranking Average Precision (LRAP) as evaluation metrics, consistent with previous research.

1) F1-score: F1-score is a commonly used metric in the field of instrument recognition. For each instrument label, F1-score is defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

and the corresponding precision and recall values for each instrument label are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

where TP is true positive, FP is false positive, and FN is false negative.

We calculate the averaged metrics over the 11 instrument labels to evaluate the overall model performance. To obtain the binary outputs, we compute the scores under a range of thresholds and report the optimal ones. The sample-wise predictions that exceed the threshold will be binarized as ones, i.e., the active labels. Since the label distribution of the IRMAS dataset is not balanced, both macro and micro-averaged metrics are calculated:

$$\text{P}_{macro} = \frac{1}{L}\sum_{l=1}^{L}\frac{\text{TP}_l}{\text{TP}_l + \text{FP}_l}, \tag{7}$$

$$\text{R}_{macro} = \frac{1}{L}\sum_{l=1}^{L}\frac{\text{TP}_l}{\text{TP}_l + \text{FN}_l}, \tag{8}$$

$$\text{P}_{micro} = \frac{\sum_{l=1}^{L}\text{TP}_l}{\sum_{l=1}^{L}(\text{TP}_l + \text{FP}_l)}, \tag{9}$$

$$\text{R}_{micro} = \frac{\sum_{l=1}^{L}\text{TP}_l}{\sum_{l=1}^{L}(\text{TP}_l + \text{FN}_l)}, \tag{10}$$

where l denotes the index of each label, and L denotes the number of labels. respectively, The macro and micro averaged F1-scores are defined as:

TABLE II
TRAINING WITH RANDOM INITIALIZATION *vs.* WITH NSYNTH PRE-TRAINING

Initialization	F1-micro	F1-macro	LRAP
Random	0.634 ± 0.0075	0.536 ± 0.0127	0.780 ± 0.0057
NSynth	0.674 ± 0.0068	0.584 ± 0.0068	0.814 ± 0.0020

$$F1_{macro} = \frac{2 \cdot \text{P}_{macro} \cdot \text{R}_{macro}}{\text{P}_{macro} + \text{R}_{macro}}, \tag{11}$$

$$F1_{micro} = \frac{2 \cdot \text{P}_{micro} \cdot \text{R}_{micro}}{\text{P}_{micro} + \text{R}_{micro}}. \tag{12}$$

The macro average considers each class to have the same weight, whereas the micro average treats each instance as having equal weight, thus assigning more weights to the classes with more samples.

2) LRAP: Proposed in [31], LRAP is a metric for multi-label classification, and it was introduced to the instrument activity detection task by Gururani *et al.* [32]. Suppose we have N samples and M labels. Given ground truth labels $y \in \{0,1\}^{N\times M}$ and predictions $\hat{f} \in \mathbb{R}^{N\times M}$, it is defined as:

$$LRAP(y,\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{||y_i||_0}\sum_{j:y_{ij}=1}\frac{|\mathcal{L}_{ij}|}{\text{rank}_{ij}}, \tag{13}$$

where $\mathcal{L}_{ij} = \left\{k : y_{ik} = 1, \hat{f}_{ik} \geq \hat{f}_{ij}\right\}$ and $\text{rank}_{ij} = \left|\left\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\right\}\right|$. $|\cdot|$ computes number of elements of the set and $||\cdot||_0$ computes the number of nonzero elements in a vector. The intuition of this metric is that our model should assign a higher score to the true labels than the other ones. Thus, LRAP does not require a threshold to have binary outputs as F1-score does.

V. EXPERIMENTAL RESULTS

We present the evaluation results on the IRMAS testing set. Note that we run each experiment three times with the same configuration except for the random seed, and report the mean value and standard deviation.

1) Effects of NSynth Pre-training: To examine the effects of NSynth pre-training, we train models from scratch with only the IRMAS training data. The comparison is shown in Table II. From this table, we can see that NSynth pre-training strongly improves the performance of downstream task, i.e. predominant instrument recognition (from 0.634 to 0.674 for F1-micro). After analyzing the average performance over all the labels, we visualize the instrument-wise micro F1-score to get more insights into how the NSynth pre-training weights benefit predominant instrument recognition. As shown in Fig. 3, NSynth pre-training boosts the system's performance in almost all the instruments except for the violin.

NSynth itself is a large dataset, and its monophonic samples share similarities with polyphonic music data to a certain extent. Thus, we are interested to see how the effectiveness changes with the amount of the IRMAS training data used

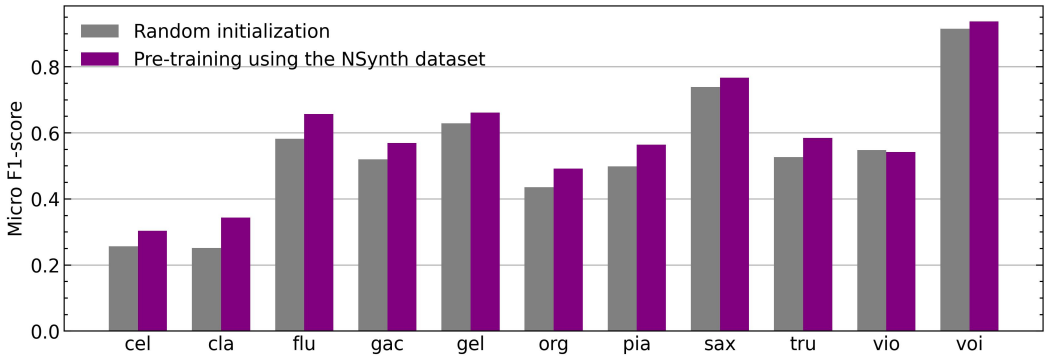


Fig. 3. Instrument-wise analysis: Training from random initialization *vs.* with NSynth pre-training.

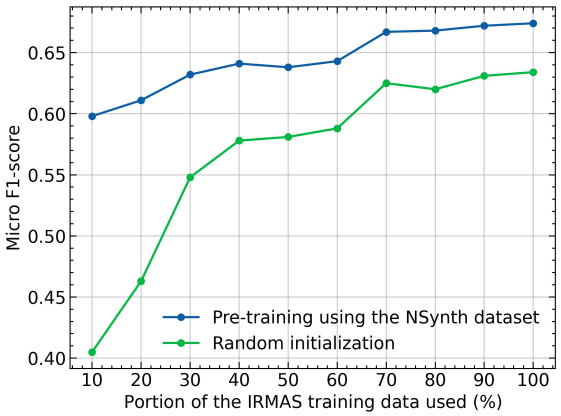


Fig. 4. Portion of the IRMAS training data used in the fine-tuning phase *vs.* evaluation results on the IRMAS testing data.

in the fine-tuning phase, in order to explore the effect of NSynth pre-trained weights a step further. Again, we report the results on the IRMAS testing set. Fig. 4 shows the micro F1-score obtained with respect to the amount of training data. Overall, the model benefits from NSynth pre-training regardless of the training data volume. The gap between pre-trained and randomly-initialized models is enlarged as the training data volume decreases. Surprisingly, with NSynth pre-trained weights and 10% of the IRMAS training data, we can train a model that is reasonable for predominant instrument recognition.

2) Effects of Data Augmentations: Next, we ablate the data augmentation methods during the pre-training phase to examine their individual contributions to the results. The ablation results are shown in Table III. While all augmentation techniques contribute to better performance, mixing two samples with soft labels has the most impact (by removing it, the micro F1-score drops from 0.674 to 0.657). And by pre-training without mixup and audio effects, the micro F1-score drops to 0.642.

3) Single Predominant Instrument Identification: As in [14], we also analyze the model's ability to identify the "most" predominant instrument of the music segment. Here we report

TABLE III
ABLATIONS OF PRE-TRAINING AUGMENTATION METHODS

Augmentations	F1-micro	F1-macro	LRAP
All	0.674 ± 0.0068	0.584 ± 0.0068	0.814 ± 0.0020
- mixup	0.657 ± 0.0029	0.560 ± 0.0045	0.804 ± 0.0040
- audio effect	0.671 ± 0.0031	0.576 ±0.0055	0.812 ± 0.0030
- both[a]	0.642 ± 0.0050	0.535 ± 0.0031	0.791 ± 0.0037
- concatenation	0.670 ± 0.0012	0.576 ± 0.0015	0.813 ± 0.0013

[a] Without mixup and audio effects

our results of the last epoch on the validation set. Note that we select the validation set to avoid similarity to the training split, so we believe that these results properly reflect the upper bound of our model's generalization ability. The overall identification accuracy is 72.6%. We present the confusion matrix in Fig. 5, and t-SNE visualization [33] of the neural embeddings, explicitly the output of the penultimate layer, in Fig. 6. From the results we can see that our model is prone to confuse the timbre of violin, flute, and clarinet with others.

4) Parameter Reduction: In the field of predominant instrument recognition, designing a reduced, lightweight model has been a research direction recently [15] [18]. In our research, considering the potential applications of our system in DAW (Digital Audio Workstation) software, we want to reduce the parameters of the model, because when producing or mixing music on our local machines, we do not want the tagging system to take up too much of the limited hardware resources.

To reduce the parameters, we run experiments with backbones that are more shallow and further adopt a shared residual block strategy (ShaResNets) proposed in [6] in consecutive residual blocks between two spatial downsampling layers. Fig. 7 presents the testing micro F1-score with respect to the models' number of parameters. We use $\{d_1, d_2, d_3, d_4\}$ to denote the number of repetitive residual blocks in n-th section in ResNet. In this figure, ResNet34 is the model we start with, ResNet28 reduces the residual layers of backbone to $\{3, 3, 4, 3\}$ and ResNet18's backbone has residual layers of $\{2, 2, 2, 2\}$. Previous works are plotted as dots for comparison. The best performance is achieved by our base model, ResNet34. Using shared blocks, we can reduce its number of parameters to 836K (64%) with 0.011 (1.6%) reduction in the

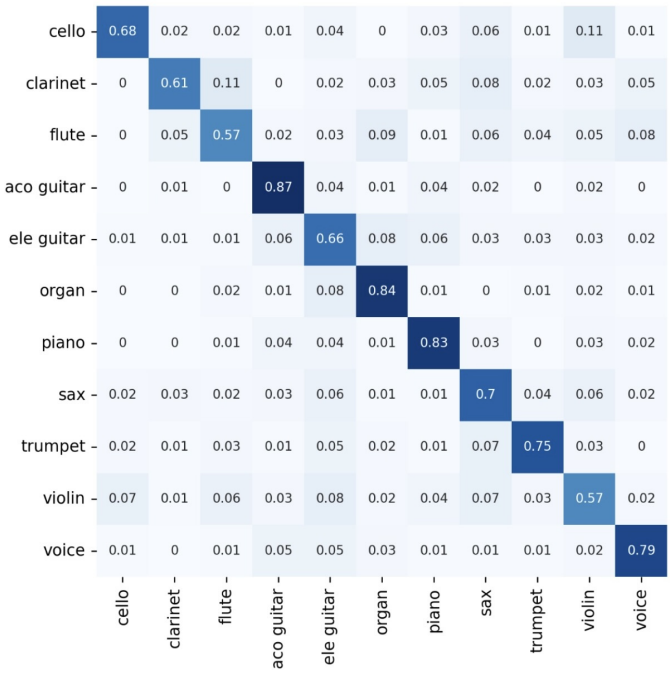


Fig. 5. Confusion matrix of single predominant instrument identification. The columns are predictions and the rows are groun truth labels. Here, "aco guitar" and "ele guitar" stand for acoustic guitar and electric guitar, respectively.

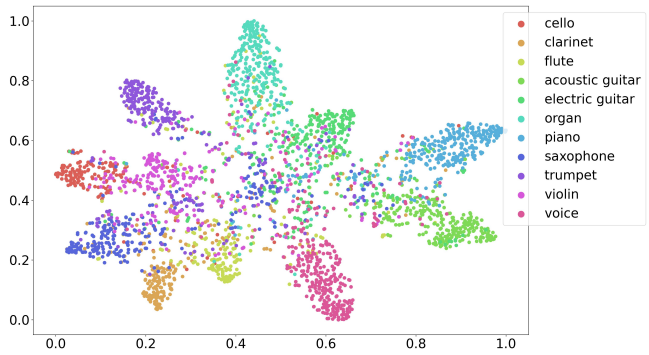
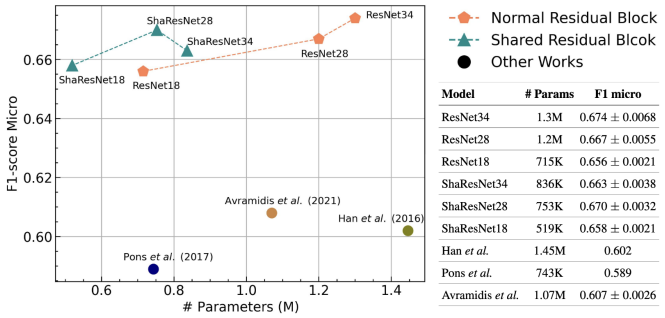


Fig. 6. t-SNE visualization of the embeddings of the validation set.

F1 score. For ResNet28 and ResNet18, sharing the weights achieves even higher performance than the original structures. ShaResNet28, which has 753K parameters, almost the same number of parameters as Pons's model in [15], has a competitive performance to our best model. Our smallest model, ShaResNet18, which only has 519K parameters, achieves a micro F1-score of 0.658, a surprising score considering its small size. Besides, from this figure, we can see a trend that as the number of parameters grows, the performance of normal residual networks grows, too. Thus we can infer that there is still room for improvement if we add more parameters to the model.

5) Comparison with Other Works: Table IV shows the micro-averaged F1-score, macro-averaged F1-score, and LRAP achieved by our proposed method and previous approaches.



Model	# Params	F1 micro
ResNet34	1.3M	0.674 ± 0.0068
ResNet28	1.2M	0.667 ± 0.0055
ResNet18	715K	0.656 ± 0.0021
ShaResNet34	836K	0.663 ± 0.0038
ShaResNet28	753K	0.670 ± 0.0032
ShaResNet18	519K	0.658 ± 0.0021
Han *et al.*	1.45M	0.602
Pons *et al.*	743K	0.589
Avramidis *et al.*	1.07M	0.607 ± 0.0026

Fig. 7. Comparison of evaluation performance between previous work and our reduced models.

TABLE IV
COMPARISON OF EVALUATION RESULTS ON THE IRMAS TESTING DATA

Methods	Features	F1-micro	F1-macro	LRAP
This work	Waveform	0.674	0.584	0.814
Avramidis *et al.* [18]	Waveform	0.608	0.543	0.747
Kratimenos *et al.* [4]	CQT	0.647	0.546	0.805
Zhong *et al.* [19][a]	Mel	0.680	0.600	0.818
Reghunath & Rajan [17]	Mel[b]	0.66	0.62	-
Yu *et al.* [16]	Mel	0.661	0.569	-
Pons *et al.* [15]	Mel	0.589	0.516	-
Han *et al.* [14][c]	Mel	0.619	0.513	-

[a] Here we show the results of single model. For the model ensemble in this work, the results were 0.688, 0.606, and 0.826 for F1-micro, F1-macro and LRAP respectively.
[b] They fused Mel-spectrogram with modgdgram and tempogram as features.
[c] Here we show their results in publication, but in preprint, the results were 0.602 and 0.503, for F1-micro and F1-macro respectively.

For all three evaluation metrics, the performance of our system produced the second best results. Our method, as an end-to-end classification approach that takes raw music data as input, outperforms the previous end-to-end system by 0.066 in micro F1-score, a 10.9% relative improvement, which is noteworthy taking into account the difficulties of this task. Also, for all three metrics, our method shows a better performance than most of the previous methods that take time-frequency representations as inputs. [19] obtained a micro F1-score of 0.680 with a vanilla ResNet-50 whose number of parameters is 25.5M. But note that in this work, the best model only has 1.3M parameters, which is no more than 5.1% of the former. Overall, this result indicates great potential in such end-to-end instrument recognition systems with trainable frontends.

VI. CONCLUSION

In conclusion, our study demonstrates the effectiveness of an end-to-end predominant instrument recognition system for polyphonic multi-instrument music. The key contributions of our research can be summarized as follows:

- A pre-training and fine-tuning approach utilizing monophonic musical note data proves effective in predominant instrument recognition.
- Data augmentation techniques during pre-training help bridge the gap between monophonic data and real-world

polyphonic music, contributing to the robustness of our model.

- Our best model achieves a micro F1-score of 0.674 and an LRAP of 0.814, marking a significant improvement of 10.9% and 8.9% relative to the previous end-to-end approach.
- A lightweight model with only 519K trainable parameters can still deliver competitive performance, demonstrating the potential for efficient deployment in various applications.

For future work, we will continue this research focusing on both model architecture as well as the challenges inherent in the data. First, we will improve the model performance by increasing the number of parameters, since the largest models we tested have only 1.3M parameters. Second, we would like to experiment with synthesized music rendered from MIDI scores as pre-training data. Moreover, we are also interested to explore this polyphonic instrument recognition problem in a zero-shot way.

ACKNOWLEDGEMENTS

We would like to thank Ms. Xuan Shi from the University of Southern California for her useful comments. This study is supported by JST CREST Grant Number JPMJCR18A6 and by MEXT KAKENHI grant 21K11951.

REFERENCES

[1] M. Schedl, E. Gómez, J. Urbano *et al.*, "Music information retrieval: Recent developments and applications," *Foundations and Trends® in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.

[2] Y.-N. Hung and A. Lerch, "Multitask learning for instrument activation aware music source separation," in *ISMIR*, 2020, pp. 748–755.

[3] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." in *ISMIR*, 2012, pp. 559–564.

[4] A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Augmentation methods on monophonic audio for instrument classification in polyphonic music," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 156–160.

[5] X. Shi, E. Cooper, and J. Yamagishi, "Use of speaker recognition approaches for learning and evaluating embedding representations of musical instrument sounds," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 367–377, 2022.

[6] A. Boulch, "Reducing parameter number in residual networks by sharing weights," *Pattern Recognition Letters*, vol. 103, pp. 53–59, 2018.

[7] I. Kaminsky and A. Materka, "Automatic source identification of monophonic musical instrument sounds," in *Proceedings of ICNN'95-International Conference on Neural Networks*, vol. 1. IEEE, 1995, pp. 189–194.

[8] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 5. IEEE, 2003, pp. V–553.

[9] J. Eggink and G. J. Brown, "Instrument recognition in accompanied sonatas and concertos," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2004, pp. iv–iv.

[10] A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Digital Audio Effects 2004*, 2004, pp. 1–1.

[11] V. Lostanlen, J. Andén, and M. Lagrange, "Extended playing techniques: the next milestone in musical instrument recognition," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, 2018, pp. 1–10.

[12] F. Fuhrmann, M. Haro, and P. Herrera, "Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music." in *ISMIR*, 2009, pp. 321–326.

[13] F. Fuhrmann *et al.*, "Automatic musical instrument recognition from polyphonic music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, 2012.

[14] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2016.

[15] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 2744–2748.

[16] D. Yu, H. Duan, J. Fang, and B. Zeng, "Predominant instrument recognition based on deep neural network with auxiliary classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 852–861, 2020.

[17] L. C. Reghunath and R. Rajan, "Transformer-based ensemble method for multiple predominant instruments recognition in polyphonic music," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–14, 2022.

[18] K. Avramidis, A. Kratimenos, C. Garoufis, A. Zlatintsi, and P. Maragos, "Deep convolutional and recurrent networks for polyphonic instrument classification from monophonic raw audio waveforms," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3010–3014.

[19] L. Zhong, N. Minematsu, and D. Saito, "Predominant instrument recognition in polyphonic music based on transfer learning with vanilla ResNet-50," in *IEICE Technical Report*. vol. 122, no. 387, EA2022-114, pp. 232-237, 2023.

[20] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5189–5193.

[23] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," in *International Conference on Learning Representations*, 2018.

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[25] ITU-R BS.1770-4, *Algorithms to measure audio programme loudness and true-peak audio level.*, Std., 2015.

[26] A. Ramires and X. Serra, "Data augmentation for instrument classification robust to audio effects," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, 2019.

[27] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1068–1077.

[28] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[31] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine learning*, vol. 39, no. 2, pp. 135–168, 2000.

[32] S. Gururani, C. Summers, and A. Lerch, "Instrument activity detection in polyphonic music using deep neural networks." in *ISMIR*, 2018, pp. 569–576.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Speech Quality Improvement Utilizing Out-of-Focus Areas in Rolling-Shutter Video on Speech Extraction

Hayata Nakano*, Tsubasa Yoshizawa*, Yuting Geng*, Kenta Iwai* and Takanobu Nishiura*
* Ritsumeikan University, Shiga, Japan
E-mail: {is0516vr@ed, is0436eh@ed, geng@fc, iwai18sp@fc, nishiura@is}.ritsumei.ac.jp

***Abstract*—We propose two methods to improve the quality of extracted speech signals utilizing the out-of-focus areas in video captured with a rolling-shutter camera. A rolling-shutter camera exposes and reads pixels from the top row of an image to the bottom, but when capturing an object that is vibrating due to speech, image distortion occurs due to the different exposure start times. The conventional method extracts a speech signal from a phase variation calculated from an image distortion. Here, we consider a case where out-of-focus areas arise in the captured video. In this case, the phase variation is not calculated correctly, which is expected to cause quality degradation of the speech signal extracted from the video. Our first proposed method weights the phase variation and the second one removes the out-of-focus areas. Experimental results show that our first method improves the quality of extracted speech signal and the second one reduces the time complexity.**

I. Introduction

In speech recording, an air-conduction microphone is typically used for converting speech signals into electrical signals when sound waves arrive at its diaphragm. As a result, the microphone acquires not only the target speech but also background noise. To solve this problem, methods [1], [2] have been proposed for extracting a speech signal from a video of an object vibrating due to speech by capturing an object with a consumer-level camera. These methods extract the speech signal by using only the captured video images. While this approach prevents the extracted speech signal from being affected by the noise near the camera, the speech recorded by the air-conduction microphone includes the noise near the microphone.

Many consumer-level cameras come equipped with complementary metal oxide semiconductor (CMOS) image sensors [3], [4] and a rolling-shutter system. The rolling-shutter system exposes and reads pixels from the top row of an image to the bottom which means that image distortion can occur due to the different exposure start times when capturing an object that is vibrating due to speech. The speech extraction method proposed in [1] utilizes this distortion by extracting a speech signal from a phase variation in each row of captured images. Here, we consider the case where out-of-focus areas arise in the captured video. Out-of-focus areas are blurred compared to an in-focus area. Therefore, when out-of-focus areas of the captured video are used for speech extraction, the phase variation cannot be calculated correctly, which degrades the quality of the extracted speech signal.

In this paper, we propose two speech extraction methods that focus on out-of-focus areas. The proposed methods calculate the weights based on the partial derivative in the captured images and process out-of-focus areas using the calculated weights. Specifically, the first method weights the phase variation to reduce the contribution rate of the out-of-focus areas and extracts the speech signal. The second method removes out-of-focus areas and extracts the speech signal. We evaluate the quality of the extracted speech signal based on short-time objective intelligibility (STOI) [5] and log-spectral distance (LSD) [6].

II. Speech Extraction Method from Rolling-Shutter Video

This section describes the conventional method [1] for speech extraction by the rolling-shutter video, which is shown in Fig. 1. This method captures RGB video of an object that is vibrating due to speech, converts it to a grayscale video $I(x, y, t)$, and then extracts a speech signal $u(n)$ from $I(x, y, t)$, where x and y are the column and row indices, t is the frame index, and n is the time index. The speech signal $u(n)$ is calculated on the basis of the phase variation of the captured video $I(x, y, t)$. The phase variation is calculated using a complex steerable pyramid [7], [8] for the captured video $I(x, y, t)$. The complex steerable pyramid consists of sub-band images for each scale r and orientation θ. Hence, $I(x, y, t)$ can be transformed by the complex steerable pyramid as

$$S(r, \theta, x, y, t) = A(r, \theta, x, y, t)e^{j\phi(r,\theta,x,y,t)}, \tag{1}$$

where j denotes the imaginary unit, $A(r, \theta, x, y, t)$ denotes the amplitude of the sub-band video, and $\phi(r, \theta, x, y, t)$ denotes its phase. $S(r, \theta, x, y, t)$ is obtained by repeating Gabor filter convolution and downsampling on $I(x, y, t)$. As the conventional method uses the phase variation of each row in each video frame, the orientation θ is fixed to 0. Therefore, $S(r, x, y, t)$, which is the decomposition of $I(x, y, t)$, is given by

$$S(r, x, y, t) = A(r, x, y, t)e^{j\phi(r,x,y,t)}, \tag{2}$$

where $S(r, x, y, t)$ is downsampled from the original resolution by 2^{-r} times the resolution for each scale r. Then, the numbers of columns and rows in the captured video $S(r, x, y, t)$ for each scale r are given by

$$W_r = 2^{-r}W, \tag{3}$$

$$H_r = 2^{-r}H, \tag{4}$$

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

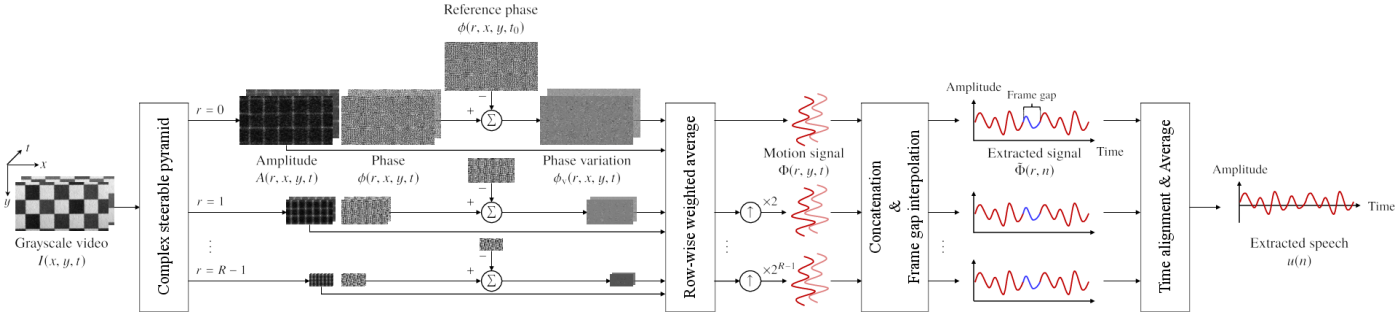


Fig. 1. Overview of speech extraction from rolling-shutter video.

where W and H denote the number of columns and rows in $I(x, y, t)$. By using the phase $\phi(r, x, y, t)$ of each row in each video frame, it is possible to extract the speech signal, as the vibration of an object will cause a phase variation [9], [10], [11]. The first step is to calculate the phase difference $\phi_v(r, x, y, t)$ from the reference frame $\phi(r, x, y, t_0)$, as

$$\phi_v(r, x, y, t) = \phi(r, x, y, t) - \phi(r, x, y, t_0), \tag{5}$$

where t_0 denotes the index representing the reference frame.

The next step is to obtain one value per row as the amplitude of the speech signal. We calculate the row-wise weighted average $\Phi(r, y, t)$ of the phase difference $\phi_v(r, x, y, t)$. Here, the square of the amplitude is used as a weight, since the phase can be accurately estimated for pixels with large amplitude. The row-wise weighted average $\Phi(r, y, t)$ is calculated as

$$\Phi(r, y, t) = \sum_{x=0}^{W_r - 1} A^2(r, x, y, t)\phi_{\mathrm{v}}(r, x, y, t). \tag{6}$$

As the length of $\Phi(r, y, t)$ is H_r for each r, the signal length H_r is aligned to H through upsampling by a factor of 2^r for each r. Here, $\Phi(r, y, t)$ for each r is a two-dimensional signal. Therefore, the averaged phase difference is transformed to a one-dimensional signal $\tilde{\Phi}(r, n)$ for each r, as

$$\tilde{\Phi}(r, n) = \Phi(r, y, t), \tag{7}$$

$$n = y + (H + N_{\mathrm{seg}})t, \tag{8}$$

where N_{seg} is the number of samples in a frame gap. The missing segment arises due to the characteristics of rolling-shutter cameras [3]. Therefore, an autoregressive model [12] is utilized to interpolate the missing segments between frames of the captured video.

The final step is to enhance the speech signal. $\tilde{\Phi}(r, n)$ from different r are synchronously added. For each scale r, a delay $d(r)$ between the reference signal $\tilde{\Phi}(r_0, n)$ and $\tilde{\Phi}(r, n)$ is calculated by

$$d(r) = \underset{m \geq 0}{\operatorname{argmin}}\ \tilde{\Phi}(r_0, n)^{\top}\tilde{\Phi}(r, n + m), \tag{9}$$

where r_0 denotes the index representing the reference scale and $\top$ denotes the transpose operator. For different r, the signals $\tilde{\Phi}(r, n)$ are added and averaged after synchronization by

$$u(n) = \frac{1}{R}\sum_{r=0}^{R-1} \tilde{\Phi}(r, n + d(r)), \tag{10}$$

where R is the number of scales of the complex steerable pyramid.

The conventional method performs speech extraction without handling out-of-focus areas. Since out-of-focus areas are blurred compared to the in-focus area as shown in Fig. 2, it is not able to correctly calculate phase variation in out-of-focus areas. In this paper, we propose methods to solve this problem.

III. SPEECH EXTRACTION METHODS UTILIZING OUT-OF-FOCUS AREAS

In this section, we propose two methods: one that weights the phase variation to reduce the contribution rate of the out-of-focus areas, and one that removes the out-of-focus areas from the captured video and extracts the speech signal.

A. Method 1:Weighted Phase Variation

As mentioned in Section Ⅱ, when out-of-focus areas of a captured video are used for speech extraction, the phase variation cannot be calculated correctly and the quality of the extracted speech signal degrades. Therefore, it is required to reduce the effect of out-of-focus area by weighting phase variation in each area. Since the absolute value of the partial derivative of an image brightness value is large at the edge of an image, weighting by the absolute value of the partial derivative leads to an emphasis on the phase variation at the edge. First, a weight $M(x, y)$ can be calculated as the absolute value of the partial derivative as

$$M(x, y) = |G_{\mathrm{X}}(x, y)|, \tag{11}$$

where $G_{\mathrm{X}}(x, y)$ is the partial derivative with respect to x and is calculated as

$$G_{\mathrm{X}}(x, y) = K(x, y) * I(x, y, t_0), \tag{12}$$

where $*$ denotes convolution and $K(x, y)$ denotes 3×3 Sobel filter in the row direction. Next, as in Section II, the phase variance $\phi_{\mathrm{v}}(r, x, y, t)$ is calculated using (1)–(5), and (6) is replaced with (13) to weight $M(x, y)$ to $\phi_{\mathrm{v}}(r, x, y, t)$. Finally,

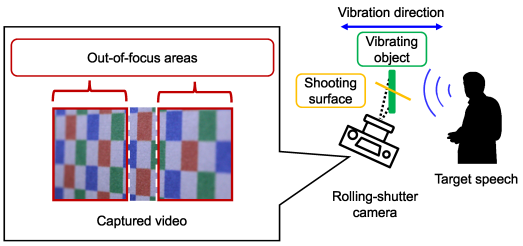


Fig. 2. Out-of-focus areas of captured video.

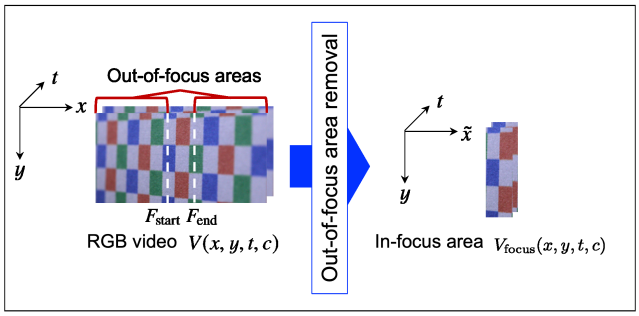


Fig. 3. Overview of out-of-focus area removal.

as in Section II, the extracted speech signal is obtained by (7)–(10).

$$\Phi(r, y, t) = \sum_{x=0}^{W_r - 1} M(x, y) A^2(r, x, y, t) \phi_{\mathrm{v}}(r, x, y, t). \quad (13)$$

B. Method 2:Removing Out-of-Focus Areas from Captured Video

As mentioned in Section Ⅱ, the phase variation cannot be calculated correctly in out-of-focus areas, thus the speech signal is extracted from the in-focus area only. When video is captured as shown in Fig. 2, the captured video is divided into in-focus area and out-of-focus areas, as shown in Fig. 3. In this case, pixels in the same column are considered to have the same weight. In addition, based on the relationship between focal length and depth of field, out-of-focus areas are composed of multiple adjacent columns. On the basis of the above, we perform out-of-focus area removal for the captured video.

First, $M(x, y)$ is calculated for each pixel of the captured video using (11) and (12). The column mean of $M(x, y)$ is calculated as

$$\bar{M}(x) = \frac{1}{H} \sum_{y=0}^{H-1} |M(x, y)|. \quad (14)$$

Next, we define a set $\boldsymbol{F}$ of column indices x such that $\bar{M}(x)$ is larger than the threshold value $\bar{M}_{\mathrm{th}}$ as

$$\boldsymbol{F} := \{x \mid \bar{M}(x) \geq \bar{M}_{\mathrm{th}}\}, \quad (15)$$

$$\bar{M}_{\mathrm{th}} = \frac{\alpha}{W} \sum_{x=0}^{W-1} \bar{M}(x), \quad (16)$$

where $\bar{M}_{\mathrm{th}}$ is calculated as α times the row means of $\bar{M}(x)$. The effect of removing out-of-focus areas can be amplified by increasing this α. Next, the column indices indicating the start point F_{start} and end point F_{end} of the in-focus region are calculated as

$$(F_{\mathrm{start}}, F_{\mathrm{end}}) = \left(\min_{x \in \boldsymbol{F}} x, \max_{x \in \boldsymbol{F}} x \right). \quad (17)$$

Finally, we remove the out-of-focus areas from the RGB video $V(x, y, t, c)$ and calculate the video $V_{\mathrm{focus}}(x, y, t, c)$ in the in-focus area as

$$V_{\mathrm{focal}}(\tilde{x}, y, t, c) = V(\tilde{x} + F_{\mathrm{start}} - 1, y, t, c), 1 \leq \tilde{x} \leq \tilde{W}, \quad (18)$$

$$\tilde{W} = F_{\mathrm{end}} - F_{\mathrm{start}} + 1, \quad (19)$$

where c denotes the index representing R, G, and B channels, $c = 0, 1$, and 2 are for the R, G, and B channels of the video, and $\tilde{W}$ denotes the number of columns in the in-focus region video. The speech signal is extracted by applying the method described in Section II to the obtained $V_{\mathrm{focus}}(x, y, t, c)$.

IV. Evaluation Experiments

We conducted evaluation experiments to investigate the effectiveness of the proposed methods. First, we show the experimental conditions for speech extraction from videos captured by a rolling-shutter camera. Next, we describe the quantitative experimental conditions and evaluation metrics for speech extraction and discuss the results. Finally, we describe the speech extraction results for one sample selected from a speech corpus [13].

A. Experimental Conditions for Speech Extraction

Figs. 4–6 respectively show the experimental setup, the equipment arrangement, and a paper with a printed pattern used as the vibrating object. Moreover, Tables I and II list the experimental conditions and equipment. As shown in Figs. 4 and 5, a paper placed in front of a loudspeaker was captured with a rolling-shutter camera, and the speech signal was extracted from the captured video. The videos were taken at an angle of 30° to the paper so that the vibrations in the row direction would be large. The pattern shown in Fig. 6 was printed on paper from the Salzburg Texture Image Database (STex) [14], which is a database of textures. In addition, we used a floodlight for illumination so that we could accurately capture the paper's vibration when the shutter speed was high. We used 50 sentences comprising five sentences for each of the ten speakers (five women, and five men) selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus [13].In this experiment, we set $t_0 = 0$ in (6) and (11), and used the initial frame taken in silence as the reference frame. We set $r_0 = 0$ in (9), with the lowest level of the complex steerable pyramid used as a reference scale. Since the speech extraction method using captured videos is based on the vibration of the object, vibration caused by non-speech elements introduces noise into the extracted speech signal. We suppressed this noise

in these experiments using the minimum mean-square error short-time spectral amplitude (MMSE-STSA) method [15].



Fig. 4. Experimental scene.

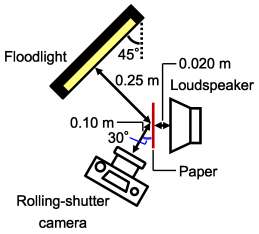


Fig. 5. Arrangement of experimental equipment.

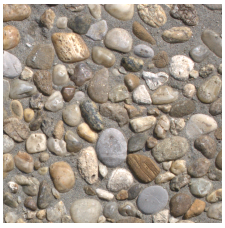


Fig. 6. Printed pattern.

TABLE I
EXPERIMENTAL CONDITION

Environment	Soundproof room (T_{60}=100 ms)
Ambient noise level	L_{A}=46.1 dB
Temperature/Humidity	21.3°C/36.0%
Sound source	10 speakers × 5 sentences from TIMIT Acoustic-Phonetic Continuous Speech Corpus[13]
Sound pressure level	90 dB at 0 m from paper
Sampling frequency / Quantization	8 kHz / 16 bits
Resolution (width × height)	1920 × 1080 px
Frame rate	60 fps
Exposure time	1/4000 s

TABLE II
EXPERIMENTAL EQUIPMENT

Camera	Canon EOS 5D MarkIV
Camera lens	Canon MP-E 65 mm f/2.8 1-5x
Loudspeaker	FOSTEX FE83En
Loudspeaker amplifier	BOSE 1705II
Floodlight	GOODGOODS GDGDS-WL02 (10000 lm)

B. Quantitative Experiments

We evaluated the quality of extracted speech by short-time objective intelligibility (STOI) [5] and log-spectral distance (LSD) [6]. STOI is a measure to evaluate the intelligibility of noise-suppressed speech. It takes a score from 0 to 1 with a higher score indicating higher intelligibility of evaluated speech. LSD is a difference between spectral envelope of a clean speech and that of an evaluated speech. The lower the LSD score, the smaller the difference between the clean speech and the evaluated speech, indicating that the evaluated spech is closer to the clean speech.

First, we evaluated the extracted speech signal extracted which by the weighted phase variation method. Figs. 7 and 8 show the results of evaluating the intelligibility of the extracted speech evaluated by STOI and the quality of the extracted speech evaluated by LSD, where the STOI and LSD values are the average scores of 50 sentences. Error bars indicate the standard deviation of each score. Paired-sample t-tests [16] on the STOI and LSD scores for the conventional and proposed methods at a significance level of 5% revealed no significant differences between the STOI scores, but significant differences between the LSD scores. In addition, the LSD score of the proposed method was on average 0.089 dB smaller than that of the conventional method, as shown in Fig. 8. This finding demonstrates that the quality of extracted speech can be improved by using the proposed method compared to the conventional method.

Next, we evaluated the speech signal extracted by the out-of-focus area removal method. In this experiment, we evaluated the cases $\alpha = 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$, and 3.5 in (16). $\alpha = 0.0$ is the case where the speech is extracted from the captured video without out-of-focus area removal. As shown in Figs. 7 and 8, both the STOI scores and LSD scores of the proposed and conventional methods were comparable for all α. Paired-sample t-tests [16] on the STOI scores for the conventional and proposed methods at a significance level of 5% revealed no significant differences and similar results were obtained for the LSD scores. This finding demonstrates that the quality of the extracted speech is not improved by the removal of out-of-focus areas.

C. Time Complexity

Table III shows that the number of columns of captured videos decreases as α increases. By reducing the number of columns, the number of convolutions between the captured images and the Gabor filter in the speech extraction process is reduced, and thus the time complexity can be reduced. Specifically, the processing time for 50 sentences of the conventional method takes about 70 minutes using MATLAB on a computer with Intel(R) Core(TM) i9-11950H @2.60GHz and 32 GB of RAM. On the other hand, the processing time for 50 sentences of the proposed method under $\alpha = 1.5$ takes about 20 minutes in the same environment as the conventional method. These results indicate that the method of out-of-focus area removal reduces the time complexity without degrading the quality of the extracted speech signal.

TABLE III
NUMBER OF ROWS OF VIDEO IN EACH α

α	0.0 (Original)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
In-focus area $\tilde{W}$	1920	1910	1052	326	232	201	123	48

D. Experimental Results of Speech Extraction

As a final evaluation, we examined one sample picked up from the speech corpus [13]. Spectrograms of the clean speech, the speech extracted by the conventional method, and the

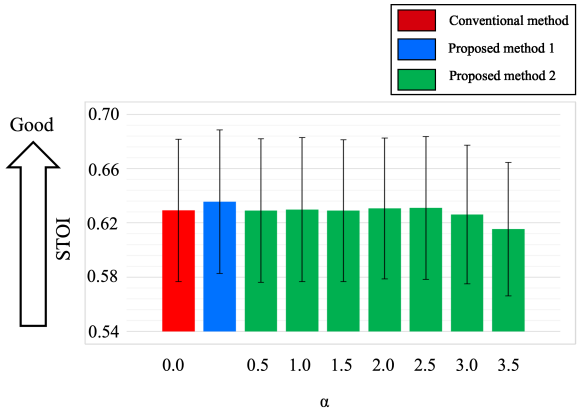


Fig. 7. Evaluation results with STOI.

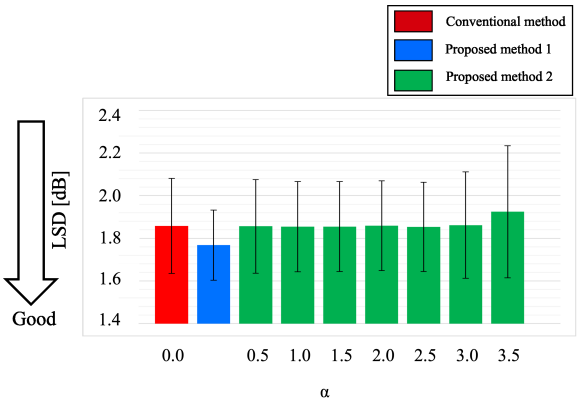
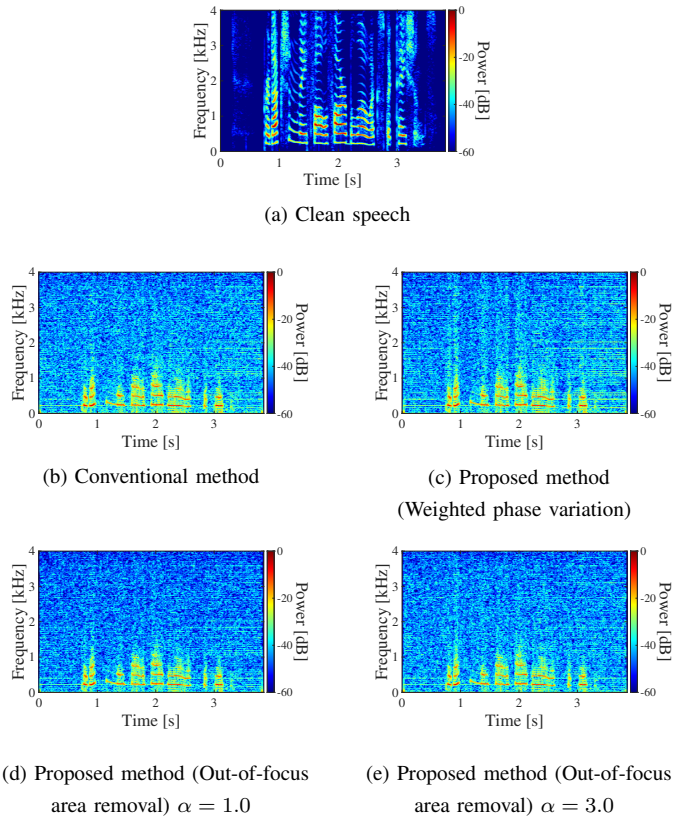


Fig. 8. Evaluation results with LSD.



(a) Clean speech

(b) Conventional method

(c) Proposed method (Weighted phase variation)

(d) Proposed method (Out-of-focus area removal) $\alpha = 1.0$

(e) Proposed method (Out-of-focus area removal) $\alpha = 3.0$

Fig. 9. Spectrograms of speech.

speech extracted by the method of weighted phase variation are shown in Fig. 9. As we can see in Fig. 9 (c), the speech extracted by the proposed method contains more harmonic components in the high-frequency band than the conventional method. These results demonstrate the difficulty of measuring high-frequency vibrations in the out-of-focus areas. In Figs. 9 (d) and (e), the structure of the extracted speech in each α is generally the same, even though the number of video columns W varies depending on the method used. These findings demonstrate that the method of out-of-focus area removal can reduce the time complexity without degrading the quality of the extracted speech.

Overall, experimental results show that proposed method 1 improves the speech quality of extracted speech signal and proposed method 2 reduces the time complexity.

V. Conclusion

In this paper, we proposed two speech extraction methods that focus on out-of-focus areas to improve the quality of the extracted speech. The first method weights the phase variation, thereby reducing the contribution of the out-of-focus areas, and the second method removes the out-of-focus areas of the captured video and extracts the speech signal. Our experimental findings demonstrate that utilizing the out-of-focus areas in speech extraction is effective in improving sound quality or reducing time complexity. In future work, we aim to improve the method of weight calculation in the proposed method to achieve even higher speech quality.

Acknowledgment

This work was partly supported by Ritsumeikan University R-GIRO and RARA, and JSPS KAKENHI Grant Numbers JP19H04142, JP21H03488, JP21H04427, and JP23H03425.

References

[1] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–10, Jul. 2014.

[2] T. Yoshizawa, Y. Geng, K. Iwai, and T. Nishiura, "Evaluation of Speech Extraction from Rolling-Shutter Video by Measuring Vibration Vertical to Shooting Surface," In *Acoustics 2022, Online*, Oct. 2022.

[3] O. Ait-Aider, N. Andreff, J. M. Lavest, and P. Martinet, "Exploiting Rolling Shutter Distortions for Simultaneous Object Pose and Velocity Computation Using a Single View," In *Proceedings of Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, p. 35, Jan. 2006.

[4] H. Sumi, T. Narabu, and S. Saito, "For the better image quality of CMOS image Sensor," *IEICE ESS Fundamentals Review*, vol. 3, no. 3, pp. 44–51, Jan. 2010.

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[6] L. R. Rabiner and B. H. Juang, "*Fundamentals of speech recognition*, "Prentice Hall, Aug. 1993.

[7] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.
[8] J. Portilla and E. P. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *International Journal of Computer Vision*, vol. 40, pp. 49–70, Oct. 2000.
[9] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-Based Video Motion Processing," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–10, Jul. 2013.
[10] T. Gautama and M. Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1127–1136, Sep. 2002.
[11] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 188–200, Mar. 2002.
[12] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, Apr. 1986.
[13] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium, Philadelphia, PA, USA*, Feb. 1993.
[14] Multimedia Signal Processing and Security Lab, University of Salzburg, Salzburg, Austria, "Salzburg Texture Image Database (STex)," https://wavelab.at/sources/STex/, Accessed: Jan. 2023.
[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[16] Student, "The Probable Error of a Mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Personalized Audio Quality Preference Prediction

Chung-Che Wang, Yu-Chun Lin, Yu-Teng Hsu, and Jyh-Shing Roger Jang
Dept. of Computer Science and Information Engineering, National Taiwan Univ., Taiwan
E-mail: geniusturtle6174@gmail.com, {yuchun.lin, pooh.hsu, jang}@mirlab.org

***Abstract*—This paper proposes to use both audio input and subject information to predict the personalized preference of two audio segments with the same content in different qualities. A siamese network is used to compare two inputs and predict the preference. Several different structures for each side of the siamese network are investigated. The baseline structure which uses only audio information involves using a pretrained audio encoder followed by fully connected layers. In several different proposed structures, the approach of concatenating subject information with audio embedding before feeding it into fully connected layers outperforms the baseline model the most, resulting in an increase in overall accuracy from 77.56% to 78.04%. Experimental results also demonstrate that utilizing the complete set of subject information, which includes age, gender, and headphone/earphone specifications such as impedance, frequency response range, and sensitivity, is more effective than using a subset of this information.**

I. Introduction

Automatic audio quality measurement can be applied in many different audio related tasks, including detecting fake-quality audio [1], monitoring speech quality for online conference [2], or assessing the naturalness of synthesized speech [3]. Besides, since many people use mobile phones in their daily life, and the quality of audio playback is one of the considerations of choosing mobile phones [4], measuring the quality of audio playback for mobile phones would also be one of the important tasks.

On the other hand, while some research focused on predicting the mean score of users[5], methods that aimed to make personalized predictions have also been studied. Saquil et al. classified users into different categories and used the categorical labels as one of the inputs of their network to rank the preference of different video segments [6]. Xu et al. added personal information and personality test results of subjects to the input features to their models to predict the preference of sad music [7].

While Saquil et al. [6] and Xu et al. [7] focused on predicting the preference of different audio or video content, Leng et al. [8] and Hunag et al. [9] focused on predicting the preference for the same content across different qualities. Leng et al. proposed mean-bias network (MBNet), which was composed of a MeanNet accepting only audio spectrum to predict the mean opinion score (MOS), and a BiasNet in parallel with MeanNet accepting both audio spectrum and subject's ID to predict the possibly biased scores judged by individual subjects [8]. To improve MBNet, Hunag et al. [9] proposed listener-dependent network (LDNet), which used an encoder-decoder structure to predict the listener-dependent score, where the encoder accepted the input speech, and the decoder accepted the encoded vector and the subject's ID to output the listener-dependent score. However, the largest limitation of their methods is that the input of the subject's information is simply their ID, thus the subjects cannot be unseen at the inference stage.

To our best knowledge, despite the impact of subject information on listening tests has been discussed [10], there is no research focused on adding subjects' personal information to the prediction model for audio quality preference. In this paper, we aim to predict the personalized preference of two audio segments played by different mobile phones, i.e. preference of two audio segments with the same content in different qualities. Model performance of using several network structures and types of personal information are investigated. The rest of this paper is organized as follows. Section II describes the process of dataset collection, Section III describes our proposed methods, Section IV shows the experimental setting and results, and Section V concludes this paper and addresses future work.

II. Dataset

To obtain audio clips of different qualities, we select some song segments and use some devices to play the segments, and record them. In this study, seven song segments selected from English, Mandarin, Japanese, and Korean pop songs are used for playback by five mobile phones with different brands. Each song segment is about 10 to 15 seconds long and is played by each mobile phone with two different volumes. The first volume is each individual phone's max volume, which is about 78 to 84 dB for different phones. The second volume is called normal volume, which is about 70 to 75 dB for different phones. A 3DIO Free Space XLR binaural microphone and an ART USB Dual Pre Project Series computer interface are used for recording. Configurations for recording, including the distance between mobile phone and microphone and volume gain, are kept the same for all song segments and all mobile phones.

The recorded files are then used to make questionnaires. The five recorded audio files of each volume of each song segment form ten pairs, where each of the two audio files in one pair corresponds to different mobile phones, and the actual order of the two audio files in the questionnaire is randomly assigned. Therefore, each of the two volumes has 70 pairs. These 70 pairs are divided into five questionnaires, where each questionnaire contains 14 pairs corresponding to the seven different songs.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

Since the preference of two audio files with similar qualities may be difficult to judge but the preference of two audio files with dissimilar qualities may not, we try to balance the total difficulties of questionnaires by dividing the 70 pairs carefully. The five files are first sorted based on the prejudged order of qualities of their corresponding mobile phones. After that, the 70 pairs can be viewed as a composition of 28, 21, 14, and 7 pairs respectively with distances of 1, 2, 3, and 4, where the 28 pairs are assumed to be the most difficult ones, and the 7 pairs are assumed to be the easiest ones. For each difficulty, the corresponding pairs are assigned into questionnaires as balanced as possible. Finally, each questionnaire contains at least 1 pair belonging to the easiest group, and at most 6 pairs belonging to the most difficult group.

In each audio pair of each questionnaire, subjects will choose which one is better according to their own preference. The preference score is originally an integer in the range of [1, 5], where 1 and 2 indicates that the subject prefers the first audio, 3 indicates that no preference for any one of the two audio files, and 4 and 5 indicates that the subject prefers the second audio. For convenience, we will translate the scores to the range of [-2, 2] for further description and analysis, where -2 and -1 indicates that the subject prefers the first audio, 0 indicates that no preference for any one of the two audio files, and 1 and 2 indicates that the subject prefers the second audio.

Finally, 31 subjects joined the test, including 27 males and 4 females. Their ages range from 21 to 46 years, where 21 of them are between the ages of 21 and 25. The total number of questions answered by subjects is 2,856. Beside, subjects are asked to provide the brand and model of headphones or earphones they used for this test, and the specifications of headphones or earphones, including impedance, lower and upper bound of frequency response, and sensitivity will be collected from the internet. After removing the data of subjects where more than half of the specifications of headphones or earphones they used cannot be found, and the questions of which preference score were answered as 0 (i.e. no preference for any one of the two audio files), the number of remaining subjects is 23, and the number of preference of audio pairs being answered is 2,000. The number of answers of ± 2 and ± 1 are respectively 936 and 1,064. The whole dataset is public available[1].

III. Methods

Since the primary objective of this study is to predict the preference between two audio segments, using a network with two input audio segments is not feasible. This is because the output preference may not remain consistent when the order of inputs is changed. Therefore, we employ the siamese network structure to ensure that the output preference remains the same, regardless of the order of input. Fig. 1 (a) shows the overall structure of the siamese network, where the inputs are passed through two networks sharing the same weights, and the outputs of the two networks are concatenated together,

[1]https://zenodo.org/record/8293869

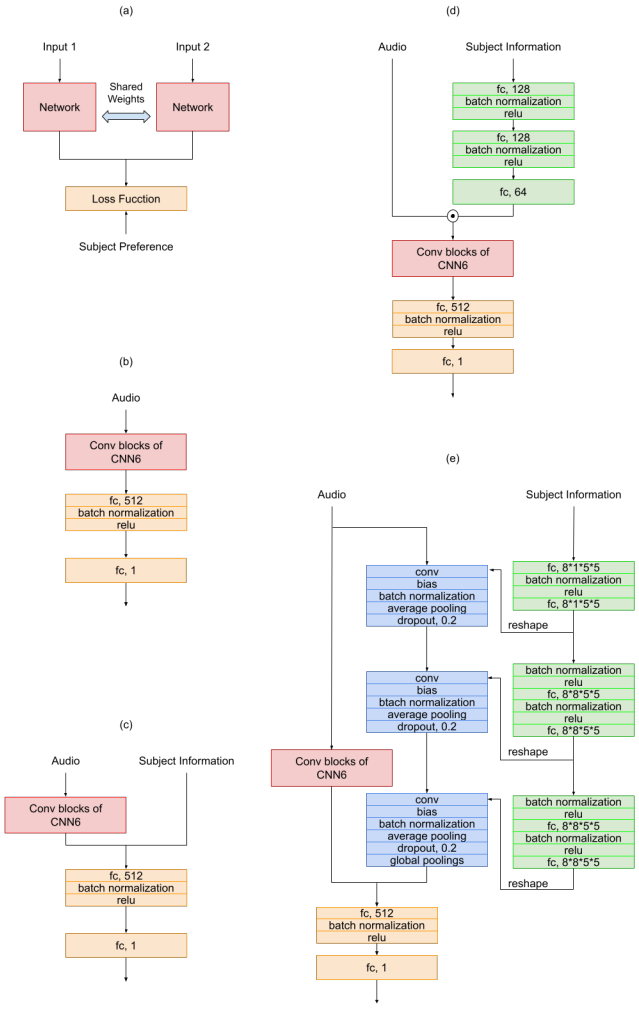


Fig. 1. Model structure. (a) Siamese network. (b) The network uses audio input only (**AO**). (c) The network uses both audio input and subject information, fused later than the CNN6 (**A+S-L**). (d) Audio input and subject information fused earlier than the CNN6 (**A+S-E**). (e) Subject information is used as the kernels of convolution to process audio input in parallel with CNN6 (**A+S-P**).

passed through a softmax function, and used for further loss calculation. For the two sides of the siamese network, the side with the larger output means that the corresponding input audio is preferred by a subject. Note that the network is trained to predict the preference of qualities of two audio segments for one subject, thus the audio segment input to the two sides are in different qualities, but the subject information input to the two sides are the same.

Fig. 1 (b) shows the baseline network which uses only audio features on each side. The pretrained model "CNN6" in PANNs [11], which accepts a log-mel spectrogram of 64 mel bins as input, is used as the basic structure. The original fully-connected layers of CNN6 were replaced with a multi-layer perceptron (MLP) block containing a linear layer with 512 output nodes (denoted by "fc, 512"), a batch normalization layer, an activation function of ReLU, and a linear layer with one output node (denoted by "fc, 1"). We name this structure

AO in this paper.

Fig. 1 (c) shows the first network using both audio input and subject information on each side. The subject information contains one dimension of age, one dimension of gender, and four dimensions of specifications of headphones or earphones, i.e. impedance, lower and upper bound of frequency response, and sensitivity. The value of -1 will be used for missing values of specifications of headphones or earphones. The six-dimensional vector is concatenated with the output of the convolutional block of CNN6, and fed into consequent layers together. Not that this network can be viewed as an LDNet with CNN6 as the encoder and an MLP block as the decoder, except that the input subject information is not only their ID, and this network will be combined with another identical one and viewed as a siamese network. We name this structure **A+S-L** (L for late fusion) in this paper.

To investigate the possibility of different methods for using both audio input and subject information, two more structures of networks are introduced in this paper. Fig. 1 (d) and Fig. 1 (e) respectively show the second and the third network using both audio input and subject information on each side. In the network in Fig. 1 (d), the subject information is passed through an MLP block, multiplied with audio input, and then fed into remaining parts. In this network, the subject information acts like an equalizer for the corresponding audio signal. We name this structure **A+S-E**, where E is for early fusion. In the network in Fig. 1 (e), the subject information is also passed through some MLP blocks (light green blocks in the figure), but the output of these MLP blocks are reshaped and used as the convolution kernel for input audio, and the convolutional blocks (light blue blicks in the figure) are designed to be similar to that of CNN6. This path is in parallel with the original CNN6, and the outputs of this path and CNN6 will be concatenated together for further processing. We name this structure **A+S-P**, where P is for parallel.

The sample rate of audio files are originally 44.1 kHz and are reampled to 32 kHz for the use of CNN6. Frame size and hop size are respectively 32 ms and 10 ms. SpecAugment [12] is applied for the training data online during the training process for data augmentation, where the number of stripes for both time and frequency axes is 2, and the maximal drop widths are 64 for the time axis and 4 for the frequency axis. The negative log-likelihood is used as the loss function, and the ADAM optimizer is used to optimize the network. The initial learning rate is 5×10^{-5} and will be decreased by a factor of 0.95 after each epoch. The batch size is 64. In all of the models, the CNN6 will be fine-tuned.

IV. Experimental Results

A. Experimental Setup

To test the ability of predicting the preference for unseen subjects, we divide all the 23 subjects into 7 folds. All the subjects are sorted by their age, and when the subjects of index k are used for testing, the subjects of index $k+1 \ mod \ 7$ are used for validation, and the remaining subjects are used for training. The number of epochs for training is 50, and the training process will be early-stopped if the validation loss does not achieve a new lower value for consequently 10 epochs. The evaluation metric is accuracy, which is defined by the number of pairs of which the preferences are correctly predicted divided by the total number of test pairs.

Experiments are conducted in a container running on a machine with Intel(R) Xeon(R) Gold 6154 CPU, occupying 4 CPU cores, 90 GB of host memory, and a Tesla V100-SXM2 GPU.

B. Results

We first compare the performances of different model. The test accuracies of the 7 folds and the overall accuracies are shown in Table I, where the results of each model in each fold are the average of 21 runs, and the overall accuracies are calculated by the weighted average of accuracies of all the folds. As shown in the table, **A+S-L** performs better than **AO** on average, where the accuracy grows from 77.56% to 78.04%. For individual folds, higher mean values are achieved in 5 out of the 7 folds, lower standard deviations are achieved in 6 out of the 7 folds, and mean values are higher and standard deviations are lower in 4 out of the 7 folds. On the other hand, performance of **A+S-E** is significantly lower than other models, and **A+S-P** achieves similar mean values to **AO** and **A+S-L**, but the standard deviations are higher, which is possibly due to that the network is more complex but the amount of data is not very large. We do not compare to other LDNet structures which use different encoders and decoders because the amount of data is currently not very large, thus may not be suitable for training more complex network structures.

To observe the importance of both audio input and subject information, the weight matrices of the linear layers of the last MLP block of **A+S-L** and **A+S-P** are respectively multiplied, and the mean absolute values are shown in Fig. 2, where the first 512 values are for audio information processed by CNN6, and the remaining 6 or 8 values are for raw subject information in **A+S-L** or subject information convolved with audio information in **A+S-P**. For **A+S-L**, the first 512 values are around 0.4, and the remaining values are around 0.02. This difference of scales may simply be due to the fact that the values of audio information processed by CNN6 are normalized but the values of raw subject information are not. Therefore the importance of audio and subject information may be similar. For **A+S-P**, both audio and subject information are normalized, but the mean absolute values of multiplied weight matrices are at different levels (0.03 for audio and 0.01 for subject information), so the subject information may not be utilized very well in **A+S-P**, but collecting more data to train a more complex model may improve the accuracy.

Test accuracies of **A+S-L** using different subsets of subject information are shown in Table II. When using only age and gender as the subject information, the model performs slightly worse than that of using other subsets of subject information in most of the folds. On the other hand, when

TABLE I

THE TEST ACCURACIES OF THE 7 FOLDS AND THE OVERALL ACCURACIES FOR DIFFERENT MODELS. NUMBER OF PREFERENCE QUESTIONS ANSWERED BY SUBJECTS IS DENOTED BY "#Q". THE RESULTS OF EACH MODEL IN EACH FOLD ARE THE AVERAGE OF 21 RUNS. THE MEAN VALUE AND STANDARD DEVIATION IS DENOTED BY THE FORM OF "MEAN±STD".

Fold (#Q) \ Model	AO	A+S-L	A+S-E	A+S-P
1 (403)	0.8029 ± 0.0151	**0.8056 ± 0.0069**	0.5757 ± 0.0462	0.8098 ± 0.0188
2 (269)	0.7836 ± 0.0121	**0.7946 ± 0.0062**	0.6887 ± 0.1071	0.7874 ± 0.0132
3 (318)	**0.7843 ± 0.0118**	0.7808 ± 0.0066	0.6499 ± 0.1241	0.7719 ± 0.0117
4 (215)	**0.6742 ± 0.0134**	0.6670 ± 0.0118	0.5826 ± 0.0698	0.6684 ± 0.0151
5 (289)	0.7569 ± 0.0157	**0.7670 ± 0.0091**	0.6753 ± 0.0537	0.7161 ± 0.0539
6 (305)	0.7761 ± 0.0063	**0.7864 ± 0.0076**	0.5633 ± 0.1340	0.7775 ± 0.0069
7 (201)	0.8316 ± 0.0171	**0.8420 ± 0.0141**	0.5714 ± 0.0538	0.8380 ± 0.0164
Overall	0.7756	**0.7804**	0.6155	0.7699

TABLE II

TEST ACCURACIES OF **A+S-L** USING DIFFERENT SUBSETS OF SUBJECT INFORMATION. **IMPD**: IMPEDANCE. **SENSIT**: SENSITIVITY. **FREQ**: FREQUENCY.

Fold (#Q) \ Subject Info	all	age & gender	all specs of head/earphones	impd & sensit of head/earphones	freq responses of head/earphones
1 (403)	0.8056 ± 0.0069	0.8117 ± 0.0177	0.8060 ± 0.0103	**0.8134 ± 0.0197**	0.8073 ± 0.0154
2 (269)	0.7946 ± 0.0062	0.7775 ± 0.0134	**0.7959 ± 0.0098**	0.7809 ± 0.0104	0.7864 ± 0.0059
3 (318)	0.7808 ± 0.0066	0.7811 ± 0.0108	0.7767 ± 0.0089	**0.7848 ± 0.0224**	0.7774 ± 0.0081
4 (215)	0.6670 ± 0.0118	**0.6712 ± 0.0107**	0.6709 ± 0.0114	0.6668 ± 0.0183	0.6633 ± 0.0135
5 (289)	0.7670 ± 0.0091	0.7330 ± 0.0113	**0.7690 ± 0.0108**	0.7631 ± 0.0255	0.7681 ± 0.0093
6 (305)	0.7864 ± 0.0076	0.7807 ± 0.0071	**0.7866 ± 0.0064**	0.7672 ± 0.0121	0.7831 ± 0.0070
7 (201)	**0.8420 ± 0.0141**	0.8224 ± 0.0138	0.8398 ± 0.0094	0.8403 ± 0.0181	0.8400 ± 0.0138
Overall	0.7804	0.7721	**0.7806**	0.7771	0.7782

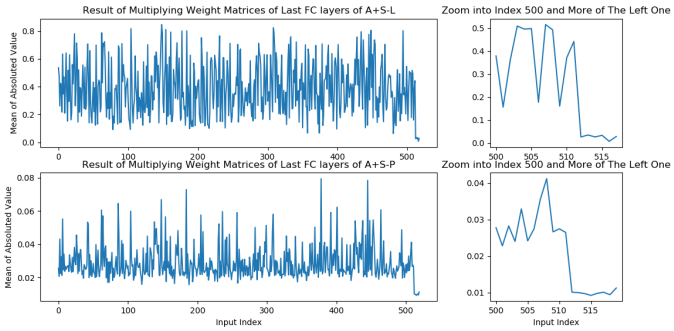


Fig. 2. Mean absolute values of multiplied results of weight matrices of the linear layers of the last MLP block of **A+S-L** (upper plots) and **A+S-P** (lower plots), where the left and right plots are respectively all the values and values of and after index 500.

using all specifications of headphones or earphones as the subject information, the accuracies are slightly higher than that of using all the available subject information, but the standard deviations are also slightly higher. Above results suggest that the model performs more stable when all the available subject information is used. However, using only specifications of headphones or earphones may also be a good choice when age and gender information of subjects is not available.

V. CONCLUSIONS AND FUTURE WORK

This paper proposes to make personalized prediction of the preference of two audio segments with the same content but are in different qualities using both audio input and subject information. Siamese network is used to compare two sets of input, and the side with larger output indicates that the corresponding input audio is preferred. For each side of the siamese network, the performances of several different structures are investigated, and concatenating subject information with audio embedding in between audio encoder and fully connected layers outperforms a baseline model using only audio input the most, where the overall accuracy grows from 77.56% to 78.04%. The effectiveness of different types of subject information is also investigated, and using all the available subject information, including age, gender, and the specifications of headphones or earphones, is found to be the most stable.

Since currently the ages of most subjects concentrate in a small range and most of the subjects are male, recruiting more subjects with different ages and genders in the future may help with a better discussion of the effectiveness of age and gender to predict the personalized preference of audio quality. Besides, since the environment and other information for listening devices including life and power of headphones or earphones may also affect the audio preference, collecting such information may also help predict the audio preference better.

VI. ACKNOWLEDGEMENTS

This work was supported by the ASUSTek Computer Inc. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

REFERENCES

[1] Jinglei Zhou, Rangding Wang, Chao Jin, and Diqun Yan, "Detecting fake-quality wav audio based on phase differences," in *International Workshop on Digital Watermarking*. Springer, 2014, pp. 525–534.

[2] Gaoxiong Yi, Wei Xiao, Yiming Xiao, Babak Naderi, Sebastian Möller, Wafaa Wardah, Gabriel Mittag, Ross Cutler, Zhuohuang Zhang, Donald S Williamson, et al., "Conferencingspeech 2022 challenge: Non-intrusive objective speech quality assessment (nisqa) challenge for online conferencing applications," *arXiv preprint arXiv:2203.16032*, 2022.

[3] Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A Saurous, and D Sculley, "Automos: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv:1611.09207*, 2016.

[4] KV Sriram, H Mahesh Prabhu, and Aditi Ajay Bhat, "Mobile phone usability and its influence on brand loyalty and re-purchase intention: An empirical," in *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2019, pp. 1–4.

[5] Pranay Manocha, Zeyu Jin, and Adam Finkelstein, "Sqapp: No-reference speech quality assessment via pairwise preference," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 891–895.

[6] Yassir Saquil, Da Chen, Yuan He, Chuan Li, and Yong-Liang Yang, "Multiple pairwise ranking networks for personalized video summarization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1718–1727.

[7] Liuchang Xu, Ye Zheng, Dayu Xu, and Liang Xu, "Predicting the preference for sad music: the role of gender, personality, and audio features," *IEEE Access*, vol. 9, pp. 92952–92963, 2021.

[8] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[9] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda, "Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 896–900.

[10] Tim Walton and Michael Evans, "The role of human influence factors on overall listening experience," *Quality and User Experience*, vol. 3, no. 1, pp. 1–16, 2018.

[11] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[12] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

AVATAR: Robust Voice Search Engine Leveraging Autoregressive Document Retrieval and Contrastive Learning

Yi-Cheng Wang, Tzu-Ting Yang, Hsin-Wei Wang, Bi-Cheng Yan, Berlin Chen
Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E-mail:{yichengwang, tzutingyang, hsinweiwang, bicheng, berlin}@ntnu.edu.tw

***Abstract*—Voice, as input, has progressively become popular on mobiles and seems to transcend almost entirely text input. Through voice, the voice search (VS) system can provide a more natural way to meet user's information needs. However, errors from the automatic speech recognition (ASR) system can be catastrophic to the VS system. Building on the recent advanced lightweight autoregressive retrieval model, which has the potential to be deployed on mobiles, leading to a more secure and personal VS assistant. This paper presents a novel study of VS leveraging autoregressive retrieval and tackles the crucial problems facing VS, viz. the performance drop caused by ASR noise, via data augmentations and contrastive learning, showing how explicit and implicit modeling the noise patterns can alleviate the problems. A series of experiments conducted on the Open-Domain Question Answering (ODSQA) confirm our approach's effectiveness and robustness in relation to some strong baseline systems.**

***Index Terms*—Voice search, Information retrieval, Autoregressive information retrieval, Contrastive learning**

I. Introduction

Compared to desktop and laptop computers, flourishing with recent advances in speech technology, voice-based input is gradually replacing text input as the primary method of interaction between humans and machines on small-screen mobile devices [1]. Through voice inputs, how to provide users with an effective mechanism to access the content they want in an overwhelming amount of information available on the Internet, viz. voice search (VS), has gained a place with other spoken language technologies at the center of the stage. VS is related to, but distinct from spoken document retrieval (SDR) [2], [3]. Compared to SDR, in which text queries are given to the system and need to search for relevant spoken documents meanwhile facing essential issues, the queries are usually too short to conveying the information needs. VS searches for relevant text documents using a more natural modality, viz. voice, that lead to longer queries, thus better-expressing information needs [4].

Over the years, many efforts have been devoted to investigating deep neural network-based retrieval methods, showing good promise in many IR and SDR tasks. These models can be applied to various VS scenarios. Recently, dense retrieval models have become the primary neural-based document retrieval approach, complementary to sparse retrieval models such as TF-IDF [5] and BM25 [6], that matches keywords efficiently with an inverted index. Dense retrieval models [7], [8] leverage the pre-trained language models to encode information in latent semantic space, thus better capturing the semantic relationships between queries and documents. A typical dense retrieval construct of a dual encoder contains two independent neural networks optimized for embedding the queries and documents. The advantage of the dual encoder design is that the entire corpus can be encoded and indexed offline. At the inference time, the score of a query-document pair can be efficiently computed as the inner product of the corresponding query and document embeddings.

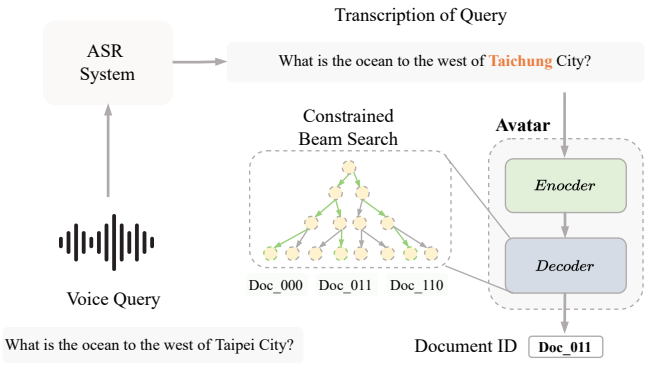


Fig. 1. Flow diagram of our proposed Avatar VS system. Given a voice query, Avatar first uses an ASR system to obtain a text query (errors generated from the ASR system were denoted in orange). Then, an autoregressive retriever directly generates the ranked list of relevant document ID (docid) through constrained beam search.

Recently, another novel autoregressive retrieval approach orthogonal to the dual encoder models dubbed Differentiable Search Index (DSI)[9] has been proposed. DSI first encodes all the information about the corpus into a single transformer model's parameters space, on top of which DSI can generate the relevant document identities (docids) in an autoregressive manner in response to a user query. DSI has many advantages over the dual encoder models: 1) It avoids using only dot products, which could miss the fine-grained interaction between the query and the document meta information. 2) It lowers memory requirements; storing dense vectors for the whole corpus requires a large memory footprint. Although

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

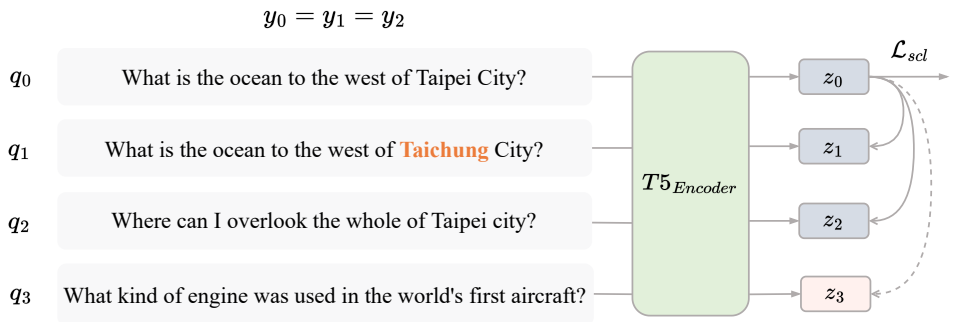


Fig. 2. Pre-training: supervised contrastive learning (SCL). SCL takes data of the same class as positive samples and consists of three query types: queries from the original dataset, pseudo queries generated from the query generation model, and augmented queries.

with the benefits mentioned above, DSI suffers serious data distribution problems during the model training and inference phase. [10]. Specifically, DSI learns to build connections between long document texts and their docids, but at inference time, relatively short queries are input into the model to retrieve their relevant docids. Differentiable search index with query generation (DSI-QG) [10] migrates this problem using another query generation model to generate relevant pseudo queries from documents. It uses these short pseudo queries instead of long text documents to build the connection with their docids in the training phase. After incorporating these pseudo queries [11], the autoregressive retrieval model has become effective while requiring less memory footprint and has potential to be deployed on the mobile or edge device, making it a more secure and personal retrieval system.

Nevertheless, as far as we are concerned, the autoregressive retrieval model has not been sufficiently and systematically studied in neither ad-hoc information retrieval (IR) nor VS, and its retrieval effectiveness is mostly unknown. Based on this background, we present an empirical VS evaluation that sheds light on the efficacy of the autoregressive retrieval model in this paper. Further, we address one of the critical problems with VS, viz. noises in the query caused by the ASR system has an enormous impact on the retrieval model [12], [13], by means of explicitly modeling the ASR noise pattern using data augmentation and implicitly teaching the model to distinguish the features invariant to noise using contrastive learning. A series of experiments conducted on the Open-Domain Spoken Question Answering dataset (ODSQA) [14] confirm our approach's effectiveness and robustness in relation to some strong baseline systems.

II. Methodology

In this paper, we propose Avatar, a robust voice search engine leveraging **a**utoregressive retrie**va**l and con**t**rastive le**ar**ning. Figure 1 shows the main workflow. Given a voice query, Avatar first uses an ASR system to transcribe voice queries into a text query. Then, an autoregressive retriever built from standard transformer architecture directly generates the relevant docids through constrained beam search. In this section, we first introduce autoregressive retrieval and then demonstrate our approach to robust the model.

A. Autoregressive Retrieval

Unlike classic retrieval techniques, autoregressive retrieval methods use a sequence-to-sequence (seq2seq) language model for encoding all the corpus information into the model's parameter space. After receiving the user query, beam search is used to generate the rank list of docids. Specifically, the autoregressive retriever ranks each document $d \in D$ in the corpus by computing a score with an autoregressive formulation:

$$score(d|q) = p_\theta(y|q) = \prod_{m=1}^{M} p_\theta(y|y_{1<m}, q), \tag{1}$$

where q is the query entered by the user, y is the set of M tokens in docid of d, and θ represents the parameters of the model. Original DSI models suffer from data distribution problems during the model training and inference phase. To migrate this problem, we follow the DSI-QG to use the pseudo queries generated from a seq2seq query generation model and incorporate them into the model training phase. In other words, we generate pseudo queries Q_{qg} from their relevant documents using a query generation model. Following that we combine them with the original queries Q_{sup} from the training data to form the new training sets $q \in Q_{seq} = \{Q_{sup} \cup Q_{qg}\}$. After that, we train the model by the general seq2seq training objective and teach forcing:

$$\mathcal{L}_{seq}(\theta) = -\sum_{q \in Q_{seq}} \log p_\theta(y|q). \tag{2}$$

A well-built docid must be able to identify the different documents while reflecting their semantic information. Since exploring different docids is not the focus of this study, we adopt the semantic docid proposed in DSI and leave the extension of docids for future work. At first, semantic docid

uses a BERT language model [15] to encode all the documents in the corpus to obtain the semantic vectors. Second, the hierarchical clustering algorithm is employed to cluster the semantically similar documents in the same group in a hierarchical fashion. Finally, we can assign each document the identifier by group number by traversing the hierarchical tree. The docids generated by the beam search do not necessarily exist in the corpus. Inspired by [16], we use the constrained beam search to guide the decoder to search in a limited tokens space for each step to generate the valid docids. Concretely, we define constraints based on a prefix tree built on all docid strings.

B. Explicit: Data Augmentation

Data augmentation (DA) is a simple and effective method often used to strengthen models by explicitly exposing them to data containing ASR noise and clean data so that model learning remains invariant to noise. Specifically, we generate three random augmentations for each query in Q_{seq}: $Q_{da} = Aug(Q_{seq})$, where $Aug(\cdot)$ is a data augmentation module similar to [17], we generate random substitution, deletion, or insertion errors. For substitution, we use similar phonological words to substitute them. Finally, we train the model using the training sets consisting of $Q = \{Q_{seq} \cup Q_{da}\}$.

C. Implicit: Contrastive Learning

Contrastive learning (CL) can help the model distinguish the invariant features in the ASR noise and clean text query. Firstly, we pre-train the model's encoder to bring the original queries closer to the augmented queries and push the others away. By closely looking into the autoregressive retrieval model, we find that it's fundamentally similar to the sequence classification task. Thus bringing the same class close together in the latent space can benefit the classification task later.

Supervised Contrastive Learning (SCL)[18] takes data of the same class as positive samples and pulls their embeddings closer together. In the end, representations from the same class form a clustering effect and discriminate different classes by the margins created between them. It is more suitable to use SCL for the autoregressive retrieval model to enhance its robustness against ASR errors. Given a min-Batch of N queries, random sample from the Q training sets, $B = \{(q_i, y_i)\}_{i=1..N}$. We first obtain the representation of each query through the model's encoder $Enc(\cdot)$ and use the $Proj(\cdot)$ projection network, $z = Proj(Enc(q))$ to obtain the sequence-level representation. Let $i \in I = \{1, ..., N\}$ be the index of B. The following equation can describe our SCL objective:

$$\mathcal{L}_{scl} = -\sum_{i \in I} \frac{1}{|S(i)|} \sum_{s \in S(i)} \log \frac{\exp(z_i \cdot z_s)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a)} \quad (3)$$

Here $\cdot$ denotes inner product operations, $A(i) = I \setminus \{i\}$ is the set of index I minus i, $S(i) = \{s \in A(i) : y_s = y_i\}$ is the index of all positive samples of index i, and $|S(i)|$ is its cardinality. The process is illustrated in Figure 2.

TABLE I
DATA STATISTICS OF ODSQA-TEST. WE EMPLOY THREE DIFFERENT ASR SYSTEMS TO UNDERSTAND BETTER THE INFLUENCES CAUSED BY DIFFERENT WER-LEVEL AND ENTITY ERROR RATES (EER).

		ODSQA(Test)	
ASR Model	**Train Sets**	**WER**	**EER**
iFLYTEK	-	10.61%	19.75%
Conformer Mask-CTC	In-house	15.06%	42.68%
Conformer	Aishell	23.89%	50.87%

D. Combine

Combining the above Explicit and Implicit approaches makes our proposed Avatar model more robust when encountering ASR noises. Overall, we first use SCL Eq. (3) to pre-train the model's encoder and then use DA with general seq2seq objective Eq. (2) to fine-tune the model further.

III. EXPERIMENTS

A. Experimental Setting

Dataset and Evaluation. We used Open-Domain Spoken Question Answering (ODSQA) for our experiments. ODSQA consists of 30,461 query-document pairs, whose the queries are natural language, and the 2,051 documents are from Delta Reading Comprehension Dataset (DRCD) [19]. Since ODSQA only releases 1,465 query audios equipped with official ASR transcriptions from the ODSQA-test, we adopt these as our testing set and the remaining 28,996 query-document pairs as our training set. To observe the influences on the model caused by different WER-level and Entity Error Rates (EER), in addition to official ODSQA ASR transcriptions, we apply another two ASR systems to obtain the testing set's transcriptions. Specifically, we used a Conformer Mask-CTC ASR system trained on In-house Data and a Conformer ASR system on Aishell [20]. Detail summarized in Table I.

Like the original DSI, we utilize Hits@1 and Hits@10 to evaluate the effectiveness of the baselines and the model. This metric reports the proportion of the correct docid ranked in the top 1 and top 10 predictions.

Baselines. We compare Avatar with the following baselines: 1) Okapi BM25: A classic sparse retrieval method bases on the inverted index. 2) DPR: A dual encoder dense retriever that trained with contrastive loss and hard negatives. 3) DSI: An autoregressive retrieval method that uses document texts as input for indexing. 4) DSI-QG: An improved version of DSI that mitigates the data distribution mismatch problems by using generated queries as inputs for indexing.

Implementation details In this study, we use two multilingual T5(mT5)[21] base models provided by huggingface for our system, one as the Avatar model and the other as the Query Generation model. We apply the same training method as that of DSI-QG for the query generation model and generate three relevant pseudo queries for each document. For the Avatar model pre-training, we employ the same

TABLE II
THE MAIN RETRIEVAL RESULTS OF AVATAR AND SOME WELL-KNOWN RETRIEVAL MODELS.

	ODSQA(Test)							
	Clean		WER 10%		WER 15%		WER 23%	
Model	**Hit@1**	**Hit@10**	**Hit@1**	**Hit@10**	**Hit@1**	**Hit@10**	**Hit@1**	**Hit@10**
BM25	34.90	53.55	29.82	45.25	22.62	37.18	19.95	30.69
DPR	49.11	**70.15**	39.59	59.59	29.46	49.96	22.76	41.55
DSI	43.16	60.17	32.71	49.04	25.76	42.03	19.27	33.01
DSI-QG	46.51	64.95	37.22	54.64	28.43	44.77	22.76	38.07
Avatar	**52.73**	69.12	**43.23**	**60.17**	**36.02**	**53.86**	**29.18**	**44.90**

TABLE III
ABLATION STUDY OF DIFFERENT WER-LEVEL.

	ODSQA(Test)					
	WER 10%		WER 15%		WER 23%	
Model	**Hit@1**	**Hit@10**	**Hit@1**	**Hit@10**	**Hit@1**	**Hit@10**
w/o Data Augm.	37.22	54.64	28.43	44.77	22.76	38.07
w/o SCL	39.75	58.33	31.85	50.85	26.17	42.31
Avatar	**43.23**	**60.17**	**36.02**	**53.86**	**29.18**	**44.90**

TABLE IV
ABLATION STUDY OF DIFFERENT EER-LEVEL.

	ODSQA(Test) - WER 10%			
	Non-entity Utterances		Entity Utterances	
Model	**Hit@1**	**Hit@10**	**Hit@1**	**Hit@10**
w/o Data Augm.	41.16	58.37	27.86	41.25
w/o SCL	43.98	62.38	27.04	46.17
Avatar	**46.90**	**64.38**	**32.24**	**47.54**

$MeanPooling$ as that of [22] to obtain the sequence-level represents. We utilize the same mT5 model as a fair comparison for all the learning-based models in the baseline. Since the document length in the corpus is too long for the model to accommodate, we only keep the document title and the first 100 tokens as the model's inputs.

B. Main Results

The evaluation performance is presented in Table 3. Based on term matching, we find that the BM25 model has the lowest performance among all models, showing the importance of assessing the semantic information. By looking at DSI-QG, incorporating the pseudo queries can effectively improve the model's retrieval ability compared to the autoregressive DSI model. Moreover, the noise caused by the ASR system can cause catastrophic impacts on all the baselines model, including the strong DPR model. The magnitude of the drop also rises with the increase of WER. Through data augmentation and supervised contrastive learning, our proposed Avatar can alleviate the influences from the imperfection of the ASR transcriptions and not only keep the performance in a clean environment but also simultaneously increase the retrieval ability in contrast to other strong baseline systems.

C. Ablation Study

Study of different WER. We now focus on evaluating the effectiveness of different ASR noise-robust methods in the model and show the performance in Table III. Not surprisingly, explicit modeling of the noise pattern, viz. data augmentation, has the biggest effect on the model's performance. Through implicitly teaching the model to distinguish the invariant features between the ASR noise and clean query text, viz. pre-trained with SCL objective, we can further enhance the model's robustness, after combine these two, the model's performance increases vastly.

Study of different EER. The entities mentioned in the query significantly impact the success of retrieving the relevant documents. To further test our proposed Avatar's robustness, we split the testing sets into two subsets, one contains the ASR noise in the entity mentioned in the queries, and the other includes the noise in the non-entity position, results shown in Table 4. First, the model's retrieval effect is dramatically reduced when ASR noise occurs in Entity queries, which illustrates the importance of the entities mentioned in the query for retrieval. We find that adding data augmentation and contrastive learning improves the effectiveness of our model.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a novel robustness method to improve the performance of the autoregressive retrieval model when exposed to noisy ASR transcriptions, whose effectiveness has also been validated and analyzed through a series of empirical experiments. The proposed model Avatar sheds light on the potential of an on-device VS engine, which can bring more convenience, security, and personal experience. In future work, we intend to first tackle the more intensive tasks which are entities error caused by ASR system and second extend the model to a large scale that requires designing the autoregressive retrieval with more capacity.

References

[1] F. Crestani, S. Mizzaro, and I. Scagnetto, "Mobile information retrieval," *CoRR*, vol. abs/1902.01790, 2019.

[2] L. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 42–60, 2005. DOI: 10.1109/MSP.2005.1511823.

[3] S. Fan-Jiang, T. Lo, and B. Chen, "Spoken document retrieval leveraging bert-based modeling and query reformulation," IEEE, 2020, pp. 8144–8148. DOI: 10.1109/ICASSP40776.2020.9052910.

[4] I. Guy, "Searching by talking: Analysis of voice queries on mobile web search," R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, and J. Zobel, Eds., ACM, 2016, pp. 35–44. DOI: 10.1145/2911451.2911525.

[5] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975. DOI: 10.1145/361219.361220.

[6] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments - part 1," *Inf. Process. Manag.*, vol. 36, no. 6, pp. 779–808, 2000. DOI: 10.1016/S0306-4573(00)00015-7.

[7] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over BERT," J. X. Huang, Y. Chang, X. Cheng, *et al.*, Eds., ACM, 2020, pp. 39–48. DOI: 10.1145/3397271.3401075.

[8] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 329–345, 2021. DOI: 10.1162/tacl_a_00369.

[9] Y. Tay, V. Q. Tran, M. Dehghani, *et al.*, "Transformer memory as a differentiable search index," *CoRR*, vol. abs/2202.06991, 2022. arXiv: 2202.06991.

[10] S. Zhuang, H. Ren, L. Shou, *et al.*, "Bridging the gap between indexing and retrieval for differentiable search index with query generation," *CoRR*, vol. abs/2206.10128, 2022. DOI: 10.48550/arXiv.2206.10128. arXiv: 2206.10128.

[11] Y. Zhou, J. Yao, Z. Dou, L. Wu, P. Zhang, and J. Wen, "Ultron: An ultimate retriever on corpus with a model-based indexer," *CoRR*, vol. abs/2208.09257, 2022. DOI: 10.48550/arXiv.2208.09257. arXiv: 2208.09257.

[12] G. Sidiropoulos, S. Vakulenko, and E. Kanoulas, "On the impact of speech recognition errors in passage retrieval for spoken question answering," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, M. A. Hasan and L. Xiong, Eds., ACM, 2022, pp. 4485–4489. DOI: 10.1145/3511808.3557662.

[13] Y. Chang and Y. Chen, "Contrastive learning for improving ASR robustness in spoken language understanding," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds., ISCA, 2022, pp. 3458–3462. DOI: 10.21437/Interspeech.2022-781.

[14] C. Lee, S. Wang, H. Chang, and H. Lee, "ODSQA: open-domain spoken question answering dataset," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, IEEE, 2018, pp. 949–956. DOI: 10.1109/SLT.2018.8639505.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[16] N. D. Cao, G. Izacard, S. Riedel, and F. Petroni, "Autoregressive entity retrieval," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[17] Y. Leng, X. Tan, L. Zhu, *et al.*, "Fastcorrect: Fast error correction with edit alignment for automatic speech recognition," M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 21 708–21 719.

[18] P. Khosla, P. Teterwak, C. Wang, *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[19] C. Shao, T. Liu, Y. Lai, Y. Tseng, and S. Tsai, "DRCD: a chinese machine reading comprehension dataset," *CoRR*, vol. abs/1806.00920, 2018. arXiv: 1806.00920.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline," IEEE, 2017, pp. 1–5. DOI: 10.1109/ICSDA.2017.8384449.

[21] L. Xue, N. Constant, A. Roberts, *et al.*, "Mt5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, *et al.*, Eds., Association for Computational Linguistics, 2021, pp. 483–498.

[22] J. Ni, G. H. Ábrego, N. Constant, *et al.*, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, 2022, pp. 1864–1874. DOI: 10.18653/v1/2022.findings-acl.146.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Regression-based Sound Event Detection with Semi-supervised Learning

Chia-Chuan Liu,* Chia-Ping Chen*
Chung-Li Lu,† Bo-cheng Chan,† Yu-Han Cheng,† Hsiang-Feng Chuang,† Wei-Yu Chen†
* National Sun Yat-sen University, Taiwan
E-mail: m103040063@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw
† Advanced Technology Laboratory, Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan
{chungli, cbc, henacheng, gotop, weiweichen}@cht.com.tw

***Abstract*—We propose an end-to-end regression-based semi-supervised approach to the problem of sound event detection (SED) in this paper. As a segment-level method, the proposed approach to SED treats the prediction of the temporal location of acoustic events as a regression problem. It alleviates issues with classification-based approaches based on frame-level classification followed by smoothing operation. Treating SED as a regression problem, we design machine-learning systems to learn the relation between onset, duration, and acoustic features. To leverage unlabeled data in model training, we propose a simple and effective training process with a mean-teacher framework to improve the utilization of training data. We implemented two regression-based SED systems with CRNN and VGGSK models. On the validation set of Domestic Environment Sound Event Detection Dataset, the regression-based systems achieve respectively the event-based F1 scores of 0.480 and 0.523, and the intersection-based F1 scores of 0.673 and 0.728. Both systems outperform the classification-based systems using the same backbone network with event-based F1 scores of 0.429 and 0.463 and intersection-based F1 scores of 0.643 and 0.704.**

I. Introduction

Sound event detection (SED) aims to identify acoustic event classes and locate the corresponding time boundaries within audio clips. SED can be used for acoustics-based applications, e.g. smart homes/speakers, where systems detect acoustic events and take appropriate actions.

Detection and Classification of Acoustic Scenes and Events task 4 (DCASE Task 4), which focuses on semi-supervised learning in SED, has been one of the major contests in recent years. Most participants in this task have adopted classification-based approaches [1]. Such approaches are based on frame-by-frame classification, in which the core machining-learning problem is to predict the event-wise probability in each frame. One issue with such approaches is that oscillatory predictions occur when the probability is near the threshold. Smoothing operations such as median filtering are often necessary to stabilize system outputs.

To locate sound events directly, we propose a regression-based approach for sound event detection. Inspired by the You-Only-Look-Once (YOLO) object detection algorithm [2], sound events are detected via bounding boxes. The proposed approach utilizes an end-to-end model which predicts frame-by-frame bounding boxes in the frame. Each bounding box prediction conveys temporal and prominent information for sound events. As onset is more distinct than center in acoustic events, bounding boxes are designed to convey the onset time and duration time. Since regression based on anchor boxes can stabilize the localization of object boundaries [3]–[6], the duration is determined with reference to anchor box priors. Furthermore, since DCASE Task 4 involves semi-supervised learning for SED, we also implement a training framework using the mean teacher framework [7]. During training, we employ a teacher-student model architecture. The teacher model is in charge of providing pseudo-labels, which are used in supervising the student model. By minimizing the difference between the outputs of the teacher and student models, we effectively leverage the unlabeled data.

SED based on a regression approach has attracted attention in recent years. You Only Hear Once (YOHO) [8], [9] is an end-to-end regression-based supervised SED system that performs well on TUT [10] and UrbanSound [11] datasets. SEDT [12] is a DETR-based [13] SED system for DCASE Task 4, which locates the temporal center and duration of sound events through regression. The semi-supervised SEDT (SS-SEDT) utilizes a teacher-student architecture with exponential moving average (EMA) updates. The training process of SS-SEDT consists of two phases: the burn-in phase and the teacher-guided phase. In the burn-in phase, SEDT is trained with labeled data. In the subsequent teacher-guided phase, the trained SEDT is duplicated into two models, and maintains consistency training on unlabeled data. Compared to YOHO, SEDT and SS-SEDT, our proposed approach has the following novelties. First, our approach is based on anchor box regression to respond to sound events with varying duration. Second, the semi-supervised training process is simpler compared to SS-SEDT, as it requires only one-stage training. Lastly, the approach is flexible and compatible with CRNN, which is the mainstream architecture in DCASE Task 4.

The rest of this paper is organized as follows. We introduce our methods in Section II. We describe the datasets and the experiment setting in Section III. The experiment results are shown in Sections IV. The concluding remarks are summarized in Section V.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

II. METHODS

We introduce classification-based and regression-based approaches for SED. They are followed by the description of the loss functions, neural-network models, and the post-processing methods that are used in our systems.

A. Classification-based Detection

In the common classification-based approach, the output prediction $\hat{Y}$ is of size $N \times C$, where N is the number of frames and C is the number of event classes, and $\hat{y}_{i,c} \in [0,1]$ represents the probability of class c at frame i. After binarizing the prediction with a threshold and smoothing it with a median filter, the event classes and the corresponding temporal boundaries can be generated by finding consecutive active frames. Additionally, the clip-level prediction $\hat{y}_w$ can be derived by aggregating the frame-level prediction in $\hat{Y}$. A common method is attention pooling [14], where the probability of event class $\hat{y}_w(c)$ is weighted average of $\hat{Y}$

$$\hat{y}_w(c) = \frac{\sum_{i=1}^{N} w_{i,c}\hat{y}_{i,c}}{\sum_{i=1}^{N} w_{i,c}} \tag{1}$$

where $w_{i,c} \in [0,1]$ is a learnable attention weight for $\hat{y}_{i,c}$.

B. Regression-based Detection

In our regression-based approach, we replace the frame-level classification with bounding box prediction. An one-dimensional bounding box is considered a basic event prediction unit for a sound event. Each frame is assigned B hypothetical bounding boxes. Duration prediction is specified via anchor box priors. The prediction $\hat{y}_{i,j}$ for bounding box j with onset in frame i consists of the following components.

- $\hat{y}_{\text{conf}}$: confidence score in $[0,1]$. A bounding box is considered as a candidate event only if $\hat{y}_{\text{conf}}$ exceeds a threshold τ_{conf} (set as 0.5 in this paper).
- $\hat{y}_s, \hat{y}_d$: output variables for the onset and duration of a bounding box. Specifically

$$\hat{y}_s = i + \sigma\left(\hat{t}_s\right), \ \hat{y}_d = p_j e^{\hat{t}_d}. \tag{2}$$

  That is, $\hat{y}_s$ is specified by offset to frame index i and $\hat{y}_d$ is specified by anchor box prior p_j.
- $\hat{y}_{\text{cls}}$: probability for each event class.

Thus, the size of the output prediction is $N \times B \times (3+C)$ for an input of N frames.

During training, each bounding box in prediction $\hat{y}$ is specified as foreground, background, or ignored case. For a ground-truth event with onset in frame i, the bounding box j with maximum Intersection-over-Union (IoU) is assigned as a foreground case. The confidence score and the class probability are set to 1. The onset and duration are set by Eq. (2). The other bounding boxes are assigned as background cases and their confidence scores are set to 0. To alleviate the imbalance between the foreground and the background cases, the background cases with a specified condition will be ignored. When the IoU between the background case and the foreground case reaches a threshold τ_{ignored} (set as 0.7 in this paper), the background case would not calculate the loss. For inference, we only consider the bounding boxes with confidence score larger than τ_{conf}, i.e. the candidates.

As with classification-based SED systems, our regression-based systems can also produce clip-level predictions $\hat{y}_w$. We apply max pooling to aggregate classification scores of all bounding boxes. The equation is as follows:

$$\hat{y}_w(c) = \max_{i \in N, j \in B} \hat{y}_{c,i,j}. \tag{3}$$

C. Loss Functions

The total loss consists of the supervised loss, weak-supervised loss and the consistency loss

$$L_{\text{total}} = L_{\text{supervised}} + L_{\text{weak-supervised}} + L_{\text{consistency}}. \tag{4}$$

The supervised loss $L_{\text{supervised}}$ is based on the YOLO loss

$$L_{\text{supervised}} = L_{\text{coord}} + L_{\text{conf}} + L_{\text{cls}} \tag{5}$$

where $L_{\text{coord}}, L_{\text{conf}}$ and L_{cls} denote the coordinate loss, confidence loss and classification loss, respectively. We modify the coordinate loss term L_{coord} to adapt to the one-dimensional localization scenario. Given a set of strong annotated labels Y and a bounding box prediction $\hat{Y}$, the coordinated loss is

$$L_{\text{coord}} = \sum_{i=1}^{N}\sum_{j=1}^{B} \mathbb{1}_{ij}^{obj}[(\hat{y}_{s,i,j} - y_{s,i,j})^2 + (\hat{y}_{d,i,j} - y_{d,i,j})^2]. \tag{6}$$

The remaining terms L_{conf} and L_{cls} are the same as in YOLO, i.e., L_{conf} is the mean-square error (MSE) of confidence scores

$$L_{\text{conf}} = \sum_{i=1}^{N}\sum_{j=1}^{B} \mathbb{1}_{ij}^{obj}(\hat{y}_{c,i,j} - y_{c,i,j})^2, \tag{7}$$

L_{cls} is the binary cross entropy (BCE) of classification scores

$$L_{\text{cls}} = \sum_{i=1}^{N}\sum_{j=1}^{B} \mathbb{1}_{ij}^{obj} \sum_{c \in C} \text{BCE}(\hat{y}_{c,i,j}, y_{c,i,j}), \tag{8}$$

where $\mathbb{1}_{ij}^{obj}$ is an indicator with value 1 when bounding box j in frame i is a foreground case and 0 otherwise. The weak-supervised loss $L_{\text{weak-supervised}}$ is based on the weak clip-level labels. Specifically,

$$L_{\text{weak-supervised}} = \text{BCE}(\hat{y}_w, y_w) \tag{9}$$

where $\hat{y}_w, y_w$ are model prediction and ground-truth label. To utilize the unlabeled data, the system is implemented with the mean teacher framework. We employ two identical models during training, referred to as the teacher model θ' and the student model θ. The parameter of the teacher model is updated by the student model with an exponential moving average, i.e.

$$\theta_t^{'} = \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_{t-1} \tag{10}$$

where t represents the training step and α is a weight in $[0,1]$. The consistency loss $L_{\text{consistency}}$ depends on student/teacher model prediction via

$$L_{\text{consistency}} = L_{\text{pos}} + L_{\text{neg}} + L_{\text{weak-consistency}}. \tag{11}$$

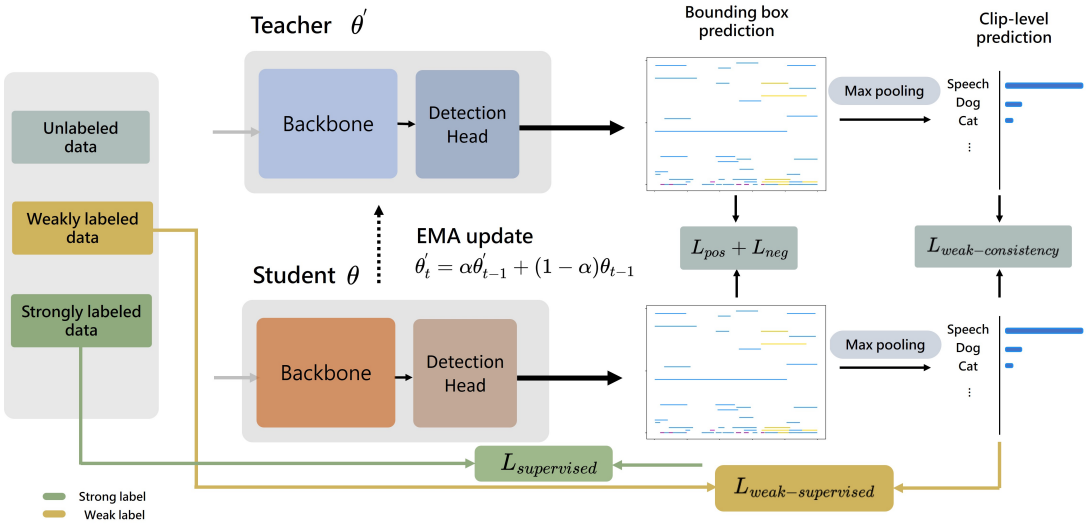


Fig. 1. The training process for the regression-based system: The parameter of the teacher model is updated by the student model with an exponential moving average. During training, teacher model is in charge of providing reliable predictions for unlabeled data, and student model is trained. The consistency loss is the difference between the predictions of the two models.

To mitigate the impact of low-quality pseudo labels, the bounding boxes produced by the teacher model are divided into positive/negative cases based on confidence threshold $\tau_p = 0.8$. For the positive cases, L_{pos} is the MSE between the model predictions for all confidence, coordinate and classification. For the negative cases, L_{neg} is the MSE for only the confidence scores. $L_{\text{weak-consistency}}$ is the BCE of the clip-level predictions. The training process is illustrated in Figure 1.

D. Model Architecture

Our regression-based SED network is composed of a backbone and a detection head. The former extracts features from input data, and the latter produces bounding boxes for each frame. In this paper, we implement two regression-based models with different backbones. One is CRNN [15], the baseline model for DCASE 2022 Task 4, that consists of a convolution neural network followed by a recurrent neural network. The other is VGGSK [16], a popular system on DCASE Task 4, that uses selective kernel units [17] to enhance performance.

E. Post-processing

The set of candidate events (bounding boxes with high confidence scores) often contains redundancy. De-duplication is necessary for regression-based systems. How to effectively select the correct bounding boxes dramatically affects performance. Non-maximum suppression (NMS) algorithm is commonly used in image object detection [18], [19] to eliminate redundant bounding boxes. However, in our system we did not apply NMS directly due to the polyphonic situation.

We propose two approaches to properly select bounding boxes for sound event detection. The first approach is called bounding box mutual exclusion, in which we select the bounding box with the highest score for each class. The second approach is called bounding box union, in which we concatenate bounding boxes with high classification scores for each class. This is shown in Figure 2.

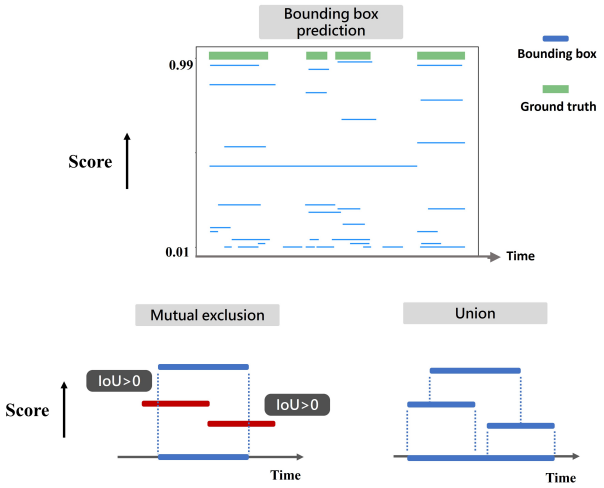


Fig. 2. Bounding box mutual exclusion and bounding box union for post-processing candidate events. The upper of the figure presents an example of a regression-based system inferring an audio clip. Bounding box mutual exclusion discards other overlapping bounding boxes. Bounding box union concatenates overlapping bounding boxes.

F. Data augmentation

We apply mixup data augmentation [20]. Artificially labelled data points created through linear combination, i.e.

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda)x_j \\ \hat{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{12}$$

where x_i, x_j are randomly sampled data, y_i, y_j are the corresponding labels, and $\hat{x}, \hat{y}$ are generated data and label.

TABLE I

RESULTS OF DIFFERENT DETECTION APPROACHES ON DESED VALIDATION SET: THE TABLE SHOWS THE RESULTS ON VALIDATION SET WITH DIFFERENT DETECTION APPROACHES. MODELS WITH * ARE PROPOSED BY OTHER TEAMS IN DCASE 2022 TASK 4. MARK "-" INDICATES THAT THE RESULT IS NOT SHOWN IN THE SOURCE PAPER.

	Method	**CB-F1**	**IB-F1**	**PSDS1**	**PSDS2**	**Model size**
FDY-CRNN* [21]	Classification	0.489	0.720	0.405	0.638	11.1M
SEDT*[12]	Regression	-	-	0.415	0.582	-
SS-SEDT*[12]		-	-	0.388	0.573	
CRNN	Classification	0.429	0.643	0.351	0.552	1.1M
	Ours	**0.480**	**0.673**	**0.384**	**0.562**	
VGGSK	Classification	0.463	0.704	0.400	**0.638**	3.4M
	Ours	**0.523**	**0.728**	**0.408**	0.604	

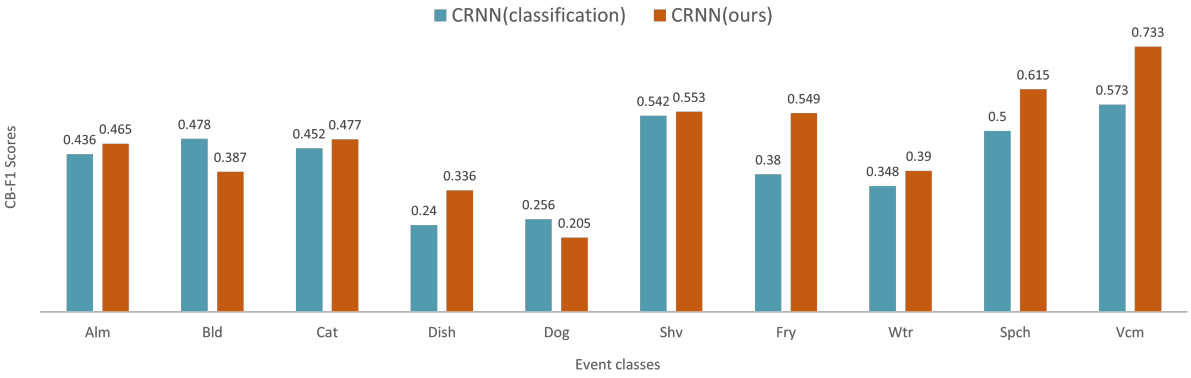


Fig. 3. Comparison of SED systems using classification and regression detection approaches in terms of class-wise event-based F1 measures. Alm, bld, cat, dsh, dog, shv, fry, wtr, spch and vcm are abbreviations of alarm/bell ringing, blender, dishes, dog, electric shaver toothbrush, frying, running water, speech and vacuum cleaner respectively.

III. EXPERIMENT SETUP

A. Datasets and Model Training

We use Domestic Environment Sound Event Detection (DESED) dataset [22], with 10 classes of sound events in home environments, for training and evaluating the proposed SED methods. The training set consists of a strongly annotated set of 10,000 clips, a weakly annotated set of 1,578 clips, and an unlabeled set of 14,412 clips. In addition, an additional strongly labeled dataset with 3,470 clips provided by DCASE 2022 task 4 is used in the training. The duration time of each audio clip is 10 seconds. We evaluated the performance of the systems on the DESED validation set, which consists of 1,168 clips with strongly annotated labels.

All experimental results in this paper are based on the same settings. Each model is trained for 200 epochs with ADAM optimizer. The maximum learning rate is 0.001. The learning rate is initially set to a minimum value and increased gradually to maximum by a warm-up schedule during the first 50 epochs.

B. Pre-processing

Each audio clip in the dataset is resampled to 16kHz and mono due to the inconsistency in audio format. Mel-spectrogram is extracted from audio clip using short-time Fourier transform with window size of 2048, hop length of 256 then applied 128 mel-scale filters. As a result, the input acoustic feature is represented with 626 frames and 128 frequencies.

C. Evaluation Metrics

We use collar-based F1 (CB-F1), intersection-based F1 (IB-F1) [23] and PSDS [24] to evaluate the systems. For the collar-based and intersection-F1 metrics, the systems were evaluated with a threshold fixed at 0.5. For PSDS, two different scenarios were measured, namely PSDS1 with $\rho_{\text{DTC}} = \rho_{\text{GTC}} = 0.7, \rho_{\text{CTTC}} = 0$ and PSDS2 with $\rho_{\text{DTC}} = 0.1, \rho_{\text{GTC}} = 0.1, \rho_{\text{CTTC}} = 0.3$. These two scenarios were evaluated on 50 thresholds linearly distributed between 0.01 and 0.99.

IV. EXPERIMENTAL RESULTS

We compare the performance of implemented SED systems based on different approaches. Furthermore, we analyze the effect of different post-processing methods.

A. Comparison of Different SED Approaches

In Table I, we show the results of SED systems using different detection approaches described in Section II,. We also compare them with FDY-CRNN, SEDT and SS-SEDT. The results of SEDT, SS-SEDT and classification-based CRNN (rows 2-4) are obtained from DCASE 2022 Task 4 technical reports, and the remaining results are re-produced from open-source code under the same configuration and training process.

TABLE II
RESULTS OF SED SYSTEMS USING ANCHOR-FREE AND ANCHOR-BASED DETECTION APPROACHES. THE NUMBER OF BOUNDING BOXES B IS 5.

	Anchor boxes	K-means	IoU	CB-F1	IB-F1	PSDS1	PSDS2
CRNN			-	0.399	0.645	0.252	**0.569**
	✓		0.590	0.424	0.641	0.288	0.553
	✓	✓	**0.835**	**0.480**	**0.673**	**0.384**	0.562

TABLE III
RESULTS OF REGRESSION-BASED SED SYSTEM USING DIFFERENT NUMBERS OF BOUNDING BOXES B.

	B	CB-F1	IB-F1	PSDS1	PSDS2
CRNN	3	0.457	0.670	0.323	**0.570**
	4	0.470	0.670	0.359	0.565
	5	**0.480**	**0.673**	**0.384**	0.562
	6	0.476	0.653	0.362	0.548
	7	0.457	0.630	0.353	0.549

As shown in the table, SED systems using our proposed approach outperform classification-based systems on almost all metrics. One of the reasons is that in the post-processing of the classification-based approach, the window size of the median filter is manually adjusted and is static. This can affect performance because the continuity of sound events is various in practice. On the other hand, our proposal provides a finer resolution and a more automated process to fit sound events. Furthermore, SED systems using our approach only require two weights to convey time boundaries for a sound event: the start time and the anchor box scaling factor. It improves the continuity of the time boundary, resulting in an increase on both CB-F1 and IB-F1. In contrast, the classification-based system requires all frames within the target period to be predicted as active. This may cause breakpoints when encountering non-distinctive parts of acoustic features. We compared the class-wise CB-F1 scores of the regression-based and classification-based CRNN in Figure 3. From the results, regression-based CRNN has improved significantly for long-duration classes such as frying, running water and vacuum cleaner. Here we give an instance in Figure 4 to illustrate the improvement. For PSDS, our approach does not improve significantly. PSDS comprehensively evaluates performance over a wide range of thresholds. However, predictions at a specific threshold are dependent on high-scoring bounding boxes according to our de-duplicate strategy. Therefore, when high-scoring bounding boxes are incorrect, there would be a knock-on effect. This results in a large number of false positives, as shown in Figure 5.

B. Number of Bounding Boxes in Regression-Based Systems

The number of bounding boxes B indicates the maximum number of sound event predictions for input data. Table III shows the results of systems with different numbers of bounding boxes. From the table, the system with $B = 5$ has better performance than other settings. A larger B allows the system to generate more predictions in each frame, which can reduces false negative cases. However, too large B might cause

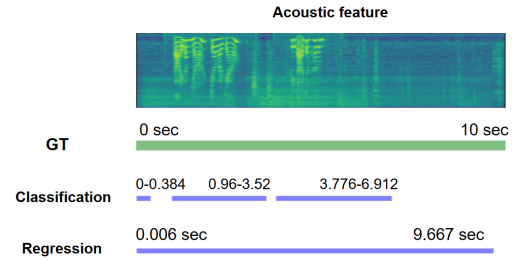


Fig. 4. An instance of CRNN inferences a 10-second audio clip with a 10 second "Frying" class in the DESED validation set: The green lines indicate the time period of events, and the blue lines are the predictions of the systems. The classification-based approach is prone to discontinuity in predicting duration, resulting in two breakpoints in this example. The regression-based approach, on the other hand, is able to overcome this problem.

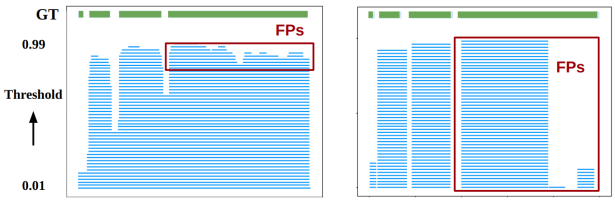


Fig. 5. The difficulties of the regression-based system in computing PSDS: The output of the segmentation-based system and the regression-based system at each threshold (0.01 - 0.99) is shown on the left and right of the figure, respectively. The vertical axis is the predicted result for each threshold. Due to the post-processing approach, when the predictions with high scores of the regression-based system is incorrect, it will significantly affect the results of other thresholds as well.

the system to produce too many false positive predictions. In our experiments, regression-based CRNN with $B = 5$ is the optimal setting for bounding box numbers.

C. Impact of Regression Based On Anchor Boxes

Anchor-free systems estimate the offset of object boundaries directly, and anchor-based systems estimate it based on predefined anchor boxes. For determining the prior lengths of anchor boxes, we first select B different lengths linearly distributed within the maximum audio length (10 sec in our training data) as the initial lengths. Next, we optimize the anchor boxes using K-means to fit the distribution of ground-truth events. Table II shows the results of SED systems using anchor-free and anchor-based detection approaches. From the table, the anchor-based systems are better than the anchor-free systems. This is because ground-truth labels are explicitly assigned to anchor-based systems based on maximum IoU during training. Anchor-free systems are prone to generate too many positive samples due to label assignment instability. After applying K-

TABLE IV
RESULTS OF THREE DIFFERENT POST-PROCESSING APPROACHES: THE TABLE COMPARES THE EFFECT OF USING DIFFERENT POST-PROCESSING APPROACHES FOR THE REGRESSION-BASED CRNN.

De-duplicate	CB-F1	IB-F1	PSDS1	PSDS2
NMS	0.362	0.511	0.325	0.457
Union	0.422	0.649	0.321	**0.564**
ME	**0.480**	**0.673**	**0.384**	0.562

means, the average IoU between the anchor boxes and the training data improved from 0.593 to 0.843, and anchor-based systems further improved performance.

D. Comparison of De-duplicate strategies

As mentioned in section 2.4, our system has three different ways to eliminate redundant bounding boxes in the output. NMS is a widely used strategy for image object detection, but it may not be suitable for polyphonic sound event detection tasks. To solve this problem, we propose two strategies, namely the bounding box mutual exclusion and the bounding box union strategy. According to the results of Table IV, we can see that the bounding box mutual exclusion strategy has superior results. The bounding boxes with higher scores tend to provide more quality predictions for classification and localization. Therefore, consistently selecting bounding boxes with the highest score can result in better coverage of target events. In comparison, the bounding box union strategy suffers from bounding boxes with low scores, which can easily overestimate the event duration. As a result, the prediction is likely to be false positive. Last, NMS discards redundant bounding boxes only based on the IoU without taking into account their class. Therefore, true positive predictions may be erroneously removed when overlapping events reach the NMS threshold and belong to different classes.

V. CONCLUSION

We propose regression-based approaches to sound event detection in this paper. This is motivated by the fact that in contrast to the mainstream classification-based approaches, regression-based approaches detects the entirety of sound events directly. The regression methodology improves the continuity of the duration prediction and alleviates the fragmentation problem, and it increases collar-based and intersection-based F1 scores. Furthermore, since the performance of the regression-based system is highly dependent on the selection of the bounding boxes, we propose two bounding-box selection methods that provide an effective strategy for selecting appropriate bounding boxes. For the proposed regression-based approach, we implemented SED systems using CRNN and VGGSK neural network models. Experiment results show that the proposed approach achieves state-of-the-art performance levels on the open and public DESED dataset.

REFERENCES

[1] T. Theodorou, I. Mporas, and N. Fakotakis, "An overview of automatic audio segmentation," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 6, no. 11, p. 1, 2014.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[5] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[8] S. Venkatesh, D. Moffat, and E. R. Miranda, "You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection," *Applied Sciences*, vol. 12, no. 7, p. 3293, 2022.

[9] S. Tiwari, K. Lakhotia, and M. Mulimani, "Evaluating robustness of you only hear once (YOHO) algorithm on noisy audios in the voice dataset," *arXiv preprint arXiv:2111.01205*, 2021.

[10] A. Mesaros, T. Heittola, A. Diment, *et al.*, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[11] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2017, pp. 344–348.

[12] Z. Ye, X. Wang, H. Liu, *et al.*, "Sound event detection transformer: An event-based end-to-end model for sound event detection," *arXiv preprint arXiv:2110.02011*, 2021.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, Springer, 2020, pp. 213–229.

[14] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for

sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 31–35.

[15] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 task 4," *Orange Labs Lannion, France, Tech. Rep*, 2019.

[16] S.-J. Huang, C.-C. Liu, and C.-P. Chen, "Sound Event Detection System Based on VGGSKCCT Model Architecture with Knowledge Distillation," *Applied Artificial Intelligence*, vol. 37, no. 1, p. 2 152 948, 2023.

[17] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[19] W. Liu, D. Anguelov, D. Erhan, *et al.*, "SSD: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 2016, pp. 21–37.

[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[21] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.

[22] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[24] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 61–65.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Proportionate NLMS with Variable Step-Size for Adaptive Feedback Cancellation in Hearing Aids

Linh Thi Thuc Tran*, Felix Albu†, Hieu Trung Nguyen* and Sven Nordholm‡
* Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
E-mail: linhtt@ptit.edu.vn, hieunt@ptit.edu.vn Tel: +84-24-38544451
† Valahia University of Targoviste, Targoviste, Romania
E-mail: felix.albu@valahia.ro Tel: +40-24521768
‡ Curtin University, Perth, Australia
E-mail: s.nordholm@curtin.edu.au Tel: +61-8-92667439

***Abstract*—It is known that the Proportionate Normalized Least-Mean-Squares (PNLMS) algorithm is an efficient solution applicable to both Adaptive Echo Cancellation (AEC) and Adaptive Feedback Cancellation (AFC) systems. It enhances the performance by individually assigning step sizes to each filter coefficient, resulting in superior initial convergence and tracking rates compared to the NLMS. The Improved PNLMS (IPNLMS) surpasses the PNLMS by incorporating a novel rule that optimally leverages the predicted feedback channel's characteristics to compute tap-specific step sizes.**

In this paper, we introduce an enhancement strategy aimed at improving the performance of the IPNLMS algorithm on the prediction error method (PEM) in the context of Adaptive Feedback Cancellation for hearing aids. This enhancement involves the implementation of variable step-size control. We introduce the Improved Practical Variable Step-Size (IPVSS) built upon the Practical Variable Step-Size (PVSS) approach that incorporates upper and lower fixed step sizes to define a bounded range for step-size adjustments. We evaluate the proposed approach using incoming signals containing both speech and music. We also employ real-world feedback channel measurements. Our simulation outcomes clearly indicate that our method delivers an important enhancement in performance when compared to the PEM-IPNLMS approach, irrespective of whether we utilize the upper or lower step-size bounds. Furthermore, our method outperforms the PEM approach when employing the NLMS and IPVSS while preserving a good signal quality.

I. Introduction

An inherent issue in closed-loop systems like public address (PA) and open-fit hearing aid (HA) is acoustic feedback. It is produced by acoustic signal from the loudspeaker returning to the microphone, accordingly the received signal at microphone comprises of both the incoming signal and the feedback signal. The acoustic feedback issue reduces signal quality and restricts the maximum stable gain that can be achieved. Under certain conditions, it delivers the system into instability and howling may happen. Acoustic feedback cancellation is more difficult in open-fit HA than in PA because of the shorter distance from the microphone to the loudspeaker. Multiple methods for canceling acoustic feedback have been developed over the past sixty years [1]–[5]. Among them, a well-known approach is adaptive feedback cancellation (AFC), in which we employ an adaptive filter to predict the feedback channel. Then we compute the predicted feedback signal and subtract it from the microphone signal. If the prediction is perfect, the error signal $e\,(n)$ retains only the incoming signal. In practice, this prediction often produces a bias because the incoming signal is highly correlated with the loudspeaker signal.

This bias can be reduced by several algorithms proposed in the literature such as a delay insertion [4], [6], phase modulation and frequency shifting [7]–[10], probe noise insertion [11], [12], variable step-size [13]–[18], fast converging adaptive filters [19]–[22], pre-filters [23]–[25], sub-band filters [26]–[28], multiple microphones [29]–[32], etc.

Recently, PNLMS and IPNLMS algorithms have attracted audience attention for both adaptive echo cancellation (AEC) [33]–[35] and adaptive feedback cancellation (AFC) [36]–[38] applications. These algorithms leverage the sparsity of the echo/feedback channels to compute tap-specific step sizes. Due to exploiting an advanced update rule, the convergence as well as tracking rates of the PEM-IPNLMS have been improved compared to PEM-PNLMS while preserving a lower steady-state misalignment for both incoming signals as speech and music [37].

In this work, we extend the PEM-IPNLMS by integrating the improved practical variable step-size control. This provides a new algorithm that maintains the benefits of both previously published approaches. We call this new algorithm PEM-IPNLMS-IPVSS. The simulation outcomes indicate that our proposed system surpasses the system employing only the upper or the lower fixed step-size. Moreover, it also converges faster than the PEM integrating NLMS and IPVSS (called PEM-NLMS-IPVSS) while preserving a high signal quality.

II. Proposed approach

Fig. 1 illustrates a structure of the proposed PEM-IPNLMS-IPVSS approach. In this approach, pre-filters have been used to lower inherent correlation between the incoming signal and the loudspeaker signal, while the IPNLMS improved convergence and tracking abilities. Performance of PEM-IPNLMS is further enhanced by integrating the IPVSS. This variable step-size algorithm allows fast convergence and tracking rates while the framework is (or near) unstable by assigning a high step-size.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

It also produces a small step-size to obtain a low steady-state misalignment while the framework is converged.

In the PEM-IPNLMS-IPVSS approach, the microphone signal is defined as

$$x(n) = u(n) + z(n), \tag{1}$$

where $x(n), u(n)$ and $z(n) = H(q)\,y(n) = \mathbf{h}^T\mathbf{y}(n)$ denote the microphone signal, incoming signal and feedback signal, respectively. $H(q)$ represents the polynomial transfer function of the real feedback path in q, $H(q) = \mathbf{h}^T\mathbf{q}$, where L_h-length impulse response (IR) of the real feedback channel is represented as $\mathbf{h} = [h_0, h_1, \ldots, h_{L_h-1}]^T$ with $\mathbf{q} = \left[1, q^{-1}, \ldots, q^{-L_h+1}\right]^T$ and q^{-1} the discrete-time delay operator. The microphone signal is delayed and then amplified while passing a signal processing path (so-called the forward path) before going back to the loudspeaker. This generates a closed-loop system. We compute the signal received at loudspeaker as $y(n) = G(q)\,e(n)$. The loudspeaker and microphone signals are pre-whitened by using a pre-filter, i.e.,

$$y_p(n) = \hat{A}(q)\,y(n), \tag{2}$$

$$x_p(n) = \hat{A}(q)\,x(n), \tag{3}$$

where $y_p(n)$ and $x_p(n)$ denote signals at loudspeaker and microphone after pre-whitening, respectively; and $\hat{A}(q)$ denotes an estimate of $A(q)$. We assume that the incoming signal is modeled by passing white Gaussian noise, $w(n)$, through an all-pole filter, $A^{-1}(q)$, i.e.,

$$u(n) = A^{-1}(q)\,w(n). \tag{4}$$

Note that this filter must be stable and inverse. Thus, the pre-whitened error signal can be computed as follows

$$e_p(n) = x_p(n) - \hat{\mathbf{h}}^T\mathbf{y}_p(n), \tag{5}$$

with $\mathbf{y}_p(n) = \left[y_p(n), y_p(n-1), \ldots, y_p\left(n - L_{\hat{h}} + 1\right)\right]^T$ a $L_{\hat{h}}$-dimensional signal. To predict the IR of the real feedback channel we minimize the mean square prediction error, $E\left\{e_p^2(n)\right\}$, i.e.,

$$\min_{\hat{\mathbf{h}}} E\left\{e_p^2(n)\right\} = \min_{\hat{\mathbf{h}}} E\left\{\left|x_p(n) - \hat{\mathbf{h}}^T\mathbf{y}_p(n)\right|^2\right\}. \tag{6}$$

The optimal solution for (6) can be recursively approximated using IPNLMS algorithm as

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \frac{\mu_{IPVSS}(n)\,\mathbf{B}(n-1)\,\mathbf{y}_p(n)\,e_p(n)}{\mathbf{y}_p^T(n)\mathbf{B}(n-1)\,\mathbf{y}_p(n) + \alpha_{IPNLMS}}, \tag{7}$$

where $\mathbf{B}(n-1) = diag\left\{b_0(n-1), \ldots, b_{L_{\hat{h}}-1}(n-1)\right\}$ is a diagonal matrix and its elements are calculated as

$$b_l(n) = \frac{1-\beta}{2L_{\hat{h}}} + (1+\beta)\frac{\left|\hat{h}_l(n)\right|}{2\left\|\hat{\mathbf{h}}(n)\right\|_{l_1} + \xi}, \tag{8}$$

with $\left\|\hat{\mathbf{h}}(n)\right\|_{l_1} = \sum_{l=0}^{L_{\hat{h}}-1}\left|\hat{h}_l(n)\right|$, ξ is a small positive constant and $-1 \le \beta < 1$. The regularization parameter of the proposed PEM-IPNLMS-IPVSS is defined as $\alpha_{IPNLMS} = \frac{1-\beta}{2L_{\hat{h}}}\alpha_{NLMS}$.

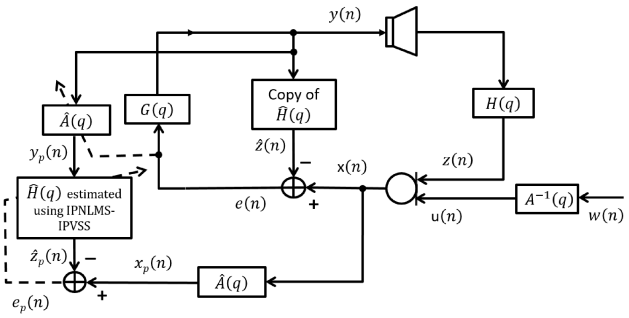


Fig. 1. Proposed PEM-IPNLMS-IPVSS model.

In addition, we implement the improved practical variable step-size algorithm for the PEM-IPNLMS to obtain further performance enhancement in both tracking and convergence. In particular, the improved practical variable step-size called $\mu_{IPVSS}(n)$ in (7) is denoted as

$$\mu_{IPVSS}(n) = \begin{cases} \mu_a & if\ \ \mu_c(n) > \mu_a \\ \mu_b & if\ \ \mu_c(n) < \mu_b \\ \mu_c(n) & otherwise \end{cases}, \tag{9}$$

with μ_a, μ_b the fixed upper and lower step-sizes respectively, and $\mu_a > \mu_b$. We have

$$\mu_c(n) = \mu_a.\left|1 - \frac{\sqrt{\hat{\sigma}_{x_p}^2(n) - \hat{\sigma}_{\hat{z}_p}^2(n)}}{\hat{\sigma}_{e_p}(n) + \epsilon}\right|, \tag{10}$$

with ϵ a small positive constant. In practice, the error signal power $\hat{\sigma}_{e_p}^2(n)$, the microphone signal power $\hat{\sigma}_{x_p}^2(n)$, the estimated feedback signal power $\hat{\sigma}_{\hat{z}_p}^2(n)$ are recursively estimated, i.e.,

$$\hat{\sigma}_{e_p}^2(n) = \lambda\hat{\sigma}_{e_p}^2(n-1) + (1-\lambda)\,e_p^2(n), \tag{11}$$

$$\hat{\sigma}_{x_p}^2(n) = \lambda\hat{\sigma}_{x_p}^2(n-1) + (1-\lambda)\,x_p^2(n), \tag{12}$$

$$\hat{\sigma}_{\hat{z}_p}^2(n) = \lambda\hat{\sigma}_{\hat{z}_p}^2(n-1) + (1-\lambda)\,\hat{z}_p^2(n), \tag{13}$$

where the weighting parameter λ is chosen close to 1.

III. Simulations

A. Simulation Setup

We select the acoustic feedback channels measured using a two-microphone hearing aid placed behind the ear [39] for the simulations. Fig. 2 demonstrates the measured feedback channels with regard to amplitude response and phase, where $H_1(f)$ represents the 1^{st} feedback channel measured in free-field and $H_2(f)$ is the 2^{nd} feedback channel measured with a telephone putting near the ear.

We utilize the recorded incoming signals, where the speech was produced based on concatenating speech segments of male and female speakers from NOIZEUS database [40]; and the music incoming signal was extracted from the John Lennon's song, Imagine. For all simulations, we chose the following

parameters. For instance, lengths of the measured and the predicted feedback channels were $L_h = 100$ and $L_{\hat{h}} = 64$, respectively. Forward path has a gain $|G| = 30$ dB and a delay $d_g = 96$ samples, while feedback canceller path has a delay $d_{fb} = 1$ sample. We chose 16 kHz for sampling frequency. The upper limit of the step-size was $\mu_a = 0.01$ and the lower limit $\mu_b = 0.001$. We set $\lambda = 0.9999$ and $\epsilon = 10^{-6}$. For PEM using NLMS we selected $\alpha_{NLMS} = 10^{-10}$. Coefficients of the pre-filter $\hat{A}(q)$ with order of 20 were estimated every 10 ms using the Levinson-Durbin algorithm.

Two metrics such as added stable gain, ASG, and normalized misalignment, MIS, [5] were chosen to evaluate the considered AFC approaches, i.e.,

$$\text{MIS} = 10\log_{10}\left(\frac{\int_0^{\pi}\left|H(\omega) - e^{-j\omega d_{fb}}\hat{H}(\omega)\right|^2 \mathrm{d}\omega}{\int_0^{\pi}|H(\omega)|^2\,\mathrm{d}\omega}\right), \quad (14)$$

$$\begin{aligned}\text{ASG} &= 10\log_{10}\frac{1}{\max_\omega |H(\omega) - \hat{H}(\omega)|^2} \\ &- 10\log_{10}\frac{1}{\max_\omega |H(\omega)|^2}, \end{aligned} \quad (15)$$

with $H(\omega)$, $\hat{H}(\omega)$ corresponding to frequency responses of the measured and predicted feedback channels, respectively and ω the normalized frequency. The better AFC system will have a lower MIS and a higher ASG. Moreover, PESQ, the perceptual evaluation of speech quality [40], was also used to assess the quality of speech signals. The PESQ is bounded in range [-0.5, 4.5], in which -0.5 means poor whereas 4.5 means excellent quality.

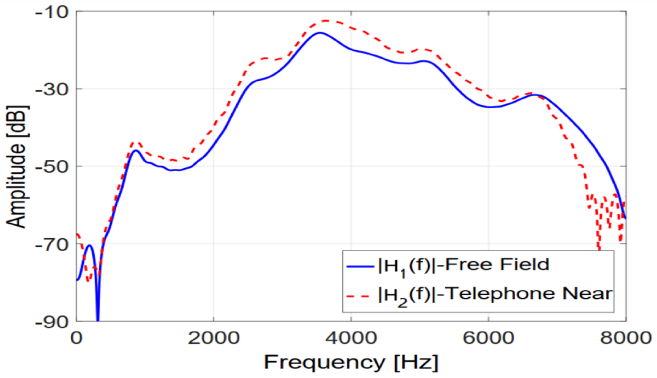
All simulations were run during 80 s, and the feedback channel suddenly changed from $H_1(f)$ to $H_2(f)$ after 40 s.

B. Simulation Results

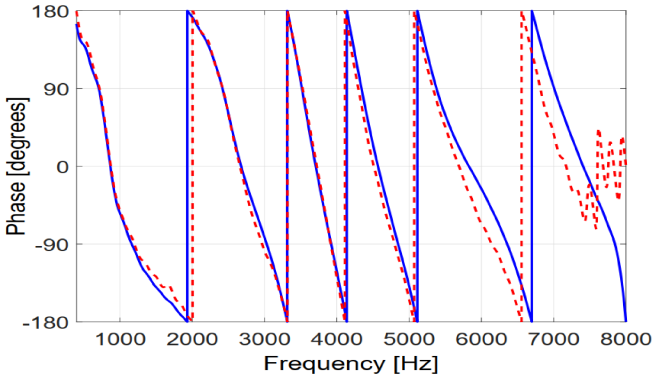
Scenario 1: recorded speech is the incoming signal.

Fig. 3 shows the comparison in terms of MIS and ASG among the proposed PEM-IPNLMS-IPVSS and some baselines such as PEM-IPNLMS with the upper step-size and the lower step-size, PEM-NLMS-IPVSS with bounded step-sizes μ_a and μ_b. We notice the proposed approach outperforms the PEM-IPNLMS with μ_b. It also obtains higher speed of convergence and tracking than PEM-NLMS-IPVSS, while achieving lower steady-state errors. Although the proposed approach converges slower than the PEM-IPNLMS with μ_a at the beginning, it achieves a similar tracking speed when the feedback channel suddenly changes and outperforms in terms of steady-state errors.

Table I demonstrates the average misalignment $(\overline{\text{MIS}_i})$, the average added stable gain $(\overline{\text{ASG}_i})$, and PESQ_i, where subscript i corresponds to the i*th* feedback channel. We compute the PESQ_i after the system has converged. Specifically, PESQ_1 is computed during the period from 20 s to 40 s; PESQ_2 is computed during the period from 59 s to 80 s. We observe the proposed system achieves the lowest average misalignment and the highest average added stable gain compared to other baselines for both measured feedback paths. Furthermore, the PEM-IPNLMS-IPVSS obtains comparable PESQ compared to the PEM-IPNLMS with μ_b and the PEM-NLMS-IPVSS while outperforms the PEM-IPNLMS with μ_a. Those results match well with the observation from Fig. 3.



(a)



(b)

Fig. 2. Amplitude and Phase responses of measured feedback channels.

Scenario 2: recorded music is the incoming signal.

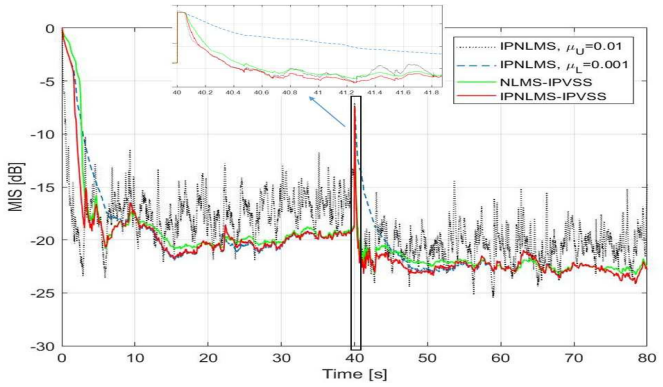
Fig. 4 compares MIS and ASG of mentioned approaches for the case of music as incoming signal. We observe the proposed approach converges quicker than PEM-IPNLMS with μ_b and PEM-NLMS-IPVSS. Although our proposed approach converges slower than PEM-IPNLMS with μ_a, it outperforms PEM-IPNLMS with μ_a in terms of steady-state errors. When we abruptly change the feedback channel from $H_1(f)$ to $H_2(f)$ after 40 s, three approaches including PEM-IPNLMS with μ_a, PEM-NLMS-IPVSS and our proposed PEM-IPNLMS-IVPSS obtain a similar tracking speed, while surpass PEM-IPNLMS with μ_b.

Table II shows that the proposed approach achieves the highest average misalignment for both feedback channels and the highest average ASG for the 1^{st} feedback channel $(\overline{\text{ASG}_1})$. For the 2^{st} feedback channel, it obtains a sightly lower average ASG $(\overline{\text{ASG}_2})$, but still higher than the PEM-IPNLMS using μ_a or μ_b.

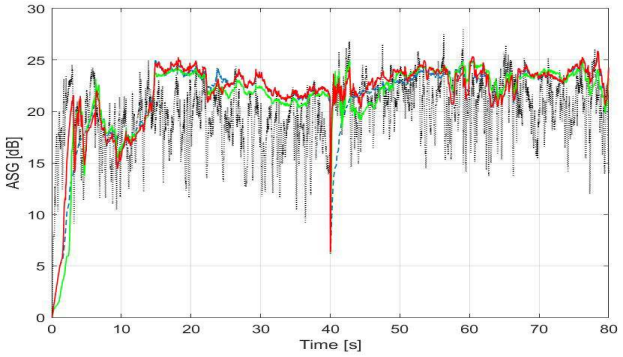
Fig. 5 illustrates the application of variable step-size to the proposed framework for the incoming signals as speech and

TABLE I
COMPARISON AMONG MENTIONED AFC APPROACHES IN AVERAGE MIS, AVERAGE ASG AND PESQ FOR SPEECH INCOMING SIGNAL. WE HIGHLIGHTED THE BEST RESULTS IN BOLD.

AFC approach	$\overline{\text{MIS}}_1$	$\overline{\text{MIS}}_2$	$\overline{\text{ASG}}_1$	$\overline{\text{ASG}}_2$	PESQ_1	PESQ_2
IPNLMS, μ_a	-17.13	-20.44	18.29	20.83	3.25	3.28
IPNLMS, μ_b	-18.33	-21.89	20.23	22.60	**4.37**	4.12
NLMS-IPVSS	-18.29	-22.12	19.79	22.89	4.36	**4.17**
IPNLMS-IPVSS	$\mathbf{-18.80}$	$\mathbf{-22.44}$	**20.40**	**23.10**	**4.37**	4.11

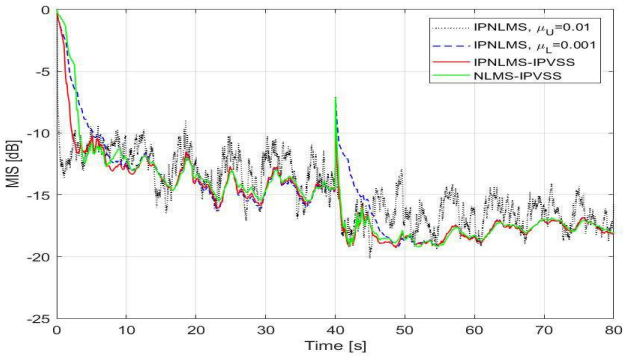


(a)

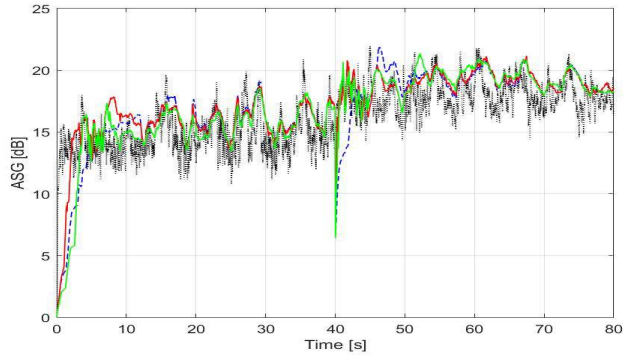


(b)

Fig. 3. Compare MIS (a) and ASG (b) of the proposed framework with PEM-IPNLMS (μ_a), PEM-IPNLMS (μ_b) and PEM-NLMS-IPNLMS; speech as input, an abrupt change of feedback channel from H_1 to H_2.



(a)



(b)

Fig. 4. Compare MIS (a) and ASG (b) of the proposed framework with PEM-IPNLMS (μ_a), PEM-IPNLMS (μ_b) and PEM-NLMS-IPNLMS; music as input, an abrupt change of feedback channel from H_1 to H_2.

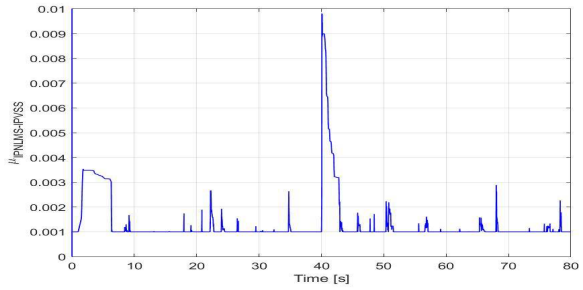
TABLE II
COMPARISON AMONG MENTIONED AFC APPROACHES IN AVERAGE MIS, AVERAGE ASG AND PESQ FOR MUSIC INCOMING SIGNAL. WE HIGHLIGHTED THE BEST RESULTS IN BOLD.

AFC approach	$\overline{\text{MIS}}_1$	$\overline{\text{MIS}}_2$	$\overline{\text{ASG}}_1$	$\overline{\text{ASG}}_2$
IPNLMS, μ_a	-12.56	-16.62	14.51	17.86
IPNLMS, μ_b	-12.73	-17.36	14.93	18.81
NLMS-IPVSS	-12.67	-17.82	14.68	**19.05**
IPNLMS-IPVSS	$\mathbf{-13.05}$	$\mathbf{-17.90}$	**15.39**	18.94

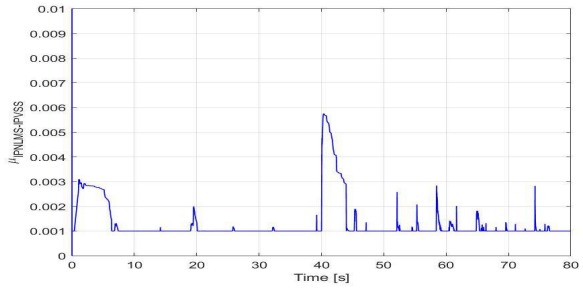
music. As we can see, the proposed approach takes a large step-size (strong peaks) at initialization and at a sudden change of the feedback channel (in these situations, the system is unstable). It also takes a low step-size during the time the system has converged. These outcomes coordinate well with the theory analysis of IPNLMS-IPVSS.

IV. CONCLUSIONS

A novel AFC approach that integrates the improved practical variable step-size on top of PEM-IPNLMS is proposed. This approach allows PEM-IPNLMS to obtain a high step-size while the framework is unstable or near unstable conditions, and a small step-size while the framework has converged. In overall, our proposed approach accomplishes an enhancement for the tracking and convergence abilities, normalized misalignment and added stable gain compared to other mentioned AFC approaches. It also provides a very good signal quality



(a)



(b)

Fig. 5. The step-size variation of the proposed method with $\mu_a = 0.01$ and $\mu_b = 0.001$: (a) for speech input, b) for music input.

with PESQ > 4 for both feedback paths.

ACKNOWLEDGMENT

This research was supported by a grant of the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam, project number 02-2023-HV-DT1. The work of F. Albu was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project number PN-III-P4-PCE-2021-0780, within PNCDI III.

REFERENCES

[1] J. Kates, "Feedback cancellation in hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 1125–1128.

[2] T. van Waterschoot and M. Moonen, "Fifty Years of Acoustic Feedback Control: State of the Art and Future Challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, 2011.

[3] J. Hellgren and U. Forssell, "Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 906–913, 2001.

[4] M. G. Siqueira and A. Alwan, "Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids," *IEEE Trans. Speech, Audio Process.*, vol. 8, no. 4, pp. 443–453, 2000.

[5] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, "Feedback control in hearing aids," in *Springer Handb. Speech Process.* Springer, 2008, pp. 979–1000.

[6] S. Laugesen, K. V. Hansen, and J. Hellgren, "Acceptable delays in hearing aids and implications for feedback cancellation," *J. Acoust. Soc. Am.*, vol. 105, no. 2, pp. 1211–1212, 1999.

[7] M. R. Schroeder, "Improvement of Acoustic-Feedback Stability by Frequency Shifting," *J. Acoust. Soc. Am.*, vol. 36, no. 9, pp. 1718–1724, 1964.

[8] M. Guo and B. Kuenzle, "On the periodically time-varying bias in adaptive feedback cancellation systems with frequency shifting," in *Proc. Int. Conf. Acoust., Speech Signal Process.* IEEE, 2016, pp. 539–543.

[9] M. Guo, S. H. Jensen, J. Jensen, and S. L. Grant, "On the use of a phase modulation method for decorrelation in acoustic feedback cancellation," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 2000–2004.

[10] L. T. T. Tran, H. Schepker, S. Doclo, H. H. Dam, and S. E. Nordholm, "Adaptive feedback control using improved variable step-size affine projection algorithm for hearing aids," in *2017 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. IEEE, 2017, pp. 1633–1640.

[11] M. Guo, T. B. Elmedyb, S. H. Jensen, and J. Jensen, "On acoustic feedback cancellation using probe noise in multiple-microphone and single-loudspeaker systems," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 283–286, 2012.

[12] C. R. C. Nakagawa, S. Nordholm, and W.-Y. Yan, "Feedback cancellation with probe shaping compensation," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 365–369, 2014.

[13] S. Thipphayathetthana and C. Chinrungrueng, "Variable step-size of the least-mean-square algorithm for reducing acoustic feedback in hearing aids," in *Proc. IEEE Asia-Pacific Conf. Circuits Syst.*, 2000, pp. 407–410.

[14] M. Rotaru, F. Albu, and H. Coanda, "A variable step size modified decorrelated NLMS algorithm for adaptive feedback cancellation in hearing aids," *Proc. ISETC. Timisoara*, pp. 263–266, 2012.

[15] F. Strasser and H. Puder, "Adaptive feedback cancellation for realistic hearing aid applications," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2322–2333, 2015.

[16] L. T. T. Tran, H. Schepker, S. Doclo, H. H. Dam, and S. Nordholm, "Improved Practical Variable Step-Size Algorithm For Adaptive Feedback Control in Hearing Aids," in *Proc. IEEE Int. Conf. Signal Process. Commun. Syst.*, 2016.

[17] S. Nordholm, H. Schepker, L. T. T. Tran, and S. Doclo, "Stability-controlled hybrid adaptive feedback cancellation scheme for hearing aids," *J. Acoust. Soc. Am.*, vol. 143, no. 1, pp. 150–166, 2018.

[18] L. T. T. Tran and S. E. Nordholm, "A Switched Algorithm for Adaptive Feedback Cancellation Using Pre-Filters in Hearing Aids," *Audiol. Res.*, vol. 11, no. 3, pp. 389–409, 2021.

[19] S. Lee, I.-Y. Kim, and Y.-C. Park, "Approximated affine projection algorithm for feedback cancellation in hearing aids," *Comp. methods programs Biomed.*, vol. 87, no. 3, pp. 254–261, 2007.

[20] K. Lee, Y.-h. Baik, Y. Park, D. Kim, and J. Sohn, "Robust adaptive feedback canceller based on modified pseudo affine projection algorithm," in *Annu. Int. Conf. IEEE Engin. Med. Biol. Soc.*, 2011, pp. 3760–3763.

[21] F. Albu, "Simplified proportionate affine projection algorithms," in *Proc. 19th Int. Conf. Syst., Signals Image Process.* IEEE, 2012, pp. 368–371.

[22] L. T. T. Tran, H. H. Dam, and S. Nordholm, "Affine Projection Algorithm For Acoustic Feedback Cancellation Using Prediction Error Method In Hearing Aids," *Proc. IEEE Int. Work. Acoust. Signal Enhan.*, 2016.

[23] J. Hellgren, "Analysis of feedback cancellation in hearing aids with filtered-X LMS and the direct method of closed loop identification," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 119–131, 2002.

[24] A. Spriet, I. Proudler, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3749–3763, 2005.

[25] A. Spriet, G. Rombouts, M. Moonen, and J. Wouters, "Adaptive feedback cancellation in hearing aids," *Elsevier J. Franklin Inst.*, vol. 343, no. 6, pp. 545–573, 2006.

[26] F. Yang, M. Wu, P. Ji, and J. Yang, "An improved multiband-structured subband adaptive filter algorithm," *IEEE Signal Process. Lett.*, vol. 19, no. 10, pp. 647–650, 2012.

[27] F. Strasser and H. Puder, "Sub-band feedback cancellation with variable step sizes for music signals in hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* IEEE, 2014, pp. 8207–8211.

[28] S. Pradhan, V. Patel, D. Somani, and N. V. George, "An improved proportionate delayless multiband-structured subband adaptive feedback canceller for digital hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1633–1643, 2017.

[29] C. R. C. Nakagawa, S. Nordholm, and W.-Y. Yan, "Analysis of Two Microphone Method for Feedback Cancellation," *IEEE Signal Process. Lett.*, vol. 22, pp. 35–39, 2015.

[30] L. T. T. Tran, H. Schepker, S. Doclo, H. H. Dam, and S. Nordholm, "Frequency Domain Improved Practical Variable Step-Size for Adaptive Feedback Cancellation Using Pre-Filters," in *2018 16th Int. Work. Acoust. Signal Enhanc.* IEEE, 2018, pp. 171–175.

[31] H. Schepker, S. E. Nordholm, L. T. T. Tran, and S. Doclo, "Null-steering beamformer-based feedback cancellation for multi-microphone

hearing aids with incoming signal preservation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 679–691, 2019.
[32] H. Schepker, S. Nordholm, and S. Doclo, "Acoustic Feedback Suppression for Multi-Microphone Hearing Devices Using a Soft-Constrained Null-Steering Beamformer," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 929–940, 2020.
[33] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 508–518, 2000.
[34] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2. IEEE, 2002, pp. II—-1881.
[35] C. Paleologu, J. Benesty, and S. Ciochin\ua, "An improved proportionate NLMS algorithm based on the l 0 norm," in *Int. Conf. Acoust. Speech Signal Process.* IEEE, 2010, pp. 309–312.
[36] F. Albu, R. Nakagawa, and S. Nordholm, "Proportionate algorithms for two-microphone active feedback cancellation," in *Proc. 23rd Eur. Signal Process. Conf.* IEEE, 2015, pp. 290–294.
[37] L. T. T. Tran, H. Schepker, S. Doclo, H. H. Dam, and S. Nordholm, "Proportionate NLMS for adaptive feedback control in hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017.
[38] F. Albu, L. T. T. Tran, and S. Nordholm, "A combined variable step size strategy for two microphones acoustic feedback cancellation using proportionate algorithms," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. IEEE, 2017, pp. 1373–1377.
[39] T. Sankowsky-Rothe, M. Blau, H. Schepker, and S. Doclo, "Reciprocal measurement of acoustic feedback paths in hearing aids," *J. Acoust. Soc. Am.*, vol. 138, no. 4, pp. EL399—-EL404, 2015.
[40] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Residual Echo Suppression using Spatial Feature for Stereo Acoustic Echo Cancellation

Hsuan-Cheng Chou*, Yih-Liang Shen*, Meng-Hsuan Wu*†, Bo-Wun Shih† and Tai-Shih Chi*
* Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan
† Realtek Semiconductor Corp., Hsinchu, Taiwan
azqlmpgghh@gmail.com, dennis831209@gmail.com, luckily75368@gmail.com, bowuns@realtek.com, tschi@nycu.edu.tw

***Abstract*—There have been some advances in deep learning based stereo-AEC (SAEC) systems in recent years. However, most of the studies focused on solving the non-uniqueness problem and did not explore the impact of spatial cues on SAEC. In this paper, we propose a composite SAEC system that combines conventional adaptive filters and Wiener filters with a NN-based residual echo suppression (RES) module. We adopt generalized cross correlation (GCC) as an additional input to allow the RES module to better analyze the embedded spatial cues. Experimental results show that the addition of GCC stably improves system performance with a little computational overhead. In addition, the proposed system achieves better results with much less computation than compared NN-based SAEC systems in all test conditions.**

I. Introduction

Teleconferencing, online meetings, and work-from-home (WFH) become popular during the COVID pandemic/endemic. High-quality voice communication through internet is in high demand. Acoustic echo cancellation (AEC), which removes unwanted loop-back sounds from the received signals at microphones, is a very critical module in the communication system. In modern systems, AEC is commonly accompanied by background noise suppression and speech de-reverberation. In hands-free communication applications, where multiple microphones and speakers are often used, the widely used single-channel AEC cannot perform well, thus facilitating the development of stereo-AEC (SAEC).

SAEC is for the communication scenario that uses at least two microphones and speakers. The main difference between SAEC and the single-channel AEC is the non-uniqueness problem, which leads to performance degradation of adaptive filters [1]. Due to the high coherence between signals of the stereo loudspeakers, adaptive methods would have difficulties in approximating multiple echo paths, which leads to a reduced convergence rate and even frequently re-converging. Many studies have attempted to mitigate this problem, such as modifying the far-end signal to de-correlate the stereophonic channels without largely impacting human listening experiences [2]. Another approach is to orthogonally transform the reference signal of the adaptive method and predict the echo paths of the transformed signal [3].

Neural networks have been successfully used to boost the performance of echo cancellation systems. An intuitive approach is to place the neural network behind the conventional AEC to replace the residual echo suppression (RES) module by suppressing both noise and residual echo [4][5][6]. Some studies integrate the tasks of AEC and RES, and consider echo cancellation as a task of supervised speech separation [7]. Most of the past studies proposed in several AEC Challenges are limited to single-channel AEC tasks [8][9]. Recently, fully neural network systems were proposed for SAEC to explicitly avoid the non-uniqueness problem [10][11]. Although these systems are capable of handling SAEC tasks, they only perform AEC on different microphone channels separately without using any spatial cues embedded in stereo microphone channels. In [12], a multiple-input-multiple-output (MIMO) convolutional recurrent network (CRN) architecture was proposed to handle information between multiple microphones. However, to preserve spatial information in the convolution process, the stride in convolution layers was set to one, resulting in heavy computational load.

In this paper, we propose a composite SAEC system consisting of adaptive filters, Wiener filters, and a spatial-aware neural RES module, without inducing heavy computational load as in [12]. The adaptive filters are implemented with robust short-time Fourier transform domain adaptive filters [13] to suppress echoes and to solve the double-talk problem, and with the de-correlation method in [2] to reduce the coherence between channels. The Wiener filters are then applied to suppress stationary noise. To make the neural RES module spatial-aware, we explicitly add the generalized cross correlation (GCC)[14] cues to the input of the neural RES module. Simulation results show explicitly adding GCC cues to the RES module is beneficial to performance of the SAEC system.

The rest of the paper is organized as follows. Section 2 describes the signal model of the SAEC problem. Section 3 describes the proposed system architecture. Section 4 presents the experimental settings and results. Section 5 concludes the paper.

II. Signal Model

In this paper, we focus on a two-loudspeaker, two-microphone scenario. The speech signal of the far-end user is picked up by the two microphones in the transmission room (far-end), transmitted through the communication system, and then broadcast by the two loudspeakers in the receiving room (near-end). Since we do not want the sound emitted from the loudspeakers to be captured by the microphones and sent back to the transmission room, we need to use an AEC system to

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

suppress the echo in the microphone signal. As shown in Fig. 1, the two channels of the far-end signal are labeled as x_1, and x_2, which are obtained by convolving a source signal in the transmission room with two room impulse responses (RIRs) for two microphones.

Without loss of generality, we assume the far-end signals contain only voices of the users in the transmission room, i.e., the echo signals in the transmission room are ideally eliminated. In the receiving room, the two loudspeakers play the far-end signals, generating echoes through the RIRs, h_{11}, h_{12}, h_{21}, and h_{22}, of the respective coupling paths between the loudspeakers and the microphones. These echoes, together with the near-end user voice and ambient background noise in the receiving room, constitute the signal received by the microphones y_1 and y_2, which can be expressed in the short-term Fourier transform (STFT) domain as:

$$\begin{aligned} Y_i(t,f) = X_1(t,f)H_{1i}(f) + X_2(t,f)H_{2i}(f) \\ +V(t,f) + R(t,f) + W(t,f), i = 1,2 \end{aligned} \tag{1}$$

where i is the microphone index; Y and X are the STFT of y and x. H is the transfer function of the corresponding echo path. V represents the STFT of the direct and early reflections (within 50 ms) of the near-end speech; R and W are the STFT of late-reverberations and the background noise in the receiving room, respectively. The goal of the proposed SAEC system includes not only echo cancellation and noise suppression, but also speech de-reverberation. Hence, V is used as the training target of the RES module.

III. Proposed System

Fig. 1 shows the architecture of the proposed system. The system consists of three components, the adaptive filters, the Wiener filters, and the spatial-aware RES module with a GCC encoder. First, the signals received by the microphones are transformed to the time-frequency (T-F) domain using STFT and then passed through the adaptive filters and the Wiener filters to respectively suppress linear echoes and the stationary part of the background noise. Afterward, the filtered signals, signals received by microphones, and far-end signals are fed into the RES module.

Although signals of two channels are fed into the RES module altogether, we cannot be sure the network extracts spatial information effectively. Therefore, we explicitly compute the GCC-PHAT from the microphone signals y_1 and y_2, and embed the spatial cues through the GCC encoder. The output of the GCC encoder is then added to the recurrent module of the RES module. In the following subsections, we present the system components separately.

A. Adaptive Filter

We use the robust STFT domain adaptive filter proposed in [13] for directly suppressing linear echo in the T-F domain. For each adaptive filter, the learning rate is set to 5e-2, the number of taps is set to 5, and the internal parameter γ, which adjusts the learning rate during double-talk, is set to 1e-2. The function of the adaptive filter is to estimate RIRs by minimizing the output, which can be expressed as follows:

$$\begin{aligned} E_i^{AEC}(t,f) = Y_i(t,f) - X_1(t,f)\hat{H}_{1i}(f) \\ -X_2(t,f)\hat{H}_{2i}(f), i = 1,2 \end{aligned} \tag{2}$$

where E^{AEC} is the output of the adaptive filter, and $\hat{H}$ is the transfer function of the estimated RIR. Note, to tackle the non-uniqueness problem, we first decorrelate the far-end signals using parameters given in [2] before engaging the adaptive filtering processes.

B. Wiener Filter

To enhance the robustness of the system to unseen noise, we use a Wiener filter to suppress the stationary noise after each adaptive filter. The Wiener filter proposed in [15] with the behavior controlling parameter β set to 0.99 is used in the system. This version of Wiener filter is based on time-frame energy estimation and two-stage noise suppression (TSNR).

$$E_i^{WF}(t,f) = E_i^{AEC}(t,f)\hat{M}_i^{WF}(t,f), i = 1,2 \tag{3}$$

where E^{WF} is the output of the Wiener filter and $\hat{M}^{WF}$ is the Wiener gain.

C. Residual Echo Suppression

The basic architecture of the proposed RES module is CRN, which consists of an encoder, a recurrent module, and a decoder as shown in Fig. 2. We set its hyperparameters by network pruning of the convolution layers [16]. The concatenation of Y_1, Y_2, X_1, X_2, and E_1^{WF}, E_2^{WF} is used as the input to the RES. The encoder consists of 5 causal 2D convolution layers followed by BN [17], and ELU [18]. Each convolution layer contains [16, 26, 38, 54, 54] kernels with kernel size = [3, 1], stride = [2, 1]. The recurrent module consists of two layers of unidirectional GRU with 128 units and fully connected layers with 216 units. The decoder is symmetric to the encoder except that each layer contains a transposed convolution layer.

In a typical CRN, the output of the encoder is skip connected to the input of the parallel decoder to compensate the information loss through the recurrent module. To further reduce computation, we adopt the method in [19] to first apply one-by-one convolution to organize output channels of the encoder and then to change the skip connection to summation. The activation function of the final layer in the decoder is Softplus, which does not restrict the range of output values to $[0, 1]$ to allow compensation for possible over-suppression by the preceding Wiener filter. The outputs of the RES are two magnitude masks, which can be expressed as:

$$\hat{M}_{1,2}^{RES} = G_{RES}(Y_1, Y_2, X_1, X_2, E_1^{WF}, E_2^{WF} \mid \phi) \tag{4}$$

where ϕ indicates the set of hyperparameters. Note that the time and frequency indices (t, f) of variables $\hat{M}_{1,2}^{RES}$, Y_1, Y_2, X_1, X_2, E_1^{WF} and E_2^{WF} are purposely omitted for a short expression. The final predicted magnitude spectrograms

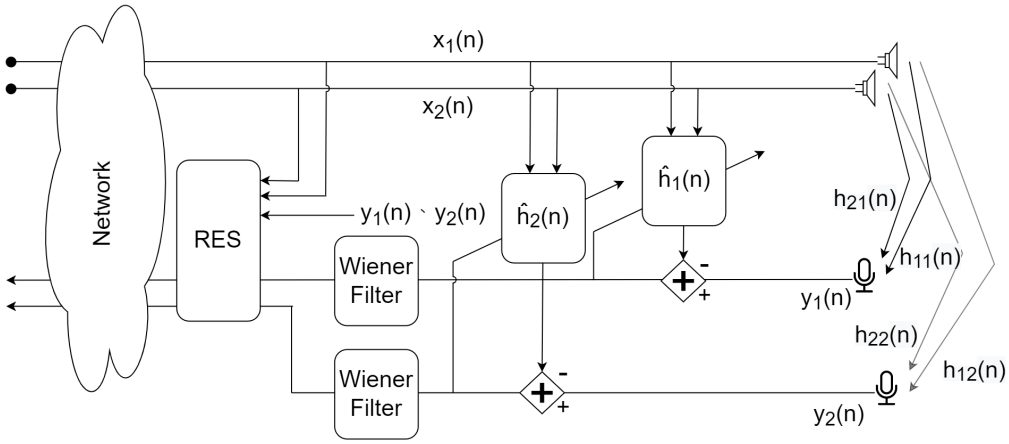


Fig. 1. Architecture of the proposed SAEC system.

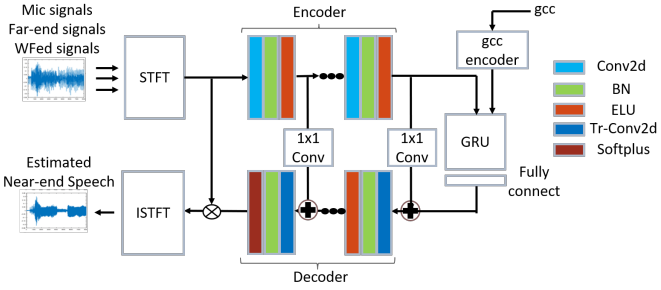


Fig. 2. Architecture of the RES module.

of speech signals are obtained by multiplying the Wiener filter outputs and the RES outputs as follows:

$$\hat{V}_i(t,\ f) = E_i^{WF}(t,\ f)\hat{M}_i^{RES}(t,\ f), i = 1, 2 \tag{5}$$

D. GCC Encoder

To explicitly provide spatial cues to the RES module, we calculate the GCC-PHAT from microphone signals as follows [14][20]:

$$\begin{aligned} GCC_{PHAT}(\psi,\ t,\ f) &= Real\{\frac{Y_1(t,\ f)Y_2^H(t,\ f)}{|Y_1(t,\ f)||Y_2^H(t,\ f)|}e^{-j2\pi\frac{f}{N}f_s\psi}\} \\ &= cos(\angle Y_1(t,\ f) - \angle Y_2(t,\ f) - 2\pi\frac{f}{N}f_s\psi) \end{aligned} \tag{6}$$

where f_s is the sampling rate, N is the number of point of FFT, and ψ are 37 possible time differences of arrival.

As shown in Fig. 2, we use the GCC-encoder, which consists of four layers of one-dimensional depth-separable convolution [21], to embed GCC-PHAT. Numbers of filters of the four layers are [64, 32, 16, 8], and the kernel size and the stride are set to 4 and 1. The encoder uses f and ψ as the channel axis and the feature axis, and performs feature extraction on each time frame. The final output features are reshaped to form the spatial embedding with the size of $(T, 37 * 8)$, where T is the total number of frames, and then concatenated to the input of GRU of the RES module.

TABLE I

TRAINING AND TEST CONDITIONS. ('MIC POS.' IS SHORT FOR 'MICROPHONE ARRAY POSITION' AND 'DIST.' INDICATES THE DISTANCE BETWEEN THE NEAR-END TALKER AND THE MICROPHONE ARRAY. SNR IS THE SIGNAL-TO-NOISE POWER RATIO AND SER IS THE SIGNAL-TO-ECHO POWER RATIO. NOISE INCLUDES DIRECTIONAL AND DIFFUSE NOISE WITH THE POWER RATIO OF 2 TO 1.)

	Training	Test
T60 (s)	0.2, 0.4, 0.6	0.3, 0.5, 0.7
[Room size, Mic pos.] (m)	[(7, 10, 3), (0.5, 5, 1)] [(12, 15, 3), (1.5, 7.5, 1)] [(17, 20, 3), (2.5, 10, 1)]	[(8, 10, 3), (0.5, 5, 1)] [(15, 18, 3), (1, 9, 1)]
Dist. (m)	2.5, 3, 3.5, 4, 4.5, 5	2.5, 3, 3.5, 4, 4.5, 5
SNR (dB)	0, 5, 10, 15, 20	0, 6, 12
SER (dB)	-10, -5, 0, 5, 10	-6, 0, 6

IV. EXPERIMENTAL RESULTS

A. Dataset Preparation

For system evaluations, we used 4620 and 1680 sentences from the TRAIN and TEST categories of the TIMIT corpus [22] to build our training, validation, and test sets. TIMIT contains 630 American English speakers from different dialect regions of the United States. In this study, we included all speakers and selected different speakers to build training and test tests. For generating various diffuse noise, 10 and 15 types of noise were selected from NOISEX-92 [23] and Freesound [24] corpora as the noise sources. 13 types of diffuse noise were included in the training set. The validation and test sets contained the remaining 12 unseen types of diffuse noise. For directional noise, sound samples of PNL 100 Nonspeech Sounds [25] were selected for training, and the remaining non-stationary types of noise of NOISEX-92 were used for validation and test.

The simulation scenario is a group teleconferencing using TV in a large room. The settings of the training and test rooms are shown in Table I. The two-channel microphone array is placed in the center of the room's long side, with a distance from the back wall of at least 50 cm. The distance between two microphones is 7 cm. To mimic TV settings,

TABLE II
EXPERIMENTAL RESULTS IN ALL TEST CONDITIONS. THE PESQ-WB SCORE IN EACH PAIR OF SNR-SER CONDITION IS AVERAGED FROM 1000 TEST SENTENCES. NOTE THAT THE OTHER TWO TEST CONDITIONS OF [ROOM SIZE, MIC POS.] AND DIST. ARE ALL RANDOMLY SELECTED FROM TEST CONDITIONS LISTED IN TABLE 1.

Metric	PESQ-WB										ERLE (dB)	MACs (G/s)	Params (MB)
SNR (dB)	0			6			12			Avg	-	-	-
SER (dB)	-6	0	6	-6	0	6	-6	0	6	Avg			
Noisy	1.09	1.09	1.09	1.09	1.12	1.16	1.10	1.15	1.24	1.13			
Proposed-M	1.26	1.34	1.39	1.38	1.52	1.61	1.48	1.69	1.84	1.50	80.59	0.112*2	**1.23**
Proposed-B	1.26	1.35	1.41	1.38	1.53	1.63	1.47	1.69	1.85	1.51	80.19	0.114	1.23
Proposed-GCC	**1.28**	**1.37**	**1.43**	**1.42**	**1.57**	**1.67**	**1.52**	**1.74**	**1.90**	**1.54**	72.29	0.135	1.73
Baseline 1 [11]	1.18	1.25	1.30	1.27	1.40	1.50	1.36	1.52	1.69	1.39	80.50	2.17*2	56.16
Baseline 2 [12]	1.27	1.33	1.37	1.39	1.52	1.60	1.51	1.73	1.86	1.51	39.09	4.78	**1.23**

the two loudspeakers are set to line up with the microphone array, and each of which is placed 50 cm apart from the center of the array on each side. Once the room size and locations of the microphones and loudspeakers are determined, the four RIRs, h_{11}, h_{12}, h_{21}, and h_{22}, of the respective coupling paths between the loudspeakers and the microphones are simulated using the image method [26][27]. The near-end talker is randomly placed at 37 possible angles (between 0 and 180 degrees on the front semicircle with a step of 5 degrees) paired with the distance shown in Table I. To generate diffuse noise, the noise source is placed at 37 angles on the largest semicircle from the microphone array with random phase shifts. For simulating directional noise, one noise source is randomly placed on the wall to imitate leaking noise from a window.

Based on these settings, we generated $10,000$ 10-sec training sentences for a total of about 28 hours. Each of the validation and test sets contained $1,000$ 10-sec sentences. In these sentences, near-end speech appeared randomly after the initial 0.5 sec to 1.5 sec, with a length between 5 sec and 7 sec. Echoes and noise were present throughout the entire 10 sec with a 10% chance of noise-free. To better reflect the real-world situations, echo-free, slight non-linear echo, and severe non-linear echo were considered. We used the slight scaled error function in [28] and severe non-linearity functions in [4] to produce corresponding echo signals. Eventually, in training samples, there was a 20% chance of echo-free, a 20% chance with slight non-linear echo, a 20% chance with severe non-linear echo, and a 40% chance with normal linear echo.

B. Training Settings and Experimental Results

The audio signals were converted into the T-F domain by the short-time Fourier transform using a 512-point Hann window and a 256-point hop size. For training, we used the AdamW optimizer with a learning rate of 1e-3 and a weight decaying rate of 1e-2. The training batch size was 32. All of the network weights were initialized by the method described in [29], the default initialization method of Pytorch [30]. The loss function was mean absolute error (MAE). If the model did not achieve better results in the validation set for 30 consecutive epochs during training, the learning rate was cut in half. At the third time of halving, the training was terminated and the best-performing checkpoint in the validation set was selected for test.

For evaluations, we compare the proposed SAEC system with other recently proposed SAEC systems using our generated datasets. There are five compared systems, Baseline 1, Baseline 2, Proposed-M, Proposed-B and Proposed-GCC. Baseline 1 is the NN based SAEC system proposed in [11]. Baseline 2 is the NN-based MIMO SAEC system proposed in [12]. Proposed-M is our proposed system, but the RES are operated separately on two channels without incorporating GCC information. In other words, it is a bi-lateral system rather than a binaural system. Proposed-B is our proposed model without the GCC-encoder, where the spatial information is implicitly provided by the received signal y_1 and y_2. Proposed-GCC is our proposed model with the GCC-encoder which explicitly provides the spatial cue of GCC-PHAT.

We used objective measures of wide-band PESQ (PESQ-WB) [31] and echo return loss enhancement (ERLE) to assess sound quality of near-end speech and the effectiveness of echo cancellation, respectively. The test results of all compared systems in different conditions are shown in Table II. The PESQ-WB scores of Proposed-GCC outperforms other systems in all conditions. The averaged PESQ-WB scores show the performance of Proposed-B is slightly better than the performance of Proposed-M and is comparable to the performance of Baseline 2. The superior performance of the binaural system Proposed-GCC to the binaural system Proposed-B and the bilateral system Proposed-M confirms that explicitly providing spatial information encoded in GCC-PHAT to the system is beneficial to its performance. The lower performance of Proposed-B clearly demonstrates the RES module cannot effectively decode spatial information from raw data received by two microphones. Moreover, as shown in the 'MACs' column, the computation of Proposed-GCC is much less than that of Baseline 1 and Baseline 2.

Regarding the ERLE measure, Proposed-M performs the best and Proposal-GCC comes fourth. However, ERLE was originally used to assess adaptive filter's capability in reducing linear echo without considering noise suppression. In literature, NN-based methods can easily achieve an ERLE higher than 30 dB due to their noise suppression capability. The correlation between ERLE scores and subjective listening tests is unclear

at such high levels. Even with such a large gap in ERLE between Baseline 1 and Baseline 2, we did not detect any significant difference between them in our informal listening tests.

C. Limitation of the Study

In this study, all RIRs were simulated using the image method. No real-world RIRs were measured and included during training. Hence, the performance of the proposed model will deteriorate in real-world environments. The model needs to be fine-tuned using a small amount of real-world data.

V. Conclusion

In this paper, we propose a composite system to tackle the SAEC problem for the condition involving a single dominant near-end speaker. The system contains conventional de-echo, speech enhancement, and the NN-based RES modules. By explicitly incorporating spatial information from GCC-PHAT into the RES module, the performance of the binaural system is stably improved. Experimental results show that the proposed system can produce higher PESQ-WB scores with much less computation than other compared NN-based SAEC systems in all test conditions. In the future, we will replace the conventional adaptive filters and Wiener filters with NNs to have a complete NN-based SAEC system, and investigate the condition of multiple dominant speakers.

References

[1] M Mohan Sondhi, Dennis R Morgan, and Joseph L Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal processing letters*, vol. 2, no. 8, pp. 148–151, 1995.

[2] Tomas Gaensler and Eric Diethorn, "Inter-channel coherence reduction method for stereophonic acoustic echo cancellation," in *IWAENC*. IEEE, 2016, pp. 1–5.

[3] Saeed Bagheri and Daniele Giacobello, "Robust stft domain multi-channel acoustic echo cancellation with adaptive decorrelation of the reference signals," in *ICASSP*, 2021, pp. 131–135.

[4] Amin Fazel, Mostafa El-Khamy, and Jungwon Lee, "Cad-aec: Context-aware deep acoustic echo cancellation," in *ICASSP*, 2020, pp. 6919–6923.

[5] Shimin Zhang, Ziteng Wang, Jiayao Sun, Yihui Fu, Biao Tian, Qiang Fu, and Lei Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9127–9131.

[6] Haoran Zhao, Nan Li, Runqiang Han, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu, "A deep hierarchical fusion network for fullband acoustic echo cancellation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9112–9116.

[7] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *Interspeech*, 2019, pp. 4255–4259.

[8] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP*, 2021, pp. 151–155.

[9] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, and Robert Aichner, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP*, 2022, pp. 9107–9111.

[10] Hao Zhang and DeLiang Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation.," in *Interspeech*, 2021, pp. 1139–1143.

[11] Linjuan Cheng, Renhua Peng, Andong Li, Chengshi Zheng, and Xiaodong Li, "Deep learning-based stereophonic acoustic echo suppression without decorrelation," *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 816–829, 2021.

[12] Chenggang Zhang, Jinjiang Liu, and Xueliang Zhang, "Lcsm: A lightweight complex spectral mapping framework for stereophonic acoustic echo cancellation," *arXiv preprint arXiv:2208.07277*, 2022.

[13] Jason Wung, Daniele Giacobello, and Joshua Atkins, "Robust acoustic echo cancellation in the short-time fourier transform domain using adaptive crossband filters," in *ICASSP*, 2014, pp. 1300–1304.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[15] Cyril Plapous, Claude Marro, Laurent Mauuary, and Pascal Scalart, "A two-step noise reduction technique," in *ICASSP*, 2004, vol. 1, pp. I–289.

[16] Zi Wang, Chengcheng Li, and Xiangyang Wang, "Convolutional neural network pruning with structural redundancy reduction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14913–14922.

[17] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[18] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[19] Sebastian Braun, Hannes Gamper, Chandan KA Reddy, and Ivan Tashev, "Towards efficient models for real-time deep noise suppression," in *ICASSP*, 2021, pp. 656–660.

[20] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, "Robust tdoa estimation based on time-frequency masking and deep neural networks.," in *Interspeech*, 2018, pp. 322–326.

[21] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[22] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[23] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[24] Frederic Font, Gerard Roma, and Xavier Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[25] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[26] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[28] Hao Zhang and DeLiang Wang, "Multi-channel and multi-microphone acoustic echo cancellation using a deep learning based approach," *arXiv preprint arXiv:2103.02552*, 2021.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[31] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

Multitaper Adaptive Time-Frequency Windowed Fourier Transform Based on the Reliable Region of Window Widths

Jen-Chieh Cheng [1] and Jian-Jiun Ding [2*]
[1] National Taiwan University, Taiwan
E-mail: f05942132@ntu.edu.tw Tel: +886-233663662
[2*] National Taiwan University, Taiwan
E-mail: jjding@ntu.edu.tw Tel: +886-233669652

Abstract — **The window width of the short-time Fourier transform (STFT) has a critical effect on the resolution of the time-frequency distribution. Until now, due to the limitation of the fast Fourier transform (FFT) algorithm, traditional adaptive methods, such as the adaptive STFT and the adaptive synchrosqueezing transform (ST), either only considerer the time-domain width or only considerer the frequency-domain width. However, a more general adaptive method usually has a higher computation complexity due to the search for optimal parameters. In this paper, we propose the reliable region of the window widths for both time and frequency axes. We also simplify the searching process for optimal parameters. Moreover, based on the reliable region of the window widths, we propose a method called the multitaper adaptive time-frequency windowed Fourier transform (MATFWFT) that has time-variant and frequency-variant window size. It can simultaneously estimate the optimal window size for both the time axis and the frequency axis and achieve a time-frequency analysis result with high energy concentration.**

I. Introduction

Time-frequency analysis (TFA) is a critical tool for biomedicine, seismology, communications, and mechanical vibration engineering. Some of the typical methods of TFA, include short-time Fourier transform (STFT), Wigner-Ville distribution (WVD), wavelet transform (WT), and S-transform (ST). However, traditional TFA methods is their restricted resolution caused by the uncertainty principle. They are unable to achieve high resolution in the time and frequency axes simultaneously. This means that it is hard to get a sharp distribution of TFAs. Therefore, to address this problem, reassignment method (RM) [1]−[5] was developed in recent years as a method to concentrate energy to the center of gravity. For frequency-varying signals only, the RM has been improved to the SST [4]. In order to tackle the issue of fast frequency-varying signals, additional techniques were proposed, including the second order SST (SST2) [6] and high order SST [7]. Furthermore, there are also multi-step recursive methods, such as the multisynchrosqueezing transform (MSST) [8]. In addition to being based on STFT, SST can be derived using different transformations. These include the WT-based synchrosqueezing transform [9] and the ST-based synchrosqueezing transform [10, 11].

However, regardless of the traditional TFAs, i.e., STFT, WT, ST, or the post-processing methods, i.e. the RM, the SST, the SET or others, the window width is still the most important and the most critical to the clarity of the time frequency distribution. Therefore, numerous adaptive methods were proposed to automatically search for the optimal window width, i.e. adaptive STFT (ASTFT) [12], adaptive ST (AST) [13, 14], adaptive SST (ASST) [15−17], adaptive SEST (ASEST) [14], multitaper ASTFT (MASTFT) [18], etc. Among them, the adaptive methods are mainly classified into two categories, one is based on maximizing the concentration measure, such as the Rényi entropy [15] and the other is based on the chirp rate (CR) to obtain the optimal variance of window widths, for example [12, 14, 16−18].

On the other hand, the multitaper, also known as Multi resolution, is a spectral density estimation technique. It is a method of superposition of multiple scales of the spectrum. Initially, multitaper was proposed to be applied in spectrum for reducing sidelobe and noise, e.g. [19]. Then, in order to increase the TF distribution clarity and robustness to noise, it was gradually applied to TFA, e.g. [20] also known as superlets [21]. Then, it was also applied to the RM method, e.g., [22]. As well as in SET, e.g. [18]. Incidentally, multitaper is also can divided into two main categories, such as the arithmetic

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

averaging [19]−[22] or the geometric averaging [18, 20].

Although many ASTFT, AST and adaptive postprocessing methods of TFA have been proposed. However, all of these methods are only adaptive for one of the axes, either the time axis or the frequency axis. This limitation is clearly due to the constraint imposed by the FFT.

Therefore, in this paper, we propose the MATFWFT, which is an adaptive method that utilizes the MT technique to overcome the limitation of FFT. It can obtain the optimal window width in both time and frequency axes simultaneously. In other words, we combine the time varying window widths and frequency varying window widths to achieve the time-frequency varying window widths under the constrain of the FFT. On the other hand, the range of the window width is the most important issue of the adaptive method. We analyze the possible minimum and maximum limits of window widths in the time domain and the frequency domain, and derive the reliable region of the window width under the constraints of the FFT, and adopt the different sampling intervals for the time and frequency axes to achieve a small number of window widths. It deals with the circumstances of many different types of input signal and reduce the complexity significantly.

II. Previous Work

The short-time Fourier transform (STFT) is a time-frequency analysis method. Its definition is

$$V(t,f)=\int_{-\infty}^{\infty}x(\tau)w(\tau-t)e^{-i2\pi f(\tau-t)}d\tau =\int_{-\infty}^{\infty}x(t+\tau)w(\tau)e^{-i2\pi f\tau}d\tau \qquad (1)$$

where $x(t)$ is the input signal and $w(t)$ is the window (mask) function. It can be chosen as a scaled Gaussian function, i.e.,

$$w(t)=e^{-\pi t^2/\sigma^2} \qquad (2)$$

where the parameter σ control the window width which affect the clarity of the STFT and then the STFT is also called the Gabor transform.

The general ST works with many different window functions in the frequency domain [13]. It can be viewed as the STFT in the frequency domain and is defined as

$$ST(t,f)=\int_{-\infty}^{\infty}x(\tau)w(t-\tau)e^{-i2\pi f\tau}d\tau =\int_{-\infty}^{\infty}X(f+\alpha)W(\alpha)e^{i2\pi\alpha t}d\alpha . \qquad (3)$$

where $X(f)$, $W(f)$ are the Fourier transform of the $x(t)$, $w(t)$ respectively. In the ST, we also adopt $W(f)$ as the scaled Gaussian function (but it is in the frequency domain), i.e.,

$$W(f)=e^{-\pi f^2/\sigma^2} . \qquad (4)$$

III. Proposed Method

A. Adaptive STFT and Adaptive S-Transform

For considering time varying and frequency vary window, we redefine the Gaussian window function with time-varying variance and in the STFT (1) as

$$w(t)=e^{-\pi t^2/\sigma^2(t)} \qquad (5)$$

and the frequency-varying variance as the window function in the ST (3) as

$$w(f)=e^{-\pi f^2/\sigma^2(f)} \qquad (6)$$

B. Rényi Entropy

The Rényi entropy [25] is a method commonly to measure energy concentration of the time-frequency representation (TFR). We adopt it to measure the distribution of TFR. The local time Rényi entropy is defined as

$$E(t)=\frac{1}{1-\alpha}\log_2\frac{\int_{t-\tau}^{t+\tau}\int_0^{\infty}\left|V(b,f,\sigma)\right|^{2\alpha}dfdb}{\left(\int_{t-\tau}^{t+\tau}\int_0^{\infty}\left|V(b,f,\sigma)\right|^{2}dfdb\right)^{\alpha}} \qquad (7)$$

where $\alpha > 2$ is recommended for TFR measures, V denotes any TFR, and τ is denote the size of the neighborhood.

Similarly, the local frequency Rényi entropy is defined as

$$E(f)=\frac{1}{1-\alpha}\log_2\frac{\int_{f-\tau}^{f+\tau}\int_{-\infty}^{\infty}\left|V(t,b,\sigma)\right|^{2\alpha}dtdb}{\left(\int_{f-\tau}^{f+\tau}\int_{-\infty}^{\infty}\left|V(t,b,\sigma)\right|^{2}dtdb\right)^{\alpha}} . \qquad (8)$$

C. Reliable Region of the Window Width

The window width is limited by the Heisenberg uncertainty principle. In other words, if the window width in the time domain is chosen to be wider, the frequency width is narrower, and conversely, if width in the time domain is chosen to be narrower, the frequency width is wider. In order to find the reasonable range of the window width, we observe the relationship between its width and Δt, Δf are the sampling time interval and sampling frequency interval respectively.

First, we temporarily change the window width of time and frequency $\sigma(t)$, $\sigma(f)$ to fixed value σ_t and σ_f, respectively.

1) Reliable Region of the Window Width by the Closed Form of the STFT and the ST. Recalling the case of an input

signal being an impulse function in the time and frequency domains respectively, the STFT and ST can be expressed as:

$$\left|V\left(t,f\right)\right| = e^{-\pi(t_0-t)^2/\sigma_t^2} = \left|V\left(t\right)\right|, \tag{9}$$

$$\left|ST\left(t,f\right)\right| = e^{-\pi(f_0-f)^2/\sigma_f^2} = \left|ST\left(f\right)\right|. \tag{10}$$

Since the variance needs to be larger than the inherent physical constraints, we have

$$\left\{\sigma_t,\sigma_f\right\} \geq C\left\{\sqrt{2\pi}\Delta_t,\sqrt{2\pi}\Delta_f\right\} \tag{11}$$

where the constant $C \geq 1$. By the same way, recalling the case of an input signal being an exponential function in the time and frequency domains, the STFT and the ST can be expressed as:

$$\left|V\left(t,f\right)\right| = e^{-\sigma_t^2\pi\left[\phi'(t)-f\right]^2} = \left|V\left(f\right)\right|, \tag{12}$$

$$\left|ST\left(t,f\right)\right| = e^{-\sigma_f^2\pi\left[t-\varphi'(f)\right]^2} = \left|ST\left(t\right)\right|. \tag{13}$$

Note that we only care about the exponential term. Since the variance need to be larger than the inherent physical constraints, we have

$$\left\{\sigma_t,\sigma_f\right\} \leq \left\{\frac{1}{C\sqrt{2\pi}\Delta_f},\frac{1}{C\sqrt{2\pi}\Delta_t}\right\}. \tag{14}$$

where the constant $C \geq 1$. We can observe that the absolute values of the STFT and the ST can be interpreted as the exponential function in the frequency and time domains, respectively. Moreover, because the STFT (1) and the ST (3) can exchanged by the FT, the window function widths of each can be related to each other, i.e.,

$$w\left(t\right) = e^{-\pi t^2/\sigma_t^2},\ W\left(f\right) = \mathrm{F}\left\{w\left(t\right)\right\} = \sigma_t e^{-\pi\sigma_t^2 t^2}, \tag{15}$$

and the window function in the ST (3) is

$$W\left(f\right) = e^{-\pi f^2/\sigma_f^2}. \tag{16}$$

For the clarity of the time frequency resolution, we only pay attention to the influence of the window width, which is the exponential term of the scaled Gaussian function. Hence, by comparing the exponential term of the scaled Gaussian function $W(f)$ in (15) and (16), we know that σ_t and σ_f are reciprocally related, i.e.,

$$\sigma_t = 1/\sigma_f. \tag{17}$$

By observing (11), (14) with (17), we know that (11) and (14) are inherently equal, which is also highly logical.

2) Reliable Region of the Window Width by Standard Deviation of the Window

On the other hand, another common way to explore the window width is the standard deviation (SD) in [22]. The width, SD of the window function or called the temporal and spectral spreads. Consider in the time domain, its definition is

$$SD_t^2 = \frac{\int_{-\infty}^{\infty} t^2\left|w\left(t\right)\right|^2 dt}{\int_{-\infty}^{\infty}\left|w\left(t\right)\right|^2 dt} = \frac{\sigma_t^2}{4\pi},\ SD_t = \sqrt{\frac{\sigma_t^2}{4\pi}} \tag{18}$$

where $w(t)$ is defined in (2). In the frequency domain, SD is defined as

$$SD_f^2 = \frac{\int_{-\infty}^{\infty} f^2\left|W\left(f\right)\right|^2 df}{\int_{-\infty}^{\infty}\left|W\left(f\right)\right|^2 df} = \frac{\sigma_f^2}{4\pi},\ SD_f = \sqrt{\frac{\sigma_f^2}{4\pi}} \tag{19}$$

where $W(f)$ is defined in (4). Note that the SDs definition omit the mean term since the mean value of the window functions of the time and frequency domain are both zero. Obviously, in practice, since the SDs need to be larger than the inherent physical constraints, we have

$$\left\{\sigma_t,\sigma_f\right\} \geq C\left\{\sqrt{4\pi}\Delta_t,\sqrt{4\pi}\Delta_f\right\} \tag{20}$$

where the constant $C \geq 1$.

Furthermore, we recall the previously derived window width restriction (20) and reciprocal relation (17), which becomes the following restriction

$$C\sqrt{4\pi}\Delta_t \leq \sigma_t \leq \frac{1}{C\sqrt{4\pi}\Delta_f},\ C\sqrt{4\pi}\Delta_f \leq \sigma_f \leq \frac{1}{C\sqrt{4\pi}\Delta_t}. \tag{21}$$

For applying the FFT algorithm, we can set

$$\Delta_t\Delta_f = 1/N. \tag{22}$$

In this case, substitute (22) into (21) and rewrite (21) as

$$C\sqrt{4\pi}\Delta_t \leq \sigma_t \leq \frac{N}{C\sqrt{4\pi}}\Delta_t,\ C\sqrt{4\pi}\Delta_f \leq \sigma_f \leq \frac{N}{C\sqrt{4\pi}}\Delta_f. \tag{23}$$

We can observe that the reliable region of the window width determined by the SD method has stricter conditions compared to the region determined by the closed form of the STFT and the ST method. Therefore, the SD width perspective, we not only provide a reasonable lower bound for the window width, but also exploit the relationship between the window function in the STFT and ST to obtain a reasonable upper bound. In our

following work, we adopt (21) or (23), which provide suitable and reliable ranges of the window width.

D. Choosing the Time-Frequency Varying Window Width

Due to the restriction of the FFT algorithm, the window width is usually selected only for the time axis or the frequency axis. However, to overcome this issue, we propose a new approach to incorporate the selection based on both the time and frequency axes within the restriction of the FFT algorithm.

Firstly, consider the case of time-varying signal. We adopt the approach of optimizing the Rényi entropy (7) for the selection of the window width and limit its range to (21). At a fixed time t, we can use (7) to find the optimal $\sigma(t)$

$$\sigma(t) = \min_{\sigma} E(t). \tag{24}$$

Note that the local time Rényi entropy E(t) in (24) can be applied to any TFR. Specifically, apply the STFT to the local time Rényi entropy E(t) in the (7), we can get the corresponding optimal window width

$$\sigma_{STFT}(t) = \min_{\sigma} E_{STFT}(t). \tag{25}$$

where E_{STFT} denotes the local Rényi entropy of the STFT.

Secondly, similar to the time-varying method, we also consider the case of frequency varying signal. We adopt the approach of optimizing the local frequency Rényi entropy (8). At a fixed frequency f, we can use (8) to find the optimal σ(f) with the most concentrated TFR

$$\sigma(f) = \min_{\sigma} E(f). \tag{26}$$

We apply the ST to the local frequency Rényi entropy $E(f)$ in the (8) and obtain the optimal window width

$$\sigma_{ST}(f) = \min_{\sigma} E_{ST}(f). \tag{27}$$

where E_{ST} denotes the local frequency Rényi entropy of TFSEST and $\sigma_{TFSEST}(f)$ denotes its optimal window width.

E. Multitaper Adaptive Time Frequency Windowed FT

After the process of optimization of the local time and frequency Rényi entropy of the STFT and the ST, $E(t)$, and $E(f)$, we can get the corresponding optimal window width, (25) and (27), respectively. Substitute the (25) and (27) into STFT, ST respectively. Refer to them as the adaptive STFT and the adaptive ST based on the Rényi entropy.

Finally, we utilize the multitaper technique to merge the adaptive STFT and the adaptive ST to define the MATFWFT:

$$MATFWFT(t,f) = \left(\left| ASTFT(t,f) \right|^2 + \left| AST(t,f) \right|^2 \right) / 2. \tag{28}$$

The ASTFT and the AST are denoted the adaptive STFT and the adaptive ST based on the Rényi entropy respectively. It addresses the long-standing issue that adaptive the window width only depends on a single axis. The optimal window width is obtained for each time-frequency point, not only for the time axis or frequency axis.

IV. Digital Implementation

In this section, we introduce the digital implementation method of the ATFWSET. For the STFT, we can set

$$t = n\Delta_t, \quad \tau = k\Delta_\tau, \quad f = m\Delta_f, \quad \Delta_\tau \Delta_f = \frac{1}{N}. \tag{29}$$

In specifically $\Delta_t = \Delta_\tau$, then the discrete STFT becomes

$$V[n,m] = \sum_k x[n+k] w[k] e^{-i2\pi \frac{km}{N}} \tag{30}$$

For the ST, we also can set

$$t = n\Delta_t, \quad \alpha = p\Delta_\alpha, \quad f = m\Delta_f, \quad \Delta_t \Delta_\alpha = \frac{1}{N}. \tag{31}$$

Specifically, $\Delta_f = \Delta_\alpha$, then the discrete general ST becomes

$$ST[n,m] = \sum_p X[p+m] W[p] e^{i2\pi \frac{pn}{N}}. \tag{32}$$

For practical implementation, recall the reliable window widths (23) proposed in Section III. It is recommended to adopt equally spaced intervals based on multiples of the sampling time interval and sampling frequency interval respectively, i.e., the range of (50) is rewritten as follows

$$\begin{aligned} \sigma_t &= \Delta_t \cdot round\left(\left[\sqrt{4\pi}C \quad 2k\sqrt{4\pi}C \quad \cdots \quad N/\sqrt{4\pi}C \right] \right) \\ \sigma_f &= \Delta_f \cdot round\left(\left[\sqrt{4\pi}C \quad 2k\sqrt{4\pi}C \quad \cdots \quad N/\sqrt{4\pi}C \right] \right) \end{aligned} \tag{33}$$

where k is denoted multiple of sampling interval and round(.) is denoted as the operator that rounds the value.

In the first step, for each window width in the reliable region of the window width range in the time domain and frequency domain respectively. Then, the ASTFT and AST are obtained by estimating the optimal window width. Note that for each window width is limited in the reliable region of the window width.

Finally, by using the mutitaper technique for ASTFT and AST, we can obtain MATFWFT.

IV. Experiments

In this section, we demonstrate the experiments to illustrate the efficiency and robustness about the proposed method. We add the comparison of the highest resolution in the time domain and frequency domain, i.e., $STFT_{SD}$, ST_{SD} which are denoted the STFT with the window width SD_t, and the ST with the window width SD_f respectively. We also compare with several currently existing adaptive time frequency methods, including the adaptive STFT (ASTFT), adaptive ST (AST) and our proposed ATFWFT, all based on the local Rényi entropy in (7) or (8) respectively, i.e., ASTFT with the window width $\sigma_{STFT}(t) = \min E_{STFT}(t)$, AST with the window width $\sigma_{ST}(f) = \min E_{ST}(f)$. In addition to the above adaptive method, Finally, with the multitaper technique, there are $MSTFT_{SD} = (|STFT_{SD}|^2 + |ST_{SD}|^2) / 2$, MATFWFT = $(|ASTFT|^2 + |AST|^2) / 2$.

In the experiment, we set $N = 200$, $\Delta_t = 1/200$, $\Delta_f = 1$, and the local Rényi entropy size of the neighborhood $\tau = 4$ in (7), (8), and the constant $C = 2$, multiple of sampling interval $k = 1$ in (33). The reliable region of the window width is rewritten as $\sigma_t = [7\ 8 \cdots 28] \cdot \Delta_t$, $\sigma_f = [7\ 8 \cdots 28] \cdot \Delta_f$. Note that because the upper and lower bounds are restricted by reliable region of the window width, the number of them in our setting is only 22. However, it is also allowed to sample with wider intervals depending on the number of samples you need. It is efficient in reducing the complexity.

In order to show that there are various signals at the same time or frequency, first we consider the three-component exponential signals

$$x(t) = e^{-0.1t+i2\pi\left[\left(10t^2+10t\right)+\left(7.5t^3+35t\right)+\left(7.5t^4+60t\right)\right]}. \tag{34}$$

Then, we also consider the three-component impulse signals

$$x(t) = \mathcal{F}\left\{e^{-0.1t+i2\pi\left[\left(7.5t^2-80t\right)+\left(7.5t^3-20t\right)+\left(7.5t^4+30t\right)\right]}\right\}. \tag{35}$$

Then, we consider a complicated test signals that contain multi-component exponential signals and impulse signals as:

$$\begin{aligned}
x(t) &= x_1(t) + x_2(t) + x_3(t) + x_4(t) + x_5(t) \\
&\quad + x_6(t) + x_7(t), \\
x_1(t) &= 0.2\exp\left(i2\pi\left(7.5t^2+10t\right)\right),\ t \in [0, 0.4), \\
x_2(t) &= 0.2\exp\left(i2\pi\left(7.5t^2+80t\right)\right),\ t \in [0.4, 0.6), \\
x_3(t) &= 0.2\exp\left(i2\pi\left(7.5t^2+10t\right)\right),\ t \in [0.6, 1], \\
x_4(t) &= \exp\left(-10000\left(t-0.25\right)^2 + i2\pi\left(80t\right)\right), \\
x_5(t) &= 1.2\exp\left(-10000\left(t-0.5\right)^2 + i2\pi\left(10t\right)\right), \\
x_6(t) &= \exp\left(-10000\left(t-0.25\right)^2 + i2\pi\left(80t\right)\right), \\
x_7(t) &= 0.35\exp\left(i2\pi \cdot \begin{pmatrix} 0.8\sin\left(4\pi t\right) \\ +7.5t^2+10t \end{pmatrix}\right).
\end{aligned} \tag{36}$$

Figs. 1-3 provide the TF plots of signal (34)-(36), respectively, obtained by (a) the ASTFT, (b) the AST, (c) the MATFWFT, (d) the STFT with the window width SDt, (e) ST with the window width SDf, and (f) the MSTFTSD. For this multi-component signal, we observe the results as following.

Firstly, we observe the STFT and the ST using the smallest window width, SD_t and SD_f, i.e., the highest resolution in the time and frequency domain, respectively. However, some components of the signal disappear because of the limitation of the uncertainty principle where high resolution implies low resolution in the corresponding frequency and time domain.

Further, we utilize the multitaper technique to improve the concentration of the energy by combining and MSTFTSD. With the multitaper effect of MATFWFT, it obviously contours the accurate time frequency distribution. That combine the advantages of both characteristic, which contain the optimal window width of the time and frequency domain respectively.

Finally, we compare the performance of our proposed method with other different methods. We measure the energy concentration through global Rényi entropy, probability energy distribution (PED).

The PED is defended

$$PED = \frac{\sum_n \sum_k \left|V(n,k)\right|^4}{\left(\sum_n \sum_k \left|V(n,k)\right|^2\right)^2}. \tag{37}$$

It is a model of the probability to measure the sharpness of the time frequency distribution. Note that smaller the PED value represents the less concentrated the energy distribution, and vice versa.

In Table I, we show the result comparing many different adaptive methods.

VI. Conclusion

In this paper, we proposed the reliable window width region which has the advantage of reducing the complexity by narrowing the range of the window width. Based on the reliable window width region and the local Rényi entropy, we also proposed the MATFWFT, which is an adaptive method that incorporates optimal time-varying and frequency-varying

window widths simultaneously. It solves the traditional issue of the dependence of the window width on a single axis only. It has the advantage of working effectively at any time-frequency points, for both exponential and impulse signals. In total, it contributes to a time frequency analysis method with low complexity and capable of estimating the time-varying and frequency-varying window sizes in a concurrent manner.

ACKNOWLEDGMENT

The authors thank for the support of the National Science and Technology Council, Taiwan, under the contract of MOST 110-2221-E-002 -092 -MY3.

REFERENCES

[1] K. Kodera, C. De Villedary, and R. Gendrin, "A new method for the numerical analysis of non-stationary signals," Phys. Earth Planet. Inter., vol. 12, no. 2, pp. 142-150, 1976.

[2] K. Kodera, R. Gendrin, and C. Villedary, "Analysis of time-varying signals with small BT values," IEEE Trans. Acoust. Speech Signal Process., vol. 26, no. 1, pp. 64-76, 1978.

[3] F. Auger, and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," IEEE Trans. Signal Process., vol. 43, no. 5, pp. 1068-1089, 1995.

[4] F. Auger, P. Flandrin, Y. T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H. T. Wu, "Time-Frequency Reassignment and Synchrosqueezing: An Overview," IEEE Signal Process. Mag., vol. 30, no. 6, pp. 32-41, 2013.

[5] S. Meignen, and N. Singh, "Analysis of Reassignment Operators Used in Synchrosqueezing Transforms: With an Application to Instantaneous Frequency Estimation," IEEE Trans. Signal Process., vol. 70, pp. 216-227, 2022.

[6] T. Oberlin, S. Meignen, and V. Perrier, "Second-Order Synchrosqueezing Transform or Invertible Reassignment? Towards Ideal Time-Frequency Representations," IEEE Trans. Signal Process., vol. 63, no. 5, pp. 1335-1344, 2015.

[7] D. H. Pham, and S. Meignen, "High-order synchrosqueezing transform for multicomponent signals analysis—With an application to gravitational-wave signal," IEEE Trans. Signal Process., vol. 65, no. 12, pp. 3168-3178, 2017.

[8]G. Yu, Z. Wang, and P. Zhao, "Multisynchrosqueezing transform," IEEE Trans. Ind. Electron., vol. 66, no. 7, pp. 5441-5455, 2019.

[9] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," Appl. Comput. Harmon. Anal., vol. 30, no. 2, pp. 243-261, 2011.

[10] Z.-l. Huang, J. Zhang, T.-h. Zhao, and Y. Sun, "Synchrosqueezing S-transform and its application in seismic spectral decomposition," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 2, pp. 817-825, 2015.

[11] Q. Wang, J. Gao, N. Liu, and X. Jiang, "High-resolution seismic time–frequency analysis using the synchrosqueezing generalized S-transform," IEEE Geosci. Remote Sens. Lett., vol. 15, no. 3, pp. 374-378, 2018.

[12] S. C. Pei, and S. G. Huang, "STFT With Adaptive Window Width Based on the Chirp Rate," IEEE Trans. Signal Process., vol. 60, no. 8, pp. 4065-4080, 2012.

[13] W. Yao, Z. Teng, Q. Tang, and P. Zuo, "Adaptive Dolph–Chebyshev window-based S transform in time-frequency analysis," IET Signal Proc., vol. 8, no. 9, pp. 927-937, 2014.

[14] B. Li, Z. Zhang, and X. Zhu, "Adaptive S-Transform with Chirp-Modulated Window and Its Synchroextracting Transform," Circuits, Syst., Signal Process., vol. 40, no. 11, pp. 5654-5681, 2021.

[15] Y.-L. Sheu, L.-Y. Hsu, P.-T. Chou, and H.-T. Wu, "Entropy-based time-varying window width selection for nonlinear-type time–frequency analysis," Int. J. Data Sci. Anal., vol. 3, no. 4, pp. 231-245, 2017.

[16] L. Li, H. Cai, H. Han, Q. Jiang, and H. Ji, "Adaptive short-time Fourier transform and synchrosqueezing transform for non-stationary signal separation," Signal Process., vol. 166, pp. 107231, 2020.

[17] L. Li, H. Cai, and Q. Jiang, "Adaptive synchrosqueezing transform with a time-varying parameter for non-stationary signal separation," Appl. Comput. Harmon. Anal., vol. 49, no. 3, pp. 1075-1106, 2020.

[18] X. Zhu, B. Li, K. Yang, Z. Zhang, and W. Li, "Parameter analysis of chirplet transform and high-resolution time-frequency representation via chirplets combination," Signal Process., vol. 205, pp. 108824, 2023.

[19] S. Karnik, J. Romberg, and M. A. Davenport, "Thomson's Multitaper Method Revisited," IEEE Trans. Inf. Theory, vol. 68, no. 7, pp. 4864-4891, 2022.

[20] S. Cheung, and J. S. Lim, "Combined multiresolution (wide-band/narrow-band) spectrogram," IEEE Trans. Signal Process., vol. 40, no. 4, pp. 975-977, 1992.

[21] V. V. Moca, H. Bârzan, A. Nagy-Dăbâcan, and R. C. Mureșan, "Time-frequency super-resolution with superlets," Nat. Commun., vol. 12, no. 1, pp. 337, 2021.

[22] I. Reinhold, and M. Sandsten, "The multitaper reassigned spectrogram for oscillating transients with Gaussian envelopes," Signal Process., vol. 198, pp. 108570, 2022.

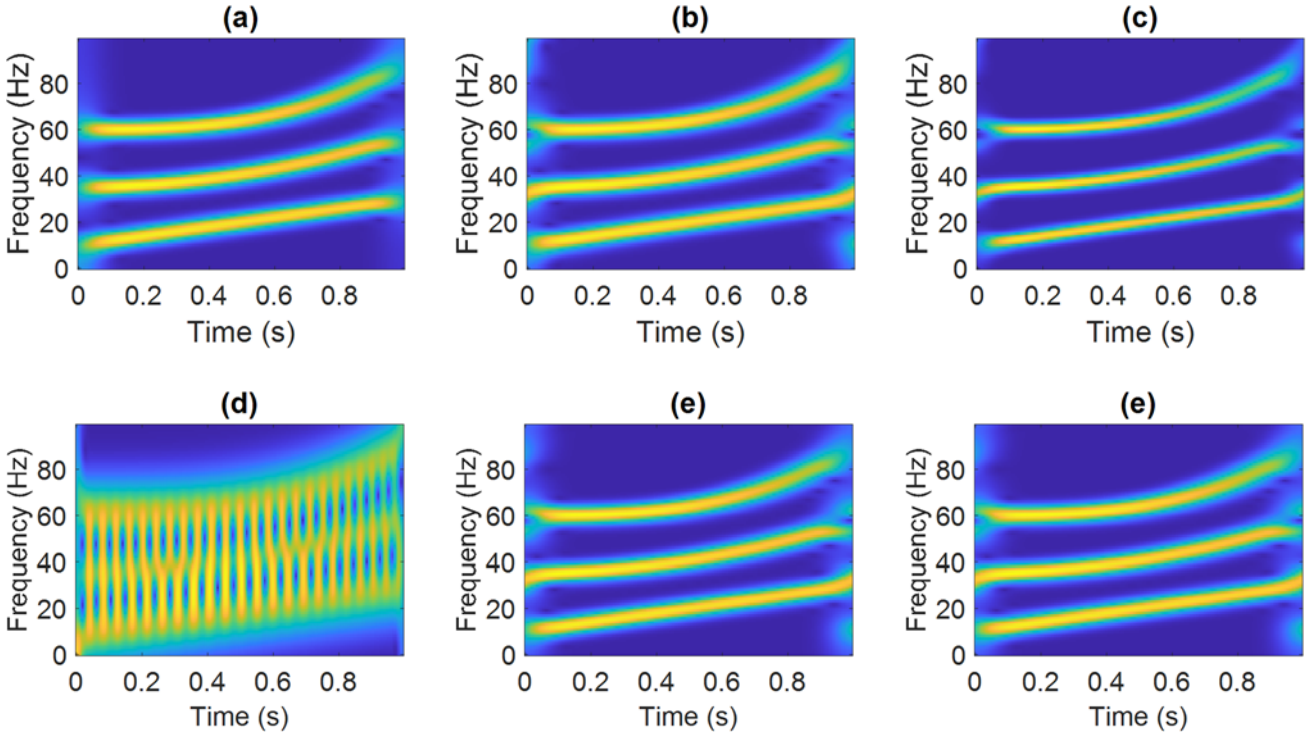


Fig. 1 The TF plots of the signal in (34). (a) ASTFT; (b) AST ; (c) MATFWFT; (d) $STFT_{SD}$; (e) ST_{SD}; (f) $MSTFT_{SD}$.

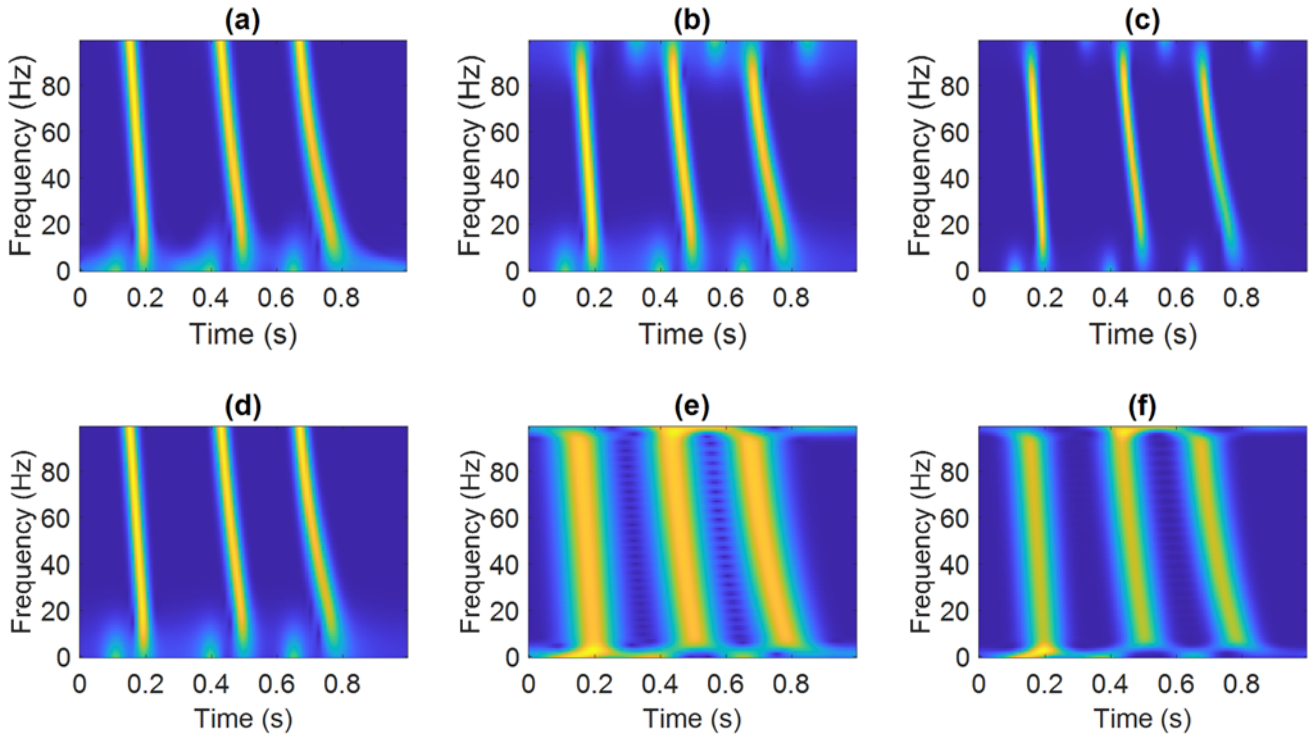


Fig. 2 The TF plots of the signal in (35). (a) ASTFT; (b) AST ; (c) MATFWFT; (d) $STFT_{SD}$; (e) ST_{SD}; (f) $MSTFT_{SD}$.

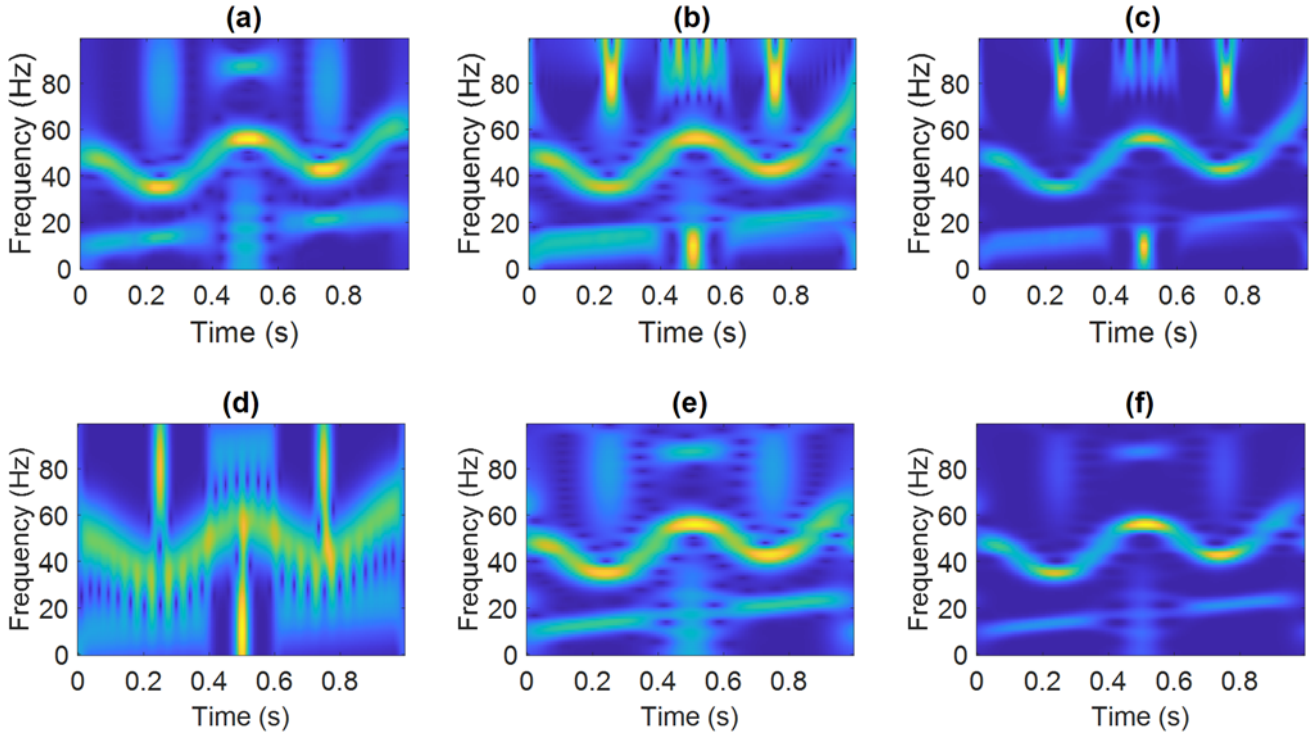


Fig. 3 The TF plots of the signal in (36). (a) ASTFT; (b) AST ; (c) MATFWFT; (d) $STFT_{SD}$; (e) ST_{SD}; (f) $MSTFT_{SD}$.

TABLE I THE RÉNYI ENTROPY AND THE PROBABILITY ENERGY DISTRIBUTION (PED) OF THE SIGNAL (34)-(36). A LOWER RENYI ENTROPY AND A HIGHER PED MEANS BETTER ENERGY CONCENTTATION.

(34)	ASTFT	AST	MATFWFT	$STFT_{SD}$	ST_{SD}	**$MSTFT_{SD}$**
Rényi Entropy	12.16	12.23	11.56	13.57	12.20	**11.52**
PED	$2.07\cdot10^{-4}$	$1.97\cdot10^{-4}$	$3.12\cdot10^{-4}$	$0.79\cdot10^{-4}$	$2.01\cdot10^{-4}$	**$3.20\cdot10^{-4}$**

(35)	ASTFT	AST	MATFWFT	$STFT_{SD}$	ST_{SD}	$MSTFT_{SD}$
Rényi Entropy	11.43	11.51	10.61	11.36	13.03	**10.22**
PED	$3.39\cdot10^{-4}$	$3.17\cdot10^{-4}$	**$5.95\cdot10^{-4}$**	$3.56\cdot10^{-4}$	$1.14\cdot10^{-4}$	$1.69\cdot10^{-4}$

(36)	ASTFT	AST	MATFWFT	$STFT_{SD}$	ST_{SD}	**$MSTFT_{SD}$**
Rényi Entropy	12.06	12.49	11.12	12.95	12.27	**10.63**
PED	$2.04\cdot10^{-4}$	$1.58\cdot10^{-4}$	$3.87\cdot10^{-4}$	$1.17\cdot10^{-4}$	$1.79\cdot10^{-4}$	**$5.39\cdot10^{-4}$**

Enhancing Retinal Disease Classification with Dual Scale Twin Vision Transformers using OCT Imaging

Prakash Kumar Karn[1*], Waleed H Abdulla[2]
[1,2] The University of Auckland, New Zealand
[*]E-mail: pkar443@aucklanduni.ac.nz
[2]E-mail: w.abdulla@auckland.ac.nz

***Abstract* — Retinal optical coherence tomography (OCT) imaging has become a valuable tool in diagnosing and monitoring retinal diseases. Recent advancements in vision transformers have demonstrated impressive performance in image classification tasks. This research article presents a novel approach, named Dual Scale Twin Vision Transformer (DST-ViT), for retinal disease classification using OCT images.**

DST-ViT combines the advantages of dual-scale representation learning and the Twin transformer architecture to improve disease classification accuracy. By leveraging dual-scale input OCT images, DST-ViT effectively captures fine-grained details and global contextual information simultaneously, enhancing classification performance. With an outstanding average accuracy of 99.92%, DST-ViT surpasses existing state-of-the-art methods.

This research contributes to the growing understanding of ViT and their application in retinal disease classification. The promising results achieved by DST-ViT open possibilities for improved diagnosis and treatment planning in the field of ophthalmology.

I. INTRODUCTION

Optical Coherence Tomography (OCT) is a powerful imaging technique used to generate cross-sectional representations of various tissues and detect pathological structures within them [1]. In the context of retinal diseases, OCT enables the study of oedemas, which are fluid accumulation and leakages caused by the erosion of retinal blood vessels. These oedemas, particularly Diabetic Macular Oedemas (DME), are the leading causes of impaired vision in advanced countries. Timely and accurate diagnosis of DME is crucial to prevent permanent damage to retina [2].

Computer-aided analysis systems have emerged to decrease the bias associated with expert interpretation and enhance the monitoring and treatment of retinal pathologies. However, many of these methodologies remain opaque, obscuring their internal logic and making it challenging for experts to comprehend their predictions' underlying rationale [3]. This lack of interpretability is particularly concerning in ophthalmology, as misguiding information from black-box systems can significantly impact patient quality of life. Consequently, researchers have published methods to elucidate the internal methods of these black-box models. For instance, Vidal et al. [4][5] proposed an intensive voting approach to create a map of recognised pathological regions. Wang et al. [6] explored through gradient-based class-activation maps, and Reza et al. [7] utilised the LIME structure.

However, existing explainable methodologies primarily focus on representing what is already present in the image without establishing connections to past learning. To refer this space, Philipps et al.[8] introduced an architecture inspired by the ViT. This study introduces a transformer-based automatic retinal lesion segmentation method in SD-OCT images. It combines the advantages of ViT and CNNs, providing efficient long-range feature extraction and aggregation capabilities with data-efficient training. While a competitor achieves a higher Dice score, the proposed method stands out for its lower computational cost. This approach has been effectively transformed to other medical imaging domains, such as chest X-Ray [9] and the diagnosis of COVID-19 [10]. In natural language processing (NLP), transformers have shown great potential for sequence-to-sequence modelling [11]. As a result, the computer vision community is now interested in

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

discovering whether transformers can compete with CNN-based designs like ResNet [12] and EfficientNet [13]. Previous research on ViT was primarily focused on combining CNNs with self-attention, resulting in hybrid approaches [14], [15]. While these hybrid models achieve promising performance, they are computationally less efficient than attention-based transformers. ViT [16] introduced a convolution-free approach that utilises a sequence of embedded image patches as input to a standard transformer, achieving comparable performance to CNN models. However, ViT requires large-scale datasets like ImageNet21K[17] and JFT300M [18] for training. Subsequent research, such as DeiT [19], demonstrated that high-performance ViT models could be trained with less data with appropriate data augmentation and model regularisation. ViTs have gained significant research interest as an alternative to convolutional neural networks (CNNs) in visual perception tasks. ViTs offer several advantages, including the ability to model long-range dependencies, handle multi-modal data, and introduce less inductive bias. These characteristics make ViTs a promising candidate to replace CNNs as the foundational component in future visual perception systems.

The computational complexity of the spatial self-attention operation, however, presents a barrier when applying transformers to vision tasks. The number of pixels in the input image causes its complexity to increase in a quadratic fashion. Utilising locally grouped self-attention or self-attention in non-overlapping windows, as shown in the Swin Transformer, is a solution to this problem [20]. This strategy lessens computing complexity, but it restricts connectivity between various windows, leading to a constrained receptive field. For dense prediction tasks like picture segmentation and object detection, a broader receptive field is important. The Swin Transformer introduces a moving window procedure to address this limitation, gradually moving the boundaries of local windows as the network progresses. But the uneven sizes of these shifted windows can pose challenges when deploying models with specific frameworks like ONNX or TensorRT, which prefer equal-sized windows[21]. However, applying this methodology to Diabetic Macular Oedema classification in OCT images proved challenging due to the resemblance between diagnostic shapes and other regular patterns found in normal retinal layers.

In this research article, we proposed a novel approach, the Dual Scale Twin Vision Transformer (DST-ViT), for retinal disease classification using OCT image data. The DST-ViT model combines two powerful techniques, the dual scale[22] and the twin ViT framework[22], to improve the accuracy and interpretability of retinal disease diagnosis. By using these advanced methodologies, DST-ViT can better understand and classify retinal diseases. The dual-scale approach allows the model to capture fine details and overall context from the OCT images, resulting in more accurate predictions. Additionally, the twin ViT framework enhances the interpretability of the model's decisions, providing insights into how it makes its diagnoses. We demonstrated the effectiveness of our approach in addressing the challenges posed by DME classification and provided comprehensive evaluations and comparisons with existing methods. By leveraging these advancements, DST-ViT offers a more reliable and understandable method for diagnosing retinal diseases, contributing to improved patient care and treatment planning.

Rest of the paper is structured as follows: Section II discuss about dataset, methodology and experiment performed. Section III discuss about the results we obtained and comparison of our proposed model with the state-of-art. It also discusses about limitation and future scope of our proposed model.

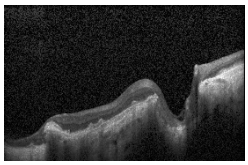
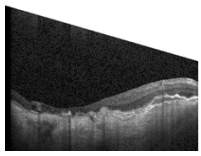
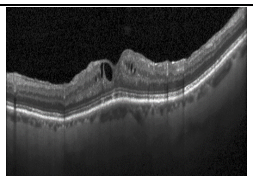
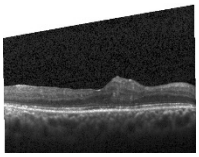
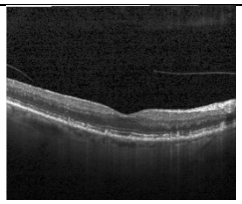
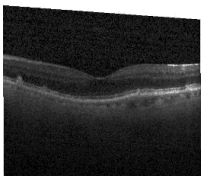
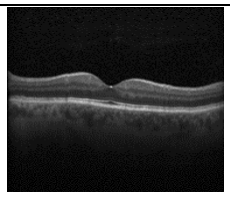
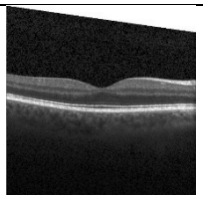
II. Materials and Methods

A. Datasets

We used a dataset of optical coherence tomography (OCT) image data for this research [23]. This publicly accessible dataset was gathered between 2013 and 2017 from five institutions. There are 1,000 test images and 108,309 total

training images in it. The collection includes four classes: normal OCT, diabetic macular oedema, and choroidal neovascularization (8,616 pictures from Drusen). The sample images of good quality and skewed image from each class is presented in Table 1. For this research, we initially combined the existing training and testing data into a single dataset. Subsequently, we followed a rigorous approach to ensure robust model evaluation by dividing each class into three distinct subsets: one for training, one for testing, and one for validation. This division was carried out in an 80:10:10 ratio, where 80% of the data from each class was allocated to training, 10% for testing, and an additional 10% for validation. This strategy allowed us to maintain a balanced and representative distribution of data across all classes while facilitating effective model training, evaluation, and hyperparameter tuning

Table 1 Sample Images from OCT-2017 Dataset

Retina Condition	Good Quality image	Skewed Image
CNV		
DME		
Drusen		
Normal		

In the case of CNV, abnormal vessel growth occurs from the choroidal vasculature to the neurosensory retina through Bruch's membrane. This results in variations below or above the RPE layer and within the intra-retinal area. These variations can manifest as macular fluid regions (depicted in black), a bumpy RPE layer, and poorly defined boundaries. For DME, intraretinal fluids are present between the Inner Limiting Membrane (ILM) and Retinal Pigment Epithelium (RPE) layers. These fluids appear as dark regions in the image. In the case of Drusen, the RPE layer exhibits a bumpy shape instead of a flat shape. Identification of Drusen can help to diagnose age-related macular degeneration[2].

Overall, this dataset provides a diverse range of OCT images showcasing different retinal conditions, enabling the development and evaluation of algorithms for retinal disease classification.

Within the dataset, the images display variations in both size and quality, accompanied by the presence of black-and-white noise, as depicted in Table 2. The figure consists of four rows, each corresponding to a specific class: CNV, DME, Drusen, and Normal.

In the first column, all the images share a square shape. However, the second column contains images with non-square shapes and a white background. Additionally, the images within the dataset possess different zoom rates. The images in the second column have lower zoom ratios than those in the first.

B. Methodology

This research article proposes a novel approach to enhance image features by combining image patches at different scales, also known as tokens in a transformer. After being linearly embedded, these tokens are input into a Twin transformer architecture that incorporates local and global self-attention mechanisms. Unlike the shifted window approach used in the original Swin transformer[20], this modified version leverages a combination of local and global self-attention to achieve more effective feature extraction. In the early ViT, single-scale non-overlapping patches were given as the input sequence to the transformer, which lacks the feature information and relation with the neighbours. To circumvent this, a swin transformer was proposed, which contains shifted window self-attention module which acts like a convolution of the image but with less computational cost. Later, a more efficient network was proposed by [21], containing local self-attention and global self-attention cascaded over each other with residual

connection. In this paper, we have modified the Twin model proposed by[21]. with a dual scale input. The detailed block diagram of the proposed model in given Fig. 1.

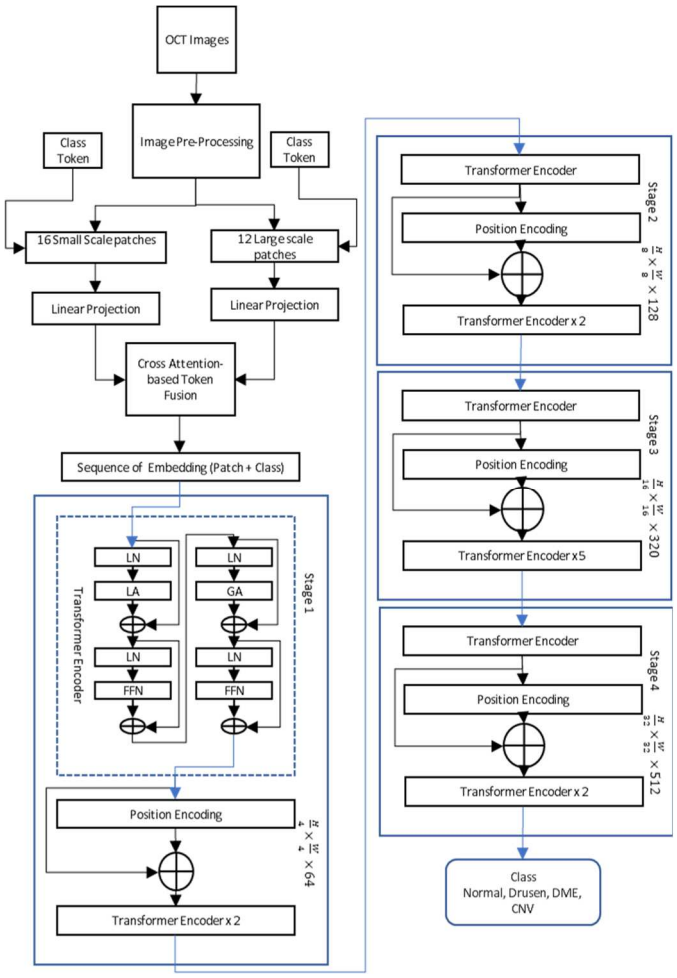


Fig. 1 Detailed Block Diagram of Proposed Dual Scale Twin Vision Transformer (DST-ViT). LN: Layer Norm, LA: Local Self-Attention, GA: Global Self-Attention, FFN: Feed Forward Network

The proposed model can be explained into two separate models, i.e., dual scale input and Local & Global self-attention-based ViT. We have seen in Table 2 that the OCT image data which we have used for this experiment contain some skewed images. These images contain some white backgrounds, which are not our subject of interest and may cause bias in the model. Also, the number of images in each dataset class is not even, which causes an imbalance in the training process. We have used some pre-processing techniques adapted from [21] to address this issue. The pre-processing involves several steps, including noise removal, rotated retina layer adjustment, and final retina layer cropping. A histogram is computed by projecting the image onto the vertical axis to ensure accurate alignment and proper orientation of the retina layer. This histogram provides information to adjust the rotated retina layer, improving the alignment and consistency of the estimated features. The final step involves cropping only the retina layer from the image, discarding irrelevant information and isolating the region of interest. The cropped retina layer serves as the input for the subsequent Transformer block.

The idea of dual-scale input for transformer block is taken from [22], which shows the importance of multi-scale input in ViT. For the sake of computational efficiency, we have used dual-scale input. As per the experiment conducted by Chun-Fu [22], images partitioned in (12, 16) patches exhibit better accuracy and have less computation burden than (8, 16). Similar to the original BERT model [24], we have added an additional class token to each small and large-scale patch. This information of position and class for both scales is fused using cross-attention-based toke fusion proposed by[22]. This approach employs the CLS token, which learns abstract information within its branch, as an agent for knowledge exchange. By interacting with patch tokens from other branches, the CLS token facilitates information integration at different scales, complementing the representation of individual patch tokens.

The Swin Transformer [25] enhances communication among patches and receptive fields by alternating local window-based attention with shifted window partitions in successive layers. However, this method lacks optimization for mobile devices and relies on memory-intensive cyclic shift operations unsupported by popular inference frameworks. In contrast, Twins models excel by eliminating the need for such operations and relying solely on optimized matrix multiplications in modern deep learning frameworks. In our proposed model, we evaluated self-attention and global

attention features on the embedded token to extract class-wise details of OCT images. Details on Local self-attention (LA) and Global self-attention (GA) and its implementation in ViT can be found in [21]

C. Experiment

The experiment was conducted on a Windows 10 core i7 processor and 32 Gb memory. We used Python 3.10 and PyTorch 1.11.0 to build the model. The models were trained on an NVIDIA Quadro P620 GPU with 512 CUDA cores. The weights were initialised using Xavier initialisation [26]. The hyperparameters β1 and β2 of the Adam optimiser represent the initial decay rates used to estimate the first and second moments of the gradient. These decay rates are multiplied by themselves exponentially at the end of each training step (batch). Reducing β1 and β2 values can slow down the learning process. Therefore, if the training progress is excessively fast, adjusting these hyperparameters can be beneficial in controlling the learning rate. In this experiment, the Adam optimiser with a β1 value of 0.900 was utilised for optimisation during training. The initial learning rate was set to 0.0002 and decayed to 0.00001 towards the end. All OCT images were resized and cropped to 224 × 224, and a batch size of 32 was used. Sixteen tokens were generated for the small-scale image patch and 12 tokens for the large-scale patch. The model was trained for 200 epochs, and the performance was evaluated using the model at the final epoch.

D. Evaluation of Model

We utilise the softmax method to evaluate the classification performance to convert logits into class probabilities. The predicted category is determined by selecting the class with the highest probability. Accuracy, precision, recall, and F1-score are employed as evaluation metrics. The formulas for these evaluation metrics are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precison = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = \frac{2(Precision \times Recall)}{Precision+Recall} \quad (4)$$

Here, TP refers to true positive, TN refers to true negative, FP refers to false positive, and FN refers to false negative. These metrics provide comprehensive insights into the performance of the classification model. Apart from this, we have also used Score-CAM[27] to visualise the prediction and decision-making capability of the model in the form of a heatmap of the OCT image.

III. Results and Discussion

The result of this experiment is represented in two forms: one with performance metrics and another with heat map generation of the OCT image. In Table 2, we have calculated evaluation parameters (Accuracy, Precision, Recall and F1 Score) to measure the robustness of our proposed model. The average combined class accuracy is 99.92% whereas individual class accuracy ranges between 99.75-100%. Out of all classes, DST-ViT has been shown to have 100% accuracy for diabetic macula oedema (DME) and normal image. This also implies that there is less chance of misdiagnosing patients with no retinal pathology.

Table 2 Evaluation metrics of the proposed Model on the OCT 2017 dataset.

Class	Accuracy	Precision	Recall	F1-Score
Normal	1.0000	0.9990	1.0000	0.9994
CNV	0.9996	0.9960	1.0000	0.9979
DME	1.0000	0.9990	0.9980	0.9984
Drusen	0.9975	0.9985	0.9975	0.9979
Average	**0.9992**	**0.9981**	**0.9988**	**0.9984**

No model can be evaluated unless it is not compared with the state-of-art. We have compared our proposed model with the literature using transformer model to classify OCT image and with other machine learning model. The comparison is shown Table 3. The highlighted values are the average output

of each class. The result illustrated in Table 3 shows that our algorithm has not only outperformed the performance of another transformer model with fair margin but with other machine learning model with greater margin.

Table 3: Comparison of the proposed model with various *ViT*s and machine learning models.

		Accuracy	Precision	Recall	F1-Score
Transformer Models	ViT[16]	99.00	99.07	99.07	99.07
	Swin Transformer[20]	99.07	99.07	99.07	99.07
	S-Poly-transformer[21]	99.80	99.70	99.80	99.80
	DST-ViT (Proposed)	**99.92**	**99.81**	**99.88**	**99.84**
Other Machine Learning	Kim et.al.[28]	98.90	98.90	98.90	98.90
	Saleh et.al.[29]	98.50	98.50	98.50	98.50
	Inception V3[30]	96.60	97.40	97.80	97.60
	Kermany et.al.[23]	96.10	96.12	96.10	96.10
	MDFF[31]	99.60	99.60	99.60	99.60

As discussed in the evaluation model section, we have visualised some of the test images using a heat map. This heat map was generated using a model proposed by [27] to visualise how the proposed model focuses on the region of greater attention with higher probabilities. A few examples of test image with corresponding heat maps are given in Fig. 2.

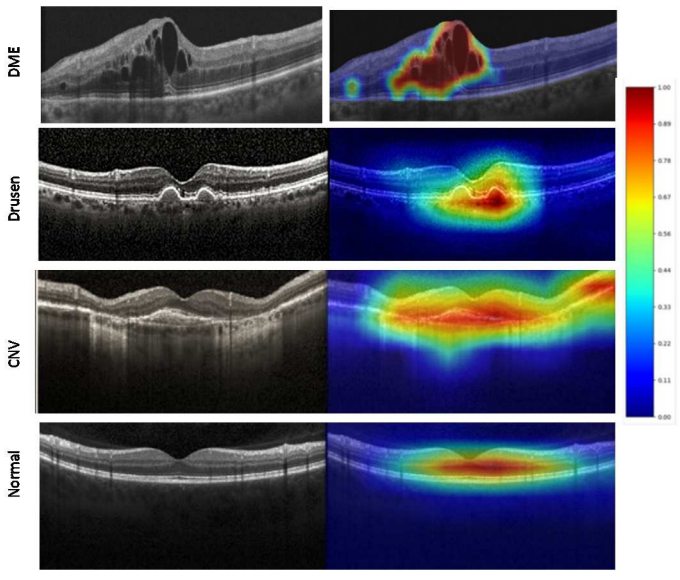


Fig. 2 Heat map of OCT image from each class.

Heatmaps provide visual evidence of the regions within an OCT image that influenced the algorithm's decision. Clinicians can visually confirm whether the algorithm focused on the correct features associated with drusen and CNV, enhancing their confidence in the diagnosis. In this research article, we proposed a novel approach, Dual Scale Twin Vision Transformer (DST-ViT), for retinal disease classification using OCT image data. Our method combines the advantages of dual-scale representation learning and the transformer architecture to enhance the accuracy of disease classification. The experimental results demonstrate the effectiveness of DST-ViT in accurately identifying various retinal diseases. Furthermore, the comparison with existing state-of-the-art methods showcases the superiority of DST-ViT in terms of classification. The incorporation of dual-scale representation allows DST-ViT to capture fine-grained details and global contextual information simultaneously, leading to improved classification accuracy. Overall, our study highlights the potential of DST-ViT as a promising tool for retinal disease classification, with implications for early diagnosis and treatment.

IV. Conclusion

In conclusion, our research article presents the Dual Scale Twin Vision Transformer (DST-ViT) as a highly effective approach for retinal disease classification using OCT image data. By leveraging dual-scale representation learning and the local and global self-attention-based transformer architecture, DST-ViT demonstrates superior accuracy in identifying various retinal diseases. The experimental results strongly support the superiority of DST-ViT over existing state-of-the-art methods based on comprehensive performance evaluation metrics.

However, it is essential to acknowledge certain limitations in our study. The performance of DST-ViT may be influenced by factors such as dataset size and class imbalance, which warrant further investigation. Additionally, exploring the interpretability of the model's decisions can provide valuable insights into the classification process and foster trust in the system.

For future work, we suggest addressing these limitations and investigating the generalizability of DST-ViT to different OCT image datasets. Efforts can be directed toward exploring

transfer learning and domain adaptation techniques to enhance the model's performance in scenarios with limited annotated data. Furthermore, integrating clinical data and longitudinal OCT scans holds promise for providing valuable information for disease prognosis and treatment planning. These avenues of research offer exciting prospects for advancing retinal disease classification using the DST-ViT framework.

The novel approach presented in this paper, DST-ViT, showcases the potential of ViT in the field of retinal disease classification. Its success highlights the importance of leveraging advanced machine-learning techniques to improve medical diagnosis and patient care in ophthalmology.

REFERENCES

[1] M. R. Hee, "Optical Coherence Tomography of the Human Retina," *Arch. Ophthalmol.*, vol. 113, no. 3, p. 325, Mar. 1995.

[2] P. K. Karn and W. H. Abdulla, "On Machine Learning in Clinical Interpretation of Retinal Diseases Using OCT Images," *Bioengineering*, vol. 10, no. 4, 2023.

[3] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electron.*, vol. 8, no. 8, pp. 1–34, 2019.

[4] P. L. Vidal, J. de Moura, M. Díaz, J. Novo, and M. Ortega, "Diabetic macular edema characterization and visualization using optical coherence tomography images," *Appl. Sci.*, vol. 10, no. 21, pp. 1–23, 2020.

[5] P. L. Vidal, J. De Moura, J. Novo, and M. Ortega, "Cystoid Fluid Color Map Generation in Optical Coherence Tomography Images Using a Densely Connected Convolutional Neural Network," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2019-July, no. July, pp. 1–8, 2019.

[6] Y. Wang, M. Lucas, J. Furst, A. A. Fawzi, and D. Raicu, "Explainable Deep Learning for Biomarker Classification of OCT Images," *Proc. - IEEE 20th Int. Conf. Bioinforma. Bioeng. BIBE 2020*, pp. 204–210, 2020.

[7] M. T. Reza, F. Ahmed, S. Sharar, and A. A. Rasel, "Interpretable Retinal Disease Classification from OCT Images Using Deep Neural Network and Explainable AI," *Proc. Int. Conf. Electron. Commun. Inf. Technol. ICECIT 2021*, pp. 1–4, 2021.

[8] D. Philippi, K. Rothaus, and M. Castelli, "A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images," *Sci. Rep.*, vol. 13, no. 1, pp. 1–14, 2023.

[9] S. Park *et al.*, "Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation," *Nat. Commun.*, vol. 13, no. 1, pp. 1–11, 2022.

[10] D. Shome *et al.*, "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, 2021.

[11] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the Real World: A Survey on NLP Applications," *Inf.*, vol. 14, no. 4, 2023.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016.

[13] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[14] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. Figure 1, pp. 16514–16524, 2021.

[15] H. Zhao, J. Jia, and V. Koltun, "Exploring Self-attention for Image Recognition," *Proc. IEEE Comput.*

Soc. Conf. Comput. Vis. Pattern Recognit., pp. 10073–10082, 2020.

[16] D. Alexey *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," *arXiv:2010.11929*, 2021.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2010.

[18] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 843–852, 2017.

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," pp. 1–22, 2020.

[20] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002.

[21] X. Chu *et al.*, "Twins : Revisiting the Design of Spatial Attention in Vision Transformers. NeurIPS, 1–14.wins : Revisiting the Design of Spatial Attention in Vision Transformers," no. NeurIPS, pp. 1–14, 2021.

[22] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classificatio."

[23] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, 2018.

[24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[25] H. Cao *et al.*, "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation," pp. 205–218, 2023.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, 2010.

[27] H. Wang *et al.*, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2020-June, pp. 111–119, 2020.

[28] J. Kim and L. Tran, "Ensemble learning based on convolutional neural networks for the classification of retinal diseases from optical coherence tomography images," *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2020-July, pp. 532–537, 2020.

[29] N. Saleh, M. Abdel Wahed, and A. M. Salaheldin, "Transfer learning-based platform for detecting multi-classification retinal disorders using optical coherence tomography images," *Int. J. Imaging Syst. Technol.*, vol. 32, no. 3, pp. 740–752, May 2022.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016.

[31] V. Das, S. Dandapat, and P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images," *Biomed. Signal Process. Control*, vol. 54, p. 101605, 2019.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Classification of Infant Sleep/Wake States: Cross-Attention among Large Scale Pretrained Transformer Networks using Audio, ECG, and IMU Data

Kai Chieh Chang*†, Mark Hasegawa-Johnson*†, Nancy L. McElwain†§, Bashima Islam‡
* Department of Electrical and Computer Engineering, University of Illinois at Urbana Champaign, USA
† Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, USA
§ Department of Human Development and Family Studies, University of Illinois at Urbana-Champaign, USA
‡ Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, USA
E-mail: kcchang3@illinois.edu, jhasegaw@illinois.edu, mcelwn@illinois.edu, bislam@wpi.edu

***Abstract*—Infant sleep is critical to brain and behavioral development. Prior studies on infant sleep/wake classification have been largely limited to reliance on expensive and burdensome polysomnography (PSG) tests in the laboratory or wearable devices that collect single-modality data. To facilitate data collection and accuracy of detection, we aimed to advance this field of study by using a multi-modal wearable device, LittleBeats (LB), to collect audio, electrocardiogram (ECG), and inertial measurement unit (IMU) data among a cohort of 28 infants. We employed a 3-branch (audio/ECG/IMU) large scale transformer-based neural network (NN) to demonstrate the potential of such multi-modal data. We pretrained each branch independently with its respective modality, then finetuned the model by fusing the transformer layers with cross-attention. We show that multi-modal data significantly improves sleep/wake classification (accuracy = 0.880), compared with use of a single modality (accuracy = 0.732). Our approach to multi-modal mid-level fusion may be adaptable to a diverse range of architectures and tasks, expanding future directions of infant behavioral research.**

I. Introduction

Sleep is a physiological and behavioral process central to brain development during infancy [1]–[4]. During the first six months of life, infants spend more time sleeping than awake [5]. Further, indices of infant sleep quality and quantity are associated with subsequent cognitive and language development [6]–[11], attention regulation [12]–[14], social-emotional functioning [10], [12], and physical health [15], [16]. Given the importance of infant sleep to development, automated and unobtrusive monitoring of sleep is critical. Thus, we focus on the task of sleep/wake classification in this study.

Laboratory-based polysomnography (PSG) is the gold-standard sleep assessment for adults and infants alike [18], [19]. However, given the high level of burden and expense of laboratory PSG, researchers have used wearable devices for sleep monitoring [20]. Notably, past studies of sleep/wake classification using wearable devices typically extract sleep features from single-modality data. For example, with actigraphy data, [21] carried out extensive studies on adult sleep/wake classification using traditional signal processing methods, classic machine learning methods, and several deep learning methods. Electrocardiogram (ECG) signals are also often used because stable QRS complex and longer R peak period can indicate sleep stages [22]. Studies such as [23] used ECG data to classify adult sleep/wake with a 5-layer convolutional neural network (CNN) concatenated with 2 fully connected (FC) layers. A signal-processing-based approach was suggested by [24], where audio recordings of adults' respiratory sounds were used to estimate sleep/wake likelihood from signal level features such as autocorrelation. All published studies of automatic sleep/wake classification use different datasets, therefore their accuracies are not strictly comparable, but the accuracies reported in all four of these studies are between 81%

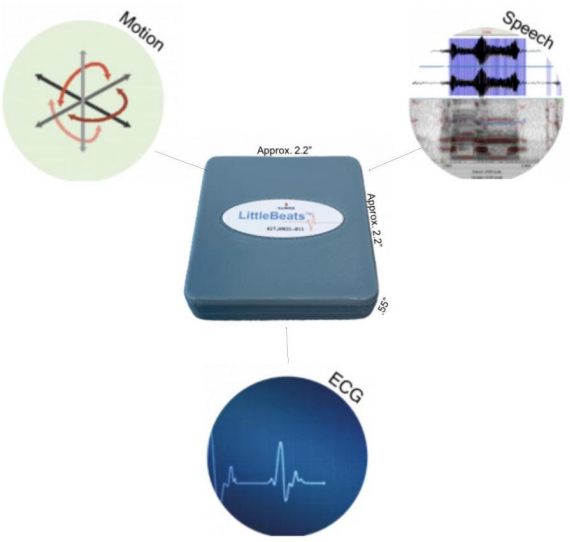


Fig. 1. LittleBeats (center) is capable of collecting audio, ECG, and IMU data. Reproduced with permission from [17]

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

and 84% (see Table I for details). A few studies have used bimodal data to further improve classification performance. Ref. [25] concatenated the frequency bins from respiratory signals and ECG to feed into a FC based network, achieving classification accuracy of 85.3%. Ref. [26] uses random forest to classify sleep/wake from acceleration and heart rate data collected by Apple Watch, achieving 87.3% accuracy. Recent work tend to focus on a more involved sleep classification with 5 stages. Ref. [27] extracted infant vocalizations from audio recordings and physical motions from video recordings, and classified sleep stages using a random forest, resulting in an average accuracy near 85%. Supratak et al.'s DeepSleepNet [28] and their subsequent stage of the art (SOTA) TinySleepNet [29] used only electroencephalogram (EEG) to reach 87.5% in accuracy.

To achieve further improvements in accuracy, we leveraged trimodal data collected by an infant wearable device, LittleBeats (LB), as shown in Fig. 1. LB can synchronously record infant sounds (via microphone [audio]), physiology (via 3-lead ECG), and motion (via inertial measurement unit [IMU]) for extended periods of time (8-10 hours) in the home context.

To perform sleep/wake classification using LB multi-modal data, we developed an ensemble of three large scale transformer networks pretrained on unimodal audio, ECG, and IMU datasets, and fine-tuned using trimodal data. This system has the potential to combine all important feature extraction done by the research mentioned above, including the extraction of vocalization and breathing patterns from audio recording, QRS complex features from ECG, and position and motion from IMU. To the best of our knowledge, our study is the first to access these three modalities jointly on sleep/wake classification. We show superior classification performance using all three modalities compared with any single modality or pairs of modalities combined.

In contributing to neural network (NN) architectures, we explore the benefits of pretraining large scale transformer networks on unlabeled audio, ECG, and IMU data, then finetuning a cross-attention based fusion architecture on a small LB labeled dataset. Unsupervised pretraining involves training a NN on a large, diverse dataset, which enables it to learn generic latent features that can be applied to a wide range of tasks. There is a wide range of pretraining schemes for audio, ECG, and IMU data. Specifically, wav2vec 2.0 (W2V2) [30] showed the benefits of pretraining audio using contrastive loss, achieving state of the art performance in tasks such as automatic speech recognition. Ref. [31] combined W2V2 and contrastive multi-segment coding to pretrain on ECG signals, achieving outstanding performance on cardiac arrhythmia classification and patient identification. LIMU-BERT [32] is derived from bidirectional encoder representations from transformers (BERT) [33], and is used to pretrain transformers with unlabeled IMU data. Inspired by the above architectures, we pretrained three branches of large scale transformer networks, one for each modality, and were able to extract latent features useful for sleep/wake classification.

To utilize features from all three modalities, we implemented a fusion strategy that relies on cross-attention among transformer layers. Multimodal fusion is often done by early fusion, where the input from each modality is concatenated and fed into the network as a single input, or late fusion, where each modality is processed separately and the outputs are combined at the final stage. We define early fusion as fusing right after feature extraction and not fusion of raw data. Much work has been done on these 2 fusion techniques. Ref. [34] employed early fusion by cascading learned features from ECG and IMU directly, and passed them through a dense network. Ref. [35] described a simple late fusion of different branches by summing or averaging the logit outputs. Recently, intermediate fusion has become more common. This type of feature fusion happens in the middle of the NN, and uses a variety of architectures depending on the tasks. Ref. [36] extensively used so-called "cross-stitch" modules for soft parameter fusion, taking multiple linear weighted sums of each feature extraction layer as input for the next layer. Cross-attention based fusion techniques were explored in recent papers such as [37] [38], where the attention layers take concatenated features from different modalities as input, and [39], where a single cross-attention layer is used to share information between branches. Our approach is innovative, in that we fuse the three transformer networks by alternating self and cross-attention at different layers, which both preserves each branch's transformer features and incorporates attention from other modalities.

In sum, in addition to being the first study to combine audio, ECG, and IMU for classification of infant sleep/wake states, we develop an innovative cross-attention based fusion for large scale pretrained transformer networks to combine three modalities. We not only reiterate the well known benefit of pretraining, but also demonstrate the ability of transformers to be trained using cross-attention-based multi-modal fusion to improve accuracy. We believe our work lays the groundwork for improved accuracy in a wide variety of signal processing tasks by the use of multimodal wearable devices such as LB.

II. Method

Our architecture is an ensemble of three similarly structured transformer networks, with each branch targeting audio, ECG, or IMU data respectively, as shown in Fig. 2.

For the audio branch, we utilize a standard 12-layer W2V2 [30] network because W2V2 repeatedly shows superb performance in various speech-related tasks such as automatic speech recognition and phoneme recognition. A recent study [40] showed an oracle W2V2 pretrained on 4300 hours of unlabeled infant-family audio data collected by LB or the Language ENvironment Analysis (LENA) system [41] increased performance of speech diarization and vocalization type classifications. As infant vocalization features may well include cues of sleep, we adopt pretrained weights from [40] for the audio W2V2 branch.

As for ECG data, we were inspired by the similarity between speech and ECG. As early as 1940, researchers [42] were

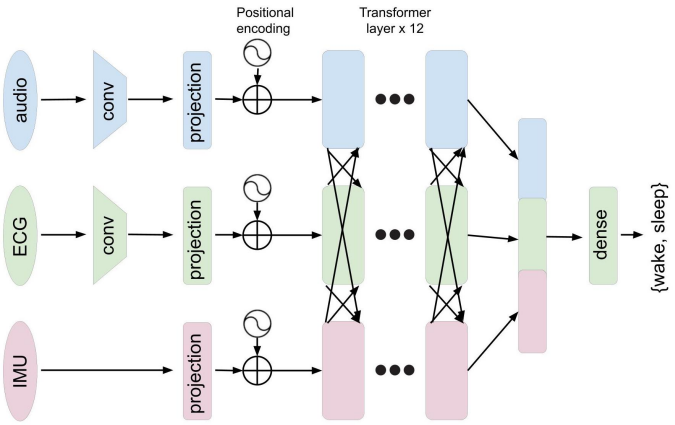


Fig. 2. The LittleBeats Model Architecture. Each of the audio (top), ECG (center), and IMU (bottom) branches consists of feature extraction (convolutional and/or feedforward projection), positional encoding, and 12 transformer layers. Each branch is pretrained individually on unlabeled data. During labeled finetuning, the three branches are combined via cross-attention at the feature level. Their outputs are then concatenated and fed into dense layers to output logits.

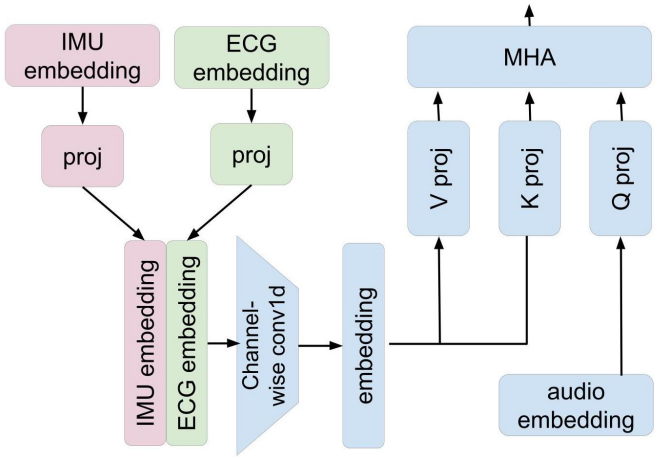


Fig. 3. Cross-attention for the audio branch. Multiheaded attention layer (MHA) takes the audio embedding as query. IMU and ECG outputs from the other two branches first get linearly projected and concatenated into a 2-channel embedding. The embedding is reduced to 1-channel by a 1-d convolution layer and passed into MHA as key and value.

actively using impulse trains to imitate glottal excitation for speech generation. R-peaks in ECG data have a similar time domain structure, so a speech-centric NN should be able to learn ECG features. Speech and ECG differ in one respect: speech contains periodicity at many frequencies (pitch and formants), while ECG is dominated by the inter-beat interval, therefore ECG features may be learnable using less pretraining than speech. We pretrain a standard 12-layer W2V2 using 574 hours of unlabeled ECG data to utilize our observations above and to match the structure of the audio branch.

IMU data does not share similar structure as audio or ECG: an IMU signal is composed of parallel 3-axis accelerometer and 3-axis gyroscope signals, whose correlations are important for signal interpretation. To take full advantage of the information contained in our IMU data, we used a pretraining paradigm based on LIMU-BERT [32]. In place of W2V2's CNN feature extraction layers, LIMU-BERT begins with a single fully-connected feature projection layer per input sample, followed by several layers of transformers. This network is pretrained with Masked Language Model objectives and a Span Masking mechanism. We pretrained 574 hours of IMU data collected from LB, on a modified LIMU-BERT architecture where we lifted the restriction that all transformers share weights, extended the number of layers to 12, and matched the transformer implementation to the one found in W2V2.

To maximize the model's ability to learn across modalities, or to extend the transformer attention beyond a single modality, we utilized linear transformation and cross-attention every 4 transformer layers, as inspired by [43] and their work using cross-attention across audio channels. An example of the cross-attention mechanism for the audio branch is shown in Fig. 3. A cross-attention layer for the audio branch consists of multiple cross-attention heads, each of which takes its query from the previous layer's audio embedding, and its values and keys from a fusion of the previous layer's ECG and IMU branches. ECG and IMU embeddings are first each passed through a linear projection layer, producing outputs with the same embedding dimension as audio, then concatenated to create a 2 channel embedding. A channel-wise 1d convolutional layer is then used to create the vector that serves as key and value of the audio cross-attention layer. The multiple audio cross-attention heads are then concatenated, added to a residual connection (from the previous audio layer), normalized, and passed through a feedforward layer. The ECG and IMU cross-attention layers have the same architecture, with each modality receiving query inputs from its own previous layer, while each modality's key and value inputs come from the other two modalities.

Mathematically, we can denote each hidden input embedding to the transformer layer as $\mathbf{H}_k \in \mathbb{R}^{N \times L_e}$, where the indices (k, l, m) are any permutation of $\{\text{audio}, \text{ECG}, \text{IMU}\}$, N is the number of samples in the input, and L_e is the architecture embedding size. Let's define the query, key, and value projection of each branch to be $\mathbf{W}^{\{Q,K,V\}} \in \mathbb{R}^{L_e \times L_e}$. Let's constrain the discussion of attention regarding one head of one branch where $k = \text{audio}$. The attention layers can be self attention or cross-attention depending on the layer index i. We found using cross-attention at layers 1, 5, and 9 (i.e. $i\%4 = 1$ for layers $0 \leq i \leq 11$) gives the best performance.

$$\text{SelfAttn}_{i,k} = \text{softmax}\left(\frac{\mathbf{H}_k\mathbf{W}_i^Q\left(\mathbf{H}_k\mathbf{W}_i^K\right)^T}{\sqrt{L_e}}\right)\mathbf{H}_k\mathbf{W}_i^V$$

$$\text{CrossAttn}_{i,k} = \text{softmax}\left(\frac{\mathbf{H}_k\mathbf{W}_i^Q\left(\mathbf{H}_{\text{cross}}\mathbf{W}_i^K\right)^T}{\sqrt{L_e}}\right)\mathbf{H}_{\text{cross}}\mathbf{W}_i^V$$

$$\mathbf{H}_{\text{cross}} = \mathbf{A}\begin{bmatrix}\text{proj}(\mathbf{H}_l^T)\\ \text{proj}(\mathbf{H}_m^T)\end{bmatrix}$$

where $\mathbf{A} \in \mathbb{R}^{L_e \times 2}$ is a trainable mixing weight. This discussion is applicable to the other heads and other branches.

The three branches' outputs are concatenated and passed through three dense layers with ReLU activations to produce binary logits for sleep and wake. With the pretrained weights loaded onto each branch and the ensemble in place, we use frame-wise cross entropy loss to finetune the network.

III. Experiment

A. Dataset

Participants were recruited via study flyers distributed to local community organizations as well as online listservs serving families of young children. In the context of larger studies of child behavioral development, study coordinators provided participating families with a LB or LENA kit, along with instructions for conducting associated recordings in the home. Families were asked to complete 2 to 3 daylong (8+ hours) recordings over a 2-week period. All study procedures were approved by the Institutional Review Board at the University of Illinois Urbana-Champaign.

Unlabeled data included (a) 4300hr of audio home recordings (1000hr collected by LB and 3200hr by LENA) from 245 families with children under 5 years of age (see [40] for further details), as well as (b) 574hr of ECG data and (c) 574hr of IMU data collected by LB from 28 families with children (54% female) under 5 years of age (Mean = 26 months, Range: 3-65 months). For the labeled data, we gathered 68.5hr of synchronized audio, ECG, and IMU data from the same 28 families. Four human annotators labeled infant sleep and wake states from 1.5hr segments of LB audio files. Annotation of sleep states was facilitated by referring to (a) parental reports of their child's activities while wearing the LB device in the home, (b) visual inspection of the wav form of the audio file, and (c) listening to audio recordings for indicators of infant sleep (e.g., slowed steady breathing, no vocalizations or movement). Human annotators were trained by the third author and achieved high inter-observer reliability on set of training tapes ($>$ 80 percent agreement). Annotators took detailed notes about any files where they had low confidence in identifying sleep versus wake states. All low-confidence files were reviewed by the team and third author; in these cases, the final labels were made by consensus, thus increasing confidence regarding reliability of the labels. Although PSG and EEG are the gold-standard for sleep detection, we did not collect these measurements in the current study. We are currently extending the work reported here, however, using PSG labels obtained during a laboratory sleep study among adult patients.

We randomly divide the labeled data into training set (50hr, 25 families), validation set (4.5hr, 1 family), and testing set (14hr, 2 families). Data from families/infants in the training, validation, and testing datasets did not overlap.

B. Data Preprocessing

As a first step, we synchronized data across modalities. LB records each modality in slightly different time segments. For each segment, LB records the timestamps in absolute UTC start time and end time, and the associated start and end sample index. Formally, let's define

$$\mathbf{T}_k^{\text{start}}, \mathbf{T}_k^{\text{end}}, \mathbf{S}_k^{\text{start}}, \mathbf{S}_k^{\text{end}} \in \mathbb{R}^{L_k}$$

where for each modality, each element in $\mathbf{T}$ is the absolute UTC start or end time, each element in $\mathbf{S}$ is starting or ending data sample index, L is the number of data chunks.

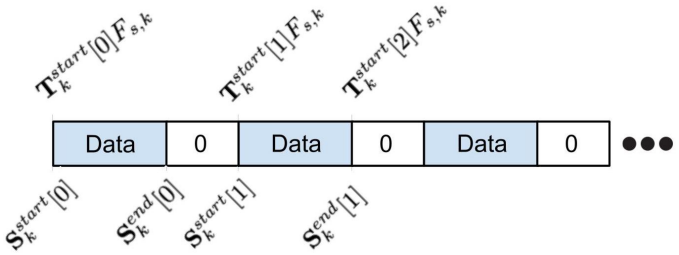


Fig. 4. Zeroing data according to timestamp. For each data segment, the recorded UTC time multiplied by sampling frequency ($F_{s,k} * (\mathbf{T}_k^{\text{end}}[i] - \mathbf{T}_k^{\text{start}}[i])$) is longer than the recorded number of samples ($\mathbf{S}_k^{\text{end}}[i] - \mathbf{S}_k^{\text{start}}[i]$). The missing data are simply filled with zeros.

LB is designed to function continuously despite variability in the number of nonzero IMU samples, therefore it sometimes fails to record all samples in a given audio or ECG segment. Such missed samples can be detected by comparing the known ground truth duration of a segment, $F_{s,k} * (\mathbf{T}_k^{\text{end}}[i] - \mathbf{T}_k^{\text{start}}[i])$, to the number of samples it contains, $(\mathbf{S}_k^{\text{end}}[i] - \mathbf{S}_k^{\text{start}}[i])$, where $F_{s,k}$ is the sampling frequency for modality k. When an incomplete segment is detected, we fill in the missing data with zeros. We obtain $\hat{\mathbf{Z}}_k$ where each of the vectors has length $F_{s,k} * (\mathbf{T}_k^{\text{end}}[i] - \mathbf{T}_k^{\text{start}}[i])$. This is shown in Fig. 4.

Timestamps of different modalities are approximately but not precisely aligned, because the writing to SD card on LB is asynchronous. In order to make multimodal processing possible, we simply discard the data that did not overlap according to the timestamp and truncate $\hat{\mathbf{Z}}_k$ to $\mathbf{Z}_k \in \mathbb{R}^L$ where L is the same across all modalities. $\mathbf{Z}_k$ is then segmented into 30s segments $\mathbf{X}_k$.

Human annotators mark the beginning and ending of each period of sleep. These times are used to assign, to each 30s segment $\mathbf{X}_k$, a binary label $\mathbf{Y}_k \in \{0, 1\}$ where 0 is wake and 1 is sleep. If both sleep and wake labels are present in segment k, $\mathbf{Y}_k$ is set equal to the label with longer duration.

C. Metric

Because this is a binary classification task, we evaluated standard accuracy, precision, recall, F1-score, and Cohen's kappa κ.

D. Implementation Details

Audio sample rate is downsampled from 24000Hz to 16000Hz and ECG sample rate is upsampled from 2381Hz to 16000Hz, to match the default sampling rate of W2V2. IMU data are kept at a sampling rate of 150Hz because upsampling to 16000Hz might introduce too many artifacts. We did not

TABLE I

RESULTS COMPARING ARCHITECTURES WITH DIFFERENT NUMBER AND TYPES OF MODALITIES. RESULTS IN THE FIRST SECTION ARE OUR WORK; RESULTS IN THE SECOND SECTION ARE BASELINES COMPUTED USING THE SAME TEST SET; RESULTS IN THE THIRD SECTION ARE COPIED FROM THE CORRESPONDING REFERENCE PAPERS, AND WERE THEREFORE COMPUTED USING INCOMPARABLE TEST SETS.

Test Set	Architecture	Modality	↑ Accuracy	↑ Precision	↑ Recall	↑ F1	↑ Kappa
LB 68.5 hr	W2V2-mod	Audio	0.797	0.794	0.909	0.848	0.546
	W2V2-mod	ECG	0.732	0.704	**0.980**	0.819	0.347
	LIMU-BERT-mod	IMU	0.786	0.837	0.815	0.826	0.549
	Proposed	Audio + ECG + IMU	**0.880**	**0.881**	0.934	**0.907**	**0.741**
LB 68.5 hr	FFT + Dense [25] mod	Audio + ECG	0.444	0.729	0.168	0.273	0.0523
	CNN [23] mod	ECG	0.565	0.494	0.434	0.634	0.0991
	Random Forest [26]	ECG + IMU	0.717	0.729	0.799	0.762	0.410
	CNN + LSTM [29]	ECG	0.843	0.825	0.778	0.748	0.639
Adult 7.1 hr	Signal Processing [24]	Audio	0.833	-	0.922	-	0.508
Adult 37k hr	LSTM [21]	Actigraphy	0.831	0.816	0.914	0.855	-
Adult 292 hr	FFT + Dense [25]	Respiratory + ECG	0.853	-	-	-	-
Adult 110 hr	CNN [23]	ECG	0.815	0.478	0.930	0.360	0.260
Adult 244 hr	Random Forest [26]	ECG + IMU	0.873	-	0.895	-	0.396

experiment with data interpolation to fill out the zeros during time synchronization nor with data augmentation to diversify the data.

For audio and ECG W2V2 branch, we used almost the same baseline structure as described in [30], with 12 transformer layers, hidden size 768, intermediete size 3072 and 16 attention heads. Convolution feature extraction structure remain the same with hidden size 512, kernel size [10,3,3,3,3,2,2] and strides [5,2,2,2,2,2,2]. For IMU branch, the feature extraction is done by a linear layer, a normalization layer, and adding a position embedding. The decoding part has 12 transformer layers, hidden size 72, intermediate size 144 and 4 attention heads. Outputs of the three branches are concatenated into a vector of size 1608, passed through a dense structure as follows: a linear layer from size 1608 to size 1608, ReLU, linear layer from size 1608 to 804, dropout of rate 0.1, linear layer from 804 to 2.

We found empirically that training becomes unstable when the attention layers are loaded with pretrained weights, so during finetuning, our code only loads the pretrained weights from the feature extractor up to positional encoder. The attention layers are initialized by default with HuggingFace.

Each experiment was run for 2 epochs with batch size 16, with the standard Adam optimizer. This was done on a single RTX A6000 GPU. Training code was based on Huggingface Transformers and the code is available for inspection and further development. [1]

E. Baseline Implementation

The majority of sleep/wake classification baselines did not release code, so we implemented some of their algorithms with modification based on their description in the paper. Specifically, for [25] we replaced the respiratory input with LB audio data, and trained with gradient descent instead of Levenberg–Marquardt back-propagation algorithm. For [23], we used increasing strides ([1, 2, 3, 4, 5, 6]) for the CNN layers instead of using [2, 2, 2, 2, 2, 2] because each LB ECG segments has more samples than theirs. We also modified the dropout value to 0.1 based on empirical testing. [26] used 3-axis acceleration and heart rate data collected on Apple Watch as their dataset. We calculated heart rate at each recorded segment from LB ECG data, using the python package HeartPy. We then input our timestamp, acceleration data, and calculated heart rate to their released code. Finally, state of the art sleep stage classifier [29] released their code online. We used ECG as input to the model instead of EEG in the original paper.

[1] https://github.com/kaichieh121/LittleBeatsPrelim

IV. RESULTS

A. Benefits of Three Modalities

The first section in Table I shows how using all modalities together compares to using each of the modalities alone for the purpose of sleep/wake classification. The single modality architecture takes each branch in Fig. 2, with pretrained weights, and adds the final three FC layers for classification. The proposed architecture is the entire model pretrained on unlabeled data and finetuned with proposed cross-attention at every 4 layers. We see that, by quite a large margin, using all three modalities together as proposed, the network is able to learn more about infants' sleep patterns and achieve better performance across the board than using just a single modality. Note that while recall for the ECG branch outperforms the proposed network by 5%, its precision is lower by 20%, suggesting that the unimodal ECG network is a worse classifier since it too frequently classifies segments as "sleep."

The second section in Table I presents results of key baseline implementation on the LB dataset, as described in Section III-E. Because [25] and [23] have not released their code and [26] and [29] are not designed to work strictly on ECG data, results need to be interpreted with caution. However, we can still observe that our proposed model using all three modalities

TABLE II
RESULTS WHEN PRETRAINING DIFFERENT BRANCHES OF THE ARCHITECTURE. FIRST COLUMN INDICATES THE BRANCH LOADED WITH PRETRAINED WEIGHTS.

Pretrained	↑ Accuracy	↑ Precision	↑ Recall	↑ F1	↑ Kappa
-	0.797	0.829	0.850	0.839	0.565
IMU	0.837	0.874	0.863	0.868	0.655
Audio+IMU	0.871	0.858	**0.949**	0.901	0.715
ECG+IMU	0.875	0.866	0.944	0.904	0.726
All	**0.880**	**0.881**	0.934	**0.907**	**0.741**

TABLE III
RESULTS OF EARLY, LATE, AND THE PROPOSED FUSION TECHNIQUE

Fusion	↑ Accuracy	↑ Precision	↑ Recall	↑ F1	↑ Kappa
Early	0.779	0.857	0.773	0.813	0.544
Late	0.798	0.781	**0.947**	0.856	0.530
Proposed	**0.880**	**0.881**	0.934	**0.907**	**0.741**

outperforms all key baselines, which only use one or two modalities.

Numbers in the last section in Table I are results reported in other papers, using other datasets, and are therefore not strictly comparable to the numbers in the first two sections. The classifiers summarized by these results are sleep/wake classifiers using audio only [24], ECG only [23], ECG and respiratory signals [25], motion [21], and ECG and IMU [26]. While all datasets are different and comparison of the accuracies is therefore not theoretically justifiable, such a comparison nevertheless supports the conclusion displayed in the top half of the table, viz., that sleep/wake classification performed using three modalities is more accurate than sleep/wake classification performed using only one or two modalities.

B. Benefits of Pretraining

Table II shows how pretraining affects the performance of the proposed architecture. Here we configured the model to use the proposed architecture with three branches and cross attention, but only load pretrained weights onto specific branches. Cross-attention is applied at every 4^{th} layer for the 12 layers of transformers. Every time we load pretrained weights for one more branch, we have a performance increase across all metrics. For example, the accuracy is 0.797 when we don't pretrain any of the branches. This increases to 0.837 when we use pretrain weights for the IMU branch. Accuracy further improves when we load pretrain weights of audio and ECG branch. This reinforces our understanding of pretrained networks. By pretraining each branch using the unlabeled data, then fine-tuning the entire system together against a small number of labeled data, we are able to get better performance.

C. Fusion Techniques

Our discussion is incomplete without comparison to early and late fusion techniques discussed in the literature [35]. Table III presents performance of 2 fusion methods against the proposed cross-attention-based intermediate fusion technique.

TABLE IV
RESULT WITH CROSS-ATTENTION APPLIED AT VARIOUS LAYERS. THE EQUATION IN THE FIRST COLUMN SPECIFIES THE LAYERS IN WHICH CROSS-MODAL ATTENTION IS APPLIED, E.G., $i\%4 = 1$ MEANS THAT CROSS-MODAL ATTENTION WAS PERFORMED IN LAYERS 1, 5, AND 9, WHERE LAYERS ARE NUMBERED $0 \leq i \leq 11$.

Cross	↑ Accuracy	↑ Precision	↑ Recall	↑ $F1$	↑ Kappa
None	0.765	0.788	0.851	0.818	0.487
i%2=1	-	-	-	-	-
i%4=1	**0.880**	**0.881**	**0.934**	**0.907**	**0.741**
i%6=1	0.836	0.870	0.865	0.868	0.652
$0 \leq i < 4$	0.755	0.811	0.791	0.801	0.484
$4 \leq i < 8$	-	-	-	-	-
$8 \leq i < 12$	-	-	-	-	-

Early fusion is inspired by [34], where convolutional features extracted from ECG and IMU data are concatenated with anthropometric data as input to a downstream neural network. In our variation, we concatenate the three outputs from the feature extractor for each branch as shown in Fig. 2, skipping the transformer layers, and pass through the dense network for fine-tuning. As for late fusion, we leave in the transformer layers, triplicate the FC layers to generate logits for each modality, and average the three logits for evaluation. The proposed cross-attention fusion achieved better performance in almost every metric except that it has a lower recall than late fusion. Similar to the trend found in Table I, this high recall for late fusion results from a disproportional confusion matrix, as shown by late fusion's low precision score. Therefore, the proposed fusion architecture is still favorable in general.

D. Ablation Study

Table IV shows the effect of using cross-attention at different layers. All models were fine-tuned on labeled multimodal data after loading pretrained weights for all three modalities separately. "-" means the architecture did not converge. We empirically found that cross-modal attention once every 4 layers ($i\%4 = 1$) gives the best performance across the board. Many models with cross-modal attention fail to converge: convergence failure occurs with cross-modal attention every second layer, and with cross-modal attention in the middle four layers ($4 \leq i < 8$) or the last four layers ($8 \leq i < 12$). Cross-modal attention in the first four layers ($0 \leq i < 4$) converged to a system with accuracy lower than that of the system with no cross-modal attention. The sensitivity of cross-modal attention to architectural configuration was an unexpected result. We defer this to future studies.

V. CONCLUSION

With the development of multi-modal wearable devices, we are able to gather synchronized audio, ECG, and IMU data for the task of infant sleep/wake classification. We demonstrated the best classification performance when using all three modalities compared with our own single modality network or single/double modality network found in the literature. In addition, we developed an ensemble of large scale pretrained

transformer neural network, by fusing the transformer layers with cross-attention at every 4 layers. This fusion method is not limited to the task of wake/sleep classification, but seems likely to generalize successfully to any multi-modal network in which all modalities have the same number of transformer layers. Our work presents exciting directions for multi-modal studies of infant and child development.

ACKNOWLEDGMENT

We would like to thank the families who participated in this research, as well as Jordan Bodway and Jenny Baldwin for their assistance with data collection and processing. This work was supported by funding from the National Institute of Mental Health (R21MH112578), National Institute of Drug Abuse (R34DA050256), the National Institute of Food and Agriculture (ILLU-793-368), and the Personalized Nutrition Initiative and Center for Social and Behavioral Science at the University of Illinois at Urbana-Champaign through the Seed Grant program.

REFERENCES

[1] A. R. Tarullo, P. D. Balsam, and W. P. Fifer, "Sleep and infant learning," *Infant Child Dev.*, vol. 20, no. 1, pp. 35–46, Jan. 2011.

[2] M. S. Blumberg, A. J. Gall, and W. D. Todd, "The development of sleep-wake rhythms and the search for elemental circuits in the infant brain," *Behav. Neurosci.*, vol. 128, no. 3, pp. 250–263, Jun. 2014.

[3] R. E. Dahl, "The regulation of sleep and arousal: Development and psychopathology," *Dev Psychopathol*, vol. 8, no. 3, pp. 3–27, 1996.

[4] E. Bathory and S. Tomopoulos, "Sleep regulation, physiology and development, sleep duration and patterns, and sleep hygiene in infants, toddlers, and preschool-age children," *Curr. Probl. Pediatr. Adolesc. Health Care*, vol. 47, no. 2, pp. 29–42, Feb. 2017.

[5] B. C. Galland, B. J. Taylor, D. E. Elder, and P. Herbison, "Normal sleep patterns in infants and children: a systematic review of observational studies," *Sleep medicine reviews*, vol. 16, no. 3, p. 213—222, June 2012.

[6] A. Hupbach, R. L. Gomez, R. R. Bootzin, and L. Nadel, "Nap-dependent learning in infants," *Dev. Sci.*, vol. 12, no. 6, pp. 1007–1012, Nov. 2009.

[7] K. Horváth, K. Myers, R. Foster, and K. Plunkett, "Napping facilitates word learning in early lexical development," *J. Sleep Res.*, vol. 24, no. 5, pp. 503–509, Oct. 2015.

[8] E. Dearing, K. McCartney, N. L. Marshall, and R. M. Warner, "Parental reports of children's sleep and wakefulness: longitudinal associations with cognitive and language outcomes," *Infant Behav. Dev.*, vol. 24, no. 2, pp. 151–170, Feb. 2001.

[9] A. Bernier, S. M. Carlson, S. Bordeleau, and J. Carrier, "Relations between physiological and cognitive regulatory systems: infant sleep regulation and subsequent executive functioning," *Child Dev.*, vol. 81, no. 6, pp. 1739–1752, Nov. 2010.

[10] M. Ednick, A. P. Cohen, G. L. McPhail, D. Beebe, N. Simakajornboon, and R. S. Amin, "A review of the effects of sleep during the first year of life on cognitive, psychomotor, and temperament development," *Sleep*, vol. 32, no. 11, pp. 1449–1458, Nov. 2009.

[11] A. Scher, "Infant sleep at 10 months of age as a window to cognitive development," *Early Hum. Dev.*, vol. 81, no. 3, pp. 289–292, Mar. 2005.

[12] A. Sadeh, G. De Marcas, Y. Guri, A. Berger, L. Tikotzky, and Y. Bar-Haim, "Infant sleep predicts attention regulation and behavior problems at 3-4 years of age," *Dev. Neuropsychol.*, vol. 40, no. 3, pp. 122–137, 2015.

[13] M. Thunström, "Severe sleep problems in infancy associated with subsequent development of attention-deficit/hyperactivity disorder at 5.5 years of age," *Acta Paediatr.*, vol. 91, no. 5, pp. 584–592, 2002.

[14] F. V. O'Callaghan, A. Al Mamun, M. O'Callaghan, A. Clavarino, G. M. Williams, W. Bor, H. Heussler, and J. M. Najman, "The link between sleep problems in infancy and early childhood and attention problems at 5 and 14 years: Evidence from a birth cohort study," *Early Hum. Dev.*, vol. 86, no. 7, pp. 419–424, Jul. 2010.

[15] J.-P. Chaput, C. E. Gray, V. J. Poitras, V. Carson, R. Gruber, C. S. Birken, J. E. MacLean, S. Aubert, M. Sampson, and M. S. Tremblay, "Systematic review of the relationships between sleep duration and health indicators in the early years (0-4 years)," *BMC Public Health*, vol. 17, no. Suppl 5, p. 855, Nov. 2017.

[16] L. Matricciani, C. Paquet, B. Galland, M. Short, and T. Olds, "Children's sleep and health: A meta-review," *Sleep Med. Rev.*, vol. 46, pp. 136–150, Aug. 2019.

[17] "About littlebeats." [Online]. Available: https://littlebeats.hdfs.illinois.edu/about-littlebeats/

[18] A. Sadeh, "Iii. sleep assessment methods," *Monogr. Soc. Res. Child Dev.*, vol. 80, no. 1, pp. 33–48, Mar. 2015.

[19] S. E. Beck and C. L. Marcus, "Pediatric polysomnography," *Sleep Med. Clin.*, vol. 4, no. 3, pp. 393–406, Sep. 2009.

[20] A. Sadeh, "Iii. sleep assessment methods," *Monogr. Soc. Res. Child Dev.*, vol. 80, no. 1, pp. 33–48, Mar. 2015.

[21] J. Palotti, R. Mall, M. Aupetit, M. Rueschman, M. Singh, A. Sathyanarayana, S. Taheri, and L. Fernandez-Luque, "Benchmark on a large cohort for sleep-wake classification with machine learning techniques," Jun 2019.

[22] C. Cajochen, J. Pischke, D. Aeschbach, and A. A. Borbély, "Heart rate dynamics during human sleep," *Physiology & Behavior*, vol. 55, no. 4, pp. 769–774, 1994.

[23] J. Malik, Y.-L. Lo, and H. tieng Wu, "Sleep-wake classification via quantifying heart rate variability by convolutional neural network," *Physiological Measurement*, vol. 39, no. 8, p. 085004, aug 2018.

[24] E. Dafna, A. Tarasiuk, and Y. Zigel, "Sleep-wake evaluation from whole-night non-contact audio recordings of breathing sounds," p. e0117382, Feb 2015.

[25] W. Karlen, C. Mattiussi, and D. Floreano, "Sleep and wake classification with ecg and respiratory effort signals," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 2, pp. 71–78, 2009.

[26] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, vol. 42, no. 12, 08 2019, zsz180.

[27] S. Cabon, F. Porée, A. Simon, B. Met-Montot, P. Pladys, O. Rosec, N. Nardi, and G. Carrault, "Audio- and video-based estimation of the sleep stages of newborns in neonatal intensive care unit," *Biomedical Signal Processing and Control*, vol. 52, pp. 362–370, 2019.

[28] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, nov 2017.

[29] A. Supratak and Y. Guo, "Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 641–644, 2020.

[30] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[31] J. Oh, H. Chung, J. myoung Kwon, D. gyun Hong, and E. Choi, "Lead-agnostic self-supervised learning for local and global representations of electrocardiogram," 2022.

[32] H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen, "Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 220–233.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[34] Z. Ni, T. Wu, T. Wang, F. Sun, and Y. Li, "Deep multi-branch two-stage regression network for accurate energy expenditure estimation with ecg and imu data," p. 3224–3233, Oct 2022.

[35] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6.

[36] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," 2021.

[37] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-level multi-modal cross-attention network for fake news detection," *IEEE Access*, vol. 9, pp. 132 363–132 373, 2021.

[38] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 938–10 947.
[39] X. Song, H. Chao, X. Xu, H. Guo, S. Xu, B. Turkbey, B. J. Wood, T. Sanford, G. Wang, and P. Yan, "Cross-modal attention for multi-modal image registration," *Medical Image Analysis*, vol. 82, p. 102612, 2022.
[40] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Towards robust family-infant audio analysis based on unsupervised pretraining of wav2vec 2.0 on large-scale unlabeled family audio," 2023.
[41] D. Xu, J. A. Richards, and J. Gilkerson, "Automated analysis of child phonetic production using naturalistic recordings," *J. Speech Lang. Hear. Res.*, vol. 57, no. 5, pp. 1638–1650, Oct. 2014.
[42] H. Dudley, "The vocoder—electrical re-creation of speech," *Journal of the Society of Motion Picture Engineers*, vol. 34, no. 3, pp. 272–278, 1940.
[43] F.-J. Chang, M. Radfar, A. Mouchtaris, B. King, and S. Kunzmann, "End-to-end multi-channel transformer for speech recognition," 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Dynamic Characteristics of Electroencephalogram Reflecting Driving-Experience-Dependent Performance Using Microstates

Yuta Iinuma*, Takuto Ozawa†, Sou Nobukawa*, Nobuhiko Wagatsuma‡ and Keiichiro Inagaki§
* Graduate School of Information and Computer Science, Chiba Institute of Technology, Japan
E-mail: s1831006fv@s.chibakoudai.jp (Y.I.), nobukawa@cs.it-chiba.ac.jp (S.N.)
† Department of Computer Science, Chiba Institute of Technology, Narashino, Japan
‡ Department of Information Science, Toho University, Japan
§ Department of Artificial Intelligence and Robotics, Chubu University, Japan
E-mail:kay@isc.chubu.ac.jp

***Abstract*—To further reduce the number of traffic accidents, it has been necessary in recent years to develop advanced driver assistance systems with functions that provide optimal assistance based on driving performance. Recent studies have used electroencephalogram (EEG) to comprehensively evaluate the complex neural process of visual, perceptual, attentional, and motor functions. Microstate analysis, which reflects the interactions of global neural networks by measuring the temporal transition of brain activity, is widely used to capture complex cognitive processes. In this context, we hypothesized that microstate analysis is a suitable method for capturing driving-experience-dependent performance using EEG. To prove our hypothesis, we evaluated the EEG data of novice and expert drivers while they viewed driving scenes using microstate analysis to detect driving-experience-dependent performance using EEG. The results showed that the dynamic characteristic of EEG concerning driving-experience-dependent performance reflected efficient cognitive control and suppressed redundant bottom-up attention. This finding may be widely utilized for estimating the driving-experience-dependent performance of advanced driver assistant systems.**

I. Introduction

Recent advancements in supporting technologies for intelligent transportation systems, such as forward collision warnings, advanced emergency braking systems, and driver monitoring, have aided in reducing traffic accidents [1]–[6]. Developing driver assistance systems with functions optimized for driving performance-dependent assistance is needed to further reduce the number of traffic accidents [7], [8].

Pupillometry and visual attention are the most prominent approaches for evaluating driving performance using eye-tracking sensors. According to studies using pupillometry, pupil diameter variations can detect driver fatigue [9], [10]. In addition to pupillometry studies, visual attention studies have shown that driving performance can be determined by the duration and number of fixations on objects such as main road areas, cars, and pedestrians [11], [12]. However, these methods have revealed that accurately estimating driving performance is problematic owing to factors such as vehicle vibration, backlights, and traffic signals [13]–[16]. Moreover, recent studies have shown that driving performance is not solely determined by visual information processing, but also by higher cognitive functions, such as decision-making and memory [17]–[19]. Previous studies have shown a strong correlation between driving performance and the neural activity detected by electroencephalography (EEG), magnetoencephalography, and functional magnetic resonance imaging (fMRI) of the brain [20]–[24]. Due to its portability, non-invasiveness, and high temporal resolution, EEG is widely used to capture brain activity during driving [25]–[27]. Therefore, EEG, which can detect neural activity directly from the brain, may provide a comprehensive assessment of complex neural processes involving vision, perception, attention, and motor functions.

Functional connectivity, which reflects the integration of information processing between brain regions as a pair-to-pair interaction of neural activities, is a method for evaluating driving performance using EEG. [26], [28], [29]. Specifically, methods focusing on functional connectivity have revealed the network characteristics of neural activity related to driving performance that are affected by fatigue, drowsiness, and driving skill level [26], [28], [29]. Using the phase lag index (PLI), which achieves high spatial resolution by avoiding volume conduction, a previous study showed that the driving experience enhances neural pathways in a gamma-band functional network supporting visual perceptual functions [26]. However, it has been pointed out that this assessment method, based on long-term temporal averages in the PLI, is inadequate for capturing dynamic network activity during driving, which is strongly associated with cognitive processes. [26], [30].

Microstate analysis, which focuses on the quasi-stable spatial patterns of EEG, effectively captures dynamic brain activity during driving [31], [32]. This method, which simultaneously considers signals from all electrodes, can capture the interactions of global neural networks by measuring temporal transitions [32]–[34]. Specifically, these microstate transitions have been shown to reflect lethargy- and fatigue-related neural network characteristics that affect driving performance [30], [35]. In this context, we hypothesized that microstate analysis,

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

which allows for the comprehensive detection of the interactions of global neural networks, is a suitable method for detecting moment-to-moment dynamics of brain activity related to driving performance, depending on driving experience as related to skill level using EEG. To prove this hypothesis, we evaluated the EEG data of novice and expert drivers while they viewed driving scenes using microstate analysis to detect driving experience-dependent EEG performance.

II. Methods

A. Participants

The EEG data were measured at the Department of Robotics Science and Technology, Chubu University, Japan. Among the 15 participants, all were healthy individuals aged 20 to 23 (comprising 1 female and 14 males). They either had natural clear vision or had it corrected to the standard using glasses or contact lenses, and each held a valid driving license. Based on their driving performance, participants were classified as novice or expert drivers. Driving frequency was used to evaluate driving performance. Participants who drove less frequently than once every two weeks were classified as the novice group, while those who drove more frequently than once every two weeks were classified as the expert group. The groups of novices and experts consisted of eight and seven members, respectively. In this study, participants were guided to view the traffic situations as though they were on their regular commute, to gauge an authentic understanding of the driving surroundings. All procedures were conducted in accordance with the principles of the Declaration of Helsinki. The study protocol was approved by tthe Ethics Committee of Chubu University. All the study participants provided informed consent. The EEG data used in this study were evaluated using the PLI and power analyses [26], [27].

B. EEG Recording and Data Preprocessing

EEG data were measured using a 21-channel system (EPOC+ produced by Emotive) at 14 electrodes (AF3, AF4, F3, F4, F7, F8, FC5, FC6, P8, P7, T7, T8, O1, and O2) in accordance with the international 10-20 system, with the two ear lobes jointly forming the reference [36]. The participants were measured for 2–3 min while sitting 60 cm away from a monitor (1920 $\times$ 1080 pixels) and watching a driving scene filmed with a video camera. The video was filmed on a highway. The scenes involve other vehicles in the same or side lanes (passing lane), signals (red, yellow, and green), road signs, pedestrians, intersections, crosswalks, highway ramps, and cars stopping at the red signal. The video only captures the forward perspective from the car's camera, excluding views from the side and rear mirrors. In a previous study, we revealed that a difference in the EEG power spectrum and PLI between expert and novice drivers can be detected through the length of the video stimulus [26], [27]. Therefore, we consider that 2 to 3 min is an appropriate movie length to evaluate the microstate considering the dynamic characteristics of EEG that are attributable to driving performance. The sampling frequency was 128 Hz. Throughout the experiment, the saline solution sustained the electrode impedance within the necessary $10\,\mathrm{k}\Omega$ to $20\,\mathrm{k}\Omega$ range, as required by the recording software. EEG data underwent artifact subspace reconstruction (ASR) to eliminate disturbances from headset movement, eye blinking, eye motion, and activity from neck and scalp muscles. Open-source MATLAB scripts to execute ASR can be found as plug-in functions for clean_rawdata in EEGLAB [37]. In this study, we used the cutoff parameter, defined as the rejection thresholds with standard deviations of $SD20\%$ to identify artifacts.

C. Microstate Analysis based on Amplitude Distribution

In this study, we employed the dynamic attributes of instantaneous amplitudes within EEG signals to categorize different brain activity states. The process for estimating instantaneous amplitude dynamics is indicated in Fig.1 (a). First, the EEG data was applied by bandpass filtering within the amplitude spectrum of 2-20 Hz [38], [39]. Next, we applied the Hilbert transform to determine the instantaneous amplitude $\theta(t)(-\pi \leq \theta \leq \pi)$ of the EEG signals. And this amplitude was converted from the mean across all electrodes to the deviation $dIA_i(t)$, where i represents the electrode position. To identify a microstate characterized by a specific spatial power distribution, we computed the global field power (GFP) based on the spatial variance among the electrodes. The spatial power distribution, corresponding to the GFP's local peak, was extracted from the $dIA_i(t)$ time series [40], [41]. As shown in Fig.1 (b), the $dIA_i(t)$ set at the GFP's local peak for both novice and expert drivers was grouped into k clusters via the k-medoid technique. Because of the small number of electrodes, this study used a cluster size of $k = 2$. By employing this approach, we hypothesized that dynamic state transitions reflected the evolving neural connections across the entire brain network [42], [43]. Furthermore, we valuated the probability of state transitions between the applied microstates using the k-medoids method.

D. Statistical Analysis

We performed a paired t-test to assess the differences in state transition probabilities between novice and expert drivers. The t-scores applied correction using the Benjamini-Hochberg false discovery rate method for multiple comparisons (involving four p-values). We defined the significance threshold at q <0.05.

III. Results

We estimated the characteristics of state transitions from the distribution of instantaneous amplitudes in EEG recordings taken from novice and expert drivers as they observed a highway driving scene. Figure 2 shows the average time for the $dIA_i(t)$ duration of each state classified by k-medoids for novice and expert drivers. For both groups, the frontal-specific activated state was identified in state #1 and the occipital-specific activated state in state #2. These microstates reflect the bottom-up and top-down neural activities involving visual attention, respectively [44], [45]. In addition, to evaluate the state transitions among the activated states in Fig.2 from

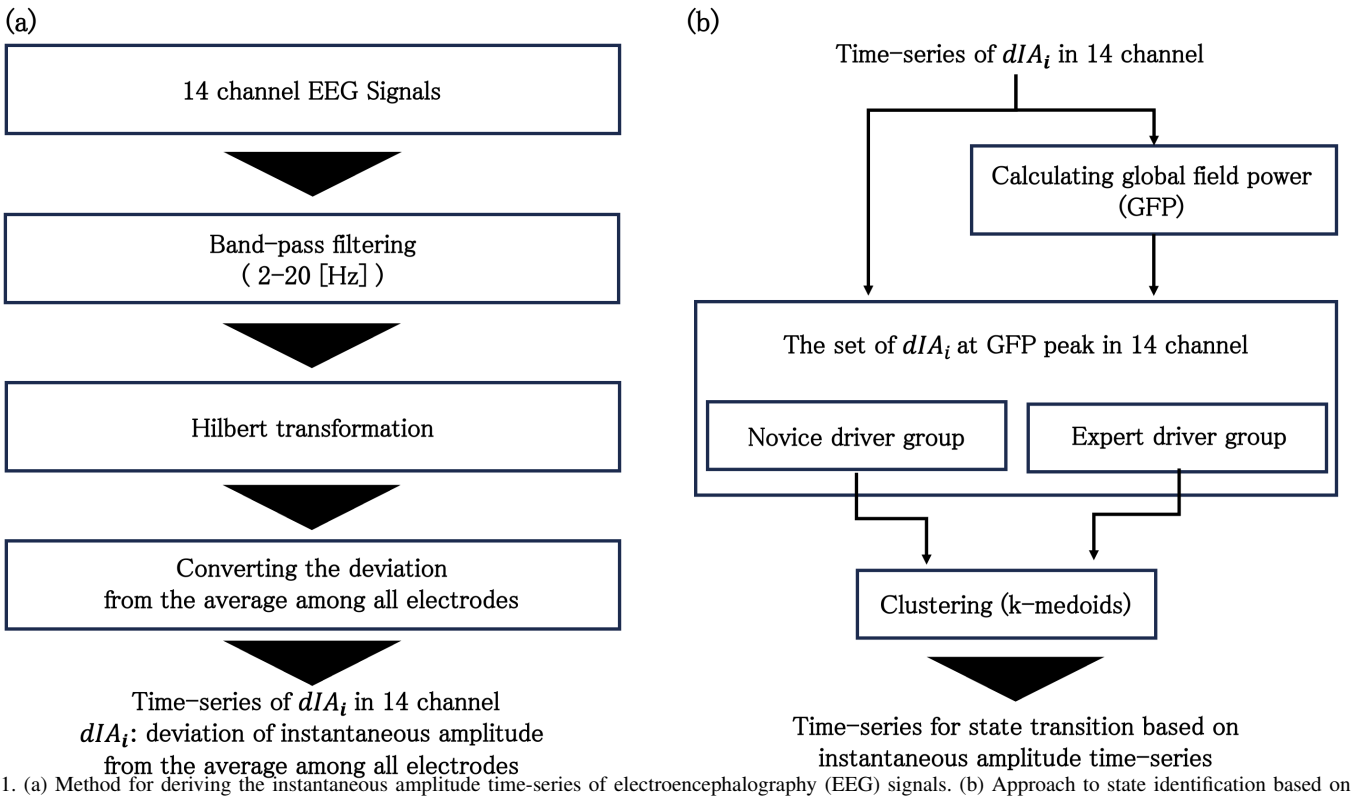


Fig. 1. (a) Method for deriving the instantaneous amplitude time-series of electroencephalography (EEG) signals. (b) Approach to state identification based on the instantaneous amplitude time-series of EEG signals, as revealed in our preceding publication

source to destination, Figure 3 (a) and (b) show the average of these state transition probabilities for novice and expert drivers. Figure 3 (c) shows the t-values of the state transition probabilities between novice and expert drivers. For the expert driver, the state transition from states #2 to #1 ($t = -3.569$ (p = 0.006 ($q < 0.05$))) was significantly lower, whereas the state transition from states #1 to #1 ($t = 3.569$ (p = 0.006 ($q < 0.05$))) was significantly higher.

IV. Discussion

In this study, we evaluated the EEG data of novice and expert drivers while watching driving scenes, using microstate analysis to detect driving performance, including driving experience. Resultantly, for both groups, the frontal-specific and occipital-specific-activated states were identified in states #1 and #2, respectively. For the expert driver, the state transition probability from #2 to #1 is significantly lower, whereas the state transition probability from #1 to #1 is significantly higher.

First, we considered why expert drivers are more active in the frontal cortex than novice drivers. Recent studies have shown that neural activity in the frontal cortex is strongly related to cognitive control during driving, including visual attention movements, decision-making, and action planning for driving behavior [46], [47]. In particular, a previous study using fMRI showed that expert drivers had more significant activation in the frontal lobe during driving than novice drivers [48]. These activated neural activities are considered when expert drivers perform cognitive control, such as visual attention movements and decision making, more efficiently than novice drivers based on their driving experience [49]–[51]. Therefore, the results of this study (Figs.3(c)) may reflect the neural activity of efficient cognitive control in expert drivers.

Second, we considered why expert drivers suppressed occipital-to-frontal transitions more than novice drivers. In previous studies, expert drivers were considered to suppress bottom-up attention because of the control of attentional functions by the frontal lobe [48], [52]. Panos *et al.* revealed that expert drivers perform efficient visual searches to anticipate danger, whereas novice drivers perform inefficient visual searches such as missing important objects or fixating on unimportant objects [50]. These differences in visual characteristics suggest that expert drivers can prioritize the visual area where hazards are most likely to occur based on past driving experience [53], [54]. Therefore, the results of this study (Fig.3(c)) may reflect the neural activity of suppressed redundant bottom-up attention of expert drivers.

Third, this study had some limitations. Although this study revealed the characteristics of microstates related to driving experience, the number of participants was small. Therefore, further studies with larger sample sizes are required. Additionally, previous studies using EEG power analysis evaluated the driving performance while viewing three different types of driving scenes (traffic jams, urban areas, and highways) [27]. However, in this study, the evaluations were conducted only while viewing highway driving scenes. It is important to address these element in the future to provide an in-depth

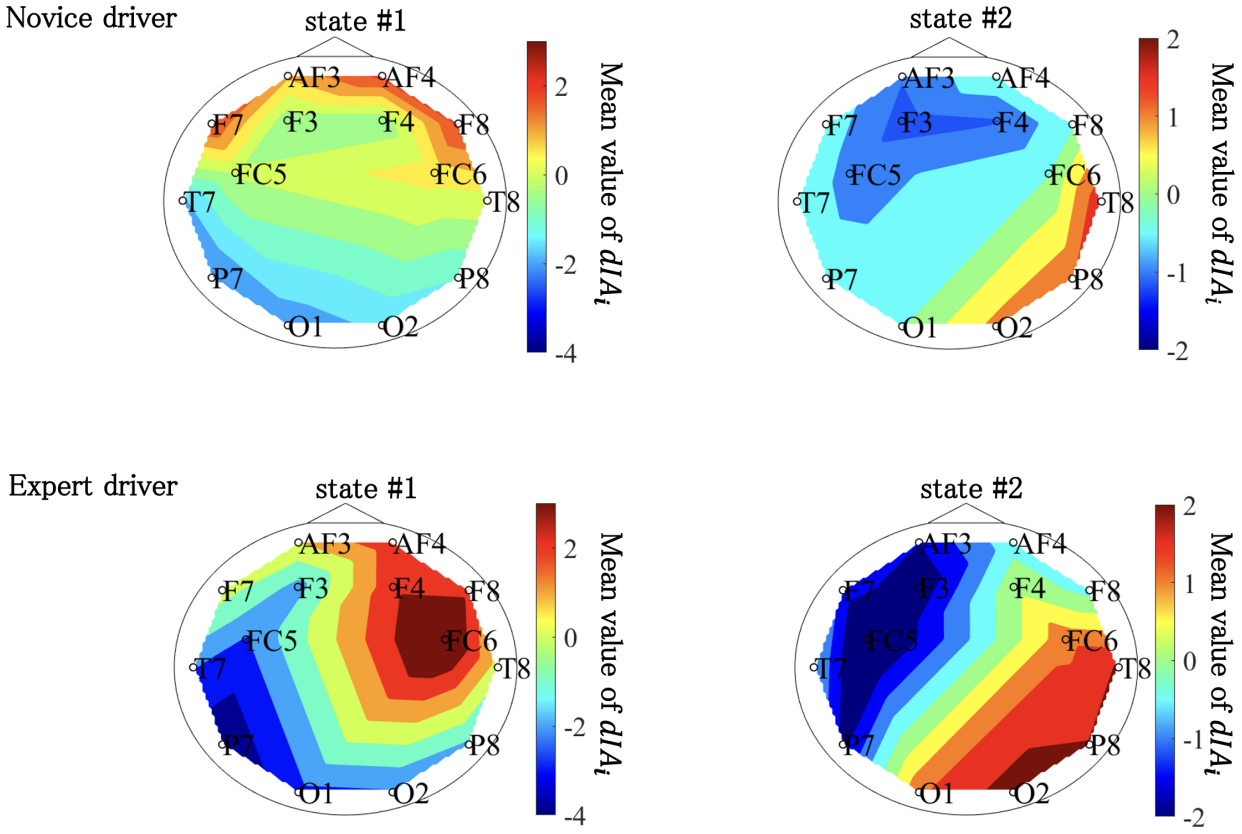


Fig. 2. Time average of $dIA_i(t)$ among all electrode durations of each state classified by k-medoids for novice and expert drivers while watching a driving scene of a highway

understanding of driving-experience-dependent performance with high accuracy in divergent situations. Additionally, this study was based on the hypothesis that driving performance, depending on driving experience related to skill level, may be reflected by the interactions of the global neural network. The EEG of the 14-electrode system used in this study is sufficient to reflect global neural network interactions, but insufficient to focus on local neural network interactions [42], [55]. Therefore, it is necessary to increase the number of electrodes and focus on local dynamics to more accurately reflect the driving experience-dependent performance from brain activity. In addition, because of the small number of electrodes used in this study, we defined two activation states to evaluate the state transitions. However, this is fewer than the number of state definitions in a general microstate analysis [32], [38]. Therefore, in this study, the results of the state definitions were not stable between expert and novice users. In the future, we intend to increase the number of electrodes and evaluate the state transitions by defining four activation states as a representative classification for microstate analysis.

In conclusion, this study identified the dynamic characteristics of EEG concerning driving experience-dependent performance, reflecting efficient cognitive control and suppressing redundant bottom-up attention. Despite the limitations of this study, this finding may be utilized to estimate driving experience-dependent performance in advanced driver assistance systems.

Acknowledgment

Part of this work was supported by JSPS KAKENHI (grant numbers JP17K12781, JP22K12138 (K.I.), and JP22K12183 (S.N.)).

References

[1] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems." *IEEE Vehicular Technology Magazine*, 5(1):77–84, 2010.

[2] S. Kaplan, M. A. Guvensan, A. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey." *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3017–3032, 2015.

[3] J. Wang, C. Yu, S. E. Li and L. Wang, "A forward collision warning algorithm with adaptation to driver behaviors." *IEEE Transactions on Intelligent Transportation Systems*, 17(4):1157–1167, 2015.

[4] R. Zhang, K. Li, Z. He, H. Wang, and F. You, "Advanced emergency braking control based on a nonlinear model predictive algorithm for intelligent vehicles." *Applied sciences*, 7(5):504, 2017.

[5] V. Milanés, S. E. Shladover, J. Spring, C. Nowakowski, H. Kawazoe, and M. Nakamura, "Cooperative adaptive cruise control in real traffic situations." *IEEE Transactions on intelligent transportation systems*, 15(1):296–305, 2013.

[6] L. Li, D. Wen, N. N. Zheng, and L. C. Shen, "Cognitive cars: A new frontier for ADAS research." *IEEE Transactions on Intelligent Transportation Systems*, 13(1):395–407, 2011.

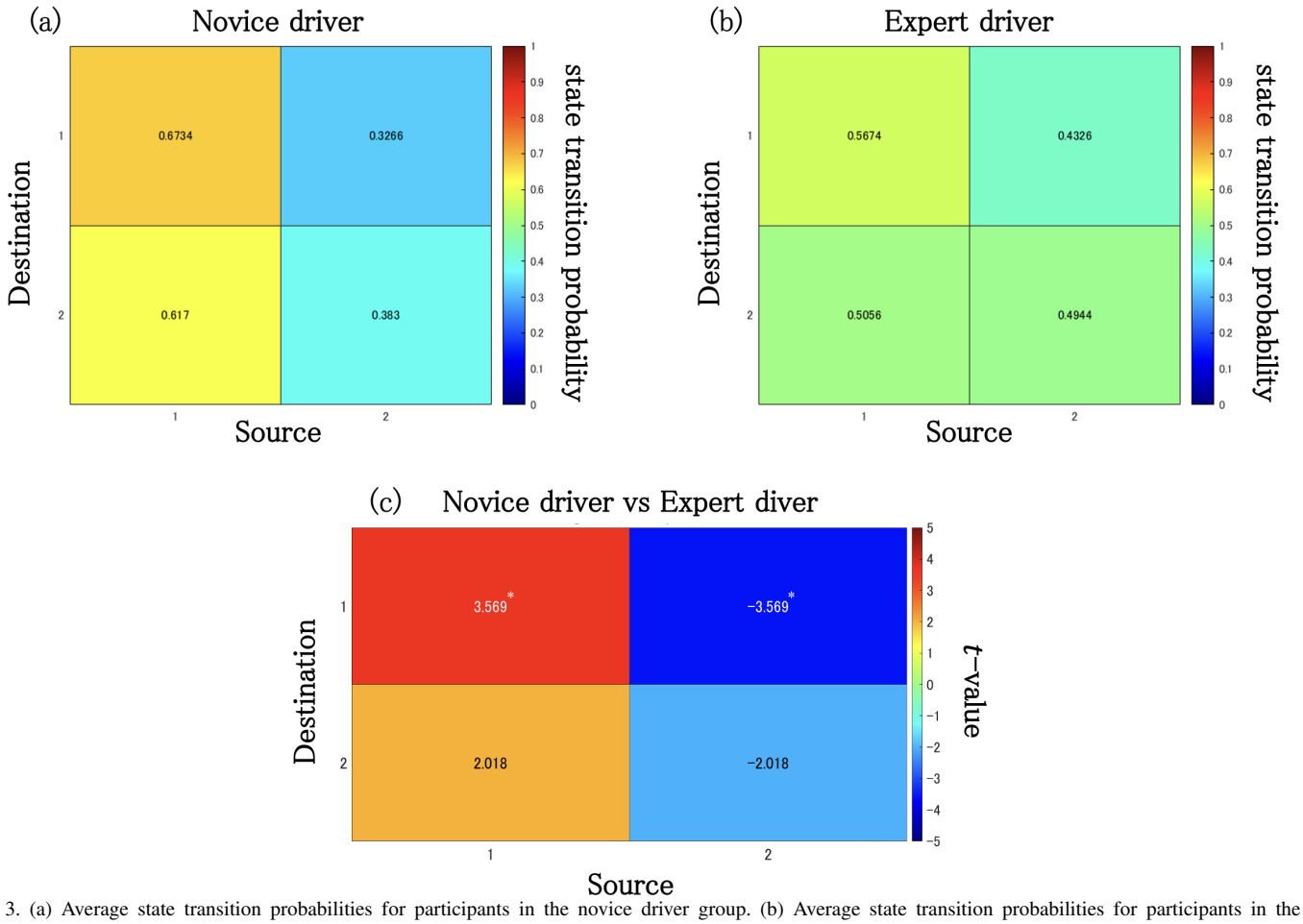


Fig. 3. (a) Average state transition probabilities for participants in the novice driver group. (b) Average state transition probabilities for participants in the expert driver group. (c) t-values comparing state transition probabilities between novice and expert driver groups. Warm (cold) colors indicate greater (lesser) probabilities for novice drivers. Regions meeting the significance threshold (q <0.05) are marked with white asterisks.

[7] W. Wang, J. Xi, H. Chen, "Modeling and recognizing driver behavior based on driving data: A survey." *Mathematical Problems in Engineering*, 2014:1–20, 02 2014.

[8] Y. Xu, Z. Ye, and C. Wang, "Modeling commercial vehicle drivers' acceptance of advanced driving assistance system (ADAS)." *Journal of Intelligent and Connected Vehicles*, 4(3):125–135, 2021.

[9] H. Yokoyama, K. Eihata, J. Muramatsu, and Y. Fujiwara, "Prediction of driver's workload from slow fluctuations of pupil diameter." In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 1775–1780, 2018.

[10] L. Shi, L. Zheng, D. Jin, Z. Lin, Q. Zhang, and Zhang, "Assessment of combination of automated pupillometry and heart rate variability to detect driving fatigue." *Frontiers in Public Health*, 10, 2022.

[11] T. Wijayanto, S. R. Marcilia, G. Lufityanto, "Visual attention, driving behavior and driving performance among young drivers in sleep-deprived condition." *KnE Life Sciences*, pages 424–434, 2018.

[12] J. Xu, J. Min, and J. Hu, "Real-time eye tracking for the assessment of driver fatigue." *Healthcare technology letters*, 5(2):54–58, 2018.

[13] G. Prabhakar, N. Madhu, and P. Biswas, "Comparing pupil dilation, head movement, and EEG for distraction detection of drivers." In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*, pages 1–5, 2018.

[14] A. S. Le, T. Suzuki, and H. Aoki, "Evaluating driver cognitive distraction by eye tracking: From simulator to driving." *Transportation research interdisciplinary perspectives*, 4:100087, 2020.

[15] A. T. Duchowski, "A breadth-first survey of eye-tracking applications." *Behav Res Methods Instrum Comput*, 34(4):455–470, Nov 2002.

[16] M. Schwalm, A. Keinath, and H. D. Zimmer, "Pupillometry as a method for measuring mental workload within a simulated driving task." *Human Factors for assistance and automation*, 1986:1–13, 2008.

[17] A. Vakulin, A. D ' Rozario, J. W. Kim, B. Watson, C. N. D. Wang, A. Coeytaux, D. Bartlett, K. Wong, and R. Grunstein, "Quantitative sleep EEG and polysomnographic predictors of driving simulator performance in obstructive sleep apnea." *Clinical Neurophysiology*, 127(2):1428–1435, 2016.

[18] J. Perrier, S. Jongen, E. Vuurman, M. L. Bocca, J. G. Ramaekers, A. Vermeeren, "Driving performance and EEG fluctuations during on-the-road driving following sleep deprivation." *Biological psychology*, 121:1–11, 2016.

[19] K. J. Anstey, T. D. Windsor, M. A. Luszcz, and G. R. Andrews, "Predicting driving cessation over 5 years in older adults: Psychological well-being and cognitive competence are stronger predictors than physical health." *Journal of the American Geriatrics Society*, 54(1):121–126, 2006.

[20] M. Awais, N. Badruddin, and M. Drieberg, "Driver drowsiness detection using EEG power spectrum analysis." In *2014 IEEE Region 10 symposium*, pages 244–247. IEEE, 2014.

[21] G. Borghini, G. Vecchiato, J. Toppi, L. Astolfi, A. Maglione, R. Isabella, C. Caltagirone, W. Kong, D. Wei, Z. Zhou, L. Polidori, S. Vitiello, F. Babiloni, "Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices." In *2012 annual international conference of the IEEE engineering in the Medicine and Biology Society* , pages 6442–6445. IEEE, 2012.

[22] S. K. Lal and A. Craig, "A critical review of the psychophysiology of driver fatigue." *Biological psychology*, 55(3):173–194, 2001.

[23] T. Nozawa, Y. Taki, A. Kanno, Y. Akimoto, M. Ihara, R. Yokoyama, Y. Kotozaki, R. Nouchi, A. Sekiguchi, H. Takeuchi, et al, "Effects of

different types of cognitive training on cognitive function, brain structure, and driving safety in senior daily drivers: a pilot study." *Behavioural neurology*, 2015, 2015.

[24] P. Chen, F. Chen, L. Zhang, X. Ma, and X. Pan, "Examining the influence of decorated sidewall in road tunnels using fMRI technology." *Tunnelling and Underground Space Technology*, 99:103362, 2020.

[25] S. Ahn and S. C. Jun, "Multi-modal integration of EEG-fNIRS for brain-computer interfaces – current limitations and future directions." *Frontiers in Human Neuroscience*, 11, 2017.

[26] M. Tsurushima, K. Otsu, S. Nobukawa, N. Wagatsuma, and K. Inagaki, "Assessment of functional connectivity induced by driving experience." *Nonlinear Theory and Its Applications, IEICE*, 14(2):559–568, 2023.

[27] K. Inagaki, T. Maruno, K. Yamamoto, "Evaluation of EEG activation pattern on the experience of visual perception in the driving." *IEICE Transactions on Information and Systems*, E103.D(9):2032–2034, 2020.

[28] J. Chen, H. Wang, Q. Wang and C. Hua, "Exploring the fatigue affecting electroencephalography based functional brain networks during real driving in young males." *Neuropsychologia*, 129:200–211, 2019.

[29] J. Chen, H. Wang, C. Hua, Q. Wang, and C. Liu, "Graph analysis of functional brain network topology using minimum spanning tree in driver drowsiness." *Cognitive neurodynamics*, 12:569–581, 2018.

[30] W. Li, S. Cheng, H. Wang, and Y. Chang, "EEG microstate changes according to mental fatigue induced by aircraft piloting simulation: An exploratory study." *Behavioural Brain Research*, 438:114203, 2023.

[31] W. Hu, Z. Zhang, L. Zhang, G. Huang, L. Li and Z. Liang, "Microstate detection in naturalistic electroencephalography data: a systematic comparison of topographical clustering strategies on an emotional database." *Frontiers in Neuroscience*, 16:49, 2022.

[32] C. M. Michel and T. Koenig, "EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review." *Neuroimage*, 180:577–593, 2018.

[33] P. Milz, P. L. Faber, D. Lehmann, T. Koenig, K. Kochi, and R. D. P. Marqui, "The functional significance of EEG microstates—associations with modalities of thinking." *Neuroimage*, 125:643–656, 2016.

[34] B. A. Seitzman, M. Abell, S. C. Bartley, M. A. Erickson, A. R. Bolbecker, and W. Hetrick, "Cognitive manipulation of brain electric microstates." *Neuroimage*, 146:533–543, 2017.

[35] M. Krylova, S. Alizadeh, I. Izyurov, V. Teckentrup, C. Chang, J. V. Meer, M. Erb, N. Kroemer, T. Koenig, M. Walter, et al, "Evidence for modulation of EEG microstate sequence by vigilance level." *Neuroimage*, 224:117393, 2021.

[36] R. W. Homan, J. Herman and P. Purdy, "Cerebral location of international 10–20 system electrode placement." *Electroencephalography and clinical neurophysiology*, 66(4):376–382, 1987.

[37] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[38] K. Kim, N. T. Duc, M. Choi, and B. Lee, "EEG microstate features for schizophrenia classification." *PloS one*, 16(5):e0251842, 2021.

[39] V. R. Kiran, S. S. Rajagopalan, S. Bhardwaj, R. P. R. V. R. Reddam, C. Ganne, R. Kenchaiah, R. C. Mundlamuri, T. Kandavel, K. K. Majumdar, S. Parthasarathy, S. Sinha, and R. D. Bharath, "Machine learning detects EEG microstate alterations in patients living with temporal lobe epilepsy." *Seizure*, 61:8–13, 2018.

[40] M. Gärtner, V. Brodbeck, H. Laufs, and G. Schneider, "A stochastic model for EEG microstate sequence analysis." *Neuroimage*, 104:199–208, 2015.

[41] A. Khanna, P. Leone, C. M. Michel, F. Farzan, "Microstates in resting-state EEG: current status and future directions." *Neuroscience & Biobehavioral Reviews*, 49:105–113, 2015.

[42] Y. Iinuma, S. Nobukawa, H. Nishimura, and T. Takahashi, "Dynamic characteristics of state transitions composed of neural activity in the brain by circadian rhythms." In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 152–157, 2022.

[43] S. Nobukawa, T. Ikeda, M. Kikuchi, and T. Takahashi, "Dynamical characteristics of state transition defined by neural activity of phase in alzheimer's disease." In *International Conference on Neural Information Processing*, pages 46–54. Springer, 2021.

[44] T. Deng, K. Yang, Y. Li and H. Yan, "Where does the driver look? top-down-based saliency detection in traffic driving environments." *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2051–2062, 2016.

[45] T. Deng, H. Yan, and Y. J. Li, "Learning to boost bottom-up fixation prediction in driving environments via random forest." *IEEE Transactions on Intelligent Transportation Systems*, 19(9):3059–3067, 2017.

[46] S. Basharpoor, F. Heidari, and P. Molavi. "EEG coherence in theta, alpha, and beta bands in frontal regions and executive functions." *Applied Neuropsychology: Adult*, 28(3):310–317, 2021.

[47] C. Li, X. Li, M. Lv, F. Chen, X. Ma, and L. Zhang, "How does approaching a lead vehicle and monitoring request affect drivers' takeover performance? a simulated driving study with functional MRI." *International journal of environmental research and public health*, 19(1):412, 2021.

[48] G. Bernardi, L. Cecchetti, G. Handjaras, L. Sani, A. Gaglianese, R. Ceccarelli, F. Franzoni, F. Galetta, G. Santoro, R. Goebel, et al, "It's not all in your car: functional and structural correlates of exceptional driving skills in professional racers." *Frontiers in human neuroscience*, 8:888, 2014.

[49] G. Underwood, "Visual attention and the transition from novice to advanced driver." *Ergonomics*, 50(8):1235–1249, 2007.

[50] P. Konstantopoulos, P. Chapman, and D. Crundall, "Driver's visual attention as a function of driving experience and visibility. using a driving simulator to explore drivers' eye movements in day, night and rain driving." *Accident Analysis & Prevention*, 42(3):827–834, 2010.

[51] A. K. Pradhan, A. Pollatsek, M. Knodler, and D. L. Fisher. "Can younger drivers be trained to scan for information that will reduce their risk in roadway traffic scenarios that are hard to identify as hazardous?" *Ergonomics*, 52(6):657–673, 2009.

[52] J. T. Bowling, K. J. Friston, and J. B. Hopfinger. "Top-down versus bottom-up attention differentially modulate frontal–parietal connectivity." *Human brain mapping*, 41(4):928–942, 2020.

[53] I. Muela, A. B. Chica, P. G. Fernandez, and C. Castro, "Visual attention in realistic driving situations: Attentional capture and hazard prediction." *Applied ergonomics*, 90:103235, 2021.

[54] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search." *Psychol Rev*, 113(4):766–786, Oct 2006.

[55] K. Zhang, W. Shi, C. Wang, Y. Li, Z. Liu, T. Liu, J. Li, X. Yan, Q. Wang, Z. Cao, et al, "Reliability of EEG microstate analysis at different electrode densities during propofol-induced transitions of brain states." *NeuroImage*, 231:117861, 2021.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Quefrency Domain Features with Residual Networks for Spoof Speech Detection

Madhu R. Kamble
Samsung Research and Development Institute, Bangalore (SRI-B), India.
E-mail: madhu.r@samsung.com

***Abstract*—Despite of having such high-end solutions for spoof speech detection task, generalization approach to the unseen and convolutive complex signals still remains a challenge? The solution should handle the complexity of signals for both replay and synthetic speech signals. The objective of this paper is to extract robust quefrency-domain features along with residual networks for SSD. In particular, we focus on the replay signal that carries the characteristics of the intermediate recording and playback device and hence, the end result is the summation of the complex signal. This makes the signal very challenging and thus, the features must be capabale enough to discriminate the differences. With the recent Neural network-based approach the synthetic signal generation are very much similar to the genuine signal which make the SSD task more complicated. We investigate the short-time cepstrum features for the SSD task on ASVspoof 2019 Challenge database. The cepstrogram results on the PA task, indicates that it achieves better performance for the replay detection. For geniune speech signal, the periodic pattern are easilt observed in the cepstrogram that are missing for the replay as well as for SS and VC signals. This observation is also reflected in the performance evaluation resulting in EER of 2.08 % and t-DCF of 0.04 on evaluation set of LA task obtained with LFCC features and Resnet architecture. Whereas, for PA task with cepstrogram features the least EER we achieved is 1.07 % with 0.056 t-DCF, respectively.**

***Index Terms*—Automatic Speaker Verification, Spoof, Quefrency, Cepstrogram.**

I. Introduction

The Automatic Speaker Verification (ASV) system is a biometric speaker authentication to verify a claimed speaker's identity with the help of machines. The biometric traits that have been successfully used in practical application includes voice, face, iris, fingerprint, palmprint, palm/finger vein, etc [1]–[3]. With the recent increasing technological developments, it is possible to generate spoofed data that resembles very close to their natural counterparts. There are several kinds of spoofing attacks for ASV system, such as replay, impersonation, speech synthesis (SS), voice conversion (VC), and twins [4]. Spoofing is used by hackers to achieve numerous goals, such as getting sensitive information from their targets, or gaining access to restricted area. Hence, when it comes to the real world scenario, a generalized solution is needed to prevent from such hackers [5].

Over the past few years, a series of ASVspoof challenges focusing on different types of attacks were developed after every two consecutive years [6]–[8]. The organizers also developed the database as per the recent trends approaching towards the neural netowrk-based algorithm along with high quality gadgets used for the recording and playing back [8].

Researchers are currently focusing on the development of Spoof Speech Detection (SSD) system that are based on neaural network , graph attention, end-to-end system [9]. Recent end-to-end models have shifted fromhand-crafted features such as LFCC [10] and CQCC [11] to rawwaveforms, achieving state-of-the-art performance [12], [13]. Othereffors have focused on training strategies, such as data augmenta-tion [14] and multi-task learning [15]. A simple attention module was proposed in [16] that infers 3-dimensional attention weights for the feature map in a convolutional layer, which then optimizes an energy function to determine each neuron's importance. An attack Agnostic Dataset was introduced as a combination of two audio DeepFakes and one anti-spoofing dataset [17]. To reduce the variance of genuine embedding space, an architecture was designed called as residual variability block, which can be flexibly integrated into usual convolutional neural networks [18]. The multi-path GMM-MobileNet model was proposed in [19], which includes the GMMs trained on genuine and spoofed speech generated using various attack algorithms or codecs.

Despite of having such high-end solutions, generalization approach to the unseen and convolutive complex replay signals along with rapid development of spoof speech synthesis techniques still remains a challenge. Hence, the solution should take care of the complexity of the signals using the neural network back-end but at the same time it should not be the complex architecture that makes it challenging while developing for the applications. With complex architecture the latency period required to provide the results is more also with big models size it will make the application to work very slow. Instend of just running behind to get the good performance on a closed set database one should also think will is the next problem that we will be facing?

In this work, we are focusing on the simple front-end feature extraction approach using a short-time cepstrum (cepstrogram) features in the quefrency domain for the spoof speech detection (SSD) task. In particular, we analyze the quefrency domain representation for the logical as well as for physical access scenarios. We observed that for the replay detection task the short-time cepstrum obtained in the quefrency domain detects the spoof signals better from the genuine data. For the synthetic signal detection the cepstrogram features did not perfrom that

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

well, however for replay signal with Resnet-18 architecture along with softmax, amsoftmax and ocsoftmax loss function thats yeilds better performance for the SSD task. We performed the experiments on the ASVSpoof 2019 challenge database that focuses on the Logical Access (LA) and Physical Access (PA) tasks, respectively.

II. Details of Cepstrogram

Before proceeding in the Quefrency domain, let us operate the signal on the frequency side. Let $x(t)$ be a signal with a simple echo and it can be represented as [20]:

$$x(t) = s(t) + \alpha s(t - \tau). \tag{1}$$

The Fourier spectral density (spectrum) of signal is given by:

$$|X(f)|^2 = |S(f)|^2[1 + \alpha^2 + 2\alpha\cos(2\pi f\tau)]. \tag{2}$$

The spectral density of a signal with an echo has the form of an envelope (the spectrum of the original signal) that modulates a periodic function of frequency [21]. By taking the logarithm of the spectrum, this product is converted to the sum of two components.

$$\begin{aligned} C(f) &= \log|X(f)|^2 \\ &= \log|S(f)|^2 + \log[1 + \alpha^2 \\ &\quad + 2\alpha\cos(2\pi f\tau)]. \end{aligned} \tag{3}$$

Thus, $C(f)$ viewed as a waveform has an additive periodic component. This new *"spectral"* representation domain was not present in the frequency domain, nor in the time domain. Hence, Bogert *'et al.'* choose to refer to it as the *quefrency domain*, and they termed the spectrum of the log of the spectrum of a time waveform the *cepstrum* [20]. We believe that the periodic patterns can be easily observed in the cepstrogram, and therefore argue that cepstrogram can be a powerful feature for the detection task [22].

The power spectral energy density along with spectrogram as well as cepstrogram representation is shown in Figure 1 for genuine and replay spoof speech signal, respectively. A signal with echoes exhibits rahmonic peaks in the quefrency domain, which reflects its pitch in the cepstrum analysis [22]. We cannot observe much of the difference in the spectral energy density between genuine and the spoofed speech. However, when we analyse the cepstrogram for the genuine and replay speech, we can observe that the replay signal have strong and more dominant peaks. From this investigation we can say that when the genuine speech is re-recorded or simulated with reverberation, additional rahmonic peaks are observed in the quefrency domain.

The replay speech signal r(t) can be modeled as the convolution of the genuine speech signal, x(t), and the room impulse response of the propagating environment, and is given by [23]:

$$r(t) = x(t) * h_r(t), \tag{4}$$

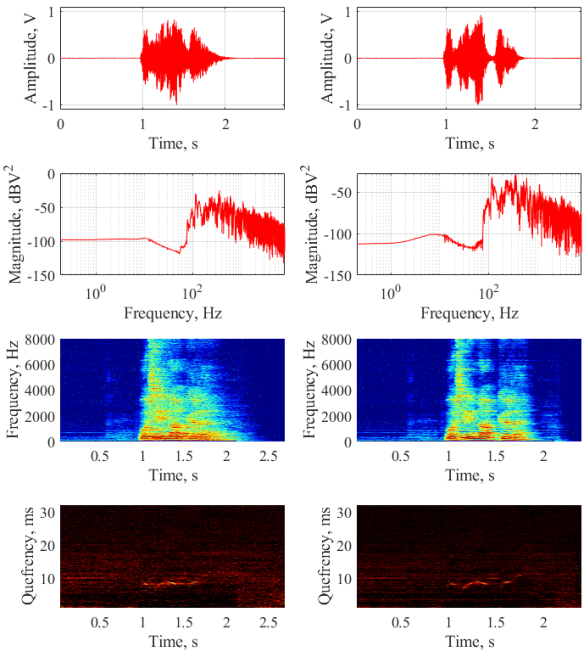


Fig. 1. (a) Time-domain speech signal of natural and replay signal corresponding representation of (b) spectral energy, (c) spectrogram, and (d) cepstrogram.

where the $h_r(t)$ is the convolution of the room impulse response obtained from the recorded room, $h_1(t)$, with room impulse response of the playback room, $h_2(t)$, i.e.,

$$h_r(t) = h_1(t) * h_2(t). \tag{3}$$

First, signals which are combined convolutionally in the time-domain have complex cepstra which are combined additively. As a result, deconvolution is reduced to subtraction in the cepstrum. Second, the complex cepstrum is a measure of the "frequency" of variation (known as quefrency) in the log spectrum, and so signals which vary slowly in the log spectrum may be separated from quickly varying signals by windowing the complex cepstrum [24].

III. Experimental Setup

In this paper, ASVspoof 2019 challenge database is used, which is mainly focused on the Logical and Physical Access (LA and PA) [8]. All the utterances in the database have a sampling rate of 16 kHz, and the resolution of 16-bits per sample. The organizers of the ASVspoof 2019 Challenge provided a baseline system that utilizes two feature sets, namely, CQCC and LFCC with 512 mixtures for modeling genuine and corresponding spoof models in Gaussian Mixture Model (GMM) [10], [11]. The ASVspoof 2019 Challenge uses the minimum Tandem-Detection Cost Function (t-DCF) as the evaluation metric along with EER [25].

We compared five different type of features, namely, Linear Frequency Cepstral Coefficients (LFCC), Constant Q Cepstral Coefficients (CQCC), Teager Energy Cepstral Coefficients (TECC), Gabor Filterbank Cepstral Coefficients (GTFB), and Cepstrogram (CPST). The feature extraction parameters for the same are mentioned below.

LFCC: The LFCC feature set is extracted using 40 linearly-spaced triangular filterbank with F_{min} = 10 Hz, and F_{max} = 8000 Hz with 40 static coefficients, resulting in total 120-D feature vector (including 40- Δ and 40 $\Delta\Delta$) [10]

CQCC: The CQCC features are extracted with a maximum frequency of $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 44.1kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \simeq 43$Hz (9 being the number of octaves). The number of bins per octave B is set to 96. Only 20 static coefficients (with log-energy) were considered, resulting in total 60-dimensional (D) feature vector (including *20*-Δ and *20*-$\Delta\Delta$) [26].

TECC: The TECC feature set was extracted using 40 Mel-spaced Gabor filterbank with f_{min}=10 Hz, and f_{max}=fs/2 Hz. For each subband filtered signals, we obtain 40-D static features appended along with their Δ and $\Delta\Delta$ coefficients resulting in 120-D feature vector [23], [27].

GTFB: The Gabor filterbank feature set were extracted using 40-dimensional linearly spaced with a maximum frequency of $F_{max} = Fs/2$ and a minimum frequency of $F_{min} = 10Hz$, with Δ and $\Delta\Delta$ resulting in total 120-D feature vector [28].

Here, we explored the two types of back-end classifier along with the above mentioned feature sets to have the better SSD system.

A. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a more popular and well known classification technique widely used in signal processing and pattern recognition literature. GMM is a generative model, which represent each class as a weighted sum of *M* multivariate Gaussian and is given by

$$p(x|\lambda) = \sum_{k=1}^{M} w_k p_k(x), \tag{5}$$

where w_k is the k^{th} mixture weight and $p_k(x)$ is a *D*-variate Gaussian density function with mean vector, μ_i and covariance matrix, Σ_i. The model parameter is defined by λ. The decision of whether the test speech being genuine or spoofed depends on the scores of Log Likelihood Ratio (LLR),

$$LLR = log\frac{P(X|H_0)}{P(X|H_1)}, \tag{6}$$

where $P(X|H_0)$ and $P(X|H_1)$ are the likelihood scores of genuine and replay speech, respectively.

B. Deep Residual Network

In this section, we explored binary classification loss functions, followed with one-class learning loss function for voice spoofing detection. The network architecture is adapted from [29] The architecture takes the extracted features as input and outputs the confidence score to indicate the classification result. We use the intermediate output before the last fully connected layer as the embedding for the speech utterances, where 256 is the embedding dimension [30].

The Softmax loss for binary classification is given as:

$$\begin{aligned}\mathcal{L}_S &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{x}_i}+e^{\boldsymbol{w}_{1-y_i}^T\boldsymbol{x}_i}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{(\boldsymbol{w}_{1-y_i}-\boldsymbol{w}_{y_i})^T\boldsymbol{x}_i}\right),\end{aligned} \tag{1}$$

AM-Softmax improves upon this by introducing an angular margin and is given as:

$$\begin{aligned}\mathcal{L}_{AMS} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}}{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^T\hat{\boldsymbol{x}}_i-m)}+e^{\alpha\hat{\boldsymbol{w}}_{1-y_i}^T\hat{\boldsymbol{x}}_i}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha\left(m-(\hat{\boldsymbol{w}}_{y_i}-\hat{\boldsymbol{w}}_{1-y_i})^T\hat{\boldsymbol{x}}_i\right)}\right),\end{aligned} \tag{2}$$

Another countermeasure we used to perform our experiments is based on one class classification technique. The One-class Softmax (OC-Softmax) is denoted as:

$$\mathcal{L}_{OCS} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha(m_{y_i}-\hat{\boldsymbol{w}}_0\hat{\boldsymbol{x}}_i)(-1)^{y_i}}\right), \tag{3}$$

where $\boldsymbol{x}_i \epsilon \mathbb{R}^D$ and $y_i \epsilon 0, 1$ are the embedding vector and label of the i-th sample respectively, $\boldsymbol{w}_0 \epsilon \mathbb{R}^D$ are the weight vectors of the two classes, and N is the number of samples in a mini-batch, α is a scale factor, m is the margin for cosine similarity, and $\hat{\boldsymbol{w}}$, $\hat{\boldsymbol{x}}$ are normalized $\boldsymbol{w}$ and $\boldsymbol{x}$ respectively. For systems trained with OC-Softmax, the CM score is the cosine similarity of $\boldsymbol{x}_i$ and $\boldsymbol{w}_0$.

C. Evaluation Metric

The ASV system uses a standard evaluation metric, i.e., Equal Error Rate (EER), which is indicated on the Detection Error Trade-off (DET) curve. When the operating point in the DET curve of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) or miss probability is *equal*, then it is referred to as EER.

IV. Experimental Result

The experiments are performed on the standard ASVSpoof 2019 Challenge database on both LA and PA task.

TABLE I
LA RESULTS IN EER AND MIN T-DCF REPORTED IN PARAENTHESIS ()

Loss Function	CQCC		LFCC		TECC		GTFB		CPST	
	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
softmax	0.15	9.69 (0.26)	0.24	9.82 (0.07)	**0.022**	9.31 (0.2199)	0.039	10.05 (0.3044)	1.49	9.81 (0.20)
amsoftmax	0.20	7.23 (0.191)	0.43	3.04 (0.06)	0.027	9.55 (0.2105)	0.03	7.09 (0.0184)	1.53	10.30 (0.22)
ocsoftmax	0.156	7.341 (0.187)	0.30	**2.08** (0.04)	0.028	11.45 (0.2145)	0.039	9.10 (0.2352)	1.56	10.17 (0.2068)

TABLE II
BREAKDOWN OF LA ATTACKS

	CQCC			LFCC			TECC			GTFB			CPST		
	SF	AM	OC	SF	AM	OC	SF	AM	OC	SF	AM	OC	SF	AM	OC
A07	0.01	0.10	0.13	0.61	0.32	0.16	1.40	1.64	0.61	0.02	0.01	**0.0**	1.34	2.03	1.26
A08	0.34	0.57	0.73	2.68	0.09	**0.04**	8.15	7.61	4.05	0.75	0.12	0.24	1.99	2.95	1.99
A09	0.34	0.99	1.05	0.08	**0.01**	0.04	1.40	1.68	0.70	0.34	0.09	0.05	0.12	0.34	1.26
A10	11.82	2.23	1.83	1.38	1.28	**0.66**	1.48	1.86	0.89	14.49	8.71	19.25	2.13	3.68	1.26
A11	2.99	1.99	1.52	0.70	0.09	**0.04**	1.32	1.42	0.81	11.81	5.43	8.32	1.60	2.21	1.26
A12	11.02	1.97	1.19	0.52	**0.34**	0.43	2.11	1.60	0.71	1.91	3.49	1.52	1.54	2.66	1.32
A13	15.60	2.11	1.10	1.72	0.30	**0.17**	1.40	1.30	0.48	14.89	9.91	18.72	1.89	3.25	1.26
A14	13.12	4.39	3.01	4.78	0.87	**0.40**	3.94	2.29	1.08	8.71	4.39	7.26	1.12	1.87	1.26
A15	7.95	4.15	3.94	6.39	1.10	1.36	2.19	1.62	1.07	10.52	14.42	8.40	**0.79**	0.95	1.26
A16	0.05	0.34	0.48	0.77	0.43	0.28	2.68	2.44	1.60	0.08	0.04	**0.0**	1.85	1.89	1.26
A17	26.14	29.18	41.91	28.75	12.82	**9.50**	26.52	24.84	34.81	19.57	20.83	9.77	22.12	25.00	20.47
A18	3.60	11.78	10.52	1.50	**0.62**	0.63	22.25	28.59	36.87	12.53	1.89	21.14	40.37	49.14	45.48
A19	0.18	0.85	1.16	2.54	1.10	0.81	13.59	16.70	17.47	0.18	0.04	**0.01**	11.12	10.13	11.62
Pooled	9.69	7.23	7.34	5.66	3.04	**2.08**	9.31	9.55	11.45	10.05	7.09	9.10	9.81	10.30	10.17

A. Results on LA Task

The results in terms of EER and t-DCF with Resnet architecture along with different loss functions for LA task is reported in Table I. We compared OC-Softmax with the conventional binary classification loss functions, under the setting of the same input features and models. In additon, we also compared the results with other feature sets instead of only focusing on the LFCC features. The three losses perform similarly for CQCC, LFCC, TECC, and GTFB on the development set, showing that they have good discrimination ability to detect known attacks. We can observe that on development set TECC feature gave a lower EER of 0.022 %, however with the unknown attacks in the evaluation set, it failed to achieve the better performance. With the quefrency-domain features (CPST) failed to achive the good performance in development as well as in evaluation set. This proves that because of the loss of the natural harmonic periodic patterns the CPST features do not perform good for the synthetic signals and voice converted signals. With LFCC feature set along with OCsoftmax loss function the SSD system achieved a low EER of 2.08 % and t-DCF of 0.04, respectively.

We also compare the performance on individual unseen attacks in the evaluation set reported in Table II. As for individual attacks, LFCC with OC-Softmax achieves good performance over all the other features. However, for A17 attack, it shows the efficacy of one-class learning to detect unknown attacks. Though it is the case for all the features too. As reported in [8], A17 is the most difficult attack in the evaluation set, but still LFCC with oc softmax shows a significant improvement resulting in 9.50 % EER.

B. Results on PA Task

This sections discusses the results on PA task as reported in Table III. The CPST feature set gave better performance with binary classification specifically with softmax loss function resulting in an EER of 0.66 % and 1.70 % on developemnt and evaluation set, respectively. The corresponding min t-DCF for CPST is 0.056. Here, we can observe from the results that the complex cepstra developed which are combined convolutionally in the time-domain generated during the replay recording phase shows its significance discrimination patterns. That yeilds to obtain a good performance for CPST features when compared with other features sets.

We also compare the performance on individual unseen attacks in the evaluation set reported inTable IV. With softmax and one class loss function, CPST features gave least EER for the individual unseen attacks compared to other feature sets. With OCsoftmax loss function, we can observe that for AC, BC, and CC unseen attacks it yeilds lower EER compared to the binary loss function for CPST features. This shows that the replay configurations used during the recording of the signals do matter for the performance degradation and improvements. The attacks with AC, BC, and CC have used low quality device for recording the speech data. Also the distance from talker to the device is more hence, the speech recorded at the final stage is the convoluted with other complexity here. This

TABLE III
PA RESULTS IN EER AND MIN T-DCF REPORTED IN PARAENTHESIS ()

Loss Function	CQCC		LFCC		TECC		GTFB		CPST	
	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
softmax	2.66	5.09 (0.1517)	1.36	1.845 (0.05)	3.11	3.84 (0.1094)	1.03	1.72 (0.0531)	**0.66**	**1.70** (0.056)
amsoftmax	3.09	5.461 (0.1717)	1.51	2.13 (0.066)	4.89	6.22 (0.182)	1.35	1.98 (0.06494)	0.81	1.83 (0.060)
ocsoftmax	3.0	4.63 (0.147)	1.61	1.79 (0.0508)	3.12	4.32 (0.1226)	0.997	2.02 0.0686	0.76	1.95 (0.062)

TABLE IV
BREAKDOWN OF PA ATTACKS

	CQCC			LFCC			TECC			GTFB			CPST		
	SF	AM	OC	SF	AM	OC	SF	AM	OC	SF	AM	OC	SF	AM	OC
AA	6.32	6.37	5.75	2.52	2.90	2.59	4.44	6.06	5.19	2.23	2.82	2.65	**1.90**	2.46	2.07
AB	5.34	5.73	4.24	1.41	1.45	**1.35**	4.59	6.19	5.06	1.42	**1.35**	1.60	1.97	2.11	3.00
AC	3.08	2.34	2.36	1.28	1.50	1.299	2.34	2.95	2.66	1.23	1..49	1.31	1.03	0.73	**0.28**
BA	6.42	6.73	6.04	2.33	2.90	2.36	4.29	7.21	4.58	2.07	2.44	2.50	1.82	1.95	**1.81**
BB	5.98	6.78	4.96	1.37	1.58	1.28	4.35	7.41	5.12	1.52	**1.33**	1.64	2.38	2.20	2.91
BC	2.93	2.43	2.27	1.30	1.55	1.24	2.26	3.68	2.44	1.14	1.35	1.37	1.07	0.87	**0.50**
CA	6.36	7.25	6.22	2.30	2.82	2.13	4.05	8.39	4.37	2.10	2.34	2.46	**1.59**	1.78	1.66
CB	5.42	6.84	4.69	1.54	1.54	**1.29**	4.64	8.73	5.38	1.91	1.44	1.68	1.97	2.00	2.81
CC	2.83	2.37	2.11	1.41	1.57	1.31	2.29	4.06	2.49	1.49	1.64	1.39	1.19	0.85	**0.49**
Pooled	5.09	5.46	4.63	1.84	2.13	1.79	3.84	6.22	4.32	1.72	1.98	2.02	**1.70**	1.83	1.95

all parameters helps to discriminate the speech signal in the quefrency domain as discussed above.

V. SUMMARY AND CONCLUSIONS

This paper focuses on the short-time cepstrum features for the spoof speech detection (SSD) task. We analysed the cepstrogram for the genuine and replayed signal and compared it with the traditional spectrogram. We observed that because of the reverberation and simulated replay signal additional dominant peaks are obtained. The periodic pattern are clearly visible in the cepstrogram for the genuine signal which is missing for the spoofed signal. This difference helps us to detect the replay signal from its natural part. Which eventually reflected it terms of the results for the PA task of ASVspoof 2019 challenge database. On the other side, for LA task, the signals are very much similar to the genuine speech because of the way they are generated using neural networks techniques that makes it to not perform so well with cepstrogram features. In future, we will focus more on the reverberation concept to have better clarity of the replay signal.

REFERENCES

[1] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *INTERSPEECH*, Lyon, France, 2013, pp. 925–929.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (ASV) system," *in Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, pp. 1047–1053, Hawaii, USA, November 12-15, 2018.

[4] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Voice anti-spoofing," *Handbook of Biometric Antispoofing, S. Marcel, SZ Li, and M. Nixon, Eds. Springer*, 2014.

[5] H. Li, H. A. Patil, and M. R. Kamble, "Tutorial on spoofing attack of speaker recognition," in *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC)*, Kuala Lumpur, Malaysia, 2017.

[6] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[7] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2017. [Online]. Available: LastAccess11-Jan-2018,http://www.asvspoof.org/

[8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *INTERSPEECH*, Graz, Austria, September 15-19, 2019, pp. 1008–1012.

[9] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint arXiv:2210.02437*, 2022.

[10] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2087–2091.

[11] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[12] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[13] Q. Fu, Z. Teng, J. White, M. E. Powell, and D. C. Schmidt, "Fastaudio: A learnable audio front-end for spoof speech detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3693–3697.

[14] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6382–6386.

[15] Y. Mo and S. Wang, "Multi-task learning improves synthetic speech detection," in *ICASSP 2022-2022 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 6392–6396.

[16] J. H. Hansen and Z. WANG, "Audio Anti-spoofing Using Simple Attention Module and Joint Optimization Based on Additive Angular Margin Loss and Meta-learning," in *INTERSPEECH*, 2022, pp. 376–380.

[17] P. Kawa, M. Plata, and P. Syga, "Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection," in *INTERSPEECH*, 2022, pp. 4023–4027.

[18] Y. Lou, S. Pu, J. Zhou, X. Qi, Q. Dong, and H. Zhou, "A Deep One-Class Learning Method for Replay Attack Detection," in *INTERSPEECH*, 2022, pp. 4765–4769.

[19] Y. Wen, Z. Lei, Y. Yang, C. Liu, and M. Ma, "Multi-Path GMM-MobileNet Based on Attack Algorithms and Codecs for Synthetic Speech and Deepfake Detection," in *INTERSPEECH*, 2022, pp. 4795–4799.

[20] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE signal processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[21] B. P. Bogert, "The quefrency alanysis of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," *Time series analysis*, pp. 209–243, 1963.

[22] S.-K. Lee, Y. Tsao, and H.-M. Wang, "A study of using cepstrogram for countermeasure against replay attacks," *arXiv preprint arXiv:2204.04333*, 2022.

[23] M. R. Kamble and H. A. Patil, "Analysis of reverberation via Teager energy features for replay spoof speech detection," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK*, pp. 2607–2611, May 12-17, 2019.

[24] D. Bees, P. Kabal, and M. Blostein, "Application of complex cepstrum to acoustic dereverberation," in *Proc. Biennial Symp. Commun*, 1990, pp. 324–327.

[25] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, pp. 312–319.

[26] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[27] Kamble, Madhu R and Patil, Hemant A, "Analysis of Teager energy profiles for spoof speech detection," in *aceptted in APSIPA-ASC*, Lanzhou, China, 2019.

[28] M. R. Kamble, S. Nayak, M. A. B. Shaik, S. P. Rath, V. Vij, and H. A. Patil, "Teager energy subband filtered features for near and far-field automatic speech recognition," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 491–496.

[29] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Computer Speech & Language*, vol. 63, p. 101096, 2020.

[30] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

PDF-NET: Pitch-adaptive Dynamic Filter Network for Intra-gender Speaker Verification

Zhenyu Piao, Hyungseob Lim, Miseul Kim, Hong-goo Kang
Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
E-mail: {jkyung, hyungseob.lim, miseul4345@dsp.yonsei.ac.kr}, hgkang@yonsei.ac.kr

***Abstract*—In this paper, we propose a method to significantly improve the performance of the intra-gender comparison task in a neural speaker verification (SV) system. Recent state-of-the-art neural SV models have a substantial performance decline in the same gender comparison trials, but few studies have comprehensively investigated this problem. To this end, we develop a light-weight pitch-adaptive dynamic filter network (PDF-NET) that generates the convolutional filters conditioned on the input pitch information, from the observation that intra-gender speakers share a similar pitch range. By doing so, speaker embeddings from the same speakers are well clustered with each other even within the same gender class. Experimental results show that our proposed method outperforms the baseline for various SV datasets. In particular, female SV trials on LibriSpeech performed herein achieve a relative improvement of 11.4% in terms of equal error rate with only 1.6k additional parameters.**

Index Terms: Intra-gender speaker verification, dynamic filter network, pitch-adaptive filter, light-weight system

I. Introduction

Voice biometrics or speaker verification (SV) is a technology that uses voice to authenticate a claimed person. SV has received considerable attention in recent years because of its ease of use, a variety of potential applications [1]–[3], and advances in neural network architectures.

Most recently developed deep learning-based SV models have similar frameworks, wherein the input raw waveform is first converted into hand-crafted acoustic features such as mel-spectrograms or mel-frequency cepstral coefficients (MFCCs). Then, those acoustic features are fed into a trunk network such as ResNet [4], [5] or TDNN-based models [6], [7] to generate frame-level speaker embeddings. In the end, the frame-level embeddings are aggregated into a fixed-dimensional utterance-level embedding through various pooling methods [8], [9]. Nowadays, the dominant research on SV focuses on finding a reliable trunk network architecture to improve the representation capacity of the models. Stacking a large number of layers, enlarging feature dimensions to the axis of depth and width are typical examples of the approaches [6], [10], [11].

Despite the development of well-established trunk networks, most SV systems still exhibit low discriminative performance on intra-gender verification tasks where the fundamental frequencies of speakers tend to be within a similar range [12]–[14]. The low discriminative performance can be attributed to two limitations of existing SV models. First, the frequency resolution of input mel-scale features is not sufficiently high for capturing speaker-specific features such as fundamental frequencies. Second, most of the SV networks use input-invariant static convolutional filters [6], [15] to extract an embedding vector from the mel features, hence it is very difficult to reflect dynamic local characteristics of input features.

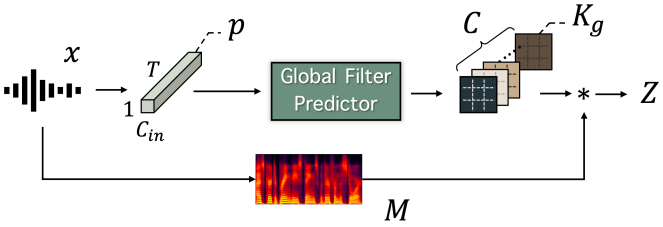


Figure 1. Illustration of the global filter predictor (GFP). Global filter K_g is obtained from an utterance-level condition $\boldsymbol{p}$. The output embedding Z is combined with the embedding obtained from static filters to be used as an input of the trunk network.

To overcome the aforementioned limitations, it was proposed to utilize high-dimensional acoustic features [16]–[20] rather than only utilizing the mel-scale features. For example, some early studies incorporate fundamental frequency (f0) or delta-f0 into the existing SV models [16]–[19], and succeed in improving the SV performance. However, we found that directly utilizing f0 features as input to the system could not contribute to the intra-gender tasks (see Section V-A for more details). On the other hand, there exists some predominant methods that stack a large number of layers or increase the channel size [6], [7], [11] thereby improving the modeling performance. However, those methods exhibit high computational complexity and require a large memory size, which can not satisfy the increasing demand for light-weight SV systems [21]–[23].

In this paper, we propose a light-weight SV model named pitch-adaptive dynamic filter network (PDF-NET) to improve the SV performance on the intra-gender comparison tasks. Our PDF-NET consists of dynamic convolutional filters that are adaptive to the pitch[1] information of the input signal. More specifically, we design two filter prediction modules named global filter predictor (GFP) and local filter predictor (LFP). The former generates global pitch-adaptive filters in the utterance-level, while the latter generates local pitch-adaptive filters in the frame-level. Experimental results show that our proposed model's equal error rate (EER) on the female comparison task on LibriSpeech is decreased by 11.4% from the baseline model, with a marginal increase in the number of

[1]In this work, we use the term "pitch" as a synonym of "fundamental frequency" for brevity, although their precise meanings are different.

model parameters.

Our contributions in this paper are summarized as follows:

- To the best of our knowledge, our approach is the first attempt to demonstrate the efficacy of pitch-adaptive dynamic convolutional filters on intra-gender SV tasks.
- Based on the famous Fast ResNet architecture, we develop a light-weight SV model which shows better performance than the baseline model while maintaining the number of parameters in the system.

II. Related Work

A. Pitch-conditioned SV

In the early exploration stage of speaker verification research, several efforts had been made to utilize pitch information to improve the performance. Yegnanarayana et al. [17] highlights the importance of the suprasegmental information and incorporates these features into a baseline SV system that only uses the spectral information. In [18], Dehak et al. additionally utilize various prosodic features (e.g., pitch, energy, and duration) using joint factor analysis to improve the SV performance. Gong et al. [19] introduce a pitch synchronous feature extraction method which utilizes pitch information to determine the size of the short-time Fourier Transform (STFT) window, demonstrating the effectiveness of the proposed method.

However, these studies only focus on the selection or extraction of the input features, and an effective way to exploit the information (especially upon the deep neural network architectures) has not been thoroughly explored. Based on the analysis of the neural SV architectures, we develop a sophisticated method to efficiently provide pitch information to a neural SV model.

B. Adaptive Convolutional Neural Network for SV

Several methods have been proposed to generate input-adaptive convolutional neural network (CNN) filters. For example, Kim and Park [24] utilizes input-dependent filters during the feature extraction stage to accurately model the dynamic nature of phonetic variations in each analysis frame. Furthermore, a temporal dynamic convolutional neural network (TDY-CNN) [25] has been proposed to apply an adaptive filter for each time sequence. Here, the adaptive filters are obtained from the weighted sum of basic filters with a softmax constraint on attention weights.

Although the aforementioned approaches can achieve high performance in various SV tasks, they have a few limitations: 1) They require a large number of parameters and involve complex computations, which make it difficult to build a light-weight real-time SV system. 2) The adaptive filter estimation process only relies on input mel-scale features with low frequency resolution. In this study, we generate filters based on more elaborate speaker-related features (i.e., pitch) to effectively estimate speaker-discriminative embeddings with a small increment of model size compared to the baseline model. In other words, our proposed SV network is designed to be more pitch-aware and distinctive to identify the differences between intra-gender speakers having similar pitch ranges.

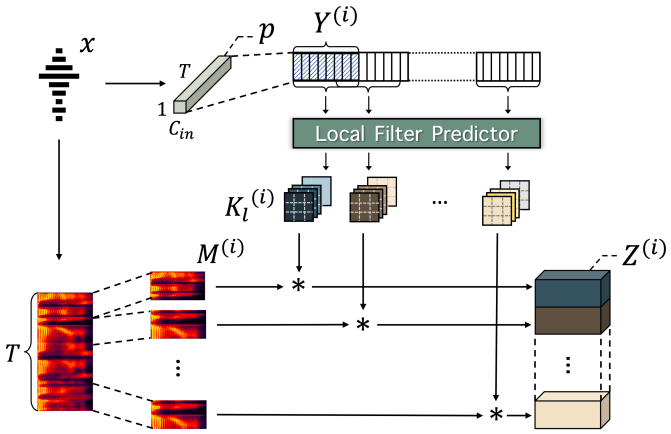


Figure 2. Illustration of the local filter predictor (LFP). Predicted filter $K_l^{(i)}$ is obtained from $Y^{(i)}$, then is convolved with $M^{(i)}$ to generate $Z^{(i)}$. Finally, the concatenated output Z is combined with features obtained from static kernel to feed into the trunk network.

III. Proposed Model

As shown in Table III, existing neural SV models exhibit relatively poor performance on the intra-gender comparison tasks[2], for the reason that those speakers have similar pitch ranges. In order to solve this problem in an efficient way, we propose PDF-NET which adaptively generates dynamic convolutional filters conditioned on the pitch information of the input signal.

We propose two different filter prediction modules—global filter predictor (GFP) and local filter predictor (LFP). Although both of them generate pitch-adaptive dynamic filters from the input fundamental frequencies, the scopes of the filters are different. The GFP generates global utterance-level pitch-adaptive filters by summarizing all the information extracted from its input pitch contour. By contrast, the LFP generates different local pitch-adaptive filters for different time frames. The latent vectors obtained from the dynamic filters are combined with the ones obtained from static convolutional filters to form a new speaker representation. The following sections describe the structures of the two pitch-adaptive filter predictors in more detail.

A. Global filter predictor (GFP)

Figure 1 illustrates the operation of GFP, which can be summarized as follows:

$$K_g = \text{Reshape}(\text{FC}(\text{Conv1D}(\boldsymbol{p}))). \quad (1)$$

First, the pitch contour $\boldsymbol{p} \in \mathbb{R}^{1\times T}$ is extracted from an input raw waveform, $\boldsymbol{x}$, where T denotes the total number of frames. Then, the pitch contour passes through 1-D CNN layers after log-compression to enlarge the channel dimension into C. A fully-connected (FC) layer is applied along the time dimension to summarize the temporal variation of the pitch contour into a single vector of length k^2C, where k denotes the filter size. The vector is reshaped into a tensor $K_g \in \mathbb{R}^{C\times 1\times k\times k}$, which is convolved with the mel-spectrogram, M, extracted from

[2]Similar observation has been found when we conducted the verification task using other SV models and intra-gender datasets.

the same input waveform, $\boldsymbol{x}$. Note that the dynamic filters are applied only to the input mel-spectrogram and not used in the intermediate layers of the trunk network.

B. Local filter predictor (LFP)

Figure 2 depicts the process of generating and applying dynamic local filters. Unlike the GFP, the LFP generates different filters for different time steps. To obtain local frame-level filters that account for temporal-dynamic speaker characteristics, 1-D CNN with filter size k and stride s is applied to the log-compressed pitch contour, $\boldsymbol{p}$. The obtained latent vectors, Y, and the mel-spectrogram, M, are aligned and segmented so that each segment includes the n number of vectors. Thereafter, local filters for the i-th segment, $K_l^{(i)}$, is generated from $Y^{(i)}$. A segment of the mel-spectrogram, $M^{(i)}$, is convolved with the filters to generate a speaker representation $Z^{(i)}$. The same process is repeated for all the segments. In summary, the process of local filter prediction is given by:

$$K_l^{(i)} = \mathrm{Reshape}(\mathrm{FC}(Y^{(i)})), \tag{2}$$

where $Y^{(i)}$ is the i-th segment of the output of the convolutional layer with stride s. The optimal number of filters, k, and the stride, s of the 1-D convolutional layer were experimentally determined to obtain localized pitch-adaptive dynamic filters.

IV. Experiment setup

A. Dataset

We use two popular datasets, named VoxCeleb [26], [27] and LibriSpeech [28], for the experiments. For training, we use the development subset of VoxCeleb2 [27] without any augmentation methods. For the performance evaluation, we use the benchmark VoxCeleb1-O test dataset. Because there is no available public dataset designed for intra-gender verification task, we construct two additional intra-gender test datasets to evaluate the performance on the task. As illustrated in table I, the newly-developed datasets are named – Vox1-F, Vox1-M, Libri-F and Libri-M. Vox1-F and Vox1-M, which consists of samples in the VoxCeleb1 development subset. Vox1-F contains the data of 561 female speakers, and 100 pairs are made for each speaker; Vox1-M contains data from 100×690 trials with male speakers. Similarly, we also sampled trials from the LibriSpeech [28] train-clean-100 set to construct LibriSpeech intra-gender dataset, named Libri-F and Libri-M, which contain 12,500 and 12,600 random pairs from 125 female and 126 male speakers, respectively.

B. Implementation details

For the input acoustic features, we use 40-dimensional mel-spectrogram and 1-dimensional pitch. We convert the time domain raw waveform into frequency domain acoustic features with a frame length of 25 ms and a hop length of 10 ms. We estimate the frame-level pitch values using the pyin algorithm [29] and then convert the pitch values into a log scale for a better performance.

Table I
Details of intra-gender datasets. Vox1 and Libri are the abbreviations of VoxCeleb1 and LibriSpeech datasets, respectively. F and M indicates the gender of intra-gender trials.

	# of speakers	Trial/speaker	Total # of trials
Vox1-F	561	100	56,100
Vox1-M	690	100	69,000
Libri-F	125	100	12,500
Libri-M	126	100	12,600

Table II
Overall performance of the proposed model compared with the baseline on Vox1-O dataset in terms of equal error rate (EER) and minimum detection cost function (MinDCF), along with the number of model parameters.

Model	Params. (M)	EER (%)	MinDCF
Baseline	1.438	2.797	0.197
GFP	1.448	2.691	0.191
LFP (s=8, k =3)	1.444	2.781	0.186
LFP (s=8, k =5)	1.449	**2.617**	**0.176**
LFP (s=10, k =3)	1.445	2.622	0.180
LFP (s=10, k =5)	1.450	2.792	0.187

Regarding the choice of the trunk network, we adopt the Fast ResNet-34 model [15] because of its light-weight architecture (about 1.4M parameters) and small number of channels. After extracting frame-level embeddings from the trunk network, self attentive pooling (SAP) [9] is used to summarize frame-level features into a 512-dimensional utterance-level speaker embedding. Then, additive margin softmax (AM-Softmax) loss [30] is applied to the embedding for speaker verification task.

During the training stage, we use a 2-second segment for a stable training, which is randomly cropped from each utterance. And we train all the baseline and proposed models for 150 epochs equally. For the inference stage, we adopt the testing scheme of [15], where ten 2-second temporal crops are generated at regular intervals. The Adam optimizer [31] with an initial learning rate of 0.001 and StepLR scheduler with a learning decay of 0.95 are used for training.

C. Evaluation metric

For the evaluation metric, we use the equal error rate (EER) and the minimum of the detection cost function (MinDCF) to compare the performance of the proposed and baseline models. EER indicates the point where both false acceptance (FA) and false rejection (FR) rates are equal. MinDCF is a weighted sum of FA and FR probabilities with given parameters $P_{\mathrm{Target}} = 0.05, C_{\mathrm{Miss}} = C_{\mathrm{FalseAlarm}} = 1$.

Table III
PERFORMANCE OF THE PROPOSED MODEL COMPARED WITH THE BASELINE ON VARIOUS INTRA-GENDER DATASETS.

Model	Vox1-F		Vox1-M		Libri-F		Libri-M	
	EER (%)	MinDCF	EER (%)	MinDCF	EER (%)	MinDCF	EER (%)	MinDCF
Baseline	4.659	**0.269**	4.135	0.242	1.824	0.108	1.222	**0.055**
Baseline (Mel+f0)	4.784	0.288	4.195	0.249	1.776	0.115	1.397	0.067
GFP	4.601	0.295	**3.901**	0.238	**1.616**	0.100	**1.206**	0.060
LFP (s=8, k =3)	4.941	0.293	4.316	0.246	2.064	0.111	1.429	0.059
LFP (s=8, k =5)	4.733	0.301	4.189	0.239	1.872	0.108	1.270	0.070
LFP (s=10, k =3)	**4.520**	0.288	4.195	**0.236**	1.984	0.130	1.254	0.076
LFP (s=10, k =5)	4.67	0.285	4.165	0.251	1.968	**0.096**	1.270	0.068

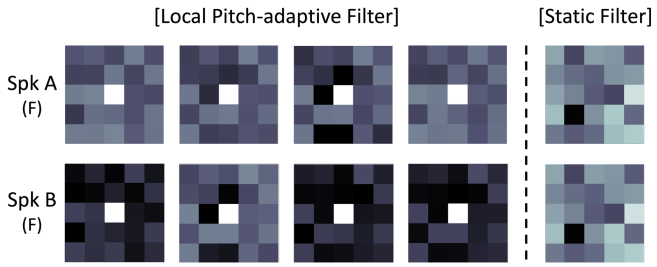


Figure 3. Visualization of local pitch-adaptive and static filters. Local pitch-adaptive filters are obtained from different speakers (vertical) and time sequences (horizontal).

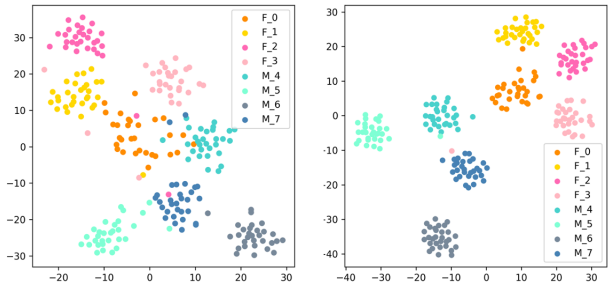


Figure 4. t-SNE plot of speaker embedding from static filters (left) and proposed pitch-adaptive filters (right).

V. EXPERIMENTAL RESULTS

A. Performance Evaluation

Results on the Voxceleb-O dataset. Table II summarizes the performances and corresponding model size of the proposed method and baseline models evaluated on the benchmark Voxceleb1-O test dataset. All trials of the proposed GFP and LFP modules outperform the baseline Fast ResNet-34 models by a large margin, even with a small increments (0.01M $\sim$ 0.04M) on the number of parameters. For GFP, it outperforms the baseline system with 3.8% and 3.0% relative improvements in terms of EER and MinDCF, respectively. Moreover, the LFP module exhibits superior performance compared with the GFP. The optimal setting of the LFP with stride 8 and filter size 5, improves the SV performance by 6.4% EER and 10.7% MinDCF relatively.

Results on the intra-gender datasets. To evaluate the performance of the proposed models more accurately, we further investigate the models in some more challenging intra-gender SV trials. As discussed in section IV-A, we generate four additional datasets: Vox1-F, Vox1-M, Libri-F and Libri-M. As shown in Table III, directly incorporating f0 into the system as an additional input does not contribute to the intra-gender verification task. Instead, our proposed model and its variants show an improvement on the intra-gender SV performance in various scenarios. Especially, the model using the GFP module exhibits good performance on the intra-gender verification task; it improves the relative EER score by 6.3% and 3.5% in the intra-female and intra-male comparison tasks, respectively. Furthermore, a significant improvement has been found in the LibriSpeech female comparison where our proposed GFP module outperforms the baseline by 11.4 % EER relatively. However, there is lack of a single configuration that works best in all domain, presumably due to the distinct characteristics of the two datasets we used. Still, our proposed models show superior performance in most cases.

B. Visualization of the pitch-adaptive dynamic filters

To investigate the role of the dynamic filter predictors, we visualize the local pitch-adaptive dynamic filters together with the static ones obtained from two female speakers as illustrated in Figure 3. We plot 4 different local filters extracted from 4 different time sequences for each speaker, and all the local pitch-adaptive and static filters are extracted from the same utterance and channel. Because the local pitch-adaptive filter predictor generates dynamic filters for each speaker in a time-dependent manner, all of the filters obtained from different speakers and time sequences are not the same. Especially, even for those speakers who are in the same gender category, our proposed local filter predictor is able to generate totally different filter patterns so that it enables better discrimination between different speakers. In contrast, the static filters are invariant for various speakers (for brevity, we only plot one filter for each speaker). Hence it could not adaptively filter the input vectors in a speaker-dependent way.

C. t-SNE plot of the speaker embedding

In order to analyze the different characteristics of the output speaker embedding extracted from static filters and proposed

pitch-adaptive dynamic filters, we visualize the embeddings using t-SNE plots, as displayed in Figure 4. The female speakers are annotated with labels 0 to 3 and depicted in light (orange, yellow, hot pink, light pink) colors, while male speakers are annotated with labels 4 to 7 and depicted in deep (blue, mint, grey, navy) colors. Comparison of the two different types of filters with respect to the density of the clusters reveal that the embeddings obtained from the static filters are well clustered within the same speaker. However, the intra-class variance is still large. On the contrary, speaker embeddings extracted from pitch-adaptive filters have smaller intra-class variance compared to that of the static filters, which demonstrate the efficiency of our pitch-adaptive dynamic filters on discriminating intra-gender speakers.

VI. Conclusion

In this study, we proposed a pitch-adaptive dynamic filter network (PDF-NET) model to improve the speaker verification performance on intra-gender trials. PDF-NET consists of two pitch-adaptive filter predictors; one generates global utterance-level dynamic filters and the other generates local frame-level dynamic filters. Since our proposed model adaptively generates filters depending on the different pitch characteristics of input speakers, it is easy to discriminate intra-gender speakers who have similar pitch ranges. Our proposed model shows better performance than the baseline model in various SV datasets only with marginal increment on the number of parameters. Specifically, our model outperforms the baseline model by 11.4 % relative EER improvement in the LibriSpeech female comparison task, indicating the efficiency of the proposed method.

VII. Acknowledgement

This work was supported by Hyundai Motor and Kia.

References

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE circuits and systems magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[2] M. G. Gomar, "System and method for speaker recognition on mobile devices," May 26 2015, uS Patent 9,042,867.

[3] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers & Electrical Engineering*, vol. 90, p. 107005, 2021.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*. IEEE, 2016, pp. 770–778.

[5] M. Kim, Z. Piao, S. Um, R. Lee, J. Joh, S. Lee, and H.-G. Kang, "Light-weight speaker verification with global context information," in *Proc. INTERSPEECH*. IEEE, 2022, pp. 5105–5109.

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. INTERSPEECH*. IEEE, 2020, pp. 3830–3834.

[7] Y.-Q. Yu and W.-J. Li, "Densely connected time delay neural network for speaker verification." in *Proc. INTERSPEECH*. IEEE, 2020, pp. 921–925.

[8] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. INTERSPEECH*. IEEE, 2018, pp. 2252–2256.

[9] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification," in *Proc. INTERSPEECH*. IEEE, 2018, pp. 3573–3577.

[10] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[11] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," in *Proc. INTERSPEECH*. IEEE, 2022, pp. 306–310.

[12] M. Latinus and M. J. Taylor, "Discriminating male and female voices: differentiating pitch and gender," *Brain topography*, vol. 25, no. 2, pp. 194–204, 2012.

[13] C. R. Pernet and P. Belin, "The role of pitch and timbre in voice gender categorization," *Frontiers in psychology*, vol. 3, p. 23, 2012.

[14] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, vol. 5, no. 2, pp. 211–225, 2012.

[15] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. INTERSPEECH*. IEEE, 2020, pp. 2977–2981.

[16] M. K. Sönmez, E. Shriberg, L. P. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification." in *ICSLP*, vol. 7, 1998, pp. 3189–9192.

[17] B. Yegnanarayana, S. M. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Transactions on speech and audio processing*, vol. 13, no. 4, pp. 575–582, 2005.

[18] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[19] W.-G. Gong, L.-P. Yang, and D. Chen, "Pitch synchronous based feature extraction for noise-robust speaker verification," in *Proc. CISP*, vol. 5. IEEE, 2008, pp. 295–298.

[20] A. Fazel and S. Chakrabartty, "An overview of statistical pattern recognition techniques for speaker verification," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 62–81, 2011.

[21] I. Velez, C. Rascon, and G. Fuentes-Pineda, "Lightweight speaker verification for online identification of new speakers with short segments," *Applied Soft Computing*, vol. 95, p. 106704, 2020.

[22] L.-C. Chang, Z. Chen, C. Chen, G. Wang, and Z. Bi, "Defending against adversarial attacks in speaker verification systems," in *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*. IEEE, 2021, pp. 1–8.

[23] Q. Li, L. Yang, X. Wang, X. Qin, J. Wang, and M. Li, "Towards lightweight applications: Asymmetric enroll-verify structure for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7067–7071.

[24] S.-H. Kim and Y.-H. Park, "Adaptive Convolutional Neural Network for Text-Independent Speaker Recognition," in *Proc. INTERSPEECH*. IEEE, 2021, pp. 66–70.

[25] S.-H. Kim, H. Nam, and Y.-H. Park, "Temporal dynamic convolutional neural network for text-independent speaker verification and phonemic analysis," in *Proc. ICASSP*. IEEE, 2022, pp. 6742–6746.

[26] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. INTERSPEECH*. IEEE, 2017, pp. 2616–2620.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. INTERSPEECH*. IEEE, 2018, pp. 1086–1090.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[29] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. ICASSP*, 2014, pp. 659–663.

[30] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. CVPR*. IEEE, 2018, pp. 5265–5274.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Subjective Evaluation of a Focused Sound Source Reproducing at the positions of a Listener's Moving Hand

Miho Hirohashi* and Yoichi Haneda†
* The University of Electro-Communications, Japan
E-mail: h2230118@edu.cc.uec.ac.jp
† The University of Electro-Communications, Japan
E-mail: haneda.yoichi@uec.ac.jp

***Abstract*—Herein, we investigated the subjective impression of a mid-air sound source whose position was directly manipulated by hand tracking. The mid-air sound source was generated as a focused sound source based on wave field synthesis with a 32-channel linear loudspeaker array. The hand position was tracked using motion capture equipment (Leap Motion). If the focused sound source is not moved within the frame depending on the motion capture acquisition period, the focused sound source cannot be moved smoothly. To achieve smoother and more natural movement of the focused sound source, we moved the focused sound source sample-by-sample in the frame using an all-pass filter. Subjective tests indicated that sound localization and enjoyment increased when the focused sound source followed the hand movements of a listener. Moreover, when the focused sound source was pinched and moved, the sense of movement was enhanced by changing the sound volume according to the moving speed of the focused sound source. The proposed system can contribute to the development of innovative 3D audio applications.**

I. Introduction

Recently, the demand for 3D audio technology has increased owing to the development of 3D video technology. For example, movies use audio technologies, e.g., the 5.1 surround sound system. However, such 3D sound is designed by professional sound creators, and the audience can only hear it. However, the importance of sound in a 3D space has attracted attention owing to recent developments in XR [1]. Spatial sound considerably influences immersion in a 3D virtual space [2]. For example, soundscapes and footsteps that match walking in a virtual space can enhance the sense of immersion in the virtual space [3]. In addition, research is being conducted to make the body feel lighter and support the learning of sports and rehabilitation by playing sounds along with movements [4], [5]. However, these studies focused on the listener's body instead of changing the perception of the sound. Some systems use motion capture to mix sound in 3D space and reproduce moving sound sources acquired by motion capture in real time in a remote room [6], [7]. However, these systems do not allow listeners, who can manipulate sound through movements, to feel sound in real time. Another system allows the listener to adjust the spatial equalizer and move the virtual sound source at the listening position [8]. However, this system uses knobs to make adjustments, which is not intuitive. In the field of virtual reality as well, research on the application of sound and research on the relationship between images and localization accuracy are being conducted, but they can still be examined [9], [10]. Therefore, in this study, we investigated how listeners perceive sounds when they actively change the sound source position. We used a hand-manipulation-focused sound source generation system that can directly change the position of the focused sound source.

We generated focused sound sources using the wave field synthesis (WFS) method [11], [12]. In conventional WFS, filters for obtaining loudspeaker driving signals are analyzed in the frequency domain. In this method, if the focused sound source position changes for each sample, the filter must be regenerated at every sample time, resulting in an enormous amount of computation. However, when the loudspeaker drive signal is calculated for each frame, it is necessary to overlap and add frames by windowing. In this case, the phases of the loudspeaker output signals between the two frames do not match, and the sound may not be natural.

Herein, we use a time-domain-focused sound source formulation instead of a frequency-domain formulation [12]. This approach helps segregate focused sound source generation filters into three parts. The first part is a filter specific to the system that can be pre-convoluted into a source signal. The other two parts are delays and scalar gains, which depend on the loudspeaker and focused sound source positions. Although the delays are fractional, they can be considered as all-pass filters to move the focused sound source sample-by-sample within a frame based on a small amount of computation.

We used Leap Motion, a hand-tracking device, to capture hand positions [13]. When a focused sound source is generated at the position of the hand acquired using Leap Motion, the listener can feel as if he/she is moving the focused sound source manually. Moreover, BoomRoom [14] uses a circular loudspeaker array; however, we use a linear loudspeaker array and simple motion capture.

Using the aforementioned system, we investigated the ex-

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

perience of people when a focused sound source moves as they manipulate it. In the experiment, we first investigated the localization accuracy of the focused sound source generated using the proposed system without hand tracking. Next, we investigated sound localization and listener enjoyment when the sound image was generated at the position of the listener's hand. Finally, we investigated whether there was a sense of moving the sound image, which was weightless and irrelevant, when the position and sound volume of the focused sound source were adjusted by simulating the movement of the focused sound source by pinching it with a hand.

The remainder of this study is organized as follows: Section II describes a hand-manipulation focused sound source generation system that can move the focused sound source position via hand tracking. Section III presents the listening experiments conducted using the aforementioned system. Section IV presents the conclusions and prospects of generating interactively focused sound sources.

II. Proposed Method

The proposed system uses a 32-channel linear loudspeaker array, motion capture technology, personal computer for signal generation, and Digital-to-Analog Converter (D/A converter). Fig. 1 shows the overview of the system. This system generates a focused sound source that moves according to hand movements, which listeners hear.

A. Smoothly moving focused sound source generation method

Using (1), a focused sound source can be generated using a linear loudspeaker array [12]. $D_{2.5\mathrm{D}}(\mathbf{r}_s, \omega)$ is a frequency domain formula that derives the loudspeaker output signal at $\mathbf{r}_s$. This equation was derived from the Rayleigh integral, and the Hankel function was replaced with an approximate exponential function.

$$D_{2.5\mathrm{D}}(\mathbf{r}_s, \omega) \approx -g_s \sqrt{\frac{\frac{\omega}{c}}{2\pi j}} \frac{y_s - y_F(n)}{|\mathbf{r}_s - \mathbf{r}_F(n)|} \frac{e^{j\frac{\omega}{c}|\mathbf{r}_s - \mathbf{r}_F(n)|}}{\sqrt{|\mathbf{r}_s - \mathbf{r}_F(n)|}} \quad (1)$$

Here, $\mathbf{r}_s = [x_s, y_s]$ denotes the loudspeaker position, s denotes the number of loudspeakers, and $s = 1 \sim S$ when the number of loudspeakers is S. ω denotes the angular frequency, c denotes the sound speed, and j denotes an imaginary unit. $\mathbf{r}_F(n) = [x_F(n), y_F(n)]$ denote the position of the focused sound source at time n. g_s is defined as $g_s = \sqrt{2\pi\,|\mathbf{r}_{\mathrm{ref}} - \mathbf{r}_s|}$, and $\mathbf{r}_{\mathrm{ref}}$ are the coordinates for approximately correcting the sound attenuation characteristics in 2.5 dimensions.

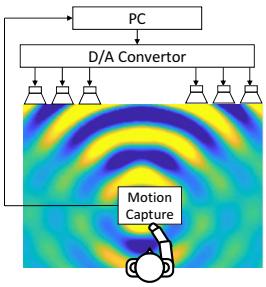


Fig. 1. Overview of proposed system.

The time-domain filter $d(\mathbf{r}_s, n)$ is obtained using the inverse Fourier transform of (1).

$$d(\mathbf{r}_s, n) = G_s(n) Q(t) \delta(n - \tau_s(n)) \quad (2)$$

$G_s(n)$ denotes the scalar gain that differs for each loudspeaker.

$$G_s(n) = -g_s \sqrt{\frac{1}{2\pi}} \frac{y_s - y_F(n)}{|\mathbf{r}_s - \mathbf{r}_F(n)|} \frac{1}{\sqrt{|\mathbf{r}_s - \mathbf{r}_F(n)|}} \quad (3)$$

$Q(t)$ denotes the Finite Impulse Respomse (FIR) filter obtained in advance using (4) and is a fixed filter that does not depend on the position of the loudspeaker or focused sound source.

$$Q(t) = \mathrm{IFFT}\left(\sqrt{\frac{\omega}{jc}}\right) \quad (4)$$

As this filter does not depend on the loudspeaker position, it can be pre-convolved with the source signal.

$\delta(n - \tau_s(n))$ is realized with a delay. However, the delay of each loudspeaker $\tau_s(n) = \frac{|\mathbf{r}_s - \mathbf{r}_F(n)|}{c}$ is fractional. Fractional delays should be implemented precisely by using a sinc function [15]. However, as this function requires a large amount of computation, we replace it with a first-order time-varying all-pass filter [16]. First, we divide $\tau_s(n)$ into integers and fractional parts $\alpha(n)$. Then, we express the fractional delay using a first-order time-varying all-pass filter, as expressed in (5).

$$H_n(z) = \frac{z^{-1} + (1 - \alpha(n))}{1 + (1 - \alpha(n)) z^{-1}} \quad (5)$$

Using the aforementioned method, the position of the focused sound source was smoothly changed for each sample.

B. Directly manipulate the focused sound source using motion capture

By generating a focused sound source at the position obtained through hand tracking, it is possible to generate a mid-air sound image following the hand position. The proposed system uses motion capture hand-tracking technology known as Leap Motion [13]. Leap Motion Controller is an optical hand tracking module that can measure hand movements in detail by holding the hand over. Leap Motion measures hand movements at a maximum of 200 fps [13]. This value is below the sound-sampling frequency of 48000 Hz. Therefore, frame processing is introduced. This system acquires the hand position once per frame and interpolates it within the frame using a uniform linear motion. To interpolate the hand position in the frame from the previous hand position, the focused sound source position was delayed by one frame from the actual hand position. Let N be the frame size and $\mathbf{h}(n)$ be the hand position obtained at time n. The focused sound source

positions at nth time and $n+N$th time are expressed in (6) and (7), respectively.

$$\mathbf{r}_F(n) = \mathbf{h}(n-N) \tag{6}$$

$$\mathbf{r}_F(n+N) = \mathbf{h}(n) \tag{7}$$

The focused sound source position $\mathbf{r}_F(n+i)$ $(i = 0, 1, 2, ..., N-1)$ is expressed in (8).

$$\mathbf{r}_F(n+i) = \frac{(\mathbf{r}_F(n+N) - \mathbf{r}_F(n))}{N} i + \mathbf{r}_F(n) \tag{8}$$

III. Experiments

We conducted subjective evaluations using the system described in Section II. First, we conducted a basic experiment to investigate the localization accuracy of the focused sound source, including the source distance from the listeners. Next, we investigated sound localization and sense of enjoyment while directly manipulating the focused sound source manually. Furthermore, the sensation obtained by pinching and moving an insubstantially focused sound source was subjectively evaluated.

A. Experiment A: basic localization without hand motion

First, we conducted a sound-localization experiment using a 32-channel linear loudspeaker array. The listening experiment did not involve hand tracking.

1) Implementation: Fig. 2 shows the positional relationship between the equipment and listener. The number of loudspeakers was 32, and the loudspeaker spacing was 10 cm. The movement range of the focused sound source is indicated by the colored area in Fig. 2 considering the possible generation area of the focused sound source. The sampling frequency was 48 kHz, signal bandwidth was 120 Hz $\sim$ 8 kHz, and frame length was 4800 samples. The sound source signal was white noise with a duration of 8 s.

2) Procedure: The experiment included 10 people aged 22–29 years. Localization experiments were first conducted with a stationary focused sound source and then with a moving focused sound source without hand tracking. Fig. 3 shows the position and movement direction of the focused sound source. Position 9 was at a distance of 0.3 m from the participant's position in the direction of the linear loudspeaker array. Positions 1–8 are eight points set at 45-degree intervals on a circle with a radius of 0.3 m centered at position 9. The participants were asked to face forward and keep their heads still while listening. In addition, the participants did not move their hands in this experiment A.

In the experiments with a stationary focused sound source, the focused sound source was reproduced at any position from 1 to 9. The focused sound source was presented twice at each position in a random order, i.e., it was presented 18 times in total. The participants listened to the sounds and answered with a number from 1 to 9 as to which position they felt was the most appropriate for the image sound position.

In the experiments, the focused sound source moved among positions 1–8, which existed in a straight line passing through position 9. The focused sound source remained stationary at the starting point for 2 s, moved at a constant speed for 4 s, and remained stationary for 2 s after reaching the end position. The focused sound source was presented twice at each position in a random order, i.e., it was presented 16 times in total. Participants listened to the presented sound and answered with a number ranging from 1 to 8 that they felt was the most appropriate arrival position for the focused sound source. For example, the participants responded with 6 when they felt that the focused sound source was at position 2, passed position 9, and arrived at position 6. The participants practiced before starting the experiment.

3) Results: Figs. 4 and 5 show the experimental results. The horizontal and vertical axes represent the reproduced

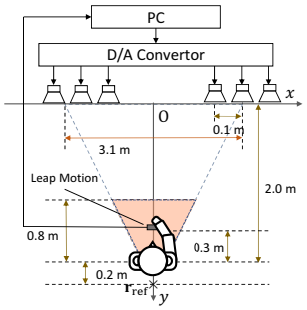


Fig. 2. Layouts of experiments A (basic localization w/o hand motion) and B (manipulating focused sound source with hand). The colored range indicates moving range of the focused sound source.

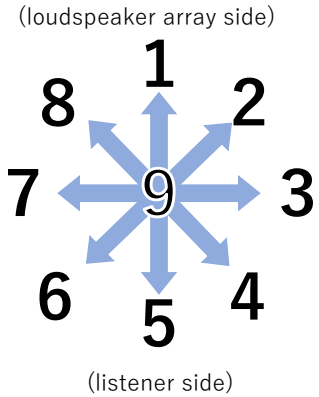


Fig. 3. Numbers indicating the positions and directions of the movement of the focused sound source for experiment A (basic localization w/o hand motion).

position of the focused sound source and listener's response, respectively. The size of each circle represents the number of responses.

According to the experimental results, we found that a moving focused sound source exhibited better sound localization than a static sound image. In addition, the participants could perceive the left and right directions for stationary focused sound sources and the front, back, left, and right directions for moving focused sound sources. Based on these results, we conclude that our system is suitable for generating focused sound sources.

B. Experiment B: directly manipulating focused sound source with hand

Next, we conducted an experiment to investigate whether the direct manipulation of the focused sound source by hand enhances sound localization and whether it is enjoyable.

1) Implementation: Fig. 2 shows the positional relationship between the experimental system and participant. In this system, there is a total delay of approximately 200 ms, which is approximately 100 ms for frame processing and approximately 100 ms for others, until the focused sound source is generated at the hand position acquired using Leap Motion. Other conditions are the same as considered in the previous experiments.

2) Procedure: The participants in the experiment were 10 people aged 22–29 years. They answered the following two questionnaires for each trial:

1) Is the sound image localized in a clear position?
2) Have you enjoyed?

Participants responded to these questions on a five-point Likert scale. Choices were as follows: 5 is "agree," 4 is "somewhat agree," 3 is "cannot say either way," 2 is "somewhat disagree," and 1 is "disagree."

Table I lists the control methods for sound images. A, B, C, and D denote the control methods for investigating whether sound localization can be improved by manually manipulating the focused sound source position. E and F are the control methods for investigating whether sound localization can be improved even if the sound volume changes with stereo playback instead of a focused sound source. Fig. 6 shows a subject conducting the experiment with 32-channel linear loudspeaker array.

In the case of the sound image that moved automatically, the participants listened only, and in other cases, the participants could move the sound image following hand movements using motion capture. There are four trajectories for moving the sound image: closer or farther from the face, left and right, clockwise, and counterclockwise. We used a display to instruct the participants whether the trial was automatic, closer or farther, left–right, clockwise, or counterclockwise for each

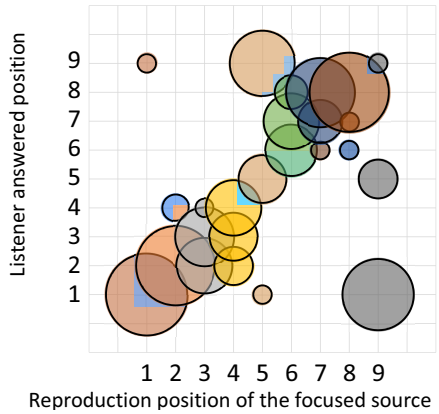


Fig. 4. Relation between the reproduced position and response of the stationary focused sound source for experiment A (basic localization w/o hand motion).

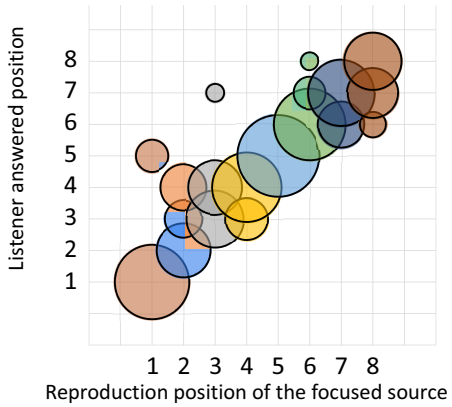


Fig. 5. Relation between the reproduced position and response of the moving focused sound source for experiment A (basic localization w/o hand motion).

TABLE I
SOUND IMAGE CONTROL METHOD FOR EXPERIMENT B.

A	Auto-moving focused sound source
B	Focused sound source following hand position
C	Focused sound source following hand position with a delay of 0.7 s
D	Randomly moving focused sound source
E	Auto-moving stereo sound source
F	Stereo sound source following hand position

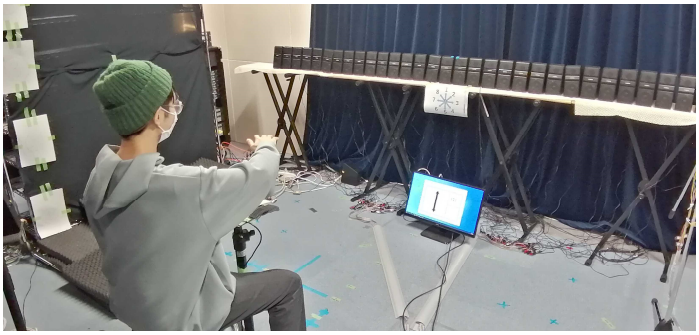


Fig. 6. Subject conducting the experiment with 32-channel linear loudspeaker array.

trial. The sound image moved slowly within a radius of 30 cm and was centered on the motion capture. For the auto-moving focused sound source (control method A), the focused sound source trajectory was fixed. The participants practiced in advance so that the operation would be similar to an automatic operation.

The participants listened to 48 sound sources consisting of 24 combinations of six types of sound-image control methods and four types of trajectories. The order in which participants listened to the sound sources was randomized. Participants were given 10 s to answer each time they listened to one stimulus and allowed to take a 1-min break after completing 24 trials.

The stereo reproduction output signal was generated by multiplying the output obtained based on the sine law by six times the gain obtained using (3). This output signal was reproduced using two left and two right loudspeakers near an equilateral triangle, including the listening position.

3) Results: Figs. 7 and 8 show the answers to questionnaires 1 and 2 of the sound image control methods, respectively. The bar graph represents the mean of the responses, and the error bar represents the standard deviation.

Fig. 7 shows that among the six control methods, B, in which the focused sound source follows the position of the hand, exhibits the highest average evaluation value, with a small number of low evaluation responses. By comparing B and F, A and E, respectively, we found that the focused sound source was evaluated to be higher than the stereo sound source. By comparing B, C, and D, we found that the evaluation decreased slightly when the focused sound source was followed and decreased significantly when the focused sound source moved randomly. Furthermore, by comparing C and F, we found that the localization of the focused sound source with a relatively long delay was better than that of the stereo sound source following the hand.

As shown in Figs. 7 and 8, the relationship between the sound image control method and enjoyment exhibited the same tendency as the relationship between the sound image control method and sound localization.

Based on the aforementioned results, we found that the focused sound source that follows the hand possessed the most localized experience and was enjoyable among all the sound image control methods. Moreover, the focused sound source was as enjoyable as the stereo sound source if there was a delay. Therefore, it is important to process minimal delays to ensure enjoyment.

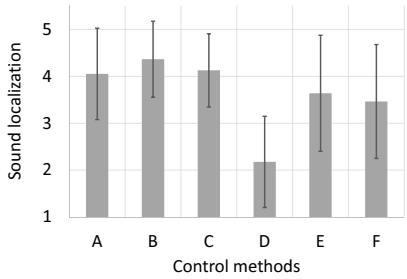


Fig. 7. Relationship between the sound image control method and sound localization for experiment B.

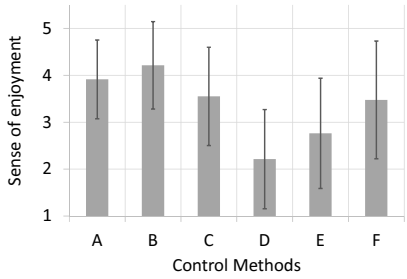


Fig. 8. Relationship between the sound image control method and enjoyment for experiment B.

C. Experiment C: pinching and moving focused sound source

The results of experiment B (directly manipulating focused sound source with hand) indicated that moving the focused sound source improved sound localization and enjoyment. Finally, we used a system that generates a focused sound source that can be moved by pinching the hand and conducted an experiment to investigate whether there was any experience of moving the focused sound source.

1) Implements: Fig. 9 represents the positional relationship between the system and participants. The previous experimental conditions were considered, and the focused sound source generation area was placed closer to the linear loudspeaker array.

In this experiment C, the focused sound source follows the hand only when the thumb and index fingers are pinched. Leap Motion can detect the shape of the user's hand in detail. Therefore, we determined whether the fingers were pinched by measuring the distance between the tip of the thumb and

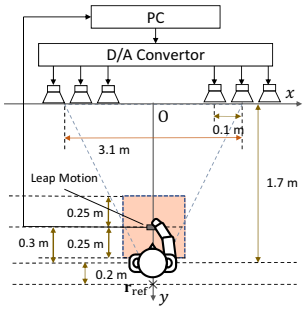


Fig. 9. Positional relationship between the system and participants of experiment (pinching and moving focused sound source). The colored range is the moving range of the sound image.

index finger.

2) Procedure: Seven people aged 22-26 years participated in the experiment. We instructed the participants to sit in the listening position, pinch, and move the focused sound source after reproducing white noise. We instructed the participants to move the focused sound source closer to them, farther from them, left, and right and move the focused sound source away from them at least once during the movement. The moving speed of the focused sound source was not specified. The participants answered the following questionnaire after each trial:

1) Did you feel like you were moving something?

Participants responded to these questions on a five-point Likert scale. Choices were as follows: 5 is "agree," 4 is "somewhat agree," 3 is "cannot say either way," 2 is "somewhat disagree," and 1 is "disagree." In addition, the participants freely wrote their impressions of the experiment after all trials finished.

Table II lists the focused sound source control methods, which control the source movement velocity and sound volume to simulate the weight. β was introduced to control the movement of the focused sound source in terms of moving ahead or delaying the position of the hand. If β is larger than 1.0, the focused sound source position moves more than the actual hand movement; if it is smaller than 1, the focused sound source position moves less than the actual hand movement. This processing was implemented by applying pseudo-haptics to the sound, in which a pseudo-tactile sense of resistance was felt by slowing the movement of the cursor [17]. Figs. 10 and 11 show the changes in the positions of the focused sound source.

For example, the focused sound source moved, as shown in Fig. 10 when $\beta = 2.0$. In this case, the sound image was localized in front of the hand position; therefore, it moved faster than the hand movement. Meanwhile, the focused sound source also moved, as shown in Fig. 11 when $\beta = 0.5$. In this case, the sound image is localized behind the position of the hand; therefore, it moves slower than the movement of the hand.

Sound volume change processing enhances and reduces the sound volume as the hand moves faster and slower, respectively. The upper and lower limits were set for the sound volume, and the maximum change was approximately 6 dB. The results of the previous study relate to loud sound with rough textures, and soft sound with smooth textures [18]. Therefore, we assume that loudness while moving an object is related to the weight of the object. Moreover, in the real world, fricative sounds tend to become louder as the moving speed increases. Therefore, to this experimental system, we add a process that changes the sound volume according to the moving speed of the focused sound source.

TABLE II
FOCUSED SOUND SOURCE CONTROL METHOD FOR EXPERIMENT C

Sound volume change	Position of focused sound source			
	ahead ($\beta = 2.0$)	follow ($\beta = 1.0$)	delay ($\beta = 0.5$)	stationary
Yes	A	C	E	G
No	B	D	F	H

3) Results: Fig. 12 shows the answers to questionnaire 1 for each focused sound source control method. The bar graph represents the mean of the responses and the error bar represents the standard deviation. As shown in Fig. 12, it is possible to present the sense of movement by moving the focused sound source according to the hand motion. In addition, we compared

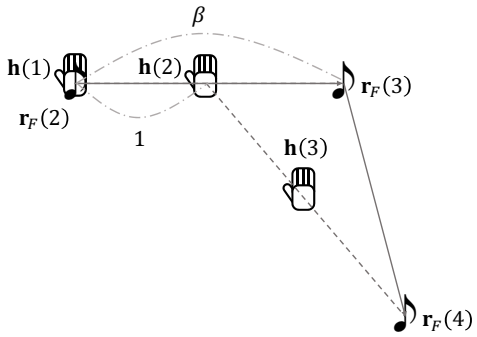


Fig. 10. Focused sound source moves ahead of the hand for experiment C ($\beta = 2.0$).

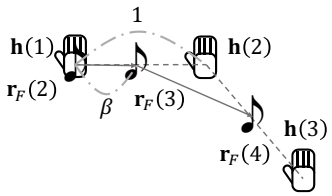


Fig. 11. Focused sound source delays the hand for experiment C ($\beta = 0.5$).

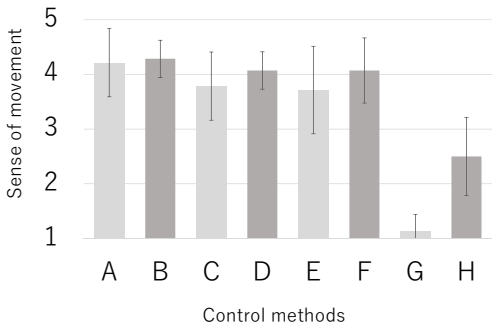


Fig. 12. The relationship between the control method of the focused sound source and the sense of movement for experiment C.

A and B, C and D, E and F, and G and H, which possessed the same moving speed but different sound volumes (changes in sound volumes). In all the combinations, when the sound volume of the focused sound source was changed according to the hand movement, the sense of movement was higher than when it was not changed. Therefore, it implies that changing the sound volume of the focused sound source according to the hand movement tends to increase the sense of movement.

There were opinions on the impression of the free description after the experiment. "The movement that followed the hand was interesting, and I got the impression that an object was there," "It felt like I was remotely controlling it rather than moving it with the hand." Based on these opinions, it may be possible to present sensations other than auditory by manipulating the focused sound source, although there are individual differences in perception.

IV. Conclusion

Herein, we investigated the experience of listeners when they interactively change their mid-air sound source position. We created a system that generates a focused sound source whose position can be manually manipulated and improved it so that the focused sound source can be pinched by the hand. The experimental results indicated that sound localization and the sense of enjoyment can be improved by moving the focused sound source as desired. Moreover, we found that by changing the sound volume of the focused sound source according to the movement speed, the sense of movement tended to increase. In the future, we plan to study methods for generating interactively focused sound sources.

References

[1] G. Kailas and N. Tiwari, "Design for immersive experience: Role of spatial audio in extended reality applications," in *Design for Tomorrow —Volume 2: Proceedings of ICoRD 2021*. Springer, 2021, pp. 853–863.

[2] S. Poeschl, K. Wall, and N. Doering, "Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence," in *2013 IEEE Virtual Reality (VR)*. IEEE, 2013, pp. 129–130.

[3] A. C. Kern and W. Ellermeier, "Audio in vr: Effects of a soundscape and movement-triggered step sounds on presence," *Frontiers in Robotics and AI*, vol. 7, p. 20, 2020.

[4] A. Tajadura-Jiménez, J. Newbold, L. Zhang, P. Rick, and N. Bianchi-Berthouze, "As light as you aspire to be: Changing body perception with sound to support physical activity," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[5] N. Schaffert, T. B. Janzen, K. Mattes, and M. H. Thaut, "A review on the relationship between sound and movement in sports and rehabilitation," *Frontiers in psychology*, vol. 10, p. 244, 2019.

[6] S. Gelineck and D. Korsgaard, "An exploratory evaluation of user interfaces for 3d audio mixing," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[7] M. Gasparini, S. Cecchi, L. Romoli, A. Primavera, P. Peretti, and F. Piazza, "Kinect application for a wave field synthesis-based reproduction system," in *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.

[8] Y. H. Kim and M. H. Song, "Design and implementation of "spatial equalizer"," in *Audio Engineering Society Conference: 43rd International Conference: Audio for Wirelessly Networked Personal Devices*. Audio Engineering Society, 2011.

[9] J. Tang, "Research on the application of sound in virtual reality," *Highlights in Science, Engineering and Technology*, vol. 44, pp. 206–212, 2023.

[10] C. Valzolgher, M. Alzhaler, E. Gessa, M. Todeschini, P. Nieto, G. Verdelet, R. Salemme, V. Gaveau, M. Marx, E. Truy *et al.*, "The impact of a visual spatial frame on real sound-source localization in virtual reality," *Current Research in Behavioral Sciences*, vol. 1, p. 100003, 2020.

[11] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[12] S. Spors, H. Wierstorf, M. Geier, and J. Ahrens, "Physical and perceptual properties of focused virtual sources in wave field synthesis," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[13] ultraleap, "Ultraleap: Digital worlds that feel human," https://www.ultraleap.com/, (accessed May. 2023).

[14] J. Müller, M. Geier, C. Dicke, and S. Spors, "The boomroom: mid-air direct interaction with virtual sound sources," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 247–256.

[15] V. Valimaki and T. I. Laakso, "Principles of fractional delay filters," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 6. IEEE, 2000, pp. 3870–3873.

[16] J. S. Lim and A. V. Oppenheim, *Advanced Topics in Signal Processing*. Prentice-Hall, Inc., 1988.

[17] A. Lécuyer, S. Coquillart, A. Kheddar, P. Richard, and P. Coiffet, "Pseudo-haptic feedback: can isometric input devices simulate force feedback?" in *Proceedings IEEE Virtual Reality 2000 (Cat. No. 00CB37048)*. IEEE, 2000, pp. 83–90.

[18] D. Peeva, B. Baird, O. Izmirli, and D. Blevins, "Haptic and sound correlations: Pitch, loudness and texture," in *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.* IEEE, 2004, pp. 659–664.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Time Sensitive Hash and Adaptive Image Recovery based Self-embedding Fragile Watermarking Scheme in Encrypted Images

Xin Wang† and Hongjie He†* and Fan Chen‡
† School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
‡ School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
Corresponding email: hjhe@swjtu.edu.cn

***Abstract*—To protect the integrity and privacy of cloud storage images, this paper proposed a self-embedding fragile watermarking scheme in encrypted images. An original image is divided into the seven least significant bit-planes (7LSBs) and the most significant bit-plane (MSB). 7LSBs plane is encrypted using pixel XOR and block scrambling. MSB plane is predicted and compressed for encoding to vacate space for reversible embedding of watermark. A time sensitive hash (TSH) method is designed to resist the collage attack. The TSH enables the hash function to generate a unique hash value under the same input, which greatly reduces the probability of missed detection under collage attack. The adaptive image recovery (AIR) method is also proposed to improve the quality of image recovery. The effectiveness of AIR is achieved by adaptively selecting the recovery method based on the estimated tamper attack type. Experimental results show that the proposed scheme effectively protects the confidentiality of image content and exhibits a strong recovery effect against various tamper attacks, even achieving lossless recovery without tamper. Compared with the existing schemes, the proposed scheme has certain effectiveness and superiority.**

I. Introduction

With the rapid development and increasing popularity of cloud storage, more and more people are inclined to store digital images in the cloud due to its advantages such as high storage efficiency, low cost, and intelligent management. However, cloud storage also entails a series of security issues, such as data tampering [1], data loss, and data leakage [2]. These security issues may arise from cloud storage service providers, hacker attacks, internal personnel and other factors. According to a survey conducted by the data security company Varonis, the number of data leakage incidents in 2020 reached 3,950, an increase of 41% compared to 2019 [3]. Therefore, ensuring the integrity and privacy protection of digital images during the process of using cloud storage services has significant social implications and broad application prospects.

Self-embedding fragile watermark is one of the effective methods for protecting the integrity of images. In [4], Zhang et al. proposed a reference sharing mechanism (RSM) to achieve perfect restoration of watermarked images within a certain tamper range. In order to obtain the best recovery performance at different tamper rates, Qin et al. [5] proposed an embedding mode adaptive selection scheme based on Zhang's work. Compared with [4], it has better flexibility and higher quality restored images. However, these schemes [4], [5] cannot resist collage attacks. [6], [7] introduced the discrete cosine transform (DCT) and integer wavelet transform (IWT) into fragile watermark, but the recovery quality of these schemes is relatively low. In addition, [8]–[12] proposed fragile watermarking schemes based on block truncation coding (BTC) and absolute moment block truncation coding (AMBTC). These methods generated the recovery watermark by compressing the image information in a lossy manner. Compared to methods that use block mean value as recovery watermark, the recovery quality for images with high texture complexity is better.

However, the above fragile watermark scheme can only guarantee the integrity of the image, but not the privacy of the image. In addition, in order to ensure the imperceptibility of the watermark, researchers usually replace the LSB of the original image with the watermark, so that the process of generating the watermarked image is irreversible. This has great limitations in military and medical applications. Therefore, fragile watermarking schemes for encrypted images have emerged. Jiang et al. [13] proposed a high-capacity fragile watermark scheme with probabilistic encryption. Although this scheme can achieve pixel-level tamper detection and localization, it has a high time complexity and cannot recover tampered images. Huang et al. [14] proposed a multi-level authentication scheme for cloud storage, which embeds watermark in the redundant space after image POB compression. Although this scheme can still recover images despite large-scale tamper, due to the introduction of POB compression method, the time complexity of this scheme is relatively high. In addition, for each image uploaded to the cloud, users need to save the number of binary digits '1' before compressing each pixel, which requires additional storage space. We call this phenomenon data expansion.

The above mentioned issues are summarized: 1) Fragile watermarking schemes in plaintext images cannot protect the privacy of the image and the process of watermark embedding is irreversible 2) Some fragile watermarking schemes cannot resist collage attack. 3) The existing fragile watermarking schemes in the encrypted images have high time complexity and data expansion.

In this paper, based on the idea of RDH-EI [15]–[17], we propose a self-embedding fragile watermarking scheme in the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

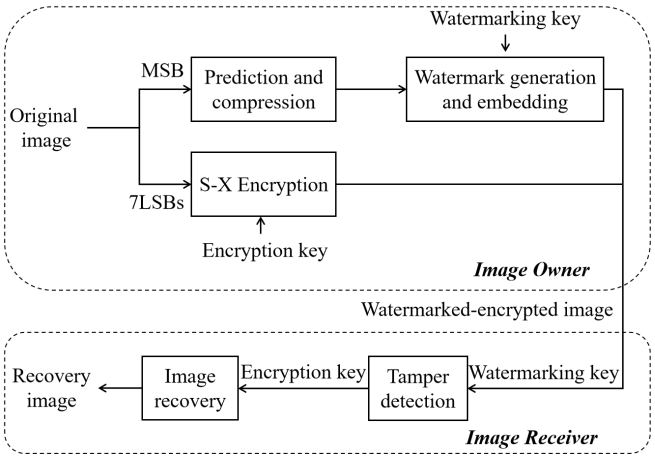


Fig. 1. Framework of the proposed scheme.

encrypted images to achieves high tamper detection performance and good recovery image quality at a low time cost. The contributions of this paper are summarized as follows.

1) A self-embedding fragile watermarking scheme in encrypted images is proposed to simultaneously solve the integrity and privacy of the images. This scheme has low time complexity and no data expansion.

2) Design a time sensitive hash (TSH) method to resist collage attack. This is achieved by combining the timestamp with the image block index as the unique hash key of the hash function. Experimental results show that the average is reduced by more than 0.9 compared with [4] under the collage attack.

3) Design an adaptive image recovery (AIR) method to improve recovery image quality. This is achieved by adaptively selecting the image recovery method according to the estimated tamper attack type. It greatly improves the quality of the recovery image under random noise attack.

The rest of this paper is organized as follows. The proposed scheme is described in detail in the section II. The section III carries on the experimental analysis and comparison. See section IV for conclusion.

II. Proposed scheme

As shown in Fig 1, the proposed self-embedding fragile watermarking scheme in encrypted images includes two participants: image owner and image receiver. The image owner encrypts the 7LSBs through the encryption key (Key_1) and embeds encrypted watermark in the redundant space vacated after MSB prediction and compression using the watermarking key (Key_2). Finally, upload the generated watermarked-encrypted image to the cloud. The image receiver, who owns Key_1 and Key_2, can download watermarked-encrypted image from the cloud, extract watermark for tamper detection and image recovery.

A. S-X Encryption

An original image $I = \{I_k | k = 1, 2, ..., 8\}$ of size $M \times N$ contains 8 bit-planes. The higher the bit-plane of an image, the stronger the correlation between adjacent pixels, so we split it into MSB I_{msb} and 7LSBs I_{lsb}. Then, we divided it into 8×8 non-overlapping image blocks $B = \{B_n | n = 1, 2, \ldots, N_B\}$, where $N_B = (M/8) \times (N/8)$. According to the Key_1, we pseudo-randomly permute all blocks in the 7LSBs and then perform a bitwise XOR operation on each pixel within the block.

B. Prediction and compression

In order to make more images can vacate enough space to embed the watermark, we designed a MAP prediction method. The MAP optimizes compression encoding performance by changing the number of '1' in the prediction bit-plane. The MAP consists of two prediction methods: MSB median prediction (MMP) and MSB flip prediction (MIP),

$$P = \begin{cases} MMP(I_{msb}), & if \quad L_{MMP} \leq Th \\ MIP(I), & if \quad L_{MMP} > Th \end{cases} \tag{1}$$

among them

$$Th = 22 \times N_B - 10 \tag{2}$$

where P represents the predicted bit-plane, L_{MMP} is the length of the prediction compression code C after using MMP, Th is a threshold value, it represents the maximum length of the original data when using RSM. MMP is an prediction method based on median prediction, which takes the bit-plane as the object. MIP is a non-iterative flip prediction method that designed based on iterative MSB-inversion prediction [16].

All images prioritize using MMP prediction because when the image is tampered, only MSB using RSM can be restored losslessly, while MIP needs to use eight planes of the image to restore the prediction, resulting in image distortion. However, after using MMP in the MSB of complex textured images, the number of '1' in the predicted bit-plane is still too large, making it difficult for compression encoding to vacate enough space to embed watermark. At this time, MIP should be utilized. Finally, adaptive compression coding [15] is used to obtain the C.

C. Watermark generation and embedding

The watermark mainly consists of three parts: recovery bit, anti-collage bit, and authentication bit. For each image block , we allocate 33 bits recovery bit, 1 bit anti-collage bit, and 30 bits authentication bit.

1) Recovery bit: In order to lossless recovery of the prediction compression code, we use the RSM to generate the recovery bit. Two reference sharing parameters α, β are chosen, where $\alpha = 256$ and $\beta = 128$. The generation process of the recovery bit is shown in Fig 2. Firstly, we divide original data into $S = \lceil (L + 10)/\alpha \rceil$ subjects $C_1, C_2, \ldots, C_S$, each group contains α bits. L is the code length of C. The x th subset C_x are denoted by $c_{x,1}, c_{x,2}, \ldots c_{x,\alpha}$, $x = 1, 2, \ldots S$. In particular, the first 10 bits of the subset C_1 are used to store S, and if the length of the subset C_S is not equal to α, the rest space will be filled with '0'. Then, equation (3) is used

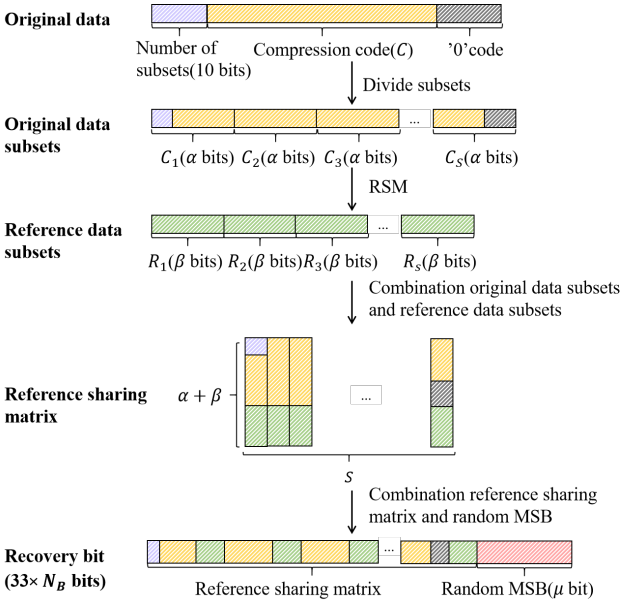


Fig. 2. Recovery bit generation.

to convert the α bits original data of subset C_x into the β bits reference data of subset R_x, $_{x,1}, r_{x,2}, \ldots r_{x,\beta}$,

$$\begin{bmatrix} r_{x,1} \\ r_{x,2} \\ \vdots \\ r_{x,\beta} \end{bmatrix} = A_x \begin{bmatrix} c_{x,1} \\ c_{x,2} \\ \vdots \\ c_{x,\alpha} \end{bmatrix} \tag{3}$$

where A_x is a pseudo-random binary matrix of size $\beta \times \alpha$ generated by the key. Then, the original data and reference data of each subset are concatenated by column. Finally, for the μ bits space remaining after using the RSM, we filling it by R_P randomly selected from the MSB to obtain the recovery bitstream C_R. C_R is scramble and XOR encryption based on Key_2 and divided into N_B groups, distribution of each image block 33 bits.

2) Anti-collage bit: To resist the collage attack, we develop the TSH method, which uses the timestamp as the random number seed to generate the anti-collage key $Ckey$. $Ckey$ is converted into a 50 bits binary sequence and concatenated repeatedly. Then, N_B bits are intercepted from sequence to generate the anti-collage bitstream C_C. C_C is scramble and XOR encrypted by Key_2 and divided into N_B groups, distribution of each image block 1 bit.

3) Authentication bit: For each image block B_n, 33 bits of recovery bit and 1 bit of anti-collage bit are concatenated, and the rest of 30 bits are filled with '0' to obtain the watermark sequence. The watermark sequence is combined with the corresponding encrypted 7LSBs block as the input of the hash function. The hash key $Hkey$ is formed by combining the $Ckey$ with the block index. We use the first 30 bits of the hash function output as authentication bit W_A, W_A is scramble and XOR encryption according to the Key_2.

$$Hkey = Ckey + n \tag{4}$$

Finally, The authentication bit replaces the 30 bits '0' of the watermark sequence in order to generate the watermarked MSB. The watermarked-encrypted image I_w is obtained by combining the watermarked MSB with the encrypted 7LSBs.

D. Tamper detection

After downloading image I_a from the cloud, the image receiver performs tamper detection on it through authentication bit, and estimates the type of tamper attack through anti-collage bit.

Firstly, the image I_a is divided into 8×8 non-overlapping image blocks. The anti-collage bits are extracted from the 34th bit of each image block in MSB. We splice it in sequence and decrypt it with Key_2 to obtain the anti-collage bitstream $C_C{}'$. Since C_C is formed by repeated splicing of $Ckey$, the sequences of each group (except the last group) grouped by $C_C{}'$ should be equal. If not, the image is tampered. In this process, we can extract the correct $Ckey$ and calculate the error proportion of the $C_C{}'$ according to equation (5),

$$ep^{C_C} = \sigma^{C_C}/N_B \tag{5}$$

Where ep^{C_C} represents the proportion of the error bit in $C_C{}'$ and σ^{C_C} represents the number of the error bit in $C_C{}'$.

Subsequently, when $Hkey$ is obtained by combining $Ckey$ and block index, the authentication bits $W_A{}'$ are regenerated by the same steps. The 35th to 64th bit of B_n in the MSB are extracted to obtain the authentication bits $W_A{}''$. If $W_A{}' = W_A{}''$, the block is a complete block, otherwise the block is a tampered block. In this way, the tamper localization $flag_1$ of the encrypted image can be obtained. After using Key_1 to recover the scrambled $flag_1$, we can also obtain the tamper localization $flag_2$ of the plaintext image.

Finally, the type of tamper attack can be estimated by using the proportion of tampered pixels ep^{flag} in $flag_1$ or $flag_2$ and ep^{C_C} through equation (6),

$$\theta = ep^{flag} - 100 \times ep^{C_C} \tag{6}$$

If $\theta > 0$, the type of tamper attack is random noise attack, otherwise, it is regional attack. We will perform the secondary tamper localization of the image to reduce the number of misjudged pixels under random noise attack. The decrypted 7LSBs plane is divided into 2×2 blocks, and the principle of correlation between pixels is used to get tamper localization $flag_3$ of random noise attack.

E. Image recovery

Image recovery can be divided into two situations: image not tampered and image tampered. If I_a is not tampered, the image receiver decrypts the 7LSBs using Key_1, then decrypts the $C_R{}'$ using Key_2 and restores the compression and prediction of the C' to obtain the MSB. The recovery image I_r is obtained by recombining the MSB and 7LSBs.

If I_a is tampered, determine whether S is tampered according to $C_R{}'$. If S is tampered, the first subset $C_1{}'$ is extracted, and the complete reference data is used to recover the tampered original data to obtain the correct S. Then, We use the same steps to recover tampered original data in all subsets. Finally, the prediction compression code C' is obtained by concatenating all the original data in sequence. The recovery equation is shown in (7),

$$\begin{bmatrix} r_{x,1} \\ r_{x,2} \\ \vdots \\ r_{x,\beta'} \end{bmatrix} - A_x^{(E,R)} \cdot C^{(R)} = A_x^{(E,T)} \cdot C^{(T)} \tag{7}$$

where $C^{(R)}$ and $C^{(T)}$ denote the complete original data and the tampered original data, respectively. $A_x^{(E,R)}$ and $A_x^{(E,T)}$ are the corresponding columns of $C^{(R)}$ and $C^{(T)}$ in $A_x^{(E)}$, $A_x^{(E)}$ is a matrix that extracts the rows corresponding to the complete reference data from A_x. We use Gaussian elimination method to solve. If there is a unique solution, we can recover the tampered original data in the subset. If the solution is not unique or there is no solution, the recovery fails.

Algorithm 1 Adaptive image recovery

Input: C' and Key_1
Output: I_r

(1) S-X decryption of the 7LSBs using the Key_1
(2) Decompress C' according to [15]
(3) Obtain the I_r
switch Image recovery situation **do**
  case MMP/MIP + regional attack + C' Recovery failure
    The complete bits from R_P are selected to generate the $I_{msb}{}'$ according to the $flag_2$;
    $I_r = I_{msb}{}' \times 2^7 + I_{lsb}{}'$;
  case MMP + regional attack + C' Recovery success
    $I_{msb}{}'$ is obtained by restoring MMP;
    $I_r = I_{msb}{}' \times 2^7 + I_{lsb}{}'$;
    Do domain recovery according to $flag_2$;
  case MIP + regional attack + C' Recovery success
    I_r is obtained by restoring MIP;
    Do domain recovery according to $flag_2$;
  case MMP/MIP + random nose attack + C' Recovery failure
    The complete bits from R_P are selected to generate the $I_{msb}{}'$ according to the $flag_3$;
    $I_r = I_{msb}{}' \times 2^7 + I_{lsb}{}'$;
  case MMP + random nose attack + C' Recovery success
    $I_{msb}{}'$ is obtained by restoring MMP;
    $I_r = I_{msb}{}' \times 2^7 + I_{lsb}{}'$;
    Do domain recovery according to $flag_3$;
  case MIP + random nose attack + C' Recovery success
    I_r is obtained by restoring MIP;
    Do domain recovery according to $flag_3$;
end
return I_r.

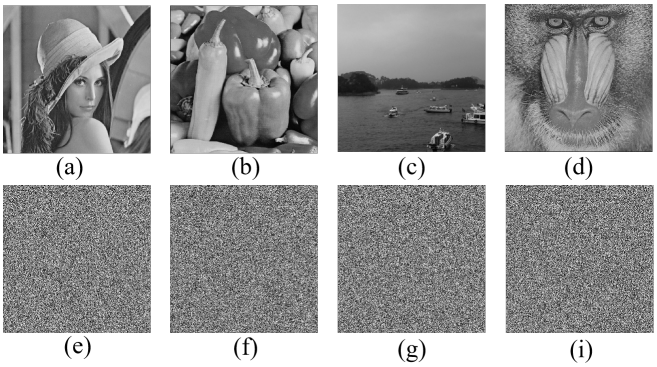


Fig. 3. The selected test images. (a) Lena, (b) Pepper, (c) Lake, (d) Baboon, (e)-(i) are the watermarked-encrypted image for (a)-(d), respectively.

Finally, the AIR method is selected adaptively according to the prediction method, the estimated tamper attack type and the recovery situation of C', so as to achieve the best recovery quality. The details are shown in Algorithm 1.

III. Experimental results

This section mainly evaluates the proposed scheme from four aspects: watermarked-encrypted image security, tamper detection, image recovery and time cost. Fig.3 (a)-(d) show four standard test images of size 512×512 as original images, Lena, Pepper, Lake, Baboon. After embedding the watermark, their corresponding watermarked-encrypted images are shown in (e)-(i). From the visual effect, the human eye cannot recognize the original image content.

A. Security analysis

1) Key space: It is known that the security of a cryptosystem depends solely on the key. In this scheme, we use Key_1 and Key_2 to encrypt the 7LSBs and C, respectively. Then the probability that an attacker correctly gets MSB for all pixel without knowing the key is $1/((1/L!) \times (1/2)^L)$. For 7LSBs, if the brute force method is applied, the probability to correctly obtain 7LSBs is $1/((1/N_B!) \times (1/2)^{7\times M\times N})$. Therefore, for an image of size $M \times N$, the probability of the attacker successfully recovering to the original image is $1/((1/L!) \times (1/2)^L \times (1/N_B!) \times (1/2)^{7\times M\times N})$.

2) Pixel adjacent correlation: A good image encryption algorithm should be able to reduce the correlation of adjacent pixels, as far as possible to achieve zero correlation. We randomly selected 1000 pixels from the original image and the watermarked-encrypted image from the horizontal, vertical and diagonal directions for correlation analysis. table I shows the average value of 100 results. It can be seen from table I that the original image has a large correlation in three directions, while the adjacent correlation of the watermarked-encrypted image is almost zero, indicating that the watermarked-encrypted image effectively hides the correlation between adjacent pixels.

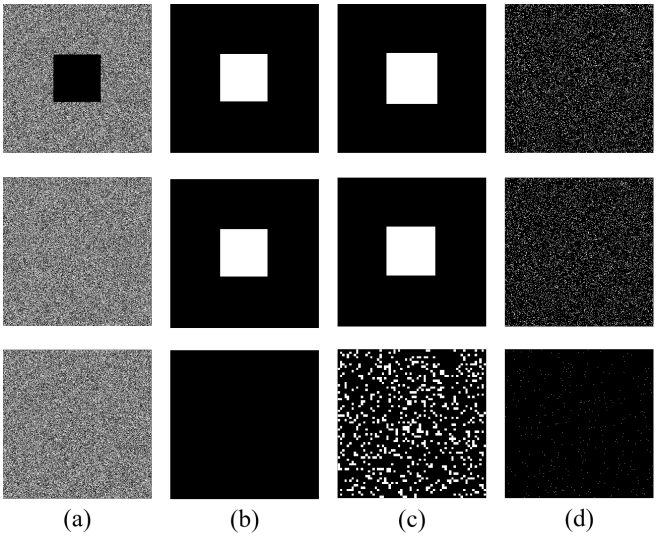


Fig. 4. Tamper detection of Lena under different types of attacks. (a) $\gamma = 10\%$ cropping attack, $\gamma = 10\%$ college attack and $\rho = 1000$ random noise attack from top to bottom, (b) true tamper location, (c) tamper detection localization of encrypted images, (d) tamper detection localization of plaintext images.

B. Tamper detection analysis

For tamper detection, we use the probability of false rejection (PFR), probability of false acceptance (PFA), and probability of false decision (PFD) to evaluate the detection performance of the watermark,

$$PFR = \frac{FT}{M \times N} \tag{8}$$

$$PFA = \frac{TF}{M \times N - \sigma^{flag}} \tag{9}$$

$$PFD = \frac{\sigma^{flag}}{M \times N} \times PFA + (1 - \frac{\sigma^{flag}}{M \times N}) \times PFR \tag{10}$$

Where TF represents the number of complete pixels judged as tampered pixels, FT represents the number of tampered pixels judged as complete pixels, σ^{flag} represents the number of the tampered pixel in $flag_1$ or $flag_2$.

As shown in Fig 4, taking the Lena image as an example. Column (a) shows the tampered image under the cropping attack with $\gamma = 10\%$ in the center region, the collage attack with $\gamma = 10\%$ in the center region and the random noise attack with $\rho = 1000$ from top to bottom, respectively. Column (b) is the true tamper location in encrypted images. Column (c) shows the tamper detection localization results of encrypted images. Column (d) shows the tamper detection localization results of plaintext images. γ represents the tampered range, rho represents the number of tampered pixels. The detailed detection results of the four images are shown in table II. As can be seen from table II, the average PFD is all less than 2% under the cropping attack and collage attack. Under the collage attack, the average PFA is 0.001, this means that the probability of missed detection is very low. However, the accuracy under random noise attack is relatively low.

TABLE I
CORRELATION OF IMAGES

Direction	Image type	Lena	Pepper	Lake	Baboon
Horizontal	Original	0.9697	0.9730	0.9907	0.8649
	Watermarked	0.0047	-0.0022	-0.0022	-0.0013
Vertical	Original	0.9843	0.9762	0.9780	0.7533
	Watermarked	-0.0004	-0.0008	-0.0078	0.0019
Diagonal	Original	0.9562	0.9575	0.9719	0.7197
	Watermarked	-0.0030	-0.0064	-0.0003	0.0007



Fig. 5. Recovery image of Lena under different types of attacks. (a) $\gamma = 10\%$ cropping attack, (b) $\gamma = 10\%$ college attack, (c) $\rho = 1000$ random noise attack.

Table III shows the results of PFR, PFA and PFD under collage attack with different γ for proposed scheme, [4] and [14]. The proposed scheme and [14] are both able to resist collage attack. Compared with the [4], which also uses hash function to generate authentication watermark, the average PFA of the proposed scheme is less than 1%. Although under a large-scale collage attack, the proposed scheme can still accurately detect the tampered area, but [4] have a high error detection rate. Therefore, TSH method performs well.

C. Image recovery analysis

In this subsection, we mainly adopt Peak Signal-to-Noise Ratio (PSNR) and Structural similarity (SSIM) to measure the recovery image quality. The recovery images corresponding to Fig 4 (a) are shown in Fig 5 (a)-(c). Table IV shows the PSNR and SSIM of the four original images and the recovery images corresponding to the above attacks. It can be seen from table IV, the recovery quality of baboon image is lower than other images. The texture complexity of Baboon image is too high, so MIP prediction is used. However, the peoposed scheme only uses compression codes on the MSB as the original data

TABLE II
PFR, PFA, PFD FOR TEST IMAGES AGAINST VARIOUS TAMPER ATTACKS

Attack		Lena	Pepper	Lake	Baboon
Cropping($\gamma = 10\%$)	PFR	0.0187	0.0187	0.0187	0.0184
	PFA	0	0	0	0
	PFD	0.0168	0.0168	0.0168	0.0168
Collage($\gamma = 10\%$)	PFR	0.0178	0.0175	0.0178	0.0182
	PFA	0.0012	0.0012	0.0012	0.0006
	PFD	0.0161	0.0158	0.0161	0.0164
Random($\rho = 1000$)	PFR	0.1609	0.1652	0.1698	0.1703
	PFA	0.2630	0.2360	0.2230	0.2140
	PFD	0.1613	0.1655	0.1700	0.1704

TABLE III
COMPARISON OF PFR, PFA AND PFD BETWEEN PROPOSED SCHEMES AND RELATED WORKS.

Tamper rate	[4]			[14]			Propose		
	PFR	PFA	PFD	PFR	PFA	PFD	PFR	PFA	PFD
$\gamma = 5\%$	0.0187	0.9635	0.1145	0.0014	0	0.0012	0.0177	0.0012	0.016
$\gamma = 10\%$	0.0246	0.9568	0.2111	0.0022	0	0.0018	0.0241	0.0009	0.0195
$\gamma = 15\%$	0.0217	0.937	0.2974	0.0031	0	0.0021	0.0217	0.0004	0.0152
$\gamma = 20\%$	0.0465	0.9695	0.4184	0.0042	0	0.0025	0.0255	0.0006	0.0157
$\gamma = 25\%$	0.028	0.9403	0.4866	0.0056	0	0.0028	0.0574	0.0009	0.03

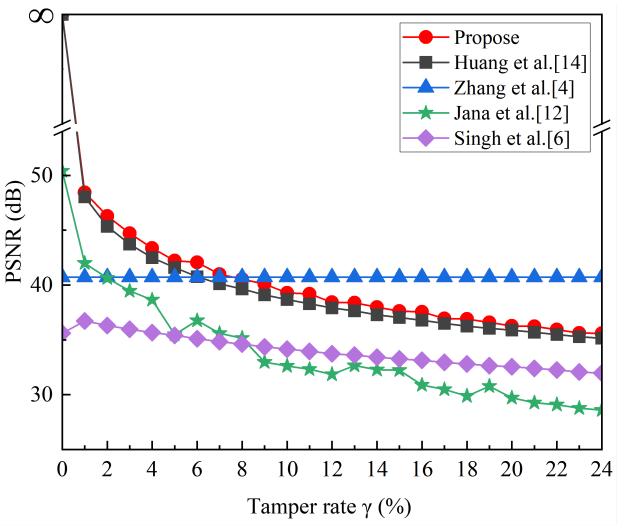


Fig. 6. PSNR of image recovery under cropping attack.

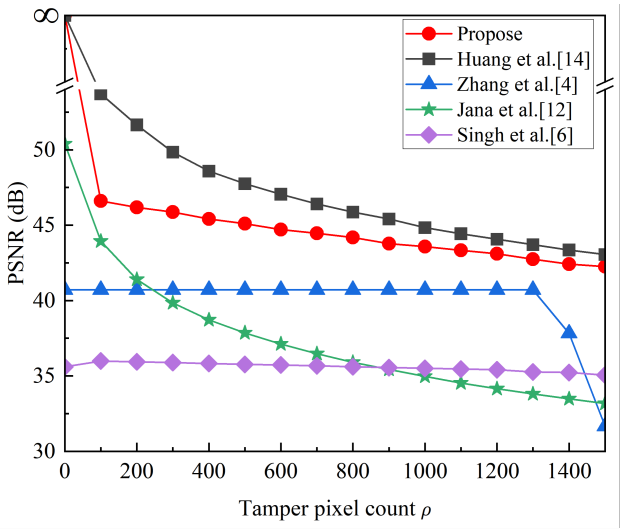


Fig. 7. PSNR of image recovery under random noise attack.

to generate recovery watermark, and 7LSBs is restored using domain complete pixels. This leads to failure of MIP prediction recovery when the image is tampered. However, the average PSNR of the remaining three images under cropping attack and collage attack reaches 41.18dB and 41.02dB, and the average SSIM is above 0.96, which indicates that the recovery effect is good. Especially for random noise attack, the average PSNR reaches 45.99dB and the average SSIM is above 0.99. Fig. 6 and Fig. 7 compare the average PSNR of the recovery image and the original image of 50 randomly selected UCID images [18] under cropping attack and random noise attack. When the image is not tampered, except the proposed scheme and [14], other schemes cannot achieve lossless recovery. This is because both schemes utilize the redundant space after compression to embed the watermark, rather than directly replacing the LSB of the image. Under the cropping attack, the proposed scheme outperforms all schemes when $\gamma < 8\%$. While , the PSNR of [4] is superior to the proposed scheme when $\gamma > 8\%$, this is because [4] uses the RSM to losslessly restore the 5MSBs of the image. Under the random noise attack, the PSNR of the proposed scheme is greater than 40dB, and the average PSNR is 44.25dB. This is better than the same 8×8 block size [4] and slightly lower than the 2×2 block size [14]. The results show that the AIR method is effective.

TABLE IV
PSNR, SSIM FOR TEST IMAGES AGAINST VARIOUS TAMPER ATTACKS

Direction		Lena	Pepper	Lake	Baboon
Cropping($\gamma = 10\%$)	PSNR	41.05	40.75	41.73	14.38
	SSIM	0.9653	0.9626	0.9589	0.5063
Collage($\gamma = 10\%$)	PSNR	40.95	40.57	41.53	14.49
	SSIM	0.9650	0.9626	0.9592	0.5558
Random($\rho = 1000$)	PSNR	47.52	46.05	44.39	25.00
	SSIM	0.9942	0.9923	0.9933	0.92427

D. Time analysis

We count the time required by the image owner and the image receiver in proposed scheme, respectively.It can be seen

TABLE V
COMPARISON OF RUNNING TIME BETWEEN PROPOSED SCHEMES AND RELATED WORKS

Tamper rate	Participant	Time(/s)			
		[14]	[4]	[12]	Propose
$\gamma = 5\%$	Image owner	114.89	18.54	16.29	2.63
	Image receiver	38.53	150.49	18.62	3.37
$\gamma = 10\%$	Image owner	115.89	18.07	16.26	2.62
	Image receiver	38.98	222.05	18.66	3.9
$\gamma = 15\%$	Image owner	115.63	18.86	16.33	2.65
	Image receiver	38.37	283.78	19.03	4.45
$\gamma = 20\%$	Image owner	115.24	18.75	16.21	2.64
	Image receiver	38.09	351.93	18.97	4.79
$\gamma = 25\%$	Image owner	115.24	18.06	16.22	2.65
	Image receiver	38.17	391.91	19.15	4.78

from table V that the average running time of the proposed scheme is the shortest. Compared with the [4], which also adopt the RSM, our scheme greatly reduces the time required by the image receiver. This is because [4] uses all bits of the 5MSBs as original data, while the proposed scheme only uses the MSB plane prediction compression code as original data, which greatly reduces the number of solutions to equation (7). Compared with scheme [14], which is also a fragile watermarking scheme in encrypted images, the time required for two participants is about 20 times slower than the proposed scheme. This is because the time complexity of the POB compression is high.

IV. Conclusion

To simultaneously protect the integrity and privacy of cloud storage images, we propose a self-embedding fragile watermarking scheme in encrypted images. In this scheme, the TSH method is designed to resist collage attacks by using timestamp to make the hash function generate unique output. Another contribution is the AIR method, which adaptively selects the best image recovery method for different types of tamper attacks. Furthermore, experimental results confirm that the proposed scheme achieves high tamper detection performance and good recovery image quality at a lower time cost. Future work will focus on increasing the tolerable tamper range of image using advanced image processing methods.

References

[1] Y. Yang, Y. Chen, F. Chen, and J. Chen, "An efficient identity-based provable data possession protocol with compressed cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1359–1371, 2022.

[2] A. Mohiyuddin, A. R. Javed, C. Chakraborty, M. Shabbir, and J. Nebhen, "Secure cloud storage for medical iot data using adaptive neuro-fuzzy inference system," *International Journal of Fuzzy Systems*, vol. 24, no. 2, pp. 1203–1215, 2022.

[3] P. Prajapati and P. Shah, "A review on secure data deduplication: Cloud storage security issue," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 3996–4007, 2022.

[4] X. Zhang, S. Wang, Z. Qian, and G. Feng, "Reference sharing mechanism for watermark self-embedding," *IEEE Transactions on Image Processing*, vol. 20, no. 2, pp. 485–495, 2011.

[5] C. Qin, H. Wang, X. Zhang, and X. Sun, "Self-embedding fragile watermarking based on reference-data interleaving and adaptive selection of embedding mode," *Information Sciences*, vol. 373, pp. 233–250, 2016.

[6] D. Singh, S. Singh, and S. Udmale, "An efficient self-embedding fragile watermarking scheme for image authentication with two chances for recovery capability," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1045–1066, 2023.

[7] D. Singh, S. Udmale, and S. Singh, "Integer wavelet transform based an effective fragile watermarking scheme for exact authentication and restoration," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 4841–4852, 2023.

[8] W. Li, C. Lin, and J. Pan, "Novel image authentication scheme with fine image quality for btc-based compressed images," *Multimedia Tools and Applications*, vol. 75, pp. 4771–4793, 2016.

[9] H. Hemida O.and He, "A self-recovery watermarking scheme based on block truncation coding and quantum chaos map," *Multimedia Tools and Applications*, vol. 79, pp. 18 695–18 725, 2020.

[10] M. Swain and D. Swain, "An effective watermarking technique using btc and svd for image authentication and quality recovery," *Integration*, vol. 83, pp. 12–13, 2022.

[11] C. Kim, D. Shin, and C. Yang, "Self-embedding fragile watermarking scheme to restoration of a tampered image using ambtc," *Personal and ubiquitous computing*, vol. 22, pp. 11–22, 2018.

[12] M. Jana, B. Jana, and S. Joardar, "Local feature based self-embedding fragile watermarking scheme for tampered detection and recovery utilizing ambtc with fuzzy logic," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, Part B, pp. 9822–9835, 2022.

[13] J. L, Z. H, and W. H, "A multipermutation superposition coding-based fragile watermarking for probabilistic encryption," *Multimedia Tools and Applications*, vol. 81, no. 21, pp. 30 025–30 048, 2022.

[14] L. Huang, D. Kuang, C.-l. Li, Y.-j. Zhuang, S.-h. Duan, and X.-y. Zhou, "A self-embedding secure fragile watermarking scheme with high quality recovery," *Journal of Visual Communication and Image Representation*, vol. 83, p. 103 437, 2022.

[15] F. Chen, Y. Yang, H. He, and Y. Yuan, "Adaptive coding and orderedindex extended scrambling based rdh in encrypted images," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[16] F. Chen, Y. Yuan, H. He, M. Tian, and H.-M. Tai, "Multi-msb compression based reversible data hiding scheme in encrypted images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 905–916, 2021.

[17] Z. Qian, X. Zhang, and G. Feng, "Reversible data hiding in encrypted images based on progressive recovery," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1672–1676, 2016.

[18] S. G and S. M, "Ucid: An uncompressed color image database," *Storage and Retrieval Methods and Applications for Multimedia*, vol. 5307, pp. 472–48, 2004.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multi-granularity Semantic and Acoustic Stress Prediction for Expressive TTS

Wenjiang Chi* Xiaoqin Feng* Liumeng Xue† Yunlin Chen* Lei Xie† Zhifei Li*
* Shanghai Mobvoi Information Technology Co., Ltd, China.
E-mail: {wenjiang.chi, xiaoqin.feng, yunlinchen, zfli}@mobvoi.com
† Audio, Speech and Language Processing Group (ASLP)
E-mail: lmxue@nwpu-aslp.org, lxie@nwpu.edu.cn

***Abstract*—Stress, as the perceptual prominence within sentences, plays a key role in expressive text-to-speech (TTS). It can be either the semantic focus in text or the acoustic prominence in speech. However, stress labels are always annotated by listening to the speech, lacking semantic information in the corresponding text, which may degrade the accuracy of stress prediction and the expressivity of TTS. This paper proposes a multi-granularity stress prediction method for expressive TTS. Specifically, we first build Chinese Mandarin datasets with both coarse-grained semantic stress and fine-grained acoustic stress. Then, the proposed model progressively predicts semantic stress and acoustic stress. Finally, a TTS model is adopted to synthesize speech with the predicted stress. Experimental results on the proposed model and synthesized speech show that our proposed model achieves good accuracy in stress prediction and improves the expressiveness and naturalness of the synthesized speech.**

I. Introduction

In recent years, text-to-speech (TTS) systems have made remarkable progress in terms of speech quality and generation speed [1]–[8]. The TTS system is even capable of producing almost human-like speech. However, the expressiveness and naturalness of the synthesized speech can still be improved [9]–[11]. Prosody, which is crucial and essential for expressive speech synthesis, is mainly composed of rhythm, stress and intonation [12]–[16]. Stress, as a perceptual prominence within words or utterances, is a critical factor in prosody [17]. Therefore, accurately modeling stress is beneficial to improve the expressiveness and naturalness of TTS.

Constructing a Mandarin dataset with stress annotations is more challenging compared to constructing one in English. English is a stress-timed language, so it is relatively easy to label the prominence using a well-established English ToBI annotation system [18]. While, Mandarin is a syllable-timed language, which is characterized by syllables with the corresponding tone and pitch contour [19]. The interaction between tone, intonation, and stress makes the stress annotation of Mandarin difficult. Chinese Mandarin stress can be categorized into sentential stress and lexical stress [20], [21]. Sentential stress always reflects on certain syllables or words in need of logical semantics or emotional expressiveness. While, lexical stress, as a part of the word phonetic structure, can be used to distinguish the word sense and part of speech (POS), which is affected by acoustic features, such as pitch, duration and energy.

Generally, there are two typical approaches to acquiring stress labels from text. One is a manual approach by listening to the speech and perceiving the prominence. Li *et al*. released ASCCD, an open-source Mandarin corpus with stress annotations that are labeled by human perception [22]. The other is an automatic approach via Continuous Wavelet Transform (CWT) analysis of the speech [23]. Talman *et al.* [24] utilized the CWT to automatically extract word-level acoustic prominence of LibriSpeech [25]. Both methods are performed based on acoustic information in speech without the consideration of semantic information in the text transcription of the speech, which may result in less accuracy for stress detection and less expressivity for speech synthesis.

Based on the dataset either with manual stress annotations or automatic stress annotations, stress can be predicted from the acoustic features extracted from the speech [26], [27], such as minimum pitch, maximum pitch, duration and so on, linguistic features [19] analyzed from the text, such as tone, POS, positions of the words, or the combination of acoustic and linguistic features [20], [26]. As for TTS, it is more practical to predict stress from only text because no speech is accessible when generating speech from text. Accordingly, word embedding extracted from the text is introduced into the model of stress prediction [17], [28]. Furthermore, with the significant improvement of pre-trained language models, such as BERT [29], the learned contextualized word representation is involved in rich semantic and syntactic cues of the text, which could be beneficial to the stress prediction and the downstream TTS task. Talman *et al.* [24] employed the contextualized word representation extracted from BERT to predict stress, outperforming other models using word embedding or linguistic features. However, although it achieved an excellent result in stress prediction, the performance of TTS with the predicted stress is unknown.

In this paper, we propose a multi-granularity stress prediction model for TTS, in which stress is predicted from the contextualized word representation extracted from text via BERT and the TTS framework of VITS [8] is adopted to generate speech with stressed words. Considering the two-level nature of Mandarin stress in terms of semantic information of text and acoustic information of speech, we first build two Chinese Mandarin datasets with multi-grained stress

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

annotations. One is labeled with coarse-grained stress which attends to the phrase based on the semantic information of the text. The other is labeled with fine-grained stress that mainly focuses on specific words according to the prominence perception of the speech. Then, we design a multi-granularity stress prediction model to progressively predict stress from coarse-grained level to fine-grained level. Finally, we feed the predicted stress into a TTS model to synthesize speech with stressed words. We objectively and subjectively evaluate the performance of the stress prediction and the corresponding synthesized speech. Experimental results demonstrate that the proposed multi-granularity stress prediction model achieves more accurate performance and the generated speech based on the predicted stress is more expressive and natural.

II. Dataset Construction

For the standard stress corpus collection, the voice actor is often required to read well-designed utterances fluently and accurately, and then professional labelers or people from crowdsourcing projects annotate the stress according to the perceived prominence from the recordings, as shown in the left part of Figure 1. The stress expression obtained by this approach depends on the voice actor's performance and is limited by the recording environment, pre-designed script, etc. It may not suitable for real-world scenarios and is unaffordable as well. We propose a new multi-grained stress annotation method as shown in the right part of Figure 1, which is comprised of coarse-grained stress with semantic information and fine-grained stress with acoustic information. It is noted that we mainly focus on stressed and unstressed behavior instead of different stress levels [17], so we explicitly annotate the stressed words and then the remaining words are classified into unstressed words. Additionally, before labeling the stress, crowdsourcing people are trained on a subset corpus multiple times in advance, aiming to ensure intra- and inter-rater annotation consistency and reliability. The consistencies of those two-stage stress labeling after training are 76% and 85% in all annotators.

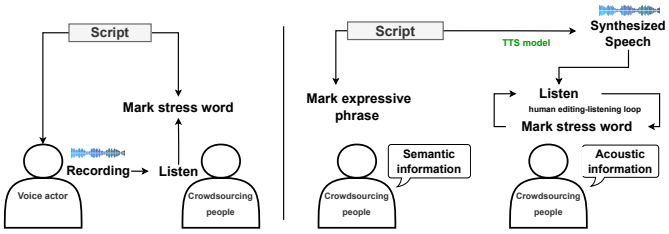


Fig. 1. The difference of stress data construction

A. Coarse-grained Semantic Stress

We first construct a coarse-grained stress corpus that centralizes semantic phrases within a sentence. To this end, people from the crowdsourcing label phrases which are potential stress according to the semantic information delivered in the text. To ensure the accuracy and consistency of the stress label, we establish semi-open annotation criteria. The specification of stress annotation are as follows:

- Mainly focus on emphasized entities, descriptive phrases, and tone conjunctions in the text. For example, the stress phrase of the sentence "年轻的母亲暴跳如雷，竟然打翻了桌子" (The young mother gets into a rage, unexpectedly knocking over the table) is "暴跳如雷" (gets into a rage) and "竟然" (unexpectedly), both of which clearly express the emotion, i.e., angry and surprise.
- To maintain the annotation consistency, we recommend annotating 1 to 3 phrases per sentence by extending or reducing the annotations based on the first rule.
- Based on our statistical results of subword length for stress character, all annotated phrases should keep a full sub-phrase length within the range of 2 to 6 characters.

B. Fine-grained Acoustic Stresss

年 轻 的 母 亲 暴 跳 如 雷， 竟 然 打 翻 了 桌 子!
年 轻 的 母 亲 暴 跳 如 雷， 竟 然 打 翻 了 桌 子!
The young mother gets into a rage, unexpectedly knocking over the table!

Fig. 2. An example of the dataset with multi-grained stress. The character in blue is semantic stress and the character in red is acoustic stress

As mentioned before, standard stress labels are annotated by listening to the recordings, in which stressed words are fixed. Consequently, such stress datasets may lack flexibility and diversity for stress prediction and speech synthesis. To solve this problem, we propose a new fine-grained stress labeling method based on acoustic information. Specifically, we first utilize a high-quality stress-controllable **TTS model** as described in Section III-C to let users set the desired stressed words in the given text to generate speech with stress. Please kindly take note that the TTS model can synthesize speech with stressed words by feeding stress tags, but it cannot predict stress. Then, we use the generated speech for fine-grained stress labeling. Additionally, we introduce a human listening-editing loop to modify the stress position and repeatedly. The human listening-editing loop consists of two circular steps: (1) Listening, where people listen to the synthesized audio generated from the above TTS model. (2) Editing, where people listen to the synthesized speech and then mark the stress word in the current transcript to make the speech perceived more natural. For example, even though "暴跳如雷" (gets into a rage) can be a semantic stress, the speech with all these 4 continuous stress characters does not natural, so finally, we're going to set the stress position at "暴". Similarly, to ensure the reliability of stress labeling, specifications of acoustic stress annotation are as follows:

- Only the locations with high stress intensity are annotated to ensure the accuracy of stress annotation.
- We recommend annotating 1 to 3 positions per sentence, and each position should be in the range of 1 to 2 characters, which maintains the annotation consistency
- To mitigate potential confusion arising from repeated listening, annotators are allowed to engage in the human

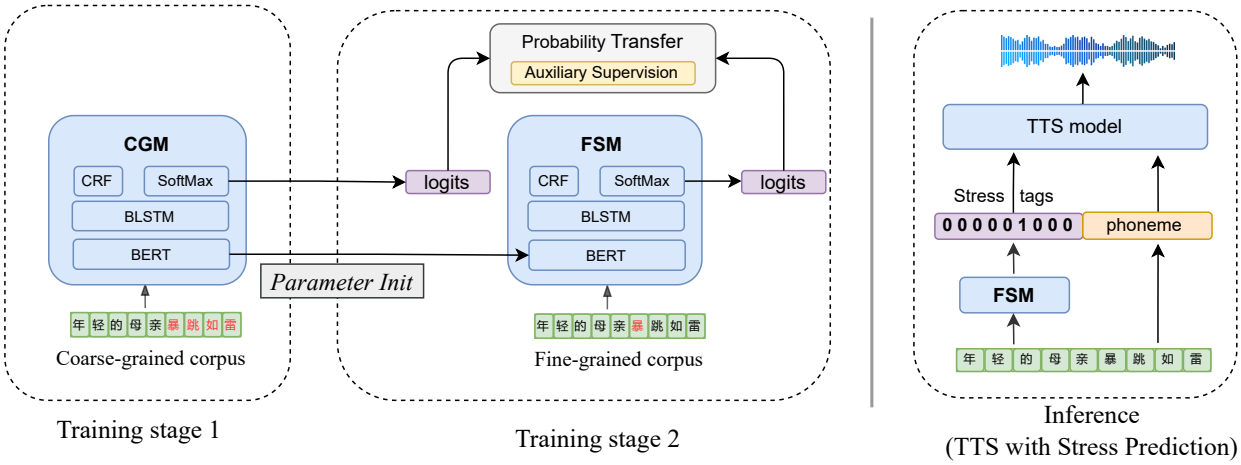


Fig. 3. The overall architecture of our proposed model, where the character in red is ground-truth stressed words. The illustration for the calculation of auxiliary supervised loss in training stage 2

listening-editing loop for each transcript a maximum of 3 times.

Figure 2 shows an example of our dataset with multi-grained stress, where the character in blue is labeled as semantic stress and the character in red is labeled as acoustic stress.

III. Multi-granularity Stress Prediction

The overall architecture of our proposed multi-granularity stress prediction model is presented in Figure 3. The coarse-grained stress and fine-grained stress provide different aspects of information within a sentence, i.e., semantic and acoustic information, respectively. To take full advantage of these two types of stress, the stress prediction model we proposed is a two-stage training model. In the first stage, we trained a coarse-grained stress model (CGM) using a coarse-grained stress corpus. In the second stage, we trained a fine-grained stress model (FSM) using a fine-grained stress corpus. We connected these two stages using two strategies: (1) initializing FSM's BERT parameter with the parameters learned from CGM, and (2) designing a coarse-grained supervision loss to maintain CGM's supervision effect and meanwhile ensure the diversity of FSM. The final stress is predicted from FSM, which is learned by transferring knowledge from CGM. Compared with the model trained on the dataset with acoustic stress only, our model incorporates stress from both semantic and acoustic information, improving the diversity of stress and also benefiting expressive speech synthesis.

A. Stage 1:Coarse-grained Semantic Stress Prediction

As shown in Figure 3, in the first stage, the coarse-grained stress model (CGM) is trained on the coarse-grained corpus as described in Section II-A to identify the salient entities within a sentence. The input of the model is Chinese Mandarin characters, which is consumed by a pre-trained Chinese BERT [30] with 12 transformer layers and 768 hidden states, followed by a Bidirectional Long Short-Term Memory (BLSTM) [31], [32] layer with 768 hidden units. In the BLSTM layer, dropout with the probability of 0.1 is adopted to prevent overfitting.

The loss function of CGM L_{CGM} includes a sequence loss of CRF [33] L_{crf} and a cross-entropy loss L_{ce}, i.e., $L_{CGM} = L_{crf} + L_{ce}$, which could better perform classification task (0 is unstressed and 1 is stressed) and meanwhile keep the contextual connection of the input character sequences. Specifically, the L_{crf} is:

$$\begin{aligned} L_{crf} &= -logP(y|x) \\ &= -(score(y) - log(\Sigma_{\hat{y}} score(\hat{y}))) \end{aligned} \tag{1}$$

where y is the predicted sequence label, x is the current input sentence, $score(y)$ is the score of the current CRF predicted path y, and $score(\hat{y})$ denotes all possible sequences of labels. Besides, L_{ce} is:

$$L_{ce} = -\sum_{i=1}^{seq_len} \sum_{k=1}^{n_class} y_{ik} log G1(\hat{y}_{ik}|x) \tag{2}$$

where seq_len is the length of the input sequence, n_class is the number of label classes, y_{ik} refers to the true label, $\hat{y}_{ik}$ is the k-th value of the predicted label and $G1$ refers to the output probability of the softmax layer in CGM.

B. Stage 2:Fine-grained Acoustic Stress Prediction

After acquiring the coarse-grained semantic stress in the stage 1, we further extract fine-grained acoustic stress.

In the stage 2, a fine-grained stress model (FSM) is utilized to predict stress from Chinese Mandarin characters using the fine-grained stress corpus as described in Section II-B. The structure of FSM is similar to that of CGM, but the BERT is initialized by the parameters learned from CGM. The dropout is also applied to the BLSTM layer with a probability of 0.3.

The loss function of FSM L_{FSM} is $L_{FSM} = L_{crf} + L_{cgce}$, where L_{crf} is CRF loss same as that of CGM and L_{cgce} is the auxiliary supervised loss which is designed based on the focal loss [34] . We use L_{cgce} to preserve the semantic information learned from the CGM so as to better connect the two training stages and predict stress as well. L_{cgce} is formulated as:

$$L_{cgce} = -\sum_{i=1}^{seq_len}\sum_{k=1}^{n_class} y_{ik} log(F2) \quad (3)$$

$$F2 = Min(\beta G2_i, 1) F1(\hat{y}_{ik}|x) \quad (4)$$

where $F1$ refers to the output probability of softmax layer in the FSM, $G2_i$ is the probability transferred from $G1_i$ for the i-th character, which is calculated by the following equation:

$$G2_i = \begin{cases} \frac{1}{\beta} & , if\ gt_i = 1 \\ G1_i & , otherwise \end{cases} \quad (5)$$

where gt_i is the ground-truth of fine-grained stress label and $G1_i$ is the predicted probability from CGM. β is the reciprocal of $1/\beta$. $1/\beta$ is the probability threshold for CGM to be 1 (if $G1_i >= 1/\beta$, then the i-th character is stressed). We use β as a regularization parameter to constrain the impact of coarse-grained supervision on the FSM.

We empirically suggest that β smaller than 10 is more effective. Moreover, the Min truncated normalization is employed to transfer the probability, preventing gradient explosion.

The process of $F2$ calculation is illustrated in Figure 4, in which $G2_i$ is transferred from $G1_i$ according to gt_i and then multiplies β and $F1$, generating $F2$. So, $F2$ is adjusted by gt_i:

- When the gt_i is 1 (stressed), the $F2_i$ will not be affected.
- When the gt_i is 0 (unstressed), the direction of $F2_i$ change is consistent with $G1_i$.

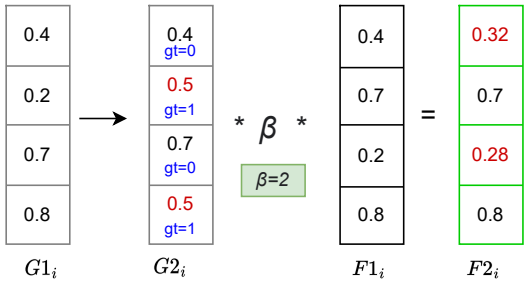


Fig. 4. Probability transfer, where **gt** means the ground-truth of fine-grained stress label and $1/\beta$ is the probability threshold for CGM to be 1.

In other words, we strengthen the L_{cgce} only when the CGM's predicted probability is larger than $1/\beta$, and vice versa, maintaining the CGM supervision effect and ensuring the diversity of the FSM itself.

C. TTS with Predicted Stress

The TTS model adopts the framework of VITS, which is able to synthesize speech with stressed words by feeding the stress tags. As shown in the right part of Figure 3, FSM takes the text as input and outputs the stress tag for each character in the text. And then, the TTS model takes the phoneme sequence extracted from the text and the predicted stress tags as input and generates speech with the corresponding stress. The TTS model is used to verify the effectiveness of the proposed model by feeding the predicted stress as the stress tag. Besides, it is also used to build the dataset with fine-grained acoustic stress (Section II-B) due to its high performance of speech generation.

IV. Experiments

A. Experimental Setups

The detailed information of the datasets, which have varying numbers of sentences, is listed in Table I. As the description of the dataset collection process in Section 2, we collected the coarse-grained stress dataset and fine-grained stress dataset separately. The coarse-grained stress dataset contains 6,817 sentences and the fine-grained stress data includes 13,615 sentences. The stress distribution of the datasets listed in the last column shows that the percentage of fine-grained stress is smaller than that of coarse-grained stress, which is consistent with our annotation specifications as described in Section II, indicating that the annotation results are valid. To train the stress prediction model, we divide each dataset into three subsets with a ratio of 8:1:1, including the training set, validation set and testing set. Moreover, the TTS model is trained on the voice actor's recordings with pre-defined stressed words. Only in this way, we can obtain a relatively reliable TTS system that can generate speech with stressed words.

TABLE I
STATISTICAL RESULTS OF THE STRESS DATASET.

Data Type	#Sentence	Stress Distribution		
		#Character	#Stress	Percentage(%)
Coarse-grained	6,817	125,645	37,063	29.50
Fine-grained	13,615	205,132	22,714	11.07

In the two stages of model training, the batch size is set as 4 for two GPUs and the learning rates of BERT and CRF are 1e-3 and 5e-5, respectively. β in the auxiliary supervised loss L_{cgce} is set to 2.

B. Experimental Results

Objective Evaluations We objectively evaluate the performance of CGM and FSM, comparing with different stress predictor models investigated in Talman *et al.* [24] using our fine-grained acoustic stress corpus. The objective evaluation results, including precision (micro), recall (micro) and F1 (micro) score, are listed in Table II. As we can see, the results show that the performance of our proposed model, i.e., FSM, is better than BERT-base, 3-layer BLSTM and CRF. We suggest that pre-trained language models, such as BERT, are more suitable for our small dataset to capture contextual semantic information. Moreover, FSM and FSM (*one-stage) has a lower F1 score than CGM, while it tends to generate more natural speech (as shown in Table III and Figure 5). The results demonstrate the challenge of directly modeling fine-grained stress, indicating the effectiveness of our proposed multi-granularity stress prediction approach.

Subjective Evaluations We carry out 5-point MOS [35] subjectively listening tests in terms of naturalness (NMOS)

TABLE II
OBJECTIVE EVALUATION RESULTS.

Model	Precision	Recall	F1
CGM	0.8471	0.9061	0.8756
FSM	0.7593	0.6762	0.7153
FSM (*one-stage)	0.6218	0.6284	0.6251
BERT-base	0.6050	0.5737	0.5890
3-layer BLSTM	0.0100	0.8471	0.0197
CRF	0.5278	0.0252	0.0482

and expressiveness (EMOS). Twenty listeners participate in the listening test, rating on 20 samples randomly selected from the testing set [1]. The subjective results are shown in Table III. Except for the first-stage stress prediction model CGM and our final stress prediction model FSM, there are other three types of generated speech for comparison: (a) **Original** represents the synthesized speech without stress; (b) **ManualSet** stands for the synthesized speech with manual annotated acoustic stress; (c) **CGM**$_{random}$ is the generated speech with a random selection of stress positions based on the results of the CGM.

TABLE III
SUBJECTIVE EVALUATION RESULTS.

Type	Original	ManualSet	CGM	CGM$_{random}$	FSM
NMOS	3.86	3.98	3.75	3.83	**3.95**
EMOS	3.9	4.08	3.83	3.88	**4.07**

From the subjective results, we can see that CGM performs poorly, resulting in unnatural and abnormal speech. This can be attributed to the fact that CGM solely focuses on semantic stress (sentential level) annotated from the text without considering of acoustic features of the speech. The final stress should be at the fine-grained level (lexical level) based on acoustic information, which is why CGM$_{random}$ outperforms CGM. The result of our proposed model FSM is close to that of the ground-truth result from ManualSet. We speculate that FSM improves the naturalness and expressiveness of TTS by leveraging both semantic and acoustic information.

C. Ablation Studies

In this section, we conduct ablation studies on the auxiliary supervised loss L_{cgce} and two-stage training strategy. Specifically, we replace L_{cgce} with L_{ce}, i.e., w/o L_{cgce}, and further use the parameters of pre-trained Chinese BERT instead of CGM to initialize FSM, i.e., w/o CGM. Please kindly take note that 'w/o CGM' denotes the utilization of a one-stage training approach solely with the fine-grained stress corpus (acoustic stress information).

The ablation results of objective evaluations are shown in Table IV. It can be seen that the initialization of CGM can enhance the effect of the FSM a little bit. However, the auxiliary supervision loss of CGM significantly improves the performance of the FSM, achieving the highest F1 score of 71.53% with 7% to 8% improvement. We guess that while BERT can provide abundant semantic information, the CGM model still offers remarkable stress-related information in the sentence, thereby facilitating the FSM in detecting fine-grained stress keywords. The subjective preference test results, as depicted in Figure 5, demonstrate that the speech produced by FSM trained on a two-stage approach with coarse-grained information exhibits greater naturalness and expressiveness. Overall, the results demonstrate that the incorporation of coarse-grained information supervision is an effective approach to preventing stress weight dispersion during training.

[1]Samples can be found at https://xqfeng-josie.github.io/stress/

TABLE IV
ABLATION RESULTS OF OBJECTIVE EVALUATION

Model	Loss	Model$_{\text{init}}$	Precision	Recall	F1
FSM	L_{cgce}	CGM	0.7593	0.6762	0.7153
FSM	w/o L_{cgce}	CGM	0.6981	0.5780	0.6335
FSM (*one-stage)	w/o L_{cgce}	w/o CGM	0.6218	0.6284	0.6251

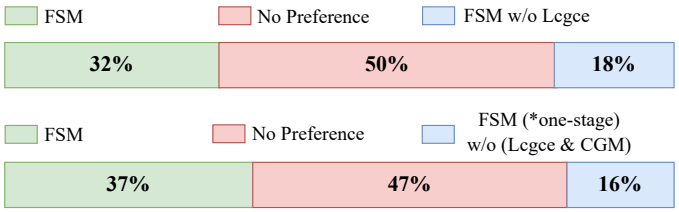


Fig. 5. Preference test results for ablation studies.

V. CONCLUSIONS

In this work, we propose a multi-granularity stress prediction model to improve the naturalness and expressiveness of TTS. In consideration of the nature of multi-grained stress, we construct the dataset with coarse-grained semantic stress and fine-grained acoustic stress to incorporate both the semantic knowledge in the text and the acoustic information in the speech. Then, the two-stage stress prediction model progressively predicts the coarse-grained stress and fine-grained stress from text, where the pre-trained language model is adopted to extract contextualized word representation from the text. Experimental results objectively and subjectively show that our proposed stress prediction model gets a higher F1 score and achieves more natural and expressive synthetic speech. In the future, we will further improve the multi-stage stress prediction model to close the performance gap between human speech and synthetic speech.

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[3] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[5] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," *arXiv preprint arXiv:2006.03575*, 2020.

[6] C. Yu, H. Lu, N. Hu, *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[7] Z.-H. Ling, S.-Y. Kang, H. Zen, *et al.*, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5530–5540.

[9] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards multi-scale style control for expressive speech synthesis," *arXiv preprint arXiv:2104.03521*, 2021.

[10] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2021, pp. 1–5.

[11] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Controllable cross-speaker emotion transfer for end-to-end speech synthesis," *arXiv preprint arXiv:2109.06733*, 2021.

[12] R. Skerry-Ryan, E. Battenberg, Y. Xiao, *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*, PMLR, 2018, pp. 4693–4702.

[13] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, PMLR, 2018, pp. 5180–5189.

[14] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6945–6949.

[15] Y. Wang, R. Skerry-Ryan, Y. Xiao, *et al.*, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[16] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis," *arXiv preprint arXiv:1904.02373*, 2019.

[17] Y. Zheng, Y. Li, Z. Wen, B. Liu, and J. Tao, "Text-based sentential stress prediction using continuous lexical embedding for mandarin speech synthesis," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2016, pp. 1–5.

[18] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, *et al.*, "Tobi: A standard for labeling english prosody.," in *ICSLP*, vol. 2, 1992, pp. 867–870.

[19] Y. Li and J. Tao, "Mandarin stress analysis and prediction for speech synthesis," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, Springer, 2015, pp. 83–95.

[20] Shao, Yanqiu, Han, *et al.*, "Study on automatic prediction of sentential stress for chinese putonghua text-to-speech system with natural style," 2007.

[21] M. Chu and M. Bao, "Comparison of sentential-stress allocation within base phrases among different reading styles," in *Speech Prosody 2004, International Conference*, 2004.

[22] L. Aijun, C. Xiaoxia, S. Guohua, *et al.*, "The phonetic labeling on read and spontaneous discourse corpora," in *proceedings of International Conference on Spoken Language Processing (ICSLP)*, Citeseer, 2000.

[23] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[24] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," *arXiv preprint arXiv:1908.02262*, 2019.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[26] C. Ni, W. Liu, and B. Xu, "Mandarin pitch accent prediction using hierarchical model based ensemble machine learning," in *2009 IEEE Youth Conference on Information, Computing and Telecommunication*, IEEE, 2009, pp. 327–330.

[27] S. Shechtman and M. Mordechay, "Emphatic speech prosody prediction with deep lstm networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5119–5123. DOI: 10.1109/ICASSP.2018.8462473.

[28] Y. Mass, S. Shechtman, M. Mordechay, *et al.*, "Word Emphasis Prediction for Expressive Text to Speech," in *Proc. Interspeech 2018*, 2018, pp. 2868–2872. DOI: 10.21437/Interspeech.2018-1159.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.

[31] P. Zhou, W. Shi, J. Tian, *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2016, pp. 207–212.

[32] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[33] T. Müller, H. Schmid, and H. Schütze, "Efficient higher-order crfs for morphological tagging," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 322–332.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[35] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1137–1144, 2006.

NADiffuSE: Noise-aware Diffusion-based Model for Speech Enhancement

Wen Wang, Dongchao Yang, Qichen Ye, Bowen Cao and Yuexian Zou*
the School of Electronic and Computer Engineering, Peking University, Shenzhen, 518055, China
E-mail: {wangw, 2001212832, yeeeqichen, cbw2021}@stu.pku.edu.cn, zouyx@pku.edu.cn

***Abstract*—The goal of speech enhancement (SE) is to eliminate the background interference from the noisy speech signal. Generative models such as diffusion models (DM) have been applied to the task of SE because of better generalization in unseen noisy scenes. Technical routes for the DM-based SE methods can be summarized into three types: task-adapted diffusion process formulation, generator-plus-conditioner (GPC) structures and the multi-stage frameworks. We focus on the first two approaches, which are constructed under the GPC architecture and use the task-adapted diffusion process to better deal with the real noise. However, the performance of these SE models is limited by the following issues: (a) Non-Gaussian noise estimation in the task-adapted diffusion process. (b) Conditional domain bias caused by the weak conditioner design in the GPC structure. (c) Large amount of residual noise caused by unreasonable interpolation operations during inference. To solve the above problems, we propose a noise-aware diffusion-based SE model (NADiffuSE) to boost the SE performance, where the noise representation is extracted from the noisy speech signal and introduced as a global conditional information for estimating the non-Gaussian components. Furthermore, the anchor-based inference algorithm is employed to achieve a compromise between the speech distortion and noise residual. In order to mitigate the performance degradation caused by the conditional domain bias in the GPC framework, we investigate three model variants, all of which can be viewed as multi-stage SE based on the preprocessing networks for Mel spectrograms. Experimental results show that NADiffuSE outperforms other DM-based SE models under the GPC infrastructure. Audio samples are available at: https://square-of-w.github.io/NADiffuSE-demo/.**

I. Introduction

In the last decade, deep learning (DL) methods [1], [2] have become the mainstream for speech enhancement (SE), which can be divided into two categories: the discriminative and generative ones [3]. The discriminative methods learn non-linear mappings[4] or estimate time-frequency masks [5]–[7] through annotated speech data pairs, but suffer from non-linear artifacts and poor generalizations [8]. The generative approaches [9]–[11] employ different infrastructures such as Generative Adversarial Networks (GANs) [12], Variational Autoencoders (VAEs) [13] and flow-based models [14] to learn the distribution of clean speech signals, which can be more robust to complex and varying noise scenarios [15].

Diffusion Denoising Probabilistic Models (DDPM) [16] is a new kind of generative model inspired by the nonequilibrium thermodynamics [17]. DDPM corrupts the data to the pre-defined Gaussian distribution by gradually adding noise at the forward process, and the random noise is progressively denoised and finally restored to the original data at the reverse stage. Many works [18]–[20] have employed the DDPM to generate natural and high-quality human voice from the given text or Mel spectrogram. Recently, DDPM-based SE methods [21]–[29] have also been explored for the promising result in speech generation.

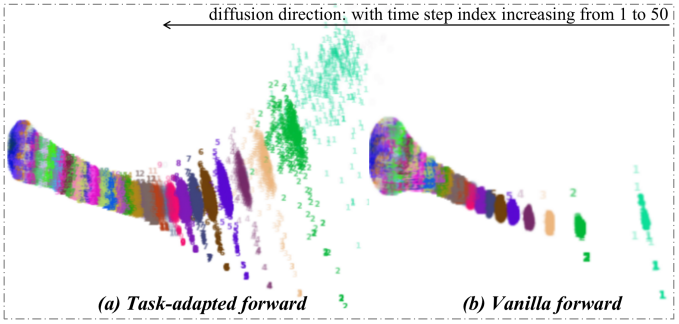


Fig. 1. To demonstrate the effect of non-Gaussian noise in diffusion-based SE, we visualize the vanilla and task-adapted diffusion process: we randomly selected 2000 utterances, which are first converted into Mel spectrograms, then cropped to 62 frames (same as the training settings), and finally visualized in two dimensions using the t-SNE algorithm. We use different colors and indexes to label the data at different time steps.

Current diffusion-based SE models can be categorized into three types. The first [21]–[23] is to change the mathematical form of the diffusion process to adapt to the task of SE, in which the mean of clean speech signals is gradually pulled towards the noisy ones by interpolating the asymptotically increasing noisy signal, so that the reverse stage is directly the speech enhancement process. The second [23]–[26] is to train a conditional network based on a well-trained pure speech generation network (Generator-plus-Conditioner, GPC). Under this setting, the enhanced speech signal is generated with the acoustic feature produced by the conditioner, which can be expressed as ***output = generator(conditioner(noisy input))***. The last is to develop a multi-stage SE where the diffusion model aims to learn the fine-grained or residual speech signal based on the coarsely enhanced one [27]–[29]. This study focuses on the first two lines, and leverages a task-adapted diffusion process under the GPC architecture to deal with the real noise.

However, there exists three problems in the existing works: (1) **Non-Gaussian Estimation**. As shown in Fig. 1, incorporating the noisy signal into the forward process makes the data

*Corresponding author.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

no longer satisfies the tight Gaussian distribution at each step, which we attribute to the effect of background noise. Noise interference varies in realistic scenarios, hence it is challenging for SE models to understand the patterns of various noise and to adaptively learn the ability of denoising in real situations [30]–[33]. Moreover, the estimation target changes from the added Gaussian noise to a combination of Gaussian noise and non-Gaussian background noise, which makes training the model harder [21]. (2) **Conditional domain bias**. In the GPC framework, the generator and conditioner determine the upper and lower bounds for the SE performance: If a conditioner can ideally map the noisy features completely to the pure ones, the best performance of SE is then obtained (upper bound). Otherwise, if the conditioner does not work at all, the worst performance is obtained (lower bound). Table I records the SE performance bounds of [23]. We observe that current conditioner is not sufficient to compensate for the gap between lower and upper bounds. We define it as conditional domain bias, which results from the change in the dimension, type and purity of the acoustic features (usually Mel spectrogram). (3) **Under-explained interpolation**. In the original inference algorithm (*cf.* Line 11, Alg. 1), the linear interpolation operation [21], [23], [26] is commonly used in the last step with a certain percentage ($r = 0.2$), which serves as an implicit post-processing method to supplement the lost speech details. Although the interpolation does improve the objective evaluation metrics (*cf.* Table III row 4 and 5), its validity is limited due to the introduced large amount of additional noise. We give a comparison between with and without the interpolation in Fig. 4 as an example of evidence: the white dashed box in (f) indicates the presence of the significant noise components.

To address the above issues, our improvements are refined into a noise-aware diffusion-based SE model (NADiffuSE). The main contributions of this work are summarized as follows:

1. To more accurately estimate the non-Gaussian noise component in the task-adapted diffusion process, we propose to use noise encodings to guide the diffusion model for adaptive noise reduction in real noisy situations.
2. To further reduce the conditional domain bias under the GPC architecture, we design three network variants based on the additional pre-processor network to improve the quality of regenerated speech signals.
3. To reduce the additional noise introduced by the interpolation operation, we construct a relatively accurate data anchor point from noisy speech signals at specified time steps, based on which we use iterative interpolation operations to refine the speech details.

II. Related Work

A. Diffusion Model

The diffusion model [16], [34], [35] contains a forward and a backward process. In the mathematical form of DDPM [16], each process is represented by a first-order Markov chain with fixed time steps. In the forward process, a series of corrupted data $x_{1:T} = x_1, x_2, ..., x_T$ can be obtained from the original clean data x_0 through the transfer distribution $q(x_t|x_0) = \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon$, where $\bar{a}_t = \prod_{i=1}^{T} a_i$, $a_t = 1-\beta_t$, $\epsilon \sim \mathcal{N}(0, I)$ and β_t is a constant defined in advance. In the reverse process, the original data x_0 can be recovered from the latent distribution x_T by iteratively performing the backward transfer step $p_\theta(x_{t-1}|x_t) = \mu_\theta(x_t) + \widetilde{\beta}_t I$, where $\widetilde{\beta}_t$ is a constant and $\mu_\theta(x_t) = \frac{1}{\sqrt{a_t}}(x_t - (\beta_t/\sqrt{1-\bar{a}_t})\epsilon_\theta(x_t))$.

TABLE I
Evaluation results of the GPC structure [26] on new (a) and original (b) VoiceBank-Demand; For WaveNet[36]-based generator we evaluated the SE performance's Upper (generator with clean Mel spectrograms) and Lower (with noisy Mel spectrograms) bounds. Task-adapted diffusion process [23] is adopted here.

(a) Evaluation results on new VoiceBank-Demand

Model	Mode	CSIG	CBAK	COVL	PESQ
Unprocessed	/	3.65	3.16	2.91	2.13
CDiffuSE	/	3.83	3.13	3.19	2.55
DiffWave	Upper	4.41	3.57	3.84	3.26
	Lower	3.45	2.68	2.78	2.16
DiffWave-Cls	Upper	**4.65**	**3.67**	**4.05**	**3.44**
	Lower	3.55	2.73	2.86	2.20
DiffWave-Emb	Upper	4.34	3.54	3.79	3.22
	Lower	**3.59**	**2.78**	**2.89**	**2.22**

(b) Evaluation results on original VoiceBank-Demand

Model	Mode	CSIG	CBAK	COVL	PESQ
Unprocessed	/	3.35	2.44	2.63	1.97
CDiffuSE	/	3.66	2.83	3.03	2.44
DiffWave	Upper	4.34	**3.53**	**3.79**	**3.21**
	Lower	3.40	**2.55**	2.71	2.08
DiffWave-Emb	Upper	**4.36**	3.42	3.78	3.19
	Lower	**3.41**	2.47	**2.74**	**2.15**

B. Diffusion-based SE Methods

The current technical routes of diffusion-based SE mothods can be divided into three categories: Task-adapted diffusion formula, GPC structures and multi-stage frameworks.

Task-adapted diffusion formula [21]–[23] adapt the task of SE by incorporating the noisy sigal y into the diffusion process: the mean of x_t is progressively pulled from the x_0 to y. The data degradation in [23] is formulated as Eq. 1, where $m_t = \sqrt{(1-\bar{\alpha}_t)/\sqrt{\bar{\alpha}_t}}$ denotes the asymptotic coefficient and $\bar{\delta}_t = 1-(1+m_t^2)\bar{\alpha}_t$ is the variance for added Gaussian noise.

$$x_t = \sqrt{\bar{a}_t}((1-m_t)x_0 + m_t \cdot y) + \sqrt{\bar{\delta}_t}\epsilon \tag{1}$$

While [21], [22] have the closed-form solution for the mean value as $\mu(x_0, y, t) = e^{-\gamma t}x_0 + (1-e^{-\gamma t})y$.

GPC structures consist of a generator which can generate the clean speech signal based on the clean conditions and a conditioner, i.e. network designs or training methods. One kind of conditioner [23], [26] works through two-stage training: the first stage uses the 80-dim pure Mel spectrogram as a condition (pre-training), and the second one adjusts the weights using the 513-dim noisy amplitude spectrogram (fine-tuning). The other [25] is replacing the submodule in the generator

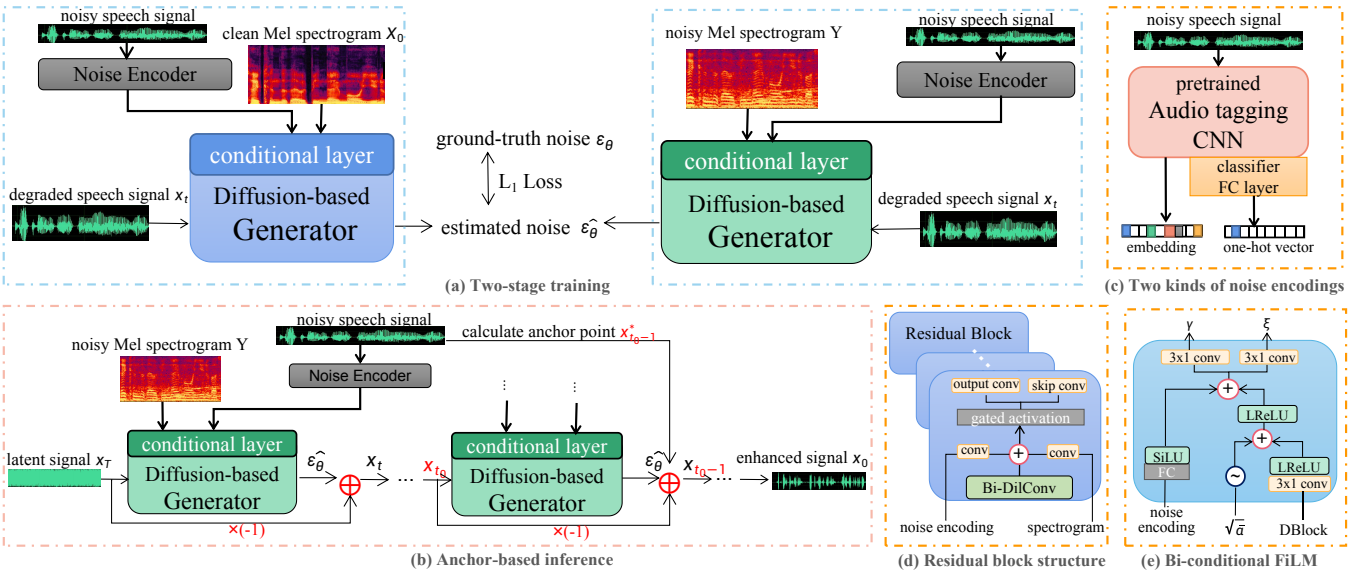


Fig. 2. An overview of the proposed NADiffuSE: (a) pre-training (left) and fine-tuning (right), both of which are conditioned on Mel spectrograms and noise encodings; (b) improved inference process where the iterative interpolation operates in the last t_0 steps based on the anchor point; (c) latent embedding (left) and category classifier (right) of noise encoding in proposed noise-aware training; (d) bi-conditional Residual blocks.

with a newly trained one to alter the degraded conditions to match the original ones.

Multi-stage frameworks take full advantage of the nature that the diffusion model is more suitable for detail refinement. The coarsely enhanced result is obtained through a discriminative model or also a diffusion-based mothod. And then conditioned on the coarsely enhanced signal, the diffusion model is used to generate the fine-grained enhanced signal [28], [29] or the residual signal [27] against the clean one.

III. Proposed Method

We make improvements to address the above issues and give an overview diagram in Fig. 2. The training process in (a) follows the same two-stage paradigm as [23], where both noise encodings and spectrograms are used as conditional information. The inference process in (b) is iterative and improved using noisy speech signals y in the last steps.

A. Noise encodings

We propose to additionally use noise coding as global conditional information in an aim to mine the priori information of the acoustic noise to guide the diffusion model for accurate combine noise estimation. We explore two different types of noise encodings as drawn in (c).

1) Category classifier: We first use the categorical properties of the background noise, in which the noise type labels are fed into the learnable embedding layer in the form of one-hot vectors. In order to get the ground truth noise label during training, we only consider a fixed number of closed sets of noise scenes. At the inference stage, a noise classifier is required to predict the noise type of the noisy speech signal. The background noise is usually labeled using the the keywords of the acoustic scene where the noise is collected, like living room, street, bathroom and etc. The concept of noise semantics is difficult to define and we propose to describe the background noise with acoustic events. We observe a one-to-many relationship between noise scenes and sound events, for example, keyboard tapping, TV background and babble noise may exist simultaneously in a living scene. Therefore, we perform transfer learning based on a large-scale pre-trained network structure [37] for the audio tagging task. We load the pre-trained weights as initialization and add a linear classification into the original structure. As described above, we build a noise classifier whose input is the noisy speech signal and output is a value belonging to a predefined set of noisy scene labels $\{0, 1, ..., N-1\}$, with N being the total number of noise categories.

2) Latent embedding: In order to extend the noise encodings to the open domain, we further explore to characterize the noise properties in a latent embedding mechanism. We directly use the output of the classifier's (described above) last hidden layer as a noise feature. The 2048-dimensional embedding is extracted from the noisy speech signal and aligned with the hidden dimension of the residual block through MLP structure. It is worth noting that we still fine-tune the convolutional feature extractor using the classification task in order to make it more relevant to the background noise characterization.

B. Anchor-based inference algorithm

The ideal case of inference is that the reverse data distribution fits the forward one exactly. We use the same conditional diffusion process as in Eq. 1, where the mean value is determined partly by x_0 and partly by y. Thus we can use the noisy signal y (known during inference) to construct a relatively accurate anchor point for the reverse process in Eq

Algorithm 1 Anchor-based Inference algorithm

Sample $x_T \sim p_{latent}$
for $t = T-1, T-2, ..., 0$ **do**
  Compute $\mu_\theta(x_{t+1}, y)$ and $\bar{\delta}_t$ as in Eq. (3)
  Sample $x_t = \mu_\theta(x_{t+1}, y) + \bar{\delta}_t I$ from $p_\theta(x_t|x_{t+1}, y)$
  if select Improved Sampling and $t < t_0$: **then**
    Calculate the anchor point x_t^* using Eq. (2)
    Do interpolation $x_t = r_t \cdot x_t^* + (1 - r_t) \cdot x_t$
  end if
end for
if select Original Sampling: **then**
  $x_0 = r \cdot x_0 + (1 - r) \cdot y$
end if
return x_0

2, where $\bar{\alpha}_t$ and $\bar{\delta}_t$ follow the previous definition.

$$x_t^* = m_t\sqrt{\bar{\alpha}_t}y + \sqrt{\bar{\delta}_t}\epsilon \tag{2}$$

Anchors contain both useful clean speech details and degraded background noise. We have described in Section I the pros and cons of the current interpolation operation used in the inference process. Rather than directly interpolating the result using noisy speech at the final step, We use anchors for the interpolation to decrease the extra noise. We use the anchor-based interpolation repeatedly in and only in the last few steps, because the data will be more sensitive to the slight inaccuracies as inference finishes. Therefore, we have used iterative interpolation operations to remove some of the noise residules contained in the anchor points by stepwise refinement. To further weaken the effect of noise, we linearly anneal the interpolation coefficients at the rate of $1/t_0$ (other annealing options are worth exploring). Our improved anchor-based inference algorithm is given in Alg. 1: we follow Eq. (3) at each step to sample x_{t-1} from the reverse probability distribution, and in the last t_0 steps perform an interpolation with the anchor point. The mean is a linear combination of x_t, y and the estimated noise term ϵ, where c_{xt},c_{yt} and $c_{\epsilon t}$ are constants and the detailed derivation follows [23].

$$\begin{aligned}\mu_\theta(x_{t+1}, y) =& c_{xt}x_t + x_{yt}y_t + c_{\epsilon t}\epsilon_\theta \\ \bar{\delta}_t =& \delta_{t-1} - \frac{1 - m_t}{1 - m_{t-1}}^2 \alpha_t \frac{\delta_{t-1}^2}{\delta_t}\end{aligned} \tag{3}$$

C. Model variants in the conditioner design

In the section I we give a definition of the conditional domain bias, a problem that exists under the generator-plus-conditioner (GPC) structure. Following the similar conditioner design in [23], we first consider using the 80-dimensional Mel spectrogram as conditional information in both training stages to reduce the difficulty of aligning the conditional domain in the fine-tuning phase. Given that the current conditional mapping layer contains only a simple up-sampling layer and a shallow MLP structure, we are inspired by [38] to additionally train a preprocessing network (preprocessor) for enhancing the Mel spectrogram. Thus we can get the pre-enhanced Mel

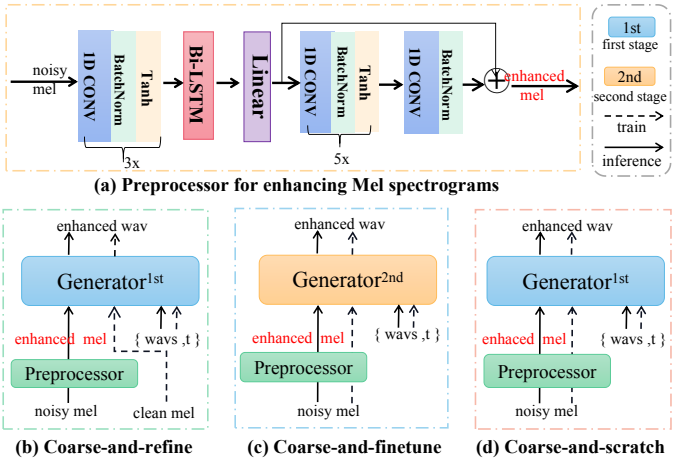


Fig. 3. Training (dashed line) and inference (solid line) pipelines of three conditioners based on the preprocessor: (a) preprocessor used for enhancing the noisy Mel spectrograms (known as *coarse*) (b) *coarse-and-refine* only needs one training stage in which the generator is trained with clean Mel spectrograms and inferenced with enhanced ones (c) *coarse-and-finetune* stills goes through the second training stage where the generator is finetuned with enhanced Mel spectrograms (d) *coarse-and-scratch* also needs only one training stage where the generator uses enhanced Mel spectrograms in both the training and inference phases.

spectrogram and name this process as *coarse*. Given the pre-enhanced Mel spectrogram: on the one hand it can be directly used as the initial result of the enhancement and then refined using the generator (coarse-and-refine). On the other, compared to the noisy speech signal's Mel spectrogram, it is less difficult to perform fine-tuning with the pre-enhanced one (coarse-and-finetune). Last, the pre-enhanced Mel spectrogram is available in both training and testing phases, which can be used to directly train a generator from scratch, without further fine-tuning. In conclusion, three feasible conditioner designs are illustrated in Fig. 3 (b), (c) and (d).

1. Coarse-and-refine: The generator which is only trained with clean Mel spectrograms can be directly used to generate the speech with the pre-enhanced Mel spectrograms.
2. Coarse-and-finetune: After training with clean Mel spectrograms, the generator is then fine-tuned with enhanced Mel spectrograms instead of the noisy ones.
3. Coarse-and-scratch: The generator is trained from scratch using the pre-enhanced Mel spectrograms.

All three network variants use the pre-enhanced Mel spectrogram as a condition for inference, and Gaussian-like speech signal sampled from latent distribution $N(x_T; \sqrt{\bar{\alpha}_T}y, \delta_T I)$ as signal input. The above proposed conditioner mechanism is applicable for all SE models under the GPC structure.

IV. Experiment

A. Experimental Setup

1) Dataset: We evaluated above all methods on the original and newly simulated VoiceBank-DEMAND [39] dataset. The new one follows the same setting which consists of 30 speakers from the VoiceBank [40] corpus and was mixed with 12

TABLE II
COMPARATIVE RESULTS OF THE PROPOSED NADIFFUSE ON AUXIALIARY INFORMATION. ALL RESULTS ADOPTED THE ORIGINAL INTERPOLATION OPERATION WHERE 20% NOISY SIGNAL IS ADDED AT THE END OF THE REVERSE PROCESS. SIGN * MEANS REPRODUCED RESULTS.

Row-Id	Method	Auxiliary		Metrics			
		Spectrogram	Noise encoding	CSIG	CBAK	COVL	PESQ
0	Unprocessed	✗	✗	3.65	3.16	2.91	2.13
1	DiffuSE*	80-dim Mel+513-dim Spec	✗	3.80	3.10	3.15	2.52
2	CDiffuSE*	80-dim Mel+513-dim Spec	✗	3.83	3.13	3.19	2.55
3	no-condition	✗	✗	2.87	2.61	2.36	1.88
4	mel-conditioned	80-dim Mel+80-dim Mel	✗	3.86	3.13	3.22	2.58
5	noise-class	✗	category classifier	3.52	2.86	2.85	2.20
6	NADiffuSE	80-dim Mel+80-dim Mel	hard classifier	**3.91**	**3.15**	**3.26**	**2.63**

noise types* from the DEMAND [41] database. It was then divided into a training, validation and test set with 26, 2 and 2 speakers, containing 10792, 770 and 824 synthesized utterances, respectively. The signal-to-Noise (SNR) range of the training and validation set is $\{0, 5, 10, 15\}$, and the test set is $\{2.5, 7.5, 12.5, 17.5\}$. All of the utterances were resampled to 16KHz sampling rates. We use PESQ, CSIG, CBAK and COVL as evaluation metrics for enhanced speech, with higher scores indicating better performance.

2) Model infrastructure: Our proposed model follows the generator-plus-conditioner architecture mentioned in the previous section. The generator is constructed based on DiffWave [18], which is a diffusion model based vocoder and has been extended to the task of SE by [23], [25], [26]. The network in above works constructs from WaveNet [36] which has 30 layers of residual blocks with dilated convolution (conv) and gated activation. As shown in Fig. 2 (d), each residual block has two 1x1 conv layers for Mel spectrograms and noise encodings.

3) Training settings: All of experiments are based on the same training configurations as CDiffuSE-base [23] in which 50-step linear noise scheduler $\beta_t \in [1 \times 10^{-4}, 0.035]$ is used. The learning rate is 2×10^{-4} and the batch size is 16 for both training stages. The dimension for the Mel spectrogram is 80, which is transformed by STFT with the window size of 1024 and the shift of 256. We uniformly train 100w iterations in the first training stage and 30w in the second stage. The full sampling scheme is used in the reverse process.

B. Evaluation results for noise encodings

To resolve the non-Gaussian estimation problem proposed in the Section I, two types of noise encodings, namely category classifier (Cls) and latent embedding (Emb) were explored in Table I. Among them, the latent embedding form cannot obtain good performance gain in upper bound probably because it contains too much redundant information about audio events in the background noise that does not need to be considered. Therefore, our following proposed model uses the noise representation in the form of one hot vectors by default. In order to get the ground-truth background noise labels, we conducted experiments on the new simulated dataset. The ablation studies are done to validate the effectiveness of noise encodings. From Table II we can observe: (1) Under the GPC archietecture, mel-spectrogram works as an important condition to improve the metric scores of restored speech signals when compared with row 3 and 4; (2) Metrics in row 5 increase a little compared to row 3, showing that the noise embeddings can improve the enhancement performance to some extent, but slightly lacks in speech detail fidelity; (3) The best performance is obtained when two conditions are both used as NADiffuSE (row 6) reach the highest scores among all combination of auxiliary information. It can be explained that the noise embedding provides the priori information about noise patterns, and thus viewed as an implicit multi-branch switch that guides the model for adaptive noise reduction. We also visualized the effects in in Fig. 4: when inference only with Mel spectrograms, more details in the input noisy speech signal are reserved, but also including residual background noise components according to the dashed box in (c); when only using noise encodings, the noise is removed more completely, owing to the explicit use of noise characteristics, but some speech details are lost as shown by the solid ellipse in (d). Another point of interest is that the difference between Table II row 2 and 4 is only in the spectral information used in the second training phase. The results show that row 4 can achieve comparable (or even a little better) results than row 2, which validates our view that there is no need to change the type and dimensionality of spectrograms before and after the two training phases, and lays the foundation for our proposal of a preprocessing network for the Mel spectrogram later on.

C. Evaluation results for the improved inference

To evaluate the anchor-based inference algorithm, we first analyze the original interpolation operation in the inference algorithm and further investigate different parameter settings for the improved inference algorithm. We can first summarize from Table III that without any interpolation operation, the performance will significantly degrade because there is a big gap between NADiffuSE- and NADiffuSE. As we have analyzed in Section III-B, we only use iterative interpolation operations in a limited number of time steps because of the increasing sensitivity to extra noise during the inference. In Fig. 5, we give an example of visualization on the inference process where the last five steps play an important role in restoring the speech signal. Therefore, we empirically choose $t = 5$ first and explore the effect of different interpolation coefficients on the final performance in the row 8, and we can find that

*TCAR, PSTATION, DLIVING, TBUS, TMETRO, OMEETING, SPSQUARE, STRAFFIC, PRESTO, OOFFICE, PCAFETER, NFIELD.

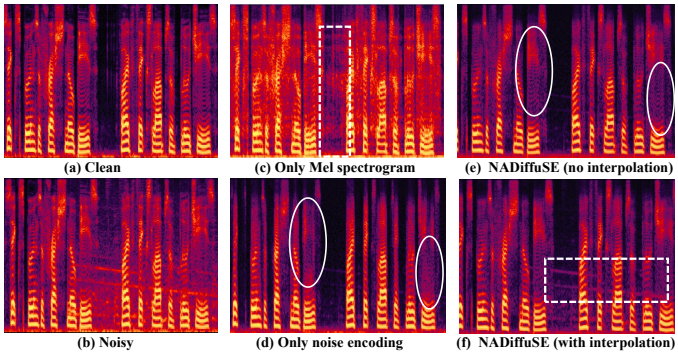


Fig. 4. Mel spectrograms of (a) clean speech, (b) noisy speech, (c) enhanced speech only using the mel-spectrogram as condition, (d) enhanced speech only using noise encodings, (e) NADiffuSE's enhanced speech without the interpolation and (f) NADiffuSE's enhanced speech with the interpolation. The dashed box indicates noise residuals and the solid ellipse indicates speech distortion.

TABLE III
COMPARATIVE RESULTS OF THE THE ANCHOR-BASED INFERENCE ALGORITHM. FOR THE SAMPLING OPTION, "$t = 0$" WITH "$r = 0.2$" MEANS THE ORIGINAL INTERPOLATION DURING INFERENCE (DENOTED BY "-"). SIGN "+" MEANS OUR IMPROVED INFERENCE ALGORITHM. ALL BASELINE ARE REPRODUCED ON THE NEW DATASET.

Row-Id	Method	Sampling		Metrics			
		Ratio (r)	Step (t)	CSIG	CBAK	COVL	PESQ
0	Unprocessed	✗	✗	3.65	3.16	2.91	2.13
1	SEGAN	✗	✗	3.71	3.12	3.04	2.35
2	SGMSE+	✗	✗	**4.18**	**3.47**	**3.58**	**3.24**
3	DiffuSE	r=0.2	t=0	3.80	3.10	3.15	2.52
4	CDiffuSE	r=0.2	t=0	3.83	3.13	3.19	2.55
5	CDiffuSE-	✗	✗	3.80	3.13	3.16	2.52
6	NADiffuSE	r=0.2	t=0	3.91	3.15	3.26	2.63
7	NADiffuSE-	✗	✗	3.84	3.09	3.18	2.55
8	NADiffuSE+	r=0.2	t=5	3.92	3.15	3.25	2.62
		r=0.5		3.81	3.11	3.09	2.37
		r=0.8		3.77	3.15	2.92	2.15
9	NADiffuSE+	r=0.1	t=5	**3.94**	**3.18**	**3.31**	**2.69**
10	NADiffuSE+	r=0.1	t=50	3.88	3.12	3.25	2.63
			t=10	3.89	3.13	3.26	2.64
			t=2	3.91	3.16	3.27	2.64

the smaller the coefficient is, the higher the performance gain can be obtained. Specially, we degrade the coefficients linearly on average with the time step, e.g. for $t = 5, r = 0.1$, we would apply coefficients of $[0.1, 0.08, 0.06, 0.04, 0.02]$ for the last five steps respectively. After we experimentally selected r = 0.1, we did a validation test in row 10 and the results showed that the final performance indeed decreases as the time step t increases. Ovarally, we have some conclusions: (1) Our proposed inference algorithm is more effective than the original one. (2) $r = 0.1, t = 5$ is the best parameter setting.

D. Evaluation results for the proposed model and its variants

1) Overal model: Through the previous experiments, our proposed model is identified as a time-domain diffusion model that uses Mel specrograms and the category classifier form of noise encodings as conditional information in both training stages. For inference, we need to train a noise classifier (detail in Section III-A1) to get the estimated noise type from the given noisy speech signal. The experimental results in Table IV show that the convergence of the noisy classifier on our newly simulated dataset is not difficult and there is no significant difference in the final SE performance between the two trained classifiers with 92% and 96% accuracy (row 1 v.s. 2). This

TABLE IV
EVALUATION RESULTS OF NADIFFUSE+ AND ITS VARIANTS. FOR THE NEEDED NOISE LABELS, WE NEED AN ACCURATE NOISE CLASSIFIER. WE DENOTE THE GROUND-TRUTH NOISE LABEL AS GT. MODELS WITH THE 92% AND 96% ACCURACY CLASSIFIER ARE ABBREVIATED AS 92%-NADIFFUSE+ AND 96%-NADIFFUSE+.

Row-Id	Method	CSIG	CBAK	COVL	PESQ
0	*gt-NADiffuSE+*	3.95	3.22	3.31	2.69
1	*92%-NADiffuSE+*	3.94	3.18	3.29	2.66
2	*96%-NADiffuSE+*	3.94	3.18	3.31	2.69
3	*Coarse-and-refine*	**4.06**	**3.22**	**3.44**	**2.84**
4	*Coarse-and-finetune*	3.93	3.20	3.24	2.54
5	*Coarse-and-scratch*	3.98	3.19	3.35	2.74

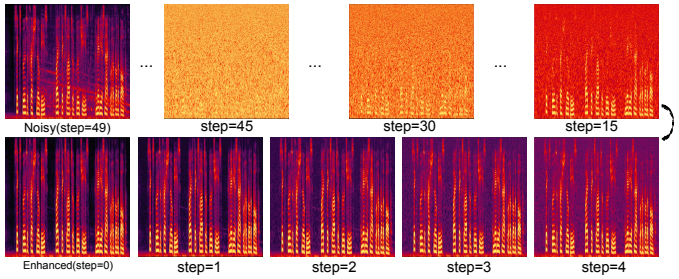


Fig. 5. Visualization of the iterative reverse stage. The spectrograms of specific time steps are given, showing the importance of the last five steps for speech restoration.

can be interpreted as redundancy in the current category labeling of background noise, which inspires us to explore more fine-grained noise encoding. Our proposed NADiffuSE and NADiffuSE+ both perform better than DiffuSE [26] and CDiffuSE [23]. But this way of diffusion-based SE models still lag behind spectral domain methods such as [22], which uses the stochastic differential equation to formulate the diffusion process. Furthermore, we also conduct out-of-domain experiments to validate NADiffuSE's generalization ability for unseen noise scenes compared with baselines. We simulated the new dataset using VoiceBank [40] and Noisex-92 with the same setting in Section IV-A1. Results in Table V show that our method achieves better scores than other time-domain models under the GPC structure.

2) Variants: We compare the performance of the three network variants described in the previous paper and all our inferences are based on the improved algorithm with the best parameter setting, $r = 0.1, t = 5$. All variants are based on 96%-NADiffuSE+. As shown in Table IV, the coarse-and-refine approach achieves the best performance. The coarse-and-scratch can get a small performance gain with only one training stage required. The coarse-and-finetune performs comparably to the originally proposed model, with no significant performance gains, which may further need a well-selected pre-trainig checkpoint and fine-tuning strategy. We can gain insight from row 3 that multi-stage iterative speech enhancement can help recover speech details better, and generative diffusion models have great potential for detail refinement in data reconstruction. All above network variants essentially rely on the training of respective sub-modules, so the quality of the pre-enhanced spectrogram feature would limit the final SE performance.

TABLE V
RESULTS OF UNSEEN NOISE SCENES FROM THE SIMULATED DATASET WITH VOICEBANK AND NOISEX-92.

Method	CSIG	CBAK	COVL	PESQ
Unprocessed	2.84	2.75	2.21	1.59
DiffuSE	2.92	2.69	2.40	1.89
CDiffuSE	2.97	2.72	2.42	1.91
NADiffuSE+	**3.04**	**2.73**	**2.46**	**1.92**

V. CONCLUSIONS

We summarize and analyze the current diffusion-based speech enhancement methods, where in the setting of generator-plus-conditional architecture (GPC), we propose a noise-aware diffusion-based SE model (NADiffuSE), which conducts denoising under the global guidance of noise encodings to help the non-Gaussian noise estimation. To reduce the additional noise introduced by the original interpolation operation, we propose the anchor-based inference algorithm to to complement speech details and reduce the residual noise. Plus, we investigate three variants of NADiffuSE which use the preprocessing network to enhance the Mel spectrogram in advance, to further bridge the gap in the performance bounds. Through experiments, we have shown that our model performs better than other diffusion-based SE models under the GPC structure.

ACKNOWLEDGMENT

This paper was partially supported by NSFC (No:62176008) and Shenzhen Science&Technology Research Program (No: GXWD20201231165807007-20200814115301001).

REFERENCES

[1] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[2] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.

[3] Z. Zhou, *Machine learning*. Springer Nature, 2021.

[4] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[5] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, Springer, 2005, pp. 181–197.

[6] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 7092–7096.

[7] D. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[8] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2010.

[9] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, pp. 3642–3646, 2017.

[10] Y. Bando, M. Mimura, K. Itoyama, . Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 716–720.

[11] M. Strauss and B. Edler, "A flow-based neural network for time domain speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5754–5758.

[12] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, and A. Sengupta B.and Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[13] D. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[14] D. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[15] L. Ruthotto and E. Haber, "An introduction to deep generative modeling," *GAMM-Mitteilungen*, vol. 44, no. 2, e202100008, 2021.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[17] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, PMLR, 2015, pp. 2256–2265.

[18] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[19] N. Chen, Y. Zhang, H. Zen, R. Weiss, and e. a. Norouzi M, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[20] D. Yang, J. Yu, H. Wang, W. Wang, and e. a. Weng C, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[21] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," *arXiv preprint arXiv:2203.17004*, 2022.

[22] J. Richter, S. Welker, J. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation

with diffusion-based generative models," *arXiv preprint arXiv:2208.05830*, 2022.

[23] Y. Lu, Z. Wang, S. Watanabe, A. Richard, and e. a. Yu C, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7402–7406.

[24] J. Serrà, S. Pascual, J. Pons, R. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[25] J. Zhang, S. Jayasuriya, and V. Berisha, "Restoring degraded speech via a modified diffusion model," in *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, International Speech Communication Association, 2021, pp. 2753–2757.

[26] Y. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, pp. 659–666.

[27] Z. Qiu, M. Fu, Y. Yu, L. Yin, and e. a. Sun F, "Srtnet: Time domain speech enhancement via stochastic refinement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[28] J. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *arXiv preprint arXiv:2212.11851*, 2022.

[29] R. Sawata, N. Murata, Y. Takida, T. Uesaka, and e. a. Shibuya T, "A versatile diffusion-based generative refiner for speech enhancement," *arXiv preprint arXiv:2210.17287*, 2022.

[30] Y. Xu, J. Du, L. Dai, and C. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[31] F. Deng, T. Jiang, X. Wang, *et al.*, "Naagn: Noise-aware attention-gated network for speech enhancement.," in *Interspeech*, 2020, pp. 2457–2461.

[32] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 676–680.

[33] J. Lee, Y. Jung, M. Jung, and H. Kim, "Dynamic noise embedding: Noise aware training and adaptation for speech enhancement," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 739–746.

[34] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8162–8171.

[35] Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, and e. a. Ermon S, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[36] A. Oord, S. Dieleman, H. Zen, K. Simonyan, and e. a. Vinyals O, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[37] Q. Kong, Y. Cao, T. Iqbal, *et al.*, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[38] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2021, pp. 166–170.

[39] C. Valentini Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *SSW*, 2016, pp. 146–152.

[40] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, IEEE, 2013, pp. 1–4.

[41] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, Acoustical Society of America, vol. 19, 2013, p. 035 081.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Multi-accent pronunciation assessment based on domain adversarial training

Binghuai Lin*, and Liyuan Wang*
* Smart Platform Product Department, Tencent Technology Co., Ltd, China
E-mail: {binghuailin, sumerlywang}@tencent.com

***Abstract*—Second language learners are usually influenced by their first languages and thus produce varied and complex accents difficult for pronunciation assessment. Meanwhile, obtaining sufficient human-labeled data for accented speech is tedious and costly, which limits the robustness and accuracy of pronunciation assessment. In this paper, we propose an end-to-end (E2E) method for multi-accent pronunciation assessment. We utilize a cross-lingual pre-trained acoustic model to ensure the discriminative capability of feature representations for the assessment of different accents. Moreover, to improve the robustness of pronunciation assessment on low-resource or unannotated accented speech, we employ domain adversary training to make representations accent-invariant for assessment. Experimental results show the proposed method outperforms the baselines across multiple accents of English, including Chinese, Korean, German, and Indonesian, in the Pearson correlation coefficient (PCC). When assessing unseen English pronunciation with Japanese accents, it also shows superiority to other baselines.**

Index Terms: pronunciation assessment, multi-accent, multi-lingual pre-trained acoustic model, low-resource accented data, domain adversarial training

I. Introduction

It has been found that the first language (L1) affects the acquisition of the second language (L2) [1][2][3], which is called accent transfer or language transfer [1][4]. Previous work has pointed out that when there are differences between L1 and L2, the learner's L1 knowledge would interfere with the L2, and when the L1 and L2 are similar, the L1 could aid L2 learning [1]. One way to help L2 learners reduce the impact of L1 accent is through Computer-Assisted Pronunciation Training (CAPT). Automatic pronunciation assessment in CAPT can serve as an essential tool to guide L2 learners. Due to data scarcity, there is an increasing need for a generic automatic pronunciation assessment method for multi-accent scenarios, which should not only be based on discriminative representations for different accent transfers, e.g., phonetic transfer, but can also be generalized to novel/unseen accents.

Much effort has been put into the investigation of L2 pronunciation assessment for single L1 accent scenarios. Classical methods normally utilize traditional feature representations such as GOP, which is defined as the posterior probability of the uttered phoneme given the corresponding acoustic segment [5]. GOP was first computed using Gaussian-mixture-model-hidden-Markov-model-based (GMM-HMM) acoustic models [5], and later was optimized based on the DNN-HMM-based acoustic model to improve the performance of CALL with the development of the deep neural network (DNN) [6]. Although it has prevailed in the related studies ever since [7], [8], there are still many limitations in the current DNN-HMM-based GOP computation, and a lot of effort has been put into resolving the problems [9]. There is also works for better utilization of GOP scores. Pronunciation scoring at the word or sentence level is obtained by averaging the phone level GOP scores [6][9]. Although GOP plays an important role in automatic pronunciation assessment, it may not provide sufficient information for the pronunciation proficiency assessment [10], [11] and doesn't perform well across multiple accents of L2 learners [12].

Other than GOP-based methods, many neural network (NN)-based methods have been investigated for better representations for pronunciation assessment [13][14]. Recently, pre-training has gained popularity in different areas such as natural language processing (NLP) and speech processing [10][15]. By fine-tuning the pre-trained representations, many downstream tasks have achieved the state-of-the-art (SOTA) performance with limited amounts of labeled data. A pre-trained model called wav2vec 2.0 was proposed, and it has achieved SOTA performance in many speech applications [15]. It has also been successfully applied to mispronunciation detection and diagnosis (MDD) with connectionist temporal classification (CTC) to decode phoneme sequences [16]. Recently, a model trained on multiple languages called XLSR-53 has been proposed [17], which learns generic cross-lingual representations in an unsupervised way. In this paper, we leverage data from different languages by utilizing the cross-lingual pre-trained representations to assess different kinds of pronunciation accents.

Cross-lingual pre-training is good for generating discriminative feature representations, which can be further enhanced, with carefully-designed training criteria, for generic pronunciation assessment of different seen/unseen accents. One popular method is training the model based on mixed data of different accents. However, it is proved to be incapable of generalizing to unseen accents. Another approach is to utilize multi-task learning (MTL) training criterion, and it has been proven effective in accented speech recognition [18]. Instead of extracting shared representations across multiple accents by MTL, previous work utilizes domain adversarial training (DAT) to extract domain-invariant representations. DAT has been applied successfully to speech recognition and computer vision tasks [19][20]. Inspired by these studies, we adopt DAT to learn accent-invariant representations for multi-accent

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

automatic pronunciation assessment.

In this paper, the proposed multi-accent automatic pronunciation model is based on cross-lingual pre-trained representations for the sake of better discriminative capability of different accents. To improve the zero-shot capability of the model, we further enhance the representations such that they can be accent-invariant through DAT. Experimental results for L2 English learners with different accents, i.e., Chinese, Korean, German, and Indonesian, show the proposed method outperforms the MTL and other baselines. The proposed method shows better zero-shot capability for unseen accents compared with the baselines.

II. Proposed method

Here, we will briefly introduce the network structure for automatic pronunciation assessment, which is similar to the previous work [21]. We will also illustrate the methodologies of pronunciation assessment for multiple accents.

A. Model for pronunciation assessment

The basic structure of the proposed model comprises a phoneme and acoustic encoder shown in Figure 1. The phoneme encoder takes canonical phonemes as input and outputs phoneme representations. The acoustic encoder takes the waveform as input and outputs acoustic representations. The acoustic representations are re-trained by CTC-based phoneme recognition. The weighted acoustic representations, which are calculated based on the attention between each phoneme and the whole acoustic features, are fused with phoneme representations to obtain the final pronunciation score.

B. Multi-accent feature representations

It is well-known that L1 has impact on the acquisition of L2, resulting in different accents of English. It has been found that if there are no or few similarities between the structure of the L1 and L2, the learners will face a lot of problems in L2 acquisition [2]. For example, as there are no vowels like 'DH' (a phoneme in English) in Chinese, Chinese learners tend to substitute it with 'S' or 'Z' (phonemes in Chinese). Different from the previous work utilizing English pre-trained representations [21], to better evaluate the transferred phonetic pronunciation, we utilize cross-lingual speech representations for multi-accent pronunciation assessment, which were pre-trained based on 56 thousand hours of speech data consisting of 53 languages and have been proved very effective for learning generic cross-lingual representations in an unsupervised way [17]. In the following experimental settings, we will show its superiority to the pre-trained model trained based on monolingual data only.

C. Accent-invariant feature representations

To improve the robustness of pronunciation assessment for low-resource or unseen accents, we adopt DAT to learn accent-invariant feature representations. Traditional DAT is composed of a feature extractor, a domain classifier, and a label predictor, where the domain classifier and label predictor share representations from the feature extractor. Gradient reverse layer (GRL) is inserted between the feature extractor and domain classifier [22]. The whole network is optimized through DAT to obtain domain-invariant features. In this paper, we adopt this mechanism for automatic pronunciation assessment. The simplified schematic is shown in Figure 2, where the accent classifier is based on the weighted acoustic representations, and the scoring regressor utilizes the fused multi-modal representations combining phoneme and acoustic representations. Considering different accents mainly depend on acoustic representations, and the pronunciation assessment needs to determine whether the pronunciation comply with the canonical reading text, we share the weighted acoustic representations between the scoring regressor and the accent classifier instead of the fused multi-modal representations. The shared weighted acoustic representations are averaged based on a pooling layer followed by a fully connected layer with SoftMax activation to calculate the final accent classification results.

Specifically, the loss function is defined in Eq. (1). In the equation, $\theta_a, \theta_p, \theta_y$, and θ_d indicate parameters for acoustic representations, phoneme representations, scoring regressor, and accent classifier, respectively. L_y^i and L_d^i are the scoring regression and accent classification loss for the ith sample. β is the hyper-parameter to balance the weight between scoring regression and accent classification. The model parameters are updated by the optimizer such as Adam or stochastic gradient descent (SGD). Different values of β also indicate different training strategies. We perform experiments for multi-task learning MTL ($\beta > 0$) and direct pronunciation assessment ($\beta = 0$). We will illustrate the comparison results for different training strategies.

$$E(\theta_a, \theta_p, \theta_y, \theta_d) = \frac{1}{N} \sum_{i=1}^{N} (L_y^i(\theta_a, \theta_p, \theta_y) - \beta \times L_d^i(\theta_a, \theta_d)) \tag{1}$$

III. Experiments

A. Corpus

We collect L2 English speech from different sources to conduct experiments for multi-accent English learners including Chinese, German, Indonesian, Japanese, and Korean. Data with Chinese accent comes from speechocean762 [23]. It comprises $5,000$ English sentences spoken by 250 Chinese speakers. It is annotated by five experts and scored at three levels, including phoneme level, word level, and sentence level. The sentence-level scores include accuracy, completeness, fluency, and prosody. In this paper, we mainly focus on sentence-level accuracy. Datasets spoken by Indonesian, Japanese and Korean learners are from the International Corpus Network of Asian Learners of English (ICNALE) [24]. They are constructed by L2 English learners in ten countries and regions of Asia as well as native English speakers. There are spoken monologue, spoken dialog, written essays, and edited essays. We utilize the speech from the spoken monologue task, and select data

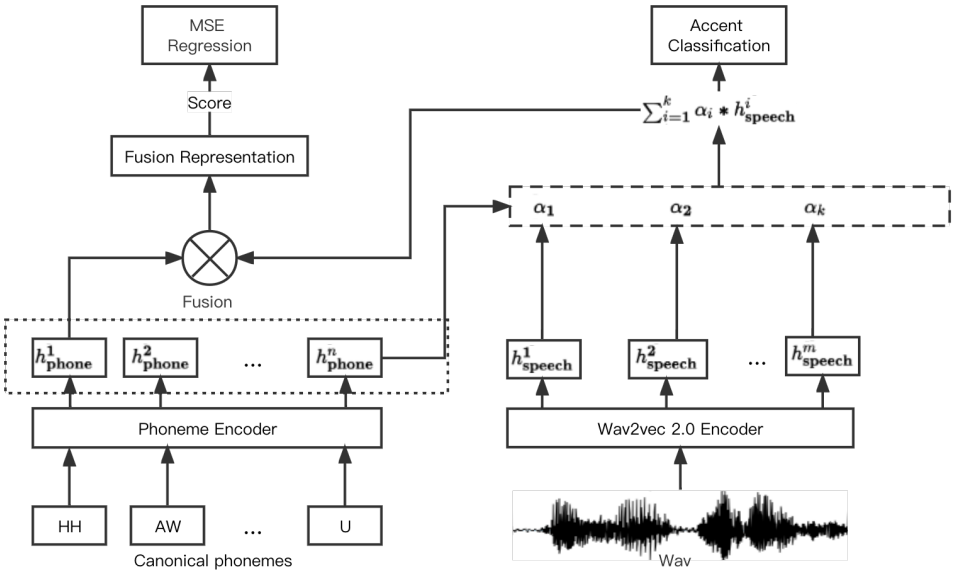


Fig. 1. Model for automatic pronunciation assessment

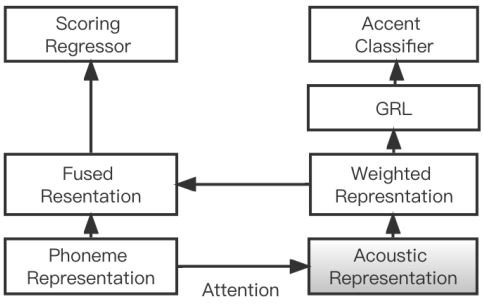


Fig. 2. DAT for automatic pronunciation assessment

of Korean, Japanese, and Indonesian accents for experiments. We cut each long speech into around 5s for the convenience of training and processing. Finally, we obtain $4,000$ utterances for each of the Korean, Japanese, and Indonesian. Data with German accent comes from Dresden Corpus [25], which was based on assessment interviews that lasted 25 minutes composed of three parts in three languages. We select the first part in English and divide it into sentences for testing. After processing, we obtain $4,000$ utterances with German accents. Except for the speechocean762 dataset, the rest datasets from L2 learners are all annotated by three experts and averaged to obtain the final pronunciation scores. The averaged inter-rater calculated by Pearson correlation coefficient (PCC) between the scores of one rater and the average scores of the other two, is 0.8. We scale all scores to the range from 0 to 1 for convenience. The processed multi-accent corpora are summarized in Table I. For Chinese, Korean, German, and Indonesian accents, we split the data into halves for training and testing. The Japanese-accented data are used for testing only to validate the zero-shot capability of the proposed method for unseen/novel accents.

TABLE I
MULTI-ACCENT CORPORA

Accent	Corpus	Train	Test
Japan	ICNALE Corpus	0	4,000
Korean	ICNALE Corpus	2,000	2,000
Indonesia	ICNALE Corpus	2,000	2,000
China	Speech Ocean	2,500	2,500
German	Dresden Corpus	2,000	2,000

B. Experimental setup

We perform experiments using the public Huggingface Transformers [26]. To demonstrate the superiority of multi-lingual representations for multi-accent pronunciation assessment, for multi-lingual representations the acoustic encoder is initialized from wav2vec2-large-xlsr-53 [*], while for mono-lingual representation we initialize it from wav2vec2-large [*]. The phoneme encoder is initialized with the dimensionality of 1024, which is the same as that of acoustic representations and based on the phoneme vocabulary of 39 according to Carnegie Mellon University (CMU) Pronouncing Dictionary [*]. The dimensionality of the fully connected layer followed the fused representations is 1024×1.

We evaluate the proposed method in the Pearson correlation coefficient (PCC), which is calculated between human labels and predicted scores.

[*]https://huggingface.co/facebook/wav2vec2-large-xlsr-53
[*]https://huggingface.co/facebook/wav2vec2-base-960h
[*]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

TABLE II
COMPARISON WITH DIFFERENT BASELINES IN PCC BETWEEN LABELS AND PREDICTED SCORES

Method	Monolingual					Multi-lingual				
	CH	KO	GER	IND	JP	CH	KO	GER	IND	JP
Separate	0.67	0.53	0.52	0.48	0.56	0.68	0.51	0.55	0.5	0.57
Joint	0.67	0.52	0.54	0.50	0.48	0.69	0.54	0.53	0.49	0.50
MTL	0.69	**0.54**	0.54	0.53	0.54	0.7	0.54	0.55	0.52	0.53
DAT (the proposed)	**0.7**	**0.54**	**0.55**	**0.55**	**0.54**	**0.72**	**0.56**	**0.57**	**0.54**	**0.58**

C. Experimental results

First, we will show the comparison results of different baselines. Then, we will visualize phoneme acoustic representations for different accents before and after DAT training to demonstrate its effectiveness.

1) Comparison of baselines: We compare the baselines including separate training with each accented dataset trained separately (Separate), joint training by simply mixing all accented datasets without any additional objective (Joint), MTL training with mixed data and an additional accent classification objective (MTL), and DAT training with additional DAT mechanism as proposed (DAT). Also, as mentioned in Section III-B, we initialize the acoustic encoder with different representations, including monolingual and multi-lingual pre-trained representations. For separate training, we divide the data of each accent into halves for training and testing. For other training strategies, we mix data of different accents for training except for Japanese, which is treated as an unseen accent here. We denote Chinese, Korean, German, Indonesian, and Japanese accents as CH, KO, GER, IND, and JP. The results are shown in Table II.

From the results, we can see there is no significant difference between separate and joint training for the seen accents, while the performance deteriorates significantly for the unseen accent, i.e., Japanese. The MTL training with additional accent classification can improve the performance compared with separate training for the seen accented data, but has similar performance for the unseen accent. When trained by DAT, it obtains the best performance for both seen and unseen accents, indicating its effectiveness and generalizability. Overall, the model initialized from multi-lingual representations performs better than that from monolingual representations, indicating better discriminative capability for multi-accent situations.

2) Visualization of phoneme acoustic representations for different accents: To show effectiveness of DAT, we visualize the acoustic representations before and after the DAT using t-SNE calculated by a sklearn package [27]. Specifically, we extract the weighted acoustic representation for each phoneme based on the attention mechanism between phoneme embeddings and the acoustic representations of the whole utterance shown in Figure 1, which is also the penultimate layer before accent classification. We choose utterances with pronunciation scoring labels higher than 0.7 for each accent, which can be treated as well-pronounced utterances. Then, We randomly select four different phonemes to visualize their acoustic representations. Here, we aggregate the representations of different accents for each of the phoneme together as shown in Figure 3, where blue and green colors indicate representations for DAT and Joint training, respectively. From the figure, we can see for a particular phoneme, the acoustic representations of different accents becomes much more concentrated with DAT training compared with joint training. There are distinct groups of points for joint training, while for DAT there is only one concentrated cluster, indicating DAT makes the representations domain-invariant regardless of accents.

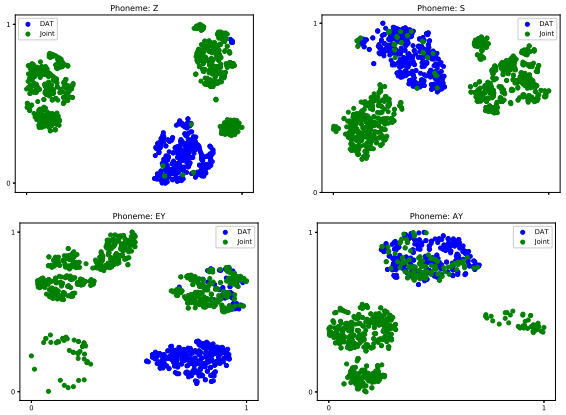


Fig. 3. Visualization of weighted phoneme acoustic representations with different accents by DAT and Joint training

IV. CONCLUSION

In this paper, we propose a multi-accent automatic pronunciation assessment approach considering its generalizability. To take the impact of L1 on the L2 English learners into account, we utilize multi-lingual pre-trained representations to make them more discriminative for assessment of different accents. To improve the zero-shot capability for low-resource or unseen accents, we adopt DAT to learn acoustic representations invariant to different accents. Experimental results based on multi-accent data demonstrate the superiority of the proposed method for seen and unseen accents in PCC.

REFERENCES

[1] L. Ortega, *Understanding second language acquisition*. Routledge, 2014.

[2] A. Derakhshan and E. Karimi, "The interference of first language and second language acquisition," *Theory and Practice in language studies*, vol. 5, no. 10, p. 2112, 2015.

[3] Z. Fu, C.-H. Ji, H. Weiss-Krumm, G. Wang, and Y. Ma, "Chinese-to-english phonetic transfer of chinese university efl students," *Asian Journal of Applied Linguistics*, vol. 7, no. 1, pp. 18–31, 2020.

[4] S. Tahta, M. Wood, and K. Loewenthal, "Foreign accents: Factors relating to transfer of accent from the first language to a second language," *Language and Speech*, vol. 24, no. 3, pp. 265–272, 1981.

[5] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[6] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)." in *Interspeech*, 2013, pp. 1886–1890.

[7] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," in *SLaTE*, 2009.

[8] H. Ryu, H. Hong, S. Kim, and M. Chung, "Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[9] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities," *Proc. Interspeech 2019*, 2019.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[12] S.-W. F. Jiang, B.-C. Yan, T.-H. Lo, F.-A. Chao, and B. Chen, "Towards robust mispronunciation detection and diagnosis for l2 english learners with accent-modulating methods," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1065–1070.

[13] K. Kyriakopoulos, K. Knill, and M. Gales, "A deep learning approach to assessing non-native pronunciation of English using phone distances," pp. 1626–1630, 2018.

[14] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 338–345.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[16] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan, "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis," *Proc. Interspeech 2021*, pp. 4448–4452, 2021.

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[18] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[19] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[20] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[21] B. Lin and L. Wang, "Exploiting information from native data for non-native automatic pronunciation assessment," in *2022 IEEE Spoken Language Technology Workshop (SLT) in press*. IEEE, 2022.

[22] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[23] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[24] S. Ishikawa, "Icnale: the international corpus network of asian learners of english," 2018.

[25] A. Kubanek-German, "Early language programmes in germany," *An early start: Young learners and modern languages in Europe and beyond*, pp. 59–70, 2000.

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

GAN-Based Time-Domain Packet Loss Concealment Method with Consistent Mapping Approach

Yunhao Zhao, Changchun Bao✉, Xue Yang and Jing Zhou
Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing
S202171060@emails.bjut.edu.cn, baochch@bjut.edu.cn, yangx11@emails.bjut.edu.cn, zhoujing@emails.bjut.edu.cn

***Abstract* — In the speech communication over the internet, due to the latency and jitter of the network, the data packets may be lost and result in a degraded user experience. Therefore, the packet loss concealment (PLC) technology plays an important role in recovering the lost information. In this paper, a PLC method only using few historical information is proposed based on generative adversarial network (GAN). In this method, the generator is composed of the improved gated attention unit (IGAU) and temporal convolution network, which can fully exploit the available time-domain information and recover the lost speech signal. Specially, a consistent mapping approach is devised to enhance the quality of the generated speech. In addition, two additional loss functions are used to assist in the training of network. The experimental results show that the proposed method has better performance in terms of both perceptual quality and speech intelligibility at different packet loss rates.**

I. INTRODUCTION

Voice over Internet Protocol (VoIP) is a technology that can package speech according to Transmission Control Protocol/Internet Protocol (TCP/IP) and send data packets to the receiver through the internet. However, due to the latency or jitter of the network, the data packets may be damaged or lost during the transmission [1], which is one of the main reasons for the degradation of speech quality in the VoIP based communication. The target of packet loss concealment (PLC) is to recover the lost speech signal for mitigating the degradation of speech quality caused by the packet loss.

The conventional PLC methods [2-4] simply used the silent speech, history speech and noise to replace the lost speech. In [5], Sanneck et al. proposed the Waveform Similarity Overlap Add (WSLOA) method to recover the lost speech by elongating a segment of speech in front of the lost frame in the time-domain. In [6], Rodbro et al. used the hidden Markov model (HMM) to predict the spectral parameters of the lost speech frame for synthesizing speech in the frequency-domain. These conventional PLC methods can be implemented in real-time because of their low computational complexity. However, it is difficult for these methods to deal with burst packet loss, especially for continuous loss (more than 40ms duration) [7].

With the development of deep learning, many PLC methods have been proposed by fully exploiting the powerful modeling ability of neural network. In [8], Lee et al. used the logarithmic power spectrum and phase spectrum of the previous speech frames as the input feature to train a deep neural network (DNN) model for recovering the lost speech frames. Besides, since the PLC algorithm based on recurrent neural network (RNN) can model the global and local dependencies of speech signal and recover the lost speech in a sequential manner, many network architectures using RNN have been proposed and applied to the task of the PLC, such as the Long Short-Term Memory (LSTM) [9], Wave-RNN [10], Wave Network Equalizer (Wave-Net-EQ) [11] and convolutional recurrent network (CRN) [12]. Meanwhile, the generative adversarial network (GAN) [13] was also widely used in the generative tasks for its parallel calculation ability, which can reduce the inference time. In [14], Shi et al. utilized the speech enhancement generative adversarial network (SEGAN) to conceal the lost speech in the time-domain. Whereas Wang et al. [15] proposed a temporal-spectral generative adversarial network (TFGAN) with a temporal discriminator and a spectral discriminator to distinguish the generated speech and ground-truth speech signal by the waveform and spectral details. Some GAN-based PLC methods [16-19] have obtained good recovery results. However, some of them still need the future information [17, 18], which is harmful for real-time processing.

In this paper, a time-domain PLC method based on GAN is proposed to recover currently lost speech frame by only using eight previous speech frames. Specifically, the generator of the proposed GAN model combines the improved gated attention unit (IGAU) and the temporal convolution network (TCN) to learn the features of the previous eight speech frames at different scales, which can better model the contextual information of lost speech frame. Besides, proposed consistent

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

mapping approach used in GAN network can improve the perceptual quality and intelligibility of the reconstructed speech. Furthermore, two additional loss terms are used to assist in the network training so that the generated speech is more closed to real one.

The remainder of this paper is organized as follows. The PLC model and the related GAN are first reviewed in Section II. Then, the network architecture and mapping approach of the proposed method is detailed in Section III. Subsequently, the experimental setup is given in Section IV and the experimental results are analyzed in Section V. Finally, the conclusions are drawn in Section VI.

II. PLC Model And Related GAN

A. PLC Model

In VoIP, speech signal is usually segmented into the frames and there is no overlap between the frames [17]. Besides, the framed speech signals are formed into the packets for efficient transmission. For simplicity, we assume that each packet contains only one speech frame, i.e.:

$$P_i = \{s_i(n), ..., s_i(n-1+M)\} \tag{1}$$

where $s_i(n)$ represents the speech signal of the i^{th} frame at the time n. P_i refers to the packet consisting of the i^{th} frame speech. M is frame length. When the current packet P_i is lost, the PLC uses the packets consisting of some previous frames to recover the current packet P_i, that is,

$$P_i = F(P_{i-1}, ..., P_{i-N}) \tag{2}$$

where $\hat{P}_i$ is the estimated result of P_i. F represents the nonlinear mapping relationship learned by the network. N denotes the number of the packets consisting of previous frames. Subscript i is the frame index.

B. TFGAN

The TFGAN [15] is a pioneering GAN model for PLC. Its block diagram is illustrated in Fig. 1, which is composed of the time-domain generator and the time-frequency discriminator. The time-domain generator takes a U-Net structure [20] with a temporal convolution network encoder (TCN-Encoder) and a temporal convolution network decoder (TCN-Decoder) [21], which is used to extract the time-domain feature of input speech. To learn the temporal dependencies and enhance the feature correlation of time-domain feature extracted by TCN, the gated linear units (GLU) [22] are used as the bottleneck layer between the encoder and decoder of the U-Net, where the dilation rates in individual GLU are different. In addition, a time-domain discriminator and a frequency-domain discriminator are employed to distinguish the speech generated by the generator from the real speech in both the time-domain and frequency-domain so that the generator could generate more realistic speech as the replacement of the lost speech.

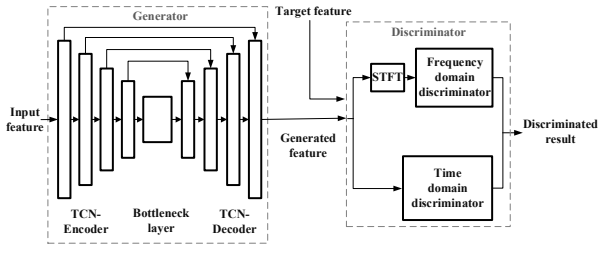


Fig. 1 Block diagram of TFGAN model.

In TFGAN, the mapping approach used during the network training is shown in Fig. 2(a). In the training phase, the input feature consists of N speech frames in front of the lost frame and one zero-padding frame, where speech samples are all set to zero. The target feature is composed of N+1 real speech frames. The TFGAN is trained to learn the mapping relationship between the input feature and the target feature. In the testing phase shown in Fig. 2(b), once the packet loss is detected, the input feature with the same structure as the training phase is fed into the network. Finally, the last frame of the output (N+1) frames is used as the recovered speech frame.

Although the TFGAN-based PLC method has achieved good performance, there is still room for the improvement. Firstly, when GLU is used as the bottleneck layer, the TFGAN could not model the global dependency effectively. Thus, the amplitude of the recovered speech frame may be inaccurate and the partial detail information may be lost. Secondly, during the training phase, the TFGAN not only aims to learn the identity mapping for the first N frames, but also trying to learn a more

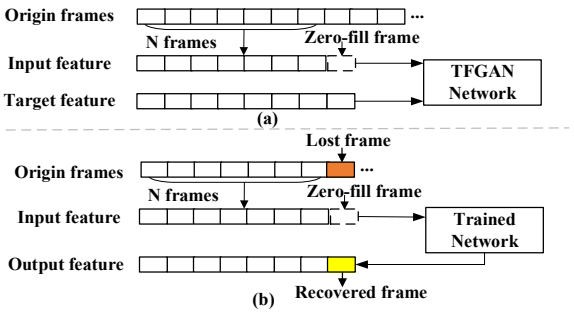


Fig. 2 Schematic diagram of mapping approach in TFGAN during the (a) training phase; (b) testing phase.

complex mapping from the zero-padding frame to the real speech frame for the last one frame. This inconsistent mapping makes the network difficult to be trained, and results in a poor quality to the generated speech waveforms. Therefore, we need to design a more appropriate network which can fully exploit the global dependencies of input speech frames. At the same time, it is necessary to propose a consistent mapping approach to improve the generated speech quality of the network for obtaining better PLC results.

III. Proposed Method

A. Generator with the Proposed IGAU

In TFGAN [15], the generator adopts the U-Net structure, which can aggregate information at different levels and alleviate the gradient vanishing problem through skip connections. However, the GLU is utilized as the bottleneck layer and it mainly consists of convolutional block. Thus, its ability of modeling the global dependency is limited. For the PLC, this may lead to a poor quality of the recovered speech. Recently, the Gate Attention Unit (GAU) [23] was proposed to better model global dependencies by utilizing an attention matrix that weights input feature based on the similarity between features at different positions. In [24], two different modules for modeling global dependency were organized in an interleaved way, which has been proven to be efficient for extracting global dependencies. Based on this idea, the IGAU is proposed in this paper to improve the GAU.

The IGAU module which consists of N_B IGAU is used as the bottleneck layer, which is designed for effectively modeling the global dependencies and obtaining better PLC results. As mentioned above, this module consists of several IGAU stacked sequentially, which is composed of a gated dilated convolution [15, 22] and a Gated Attention Unit. The gated dilated convolution primarily employs a gated mechanism to regulate the flow of information, highlighting the important parts within its input feature. In addition, GAU possesses the capability to model global dependencies by incorporating attention mechanism. Specifically, assume that the input of GAU is $F_{in} \in R^{C\times T}$ and is transformed by two multilayer perceptron (MLP). The obtained feature tensors are denoted as $U \in R^{C\times 2T}$ and $V \in R^{C\times 2T}$. The feature F_{in} is also fed into a single head attention module to obtain the query $Q \in R^{C\times E}$ and the key $K \in R^{C\times E}$, where E denotes the feature dimension of Q and K. These query and key are utilized to obtain the attention weight $A \in R^{C\times C}$, which is expressed as:

$$A = relu^2(Q \bullet K^T + B) \tag{3}$$

where B and $relu^2$ represent the relative positional embedding and squared ReLU activation function, respectively. The superscript T is the transpose operation. Combing the feature tensors U, V and the attention weight A, the output of GAU is expressed as:

$$O = F_{MLP}(U \odot AV) \tag{4}$$

where F_{MLP} denotes the transform achieved by MLP. The symbol $\odot$ represents the Hadamard product and $O \in R^{C\times T}$ denotes the output of GAU.

By stacking several IGAU sequentially, the gated dilated convolution and the GAU are combined alternately and iteratively in the IGAU module. Furthermore, the residual connection is used in proposed IGAU, as illustrated in Fig. 3. This is the first work that employs GAU for the PLC.

B. Consistent Mapping Approach of the Network

In this paper, a consistent mapping approach in the time-domain is proposed. Different from the mapping approach used in TFGAN, this consistent mapping approach effectively models the nonlinear function from each speech frame to its next speech frame, which improves the quality of the recovered speech for the task of PLC. The effectiveness of this consistent mapping approach will be demonstrated in the section V-B.

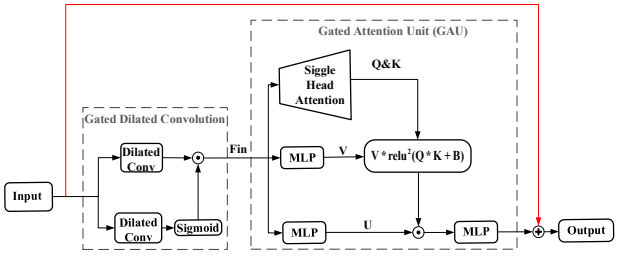


Fig. 3 Diagram of the IGAU.

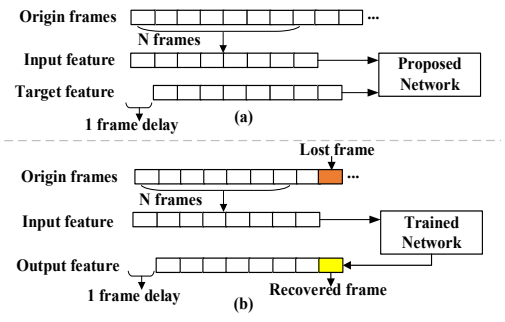


Fig. 4 Schematic diagram of the mapping approach during the (a) training phase; (b) testing phase.

In the training phase shown in Fig. 4(a), the proposed network concatenates the *N* frames in front of the lost frame as the input feature, whereas another *N*-frame set consists of (*N*–1) frames in front of the lost frame and the real value of the lost frame are used as the training target.

During the testing phase, the proposed method detects the packet loss frame-by-frame, and the relationship between the input and output features are illustrated in Fig. 4(b). That is, when a packet loss is detected, the trained network model is fed with the feature consists of *N* frames in front of the lost frame, and the trained model generates an output feature consists of *N* frames too, from which the last one frame is extracted as the recovered speech frame.

C. Loss Function

The original loss function in [13] is difficult to converge with the unbalanced generator and discriminator, resulting in a decrease of the speech quality recovered by PLC algorithm. To address this issue, the least squared loss function for GAN (LSGAN) [25] is used in proposed PLC method. The loss functions of LSGAN are shown in (5) and (6).

$$min_D L(D) = \mathrm{E}_{x\sim p_x}(D(x)-b)^2 + \mathrm{E}_{z\sim p_z}(D(G(z))-a)^2 \tag{5}$$

$$min_G L(G) = \frac{1}{2}\mathrm{E}_{x\sim p_x}(D(x)-c)^2 + \frac{1}{2}\mathrm{E}_{z\sim p_z}(D(G(z))-c)^2 \tag{6}$$

where *x* and *z* are the target feature and input feature as shown in Fig. 4. p_x and p_z are the target distribution and input distribution, respectively. *a*, *b* and c are the constants same as TFGAN.

Apart from the LSGAN, two additional loss functions are introduced to assist the training. The loss functions L_{time} computes the difference between the output feature and the target feature in the time-domain, as shown in (7):

$$L_{time} = \frac{1}{L}\|x-\tilde{x}\|_2^2 \tag{7}$$

where $\tilde{x}$ is the output feature generated by the generator. *L* and $\|\cdot\|_2$ represent the length of the input feature and L_2 norm.

According to the [26], multi-resolution STFT loss can effectively measure the spectral difference between *x* and $\tilde{x}$, as shown in (8):

$$L_{freq} = \frac{1}{R}\sum_{r=1}^{R}\left[\left(\frac{\left\|f^r(x)-f^r(\tilde{x})\right\|_F}{\left\|f^r(x)\right\|_F}\right)+\frac{1}{L}\left\|\log\frac{\left|f^r(x)\right|}{\left|f^r(\tilde{x})\right|}\right\|_1\right] \tag{8}$$

where $|\cdot|$ is modulus operation. $\|\cdot\|_1$ and $\|\cdot\|_F$ are L_1 norm and Frobenius norm, respectively. $f^r(x)$ and $f^r(\tilde{x})$ are the time-frequency representation of the *x* and $\tilde{x}$ at resolution *r*, using short-time Fourier transform (STFT). *R* is the total number of different resolutions. The same STFT configurations of [26] is used in this paper. The final loss functions used in our proposed method are expressed as:

$$L_D^{final} = L_D^{LS} \tag{9}$$

$$L_G^{final} = L_G^{LS} + C_{freq}L_{freq} + C_{time}L_{time} \tag{10}$$

where L_D^{LS} and L_G^{LS}, are the loss functions of LSGAN, as given in (5) and (6). The C_{freq} and C_{time} are the weighting coefficients determined in the experiments.

IV. Experiment Setup

A. Dataset

A total number of 11500 utterances are selected from Librispeech dataset [27] in this paper. Note that only the clean speech in this dataset is chosen to generate the training and testing data in the experiment. All speech signals are re-sampled to 16 kHz. The training set, validation set, and testing set consist of 10000, 1000, and 500 utterances, respectively. The Gilbert-Elliott packet loss model [28] is used to simulate random packet loss with the packet loss rates (PLR) varying from 10% to 40%.

B. Parameter Settings and Training Details

The frame length is set to 320 samples without overlap. The time-frequency discriminator and the parameters of TCN are same with those used in [15]. In the IGAU module, the number of IGAU N_B is set to 5. The dilated rates of 5 IGAU are set to 1, 2, 4, 8 and 16. The channel dimension *C* and time dimension *T* are 128 and 160, whereas the feature dimension *E* of the query and key is 80. The number *N* of the frames used for our proposed mapping approach is 8. The weighting coefficients C_{freq} and C_{time} are 0.5 and 100, respectively. The Adam optimizer is used for optimizing the generator and time-frequency discriminator. The initial learning rates are set to 0.0001 and 0.00005 with beta coefficients of 0.5 and 0.9. The number of the training epoch is set to 80.

V. Experiment Result

In this section, the quality of the reconstructed speech is evaluated through two objective metrics: the wide-band perceptual evaluation of speech quality (WB-PESQ) [29] and the short-term objective intelligibility (STOI) [30].

A. Ablation Study

In order to verify the effectiveness of the proposed IGAU, three different structures (GLU, GAU and IGAU) were used as the bottleneck layers. The mapping approach of ablation study has been described in Section III.B. The result of ablation study is shown in Tab. I and Tab. II.

TABLE I. WB-PESQ RESULTS OF THE ABLATION STUDY

Method	PLR			
	10%	20%	30%	40%
GLU	2.54	1.93	1.63	1.46
GAU	2.59	1.95	1.65	1.48
IGAU	**2.65**	**2.02**	**1.70**	**1.52**

TABLE II. STOI RESULTS OF THE ABLATION STUDY

Method	PLR			
	10%	20%	30%	40%
GLU	0.971	0.949	0.925	0.904
GAU	0.975	0.952	0.931	0.911
IGAU	**0.976**	**0.955**	**0.933**	**0.914**

As can be seen from Tab I and Tab II, the method IGAU, which takes IGAU as the bottleneck layer of U-Net network, could achieve better performance than those using GLU and GAU at different packet loss rates. This implies that the proposed method effectively models the global dependencies of speech signal. This is achieved through the alternation and iteration [24] of gated dilated convolution and GAU in our IGAU module.

B. Comparative Experiments

To evaluate the effectiveness of our proposed PLC method, several methods are used as the references including DNN-based method [8], CRN-based method [12], TFGAN-based method [15] and a variant of TFGAN, $\text{TFGAN}_{\text{proposed}}$, which is trained using the mapping approach described in Section III.B. The WB-PESQ and STOI results are shown in Tab. III and Tab. IV, respectively.

In Tab III, the DNN [8], CRN [12] and TFGAN [15] could recover the lost speech frames to some extent and obtain higher WB-PESQ scores than the impaired speech. Compared with the TFGAN, the variant $\text{TFGAN}_{\text{proposed}}$ achieves a WB-PESQ improvement of 0.26 at a packet loss rate 10% and of 0.1 at other loss rate, which demonstrated the effectiveness of our proposed consistent mapping approach. Compared with the $\text{TFGAN}_{\text{proposed}}$, our proposed network shows another 0.1 improvement at low packet loss rate. Thus, our proposed method achieves the highest performance and its effectiveness is validated. Similar results can be seen for STOI in Tab IV. Note that both results decrease with the increase in the packet loss rate, which is reasonable since more information is lost.

TABLE III. WB-PESQ RESULTS AT DIFFERENT PACKET LOSS RATES

Method	PLR			
	10%	20%	30%	40%
Impaired Speech	1.65	1.24	1.12	1.07
DNN	1.68	1.52	1.36	1.21
CRN	2.01	1.63	1.43	1.29
TFGAN	2.28	1.85	1.55	1.37
$\text{TFGAN}_{\text{proposed}}$	2.54	1.93	1.63	1.46
Proposed	**2.65**	**2.02**	**1.70**	**1.52**

TABLE IV. STOI RESULTS AT DIFFERENT PACKET LOSS RATES

Method	PLR			
	10%	20%	30%	40%
Impaired Speech	0.919	0.840	0.761	0.687
DNN	0.932	0.867	0.825	0.809
CRN	0.941	0.884	0.842	0.827
TFGAN	0.954	0.942	0.913	0.886
$\text{TFGAN}_{\text{proposed}}$	0.971	0.949	0.925	0.904
Proposed	**0.976**	**0.955**	**0.933**	**0.914**

The examples of the recovered voiced frame and unvoiced frame using different methods are depicted in Fig. 5 and Fig. 6, respectively. The voiced frames contain crucial speech information. Their restoration depends on the PLC algorithms. Fig. 5 (c), (d) and (e) show the recovered results of TFGAN, $\text{TFGAN}_{\text{proposed}}$ and our proposed method in the case of losing two frames (40ms), respectively. Compared with the TFGAN, the waveforms recovered by the $\text{TFGAN}_{\text{proposed}}$ and our proposed method are better aligned with the original speech signal. This demonstrates that the mapping approach proposed in Section III.B is more effective. In addition, by analyzing Fig. 5 (d) and (e), it can be observed that the proposed method better preserves the waveform details and amplitude values compared to the $\text{TFGAN}_{\text{proposed}}$, confirming that our proposed network structure can better model the dependencies contained in the speech signal and is more effective than the TFGAN.

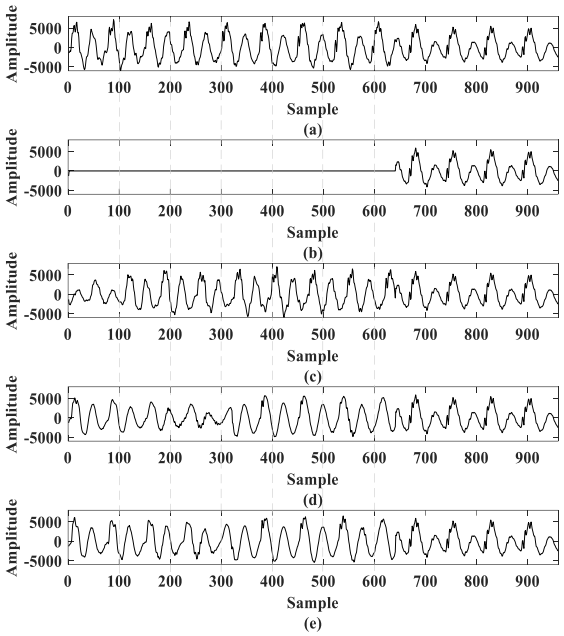


Fig. 5 Voiced speech comparison. (a) Origin speech; (b) Impaired speech; (c) Recovered speech by TFGAN; (d) Recovered speech by $\text{TFGAN}_{\text{proposed}}$; (e) Recovered speech by proposed method.

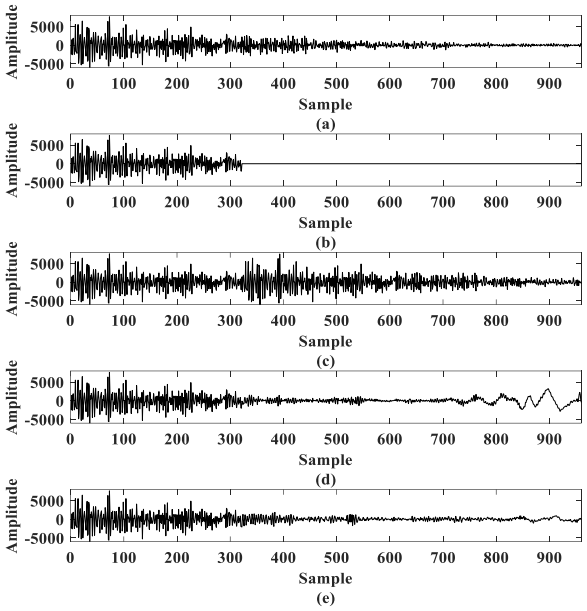


Fig. 6 Unvoiced speech comparison. (a) Origin speech; (b) Impaired speech; (c) Recovered speech by TFGAN; (d) Recovered speech by $TFGAN_{proposed}$; (e) Recovered speech by proposed method.

Unlike voiced speech, unvoiced speech is much more difficult to be recovered. As shown in Fig. 6(c), the restoration of unvoiced speech by TFGAN is poor since the recovered waveform of unvoiced frames is almost the same as the waveform in front of lost speech frames. However, our proposed method could get better restoration in amplitude.

Since our proposed method does not use of any future information, there may be discontinuities at the connection between the recovered frame and its next frame. To address this issue, we plan to design and incorporate smoothing methods without additional latency into the network training to ensure smoother transition between the frames in future work instead of a post-processing.

VI. Conclusions

In this paper, we proposed a novel approach for packet loss concealment, in which only eight historical speech frames in front of the lost frame are used. The TCN and IGAU were combined in the generator for exploiting the available time-domain information so that the amplitude and detail information of lost speech are effectively recovered. In addition, a consistent mapping approach and two additional loss terms were designed to improve speech quality. Experimental results show that the proposed PLC method has the higher performance than the reference methods at different packet loss rates.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 61831019).

References

[1] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," in *IEEE Communications Magazine*, vol. 42, no. 7, pp. 28-34, July. 2004.

[2] N. Jayant and S. Christensen, "Effects of Packet Losses in Waveform Coded Speech and Improvements Due to an Odd-Even Sample-Interpolation Procedure," in *IEEE Transactions on Communications*, vol. 29, no. 2, pp. 101-109, February. 1981.

[3] D. Goodman, G. Lockhart, O. Wasem, and W. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1440-1448, December. 1986.

[4] G. Miller and J. Licklider, "The intelligibility of interrupted speech," *The Journal of the Acoustical Society of America*, vol. 20, no. 2, pp. 167-173, March. 1950.

[5] H. Sanneck, A. Stenger, K. Ben Younes, and B. Girod, "A new technique for audio packet loss concealment," *Proceedings of GLOBECOM'96. 1996 IEEE Global Telecommunications Conference*, London, UK, 1996, pp. 48-52.

[6] C. Rodbro, M. Murthi, S. Andersen, and S. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1609-1623, September. 2006.

[7] B. Liu, Q. Song, M. Yang, W. Yuan, and T. Wang, "PLCNet: Real-time Packet Loss Concealment with Semi-supervised Generative Adversarial Network," in *Interspeech 2022*, pp. 575-579.

[8] B. Lee and J. Chang, "Packet Loss Concealment Based on Deep Neural Networks for Digital Speech Transmission," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 378-387, February. 2016.

[9] R. Lotfidereshgi and P. Gournay, "Speech Prediction Using an Adaptive Recurrent Neural Network with Application to Packet Loss Concealment," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5394-5398.

[10] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N.Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. kavukcuoglu, "Efficient neural audio synthesis," *International Conference on Machine Learning (PMLR)*, 2018, pp. 2410-2419.

[11] F. Stimberg, A. Narest, A. Bazzica, L. Kolmodin, P. Gonzalez, O. Sharonove, H. Lundin, and T. Walters, "WaveNetEQ — Packet Loss Concealment with WaveRNN," *2020 54th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2020, pp. 672-676.

[12] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A Time-Domain Convolutional Recurrent Network for Packet Loss Concealment," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 7148-7152.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courvile, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, October. 2020.

[14] Y. Shi, N. Zheng, Y. Kang, and W. Rong, "Speech Loss Compensation by Generative Adversarial Networks," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, 2019, pp. 347-351.

[15] J. Wang, Y. Guan, C. Zheng, R. Peng, and X. Li, "A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2577-2588, October. 2021.

[16] Y. Guan, G. Yu, A. Li, C. Zheng, and J. Wang, "TMGAN-PLC: Audio Packet Loss Concealment using Temporal Memory Generative Adversarial Network," *arXiv preprint arXiv:2207.01255*, 2022.

[17] N. Li, X. Zheng, C. Zhang, G. Liang, and Y. Bing, "End-to-End Multi-Loss Training for Low Delay Packet Loss Concealment," in *Interspeech 2022*, pp. 585-589.

[18] J. Valin, A. Mustafa, C. Montgomery, T. Terriberry, M. Klingbeil, P. Smaragdis, and A. Krishnaswamy, "Real-time packet loss concealment with mixed generative and predictive model," *arXiv preprint arXiv:2205.05785*, 2022.

[19] S. Pascual, J. Serrà, and J. Pons, "Adversarial Auto-Encoding for Packet Loss Concealment," *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2021, pp. 71-75.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015: 18th International Conference*, Munich, Germany, 2015, pp. 234–241.

[21] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. Dauphin, "Convolutional sequence to sequence learning," *International conference on machine learning (PMLR)*, 2017, pp. 1243-1252.

[22] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *International conference on machine learning (PMLR)*, 2017, pp. 933-941.

[23] W. Hua, Z. Dai, H. Liu, and Q. Le, "Transformer quality in linear time," *International Conference on Machine Learning (PMLR)*, 2022, pp. 9099-9117.

[24] H. Xue, X. Peng, X. Jiang, and Y. Lu, "Towards Error-Resilient Neural Speech Coding," *arXiv preprint arXiv:2207.00993*, 2022.

[25] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. Smolley, "Least squares generative adversarial networks," *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2794-2802.

[26] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-Band Melgan: Faster Waveform Generation For High-Quality Text-To-Speech," *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 492-498.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206-5210.

[28] E. Elliott. "Estimates of error rates for codes on burst-noise channels," in *The Bell System Technical Journal*, vol. 42, no. 5, pp. 1977-1997, September. 1963.

[29] A. Rix, J. Beerend, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, USA, 2001, pp. 749-752.

[30] C. Taal, R. Hendriks, R. Heusdens, and J. Jense, "An algorithm for intelligibility prediction of time frequency weighted noisy speech,". In *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.7, pp. 2125-2136. September. 2011.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

MHM-based reversible data hiding with adaptive feature selection

Shaowei Weng*, Tanshuai Hou† and Haiyang Rao†
* School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China
E-mail: wswweiwei@126.com
† School of Electronic, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou 350108, China
E-mail: htshoutanshuai@foxmail.com

***Abstract*—Considering that traditional multiple histogram modification (MHM)-based reversible data hiding (RDH) methods utilize fixed features to generate multiple prediction-error histograms (PEHs), we propose a novel MHM-based RDH method by adaptively selecting different features with considering the statistical properties of images, so that multiple sharper PEHs are constructed. Moreover, two predictors, namely the rhombus and gradient-based edge direction predictors, are exploited to further improve the prediction performance. The experimental results also demonstrates that the proposed method provides better embedding performance in terms of the visual quality and payload than several related works.**

I. Introduction

Data hiding refers to the process of hiding data into a carrier in an imperceptible manner to convey the presence of embedded data, or to protect the transmitted contents **HongINS2014**. Among various carriers such as images, videos, and audios, digital images are one of the most popular carriers because they are widely transmitted over the Internet. The image used for carrying data is called the cover image while the image with embedded data is called the stego image. When data are embedded into a cover image, the pixels will be changed, and thus the distortion is inevitably introduced. If the pixels are distorted permanently after data embedding, they cannot be recovered to their original states after completely extracting the embedded data. However, for some sensitive applications such as medical or military imaging, any permanent distortion caused by data embedding is unacceptable due to high requirements for legal consideration or high-precision nature. To this end, reversible data hiding (RDH) is introduced to fulfill the requirements by restoring the original image from the stego image along with the extraction of the embedded data. The embedding performance of an RDH method is evaluated using the tradeoff between the embedding capacity versus the image quality.

Early RDH methods mainly achieve data embedding by lossless compression **Fridrich2002EURASIP**, **Celik2005**. A common feature of these methods is to losslessly compress features extracted from the cover image for leaving embedding space; however, the obtained embedding capacity is limited. In 2003, Tian proposed an RDH method based on difference expansion (DE), in which the difference value between two adjacent pixels is expanded to carry 1 bit of data. By expanding difference values to embed data, Tian can achieve the maximum capacity of 0.5 bit per pixel (bpp) but introduces relatively large distortion. Ni *et al.*'s method **Ni2006** proposed to embed data using histogram shifting (HS). The primary idea of HS is to embed data into peak bins of an image histogram while shifting the other bins to leave embedding spaces for data embedding. Since pixels are modified by 1 at most, Ni *et al.*'s method can offer high visual quality. However, the embedding capacity is limited by the peak bin heights. Afterwards, Tian's method are extended from the following three aspects, namely DE **HuTSVT2009**, **TaiTSVT2009**, **LiTIP2011**, prediction error expansion (PEE) **Thodi07** and integer-to-integer transform **Alattar04**, **ColtucSPL2007**, **Weng08**, **WangSPL2010**, **PengSP2012**, **ColtucTIP2012**. PEE is the most effective RDH method owing to its efficient payload-distortion tradeoff. PEE was firstly proposed by Thodi *et al.* **Thodi07**, which performs DE and HS on the prediction errors of pixels generated by a median edge detection (MED) predictor **WeinbergerTIP2000**, rather than difference value of two adjacent pixels in **Tian03**. Subsequently, PEE-based RDH has been widely investigated in research such as two-dimensional (2D) prediction error histogram(PEH) modification **Ou.BIEEE2013**, two-layer embedding **LuoTIFS2008**, **Sachnev09**, **WengIS2016**, **KouhINS2020**. In addition, some PEE-based methods focus on generating sharply distributed prediction error histogram (PEH) by improving predictors such as MED **WeinbergerTIP2000**, gradient-adjusted predictor **FallahpourIEE2008**, rhombus predictor **Sachnev09**, **YangJSS2013**, interpolation technique **LuoTIFS2008**, asymmetric predictor **ChenJSS2013**, **KouhINS2020** and multiple predictors **MaJVCIR2015**. Among the aforementioned techniques, Chen *et al.* **ChenJSS2013** proposed to generate a asymmetric PEH using a predictor based on MED, thus reducing the embedding distortion by decreasing the number of shifted prediction errors. Kouhi *et al.* **KouhINS2020** proposed to generate two asymmetric PEHs using weibull generalized exponential distribution (WGED) **Mustafa2016arXiv**, thereby increasing the visual quality by reducing the number of invalid modifications.

Note that, most existing PEE-based RDH methods are based on one-dimensional (1D) or two-dimensional (2D) PEH. Given that the 1D or 2D PEH modifies all prediction errors uniformly without considering their local properties, Li *et al.* **LiXLTifs2015** proposed an embedding mechanism based

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

on multiple histograms modification (MHM), in which the embedding performance is optimized by adaptively determining two expansion bins (i.e., a pair of expansion bins) for each PEH according to the local properties of each PEH. The optimal expansion bin determination is formulated as a rate-distortion optimization problem. However, directly exploiting exhaustive search to solve the rate-distortion optimization would result in huge computational complexity. To reduce the computational complexity, Li *et al.* utilize exhaustive search to determine two optimal bins for each histogram after decreasing the solution space. Afterwards, MHM has received increasing attention **OuITCSVT18**, **QiTCSVT2020**, **Wang2019Multiple**, **WengIS2021**, **WangTCSVT2021**. Among them, Ou *et al.* **OuITCSVT18** proposed to expand multiple pairs in each PEH for achieving high embedding capacity, rather than expanding a single pair of each PEH like in Li *et al.*'s method **LiXLTifs2015**. Moreover, an advisable expansion bin selection strategy is used to optimize the embedding performance. Differently from the expansion bin selection strategy used in Ou *et al.*'s method **OuITCSVT18**, Qi *et al.* **QiTCSVT2020** proposed a computationally efficient algorithm to adaptively determine optimal multiple pairs for each histogram. Wang *et al.* proposed to utilize the fussy C-means (FCM) clustering with multiple features for constructing multiple sharper PEHs. Weng *et al.* **WengIS2021** proposed to employ K-means clustering with multiple features for constructing multiple sharper PEHs and utilize an improved crisscross optimizing algorithm to determine optimal expansion bins for each histogram. Wang *et al.* **WangTCSVT2021** proposed to construct an RDH general framework based on MHM, which involves several key issues, namely generating multiple PEHs based on optimized multi-features, allocating the embedding capacity among multiple histograms and solving the optimization problem using the evolutionary algorithms.

Different from traditional MHM-based RDH methods that select fixed features of pixels to construct multiple PEHs, this paper focuses on adaptively selecting different features based on the statistical properties of images to construct sharper PEHs. Moreover, two predictors, including the rhombus and gradient-based edge direction predictors, are conductive to generate higher peak points for each PEH.

The rest of this paper is organized as follows. The proposed method is presented in Section II. Section III provides the experimental results and the conclusions are given in Section IV.

II. The Proposed Method

This paper focuses on generating sharper PEHs by adapting various strategies like multiple predictors or adaptive feature selection, thereby reducing unnecessary pixel modifications while improving the embedding performance. Different from traditional MHM-based RDH methods that construct multiple PEHs using fixed features of pixels, the proposed method generates sharper PHEs by adaptively choosing features based on the statistical properties of images, thereby enhancing the embedding performance. The proposed method is mainly composed of the following two parts: two predictors and adaptive feature selection. Subsequently, we will introduce each part separately in the following subsections.

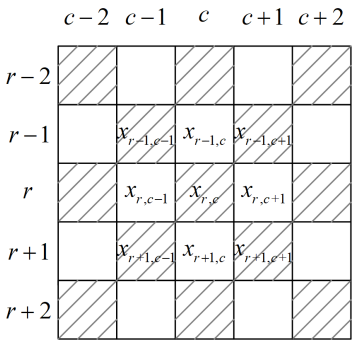


Fig. 1. Two non-overlapping sets Ψ_+ and Ψ_-.

A. Two predictors

A cover image $\mathcal{O}$ is split into two non-overlapping sets Ψ_+ and Ψ_- according to the layout of a chessboard. The pixels in Ψ_+ are used for predicting the pixels in Ψ_- and the prediction errors of the pixels in Ψ_- are embedded with data, and vice versa. Since two sets are processed in the same manner, we only discuss the case of Ψ_+.

Different from traditional RDH methods that exploit an advanced predictor to generate a symmetric PEH, the proposed method utilizes two predictors, including the rhombus predictor **Sachnev09** and gradient-based edge direction predictor (GEDP) **YangJSS2013**, to generate two asymmetric PEHs with long and short tails. Specifically, for a pixel $x_{r,c}$ in Ψ_+, suppose its four neighboring pixels constitute a set denoted by $\mathcal{S}_\mathcal{N}$, namely, $\mathcal{S}_\mathcal{N} = \{x_{r-1,c}, x_{r,c-1}, x_{r,c+1}, x_{r+1,c}\}$. Let two prediction values be $\hat{x}^1_{r,c}$ and $\hat{x}^2_{r,c}$, namely $\hat{x}^1_{r,c} = \lfloor (x_{r-1,c} + x_{r+1,c} + x_{r,c-1} + x_{r,c+1})/4 \rfloor$, and $\hat{x}^2_{r,c}$ is calculated below:

$$\hat{x}^2_{r,c} = \left\lfloor \sum_{s,t\in\Phi} w_{s,t} \times x_{s,t} \Big/ \sum_{s,t\in\Phi} w_{s,t} \right\rfloor, \tag{1}$$

where $\Phi = \{(r \pm 1, c), (r, c \pm 1)\}$, $w_{s,t}$ is the weight assigned for $x_{s,t}$, i.e., $w_{r-1,c} = 1/(1 + |\Delta v_{r,c}| + 2|\Delta v_{r-1,c}| + |\Delta v_{r-2,c}|)$, $w_{r+1,c} = 1/(1 + |\Delta v_{r,c}| + 2|\Delta v_{r+1,c}| + |\Delta v_{r+2,c}|)$, $w_{r,c-1} = 1/1 + |\Delta h_{r,c}| + 2|\Delta h_{r,c-1}| + |\Delta h_{r,c-2}|$, and $w_{r,c-1} = 1/1 + |\Delta h_{r,c}| + 2|\Delta h_{r,c-1}| + |\Delta h_{r,c-2}|$.

$\Delta h_{r,c}$ is defined below:

$$\Delta h_{r,c} = \begin{cases} \frac{3}{2} \sum\limits_{u\in\{\pm1\}} (x_{r+u,c+1} - x_{r+u,c-1}) \\ +\frac{1}{2}(x_{r,c+2} - x_{r,c-2}), \quad \text{if } x_{r,c} \in \Psi_+; \\ \frac{1}{4} \left(\sum\limits_{u\in\{\pm2\}} (x_{r+u,c+1} - x_{r+u,c-1}) \right. \\ \left. + \sum\limits_{u\in\{\pm1\}} (x_{r+u,c+2} - x_{r+u,c-2}) \right) \\ +\frac{5}{2}(x_{r,c+1} - x_{r,c-1}), \quad \text{otherwise.} \end{cases} \tag{2}$$

And, $\Delta v_{r,c}$ is calculated as follows:

$$\Delta v_{r,c}=\begin{cases} \frac{3}{2}\sum\limits_{u\in\{\pm1\}}(x_{r+1,c+u}-x_{r-1,c+u}) \\ +\frac{1}{2}\left(x_{r+2,c}-x_{r-2,c}\right), \text{ if } x_{r,c}\in\Psi_{+}; \\ \frac{1}{4}\left(\begin{array}{l}\sum\limits_{u\in\{\pm2\}}(x_{r+1,c+u}-x_{r-1,c+u}) \\ +\sum\limits_{u\in\{\pm1\}}(x_{r+2,c+u}-x_{r-2,c+u})\end{array}\right) \\ +\frac{5}{2}\left(x_{r+1,c}-x_{r-1,c}\right), \text{ otherwise.}\end{cases} \tag{3}$$

After comparing $\hat{x}^1_{r,c}$ with $\hat{x}^2_{r,c}$, the larger and smaller numbers are denoted by $\hat{x}_{\max}$ and $\hat{x}_{\min}$, respectively. The final prediction value of $x_{r,c}$, denoted by $x^*_{r,c}$, is obtained using the following rule:

$$x^*_{r,c}=\begin{cases}\hat{x}_{\min}, \text{ if } \hat{x}_{\min}==\hat{x}_{max}; \\ \hat{x}_{\min}, \text{ if } x_{r,c}\le\hat{x}_{min}; \\ \hat{x}_{\max}, \text{ if } x_{r,c}\ge\hat{x}_{max}; \\ null, \text{ otherwise.}\end{cases} \tag{4}$$

The prediction error $e_{r,c}$ is calculated as $e_{r,c}=x_{r,c}-x^*_{r,c}$ when $\hat{x}_{min}\ge x_{r,c}$ or $x_{r,c}\ge\hat{x}_{max}$.

B. Adaptive feature selection

To efficiently estimate the local smoothness, 16 features $\{f_1,f_2,\cdots,f_{16}\}$ derived from Weng *et al.*'s method **WengJVCIR2022** are generated for each pixel $x_{r,c}$. 16 features $\left\{f^k_{r,c}\right\}^{16}_{k=1}$ are extracted for $x_{r,c}$, and the k^{th} ($\omega_k\in\{1,2,\cdots,16\}$) feature ω_k is allocated to the k^{th} weight termed ω_k.

$$\ell s(x_{r,c})=\sum_{k=1}^{16}\omega_k\times f^k_{r,c}, \tag{5}$$

where $\ell s(x_{r,c})$ denotes the weighted summation of 16 features.

For the cover image $\mathcal{O}$, the correlation between the absolute value of each prediction error and the corresponding is calculated using the following formula:

$$\begin{aligned}R_{\mathbf{p_e},\mathbf{l_s}}(\mathbf{w})&=\frac{\mathrm{cov}(\mathbf{p_e},\mathbf{l_s})}{\sigma_{\mathbf{p_e}}\times\sigma_{\mathbf{l_s}}} \\ &=\frac{\sum\limits_{e_{r,c}\in\Psi_+}(e_{r,c}-\mu_e)\times(\ell s(x_{r,c})-\mu_{\ell s})}{\sqrt{\sum\limits_{e_{r,c}\in\Psi_+}(e_{r,c}-\mu_e)^2}\sqrt{\sum\limits_{x\in\Psi_+}(\ell s(x_{r,c})-\mu_{\ell s})^2}},\end{aligned} \tag{6}$$

where $\mathbf{w}=\{\omega_k\}^{16}_{k=1}$, μ_e denotes the mean value of the total prediction errors in Ψ_+, $\mu_{\ell s}$ is the mean of the total weighted features in Ψ_+. $\mathbf{p_e}$ is the vector containing all the prediction errors in Ψ_+, and $\mathbf{l_s}$ denotes the vector containing all the weighted features in Ψ_+.

The sequential quadratic programming (SQP) algorithm **Boggs1995** is used to calculate the maximum value of Eq. 6, and therefore, the corresponding 16 weights are obtained accordingly. Each of 16 weights is normalized to a real number in the range of -255 through 255. Subsequently, the absolute values of the 16 normalized weights are sorted in descending order, so that the features corresponding to the first 8 sorted weights are chosen. For some cover image, there exist $C^1_8+C^2_8+C^3_8+C^4_8+C^5_8+C^6_8+C^7_8+C^8_8=255$ combinations of 8 features. 255 combinations are sequentially visited to search for the optimal feature combination for each image, which generates the best power signal-to-noise ratio (PSNR).

After obtaining the optimal feature combination, all the pixels in Ψ_+ are sorted in ascending order of weighted summation of the selected feature combination and then split into 16 classes $\{h_1,h_2,\cdots,h_{16}\}$. The genetic algorithm (GA) is exploited to search for the optimal embedding points for each histogram. The detailed implementation details of GA are mentioned in **ZhangJVCIR2022**.

C. Data hiding

1) The additional information: For the κ^{th} histogram h_κ, the improved discrete particle swarm optimization (IDPSO) **WengJVCIR2022**, rather than the exhaustive search, is used to search for two optimal bins. Suppose that the two optimal bins of h_κ are ℓ^*_κ and γ^*_κ. For blind data extraction and image restoration, the additional information has to be embedded into the cover image. The additional information for each of two non-overlapping sets, i.e., Ψ_+ and Ψ_-, is mainly composed of the following several parts.

- The selected feature combination of 8 features;
- The parameters associated with IDPSO;
- $\{\ell^*_\kappa\}^{16}_{\kappa=1}$ and $\{\gamma^*_\kappa\}^{16}_{\kappa=1}$ for all 16 histograms;
- The compressed location map (LM) used to avoid overflow and underflow. If a pixel in Ψ_+ or Ψ_- is valued at 0 or 255, the corresponding bit in the LM is recorded by 1; otherwise, the corresponding bit is 0. Simultaneously, the pixel valued 0 or 255 is changed to be 1 or 254, respectively.

The additional information is embedded into the cover image along with the payload. The border pixels having no sufficient neighbors to predict, are simply skipped, and their least significant bits (LSBs) are used to carry the additional information. The replaced LSBs are appended to the payload. The detailed data embedding of $e_{r,c}$ in h_κ is given by

$$e'_{r,c}=\begin{cases}e_{r,c}-b, \text{ if } e_{r,c}=\ell^*_\kappa; \\ e_{r,c}-1, \text{ if } e_{r,c}<\ell^*_\kappa; \\ e_{r,c}+1, \text{ if } e_{r,c}>\gamma^*_\kappa; \\ e_{r,c}+b, \text{ if } x_{r,c}=\gamma^*_\kappa,\end{cases} \tag{7}$$

where b is one bit of data and $b\in\{0,1\}$.

D. Data extraction

The additional information must be extracted prior to the payload from the border pixels. With the aid of the additional information, e.g., ℓ^*_κ and γ^*_κ, $e_{r,c}$ is retrieved by

$$e_{r,c}=\begin{cases}e'_{r,c}-1, \text{ if } e'_{r,c}\le\ell^*_\kappa; \\ e'_{r,c}+1, \text{ if } e'_{r,c}\ge\gamma^*_\kappa.\end{cases} \tag{8}$$

Simultaneously, the embedded bit b is extracted as

$$b=\begin{cases}0, \text{ if } e'_{r,c}=\ell^*_\kappa \text{ and } e'_{r,c}=\gamma^*_\kappa; \\ 1, \text{ if } e'_{r,c}=\ell^*_\kappa-1 \text{ and } e'_{r,c}=\gamma^*_\kappa+1.\end{cases} \tag{9}$$

III. Experimental results

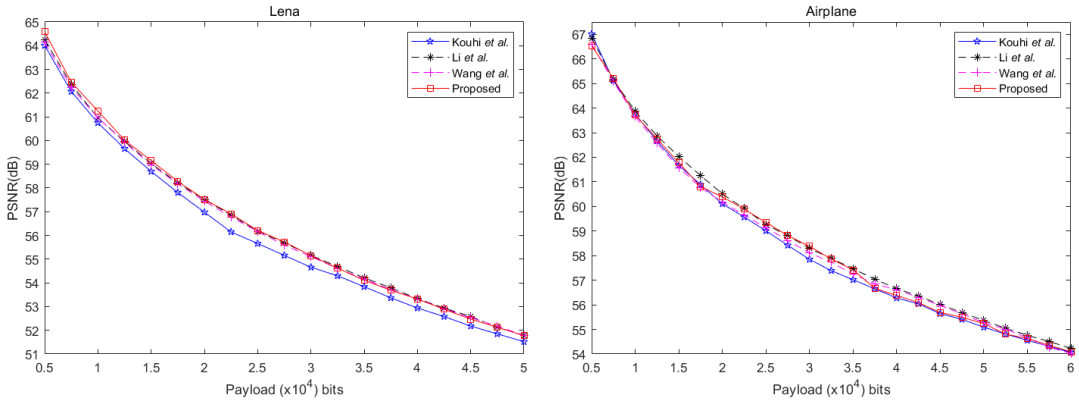
In this section, several experiments are conducted to demonstrate the superiority of the proposed method. The proposed method is compared with four related works including Kouhi *et al.*'s **KouhINS2020**, Li *et al.*'s **LiXLTifs2015**, Wang *et al.*'s methods **WangTCSVT2021**. Eight images taken from USC-SIPI **USID** are utilized as test images. From Fig. 2, it can be clearly observed that the proposed method achieves higher or comparable embedding capacity than the three compared methods.

IV. Conclusion

This paper improves the embedding performing by constructing sharp PEHs from two aspects: two predictors and adaptive feature selection. Unlike fixed feature selection utilized in traditional MHM-based RDH methods, the adaptive feature selection strategy adaptively selects different features for images by considering the statistical characteristics of images. Two predictors are jointly utilized to include the smaller predictor errors into data embedding while excluding larger prediction errors from data embedding. The experimental results also demonstrate the proposed method offer higher or comparable embedding performance than several related works.

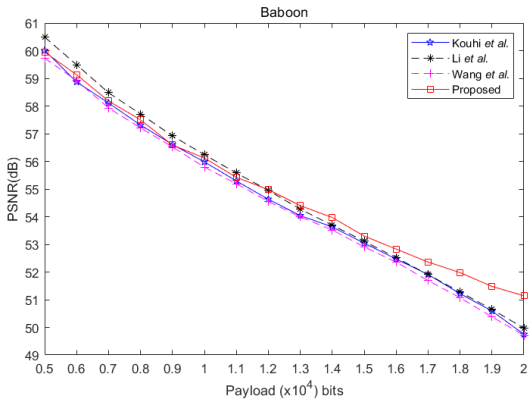
Acknowledgments

This work was supported in part by the National NSF of China under Grant 61872095, Grant 61571139, Grant 61872128, in part International Scientific and Technological Cooperation of Guangdong Province under Grant 2019A050513012, in part by the Open Project Program of Shenzhen Key Laboratory of Media Security under Grant ML-2018-03, in part by Fujian Science Fund for Distinguished Young Scholars under Grant 2020J06043.

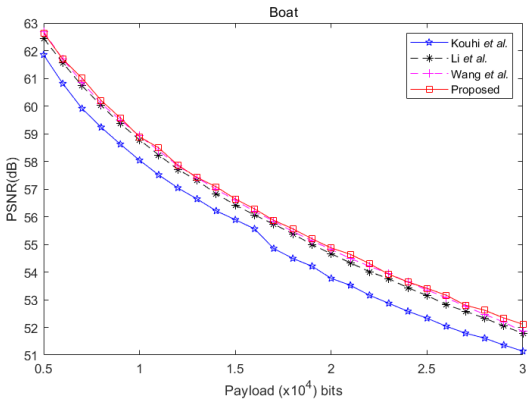


(a)

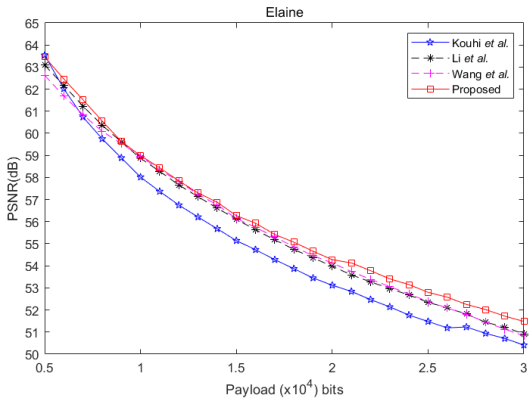
(b)



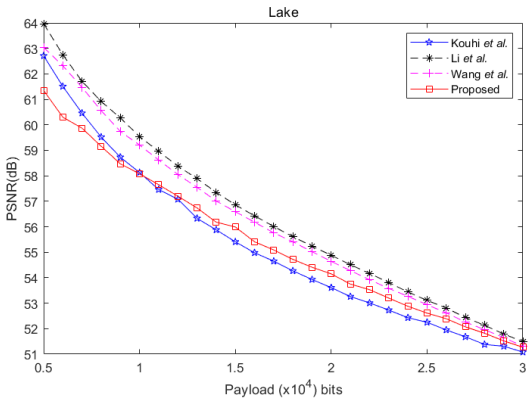
(c)



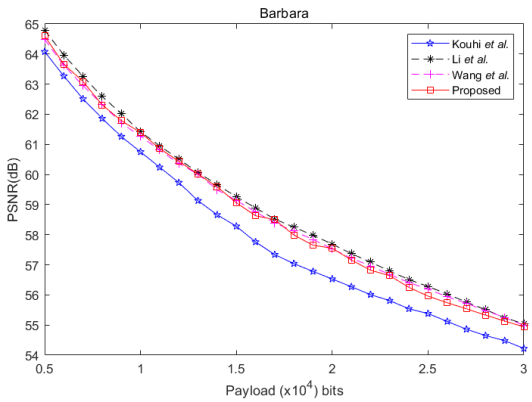
(d)



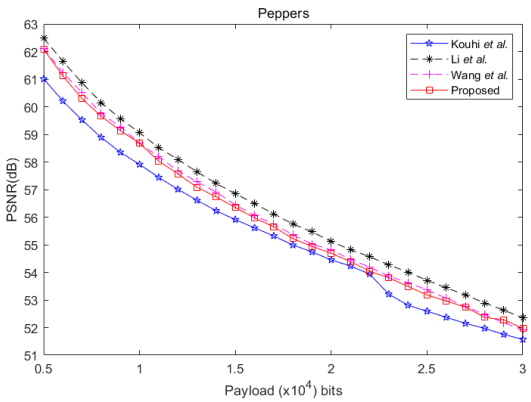
(e)



(f)



(g)



(h)

Fig. 2. Comparison of PSNR versus payload among four methods.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Feature Selection Based on Clonal Selection Algorithm for Image Steganalysis

Yu Liu* and Hongxia Wang†‡
*† School of Cyber Science and Engineering, Sichuan University, Chengdu, China
‡ Zhengzhou Xinda Institute of Advanced Technology, Zhengzhou, China
* E-mail: liuyu_scu@stu.scu.edu.cn † E-mail: hxwang@scu.edu.cn

***Abstract*—With the rapid development of steganography, researches on the adaptive steganography have made steganalysis more and more difficult. In order to improve the detection accuracy, there are plenty of works introducing complex high-dimensional features to steganalysis, among which the rich model steganalysis features are particularly classical. However, there are many redundant features in high-dimensional features, which greatly increase the time and computational cost. How to eliminate redundant features while maintaining detection accuracy is still a research hotspot. In this paper, we propose a novel feature selection method based on Clonal Selection Algorithm (FS-CSA) for image steganalysis, which separates features into multiple populations according to the idea of niche technique, and proposes a feature similarity function to calculate the affinity of features in each population, and finally uses FS-CSA to select features based on the affinity in each population in turn. The experimental results show that FS-CSA can significantly reduce the feature dimensionality while maintaining the detection accuracy, where the Spatial Rich Model (SRM) features can be reduced to about 8000 dimensions, and the Gabor Filter Residual (GFR) features can be reduced to about 7000 dimensions. The crossover experiments show that the steganalysis features selected by FS-CSA are also generalizable to any steganography algorithm.**

I. Introduction

Steganography embeds secret information into redundant parts of cover images, and there is no obvious visual difference between the cover image and the stego image, thus it can achieve the purpose of covert communication. As the opponent of steganography, steganalysis aims to detect steganography signals in an image to determine whether the image contains secret information or not. Steganography has evolved from non-adaptive steganography algorithms, such as LSB (Least Significant Bit) and Jsteg, to adaptive steganography algorithms. The representative adaptive steganography algorithms for the spatial domain are WOW [1], S-UNIWARD [2], MiPOD [3], etc. The representative adaptive steganography algorithms for the JPEG domain are nsF5 [4], J-UNIWARD [2], UERD [5], etc. The development of adaptive steganography algorithms has greatly improved the security of steganography, and brought great challenges to steganalysis, making it difficult to achieve better detection performance with the simple statistic-based features.

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61972269 and 62272331 and Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under Grant HNTS2022003. (Corresponding Author: Hongxia Wang.)

Therefore, there are plenty of works introducing complex high-dimensional features to steganalysis in order to improve detection accuracy. In the spatial domain, Fridrich et al [6] proposed the Spatial Rich Model (SRM). SRM features have high detection performance and are widely used in current image steganalysis. In the JPEG domain, Kodovsky et al [7] who drew on the idea of co-occurrence matrix of SRM proposed JPEG Rich Model (JRM) features. In addition, there are high-dimensional feature steganalysis based on random projection and phase perception. The representative ones are Discrete Cosine Transform Residual (DCTR) [8] and Gabor Filter Residual (GFR) [9]. Compared with the 64 convolution kernels in DCTR, GFR can better capture steganography signals and has better detection performance for adaptive steganography algorithms.

Although high-dimensional steganalysis features have shown better performance in detecting steganography signals, they consume a lot of time and computational cost in performing feature extraction and using classifiers to train models. In addition, there are redundant features in high-dimensional features. Removing redundant features does not degrade detection accuracy, and may even improve it slightly. Therefore, feature selection which can reduce the feature dimensionality while maintaining detection accuracy is an effective method.

Recently, with the rapid development of deep learning, steganalysis methods based on deep learning have been extensively investigated, and a considerable number of deep steganalysis networks (DSNs) have been proposed. Among them, DSNs such as Xu-Net [10], SRNet [11], Zhu-Net [12], SiastegNet [13], and SESNet [14] have achieved desirable detection performance. However, most DSNs, like SRNet, Zhu-Net, focus on building deeper and more complex networks to improve detection performance, which leads to excessive model training time and huge computational cost. Moreover, DSNs also has the problem of unsatisfactory generalization, which leads to the lack of generality when facing complex and changeable images in the real world.

Meanwhile, with the continuous development of computational intelligence, the branch, Clonal Selection Algorithm (CSA), has attracted much attention, The CSA ensures fast convergence to the global optimal solution by using the memory mechanism and has excellent search capability, which has achieved excellent results in pattern recognition, character recognition, and classification problems. Therefore, in this

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

paper, we propose a novel feature selection method based on Clonal Selection Algorithm (FS-CSA) for image steganalysis. The proposed method is able to select discriminative features from the high-dimensional features while maintaining detection accuracy. The main contributions of our proposed method are as follows:

- A novel feature selection algorithm based on Clonal Selection Algorithm for image steganalysis is proposed. And the proposed algorithm achieves about 80% feature reduction for SRM and about 60% feature reduction for GFR while maintaining detection accuracy.
- Since CSA has the ability to find the global optimal solution while converging quickly, FS-CSA can quickly select the feature subset with the highest discriminative power in high-dimensional features, and the feature subset is the optimal subset in high-dimensional features. In general, the distribution of images in the same dataset is similar, so the selected feature subset should be the approximate optimal solution on that dataset. In this paper, extensive experiments also confirm that the feature subset selected by FS-CSA is generalizable to different steganography algorithms.

The rest of this paper is organized as follows. In Section II, we summarize the previous work on feature selection for image steganalysis. In Section III, the design details of our proposed algorithm will be shown. In Section IV, the results of extensive experiments are analyzed and summarized. Finally, this paper is concluded in Section V.

II. Related Work

The purpose of feature selection is to select discriminative features from high-dimensional features, which have better performance for classifying cover images and stego images. Feature selection removes irrelevant and redundant features in high-dimensional features, which are not effective or may even be counterproductive for the classifier in order to reduce feature dimensions.

In recent years, researchers have done a lot of researches on feature selection for image steganalysis. Kaur et al [15] used Mutual Information (MI) to calculate the differentiability of features, and selected features with high differentiability. In addition, adaptive inertia-weight based Particle Swarm Optimization (APSO) was used further with Area Under Curve (AUC) as a fitness function to improve feature reduction. Ma et al [16] proposed a feature selection method based on decision rough set α-positive region reduction. A quantitative method of attribute separability was proposed to evaluate the separability of features, and features were sorted according to the separability. This method could reduce the sorted features while maintaining the detection accuracy. Pathak et al [17] proposed a variant form, levy flight-based gray wolf optimization (LFGWO). The idea of the algorithm is using the collective behavior of wolves to find the optimal solution. Several experiments were conducted on two methods of extracting features using AlexNet and Subtractive Pixel Adjacency Model (SPAM) [18], and the results showed that the algorithm had a faster convergence rate while maintaining detection accuracy. Liu et al [19] proposed a nature-inspired feature selection method using the Binary Bat Algorithm (BBA) based on the idea of echolocation behavior of bats to search for discriminative features, and the selected feature subset was used to train multiple classifiers to improve detection accuracy. Gupta et al [20] proposed an improved algorithm with Fisher linear discriminant classifier, which identified the feature submodels of SRM based on binary particle swarm optimization and local best particle swarm optimization (GLBPSO) to achieve feature reduction.

The Clonal Selection Algorithm (CSA) can converge to the global optimal solution quickly by using the affinity function, clone proliferation and mutation, and memory mechanism. And there is a correspondence in CSA and feature selection. Feature selection is to select discriminative features from high-dimensional features, in order to find the subset of features with the strongest classification ability. In CSA, antibody selection is to select the combination of antibodies with the strongest recognition ability to antigens. Feature selection based on CSA can accelerate the search for the best feature subset and maintain the fast convergence property while enhancing the global search capability. Therefore, we propose a novel feature selection method based on Clonal Selection Algorithm for image steganalysis to find the best feature subset while maintaining detection accuracy.

III. Proposed Method

In this section, we will give the feature selection based on Clonal Selection Algorithm (FS-CSA) for image steganalysis in detail. Firstly, the feature values are preprocessed to facilitate the subsequent population division. Secondly, we will determine the number of iterations, i.e., the number of populations. Then, the affinity of features in each population is calculated based on the feature similarity function. Finally, we will use FS-CSA to perform feature selection and feature update on high-dimensional features. The framework of FS-CSA is shown in Fig. 1.

A. Preprocess the feature value

As we can see from Fig. 1, after finishing feature extraction on the cover $\boldsymbol{X}$ and the stego $\boldsymbol{Y}$, the feature values will be preprocessed. First, considering that there may be regional differences between different images, in order to avoid the subsequent normalization of feature values to offset such differences, the dataset of cover and stego images is randomly disrupted and then equally and non-overlappingly divided into $n\,\{n \in N^{+}\}$ parts. And then for each of them, we preprocess features values to facilitate the subsequent population division. In the experimental stage, the value of n is fixed at 100.

First, we perform a linear normalization operation on the features to avoid the adverse effects of singular samples. The normalization operation is computed by Min-Max Normalization as

$$norm_{fc} = \frac{fc - min_c}{max_c - min_c} \tag{1}$$

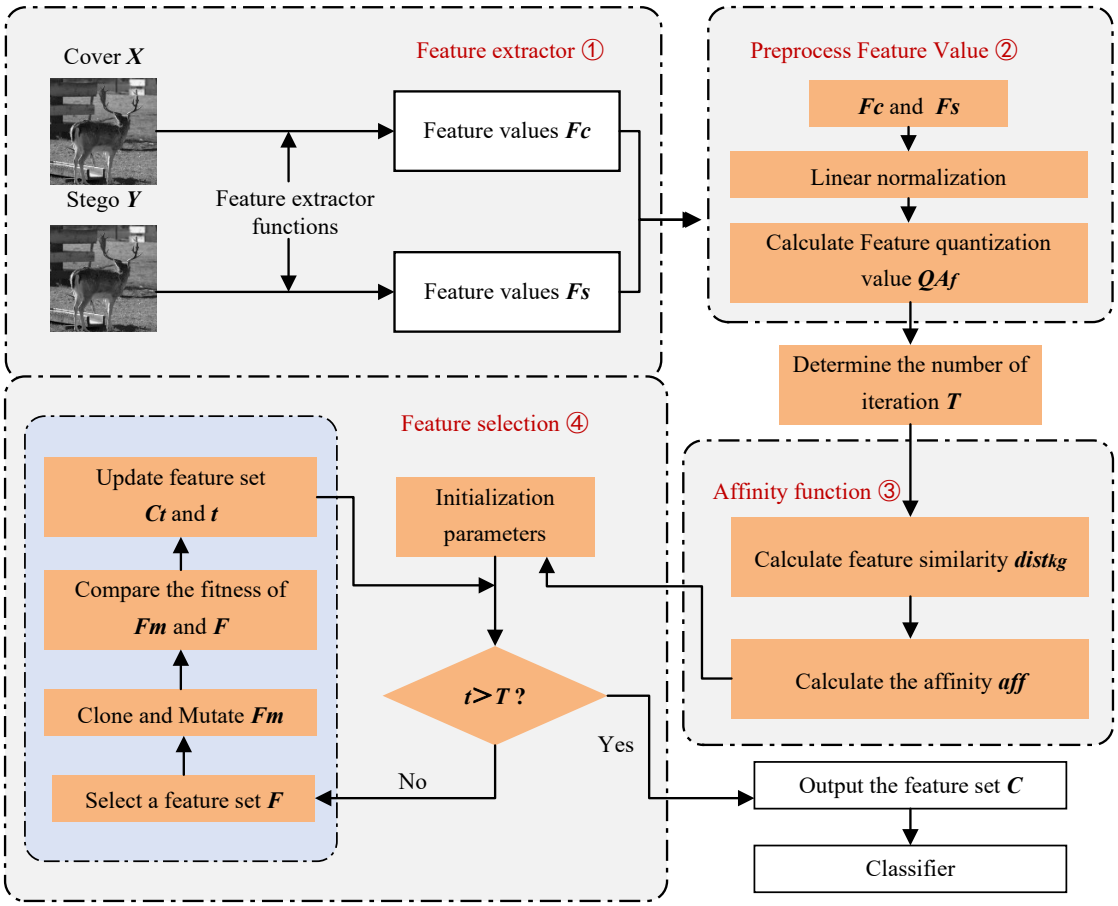


Fig. 1. The framework of FS-CSA.

where the $norm_{fc}$ is obtained by normalizing each feature of $\boldsymbol{X}$. The min_c and max_c denote the minimum value and the maximum value of all $\boldsymbol{X}$ features separately. Similarly, we can get the $norm_{fs}$ by normalizing each feature of $\boldsymbol{Y}$.

Then, according to the core idea of Fisher's Linear Discriminant and the attribute separability measurement (ASM) in [16], the $\boldsymbol{X}$ and $\boldsymbol{Y}$ features are projected onto the X-axis, and the difference value of the intra-class mean of each dimensional feature is used as the quantifier of inter-class dispersion, and the sum value of the intra-class standard deviation of each dimensional feature is used as the quantifier of intra-class aggregation. Finally, the two are divided to obtain the quantified values QA_f of each dimensional feature.

B. Determine the number of iterations

There are large differences between different feature types of feature values and feature dimensions. For example, GFR has 17000 dimensional features while SRM has 34671 dimensional features. In order to avoid a large number of features in the feature subset of different feature types obtained in each round of feature selection, which will have an impact on filtering out redundant features, the number of iterations T is determined dynamically based on the quantified values QA_f as

$$T = \lceil \frac{min(QA_b) - max(QA_l)}{niche} \rceil \tag{2}$$

where the QA_b denotes the maximum value of $n\{n \in N^+\}$ quantified values, and the QA_l denotes the minimum value of $n\{n \in N^+\}$ quantified values. Then, we take the minimum value of QA_b and the maximum value of QA_l. The purpose of this is to ensure that all features of each share can be selected in the subsequent iterations. The $niche$ refers to the iteration radius. In biology, niche refers to a role or organizational function with common characteristics in a specific environment. The niche provides the potential for species diversity and has been introduced into several optimization algorithms. In this paper, niche is used to ensure the diversity of the extracted features. Features can be divided into different groups according to the quantified values to select the excellent features in each group, which can avoid falling into the local optimal solution, and ensure that every feature can be selected. The value of $niche$ will change dynamically because the dispersion of the quantified values for different feature sets

is varied. In the experimental phase, the preset value of $niche$ is 0.008.

C. Calculate the feature affinity

After determining the number of iterations T, calculate the affinity of features which are in a specific niche. In order to filter out the redundant features, we give a feature similarity function as the affinity function to sort the importance of the features in the niche.

For the features of each iteration which are in a specific niche, the similarity between feature k and feature g is computed as

$$dist_{kg} = \sqrt{\sum_{i=1}^{n}(QA_i(k) - QA_i(g))^2}, \quad n \in N^+ \qquad (3)$$

where the similarity is calculated based on the sum of the Euclidean Distance between the quantified values of each feature k and feature g in $n\,\{n \in N^+\}$.

Then, for the calculated $dist_{kg}$, find all features that are smaller than the similarity threshold S_t $\{0 < S_t < 1\}$ and count the number of similar features for each feature. Features with a higher number of similar features are less important, because they can be replaced by other features to achieve the same effect. Therefore, we define the affinity of a feature as the reciprocal of its number of similarities, and the greater the affinity, the better the feature performs in classification.

D. Feature Selection by FS-CSA

After calculating the feature affinity, feature selection is performed on the features of each iteration which are in a specific niche. Select out the features with the highest affinity, and use its similar features as redundant features to form the similar subset until the similar subset is the same as all features which are in the niche. Then, output the subset of features $\boldsymbol{F}$

Feature update is performed on the selected feature subset $\boldsymbol{F}$ in each round of iteration. Feature update mainly includes clonal proliferation and clonal mutation. The purpose of this paper is to reduce features. Thus, this paper only performs clonal mutation on $\boldsymbol{F}$. The mutation operation refers to randomly selecting a certain number of features from the similar feature set of $\boldsymbol{F}$ and adding them to the feature subset to get the mutated feature subset $\boldsymbol{F_m}$. If the classification effect of $\boldsymbol{F_m}$ is better than $\boldsymbol{F}$, the $\boldsymbol{F_m}$ is added to the updated feature subset $\boldsymbol{C_t}$, otherwise, the $\boldsymbol{F}$ is added to the updated feature subset $\boldsymbol{C_t}$. Then, update $\boldsymbol{t}$ and perform feature selection on the features in the other niche. After the finish of the iteration, select the best feature subset $\boldsymbol{C_t}$ in each iteration as the final feature subset $\boldsymbol{C}$. Finally, use the classifier to obtain the detection accuracy of the feature subset $\boldsymbol{C}$.

IV. Experimental Results and Analysis

In order to verify the performance of the FS-CSA proposed in this paper, extensive experiments are performed on SRM and GFR.

A. Experimental setting

All experiments in this paper are performed on the dataset BOSSbase 1.01, which contains 10,000 grayscale images. The image size used in the experiments is 256×256. First, in the spatial domain, S-UNIWARD is used to construct 0.2, 0.4, 0.5, 0.8 and 1.0bpp (bits per pixel) stego images, and WOW, MiPOD are used to construct 0.4, 0.8bpp stego images. In the JPEG domain, J-UNIWARD is used to construct 0.2, 0.4, 0.5, 0.8 and 1.0bpnzAC (bits per non-zero cover AC DCT coefficient) stego images, and nsF5, UERD are used to construct 0.4, 0.8bpnzAC stego images. Specifically, UERD utilizes all DCT coefficients. The quality factor for JPEG domain methods is 95. Thus, 1 group of cover images and 18 groups of stego images are obtained. Then, SRM features (34671-D) and GFR features (17000-D) are extracted by using corresponding extraction algorithms. Finally, in this paper, we adapt the ensemble classifier to classify the images and evaluate the performance of FS-CSA using P_E (minimum total testing error under equal priors). $P_E = \min_{P_{FA}}[P_{FA} + P_{MD}(P_{FA})]/2$ and $P_A = 1 - P_E$, where P_{FA} denotes the false alarm rate and P_{MD} denotes the missed-detection rate, P_A denotes the detection accuracy. During the experiment, 10 repetitions of the experiment are conducted for each group of images, and the average of 10 P_A is taken as the final comparison result. In the experiments, 5000 pairs of images are used to train the classifier, and other 5000 pairs of images are used to test and verify the classification effect.

B. Feature Selection of SRM

SRM features are extracted from stego images which are obtained by S-UNIWARD with different payloads, then use FS-CSA to reduce SRM. Fig. 2 shows the comparison results of the average detection accuracy $\overline{P_A}$ between each iteration and the 34671-dimensional SRM features. As we can see from Fig. 2, SRM features can be efficiently reduced by FS-CSA. Under different payloads, FS-CSA can reduce SRM features to less than 7000 dimensions and to about 3000 dimensions at 0.8bpp while maintaining detection accuracy.

The average detection accuracy $\overline{P_A}$ of SRM features extracted from stego images obtained by MiPOD with different payloads before and after reduction by FS-CSA is shown in Fig. 3, and it can be seen that SRM features can be reduced to about 8,000 dimensions at 0.4bpp and 7,000 dimensions at 0.8bpp by FS-CSA.

The average detection accuracy $\overline{P_A}$ of SRM features extracted from stego images obtained by WOW with different payloads before and after reduction by FS-CSA is shown in Fig. 4, and it can be seen that SRM features can be reduced to about 9,000 dimensions at 0.4bpp and 3,000 dimensions at 0.8bpp by FS-CSA.

Extensive experiments show that FS-CSA can effectively reduce SRM features. From the experimental results, the detection accuracy of the feature subset obtained by FS-CSA can even exceed that of SRM features, which also verifies that there are indeed redundant or irrelevant features in SRM

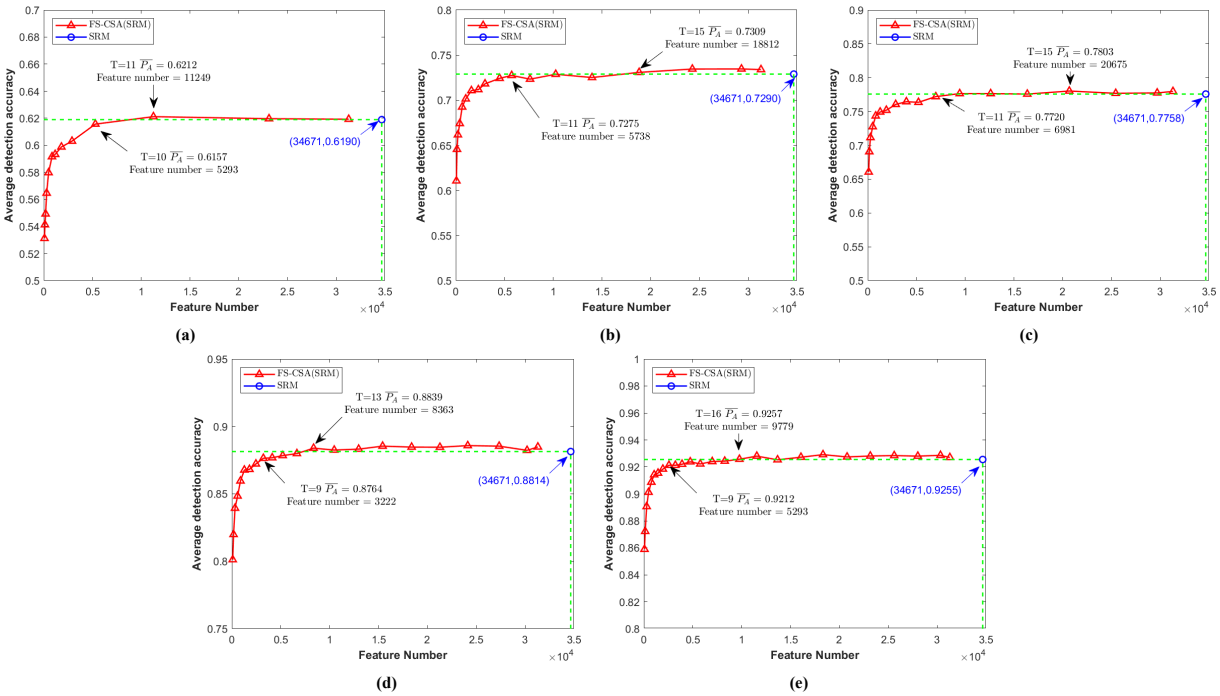


Fig. 2. Average detection accuracy $\overline{P_A}$ of SRM extracted from stego images obtained by S-UNIWARD with different payloads before and after reduction by FS-CSA (a) 0.2bpp, (b) 0.4bpp, (c) 0.5bpp, (d) 0.8bpp, (e) 1.0bpp.

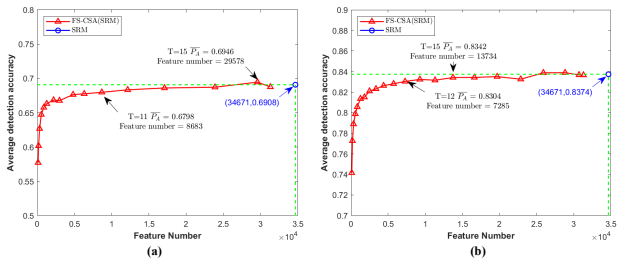


Fig. 3. Average detection accuracy $\overline{P_A}$ of SRM extracted from stego images obtained by MiPOD with different payloads before and after reduction by FS-CSA (a) 0.4bpp, (b) 0.8bpp.

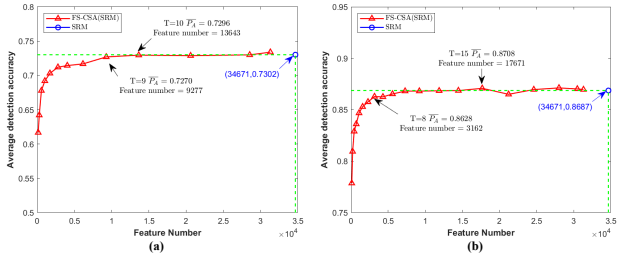


Fig. 4. Average detection accuracy $\overline{P_A}$ of SRM extracted from stego images obtained by WOW with different payloads before and after reduction by FS-CSA (a) 0.4bpp, (b) 0.8bpp.

features, and FS-CSA can effectively delete these features to achieve the purpose of reducing the feature dimension.

C. Feature Selection of GFR

The dimension number of the GFR features extracted by the GFR algorithm is 17,000, and FS-CSA proposed in this paper is used for feature selection of the GFR features. Fig. 5 shows the comparison results of the average detection accuracy $\overline{P_A}$ between each iteration and the 17000-dimensional GFR features. As we can see from Fig. 5, GFR features extracted by J-UNIWARD can be reduced to about 7000 dimensions.

The average detection accuracy $\overline{P_A}$ of GFR extracted from stego images obtained by nsF5 with different payloads before and after reduction by FS-CSA is shown in Fig. 6, and it can be seen that GFR features extracted by nsF5 can be even reduced to about 400 dimensions at 0.8bpnzAC.

The average detection accuracy $\overline{P_A}$ of GFR extracted from stego images obtained by UERD with different payloads before and after reduction by FS-CSA is shown in Fig. 7, and it can be seen that GFR features extracted by UERD can be reduced to about 10,000 dimensions.

The experimental results show that the proposed method can considerably reduce feature dimension while maintaining the detection accuracy.

D. Crossover experiments on feature selection of GFR

In this subsection, a series of crossover experiments are performed on GFR features with the same experimental setup as Section IV-C. First, crossover experiments are performed

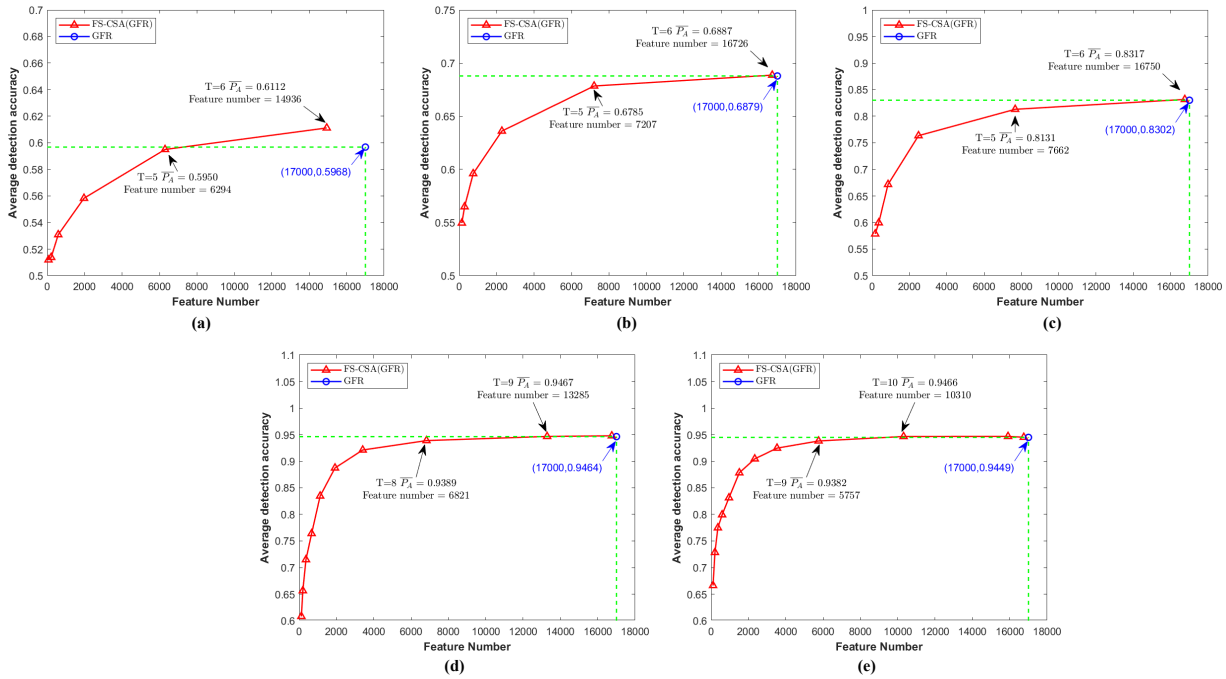


Fig. 5. Average detection accuracy $\overline{P_A}$ of GFR extracted from stego images obtained by J-UNIWARD with different payloads before and after reduction by FS-CSA (a) 0.2bpnzAC, (b) 0.4bpnzAC, (c) 0.5bpnzAC, (d) 0.8bpnzAC, (e) 1.0bpnzAC.

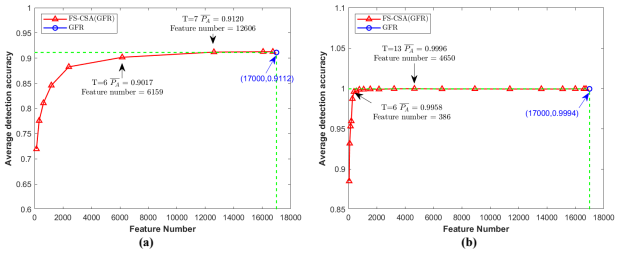


Fig. 6. Average detection accuracy $\overline{P_A}$ of GFR extracted from stego images obtained by nsF5 with different payloads before and after reduction by FS-CSA (a) 0.4bpnzAC, (b) 0.8bpnzAC.

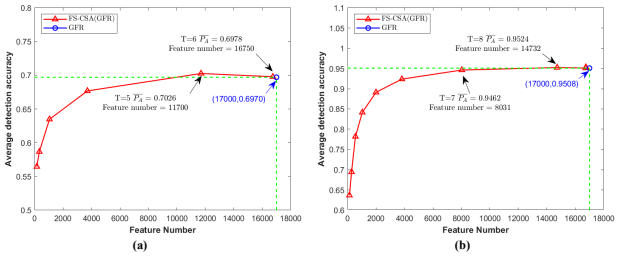


Fig. 7. Average detection accuracy $\overline{P_A}$ of GFR extracted from stego images obtained by UERD with different payloads before and after reduction by FS-CSA (a) 0.4bpnzAC, (b) 0.8bpnzAC.

on GFR features with different steganography methods and the same payload. Then, crossover experiments are performed on GFR features with the same steganography method and different payloads. Finally, crossover experiments are performed on GFR features with different steganography methods and different payloads. Table I shows the comparison results of the average detection accuracy between GFR features and the crossover features obtained by J-UNIWARD at 0.4bpnzAC. As we can see from Table I, the average detection accuracy of the crossover features is considerably close to the average detection accuracy of GFR.

The crossover experiments verify that the discriminative features in GFR features perform similarly under different steganography algorithms with different payloads. It can be proved that the feature subset obtained by J-UNIWARD (J-UNI, 0.4bpnzAC) can be directly applied to the classification detection of another steganography algorithm with different payloads, while maintaining detection accuracy.

V. Conclusion

In this paper, we propose a feature selection method based on Clonal Selection Algorithm for image steganalysis (FS-CSA). Extensive experiments show that the proposed method achieves about 80% feature reduction for SRM and about 60% feature reduction for GFR while maintaining detection accuracy. In addition, the crossover experiments show that the steganalysis features selected by FS-CSA are also generalizable to different steganography algorithm. In the future, the feature selection for steganalysis based on CSA will be further

TABLE I
THE COMPARISON RESULTS OF THE AVERAGE DETECTION ACCURACY $\overline{P_A}$ BETWEEN GFR FEATURES AND THE CROSSOVER FEATURES OBTAINED BY J-UNIWARD AT 0.4BPNZAC.

Average detection accuracy	**Steganography method (payload/bpnzAC)**					
	J-UNI(0.2)	J-UNI(0.8)	nsF5(0.4)	nsF5(0.8)	UERD(0.4)	UERD(0.8)
$\overline{P_A}$ of crossover features	**0.6005**	0.9347	0.9455	0.9993	0.7656	0.9349
$\overline{P_A}$ of GFR features	0.5968	**0.9464**	**0.9641**	**0.9994**	**0.7909**	**0.9508**

investigated, and we will try to combine the CSA with deep learning to improve the detection accuracy.

REFERENCES

[1] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012, pp. 234–239.

[2] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 1, pp. 1–13, 2014.

[3] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 221–234, 2016.

[4] J. Fridrich, T. Pevný, and J. Kodovsky, "Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities," in *Proceedings of the 9th Workshop on Multimedia & Security*, 2007, pp. 3–14.

[5] L. Guo, J. Ni, W. Su, C. Tang, and Y.-Q. Shi, "Using statistical image model for JPEG steganography: Uniform embedding revisited," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2669–2680, 2015.

[6] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[7] J. Kodovsky and J. Fridrich, "Steganalysis of JPEG images using rich models," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8303, 8303A, 2012.

[8] V. Holub and J. Fridrich, "Low-complexity features for JPEG steganalysis using undecimated DCT," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 219–228, 2015.

[9] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang, "Steganalysis of adaptive JPEG steganography using 2D Gabor filters," in *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, 2015, pp. 15–23.

[10] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[11] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2019.

[12] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1138–1150, 2020.

[13] W. You, H. Zhang, and X. Zhao, "A siamese CNN for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 291–306, 2021.

[14] J. Li, X. Wang, Y. Song, and S. Wang, "Image steganalysis via squeeze-and-excitation and separable convolutional neural network," in *The Second International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, 2021, pp. 1–4.

[15] J. Kaur and S. Singh, "Feature selection using mutual information and adaptive particle swarm optimization for image steganalysis," in *The 7th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, 2018, pp. 538–544.

[16] Y. Ma, X. Luo, X. Li, Z. Bao, and Y. Zhang, "Selection of rich model steganalysis features based on decision rough set α -positive region reduction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 336–350, 2019.

[17] Y. Pathak, K. V. Arya, and S. Tiwari, "Feature selection for image steganalysis using levy flight-based grey wolf optimization," *Multimedia Tools and Applications*, vol. 78, no. 2, pp. 1473–1494, 2019.

[18] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[19] F. Liu, X. Yan, and Y. Lu, "Feature selection for image steganalysis using binary bat algorithm," *IEEE Access*, vol. 8, pp. 4244–4249, 2020.

[20] A. Gupta, R. Chhikara, and P. Sharma, "Feature reduction of rich features for universal steganalysis using a metaheuristic approach," *International Journal of Computational Science and Engineering*, vol. 25, no. 2, pp. 211–221, 2022.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

ScaleFormer: Transformer-based speech enhancement in the multi-scale time domain

Tianci Wu*, Shulin He, Hui Zhang, XueLiang Zhang
* College of Computer Science, Inner Mongolia University, China
E-mail: wutianci@mail.imu.edu.cn heshulin@mail.imu.edu.cn cszh@imu.edu.cn cszxl@imu.edu.cn

***Abstract*—Processing speech at multiple temporal scales greatly improves the performance of automatic speech recognition, but its effect has not been fully exploited in speech enhancement tasks. In this study, we propose a novel Transformer-based neural network termed ScaleFormer, which analyzes speech at multiple temporal resolutions. In ScaleFormer, we utilize an encoder that employs multi-scale convolution to extract different temporal scale features. Then, an intra-scale transformer is used to extract the representation within each scale. After obtaining the output of the intra-scale transformer, an inter-scale transformer is used to model the relationship between multiple scales. All transformer block in ScaleFormer is designed with a dual-path framework to learn short and long-term dependencies. We conduct the experiments on the WSJ0 SI-84 corpus. Experimental results show that our approach outperforms previous representative systems in terms of objective metrics.**

I. Introduction

Noise degrades the speech quality and intelligibility. Additionally, it can severely degrade the performance of speech-based applications, such as automatic speech recognition, telecommunication, and hearing aids. The goal of speech enhancement is to restore clean speech from a mixture that is corrupted by noise, thereby improving speech quality and intelligibility. It is typically used as a pre-processing step in speech-related applications.

In recent years, deep neural networks (DNNs) have triggered breakthroughs in speech enhancement, primarily due to their powerful capability to learn complicated relations from supervised data. Conventional DNN-based speech enhancement methods [1]–[3] are based on the time-frequency domain and use short-time Fourier transform (STFT) to extract input features and training targets. However, most of the time-frequency domain approaches aim at enhancing only the spectral magnitude and the noisy phase is employed unaltered for time domain signal reconstruction. A recent study has demonstrated that phase plays an important role in the quality of enhanced speech, especially in low SNR conditions [4]. Consequently, researchers have started to explore methods to enhance magnitude and phase jointly.

Two types of approaches enhance magnitude and phase jointly: complex spectrogram enhancement and time domain enhancement. The complex spectrogram is decomposed into real and imaginary parts in the Cartesian coordinates and jointly enhanced. In this way, both magnitude and phase can be implicitly estimated. The training targets also can be further categorized as complex ratio masking [5], [6] and complex spectral mapping [7], [8].

On the other hand, the time domain methods [9]–[12] aim to predict the clean speech signal from noisy speech directly. In this process, the trouble of estimating phase information in the time-frequency domain is circumvented. As a typical approach in the time domain, the convolutional time-domain audio separation network (Conv-Tasnet) [13] can implicitly encode phase information of speech into a latent representation. The main idea behind Conv-Tasnet is to use convolutions to replace transformations like STFT to extract features directly from the raw speech signal. Based on Conv-Tasnet, SpEx+ [14] employs multiple parallel 1-D CNNs with different kernel sizes to encode the noisy speech into multi-scale speech embeddings. It has been proved that multi-scale encoding can improve the performance of automatic speech recognition [15].

Motivated by SpEx+[14], we propose a novel framework called ScaleFormer for speech enhancement task, mainly consisting of a multi-scale encoder, an intra-scale transformer, and an inter-scale transformer. The multi-scale encoder can extract speech features at multiple temporal scales through three parallel 1-D CNNs with different kernel sizes. The intra-scale transformer is utilized to extract the representation at each certain scale. And the inter-scale transformer can model the relationship between multiple temporal scales. For all the transformer layers of ScaleFormer, we adopt the dual-path framework introduced by DPRNN [16] that can learn both short and long dependencies.

This paper is organized as follows: We first describe the ScaleFormer in the next section. Section 3 describes the experimental details. Results and comparisons are given in Section 4. Section 5 concludes the paper.

II. The Model

The proposed speech enhancement network consists of a multi-scale encoder, a ScaleFormer block, and a decoder, as shown in Fig. 1. The multi-scale encoder is responsible for transforming the mixture waveform into a multi-scale representation. The ScaleFormer block can extract features on a certain temporal scale and model the relationships between different temporal scales. The decoder reconstructs the clean signals in the time domain using the mask predicted by the ScaleFormer block. In the following sections, we will introduce key components and loss functions used in our model.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

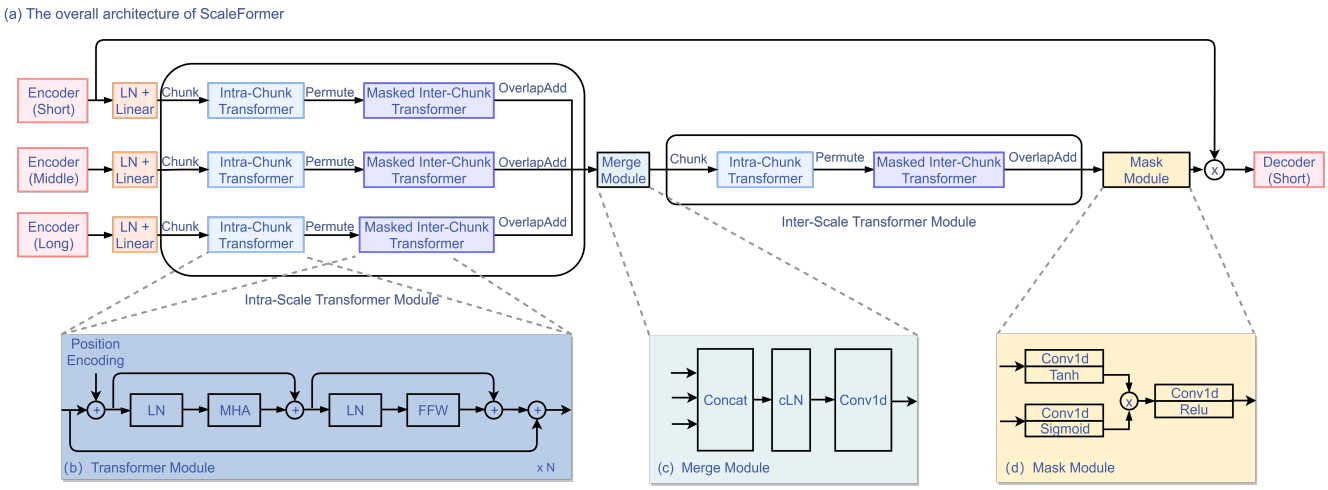


Fig. 1. The architecture of ScaleFormer.

A. multi-scale encoder

Speech contains rich feature information over multiple temporal scales such as phonemic and prosodic so the performance of speech recognition systems can be greatly improved by multiple temporal resolutions [14]. Motivated by SpEx+, we use three parallel 1-D convolutions with different kernel size, l_1 (short), l_2 (middle), l_3 (long) to extract the different temporal scale features. The multi-scale encoder can transform the speech signal $x \epsilon R^T$ directly into a higher level representation, and a rectified linear unit (ReLU) enforced a non-negative constrain, so we have

$$W_{short} = ReLU(Conv1d(x; l_1)) \quad (1)$$

$$W_{middle} = ReLU(Conv1d(x; l_2)) \quad (2)$$

$$W_{long} = ReLU(Conv1d(x; l_3)) \quad (3)$$

where W_{short},W_{middle}, and W_{long} $\epsilon R^{F \times T'}$. l_1, l_2, and l_3 are the different kernel size in the 1-D CNNs, respectively. The outputs of the multi-scale encoder are further processed by layer normalization and a linear layer.

B. ScaleFormer Block

The ScaleFormer Block consists of three modules: an intra-scale transformer module, an inter-scale transformer module, and a masking module. The intra-scale transformer module can perform feature extraction on a certain scale. Afterward, the inter-scale transformer module is utilized to model the relationship between different temporal scale features. At last, a masking module is used to predict a mask for the output of the encoder.

1) Intra-Scale and Inter-Scale Transformer: SepFormer [17] splits the input into smaller chunks and uses a dual-path transformer to process intra-chunk and inter-chunk data. The intra-chunk transformer can model the short-term dependencies by processing each chunk independently and the inter-chunk transformer can model the long-term dependencies across the chunk. The dual-path transformer framework also makes it possible to reduce the quadratic complexity of the transformer by processing smaller chunks.

In this work, both the intra-scale and inter-scale transformers are based on SepFormer, which works on different scales. The intra-scale transformer is responsible for modeling a specific scale representation. Three intra-scale transformers correspond to the outputs of three different scales of the multi-scale encoder respectively. A specific scale feature is fed into a dual-path transformer to extract representation within a certain scale by capturing short-term and long-term dependencies. Fig. 1(b) shows the architecture of the transformer employed in the intra-scale transformer. The building blocks of the transformer are layer normalization, multi-head attention block (MHA), and feed-forward block (FFW). Sinusoidal positional encoding is added to the input, which can introduce position information to the model. The procedure is computed as:

$$w_1 = w + pe \quad (4)$$

$$w_2 = MultiHeadAttention(LayerNorm(w_1)) + w_1 \quad (5)$$

$$w_3 = FeedForward(LayerNorm(w_2)) + w_2 + w \quad (6)$$

where w is the output of the multi-scale encoder and pe is the positional encoding. This producer is repeated N times. To ensure the causality of the system, masked multi-head attention is used in the inter-chunk transformer module, as shown in Fig. 2(b). The output is merged back via the overlap-add method.

Merge Module is employed to preliminarily fuse features of different scales. The architecture is shown in Fig. 1(c). The outputs of the intra-scale transformer are concatenated along the channel dimension, and then cumulative layer normalization (cLN)[13] is applied. Finally, a 1-D convolution is used to adjust the number of channels.

Due to the Merge Module just can fuse features in channel dimension, it cannot model accurately the relationship between different scales. So the outputs of the Merge Module are further fed into an inter-scale transformer to model the relationship between different scales. The inter-scale transformer module

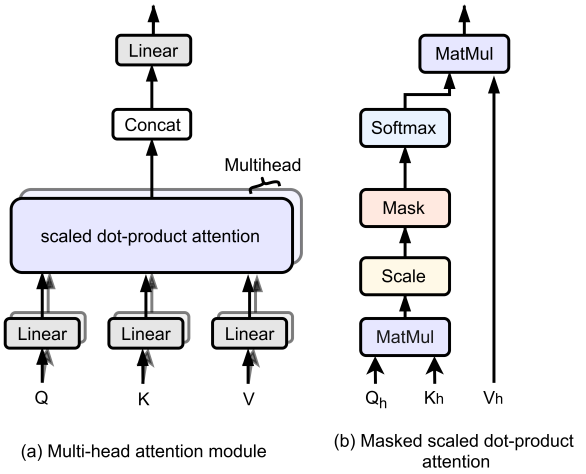


Fig. 2. The architecture of attention [18].

has the same structure as the intra-scale transformer module, but its input aggregates information from all temporal scales.

2) masking module: The structure of the masking module is shown in Fig. 1(d), which uses sigmoid and hyperbolic tangent(Tanh) branches to perform gating and restricts the mask to the [-1,1]. And a ReLu is also employed to enforce a non-negative constraint.

C. Decoder

The decoder merely uses a transposed convolution layer to reconstruct the time-domain signal whose kernel size and stride are the same as the short encoder. Note that we only use one speech decoder to recover speech as estimating a single mask can make the network more focused on synthesizing speech at a specific scale and further reduce the parameters and computational cost of the model.

$$\hat{s} = ConvTranspose1d(M \otimes W_{short}) \tag{7}$$

where $\otimes$ is an operation for element-wise multiplication. M and W_{short} are the mask predicted by the ScaleFormer block and the output of the encoder, respectively.

D. Loss Function

To train the model, we utilize the time-domain loss which is based on the mean square error (MSE) between the denoised speech and clean speech. The time-domain loss is defined as:

$$L\left(s, \hat{s}\right) = \frac{1}{N} \sum_{i=0}^{N-1} \left(s\left(i\right) - \hat{s}\left(i\right)\right)^2 \tag{8}$$

where s and $\hat{s}$ are the sample of the clean speech and the denoised speech, respectively. And N denotes the number of samples in the waveform.

III. Experiment

A. Datasets

In our experiments, we evaluate the performance of our proposed network on the WSJ0 SI-84 dataset [19] which includes 7138 utterances from 83 speakers (42 males and 41 females). Among these speakers, we use the utterances of 77 speakers for training and the rest for the test. we employ 10000 non-speech sounds from a sound effect library (available at www.sound-ideas.com)[20] and generate 320,000 and 3,000 utterances. Specifically, we mix randomly selected training utterances with a random cut from the 10000 training noises at a signal-to-noise ratio (SNR) than is randomly chosen from {-5,-4,-3,-2,-1,0}dB for training and validation, respectively. For the test set, two noises (babble and cafeteria) from Auditec CD (available at http://www.auditec.com) are used to generate 300 mixtures at each SNR of -5dB,0dB, and 5dB.

B. Experimental Setup

In order to facilitate the comparison with SpEx+, we set the kernel size of the convolution in the multi-scale encoder to be the same, l_1 = 2.5ms, l_2 = 10ms, and l_3 = 20ms for the speech of 16kHz sampling rate, respectively, and the stride is half the size of the short convolution kernel (l_1).

The chunk size is 250 with a 50 $\%$ overlap. The number of global and local repeats is R = 2 and N = 4 in ScaleFormer, respectively. We use 2 parallel attention heads and a 256-dimensional feed-forward network within each Transformer layer. The number of channels is 64 and remains unchanged.

All utterances are sampled at 16kHz. The models have trained with Adam [21] optimizer with a learning rate of 0.001. And the batch size is set to 4 at the utterance level. Note that a random segment of 7 seconds is intercepted from an utterance if it is larger than 7 seconds. The smaller utterances are zero-padded to match the size of the largest utterance in the batch. We trained all models for 15 epochs.

C. Baselines

In our experiments, we compare our proposed ScaleFormer with the other baselines, namely CRN [1], GCRN [7], DCCRN-E [6], AECNN [22], and SpEx+ [14]. We keep the best configuration in the original papers.

- CRN: The network is an encoder-decoder structure and 2 LSTM layers are used for sequence modeling. The input of CRN is a magnitude spectrum.
- GCRN: It is a gated convolutional recurrent network for complex spectral mapping. It has a similar structure to CRN except that GCRN has two decoders to model real and imaginary, respectively. The input of GCRN is a complex spectrum.
- AECNN: The network is an autoencoder-based fully convolutional neural network for speech enhancement in the time domain. The input of AECNN is a raw waveform.
- DCCRN-E: It has a similar architecture to CRN but the real-valued convolution is replaced with complex-valued convolution to simulate the complex-valued operation and LSTM layers are not changed. The input of DCCRN-E is a complex spectrum.
- SpEx+: It is a speaker extraction network that extracts multiple temporal resolutions through three parallel CNNs with different filter lengths. In order to apply it in the

TABLE I
STOI AND PESQ COMPARISONS BETWEEN SCALEFORMER AND THE BASELINE MODELS.

Metrics	STOI								PESQ								Parameters
Test Noise	Babble				Cafeteria				Babble				Cafeteria				
Test SNR	-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG	
Mixture	59.1	71.6	83.3	71.3	57.5	70.1	82.4	70.0	1.52	1.76	2.07	1.78	1.40	1.69	2.07	1.72	
CRN	77.9	88.0	93.2	86.4	75.7	86.6	92.7	85.0	1.99	2.50	2.91	2.47	2.01	2.47	2.89	2.46	17.58M
GCRN	81.1	90.7	94.7	88.8	78.0	88.8	94.1	87.0	2.03	2.65	3.07	2.58	1.97	2.56	3.03	2.52	9.76M
AECNN	80.5	90.6	94.2	88.4	80.0	89.4	94.0	87.8	2.00	2.57	2.88	2.48	2.03	2.55	2.93	2.50	18M
DCCRN-E	78.6	89.6	94.7	87.6	77.2	88.5	94.3	86.7	1.98	2.55	2.98	2.50	2.02	2.53	2.97	2.51	3.7M
SpEx+	83.2	91.8	95.2	90.1	79.9	89.8	94.6	88.1	2.15	2.71	3.09	2.65	2.09	2.61	3.06	2.59	9.21M
ScaleFormer	**84.8**	**92.7**	**95.9**	**91.1**	**81.8**	**90.8**	**95.2**	**89.2**	**2.26**	**2.8**	**3.12**	**2.73**	**2.2**	**2.71**	**3.09**	**2.67**	3.3M

TABLE II
STOI AND PESQ SCORE IN THE ABLATION STUDY.

Metrics	STOI								PESQ								Parameters
Test Noise	Babble				Cafeteria				Babble				Cafeteria				
Test SNR	-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG	-5dB	0dB	5dB	AVG	
Mixture	59.1	71.6	83.3	71.3	57.5	70.1	82.4	70.0	1.52	1.76	2.07	1.78	1.40	1.69	2.07	1.72	
ScaleFormer-Inter	80.2	90.1	94.3	88.2	77.3	88.2	93.7	86.4	1.99	2.50	2.87	2.45	1.92	2.42	2.84	2.40	3.4M
ScaleFormer-Intra	84.1	92.3	95.7	90.7	81.1	90.6	95.2	89.0	2.20	2.73	3.10	2.68	2.14	2.65	3.07	2.62	3.3M
ScaleFormer	**84.8**	**92.7**	**95.9**	**91.1**	**81.8**	**90.8**	**95.2**	**89.2**	**2.26**	**2.8**	**3.12**	**2.73**	**2.2**	**2.71**	**3.09**	**2.67**	3.3M

speech enhancement task, we remove the speaker encoder which is designed to extract the speaker embedding of the target speaker from the reference speech. For a fair comparison, we use the MSE loss function instead of SI-SDR in SpEx+ and ScaleFormer.

D. Evaluate Metrics

The performance is evaluated with two objective metrics: perceptual evaluation of speech quality (PESQ)[23] and short time objective intelligibility (STOI)[24]. STOI values typically range from 0 to 1, which can be roughly viewed as percent correct. PESQ values range from -0.5 to 4.5. For both of the STOI and PESQ metrics, the higher number indicates better performance.

IV. RESULT

A. Comparison with Baselines

We show the evaluation results on the WSJ0 SI-84 dataset and compare the performance of our proposed ScaleFormer with baselines. Table I presents the STOI and PESQ score on the test sets and the best result is highlighted by a boldface number. According to the table, we can observe that the ScaleFormer achieves competitive performance compared with other advanced baselines. When compared with GCRN [7], an average improvement of 2.3% and 2.2% for babble and cafeteria is achieved, respectively. For PESQ, ScaleFormer has a 0.15 improvement over GCRN in the case of babble and cafeteria, respectively. Similarly, when compared with the AECNN [22], the STOI is improved by 2.7% and 1.4% and PESQ is improved by 0.25 and 0.17 for babble and cafeteria, respectively.

Next, we compare ScaleFormer with SpEx+. Although both SpEx+ and the proposed ScaleFormer use multi-scale representation in speech enhancement, structures are totally different. SpEx+ is based on TCN. While ScaleFormer is based on the transformer in which we use intra-scale and inter-scale transformers to fully explore the information within each scale and the relationship between different temporal scales. For STOI, ScaleFormer has a 1.0% and 1.1 % improvement over SpEx+ in the case of babble and cafeteria, respectively. And for PESQ, ScaleFormer also has a 0.11 on -5dB improvement over SpEx+ in the case of babble and cafeteria, respectively. This demonstrates the effectiveness of both intra-scale and inter-scale modeling for different scales. We also compare the number of trainable parameters of our proposed network. The proposed network has much fewer parameters, and its results are better than other baselines.

B. Ablation Study

To further demonstrate the validity of our proposed model, we conducted an ablation analysis and present the results in Table II on the test sets. ScaleFormer-Inter represents the utilization of a multi-scale encoder and the exclusion of three intra-scale transformers, including only the inter-scale transformer. On the other hand, ScaleFormer-Intra utilizes multi-scale encoders and three intra-scale transformers but excludes the inter-scale transformers. To ensure a fair comparison, the number of parameters of the three models was kept essentially the same.

After comparing ScaleFormer-Inter, ScaleFormer-Intra, and ScaleFormer, it is evident that ScaleFormer outperforms the other models, demonstrating the importance of modeling both

intra-scale and inter-scale relationships. Specifically, in the case of babble and cafeteria, ScaleFormer shows a 0.4% and 0.2% improvement over ScaleFormer-Intra in terms of STOI and an average improvement of 0.05 over ScaleFormer-Intra in terms of PESQ.

V. Conclusions

In this study, we propose a novel framework for single-channel speech enhancement, which consists of a multi-scale encoder, intra-scale transformer, and inter-scale transformer. The multi-scale encoder can transform the raw speech into a multi-scale representation. The intra-scale transformer can extract the representation at each certain scale and the inter-scale transformer can model the relationship between multiple temporal scales. The results show that the proposed model outperforms other advanced models in terms of objective intelligibility and quality scores.

References

[1] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[2] S. Pirhosseinloo and J. S. Brumberg, "Monaural speech enhancement with dilated convolutions.," in *Interspeech*, 2019, pp. 3143–3147.

[3] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 189–198, 2018.

[4] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[5] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[6] Y. Hu, Y. Liu, S. Lv, *et al.*, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[7] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[8] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6857–6861.

[9] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6629–6633.

[10] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6875–6879.

[11] A. Pandey and D. Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, 2021.

[12] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7842–7846.

[13] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[14] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," *arXiv preprint arXiv:2005.04686*, 2020.

[15] X. Teng, X. Tian, and D. Poeppel, "Testing multi-scale processing in the auditory system," *Scientific reports*, vol. 6, no. 1, pp. 1–13, 2016.

[16] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[17] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 21–25.

[18] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.

[19] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[20] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[23] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality

(pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

UniVR: A Unified Framework for Pitch-Shifted Voice Restoration in Speaker Identification

Yangfu Li* and Xiaodan Lin* †
*School of Information Science and Engineering, National Huaqiao University, Xiamen, China
E-mail: 21013082029@stu.hqu.edu.cn; xd_lin@hqu.edu.cn

***Abstract*—As an efficient spoofing attack, pitch scaling can conceal speakers' identities at a low cost by raising or lowering the pitch in a linear style, making it harmful to the security of speaker identification systems. Several studies have explored the possibility of recognizing pitch-shifted voices and also tried to restore them to their original versions. However, they primarily focus on improving the recognition process, while neglecting the restoration process, resulting in poor acoustic quality in the restored voice. In this paper, we propose a unified framework (UniVR) that includes a transformer-based estimator for recognizing pitch-shifted voices and a specially designed voice restoration network for restoring them in high quality. Experiments on AISHELL-1 and AISHELL-3 datasets demonstrate that UniVR achieves state-of-the-art results compared to current anti-pitch-scaling methods, in terms of both recognition accuracy and restoration quality of pitch-shifted voices. Moreover, UniVR can protect current speaker identification systems from pitch-scaling attacks.**

I. Introduction

Speaker Identification (SI) aims to determine an anonymous speaker's identity from a set of enrolled speakers depending on the similarity between the input utterance and the utterance of enrolled speakers [1]. Due to the recent emergence of SI in high-security-required fields, such as multimedia forensics, there is growing concern about its security risks [2]–[4]. One of the main threats to SI is *pitch scaling*, which enables real-time pitch raising or pitch lowering in a linear style [5], [6]. This process can be formulated as follows:

$$p_s = 2^{\alpha/12} \cdot p_o, \tag{1}$$

where p_o and p_s represent the original pitch and scaled pitch, which are marked in Hz. α is the disguising parameter, which is marked via semitone to describe the degree of disguise. Pitch scaling can be easily implemented via numerous software programs, such as Audacity, Audition, iZotope, etc. Besides, several studies reveal that pitch information plays a crucial role in SI [7]–[9]. Consequently, attackers can exploit pitch scaling to conceal their identities and deceive the advanced SI systems at very low cost. Hence, restoration of the pitch-shifted voice is of great significance for trusted speaker identification and voiceprint forensics in digital crime.

As revealed in Equ 1, the original pitch can be uniquely determined by the disguising parameter. Therefore, restoring pitch-shifted voices typically involves two stages: estimating the disguising parameters and resolving the original voice based on the estimated parameters. However, existing studies primarily focus on developing a robust estimator, while simply performing the opposite-degree pitch scaling to restore the pitch [10]–[14], as shown in Fig. 1. This lack of exploration of the restoration process results in several issues. Firstly, due to the artifacts introduced during pitch scaling [15], [16], the acoustic quality of the speech restored by existing methods through multiple pitch shifting is generally unsatisfactory. Secondly, conventional pitch-scaling algorithms are undifferentiable, making it difficult for current methods to utilize multiple loss information in the restoration process to further supervise the estimator, thus increasing the training difficulty. Furthermore, most current methods directly train the estimator on handcrafted datasets in supervised mode to learn a mapping between the pitch-shifted speech and the disguising parameter. Unfortunately, the scale of handcrafted datasets is generally small, and the patterns of pitch-shifted speech disguised by various algorithms are also different, making it challenging for them to build generalizable estimators.

To address these issues, we propose a unified framework termed *UniVR* for pitch-shifted voice retoration. As shown in Fig. 1, UniVR is mainly composed of two components: a pre-trained estimator for adaptive estimation of disguising

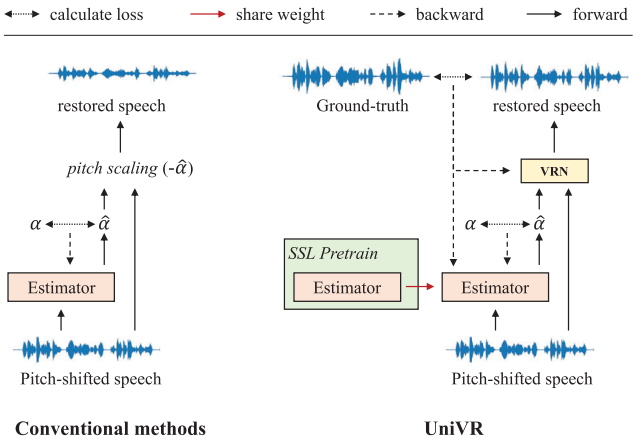


Fig. 1: Comparison of the methods for pitch-shifted voice restoration.

† Coresponding author

This work was supported by the National Natural Science Foundation of China under Grant 62002124 and the Fundamental Research Funds for the Central Universities under Grant ZQN-1112

Code is available on: https://github.com/YChenL/UniVR.

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

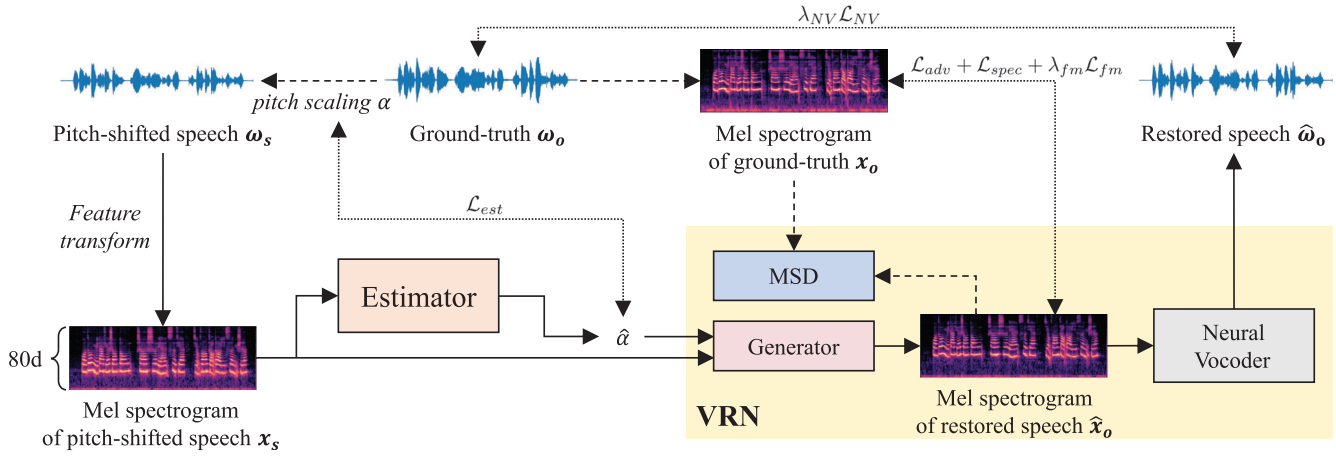


Fig. 2: Overview of UniVR, which includes an estimator and a voice restoration network (VRN). The VRN is composed of three subnetworks: a generator, an associated multi-scale discriminator (MSD), and a neural vocoder. The dashed lines denote the training process.

parameters and a learnable *Voice Restoration Network* (VRN) for high-quality restoration of pitch-shifted voice, which can be jointly optimized end-to-end. Experiments on AISHELL-1 and AISHELL-3 demonstrate that UniVR achieves state-of-the-art results when compared with popular baselines [12]–[14] in terms of both the estimation accuracy of the disguising parameter and the quality of the restored speech. The main contributions of our work are listed as follows:

- To our best knowledge, this study is the first to leverage transformer blocks together with self-supervised learning (SSL) pretrain to enhance the performance and generalization of the estimator trained on handcrafted datasets.
- A Voice Restoration Network (VRN) is elaborately designed to restore the pitch of disguised speech and remove the distortions introduced during pitch scaling, thus improving the acoustic quality of restoration.
- Incorporating the transformer-based estimator and VRN, we build a unified framework (UniVR) for pitch-shifted voice restoration, which can be jointly optimized end-to-end. We conduct numerous experiments on AISHELL1 and AISHELL3k, and the results demonstrate that UniVR achieves state-of-the-art anti-pitch-scaling performance.

II. Related work

Existing methods for disguising parameter estimation on pitch scaling can be divided into two categories: reference-free methods [10]–[13] and reference-based methods [14]. Specifically, [10] builds a support vector machine (SVM) to classify whether the pitch has been shifted based on several handcrafted features of the speech. [11] proposes the first convolutional neural network-based (CNN) estimator for adaptively extracting features from pitch-shifted speech and mapping them to the disguising parameters. In [12], a data augmentation strategy is proposed, which is helpful for improving the generalization ability of the estimator. Later, authors in [13] propose a multi-task learning strategy, which boosts the performance of the estimator by simultaneously identifying pitch-shifted speech and estimating the disguising parameter. On the other hand, [14] proposes a reference-based method. It produces a series of restored samples based on assumed disguising parameters, and then determines the correct disguising parameters from the assumations based on the similarity between the restorations and the original pitch reference.

III. Methodology

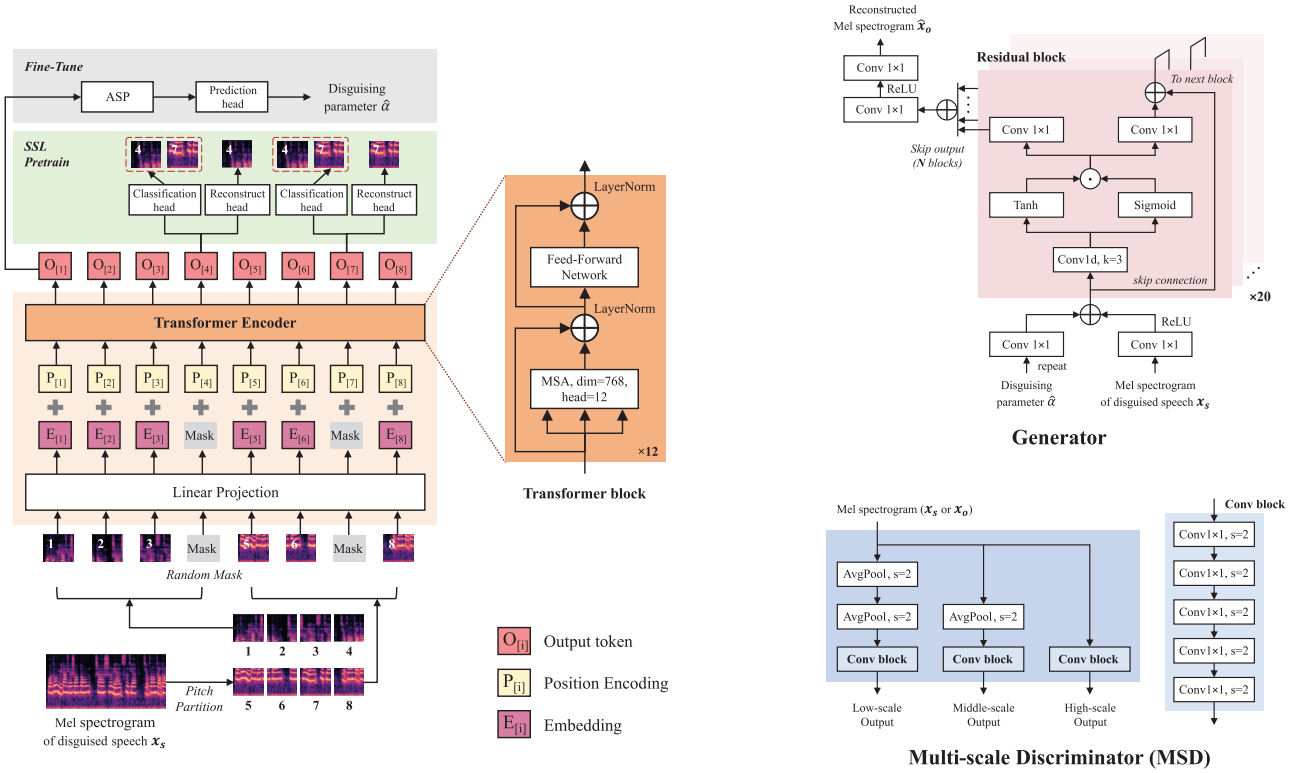
The overall architecture and data pipelines of UniVR are shown in Fig. 2. In general, UniVR is made up of a reference-free estimator and a voice restoration network (VRN), which are jointly optimized end-to-end. We detail these components in subsections *A* and *B*, respectively.

A. Estimator

Recently, many studies reveal the capability of transformers in learning robust speech representations [17]–[19]. Therefore, we develop a transformer-based estimator E that receives 80-dimensional Mel spectrograms of the pitch-shifted speech, i.e., $\boldsymbol{x}_s \in \mathbb{R}^{80\times T}$, as inputs and then outputs the estimation of disguising parameters $\hat{\alpha}$. It can be formulated as follows:

$$\hat{\alpha} = E(\boldsymbol{x}_s). \tag{2}$$

Specifically, the estimator mainly consists of a linear projection and a transformer encoder composed of 12 transformer blocks, as shown in Fig. 3 (a). Firstly, the input is partitioned into a series of 16×16 patches. Subsequently, these patches are flattened and fed to the linear projection layer for feature embeddings. As the standard transformer [20], cosine function is applied to preserve the position information of the embeddings. Finally, an Attentive Statistics Pooling (ASP) [21] together with the prediction head composed of two fully connected layers are applied to yield the estimated disguising parameter according to the output tokens. To ease the training difficulty, we also introduce a novel self-supervised learning (SSL) pretrain term *MSPM* [22], which randomly masks some of the input patches, and then utilizes the estimator to recognize and



(a) Estimator (b) VRN

Fig. 3: Detailed structure of the components in UniVR, where MSA denotes Multi-head Self-Attention, ASP denotes Attentive Statistics Pooling, and Conv1 $\times$ 1 denotes a one-dimensional (1D) convolutional operation with kernel size 1.

reconstruct the masked patches. MSPM can help the estimator learn a strong speech representation that is equally effective for the pitch-shifted samples.

B. Voice Restoration Network

VRN is built on the *Generative Adversarial Network* [23], where a generator G is employed to reconstruct the original-pitch mel spectrogram in high quality, an associated multi-scale discriminator (MSD) D_ϕ is utilized to guide the generator, and a neural vocoder V is applied to transform the reconstructed mel spectrogram into the corresponding waveform.

1) Generator: As shown in Fig. 3 (b), the generator mainly consists of 20 gated residual blocks with 256 channels. It receives the element-wise summation of the estimated disguising parameter $\hat{\alpha}$ and the disguised mel spectrogram $\boldsymbol{x_s}$ as the input and reconstructs the original-pitch mel spectrogram $\hat{\boldsymbol{x}}_o$:

$$\hat{\boldsymbol{x}}_{\boldsymbol{o}} = G(\boldsymbol{x_s}, \hat{\alpha}) \in \mathbb{R}^{80\times T}. \tag{3}$$

2) multi-scale discriminator: MSD aims to distinguish the reconstructed mel spectrogram from the real mel spectrogram. As shown in Fig. 3 (b), MSD can be viewed as a combination of three branches with different receptive fields, where each branch employs five convolutional layers to extract the features of the mel spectrograms. It can simultaneously focus on different scales of temporal information in the speech. It accepts the mel spectrogram as the input and outputs the distribution of input authenticity at different scales:

$$D^i_\phi(\boldsymbol{x}) := [D^i_{ls}(\boldsymbol{x}), D^i_{ms}(\boldsymbol{x}), D^i_{hs}(\boldsymbol{x})], i \in 1, 2, ...5, \tag{4}$$

where $D^i_\phi(\boldsymbol{x})$ is the i-th feature map in D_ϕ. $D^i_{ls}(\boldsymbol{x})$, $D^i_{ms}(\boldsymbol{x})$, and $D^i_{hs}(\boldsymbol{x})$ is the i-th feature map in the low-scale, middle-scale, and high-scale branches, respectively.

3) Neural Vocoder: We directly applied the advanced neural vocoder term HiFiGAN [24] in UniVR to produce high-quality restored speech at very low latency. It accepts the reconstructed mel spectrogram $\hat{\boldsymbol{x}}_{\boldsymbol{o}}$ as the input and outputs the corresponding waveform $\hat{\boldsymbol{\omega}}_{\boldsymbol{o}}$. Formally, this process is written as:

$$\hat{\boldsymbol{\omega}}_{\boldsymbol{o}} = V(\hat{\boldsymbol{x}}_{\boldsymbol{o}}) \in \mathbb{R}^{1\times T}. \tag{5}$$

Notably, the neural vocoder is well trained on CN-Celeb 1&2 [25], [26] and frozen during the training process of UniVR.

C. Objective function

We design a set of objective functions for training UniVR. To be specific, the pretrained estimator E is first fine-tuned with the $\mathcal{L}_{est}$ defined as the L2 loss between the estimated disguising parameter and the ground truth:

$$\mathcal{L}_{est}(E) = ||\alpha - E(\boldsymbol{x}_s)||_2. \tag{6}$$

Then, the combination of adversarial loss $\mathcal{L}_{adv}$, spectrogarm reconstruction loss $\mathcal{L}_{spec}$, and feature matching loss $\mathcal{L}_{fm}$ is

applied to simultaneously constrain the estimator, generator, and discriminator. The $\mathcal{L}_{adv}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}(E,G,D) =& \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x_o})}[D_\phi^N(\boldsymbol{x})^2]+ \\ & \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{x_s})}[(D_\phi^N(G(\boldsymbol{z},E(\boldsymbol{z})))-1)^2].\end{aligned} \tag{7}$$

$\mathcal{L}_{spec}$ is measured by L2 distances between the spectrogram of original speech and its reconstructed counterpart, which can be formulated as follows:

$$\mathcal{L}_{spec}(E,G) = ||\boldsymbol{x_o} - G(\boldsymbol{x_s}, E(\boldsymbol{x_s}))||_2. \tag{8}$$

$\mathcal{L}_{fm}$ is computed by summing L1 distances between every discriminator feature map of real and generated samples in D_ϕ, which is defined as follows:

$$\mathcal{L}_{fm}(E,G,D) = \sum\nolimits_{i=1}^{N} ||D_\phi^i(\boldsymbol{x_o}) - D_\phi^i(G(\boldsymbol{x_s}, E(\boldsymbol{x_s})))||_1. \tag{9}$$

To further improve the naturalness of the restored speech, we also present a vocoder loss $\mathcal{L}_{NV}$, which is measured by the L2 loss between the restored speech and the ground truth:

$$\mathcal{L}_{NV}(E,G) = ||\boldsymbol{\omega_o} - V(G(\boldsymbol{x_s}, E(\boldsymbol{x_s})))||_2. \tag{10}$$

Notably, the vocoder loss only propagates to the estimator and generator to prevent overfitting of the neural vocoder. Finally, the total loss function of UniVR is formulated as follows:

$$\begin{aligned}\mathcal{L}_{total}(E,G,D) =& \mathcal{L}_{est}(E) + \mathcal{L}_{adv}(E,G,D) + \mathcal{L}_{spec}(E,G) \\ & + \lambda_{fm}\mathcal{L}_{fm}(E,G,D) + \lambda_{NV}\mathcal{L}_{NV}(E,G),\end{aligned} \tag{11}$$

where λ_{fm}, λ_{NV} are scalars set to 0.5, 0.1. E, G aim to minimize $\mathcal{L}_{total}$, while the objective of D is opposite to others.

IV. Experimential setup

A. Datasets

Several multi-speaker Mandarin datasets: CN-Celeb 1&2 [25], [26] and AISHELL-1&3 [27], [28], are used in our experiments, where the CN-Celeb is used for the SSL pre-train of the estimator and vocoder, the development part of AISHELL-1 is employed for the training of UniVR, and the test part of AISHELL-1&3 is used for evaluation. As shown in Table I, CN-Celeb consists of about 660,000 utterances from a total of 3000 speakers, while the train set of AISHELL-1 includes about 60,000+ utterances from a total of 340 speakers. The utterances in the train set are resampled to 16 kHz, and 3 extra samples are generated for each utterance using the pitch-scaling algorithm randomly selected from *Phase Vocoder* [29], *Frequency-domain Pitch Synchronous Overlap Add* (FD-PSOLA) [30], *Time-domain Pitch Synchronous Overlap Add* (TD-PSOLA) [31], and *Waveform Similarity Overlap Add* (WSOLA) [32]. The disguising parameter α follows a uniform distribution $\mathcal{U}(-8,8)$. A small subset of the synthetic data (about 2%) is reserved for hyperparameter optimization.

To evaluate different methods on more challenging and practical scenarios, the test sets of AISHELL-1&3 are disguised by not only the Phase Vocoder, FD-PSOLA, TD-PSOLA, and WSOLA, but also three commercial software programs, i.e., Audacity, Audition, and iZotope, which generate a total of 14 evaluation sets marked as A_1D_1–A_1D_7 and A_3D_1–A_3D_7, respectively. Each evaluation set is further divided into 33 subsets, where the applied disguising parameter α is increased from -8 to 8 with a step of 0.5.

TABLE I: Speech attribute distribution

Dataset	#Spks	#Utters	#α Range	#Duration
CN-Celeb1	1000	130109	-	7.57 sec
CN-Celeb2	2000	529485	-	7.41 sec
AISHELL-1 train	340	60209	$\mathcal{U}(-8,8)$	4.82 sec
AISHELL-1 test	60	651×33	$[-8,8,0.5]$	4.41 sec
AISHELL-3 test	214	750×33	$[-8,8,0.5]$	3.16 sec

B. Baseline systems

We assess the performance of UniVR in terms of both the accuracy of estimated disguising parameters and the quality of restored speech. To establish the baselines for estimation accuracy, we consider two advanced reference-free estimators [12], [13] and a reference-based estimator [14], which are introduced in Section II. The reference-free estimators are trained on the same dataset as UniVR, while the reference-based estimator uses a TDNN [33] trained on AISHELL-1 to extract speaker features. It takes a pair of speech signals $\{x, y\}$ as input, where x is the pitch-shifted speech and y is the original-pitch reference. As for restoration quality, we use the popular sound processing library, librosa [29], as the baseline to restore the pitch-shifted speech based on the estimated disguising parameter from the UniVR estimator.

Notably, pitch scaling aims to conceal the attacker's identity and deceive the SI system, which is distinct from voice conversion. Thus, the SI system can be degenerated into a speaker verification (SV) system when investigating the impact of pitch scaling on speaker characteristics. In this work, we adopt a classical SV system that employs ECAPA-TDNN [34] as the front-end to extract speaker features and applies cosine similarity as the back-end to calculate the trial scores. Subsequently, we use adaptive score normalization [35] to normalize the trial scores. The imposter cohort is constructed by averaging the embeddings from the same speaker in AISHELL-1.

C. Training strategy

The estimator is first pre-trained by MSPM on CN-Celeb 1&2 [25], [26] with 400 epochs and then fine-tuned together with the VRN on AISHELL-1 with 150 epochs. The batch size is 512. Both of them are trained with an exponentially decreased learning rate from the initial 10^{-3} to the final 10^{-5} with the Adam optimizer [36]. Furthermore, the augmentations are also applied during the fine-tuning, which follows the Kaldi recipe [37] in combination with the publicly available MUSAN dataset (music, speech, and noise) [38] and the RIR dataset (reverb) provided in [39]. The input of UniVR is an 80-dimensional Mel spectrogram calculated by a 25 ms window with a 10 ms frame shift. The experiments are implemented with an NVIDIA A100 80GB.

TABLE II: MAE results on A_1D_1–A_1D_7, where red is the best, and blue is the second.

Systems	A_1D_1	A_1D_2	A_1D_3	A_1D_4	A_1D_5	A_1D_6	A_1D_7
L. Wang [12]	0.937	1.225	1.109	1.055	1.563	1.821	1.767
Pilia [13]	0.751	0.896	0.875	0.920	1.313	1.489	1.375
UniVR	0.688	0.652	0.706	0.737	0.981	1.024	0.997
+ SSL pretrain	0.309	0.311	0.324	0.388	0.753	0.779	0.776
L. Zheng† [14]	0.675	0.684	0.619	0.607	0.731	0.786	0.774

†reference-based method.

TABLE III: Cross-dataset MAE results on A_3D_1-A_3D_7, where red is the best, and blue is the second.

Systems	A_3D_1	A_3D_2	A_3D_3	A_3D_4	A_3D_5	A_3D_6	A_3D_7
L. Wang [12]	1.064	1.372	1.301	1.158	1.635	1.911	1.807
Pilia [13]	0.884	1.033	0.996	1.015	1.439	1.561	1.419
UniVR	0.751	0.809	0.780	0.842	1.052	1.103	1.047
+ SSL pretrain	0.437	0.468	0.445	0.493	0.784	0.858	0.820
L. Zheng† [14]	0.855	0.846	0.825	0.805	0.912	0.942	0.978

†reference-based method.

V. Experimental Results

A. Accuracy of estimation

We employ the mean absolute error (MAE) measured between the estimated disguising parameter and the ground truth to evaluate the accuracy of different systems. The experimental results on the test sets of AISHELL-1 and AISHELL-3 are reported in Tables II and III, respectively. It is easy to find that UniVR achieved the highest accuracy and the best generalization ability for estimating disguising parameters in all cases compared with the existing reference-free estimators. In addition, SSL pre-training can significantly boost the performance of UniVR. Furthermore, even though the reference-free estimation is more challenging than the reference-based estimation, pretrained UniVR still obtains very competitive results compared to the reference-based baseline. Noteworthily, since the reference-based method essentially estimates the disguising parameters by solving Equ 1, its performance is almost not affected by the type of pitch-scaling algorithm. As a result, the reference-based baseline achieves better results than UniVR when facing unseen disguising methods, i.e., D_5–D_7. However, the gap between the reference-free method and the reference-based method will be narrowed in the cross-dataset evaluation, i.e., A_3D_1–A_3D_7. The reason is that the reference-free methods estimate the disguising parameters by recognizing the distortions introduced during pitch scaling, which are not actually linked to the speaker characteristics. Thus, they can be transferred well to unseen speakers. In the most complex case, i.e., cross-dataset evaluation with unseen disguising methods (A_3D_5–A_3D_7), the pretrained UniVR has achieved the best performance, giving it the most promising prospects for applications in real life.

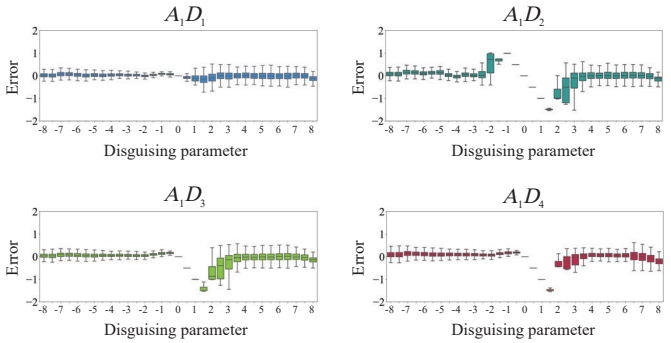


Fig. 4: Box-plot of the estimation deviation on the subsets of A_1D_1–A_1D_4 with various disguising parameters.

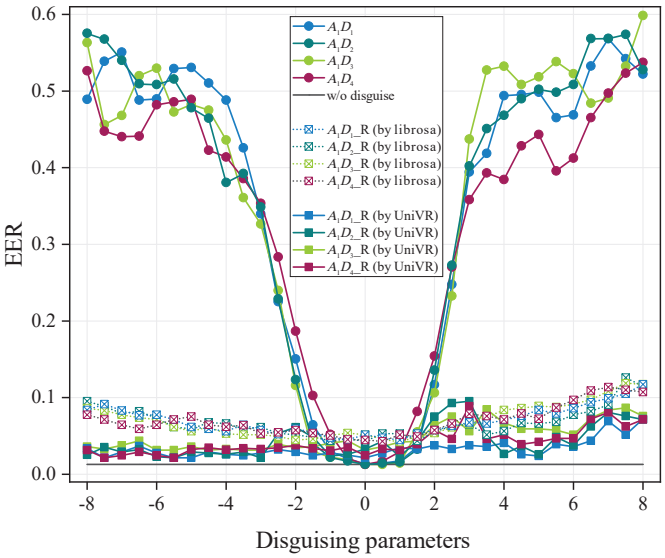
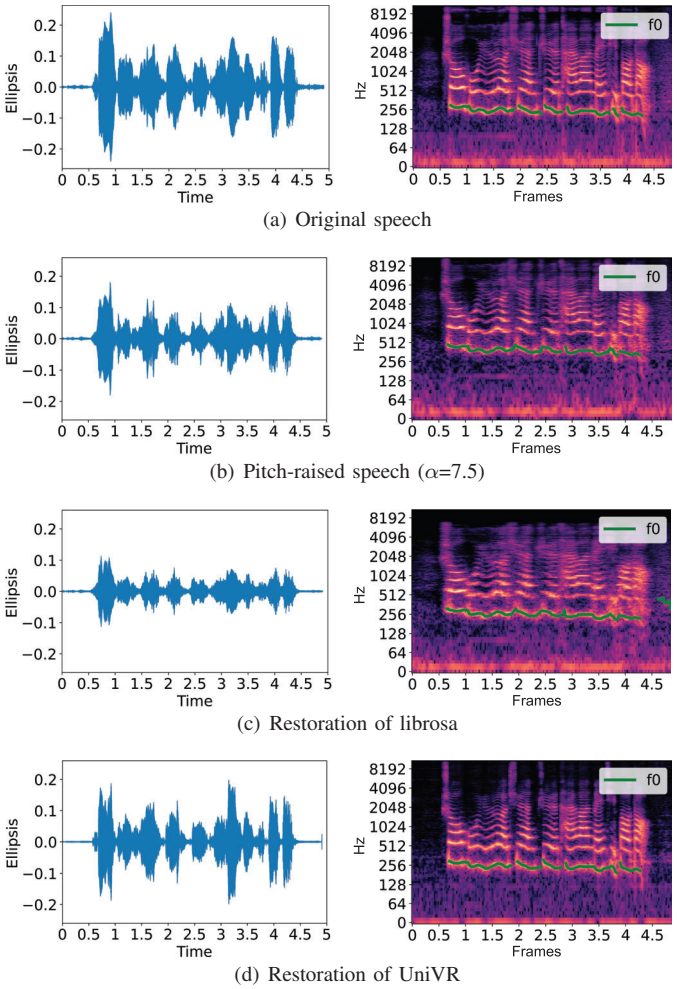


Fig. 5: EER of the SV system when faced with the pitch-shifted / restored samples from A_1D_1–A_1D_4.

We also further investigate the link between pitch-scaling degree and the performance of UniVR. Specifically, we measure the estimation deviation of UniVR on A_1D_1–A_1D_4 under different disguising parameters, which is shown in Fig. 4. The estimation of negative α is more accurate than that of positive α, which is consistent with the conclusion in previous studies [11], [12] that pitch-lowered speech is easier to distinguish than pitch-raised speech. Notably, the performance of the esitimator in UniVR shows the most serious decline when facing tiny disguising parameters within $[0, 3]$.

B. Quality of restoration

The impact of the pitch-scaling attack on the speaker verification system (SV) is shown in Fig. 5, which is measured by the Equal Error Rate (EER). Without restoration, pitch-shifted speech with the disguising parameter exceeding $[-2, 2]$ will obviously affect the performance of the SV system, making it unable to determine the speaker's identity. This problem can be effectively alleviated by restoring the pitch-shifted speech. In addition, although various pitch-shifting algorithms may cause



(a) Original speech

(b) Pitch-raised speech (α=7.5)

(c) Restoration of librosa

(d) Restoration of UniVR

Fig. 6: Waveforms and spectrograms of a pitch-shifted speech from A_1D_1 with the corresponding restorations.

different patterns of distortion, UniVR always provides higher restoration quality than the baseline, which is reflected by the lower EER of the SV system.

To demonstrate the superiority of UniVR in the acoustic quality of the restored speech, we also provide a visualization of a pitch-raised example with the corresponding restoration obtained via different methods. Notably, we only present the visualization of a pitch-raised sample from A_1D_1, as shown in Fig. 6, while the results in other cases exhibit the same trend. It is easy to find that although both librosa and UniVR can precisely restore the pitch of the speech, i.e., the f0 highlighted in green lines, librosa is unable to remove the artifacts or restore the missing high-frequency information in pitch-shifted speech and even distorts the speech more severely. In contrast, UniVR employs the specially designed VRN to remove the distortions in disguised speech by reconstructing its original mel spectrogram, yielding a more reasonable restoration. This issue can be reflected by the fact that both the mel spectrogram and the waveform of the restored speech of UniVR are closer to the ground truth than those of the librosa.

VI. Conclusions

In this study, we propose a unified framework (UniVR) for restoring pitch-shifted voice in high quality. UniVR consists of a reference-free estimator and a meticulously designed VRN that can effectively mitigate the distortion introduced during pitch scaling. Our experiments on AISHELL-1&3 demonstrate that the estimator in UniVR outperforms popular reference-free baselines, while also being extremely competitive with the reference-based baseline. Furthermore, the VRN has significantly better restoration quality than the popular method. As a result, UniVR achieves a state-of-the-art result for anti-pitch-scaling, which has significant implications for improving the security of speaker identification systems.

References

[1] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[2] Y.-H. Chin, J.-C. Wang, C.-L. Huang, K.-Y. Wang, and C.-H. Wu, "Speaker identification using discriminative features and sparse representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1979–1987, 2017.

[3] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114 591, 2021.

[4] Z. Wang and J. H. Hansen, "Multi-source domain adaptation for text-independent forensic speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 60–75, 2021.

[5] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When automatic voice disguise meets automatic speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 824–837, 2021. DOI: 10.1109/TIFS.2020.3023818.

[6] J. Laroche, "Time and pitch scale modification of audio signals," *Applications of digital signal processing to audio and acoustics*, pp. 279–309, 2002.

[7] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, IEEE, vol. 3, 1996, pp. 1800–1803.

[8] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy, "Towards combining pitch and mfcc for speaker identification systems," in *Eurospeech*, 2001, pp. 2825–2828.

[9] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "Build a sre challenge system: Lessons from voxsrc 2022 and cnsrc 2022," *arXiv preprint arXiv:2211.00815*, 2022.

[10] H. Wu, Y. Wang, and J. Huang, "Blind detection of electronic disguised voice," in *ICASSP*, IEEE, 2013, pp. 3013–3017.

[11] H. Liang, X. Lin, Q. Zhang, and X. Kang, "Recognition of spoofed voice using convolutional neural networks," in *GlobalSIP*, IEEE, 2017, pp. 293–297.

[12] L. Wang, H. Liang, X. Lin, and X. Kang, "Revealing the processing history of pitch-shifted voice using cnns," in *WIFS*, IEEE, 2018, pp. 1–7.

[13] M. Pilia, S. Mandelli, P. Bestagini, and S. Tubaro, "Time scaling detection and estimation in audio recordings," in *WIFS*, IEEE, 2021, pp. 1–6.

[14] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When automatic voice disguise meets automatic speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 824–837, 2020.

[15] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic+ noise model," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[16] S. N. Levine and J. O. Smith III, "A sines+ transients+ noise audio representation for data compression and time/pitch scale modifications," in *Audio Engineering Society Convention 105*, Audio Engineering Society, 1998.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[18] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[19] A. Baade, P. Peng, and D. Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," *arXiv preprint arXiv:2203.16691*, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[22] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *AAAI*, vol. 36, 2022, pp. 10 699–10 709.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[25] Y. Fan, J. Kang, L. Li, *et al.*, "Cn-celeb: A challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7604–7608.

[26] L. Li, R. Liu, J. Kang, *et al.*, "Cn-celeb: Multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[27] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 O-COCOSDA*, IEEE, 2017, pp. 1–5.

[28] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[29] B. McFee, C. Raffel, D. Liang, *et al.*, "Librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[30] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.

[31] *Praat: Doing phonetics by computer*, [Online]. Available: https://www.fon.hum.uva.nl/praat/, Accessed: Jul. 2022.

[32] *Soundtouch audio processing library*, [Online]. Available: http://www.surina.net/soundtouch/, Accessed: Jul. 2022.

[33] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[34] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[35] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications.," in *INTERSPEECH*, 2011, pp. 2365–2368.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP*, IEEE, 2019, pp. 5796–5800.

[38] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[39] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, IEEE, 2017, pp. 5220–5224.

LET AI FIND THE SIGNAL IN THE NOISE



THE MAX78000
AI MICROCONTROLLER

The world's first AI microcontroller designed for IoT and the intelligent edge.

A beautifully compact system on chip with convolutional neural network and ARM Cortex M4F processor able to receive A/V and analog signals for processing complex deep learning algorithms.

All possible using battery power.

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

YAMAHA
Make Waves



GLOBAL INTERNSHIP PROGRAM 2024
Research and Development Division
in Hamamatsu Japan

INTERNSHIP
Yamaha Corporation is recruiting a wide range of competitive R&D engineers for its internship programs. Interns will be working at the Innovation Center, which is Yamaha's dedicated R&D headquarters located in Hamamatsu Japan. The program will span approximately 12 weeks in 2024 summer season and will support your travel and accommodations in case of working at Yamaha. Interns will be members of teams which include other Yamaha researchers and engineers.

WHO WE ARE
With over 130 years of proud history, Yamaha Corporation is the largest and most dynamic manufacturer of acoustic and digital musical instruments and professional audio products in the world.

ABOUT RECRUITMENT
Doctoral or master's students in engineering, computer science, and/or mathematics.
Apply for our research internship
and enjoy your stay in Japan's city of music!

Please visit our website for more information on our Global Internship Program.
www.yamaha.com/en/recruitment/internship/






CyberAgent AI Lab

An R&D organization based in Japan that focuses on the utilization of a wide range of AI technologies in digital marketing

Following an R&D approach, we address a variety of challenging topics at the business level, while contributing to the academic community.

In the AI Lab, Research Scientists and Research Engineers are responsible of creating new knowledge through academic publications, and applying that knowledge to practical solutions, respectively.

Accepted
>50
papers in 2022

▼ Past accepted conferences

AAAI	ICLR
CVPR	ICML
EMNLP	IROS
HRI	NAACL
ICASSP	NeurIPS etc…

AI Implementation
>100
Projects

Our mission is contributing to academia and implementing our research results into society. We have already implemented more than 100 projects.

Audio Team

The Audio team is engaged in the research and development of speech synthesis/voice conversion methods to impersonate a specific human or to generate a variety of voices. In addition, speech recognition /enhancement methods is used for robust voice recognition regardless of the environment.

We aim to develop speech synthesis for CG models of celebrities, voices for advertisements, and other applications.
We are also engaged in research and development of speech recognition/synthesis for automatic telephone answering.


Speech synthesis for CG models


Automatic telephone answering

【Keywords】#Speech Synthesis #Lip Sync #Speech Recognition #Speech Enhancement #Source Separation #Auditory Analysis #Dialogue Systems #Voice Conversion

Recruitment Information

We are currently looking for talented Research Scientists and Research Engineers to join our team.
We will also announce internship opportunities for Ph.D. students in the upcoming winter term.



Entry

＜Experienced professionals＞



Entry

＜Internship for Ph.D. students＞

Google University Relations Programs
(research.google/outreach)

To cultivate collaborations with the academics, Google administers a variety of programs that provide resources and support to the academic and external research communities.

Opportunities for Faculty

/ Google's Research Grant
Not for open calls. Please reach out to Google contact to apply for.

/ Award for Inclusion Research
[research.google/outreach/air-program]
Supports research addressing the needs of populations historically underrepresented in the tech industry (e.g. women, people with disabilities etc.). Applications open in Summer.

/ Research Scholar Program
[research.google/outreach/research-scholar-program/]
Supports research conducted by early-career faculty globally. Applications open in Fall.

/ Google Cloud Platform (GCP)
[edu.google.com/programs/credits/research/]
Provide researchers and PhD students with computing power, storage and AI tools to develop bold ideas and make cutting-edge discoveries. Applications open any time.

/ Academic Conference Sponsorship
Sponsorship to both international and domestic academic conferences on areas relevant to Google research priorities. Not for open calls. Please reach out to Googler contact to apply for.

/ exploreCSR (CSR diversity)
[research.google/outreach/explore-csr/]
Awards that support universities to host computer science research (CSR) workshops in order to increase the number of students from groups traditionally underrepresented in CS. Applications open in Spring.

/ Visiting Researcher Program
[research.google/outreach/visiting-researcher-program/]
Academic researchers work with research and engineering teams at Google for 3-12 months.

/ Research Exchange
Research talks by people from both sides to foster research interactions and sharing.

Opportunities for Student

/ PhD Fellowship Program
[google.com/intl/en_cn/university/research/phdfellowship/]
Fellowships recognize outstanding PhD students doing exceptional and innovative research. Applications open in March.

/ Student Travel Grants
[buildyourfuture.withgoogle.com/scholarships/google-conference-scholarships]
Support outstanding research students to attend leading conferences. Applications open any time (APAC tab, East Asia form).

/ Student Researcher
[email to eastasia-ur@google.com]
Research internship to work with Google research / engineering teams at Google. Flexible internship period. BS, MS or PhD students are all eligible.

/ Visit to Google Office
Provides opportunities for students to visit Google offices and meet Googlers from different functions.

For any questions, please reach out to eastasia-ur@google.com

Google

Fairy Devices

We develop “magical” devices
that make people feel as if
a fairy is inside.



rd@fairydevices.jp

@fairyresearch

